https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

Topology-Preserving State Space Representation for Diagnosing Weather Forecast Models

Hyoungnyoun Kim[1] and Jeong Hoon Cho[1]

[1]National Institute of Meteorological Sciences, Jeju, Republic of Korea

Correspondence: Hyoungnyoun Kim (nyoun2@korea.kr)

Abstract. The evaluation of AI-based medium-range weather forecast models commonly relies on pointwise error metrics such as the root-mean-square error (RMSE), which provide useful quantitative summaries but do not directly show how multivariate forecast states evolve in relation to the observed atmospheric state. This study proposes a trajectory-based diagnostic framework built on a topology-preserving state space for visually and quantitatively comparing forecast behavior across different AI weather forecast models. Multivariate atmospheric states are represented using 850 hPa temperature, geopotential height, zonal wind, and meridional wind fields, and a shared state space is constructed through image-based feature extraction, contrastive learning, and manifold embedding. The framework is applied to forecasts from FourCastNet2, GraphCast, and Pangu-Weather under different initial-condition settings, with forecast sequences represented as trajectories in the learned state space. The results show that the proposed representation provides an intuitive way to compare model-dependent forecast evolution, identify regions of relatively good or poor trajectory behavior, and examine state- or season-dependent differences that are not readily summarized by variable-wise RMSE curves alone. The proposed framework is therefore intended as a complementary diagnostic tool for interpreting AI-based weather forecast models, rather than as a replacement for conventional verification metrics.

1 Introduction

The performance of weather forecast models is commonly evaluated using pointwise error metrics calculated for selected variables, pressure levels, regions, and forecast lead times. Metrics such as the root-mean-square error (RMSE), mean-square error (MSE), and mean absolute error (MAE) are widely used because they are simple to compute, easy to compare across models, and directly interpretable as error magnitudes in the original physical variables. These metrics remain essential for both numerical weather prediction and AI-based weather forecast models. However, they provide only a limited view of forecast behavior when the objective is to understand how a model evolves as a multivariate atmospheric system.

A medium-range weather forecast is not only a collection of independent fields at separate lead times. It is also a sequence of atmospheric states evolving from an initial condition. Each state reflects the combined structure of multiple variables, including temperature, geopotential height, and wind fields. When forecast performance is examined only through separate RMSE curves, it can be difficult to obtain an integrated view of how the predicted atmospheric state moves through time, how different models produce distinct forecast paths, or whether model performance depends on the background atmospheric state.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

This limitation is particularly relevant for AI-based forecast models, because different architectures and training strategies can produce forecasts with different structural behaviors even when their variable-wise error magnitudes appear similar.

Recent AI-based medium-range forecast models, including FourCastNet, GraphCast, Pangu-Weather, AIFS, and GenCast, have demonstrated substantial progress in data-driven weather prediction (Pathak et al., 2022; Lam et al., 2023; Bi et al., 2023; Lang et al., 2024; Price et al., 2025). These models differ in their network architectures, spatial representations, training datasets, and forecast-generation procedures. As a result, their relative performance may vary depending on the variable, region, season, initial condition, and forecast lead time. Conventional RMSE-based evaluation can summarize these differences quantitatively, but it does not directly provide a common visual or structural space in which the forecast evolution of different models can be compared. A complementary diagnostic representation is therefore useful for examining model behavior beyond separate variable-wise error curves.

Forecast verification has long recognized that forecast quality cannot be fully characterized by a single pointwise error score, particularly when spatial displacement, temporal mismatch, object structure, or multivariate dependence is important (Casati et al., 2008; Rife and Davis, 2005; Bjerregård et al., 2021). Spatial and neighborhood-based verification methods have been developed to evaluate forecast fields in ways that are less dependent on exact grid-point matching, particularly when small spatial displacements can strongly affect conventional scores (Ebert, 2008; Gilleland et al., 2009; Wernli et al., 2009; Gilleland et al., 2010). Object-based verification methods provide another structural perspective by identifying coherent meteorological features and comparing their properties rather than only grid-point values (Davis et al., 2006). Multivariate verification studies further emphasize that forecast quality may depend not only on individual variables, but also on the dependence structure among forecast components (Scheuerer and Hamill, 2015; Bjerregård et al., 2021). In parallel, representation-learning approaches have shown that learned embeddings can provide informative reduced representations of high-dimensional atmospheric states (Hoffmann and Lessig, 2023). These studies indicate that forecast evaluation can benefit from structural, spatial, multivariate, and representation-based perspectives. However, there remains a need for a practical framework that represents multivariate atmospheric states in a shared space and uses that space to visually compare the forecast evolution of different AI models.

This study addresses this need by constructing a topology-preserving state space for trajectory-based diagnosis of AI-based medium-range weather forecast models. In the proposed framework, each multivariate atmospheric state is represented as a point in a learned state space, and a forecast sequence is interpreted as a trajectory formed by connecting these points over forecast lead time. The goal is not to replace RMSE-based evaluation, but to provide an additional view in which forecast evolution can be inspected visually and compared structurally across models and initial-condition settings. Figure 1 provides a schematic overview of the framework. The modeling branch constructs the topology-preserving state space from ERA5 reanalysis data using multivariate image representation, physically motivated positive relationships, feature encoding, and manifold embedding. The evaluation branch projects IFS analysis and AI forecast outputs into the same state space, where model behavior is compared through forecast trajectories.

The key methodological requirement is that the learned state space should preserve meaningful relationships among atmospheric states. Physically similar states should be located close to one another, small perturbations such as noise or spatial shifts

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

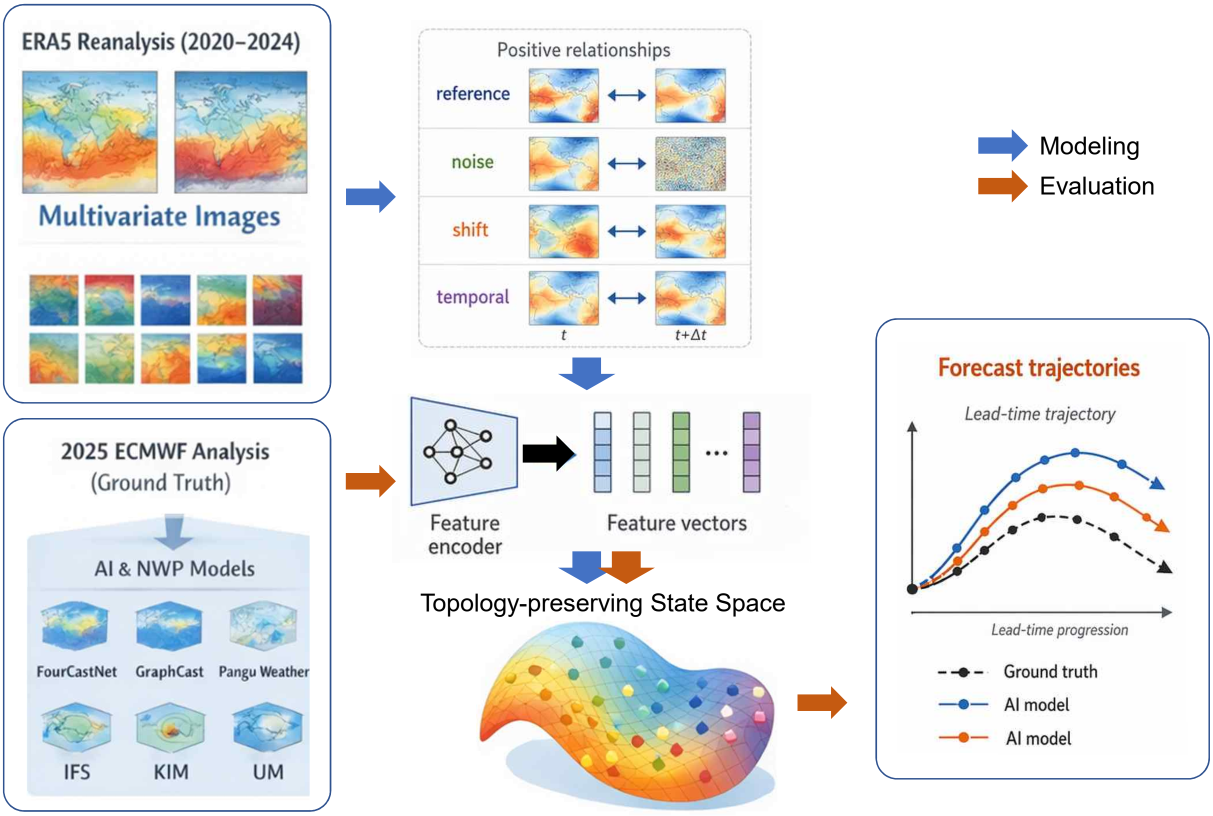

Figure 1. Schematic overview of the proposed framework for topology-preserving state-space construction and trajectory-based forecast evaluation. The modeling branch constructs a shared state space from ERA5 reanalysis data using multivariate image representation, physically motivated positive relationships, feature encoding, and manifold embedding. The evaluation branch projects IFS analysis and AI/NWP forecast outputs into the same state space and compares forecast evolution as trajectories.

should not dominate the representation, and temporally adjacent states should form reasonably continuous paths. To support these properties, the proposed framework combines image-based feature extraction, contrastive learning based on physically motivated positive relationships, and manifold embedding. The resulting topology-preserving state space is used as a common diagnostic space for comparing ground truth and forecast trajectories.

The evaluation is conducted using ERA5 reanalysis data from 2020 to 2024 to construct the state space and 2025 IFS analysis as the ground truth reference for forecast evaluation. Three AI-based medium-range forecast models, FourCastNet2, GraphCast, and Pangu-Weather, are evaluated under three initial-condition settings, IFS, KIM, and UM, resulting in nine forecast configurations. Forecast trajectories are analyzed in two ways: lead-time trajectory analysis, which visualizes the forward evolution of forecasts from a fixed initial condition, and valid-time trajectory analysis, which examines how forecasts initialized at different earlier times approach a common target verification time.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

The main contributions of this study are as follows. First, it constructs a topology-preserving state space that represents multivariate atmospheric states as comparable points in a common diagnostic space. Second, it shows how forecast sequences from different AI-based weather forecast models can be represented and compared as trajectories in that space. Third, it demonstrates that this trajectory-based view provides intuitive and visually interpretable information about model-dependent, state-dependent, and seasonally varying forecast behavior that complements conventional RMSE-based evaluation.

2 Method

The construction of the feature-based state space consists of four main steps. First, multivariate meteorological fields are converted into image-based inputs and encoded as high-dimensional feature vectors. Second, contrastive learning is applied to reorganize the feature representation according to physically motivated relationships among atmospheric states. Third, a parametric manifold embedding is trained to construct a two-dimensional state space that preserves local relationships among high-dimensional features and the smoothness of temporal evolution. Finally, the resulting state space is evaluated from several perspectives, including seasonal organization, temporal continuity, positive-pair proximity, and distortion between the learned state space and the original feature space. This pipeline is designed not as a generic dimensionality-reduction or visualization procedure, but as a way to construct a shared diagnostic space in which structural similarity and temporal evolution of atmospheric states can be examined consistently.

In this study, the term “feature space” refers to the high-dimensional representation obtained after image-based feature extraction and contrastive learning, whereas the term “state space” refers to the final two-dimensional topology-preserving diagnostic space obtained through manifold embedding. Forecast trajectories are analyzed only in this final state space.

2.1 Data definition and feature extraction

ERA5 reanalysis data were used to construct the topology-preserving state space. ERA5 provides a temporally consistent representation of atmospheric states and has been widely used for the development and evaluation of AI-based weather forecast models (Hersbach et al., 2020). In this study, ERA5 fields were extracted at 6-h intervals from January 2020 to December 2024, yielding 7308 atmospheric-state samples. The spatial domain was restricted to a region centered on East Asia. Each atmospheric field was converted into an image-like array of approximately 180×180 grid points so that image-based feature extraction could be applied while retaining the major spatial structures of the regional atmospheric state. Although the selected region represents only a small fraction of the global domain, it provides a test domain in which seasonal variability, temporal continuity, and medium-range forecast evolution can be examined.

Four variables at the 850 hPa pressure level were used: temperature, geopotential height, zonal wind, and meridional wind. These variables describe complementary aspects of the lower- to mid-tropospheric state, including thermal structure, large-scale circulation, and horizontal flow. They are also commonly available in ERA5, IFS analysis, and AI forecast outputs, allowing the same representation and evaluation procedure to be applied consistently across the training and forecast-evaluation datasets.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

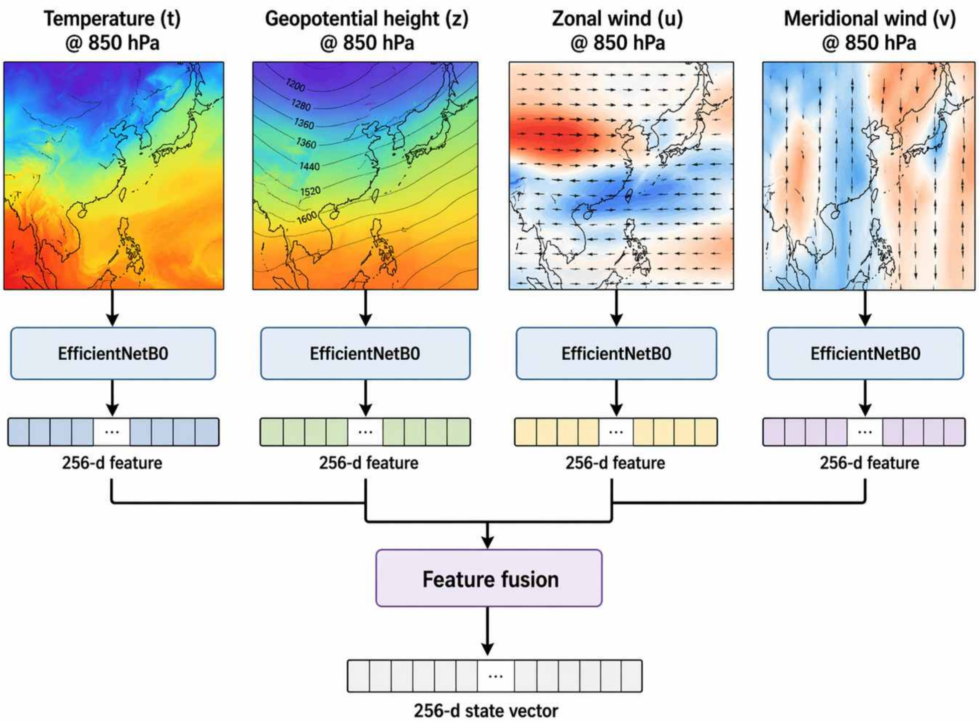

Figure 2. Example multivariate weather fields at 850 hPa used as inputs for feature extraction, together with the image-based feature extraction pipeline and feature-fusion process used to construct a 256-dimensional state vector for each atmospheric state.

For each time step, the four variables were treated as separate image-like fields and then fused into a single representation of the multivariate atmospheric state. Figure 2 shows an illustration of the input fields and the feature-extraction pipeline. Each variable field was passed through a pre-trained EfficientNetB0 encoder (Tan and Le, 2019), from which a 256-dimensional feature vector was extracted. EfficientNetB0 was used to provide a compact representation of spatial patterns in each variable field. This step is intended to encode spatial structures in a form that can subsequently be reorganized by contrastive learning, rather than to define the final diagnostic space directly.

The four variable-specific feature vectors were then combined through feature fusion to obtain a single 256-dimensional feature vector for each atmospheric state. This fusion step avoids directly averaging variables with different physical units and value ranges. Instead, each variable is first represented in a common feature form, and the variable-specific representations are then combined into one feature vector. As a result, each point in the subsequent state space represents the multivariate atmospheric state at a specific time and over the selected region.

In this framework, feature extraction serves as the representation basis for the following contrastive-learning and manifold-embedding steps. By transforming different physical variables into a unified feature vector, the procedure provides an initial state space in which distances can be further adjusted to reflect meteorologically meaningful similarity relationships.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

2.2 Topology-preserving contrastive learning

The purpose of the contrastive-learning step is to reorganize the feature representation so that related atmospheric states are located close to one another, while unrelated states are separated. Pixel-level similarity in meteorological fields does not necessarily correspond to physical similarity. For example, a small spatial displacement of a coherent weather pattern may produce a large pixel-wise difference, even though the two states remain structurally similar. Therefore, feature extraction alone does not guarantee that the learned space has a meteorologically meaningful topology.

To address this issue, four types of positive relationships were defined: reference consistency, noise robustness, shift invariance, and temporal continuity. These relationships were incorporated into a supervised contrastive-learning framework (Khosla et al., 2020).

The first relationship is reference consistency. ERA5 and the corresponding IFS analysis at the same time represent closely related atmospheric states, although they may differ in their spectral characteristics and high-frequency components. ERA5 is suitable for constructing a consistent training dataset, whereas IFS analysis is used as the reference state for forecast evaluation in this study. Therefore, each ERA5 sample was used as an anchor, and the corresponding IFS analysis was treated as a positive sample. This constraint encourages the representation to place the two data sources close to one another in the learned state space, which is important because IFS analysis is later used as the ground truth reference for forecast evaluation.

The second relationship is noise robustness. ERA5 and forecast fields may differ in terms of noise level, effective resolution, and smoothing characteristics. If the representation is overly sensitive to such appearance-level differences, trajectory-based diagnosis may become unstable. To reduce this sensitivity, noise-perturbed and blurred versions of the ERA5 fields were generated and treated as positive samples with respect to the original fields. This encourages the state space to preserve the proximity of states that are meteorologically similar despite small appearance-level perturbations.

The third relationship is shift invariance. Atmospheric patterns may remain structurally similar even when they are displaced by a small number of grid points. Such displacement can lead to a large difference in raw pixel space and may also affect CNN-based features. Because part of the forecast error in medium-range prediction is associated with phase or location errors, a small spatial displacement should not necessarily be interpreted as a complete failure to reproduce the atmospheric state. In this study, translated samples were generated by shifting the image around its center in the horizontal and vertical directions. These translated samples were paired with the original samples as positives, thereby encouraging the state space to be more robust to small phase shifts and positional errors.

The fourth relationship is temporal continuity. Atmospheric states evolve continuously in time except during rapid transitions, and temporally adjacent states should therefore form smooth paths in the learned space. To reflect this property, the base sample at time t was used as an anchor, and samples at $t+1$ and $t+2$ were used as temporal positive samples. Samples that were far apart in time were treated as negative samples. This constraint is intended to support the formation of continuous trajectories in the learned state space.

The supervised contrastive-learning loss was formulated to integrate these positive relationships. Training was performed using mini-batches, where each mini-batch consisted of B anchor states. An anchor state corresponds to the atmospheric state

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

at a specific time, and multiple views are generated from the same anchor. These views include the base state, the reference state, appearance-perturbed states, phase-shifted states, and temporally adjacent states.

All views in a mini-batch were treated as individual samples. Let M denote the total number of samples in the mini-batch. Views generated from the same anchor state share the same label. For each sample i, the positive set $P(i)$ consists of the other views associated with the same anchor state, and the comparison set $A(i)$ consists of all samples in the mini-batch except i. The supervised contrastive loss is defined as

$$L_{\text{sup}} = \frac{\sum_{i=1}^{M} w_i L_i}{\sum_i w_i}, \tag{1}$$

$$L_i = -\frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\text{sim}(z_i, z_p)/\tau)}{\sum_{a \in \mathcal{A}(i)} \exp(\text{sim}(z_i, z_a)/\tau)}. \tag{2}$$

Here, $sim(z_i, z_j)$ denotes cosine similarity, z_i,z_j are the L2-normalized feature embeddings of sample i and j obtained from the projection head, respectively. The parameter τ is the temperature. The weight w_i is assigned according to the view type and controls the relative contributions of the base, reference, appearance, phase, and temporal views.

In addition to the supervised contrastive loss, explicit alignment terms were used to reinforce reference consistency and temporal continuity:

$$L_{\text{ref}} = \frac{1}{B} \sum_{i=1}^{B} \left\| z_i^{\text{base}} - z_i^{\text{ref}} \right\|_2^2 \tag{3}$$

$$L_{\text{temp}} = \frac{1}{B} \sum_{i=1}^{B} \left\| z_i^{\text{base}} - z_{i+\Delta}^{\text{temp}} \right\|_2^2, \quad \Delta \in \{1, 2\} \tag{4}$$

The total loss is given by

$$L_{\text{total}} = \lambda_{\text{sup}} L_{\text{sup}} + \lambda_{\text{ref}} L_{\text{ref}} + \lambda_{\text{temp}} L_{\text{temp}}. \tag{5}$$

The coefficients λ_{sup}, λ_{ref}, and λ_{temp} control the relative contributions of the supervised contrastive, reference-alignment, and temporal-smoothness terms, respectively. The contrastive-learning procedure was implemented using a three-stage training policy consisting of warm-up, main, and fine-tuning stages (Table 1). In the warm-up stage, temporal positives and reference positives were not used; the objective was to stabilize the initial feature topology using augmentation-based positive relationships, mainly appearance, phase, and auxiliary scale perturbations. In the main stage, reference positives and temporal positives were introduced, with the temporal neighborhood size set to $k = 2$. This stage jointly optimized supervised contrastive learning, reference alignment, and temporal smoothness. In the fine-tuning stage, augmentation-based positive families were removed, while the temporal-smoothness weight was increased. This schedule was designed to first establish a stable feature topology and then progressively impose reference consistency and temporal continuity. Through this process, the initial 256-dimensional fused features were reorganized into a 128-dimensional feature representation that reflects reference consistency, noise robustness, shift invariance, and temporal continuity.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

Table 1. Stage-wise hyperparameters used for topology-preserving contrastive learning. The training procedure consists of warm-up, main, and fine-tuning stages. The warm-up stage stabilizes the initial representation without temporal positives or reference positives. The main stage introduces reference consistency and temporal continuity, while the fine-tuning stage suppresses augmentation-based positives and places greater emphasis on temporal smoothness.

Stage	**Warm-up**	**Main**	**Fine-tuning**
Steps	200	400	200
Learning rate	1.00×10^{-3}	1.00×10^{-3}	3.00×10^{-4}
Temperature	0.1	0.1	0.1
Temporal positives	No	Yes, $(k = 2)$	Yes, $(k = 2)$
Reference as positive	No	Yes	Yes
Family weights	appearance: 1.0 phase: 0.6 scale: 0.3	appearance: 0.2 phase: 0.6 scale: 0.2	appearance: 0.0 phase: 0.0 scale: 0.0
Loss weights	$\lambda_{\text{sup}} = 1.0$ $\lambda_{\text{ref}} = 0.0$ $\lambda_{\text{temp}} = 0.0$	$\lambda_{\text{sup}} = 1.0$ $\lambda_{\text{ref}} = 0.5$ $\lambda_{\text{temp}} = 0.4$	$\lambda_{\text{sup}} = 0.0$ $\lambda_{\text{ref}} = 0.3$ $\lambda_{\text{temp}} = 0.4$

2.3 Parametric manifold embedding

The 128-dimensional features obtained from contrastive learning reflect structural similarity among atmospheric states, but the trajectories are difficult to inspect directly in the high-dimensional space. Therefore, a two-dimensional state space is required to support visual diagnosis and trajectory-based comparison. This step is related to previous work on neighborhood-preserving and topology-aware low-dimensional representations, including UMAP, parametric UMAP, and topological autoencoders (McInnes et al., 2018; Sainburg et al., 2021; Moor et al., 2020). These methods provide useful foundations for representing high-dimensional data in a lower-dimensional space while retaining local or topological relationships.

In this study, however, the embedding is designed specifically for forecast diagnosis rather than for general-purpose visualization alone. The objective is not only to preserve local neighborhood relationships among atmospheric states, but also to obtain a state space in which temporally adjacent states can form interpretable trajectories. Accordingly, a parametric manifold embedding $f_\theta : \mathbb{R}^{128} \rightarrow \mathbb{R}^2$ was trained to map each feature vector $x_i \in \mathbb{R}^{128}$ to a two-dimensional coordinate $z_i \in \mathbb{R}^2$. This step has two objectives. The first is to preserve the local neighborhood structure of the high-dimensional feature space in the two-dimensional state space. The second is to avoid excessive discontinuities in trajectories formed by temporally adjacent states.

A k-nearest neighbor graph was constructed in the 128-dimensional feature space. Positive pairs (i, j) were defined from neighboring samples, and negative pairs (i, k) were defined from non-neighboring samples. The high-dimensional similarity

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

p_{ij} and the two-dimensional similarity q_{ij} were defined as

$$p_{ij} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{\sigma_i \sigma_j + \epsilon}\right), \quad q_{ij} = \frac{1}{1 + \|z_i - z_j\|_2^2} \tag{6}$$

Here, σ_i is the local distance scale of sample i, defined as the distance to its k-th nearest neighbor in the 128-dimensional feature space, and ϵ is a small constant for numerical stability. Based on these similarities, the distance-preserving loss was defined as

$$L_{\text{dist}} = -\mathbb{E}_{(i,j)\in E^+}\left[p_{ij} \log q_{ij}\right] - \mathbb{E}_{(i,k)\in E^-}\left[\log(1 - q_{ik})\right] \tag{7}$$

The first term encourages neighboring samples in the high-dimensional feature space to remain close in the two-dimensional state space. The second term discourages negative pairs from being placed too close to one another. The notation $\mathbb{E}_{(i,j)\in S}[.]$ used in this section denotes the average over the pair set S, and is defined as follows. Specifically, $\mathbb{E}_{(i,j)\in E^+}[.]$ denotes the average over the positive pair set E^+, $\mathbb{E}_{(i,k)\in E^-}[.]$ denotes the average over the negative pair set E^-, and $\mathbb{E}_{(i,j)\in \tau}[.]$ denotes the average over the temporally adjacent pair set τ.

Preserving neighborhood relationships too strongly may cause excessive compression of positive pairs in the two-dimensional space. To reduce this crowding effect, a stress term was introduced to regularize the local distance ratio between the original feature space and the embedding space:

$$L_{\text{stress}} = \mathbb{E}_{(i,j)\in E^+}\left[\log\left(\|z_i - z_j\|_2 + \delta\right) - \log\left(\frac{\|x_i - x_j\|_2}{\sigma_i + \epsilon} + \delta\right)\right]^2 \tag{8}$$

Here, δ and ϵ are constants for numerical stability. This term acts as a regularization component that prevents local distance relationships in the 128-dimensional feature space from collapsing excessively in the two-dimensional state space.

Temporal smoothness was also included because the subsequent evaluation relies on the interpretation of trajectories. However, enforcing all temporally adjacent states to be close may distort genuine transitions between atmospheric regimes. Therefore, the temporal-smoothness term was gated by the high-dimensional similarity p_{ij}, so that the constraint is strong only when temporally adjacent states are also similar in the original feature space:

$$L_{\text{temp}} = \mathbb{E}_{(i,j)\in \tau}\left[p_{ij} \log\left(1 + \|z_i - z_j\|_2^2\right)\right] \tag{9}$$

When p_{ij} is large, the two-dimensional distance between adjacent states is penalized more strongly, resulting in smoother trajectories. When the original feature space distance is large, p_{ij} becomes small and the temporal constraint is weakened. This design reduces unnecessary oscillations within similar regimes while allowing larger displacements during genuine structural transitions.

The final objective for the parametric manifold embedding is

$$L_{\text{total}} = L_{\text{dist}} + \lambda_{\text{stress}} L_{\text{stress}} + \lambda_{\text{temp}} L_{\text{temp}} \tag{10}$$

The hyperparameters λ_{stress} and λ_{temp} control the relative contributions of distance-ratio regularization and temporal smoothness. In this study, L_{dist} was treated as the main objective because preserving the high-dimensional feature topology was the primary goal, while the stress and temporal terms were used as auxiliary regularization terms.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

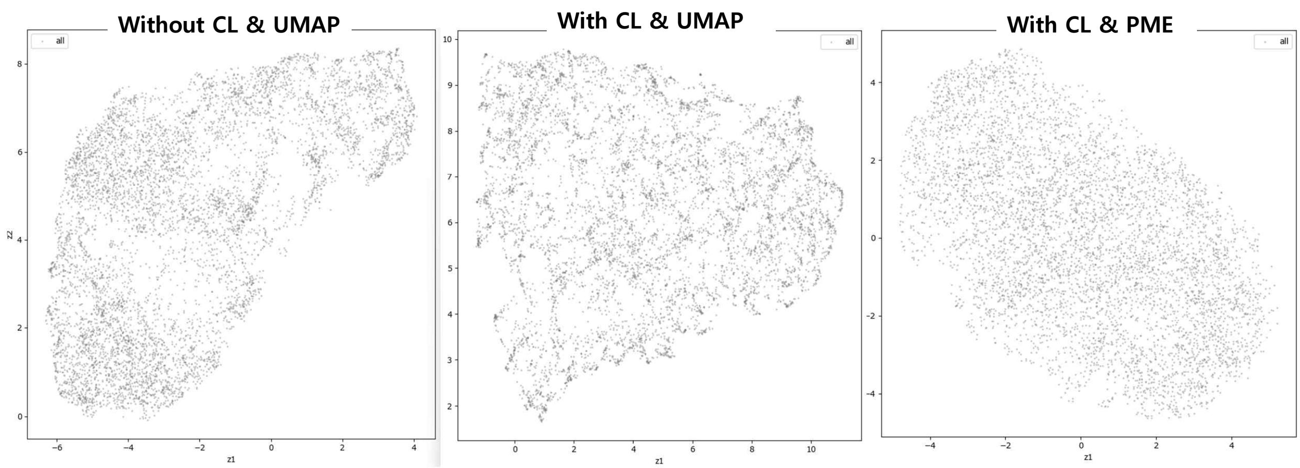

Figure 3. Comparison of the three state spaces considered in this study: *WO_CL_UMAP*, *W_CL_UMAP*, and *W_CL_PME*.

The parametric manifold embedding was trained with the same configuration for all experiments. The embedding function mapped the 128-dimensional contrastive features to a two-dimensional state space using a multilayer perceptron with hidden dimensions of 256, 256, and 128 and a dropout rate of 0.1. A k-nearest neighbor graph was constructed using the Euclidean distance with k=30, and 15 negative samples were drawn for each positive edge. Temporal smoothness was imposed only between adjacent 6-h states. The stress ($\lambda_{\mathrm{stress}}$) and temporal-smoothness (λ_{temp}) weights were set to 0.1 and 0.02, respectively, so that neighborhood preservation remained the dominant constraint while temporal continuity was used as a weak regularization term.

To compare different state-space construction procedures, three conditions were defined. The first condition, *WO_CL_UMAP*, applies conventional UMAP without contrastive learning (McInnes et al., 2018). The second condition, *W_CL_UMAP*, applies conventional UMAP after contrastive learning. The third condition, *W_CL_PME*, applies the proposed parametric manifold embedding after contrastive learning, following the general motivation of parametric neighborhood-preserving embeddings (Sainburg et al., 2021). Figure 3 compares the state spaces obtained under these three conditions. This comparison was used to assess the effect of contrastive learning on feature topology and the extent to which parametric manifold embedding further improves distance consistency and trajectory smoothness.

2.4 Validation of the state space

Before applying the constructed state space to forecast-model evaluation, it is necessary to examine whether the space satisfies the intended design properties. This is particularly important because nonlinear dimensionality reduction can distort neighborhood and distance relationships, and the quality of a low-dimensional representation should therefore be assessed using explicit diagnostic criteria rather than visual inspection alone (Lee and Verleysen, 2009; Venna et al., 2010). In this study, the state space was evaluated from four perspectives: seasonal organization, temporal continuity, positive-pair proximity, and distortion analy-

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

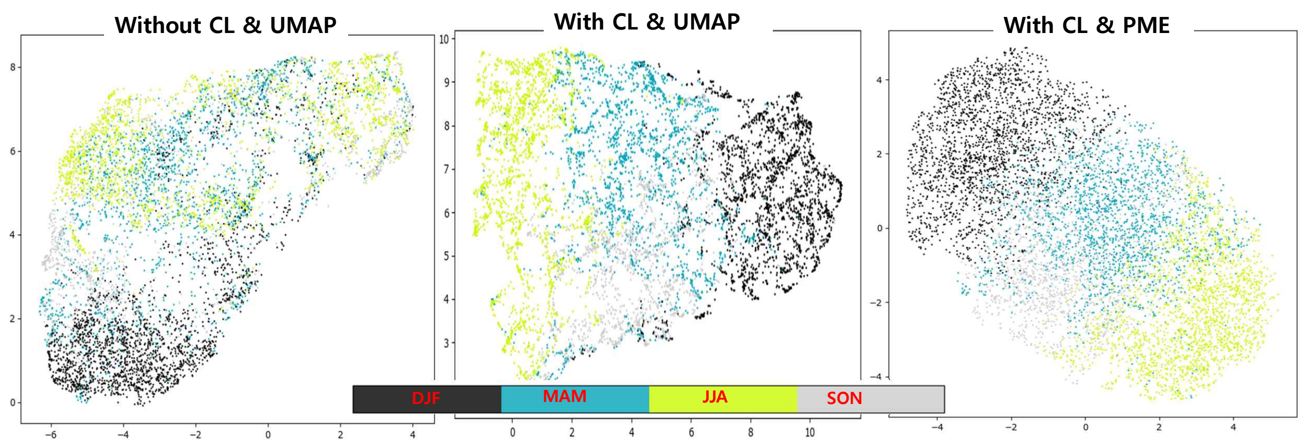

Figure 4. Seasonal distribution of ERA5 atmospheric states in the three state spaces. The panels show the same samples colored by season in *WO_CL_UMAP*, *W_CL_UMAP*, and *W_CL_PME*.

sis. These evaluations were conducted for the three state spaces, *WO_CL_UMAP*, *W_CL_UMAP*, and *W_CL_PME*, to examine the effects of contrastive learning and parametric manifold embedding.

The first validation examines seasonal organization. ERA5 samples from 2020 to 2024 were labelled by season and projected into each state space, as shown in Figure 4. In *WO_CL_UMAP*, the seasonal labels were broadly mixed, and temporal organization was not clearly expressed. In *W_CL_UMAP*, seasonal coherence improved to some extent after contrastive learning, but discontinuities remained in several regions of the space. In *W_CL_PME*, winter and summer states were more clearly separated, while spring and autumn appeared as transitional regions between them. Although seasonal classification was not the direct training target, this structure provides an indirect indication that the state space reflects both temporal progression and structural similarity among atmospheric states.

To quantify this seasonal organization, the mean within-class distance and between-class distance were computed using the seasonal labels. To reduce the influence of scale differences among embedding spaces, the two-dimensional coordinates of each space were standardized using axis-wise z-scores before distance calculation. The within-class distance, between-class distance, and between/within distance ratio were 1.4287, 1.8676, and 1.3072 for *WO_CL_UMAP*; 1.2316, 1.9962, and 1.6207 for *W_CL_UMAP*; and 1.0377, 2.0266, and 1.9530 for *W_CL_PME*, respectively. *W_CL_PME* showed the smallest within-class distance, the largest between-class distance, and the highest between/within ratio, indicating that same-season samples were more compact while different-season samples were more clearly separated. This result suggests that contrastive learning contributes to seasonal coherence and that parametric manifold embedding further organizes the learned representation into a temporally continuous state-space structure.

The second validation examines temporal continuity. Because atmospheric states usually evolve gradually over 6-h intervals, temporally adjacent samples should form continuous paths in the state space. To evaluate this property, temporally consecutive

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

Table 2. Summary statistics of trail path lengths in the three feature spaces for trail lengths of 8, 12, and 16, including the mean, median, standard deviation, minimum, and maximum.

State Space	Trail Length	Mean	Median	Std	Min	Max
WO_CL_UMAP	8	3.450	3.271	2.317	0.087	16.703
	12	5.423	5.293	3.064	0.235	18.818
	16	7.398	7.325	3.699	0.496	20.643
W_CL_UMAP	8	**1.061**	**0.755**	0.993	0.035	9.180
	12	**1.667**	**1.331**	1.292	0.069	12.066
	16	**2.273**	**1.937**	1.555	0.258	12.388
W_CL_PME	8	1.205	1.098	**0.534**	0.190	**5.397**
	12	1.893	1.777	**0.703**	0.441	**5.965**
	16	2.582	2.469	**0.842**	0.742	**6.891**

ERA5 samples were connected to form trails, and the path-length distributions of these trails were compared across the three spaces.

The total path length was computed for trails consisting of 8, 12, and 16 consecutive time steps. As in the seasonal analysis, the coordinates in each state space were standardized using axis-wise z-scores before comparison. Table 2 summarizes the resulting path-length statistics. For all three trail lengths, *W_CL_PME* showed the smallest standard deviation and the lowest maximum path length. The standard deviations were 0.5335, 0.7026, and 0.8418 for trail lengths of 8, 12, and 16, respectively. This indicates that *W_CL_PME* produced the most stable temporal connections among the three spaces. In contrast, *W_CL_UMAP* had the smallest mean path length, but its standard deviation and maximum values were larger than those of *W_CL_PME*, indicating that abrupt jumps occurred in some time intervals. Therefore, temporal continuity should be assessed not only by the mean displacement but also by the spread and extreme values of the path-length distribution.

The same tendency is visible in the path-length histograms shown in Figure 5, which summarize the trail path-length distributions for a trail length of 12. The distribution for *W_CL_PME* is concentrated within a relatively narrow range and has a shorter right tail, indicating that temporally adjacent states are connected without large jumps in most cases. *W_CL_UMAP* has a smaller central path length but a longer right tail, suggesting that it produces short movements in many intervals but discontinuous jumps in some intervals. *WO_CL_UMAP* has the broadest distribution and the most pronounced right tail, indicating the least stable temporal continuity among the three spaces.

The third validation examines whether the positive pairs defined during contrastive learning remain close in the final state space. For this purpose, empirical cumulative distribution functions (ECDFs) of two-dimensional distances were computed between the base samples and their corresponding positive pairs, including reference, noisy, and translated pairs. The ECDFs are shown in Figure 6. An ECDF that rises more rapidly and approaches one at shorter distances indicates better preservation of positive-pair proximity.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

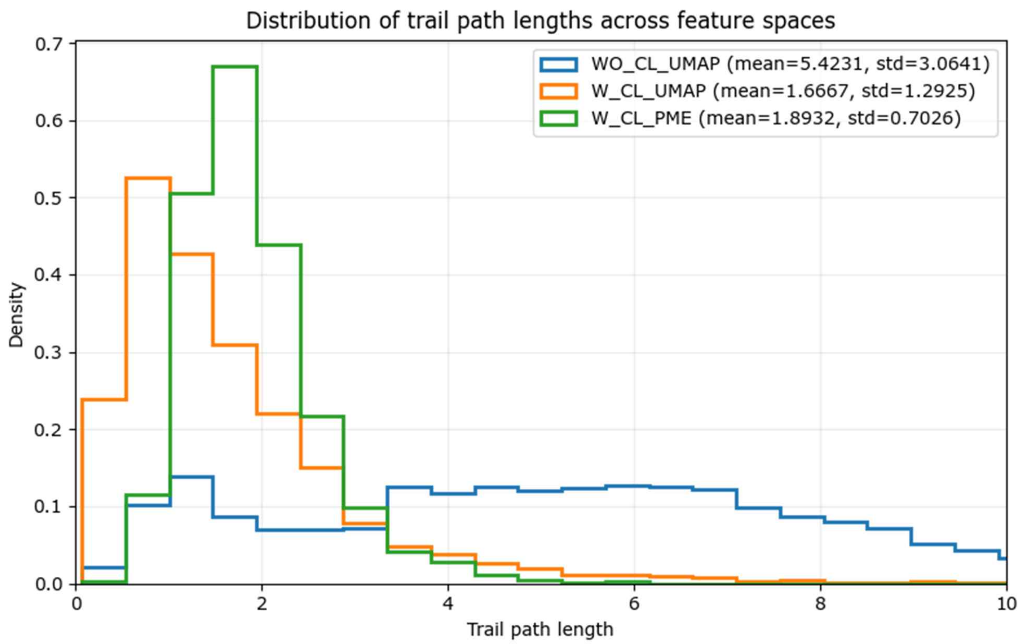

Figure 5. Distributions of trail path lengths in the three state spaces for a trail length of 12. The trajectories were formed by temporally consecutive atmospheric states.

WO_CL_UMAP showed the slowest increase in the cumulative curves for all three positive relationships, indicating relatively weak positive-pair proximity. In contrast, the two spaces constructed after contrastive learning showed much steeper increases. *W_CL_PME* showed better proximity preservation than *W_CL_UMAP* for the noise relationship and more stable convergence for the reference and translated relationships across both short and broader distance ranges. This indicates that the positive relationships learned during contrastive learning were maintained after the parametric manifold embedding step. It also suggests that *W_CL_PME* does not overemphasize a single positive relationship, but instead arranges different types of positive pairs consistently as neighboring states in the final space.

The fourth validation examines manifold distortion. To evaluate how well the distance relationships between the 128-dimensional feature space and the final two-dimensional state space were preserved, local neighbor distances were compared around each point in the two-dimensional state space. Figure 7 shows the mean 128-dimensional neighbor distance and the corresponding feature-distortion ratio. The outer regions of the state space tend to have larger 128-dimensional neighbor distances, whereas the central region tends to have smaller neighbor distances. This indicates that states with relatively large structural differences in the 128-dimensional feature space may be represented as closer points in some regions of the two-dimensional state space. When this relationship is expressed as a normalized distortion ratio, values closer to one indicate regions where the two-dimensional representation is relatively compact compared with the original feature space distance, whereas values closer to zero indicate regions where the two-dimensional distance is relatively large compared with the feature space distance.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

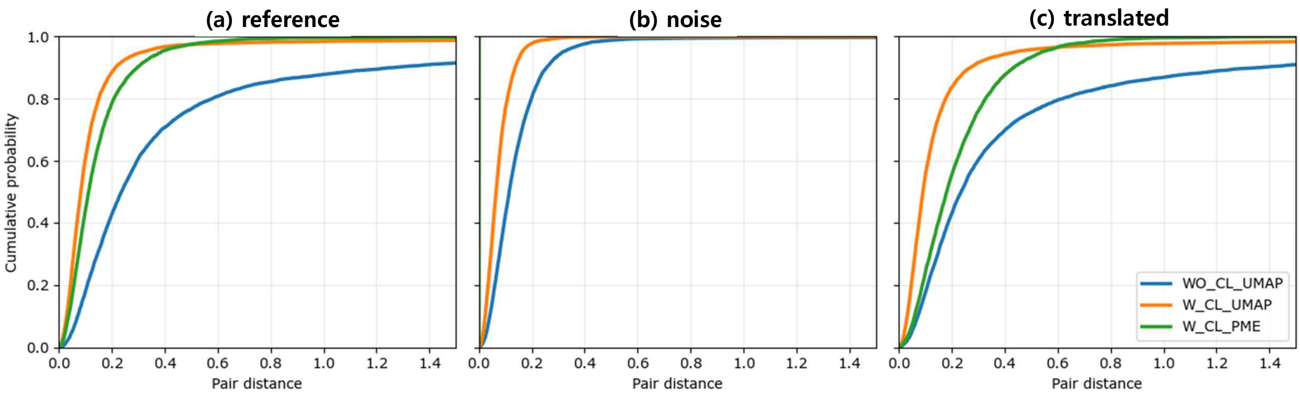

Figure 6. Empirical cumulative distribution functions of positive-pair distances in the three state spaces for (a) reference pairs, (b) noise-perturbed pairs, and (c) translated pairs.

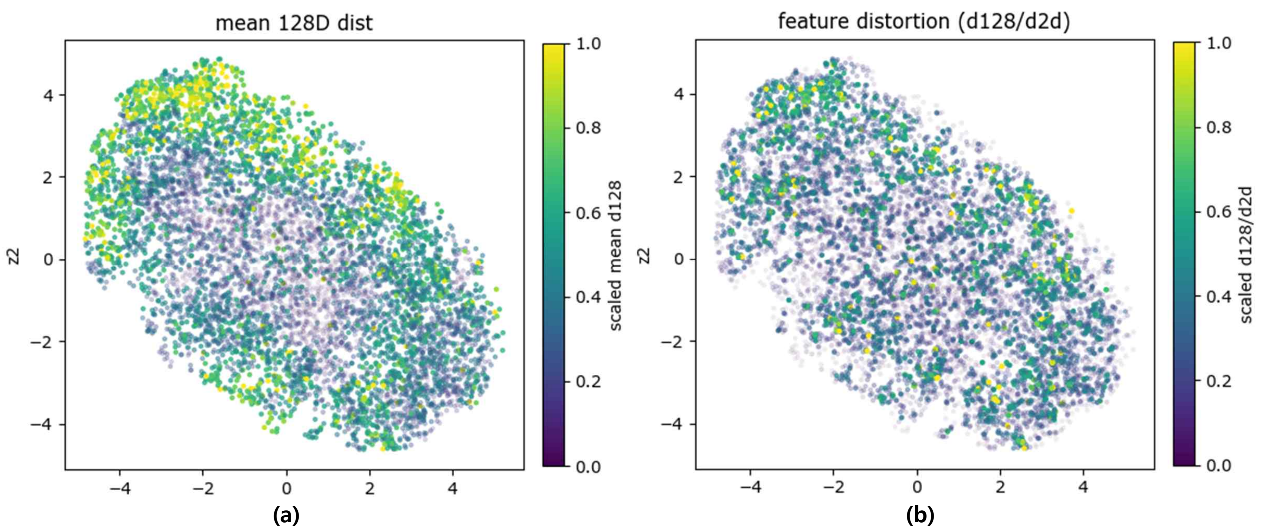

Figure 7. Distortion analysis of the shared two-dimensional state space based on feature space distances. The panels show (a) the mean 128-dimensional neighbor distance and (b) the corresponding feature-distortion ratio.

This analysis suggests that the same two-dimensional displacement may have different structural meanings depending on its location in the state space.

Taken together, these validation results indicate that the proposed state space is not merely a low-dimensional visualization. Rather, it provides a structured diagnostic space in which physically related states remain close, temporally adjacent states form smooth trajectories, and local distortions can be explicitly examined. These properties are important for trajectory-based forecast diagnosis, because the evaluation relies on the learned state space to preserve meaningful structural differences beyond

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

pointwise RMSE. In the following section, this validated state space is used to compare forecast trajectories from different models in both lead-time and valid-time settings.

3 Evaluation framework

3.1 Trajectory-based evaluation

In this study, the constructed topology-preserving state space is used to interpret AI-based medium-range forecast outputs as trajectories and to compare them with the corresponding ground truth trajectory. Two trajectory-based evaluation settings are defined. The first is lead-time trajectory evaluation, in which the forecast initial time is fixed and the evolution of the forecast state is examined as the forecast lead time increases. The second is valid-time trajectory evaluation, in which the target verification time is fixed and forecasts initialized at different earlier times are compared in terms of their convergence toward the target state. These two settings provide complementary views of forecast behavior: the former describes the forward evolution of a forecast from a given initial condition, whereas the latter describes the convergence of forecasts toward a common verification state.

Lead-time trajectory evaluation compares the path formed by a forecast initialized at a specific time over subsequent medium-range lead times. Each forecast sequence consists of state-space points at regular temporal intervals, and connecting these points in chronological order forms a forecast trajectory. By plotting the ground truth trajectory and the AI-model forecast trajectory in the same state space, it is possible to examine how closely the forecast follows the evolution of the observed atmospheric state in terms of trajectory shape and displacement, rather than only through pointwise errors at individual lead times.

Valid-time trajectory evaluation takes the opposite perspective. A single target verification time is fixed, and forecasts initialized at different lead times before that target time are projected into the same state space. This allows the evaluation to examine how forecast states approach the ground truth state as the initial time becomes closer to the target time. This setting is useful for comparing the convergence behavior and structural stability of different forecast models, particularly when the interest is not only the error at a fixed lead time but also the manner in which forecasts approach the observed atmospheric state.

3.2 Evaluation metrics

To quantify trajectory similarity, two distance metrics are used: index-aligned distance and alignment-based distance. The index-aligned distance compares points with the same temporal index along two trajectories. In this approach, the Euclidean distance is computed between corresponding points at the same lead-time index and then averaged over the trajectory segment. This metric is straightforward and directly comparable to conventional lead-time-based verification. However, it can penalize forecasts that follow a similar trajectory shape with a small timing shift. In such cases, a forecast may reproduce the structural evolution of the observed atmospheric state but still receive a large index-aligned distance because corresponding points are compared at fixed time indices.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

The alignment-based distance relaxes this fixed-index correspondence. Instead of enforcing one-to-one matching at identical indices, it searches for an optimal monotonic alignment between the two trajectories. In this study, dynamic time warping (DTW) is used for this purpose (Sakoe and Chiba, 1978) . More broadly, this follows the idea that trajectory similarity can be evaluated by considering the shape and alignment of paths rather than only point-to-point correspondence (Toohey and Duckham, 2015). DTW allows two trajectories to be compared in terms of their overall path similarity even when one trajectory leads or lags the other. This is relevant for medium-range forecast diagnosis because phase errors, timing shifts, and delayed pattern propagation are common sources of forecast discrepancy. The alignment-based distance is therefore used as a complementary metric to the index-aligned distance, providing information about trajectory shape and structural evolution that may not be captured by fixed-index comparison alone.

3.3 Experimental setup

The ground truth data for forecast evaluation were defined as the ECMWF IFS analysis for 2025. IFS analysis was used instead of ERA5 because the forecast evaluation in this study is more directly related to operational analysis fields. ERA5 provides a temporally consistent reanalysis dataset and is suitable for constructing the training state space, whereas IFS analysis provides a practical reference state for evaluating forecast outputs in an operational context.

Three AI-based medium-range forecast systems were evaluated: FourCastNet2, GraphCast, and Pangu-Weather. These models represent different network architectures and representation strategies among recent AI-based medium-range forecast models. For each model, three initial condition sources were used: IFS, KIM, and UM. This resulted in nine forecast configurations in total. Forecast outputs were produced at 6-h intervals up to 288 h, corresponding to 48 forecast lead times. The evaluation period covered 1 January to 31 December 2025, with forecast initial times sampled at 24-h intervals. This configuration provided 365 forecast cases and included seasonal variability and a range of atmospheric states.

For lead-time trajectory evaluation, each forecast sequence was represented as a trajectory from the forecast initial time to a future target time within the 12-day forecast range. In conventional RMSE-based evaluation, forecast skill is assessed by comparing the error of a specific variable at a specific lead time, for example the RMSE of 850 hPa geopotential height at a 5-day lead time. In contrast, trajectory-based comparison evaluates the path from the initial state to the target lead time. Therefore, the trajectory length increases as the evaluation lead time increases. In this study, the 288 h forecast range was divided into four trajectory windows ending at 3, 6, 9, and 12 days. This design allows short-range structural agreement and longer-range accumulated phase or trajectory deviations to be examined separately. For each window, distances between the forecast trajectory and the ground truth trajectory were computed using both the index-aligned and alignment-based metrics.

For valid-time trajectory evaluation, the target verification time was fixed, and forecasts initialized from 12 days to 1 day before the target time were selected at 1-day intervals. These 12 forecast states were connected in order to form a valid-time trajectory. Although valid-time trajectories may vary substantially for individual cases depending on the target date, their behavior over a full annual sample can be used to compare the overall convergence characteristics of forecast models. In this study, valid-time distance changes were analyzed for the full year of 2025 to evaluate how systematically each model converges toward the same target atmospheric state.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

4 Results

4.1 Lead-time trajectory

Using the feature-based state space constructed with contrastive learning and manifold embedding, the 2025 forecast cases from the nine forecast configurations were evaluated. The comparison showed that the relative behavior of the medium-range forecast models differed between RMSE-based comparison and trajectory-based comparison. Figure 8 presents representative examples for three forecast models, showing both their lead-time trajectories relative to the same ground truth trajectory and the corresponding variable-wise RMSE evolution. In the RMSE-based comparison (Figure 8, bottom), the relative ranking of the models varied depending on the meteorological variable and forecast lead time. For example, a model could show relatively low RMSE for temperature but higher RMSE for wind variables, or it could perform favorably at short lead times but lose this relative advantage at longer lead times. Thus, the four separate RMSE curves do not provide a direct view of which model more consistently reproduces the overall forecast path of the ground truth.

In contrast, the trajectory-based comparison (Figure 8, top) more directly shows whether each forecast path maintains a similar structure to the ground truth trajectory. The trajectory in the state space integrates multivariate information into a single continuous path. This representation makes it possible to examine when the forecast begins to deviate from an acceptable structural range, how consistently it follows the ground truth over a given period, and how drift accumulates at longer forecast lead times. Even when RMSE values were comparable, the trajectory shapes differed among models. Some models maintained the overall path structure relatively consistently, whereas others showed local departures or abrupt changes in direction. These differences indicate that the topology-preserving state space can reveal structural behavior that is not readily identified from pointwise error metrics alone.

The interpretation also depended on how the distance between trajectories was defined. Figure 9 compares index-aligned and alignment-based trajectory distances for the first example shown in Figure 8. In the distance matrix in Figure 9(a), the optimal alignment path generally follows the diagonal, but it deviates from the diagonal in several intervals with horizontal or vertical segments. This indicates that the forecast trajectory can retain a broadly similar structure to the ground truth trajectory while showing positional differences at the same lead-time index. In such intervals, the discrepancy can be interpreted as being related to a difference in trajectory progression speed or timing mismatch, rather than only to a collapse of the trajectory structure.

This difference is further illustrated by the cumulative distance curves in Figure 9(b). When cumulative distance was calculated along the ground truth index, the alignment-based distance increased more slowly than the index-aligned distance throughout the trajectory. The difference between the two curves became larger in the middle and later lead-time ranges. This suggests that, although errors accumulate substantially when states are compared at identical time indices, the overall trajectory shape can still be partly preserved when flexible alignment is allowed. Conversely, index-aligned comparison combines timing delays or differences in progression speed with structural deviations, and may therefore overestimate or underestimate the actual similarity between two trajectories.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

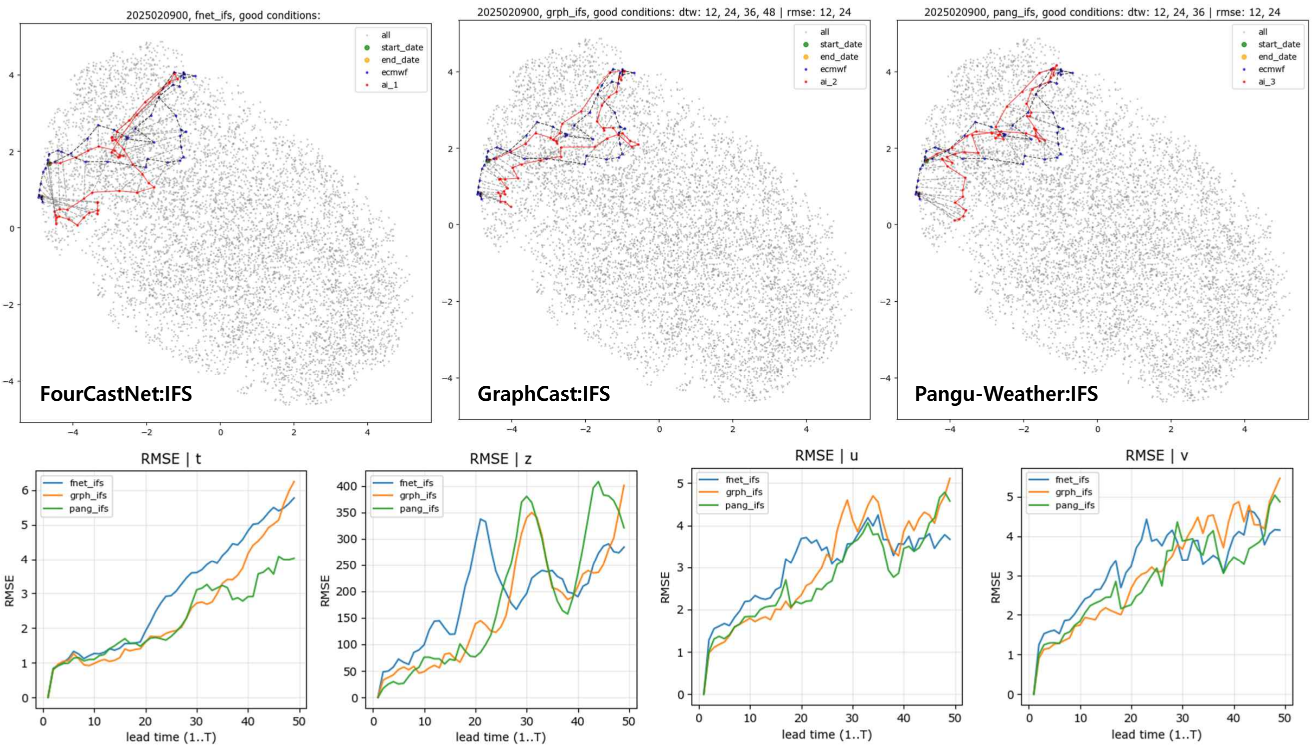

Figure 8. Representative examples of lead-time trajectory evaluation for three forecast models. The top panels show forecast trajectories on the shared state space, and the bottom panels show the corresponding lead-time evolution of variable-wise RMSE.

Taken together, this lead-time case shows that trajectory-based comparison provides structural information that is not available from variable-wise RMSE curves alone. When combined with alignment-based distance, the approach allows trajectory-shape similarity and temporal index mismatch to be interpreted separately. This result supports the use of both index-aligned pointwise comparison and trajectory-based structural comparison when diagnosing AI-based medium-range forecast models.

To summarize structural trajectory quality across all forecast cases, a quantitative criterion for identifying good trajectories was required. In this study, the threshold for a good trajectory was defined using the distance distribution of validation spatial-shift positive pairs used during contrastive learning. ECMWF-reference pairs and noise-perturbed pairs mainly represent small modifications of nearly identical states, whereas spatial-shift pairs include small phase displacements that are frequently observed in forecast errors. Spatial-shift pairs therefore provide a practical reference for states that remain structurally similar while allowing realistic positional errors. Accordingly, the 90th percentile of the validation shift-pair distance distribution, 0.4642, was used as the common threshold for defining a good trajectory. Although the threshold affects the absolute counts, it is used here as a summary device for aggregate comparison rather than as a standalone deterministic performance criterion.

Table 3 summarizes the number of good trajectories obtained from the 365 forecast initial dates for the 3-, 6-, 9-, and 12-day lead-time windows. Overall, the number of good trajectories was larger when using alignment-based distance than

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

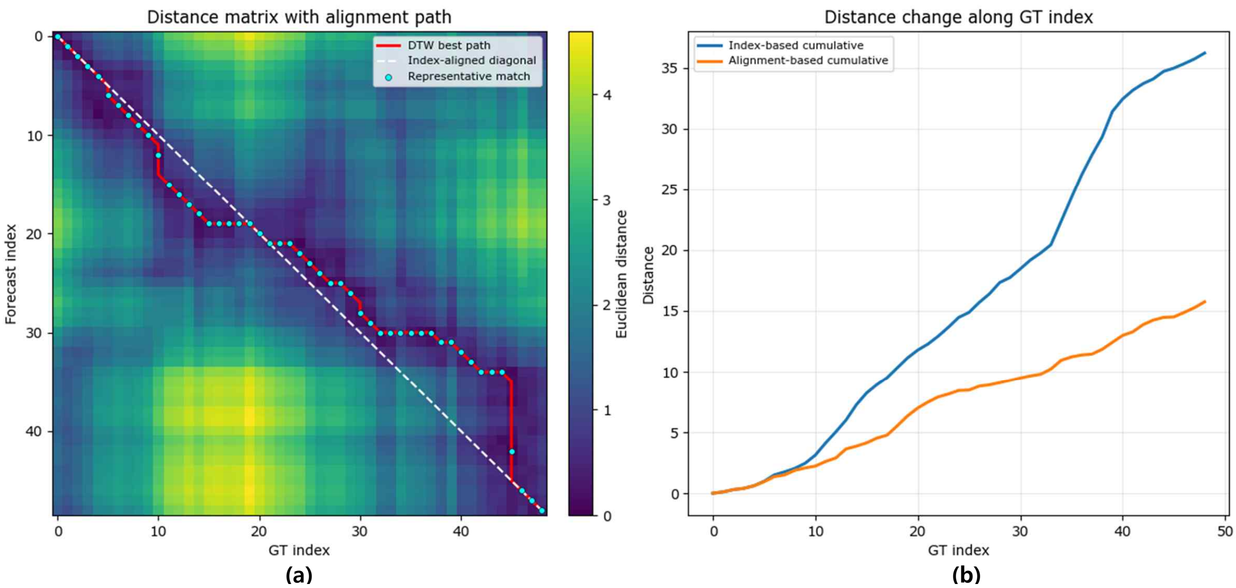

Figure 9. Example comparison of trajectory distances for a lead-time case using two matching strategies. Panel (a) shows the trajectory distance matrix with the optimal alignment path, and panel (b) shows cumulative trajectory distance for index-aligned and alignment-based comparisons.

when using index-aligned distance. This indicates that, in many cases, forecast paths evolve in a structurally similar manner to the ground truth trajectory but are penalized under index-aligned comparison because of small differences in temporal alignment. Among the model configurations, the Pangu-Weather configurations showed the largest number of good trajectories under alignment-based distance, with pang_ifs giving the highest count, whereas the GraphCast configurations, particularly graph_ifs, showed the largest number under index-aligned distance. Differences were also observed among the IFS, KIM, and UM initial-condition settings for the same backbone model. These results show that trajectory-based structural evaluation can diagnose not only average forecast behavior but also sensitivity to the initial condition.

By projecting the forecast initial states onto the state space and indicating whether each case is classified as a good trajectory, it is possible to identify state-space regions where each model performs relatively well. Based on the high good-trajectory frequencies in Table 3, the distributions for GraphCast:IFS and Pangu-Weather:IFS were compared. As shown in Figure 10, both models showed high overall densities of good trajectories, but their spatial distributions differed across seasonal state-space regions. GraphCast showed generally lower trajectory distance, but the density of good trajectories was relatively reduced in regions corresponding to spring and autumn states. In contrast, Pangu-Weather showed relatively weaker behavior in the winter region. Thus, although the two models may appear similarly favorable in terms of overall average performance, they exhibit different vulnerable regions within the topology-preserving state space. These differences are not easily identified from conventional RMSE time series alone. Mapping good-trajectory occurrence directly onto the state space provides a way to examine seasonality, state dependence, and structural bias more explicitly.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

Table 3. Counts of good lead-time trajectories for the nine forecast experiments for lead-time windows of 3, 6, 9, and 12 days, reported for alignment-based and index-aligned comparisons.

Lead-time	∼3 Days		∼6 Days		∼9 Days		∼12 Days		Mean count	
Metric	aligned	index	aligned	index	aligned	index	aligned	index	aligned	index
fnet:ifs	164	78	116	20	46	0	14	0	85.0	24.5
fnet:kim	157	64	106	19	37	0	10	0	77.5	20.8
fnet:um	156	70	102	15	49	1	11	0	79.5	21.5
graph:ifs	254	165	209	75	125	14	43	0	157.8	63.5
graph:kim	256	153	198	55	103	9	32	0	147.3	54.3
graph:um	224	118	156	42	78	6	26	0	121.0	41.5
pang:ifs	271	162	229	66	128	10	55	1	170.8	59.8
pang:kim	228	101	182	29	84	4	29	0	130.8	33.5
pang:um	227	115	170	39	81	1	29	0	126.8	38.8

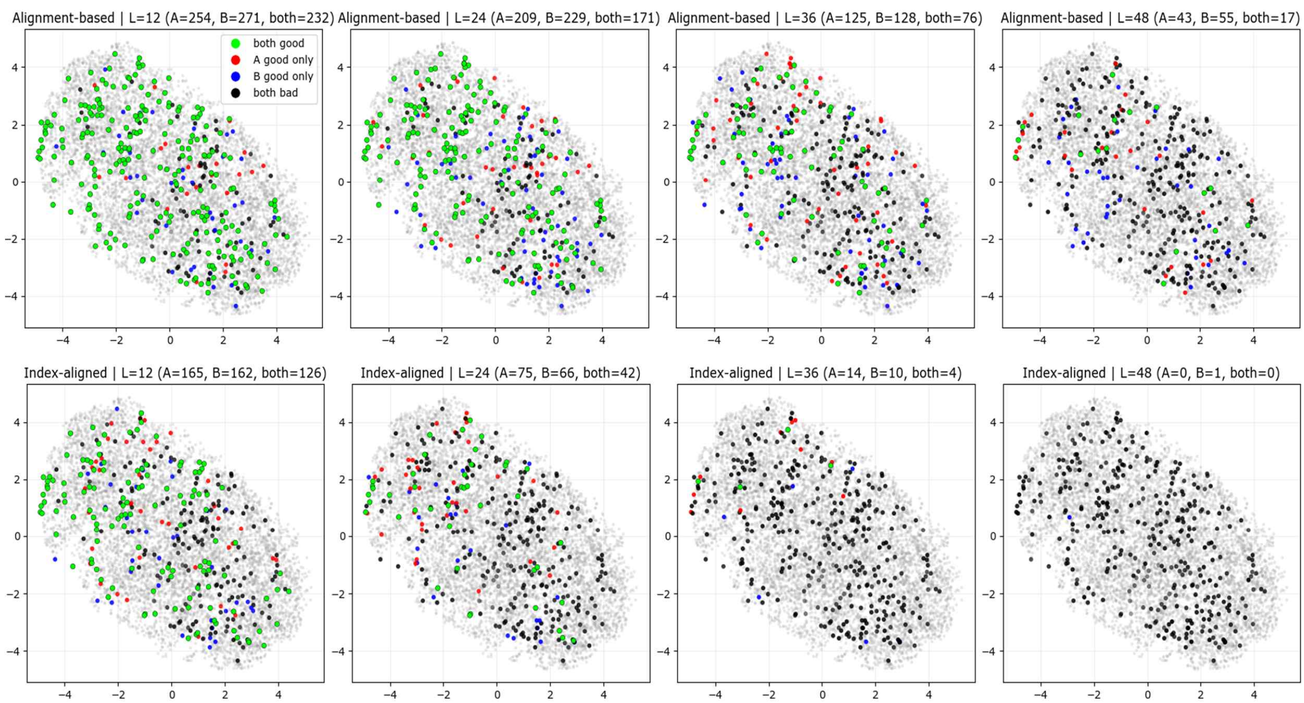

Figure 10. Distribution of good and bad lead-time trajectories on the shared state space for GraphCast:IFS (A) and Pangu-Weather:IFS (B). The top row shows alignment-based results, and the bottom row shows index-aligned results.

This seasonal dependence was also confirmed quantitatively by aggregating the seasonal proportions of good-trajectory outcomes (Table 4). When the four lead-time windows were combined, the proportion of cases in which both models produced

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

Table 4. Seasonal proportions of trajectory outcomes for GraphCast:IFS and Pangu-Weather:IFS, aggregated over all lead-time windows and categorized as both good, GraphCast only, Pangu only, and both bad.

Season	Both good	GraphCast-only	Pangu-Weather-only	All bad
DJF	48.1%	12.0%	6.0%	33.2%
MAM	33.1%	7.6%	13.6%	47.0%
JJA	24.7%	9.5%	17.1%	49.7%
SON	31.9%	7.6%	14.1%	46.4%

good trajectories was highest in winter (DJF), at 48.1 %, while the proportion of all-bad cases was lowest, at 33.2 %. This suggests that both models produced relatively stable structural forecasts in the winter state-space region. In spring (MAM), summer (JJA), and autumn (SON), the both-good proportions decreased to 33.1 %, 24.7 %, and 31.9 %, respectively, while the all-bad proportions increased to more than 46 %. In the relative comparison between the two models, the GraphCast-only proportion was higher than the Pangu-Weather-only proportion only in winter, with values of 12.0 % and 6.0 %, respectively. In contrast, the Pangu-Weather-only proportion was consistently higher in spring, summer, and autumn. These results indicate that the two models have different relative strengths and weaknesses depending on the seasonal state-space region, even when their overall performance appears similar.

4.2 Valid-time trajectory

The valid-time trajectories of the nine forecast configurations were also projected onto the state space. In general, the forecast states tended to converge toward the ground truth point as the target date approached. The representative case in Figure 11 shows this behavior in three ways. First, the state space trajectories in Figure 11(a) show forecasts initialized at different earlier times moving toward the neighbourhood of the ground truth state corresponding to the common target date. Second, the state space distance curves in Figure 11(c) generally decrease as the target date approaches, quantitatively confirming structural convergence in the state space. Third, the variable-wise RMSE curves in Figure 11(b) show the common tendency for errors to decrease when the forecast initialization time is closer to the target time. However, whereas RMSE shows separate error magnitudes for individual variables, the state space trajectory and distance curves provide an integrated view of how the multivariate atmospheric state approaches the ground truth.

This tendency was also observed in the mean valid-time distance curves calculated from 353 target dates (Figure 12). The valid-time analysis includes 353 target dates because the first 12 days of the year cannot be used when the full 12-day forecast history is required. For all models, the distance decreased as the target date approached, but differences in both the magnitude and slope of the decrease remained among model configurations. Overall, the GraphCast and Pangu-Weather configurations showed smaller mean distances than the FourCastNet2 configurations, and they tended to approach the ground truth more closely several days before the target date. This indicates that the convergence differences observed in individual cases were not restricted to isolated examples, but were also present in the annual mean.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

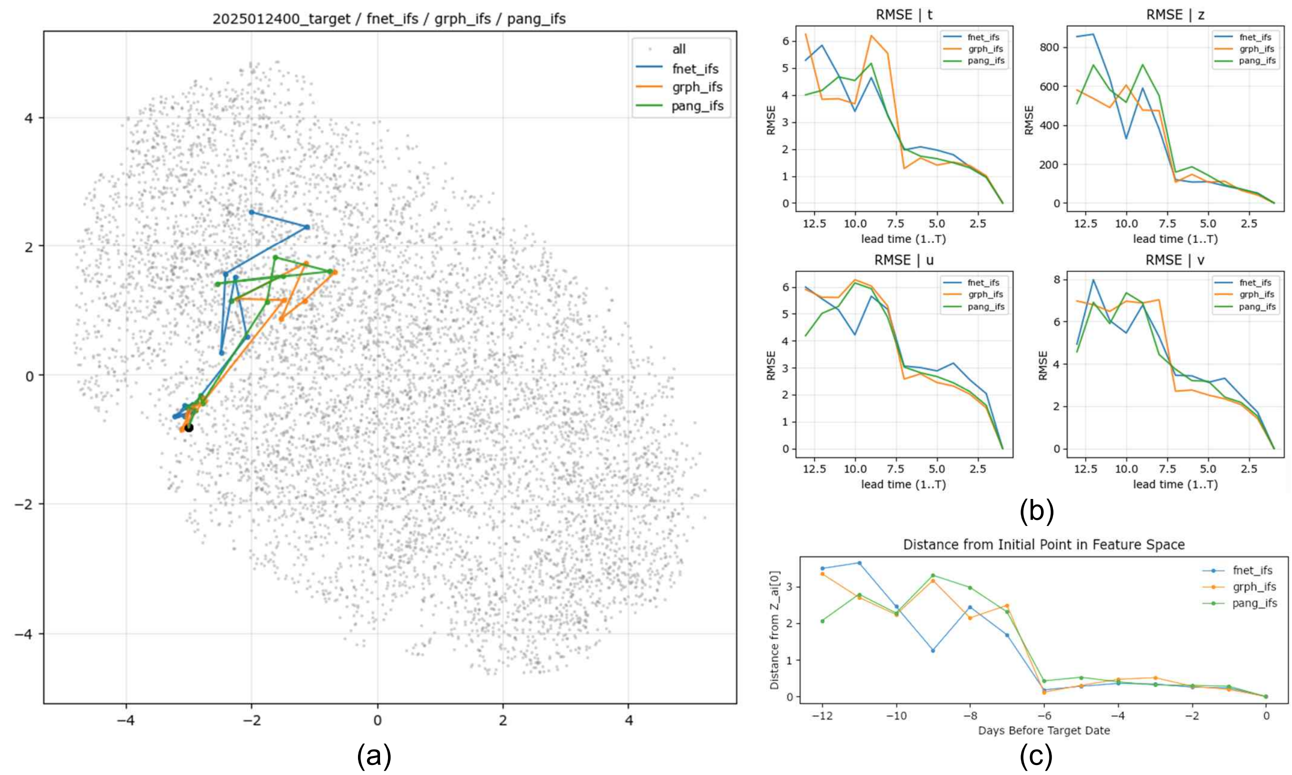

Figure 11. Representative example of valid-time trajectory evaluation for a common target date. Panel (a) shows state space trajectories, panel (b) shows variable-wise RMSE curves, and panel (c) shows state space distance curves.

To compare the distributional characteristics of the models, state space distances near the target date were summarized using boxplots (Figure 13). The tendencies observed in the mean curves were broadly consistent at the distributional level. For the final 1-step distance, final 6-step mean distance, and final 12-step mean distance, the GraphCast and Pangu-Weather configurations generally showed lower medians than the FourCastNet2 configurations. Configurations using IFS initial conditions also tended to show smaller distances. These results indicate that valid-time structural convergence depends on both the forecast-model backbone and the initial-condition setting. The relative tendencies are broadly consistent with those observed in the lead-time trajectory analysis.

5 Discussion

This study shows that representing multivariate atmospheric states in a topology-preserving state space and interpreting forecast outputs as trajectories can provide a useful complement to conventional RMSE-based evaluation. RMSE remains a simple and widely used metric because it summarizes pointwise differences at a given verification time. However, it does not readily

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

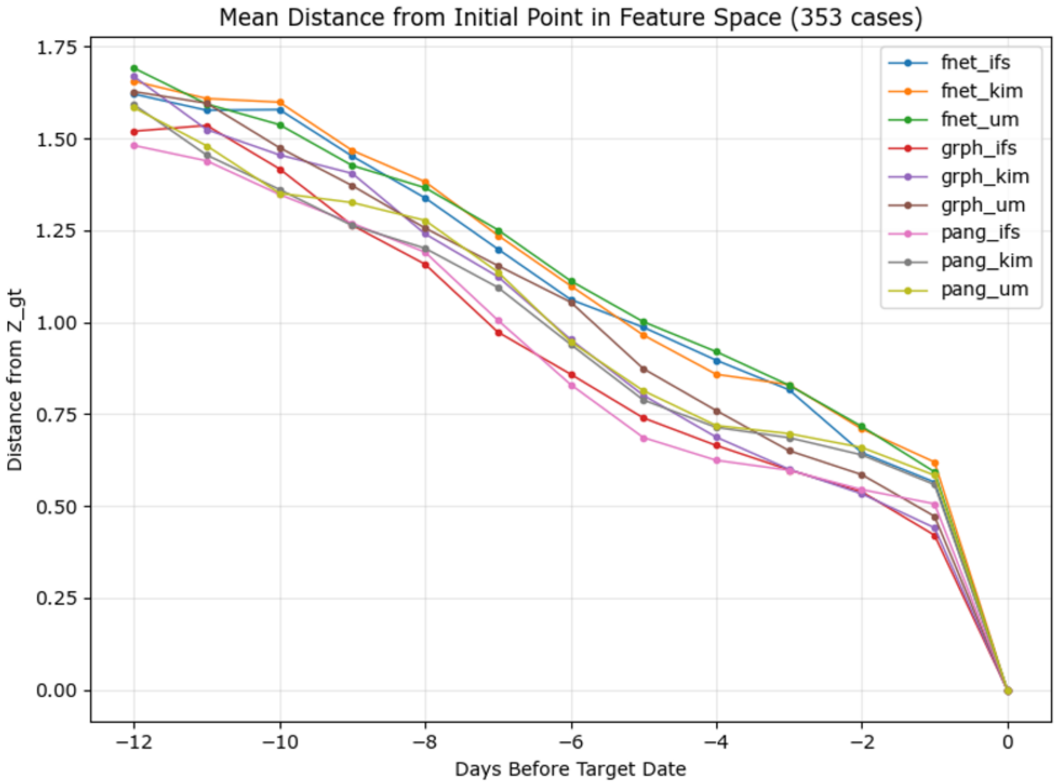

Figure 12. Mean valid-time distance to the target state in the shared state space for the nine forecast experiments, averaged over all valid target dates used in the analysis.

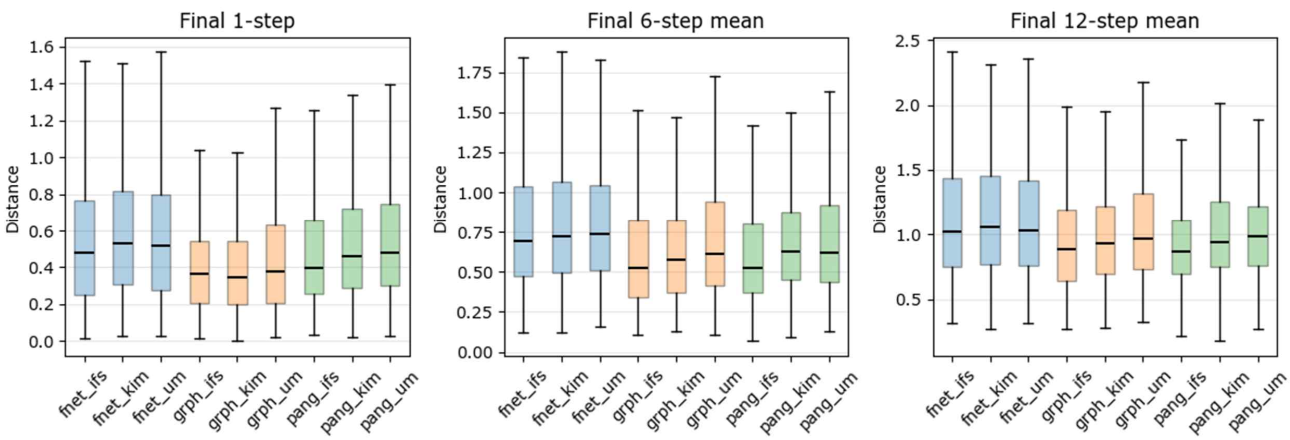

Figure 13. Boxplot summaries of valid-time state space distances for the nine forecast experiments, shown for the final 1-step distance, the final 6-step mean distance, and the final 12-step mean distance.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

distinguish between a forecast that follows the observed atmospheric evolution with a small timing offset and a forecast that evolves along a structurally different path. In contrast, the trajectory-based evaluation used in this study considers structural similarity, trajectory shape, and temporal alignment within a common state space. This makes it possible to examine whether forecast discrepancies are associated mainly with timing shifts, deformation of the spatial pattern, or accumulated drift at longer forecast lead times.

In this study, the term topology-preserving is used in an operational sense. It refers to the preservation of neighborhood relationships, positive-pair proximity, and temporal continuity that are relevant to forecast diagnosis, rather than to a formal topological equivalence between the original high-dimensional atmospheric state space and the two-dimensional embedding. The proposed state space should therefore be interpreted as a diagnostic representation designed to retain meaningful structural relationships among atmospheric states, not as a complete or lossless representation of the full atmospheric phase space.

The reference, noise, shift, and temporal relationships incorporated into the contrastive-learning procedure were central to the construction of a physically interpretable diagnostic space. Reference consistency encouraged ERA5 and IFS analysis fields at the same time to be represented near one another despite differences between data sources. Noise robustness and shift invariance reduced excessive sensitivity to appearance-level perturbations and small phase displacements, which are common in forecast fields. Temporal continuity encouraged temporally adjacent atmospheric states to form smooth paths in the state space, thereby supporting trajectory-based comparison. The positive-pair validation results indicated that these design elements were most consistently preserved in the *W_CL_PME* space. This suggests that the state-space construction was not simply a visualization procedure, but was shaped to support structural diagnosis of forecast behavior.

The proposed state space also has several interpretative limitations. First, a two-dimensional embedding cannot preserve all high-dimensional distance relationships exactly. The distortion analysis shown in Figure 7 indicates that the central and outer regions of the state space exhibit different distortion characteristics. Therefore, the same two-dimensional distance may have different structural meanings depending on its location in the state space. This implies that trajectory interpretation should consider not only absolute path length, but also the local context of the trajectory within the state space. In future extensions, distortion-aware weighting could be incorporated when computing trajectory distance, so that regions with stronger embedding distortion contribute differently to the trajectory-based comparison.

A second limitation concerns the possible tension between temporal continuity and topological consistency. Temporally adjacent states are not always structurally similar, particularly during rapid transitions. Conversely, structurally similar atmospheric states may occur at seasonally distant times. In such cases, temporal attraction and topological preservation can impose competing constraints on the embedding, potentially introducing spatial distortion. This is a general limitation of representing complex high-dimensional atmospheric variability in a two-dimensional plane. The primary purpose of the state space in this study is therefore not to assign a direct physical meaning to every coordinate, but to preserve meaningful structural relationships among observed atmospheric states. Temporal continuity should be imposed only to the extent that it does not substantially distort the underlying state space topology. In this sense, the proposed state space is designed to support smooth trajectory interpretation while giving priority to the structural organization of the high-dimensional feature space.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

The present implementation was trained for a limited East Asian domain, a single pressure level at 850 hPa, and four meteorological variables. This configuration was appropriate for evaluating the feasibility of the proposed framework, but the resulting seasonal structure, trajectory geometry, and relative model behavior may change when the method is extended to global domains, multiple pressure levels, or additional variables. Because dominant atmospheric structures differ by altitude, incorporating multiple pressure levels or linking level-specific state spaces may provide a richer basis for structural diagnosis. The method may also be useful for event-centered analyses, such as extreme-weather or typhoon cases, where distinguishing positional error from structural error is particularly important. These extensions were not evaluated here and should be treated as directions for future work rather than conclusions of the present study.

The distinction between RMSE-based evaluation and trajectory-based evaluation is also important when interpreting smoothed long-lead forecasts. AI-based medium-range forecasts often become smoother than the corresponding ground truth at longer lead times. Such smoothing can lead to larger raw RMSE, especially when small-scale features or local extrema are displaced or weakened. In the proposed state space, however, a smoothed forecast may still be located near the observed atmospheric state if it preserves the large-scale structural pattern. This behavior is not a contradiction, but reflects the different emphasis of the two evaluation approaches. RMSE measures pointwise amplitude and location errors in the original variable fields, whereas trajectory-based evaluation measures the evolution of multivariate structural similarity in the learned state space. Therefore, the proposed framework is suitable for diagnosing structural consistency and evolution, but it does not replace evaluations of small-scale detail, local extremes, or variable-specific error magnitude. It should be interpreted as a complementary diagnostic tool that provides information not fully captured by RMSE alone.

Beyond the experiments considered here, the proposed state space may provide a basis for regime-oriented analysis, analog-state retrieval, and structural interpretation of ensemble spread. This direction is consistent with recent studies showing that learned or regime-based representations can support the analysis of atmospheric predictability and state-dependent forecast behavior (Molina et al., 2023; Hoffmann and Lessig, 2023). Although these applications were not directly evaluated in the present study, the results suggest that the learned representation could be extended to these problems in future work.

6 Conclusions

This study presented a trajectory-based diagnostic framework for comparing AI-based medium-range weather forecast models in a topology-preserving state space. The framework constructs a shared representation of multivariate atmospheric states using ERA5 reanalysis data, image-based feature extraction, contrastive learning, and manifold embedding. In this space, forecast sequences from FourCastNet2, GraphCast, and Pangu-Weather can be represented as trajectories and compared with the corresponding ground truth trajectories. The lead-time and valid-time trajectory analyses showed that this representation provides an intuitive and visually interpretable way to examine how different models evolve from their initial conditions and how their forecasts approach target atmospheric states.

The results indicate that the proposed state space can reveal model-, state-, and season-dependent differences in forecast behavior that are not readily summarized by variable-wise RMSE curves alone. The framework is not intended to replace

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

conventional verification metrics, but to complement them by providing a structural view of multivariate forecast evolution. The present implementation is limited to an East Asian domain, four variables at 850 hPa, and a two-dimensional representation, and future work should examine extensions to broader domains, additional vertical levels, other variables, and ensemble-based applications. Overall, the proposed topology-preserving state space offers a practical diagnostic tool for visualizing and comparing the behavior of AI-based weather forecast models.

Code and data availability. The source code, trained models, configuration files, processed trajectory outputs, and precomputed RMSE curves used for the public reproduction workflow are archived on Zenodo at 10.5281/zenodo.20483795 under the licence specified in the archived repository (Kim and Cho, 2026). The archive corresponds to the exact repository version associated with this manuscript. The archived repository contains the materials required to reproduce the lead-time and valid-time trajectory figures, RMSE curves, and related diagnostic outputs using the default show-only workflow described in the repository README. The repository also includes the source code for the main state-space construction procedures, including feature extraction, contrastive representation learning, and parametric manifold embedding. However, the public reproduction workflow does not rerun the full state-space construction process, because it requires the original gridded meteorological fields and substantial preprocessing and training. The raw IFS analysis and forecast fields used in the full experiments were obtained through institutional access and cannot be publicly redistributed. To support reproducibility within these restrictions, the archived package provides trained model weights, processed state-space outputs, precomputed RMSE curves, and figure-generation inputs rather than the restricted raw gridded fields.

Author contributions. HK: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft preparation, Writing – review and editing. JC: Supervision, Project administration, Resources

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

Acknowledgements. The author(s) declare that financial support was received for the research and/or publication of this article. This work was funded by the Korea Meteorological Administration Research and Development Program “Development of AI-based Weather Forecasting Technology” under Grant (KMA 2021-00121).

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

References

Ben Bouallègue, Z., Alexe, M., Lang, S., Clare, M. C. A., Prieto-Nemesio, A., Dramsch, J., Pinault, F., and Raoult, B.: Data-driven ensemble forecasting with the AIFS, ECMWF Newsletter, 181, 2024, https://doi.org/10.21957/ma3p95hxe2, 2024.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Bjerregård, M. B., Møller, J. K., and Madsen, H.: An introduction to multivariate probabilistic forecast evaluation, Energy and AI, 4, 100058, https://doi.org/10.1016/j.egyai.2021.100058, 2021.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown, B. G., and Mason, S.: Forecast verification: current status and future directions, Meteorol. Appl., 15, 3–18, https://doi.org/10.1002/met.52, 2008.

Davis, C. A., Brown, B. G., and Bullock, R.: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas, Mon. Weather Rev., 134, 1772–1784, https://doi.org/10.1175/MWR3145.1, 2006.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorol. Appl., 15, 51–64, https://doi.org/10.1002/met.25, 2008.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of spatial forecast verification methods, Weather Forecast., 24, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1, 2009.

Gilleland, E., Brown, B. G., and Ebert, E. E.: Verifying forecasts spatially, Bull. Amer. Meteor. Soc., 91, 1365–1373, https://doi.org/10.1175/2010BAMS2819.1, 2010.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hoffmann, S. and Lessig, C.: AtmoDist: self-supervised representation learning for atmospheric dynamics, Environ. Data Sci., 2, e6, https://doi.org/10.1017/eds.2023.7, 2023.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D.: Supervised contrastive learning, Adv. Neural Inf. Process. Syst., 33, 18661–18673, 2020.

Kim, H. and Cho, J. H.: Topology-preserving state space representation for diagnosing weather forecast models: code and reproduction package, Zenodo [code], https://doi.org/10.5281/zenodo.20483795, 2026

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Ben Bouallègue, Z., Prieto-Nemesio, A., Dueben, P. D., Brown, A., Pappenberger, F., and Rabier, F.: AIFS – ECMWF's data-driven forecasting system, arXiv [preprint], https://doi.org/10.48550/arXiv.2406.01465, 2024.

Lee, J. A. and Verleysen, M.: Quality assessment of dimensionality reduction: rank-based criteria, Neurocomputing, 72, 1431–1443, https://doi.org/10.1016/j.neucom.2008.12.017, 2009.

https://doi.org/10.5194/egusphere-2026-3142
Preprint. Discussion started: 31 July 2026

McInnes, L., Healy, J., and Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv [preprint], https://doi.org/10.48550/arXiv.1802.03426, 2018.

Molina, M. J., Richter, J. H., Glanville, A. A., Dagon, K., Berner, J., Hu, A., and Meehl, G. A.: Subseasonal representation and predictability of North American weather regimes using cluster analysis, Artif. Intell. Earth Syst., 2, e220051, https://doi.org/10.1175/AIES-D-22-0051.1, 2023.

Moor, M., Horn, M., Rieck, B., and Borgwardt, K.: Topological autoencoders, Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, 119, 7045–7054, 2020.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A.: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators, arXiv [preprint], https://doi.org/10.48550/arXiv.2202.11214, 2022.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M.: Probabilistic weather forecasting with machine learning, Nature, 637, 84–90, https://doi.org/10.1038/s41586-024-08252-9, 2025.

Rife, D. L. and Davis, C. A.: Verification of temporal variations in mesoscale numerical wind forecasts, Mon. Weather Rev., 133, 3368–3381, https://doi.org/10.1175/MWR3052.1, 2005.

Sainburg, T., McInnes, L., and Gentner, T. Q.: Parametric UMAP embeddings for representation and semi-supervised learning, Neural Comput., 33, 2881–2907, https://doi.org/10.1162/neco_a_01434, 2021.

Sakoe, H. and Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans. Acoust. Speech Signal Process., 26, 43–49, https://doi.org/10.1109/TASSP.1978.1163055, 1978.

Scheuerer, M. and Hamill, T. M.: Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities, Mon. Weather Rev., 143, 1321–1334, https://doi.org/10.1175/MWR-D-14-00269.1, 2015.

Tan, M. and Le, Q. V.: EfficientNet: rethinking model scaling for convolutional neural networks, Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, 97, 6105–6114, 2019.

Toohey, K. and Duckham, M.: Trajectory similarity measures, SIGSPATIAL Special, 7, 43–50, https://doi.org/10.1145/2782759.2782767, 2015.

Venna, J., Peltonen, J., Nybo, K., Aidos, H., and Kaski, S.: Information retrieval perspective to nonlinear dimensionality reduction for data visualization, J. Mach. Learn. Res., 11, 451–490, 2010.

Wernli, H., Hofmann, C., and Zimmer, M.: Spatial forecast verification methods intercomparison project: application of the SAL technique, Weather Forecast., 24, 1472–1484, https://doi.org/10.1175/2009WAF2222271.1, 2009.