**Review 2 of Topology-Preserving State Space Representation for Diagnosing Weather Forecast Models**

This manuscript presents an interesting and potentially useful trajectory-based diagnostic framework for AI medium-range weather forecasts. By constructing a learned multivariate state space from ERA5 and IFS analysis and viewing forecasts from FourCastNet2, GraphCast, and Pangu-Weather as trajectories, the authors provide a more structural picture of model behavior than standard RMSE curves can offer. The physically motivated contrastive-learning setup, the careful checks on seasonal organization and temporal continuity, and the ability to highlight regimes where different AI models behave differently are clear strengths.

At the same time, the current implementation is quite specific to one region, level, and variable set, and the paper needs to state more clearly that the results are domain- and variable-dependent and to outline how the framework might generalize. The notion of a "topology-preserving" 2D embedding also needs more down-to-earth explanation, so that readers don't over-interpret distances without consulting the distortion analysis. In addition, it would help to show more explicitly how a model developer might use these diagnostics in practice, and to tie the "good trajectory" concept to familiar synoptic systems through a concrete case study. Finally, the title should be brought into line with the AI-focused scope (or the experiments broadened to include an NWP model), and the reliance on ERA5/IFS rather than direct observations as the reference state should be acknowledged and discussed, ideally with a brief outlook on how observational data could be brought into the framework in future work. On balance, I recommend major revision.

Author response:

We sincerely thank the reviewer for the thoughtful and constructive review. We appreciate the recognition of the diagnostic value of the trajectory-based framework and its physically motivated learning approach. The comments have helped us identify where further explanation and supporting evidence would make the manuscript clearer to readers.

In response, we have added a 500-hPa sensitivity experiment and an additional analysis of embedding distortion, clarifying both the framework's applicability and the interpretation of state-space distances. We have also included a weather-map case study linking good/bad trajectory classifications to synoptic structures and expanded the discussion of potential uses in model development. The Discussion and Conclusions now address configuration-dependent findings, the use of ERA5/IFS reference data, and future extensions incorporating observations, additional levels, variables, and domains. We have revised the title to reflect the AI-focused scope and supplemented the methodological descriptions and quantitative summaries.

We provide detailed responses to each major and specific comment below, together with the corresponding manuscript changes.

Major comments:

1. Domain and variable generality

The implementation focuses on a regional domain centered on East Asia, 850 hPa only, and four variables (temperature, geopotential height, zonal wind, meridional wind). This configuration is well-justified as a testbed for medium-range evolution, and the authors note that it represents a small fraction of the global domain. However, readers may wonder how general the conclusions are. For

example, whether findings about GraphCast vs Pangu-Weather seasonal behavior and trajectory quality extend to other regions, levels, or variable combinations. Therefore, I recommend more explicitly emphasizing in the Conclusions that the diagnostic insights reported here are domain- and variable-specific and should not be interpreted as global performance rankings. Additionally, it would also be valuable to briefly discuss how the framework could be extended to multiple levels or larger domains (e.g., by stacking variables across levels or training separate embeddings), and any anticipated challenges (e.g., higher dimensionality, more complex seasonal organization). This would help communicate that the method is not fundamentally limited to the chosen configuration.

Author response:

Thank you for this comment. We have added a 500-hPa sensitivity experiment and expanded the discussion of generalizability and possible extensions.

Using the same regional domain and four variable types, we trained a separate state-space representation at 500 hPa. The resulting representation exhibited comparable seasonal organization and qualitatively similar spatial patterns of local neighbor distances and embedding distortion (Figs. R2-F1 and R2-F2). Furthermore, GraphCast-IFS and Pangu-Weather-IFS remained the two leading configurations among the nine forecast configurations in mean good-trajectory counts across the four lead-time windows under both distance metrics. These findings provide experimental support for the applicability of the framework across pressure levels and the consistency of its broad diagnostic findings. The joint good/bad classifications of GraphCast:IFS and Pangu-Weather:IFS at 500 hPa are shown in Fig. R2-F3.

The detailed seasonal distributions of trajectory quality nevertheless differed between the two levels (Table R2-T1). We have therefore clarified in the Conclusions that the seasonal strengths and relative performance reported for the original configuration are conditional on the selected atmospheric representation and should not be interpreted as absolute or global model rankings. Generality across other geographical domains and alternative variable combinations remains to be evaluated.

Extension to additional variables or multiple pressure levels within the same domain is conceptually straightforward: variable- and level-specific features can be combined through the existing feature-fusion approach, followed by contrastive learning and manifold embedding. This would require retraining and appropriate normalization and balancing of the inputs, but would retain the trajectory-based diagnostic procedure. For larger domains, a more substantial consideration is interpretability. Multiple weather systems may evolve simultaneously and differently across the domain, so a trajectory representing their combined evolution may be harder to associate with a particular atmospheric movement or synoptic process.

The contrastive-learning relationships would also need to be adjusted to the spatial and temporal characteristics of the data. At the current 0.25° resolution, a given translation in grid cells represents a particular geographical displacement. With coarser data, such as 0.5° or 1°, fewer grid cells would generally be needed to retain a comparable physical displacement range for spatial-shift positive pairs.

Similarly, temporal neighborhoods should be specified in elapsed time rather than a fixed number of steps. For example, a 12-hour neighborhood corresponds to two steps with 6-hourly data, four steps with 3-hourly data, and one step with 12-hourly data. The appropriate temporal window would also need to be validated for the atmospheric evolution being represented. We have added these considerations to explain how the framework can be extended while adapting its positive-pair definitions to the characteristics of each dataset.

2. Interpretation of “topology-preserving” and embedding distortions

The paper uses the term “topology-preserving state space” and provides a thoughtful distortion analysis showing that local neighbor distances in the 128D feature space and the 2D embedding can differ, with outer regions having larger 128D neighbor distances and variable distortion ratios. The authors note that low-dimensional representations can distort neighborhood relationships and that explicit diagnostics are needed. Given the potential for readers to over-interpret 2D coordinates and distances, I suggest further emphasizing that the embedding preserves key neighborhood and continuity properties relevant to trajectory diagnostics, but not all aspects of the high-dimensional topology, and that the distortion maps should be consulted when interpreting large displacements or aggregating trajectories.

Author response:

Thank you for this helpful suggestion. We agree that the meaning and limitations of “topology-preserving” should be stated more explicitly so that readers do not over-interpret the 2D representation.

To further support this clarification, we conducted an additional analysis comparing temporal transitions with similar 2D displacements across regions with different local distortion levels (Fig. R2-F4). The results showed that comparable movements in 2D could correspond to larger changes in the 128D feature space in regions with greater relative compression. This provides a concrete illustration of why trajectory distances should be interpreted in their local state-space context.

We have accordingly expanded the Discussion to clarify that the embedding is designed to preserve neighborhood relationships and temporal continuity relevant to trajectory diagnostics, without implying complete preservation of the high-dimensional topology or distances. We also emphasize that similar reference–forecast distances in different parts of the 2D space should not automatically be interpreted as equivalent feature-space discrepancies. Following the reviewer’s recommendation, we advise consulting the distortion maps when interpreting large displacements or aggregating trajectory diagnostics, while retaining the representation’s value for examining structural relationships and forecast evolution.

3. Practical use cases for model developers

The trajectory diagnostics reveal interesting patterns. For instance, cases where models have similar RMSE but differing trajectory quality, seasonal regimes where GraphCast or Pangu-Weather trajectories are more often classified as “good”, and differences between timing errors (index-based distances) and structural differences (alignment-based distances). These insights have clear value for model developers and possibly for operational forecast verification, but the manuscript could spell out more about how the framework might be used in practice. For example, you may add one use cases such as a scenario where trajectory diagnostics highlight regimes (season/region) with consistently poor structural evolution for a given AI model, guiding targeted retraining or postprocessing.

Author response:

Thank you for highlighting this practical aspect. We agree that explaining how the diagnostic results could inform model development is important, and we have expanded the Discussion to include concrete potential applications.

Following the reviewer’s suggestion, we describe how developers could identify seasons or state-

space regions in which a model repeatedly produces trajectories that deviate structurally from the reference evolution. Examining the corresponding atmospheric fields could then help characterize the conditions associated with these weaknesses and guide targeted training-data enrichment, retraining, or post-processing. The diagnostics would provide a basis for selecting cases and development priorities, while the effectiveness of any intervention would require subsequent evaluation. The state-space and seasonal classifications in Fig. R2-F3 and Table R2-T1, together with the trajectory–weather-map comparison in Fig. R2-F5, illustrate the diagnostic information that could support such case selection.

We also discuss comparing successive model versions or alternative training configurations within the same fixed state-space representation. For example, two versions with similar RMSE could exhibit different agreement with the reference trajectory, allowing developers to assess whether a modification improves the consistency of forecast evolution in addition to pointwise accuracy. Applying conventional verification and trajectory diagnostics together could therefore provide a more informative assessment of model changes.

These applications are presented as development scenarios supported by the diagnostic capabilities of the framework; their effectiveness in improving forecast models has not been evaluated in the present study.

4. Scope and title alignment: AI vs. NWP models

Although the abstract and introduction emphasize “AI-based medium-range weather forecast models,” the current title (“Topology-Preserving State Space Representation for Diagnosing Weather Forecast Models”) is relatively broader and could be interpreted as covering both AI and conventional numerical weather prediction (NWP) models. In the experiments, all forecast systems are AI-based (FourCastNet2, GraphCast, Pangu-Weather), with IFS, KIM, and UM providing initial conditions and reference fields rather than being evaluated as forecast models in the same trajectory framework. I recommend either (a) updating the title to explicitly reflect the current scope, e.g., “Topology-Preserving State Space Representation for Diagnosing AI-Based Weather Forecast Models,” or (b) expanding the evaluation to include at least one physics-based NWP model as a forecast system in the trajectory analysis. The latter option would be particularly powerful as it would demonstrate how the diagnostic framework can compare AI and NWP models within a common state space, potentially revealing regime-dependent differences in structural predictability and drift that are of high interest to the broader GMD readership.

Author response:

Thank you for this helpful suggestion. We fully agree that the title should explicitly reflect the AI-focused scope of the experiments presented in this manuscript. Accordingly, we have revised the title to:

“Topology-Preserving State Space Representation for Diagnosing AI-Based Weather Forecast Models”

This revision aligns the title more closely with the manuscript’s content and clarifies that the forecast models evaluated in this study are AI-based. We appreciate the reviewer’s recommendation to make the scope clearer to readers.

5. Linkage between trajectories and synoptic patterns/weather systems

The definition of a “good trajectory” is clear and methodologically grounded: trajectories whose

distances fall below a threshold derived from the 90th percentile of spatial-shift positive pairs are considered structurally similar to the reference, allowing realistic positional errors. However, from a meteorological perspective, it remains somewhat abstract for readers to connect “good trajectories” to the predictability of specific weather systems (e.g., fronts, cyclones, jets). To make the diagnostic more tangible, I suggest including at least one case-study demonstration where trajectory classification is explicitly linked to synoptic features of interest. For example, select a medium-range event involving a frontal passage, a midlatitude cyclone, or a strong jet and show the corresponding trajectories in the state space, alongside maps of the relevant features. You could then illustrate whether “good trajectories” correspond to accurate evolution of the synoptic pattern (even with timing shifts) and whether “bad trajectories” reflect structural failures (e.g., misrepresentation of cyclone deepening or frontal orientation). This would help readers see how the abstract trajectory metrics translate into familiar synoptic concepts and how the framework might inform regime-dependent predictability assessments.

Author response:

Thank you for this helpful suggestion. We agree that connecting trajectory classification to recognizable synoptic features is important for making the diagnostic results physically interpretable.

We have added a case study initialized at 00 UTC on 9 October 2025, showing the reference trajectory and two forecast trajectories together with their corresponding weather maps at selected lead times through +144 h (Fig. R2-F5). Pangu-Weather:IFS, classified as a good trajectory over this window, reproduces the approximate position and subsequent eastward movement of the closed low evident in the reference 850-hPa geopotential-height and wind fields. In contrast, FourCastNet2:IFS, classified as a bad trajectory, does not maintain the distinct closed circulation evident in the reference at +72 and +108 h, alongside a more pronounced departure from the reference trajectory.

We have added an accompanying discussion linking these differences in the weather maps to the trajectory classification. This case provides a concrete example of how structural agreement or departure in the learned state space can relate to the evolution of an identifiable weather system, complementing the aggregate trajectory statistics presented in the manuscript.

6. Use of reanalysis and IFS analysis as reference vs. direct observations

The framework constructs the state space using ERA5 reanalysis (2020–2024) and evaluates forecasts against 2025 IFS analysis fields as the ground truth. This is reasonable from a modeling and operational perspective, and ERA5/IFS are widely used benchmarks. However, it does mean that both the training and evaluation reference are model-based products rather than direct observations, which may raise questions for readers who view observations as closer to the “truth.” I suggest more explicitly discussing this limitation. For example, note that the state space and trajectory diagnostics are currently referenced to ERA5/IFS because of data availability, consistency, and resolution, and that incorporating direct observations (e.g., radiosondes, satellite-derived fields, or observation-based analyses) would require additional preprocessing, quality control, and possibly different feature encoders. You could also comment on how discrepancies between IFS and observations might affect interpretation of “good trajectories”. If feasible, potential of integration with observational datasets for model evaluation in future work would be also beneficial.

Author response:

Thank you for this helpful comment. We agree that the choice of reference data and its implications should be explained more explicitly. We have added a paragraph to the Discussion addressing these points.

ERA5 and IFS analysis were selected because of their availability, spatial and temporal consistency, and provision of multivariate gridded fields at a suitable resolution for the common input representation. Direct observations do not generally provide equivalent coverage across all variables, locations, and evaluation times. We now explicitly acknowledge that reanalysis and analysis fields are model-informed estimates incorporating observations, rather than direct observational truth.

We also clarify that a "good trajectory" indicates structural agreement with the IFS-analysis reference within the learned representation. Errors or biases in that reference can affect trajectory distances and classification, so agreement with the analysis does not necessarily imply equally close agreement with observations. Some classifications could therefore change if an observational reference were used.

As a future extension, we discuss constructing observation-based reference trajectories using measurements corresponding to the variables represented in the state space. For example, radiosonde datasets provide temperature, geopotential height, and winds that can be matched to the selected pressure levels. Incorporating these observations would require harmonizing units and observation times, accommodating irregular coverage and missing data, and adapting the input representation so that observations and forecasts can be compared in a common space. This would allow us to examine whether analysis-based good/bad classifications remain supported by observations and could ultimately provide an alternative to IFS-analysis reference trajectories.

Specific comments:

1. Section 1 and title: Consider revising the title to reflect the current focus on AI-based models (e.g., "AI-Based Weather Forecast Models") unless you plan to extend the evaluation to physics-based NWP models as additional forecast systems.

Author response:

Thank you for this suggestion. We agree that the title should clearly reflect the AI-focused scope of the study. As also noted in our response to Major Comment 4, we have revised the title to:

"Topology-Preserving State Space Representation for Diagnosing AI-Based Weather Forecast Models". This change makes the scope explicit and aligns the title more closely with the forecast models evaluated in the manuscript.

2. Section 2.1: It would be helpful to comment more explicitly on why 850 hPa and the chosen four variables were selected, and whether other levels or additional variables (e.g., humidity) were tested or considered. A brief discussion of how adding more levels or variables might affect the feature, and state spaces would be useful.

Author response:

Thank you for this helpful suggestion. We have clarified in Section 2.1 why 850 hPa and the four

variables were selected. This level provides a testbed for examining lower-tropospheric evolution, while temperature, geopotential height, and the two horizontal wind components describe complementary aspects of thermal structure and circulation. Their availability across the datasets also enables consistent representation and comparison.

As noted in our response to Major Comment 1, we have additionally conducted a sensitivity experiment at 500 hPa using the same four variable types (Figs. R2-F1–R2-F3 and Table R2-T1). We report these results in the model-evaluation section, where they can be considered alongside the original trajectory diagnostics.

We have expanded the Discussion to explain how additional levels or variables could affect the feature representation and learned state-space organization. Such extensions can be accommodated through feature fusion, but would require retraining and validation for the new input configuration. Additional variables such as humidity were not tested in the present study and are identified as a direction for future work. This separates the input-selection rationale in Section 2.1 from the experimental evidence and broader discussion of extensions.

3. Section 2.2: It may help to include a short illustrative example or diagram showing how a single ERA5 field generates its various positive views (reference, perturbed, translated, temporal neighbors) to make the construction more concrete for readers less familiar with supervised contrastive learning.

Author response:

Thank you for this suggestion. We have added an illustrative figure to Section 2.2 (Fig. R2-F6) showing representative examples of the positive views used in contrastive learning, using temperature at 850 hPa.

For each ERA5 base field, the figure presents the corresponding IFS analysis at the same valid time (reference view), a noise-perturbed version, a spatially translated version, and a temporally adjacent ERA5 field (illustrated here at t + 6 h). The reference and temporal views are therefore selected from corresponding analysis data and neighboring times, respectively, whereas the noise and translated views are generated by modifying the base field.

The figure also includes difference maps calculated by subtracting the base field from each view. The original fields share a common color scale within each variable, while each difference map uses its own zero-centered color scale to make differences of varying magnitudes visible. These examples clarify the construction of the positive views and their differences from the base field, providing a concrete illustration of the relationships used to encourage reference consistency, noise robustness, tolerance to small spatial displacements, and temporal continuity.

4. Equations (1) to (5): The loss definitions are clear, but the notation for the pair sets and weights could be supported by a brief table or bullet list summarizing which view types correspond to which weights and how they are combined in each training stage. This would improve readability, especially for readers trying to replicate the approach.

Author response:

Thank you for this helpful suggestion. We agree that a compact summary would make the pair-set and weighting notation in Eqs. (1)–(5) easier to follow.

We have expanded the existing Table 1 (shown in Table R2-T2 of this reply) to summarize the positive-pair types, their associated weighting terms, and the loss-component weights used during the warm-up, main-training, and fine-tuning stages. The accompanying explanations clarify how the reference, noise-perturbed, spatially translated, and temporally neighboring views correspond to the pair sets defined in the equations.

We have also revised the accompanying text and table notes to make this correspondence explicit. These additions allow readers to connect the mathematical definitions with the training configuration more directly and improve the clarity and reproducibility of the method.

5. Figures 3 and 4: The comparison of WO_CL_UMAP, W_CL_UMAP, and W_CL_PME is a key result. Consider adding a short table summarizing the main quantitative differences (e.g., within-/between-season distances, path-length statistics) alongside the figures, so readers can connect the visual impression to numeric metrics more easily.

Author response:

Thank you for this suggestion. We have expanded the existing Table 2 in Section 2.4 (shown in Table R2-T3 of this reply) to summarize the quantitative seasonal-organization results previously reported in the text. The table compares the within-season distance, between-season distance, and between/within ratio for WO_CL_UMAP, W_CL_UMAP, and W_CL_PME. These distances were calculated after axis-wise z-score standardization of each embedding to reduce the influence of differences in coordinate scale.

We have also revised the accompanying text to connect these statistics more explicitly with the visual patterns in Figs. 3 and 4. The path-length statistics for temporal continuity are retained in the same table alongside the seasonal-organization results. Together, the two parts provide quantitative support for comparing seasonal organization and temporal continuity across the three state spaces.

6. Section 2.4: The discussion of distortion and neighbor distances is insightful. You might consider adding one explicit example region in the state space where distortion is relatively high and briefly explaining how that should influence interpretation of trajectories passing through that region.

Author response:

Thank you for this suggestion. We have added an explicit example and a supporting analysis to clarify how local distortion affects the interpretation of trajectories (Section 2.4; Fig. R2-F4).

Using the local distance-ratio measure underlying Fig. 7(b), we identified states in the upper and lower 20% of the distribution. We then extracted consecutive 6-hour ERA5 transitions whose endpoints both belonged to the same group. To compare transitions with similar apparent movement in the embedding, we matched their 2D displacement lengths one-to-one within a 5% tolerance, without using their 128D distances as a matching criterion.

Among the resulting 549 matched pairs, the transition in the higher-ratio region had a larger 128D distance in 77.8% of cases. The median within-pair ratio of the 128D distances was 1.64. The illustrated example, selected to be closest to this median ratio, has similar 2D displacements of 0.5132 and 0.5326, but corresponding 128D distances of 0.04294 and 0.02617.

These results illustrate that comparable trajectory displacements in 2D can represent different magnitudes of change in the original 128D feature space. We have therefore clarified that trajectory lengths should be interpreted together with the local distortion map. This is a descriptive tendency rather than a universal relationship, and a lower local ratio does not imply distortion-free distance preservation.

7. Section 4: In addition to the aggregate counts and seasonal distributions of good trajectories, consider including a case-study figure that illustrate how trajectory classification links with primary synoptic patterns in the analysis domain. This would concretely link structural trajectory behavior to familiar synoptic features.

Author response:

Thank you for this helpful suggestion. We agree that a concrete case study would make the meteorological meaning of the trajectory classification clearer.

As also described in our response to Major Comment 5, we have added a case-study figure to Section 4 for the forecast initialized on 9 October 2025 (Fig. R2-F5). It presents the reference, good, and bad trajectories alongside the corresponding weather maps at selected lead times. The good trajectory captures the approximate location and evolution of the low-level cyclonic circulation, whereas the bad trajectory is associated with a failure to maintain its distinct structure as lead time increases.

The accompanying explanation links the trajectory classification to these identifiable synoptic features, complementing the aggregate counts and seasonal distributions already presented in Section 4.

8. Conclusions: The paper ends with a well-balanced summary, but adding a short paragraph on potential future work—such as extending to multi-level representations, incorporating additional variables or regions, integrating observational datasets into the state space, or coupling the diagnostics with data-assimilation analyses—would provide a clearer roadmap and emphasize the broader relevance of the framework for Earth-system modeling and AI-based forecasting.

Author response:

Thank you for this constructive suggestion. We have reorganized the Conclusions to incorporate the points raised throughout the review and to provide a clearer outlook for future work.

The revised Conclusions summarize the main diagnostic contributions and clarify the scope of the findings. We have also added a dedicated paragraph outlining extensions to multilevel representations, additional variables and geographical domains, the incorporation of observations as reference data, and potential connections with data-assimilation diagnostics.

These revisions place the present results in a broader research context and provide a more explicit roadmap for developing the framework further. We appreciate the reviewer's suggestion to make these directions clearer to readers.

Figure R2-F1

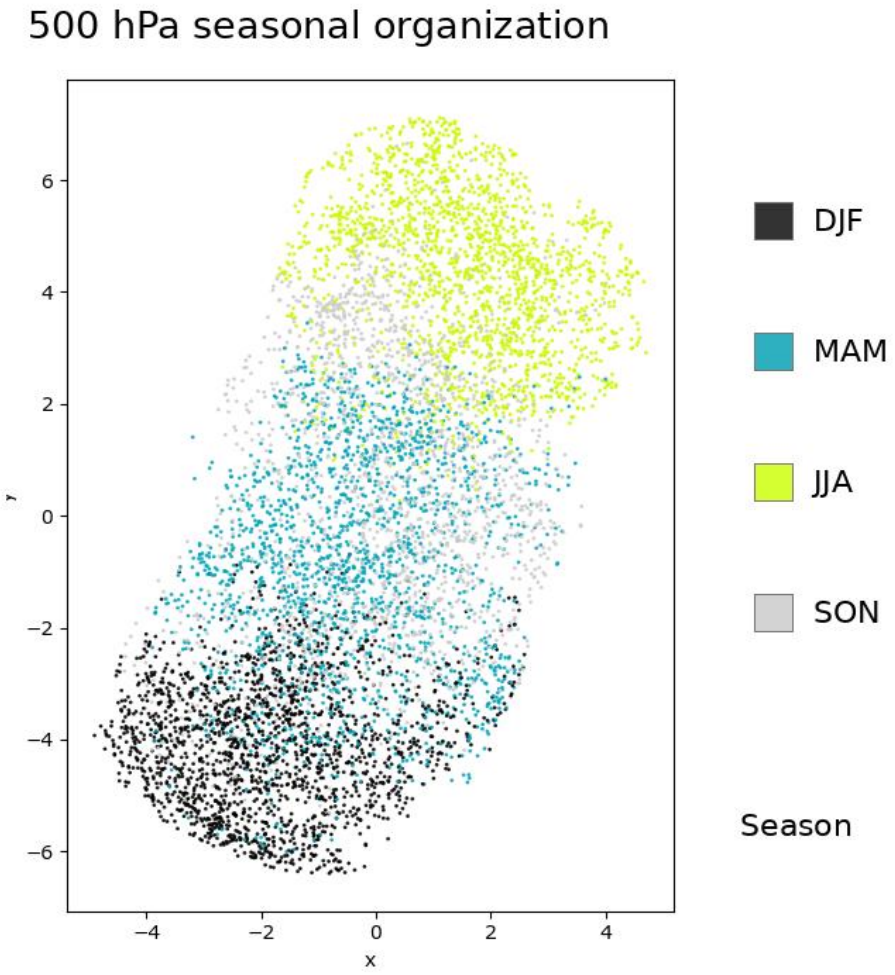



Figure R2-F1. Seasonal organization of the separately trained 500 hPa state space, using temperature, geopotential height, zonal wind, and meridional wind over the East Asian domain. Points represent ERA5 atmospheric states; the season colors follow the convention used in the submitted manuscript. Both coordinates are dimensionless embedding coordinates and have no individual physical units.

Figure R2-F2

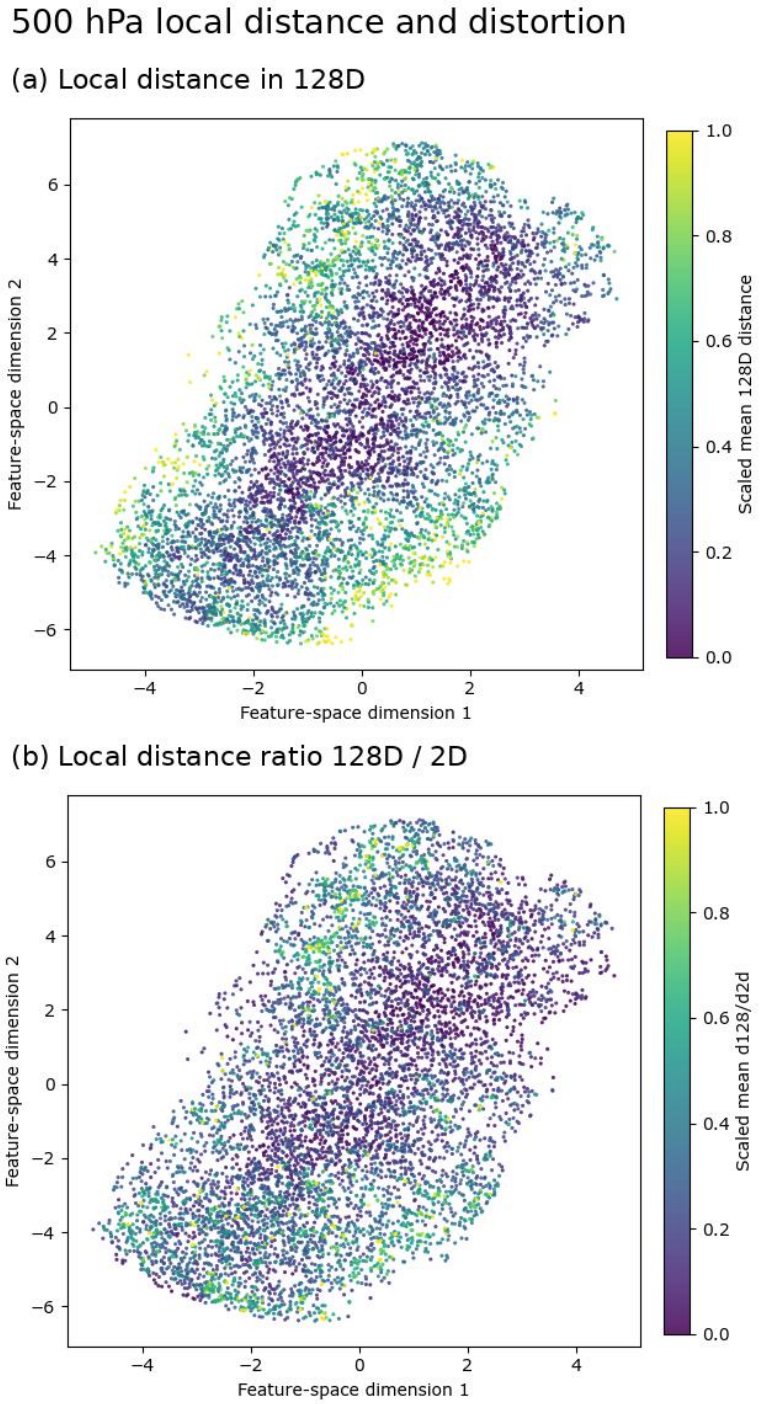



Figure R2-F2. Local diagnostics in the 500 hPa representation. (a) Mean distance to neighboring states in the 128D feature space, displayed at their 2D locations. (b) Local 128D-to-2D distance ratio. Both colorbars show scaled display values from 0 to 1, rather than absolute physical distances. The diagnostics help identify where a given 2D displacement may correspond to a different feature-space change. The two panel maps retain the supplied data and color scales.

Figure R2-F3

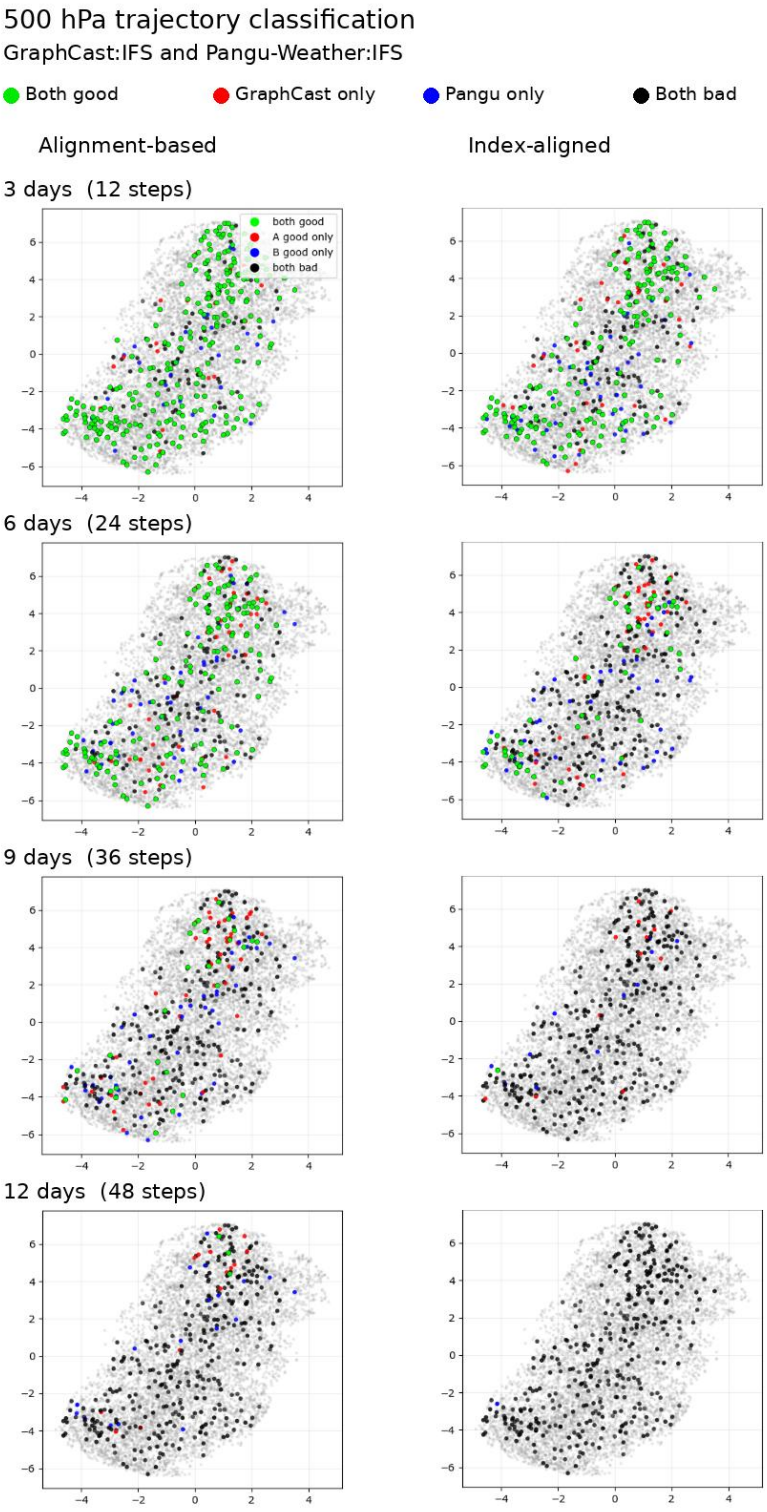



Figure R2-F3. Joint trajectory classifications for GraphCast:IFS and Pangu-Weather:IFS at 500 hPa. Rows show 3-, 6-, 9-, and 12-day windows (12, 24, 36, and 48 six-hour steps); columns show alignment-based and index-aligned distances. Green: both good; red: GraphCast:IFS only good; blue: Pangu-Weather:IFS only good; black: both bad. The gray background shows the training-state distribution. The displayed source classification uses threshold 0.4666. Embedding coordinates are dimensionless.

Table R2-T1

Table R2-T1. Seasonal percentages of joint good/bad classifications for GraphCast:IFS and Pangu-Weather:IFS at 500 and 850 hPa. GraphCast only and Pangu only indicate that only the named configuration is classified as good. Percentages are reproduced from the supplied seasonal summary; small deviations of row totals from 100% reflect rounding.

| Level | Season | Both good | GraphCast-only | Pangu-Weather-only | All bad |
|---|---|---|---|---|---|
| 500 hPa | DJF | 34.7% | 6.7% | 10.3% | 48.3% |
| | MAM | 20.9% | 4.9% | 9.2% | 64.9% |
| | JJA | 37.0% | 13.0% | 9.2% | 40.8% |
| | SON | 24.5% | 6.9% | 6.9% | 61.8% |
| 850 hPa | DJF | 48.1% | 12.2% | 6.1% | 33.6% |
| | MAM | 32.3% | 7.6% | 13.6% | 46.5% |
| | JJA | 24.2% | 9.5% | 17.1% | 49.2% |
| | SON | 31.6% | 7.7% | 14.3% | 46.4% |

DJF: December–February; MAM: March–May; JJA: June–August; SON: September–November.

Figure R2-F4

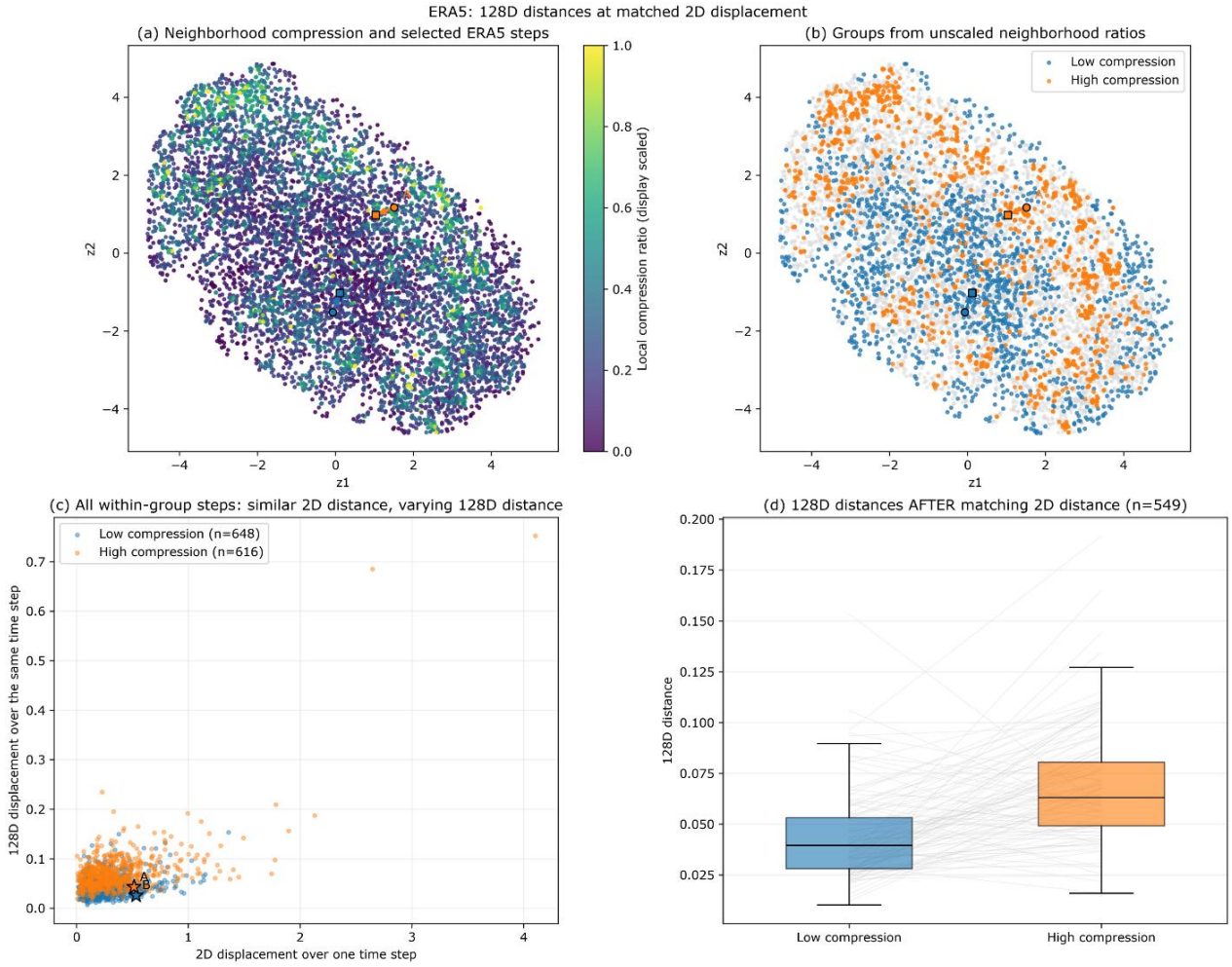



Figure R2-F4. Local distortion and matched 6-hour ERA5 transitions in the 850 hPa representation. (a) Scaled local 128D-to-2D distance ratio and representative transitions; (b) upper- and lower-fifth groups defined using the unscaled ratios; (c) 2D and 128D displacements for transitions whose endpoints both lie in the same group; (d) 128D distances after matching 2D displacement one-to-one within 5% (549 pairs). The higher-ratio transition has a larger 128D distance in 77.8% of matched pairs, with a median within-pair distance ratio of 1.64. The representative pair has 2D displacements 0.5132 and 0.5326 and 128D distances 0.04294 and 0.02617. All distances are dimensionless. Lower ratio does not imply absence of distortion.

Figure R2-F5

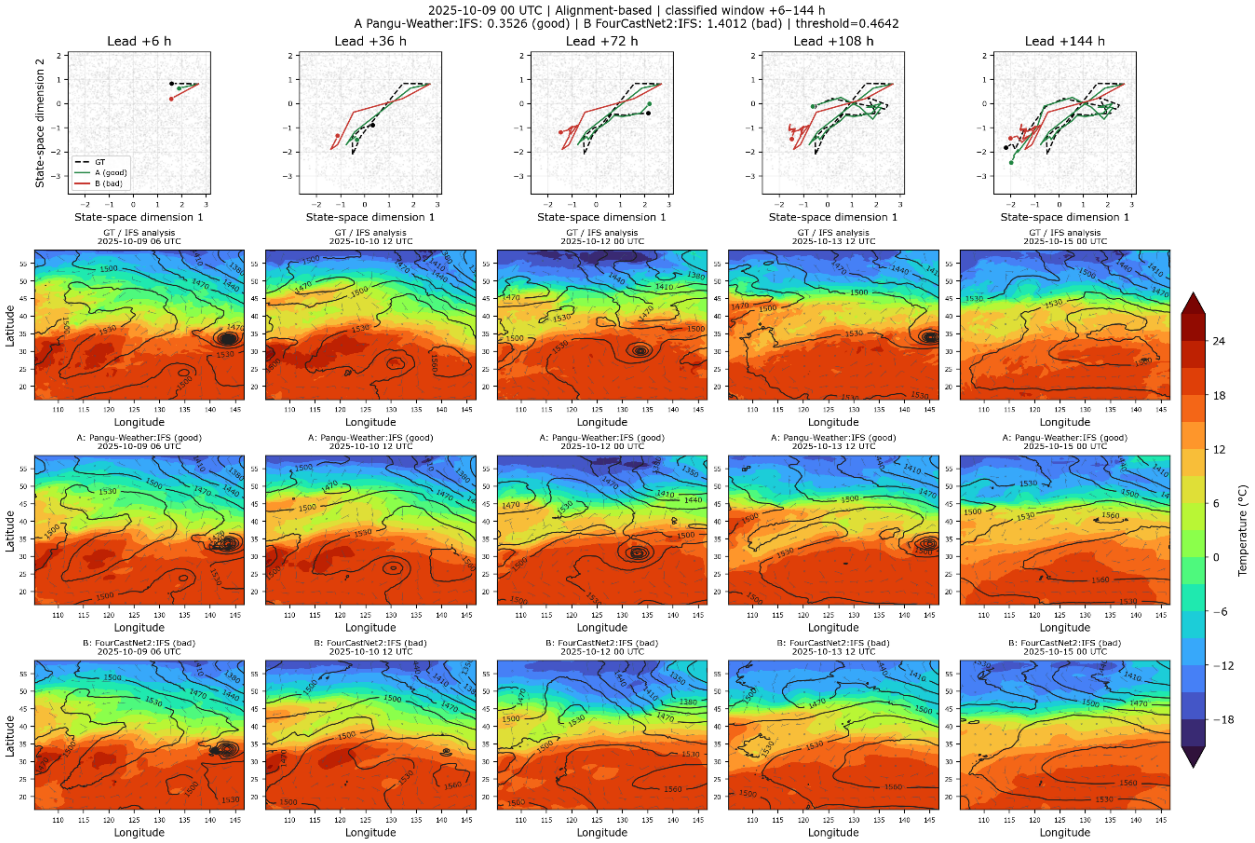



Figure R2-F5. Trajectory–weather-map case initialized at 00 UTC on 9 October 2025. The top row shows evolving reference and forecast trajectories at +6, +36, +72, +108, and +144 h. The remaining rows show IFS analysis, Pangu-Weather:IFS, and FourCastNet2:IFS. Shading indicates 850 hPa temperature (°C), contours geopotential height, and arrows horizontal wind. Good and bad refer to alignment-based classification over the +6–144 h window, rather than to independent verification of every weather feature.

Figure R2-F6

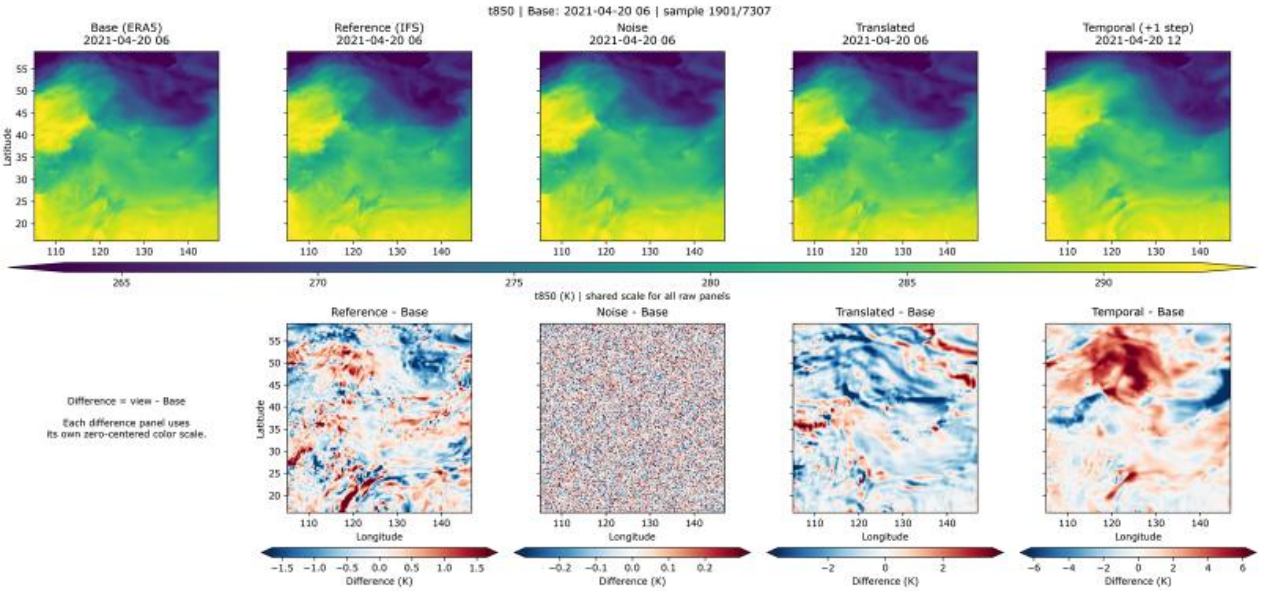



Figure R2-F6. Example positive views for ERA5 temperature at 850 hPa, valid at 06 UTC on 20 April 2021. Top: base ERA5, same-time IFS analysis, noise-perturbed base, translated base, and ERA5 at t + 6 h. The common raw-field colorbar is in K. Bottom: each view minus the base, with a separate zero-centered K scale for each difference map; difference magnitudes must therefore be read from their individual colorbars.

Table R2-T2

Table R2-T2. Training configuration to be incorporated into manuscript Table 1. The values reproduce the supplied configuration. Temporal neighborhood k = 2 corresponds to offsets of 6 and 12 h in the current sampling.

| Stage | Warm-up | Main | Fine-tuning |
|---|---|---|---|
| **Steps** | 200 | 400 | 200 |
| **Learning rate** | $1.00 \times 10^{-3}$ | $1.00 \times 10^{-3}$ | $3.00 \times 10^{-4}$ |
| **Temperature** | 0.1 | 0.1 | 0.1 |
| **Temporal positives** | No | Yes, $(k = 2)$ | Yes, $(k = 2)$ |
| **Reference as positive** | No | Yes | Yes |
| **Augmentation positives** | Included | Included | Excluded |
| **paired view weights** | Base: 1.0<br>Reference: 1.0<br>Temporal: 1.0 | Base: 0.0<br>Reference: 1.0<br>Temporal: 1.0 | Base: 0.0<br>Reference: 1.0<br>Temporal: 1.0 |
| **Family weights** | appearance: 1.0<br>phase: 0.6<br>scale: 0.3 | appearance: 0.2<br>phase: 0.6<br>scale: 0.2 | appearance: 0.0<br>phase: 0.0<br>scale: 0.0 |
| **Loss weights** | $\lambda_{\text{sup}} = 1.0$<br>$\lambda_{\text{ref}} = 0.0$<br>$\lambda_{\text{temp}} = 0.0$ | $\lambda_{\text{sup}} = 1.0$<br>$\lambda_{\text{ref}} = 0.5$<br>$\lambda_{\text{temp}} = 0.4$ | $\lambda_{\text{sup}} = 0.0$<br>$\lambda_{\text{ref}} = 0.3$<br>$\lambda_{\text{temp}} = 0.7$ |

View weights contribute to the weighted contrastive objective in Eq. (1); the λ coefficients scale the three objective terms in Eq. (5). Reference alignment compares ERA5 with IFS analysis at the same valid time. Temporal alignment compares neighboring ERA5 states. Appearance and phase families include noise perturbations and spatial translations, respectively. “Inactive” means the view is not used as a positive in that stage. An em dash retains the unspecified base-view entry in the supplied configuration.

Table R2-T3

Table R2-T3. Combined quantitative summary proposed for the existing manuscript Table 2. Coordinates in each embedding are standardized axis by axis before distance calculation. (a) Seasonal organization; (b) temporal continuity measured by path length over 8, 12, and 16 consecutive six-hour states. All distances are dimensionless.

**(a) Seasonal organization**

| State space | Within-season distance | Between-season distance | Between/within ratio |
|---|---|---|---|
| WO_CL_UMAP | 1.4287 | 1.8676 | 1.3072 |
| W_CL_UMAP | 1.2316 | 1.9962 | 1.6207 |
| W_CL_PME | 1.0377 | 2.0266 | 1.9530 |

**(b) Temporal continuity: Trail path-length statistics**

| 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** |

WO_CL_UMAP: without contrastive learning with UMAP; W_CL_UMAP: with contrastive learning with UMAP; W_CL_PME: with contrastive learning with parametric manifold embedding. Lower mean path length alone does not establish better temporal continuity; spread and extreme displacements also matter.