**Review 1 of Topology-Preserving State Space Representation for Diagnosing Weather Forecast Models by Kim and Cho**

## Overall comment

This paper proposes a trajectory-based diagnostic framework for evaluating AI-based medium-range weather forecast models in a learned two-dimensional “topology-preserving” state space. The proposed trajectory-based state-space method seems useful as an additional diagnostic lens for AI weather forecasts. It helps show whether forecasts evolve along realistic atmospheric paths, whether errors are structural or timing-related, and whether model performance depends on season, atmospheric state, or initial condition. While I think the message is interesting, many of the figures are not of suitable quality for publication. I am also concerned about whether the results are generalizable when they are based only on a single atmospheric level (850 hPa), a specific region of the globe, and a learned 2D embedding whose physical meaning and distance properties require stronger justification. On that basis, I recommend further analysis and major revisions in order to bring this manuscript up to a suitable standard.

Author response:

We thank the reviewer for the careful and constructive assessment of our manuscript and for recognizing the potential of the proposed framework as a complementary diagnostic tool. We agree that the generalizability of the findings, the interpretation of the learned state space, and the clarity of the figures require stronger support and explanation.

To address these concerns, we have conducted additional experiments at 500 hPa, prepared weather-map case studies linking trajectory behavior to meteorological structures, and examined the relationship between trajectory-based comparisons and RMSE through quantitative analyses and representative cases. This includes an analysis using 128-dimensional features to reduce the influence of two-dimensional projection distortion. We have also improved the figures and captions and clarified the interpretation and limitations of the learned representation, including the scope of conclusions drawn from the tested region and pressure levels.

Detailed responses to each comment are provided below. These additional analyses and presentation improvements will be incorporated into the revised manuscript.

## Major comments

1. The framework is demonstrated only for four variables at 850 hPa over an East Asian domain. This raises questions about how general the conclusions are. It is unclear whether the learned state-space structure, trajectory behavior, and model rankings would remain similar at other pressure levels, for vertically integrated/multilevel states, or over other regions. While the authors acknowledge this limitation, I do not think there is sufficient justification for why 850 hPa was chosen. If possible, a sensitivity test using at least one additional level or a multilevel input would strengthen the manuscript substantially.

Author response:

We thank the reviewer for this suggestion. We selected 850 hPa as an initial testbed for characterizing lower-tropospheric thermal structure and circulation using temperature, geopotential height, and horizontal wind components, which are consistently available across the datasets.

To assess sensitivity to pressure level, we additionally constructed a separate state space using the same four variables at 500 hPa over the same East Asian domain and evaluated the nine forecast configurations. The seasonal organization, local-distance and distortion diagnostics, and joint trajectory classifications are shown in Figs. R1-F1, R1-F2, and R1-F3, respectively. The 500 hPa representation also exhibits seasonal organization, but the relative model performance varies with pressure level. At 850 hPa, GraphCast:IFS produces more good trajectories than Pangu-Weather:IFS in winter, whereas Pangu-Weather:IFS produces more in summer. These relative advantages reverse at 500 hPa (Table R1-T1).

These results support applying the framework beyond 850 hPa while showing that the resulting model rankings depend on pressure level and season. We will include this sensitivity analysis in the revised manuscript and clarify both the rationale for selecting 850 hPa and the scope of the conclusions. Generalization to multilevel inputs and other regions remains to be evaluated.

2. The trajectory examples show that the method can distinguish forecast behavior, but it is still difficult to connect the state-space trajectories to actual meteorological structures. A case study showing the corresponding weather maps at selected points along the trajectory would help readers understand what it means physically when a forecast follows, lags, or diverges from the observed trajectory. This would make the diagnostic value of the method much more convincing.

Author response:

We thank the reviewer for this suggestion. We have prepared a case-study figure linking state-space trajectories directly to the corresponding weather maps (Fig. R1-F4). For forecasts initialized on 9 October 2025, the figure compares IFS analysis with Pangu-Weather:IFS and FourCastNet2:IFS, classified as good and bad, respectively, using the alignment-based distance over the +6–144 h window. A single figure combines the evolving trajectories with 850 hPa temperature, geopotential-height contours, and wind vectors at five selected lead times.

The maps show that Pangu-Weather:IFS largely reproduces the closed low and its eastward progression visible in the reference at +72 and +108 h, whereas FourCastNet2:IFS fails to retain this feature. These differences provide a concrete meteorological interpretation of the contrasting trajectory classifications.

We will include this figure and the accompanying discussion in the revised manuscript to clarify how trajectory agreement and divergence relate to the evolution of meteorological structures.

3. The manuscript argues that the trajectory method provides information not captured by RMSE, but this point is not quantified clearly enough. The authors should explicitly show where the trajectory approach agrees with RMSE and where it disagrees. For example, are there cases where RMSE suggests similar performance but the trajectory method clearly separates the models? Or are there cases where RMSE is poor but the trajectory remains structurally close to the observed evolution? A more direct comparison would make the value of the proposed method much clearer.

Author response:

We thank the reviewer for this comment. We conducted an additional quantitative comparison between trajectory distance and RMSE and selected two illustrative cases. For this analysis, trajectory distances and temporal alignments were computed in the 128-dimensional feature space to avoid additional distortion introduced by the 2D projection. The 2D representation was retained

for visualization.

The scatter plots in Fig. R1-F5 compare reductions in trajectory distance with reductions in endpoint standardized multivariate RMSE for 3-, 6-, 9-, and 12-day windows. The corresponding correlation coefficients (ρ) are −0.59, −0.56, −0.23, and 0.03, respectively. These results show that reducing trajectory distance does not consistently reduce RMSE. Because alignment is determined from feature-space distances rather than by minimizing RMSE, improvements in the two measures need not coincide.

The selected cases illustrate (i) two forecasts with similar RMSE but different trajectory distances (Case A; Fig. R1-F6) and (ii) a forecast with relatively large same-index RMSE but comparatively favorable alignment-based trajectory agreement (Case B; diagnostic curves in Fig. R1-F7 and weather maps in Fig. R1-F8). The accompanying weather-map snapshots in Figs. R1-F6 and R1-F8 compare reference fields with forecasts at the original and aligned lead times, linking these contrasting assessments to meteorological patterns and timing differences.

We will include these analyses and figures in the revised manuscript and clarify when RMSE and trajectory diagnostics provide consistent or different assessments, thereby making their complementary roles more explicit.

4. Several figures are difficult to interpret in their current form. Axis labels, legends, colorbars, and font sizes need improvement throughout. Some figures, especially those showing the learned state space, are central to the paper but are not self-explanatory. The authors should ensure that each figure can be understood without excessive reference to the text. In particular, the meaning of the embedding axes, color scales, trajectory markers, and model labels should be made clearer.

Author response:

We thank the reviewer for this important comment. We agree that clear, self-contained figures are essential because visual interpretation of the learned state space and forecast trajectories is central to the proposed framework.

We have revised the figures and captions to improve readability and clarify what is being displayed. Figure 2 (shown as Fig. R1-F10 in this reply) now shows actual ERA5 fields with colorbars and physical units. Figure 3 (Fig. R1-F11 in this reply) uses darker points, and its caption explains that the axes represent dimensionless embedding coordinates, with axis ranges set independently for each representation. Figures 8 and 11 (Figs. R1-F12 and R1-F13 in this reply, respectively), together with other related figures, have been improved through larger text and legend labels and consistent model colors. The RMSE units are now explicitly indicated, and the captions clarify the common evaluation domain and the 850 hPa pressure level.

We have also clarified the meanings of colors, trajectory markers, and model labels in the relevant captions so that the figures can be understood without excessive reference to the main text. Further details are provided in our responses to Specific Comments 3, 4, and 6. These figure and caption improvements will be incorporated into the revised manuscript.

**Other comments**

1. L28: AIFS should be defined at first use.

Author response:

Thank you for pointing this out. We will spell out the full name as “Artificial Intelligence Forecasting System (AIFS)” at its first occurrence in the Introduction.

2. L334–349: It would be useful to show a conceptual diagram of the index-aligned and alignment-based distance metrics as they relate to two example trajectories. At present, the distinction between the two metrics is described in the text, but it may not be immediately clear to readers how the distances are calculated or why the alignment-based distance can produce a different interpretation. A simple schematic showing two trajectories with corresponding same-time-point comparisons for the index-aligned metric, and a warped/shifted matching path for the alignment-based metric, would help clarify the difference. This would also make it easier to understand the later interpretation of Figure 9, where the authors argue that some forecast errors may reflect timing offsets rather than completely different trajectory evolution.

Author response:

We thank the reviewer for this helpful suggestion. We agree that a schematic would make the distinction between the two metrics easier for readers to understand.

We have prepared a schematic illustrating two trajectories with similar shapes but different temporal progressions (Fig. R1-F9). The figure contrasts same-index matching with order-preserving DTW matching, with dashed lines indicating the point pairs used to calculate each distance. It shows how timing offsets can increase the index-aligned distance even when the forecast follows a similar state-space path. We will include this schematic and an explanatory caption in Section 3.2 to clarify the two metrics and support the interpretation of Figure 9.

3. Figure 2: The meaning of the colors for zonal and meridional wind is unclear, and they do not seem to correspond directly to the vectors shown. Please add colorbars and clarify what is being plotted.

Author response:

We thank the reviewer for pointing this out. We have revised Figure 2 (shown in Fig. R1-F10 of this reply) to display actual ERA5 fields for all four variables at 850 hPa and added horizontal colorbars with physical units. The u and v panels now show the signed zonal and meridional wind components as color-shaded fields, without vector arrows. We will also clarify these color meanings in the caption so that the figure can be understood independently of the main text.

4. Figure 3: What do the axis labels Z1 and Z2 represent? Why are the axis ranges different across the three panels? If these are arbitrary embedding coordinates, this should be stated clearly.

Author response:

We thank the reviewer for pointing this out. We have revised Figure 3 (shown in Fig. R1-F11 of this

reply) using darker, single-color points and removed the redundant legend. The revised caption explains that the horizontal and vertical axes represent dimensionless embedding coordinates, rather than individual meteorological variables. It also clarifies that axis limits are set independently for each representation and that coordinate values and absolute distances are not directly comparable across panels. The coordinate meanings are explained in the caption to avoid repeating lengthy axis titles.

5. Temporal sampling: Why are 6-hourly states used? Were shorter time intervals considered? The authors should discuss whether important transitions or trajectory behavior could be missed at 6-hour resolution.

Author response:

We thank the reviewer for raising this point. We used 6-hourly ERA5 states to match the temporal resolution of the forecast datasets evaluated in this study and to examine large-scale atmospheric evolution over forecast windows of up to 12 days.

Although each state is represented by its multivariate spatial fields, temporal sampling also influences the density of sampled states and the temporal relationships used during training. We acknowledge that 6-hourly sampling may miss rapid transitions occurring between successive states. Finer sampling could resolve additional intermediate evolution, but we have not tested 1- or 2-hourly sampling and therefore cannot assume that the learned representation would remain unchanged.

The framework can accommodate finer temporal sampling. In the present configuration, temporal positive pairs at separations of 6 and 12 hours encourage continuity in the learned representation. With finer sampling, these relationships should be specified in physical time, with the sample offsets and temporal-loss weighting reassessed accordingly. We will clarify the rationale for the 6-hour interval and acknowledge its limitations for diagnosing rapid transitions in the revised manuscript.

6. Figure 8: The legend labels and text are too small (this is true for many of the figures). The RMSE panels should also clearly state the units and whether the RMSE is computed over the same regional domain and 850 hPa level.

Author response:

We thank the reviewer for highlighting these issues. We agree that the small text and legend labels reduced the readability of Figure 8 and that the RMSE panels required clearer information on units and evaluation settings. We have therefore improved Figure 8, the related Figure 11 (shown in Figs. R1-F12 and R1-F13 of this reply, respectively), and other figures by enlarging text and legend labels and improving the clarity and consistency of their presentation. The RMSE units are now explicitly indicated, and the captions clarify that RMSE is computed against ECMWF analysis over the same East Asian domain at 850 hPa used for the trajectory analysis. The captions have also been revised to make the figures easier to understand independently of the main text. These improvements will be incorporated into the revised manuscript.

7. RMSE comparison: The manuscript should more clearly identify when RMSE and the trajectory method give consistent conclusions and when they differ. This is central to the

claim that the new method provides complementary information.

Author response:

We agree that the complementary roles of the two approaches should be stated more clearly. RMSE quantifies pointwise error magnitudes in the meteorological fields, whereas trajectory diagnostics assess multivariate evolution and temporal correspondence in the learned representation.

The additional analysis and cases described in our response to Major Comment 3 (Figs. R1-F5–R1-F8) illustrate that similar RMSE can accompany different trajectory behavior, and that relatively large RMSE can coexist with close agreement after temporal alignment. We will make this complementary interpretation explicit in the revised manuscript.

Figure R1-F1

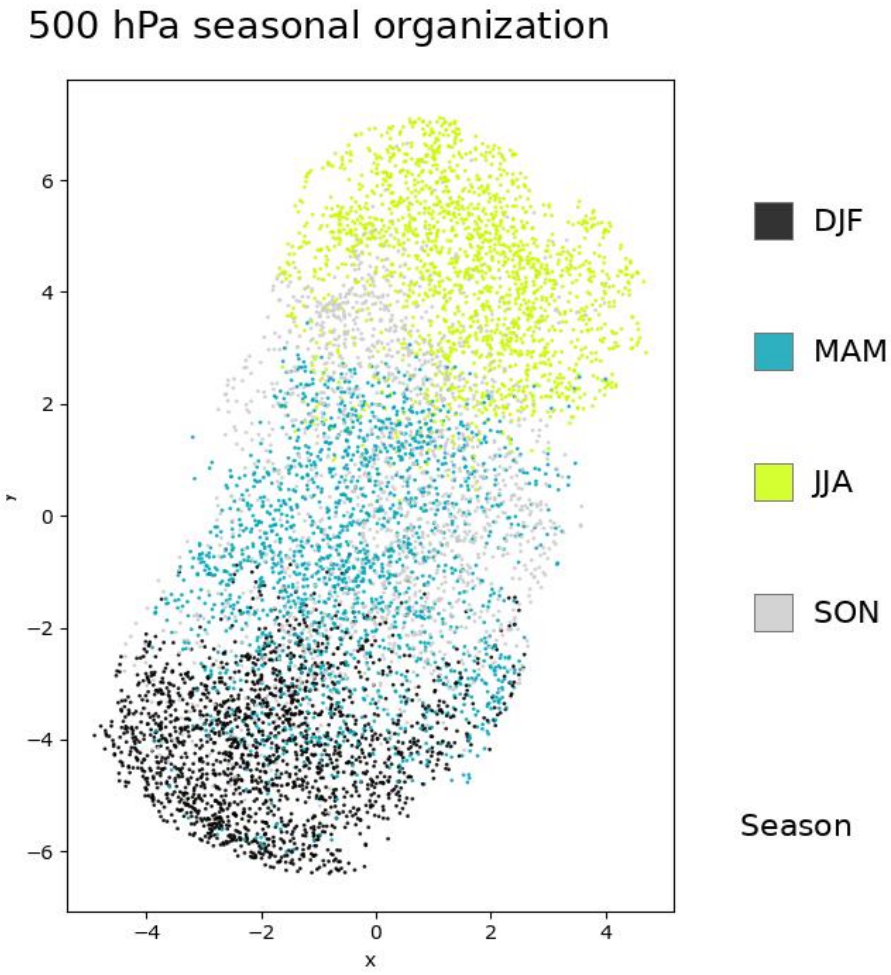



Figure R1-F1. Seasonal organization of the separately trained 500 hPa state space, using temperature, geopotential height, zonal wind, and meridional wind over the East Asian domain. Points represent ERA5 atmospheric states; the season colors follow the convention used in the submitted manuscript. Both coordinates are dimensionless embedding coordinates and have no individual physical units.

Figure R1-F2

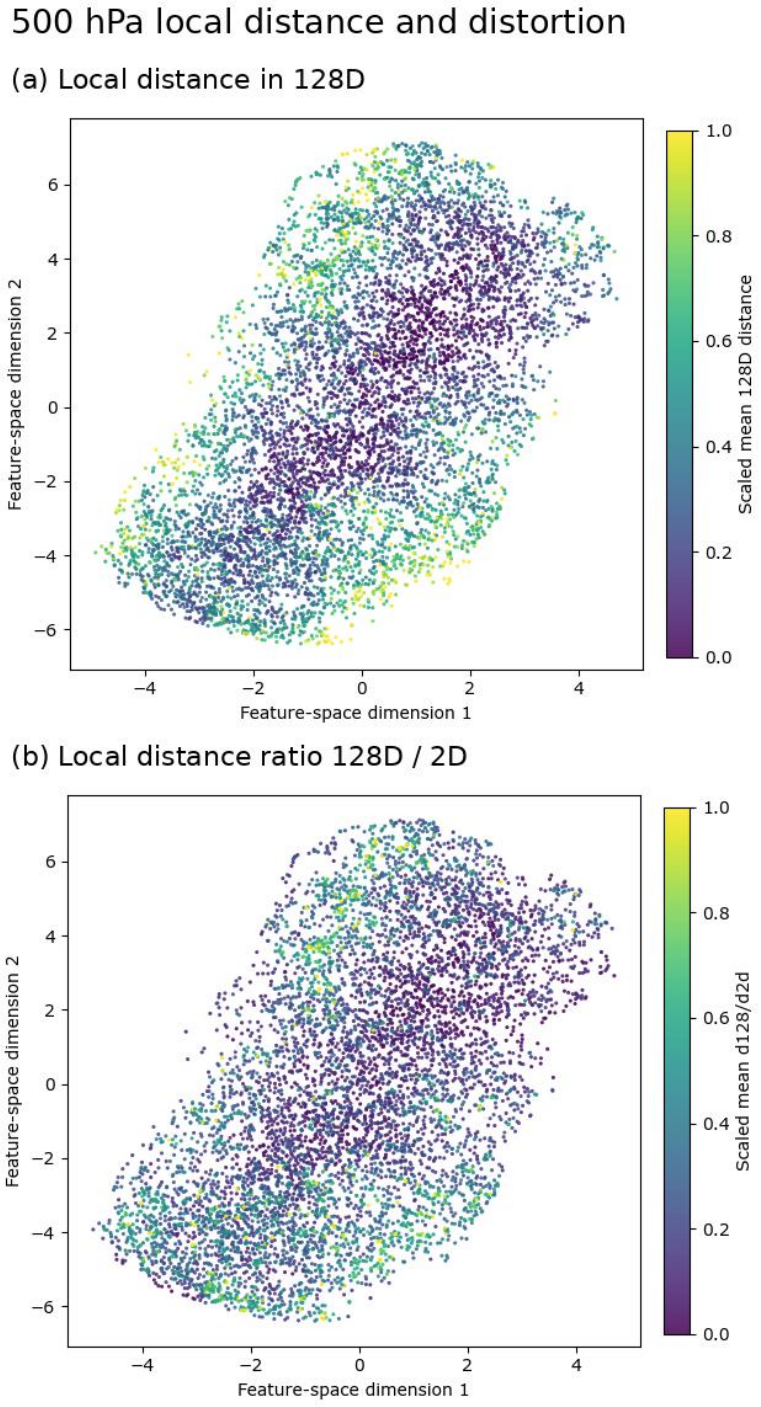



Figure R1-F2. Local diagnostics in the 500 hPa representation. (a) Mean distance to neighboring states in the 128D feature space, displayed at their 2D locations. (b) Local 128D-to-2D distance ratio. Both colorbars show scaled display values from 0 to 1, rather than absolute physical distances. The diagnostics 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 R1-F3

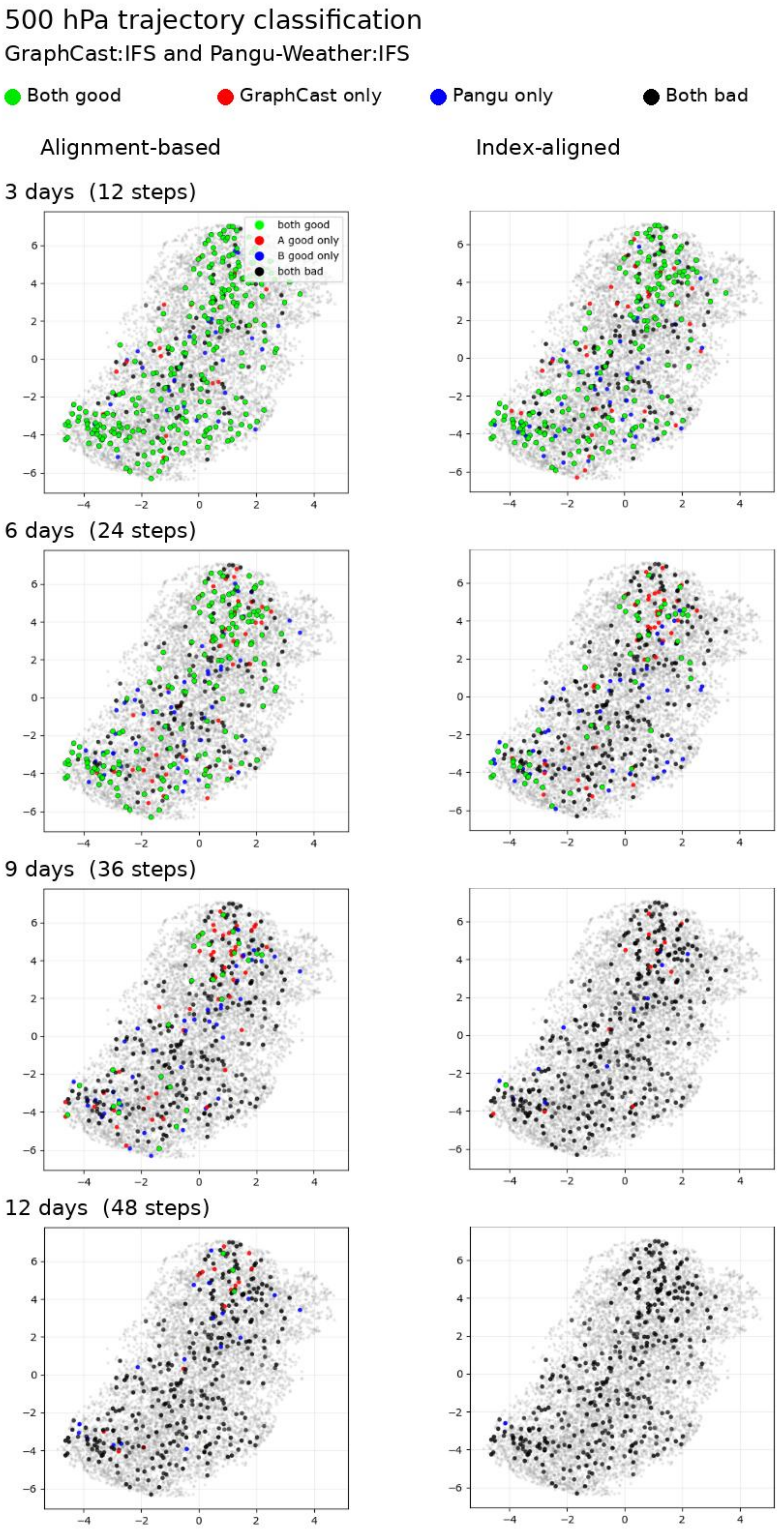



Figure R1-F3. Joint trajectory classifications for GraphCast:IFS and Pangu-Weather:IFS at 500 hPa. Rows show 3-, 6-, 9-, and 12-day windows (12, 24, 36, and 48 six-hour steps); columns show alignment-based and index-aligned distances. Green: both good; red: GraphCast:IFS only good; blue: Pangu-Weather:IFS 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 R1-T1

Table R1-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 (hPa) | Season | Both good (%) | GraphCast only (%) | Pangu only (%) | Both bad (%) |
|---|---|---|---|---|---|
| 500 | 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 | 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 R1-F4

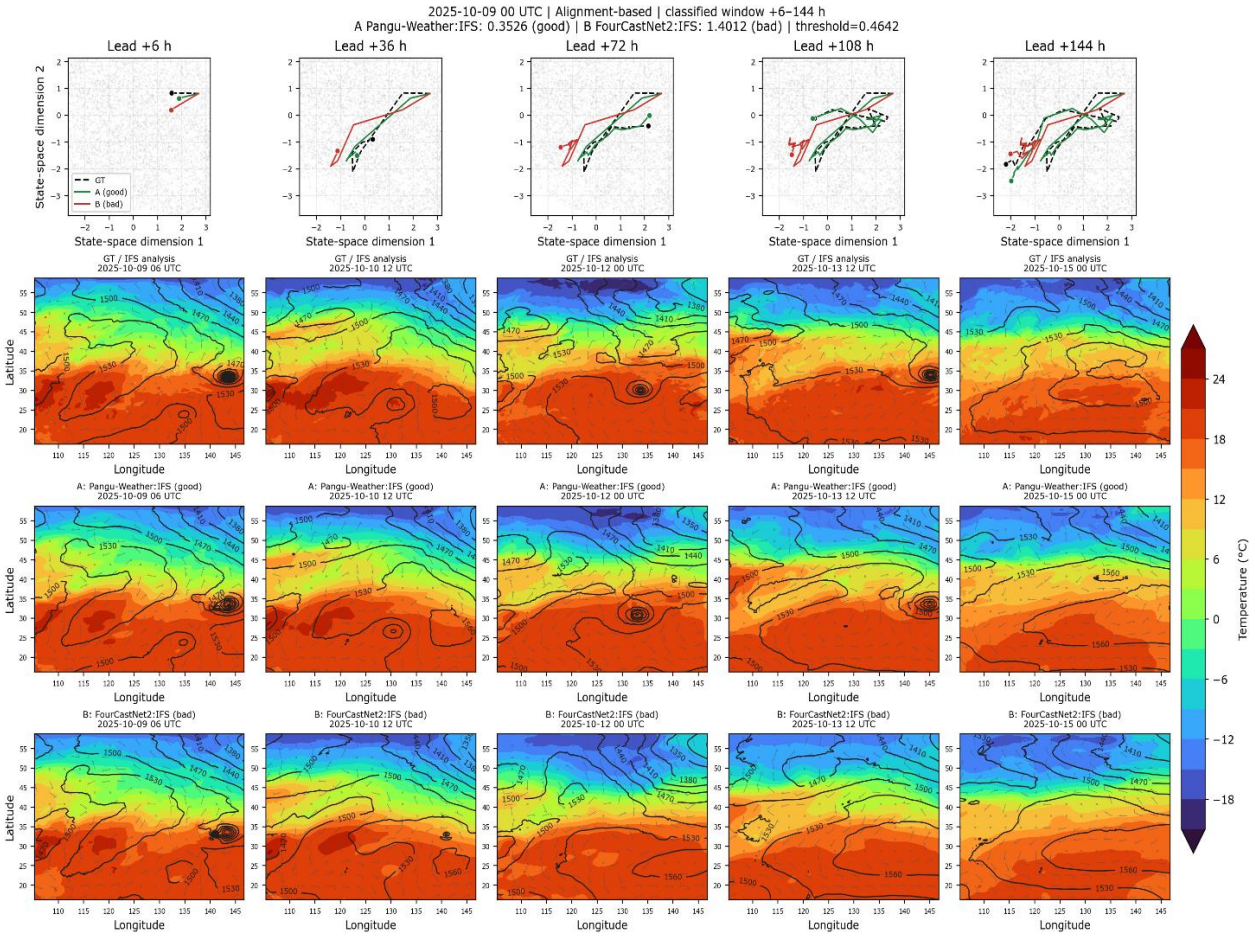



Figure R1-F4. Trajectory–weather-map case initialized at 00 UTC on 9 October 2025. The top row shows evolving reference and forecast trajectories at +6, +36, +72, +108, and +144 h. The remaining rows show IFS analysis, Pangu-Weather:IFS, and FourCastNet2:IFS. Shading indicates 850 hPa temperature (°C), 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 R1-F5

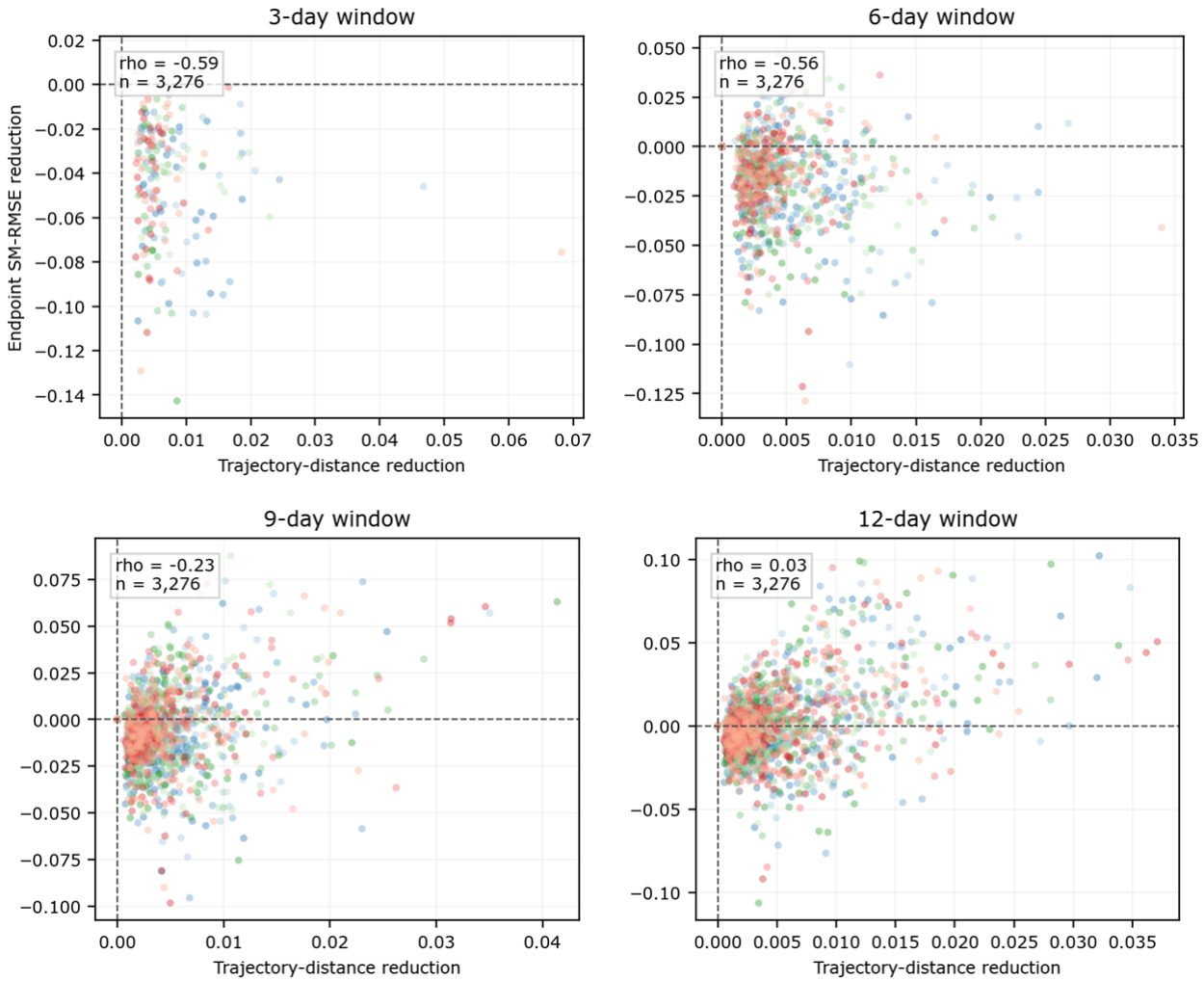



Figure R1-F5. Effect of 128D temporal alignment on standardized multivariate endpoint RMSE. Panels show 3-, 6-, 9-, and 12-day windows in reading order. The horizontal axis is the reduction from index-aligned to alignment-based trajectory distance; the vertical axis is the reduction in endpoint standardized multivariate RMSE using the aligned forecast and the same reference endpoint. Positive vertical values indicate lower RMSE, and negative values higher RMSE. The displayed correlations are −0.59, −0.56, −0.23, and 0.03, respectively; each panel contains 3276 pairs. Distances and RMSE are dimensionless. Colors distinguish the pooled forecast configurations.

Figure R1-F6

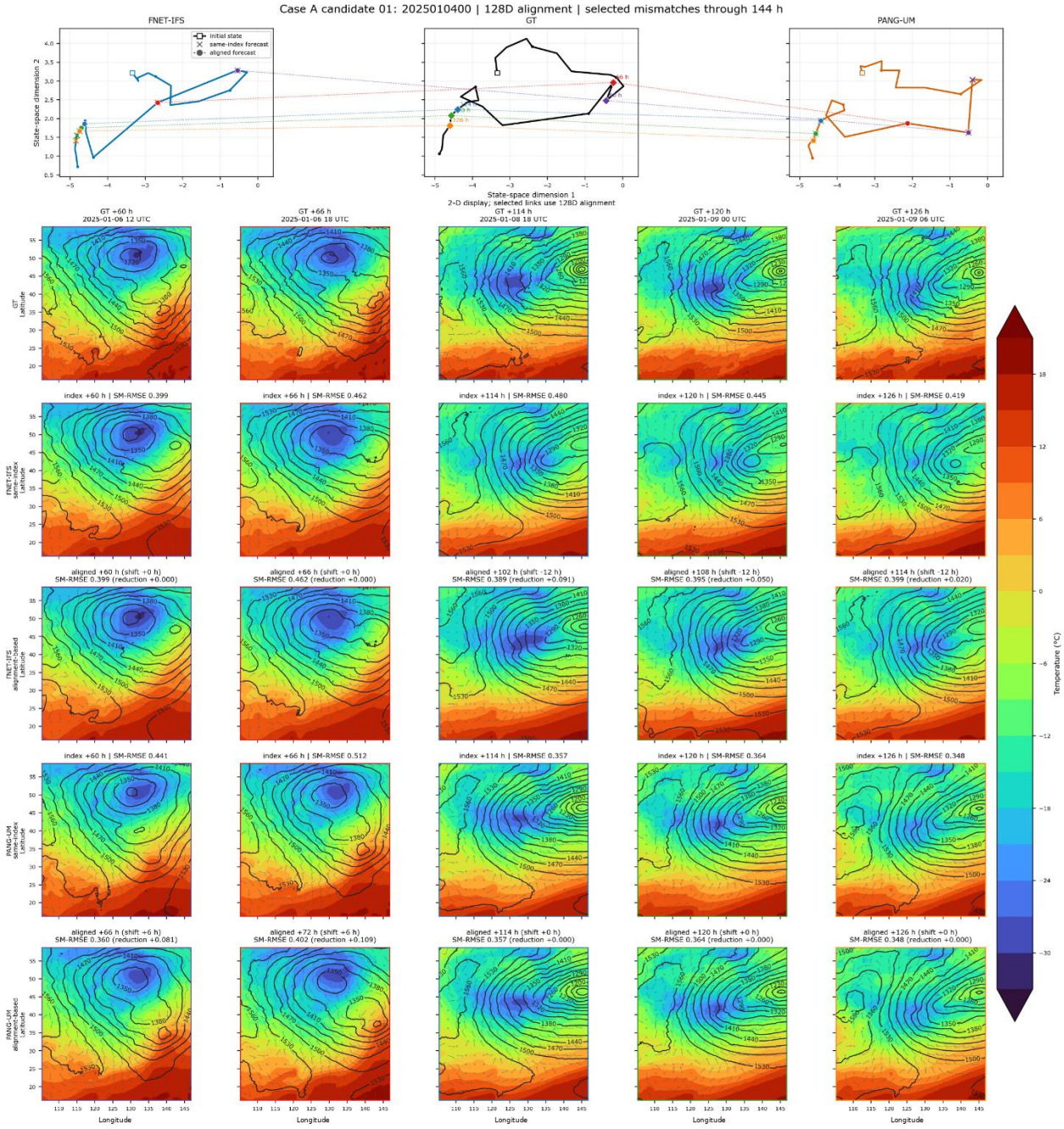



Figure R1-F6. Case A, initialized at 00 UTC on 4 January 2025, comparing FourCastNet2:IFS (FNET-IFS) with Pangu-Weather:UM (PANG-UM) through +144 h. The top panels show 2D trajectories; correspondences were selected in 128D. The map rows show the IFS-analysis reference, same-index and aligned FourCastNet2:IFS, and same-index and aligned Pangu-Weather:UM. Each map column uses the same reference valid time; aligned forecast lead times are printed above the maps. Shading is 850 hPa temperature (°C), contours geopotential height, and arrows wind. SM-RMSE denotes standardized multivariate RMSE.

Figure R1-F7

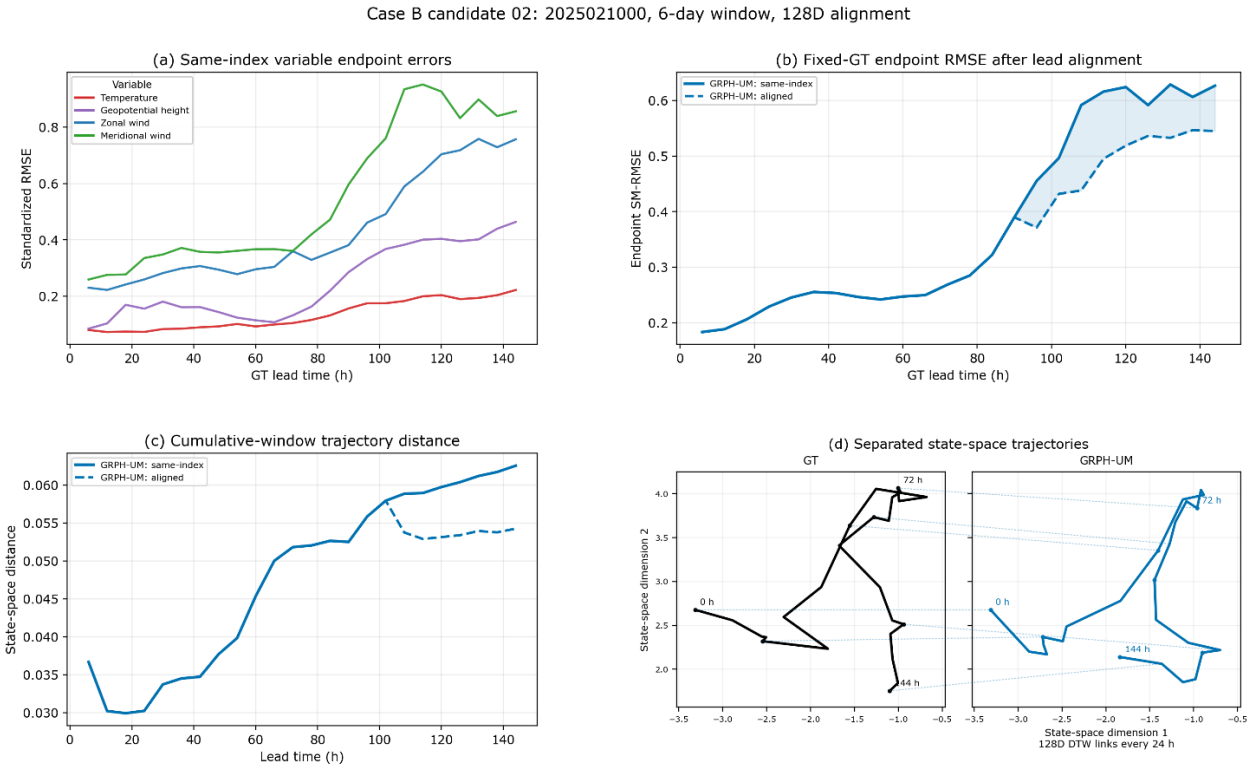



Figure R1-F7. Case B, GraphCast:UM (GRPH-UM), initialized at 00 UTC on 10 February 2025, for the 6-day window. (a) Standardized variable-wise same-index endpoint RMSE; (b) standardized multivariate endpoint RMSE with and without 128D alignment; (c) cumulative-window trajectory distance; (d) 2D display of the trajectories with selected 128D correspondences. The same reference time is used in each RMSE comparison. A lower aligned trajectory distance is not by itself evidence of improved forecast accuracy.

Figure R1-F8

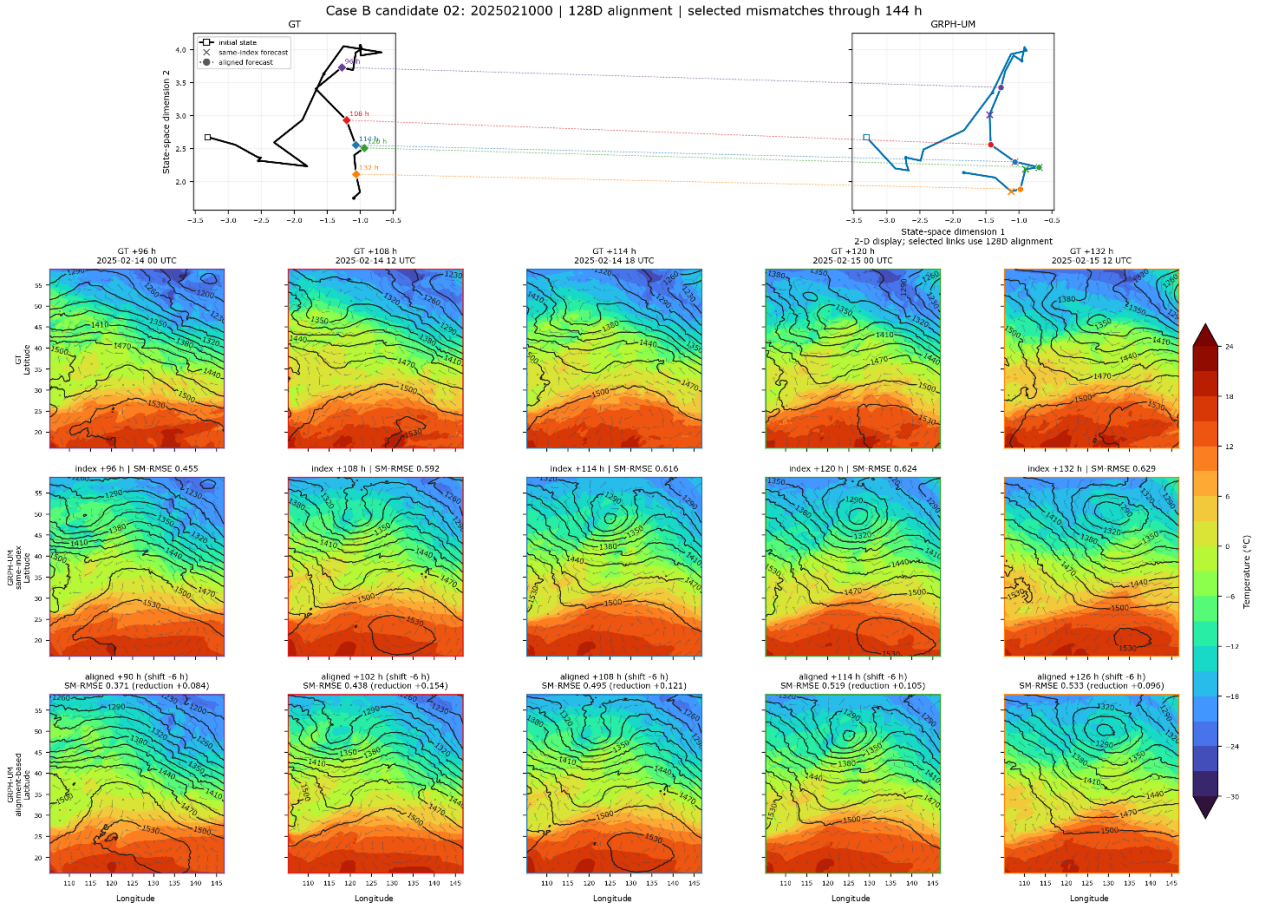



Figure R1-F8. Weather maps for Case B at selected reference leads from +96 to +132 h. Rows show the IFS-analysis reference, GraphCast:UM at the same lead, and GraphCast:UM at the lead selected by the 128D alignment. The selected leads in the last row are 6 h earlier than the reference leads in these examples. Shading is 850 hPa temperature (°C), contours geopotential height, and arrows wind. The corresponding standardized multivariate RMSE values appear above the forecast panels. The upper trajectories are a 2D display of the 128D correspondences.

Figure R1-F9

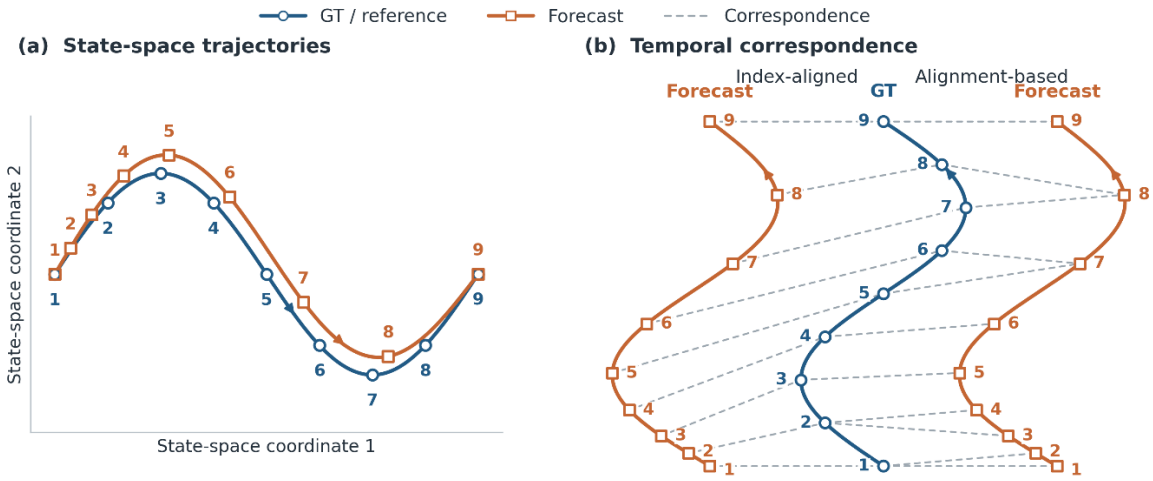



Figure R1-F9. Conceptual comparison of index-aligned and alignment-based distances. (a) The reference and forecast follow similar paths with different temporal progression. (b) Dashed correspondences contrast fixed-index matching with monotonic temporal alignment. Numbers identify successive time indices; the schematic explains the matching principle and is not a quantitative result. State-space coordinates are dimensionless.

Figure R1-F10

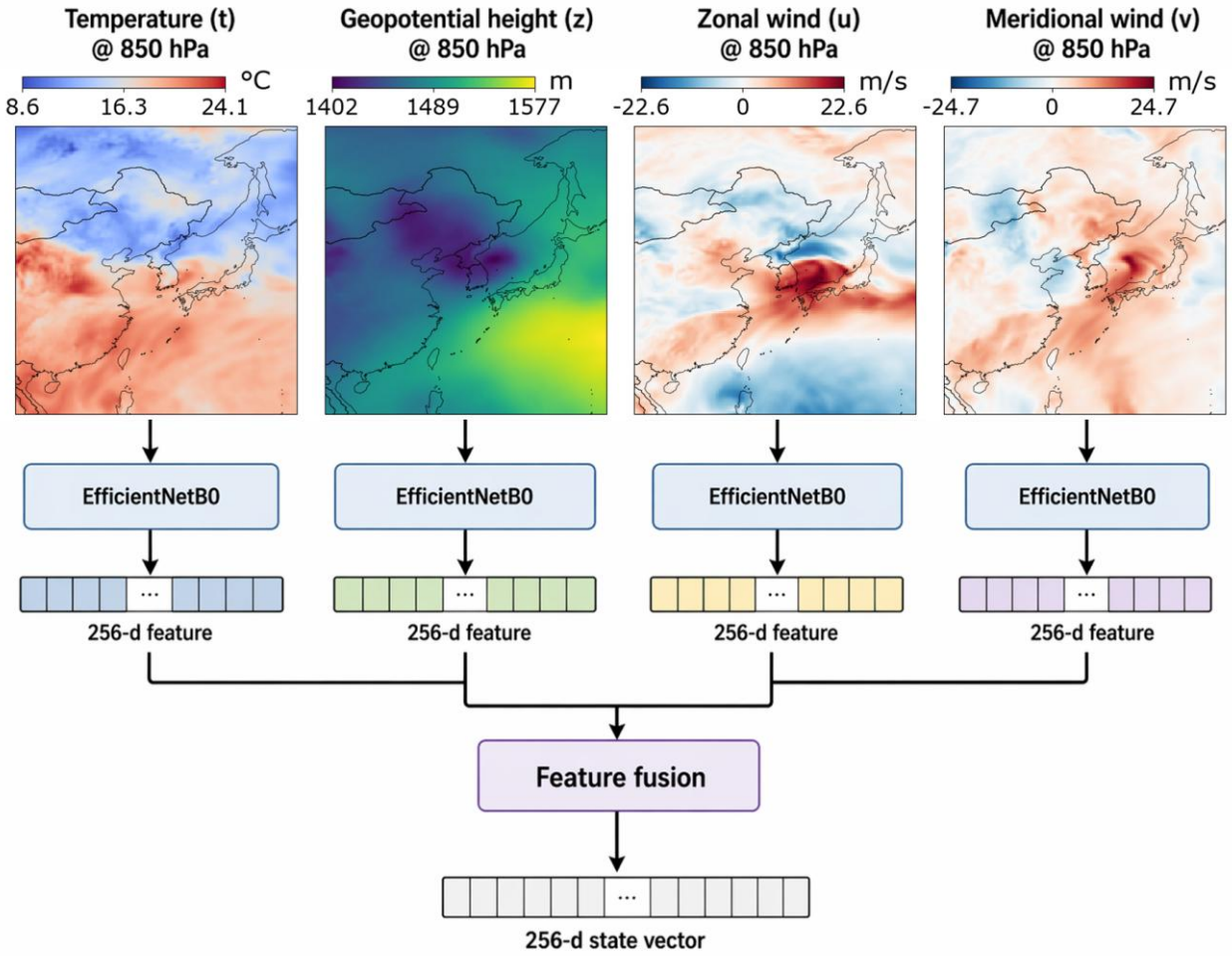



Figure R1-F10. Revised version of the submitted manuscript's Figure 2. Actual ERA5 fields at 850 hPa are displayed as temperature (°C), geopotential height (m), zonal wind (m $s^{-1}$), and meridional wind (m $s^{-1}$). The wind panels show signed scalar components. Separate variable encoders produce 256D features, which are fused into a 256D state vector.

Figure R1-F11

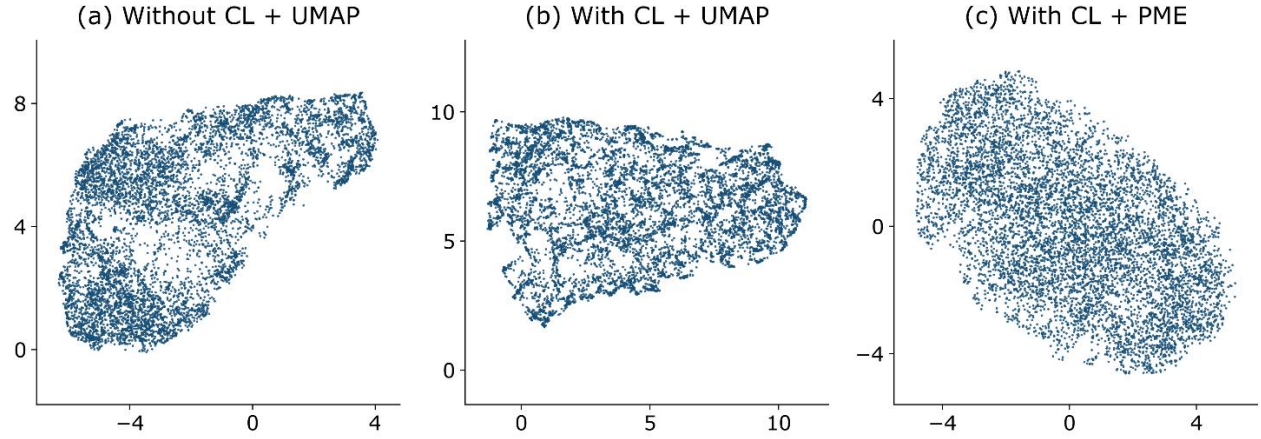



Figure R1-F11. Revised version of the submitted manuscript's Figure 3. The same atmospheric samples are displayed without contrastive learning with UMAP, with contrastive learning with UMAP, and with contrastive learning with parametric manifold embedding. Each axis is a dimensionless embedding coordinate. Axis ranges are set independently, so absolute positions and distances are not directly comparable across panels.

Figure R1-F12

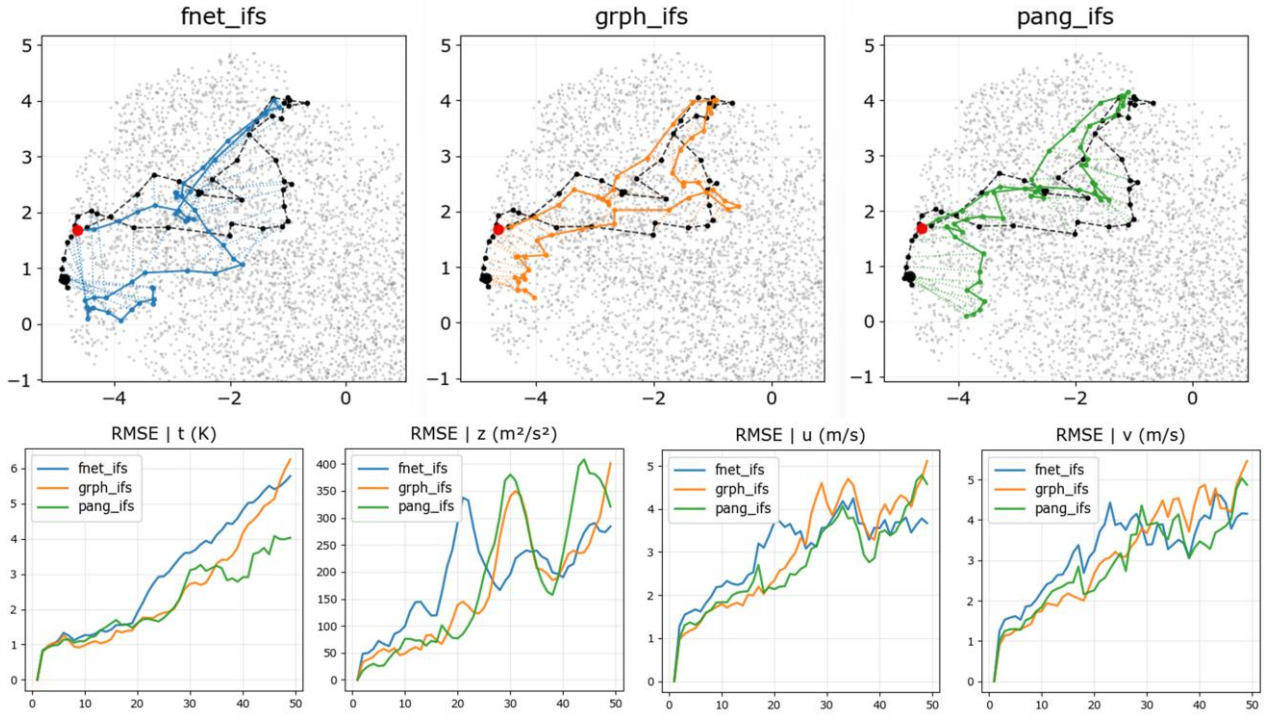



Figure R1-F12. Revised version of the submitted manuscript's Figure 8. Upper panels compare lead-time trajectories for FourCastNet2:IFS (fnet_ifs), GraphCast:IFS (grph_ifs), and Pangu-Weather:IFS (pang_ifs). Lower panels show variable-wise RMSE against IFS analysis at 850 hPa over the same East Asian evaluation domain. Units are K for temperature, $m^2\ s^{-2}$ for geopotential, and $m\ s^{-1}$ for the two wind components.

Figure R1-F13

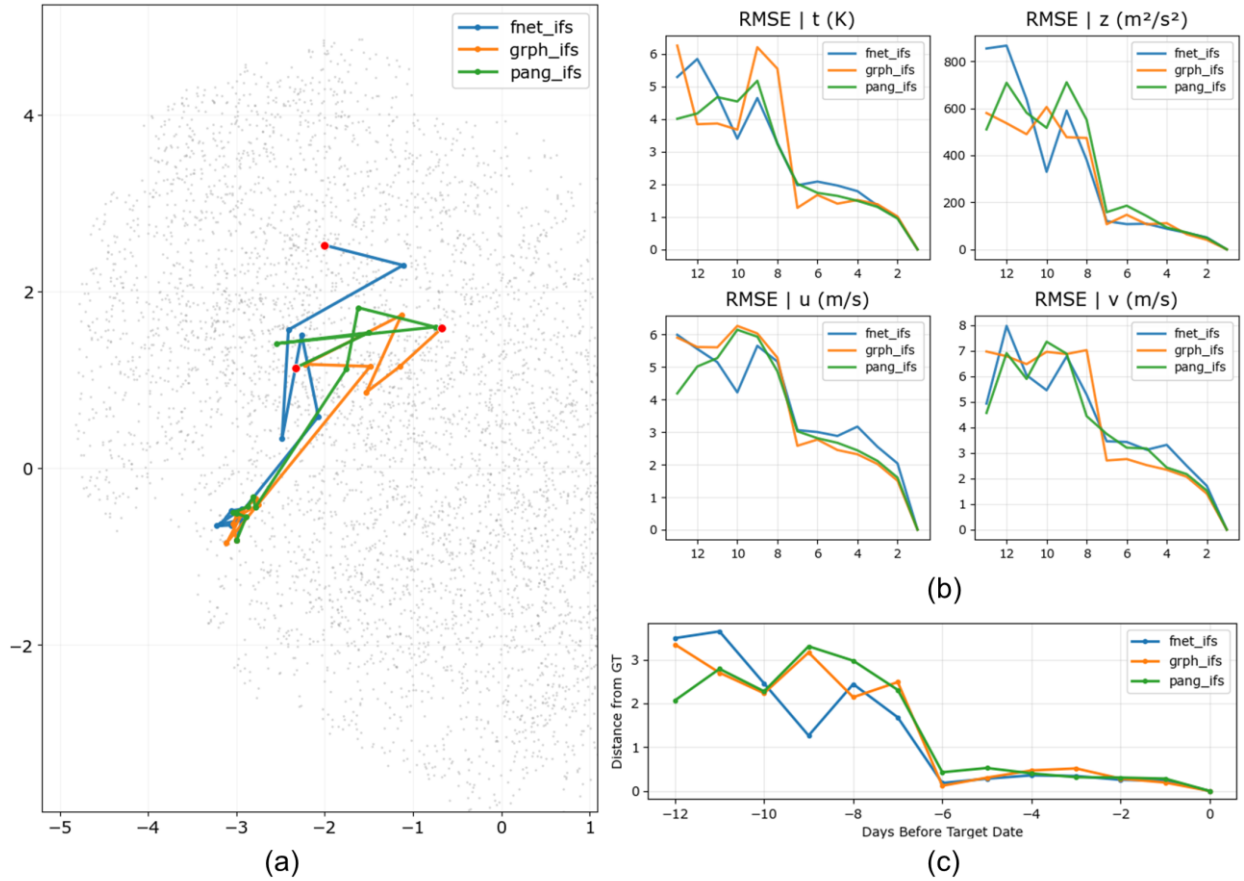



Figure R1-F13. Revised version of the submitted manuscript's Figure 11. Left: valid-time trajectories for FourCastNet2:IFS, GraphCast:IFS, and Pangu-Weather:IFS. Right: the corresponding RMSE and state-space distance as forecast initialization approaches the target time. RMSE uses IFS analysis over the same East Asian domain at 850 hPa, with units K, $m^2\ s^{-2}$, and $m\ s^{-1}$ for temperature, geopotential, and wind components, respectively.