Review of "Topology-Preserving State Space Representation for Diagnosing Weather Forecast Models" by Kim and Cho (egusphere-2026-3142)

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.

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.

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.

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.

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.

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.

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.

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.
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.
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.
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.

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.
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.
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.
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.