npj Climate and Atmospheric Science https://doi.org/10.1038/s41612-026-01380-1

Article in Press

# A spectral test of the butterfly effect and physical consistency in the diffusion-based GenCast's ensembles

Received: 3 October 2025

Accepted: 4 March 2026

Cite this article as: Kim, H., Ryu, J., Son, S.-W. *et al.* A spectral test of the butterfly effect and physical consistency in the diffusion-based GenCast's ensembles. *npj Clim Atmos Sci* (2026). https://doi.org/10.1038/s41612-026-01380-1

Hisu Kim, Jihun Ryu, Seok-Woo Son, Jee-Hoon Jeong, Hyungjun Kim & Jin-Ho Yoon

# A SPECTRAL TEST OF THE BUTTERFLY EFFECT AND PHYSICAL CONSISTENCY IN THE DIFFUSION-BASED GENCAST'S ENSEMBLES

**Hisu Kim**[1], **Jihun Ryu**[2], **Seok-Woo Son**[3], **Jee-Hoon Jeong**[4], **Hyungjun Kim**[5], **& Jin-Ho Yoon**[1,*]
[1] Department of Environment and Energy Engineering, Gwangju Institute of Science and Technology, Gwangju, South Korea
[2] Plants, Soils and Climate Department, Utah State University, Logan, UT, USA
[3] School of Earth and Environmental Sciences, Seoul National University, Seoul, South Korea
[4] Department of Environment and Energy, Sejong University, Seoul, South Korea
[5] Moon Soul Graduate School of Future Strategy, Korea Advanced Institute of Science and Technology, Daejeon, South Korea
*Corresponding author: yjinho@gist.ac.kr



**ABSTRACT**

With the rapid development of deep learning weather prediction (DLWP) models like GenCast, rigorous evaluation of their physical consistency is essential. This study investigates the dynamical fidelity of GenCast against ECMWF IFS-HRES and IFS-ENS using comprehensive kinetic energy (KE) and difference kinetic energy (DKE) spectra over 2021. Unlike the physically consistent error growth in IFS-ENS, GenCast exhibits weak planetary-scale growth and a persistent, flattened KE tail at high wavenumbers starting from the first forecast step. These mesoscale artifacts persist across multiple GenCast variants and AIFS-ENS, indicating a broader challenge for noise-conditioned generation. Helmholtz decomposition further reveals white-noise-like variance rather than balanced dynamics. Spatially, weak interactions between large-scale and mesoscale wind fields suggest a misrepresentation of topography-flow interactions. Furthermore, analyses of KE gradient ($|\nabla \mathrm{KE}|$) revealed that GenCast fails to reproduce the sharp, filamentary structures, instead generating broad, isotropic, and noisy patterns. These findings suggest that current noise injection mechanisms in DLWPs produce noisy artifacts mimicking variance without reproducing realistic error growth physics. Improving these mechanisms is vital for developing physically consistent DLWPs.

## Introduction

The butterfly effect, first proposed by Edward Lorenz, establishes a fundamental limit to atmospheric predictability [1]. Since tiny initial errors rapidly amplify and alter the large-scale state, ensemble forecasting with perturbed initial conditions was introduced to represent forecast uncertainty. Each perturbed member evolves toward a distinct but plausible trajectory of the future atmosphere [2]. This probabilistic framework supports event-likelihood estimation and more reliable risk assessment. For these reasons, ensemble forecasting has become a cornerstone of operational Numerical Weather Prediction (NWP) systems, with major centers routinely providing ensemble-based forecasts for medium-range prediction.

An NWP model, such as Integrated Forecasting System (IFS) from ECMWF, has conducted ensemble forecasts comprising more than 50 members [3]. However, producing large ensembles with high-resolution NWP models requires substantial computational resources. Deep learning has therefore emerged as an efficient alternative for global ensemble forecasting. Over the past five years have seen the emergence of several deep learning weather prediction (DLWP) models, many of which surpass conventional NWP systems across multiple evaluation metrics. Moreover, DLWPs have offered notable reductions in computational expense, while increasing accessibility of advanced forecasting tools [4, 5, 6, 7, 8, 9].

Early DLWPs, such as Pangu-Weather or GraphCast, began as deterministic forecasting systems, produce only a single prediction without ensembles [7, 4]. The fidelity of these deterministic DLWPs has been evaluated across multiple

dynamical diagnostics. [10] conducted experiments to test whether Pangu reproduces the butterfly effect through Difference Kinetic Energy (DKE). They found that, unlike the PDE-based ICON (including a convection-permitting configuration), Pangu-Weather did not reproduce the rapid error growth expected from tiny initial perturbations. Instead, its growth rates and structures remained synoptic-scale, even under tiny perturbations. This highlights a critical limitation and therefore motivates testing whether DLWPs can capture mesoscale error cascades and physically consistent dynamics. Also, [11] emphasized that the model was not fully dynamically balanced, highlighting inconsistencies in wind decomposition and vertical velocity compared with IFS-HRES forecasts and ERA5 analyses. [12] observed that Pangu-Weather lacked physical fidelity at the mesoscale, as it consistently underestimated KE required for realistic atmospheric variability. On the other hand, [13] provided empirical evidence that Pangu-Weather showed physically meaningful behavior, as artificial perturbations in geopotential or temperature fields elicited responses that are consistent with atmospheric theory. Yet these studies primarily evaluated deterministic DLWPs such as Pangu-Weather, ForecastNet, or GraphCast, often focusing on a limited number of case studies.

In contrast to deterministic DLWPs, some recent DLWPs generate ensembles and perform probabilistic forecasts that do not rely on initial condition perturbations. For instance, GenCast generates ensembles via a conditional diffusion process that iteratively denoises isotropic Gaussian noise sampled from the latent representation conditioned on the previous step [9]. Because sampling in diffusion is stochastic, drawing fresh noise realizations at inference time allows arbitrarily large ensembles without initial condition perturbations. This generative approach has reported ensemble skill competitive with IFS-ENS, a probabilistic NWP model from ECMWF, on most verified targets, and influenced subsequent probabilistic DLWPs [14, 15].

Despite the strong skill of diffusion-generated DLWPs, it remains unclear whether their ensembles are dynamically equivalent to those produced by initial condition-perturbed NWP systems, whose spread reflects the upscale amplification of initial uncertainties and adheres to physical constraints. The butterfly effect, a fundamental property of atmospheric dynamics, and its related processes can be used as a benchmark to assess the physical fidelity of ensemble generation methods such as GenCast. In geophysical turbulence, this effect appears as an upscale error cascade, which leaves a clear spectral signature traceable in the DKE spectrum. Specifically, small initial uncertainties are amplified through moist convective processes (*e.g.*, latent heat release), and then propagate upscale via gravity waves and adjustments to geostrophic balance. Ultimately, these errors saturate at large scales, setting an intrinsic limit on forecast skill [16, 17, 18, 19, 20]. Moreover, the kinetic energy (KE) spectrum provides scale-dependent saturation bounds on forecastable variance, so approaching these bounds at the appropriate wavenumbers indicates a physically consistent energy dynamics. Accordingly, prior studies have widely used DKE and KE spectral diagnostics to assess both reanalyses and climate simulation models, including NWPs and DLWPs [21, 22, 23, 24, 25, 26, 10, 27]. In this study, we used both KE and DKE diagnostics to compare diffusion-generated and initial condition-perturbed ensembles. In addition, we included the deterministic IFS-HRES forecast as a baseline, ensuring no single ensemble member is privileged, consistent with the exchangeability requirement in ensemble evaluation [28]. The methodological details for the two ECMWF models (IFS-ENS, IFS-HRES) and GenCast are provided in Methods. We began from 300 hPa DKE analysis following the previous study ([10]) in **300 hPa Difference Kinetic Energy**, then analyzed the 300 hPa KE spectra in **300 hPa Kinetic Energy Spectra**, performed Fourier scale separation in **Fourier Analysis of 300 hPa Wind Kinetic Energy**, and conducted $|\nabla \mathrm{KE}|$ diagnostics in **Magnitude of Kinetic Energy Gradient**. We mainly discuss the first ensemble member of the two probabilistic models for easier comparison with IFS-HRES. Results of the ensemble mean are also provided in **Supplementary Material**.

## Results

### 300 hPa Difference Kinetic Energy

To evaluate the error growth characteristics of GenCast relative to the dynamical baseline, we analyzed the time evolution of DKE between ensemble members. Figure 1 (a) presents the time series of globally integrated DKE. Note that, since GenCast is initialized from a single ERA5 reanalysis (*i.e.*, without ensemble members), GenCast's DKE is not defined at initialization (lead time 0). GenCast exhibited a slightly slower increase in DKE and somewhat lower overall DKE values compared to IFS-ENS during the first week, indicating a marginally weaker initial error growth rate. Until 192 hours, GenCast maintained lower DKE values than IFS-ENS, but at this lead time, the two curves intersect, and GenCast exhibited higher DKE values than IFS-ENS thereafter. Nevertheless, the overall similarity in the initial growth rates suggests that GenCast successfully captures the essential large-scale error dynamics characteristic of a practical forecast system.

However, the spectral evolution of DKE (Figure 1 (b, c)) reveals a fundamental divergence in how this error propagates across scales. It is important to distinguish between the intrinsic predictability limit, which requires infinitesimal "butterfly-sized" perturbations to observe the saturation of microscale errors and their subsequent upscale cascade

[21, 10], and the practical predictability limit observed here with full-magnitude perturbations. While our setup utilizes operational-scale perturbations, a physically consistent model must still exhibit upscale error transfer, where uncertainty at smaller scales eventually saturates and drives growth at larger scales, or error grows at all scales (up-amplitude growth).

IFS-ENS demonstrated this physical behavior: the spectral peak of DKE shifted progressively towards smaller wavenumbers with lead time, and error growth was observed across the entire scale, including the planetary scale (low wavenumber), as shown in Figure 1 (b), indicating a physically consistent up-amplitude error growth of uncertainty throughout all spatial scales. Similarly, on the synoptic scales, GenCast mimicked the behavior of IFS-ENS. The spectral peak of DKE progressively shifted towards lower wavenumbers with lead time. Despite this synoptic similarity, two anomalies indicate a fundamental disconnection in GenCast's cross-scale interactions. First, in the small-scale regime ($k > 300$), GenCast exhibited a flat, constant distribution that remained virtually unchanged across all lead times. The estimated spectral slope in the band $k > 300$ by linear regression was -3.16 for IFS-ENS and -0.52 for GenCast at saturation. This suggests that perturbations at these scales behave less like dynamically evolving fluid errors. Another difference appeared to impede error growth at the planetary scale ($k \in [1, 4]$). While IFS-ENS showed a robust amplification of large-scale errors, increasing by approximately two orders of magnitude over the 10-day forecast, GenCast exhibited stunted growth, increasing by less than a single order of magnitude.

These findings indicate that while GenCast injects perturbations internally, such perturbations may not facilitate the spread of perturbations across all scales. Another possibility is that, because GenCast learned from ERA5 data with a relatively coarse 12-hour time step, it may encounter inherent structural difficulties in replicating the sophisticated error cascade processes that emerge from the continuous time integration used in NWP systems. Regardless of the underlying cause, this raises questions regarding the physical fidelity of the simulated energy cascade mechanism. To further investigate this issue, we provided an explicit analysis of the full KE spectra in the following section.



**300 hPa Kinetic Energy Spectra**

The jet stream at 300 hPa is particularly susceptible to rapid error accumulation due to its highly variable and energetic dynamics so the temporal evolution of global KE spectra and spectra change ($\Delta E(k, \tau)$) were examined. The KE spectra and their changes from the initial state ($\tau = 0$) in Figure 2 reveal how forecasts develop KE across different scales for each model. At low wavenumbers, all three models maintained their KE, as seen in Figure 2 (a, c, e). This indicates that large-scale flow is generally conserved over the forecast period. However, the results diverged in how the models handle energy dynamics at smaller scales. The spectral change plots for IFS-HRES and IFS-ENS show distinct, dynamically consistent energy transfer.

IFS-HRES exhibited oscillatory behavior at low wavenumbers (Figure 2 (b)). In contrast, at high wavenumbers, the differences progressively decreased and converged toward a narrow range. It is noteworthy that IFS-ENS lost KE on small scales from the early stage of the forecast. As shown in Figure S5, a more distinct separation between models emerged starting at $k \approx 30$. This rapid dissipation indicates that the model's physical parameterizations act as a damping mechanism on these initial perturbations. Specifically, IFS adopts a sequential calling procedure for its moist physical parameterizations: radiation, turbulent diffusion, convection, and cloud schemes. In this sequence, the convection parameterization acts not only as a physical representation of subgrid-scale activity but also as a numerical filter. These schemes are designed to enhance atmospheric stability and suppress numerical noise efficiently, particularly at the mesoscale where resolved motions begin to transition into subgrid-scale processes [29]. The KE dissipated by turbulent diffusion, orographic form drag, and convective momentum transport is converted back into heat to satisfy the conservation of moist static energy (enthalpy) [30]. As a result, the initial condition perturbations in IFS-ENS, particularly those generated by the singular-vector initialization designed to amplify perturbations within 48 hours [3], likely introduced an unbalanced state. The model's moist physics and turbulent diffusion then act to immediately stabilize this imbalance by rapidly converting the excess KE into heat to satisfy enthalpy conservation, thereby resulting in a faster mesoscale convergence compared to the deterministic IFS-HRES.

On the other hand, GenCast exhibited a fundamentally different and less realistic pattern of energy transfer. The most critical issue is visible in its change, a tail-shaped feature at high wavenumbers in Figure 2 (f), unlike IFS. This can also be seen as a noticeable bending at the high-wavenumber end, where the slope flattens beyond a wavenumber of 300 ($k \geq 300$). Rather than showing upscale energy transfer and moving toward larger scales or dissipation, the mesoscale KE remains static and unchanged after the first forecast step, as shown by the nearly perfect overlap of all lead times. GenCast's flattened tail suggests that its mesoscale structures are not generated or dissipated in a physically realistic manner, unlike those of IFS. Figure 3 illustrates a quantitative summary of this behavior, showing that GenCast's value exhibits minimal variation around -0.4, whereas $E_{\text{meso}}(\tau)$ of both IFS models decays until approximately $\tau = 120$h.

These results were similar when ensembles were averaged, as shown in Figures S3 and S4. Three other pretrained GenCast models, which have different resolutions and fewer noise refinement steps, also exhibited similar results (Figure S6), suggesting that all GenCast variants tend to generate unrealistic mesoscale features and spurious mesoscale KE. Moreover, in **Pilot Experiment: initial condition and additional model**, which tested different initial conditions and an additional ensemble DLWP model, initializing GenCast by the initial condition of the IFS control member also produced the same flat tail (Figure S1). A recent study ([31]) showed that Denoising Diffusion Probabilistic Models can undergo "spectral collapse," in which the model "learn an increasingly biased representation of the data: large-scale structures remain visible throughout the diffusion horizon, while inertial-range and dissipation-range features are overwhelmed by noise." More precisely, they defined "signal-to-noise spectral collapse" as occurring when "standard Gaussian noise schedules disproportionately obscure high-wavenumber dynamics early in the forward process, preventing the model from learning the fine-scale score components required for physically accurate long-term emulation." This theoretical finding is consistent with the anomalous mesoscale behavior we observed in GenCast, whose backbone is conditional denoising diffusion. However, AIFS-ENS (ECMWF's ensemble DLWP model published as AIFS-CRPS) from the pilot experiment exhibited similar mesoscale behavior regardless of the choice of initial condition, even though AIFS-ENS integrates Gaussian noise embeddings into its transformer layers, not a denoising diffusion-based noise refinement (Figure S2) [32]. As both GenCast and AIFS-ENS use noise to represent uncertainty and generate ensembles, the presence of a "flat tail" in the spectra is likely related to this noise injection, based on the results above. Thus, we separated KE into rotational and divergent components by Helmholtz decomposition and reexamined their spectra.

According to [33], rotational and divergent KE spectra are expected to follow distinct power laws: rotational KE typically has a slope near -3, while divergent KE follows a -5/3 slope. Figures 4 (a) and (b) show that both IFS-HRES and IFS-ENS adhered to these theoretical expectations. In contrast, GenCast in Figure 4 (c) exhibits flat slopes in the mesoscale range for both rotational and divergent KE spectra, mirroring the flatness observed in Figure 2. This lack of scale dependence is physically unrealistic and implies an underlying issue with GenCast's mesoscale representation. Second, we observed that in the problematic tail region (high wavenumbers), rotational and divergent KE become nearly equal in GenCast. While it is expected that these quantities converge at very small scales [33], GenCast displays this equipartition beginning at much lower wavenumbers around $k = 100$, which is atypical. The combination of flat slopes and equipartition between rotational and divergent energies more closely resembled the theoretical behavior of white noise than realistic atmospheric dynamics.

Meanwhile, GenCast produces realistic energy variability in low wavenumbers ($k < 100$) without energy decay over time, except $k \leq 10$, where their power converged before 5 days (Figure 2 (f)). Notably, for $k \in [4, 10]$, the spectrum converged to values higher than the initial condition, resulting in the accumulation of KE within this band. Figure 5 summarizes the KE variation across lead times for the three models. While both numerical models exhibited a consistent decrease in KE across all spatial bands except the planetary scale, GenCast showed a persistent increase in energy within the planetary to synoptic scale range ($k \in [4, 10]$), as shown in Figure 5 (b). This accumulation contributes directly to the rise in total KE (Figure 5 (g)), with a Pearson correlation coefficient of $r = 0.93$. Whereas IFS models exhibit a steady decrease in KE due to numerical damping at high wavenumbers, GenCast might not learn such mechanisms from the reanalysis data, which can be expected; instead, energy continues to accumulate between the large and synoptic scales, leading to a steady increase in total KE.

Consequently, these results imply that GenCast's representation of KE is dynamically inconsistent. The persistent accumulation of energy at synoptic scales points toward a potential deficiency in physical dissipation while the spectral flattening and equipartition at mesoscales hint residual effects of noise injection, rather than coherent flow. While the spectra reveal that suspicious energy exists, they do not explain where these errors are located or how they interact with the other scales yet. To address this, we now shift our focus from the spectral domain to the physical domain. In the next section, we analyzed the spatial distribution of mesoscale KE to identify the localized signatures of these artifacts.

**Fourier Analysis of 300 hPa Wind Kinetic Energy**

To localize the scale contributions, we reconstructed the KE fields in physical space after wavenumber separation at $k$ =100. Figure 6 separately depicts KE components of the first ensemble member with ERA5. On the large scale (left panels of Figure 6), GenCast closely resembles the spatial structure of IFS-HRES, with enhanced KE at the eastern termini of the Northern Hemisphere continents and south of Africa, where the absence of major topographic barriers allows jet intensification. It also captured a weak KE around Maritime Southeast Asia, which is more visible in Figure 6 (k). In general, regions with elevated $\mathrm{KE}_{\mathrm{high}}$ corresponded well to the large-scale jet streams where total KE was high, as shown in the middle panels of Figure 6. Strong $\mathrm{KE}_{\mathrm{high}}$ around the equator of the eastern Pacific implies the conversion of available potential energy into KE driven by strong convection within the intertropical convergence zone was well simulated in all three models, albeit not as distinctly as in ERA5. However, a difference lies in the magnitude:

the $\mathrm{KE}_{\mathrm{high}}$ values in GenCast were approximately three times larger. This characteristic was more evident in Figure 7 (e), which indicates that the distribution of GenCast was broader than that of the other three models.

The cross-term represents the interaction between scales; specifically, how well the large-scale and mesoscale wind components are aligned, since the cross-term is essentially the inner product of the two wind vectors. In the right panels of Figure 6, a notable contrast emerged in cross-term over the Andes (this region is magnified for closer examination in the figure). While ERA5 and both IFS models displayed strong alternating cross-term values in this region, GenCast did not. As KE can be dissipated by inertial gravity wave drag generated by orography, we further investigated whether this was due to a lack of interaction between topography and the jet stream in GenCast [30]. The similar signals, however, were not observed from Figure 6 (l). Instead, cross-term spatial maps of ensemble mean displayed in Figure S7 (b) and (e), IFS-ENS exhibited strong $\mathrm{KE}_{\mathrm{high}}$ values primarily over mountainous and high-altitude regions, with most other areas appearing smoothed out due to averaging, but GenCast did not show such pronounced enhancement in these regions compared to IFS-ENS. Nevertheless, the distribution of values in the first ensemble member (Figure 7 (f)) revealed that the standard deviation for GenCast ($4.50\ \mathrm{m}^2/\mathrm{s}^2$) remained more than four times larger than that of ERA5 ($0.82\ \mathrm{m}^2/\mathrm{s}^2$), IFS-HRES ($1.13\ \mathrm{m}^2/\mathrm{s}^2$), and IFS-ENS ($1.04\ \mathrm{m}^2/\mathrm{s}^2$). The elevated distributions in $\mathrm{KE}_{\mathrm{High}}$ and cross-term can be interpreted from Figure 2 (e): since the spectral density of GenCast above $k = 200$ was lifted so that it had stronger power in mesoscale KE than IFS. This excess and broad power at high wavenumbers align with the expected signature of injected noise used for ensemble generation, as discussed in the previous section. The wide ranges of $\mathrm{KE}_{\mathrm{High}}$ and cross-term were greatly suppressed and became narrower in the ensemble-averaged distribution, though it still remained wider than that of ERA5 and IFS (Figure 7 (c)). This suggests that the extreme values observed in the tails are incoherent across the ensemble; they appear sporadically rather than systematically, indicating a high likelihood of them being noise.

Above all, the lack of coherent signals over complex terrain leads us to infer that GenCast may not have fully learned the dynamical interactions between topography and atmospheric flow, despite the inclusion of surface geopotential as an input [9]. Distributions of $\mathrm{KE}_{\mathrm{high}}$ and cross-term suggest residual noise within the generated fields. To determine the extent to which these artifacts degrade the flow's structural integrity, we turned our attention to the magnitude of the kinetic energy gradient.

**Magnitude of Kinetic Energy Gradient**

As a complement to the spectra and scale-separated maps, jet-core sharpness and alignment using the magnitude of the kinetic-energy gradient, $|\nabla \mathrm{KE}|$, at 300 hPa were evaluated. In a physically realistic atmosphere, sharp gradients typically delineate the boundaries of the jet stream core. IFS-HRES and IFS-ENS demonstrated a superior ability to reproduce the physically consistent structures of atmospheric flow, as shown in Figure 8 (b, c). Both models exhibited a narrow, filamentary flow of $|\nabla \mathrm{KE}|$ that meandered around the mid-latitudes, precisely aligning with the core of the jet stream. These coherent, jet-aligned filaments were particularly effective at highlighting the entrance and exit regions of jet streaks and boundaries of the jet. A prominent feature in both NWP models was the clear preservation of jet cores, most notably in the distinct jet stream structure over East Asia. The structures produced by these models were anisotropic, with a clear directional alignment that followed the zonal orientation of the jet stream. While sharing a common fidelity to atmospheric dynamics, IFS-ENS and IFS-HRES displayed distinct differences. As an ensemble system, IFS-ENS reflects the rapid amplification of initial perturbations imposed on its members. This amplification leads to stronger amplitudes of $|\nabla \mathrm{KE}|$ and broader gradient filaments compared to the single control forecast, IFS-HRES. Consequently, IFS-ENS predicted more intense jet-aligned gradients and a wider latitudinal extent of storm-track variance than its HRES counterpart.

In contrast, GenCast failed to reproduce the sharp, dynamically consistent features observed in the NWP forecasts. Instead of distinct filaments, GenCast's output exhibited broad noisy gradient belts, smoothing out the sharp storm-track peaks into wide belts at 30-40° latitude, as is visible in the profile in Figure 8 (d). The most striking difference lay in the texture of the generated gradients. GenCast produced a nearly isotropic, granular texture across the mid-latitudes, lacking the coherent directional alignment associated with jet dynamics. This texture was consistent with sampling noise introduced by the diffusion-based ensemble generation. After ensemble averaging, as shown in Figure S8, most of the texture present in Figure 8 (d) disappeared, indicating that the noisy patterns were incoherent and not associated with persistent physical structures. This absence of spatial continuity in the magnitude of KE gradient implies that the wind fields generated by GenCast were not dynamically smooth, which undermines their physical realism. This fragmentation is particularly problematic given that it persisted even at $\tau = 240$h, when some loss of predictability and fine detail is expected.

Furthermore, GenCast's temporal evolution was inconsistent with the progressive loss of forecast accuracy observed in NWPs; its fragmented, band-like gradients appear to be already established and remain largely unchanged across

lead times. The model also yielded anomalously high gradient magnitudes near the equatorial Indian Ocean and the Maritime Continent, regions where such signals were not as pronounced in the NWP forecasts.

## Discussion

Our analysis suggests a deficiency in GenCast's handling of mesoscale dynamics. GenCast appears to simulate realistic up-amplitude error growth, but only on the synoptic scale. Error growth in the planetary scale was suppressed, while mesoscale DKE remained static. Also, the model's KE spectra at the mesoscale exhibited a static, flattened tail, raising the possibility that the spectral power at these scales primarily reflects noise via Helmholtz decomposition and the pilot experiment. The observed accumulation of total KE, along with $\mathrm{KE}_{\mathrm{high}}$ and cross-term values at least three times greater than those in ERA5, is highly suggestive of a lack of mesoscale KE dissipation in GenCast. The vanishing cross-term signals over the Andes could potentially be linked to an underrepresentation of essential orographic drag processes, such as mountain wave breaking, which may in turn contribute to the unchecked accumulation of KE. Lastly, the isotropic, noisy texture of GenCast's KE gradient, in contrast to the coherent, jet-aligned structures in IFS-HRES and IFS-ENS, further suggests a structural difficulty in generating a dynamically consistent flow and its evolution. Taken together, the persistence of non-physical, noise-like structures at the mesoscale supports the hypothesis that the model might be operating without the necessary dissipative processes to properly control energy growth, potentially preventing the physically consistent upscale propagation of errors required to simulate the true butterfly effect. These apparent structural deficiencies suggest challenges to the physical realism of its jet stream representation and its ability to capture crucial aspects of weather system evolution.

This study targets one of the probabilistic DLWPs, GenCast, which autoregressively generates forecasts from the previous prediction. While a recent ensemble emulation model, such as Scalable Ensemble Envelope Diffusion Sampler (SEEDS), generates ensembles from one or two steps of NWP forecasts, making it challenging to track initial error propagation, some state-of-the-art DLWPs still produce ensembles based on a denoising algorithm similar to GenCast or adopt GenCast for more advanced tasks [14, 32, 34, 35, 36]. We emphasize that this study does not argue that GenCast's forecasts are unreliable. In fact, its superior performance on many conventional evaluation metrics compared to NWPs is a well-established finding and a promising supplement to current NWPs [9]. While DLWPs achieve high skill scores in metrics such as root mean square error (RMSE) or anomaly correlation coefficient (ACC), these metrics do not fully capture the models' overall performance, particularly regarding physical feasibility [11, 37, 12]. Further, as the field advances, it becomes increasingly important to evaluate these models not only through numeric metrics but also by assessing the physical plausibility of their weather maps. Our analysis method focuses on key meteorological features and their dynamic consistency. We acknowledge, however, that a full spectral KE budget analysis, as proposed by [38], would provide even deeper insights into the processes underlying upscale energy transfer. We suggest that future studies incorporate such an analysis to further advance our understanding of these results. Even so, our relatively simple approach can be integrated into benchmarks such as WeatherBench2 or ChaosBench to provide a deeper understanding of whether models produce physically feasible forecasts [39, 27]. We anticipate that addressing these specific physical and dynamical deficiencies will lead to an improvement in the predictive performance of DLWPs.

## Methods

### Data

Our study spanned the full calendar year of 2021 with 52 Monday initializations. For each initialization, we used 20 forecast steps at a uniform 12-hour interval, covering a lead time of 240 hours. We analyzed 300 hPa wind fields ($u$ and $v$) from three forecasting systems, evaluated on a 0.25° spatial resolution: ECMWF's Integrated Forecasting System in its high-resolution deterministic (IFS-HRES) and ensemble (IFS-ENS) configurations, and GenCast.

### IFS

IFS-HRES is initialized by a 4D-Var data assimilation system that ingests conventional and satellite observations [3]. We downloaded IFS-HRES for our 2021 period from WeatherBench2 [39]. IFS-ENS comprises 51 members (50 perturbed + 1 control); perturbed initial conditions combine an ensemble of data assimilations with singular-vector perturbations [3]. IFS-ENS was obtained from TIGGE archive, a THORPEX component aggregating global ensemble forecasts from 13 centres and distributed via the ECMWF portal [40]. The archived TIGGE IFS-ENS has higher native resolution (O640, corresponding to approximately 16 km or 0.14° at the equator) than GenCast, so IFS-ENS was regridded to a common resolution of 0.25° for our analysis [41].

We used the first 10 ensemble members from the TIGGE dataset. Note that we excluded the control member from IFS-ENS ensembles and use IFS-HRES for deterministic comparisons, *i.e.*, the first ensemble member of IFS-ENS does not correspond to a control forecast but is the first perturbed forecast.

**GenCast**

GenCast is an open-source deep learning weather prediction model that consumes two ERA5 reanalyses separated by 12h (at $t - 12$h and $t$) and predicts the subsequent $t + 12$h state and auto-regressively predicts future states [42]. Internally, each 12h forecast is obtained by a multi-step refinement of a noise-initialized candidate. At the start of the step, isotropic Gaussian noise on the sphere is sampled and projected to the latitude-longitude grid via an inverse spherical-harmonic transform, forming the initial candidate for $t + 12$h. The encoder maps this candidate together with the conditioning from the grid to a latent field on a refined icosahedral mesh. Next, the processor updates that representation with a graph transformer, and then the decoder projects the updated features back to the grid to produce a 12h forecast increment, which is added to the previous state. Details can be found in [9].

Four pretrained versions are publicly released, and we primarily used *GenCast 0p25deg <2019*, which has a 0.25° resolution and a 6-time refined icosahedral mesh, trained on 1979-2018 ERA5 reanalysis, tested using 2019 and later years. *GenCast 1p0deg <2019* and *GenCast 1p0deg Mini <2019* are also used for an additional experiment, which both have 1.0° resolution but 5-time and 4-time refined icosahedral mesh, respectively. We run GenCast on two NVIDIA H100 GPUs to produce forecasts with 10 ensemble members. Since the original *splashattention* is TPU-only, the implementation uses *triblockdiag_mha* for GPU execution instead: a tri-block-diagonal multi-head attention layer that is algebraically equivalent to *splashattention*. The authors noted a slight performance drop attributable to numerical precision differences between GPUs and TPUs [43].

It should be noted that these three models are initialized in fundamentally different ways. IFS-HRES utilizes a 4D-Var data assimilation system for its single deterministic initialization, IFS-ENS relies on ensembles of data assimilations and singular-vector perturbations to create its ensemble members, while GenCast starts with two ERA5 reanalyses.

**Kinetic Energy and Difference Kinetic Energy**

The butterfly effect has often been investigated in previous studies using spectral diagnostics by examining error saturation and the upscale cascade of energy in the spectral domain. The temporal evolution of DKE spectra is effective for diagnosing the growth of forecast errors and the limits of atmospheric predictability [21, 19, 20, 10]. In addition, atmospheric physics is closely linked to flow dynamics; the KE spectrum captures the multiscale interplay of energy generation, cascade, and dissipation [44]. Consequently, KE spectra are widely used to evaluate model formulation, scale interactions, and the physical realism of atmospheric models [25, 24, 22, 33]. In this study, we tracked the time evolution of global DKE and KE and their spectra to reveal the behavior of forecast error and to assess the realism of the model forecasts.

We focused on the 300 hPa level, near the climatological jet core, where wind speeds, and thus KE, are largest and small perturbations amplify most rapidly [10]. At each grid point on that pressure surface, KE and DKE are defined as

$$\mathrm{KE} = \frac{1}{2}(u^2 + v^2) \tag{1}$$

$$\mathrm{DKE} = \mathrm{var}(u) + \mathrm{var}(v) \tag{2}$$

with $u$ and $v$ the zonal and meridional wind components, respectively, and var denotes the variance across ensemble members. To obtain KE spectra $E(k, \tau)$ for a given wavenumber $k$ and lead time $\tau$, we employed the spherical-harmonic methodology of NCAR Technical Note NCAR/TN-388+STR [45]. Before the transform, all wind fields were conservatively remapped to a Gaussian grid with 360 latitudes (N360), ensuring a common representation for spectral diagnostics. For DKE spectra, our diagnostics followed the framework of [19, 10].

Spectral KE evolution from the initial valid time, $\Delta E$ is defined as

$$\Delta E(k, \tau) = E(k, \tau) - E(k, 0). \tag{3}$$

All spectra ($E$ and $\Delta E$) were computed for each forecast initialized on Mondays in 2021 (total of 52 cases). For IFS, $\tau = 0$ corresponds to the 12 UTC Monday initialization. For GenCast, $\tau = 0$ uses the ERA5 reanalysis at $t$ (the second input), *i.e.*, the valid time immediately preceding the first 12h forecast step $t + 12$h. Although GenCast ingests two analyses ($t - 12$h, $t$), the spectral difference between these inputs was negligible at the scales considered; using $t - 12$h instead of $t$ as the reference yields indistinguishable $\Delta E$.

Further, to quantify the evolution with lead time in specific spatial scales, we used the following notations for the partial sum and relative spectral change:

$$E_{n,m}(\tau) = \sum_{k=n}^{m} E(k,\tau) \tag{4}$$

$$\Rightarrow \Delta E_{n,m}(\tau) = \frac{E_{n,m}(\tau) - E_{n,m}(\tau=0)}{E_{n,m}(\tau=0)} \times 100, \tag{5}$$

and accordingly,

$$E_{\text{meso}}(\tau) = \int_{100}^{k_{\max}} \Delta E(k,\tau)\, dk \tag{6}$$

where the integral spans from $k = 100$ (approximately 400km in wavelength at the equator) to the maximum wavenumber, $k_{\max}$. Following seminal theoretical and observational work on atmospheric turbulence, the KE spectrum is known to exhibit a distinct transition in slope around $k \approx 100$. At scales larger than this, the spectrum typically follows a $k^{-3}$ power law (associated with quasi-geostrophic turbulence) [46, 47], while at smaller, mesoscale ranges, it follows a $k^{-5/3}$ slope. This mesoscale regime has been attributed to forward energy cascades of stratified turbulence [48] or inertia-gravity waves [33], consistent with aircraft observations [49], although, as highlighted by [33], the precise origin of the mesoscale spectral slope remains questionable and controversial. Although the universality of these slopes of the atmosphere has been debated [22], the transition near $k = 100$ remains a standard reference for distinguishing dynamical regimes in atmospheric modeling [38, 25, 24, 19, 33, 50]. Accordingly, we used $k = 100$ as a cutoff to separate synoptic and mesoscale bands in our spectral diagnostics.

**Scale Seperation**

Next, we computed KE in "large scale (from planetary to synoptic-scale)," "mesoscale," and cross-term components to visualize the KE distribution on a map. To do this, 300 hPa wind fields were first split in the wavenumber domain: coefficients with $|k| \le k_c = 100$ (respectively $|k| > k_c$) are retained for the large-scale (mesoscale) component. All complementary coefficients were set to zero, and an inverse transform restored the filtered wind fields to the latitude-longitude grid. Note that a cosine-tapered window was implemented at the cutoff. Then KEs for each scale were calculated as follows:

$$\text{KE} = \frac{1}{2}(u^2 + v^2) \quad (\text{where } u \approx u_{\text{low}} + u_{\text{high}},\ v \approx v_{\text{low}} + v_{\text{high}}) \tag{7}$$

$$\approx \underbrace{\frac{1}{2}\left(u_{\text{low}}^2 + v_{\text{low}}^2\right)}_{\text{KE}_{\text{low}}} + \underbrace{\frac{1}{2}\left(u_{\text{high}}^2 + v_{\text{high}}^2\right)}_{\text{KE}_{\text{high}}} + \underbrace{\left(u_{\text{low}}u_{\text{high}} + v_{\text{low}}v_{\text{high}}\right)}_{\text{cross-term}} \tag{8}$$

where subscripts "low" and "high" refer to low-wavenumber (large scale) and high-wavenumber (mesoscale) components, respectively.

**Gradient of Kinetic Energy**

On the 300 hPa surface, we also diagnosed flow sharpness via the magnitude of KE gradient. The use of gradient-based methods is well established in computer vision, where they are commonly employed to detect edges and boundaries in images [51, 52]. Recently, similar gradient-based diagnostics were applied to quantitatively assess the sharpness of meteorological images generated by AI models [53]. Since the gradient can amplify grid-scale high-frequency noise, a Gaussian filter was applied before the computation [52]. Horizontal derivatives were computed with a coordinate-aware geospatial gradient operator that accounts for the Earth's spherical geometry, yielding

$$|\nabla \text{KE}| = \sqrt{(\partial_x \text{KE})^2 + (\partial_y \text{KE})^2} \tag{9}$$

at each grid point.

**Pilot Experiment: initial condition and additional model**

Changing ICs can influence error propagation and the behavior of models, as IFS does. Recently, probabilistic DLWP models have been developed, so it is worth examining an additional state-of-the-art ensemble DLWP model to verify whether the issues observed are unique to GenCast. To assess the effect of ICs and the additional ensemble DLWP model, we designed this pilot experiment.

### AIFS-ENS

ECMWF's AIFS-ENS (also known as AIFS-CRPS) was tested as an additional DLWP ensemble baseline of **Pilot Experiment: initial condition and additional model** in **Supplementary Material**. AIFS-ENS was trained on ERA5 reanalysis (1979-2017) and fine-tuned on operational IFS data (2016-2023), and is publicly available `https://huggingface.co/ecmwf/aifs-ens-1.0`. AIFS-ENS and GenCast share several similarities: both are trained on ERA5 reanalysis, have a spatial resolution of 0.25°, and adopt an encoder-processor-decoder architecture in which the processor stage centrally employs a transformer-based graph neural network (GNN) as ensemble DLWPs [9]. However, two models utilize noise for ensemble generation in different ways. In AIFS-ENS, a sampled Gaussian noise tensor is passed through a two-layer perceptron followed by layer normalization to obtain a noise embedding, which is then used as the conditioning signal in conditional layer normalization layers in the processor transformer blocks. Further details can be found in [32].

### Experimental design

12 forecasts were produced, each initialized on the first Monday of each month, from December 2024 to November 2025, by the same version of GenCast (*GenCast 0p25deg <2019*) and AIFS-ENS. We chose 2024-2025 for this experiment in order to avoid overlap with the training period of AIFS-ENS. In addition, 0.25-degree resolution initial conditions from Open data are only available from February 2023 onward. The TIGGE dataset and WeatherBench2, where IFS-HRES and IFS-ENS were downloaded, do not fully cover the pressure levels and variables GenCast demands. Furthermore, the original period, 2021, overlaps with the AIFS-ENS fine-tuning data. Each forecast was initialized using the IFS-ENS control member (ensemble 0) or ERA5 reanalysis, and as in the main experiments, we generated 10 ensemble members for each case, each with a 10-day lead time. The initial condition data for IFS-ENS was obtained from ECMWF's Open data.

**Data Availability** ERA5 reanalysis and IFS-HRES data were obtained from the WeatherBench2 dataset `https://console.cloud.google.com/storage/browser/weatherbench2/datasets`. ERA5 reanalysis is also available at Climate Data Store (`https://cds.climate.copernicus.eu/`). IFS-ENS were downloaded from the ECMWF TIGGE archive `https://apps.ecmwf.int/datasets/data/tigge/levtype=pv/type=pf/`). ECMWF's Open data is available at `https://www.ecmwf.int/en/forecasts/datasets/open-data`

**Code Availability** The GenCast model code is available at `https://github.com/google-deepmind/graphcast`. AIFS-ENS is accessible at `https://huggingface.co/ecmwf/aifs-ens-1.0`.

**Acknowledgments** This research was supported by the National Research Foundation (NRF) of Korea under RS-2025-0236304, and the High-Performance Computing Support Project, funded by the Government of the Republic of Korea (Ministry of Science and ICT).

**Author Contributions** Hi. Kim designed and conducted the study, wrote the initial draft of the manuscript, and carried out the revisions. J. Ryu contributed to the computation of KE spectra, discussed the results, provided comments on the manuscript, and was involved in the revision process. S.-W. Son, J.-H. Jeong, and Hy. Kim contributed to writing and editing the manuscript. J.-H. Yoon supervised the overall research, secured the funding, and reviewed the manuscript. All authors have contributed to a comprehensive review to ensure the depth and rigor of the study and approved the final version of the manuscript.

**Competing Interests** The authors declare no competing financial or non-financial interests.

## References

[1] Lorenz, E. N. Deterministic Nonperiodic Flow. *Journal of The Atmospheric Sciences* (1963). URL `https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml`.

[2] Kalnay, E. *Atmospheric modeling, data assimilation and predictability* (Cambridge University Press, Cambridge, 2012), 7. print edn.

[3] Owens, R. & Hewson, T. ECMWF Forecast User Guide. *ECMWF* (2018). URL `https://www.ecmwf.int/node/16559`.

[4] Bi, K. *et al.* Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023). URL `https://www.nature.com/articles/s41586-023-06185-3`.

[5] Sun, X. *et al.* FuXi Weather: An end-to-end machine learning weather data assimilation and forecasting system (2024). URL http://arxiv.org/abs/2408.05472. ArXiv:2408.05472 [cs].

[6] Chen, K. *et al.* FengWu: Pushing the Skillful Global Medium-range Weather Forecast beyond 10 Days Lead (2023). URL http://arxiv.org/abs/2304.02948. ArXiv:2304.02948 [cs].

[7] Lam, R. *et al.* GraphCast: Learning skillful medium-range global weather forecasting (2023). URL http://arxiv.org/abs/2212.12794. ArXiv:2212.12794 [cs].

[8] Pathak, J. *et al.* FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators (2022). URL http://arxiv.org/abs/2202.11214. ArXiv:2202.11214 [physics].

[9] Price, I. *et al.* Probabilistic weather forecasting with machine learning. *Nature* 1–7 (2024). URL https://www.nature.com/articles/s41586-024-08252-9.

[10] Selz, T. & Craig, G. C. Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect? *Geophysical Research Letters* **50**, e2023GL105747 (2023). URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL105747.

[11] Bonavita, M. On Some Limitations of Current Machine Learning Weather Prediction Models. *Geophysical Research Letters* **51**, e2023GL107377 (2024). URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL107377.

[12] Li, Z. *et al.* Exploring the differences in atmospheric mesoscale kinetic energy spectra between AI based and physics based models. *Scientific Reports* **15**, 15504 (2025). URL https://www.nature.com/articles/s41598-025-99815-x.

[13] Hakim, G. J. & Masanam, S. Dynamical Tests of a Deep Learning Weather Prediction Model. *Artificial Intelligence for the Earth Systems* **3**, e230090 (2024). URL https://journals.ametsoc.org/view/journals/aies/3/3/AIES-D-23-0090.1.xml.

[14] Alet, F. *et al.* Skillful joint probabilistic weather forecasting from marginals (2025). URL http://arxiv.org/abs/2506.10772. ArXiv:2506.10772 [cs].

[15] Mihai Alexe, S. L. Data-driven ensemble forecasting with the AIFS. *ECMWF Newsletter* (2024). URL https://www.ecmwf.int/en/elibrary/81620-data-driven-ensemble-forecasting-aifs.

[16] Leith, C. E. Atmospheric Predictability and Two-Dimensional Turbulence. *Journal of the Atmospheric Sciences* (1971). URL https://journals.ametsoc.org/view/journals/atsc/28/2/1520-0469_1971_028_0145_apatdt_2_0_co_2.xml.

[17] Tribbia, J. J. & Baumhefner, D. P. Scale Interactions and Atmospheric Predictability: An Updated Perspective. *Monthly Weather Review* (2004). URL https://journals.ametsoc.org/view/journals/mwre/132/3/1520-0493_2004_132_0703_siaapa_2.0.co_2.xml.

[18] Zhang, F., Bei, N., Rotunno, R., Snyder, C. & Epifanio, C. C. Mesoscale Predictability of Moist Baroclinic Waves: Convection-Permitting Experiments and Multistage Error Growth Dynamics. *Journal of the Atmospheric Sciences* (2007). URL https://journals.ametsoc.org/view/journals/atsc/64/10/jas4028.1.xml.

[19] Selz, T., Riemer, M. & Craig, G. C. The Transition from Practical to Intrinsic Predictability of Midlatitude Weather. *Journal of the Atmospheric Sciences* (2022). URL https://journals.ametsoc.org/view/journals/atsc/79/8/JAS-D-21-0271.1.xml.

[20] Zhang, Y. Sensitivity of Intrinsic Error Growth to Large-Scale Uncertainty Structure in a Record-Breaking Summertime Rainfall Event. *Journal of The Atmospheric Sciences* (2023). URL https://journals.ametsoc.org/view/journals/atsc/80/5/JAS-D-22-0231.1.xml.

[21] Durran, D. R. & Gingrich, M. Atmospheric Predictability: Why Butterflies Are Not of Practical Importance. *Journal of The Atmospheric Sciences* (2014). URL https://journals.ametsoc.org/view/journals/atsc/71/7/jas-d-14-0007.1.xml.

[22] Wang, J.-W. A. & Sardeshmukh, P. D. Inconsistent Global Kinetic Energy Spectra in Reanalyses and Models. *Journal of The Atmospheric Sciences* (2021). URL https://journals.ametsoc.org/view/journals/atsc/78/8/JAS-D-20-0294.1.xml.

[23] Rotunno, R., Snyder, C. & Judt, F. Upscale versus "Up-Amplitude" Growth of Forecast-Error Spectra. *Journal of the Atmospheric Sciences* (2022). URL https://journals.ametsoc.org/view/journals/atsc/80/1/JAS-D-22-0070.1.xml.

[24] Lauritzen, P. H. *et al.* NCAR Release of CAM-SE in CESM2.0: A Reformulation of the Spectral Element Dynamical Core in Dry-Mass Vertical Coordinates With Comprehensive Treatment of Condensates and Energy. *Journal*

*of Advances in Modeling Earth Systems* **10**, 1537–1570 (2018). URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2017MS001257. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2017MS001257.

[25] Skamarock, W. C., Park, S.-H., Klemp, J. B. & Snyder, C. Atmospheric Kinetic Energy Spectra from Global High-Resolution Nonhydrostatic Simulations. *Journal of The Atmospheric Sciences* (2014). URL https://journals.ametsoc.org/view/journals/atsc/71/11/jas-d-14-0114.1.xml.

[26] Sun, Y. Q. & Zhang, F. A New Theoretical Framework for Understanding Multiscale Atmospheric Predictability. *Journal of the Atmospheric Sciences* (2020). URL https://journals.ametsoc.org/view/journals/atsc/77/7/jasD190271.xml.

[27] Nathaniel, J. *et al.* ChaosBench: A Multi-Channel, Physics-Based Benchmark for Subseasonal-to-Seasonal Climate Prediction (2024). URL http://arxiv.org/abs/2402.00712. ArXiv:2402.00712 [cs].

[28] Fortin, V., Abaza, M., Anctil, F. & Turcotte, R. Why Should Ensemble Spread Match the RMSE of the Ensemble Mean? *Journal of Hydrometeorology* (2014). URL https://journals.ametsoc.org/view/journals/hydr/15/4/jhm-d-14-0008_1.xml.

[29] Bechtold, P. Convection Parametrization. In *Proceedings seminar on parameterization of subgrid physical processes*, 63–86 (2008).

[30] ECMWF. IFS documentation CY47R1 - part VI: Technical and computatinal procedures. In *IFS Documentation CY47R1* (ECMWF, 2020). URL https://www.ecmwf.int/node/19750. Number: 6.

[31] Sambamurthy, A. & Chattopadhyay, A. Lazy Diffusion: Mitigating spectral collapse in generative diffusion-based stable autoregressive emulation of turbulent flows (2025). URL http://arxiv.org/abs/2512.09572. ArXiv:2512.09572 [physics].

[32] Lang, S. *et al.* AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the Continuous Ranked Probability Score (2024). URL http://arxiv.org/abs/2412.15832. ArXiv:2412.15832 [physics].

[33] Li, Z., Peng, J. & Zhang, L. Spectral Budget of Rotational and Divergent Kinetic Energy in Global Analyses. *Journal of the Atmospheric Sciences* **80**, 813–831 (2023). URL https://journals.ametsoc.org/view/journals/atsc/80/3/JAS-D-21-0332.1.xml.

[34] Andrae, M., Landelius, T., Oskarsson, J. & Lindsten, F. Continuous Ensemble Weather Forecasting with Diffusion models (2025). URL http://arxiv.org/abs/2410.05431. ArXiv:2410.05431 [cs].

[35] Antonio, B., Strommen, K. & Christensen, H. M. Seasonal forecasting using the GenCast probabilistic machine learning model (2025). URL http://arxiv.org/abs/2509.06457. ArXiv:2509.06457 [physics].

[36] Hatanpää, V. *et al.* AERIS: Argonne Earth Systems Model for Reliable and Skillful Predictions (2025). URL http://arxiv.org/abs/2509.13523. ArXiv:2509.13523 [cs].

[37] Chattopadhyay, A., Sun, Y. Q. & Hassanzadeh, P. Challenges of learning multi-scale dynamics with AI weather models: Implications for stability and one solution (2024). URL http://arxiv.org/abs/2304.07029. ArXiv:2304.07029 [physics].

[38] Augier, P. & Lindborg, E. A New Formulation of the Spectral Energy Budget of the Atmosphere, with Application to Two High-Resolution General Circulation Models. *Journal of the Atmospheric Sciences* (2013). URL https://journals.ametsoc.org/view/journals/atsc/70/7/jas-d-12-0281.1.xml.

[39] Rasp, S. *et al.* WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *Journal of Advances in Modeling Earth Systems* **16**, e2023MS004019 (2024). URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023MS004019.

[40] Bougeault, P. *et al.* The THORPEX Interactive Grand Global Ensemble. *Bulletin of the American Meteorological Society* (2010). URL https://journals.ametsoc.org/view/journals/bams/91/8/2010bams2853_1.xml.

[41] Daniel Varela, Santoalla & Bojan Kasic. TIGGE archive (2024). URL https://confluence.ecmwf.int/x/JQTs.

[42] Hersbach, H. *et al.* The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* **146**, 1999–2049 (2020). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803.

[43] Google Deepmind. graphcast/docs/cloud_vm_setup.md (2024). URL https://github.com/google-deepmind/graphcast/blob/main/docs/cloud_vm_setup.md.

[44] Wiin-Nielsen, A. & Chen, T.-C. *Fundamentals of Atmospheric Energetics* (Oxford University Press, 1993). Google-Books-ID: 31udXHRjFbIC.

[45] Jakob, A. R., Hack, A. J. & Williamson, A. D. Solutions to the Shallow Water Test Set Using the Spectral Transform Method. *University Corporation for Atmospheric Research* (1993). URL https://opensky.ucar.edu/islandora/object/%3A3430.

[46] Kraichnan, R. H. Inertial Ranges in Two-Dimensional Turbulence. *The Physics of Fluids* **10**, 1417–1423 (1967). URL https://doi.org/10.1063/1.1762301.

[47] Charney, J. G. Geostrophic Turbulence. *Journal of The Atmospheric Sciences* (1971). URL https://journals.ametsoc.org/view/journals/atsc/28/6/1520-0469_1971_028_1087_gt_2_0_co_2.xml.

[48] Lindborg, E. Can the atmospheric kinetic energy spectrum be explained by two-dimensional turbulence? *Journal of Fluid Mechanics* **388**, 259–288 (1999). URL https://www.cambridge.org/core/journals/journal-of-fluid-mechanics/article/abs/can-the-atmospheric-kinetic-energy-spectrum-be-explained-by-twodimensional-turbulence/2835300488AEAA21B17FEF0B372FA3CC.

[49] Nastrom, G. D. & Gage, K. S. A Climatology of Atmospheric Wavenumber Spectra of Wind and Temperature Observed by Commercial Aircraft. *Journal of the Atmospheric Sciences* **42**, 950–960 (1985). URL https://journals.ametsoc.org/view/journals/atsc/42/9/1520-0469_1985_042_0950_acoaws_2_0_co_2.xml.

[50] Niranjan Kumar, K. *et al.* Atmospheric kinetic energy spectra from global and regional NCM-RWF unified modelling system. *Quarterly Journal of the Royal Meteorological Society* **149**, 2784–2799 (2023). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.4531. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.4531.

[51] Gupta, S. Image Edge Detection: A Review. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 7, July 2013* **2** (2013).

[52] Vikram Mutneja, D. Methods of Image Edge Detection: A Review. *Journal of Electrical & Electronic Systems* **04** (2015). URL http://www.omicsgroup.org/journals/methods-of-image-edge-detection-a-review-2332-0796-1000150.php?aid=57249.

[53] Ebert-Uphoff, I. *et al.* Measuring Sharpness of AI-Generated Meteorological Imagery. *Artificial Intelligence for the Earth Systems* (2025). URL https://journals.ametsoc.org/view/journals/aies/4/3/AIES-D-24-0083.1.xml.

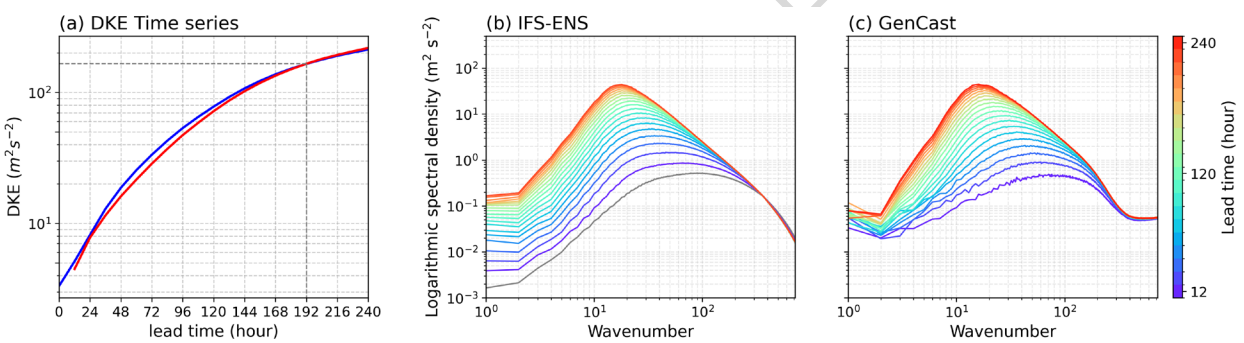


Figure 1: **Difference Kinetic Energy (DKE) of IFS-ENS and GenCast.** (a) The time series of global-averaged DKE. The blue line corresponds to IFS-ENS, and the red line corresponds to GenCast. Gray dashed line marks the intersection between IFS-ENS and GenCAst at $\tau = 192$h. (b) DKE spectra of IFS-ENS and (c) that of GenCast. The gray line in (b) represents DKE of the initial condition of IFE-ENS. The colorbar on the right in (b) and (c) indicates the forecast lead time.

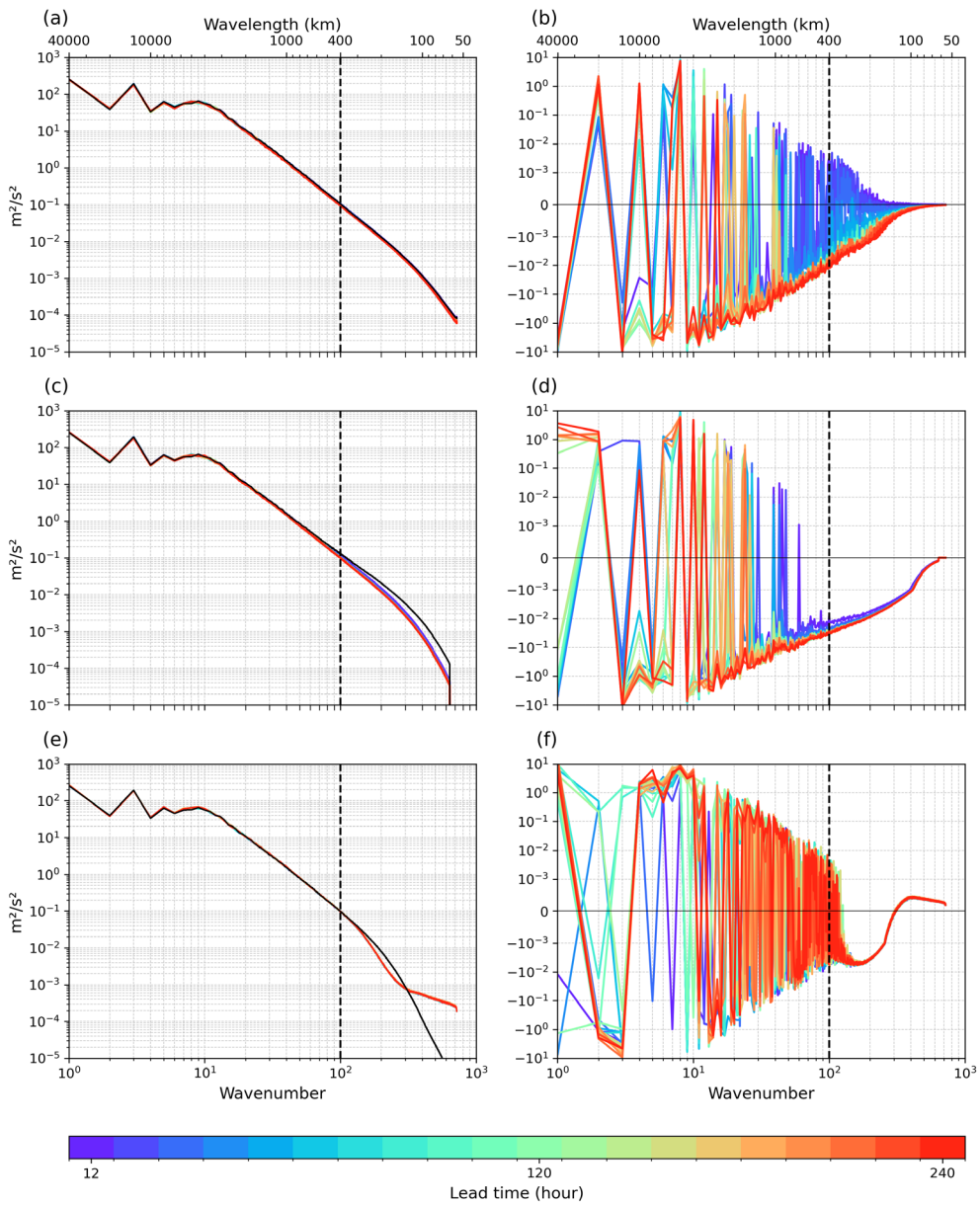


Figure 2: **Kinetic Energy (KE) spectra evolution as function of lead time for three models.** The left column (a, c, e) shows the first ensemble member 300 hPa KE spectra, and the right column (b, d, f) shows spectra change ($\Delta E(k, \tau)$). From top to bottom, rows correspond to IFS-HRES, IFS-ENS, and GenCast. Colors denote lead time from $\tau = 12$h to $\tau = 240$h. Black solid lines in (a, c, e) show the KE spectrum of the initial condition of each model: the spectrum of step 0 for IFS, and that of ERA5 reanalysis for GenCast. Black vertical dashed lines mark the wavenumber $k = 100$ ($\approx$ 400 km), separating synoptic and mesoscale ranges. Ensemble mean results are shown in Figure S3.

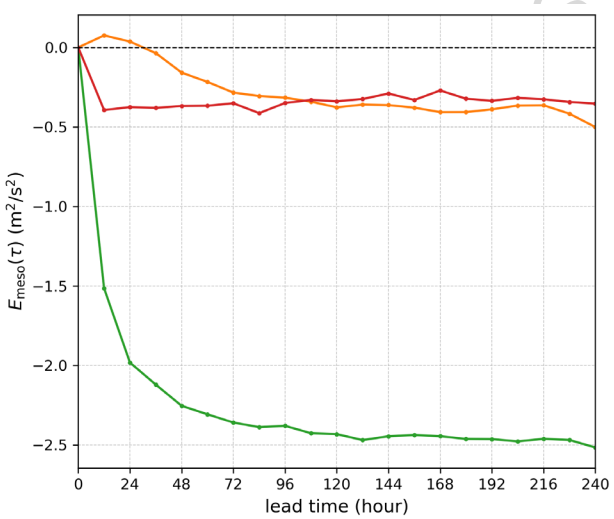


Figure 3: **Time evolution of integrated mesoscale spectral change for the first ensemble member.** The plot quantifies the cumulative energy change within the mesoscale wavenumber band ($k > 100$) over a 10-day forecast. The solid orange line represents the deterministic IFS-HRES forecast, the solid green line indicates the first IFS-ENS ensemble member, and the solid red line shows the results for the first ensemble member of GenCast. Ensemble mean results are shown in Figure S4.

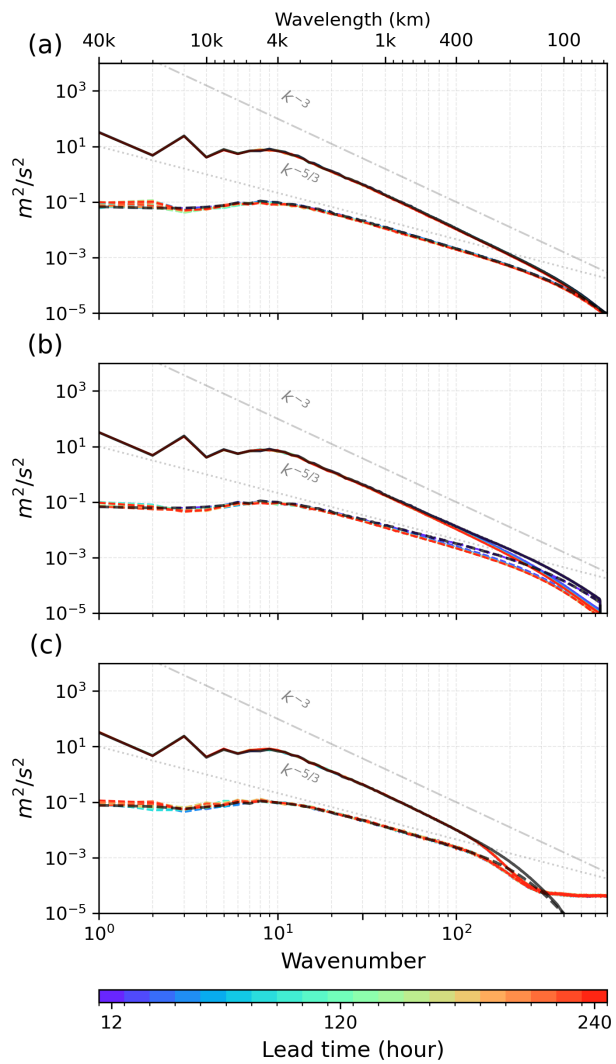


Figure 4: **Rotaional and divergent components of Kinetic Energy (KE) spectra.** Solid lines indicate rotational components and dashed lines indicate divergent components of KE for (a) IFS-HRES, (b) the first ensemble member of IFS-ENS, and (c) GenCast. Black lines for each model present decomposed spectra of the initial condition. Gray dashed lines with slopes of -3 and -5/3 are shown as reference power laws.

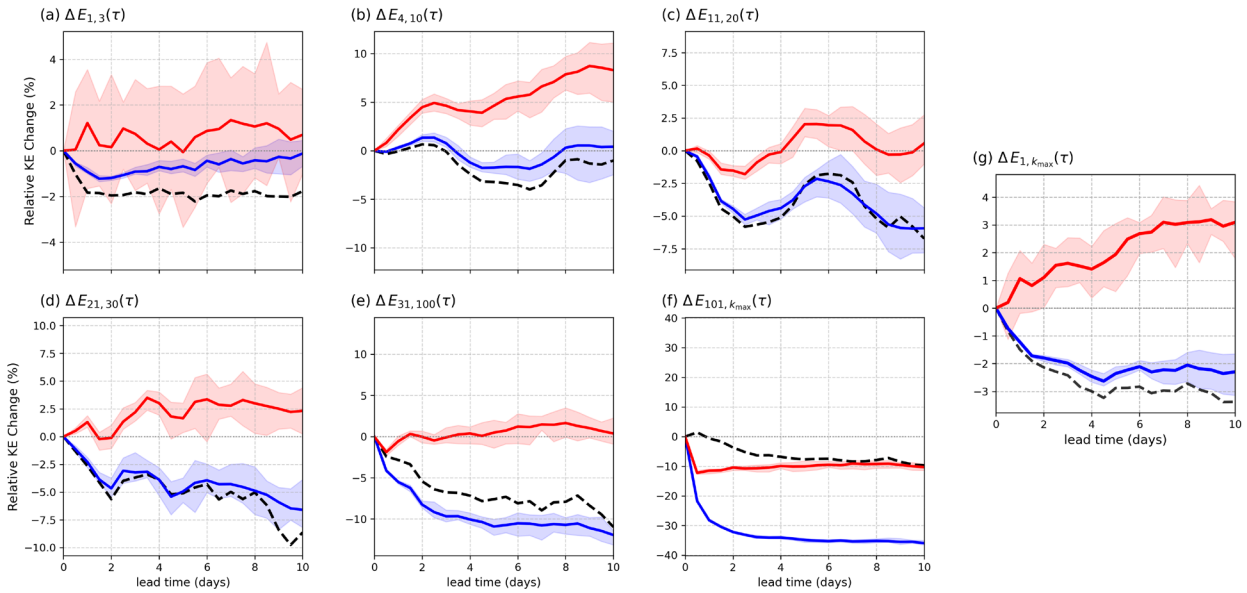


Figure 5: **Relative Kinetic Energy (KE) change against the initial condition.** Black dashed lines represent IFS-HRES, blue solid lines indicate the ensemble mean of IFS-ENS, and red solid lines correspond to the ensemble mean of GenCast. Each panel presents (a-f) integrated KE change for specific wavenumber ($k$) bands and (g) the total integrated KE change. Shaded areas denote the full ensemble spread (min-max).

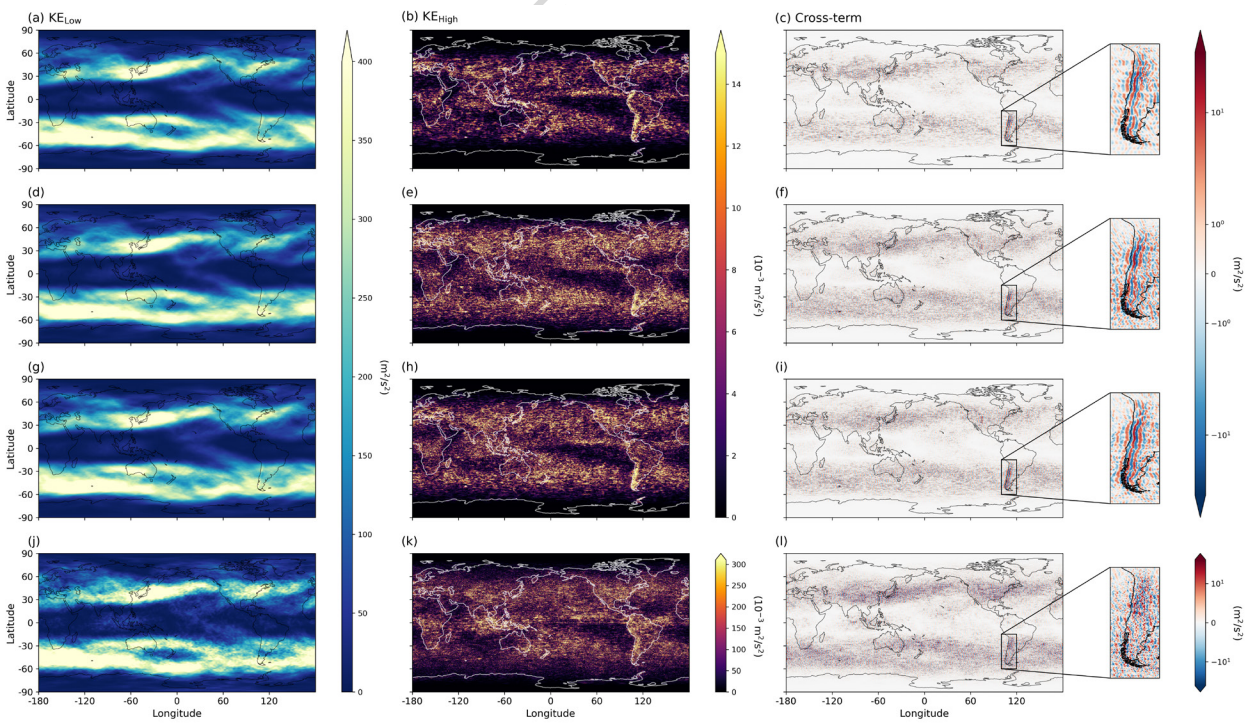


Figure 6: **Spatial decomposition of 300 hPa Kinetic Energy (KE) at lead time $\tau = 240$h** Rows correspond to (top to bottom) ERA5 reanalysis(a-c), IFS-HRES (d-f), the first member of IFS-ENS (g-i), and GenCast (j-l). Columns (from left to right) show $KE_{low}$ ($k \leq 100$), $KE_{high}$ ($k > 100$), and cross-term representing the interaction between these scales, respectively. For the cross-term panels (c, f, i, and l), magnified views of the Andes region are provided as insets to represent atmospheric flow behavior over high terrain.

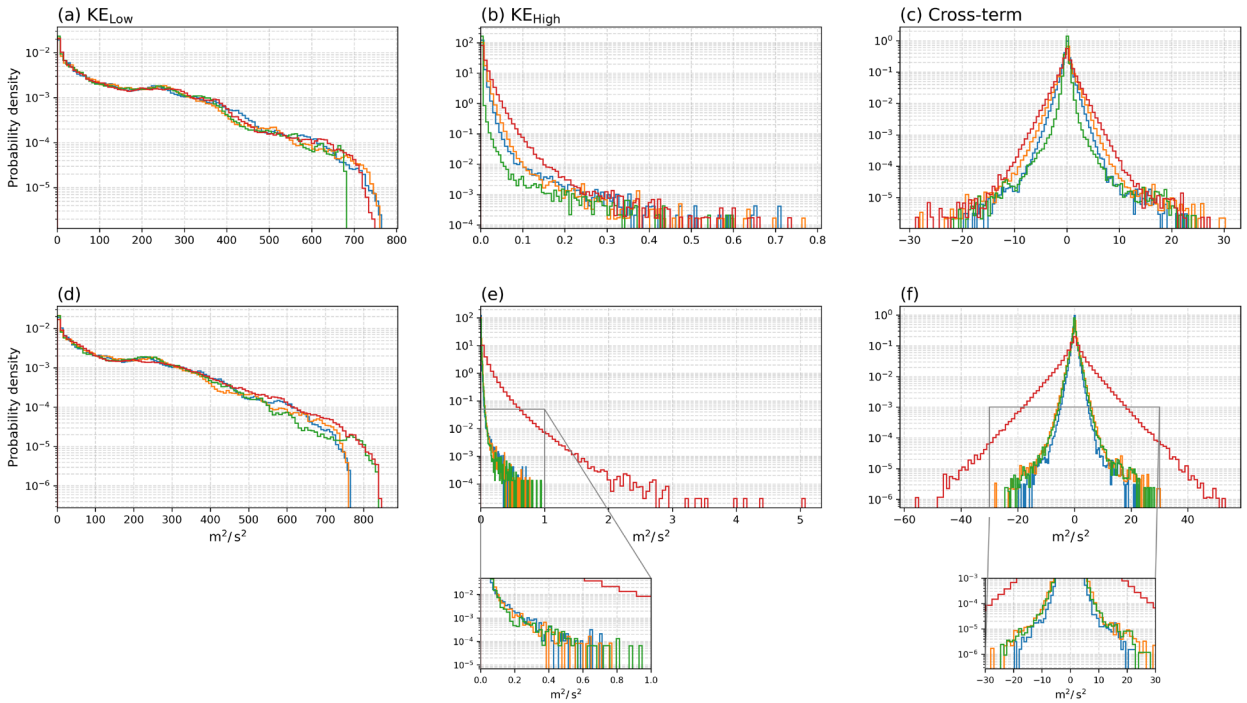


Figure 7: **Probability density functions of the decomposed Kinetic Energy (KE) terms at lead time $\tau = 240$h.** Columns (from left to right) show $\mathrm{KE}_{\mathrm{low}}$, $\mathrm{KE}_{\mathrm{high}}$, and cross-term, respectively. Line colors correspond to the following: ERA5 reanalysis (blue), IFS-HRES (orange), IFS-ENS (green), and GenCast (red). The top row shows the ensemble mean, while the bottom row shows the first ensemble member (corresponding to Figure 6). (e) and (f) include insets that show zoomed-in views of the distributions, allowing for a more detailed comparison.

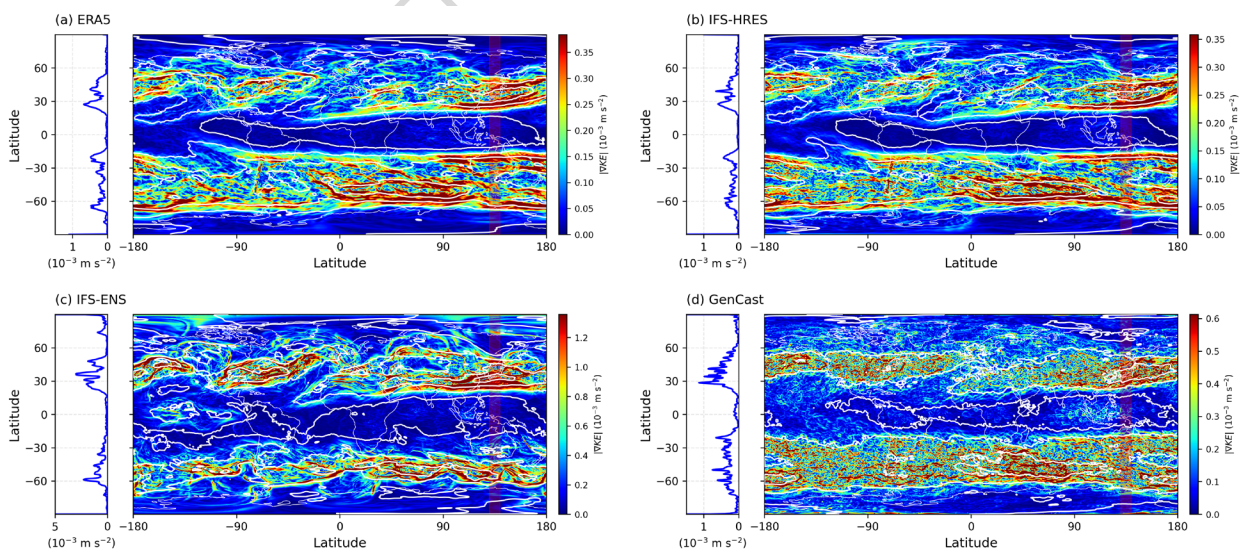


Figure 8: **Magnitude of Kinetic Energy (KE) gradient at 300 hPa at lead time $\tau = 240$h.** The profiles and maps demonstrate the sharpness of KE distribution across (a) ERA5 reanalysis, (b) IFS-HRES, (c) the first member of IFS-ENS, and (d) GenCast. The left profiles show the zonal averaged $|\nabla \mathrm{KE}|$ at longitude band 130.5°-139.5°E (red-highlighted band on global map), while global maps on the right panels display the $|\nabla \mathrm{KE}|$ using color shading. White contours overlaid on the maps denote the 300 hPa zonal wind from the corresponding model. Ensemble mean results are shown in Figure S8.