# Foundational Deep Learning Models for Global Weather Forecasting: A Review

Benjamin Richards
IMT Atlantique
Email: benjamin.richards@imt-atlantique.net

***Abstract*—Recent years have seen a surge of deep learning models for global weather forecasting, challenging the dominance of traditional numerical weather prediction. Models such as FourCastNet, Pangu-Weather, GraphCast, Aurora, and FourCastNetv2 offer competitive skill and major speed-ups. This review critically synthesises these foundational models, focusing on two underexplored dimensions: forecast skill and physical interpretability. We identify three core tensions: (1) high average skill versus reduced operational confidence, (2) physical plausibility versus limited effective resolution, and (3) growing interest in interpretability versus lack of standardised benchmarking. While DL models match or exceed NWP in global metrics (e.g. RMSE at 2–10 day lead times), they consistently underestimate extremes, limiting trust for high-impact events. They capture large-scale dynamics and sensitivity patterns similar to those from NWP adjoints, yet fail to resolve mesoscale variability due to spatial smoothing. Current interpretability efforts remain largely qualitative. We argue for a quantitative, multi-dimensional framework, including latent space analysis and spectral diagnostics, and propose extending benchmarks like WeatherBench to standardise physical realism assessments. These advances are essential for improving forecast reliability and scientific insight.**

## I. Introduction

For decades, numerical weather prediction (NWP) has been the foundation of weather forecasting. These models solve the governing physical equations on discretised grids, and their outputs are used daily for decision-making in sectors such as transportation, agriculture, and disaster risk planning. Accurate forecasts, particularly for extreme weather events, are critical for reducing human and economic losses [1]. In recent years, a paradigm shift has taken place. Deep learning (DL) models have emerged as an alternative to traditional NWP. Several factors have underpinned this transition. First, the volume and quality of reanalysis and forecast data have dramatically improved, particularly with datasets such as ERA5 from European Centre for Medium-Range Weather Forecasts (ECMWF) [2]. Second, modern deep learning frameworks and access to GPU/TPU clusters have made it feasible to train large data-driven forecasting models at global scales. Despite the fast-moving pace of this field, there remains a lack of critical reviews. Existing surveys tend to focus on architectural innovations or high-level application trends [3, 4, 5], without offering a comparative synthesis of model performance and physical consistency. The current review addresses this gap by evaluating five DL-based global forecasting models (FourCastNet [6], Pangu-Weather [7], FourCastNetv2 [8], GraphCast [9], and Aurora [10]) with a focus on forecast skill and physical interpretability. While architectural choices are also discussed, the emphasis is on understanding how model design and training affect real-world performance and alignment with physical principles. As these models are increasingly considered for operational deployment, it is crucial to assess not just their headline metrics, but also their behaviour under domain-specific diagnostics such as error growth, sensitivity to initial conditions, and response to extreme events. We structure the review as follows: In Section II, we describe the input and output formats. Section III provides an overview of the architectures of the five models. In Section IV, we discuss findings from multiple independent studies comparing forecast skill across timescales, variables, and regions. Section V examines the emerging literature on physical interpretability, including sensitivity analyses, error growth behaviour, and model consistency with dynamical principles. Finally, Section VI identifies unresolved tensions and future research directions.

## II. Inputs and Outputs

Most recent DL weather models are pretrained on ERA5 reanalysis, though Aurora additionally incorporates nine other datasets with varying spatiotemporal resolutions, number of variables, and pressure levels. All models take surface and atmospheric variables as input, but PanguWeather and Aurora encode these separately as two tensors. Most models (except FourCastNet and FourCastNetv2) also ingest static inputs such as land–sea masks; these are usually concatenated to surface variables [6, 8, 7, 10].

Forecast targets differ across models. Most predict variables at 6-hour lead times, but PanguWeather consists of four distinct models trained for 1h, 3h, 6h, and 24h lead times. Arbitrary forecast horizons are produced by chaining these models, prioritising longer steps. FourCastNet and Aurora use separate precipitation models, fine-tuned on top of their respective backbones, while PanguWeather does not predict precipitation at all [6, 7, 10].

Importantly, GraphCast's input is projected into a graph representation. Grid points are considered as grid nodes, where each node contains all relevant features: previous time step information, and the atmospheric, surface, and static variables [9].

The mathematical formulation of the learning objective, called the problem statement, also varies. FourCastNet and FourCastNetv2 are pretrained to map $X_t \rightarrow X_{t+1}$, and

fine-tuned to predict both $X_{t+1}$ and $X_{t+2}$ autoregressively, using their own output as input; the training loss includes errors at both steps. PanguWeather also maps $X_t \rightarrow X_{t+1}$ but trains a separate model for each lead time. GraphCast uses two time steps, $X_{t-1}$ and $X_t$, to predict a residual $Y_t$, which is added to $X_t$ to obtain $X_{t+1}$. Aurora uses a similar input configuration but directly predicts $X_{t+1}$ without modeling an intermediate $Y_t$. Despite these differences, all models can be used autoregressively to generate multi-step forecasts [6, 8, 7, 9, 10]. Further input configurations for each model are summarised in Table 1.

## III. Architecture

### A. Encoders

All reviewed models follow an encoder-processor-decoder framework. The encoder's primary role is to transform high-dimensional spatiotemporal input fields into lower-dimensional latent representations suitable for downstream processing. FourCastNet and FourCastNet2 use shallow convolutional encoders. The original FourCastNet applies a single convolution to produce a patch embedding, then applies an additive learnable positional embedding [6], while FourCastNet2 increases depth with two consecutive pointwise convolutions, followed by an additive positional embedding [8]. Similarly, Pangu-Weather uses a convolution layer to produce its patch embedding. However, the surface and atmospheric variables are encoded separately, applying a 2D and 3D convolution, respectively. The static variables are concatenated to the surface variables before encoding. The two resulting embeddings are then concatenated together before being passed to the processor [7]. Aurora, in contrast, implements a more structured treatment of variable types and pressure levels. Similar to Pangu-Weather, separate encoders are applied to atmospheric and surface variables, with each variable encoded using an MLP with variable-specific weights, allowing the model to learn from datasets with different variables. Pressure levels are encoded using additive Fourier embeddings (for atmospheric variables) or a learned vector (for surface variables). These representations are then aggregated through a Perceiver module, where $C_L = 3$ latent query vectors attend over $C$ pressure-level embeddings. This allows the model to learn from datasets with varying pressure levels. After this step, the atmospheric and surface embeddings are concatenated and enriched with a positional embedding, an area-size embedding, and a continuous time embedding (measured as seconds since 1 January 1970) [10]. In GraphCast, the input grid nodes are connected to the processor nodes via unidirectional edges placed between them if the distance is less than 0.6 times the edge length. Embedding is performed by first encoding all features of all nodes and edges in the network. Processor nodes contain three features: the latitude cosine, the longitude sine, and the longitude cosine. All edges contain four features: the edge length and the three values that comprise the 3-dimensional vector difference between its two nodes. A different MLP is applied for each node and edge type. Message passing using a standard interaction network is done for input grid nodes, multi-mesh nodes, and the edges between them. Specifically, the edge MLP takes the edge itself and its two nodes. The multi-mesh node MLP takes the node itself and the sum of all edges terminating at it. The grid node MLP only accepts grid nodes, as there are no edges terminating at these nodes [9].

### B. Processers

The FourCastNet processor comprises 8 blocks, each consisting of an Adaptive Fourier Neural Operator (AFNO) layer followed by an MLP. The AFNO layer performs token mixing by applying the discrete Fourier transform (DFT) to the input, projecting it into the frequency domain. Channel mixing is then conducted via a two-layer MLP with block-diagonal weight matrices shared uniformly across all spatial tokens, reducing the number of trainable parameters. A soft-thresholding shrinkage operation is applied to the MLP output to enforce sparsity in the Fourier domain. Finally, token demixing is achieved by applying the inverse discrete Fourier transform (IDFT) to reconstruct the spatial-domain representation [6].

The main difference between FourCastNetv2 and its predecessor is the use of the spherical harmonic transform (SHT) instead of the DFT. The core idea remains the same: projecting data onto a lower-dimensional spectral basis. However, while the DFT uses sinusoidal bases suited to Euclidean grids, the SHT projects onto spherical harmonics, which are better suited for data defined on a sphere. This is more natural for global weather data, since the Earth is approximately spherical. Additionally, the SHT offers rotational and translational equivariance on the sphere, making it well aligned with the symmetries of the physical domain. FourCastNetv2's processor consists of 12 SHT blocks. The first and last blocks perform spectral upscaling and downscaling, respectively. Within each block, the input is transformed into the spectral domain via the SHT, processed by complex-valued MLPs, and then projected back to physical space via the inverse SHT [8].

The Pangu-Weather processor contains four EarthSpecificLayers, which vary by the number of EarthSpecificBlocks they encapsulate. Downsampling via an MLP is applied between the first two layers, while upsampling of similar design occurs between the last two layers. Each EarthSpecificBlock centres on an attention mechanism adapted for Earth data. Inspired by the Swin Transformer's shifted window attention, this mechanism is extended to operate on 3D data. The Swin Transformer's relative position bias is replaced by an Earth-specific absolute bias that accounts for Earth's physical structure and the dependence of meteorological phenomena on absolute location. This bias modifies distances based on latitude and remains learnable; however, the bias matrix is partitioned into submatrices indexed by pressure levels and latitudes. Longitudes do not directly influence the bias, as they are assumed to be periodic, uniformly distributed, and invariant across latitudes. A bias value is extracted from a specific submatrix, identified by the window's pressure level and latitude indices, which are then combined with the local token positions within the window to compute the final bias index [7].

| Category | FourCastNet | FourCastNetV2 | PanguWeather | GraphCast | Aurora |
|---|---|---|---|---|---|
| **Period** | 1979–2022 | 1979–2015 | 1979–2017 | 1979–2022 | Multiple |
| **Lead Time** | 6h | 6h | 1/3/6/24h | 6h | 6h |
| **Surface** | U10, V10, T2M, SP, MSLP | U10, V10, T2M, MSLP | U10, V10, T2M, MSLP | T2M, U10, V10, MSLP, TP | 2T, U10, V10, MSL |
| **Atmos. Vars** | T, U, V, Z, RH | T, U, V, Z, RH | Z, Q, T, U, V | Z, Q, T, U, V, W | V, T, U, Q, Z |
| **Levels** | 4 | 13 | 13 | 37 | 3, 7, 13 |
| **Static Inputs** | N/A | N/A | Topo., Land, Soil | Geopot., Land, Time | Geopot., Land, Soil, Time |
| **Precipitation** | Yes | No | Yes | Yes | Yes |
| **Dimensions** | 721×1440×20 | 721×1440×73 | 721×1440×5, 1440×721×4 | 721×1440×(5 + 6×37) | Variable |

TABLE I
MODEL INPUTS

Aurora's processor is a 48-layer U-Net organised into multiple stages. The first stage comprises 6 layers, followed by patch merging via an MLP that reduces spatial resolution by a factor of 2. The second stage contains 10 layers, after which another patch merging further reduces resolution. The subsequent two stages include 8 layers each: the first without patch merging, and the second followed by patch splitting via an MLP that doubles the resolution. This is succeeded by 10 layers, another patch splitting, and a final 6 layers to complete the network. Each layer implements a 3D Swin Transformer block performing self-attention within windows of size (2, 12, 6), shifted along all dimensions by half the window size (e.g., (1, 6, 3)). Aurora's Swin Transformers adopt Swin2's residual post-normalization but retain Swin1's standard attention instead of cosine attention. Unlike many ViT architectures, no positional bias is used; positional information is fully encoded within the input embeddings. The use of multiple resolutions prevents fixed positional biases, which require a constant resolution [10].

GraphCast's processor uses a multi-mesh representation built from an icosahedron. Starting with a regular icosahedron composed of 20 triangular faces and 12 nodes, each face is subdivided into four smaller triangles. The original icosahedron is modelled as a bidirectional graph with 60 edges; after subdivision, the resulting mesh has 80 faces and 240 edges. Rather than replacing the original mesh, both levels are combined into a multi-mesh: the first icosahedron contributes 60 edges, the subdivided one 240 edges, totalling 300 edges. The overlap of edges between the original and subdivided meshes facilitates efficient global information mixing. This subdivision process is iterated six times, producing an icosahedron with 81,920 faces. Edge overlap across successive meshes enables representation of long-range relationships. The processor operates exclusively on multi-mesh nodes and edges, with feature definitions detailed in the encoder section. Message passing is performed using a standard interaction network with residual connections. This update process is repeated 16 times, with each layer's MLPs having distinct weights and the output of one layer feeding into the next [9].

### C. Decoders

FourCastNet's decoder consists of a single linear layer [6]. FourCastNetv2 applies two point-wise convolutional layers [8]. In Pangu-Weather, decoding is symmetric to encoding and involves patch recovery via two transposed convolutions, separately applied to surface and atmospheric variables [7]. Aurora's decoder mirrors the encoder in reverse, and includes a Perceiver module to extract atmospheric level representations from the latent space. In this module, sine/cosine embeddings of the pressure levels are used as queries in cross-attention. The latent representations are then mapped to patch embeddings via an MLP, after the Perceiver for atmospheric variables, and directly for surface variables. The decoding MLPs are variable dependent [10]. GraphCast's decoder establishes unidirectional edges between the multi-mesh and output grid nodes. For each grid point, the three multi-mesh nodes forming the corresponding triangle are connected to the grid node, and edge features are constructed as in the multi-mesh. Message passing occurs in two stages: first, edge messages from mesh nodes to grid nodes are computed using an MLP over the edge and node features; second, grid node updates are performed using the node features and the sum of incoming edge messages. A residual connection is used for updating grid nodes, while the mesh-to-grid edges are discarded after use. Finally, each grid node is passed through an MLP, and its output is added to the initial state to yield the next-step prediction [9].

## IV. MODEL SKILL

Metrics such as root mean squared error (RMSE) and anomaly correlation coefficient (ACC) are typically used to measure model skill, defined as a normalised estimate of the model error variance [11, 12]. Model skill is used in in-depth studies as well as high-impact case studies, offering insight into model performance in out-of-distribution regimes.

A recent benchmark by [13] examined four DL models (FourCastNet, Pangu-Weather, GraphCast, FourCastNetv2) alongside six NWP models in forecasting Storm Ciaran, an out-of-sample extratropical cyclone over Western Europe in November 2023. For storm track prediction, (initialised before landfall at 00 UTC on October 31) all DL models reproduced the minimum mean sea level pressure (MSLP), matching the IFS analysis well over the storm's evolution. In contrast, after 06 UTC on November 1, when wind speeds began to intensify, the maximum wind speed was underestimated by all DL models, indicating difficulty in reproducing mesoscale pressure gradients. A second initialisation (00 UTC on November 1) focused on synoptic structure during peak storm development.

DL forecasts reproduced broad frontal structures and key features such as the warm sector and cloud head. Nevertheless, differences emerged. The cloud head curvature in FourCastNetv2 appeared less pronounced, and all DL models exhibited blunted frontal gradients compared to analysis fields. These discrepancies point to challenges in accurately resolving sharp thermodynamic structures, even when large-scale alignment is preserved.

In a recent evaluation, [14] assessed the performance of FourCastNet, Pangu-Weather, and GraphCast against the HRES model across three extreme events: two heatwaves (Pacific Northwest, South Asia) and a Texas winter storm. During the Pacific Northwest heatwave, both NWP and DL models significantly underestimated peak $T_{2M}$ intensity, with forecast RMSEs roughly double their typical 10-day error. FourCastNet exhibited the largest bias during the heatwave peak, while HRES errors exceeded those of Pangu-Weather and GraphCast in post-peak stages. Performance ranking varied by lead time: GraphCast and HRES were the most accurate at short-range forecasts, but no model was consistently dominant. For the South Asian heatwave, heat index (HI), calculated from $T_{2M}$ and relative humidity at 1000hPa pressure level, was used. HI was underestimated by all models, but the DL systems performed notably worse than HRES. This suggests a limitation in how current DL models represent humidity–temperature interactions, especially in moisture-laden regimes. Evaluation of the winter storm was done using the wind chill index ($T_{WC}$), calculated from $T_{2M}$ and $v_{10}$. FourCastNet again showed the largest RMSEs, while all models struggled with $T_{WC}$. Similar to the first heatwave, Pangu-Weather and GraphCast showed smaller errors than HRES in post-peak conditions.

Moving beyond individual case studies, [15] assessed Pangu-Weather's forecast skill across 13 tropical cyclones in the Northwest Pacific, benchmarking it against two NWP systems, IFS and GFS. Forecasts were initialised using three sets of initial conditions (ICs): ECMWF, NCEP, and ERA5. In terms of global RMSE for variables such as geopotential height of 850 hPa, 925 hPa temperature, and MSLP, Pangu-Weather generally outperformed the NWP models across lead times beyond 12 hours. However, it consistently showed inferior accuracy at the shortest lead times. Among the ICs tested, ECMWF-based initialisations yielded the most accurate forecasts, suggesting sensitivity to IC quality. For cyclone track prediction, Pangu-Weather's relative skill improved with lead time, but its 12-hour lead time performance lagged both IFS and GFS. IFS exhibited a smaller median track error overall, indicating a tighter distribution of forecast quality. In contrast, for intensity prediction, Pangu-Weather significantly underestimated peak wind speeds across all cases. This points to a key limitation in ML-based cyclone intensity modeling, possibly linked to resolution constraints or insufficient representation of surface layer processes.

In a similar vein, [16] evaluated FourCastNetv2, GraphCast, Aurora, and Pangu-Weather against operational NWP models (GFS, IFS, UM) in forecasting 50 tropical cyclones across five ocean basins. The mean track error increases with forecast lead time across all ocean basins. Aurora, GraphCast, Pangu-Weather, and FourCastNetv2 consistently outperformed NWP systems in mean track error across most basins. However, the Western Pacific exhibited the steepest rise in track error and the largest inter-model variability, with GFS, IFS, UM, and GraphCast performing comparatively worse, while Aurora, Pangu-Weather, and FourCastNetv2 retained a relative advantage. In terms of intensity, all DL models substantially underestimated maximum sustained wind speeds and minimum MSLP. Whereas the best NWP models predicted central pressures in the 982–988 hPa range, DL forecasts yielded significantly higher values: 996 hPa for FourCastNetv2 and 992 hPa for the others, implying weaker storm representations. DL models also retained warm-core structures but with muted temperature anomalies relative to NWP forecasts. This aligns with their underestimation of storm intensity and suggests a broader limitation in capturing cyclone energetics.

To assess model skill across distinct climate regimes, [17] evaluated 2-metre temperature forecasts from AIFS, GraphCast, and Pangu-Weather on three out-of-sample years representing pre-industrial (1955), present-day (2023), and future warming (2049) conditions. Each model performed a 10-day forecast from daily ICs across the full year in each regime, and RMSE and mean bias were computed to quantify performance. Results indicate that while overall skill remains strong across regimes, systematic biases emerge. In the future climate year (2049), GraphCast exhibited compensating cold and warm biases that approximately cancel in global means, whereas AIFS and Pangu-Weather showed persistent cold biases. Notably, all models developed cold biases over land when initialised from anomalously warm states, suggesting a tendency to revert toward the climatological distribution seen in their training data.

## V. Physical Interpretability

A central question in evaluating DL weather models is whether their forecasts retain physically meaningful structure beyond minimising statistical loss. One method to probe this is spectral energy analysis. [18] investigated the spectral properties of Pangu-Weather forecasts at lead times of 12, 24, and 120 hours, comparing them to IFS forecasts and ERA5 reanalyses across a range of wavenumbers (up to 300) for variables such as wind speed and temperature. While IFS spectra remained relatively consistent with ERA5 across lead times, Pangu-Weather showed a marked drop in spectral energy beyond wavenumbers 60–80, with the discrepancy growing with lead time. This implies a loss of fine-scale variance and reduced spectral fidelity at smaller spatial scales. The authors estimate the model's effective resolution to be closer to 500–700 km, significantly coarser than the nominal 0.25° grid, suggesting that Pangu-Weather generates overly smoothed forecasts at small scales, particularly after 24 hours. The study also compared the spectral behaviour of Pangu-Weather to that of the ensemble mean forecasts from ECMWF's ensemble prediction system. Surprisingly, Pangu-Weather did not resemble the ensemble mean; its spectral

energy was lower and its structure distinct, indicating that the model does not simply average out ensemble spread, but rather produces a smoothed yet dynamically different solution. These results raise concerns about the model's ability to preserve multiscale variability and suggest a need to better retain physically grounded features across lead times.

[19] proposed a novel method to assess physical interpretability in DL weather models by analysing how localised perturbations evolve over time. The study evaluates whether Pangu-Weather can reproduce known dynamical responses to small initial disturbances, despite not being explicitly trained on physical governing equations. To isolate the perturbation response, the authors subtract a steady-state background field, typically defined as a seasonal climatology (e.g., DJF mean from ERA5), from the model's outputs. This allows the evolution of injected anomalies to be tracked independently of large-scale climatological structure. Four canonical scenarios were tested: steady tropical heating (to excite equatorial waves), extratropical cyclone development, geostrophic adjustment (to test recovery of balance), and hurricane evolution. In each case, the perturbation fields generated by Pangu-Weather qualitatively reproduced structural features documented in previous numerical studies. For example, the tropical heating experiment produced westward Rossby gyres consistent with linear wave theory. These results suggest that, at least for certain classes of perturbations, the model internalises aspects of atmospheric dynamics beyond mere statistical correlation, pointing to a form of emergent physical behaviour within the learned system.

[20] examined the sensitivity of PanguWeather to small perturbations in initial conditions, probing whether it exhibits the chaotic behaviour characteristic of the atmosphere (i.e., the butterfly effect). They used ICs from ECMWF's ensemble data assimilation system (EDA), which provides perturbations consistent with realistic analysis uncertainty. Two perturbation amplitudes were tested: 100% (unaltered EDA spread) and 0.01% (perturbations scaled down by a factor of 1000). This setup allows the isolation of intrinsic dynamical instability from external noise. The key diagnostic was the difference kinetic energy (DKE) at 300 hPa, defined as $DKE = var(u) + var(v)$, which measures ensemble spread in horizontal wind and serves as a proxy for perturbation growth. The NWP model included in the study showed expected behaviour: in the 100% case, DKE grew exponentially with lead time, consistent with synoptic instability; in the 0.01% case, DKE also grew rapidly at first, then saturated after 2 days. PanguWeather, in contrast, showed lower DKE in the 100% case and almost identical DKE evolution in both experiments — scaled linearly by the perturbation amplitude, but without divergence due to internal dynamics. This indicates that PanguWeather does not respond to infinitesimal IC differences, failing to reproduce the atmosphere's chaotic growth of uncertainty. In effect, it underestimates the butterfly effect by several orders of magnitude, raising concerns about its physical interpretability.

[21] investigates the physical interpretability of FourCastNetv2 by comparing its gradient-based IC sensitivities with those from the adjoint of a physics-based NWP model. The analysis targets forecast kinetic energy at 36h, computed over the Bay of Biscay as the sum of squared zonal and meridional winds. Sensitivities are computed via backpropagation using PyTorch's automatic differentiation engine, yielding $\partial K/\partial IC$ fields for various input variables (e.g., temperature, water vapor, and meridional wind). Across several variables, FourCastNetv2 and the adjoint model produce qualitatively similar sensitivity fields. Spatial patterns are largely consistent, and while FourCastNetv2 tends to underestimate local sensitivity magnitudes, the regionally aggregated values are comparable. In geopotential height, FourCastNetv2 exhibits wave-like sensitivity structures aligned with expected synoptic-scale propagation patterns, reinforcing its ability to capture coherent dynamical features. Notably, both models exhibit similar sensitivities in water vapor: when comparing FourCastNetv2's specific humidity gradients with the NWP model's relative humidity fields, spatial alignment is observed. However, FourCastNetv2 also exhibits nontrivial positive sensitivities in upper-tropospheric humidity, which are absent in the adjoint model and lack a clear physical basis. These unexplained features may reflect artefacts of the model's training or an overfitting to spurious statistical correlations, pointing to potential limitations in its physical consistency.

## VI. Discussion and Conclusion

Our review of recent ML-based weather forecasting models reveals three tensions: (1) high average model skill versus reduced operational confidence, (2) physical plausibility versus limited effective resolution, and (3) interpretability versus the absence of standardised benchmarking.

### *A. Skill vs operational confidence*

While ML models have demonstrated superior or comparable performance to NWP in terms of global RMSE at 2-10 day lead times across surface temperature, wind speed, and other variables [17, 15, 16], these metrics do not guarantee reliability in operational forecasting. For instance, [15] found that more NWP ensemble members exhibited smaller errors than their DL counterparts, even when the DL ensemble mean was more accurate. Furthermore, DL models consistently underestimate extreme weather intensities, particularly in maximum wind speed [13, 15, 16] and temperature extremes [14, 15, 16]. For example, [13] showed that during Storm Ciarán the DL models predicted peak wind speeds of 90km/h while IFS analyses reached 130km/h, an underestimation of over 40km//h. Such systematic underestimations compromises their operational value during high-impact events. Notably, this underestimation is not present in ERA5, the reanalysis dataset used to train the model, suggesting that the bias arises from the ML model itself rather than its training data [13]. [16] attributes this to excessive spatial smoothing, which damps sub-synoptic and mesoscale structures vital for capturing extremes [18].

### *B. Physical plausibility vs resolution*

The effective resolution of DL models—i.e., the smallest spatial scale at which meaningful variability is captured—is

often much coarser than their grid spacing. [18] used spectral energy diagnostics to show that although PanguWeather is trained on 0.25° grids, its effective resolution is closer to 4.5°. Despite this limitation, ML models learn and reproduce synoptic scale structures. [13] showed that jet stream positioning and frontal boundaries of a cyclone are well captured by the ML model. [19] demonstrated that tropical heating anomalies in Pangu-Weather propagate as nearly stationary Rossby waves under perturbation, in line with NWP dynamics. [21] found that sensitivity fields from backpropagation in the ML model closely resembled those from the COAMPS adjoint model, implying that the ML model learned gradients align with physical dynamics. These findings point to a key tension: while ML models display global physical plausibility, they lack the representational capacity to reproduce mesoscale and sub-synoptic structures.

### C. Interpretability vs benchmarking

Several studies have attempted to evaluate whether DL models have learnt physical dynamics, rather than simply optimising a loss function. The term "physical realism" is commonly found in the literature, yet its definition is unclear. From the literature reviewed, physical realism is a multi-dimensional construct with axes such as: (1) statistical similarity to NWP, as exemplified by [21] in the comparison of sensitivity fields, and also by [18] in comparing energy spectra; and (2) dynamical behaviour under perturbation, such as steady state responses [19], and chaotic error growth [20]. However, current analyses remain largely qualitative and limited in scope. For example, [21] only provided a visual comparison of sensitivity fields, without quantitative metrics. While methodological limitations are present, such as differences in the variables represented by each model, robust quantitative tools are needed. Since these fields are spatial images, metrics like the Structural Similarity Index Measure [22] could provide objective evaluations. Similarly, wave-like patterns observed in DL forecasts [21], possibly linked to Rossby wave packets, could be analysed using spectral diagnostics or empirical orthogonal function analysis [23]. These gaps in interpretability and quantitative rigor underscore the need for a unified framework that can measure, and ultimately improve, the physical realism of DL weather models.

Interpreting latent spaces is one promising direction toward this goal, but it is currently unexplored for enhancing physical interpretability. Latent representations are lower-dimensional embeddings that capture essential information from high-dimensional input data [24]. In computer vision, similarity maps have been used to understand what features a network has learned [25][26], while in natural language processing, latent concept attribution identifies and clusters latent concepts within the model's neural representations [27]. To our knowledge, no studies have analysed latent representations in DL weather architectures. A key question is whether latent structures or internal activations of data-driven weather models can reveal physically meaningful concepts. Outside weather forecasting, [28] applied symbolic regression to embedded messages of a graph neural network trained to predict instantaneous acceleration of a particle system. The results showed that the messages were highly correlated with the known force dynamics, allowing the messages to be interpreted as forces. More recent work by [29] developed a framework for detecting closed-form interpretations of latent spaces, with the technique being successfully applied to systems under Lorentz transformations. These methods may enable the discovery of correlations between latent spaces and known physical fields. Additionally, many DL architectures include visual transformers, suggesting that attention maps could provide insights into causal roles of internal representations in forecasts [30]. While standardising latent space quantification would be challenging, these methods hold potential to constitute an additional dimension to physical realism in data-driven weather models.

A well defined multi-dimensional definition of physical realism based on quantitative assessments is needed to increase the effective resolution of data-driven weather models. Our interpretability discussion leads directly to the need for standardised benchmarks. Current benchmarks such as WeatherBench [31] include energy spectra as a quantitative physical metric, but this is insufficient for comprehensive evaluation. This highlights the third tension: while benchmarking model skill is widespread, and limitations concerning weather intensities are well-known, efforts to quantify and standardise physical interpretability are limited. We propose developing WeatherBench to include a comprehensive view of physical realism. This will enable systematic assessment and improvement of DL weather models' ability to represent mesoscale and sub-synoptic features accurately. This advancement will not only deepen scientific understanding but also enhance operational trust and improve extreme event forecasting.

## References


[1] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," *Nature*, vol. 525, no. 7567, pp. 47–55, 2015. [Online]. Available: https://doi.org/10.1038/nature14956

[2] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803

[3] S. A. Siddiqui, J. Kossaifi, B. Bonev, C. Choy, J. Kautz, D. Krueger, and K. Azizzadenesheli, "Exploring the design space of deep-learning-based weather forecasting

systems," 2024. [Online]. Available: https://arxiv.org/abs/2410.07472

[4] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019. [Online]. Available: https://doi.org/10.1038/s41586-019-0912-1

[5] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021. [Online]. Available: https://doi.org/10.1038/s42254-021-00314-5

[6] J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, P. Hassanzadeh, K. Kashinath, and A. Anandkumar, "Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators," 2 2022. [Online]. Available: http://arxiv.org/abs/2202.11214

[7] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3d neural networks," *Nature*, vol. 619, pp. 533–538, 7 2023.

[8] B. Bonev, T. Kurth, C. Hundt, J. Pathak, M. Baust, K. Kashinath, and A. Anandkumar, "Spherical fourier neural operators: Learning stable dynamics on the sphere," 2023. [Online]. Available: https://arxiv.org/abs/2306.03838

[9] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia, "Graphcast: Learning skillful medium-range global weather forecasting," 12 2022. [Online]. Available: http://arxiv.org/abs/2212.12794

[10] C. Bodnar, W. P. Bruinsma, A. Lucic, M. Stanley, A. Vaughan, J. Brandstetter, P. Garvan, M. Riechert, J. A. Weyn, H. Dong, J. K. Gupta, K. Thambiratnam, A. T. Archibald, C.-C. Wu, E. Heider, M. Welling, R. E. Turner, and P. Perdikaris, "A foundation model for the earth system," 5 2024. [Online]. Available: http://arxiv.org/abs/2405.13063

[11] E. Blanchard-Wrigglesworth, R. I. Cullather, W. Wang, J. Zhang, and C. M. Bitz, "Model forecast skill and sensitivity to initial conditions in the seasonal sea ice outlook," *Geophysical Research Letters*, vol. 42, no. 19, pp. 8042–8048, 2015. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL065860

[12] R. D. Hetland, "Event-driven model skill assessment," *Ocean Modelling*, vol. 11, no. 1, pp. 214–223, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S146350030500003X

[13] A. J. Charlton-Perez, H. F. Dacre, S. Driscoll, S. L. Gray, B. Harvey, N. J. Harvey, K. M. Hunt, R. W. Lee, R. Swaminathan, R. Vandaele, and A. Volonté, "Do ai models produce better weather forecasts than physics-based models? a quantitative evaluation case study of storm ciarán," *npj Climate and Atmospheric Science*, vol. 7, 12 2024.

[14] O. C. Pasche, J. Wider, Z. Zhang, J. Zscheischler, and S. Engelke, "Validating deep learning weather forecast models on recent high-impact extreme events," *Artificial Intelligence for the Earth Systems*, vol. 4, no. 1, p. e240033, 2025. [Online]. Available: https://journals.ametsoc.org/view/journals/aies/4/1/AIES-D-24-0033.1.xml

[15] Y. Shi, R. Hu, N. Wu, H. Zhang, X. Liu, Z. Zeng, J. Zhu, P. Han, C. Luo, H. Zhang, J. He, and X. Shi, "Comparison of ai and nwp models in operational severe weather forecasting: A study on tropical cyclone predictions," *Journal of Geophysical Research: Machine Learning and Computation*, vol. 2, no. 2, p. e2024JH000481, 2025, e2024JH000481 2024JH000481. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024JH000481

[16] P. L. Sahu, S. Sandeep, and H. Kodamana, "Evaluating global machine learning models for tropical cyclone dynamics and thermodynamics," *Journal of Geophysical Research: Machine Learning and Computation*, vol. 2, no. 2, p. e2025JH000594, 2025, e2025JH000594 2025JH000594. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025JH000594

[17] T. Rackow, N. Koldunov, C. Lessig, I. Sandu, M. Alexe, M. Chantry, M. Clare, J. Dramsch, F. Pappenberger, X. Pedruzo-Bagazgoitia, S. Tietsche, and T. Jung, "Robustness of ai-based weather forecasts in a changing climate," 2024. [Online]. Available: https://arxiv.org/abs/2409.18529

[18] M. Bonavita, "On some limitations of current machine learning weather prediction models," *Geophysical Research Letters*, vol. 51, 6 2024.

[19] G. J. Hakim and S. Masanam, "Dynamical tests of a deep learning weather prediction model," *Artificial Intelligence for the Earth Systems*, vol. 3, no. 3, p. e230090, 2024. [Online]. Available: https://journals.ametsoc.org/view/journals/aies/3/3/AIES-D-23-0090.1.xml

[20] T. Selz and G. C. Craig, "Can artificial intelligence-based weather prediction models simulate the butterfly effect?" *Geophysical Research Letters*, vol. 50, no. 20, p. e2023GL105747, 2023, e2023GL105747 2023GL105747. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023GL105747

[21] J. Baño-Medina, A. Sengupta, J. D. Doyle, C. A. Reynolds, D. Watson-Parris, and L. D. Monache, "Are ai weather models learning atmospheric physics? a sensitivity analysis of cyclone xynthia," *npj Climate and Atmospheric Science*, vol. 8, 12 2025.

[22] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13,

no. 4, pp. 600–612, 2004.

[23] G. N. Kiladis, J. Dias, and M. Gehne, "The relationship between equatorial mixed rossby–gravity and eastward inertio-gravity waves. part i," *Journal of the Atmospheric Sciences*, vol. 73, no. 5, pp. 2123 – 2145, 2016. [Online]. Available: https://journals.ametsoc.org/view/journals/atsc/73/5/jas-d-15-0230.1.xml

[24] Y. Liu, E. Jun, Q. Li, and J. Heer, "Latent space cartography: Visual analysis of vector space embeddings," *Computer Graphics Forum (Proc. EuroVis)*, 2019. [Online]. Available: https://idl.uw.edu/papers/latent-space-cartography

[25] P. E. Rauber, S. G. Fadel, A. X. Falcão, and A. C. Telea, "Visualizing the hidden activity of artificial neural networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 101–110, 2017.

[26] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. Chau, "Activis: Visual exploration of industry-scale deep neural network models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 88–97, 2018.

[27] X. Yu, F. Dalvi, N. Durrani, M. Nouri, and H. Sajjad, "Latent concept-based explanation of nlp models," 2024. [Online]. Available: https://arxiv.org/abs/2404.12545

[28] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho, "Discovering symbolic models from deep learning with inductive biases," Red Hook, NY, USA, 2020.

[29] Z. Patel and S. J. Wetzel, "Closed-form interpretation of neural network latent spaces with symbolic gradients," 2025. [Online]. Available: https://arxiv.org/abs/2409.05305

[30] M. Chung, J. B. Won, G. Kim, Y. Kim, and U. Ozbulak, *Evaluating Visual Explanations of Attention Maps for Transformer-Based Medical Imaging*. Springer Nature Switzerland, 2025, p. 110–120. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-77610-6_11

[31] S. Rasp, S. Hoyer, A. Merose, I. Langmore, P. Battaglia, T. Russell, A. Sanchez, V. Yang, R. Carver, S. Agrawal, M. Chantry, Z. B. Bouallegue, P. Dueben, C. Bromberg, J. Sisk, L. Barrington, A. Bell, and F. Sha, "Weatherbench 2: A benchmark for the next generation of data-driven global weather models," *arXiv*, 2023.