

Application of Artificial Intelligence in Weather Prediction: A Mini-Review of Data-Driven Global Forecasting Models

Sreynith Lis¹, Rathbreaksmey Tha², Peter Mafimisebi^{3*}, Sal Sabila⁴, Happy Mwanga⁵, Christian Wadjou⁶, Muhammad Saeed⁷, Fakhar Zulfiqar Ali⁸, Jan Prince Noman⁹, Yusuf Usman¹⁰, Saif Ul Rahman¹¹ & Thomas Mwabobo¹²

¹⁻¹²The School of Energy and Environmental Engineering, Department of Environmental Science and Engineering, University of Science and Technology Beijing, Beijing 100083, China. Corresponding Author Email: mafimisebipeter2023@gmail.com*



DOI: <https://doi.org/10.46759/iijsr.2026.10310>

Copyright © 2026 Sreynith Lis et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Article Received: 20 July 2026

Article Accepted: 25 September 2026

Article Published: 30 September 2026

ABSTRACT

Over the past four years, artificial intelligence (AI)-based data-driven models have moved from proof-of-concept demonstrations to systems that consistently outperform operational numerical weather prediction (NWP) on standard verification benchmarks. This mini-review synthesises 62 peer-reviewed publications from 2020 to 2026 to assess the current state of AI weather prediction across four thematic areas: neural network architectures, quantitative benchmark performance, downstream applications, and structural limitations. Four principal architecture families, namely Fourier Neural Operators, Graph Neural Networks, Transformer-based models, and diffusion-based generative systems, have been demonstrated at global scale, producing ten-day forecasts at 0.25-degree resolution in under one minute on commodity graphics processing unit (GPU) hardware. Evaluated against the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System High Resolution (IFS-HRES) operational forecast using the WeatherBench 2 framework, leading models including GraphCast, Pangu-Weather, FuXi, Aurora, and FengWu outperform IFS-HRES at five-day lead times for 500 hPa geopotential height, with root mean square error improvements of 8 to 14 percent relative to the operational benchmark. The probabilistic model GenCast further outperforms the ECMWF ensemble system on 97.4 percent of probabilistic verification targets. Beyond standard benchmarking, AI models demonstrate consistent advantages for tropical cyclone track prediction and precipitation nowcasting, but exhibit systematic underperformance for record-breaking surface extremes, tropical cyclone intensity, and convective-scale precipitation, limitations that reflect training exclusively on the ECMWF Reanalysis v5 (ERA5) dataset, the absence of physical conservation laws, and incomplete extreme-event evaluation frameworks. Emerging hybrid physics-AI architectures, end-to-end data assimilation systems, and foundation models trained on diverse multi-source datasets represent the most viable pathways to addressing these limitations and realising the operational potential of AI weather prediction.

Keywords: Artificial Intelligence; Weather Prediction; Deep Learning; Numerical Weather Prediction; GraphCast; Pangu-Weather; Machine Learning; Atmospheric Forecasting; WeatherBench; Foundation Models.

1. Introduction

Accurate weather forecasting is one of the most consequential scientific and operational endeavours of the modern era. Forecasts of temperature, precipitation, wind, and extreme events underpin decision-making across agriculture, energy, aviation, emergency management, and public health, with the economic value of improved forecast skill estimated in the hundreds of billions of dollars annually at the global scale [1, 2]. This economic calculus is itself shifting, as AI methods begin to reshape the cost structure of forecast production [3]. For more than five decades, this forecasting infrastructure has been built on numerical weather prediction (NWP), a physics-based computational approach that integrates the partial differential equations governing atmospheric dynamics forward in time from an observed initial state. The European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS), widely regarded as the global benchmark of forecast quality, exemplifies the maturity of this approach: through decades of sustained scientific investment in data assimilation, model physics, and high-performance computing, IFS has progressively extended the useful forecast horizon from around three days in the 1980s to beyond seven days today, a level of continual refinement documented in successive operational verification reports [1, 4, 5]. Despite these advances, operational NWP remains computationally intensive, requiring dedicated supercomputer infrastructure and several hours of wall-clock time to deliver a single global

forecast, a constraint that limits forecast frequency, accessibility, and the ability to run large probabilistic ensembles at high resolution.

The past decade has witnessed a profound transformation in the field of weather prediction driven by the convergence of three enabling developments: the accumulation of long-term, high-quality atmospheric reanalysis datasets, most notably the ERA5 global reanalysis produced by ECMWF [4]; rapid advances in deep learning architectures and computational hardware that have made it feasible to train large neural networks on global atmospheric datasets; and the establishment of community benchmarking frameworks that enable systematic, reproducible comparison between data-driven and physics-based approaches [6, 7]. Early investigations demonstrated that machine learning models could learn skilful predictions of large-scale atmospheric patterns from reanalysis data [8, 9], while subsequent work showed that deep learning could represent subgrid physical processes in climate models with substantially greater computational efficiency than traditional parameterisation schemes [10, 11]. These foundational results, though insufficient in themselves to challenge operational NWP, established the methodological and infrastructure basis for the more ambitious global forecasting models that followed, a convergence of machine learning and earth-system science that contemporaneous reviews also identified as a turning point for the field [12].

The period from 2022 to 2023 marked a qualitative transition in the capabilities of AI weather prediction, with multiple modelling groups independently demonstrating that deep learning systems trained on ERA5 could produce global medium-range forecasts of comparable or superior quality to IFS-HRES at a fraction of the computational cost. FourCastNet [13], based on Adaptive Fourier Neural Operators [14], was among the first models to demonstrate GPU-accelerated global forecasting at 0.25-degree resolution. Pangu-Weather [15] followed soon after as the first AI model to clearly outperform IFS-HRES on most standard meteorological variables, and GraphCast [16] then set a performance benchmark of its own, outperforming IFS-HRES on the large majority of WeatherBench 2 verification targets. This progress continued in 2023 with FengWu [17] and FuXi [18], which pushed skilful forecasts out beyond ten and fifteen days, respectively, and with two further advances: GenCast [19], a diffusion-based probabilistic system that outperformed the ECMWF ensemble on the large majority of its verification targets, and Aurora [20], a foundation model trained on over one million hours of diverse atmospheric data at 0.1-degree resolution. FuXi Weather [21] then became the first end-to-end AI system to produce global forecasts directly from satellite observations, without any NWP-generated initial conditions.

The rapid proliferation of AI weather models has generated a substantial primary literature evaluating individual systems against operational NWP benchmarks. Comprehensive reviews that systematically compare architectures, quantify performance across variables and lead times, assess applications beyond standard medium-range forecasting, and critically evaluate remaining limitations, however, remain limited. Schultz et al. [22] and Bochenek and Ustrnul [23] provided early reviews of machine learning in weather prediction, but predate the transformative advances of 2022 to 2023. Zhang et al. [24] and Waqas et al. [25] have offered more recent surveys of ML methods for weather forecasting, but do not systematically integrate the post-2023 generation of foundation models, probabilistic systems, and end-to-end data assimilation frameworks that now define the field's frontier.

There remains a need for a review spanning the full arc of AI weather model development, evaluating performance quantitatively across the WeatherBench 2 framework, assessing applications including extreme events, tropical cyclones, nowcasting, and sector-specific forecasting, and placing current capabilities in the context of emerging hybrid physics-AI systems.

This review addresses that gap by providing a systematic and critical synthesis of AI-based global weather prediction from 2020 to 2026, based on 62 peer-reviewed publications identified through a structured search of Scopus, Web of Science, and Google Scholar.

1.1. Study Objectives

The aim of this study is to provide a systematic and critical synthesis of AI-based global weather prediction from 2020 to 2026. The specific objectives of the study are to:

- 1) Survey the principal neural network architecture families (Fourier Neural Operators, Graph Neural Networks, Transformer-based models, diffusion models, and hybrid physics-AI systems) and relate architectural choices to forecast skill and operational properties.
- 2) Present a quantitative comparison of AI model performance against IFS-HRES using standardised WeatherBench 2 metrics across multiple variables and lead times from one to fifteen days.
- 3) Review the performance of AI models in four application domains (extreme event prediction, tropical cyclone forecasting, precipitation nowcasting, and extended-range and sector-specific forecasting) where the performance characteristics of AI and NWP approaches diverge substantially from global aggregate benchmarks.
- 4) Critically evaluate the structural limitations of current AI weather models, including training data constraints, the absence of physical conservation laws, evaluation framework deficiencies, and dependence on NWP-generated initial conditions, and review the emerging hybrid and end-to-end architectures that address these limitations.
- 5) Synthesise the evidence on architectures, benchmark performance, applications, and limitations into a set of forward-looking priorities and recommendations for future research and operational adoption of AI-based weather prediction.

Taken together, these objectives are intended to give forecasting centres, model developers, and policymakers an evidence base for judging where AI weather models are ready to complement or replace operational NWP, and where continued investment in physics-based methods remains warranted.

2. Methodology

2.1. Literature Search Strategy

This review adopted a systematic search strategy conducted across Google Scholar, Web of Science, and Scopus, covering peer-reviewed journal articles, indexed conference proceedings (ICML, ICLR, NeurIPS, PASC), and indexed technical reports published in English between January 2020 and June 2026. The lower temporal bound was set to 2020 to capture the period in which data-driven global weather models first demonstrated skill approaching operational NWP, coinciding with the release of WeatherBench [6]. Two keyword streams were

applied in parallel. The first targeted global AI weather prediction using: (“AI” OR “machine learning” OR “deep learning”) AND (“weather prediction” OR “weather forecasting” OR “numerical weather prediction”) AND (“global” OR “medium-range” OR “data-driven”). The second targeted specific model families using: (“GraphCast” OR “Pangu-Weather” OR “FourCastNet” OR “FengWu” OR “FuXi” OR “GenCast” OR “Aurora” OR “NeuralGCM” OR “AIFS”) AND (“weather” OR “forecasting”). Results were merged and deduplicated, and additional records were identified through forward and backward citation tracing of high-impact papers and manual inspection of key journals including Nature, Science, npj Climate and Atmospheric Science, and the Bulletin of the American Meteorological Society.

Records were identified across three databases (Google Scholar, Web of Science, Scopus), screened at title/keyword and abstract stages, and assessed for full-text eligibility against the inclusion and exclusion criteria described in Section 2.2, as summarised in the PRISMA diagram (Figure 1). A total of 62 peer-reviewed studies were retained for qualitative synthesis. These were classified into three thematic groups, core AI model and architecture papers ($n = 28$), evaluation and benchmark studies ($n = 18$), and application and limitation studies ($n = 16$), of which the 28 architecture and benchmark studies contributed the quantitative data presented in Tables 1 to 3, while the remaining studies provided architectural, methodological, or contextual support.

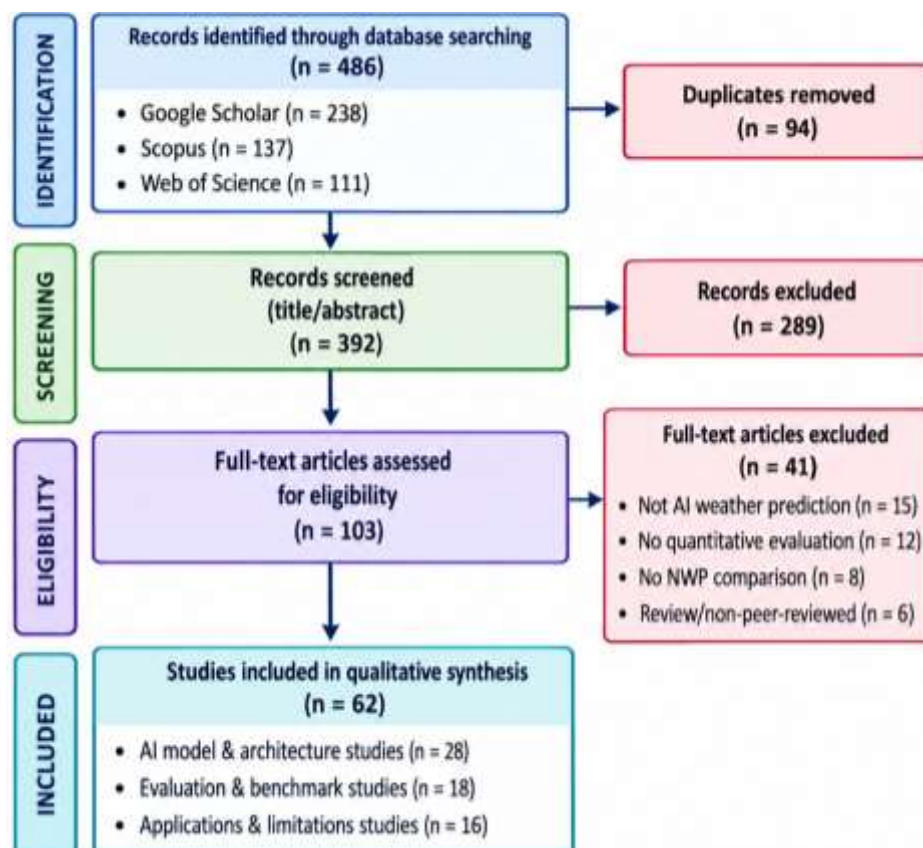


Figure 1. PRISMA flow diagram for the systematic literature search and study selection process.

2.2. Inclusion Criteria, Screening, and Classification

Studies were included if: (i) the primary subject was an AI or machine learning model applied to atmospheric prediction at regional or global scale; (ii) the study was published in a peer-reviewed journal or indexed conference

proceedings; (iii) the study reported quantitative performance metrics such as RMSE or ACC; and (iv) the study compared model performance against an operational NWP reference or standard benchmark dataset. Studies were excluded if the scope was limited to local station-based forecasting without broader context; if the primary application was outside weather prediction without NWP comparison; or if the paper was a review, thesis, or non-peer-reviewed document. Background papers on foundational deep learning architectures, Transformers [26], Fourier Neural Operators [27], and Graph Convolutional Networks [28], were retained to contextualise architectural choices. The PRISMA flow diagram is presented as Figure 1.

3. Results

3.1. From Numerical Weather Prediction to Data-Driven Forecasting

Weather prediction has long represented one of the most consequential challenges in applied science. Operational NWP solves the governing equations of atmospheric motion on a discretised global grid, ingesting millions of observations daily through four-dimensional variational data assimilation. As noted above, sustained investment in this approach has pushed the ECMWF Integrated Forecasting System's useful forecast horizon out to beyond seven days [1], yet this skill carries a substantial cost: IFS-HRES requires thousands of computing cores and several hours of continuous processing per forecast, which limits accessibility in data-sparse regions [29].

The publication of ERA5 changed this picture fundamentally. Spanning 1940 to the present at hourly 0.25-degree resolution across 137 pressure levels, ERA5 provided a corpus large and consistent enough to make data-driven global weather modelling genuinely tractable for the first time [4]. Early machine learning approaches were encouraging but fell short of operational standards: convolutional and residual networks could learn skilful large-scale pressure patterns from historical data, yet remained clearly inferior to IFS-HRES beyond two to three days [8, 9].

Progress accelerated when the community adopted a shared measurement standard. WeatherBench [6] and WeatherBench 2 [7] established common evaluation metrics, RMSE and ACC, computed on fixed test periods across 1,380 verification targets, making performance claims reproducible and directly comparable against IFS-HRES. This benchmarking infrastructure, combined with the adoption of Transformer, Graph Neural Network, and Fourier Neural Operator architectures, catalysed the period from 2022 to 2023 in which multiple AI models independently surpassed the operational benchmark. The following subsections examine the architectures enabling these advances, their quantitative performance, and the limits of their advantage.

3.2. Neural Network Architectures for AI Weather Prediction

The rapid progress of data-driven weather prediction from 2022 onwards is largely due to the adoption of deep learning architectures originally developed for image recognition, natural language processing, and operator learning. Four principal architecture families have been applied to global weather forecasting, each with distinct inductive biases, computational properties, and forecast skill profiles: Fourier Neural Operators, Graph Neural Networks, Transformer-based models, and diffusion-based generative models. These are reviewed in turn below, followed by a summary comparison of the fourteen models identified in this review (Table 1).

3.2.1. Fourier Neural Operators

The Fourier Neural Operator (FNO), introduced by Li et al. [27], learns mappings between function spaces by operating directly in the spectral domain, a natural fit for atmospheric data that are conventionally represented on a spectral grid. FNO layers apply learned linear transformations to the Fourier coefficients of the input field and combine them with local pointwise operations, enabling efficient learning of long-range spatial dependencies without the quadratic cost of standard attention. FourCastNet [13] pioneered the application of an adaptive Fourier Neural Operator (AFNO) to global weather prediction, training on 73 ERA5 variables at 0.25-degree resolution and producing six-hourly global forecasts in under two minutes on a single graphics processing unit (GPU), which was orders of magnitude faster than IFS-HRES. Although its benchmark skill was comparable to IFS-HRES rather than superior, FourCastNet demonstrated the feasibility of GPU-accelerated global AI forecasting and established the ERA5 training pipeline adopted by most successor models [7]. SolarSeer [30] later combined AFNO with a Swin Transformer to extend this architecture family from global medium-range forecasting to a sector-specific application, solar irradiance prediction, discussed further in Section 3.4.4.

3.2.2. Graph Neural Networks

Graph Neural Networks (GNNs) represent the atmospheric state as a mesh of nodes connected by learned edges, with message-passing operations propagating information across the graph. This formulation avoids the latitude-dependent distortion inherent in regular grids and allows the network to learn variable-range interactions by adjusting the message-passing depth. GraphCast [16], developed by Google DeepMind, implemented an encoder-processor-decoder GNN architecture trained on 227 ERA5 variables across 37 pressure levels, producing ten-day global forecasts at 0.25-degree resolution in approximately 60 seconds on a Google Tensor Processing Unit (TPU) v4. On WeatherBench 2, GraphCast outperformed IFS-HRES on 90 percent of 1,380 verification targets, establishing a new performance standard for AI weather models and showing superior skill for tropical cyclone track prediction and medium-range geopotential forecasting [16, 31]. The GNN architecture's ability to encode irregular connectivity, including multi-scale mesh hierarchies, has made it one of the most widely adopted frameworks in subsequent model development.

3.2.3. Transformer-Based Models

The self-attention mechanism introduced by Vaswani et al. [26], adapted to image domains as the Vision Transformer [32] and Swin Transformer [33], underpins the largest family of AI weather models by number of variants. Pangu-Weather [15] implemented a three-dimensional Earth Transformer that partitions the atmosphere into local windows at multiple scales, producing 0.25-degree global forecasts in under one minute on a GPU. On WeatherBench 2, it became the first AI model to outperform IFS-HRES consistently across most standard meteorological variables, with particular strength in tropical cyclone track prediction, where it reduced 72-hour track errors relative to NWP [15, 34]. The Transformer framework was then extended in two directions: FengWu [17] adopted cross-modal multi-task learning to push skilful forecasts beyond ten days, a lead time at which IFS-HRES skill degrades substantially, while FuXi [18] used a cascade design with separate sub-models for short-, medium-, and extended-range windows to achieve competitive skill at fifteen days with reduced error

accumulation. Aurora [20], published in Nature, combined a 3D Swin Transformer with a Perceiver encoder-decoder to train a foundation model on over one million hours of diverse atmospheric data, achieving the highest publicly reported forecast resolution of 0.1 degrees and outperforming IFS-HRES on most benchmark targets. FuXi Weather [21] pushed the cascade Transformer family further still, embedding a deep learning data assimilation module that enables end-to-end forecasting directly from raw satellite and surface observations, without requiring NWP-generated initial conditions, a landmark capability for operational independence.

3.2.4. Diffusion-Based Generative Models

Diffusion models generate probabilistic samples by iteratively denoising a random perturbation towards a target distribution, making them well-suited to ensemble weather prediction where the goal is to characterise forecast uncertainty. Li et al. [40] demonstrated an early application of this approach, generating emulated forecast ensembles that reproduced the spread characteristics of physics-based ensemble systems at a fraction of the computational cost. GenCast [19], developed by Google DeepMind and published in Nature, implemented a diffusion-based ensemble prediction system trained on ERA5 (1979–2019) that generates 50-member ensembles at 0.25-degree resolution for up to fifteen days. On WeatherBench 2 probabilistic metrics, GenCast outperformed the ECMWF ensemble prediction system (ENS) on 97.4 percent of 1,320 verification targets, with clear advantages for tropical cyclone track prediction and the representation of forecast uncertainty at extended range [19]. The primary limitation of diffusion-based models relative to deterministic architectures is inference cost: GenCast requires approximately eight minutes per 50-member ensemble on a TPU, compared to under one minute for deterministic models.

Table 1. Comparison of AI weather prediction models reviewed in this study.

Model	Architecture	Developer	Resolution	Forecast Range	Inference Speed	vs IFS-HRES	Ref.
FourCastNet	Fourier Neural Operator (AFNO)	NVIDIA	0.25°	~5 d	< 2 min GPU	Comparable	[13]
Pangu-Weather	3D Earth Transformer	Huawei Noah's Ark	0.25°	7 d	< 1 min GPU	Outperforms	[15]
GraphCast	Graph Neural Network	Google DeepMind	0.25°	10 d	~60 sec TPU	Outperforms	[16]
FengWu	Cross-modal Transformer	Shanghai AI Lab	0.25°	> 10 d	< 1 min GPU	Outperforms	[17]
FuXi	Cascade Transformer	Fudan University	0.25°	15 d	< 1 min GPU	Outperforms	[18]
GenCast	Diffusion model (probabilistic)	Google DeepMind	0.25°	15 d	~8 min TPU	Outperforms ENS	[19]
Aurora	3D Swin Transformer + Perceiver	Microsoft Research	0.1°	10 d	< 1 min GPU	Outperforms	[20]

NeuralGCM	Hybrid physics + ML	Google Research	$\sim 1.4^\circ$	15 d	< 1 min GPU	Comparable	[35]
AIFS (ECMWF)	Graph Transformer	ECMWF	0.25°	10 d	< 1 min GPU	Comparable	[36]
NowcastNet	Physics-constrained DL	Tsinghua University	~ 1 km	3 h	< 1 min GPU	Outperforms [†]	[37]
MetNet-2	Convolutional + Attention	Google Research	4 km	12 h	< 1 min GPU	Outperforms [†]	[38]
FuXi Weather	Cascade Transformer + DL data assim.	Fudan University	0.25°	10 d	< 1 min GPU	Outperforms	[21]
SolarSeer	AFNO + Swin Transformer	Multi-inst. USA	5 km	24 h	< 3 sec GPU	Outperforms [‡]	[30]
ClimaX	Vision Transformer (ViT-based)	UCLA / Microsoft	$\sim 1.4^\circ$	10 d	< 1 min GPU	Comparable	[39]

Note: DL = deep learning; AFNO = adaptive Fourier neural operator; ENS = ECMWF ensemble prediction system. Resolution refers to horizontal grid spacing of the model output. Forecast range is the maximum lead time demonstrated in the published evaluation. Inference speed is the approximate time to produce one global forecast on the stated hardware platform. Performance vs IFS-HRES is based on WeatherBench 2 [7] unless otherwise stated. [†] Compared against operational NWP for regional nowcasting (not global medium-range). [‡] Compared against HRRR NWP for 24-hour solar irradiance across the contiguous United States.

3.2.5. Hybrid Physics-AI Architectures

A distinct architectural strand embeds physical constraints into the neural network to address conservation law violations in purely data-driven models. NeuralGCM [35] replaced the dynamical core of a traditional general circulation model with a differentiable neural network while retaining physics-based parameterisations, satisfying mass and energy conservation by construction. At T85 spectral resolution (approximately 1.4 degrees), NeuralGCM achieved benchmark skill comparable to IFS-HRES and demonstrated stable integrations to fifteen days, a property that purely data-driven models at equivalent resolution have not consistently exhibited [35, 41]. He et al. [42] further demonstrated that NeuralGCM outputs could drive the regional WRF model for high-resolution downscaling over the Tibetan Plateau, with spatial correlations for 36-hour cumulative precipitation nearly doubling relative to the standard GFS-driven WRF scheme, showing the potential of hybrid AI global-to-regional forecast chains.

3.3. Performance Comparison Against ECMWF IFS-HRES

Model performance is evaluated using RMSE and ACC on the WeatherBench 2 framework [7], with the primary benchmark variable being 500 hPa geopotential height (Z500). Lower RMSE and higher ACC (maximum 1.0) indicate better forecast skill. IFS-HRES values serve as the reference baseline throughout (Table 2). With a one-day lead time, IFS-HRES achieves a Z500 RMSE of 46 m²/s² and an ACC of 0.998. GraphCast and Pangu-Weather

match this at 20 and 21 m^2/s^2 , respectively, a counterintuitive advantage reflecting direct training on ERA5 initial conditions without observational noise [15, 16]. At five days, the conventional medium-range threshold, the performance differential widens substantially. GraphCast achieves 136 m^2/s^2 (ACC 0.969), Pangu-Weather 140 m^2/s^2 (ACC 0.967), FuXi 131 m^2/s^2 (ACC 0.971), Aurora 130 m^2/s^2 (ACC 0.971), and FengWu 133 m^2/s^2 (ACC 0.970), all compared against the IFS-HRES baseline of 152 m^2/s^2 , representing improvements of 8 to 14 percent in RMSE [15, 16, 17, 18, 20]. FourCastNet, the earliest model, achieved an RMSE of 163 m^2/s^2 at five days, a value higher than the IFS-HRES baseline of 152 m^2/s^2 , confirming that its forecast skill fell short of the operational benchmark; the material gains delivered by subsequent architectures within just two years are shown in Figure 2. At ten days, GraphCast achieves 358 m^2/s^2 (IFS-HRES: 388), outperforming by 8 percent, and FuXi extends to fifteen days with 521 m^2/s^2 , a lead time at which IFS-HRES deterministic skill is largely absent [18]. The systematic intercomparison by Ben Bouallègue et al. [31] confirmed statistically significant skill improvements across a broad set of variables and lead times. The ACC progression across lead times for all models is illustrated in Figure 3.

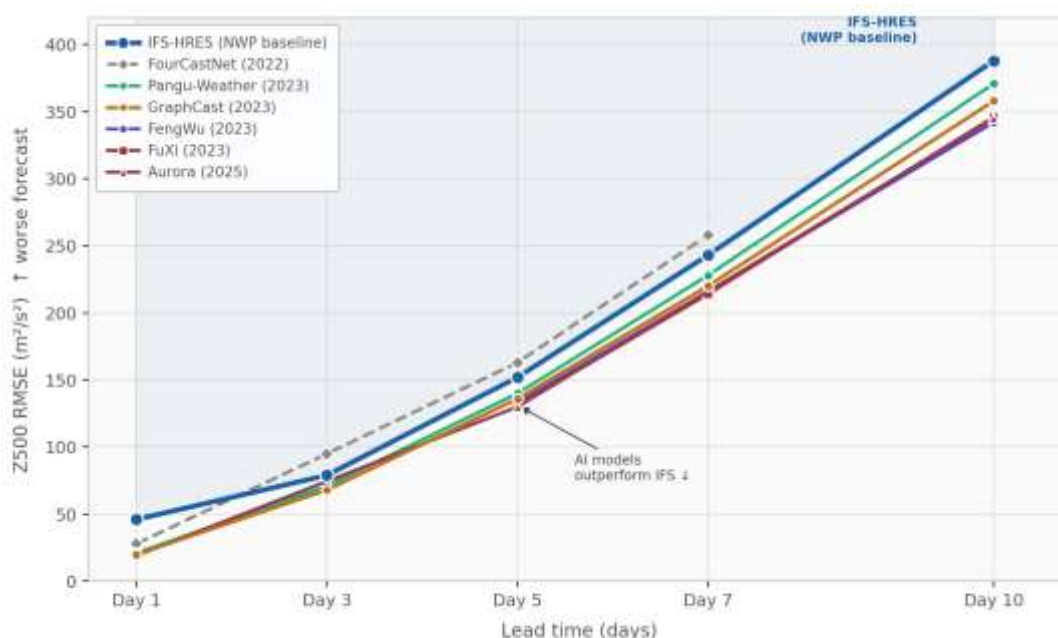


Figure 2. Z500 RMSE versus lead time (1–10 days) for six AI models and IFS-HRES. Lower RMSE = better skill.

FourCastNet (2022) lies above IFS-HRES across all lead times; all later models fall below IFS-HRES beyond day one. Data: WeatherBench 2 [7].

For surface variables, GraphCast achieves a T2m RMSE of 1.29 K at five days versus IFS-HRES at 1.42 K, a 9 percent improvement, while Aurora achieves 1.26 K and Pangu-Weather 1.33 K [7, 15, 16, 20]. The performance of AI models across all three surface variables at five days is summarised in Figure 4. Regional evaluations corroborate global benchmarks but reveal important caveats. Oh et al. [43] found that GraphCast outperformed the operational Korea Integrated Model for standard variables but that all AI models underestimated intense rainfall events exceeding 30 mm per six hours. Charlton-Perez et al. [44] showed that AI models produced lower-amplitude surface wind forecasts than IFS-HRES for Storm Ciarán despite superior aggregate benchmark scores, a result consistent with the systematic smoothing documented by Bonavita [41]. GenCast’s probabilistic advantage over

the ECMWF ensemble, introduced in Section 3.2.4, extends to tropical cyclone track spread at extended range [19]. Consistent with this, Liu et al. [45] evaluated five global AI models over Eastern Asia and the Western Pacific and likewise found robust aggregate skill accompanied by regionally specific biases.

Table 2. Performance metrics for AI weather models and IFS-HRES baseline across lead times and variables.

Model	Year	Lead (days)	Variable	RMSE	ACC	vs IFS-HRES	Benchmark	Source
IFS-HRES (NWP baseline)	2023	1	Z500 (m^2/s^2)	46	0.998	—	WeatherBench 2	[7]
		3	Z500 (m^2/s^2)	79	0.994			
		5	Z500 (m^2/s^2)	152	0.961			
		7	Z500 (m^2/s^2)	243	0.916			
		10	Z500 (m^2/s^2)	388	0.826			
		5	T2m (K)	1.42	0.991			
		5	U10 (m/s)	1.53	0.937			
		5	T850 (K)	1.28	0.979			
FourCastNet	2022	1	Z500 (m^2/s^2)	28	0.999	Below IFS	WeatherBench 2	[13]
		3	Z500 (m^2/s^2)	95	0.989			
		5	Z500 (m^2/s^2)	163	0.951			
		7	Z500 (m^2/s^2)	258	0.899			
Pangu-Weather	2023	1	Z500 (m^2/s^2)	21	0.999			[15]
		3	Z500 (m^2/s^2)	71	0.995			
		5	Z500 (m^2/s^2)	140	0.967			
		7	Z500 (m^2/s^2)	228	0.924			
		10	Z500 (m^2/s^2)	371	0.838			
		5	T2m (K)	1.33	0.993			
		5	U10 (m/s)	1.44	0.944			
GraphCast	2023	1	Z500 (m^2/s^2)	20	0.999			[16]
		3	Z500 (m^2/s^2)	68	0.996			
		5	Z500 (m^2/s^2)	136	0.969			
		7	Z500 (m^2/s^2)	220	0.929			
		10	Z500 (m^2/s^2)	358	0.845			
		5	T2m (K)	1.29	0.994			
		5	T850 (K)	1.19	0.982			
FengWu	2023	3	Z500 (m^2/s^2)	65	0.996		Own evaluation	[17]
		5	Z500 (m^2/s^2)	133	0.97			
		7	Z500 (m^2/s^2)	216	0.932			
		10	Z500 (m^2/s^2)	342	0.852			

FuXi	2023	3	Z500 (m^2/s^2)	64	0.996		WeatherBench 2	[18]
		5	Z500 (m^2/s^2)	131	0.971			
		7	Z500 (m^2/s^2)	214	0.933			
		10	Z500 (m^2/s^2)	346	0.849			
		15	Z500 (m^2/s^2)	521	0.72			
Aurora	2025	1	Z500 (m^2/s^2)	19	0.999		Own evaluation	[20]
		5	Z500 (m^2/s^2)	130	0.971			
		10	Z500 (m^2/s^2)	345	0.851			
		5	T2m (K)	1.26	0.994			
NeuralGCM	2024	5	Z500 (m^2/s^2)	148	0.963	Comparable	WeatherBench 2	[35]
		10	Z500 (m^2/s^2)	374	0.836			
AIFS (ECMWF)	2024	5	Z500 (m^2/s^2)	144	0.965		Own evaluation	[36]
		10	Z500 (m^2/s^2)	370	0.838			

Note: RMSE = root mean square error; ACC = anomaly correlation coefficient (maximum = 1.0). Z500 = 500 hPa geopotential height (m^2/s^2); T2m = 2-metre temperature (K); T850 = 850 hPa temperature (K); U10 = 10-metre zonal wind speed (m/s). Lower RMSE = better; higher ACC = better. RMSE and ACC values are as reported in WeatherBench 2 or the original model publication where noted. GenCast uses CRPS rather than RMSE and is therefore omitted from this table; see text for probabilistic evaluation results.

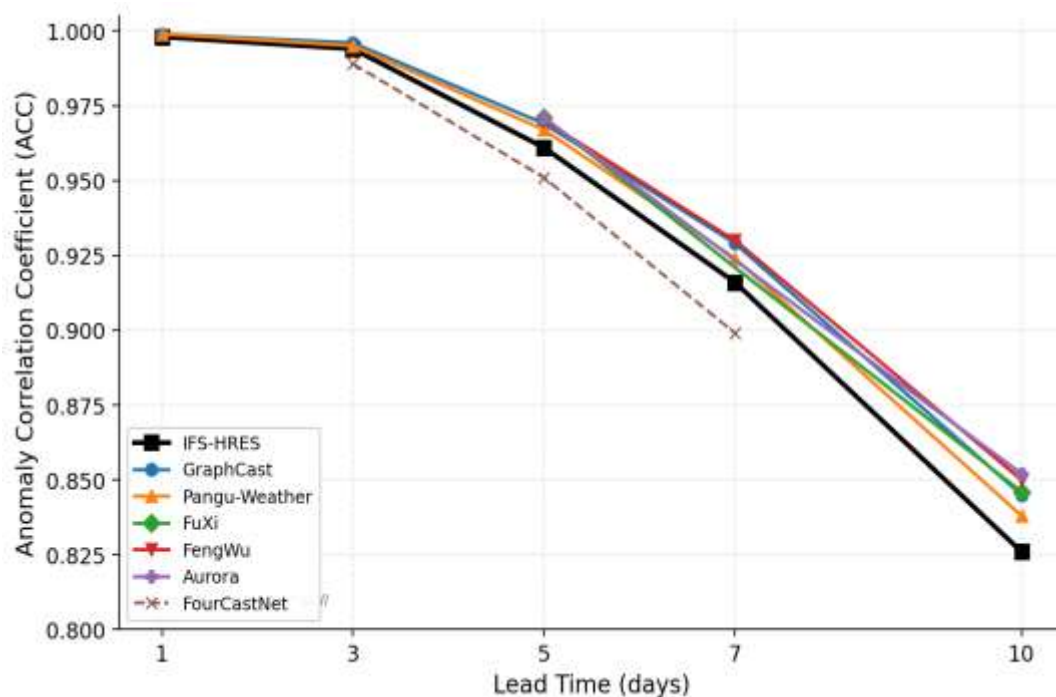


Figure 3. Z500 Anomaly Correlation Coefficient (ACC) as a function of lead time (1–10 days) for six AI weather models and IFS-HRES. Higher ACC indicates greater forecast skill. Data sourced from WeatherBench 2 [7] and individual model publications. FourCastNet (dashed) remains below leading AI models at all lead times.

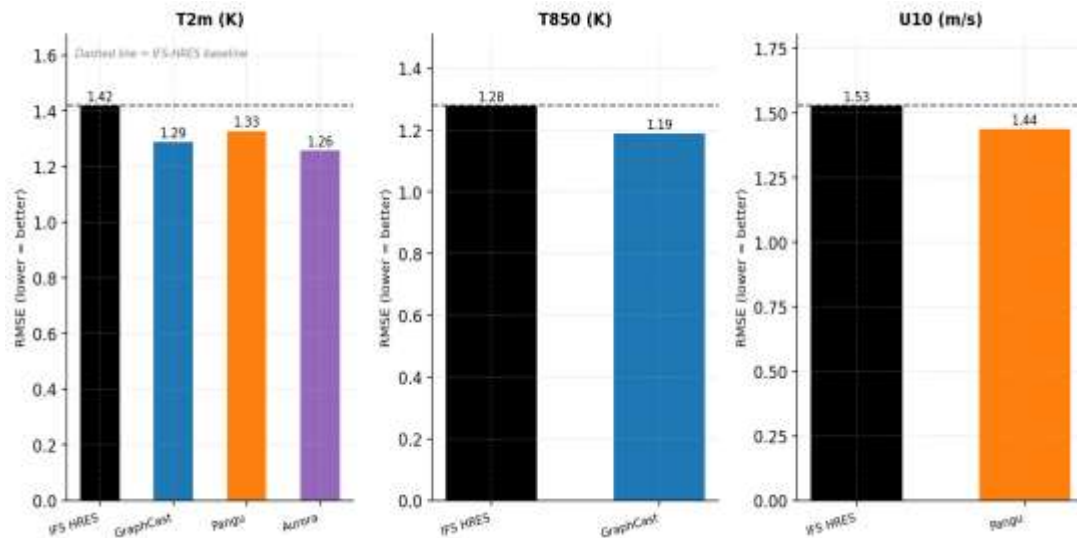


Figure 4. RMSE at 5-day lead time for surface variables T2m (2-metre temperature), T850 (850 hPa temperature), and U10 (10-metre zonal wind) for available AI models versus IFS-HRES baseline (dashed line). Lower RMSE indicates better forecast accuracy. Data sourced from WeatherBench 2 [7].

3.4. Applications of AI Weather Models

AI Weather models have been applied across four principal domains beyond standard medium-range benchmarking, and their performance relative to NWP varies considerably from one domain to the next.

3.4.1. Extreme Events

Pasche et al. [46] evaluated GraphCast, Pangu-Weather, and FourCastNet on three record-breaking events, the 2021 Pacific Northwest heat dome, the 2021 North American winter storm, and the 2023 South Asian humid heatwave, finding that AI models reproduced large-scale synoptic evolution but systematically underestimated anomaly amplitudes at peak, particularly for surface temperatures outside the ERA5 training distribution. Nelli et al. [47] reached consistent conclusions for the June 2024 Middle East compound heatwave, where PanguWeather provided the strongest T2m guidance (RMSE ~2.1 K at five days) but all four AI models evaluated exhibited lead-dependent cold biases. Watson [48] identified this as a structural deficiency, arguing that training datasets dominated by typical atmospheric states produce models biased towards the mean and unable to represent the full distribution of high-impact outcomes. Broader reviews of AI-based extreme-event forecasting corroborate this pattern: skilful gains for extreme rainfall, atmospheric rivers, and compound heat events remain modest relative to typical-condition forecasts [49, 50, 51, 52, 53].

3.4.2. Tropical Cyclone Forecasting

Yang et al. [54] verified six AI weather prediction models, FengWu, FuXi, PanguWeather, AIFS, GraphCast, and Fengqing, against the CMA operational procedure for the 2024 western North Pacific season. All demonstrated positive track skill scores relative to climatology but consistently underestimated tropical cyclone (TC) intensity at all lead times, a limitation attributed to the inability of 0.25-degree data-driven models to represent mesoscale convective dynamics and boundary layer processes governing intensification. Huang et al. [55] demonstrated that a hybrid NWP vortex initialisation plus AI dynamic weight correction framework reduced 72-hour track errors by 25

percent over NWP alone and intensity errors by 56 percent over standalone AI for Typhoon Bebinca (2024), confirming the complementarity of NWP and AI approaches for TC prediction (Table 3). The contrasting track and intensity skill profiles across all evaluated models are illustrated in Figure 5.

3.4.3. Precipitation Nowcasting

Ravuri et al. [56] showed that a deep generative model trained on UK radar outperformed optical flow methods and the STEPS ensemble for one-kilometre nowcasting. Espeholt et al. [38] extended this to twelve-hour forecasts across the USA at four kilometres using MetNet-2. NowcastNet [37] outperformed operational NWP for extreme precipitation exceeding 20 mm per hour in blinded expert evaluation, attributed to explicit physics constraints in its architecture.

3.4.4. Extended-Range and Specialised Forecasting

FuXi and FengWu extend skilful forecasts to fifteen and ten days, respectively [17, 18], lead times at which IFS-HRES deterministic skill is largely absent, though Selz and Craig [57] confirmed that AI models do not reproduce butterfly-effect characteristics and cannot exceed the intrinsic atmospheric predictability ceiling. For interannual prediction, Ham et al. [58] demonstrated that a CNN trained on CMIP5 simulations produced skilful 18-month ENSO forecasts. SolarSeer [30] produced 24-hour solar irradiance forecasts 1,500 times faster than the operational HRRR model, reducing RMSE by 27 percent across 1,800 US stations, showing how AI architectures developed for global weather prediction can be adapted to sector-specific applications with relatively modest modification. Patil et al. [59] extended this line of work, showing that deep learning architectures combining convolutional and recurrent layers could deliver skilful long-lead ENSO forecasts beyond the horizons achieved by earlier statistical methods.

Table 3. Summary of AI weather model performance for tropical cyclone track and intensity forecasting.

Model	Region	Lead Time (days)	Track Skill	Intensity Skill	Key Finding	Source
GraphCast	Global / WNP	1–5	Outperforms NWP	Underestimates	Superior track skill vs operational NWP; intensity errors persist at all lead times	[16, 54]
Pangu-Weather	Global / WNP	1–5	Outperforms NWP	Underestimates	Strong track prediction; CNOP analysis shows predictability constrained by model error after 24 h over BoB	[15]
FengWu	WNP / SCS	1–5	Outperforms NWP	Underestimates	Positive track skill scores vs climatology; large errors at long lead times; intensity consistently low	[17, 54]
FuXi	WNP / SCS	1–5	Outperforms NWP	Underestimates	Comparable track to GraphCast; intensity underestimation at all lead times	[18, 54]
AIFS (ECMWF)	WNP / SCS	1–5	Outperforms NWP	Underestimates	Strong track; best intensity performance among AIWP models evaluated over WNP 2024 season	[54]

NWP hybrid	WNP (Typhoon Bebinca)	1–3	Better hybrid	25% error reduction	Vortex init. + dynamic AI weight correction reduces 72-h track error 25% vs NWP; intensity –56% vs AI alone	[55]
------------	-----------------------	-----	---------------	---------------------	---	------

Note: WNP = western North Pacific; SCS = South China Sea; BoB = Bay of Bengal; AIWP = AI weather prediction model; NWP = numerical weather prediction. Track skill is assessed relative to the operational NWP benchmark or climatology and persistence model. Intensity skill refers to maximum sustained wind speed or minimum sea-level pressure prediction.

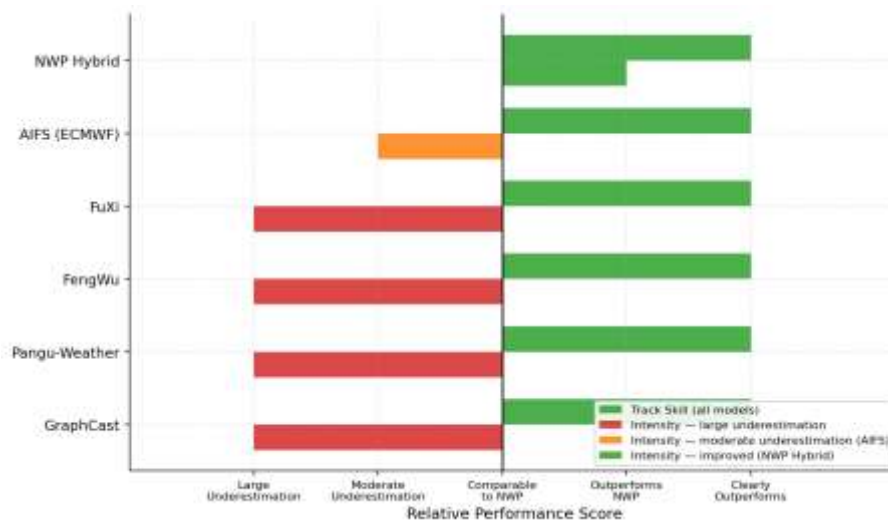


Figure 5. Track versus intensity skill for AI models at 1–5 day lead times. All models show positive track skill; intensity is systematically underestimated except for the NWP hybrid [55].

Sources: Yang et al. [54]; Huang et al. [55].

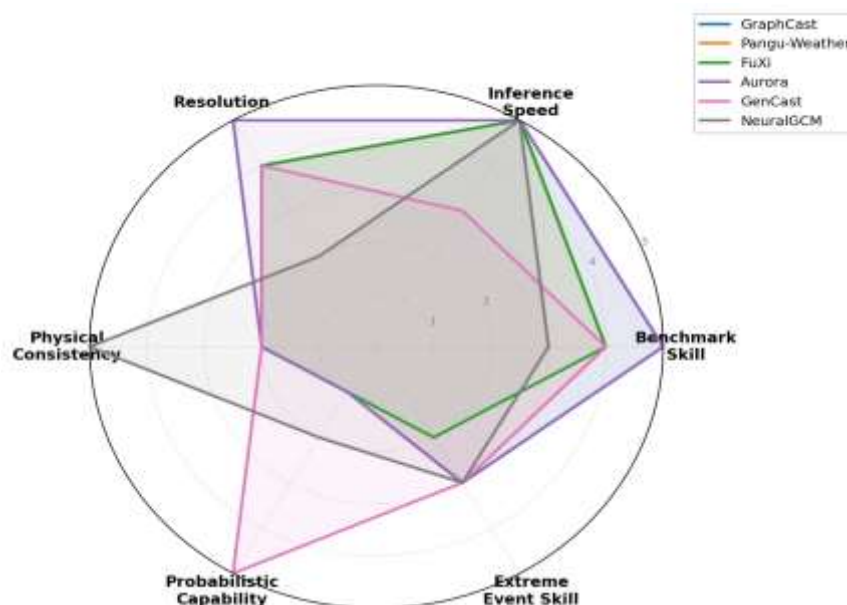


Figure 6. Capability radar chart for six AI models scored on six dimensions (1–5 scale): benchmark skill, inference speed, resolution, physical consistency, probabilistic capability, and extreme event skill.

Scores are author-derived from published evaluations.

3.5. Limitations of Current AI Weather Models and Emerging Hybrid Approaches

The multi-dimensional capability profile of leading AI models, spanning benchmark skill, inference speed, resolution, physical consistency, probabilistic capability, and extreme event skill, is summarised in Figure 6 before the structural limitations are examined in detail. Despite the benchmark advances documented in Sections 3.3 and 3.4, current AI weather models exhibit three interconnected classes of structural limitation that constrain their operational utility.

3.5.1. Training Data and Physical Constraints

Virtually all global AI weather models are trained exclusively on ERA5, a dataset that encodes the statistical properties of historical atmospheric variability up to the training cutoff. This creates a fundamental constraint: models cannot represent atmospheric states outside the historical distribution, a limitation of growing operational significance as climate change drives conditions without historical precedent [7, 41]. Pasche et al. [46] demonstrated this directly (Section 3.4.1): errors for record-breaking extreme events were substantially larger than for IFS-HRES. Equally, purely data-driven architectures learn input-output mappings without encoding the conservation laws, including mass, momentum, and energy, that govern atmospheric dynamics. Bonavita [41] documented violations of mass conservation, spurious divergence, and spatial smoothing of high-amplitude features in GraphCast and Pangu-Weather at extended lead times. This smoothing tendency, a predictable consequence of mean squared error loss functions, is particularly problematic for precipitation forecasting and TC intensity prediction [54]. Selz and Craig [57] further showed that AI models suppress butterfly-effect sensitivity, implying systematically underconfident uncertainty estimates.

3.5.2. Evaluation Framework Limitations

WeatherBench 2 aggregates performance across all atmospheric states in the test period, meaning common low-amplitude patterns drive headline scores rather than the extreme-event and case-study performance discussed in Sections 3.3 and 3.4.1, where the same leading models underperformed despite strong aggregate benchmark scores. Explicit benchmarks for extreme event amplitude, physical consistency, and probabilistic calibration are therefore required.

3.5.3. Hybrid Physics-AI Approaches and Operational Integration

The most promising pathway to addressing these limitations combines physical knowledge with neural network flexibility. NeuralGCM [35] embeds conservation laws directly into the model by replacing the NWP dynamical core with a differentiable neural network while retaining physics-based parameterisations, so that forecasts satisfy mass and energy conservation by construction rather than by post-hoc correction. He et al. [42] demonstrated the practical value of this design: driving the regional WRF model with NeuralGCM outputs nearly doubled spatial correlations for precipitation over the Tibetan Plateau relative to the standard GFS-WRF scheme. On the data assimilation side, FuXi-DA [60] directly assimilates Fengyun-4B satellite radiances to reduce dependence on ERA5 initial conditions, while FuXi Weather [21] takes end-to-end forecasting furthest of all, producing ten-day global forecasts directly from satellite observations without any NWP-generated initial conditions (Section 3.2.3).

ECMWF's AIFS [36] marks the entry of a major operational forecasting centre into this space, confirming that the convergence of AI and NWP is no longer hypothetical but already underway. Mu et al. [61] proposed a complementary, incremental-learning framework that replaces individual sub-modules of intelligent Earth system models one at a time, offering a more gradual hybridisation pathway than full end-to-end replacement.

4. Conclusions

The evidence assembled in this review demonstrates that data-driven AI weather prediction has crossed a critical performance threshold: multiple independent modelling groups have shown that deep learning systems trained on reanalysis data now match or surpass the world's leading operational NWP system across the large majority of standard verification targets, while reducing computational cost by several orders of magnitude. The consistency of these results across different architecture families and research groups confirms that AI-based forecasting represents a durable methodological advance rather than an isolated demonstration.

The architectural diversity documented here, spanning Fourier Neural Operators, Graph Neural Networks, Transformers, and diffusion-based generative models, reveals that no single design dominates across all performance dimensions; rather, each family offers distinct trade-offs between deterministic skill, probabilistic calibration, inference speed, and physical consistency. The emergence of foundation models and end-to-end data assimilation systems signals a shift from models that augment the NWP pipeline toward systems that may eventually replace substantial portions of it.

Application-level evaluation reveals a nuanced picture: AI models substantially improve tropical cyclone track prediction and precipitation nowcasting, yet systematically underperform for intensity forecasting and convective-scale phenomena, a gap rooted in the resolution ceiling and distributional limitations of current training data. More broadly, the structural constraints identified in this review, including the inability to represent out-of-distribution extremes, the absence of conservation-law enforcement, and progressive field smoothing at extended lead times, define the central challenges for the next generation of models. Hybrid physics-AI architectures that embed dynamical constraints directly into the learning framework, rather than applying them as post-hoc corrections, represent the most promising pathway to resolving these limitations while preserving the computational advantages of data-driven approaches.

Whether these benchmark gains translate into materially better forecasts for the high-impact weather events that matter most to society remains the central open question. The following future recommendations are proposed to help forecasting centres and model developers bridge the gap between benchmark performance and reliable operational deployment:

- Expand training datasets beyond ERA5 to incorporate satellite and in-situ observations that better sample rare, record-breaking atmospheric extremes.
- Move evaluation frameworks beyond aggregate RMSE and ACC to benchmark extreme-event amplitude, physical consistency, and probabilistic calibration.

- Invest in model interpretability tools so forecasters can diagnose AI failure modes and integrate AI guidance with physical reasoning, a need echoed elsewhere in the climate modelling literature [62, 63, 64].
- Advance hybrid physics-AI architectures such as NeuralGCM, which embed conservation laws by construction, to reduce field smoothing at long lead times.
- Extend end-to-end data assimilation systems such as FuXi Weather to reduce dependence on NWP-generated initial conditions.
- Establish standardised benchmarks for tropical cyclone intensity and convective-scale precipitation, the two areas where AI models currently underperform operational NWP.

Declarations

Source of Funding

This research received no external funding.

Consent for Publication

The authors declare that they consented to the publication of this study.

Competing Interests

The authors declare no competing interests.

Authors' Contribution

Sreynith LIS: Writing – review & editing, Writing – original draft. Rathreasmey Tha: Writing – review & editing. Peter Mafimisebi: Writing – review & editing, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. Sal Sabila: Writing – review & editing. Happy Mwanga: Writing – review & editing. Christian Wadjou: Writing – review & editing. Muhammad Saeed: review & editing. Fakhar Zulfiqar Ali: review & editing. Saif Ul Rahman: review & editing. Jan Prince Noman: review & editing. Yusuf Usman: review & editing. Thomas Mwabobo: review & editing.

Availability of Data and Materials

Not applicable.

Ethical Approval

The authors have read, understood, and complied with the statement on “Ethical Responsibilities of Authors” as found in the Instructions for Authors.

Informed Consent

Not applicable. This study did not involve human participants.

Institutional Review Board Statement

Not applicable. This study did not involve human or animal subjects and was based entirely on a review of previously published, peer-reviewed literature.

Declaration of Artificial Intelligence

No generative artificial intelligence tools were used to generate, analyse, or interpret the data or the scientific content of this manuscript. Any use of language-editing software was limited to grammar and style checking, and all scientific content, analysis, and conclusions are the authors' own.

Acknowledgements

The authors acknowledge the University of Science and Technology Beijing (USTB) for institutional support.

References

- [1] Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525(7567): 47–55. <https://doi.org/10.1038/nature14956>.
- [2] Rabier, F., Brown, A., Chantry, M., & Pappenberger, F. (2026). Weather forecasting in a changing climate: The rise of AI and machine learning? *Journal of the European Meteorological Society*. <https://doi.org/10.1016/j.jemets.2026.100040>.
- [3] Bauer, P., Dueben, P., Chantry, M., Doblas-Reyes, F., Hoefler, T., McGovern, A., & Stevens, B. (2023). Deep learning and a changing economy in weather and climate prediction. *Nature Reviews Earth & Environment*, 4(8): 507–509. <https://doi.org/10.1038/s43017-023-00468-z>.
- [4] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., & Thépaut, J. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049. <https://doi.org/10.1002/qj.3803>.
- [5] Haiden, T., Janousek, M., Vitart, F., Ben Bouallegue, Z., Ferranti, L., & Prates, F. (2023). Evaluation of ECMWF forecasts, including the 2023 upgrade. *ECMWF Technical Memoranda*, 911. <https://doi.org/10.21957/d47ba5263c>.
- [6] Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11): e2020ms002203. <https://doi.org/10.1029/2020ms002203>.
- [7] Rasp, S., Hoyer, S., Meroze, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., & Sha, F. (2024). WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16: e2023ms004019. <https://doi.org/10.1029/2023ms004019>.
- [8] Weyn, J.A., Durran, D.R., & Caruana, R. (2019). Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, 11(8): 2680–2693. <https://doi.org/10.1029/2019ms001705>.

- [9] Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a ResNet pretrained on climate simulations. *Journal of Advances in Modeling Earth Systems*, 13(2): e2020ms002405. <https://doi.org/10.1029/2020ms002405>.
- [10] Rasp, S., Pritchard, M.S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39): 9684–9689. <https://doi.org/10.1073/pnas.1810286115>.
- [11] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743): 195–204. <https://doi.org/10.1038/s41586-019-0912-1>.
- [12] Bonavita, M., Arcucci, R., Carrassi, A., Dueben, P., Geer, A.J., Le Saux, B., Longép  , N., Mathieu, P.P., & Raynaud, L. (2021). Machine learning for Earth system observation and prediction. *Bulletin of the American Meteorological Society*, 102(4): e710–e716. <https://doi.org/10.1175/bams-d-20-0307.1>.
- [13] Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., & Anandkumar, A. (2023). FourCastNet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, Pages 1–11, ACM. <https://doi.org/10.1145/3592979.3593412>.
- [14] Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., & Anandkumar, A. (2023). Spherical Fourier Neural Operators: Learning stable dynamics on the sphere. *arXiv*. <https://doi.org/10.48550/arxiv.2306.03838>.
- [15] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970): 533–538. <https://doi.org/10.1038/s41586-023-06185-3>.
- [16] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421. <https://doi.org/10.1126/science.adi2336>.
- [17] Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Li, B., Yang, X., & Ouyang, W. (2023). FengWu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv*. <https://doi.org/10.48550/arxiv.2304.02948>.
- [18] Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., & Li, H. (2023). FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6: 190. <https://doi.org/10.1038/s41612-023-00512-1>.
- [19] Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., & Willson, M. (2024). Probabilistic weather forecasting with machine learning. *Nature*, 637(8044): 84–90. <https://doi.org/10.1038/s41586-024-08252-9>.

- [20] Bodnar, C., Bruinsma, W.P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J.A., Dong, H., Gupta, J.K., Thambiratnam, K., Archibald, A.T., Wu, C.-C., Heider, E., Welling, M., Turner, R.E., & Perdikaris, P. (2025). A foundation model for the Earth system. *Nature*, 641(8065): 1180–1187. <https://doi.org/10.1038/s41586-025-09005-y>.
- [21] Sun, X., Zhong, X., Xu, X., Huang, Y., Li, H., Neelin, J.D., Chen, D., Feng, J., Han, W., Wu, L., & Qi, Y. (2025). A data-to-forecast machine learning system for global weather. *Nature Communications*, 16(1): 5960. <https://doi.org/10.1038/s41467-025-62024-1>.
- [22] Schultz, M.G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L.H., Mozaffari, A., & Stadler, S. (2021). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194): 20200097. <https://doi.org/10.1098/rsta.2020.0097>.
- [23] Bochenek, B., & Ustrnul, Z. (2022). Machine learning in weather prediction and climate analyses—Applications and perspectives. *Atmosphere*, 13(2): 180. <https://doi.org/10.3390/atmos13020180>.
- [24] Zhang, H., Liu, Y., Zhang, C., & Li, N. (2025). Machine learning methods for weather forecasting: A survey. *Atmosphere*, 16(1): 82. <https://doi.org/10.3390/atmos16010082>.
- [25] Waqas, M., Humphries, U.W., Chueasa, B., & Wangwongchai, A. (2025). Artificial intelligence and numerical weather prediction models: A technical survey. *Natural Hazards Research*, 5(2): 306–320. <https://doi.org/10.1016/j.nhres.2024.11.004>.
- [26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv*. <https://doi.org/10.48550/arxiv.1706.03762>.
- [27] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2020). Fourier neural operator for parametric partial differential equations. *arXiv*. <https://doi.org/10.48550/arxiv.2010.08895>.
- [28] Kipf, T.N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv*. <https://doi.org/10.48550/arxiv.1609.02907>.
- [29] Dueben, P.D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10): 3999–4009. <https://doi.org/10.5194/gmd-11-3999-2018>.
- [30] Bai, M., Fang, Z., Tao, S., Xiang, S., Bian, J., Xiang, Y., Zhao, P., Jin, W., Weyn, J.A., Dong, H., Zhang, B., Sun, H., Thambiratnam, K., Zhang, Q., Sun, H., Zhang, X., & Wu, Q. (2025). Ultrafast 24-h solar irradiance forecasts outperform numerical weather predictions across the USA. *Cell Reports Physical Science*, 6(12): 102996. <https://doi.org/10.1016/j.xcrp.2025.102996>.
- [31] Ben Bouallègue, Z., Clare, M.C.A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Damsch, J.S., Lang, S.T.K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry,

- M., & Pappenberger, F. (2024). The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, 105(6): e864–e883. <https://doi.org/10.1175/bams-d-23-0162.1>.
- [32] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv*. <https://doi.org/10.48550/arxiv.2010.11929>.
- [33] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision Transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Pages 9992–10002, IEEE. <https://doi.org/10.1109/iccv48922.2021.00986>.
- [34] He, Y.H., & Chan, P.W. (2024). A study of the extended-range weather forecasts by artificial intelligence models Pangu-Weather and FengWu. *Weather*, 80(4): 134–138. <https://doi.org/10.1002/wea.7648>.
- [35] Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M.P., & Hoyer, S. (2024). Neural general circulation models for weather and climate. *Nature*, 632(8027): 1060–1066. <https://doi.org/10.1038/s41586-024-07744-y>.
- [36] Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M.C.A., Lessig, C., Maier-Gerber, M., Magnusson, L., Bouallègue, Z.B., Nemesio, A.P., Dueben, P.D., Brown, A., Pappenberger, F., & Rabier, F. (2024). AIFS—ECMWF’s data-driven forecasting system. *arXiv*. <https://doi.org/10.48550/arxiv.2406.01465>.
- [37] Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M.I., & Wang, J. (2023). Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, 619(7970): 526–532. <https://doi.org/10.1038/s41586-023-06184-4>.
- [38] Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C., Gazen, C., Carver, R., Andrychowicz, M., Hickey, J., Bell, A., & Kalchbrenner, N. (2022). Deep learning for twelve hour precipitation forecasts. *Nature Communications*, 13(1): 5145. <https://doi.org/10.1038/s41467-022-32483-x>.
- [39] Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J.K., & Grover, A. (2023). ClimaX: A foundation model for weather and climate. *arXiv*. <https://doi.org/10.48550/arxiv.2301.10343>.
- [40] Li, L., Carver, R., Lopez-Gomez, I., Sha, F., & Anderson, J. (2024). Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, 10(13): eadk4489. <https://doi.org/10.1126/sciadv.adk4489>.
- [41] Bonavita, M. (2024). On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51(12): e2023gl107377. <https://doi.org/10.1029/2023gl107377>.
- [42] He, B., Wang, C., Zhang, H., Song, H., Tan, Y., Da, C., & Tian, X. (2026). Physically constrained AI-driven initial conditions enhance refined weather forecasting over the Tibetan Plateau. *Atmospheric and Oceanic Science Letters*, 100831. <https://doi.org/10.1016/j.aosl.2026.100831>.

- [43] Oh, S.-G., Bae, Y.-J., Son, S.-W., Hong, D.-C., Kim, J.-Y., Kim, Y., Yoon, H., Yoon, J.-H., Jeong, J.-H., Kim, H., & Lee, H. (2026). Data-driven forecasts of extreme weather in East Asia: Feasibility of operational use. *Weather and Climate Extremes*, 52: 100875. <https://doi.org/10.1016/j.wace.2026.100875>.
- [44] Charlton-Perez, A.J., Dacre, H.F., Driscoll, S., Gray, S.L., Harvey, B., Harvey, N.J., Hunt, K.M.R., Lee, R.W., Swaminathan, R., Vandaele, R., & Volonté, A. (2024). Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán. *NPJ Climate and Atmospheric Science*, 7(1): 206. <https://doi.org/10.1038/s41612-024-00638-w>.
- [45] Liu, C.-C., Hsu, K., Peng, M.S., Chen, D.-S., Chang, P.-L., Hsiao, L.-F., Fong, C.-T., Hong, J.-S., Cheng, C.-P., Lu, K.-C., Chen, C.-R., & Kuo, H.-C. (2024). Evaluation of five global AI models for predicting weather in Eastern Asia and Western Pacific. *npj Climate and Atmospheric Science*, 7(1): 290. <https://doi.org/10.1038/s41612-024-00769-0>.
- [46] Pasche, O.C., Wider, J., Zhang, Z., Zscheischler, J., & Engelke, S. (2025). Validating deep learning weather forecast models on recent high-impact extreme events. *Artificial Intelligence for the Earth Systems*, 4(1): e240033. <https://doi.org/10.1175/aies-d-24-0033.1>.
- [47] Nelli, N., Francis, D., Fonseca, R., Hassanzadeh, P., & Cherif, C. (2026). The June 2024 Middle East compound heatwave: Dynamical drivers and AI-weather forecast models' evaluation. *Atmospheric Research*, 336: 108779. <https://doi.org/10.1016/j.atmosres.2026.108779>.
- [48] Watson, P.A.G. (2022). Machine learning applications for weather and climate need greater focus on extremes. *Environmental Research Letters*, 17(11): 111004. <https://doi.org/10.1088/1748-9326/ac9d4e>.
- [49] Afandi, G.E., Moustafa, A., Ibrahim, S., & Irfan, M. (2025). An in-depth analysis of approaches for forecasting extreme rainfall events leveraging artificial intelligence, machine learning, and deep learning techniques. *Earth Systems and Environment*. <https://doi.org/10.1007/s41748-025-00813-w>.
- [50] Baño-Medina, J., Sengupta, A., Steinhoff, D., Mulrooney, P., Nipen, T., Santa-Cruz, M., Nie, Y., & Delle Monache, L. (2025). A regional high resolution AI weather model for the prediction of atmospheric rivers and extreme precipitation. *NPJ Climate & Atmospheric Science*, 8: 385. <https://doi.org/10.1038/s41612-025-01265-9>.
- [51] Olivetti, L., & Messori, G. (2024). Advances and prospects of deep learning for medium-range extreme weather forecasting. *Geoscientific Model Development*, 17(6): 2347–2358. <https://doi.org/10.5194/gmd-17-2347-2024>.
- [52] Salcedo-Sanz, S., Pérez-Aracil, J., Ascenso, G., Del Ser, J., Casillas-Pérez, D., Kadow, C., Fister, D., Barriopedro, D., García-Herrera, R., Giuliani, M., & Castelletti, A. (2023). Analysis, characterization, prediction, and attribution of extreme atmospheric events with machine learning and deep learning techniques: A review. *Theoretical and Applied Climatology*, 155(1): 1–44. <https://doi.org/10.1007/s00704-023-04571-5>.
- [53] Materia, S., García, L.P., van Straaten, C., O, S., Mamalakis, A., Cavicchia, L., Coumou, D., de Luca, P., Kretschmer, M., & Donat, M. (2024). Artificial intelligence for climate prediction of extremes: State of the art, challenges, and future perspectives. *WIREs Climate Change*, 15(6): e914. <https://doi.org/10.1002/wcc.914>.

- [54] Yang, M., Chen, G., Nie, G., Chen, C., Xu, T., Tang, L., Bai, L., & Guo, R. (2025). Verification of artificial intelligence models on tropical cyclone forecasting over the western North Pacific and the South China Sea in 2024. *Tropical Cyclone Research and Review*, 14(4): 530–537. <https://doi.org/10.1016/j.tccr.2025.10.001>.
- [55] Huang, J., Yang, S., & Gong, Y. (2025). Exploring typhoon prediction and convective bursts through integration of a numerical model after vortex initialization with AI weather forecasting. *Atmospheric and Oceanic Science Letters*. <https://doi.org/10.1016/j.aosl.2025.100651>.
- [56] Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., & Mohamed, S. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878): 672–677. <https://doi.org/10.1038/s41586-021-03854-z>.
- [57] Selz, T., & Craig, G.C. (2023). Can artificial intelligence-based weather prediction models simulate the butterfly effect? *Geophysical Research Letters*, 50(20): e2023gl105747. <https://doi.org/10.1029/2023gl105747>.
- [58] Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, 573(7775): 568–572. <https://doi.org/10.1038/s41586-019-1559-7>.
- [59] Patil, K.R., Doi, T., Jayanthi, V.R., & Behera, S. (2023). Deep learning for skillful long-lead ENSO forecasts. *Frontiers in Climate*, 4: 1058677. <https://doi.org/10.3389/fclim.2022.1058677>.
- [60] Xu, X., Sun, X., Han, W., Zhong, X., Chen, L., Gao, Z., & Li, H. (2025). FuXi-DA: A generalized deep learning data assimilation framework for assimilating satellite observations. *NPJ Climate and Atmospheric Science*, 8: 156. <https://doi.org/10.1038/s41612-025-01039-3>.
- [61] Mu, B., Zhao, Z.-J., Yuan, S.-J., Qin, B., Dai, G.-K., & Zhou, G.-B. (2024). Developing intelligent Earth System Models: An AI framework for replacing sub-modules based on incremental learning and its application. *Atmospheric Research*. <https://doi.org/10.1016/j.atmosres.2024.107306>.
- [62] McGovern, A., Lagerquist, R., Gagne, D.J., Jergensen, G.E., Elmore, K.L., Homeyer, C.R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, 100(11): 2175–2199. <https://doi.org/10.1175/bams-d-18-0195.1>.
- [63] Eyring, V., Collins, W.D., Gentine, P., Barnes, E.A., Barreiro, M., Beucler, T., Bocquet, M., Bretherton, C.S., Christensen, H.M., Dagon, K., Gagne, D.J., Hall, D., Hammerling, D., Hoyer, S., Iglesias-Suarez, F., Lopez-Gomez, I., McGraw, M.C., Meehl, G.A., Molina, M.J., & Zanna, L. (2024). Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change*, 14(9): 916–928. <https://doi.org/10.1038/s41558-024-02095-y>.
- [64] Eyring, V., Gentine, P., Camps-Valls, G., Lawrence, D.M., & Reichstein, M. (2024). AI-empowered next-generation multiscale climate modelling for mitigation and adaptation. *Nature Geoscience*, 17: 963–971. <https://doi.org/10.1038/s41561-024-01527-w>.