Artificial intelligence for weather and climate: a survey of methods, benchmarks, and scientific machine learning challenges

Andreas Holzinger · Sandro Fiore · Tullio Degiacomi · Fabrizio Antonio · Heimo Müller

Received: 13 August 2026 / Accepted: 21 August 2026
© The Author(s) 2026

Abstract

Weather and climate science is a uniquely demanding testbed for artificial intelligence, combining petabyte-scale spatiotemporal complex data on a rotating sphere, governing partial differential equations coupled to global observational and reanalysis archives, climate non-stationarity that distinguishes it from stationary fluid-dynamics or seismology benchmarks, and operational forecast verification against rare extremes. This survey makes four contributions. *First*, it organises AI architectures, i.e. convolutional networks, graph neural networks, transformers, neural operators, and generative models, by their physical inductive biases, with geometric deep learning and operator learning as unifying themes. *Second*, it gives systematic treatment to representation learning and foundation models, uncertainty quantification, and causal discovery as a complement to prediction. *Third*, it names and analyses evaluation pathologies in scientific machine learning, including the RMSE smoothness bias under which mean-squared-error training systematically rewards blurred forecasts and penalises sharp but slightly displaced predictions, ERA5 training–evaluation circularity, and benchmark overfitting. *Fourth*, it frames the open problems, i.e. distribution shift under non-stationary climate, physical consistency of learned operators, sample efficiency in data-sparse regimes, and intrinsic interpretability, as scientific machine learning problems generalisable beyond meteorology.

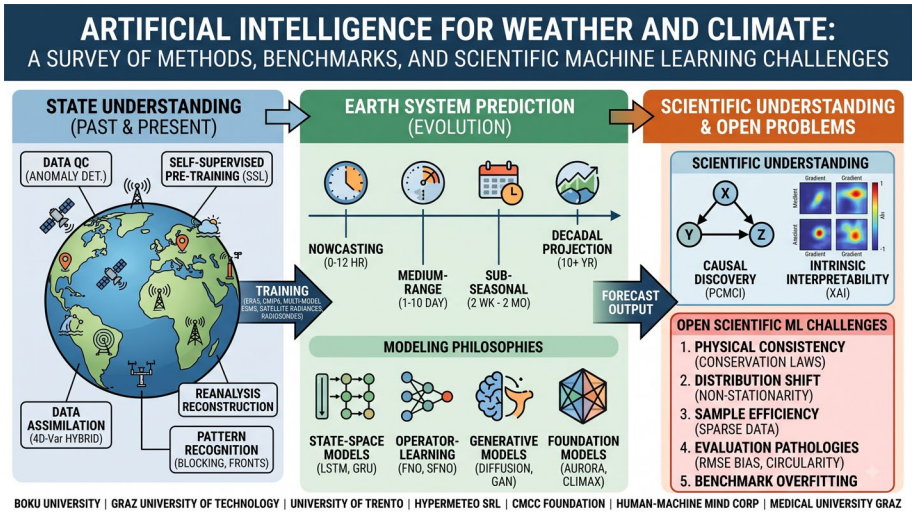
This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

https://doi.org/10.1007/s10462-026-11690-8

Published online: 11 September 2026

Springer

Graphical Abstract



Keywords Scientific machine learning · Neural operators · Geometric deep learning · Uncertainty quantification · Benchmark design · Distribution shift · Spatiotemporal learning · Foundation models

List of abbreviations

ACC	Anomaly correlation coefficient
AI	Artificial intelligence
CMIP	Coupled model intercomparison project
CNN	Convolutional neural network
CRPS	Continuous ranked probability score
DA	Data assimilation
DLWP	Deep-learning weather prediction
ECMWF	European centre for medium-range weather forecasts
ENSO	El Niño–Southern oscillation
ERA	ECMWF reanalysis
ESGF	Earth system grid federation
ESM	Earth system model
FNO	Fourier neural operator
FSS	Fractions skill score
GAN	Generative adversarial network
GNN	Graph neural network
GRU	Gated recurrent unit
LSTM	Long short-term memory
MJO	Madden–Julian oscillation
ML	Machine learning
NWP	Numerical weather prediction
PDE	Partial differential equation
PINN	Physics-informed neural network
QC	Quality control
RMSE	Root mean square error
S2S	Sub-seasonal to seasonal

SciML	Scientific machine learning
SFNO	Spherical Fourier neural operator
SHAP	Shapley additive explanations
SHT	Spherical harmonic transform
SSL	Self-supervised learning
UQ	Uncertainty quantification
XAI	Explainable artificial intelligence

1 Introduction

Artificial Intelligence (AI) has transitioned into a pervasive technology that permeates nearly all domains of modern society. At the core of this transformation lies machine learning, which functions as the principal operational paradigm enabling systems to learn patterns from data, generalise across contexts, and support or even automate decision-making processes. Contemporary AI systems are therefore less characterised by symbolic reasoning alone and more by data-driven learning mechanisms, where machine learning serves as the central workhorse driving performance, scalability, and adaptability across diverse real-world applications (Holzinger et al. 2026).

Weather and climate science is one of the most demanding testbeds for these methods (Reichstein et al. 2019; Rolnick et al. 2023). The data are global, spatiotemporal, and physically constrained: ERA5 reanalysis spans 1940 to the present at hourly resolution across hundreds of atmospheric variables (Hersbach et al. 2020); the CMIP6 archive holds roughly 28 PB of multi-model projections (Eyring et al. 2016; Jones et al. 2024; Durack et al. 2025); satellite networks produce of order one petabyte of imagery per day (Warren et al. 2015).

However, volume alone does not make this domain interesting for data-driven AI research. What makes it revealing is the full combination of challenges it presents: strong prior structure (known governing partial differential equations (PDEs)), irregular spherical geometry, non-stationary distributions driven by anthropogenic forcing, extreme evaluation difficulty (rare events, compound hazards), and high societal stakes that enforce rigour in deployment (Camps-Valls et al. 2025).

The presence of strong prior structures fundamentally differentiates this domain from typical machine learning benchmarks (Von Rueden et al. 2021). The governing dynamics are not unknown but encoded in systems of nonlinear PDEs (e.g. conservation of mass, momentum, and energy). This creates a setting where purely data-driven models must either approximate or respect well-established physical laws. Consequently, inductive biases become central: architectures such as neural operators, equivariant networks, or hybrid dynamical–statistical models are not optional refinements but necessary design choices to ensure physical consistency, stability, and extrapolation capability.

The geometric structure further complicates learning. Atmospheric and oceanic processes evolve on a rotating sphere with heterogeneous spatial resolution, irregular grids, and coordinate singularities (e.g. at the poles). Standard Euclidean assumptions underlying convolutional architectures break down, requiring spherical harmonics, graph-based representations, or coordinate-aware transformations. This introduces non-trivial challenges for scalability and numerical stability, particularly when attempting to preserve rotational equivariance and long-range dependencies across scales.

Few scientific domains simultaneously require all of: operator learning, geometric deep learning, distributional robustness, causal reasoning, and calibrated uncertainty quantification. Weather and climate science requires all of them, which is why it has become a central proving ground for scientific machine learning.

This survey approaches the literature from the AI side: not *"what AI is doing for weather and climate"* but *"what weather and climate reveal about AI methods for spatiotemporal physical systems"*.

1.1 Scope of scientific machine learning

In this survey, *scientific machine learning* (SciML) refers to machine-learning methods whose design, training, or evaluation is structured by the known physical organisation of the target system. This usage is narrower than "machine learning applied in science" and broader than "physics-informed neural networks" (Cuomo et al. 2022; Willard et al. 2023). Four epistemic roles recur throughout the manuscript and are kept distinct: (i) *learning as approximation*, where a neural component replaces a known but expensive operator (e.g. radiative transfer); (ii) *learning as surrogate modelling*, where the network emulates a full forward model (e.g. ESM emulation); (iii) *learning as inference under partial observation*, where the model recovers hidden state from sparse data (e.g. data assimilation, reanalysis reconstruction); and (iv) *learning as discovery*, where the goal is to recover relations, operators, or causal structure not contained in the original model (e.g. causal discovery of teleconnections, sub-grid parameterisations). Different roles call for different validation criteria, and we mark which role is at issue when this distinction matters.

This survey treats predictive skill, causal understanding, and explanatory interpretability as three distinct and not interchangeable objectives.

Predictive skill is operationalised through forecasting metrics (RMSE, ACC, CRPS); causal understanding refers to inferred driver–response relations in the physical system; and explanatory interpretability refers to whether a model's internal computations can be related to identifiable physical processes. The three are evaluated against different standards, and we preserve the distinction in Sects. 5 and 6. The emphasis is on the workhorses of AI, including representation learning, operator learning, geometric and equivariant deep learning, transformers for physical systems, uncertainty quantification and calibration, benchmark design and evaluation pathologies, distribution shift and non-stationarity, and interpretability as a route to scientific discovery. Individual forecast systems are described where they anchor methodological arguments, but they are not the primary topic. The past decade has produced landmark results that make these methodological questions concrete. Neural operators (Li et al. 2021; Lu et al. 2021; Bonev et al. 2023) provide a principled function-space framework for learning solution operators of PDEs, with zero-shot super-resolution as a practically testable consequence. Foundation models pre-trained on heterogeneous climate data (Nguyen et al. 2023; Bodnar et al. 2025) demonstrate transfer learning across scientific tasks. Generative ensemble systems (Ravuri et al. 2021; Price et al. 2025) expose the calibration–sharpness tension in a setting where both matter operationally. Hybrid differentiable models (Kochkov et al. 2024) dissolve the boundary between learned and physics-based computation. Each of these advances poses open questions that extend well beyond meteorology.

By a *testbed* we mean a scientific domain in which AI methods can be evaluated against four conditions that are rarely co-present elsewhere: (i) known governing equations whose violation is physically interpretable, (ii) globally available observational and reanalysis archives that fix a reproducible ground-truth pipeline, (iii) clearly defined skill metrics adopted by an operational community, and (iv) practical stakes that punish overconfident extrapolation. Weather and climate satisfy all four; pure fluid dynamics typically lacks (ii) and (iv), seismology typically lacks (iii), and standard vision or language benchmarks lack (i).

1.2 Relation to prior surveys

Three earlier surveys define the immediate baseline. Dueben and Bauer (2018) articulated the design choices for purely data-driven global weather and climate models before the transformer and neural-operator generations, focusing on feasibility and architectural constraints. Schultz et al. (2021) examined whether deep learning could match operational NWP and concluded that competitive skill was within reach but lacked physical guarantees. Rolnick et al. (2023) mapped machine learning across the climate mitigation and adaptation landscape, covering a wide application range with comparatively brief methodological depth on any one area. Zhang et al. (2025) provides an updated application-centred overview, and topic-specific reviews exist for nowcasting (Ravuri et al. 2021; De Luca et al. 2025), extremes (Salcedo-Sanz et al. 2024), and downscaling (Rampal et al. 2024). The

present survey differs from these in four respects: (a) it is organised by *physical inductive bias* rather than by application; (b) it analyses geometric deep learning and operator learning as paradigm-level contributions, which the earlier surveys mention only in passing; (c) it gives a dedicated, named treatment of evaluation pathologies (Sect. 5); and (d) it positions the open challenges as scientific machine learning problems generalisable to other physically constrained domains.

The specific contributions are: (1) framing of weather and climate as a scientific ML testbed, with explicit identification of the AI challenges the domain poses; (2) survey of AI architectures through their physical inductive biases, with geometric deep learning and operator learning as central themes; (3) systematic treatment of representation learning, foundation models, and self-supervised pre-training; (4) a standalone section on evaluation pathologies in scientific ML, including benchmark design and distribution-shift issues; (5) causal discovery and interpretability as routes to scientific understanding rather than model transparency alone; and (6) open challenges framed as scientific ML research problems, not meteorological forecasting problems.

1.3 Roadmap

The roadmap below states, for each section, the single point the reader should take away. Section 2 introduces architectural families and their physical inductive biases, geometric deep learning on the sphere, and four cross-cutting modelling philosophies (Table 2, Fig. 1); the lesson is that architecture choice in this domain is a statement about which physical symmetries and operators are built in versus learned. Section 3 surveys AI for state understanding—quality control, self-supervised pre-training, data assimilation, reanalysis, pattern recognition, and downscaling—and the reader should come away able to distinguish the four epistemic roles of learning (approximation, surrogate modelling, inference under partial observation, discovery) introduced in Sect. 1. Section 4 covers AI for state prediction across nowcasting, medium-range, sub-seasonal, and decadal timescales; the take-away is that the AI-versus-NWP framing is being replaced by hybrid differentiable systems, and that prediction skill must be read together with physical consistency. Section 5 is a standalone treatment of evaluation pathologies, and is the section whose lesson is most portable: standard metrics (above all RMSE) can reward the wrong behaviour, so the reader should learn to recognise the smoothness bias, training–evaluation circularity, and benchmark overfitting before trusting a leaderboard. Section 6 surveys causal discovery and explainability, and the reader should come away with a clear separation between predictive reliance, explanatory transparency, and genuine physical causation, together with calibrated scepticism about confirmatory causal claims. Section 7 examines deployment and infrastructure; the message is that energy, data provenance, and reproducibility are first-class infrastructure requirements, not afterthoughts. Section 8 synthesises the open challenges in scientific

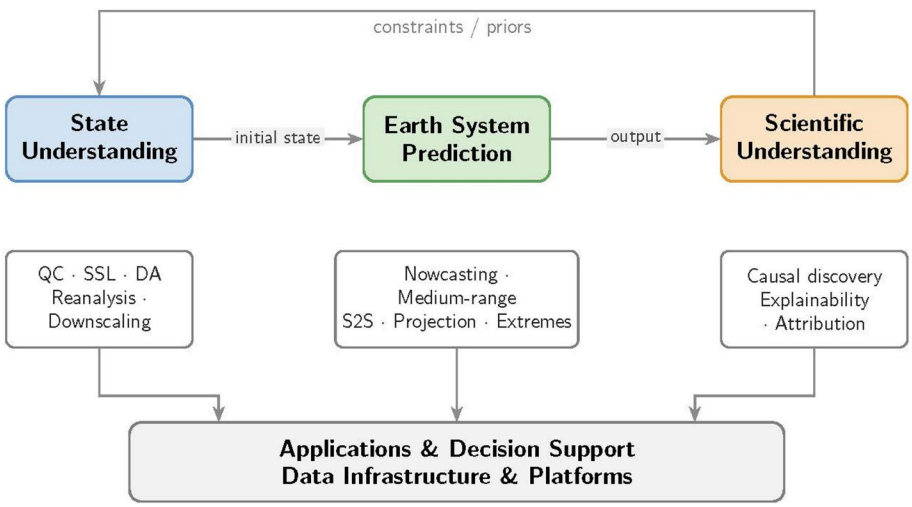


Fig. 1 The three primary AI task categories in weather and climate science, framed as scientific ML problems. *State understanding* poses representation learning, data assimilation, and reconstruction challenges. *Earth system prediction* poses operator learning, calibrated UQ, and distribution-shift challenges. *Scientific understanding* poses causal discovery and interpretability challenges. The feedback arc reflects that causal understanding informs architectural choices and physical constraints upstream (Sects. 3–6)

machine learning that this domain exposes, and is intended to leave the reader with a concrete agenda—distribution shift, physical consistency, sample efficiency, and intrinsic interpretability—generalisable beyond meteorology. Section 9 concludes with forward-looking recommendations.

1.4 Literature methodology

This article is a *critical, methods-oriented survey* rather than a systematic review in the PRISMA sense: we do not claim a fixed corpus defined by pre-registered inclusion and exclusion criteria, screening counts, or an exhaustive database census. Instead, the selection is expert-curated and argument-driven, prioritising works that anchor a methodological point about physical inductive bias, evaluation, or scientific machine learning. The survey covers approximately 2015–2026 (publication years of cited works), with pre-2015 references retained where they introduce foundational architectures (e.g. LSTM, ConvLSTM, ResNet). A small number of 2026 entries are included for items in press at the time of writing; their status as accepted but not-yet-archival is noted on first citation where relevant. Candidate literature was identified through structured (but non-exhaustive) searches of Web of Science, Scopus, and Google Scholar using terms including *neural operator atmosphere*, *Fourier neural operator PDE*, *geometric deep learning weather*, *equivariant network climate*, *foundation model earth system*, *scientific machine learning*, *distribution shift climate*, *causal inference teleconnections*, and *benchmark scientific ML*, supplemented by citation-chaining from key papers. Abstracts were read and DOIs verified where available. We make this scope explicit so that readers do not over-interpret the coverage as a systematic or quantitatively complete census of the field.

2 Background and categorisation

2.1 Characteristics of weather and climate data

Weather and climate datasets have several features that distinguish them from standard machine learning benchmarks and that shape architecture choices throughout this paper.

Spatiotemporal structure. Atmospheric fields are defined on a rotating sphere at multiple vertical levels, evolving in time under nonlinear dynamics. Any model that ignores either spatial geometry or temporal evolution loses physically meaningful structure.

Physical constraints. Valid atmospheric states must conserve mass, momentum, and energy. Standard MSE-trained networks have no mechanism to enforce these constraints, motivating physics-informed loss functions and equivariant architectures.

Scale heterogeneity. Atmospheric processes span from cloud microphysics at metres to planetary waves at thousands of kilometres. Multi-scale representation and generalisation across resolutions are both challenging.

Observational sparsity and bias. Historical reanalyses concentrate skill in data-rich mid-latitude regions (Liu et al. 2024). The tropics, Southern Ocean, and polar areas are undersampled, creating systematic biases that propagate into AI models trained on reanalysis (Ben Bouallègue et al. 2024).

Volume and heterogeneity. ERA5 spans several petabytes (Hersbach et al. 2020); CMIP6 roughly 28 PB (Eyring et al. 2016; Durack et al. 2025). Training on subsets of these requires infrastructure that "co-locates data with compute" (Elia et al. 2023; Fiore et al. 2018), where data storage and computational resources are placed in close physical proximity; this is discussed in Sect. 7.2.

2.2 AI architectures and their physical inductive biases

The choice of architecture encodes assumptions about the problem structure. Matching those assumptions to the physics—the *inductive bias*—is often decisive for generalisation from limited training data.

2.2.1 Convolutional neural networks

CNNs enforce *translational equivariance* through locally connected weight-shared filters (LeCun et al. 1998). For atmospheric fields on a regular grid, this reflects the observation that frontal gradients, convective cells, and cyclone signatures can appear at any horizontal location. Applications span downscaling (Vandal et al. 2017; Baño-Medina et al. 2020), precipitation nowcasting (Sønderby et al. 2020; Shi et al. 2017), and pattern classification (Liu et al. 2016). On a global spherical grid the flat-grid assumption introduces distortion, motivating two remedies: spherical CNNs that extend equivariance to the rotation group SO(3) (Cohen et al. 2018), and cubed-sphere remappings that preserve boundary-condition regularity at pole regions (Weyn et al. 2020).

2.2.2 Recurrent architectures

LSTM networks and GRUs encode *sequential temporal dependence* via gated memory (Hochreiter and Schmidhuber 1997). This bias matches station-level time series and slowly evolving boundary conditions. Applications include ENSO prediction at multi-year leads (Ham et al. 2019), streamflow modelling (Kratzert et al. 2018), and early data-driven global forecasting (Weyn et al. 2019); earlier feasibility studies had already applied generic deep networks to station-level forecasting (Salman et al. 2015). ConvLSTM (Shi et al. 2015) encodes spatial and temporal structure jointly, becoming the standard baseline for radar nowcasting.

2.2.3 Graph neural networks

GNNs represent the atmosphere as a graph whose nodes carry state variables and whose edges encode spatial or physical connectivity (Scarselli et al. 2009). Their key property is *invariance to grid geometry*: they operate natively on irregular meshes and encode physical adjacency directly in the edge structure. Keisler (2022) first demonstrated that a message-passing GNN on an icosahedral mesh can produce competitive global forecasts; GraphCast (Lam et al. 2023) and AIFS (Lang et al. 2024) both use GNN encoder–processor–decoder architectures on multi-scale icosahedral meshes, enabling global forecasting at 0.25° resolution.

2.2.4 Transformers and foundation models

Transformers (Vaswani et al. 2017) capture *global long-range dependencies* via self-attention with no locality constraint, making them naturally suited to atmospheric teleconnections. Pangu-Weather (Bi et al. 2023) uses a 3D Earth Attention Transformer; Aurora (Bodnar et al. 2025) is a foundation model pre-trained on a diverse corpus spanning ERA5, CMIP6, and operational analyses. ClimaX (Nguyen et al. 2023) applies masked autoencoding pre-training on heterogeneous CMIP6 data before fine-tuning for downstream tasks. Transformers are data-hungry relative to GNNs or CNNs—their weak inductive bias gives flexibility at the cost of higher compute.

2.2.5 Neural operators

Neural operators constitute a principled framework for learning *mappings between function spaces* rather than between finite vectors (Lu et al. 2021; Li et al. 2021). The distinction matters: the physical problem is inherently continuous, and the solution operator of the governing PDE should generalise across discretisations.

A classical network trained at 0.25° learns a fixed finite-dimensional map tied to that grid; a neural operator learns a continuous mapping that can in principle be queried at resolutions finer than those seen during training, enabling *zero-shot super-resolution* at inference time, though practical performance depends on training resolution and spectral truncation.

DeepONet (Lu et al. 2021) encodes the input function on a set of sensors (branch network) and the output domain (trunk network), proving universal approximation in function space.

The Fourier Neural Operator (FNO) (Li et al. 2021) parameterises the integral kernel in Fourier space via fast Fourier transforms, achieving a global receptive field at $\mathcal{O}(N \log N)$ cost. FourCastNet (Pathak et al. 2022) demonstrated that the Adaptive FNO (AFNO) scales to full global atmospheric prediction. The Spherical FNO (SFNO) (Bonev et al. 2023) extends this to spherical geometry using spherical harmonic transforms, ensuring rotational equivariance in the spectral domain.

2.2.5.1 When zero-shot super-resolution fails The claim that operator-learning models query the trained map at arbitrary resolution is, strictly, a property of the function-space objective rather than an empirical guarantee. In practice, performance at unseen resolutions is bounded by the training spectrum: scales not represented in the training data cannot be recovered without strong prior structure, and FNO-style models are vulnerable to spectral truncation when the test-time field contains energy beyond the trained band.

The Spherical FNO of Bonev et al. (2023) mitigates this on the sphere by working in spherical-harmonic space, but does not eliminate it. Independent evaluation evidence from Olivetti and Messori (2024) on medium-range extreme weather forecasting shows that AI systems trained on smoothed reanalysis underrepresent fine-scale extremes regardless of inference resolution, indicating that "zero-shot super-resolution" should be read as a function-space property with empirical limits, not as a routine deployment capability.

2.2.6 Generative models

GANs, VAEs, and diffusion models share the inductive bias of *distributional fidelity*: outputs are plausible samples from the data distribution rather than conditional means. This preserves high-frequency variability and extreme values that MSE-trained models smooth out. GenCast (Price et al. 2025) and AIFS-Diffusion (Lang et al. 2024) apply diffusion-based ensemble generation at global operational scale.

2.3 Geometric deep learning and equivariant architectures

A unifying perspective across architecture families is provided by *geometric deep learning*: the study of neural networks that respect the symmetry group of the domain (Bronstein et al. 2017; Cohen et al. 2018). The same perspective extends beyond gridded fields: 3D point cloud data can be transformed into graph-based representations that form digital twins, enabling the application of geometric deep learning and graph neural networks for structured spatial analysis. The integration of explainable AI and human-in-the-loop approaches with counterfactual reasoning is essential to ensure interpretability, validation, and reliable deployment of such models (Saranti et al. 2024; Holzinger et al. 2025a).

CNNs are equivariant under translation on a flat grid. Atmospheric data lives on the sphere, a domain with a different symmetry group (SO(3)), and physical laws are equivariant under rotation: in the absence of longitudinal heterogeneities, the same meteorological pattern rotated to any longitude should produce the same dynamical response. Architectures that violate this symmetry must learn it from data, at the cost of reduced sample efficiency and potential failure outside the training latitude range.

Three families address spherical geometry. *Spherical CNNs* (Cohen et al. 2018) extend convolution to the rotation group using generalised Fourier transforms on SO(3), achieving exact equivariance. *Cubed-sphere CNNs* (Weyn et al. 2020) tile the sphere with six cube faces, preserving boundary conditions and reducing pole distortion. *SFNO* (Bonev et al. 2023) works directly in spherical harmonic space via the real SHT, combining equivariance with the function-space properties of neural operators.

The practical consequence of correct geometric structure is not just accuracy but *generalisation*: a model equivariant under sphere rotation will correctly handle novel storm tracks at any longitude without requiring training examples at every location. The expected gain from imposing the correct symmetry group is structural rather than a fixed multiplier: the function class is restricted to maps that respect the symmetry, so spurious longitudinal dependencies cannot be learnt as nuisance variables.

Quantitatively, Bonev et al. (2023) report that SFNO produces stable autoregressive rollouts of order one year, in contrast to instability observed in FNO without spherical equivariance under the same training budget, which provides one of the few controlled comparisons of equivariance versus non-equivariance in this domain. A direct, quantified measure of the *sample-efficiency* benefit of equivariance is harder to obtain for global weather models specifically, but the most cited controlled evidence comes from an adjacent geometric-deep-learning domain: for interatomic potentials, the E(3)-equivariant NequIP architecture attains accuracy comparable to non-equivariant baselines using up to three orders of magnitude (i.e. on the order of 1000×) fewer training configurations (Batzner et al. 2022). Two caveats are essential. First, this benefit is not a fixed multiplier transferable across tasks: empirically, imposing the correct symmetry changes the *exponent* of the learning curve (test error versus training-set size), not merely its offset, so the effective data saving grows with the target accuracy rather than being a constant factor (Batzner et al. 2022). Second, the 1000× figure is established for molecular systems, not for atmospheric forecasting, where the corresponding controlled measurement has not yet been published; we therefore cite it as an order-of-magnitude indication of what correct geometric structure can buy, not as a weather-specific result. This is analogous to the role of CNNs in image recognition: the key contribution was not that CNNs were accurate on ImageNet, but that they required far less data than fully-connected networks to achieve that accuracy, because translational equivariance eliminated the need to learn the same pattern at every pixel.

2.4 Four modelling philosophies

Beyond architecture families, four *modelling philosophies* cut across the literature and shape both capabilities and limitations (see Table 1).

State-space models (LSTMs, recurrent architectures, state-space networks) treat the atmosphere as a dynamical system with latent hidden state. Their advantage is explicit temporal coherence; their limitation is propagation of errors in the latent state during long rollouts.

Operator-learning models (FNO, SFNO, DeepONet) learn the solution map of the governing PDEs in function space. Their advantage is resolution-invariant generalisation; their limitation is that standard operator-learning objectives have no mechanism to enforce conservation laws.

Generative models (diffusion, flow-matching, GANs) learn the joint data distribution and produce stochastic samples. Their advantage is sharp, physically plausible ensemble members; their limitation is that per-member compute is higher and calibration is more complex.

Foundation/pre-training models (Aurora, ClimaX, Earth observation foundation models) extract reusable representations from large heterogeneous archives via self-supervised objectives. Their advantage is transfer to data-scarce downstream tasks; their limitation is that pre-training biases propagate to all fine-tuned applications.

Table 1 When to prefer each modelling philosophy

Philosophy	Prefer when	Avoid when
State-space / recurrent	long-range temporal dependence dominates; data are sequential; rollout horizons are short	spatial structure dominates; rollouts are long and error accumulates
Operator-learning	PDE structure is well-understood; resolution-invariant inference is required; multiple discretisations matter	strict conservation laws must hold; out-of-band frequencies must be recovered
Generative	ensemble realism, multivariate dependence, or sharp extremes are required; risk is concentrated in tails	strict deterministic reproducibility is needed; calibration evaluation is expensive
Foundation / SSL	labelled data are scarce; many downstream tasks share signal; transfer learning is the bottleneck	a single high-value task with abundant labels; pre-training corpus biases are unacceptable

The table is a decision guide, not a ranking; many production systems combine philosophies (e.g. NeuralGCM combines a differentiable physical core with learned parameterisations, GenCast combines operator-style backbones with generative ensembling)

Table 2 AI architectures, physical inductive biases, typical strengths and failure modes in weather and climate applications

Architecture	Physical inductive bias	Strength	Typical failure mode
CNN (flat)	Local spatial correlation	Efficient pattern detection	Distortion on sphere
Spherical CNN	SO(3) equivariance	Global classification	High compute; limited depth
LSTM / GRU	Sequential temporal memory	Long-range temporal dependencies	Error accumulation in rollout
ConvLSTM	Spatial + temporal	Spatiotemporal nowcasting	Blurring at long lead
GNN	Irregular grid, physical adjacency	Multi-scale global forecast	Sensitive to mesh design (e.g. icosahedral mesh resolution and neighbourhood radius materially affect skill in Lam et al., 2023; Lang et al., 2024)
Transformer	Global long-range attention	Teleconnection modelling	Data-hungry; slow inference
FNO / SFNO	Discretisation-invariant operator	Zero-shot super-resolution	No conservation guarantee
DeepONet	Universal operator approximation	Multi-physics generalisation	Large sensor count needed
Diffusion / GAN	Distributional fidelity	Sharp ensemble members	Calibration complexity
Foundation (SSL)	Heterogeneous pre-training	Transfer to scarce tasks	Pre-training bias

2.5 Decision guide: choosing among philosophies

The four philosophies just described are not exclusive; many operational systems combine them. Table 1 summarises the conditions under which each is the natural default.

3 AI for understanding and refining the earth system state

The first primary function of AI in atmospheric science is extracting, correcting, reconstructing, and refining estimates of the current or past Earth system state from available observations and model output. Six task families constitute this function, as explained in the following subsections.

3.1 A categorisation of weather and climate AI tasks

Table 2 maps architectures to applications. Figure 1 illustrates the three-function pipeline and the relationships among its parts. The survey is organised around the first two of these functions, state understanding and state prediction, followed by scientific understanding, applications, and infrastructure.

Unlike prediction, which is dominated by forecasting skill metrics such as RMSE, ACC, and CRPS, state-understanding tasks are evaluated through reconstruction fidelity, physical consistency, and downstream utility in analysis or Data Assimilation (DA) pipelines. This evaluation contrast matters for architecture choice: predictive architectures optimise the conditional-mean output, whereas state-understanding tasks often require uncertainty estimates (for QC decisions), spatial coherence (for reconstruction), or physical balance (for DA initial conditions). A further distinguishing feature is that state-understanding models are rarely applied in isolation: their outputs enter DA cycles, calibration pipelines, or downstream impact models, and errors propagate through these chains in ways that single-task evaluation cannot diagnose. This systemic perspective motivates several of the methodological tensions discussed at the end of each subsection below.

3.2 Data quality control and anomaly detection

QC and anomaly detection span three distinct but related tasks: *observational QC* (detecting instrument errors, gross outliers, and encoding failures in station or satellite records), *event and regime anomaly detection* (identifying unusual atmospheric states in gridded fields or model output), and *infrastructure anomaly detection* (detecting failures in data pipelines and archive access patterns).

For observational QC, convolutional autoencoders detect anomalous patterns in gridded reanalysis fields by learning compact representations of typical atmospheric states and flagging reconstructions with large residuals (Giffard-Roisin et al. 2020). LSTM networks applied to station time series detect sensor drift and step changes that evade threshold-based methods; related multi-model recurrent architectures have also been used for short-term sea-level forecasting (Accarino et al. 2021). Ensemble-model approaches combining random forests with physical consistency checks now underpin operational QC at several national meteorological services, replacing manual inspection workflows that could not scale to modern observational volumes.

Event anomaly detection has been applied to blocking identification, tropical cyclone precursor detection, and atmospheric river classification. Deep convolutional classifiers applied to ERA5 vorticity fields identify blocking events with higher skill than rule-based indices, while being robust to inter-model variability in CMIP6 archives (Gagne et al. 2019). A recurring methodological tension in QC and anomaly detection is the distinction between a model flagging a record as *anomalous relative to training data* and a record being physically *impossible*. Rarity alone is not a detection criterion—extreme events are rare by definition and must not be filtered. Architectures that encode physical consistency constraints, or that output calibrated anomaly probabilities rather than hard flags, are better positioned to resolve this tension.

Infrastructure anomaly detection models trained on historical access statistics from Earth System Grid Federation (ESGF) data nodes[1] can detect hardware failures and network congestion in near-real time (Cinquini et al. 2014; Petrie et al. 2021). While a relatively niche application, it illustrates a broader class of AI-for-operations use cases in which the data themselves, rather than their physical content, are the subject.

3.2.1 Evaluation of QC and anomaly detection

A central evaluation difficulty is that ground truth for rare and extreme atmospheric events is itself uncertain: the very events of interest are underrepresented in any labelled training corpus. Three pragmatic strategies are used in the literature. *(i) Synthetic injection*, where known artefacts (sensor drift, instrument spikes, gross outliers) are programmatically inserted into clean records and the model's recovery rate is reported as precision and recall; this gives reproducible numbers but only evaluates the failure modes the analyst already understands. *(ii) Expert-curated case studies*, where meteorologists annotate historical periods with known quality issues (e.g. documented station relocations) and detection metrics are computed against these annotations; this is closer to operational deployment but small-sample. *(iii) Downstream impact metrics*, where the value of QC is measured by the improvement in a subsequent forecast or assimilation step rather than directly. For physically impossible states (e.g. negative specific humidity, super-saturation beyond physical bounds), hard constraints derived from thermodynamics provide deterministic ground truth and should be reported separately from distributional anomaly metrics.

3.3 Self-supervised learning and representation pre-training

Self-supervised learning (SSL) extracts reusable representations from large unlabelled climate archives, reducing dependence on scarce labelled data. Masked autoencoder objectives, where a fraction of input patches is masked and the model is trained to reconstruct them, are now the dominant SSL paradigm for spatiotemporal data.

ClimaX (Nguyen et al. 2023) applies this approach to heterogeneous CMIP6 model output, pre-training a transformer encoder on reconstruction before fine-tuning for forecasting, downscaling, extreme event detection, or climate mode identification. A single pre-trained encoder handles all downstream tasks, mirroring the transfer-learning paradigm that transformed natural language processing. Aurora (Bodnar et al. 2025) extends this to a more diverse corpus spanning ERA5, operational analyses, and CMIP6.

[1] ESGF data usage and data publication metrics. https://esgf-ui.cmcc.it/

In Earth observation, SatMAE and related geospatial masked autoencoders are pre-trained on multi-spectral and multi-temporal satellite archives (Sentinel-2, Landsat), generating general-purpose geospatial encoders fine-tunable for drought monitoring, flood mapping, and land-cover change (Camps-Valls et al. 2021). The Prithvi foundation model (NASA/IBM) (Hsu et al. 2025) demonstrates similar transfer from Harmonised Landsat–Sentinel (HLS) satellite data to a range of downstream geoscience tasks, establishing that SSL pre-training is not limited to atmospheric grid data.

A key methodological tension in SSL and foundation models is between *transfer* and *bias propagation*. A foundation encoder pre-trained on ERA5 brings ERA5 biases to every fine-tuned application, including operational post-processing systems and climate attribution tools that are supposed to correct those biases. This is analogous to the circularity problem in ERA5-trained forecasting models (Ben Bouallègue et al. 2024). The mechanism by which a more diverse pre-training corpus mitigates, but does not remove, this circularity operates through three pathways. First, when the corpus combines ERA5 with operational analyses, in-situ observations, and other NWP centres' reanalyses, the encoder cannot trivially memorise ERA5-specific assimilation artefacts because they are inconsistent with the other sources at the same locations and times. Second, multi-source pre-training provides regularisation through implicit ensemble averaging: features that appear only in one reanalysis are downweighted relative to those that recur across systems. Third, masked-reconstruction objectives on heterogeneous data force the model to learn variables that transfer across sources rather than source-specific decoration. The mitigation is partial because all modern reanalyses share many of the same underlying observation streams and NWP traditions; truly independent verification requires direct comparison against raw radiosonde or satellite radiances that no reanalysis has ingested. Foundation models pre-trained on more diverse corpora are correspondingly more demanding in data engineering and storage cost.

3.4 Machine learning data assimilation

Data assimilation (DA) combines model state estimates with observations to produce optimal atmospheric initial conditions for NWP, and is among the most computationally expensive components of operational forecasting. ML addresses this cost through several routes.

Neural network observation operators replace expensive radiative transfer calculations in satellite radiance assimilation, mapping measured radiances to atmospheric state increments orders of magnitude faster than physical codes (Han et al. 2021). Deep learning surrogates for ensemble Kalman filter covariance propagation reproduce ensemble spread at substantially lower cost than full-ensemble integrations.

A parallel line of research applies ML specifically to model-error estimation within the 4D-Var framework. Bonavita and Laloyaux (2020) demonstrated that a neural network trained on weak-constraint 4D-Var analysis increments in the ECMWF IFS can learn systematic model errors and correct them in situ, reducing forecast bias in the stratosphere. Farchi et al. (2021) extended this to online model-error correction for a quasi-geostrophic channel model, showing that online training yields better correction than offline pre-training. These results establish ML as a tractable route to operationally correcting systematic model errors that conventional bias-correction schemes cannot easily estimate.

End-to-end differentiable DA systems optimise the forecast model and assimilation update jointly via back-propagation. Brajard et al. (2020) demonstrated this principle on the Lorenz 96 model, combining an ensemble Kalman filter with a neural network emulating unresolved dynamics. NeuralGCM (Kochkov et al. 2024) scaled the differentiable-hybrid principle to real Earth-system data: its machine-learning components are trained end-to-end on up-to-five-day weather trajectories sampled from the ERA5 reanalysis, with the learned parameterisations optimised in the context of the governing dynamical equations. A subsequent, distinct line of work built on the same framework—Yuval et al. (2026)—instead trains directly on satellite-based precipitation observations and substantially improves precipitation simulation; it is this follow-up, not the original 2024 model, that bypasses reanalysis on the precipitation side. A central methodological tension in ML-DA research is between the modular approach—replacing individual DA components with learned surrogates while preserving the overall 4D-Var

structure—and the end-to-end approach, which allows gradient-based optimisation through the full system but makes individual components harder to diagnose and audit. The modular approach is operationally safer; the end-to-end approach is more powerful.

3.4.1 Why end-to-end systems are hard to debug

The debugging difficulty of end-to-end differentiable DA is structural rather than incidental. Three mechanisms are responsible. (i) Error attribution is lost: when a gradient flows through a learned observation operator, a learned dynamical core, and a learned background-covariance proxy in a single optimisation, a forecast failure cannot be cleanly traced to which component is at fault. (ii) Modular operational practices fail: conventional NWP centres inspect background departures, analysis increments, and observation rejections as separate diagnostic streams to identify regression sources before deployment; in an end-to-end model these streams are no longer separated from the model parameters themselves. (iii) Failures are silent: a model can produce physically plausible but systematically wrong analyses when the joint optimum reflects compensating errors in different components, so routine monitoring (loss curves, validation skill) does not flag the underlying inconsistency. These mechanisms motivate the modular approach for operational deployment even where end-to-end is feasible in research settings. Extending the end-to-end principle to the full variational 4D-Var framework at global operational resolution remains an open challenge.

3.5 Reanalysis generation and historical reconstruction

Reanalyses combine observations with NWP models but inherit biases from the underlying dynamical core and parameterisation schemes. ML addresses both the bias problem and the coverage problem through two distinct routes.

The first route is *gap-filling and historical reconstruction*. Kadow et al. (2020) applied image-inpainting neural networks to HadCRUT4 global temperature observations, filling spatial gaps in the pre-satellite era via transfer learning from CMIP5 and 20CR and revealing a stronger global warming trend than previous estimates. Related inpainting approaches have been applied to precipitation extremes (Zscheischler et al. 2020), soil moisture, and sea-surface temperature records. A key advantage over kriging and EOF-based infilling is that neural inpainting learns nonlinear covariance structure from model climatology before applying it to observational gaps.

The second route is *fully data-driven reanalysis*: interpolating historical observations directly onto a global grid without an NWP forward model. The motivation is to eliminate the circular dependence on NWP that affects ERA5-trained AI forecast models (Ben Bouallègue et al. 2024). Transformer-based approaches that assimilate heterogeneous observational types—radiosonde, satellite, surface—in a single pass have begun to appear. Aurora (Bodnar et al. 2025), while pre-trained on a corpus including ERA5, demonstrates the architectural principle of ingesting multi-source heterogeneous observations through a single transformer encoder. On the training side, the precipitation-focused extension of NeuralGCM (Yuval et al. 2026) optimises directly against satellite-based precipitation observations rather than reanalysis; the original Kochkov et al. (2024) model, by contrast, is trained on ERA5 trajectories. Beyond these published systems, the European Horizon project WeatherGenerator (Sect. 7.2) and the corresponding ECMWF-led foundation-model activity are actively pursuing observation-first training pipelines. We refrain from naming further unpublished efforts because the field is moving faster than peer review, and because indiscriminate citation of preprints on a topic where independent verification is critical would conflict with the present survey's stated commitment to evidentiary rigour. This area is constrained by three overlapping challenges: (1) the historical observational record is sparser and noisier than modern reanalyses, requiring models to generalise far outside their training distribution; (2) there is no NWP forward model to enforce physical consistency across variables, so learned reconstructions can violate atmospheric balance; (3) evaluation is difficult because ground truth itself is uncertain in data-sparse historical periods. Progress on each is necessary before data-driven reanalysis can complement or replace NWP-based products at global scale.

3.6 Pattern recognition in observational archives

3.6.1 Satellite and geospatial imagery

Deep CNNs have been applied at scale to cloud-type classification, tropical cyclone intensity estimation from passive microwave imagery (Wimmers et al. 2019), surface temperature anomaly detection in MODIS records (Barnes et al. 2019), and automated identification of atmospheric rivers, weather fronts, and blocking events (Liu et al. 2016; Gagne et al. 2019). Transfer learning from large natural image pre-trained networks (ResNet, VGG) reduces annotation burden when labelled climate examples are scarce (Han et al. 2021). Geospatial machine learning—AI applied to satellite and Earth observation data at scale—has emerged as a distinct research field with its own benchmarks and foundation models (Camps-Valls et al. 2021), and provides upstream representations that atmospheric science is beginning to exploit.

A persistent limitation across satellite-based pattern recognition is that high classification skill does not translate into mechanistic understanding. A model can identify atmospheric river objects with high precision and recall while providing no information about the dynamical mechanisms responsible. Architectures that integrate physical constraints (mass flux conservation, moisture transport budgets) into detection objectives are beginning to address this, but the field remains dominated by discriminative approaches evaluated on held-out ERA5 subsets. The choice of training archive also matters: classifiers trained on ERA5-derived labels inherit ERA5's representation of the very features they are learning to detect.

3.6.1.1 Inference cost Order-of-magnitude figures help locate where computational cost is, and is not, a binding constraint. A ResNet-50-class classifier applied to a typical full-disk satellite image (of order a few megapixels) requires on the order of tens of milliseconds on a contemporary GPU per scene at inference, so a single GOES- or Meteosat-disk image cycle (typically 10–15 min between full disks) is classified in well under a second on a single accelerator. Heavier transformer-based segmentation architectures or 3D models operating on multi-spectral channels scale up into the few-seconds-per-scene regime. At national-archive scale, however, the binding cost is data movement rather than compute: moving petabyte-scale satellite archives to a training cluster is the dominant cost, which is why the infrastructure layer (Sect. 7.2) emphasises co-locating compute with the archives. We deliberately avoid quoting precise wall-clock numbers from individual papers because reported speeds depend strongly on accelerator generation and batch configuration; the qualitative ordering (inference cheap, training and data movement expensive) is the durable point.

3.6.2 Radar and ground-based networks

Ground-based radar networks provide high-resolution precipitation and wind profiles over land areas. ConvLSTM-based event-sequence models applied to NEXRAD mosaics perform automated severe-storm detection and hail nowcasting at sub-hour lead times. In space weather, incoherent scatter radars such as EISCAT provide four-decade archives of ionospheric electron density, temperature, and plasma velocity. AI applied to these archives enables automated detection of rare phenomena including sporadic-E layers, polar mesospheric echoes, and near-Earth space objects (Kvammen et al. 2020). The volume of multi-decade radar archives makes unsupervised and SSL approaches particularly attractive for initial discovery tasks.

3.7 Spatial downscaling and super-resolution

Downscaling—translating coarse-resolution model output to fine-resolution fields—functions as *state refinement*: the input contains the large-scale atmospheric state and the output resolves sub-grid structure absent from the coarse model. CNN-based statistical downscaling was pioneered by Vandal et al. (2017); the DeepESD models (Baño-Medina et al. 2020) systematically compared CNN architectures across European variables, establishing

benchmark baselines still in use. Physics-informed objectives constrain outputs to satisfy area-mean conservation (Beucler et al. 2021), addressing the mass-budget violations that unconstrained networks produce.

Generative approaches—GANs, diffusion models, and flow-matching networks—preserve spatial extremes that deterministic models smooth away (Leinonen et al. 2021; Price et al. 2025). The tension between marginal skill metrics and spatial realism is a recognised methodological issue: a model can improve RMSE by smoothing fine structure, while a generative model scores worse on RMSE but produces physically plausible realisations. Evaluation therefore requires both pointwise skill metrics and spatial pattern diagnostics. Rampal et al. (2024) provide a comprehensive review including explainability methods applied to downscaling models, and conclude that uncertainty quantification is consistently the most underdeveloped component of current systems. Specifically on out-of-distribution generalisation, Rampal et al. (2024) argue that empirical downscaling algorithms learn statistical relationships from a fixed historical climate and that the validity of transferring these relationships to future warmer climates depends on the assumption of stationarity in the large-scale-to-local-scale mapping, an assumption that cannot be verified against present-day observations and that may fail when thermodynamic or circulation regimes shift outside the training range.

4 AI for predicting earth system evolution

The second primary function is generating forecasts, projections, or probabilistic ensembles of how the atmosphere and climate system will evolve. This function spans timescales from minutes (nowcasting) to centuries (climate projection), and has attracted the largest fraction of research effort in recent years.

4.1 Deterministic, probabilistic, and generative forecasting

Three approaches have emerged with distinct properties.

Deterministic models produce a single best-estimate forecast, typically trained by minimising mean squared error (MSE). Pangu-Weather (Bi et al. 2023), GraphCast (Lam et al. 2023), FuXi (Chen et al. 2023b), and FengWu (Chen et al. 2023a) are deterministic systems. MSE training produces conditional-mean forecasts that sharpen at short leads but blur at long leads where atmospheric states diverge, systematically underestimating extreme values.

Probabilistic models output distributions over future states via Monte Carlo dropout (Gal and Ghahramani 2016), deep ensembles, or distribution-parametrising output heads; an early atmospheric demonstration predicted forecast uncertainty directly from the initial state (Scher and Messori 2018). They provide calibrated marginal uncertainty estimates but may still misrepresent the multivariate dependence structure across variables—a failure mode particularly consequential for compound event prediction.

Generative ensemble models use diffusion models or flow-matching to produce physically plausible ensemble members that collectively represent forecast uncertainty. GenCast (Price et al. 2025) and AIFS-Diffusion (Lang et al. 2024) are operational-scale examples. Their key advantage is that each member is individually sharp and captures multivariate dependence structure; the cost is more complex training and calibration.

A critical conceptual point, noted by Zscheischler et al. (2020) in the compound-events context, is that marginal skill metrics cannot diagnose failures in joint probability structure. Compound events require ensemble members whose co-occurrence statistics are correct, not just members with correct marginal distributions.

Several methodological tensions cut across all three approaches and recur throughout Sect. 4.

Sharpness versus calibration. Generative models produce sharp members but can be overconfident; deterministic models are smooth but never sharp enough for impact assessment. Calibration is a necessary but not sufficient condition for useful ensemble output: a well-calibrated but uninformative ensemble has correct coverage but no forecast value.

Expressivity versus inductive bias. Transformers and large foundation models achieve the highest deterministic skill by virtue of their global attention and massive parameter counts, but their weak physical inductive bias means they can produce unphysical states outside their training distribution. Neural operators and equivariant architectures impose stronger physical structure at the cost of flexibility. The right trade-off depends on the target application: operational medium-range forecasting favours expressivity; downscaling applications in data-sparse regions favour inductive bias.

Full AI versus hybrid differentiable. Pure AI models (Pangu-Weather, GraphCast, FuXi) are fast and competitive on aggregate scores but have no mechanism to enforce conservation laws. Hybrid differentiable models (NeuralGCM, ClimODE) enforce physical structure through their dynamical cores but are harder to train and operationalise. The field is moving towards hybrid systems, but the convergence is not yet complete.

4.1.1 Benchmarks and deterministic metrics

WeatherBench (Rasp et al. 2020) established the first standardised ERA5-derived dataset and RMSE metrics for global medium-range prediction, enabling reproducible comparison between data-driven models. WeatherBench 2 (Rasp et al. 2024) updated this to accommodate higher-resolution models and added probabilistic metrics and comparison against operational NWP baselines. RMSE and Anomaly Correlation Coefficient (ACC) remain standard deterministic metrics, but carry a known bias towards smooth forecasts: a blurry but unbiased forecast incurs lower RMSE than a sharp but slightly displaced one. Also of interest is ClimateBench (Watson-Parris et al. 2022), which introduces a benchmarking framework for climate model emulation, combining outputs from comprehensive Earth System Models with baseline machine learning models to approximate climate responses across diverse emission scenarios. By enabling efficient prediction of global temperature and precipitation patterns under varying greenhouse gas forcings, it provides a scalable alternative to computationally expensive simulations and establishes a foundation for advancing physically consistent and interpretable climate emulators.

4.1.2 Probabilistic metrics

The Continuous Ranked Probability Score (CRPS) is the standard probabilistic metric, measuring the average distance between the forecast cumulative distribution function and the step-function at the observation. AIFS-Diffusion and GenCast are competitive with the ECMWF IFS ensemble under CRPS on upper-air variables (Lang et al. 2024; Price et al. 2025). Reliability diagrams and spread–skill plots assess calibration.

4.1.3 Verification of extremes and compound events

Standard aggregate metrics are dominated by typical conditions and can mask poor skill on rare extremes. Olivetti and Messori (2024) demonstrate that most published AI weather models have not been rigorously evaluated on extreme events. Threshold exceedance scores, quantile-based metrics, and case studies on high-impact historical events are needed alongside aggregate scores. Compound events—simultaneous or cascading hazards whose combined impact exceeds any single driver (Zscheischler et al. 2020)—require multivariate evaluation capturing dependence structure, not just marginal distributions. AI models trained on univariate MSE objectives can perform well on marginal scores while misrepresenting the joint probability critical for risk assessment.

A more extended treatment of evaluation pathologies in scientific machine learning is given in its own standalone Sect. 5, placed between the prediction and scientific-understanding sections so that it can serve as a reference point for both.

4.2 Precipitation nowcasting

Precipitation nowcasting was one of the earliest AI applications in operational meteorology. ConvLSTM (Shi et al. 2015) established that recurrent convolutional networks outperform optical-flow methods for radar-based nowcasting. MetNet (Sønderby et al. 2020) demonstrated that self-attention architectures with large spatial context windows produce probabilistic precipitation maps up to 8 h ahead at 1 km resolution, outperforming NOAA HRRR at forecast leads exceeding 7 h. MetNet-2 (Espeholt et al. 2022) extended this to 12 h by quadrupling the input context and adopting dilated receptive fields; on both CRPS and the Critical Success Index against the HREF ensemble baseline, MetNet-2 outperforms NWP at all 12 lead times while running in seconds rather than the hour required by physics-based ensembles. The generative DGMR (Ravuri et al. 2021) produces ensemble nowcasts preferred by operational meteorologists over both deep learning baselines and NWP under the Fractions Skill Score and CRPS. UNet-based post-processing of NWP gridded output for precipitation at kilometre scales has moved from research into early operational and pre-operational use at several European meteorological services, though comprehensive, independently documented deployment inventories remain scarce in the peer-reviewed literature.

A methodological tension runs through this subsection: MetNet and its successors are trained and evaluated on CONUS radar and satellite data, while DGMR is trained on UK Met Office radar. Both regional and continental-scale systems achieve strong results within their training domains, but neither has yet demonstrated consistent skill at global scale or in observation-sparse regions. Generalisation across climate regimes—where the training distribution differs from the target—remains an open problem for nowcasting systems.

4.3 Medium-range weather forecasting

Medium-range (1–10 day) forecasting has been the most intensively studied AI timescale, and the one where the transition from research to operational deployment is most advanced.

Early baselines by Weyn et al. (2019, 2020) and the design analysis by Dueben and Bauer (2018) established that purely data-driven global forecasting was feasible, though below operational NWP skill. FourCastNet (Pathak et al. 2022) scaled to full ERA5 resolution via AFNO. Pangu-Weather (Bi et al. 2023) and GraphCast (Lam et al. 2023) then demonstrated that AI models match or exceed ECMWF HRES on RMSE against ERA5 at 1–10 day leads. FuXi (Chen et al. 2023b) was the first to reach ECMWF ensemble-mean skill at 15-day leads using a cascade architecture; FengWu (Chen et al. 2023a) extends skilful lead time beyond 10 days.

The step from research to deployment followed quickly. AIFS (Lang et al. 2024) became fully operational in February 2025, running at 0.25° with approximately 1,000× lower energy per forecast than IFS. Aurora (Bodnar et al. 2025) reports superior RMSE against ERA5 on more than 91% of evaluated targets. Together these systems establish that data-driven global forecasting has matured from a research concept to an operational product in under a decade.

A critical caveat applies to all ERA5-trained, ERA5-evaluated models: verification against the same reanalysis used for training creates circularity (Ben Bouallègue et al. 2024). Several groups are developing AI systems trained directly on observations to break this dependency.

4.4 Hybrid and differentiable models

The framing of AI versus NWP is increasingly obsolete. The field is converging on *hybrid differentiable models* that couple learned components with physical dynamical solvers (Kochkov et al. 2024). NeuralGCM combines a differentiable spectral dynamical core for large-scale atmospheric flow with neural network parameterisations for sub-grid processes, trained end-to-end on up-to-five-day weather trajectories sampled from ERA5. Because all components are differentiable, gradient-based optimisation connects the training targets through the dynamical core to the learned components; a later application of the same framework trains directly on satellite-based

precipitation observations (Yuval et al. 2026). ClimODE (Verma et al. 2024) implements atmospheric advection as a neural ODE, enforcing value-conserving dynamics while enabling uncertainty estimation. Physics-informed neural networks (Raissi et al. 2019) embed PDE residuals directly in the training loss, a conceptually important approach though currently limited in operational scale.

At the parameterisation level, ML is embedded in conventional NWP through learned convection schemes (Rasp et al. 2018; Gentine et al. 2018), radiation emulators, and gravity-wave drag networks. These integrations are often easier to operationalise than full AI systems. They do, however, raise questions about error propagation and physical consistency, and carry the risk of failure when learned components encounter out-of-distribution inputs.

4.4.1 Why PINNs do not yet scale to operational global forecasting

Despite their conceptual appeal, physics-informed neural networks (PINNs) have not reached operational global atmospheric prediction. Two structural reasons are responsible. First, PINNs are subject to a documented *spectral bias*: training with PDE-residual losses encourages convergence on low-frequency components first, so the high-wavenumber content that carries cyclone-scale and convective-scale signal in operational forecasts is learnt slowly and incompletely without specialised loss balancing. Second, the *training cost* of soft-constrained PDE-residual objectives at global, high-resolution, multi-variable configurations is prohibitive: each gradient step requires evaluating the PDE residual across the full state-space at hundreds of collocation points per grid cell, which compounds with the already large state-size of operational forecasts. Hybrid differentiable models such as NeuralGCM (Kochkov et al. 2024) sidestep both issues by retaining a conventional dynamical core for the well-understood large-scale dynamics and using neural components only for the sub-grid processes that the PDE representation does not resolve.

4.5 Seasonal and sub-seasonal prediction

Sub-seasonal-to-seasonal (S2S) prediction (two weeks to three months) sits at the boundary of atmospheric predictability. Initial atmospheric state—relevant for weather—has largely decorrelated by week three; from that point, predictability derives from slowly evolving boundary conditions (sea-surface temperature, soil moisture, sea ice) and organised tropical modes. AI exploits these sources through high-dimensional pattern matching that dynamical models can also represent but at greater computational cost.

For ENSO prediction, Ham et al. (2019) demonstrated LSTM-based skill at up to 18-month leads, outperforming dynamical models at long horizons, in part because the network can exploit precursor patterns in the Pacific that NWP models underweight. GNN approaches exploit teleconnection graphs to improve predictions of ENSO diversity —distinguishing central-Pacific from eastern-Pacific events—which matters for downstream impacts (Cachay et al. 2021).

The Madden–Julian Oscillation (MJO), the dominant mode of tropical intraseasonal variability, is the primary predictability source for the 3–4 week window. Deep learning bias correction of multi-model S2S forecasts reduces four-week-mean MJO amplitude and phase errors by approximately 90% and 77% respectively relative to uncorrected dynamical forecasts (Kim et al. 2021). This hybrid dynamical–statistical approach—applying ML post-processing to state-of-the-art NWP output rather than replacing it—consistently outperforms purely data-driven MJO prediction at sub-seasonal leads.

Sub-seasonal large-ensemble forecasting using computationally cheap deep-learning weather prediction (DLWP) models has been demonstrated at near-ECMWF-ensemble quality with a 320-member ensemble computed in seconds rather than hours (Weyn et al. 2020). A central methodological tension in S2S work is between models that learn directly from observations and models that apply ML post-processing to dynamical output. The former avoids NWP biases but faces severe training-sample limitations; the latter inherits NWP skill but risks learning systematic model errors. Hybrid approaches—where ML corrects model error while dynamics provides

physical structure—appear to offer the best current trade-off. A deeper limit applies across all S2S methods: the predictability horizon is set by ocean–atmosphere coupling timescales, not by model architecture. AI can more efficiently exploit the signals already present in SST, soil moisture, and sea-ice anomalies, but cannot extend predictability beyond the physical window that those boundary conditions afford.

4.6 Decadal and centennial projection

Climate projection differs fundamentally from forecasting: initial atmospheric state is irrelevant over multi-decadal horizons and the forced response to greenhouse gas concentrations drives the signal. AI contributions include ESM emulation of full simulation outputs at a fraction of the cost (Rasp et al. 2018; Gentine et al. 2018), bias correction and model weighting for multi-model ensembles (Pan et al. 2021; Brunner et al. 2020), and detection and attribution of observed anomalies (Barnes et al. 2019; Trok et al. 2024). NeuralGCM (Kochkov et al. 2024) reproduced four-decade global mean temperature trends more accurately than traditional atmospheric models in hindcast experiments. These results suggest that end-to-end differentiable hybrid models may eventually replace offline bias-correction pipelines.

The distinction among the three use cases merits clarity: *emulation* (replacing an ESM forward model with a fast surrogate), *bias correction and weighting* (improving multi-model projections at regional scales), and *detection and attribution* (estimating the probability that observed anomalies are inconsistent with natural variability). Each calls for different training objectives and evaluation protocols (Ullrich et al. 2025).

4.7 Extremes, attribution, and compound risk

Accurate prediction and attribution of extreme events is where AI could deliver the highest societal impact and where evaluation is most demanding. Graph-based models track and estimate the intensity of tropical cyclones, atmospheric rivers, and blocking events (Gagne et al. 2019). Chattopadhyay et al. (2020) identified multi-week heatwave precursor patterns in ERA5, exceeding climatological baselines at 3–4 weeks' lead. Counterfactual CNN attribution of observed extreme heat events to anthropogenic forcing has been demonstrated on multiple historical events (Trok et al. 2024).

Compound events (Zscheischler et al. 2020) deserve explicit attention that the univariate extremes literature does not provide. The four compound types—multivariate, pre-conditioned, temporally compounding, and spatially compounding—require different modelling approaches. A fundamental constraint is sample size: Bevacqua et al. (2023) demonstrate that robust estimation of compound event probabilities requires data volumes far larger than those needed for univariate extremes, motivating the use of large ensemble climate model simulations that provide thousands of years of synthetic weather. This is where AI and large-ensemble simulation are particularly complementary: generative emulators can produce large synthetic ensembles for frequency and attribution analysis at a fraction of the cost of dynamical models. The qualifier matters: as ensemble size grows, the binding constraint shifts from sampling variance to the fidelity of the generator itself. A generative emulator that misrepresents the tail or the multivariate dependence structure of the underlying climate produces large numbers of *wrong* samples just as fast as it produces large numbers of correct ones, and may give a false impression of statistical confidence through narrow quantile uncertainty intervals. Quantity therefore does not substitute for quality: a 10^5-member emulator ensemble is only useful for compound-risk estimation if independent verification on held-out events establishes that the tail and dependence structure are sufficiently calibrated.

Active ML research directions include generative models that learn multivariate dependence structure, copula-integrated neural networks for joint probability estimation, and sequence models for cascading hazard chains. Evaluation must go beyond marginal scores: AI systems can achieve correct marginal skill while misrepresenting simultaneous heatwave–drought probability, concurrent wind–precipitation events, or fire–drought–heat cascades. The sample-size constraint (Bevacqua et al. 2023) means progress requires deliberate use of generative

emulators or large climate ensembles in the training regime; the field cannot rely on the observational record alone.

5 Evaluation pathologies in scientific machine learning

Scientific ML applied to weather and climate exhibits a set of recurring evaluation pathologies that the broader machine learning community rarely confronts. Naming and analysing them is a core contribution of this survey, and they are therefore treated in a section of their own. The pathologies recur across nowcasting, medium-range forecasting, S2S, climate projection, and downscaling; locating the analysis in one place avoids repetition and makes the diagnosis available as a reference for the open-challenges section (Sect. 8).

5.1 The pathologies

5.1.1 Training–evaluation circularity

The majority of published AI weather models are trained on ERA5 and evaluated against ERA5 (Ben Bouallègue et al. 2024). This creates a circular evaluation: the models learn the spectral characteristics, noise structure, and biases of ERA5, and are then rewarded for reproducing them. A model that produces ERA5-like output is not necessarily accurate; it may have learned ERA5's artefacts. Verification against independent radiosonde observations partially breaks this circularity, but is rarely reported as the primary evaluation.

5.1.2 RMSE bias towards smooth forecasts

MSE/RMSE training penalises large pointwise errors regardless of physical coherence. A forecast that is spatially smooth but close in magnitude incurs lower RMSE than one that is sharp but slightly displaced, even when the sharp forecast is physically correct and the smooth one is not. This is not a bug unique to AI; NWP ensembles exhibit the same property when verified by ensemble mean. But AI models optimised purely against RMSE learn to over-smooth, systematically underestimating extremes (Olivetti and Messori 2024). A concrete consequence is that the conditional-mean forecast of a bimodal predictive distribution (e.g. along a sharp frontal boundary) falls between the two modes, producing an outcome that occurs with zero probability in nature.

5.1.3 Data leakage via climatological normalisation

Models normalised using statistics computed over the full dataset, including the test period, have access to future climatological information during training. In a stationary climate this is a minor issue; under non-stationary warming trends it can lead to artificially optimistic skill scores.

5.1.4 Metric-dataset confounding

WeatherBench (Rasp et al. 2020) made global forecasting evaluation reproducible. But successive model generations have been optimised against its specific variable set (Z500, T850, T2M, U10), spectral resolution, and ERA5-derived targets. Models that maximise WeatherBench scores may be "overfitting to the benchmark", improving on the specific metric without improving on the underlying physical prediction problem. WeatherBench 2 (Rasp et al. 2024) broadened the evaluation surface, but the general risk of benchmark overfitting in scientific ML is structural, not solved by any single update.

5.1.5 Distribution shift under non-stationarity

Standard train–test splits use temporally adjacent data under the assumption that the data-generating distribution is stationary. Climate is not: anthropogenic warming shifts temperature distributions, alters storm-track statistics, and changes the frequency of compound events. A model trained on 1979–2018 ERA5 and evaluated on 2019–2022 is technically in-distribution; evaluated on 2050 CMIP6 output under SSP5−8.5, it is substantially out-of-distribution. Evaluation protocols that quantify degradation under controlled covariate shift are not yet standard in the field.

5.2 Towards replacement and supplementary metrics

Naming the RMSE smoothness bias is only useful if a replacement or supplement is identifiable. We do not propose that RMSE be retired; it is well-understood, communicates clearly to non-specialists, and is operationally entrenched. We do propose that it be reported *alongside* a small set of metrics that diagnose the failure modes RMSE conceals.

Spatial–displacement-tolerant skill. The Fractions Skill Score and other neighbourhood-based metrics evaluate a forecast against observations within a tolerance neighbourhood rather than pointwise, removing the penalty that RMSE applies to sharp but slightly displaced forecasts. The neighbourhood radius is a tunable parameter, which is a feature (it expresses what spatial accuracy the application demands) rather than a defect.

Probabilistic, distribution-aware metrics. The Continuous Ranked Probability Score (CRPS), Wasserstein distance between forecast and observation distributions, and quantile-based scoring rules each penalise mis-specification of the full predictive distribution rather than only its mean, removing the smoothness incentive.

Spectral and physical diagnostics. Power-spectral comparisons diagnose whether a model reproduces the energy distribution across spatial scales; conservation residuals (mass, momentum, dry-air mass) diagnose physical consistency. Neither replaces RMSE, but neither can be gamed by smoothing.

Multivariate dependence diagnostics. For compound-event relevance, copula-based scores and rank-correlation-preserving metrics check that joint probabilities are correct, not just the marginals (Zscheischler et al. 2020; Bevacqua et al. 2023). Allen et al. (2024) develop calibration diagnostics that explicitly probe multivariate structure.

A practical recommendation is that papers reporting new global forecasting systems should publish, at minimum: pointwise RMSE/ACC against a held-out reanalysis period *and* (i) CRPS for the ensemble version, (ii) a neighbourhood metric (FSS) at one or more relevant thresholds, (iii) tail-conditional skill on a small set of documented extreme events, and (iv) a power-spectrum comparison against the verification field at a representative pressure level. Reporting any one of these in isolation continues the pathology under a different name.

5.3 Critical application to flagship systems

The pathologies above apply with as much force to the celebrated systems of the past three years as to less prominent ones. The survey would be inconsistent with its own standards if it identified the pathologies in general while exempting the flagship systems from the same critical pressure.

For GraphCast (Lam et al. 2023), Pangu-Weather (Bi et al. 2023), FuXi (Chen et al. 2023b), FengWu (Chen et al. 2023a), AIFS (Lang et al. 2024), and Aurora (Bodnar et al. 2025), the reported gains against operational baselines are almost entirely on RMSE or related deterministic scores computed against ERA5 (training–evaluation circularity), are concentrated in large-scale upper-air fields (masking weaker performance on near-surface and convective scales), and have been independently verified on extremes only sparsely (Olivetti and Messori 2024). Ben Bouallègue et al. (2024) explicitly examine these systems against operational ECMWF analyses and conclude that headline RMSE numbers overstate the gap relative to physics-based forecasting once verification is shifted to an independent reference. None of this means the systems are not genuine advances; it means the size of

the advance, especially on the variables and regimes that matter most for impact, is smaller and less uniform than aggregate RMSE comparisons against ERA5 suggest. We retain the systems as landmark results while attaching the same evaluative caveats that the survey applies to the field as a whole.

6 Scientific understanding beyond prediction

Predictive skill and scientific understanding are distinct goals and sometimes in tension. A model can forecast a teleconnection pattern accurately without revealing the causal mechanisms responsible. Conversely, causal discovery can quantify physical relationships that predictive models treat as latent. Both goals are important; this section surveys methods aimed at understanding rather than pure prediction.

6.1 Causal discovery in earth system science

Causal discovery algorithms recover directed causal graphs from observational time series, going beyond correlation-based analysis to identify driver–response relationships. In the climate context, this is essential for process understanding, attribution, and evaluation of climate model fidelity. We stress at the outset, however, that the outputs of these algorithms on climate data are best read as *candidate hypotheses* rather than confirmed mechanisms: the assumptions on which their guarantees rest (sufficiency of the observed variables, stationarity, adequate sampling) are systematically violated in the Earth system, as detailed below. The reliability of algorithmically discovered edges for confirmatory claims is therefore limited, and we are careful throughout not to present recovered graphs as established causal fact.

The PCMCI algorithm (Runge et al. 2019), named for its two-stage procedure combining a PC (Peter–Clark) condition-selection step with the Momentary Conditional Independence (MCI) test, embeds conditional independence testing in a causal discovery framework designed for the characteristics of Earth system data: high dimensionality, autocorrelation, and limited sample sizes relative to the number of variables. Applied to reanalysis, PCMCI recovers causal networks of global teleconnections and enables process-oriented evaluation of climate models by comparing their causal fingerprints with observations (Nowack et al. 2020). Runge et al. (2023) provide a technical review of causal inference methods for Earth system time series, covering linear and nonlinear conditional independence tests, causal effect estimation, and attribution of extremes. Granger causality and its nonlinear generalisations provide a complementary approach particularly suited to ENSO and monsoon teleconnection analysis.

6.1.1 Limitations of conditional-independence testing in climate data

The conditional-independence machinery underlying PCMCI is well-developed for low-dimensional, weakly autocorrelated, identically distributed data; climate data violate all three of these conditions. Four limitations deserve explicit acknowledgement. *(i) Dimensionality and sample size*: gridded climate fields contain millions of variables, while the observational record for many teleconnection-relevant signals contains only decades of independent realisations, so the conditioning set is poorly sampled. *(ii) Autocorrelation*: atmospheric and oceanic variables exhibit long memory at the relevant timescales, which inflates the false-discovery rate of standard independence tests unless explicitly controlled. *(iii) Latent confounding*: unobserved variables (e.g. slowly evolving ocean heat content) can induce apparent direct dependencies between observed variables and produce edges that no algorithm using only the observed variables can correctly orient. *(iv) Non-stationarity*: anthropogenic forcing changes both marginal distributions and dependence structure across the record, so the assumption of a single underlying causal graph is, strictly, false. Runge et al. (2023) survey methods that partially address each, but no current method handles all four simultaneously at the spatial scale of climate model output. The reasonable conclusion is that algorithmic causal discovery is a *tool of triage*, useful for proposing causal hypotheses

and for ruling out claimed links that fail elementary independence tests, but *not* a means of confirming causal claims about the climate system. Confirmatory claims still require domain knowledge, mechanistic constraints, and intervention-like variation (e.g. controlled experiments in coupled models). Reported causal graphs from observational climate data should accordingly be presented with their assumption violations stated, and treated as provisional pending mechanistic or interventional corroboration; the strength of a discovered edge is bounded by the weakest of the four conditions above, not by the nominal significance level of the independence test that produced it.

6.1.2 Tool-augmented LLM agents for causal discovery

A methodologically adjacent direction couples conditional-independence-based causal discovery with tool-augmented large language model agents that integrate semantic priors and multi-modal context (Shen et al. 2025). The motivation is to lift weak observational signals through model-internal knowledge and external retrievals, rather than from data alone. For Earth-system applications, the appeal is that domain knowledge about plausible mechanisms, which is what dimension-reduction methods discard, can be re-injected through the LLM's pre-training corpus. The same scepticism that we apply to algorithmic causal discovery applies with at least equal force here: the LLM's priors are themselves derived from the published literature, so causal claims discovered in this manner are constrained by, and confounded with, what the literature already believed. The approach is most defensible as a hypothesis-generation tool to be confirmed by intervention-style analysis, not as a primary mechanism for causal claims.

6.2 Explainability and physical interpretability

6.2.1 Three distinct interpretability aims

The interpretability literature applied to scientific ML conflates three aims that should be kept separate. *Transparency of model behaviour* asks which inputs the model relies on for a given prediction; saliency maps, integrated gradients, and attention visualisation address this aim. *Semantic decomposability of learned representations* asks whether the latent features of the model align with identifiable physical structures (e.g. whether a hidden unit responds selectively to atmospheric river events). *Correspondence to physical mechanism* asks whether the input–output mapping the model implements reflects the actual causal mechanism of the physical system. These are different standards. A model can be transparent (we know which inputs it uses) without being decomposable (the latent representation is opaque) and without being mechanistically correct (it relies on a confound). The present survey uses "explainability" only for the first aim, "interpretability" for the second, and reserves "physical correspondence" for the third.

Post-hoc XAI methods—saliency maps, SHAP values, attention visualisation, and integrated gradients—identify which atmospheric patterns drive AI predictions (McGovern et al. 2019). These methods explain a model's internal computation; they do not by themselves reveal causal structure in the physical system. The distinction matters: a saliency map identifying a sea-surface temperature pattern as predictive for a heatwave outcome shows model reliance, not physical causality. Both are useful but different. Graph-specific explanation methods can additionally identify prediction-relevant nodes, edges, motifs, or coherent subgraphs (Pfeifer et al. 2022).

6.2.2 Methods closer to causality

For the third aim, post-hoc attribution is structurally insufficient and explanations constructed under causal assumptions are required. Two families are particularly relevant. *Counterfactual explanations* ask: what minimal modification to the input would have changed the prediction? When the modifications are constrained to lie on the manifold of physically realisable atmospheric states, counterfactuals approximate intervention-style explanations

and yield a notion of explanation substantially closer to causal mechanism than saliency. The integration of conformal prediction with counterfactual explanations discussed in Norinder et al. (2026) extends this with a validity guarantee under exchangeability. *Causal abstractions*, in which a high-dimensional learned model is shown to behave equivalently to a low-dimensional causal mechanism for a class of interventions, provide a stronger but harder-to-instantiate alternative.

Physics-informed architectures (Raissi et al. 2019) and equivariant networks (spherical CNNs (Cohen et al. 2018), SFNO (Bonev et al. 2023)) reduce the explanatory burden on post-hoc methods by constraining model structure to be physically consistent from the outset. Where invariances are encoded architecturally, post-hoc analysis can focus on the learned content rather than diagnosing spurious correlations.

6.2.3 A concrete application: detecting and explaining heatwave predictors

Chattopadhyay et al. (2020) applied a CNN classifier to ERA5 fields to identify precursor patterns of North American summer heatwaves at lead times of 1–5 weeks, then used layer-wise relevance propagation (a SHAP-related attribution method) to identify the spatial regions on which the classifier relied. The attribution recovered the well-known mid-tropospheric circulation patterns that synoptic meteorology associates with heatwaves: blocking anticyclones over the affected region and upstream wave-train signatures across the Pacific. The interpretation is illuminating in two directions. The attribution agreement with established meteorology is reassuring; it indicates that the classifier exploits genuine atmospheric signal rather than a spurious confound. It also illustrates the limit of the method: the attribution shows that the model uses these patterns, not that these patterns cause the heatwave. The causal interpretation would require an intervention argument (e.g. counterfactual experiments in a coupled model) that the attribution method itself does not provide. The combination of high-skill prediction, post-hoc attribution consistent with mechanism, and downstream confirmation through causal experiments is the workflow that the present field should normalise.

For operational deployment and regulatory approval of AI-based early warning systems, auditable model behaviour is increasingly an institutional requirement. A unified framework connecting learned atmospheric representations to physically interpretable mechanisms remains an open problem.

7 Deployment, applications, and infrastructure

7.1 Operational deployment and impact

In specific operational contexts, AI systems now run alongside traditional NWP. ECMWF's AIFS (Lang et al. 2024) has been in operational service since February 2025, running four times daily at 0.25° resolution (Sect. 4). AI-based and statistical NWP post-processing—calibration, precipitation downscaling, wind gust correction—is reported in operational or pre-operational use at a number of national meteorological services, although the specific systems and their operational status are not always documented in the open literature. Google Flood Hub issues 7-day inundation forecasts in over 80 countries covering an estimated 460 million people, with skill on five-day lead time predictions in ungauged basins reported to be comparable to or better than the nowcast skill of the Copernicus Emergency Management Service Global Flood Awareness System (Nearing et al. 2024); the Copernicus EMS itself applies AI-based damage mapping within hours of disaster events.

At the sectoral level, graph-network solar and wind forecasting improves grid-balancing skill over statistical baselines; ML flood models have been validated for ungauged catchments globally; CNN-based wildfire risk nowcasting combines fuel, topography, and meteorological predictors; and AI downscaling of CMIP6 output serves agricultural, insurance, and infrastructure risk applications. LLM-based data discovery tools connected to ESGF and CDS archive APIs may reduce barriers to navigating multi-petabyte climate holdings.

7.2 Data infrastructure for scientific ML

Large-scale AI training on climate data requires infrastructure that co-locates petabyte-scale archives with scalable compute. This approach has been advanced within the European Open Science Cloud (EOSC) landscape, where federated data spaces and interoperable services have been developed to support data-intensive climate workflows (Elia et al. 2023; Manzi et al. 2026), thus bridging the gap between large-scale data repositories and HPC/cloud resources required for efficient AI model training (Ejarque et al. 2022).

Key open-science platforms include the Copernicus CDS (ERA5 and Copernicus datasets), ECMWF open data (real-time IFS/AIFS under open licence), NCAR Geoscience Data Exchange (e.g. NCEP GFS, JRA-3Q, CESM), NOAA NODD (GOES, NEXRAD, GFS on AWS/GCP/Azure), and Google Earth Engine (planetary-scale EO data). Moreover, in the Climate Science domain, the Earth System Grid Federation (ESGF) represents a key federated infrastructure providing access to multi-petabyte large-scale climate model outputs, including datasets from the Coupled Model Intercomparison Project (CMIP) (Petrie et al. 2021).

Building on these federated data infrastructures, emerging cloud-native and HPC-integrated approaches are reshaping how climate data are accessed and analysed. Pangeo's Zarr-based data model, chunk-addressable arrays stored in cloud object storage, has emerged as a de facto standard for scalable random-access workflows in climate AI training and analysis. For example, in this context, CMIP datasets are increasingly replicated and made accessible on cloud platforms such as Google Cloud, enabling efficient, on-demand analysis within the same cloud-native ecosystem. In parallel, Destination Earth leverages EuroHPC infrastructures to develop high-resolution digital twins of the Earth system, supported by dedicated AI-ready data pipelines and interoperability layers.

In the context of broader European Commission-funded initiatives targeting the development of interoperable and AI-ready research infrastructures for data-intensive science, the Horizon Europe RI-SCALE project[2] illustrates multi-institutional Data Exploitation Platform (DEP) deployment, demonstrating that federated identity, data provenance, and energy accounting are necessary infrastructure components alongside AI models. Another relevant effort in this area is represented by the Horizon Europe WeatherGenerator project,[3] which aims to develop a European foundation model for weather and climate applications, also drawing on EuroHPC resources.

7.2.1 Order-of-magnitude infrastructure cost

Concrete cost figures are scarce in the peer-reviewed literature and date rapidly, so the values below should be read as orders of magnitude useful for planning rather than precise budget estimates. Storing a multi-petabyte climate archive on a major commercial cloud at standard object-storage tariffs is on the order of low tens of thousands of US dollars per petabyte per year in storage charges alone; access (egress) costs frequently dominate total expenditure when training workloads pull data across regions or out of the cloud, and can equal or exceed storage cost over a training campaign. For the full CMIP6 archive of roughly 28 PB, an unmitigated single-cloud replication therefore implies recurring costs in the hundreds of thousands to low millions of US dollars per year for storage, plus comparable or larger egress charges if researchers do not co-locate their compute with the storage. These figures explain the operational emphasis on *compute-near-data* architectures (Pangeo, ESGF replica nodes, EuroHPC pre-staging) rather than ad-hoc downloads, and they motivate the inclusion of energy and storage accounting as first-class deliverables in projects such as RI-SCALE.

[2] RI-SCALE: Unlocking Research Infrastructure Potential with Scalable AI and Data. https://www.riscale.eu/

[3] WeatherGenerator: Advancing Earth system science with innovative AI. https://weathergenerator.eu/

7.2.2 Open vs. proprietary outputs

Both RI-SCALE and WeatherGenerator commit to open-science principles consistent with European Open Science Cloud policy: trained models, scientific software, and derived datasets produced within the projects are intended for open release under FAIR-compliant licences (project-specific repositories and DOIs are published on the respective sites cited above). Inputs that are themselves proprietary (e.g. certain commercial satellite products, licensed operational analyses) cannot be redistributed by the projects and remain subject to the licences of their original providers. Readers considering reuse should consult the projects' data-management plans for the licence status of any specific output of interest.

8 Open challenges in scientific machine learning

Weather and climate science exposes a set of open problems that are general to scientific ML but are here posed with unusual sharpness. They are organised into three groups: challenges of learning from physical systems, challenges of evaluation and benchmarking, and challenges of deployment.

8.1 Learning on physical systems

8.1.1 Physical consistency of learned operators

Most high-performing AI weather models produce physically inconsistent states in some regimes. Concrete documented examples include data-driven global forecasting systems producing trace *negative* precipitation values in regions of light convection (physically impossible since precipitation flux is non-negative), small but systematic violations of dry-air mass conservation during multi-day rollouts (so the modelled atmosphere "leaks"), and locally super-saturated humidity values that exceed the Clausius–Clapeyron bound. Bonev et al. (2023) explicitly motivate SFNO partly to remove the rotational-symmetry violations that unconstrained Fourier neural operators produce on the sphere, and Ben Bouallègue et al. (2024) report that purely data-driven systems compared against operational analyses retain quantifiable biases in the residuals of mass and energy budgets even when their RMSE on individual variables is competitive.

8.1.1.1 Notions of physical consistency are not interchangeable The literature uses "physical consistency" to refer to several distinct properties; conflating them obscures progress. We distinguish five. *Conservation of integral quantities*: globally integrated mass, dry-air mass, moisture, energy, and angular momentum should be invariant or budget-closed under the modelled dynamics. *Symmetry respect*: invariances and equivariances of the governing equations (rotational equivariance on the sphere, time-translation in the absence of forcing) should be respected by the learned map. *Balance relations across variables*: geostrophic, hydrostatic, and thermal-wind balances that the atmosphere obeys to leading order in the relevant regimes should not be violated more strongly than physical fluctuations around them would suggest. *Thermodynamic admissibility*: outputs should satisfy bounds (non-negativity of mass, temperature above absolute zero, humidity below saturation under the Clausius–Clapeyron relation). *Long-term statistical fidelity*: integrated over the climatological record, modelled distributions should match the target distribution in moments and tail behaviour. A model can be strong on some of these and weak on others; reporting which of these notions a system has been audited against, and which it has not, should be a baseline expectation. Physics-informed architectures (Raissi et al. 2019), equivariant networks (Cohen et al. 2018; Bonev et al. 2023), and hybrid differentiable models (Kochkov et al. 2024) address differ-

ent subsets of these notions, and a unified framework auditing all five across the full output distribution remains absent. This is a scientific ML problem that happens to be well-posed in atmospheric science.

8.1.2 Distribution shift and non-stationarity

Climate is a non-stationary system. AI models trained on historical reanalysis face covariate shift at two timescales: near-term shift driven by anthropogenic warming trends already present in the training period, and long-horizon shift when applied to future climate projections far outside the training distribution. Standard train–test temporal splits do not measure this degradation; formal evaluation under controlled covariate shift is not yet standard. This is the climate instance of a general challenge in scientific ML: how to quantify and mitigate distribution shift when the data-generating process is physically non-stationary.

8.1.3 Sample efficiency in sparse regimes

The tropics, Southern Ocean, and polar regions are observationally undersampled relative to mid-latitude land areas. AI models trained predominantly on ERA5 inherit these coverage biases. Training directly on sparse, irregular observations requires methods tolerant of missing data and heterogeneous sampling—a geometric learning problem that the atmospheric setting poses in a practically important form.

8.2 Evaluation and benchmark design

8.2.1 Benchmark overfitting: the problem

WeatherBench (Rasp et al. 2020, 2024) established reproducible evaluation for medium-range forecasting, but successive model generations have been optimised against its specific variable set, resolution, and ERA5-derived targets. The general risk of benchmark overfitting in scientific ML, where models improve on a specific metric without improving on the underlying physical problem, is structural rather than incidental (Thiyagalingam et al. 2022).

A key mechanism underlying this phenomenon is the implicit alignment between training objectives and benchmark-specific scoring rules. When loss functions (e.g., mean squared error) closely mirror evaluation metrics, optimisation converges towards statistical fidelity with respect to the benchmark distribution, not necessarily towards physically meaningful representations. This induces a form of shortcut learning, where models exploit dominant spatiotemporal correlations present in reanalysis data, while underrepresenting rare regimes, nonlinear interactions, and regime transitions that are critical for scientific validity and decision-making (Geirhos et al. 2020).

Moreover, the reliance on ERA5-derived targets introduces a form of evaluation endogeneity. As a data assimilation product, ERA5 integrates numerical weather prediction models with observational data, thereby embedding model-based priors, smoothing effects, and assimilation artefacts. Models trained and evaluated exclusively on ERA5 may therefore learn to emulate the statistical structure of the assimilation system itself rather than the underlying atmospheric processes. This raises concerns about the external validity of reported performance gains, particularly when transferring models to operational forecasting or climate-scale applications (Hersbach et al. 2020; Bi et al. 2023).

Metric gaming further exacerbates this issue by incentivising improvements along narrow scalar indicators. Aggregate scores such as RMSE or anomaly correlation coefficient are dominated by large-scale, high-variance structures, systematically underweighting local extremes and compound events. As a consequence, models can achieve state-of-the-art benchmark performance while failing to capture high-impact phenomena such as convective precipitation, heatwaves, or coupled hazards. This disconnect is particularly problematic in risk-sensitive

domains, where decision-relevant performance is concentrated in the tails of the distribution rather than in the mean behaviour (Pasche et al. 2025).

8.2.2 Benchmark overfitting: proposed solutions and a benchmark-suite blueprint

A principled response requires extending evaluation protocols towards distribution-aware and process-aware diagnostics. This includes regime-conditioned evaluation, tail-sensitive metrics derived from extreme value theory, and multivariate calibration measures that assess the joint behaviour of atmospheric variables (Allen et al. 2024). In addition, cross-benchmark validation and temporal extrapolation tests are necessary to probe robustness under distributional shift, especially in the presence of climate non-stationarity.

Addressing benchmark overfitting is inseparable from advances in reproducibility and transparency (Semmelrock et al. 2025). Open access to trained models, training data pipelines, and evaluation code is essential for disentangling genuine methodological progress from artefacts of implementation (Bochenek and Ustrnul 2022). In this context, reproducibility should be understood not only as the ability to replicate numerical results, but as a broader epistemic requirement enabling falsifiability in the sense of Popper (1963), comparative analysis, and cumulative scientific progress (Gundersen 2021).

8.2.2.1 Blueprint for a future benchmark suite A survey of this scope is well-positioned to specify, rather than only criticise, what an adequate benchmark suite for scientific-ML weather and climate models should contain. We propose the following six ingredients as a minimum blueprint; none is novel in isolation, but they are rarely combined in published evaluations.

(i) Independent observational verification. At least one verification surface independent of the training reanalysis. For upper-air variables, this means radiosonde and aircraft observations; for surface variables, surface station networks and direct satellite-radiance hindcasts; for precipitation, gauge and radar data where calibration permits.

(ii) Regime-conditioned skill. Skill scores reported separately for at least the following regimes: tropics vs. mid-latitudes vs. polar, ocean vs. land, winter vs. summer hemispheres, and ENSO phases (positive, negative, neutral) over a multi-decadal evaluation window.

(iii) Extremes and compound events. A documented panel of historical events (e.g. Pacific Northwest heatwave June 2021, Hurricane Ian September 2022, European drought summer 2018, spring 2020 floods in East Africa) with case-study skill metrics, plus threshold-conditional and quantile-based skill at appropriate percentiles (95th, 99th, 99.9th), and multivariate dependence diagnostics for at least one canonical compound regime (e.g. heatwave–drought concurrence).

(iv) Physical-consistency diagnostics. Mass, energy, and dry-air mass budget closure measured over the rollout horizon; spectral fidelity at representative pressure levels; non-negativity and admissibility audits across all output variables (Sect. 5).

(v) Controlled distribution-shift tests. Held-out evaluation periods chosen so they fall outside the training period and reflect climate trends (e.g. post-2018 evaluation for a model trained on 1979–2018), cross-system evaluation against an independent reanalysis, and where feasible evaluation on CMIP6 emission-scenario output for stress-testing.

(vi) Reproducibility envelope. Open model weights, training code, evaluation code, and a manifest of data versions and random seeds, hosted under a long-term archival DOI; a documented recomputation protocol that allows an independent reviewer to reproduce the headline numbers within stated tolerance.

8.2.2.2 A worked instantiation To make the blueprint operational rather than aspirational, we sketch one concrete instantiation that could be assembled today from existing public components; the point is not that these specific choices are canonical, but that each ingredient maps to a named, available resource and a definite protocol. *Step 1 (substrate).* Adopt WeatherBench 2 (Rasp et al. 2024) as the scoring substrate and grid, since it already fixes a common 0.25° evaluation grid, deterministic and probabilistic scoring code, and ERA5 as the analysis reference. *Step 2 (ingredient i).* Add an independent verification layer by scoring against IGRA radiosonde pro-

files and against surface networks (e.g. HadISD/ISD stations) and, for precipitation, against gauge-radar products (e.g. MRMS over CONUS, IMERG globally), so that headline skill is never reported solely against the training reanalysis. *Step 3 (ingredient ii).* Stratify every reported score by the regime partitions named above (tropics/mid-latitude/polar, land/ocean, season, ENSO phase from the Niño−3.4 index), and publish the stratified table, not just the global mean. *Step 4 (ingredient iii).* Fix the named historical-event panel as a versioned case-study set with, per event, an evaluation window, a target field, and threshold-conditional scores at the 95th/99th/99.9th percentiles plus at least one multivariate (e.g. heat–drought) dependence diagnostic. *Step 5 (ingredient iv).* Compute mass, moisture, and energy budget residuals and power spectra at fixed pressure levels over the full rollout, with non-negativity and admissibility audits run automatically as part of scoring (Sect. 5). *Step 6 (ingredient v).* Run two held-out stress tests—temporal extrapolation (train $\leq$2018, test $\geq$2019) and cross-reanalysis evaluation against an independent analysis (e.g. JRA-3Q or MERRA-2)—and, where a model ingests boundary forcings, an optional CMIP6 SSP5−8.5 out-of-sample probe. *Step 7 (ingredient vi).* Release weights, training and scoring code, a data-version manifest, and a one-command recomputation script under an archival DOI, with a stated numerical tolerance for the headline metrics. Each step is individually routine; the contribution of a benchmark suite is to require all seven together and to report the stratified, physically-audited results rather than a single leaderboard scalar.

A benchmark satisfying all six is a higher bar than current practice. We expect a transitional period in which papers report (i), (ii), and (iii) routinely, with (iv)–(vi) becoming standard as community infrastructure catches up. WeatherBench 2 (Rasp et al. 2024) moves in this direction on (i) and (iii); ClimateBench (Watson-Parris et al. 2022) covers some of (v) for emulation. A successor effort that explicitly targets all six axes simultaneously would substantially raise the evidentiary standard of the field.

8.2.3 Calibration and uncertainty quantification

Calibrated uncertainty quantification at regional scales is an open scientific ML problem. Diffusion-based ensemble generation (Price et al. 2025) and conformal prediction (Angelopoulos and Bates 2023) are scalable directions, but conformal guarantees degrade under non-stationarity, and diffusion ensembles require careful calibration beyond standard coverage metrics. The compound-event context adds a multivariate calibration requirement that marginal calibration metrics cannot diagnose.

8.2.3.1 Why conformal guarantees degrade under non-stationarity The marginal coverage guarantee of standard split conformal prediction rests on the assumption that the calibration data and test data are *exchangeable*. In plain terms, exchangeability means that the calibration examples and the future test example are statistically interchangeable—one could shuffle the order of all of them, or swap a calibration point for the test point, without changing the joint probability of the data. Intuitively, it is the requirement that the future the model will be used on looks like a random draw from the same population as the data used to calibrate it (it is a slightly weaker condition than "independent and identically distributed", but for the present purpose the i.i.d. picture is an adequate mental model). Under climate non-stationarity this assumption fails by construction: the data-generating distribution of (say) summer surface temperature in 2050 differs from that of 1990, so a 1990 record and a 2050 record are *not* interchangeable draws from one fixed population, and the two periods are therefore not exchangeable. The calibrated quantile estimated from a 1979–2018 calibration set is an unbiased estimate of the 1979–2018 quantile, not of the 2025–2030 quantile that operational use requires. Adaptive conformal procedures, time-series-aware conformal variants, and distribution-shift-robust extensions partly mitigate this, but each introduces its own assumptions (bounded shift, known drift class, or recalibration access) that themselves must be validated against the climate-trend signal. We therefore treat conformal prediction as useful for short-horizon operational quantification and a candidate for adaptation in the climate-projection setting, rather than as a deployable guarantee at decadal lead times.

In compound-event regimes, uncertainty must be calibrated jointly across variables, space, and time, rendering marginal coverage metrics insufficient. This motivates a shift towards structured uncertainty representations that capture dependencies and conditional risk. A promising research direction is the integration of conformal prediction with counterfactual explanations, where conformal sets define uncertainty-aware feasibility regions for minimal perturbations. Such a framework could enable counterfactuals that are not only plausible and interpretable

but also statistically valid under distributional uncertainty, thereby linking predictive reliability with actionable explanation in non-stationary, high-dimensional systems (Norinder et al. 2026). Another direction is the use of transfer learning to reduce climate model projection uncertainties by constraining simulations with real-world observations (Immorlano et al. 2025).

8.2.4 Interpretability and explainability as scientific discovery

We retain here the terminological convention fixed in Sect. 6: *explainability* denotes transparency of model behaviour (which inputs a model relies on), *interpretability* denotes semantic decomposability of learned representations (whether latent features align with identifiable physical structures), and *physical correspondence* denotes agreement with the actual causal mechanism. The three open directions below are organised by which of these aims they target, and the terms are used in that fixed sense rather than interchangeably.

Post-hoc XAI provides model-level transparency (the explainability aim) (McGovern et al. 2019; Retzlaff et al. 2024); causal discovery provides system-level physical understanding (bearing on physical correspondence) (Runge et al. 2019, 2023; Emmert-Streib et al. 2026). The open challenge is whether learned representations can be connected to physically interpretable mechanisms without post-hoc analysis—whether, in other words, architectural choices (equivariance, operator learning) can make interpretability, in the semantic-decomposability sense, intrinsic rather than added. This is a general question about inductive bias in scientific ML, not a domain-specific question about meteorology.

A forward-looking research direction is the development of intrinsically interpretable architectures in scientific machine learning. Approaches based on equivariance, symmetry-preserving networks, and operator learning already move in this direction by embedding known physical invariances and functional relationships directly into the hypothesis space. Extending these ideas, future work should investigate modular and concept-aligned representations, where latent components correspond to identifiable physical processes (e.g. advection, convection, or teleconnection patterns) rather than abstract feature embeddings. Complementary to concept-aligned neural architectures, explainable-by-design classification frameworks based on ordinal sums of conjunctive, disjunctive and averaging aggregation functions can combine expert-defined conditions with data-driven flexibility while retaining globally retraceable decision logic (Hudec et al. 2021). Such glass-box approaches may be particularly relevant to bounded weather- and climate-decision tasks, including categorical warning generation, observational quality control and risk-level classification, although their scalability to continuous, high-dimensional Earth-system forecasting remains to be established.

This connects to recent efforts in mechanistic interpretability and concept bottleneck models, but requires adaptation to continuous, high-dimensional spatiotemporal systems. Recent work on structurally regularised interpretable machine learning further supports this direction, demonstrating that post-hoc smoothing and regularisation strategies can improve the interpretability and stability of ensemble-based models while preserving predictive performance, thereby highlighting the broader importance of embedding interpretability into model structure rather than relying exclusively on external explanation methods (Pfeifer et al. 2025).

A second research avenue concerns the integration of causal structure into representation learning. Rather than applying causal discovery post hoc, models could be trained with causal regularisation objectives that enforce invariance under interventions or distribution shifts, thereby encouraging representations that reflect stable physical relationships. This includes combining neural operators with causal constraints, learning structural causal models in function space, and exploiting multi-environment or multi-model climate data (e.g. CMIP ensembles) as natural interventions. Such approaches would enable a tighter coupling between predictive performance and scientific insight, addressing the long-standing tension between accuracy and interpretability.

A third direction lies in bridging scales of explanation. Current XAI methods typically operate at the level of input features or saliency maps, while causal discovery operates at the level of variables and processes. Scientific applications require intermediate representations that link local model decisions to global system dynamics. This calls for hierarchical explanation frameworks that connect micro-scale features (e.g. grid-cell anomalies)

to meso-scale patterns (e.g. circulation regimes) and ultimately to macro-scale causal mechanisms. Developing such multi-scale explanatory frameworks remains largely unexplored.

Finally, evaluation protocols for interpretability itself require substantial advancement. Existing benchmarks focus on predictive skill, while interpretability is often assessed qualitatively or through proxy metrics. We suggest at minimum two operational metrics that the field could adopt immediately. The first is a *faithfulness* score: the correlation between an attribution method's importance ranking of inputs and the actual change in model output when those inputs are perturbed in a controlled manner; high-faithfulness explanations recover the model's true dependencies, low-faithfulness ones do not. The second is *stability under perturbation*: the variation in the produced explanation when the input is perturbed within a physically plausible neighbourhood (e.g. small atmospheric noise consistent with observational uncertainty); stable explanations are reproducible, unstable ones are not. Adopting these two as reportable numbers, alongside predictive skill, would convert interpretability from a qualitative attribute to an empirical property subject to benchmarking, enabling systematic comparison of intrinsically interpretable models.

8.3 Deployment and sustainability

8.3.1 Energy and compute efficiency

Training large-scale AI weather models consumes substantial energy relative to the forecast skill gained. It is worth noting, however, that although they are computationally expensive at training time, ML-based weather prediction models can be orders of magnitude faster than numerical models at deployment. Model distillation, sparse architectures, and dedicated AI accelerators offer improvements, but the field lacks a standard efficiency metric. Energy per unit of forecast skill improvement, as opposed to absolute energy per training run, would provide a meaningful comparison baseline.

8.3.1.1 Order-of-magnitude estimates To illustrate scale, training a modern foundation-class atmospheric model with on the order of 10^8 parameters consumes tens to hundreds of thousands of accelerator-hours: GraphCast was trained for roughly four weeks on 32 Cloud TPU v4 devices (of order 2×10^4 device-hours) (Lam et al. 2023), and Pangu-Weather reports training four models over about 60 days on 192 V100 GPUs (of order 3×10^5 GPU-hours) (Bi et al. 2023). At a full-system power draw of 0.3–0.7 kW per accelerator—which, beyond the chip itself, includes host CPUs, memory, interconnect, and data-centre cooling overhead—this corresponds to roughly tens to a few hundred MWh per flagship training run, depending on configuration, accelerator generation, and parallel scale. (For comparison, a single accelerator running for a few thousand chip-hours in isolation would consume only $\mathcal{O}(1)$ MWh; the larger figures arise because many accelerators run in parallel for weeks and because full-system power exceeds bare chip power.) Translating energy to carbon depends entirely on the local grid mix: at 0.05 kg CO_2e per kWh (low-carbon grid) a 100 MWh training run produces $\sim$5 tonnes CO_2e; at 0.4 kg CO_2e per kWh (typical high-carbon grid) the same run produces $\sim$40 tonnes. These figures are deliberately given as ranges rather than precise numbers because exact published values differ across reporting conventions and accelerator generations; the point is that the carbon cost of a single training run of a flagship atmospheric model is non-trivial and depends as much on where the training happens (grid carbon intensity) as on the model itself. Inference cost, by contrast, is approximately three to four orders of magnitude lower per forecast cycle.

Recent work has begun to quantify the environmental externalities associated with large-scale AI systems, highlighting that training and deploying modern models can incur substantial energy consumption and carbon emissions depending on model size, hardware configuration, and energy source (Padovani et al. 2024). In particular, systematic assessments demonstrate that the carbon footprint of large language models scales nonlinearly with model complexity and training duration, and that these costs are often ignored in the literature. Such findings reinforce the need to move beyond raw performance metrics towards sustainability-aware evaluation frameworks that explicitly account for energy consumption and emissions (Padovani et al. 2025).

Incorporating carbon accounting into model development and benchmarking, alongside metrics such as forecast skill, would enable more transparent trade-off analyses and support the design of energy-efficient architectures, training strategies, and deployment pipelines (Jeanquartier et al. 2026).

8.3.2 Human–AI interaction and accountability

Human-in-the-loop design requires co-development between human factors researchers and AI developers from the outset, not as post-deployment integration.

As AI systems transition from experimental tools to primary forecast products, the associated institutional and societal requirements, including accountability, auditability, transparency, and public communication of uncertainty, become first-order design constraints rather than downstream considerations. This shift necessitates a research agenda that integrates technical model development with governance, human factors, and decision science.

In particular, future work should formalise auditability-by-design, where model architectures, training pipelines, and data provenance are structured to enable traceable and reproducible decision pathways. This includes the development of verifiable logging mechanisms, standardised model cards for scientific ML systems, and interfaces that expose uncertainty in forms interpretable by domain experts and the public. Closely related is the need for operational frameworks for human–AI teaming, where human expertise is not merely used for validation but actively shapes model behaviour through interactive learning, adaptive feedback loops, and context-aware overrides. Such systems require rigorous evaluation protocols that go beyond predictive skill to include trust calibration, cognitive load, and decision impact under uncertainty.

A forward-looking research direction lies in embedding regulatory and ethical constraints directly into AI system design, aligning with emerging frameworks such as the EU AI Act (Holzinger et al. 2025b). The applicability of the Act to weather forecasting is mediated by Annex III, point 2, which classifies as high-risk those AI systems intended to be used as safety components in the management or operation of critical infrastructure (including the supply of water, gas, heating, and electricity). AI-driven weather forecasts are typically not themselves a safety component, but they routinely feed systems that are, for example, AI-driven flood early warnings used in operational disaster response, grid-balancing forecasts in electricity supply, and water-resource management. To the extent that an AI forecasting system is integrated into such a safety component, the high-risk regime of the Act applies, and two of its requirements are particularly demanding for weather forecasting practice. Article 14 (human oversight) requires that high-risk AI systems be designed so that natural persons can effectively oversee their operation, and Article 15 (accuracy, robustness, and cybersecurity) requires that the system perform consistently throughout its lifecycle, including under distributional shift. The distribution-shift question (Sect. 5) is therefore not merely an academic concern: under Article 15 it becomes a compliance-relevant property that operators of safety-coupled forecasting systems will need to demonstrate. This implies the development of compliance-aware architectures that integrate risk assessment, bias monitoring, and explainability as intrinsic components rather than external audits. Furthermore, the communication of uncertainty remains an open scientific challenge: probabilistic forecasts must be translated into actionable insights without oversimplification, requiring interdisciplinary advances at the interface of AI, cognitive science, and risk communication. Ultimately, the co-evolution of AI systems and institutional practices will define their societal impact, making human-in-the-loop design not an optional enhancement but a foundational principle of trustworthy scientific machine learning.

9 Conclusions

This survey has examined weather and climate science as a testbed for AI, focusing on what the domain reveals about AI methods rather than the converse. Two takeaways stand out.

First, geometric deep learning and operator learning are paradigm-level contributions of this field to scientific ML: matching a model's symmetry group and function space to the structure of the physical domain reduces what the model needs to learn from data and improves out-of-distribution generalisation in a way that data augmentation alone cannot replicate (Cohen et al. 2018; Bonev et al. 2023; Li et al. 2021; Lu et al. 2021).

Second, the evaluation pathologies named and analysed in Sect. 5—training–evaluation circularity, RMSE smoothness bias, benchmark overfitting, distribution shift under non-stationarity, and inadequate verification on extremes and compound events—are the field's most generalisable methodological contribution: every scientific ML field that trains on assimilated products or evaluates on aggregate metrics will encounter them, and the diagnostic vocabulary developed here transfers directly to fluid dynamics, seismology, ecology, and other physically constrained domains.

The remaining themes of the survey—foundation models and their bias propagation, hybrid differentiable systems, causal discovery and intrinsic interpretability—support and elaborate these two but do not displace them.

9.1 Forward-looking recommendations

Looking to the next two to three years, we identify six concrete priorities the field should adopt as standard practice.

R1. Observational benchmarks independent of ERA5

The community should curate and adopt at least one verification surface that does not derive its labels from the same reanalysis the models train on, beginning with radiosonde and aircraft observations for upper-air variables and direct satellite radiances for column-integrated quantities. "Independent observational verification" has been discussed for some years; the practical step now is to publish curated, versioned verification datasets with DOIs and to require their use in flagship comparison studies.

R2. Reportable metric panel

New global forecasting systems should report, as a minimum: RMSE/ACC against a held-out reanalysis period, CRPS for the ensemble version, a neighbourhood metric (e.g. FSS) at threshold-relevant scales, tail-conditional skill on a small set of documented extremes, and a power-spectrum comparison against the verification field (Sect. 5.2). The incremental cost is modest; the diagnostic value is high.

R3. Compute-near-data as default infrastructure

The cost arithmetic of multi-petabyte archives (Sect. 7.2) makes ad-hoc cross-region data movement an unsustainable training pattern. Funded projects should default to compute-near-data architectures (Pangeo-style cloud-native pipelines, ESGF replica nodes, EuroHPC pre-staging) and report storage and egress costs alongside compute.

R4. Distribution-shift evaluation as a first-class metric

Held-out evaluation periods chosen to fall outside the training period, cross-system evaluation against an independent reanalysis, and where feasible evaluation on CMIP6 scenario output should be reported alongside in-distribution skill, not appended as an optional robustness check.

R5. Auditable physical-consistency reporting

Authors should report which of the five notions of physical consistency (Sect. 8) their system has been audited against and which it has not. This includes mass and energy budget closure, non-negativity audits, and spectral fidelity at representative pressure levels.

R6. Compliance-aware development for safety-coupled deployments

For systems intended to feed flood early warning, grid balancing, or water resource management, the EU AI Act Annex III high-risk regime applies; the corresponding compliance requirements (human oversight under Article 14, robustness under distribution shift under Article 15) should be addressed during model development rather than retrofitted after deployment.

The open challenges identified in Sect. 8—physical consistency of learned operators, distribution shift under non-stationarity, benchmark overfitting, calibrated UQ, and intrinsic interpretability—are not the meteorological

forecasting problems of a decade ago. They are the scientific ML problems of the next decade, posed here in a domain where solving them has direct and measurable consequence.

Whether the field meets them well will depend less on further architectural novelty than on the disciplinary maturity of its evaluation practices, which is precisely where the recommendations above seek to apply pressure.

Acknowledgements We sincerely thank the editors and the anonymous reviewers for their exceptionally thorough comments and constructive feedback, which substantially contributed to strengthening and improving the quality of this manuscript.

Author contributions A.H.: Conceptualization, Investigation, Funding Acquisition, Writing - Original Draft, Writing - Review & Editing; S.F.: Investigation, Validation, Writing - Original Draft, Writing - Review & Editing; T.D.: Investigation, Validation, Writing - Original Draft, Writing - Review & Editing; F.A.: Investigation, Validation, Writing - Original Draft, Writing - Review & Editing; H.M.: Investigation, Funding Acquistion, Visualization, Writing - Original Draft, Writing - Review & Editing; All authors reviewed the final manuscript. All authors approved the final version.

Funding This work has been partly funded by the Austrian Science Fund (FWF), Project: 10.55776/P-32554, explainable AI. This work has been partly supported by the RI-SCALE project, which has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 10188168. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

Data availibility No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that there are no conflict of interest.

Ethical approval The authors declare that this work does not raise any ethical issues.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Accarino G, Chiarelli M, Fiore S, Federico I, Causio S, Coppini G, Aloisio G (2021) A multi-model architecture based on Long Short-Term Memory neural networks for multi-step sea level forecasting. Futur Gener Comput Syst 124:1–9. https://doi.org/10.1016/j.future.2021.05.008

Allen S, Ziegel J, Ginsbourger D (2024) Assessing the calibration of multivariate probabilistic forecasts. Q J R Meteorol Soc 150(760):1315–1335. https://doi.org/10.1002/qj.4647

Angelopoulos AN, Bates S (2023) Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning 16(4):494–591. https://doi.org/10.1561/2200000101

Baño-Medina J, Manzanas R, Gutiérrez JM (2020) Configuration and intercomparison of deep learning neural models for statistical downscaling. Geoscientific Model Development 13(4):2109–2124. https://doi.org/10.5194/gmd-13-2109-2020

Barnes EA, Hurrell JW, Ebert-Uphoff I, Anderson C, Anderson D (2019) Viewing forced climate patterns through an AI lens. Geophys Res Lett 46(22):13389–13398. https://doi.org/10.1029/2019GL084944

Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, Molinari N, Smidt TE, Kozinsky B (2022) E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nat Commun 13(1):2453. https://doi.org/10.1038/s41467-022-29939-5

Ben Bouallègue Z, Clare MCA, Magnusson L, Gascón E, Maier-Gerber M, Janoušek M, Rodwell M, Pinault F, Dramsch JS, Lang STK, 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. Bull Am Meteor Soc 105(6):E864–E883. https://doi.org/10.1175/BAMS-D-23-0162.1
Beucler T, Pritchard M, Rasp S, Ott J, Baldi P, Gentine P (2021) Enforcing analytic constraints in neural networks emulating physical systems. Phys Rev Lett 126(9):098302. https://doi.org/10.1103/PhysRevLett.126.098302
Bevacqua E, Suarez-Gutierrez L, Jézéquel A, Lehner F, Vrac M, Yiou P, Zscheischler J (2023) Advancing research on compound weather and climate events via large ensemble model simulations. Nat Commun 14(1):2145. https://doi.org/10.1038/s41467-023-37847-5
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
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
Bodnar C, Bruinsma WP, Lucic A, Stanley M, Allen A, Brandstetter J, Garvan P, Riechert M, Weyn JA, Dong H, Gupta JK, Thambiratnam K, Archibald AT, Wu CC, Heider E, Welling M, Turner RE, Perdikaris P (2025) A foundation model for the Earth system. Nature 641(8065):1180–1187. https://doi.org/10.1038/s41586-025-09005-y
Bonavita M, Laloyaux P (2020) Machine learning for model error inference and correction. Journal of Advances in Modeling Earth Systems 12(12):e2020MS002232. https://doi.org/10.1029/2020MS002232
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. In Proceedings of the 40th International Conference on Machine Learning (ICML), Volume 202 of Proceedings of Machine Learning Research, pp. 2806–2823. PMLR
Brajard J, Carrassi A, Bocquet M, Bertino L (2020) Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. Journal of Computational Science 44:101171. https://doi.org/10.1016/j.jocs.2020.101171
Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P (2017) Geometric deep learning: Going beyond Euclidean data. IEEE Signal Process Mag 34(4):18–42. https://doi.org/10.1109/MSP.2017.2693418
Brunner L, Pendergrass AG, Lehner F, Merrifield AL, Lorenz R, Knutti R (2020) Reduced global warming from CMIP6 projections when weighting models by performance and independence. Earth System Dynamics 11(4):995–1012. https://doi.org/10.5194/esd-11-995-2020
Cachay R, S, Erickson E, Bucker AFC, Pokropek E, Potosnak W, Bire S, Osei S, Lütjens B (2021) The world as a graph: Improving El Niño forecasts with graph neural networks.arXiv:2104.05089
Camps-Valls G, Tuia D, Zhu XX, Reichstein M (2021) Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences. Wiley, Hoboken, NJ
Camps-Valls G, Fernández-Torres MÁ, Cohrs KH, Höhl A, Castelletti A, Pacal A, Robin C, Martinuzzi F, Papoutsis I, Prapas I, Pérez-Aracil J, Weigel K, Gonzalez-Calabuig M, Reichstein M, Rabel M, Giuliani M, Mahecha MD, Popescu OI, Pellicer-Valero OJ, Ouala S, Salcedo-Sanz S, Sippel S, Kondylatos S, Happé T, Williams T (2025) Artificial intelligence for modeling and understanding extreme weather and climate events. Nat Commun 16(1):1919. https://doi.org/10.1038/s41467-025-56573-8
Chattopadhyay A, Hassanzadeh P, Pasha S (2020) Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. Sci Rep 10(1):1317. https://doi.org/10.1038/s41598-020-57897-9
Chen K, Han T, Gong J, Bai L, Ling F, Luo JJ, 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:2304.02948
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(1): 190. https://doi.org/10.1038/s41612-023-00512-1
Cinquini L, Crichton D, Mattmann C, Harney J, Shipman G, Wang F, Ananthakrishnan R, Miller N, Denvil S, Morgan M, Pobre Z, Bell GM, Doutriaux C, Drach R, Williams D, Kershaw P, Pascoe S, Gonzalez E, Fiore S, Schweitzer R (2014) The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. Futur Gener Comput Syst 36:400–417. https://doi.org/10.1016/j.future.2013.07.002
Cohen TS, Geiger M, Köhler J, Welling M (2018) Spherical CNNs. In International Conference on Learning Representations (ICLR
Cuomo S, Di Cola VS, Giampaolo F, Rozza G, Raissi M, Piccialli F (2022) Scientific machine learning through physics-informed neural networks: Where we are and what's next. J Sci Comput 92(3):88. https://doi.org/10.1007/s10915-022-01939-z
De Luca DL, Napolitano F, Kim D, Onof C, Biondi D, Wang LP, Russo F, Ridolfi E, Moccia B, Marconi F (2025) Rainfall nowcasting models: State of the art and possible future perspectives. Hydrol Sci J 70(9):1419–1438. https://doi.org/10.1080/02626667.2025.2490780
Dueben PD, 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
Durack PJ, Taylor KE, Gleckler PJ, Meehl GA, Lawrence BN, Covey C, Stouffer RJ, Levavasseur G, Ben-Nasser A, Denvil S, Stockhause M, Gregory JM, Juckes M, Ames SK, Antonio F, Bader DC, Dunne JP, Ellis D, Eyring V, Fiore SL, Joussaume S, Kershaw P, Lamarque JF, Lautenschlager M, Lee J, Mauzey CF, Mizielinski M, Nassisi P, Nuzzo A, O'Rourke

E, Painter J, Potter GL, Rodriguez S, Williams DN (2025) The Coupled Model Intercomparison Project (CMIP): Reviewing project history, evolution, infrastructure and implementation. EGUsphere preprint egusphere-2024-3729, in open discussion

Ejarque J, Badia RM, Albertin L, Aloisio G, Baglione E, Becerra Y, Boschert S, Berlin JR, D'Anca A, Elia D, Exertier F, Fiore S, Flich J, Folch A, Gibbons SJ, Koldunov N, Lordan F, Lorito S, Løvholt F, Macías J, Marozzo F, Michelini A, Monterrubio-Velasco M, Pienkowska M, de la Puente J, Queralt A, Quintana-Ortí ES, Rodríguez JE, Romano F, Rossi R, Rybicki J, Kupczyk M, Selva J, Talia D, Tonini R, Trunfio P, Volpe M (2022) Enabling dynamic and intelligent workflows for HPC, data analytics, and AI convergence. Futur Gener Comput Syst 134:414–429. https://doi.org/10.1016/j.future.2022.04.014

Elia D, Antonio F, Fiore S, Nassisi P, Aloisio G (2023) A data space for climate science in the European Open Science Cloud. Computing in Science & Engineering 25(1):7–15. https://doi.org/10.1109/MCSE.2023.3274047

Emmert-Streib F, Tripathi S, Farea A, Holzinger A (2026) Implementing physics-informed neural networks with deep learning for differential equations. Frontiers in Artificial Intelligence 9:1717117. https://doi.org/10.3389/frai.2026.1717117

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. Nat Commun 13(1):5145. https://doi.org/10.1038/s41467-022-32483-x

Eyring V, Bony S, Meehl GA, Senior CA, Stevens B, Stouffer RJ, Taylor KE (2016) Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development 9(5):1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Farchi A, Laloyaux P, Bonavita M, Bocquet M (2021) Using machine learning to correct model error in data assimilation and forecast applications. Q J R Meteorol Soc 147(739):3067–3084. https://doi.org/10.1002/qj.4116

Fiore S, Elia D, Palazzo C, D'Anca A, Antonio F, Williams DN, Foster I, Aloisio G (2018) Towards an open (data) science analytics-hub for reproducible multi-model climate analysis at scale. In 2018 IEEE International Conference on Big Data (Big Data), pp. 3226–3234. IEEE

Gagne DJ II, Haupt SE, Nychka DW, Thompson G (2019) Interpretable deep learning for spatial analysis of severe hailstorms. Mon Weather Rev 147(8):2827–2845. https://doi.org/10.1175/MWR-D-18-0316.1

Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (ICML), Volume 48 of Proceedings of Machine Learning Research, pp. 1050–1059. PMLR

Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA (2020) Shortcut learning in deep neural networks. Nature Machine Intelligence 2(11):665–673. https://doi.org/10.1038/s42256-020-00257-z

Gentine P, Pritchard M, Rasp S, Reinaudi G, Yacalis G (2018) Could machine learning break the convection parameterization deadlock? Geophys Res Lett 45(11):5742–5751. https://doi.org/10.1029/2018GL078202

Giffard-Roisin S, Yang M, Charpiat G, Kumler Bonfanti C, Kégl B, Monteleoni C (2020) Tropical cyclone track forecasting using fused deep learning from aligned reanalysis data. Frontiers in Big Data 3:1. https://doi.org/10.3389/fdata.2020.00001

Gundersen OE (2021) The fundamental principles of reproducibility. Phil Trans R Soc A 379(2197):20200210. https://doi.org/10.1098/rsta.2020.0210

Ham YG, Kim JH, Luo JJ (2019) Deep learning for multi-year ENSO forecasts. Nature 573(7775):568–572. https://doi.org/10.1038/s41586-019-1559-7

Han L, Chen M, Chen K, Chen H, Zhang Y, Lu B, Song L, Qin R (2021) A deep learning method for bias correction of ECMWF 24–240 h forecasts. Adv Atmos Sci 38(9):1444–1459. https://doi.org/10.1007/s00376-021-0215-y

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 et al (2020) The ERA5 global reanalysis. Q J R Meteorol Soc 146(730):1999–2049. https://doi.org/10.1002/qj.3803

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holzinger A, Lukač N, Rozajac D, Johnston E, Kocic V, Hörl B, Gollob C, Nothdurft A, Stampfer K, Schweng S, Del Ser J (2025) Enhancing trust in automated 3D point cloud data interpretation through explainable counterfactuals. Information Fusion 119:103032. https://doi.org/10.1016/j.inffus.2025.103032

Holzinger A, Zatloukal K, Müller H (2025) Is human oversight to AI systems still possible? New Biotechnol 85:59–62. https://doi.org/10.1016/j.nbt.2024.12.003

Holzinger A, Longo L, Cangelosi A, Del Ser J (2026) Research frontiers in machine learning and knowledge extraction. Machine Learning and Knowledge Extraction 8(1):6. https://doi.org/10.3390/make8010006

Hsu CY, Li W, Wang S (2025) Geospatial foundation models for image analysis: Evaluating and enhancing NASA-IBM Prithvi's domain adaptability. Int J Geogr Inf Sci 39(9):2096–2125. https://doi.org/10.1080/13658816.2024.2397441

Hudec M, Minarikova E, Mesiar R, Saranti A, Holzinger A (2021) Classification by ordinal sums of conjunctive and disjunctive functions for explainable ai and interpretable machine learning solutions. Knowledge Based Systems 220:106916. https://doi.org/10.1016/j.knosys.2021.106916

Immorlano F, Eyring V, le Monnier T, de Gouville G, Accarino D, Elia S, Mandt GA, Gentine P (2025) Transferring climate change physical knowledge. Proc Natl Acad Sci 122(15):e2413503122. https://doi.org/10.1073/pnas.2413503122
Jeanquartier F, Jean-Quartier C, Rieder P, Misirlić V, Pasero C, Hohensinner R, Müller H, Holzinger A (2026) Assessing the carbon footprint of language models: Towards sustainability in AI. Resour Conserv Recycl 226:108670. https://doi.org/10.1016/j.resconrec.2025.108670
Jones CG, Adloff F, Booth BBB, Cox PM, Eyring V, Friedlingstein P, Frieler K, Hewitt HT, Jeffery HA, Joussaume S, Koenigk T, Lawrence BN, O'Rourke E, Roberts MJ, Sanderson BM, Séférian R, Somot S, Vidale PL, van Vuuren D et al (2024) Bringing it all together: Science priorities for improved understanding of Earth system change and to support international climate policy. Earth System Dynamics 15(5):1319–1351. https://doi.org/10.5194/esd-15-1319-2024
Kadow C, Hall DM, Ulbrich U (2020) Artificial intelligence reconstructs missing climate information. Nat Geosci 13(6):408–413. https://doi.org/10.1038/s41561-020-0582-5
Keisler R (2022) Forecasting global weather with graph neural networks.arXiv:2202.07575
Kim H, Ham YG, Joo YS, Son SW (2021) Deep learning for bias correction of MJO prediction. Nat Commun 12(1):3087. https://doi.org/10.1038/s41467-021-23406-3
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 MP, 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
Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M (2018) Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol Earth Syst Sci 22(11):6005–6022. https://doi.org/10.5194/hess-22-6005-2018
Kvammen, A., K. Wickstrøm, D. McKay, and N. Partamies. 2020. Auroral image classification with deep neural networks. Journal of Geophysical Research: Space Physics 125(10): e2020JA027808. https://doi.org/10.1029/2020JA027808
Lam R, Sanchez-Gonzalez A, Willson M, Wirnsberger 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
Lang, S., M. Alexe, M. Chantry, J. Dramsch, F. Pinault, B. Raoult, M.C.A. Clare, C. Lessig, M. Maier-Gerber, L. Magnusson, Z. Ben Bouallègue, P.D. Dueben, F. Pappenberger, F. Engel, F. Rabier, and F. Vitart. 2024. AIFS—ECMWF's data-driven forecasting system. arXiv:2406.01465
LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791
Leinonen J, Nerini D, Berne A (2021) Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. IEEE Trans Geosci Remote Sens 59(9):7211–7223. https://doi.org/10.1109/TGRS.2020.3032790
Li, Z., N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar 2021. Fourier neural operator for parametric partial differential equations. In International Conference on Learning Representations (ICLR)
Liu Y, Racah E, Prabhat J, Correa A, Khosrowshahi D, Lavers K, Kunkel M, Wehner W, Collins (2016) Application of deep convolutional neural networks for detecting extreme weather in climate datasets.arXiv:1605.01156
Liu R, Zhang X, Wang W, Wang Y, Liu H, Ma M, Tang G (2024) Global-scale ERA5 product precipitation and temperature evaluation. Ecol Ind 166:112481. https://doi.org/10.1016/j.ecolind.2024.112481
Lu L, Jin P, Pang G, Zhang Z, Karniadakis GE (2021) Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. Nature Machine Intelligence 3(3):218–229. https://doi.org/10.1038/s42256-021-00302-5
Manzi A, Bardaji R, Rodero I, Moltó G, Fiore S, Campos I, Elia D, Sarandrea F, Millar AP, Spiga D, Bunino M, Accarino G, Asprea L, Bernardo S, Caballer M, Chatzikyriakou C, Ciangottini D, Claus M, Cristofori A et al (2026) interTwin: Advancing scientific digital twins through AI, federated computing and data. Futur Gener Comput Syst 179:108312. https://doi.org/10.1016/j.future.2025.108312
McGovern A, Lagerquist R, Gagne DJ II, Jergensen GE, Elmore KL, Homeyer CR, Smith T (2019) Making the black box more transparent: Understanding the physical implications of machine learning. Bull Am Meteor Soc 100(11):2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1
Nearing G, Cohen D, Dube V, Gauch M, Gilon O, Harrigan S, Hassidim A, Klotz D, Kratzert F, Metzger A, Nevo S, Pappenberger F, Prudhomme C, Shalev G, Shenzis S, Tekalign TY, Weitzner D, Matias Y (2024) Global prediction of extreme floods in ungauged watersheds. Nature 627(8004):559–563. https://doi.org/10.1038/s41586-024-07145-1
Nguyen, T., J. Brandstetter, A. Kapoor, J.K. Gupta, and A. Grover 2023. ClimaX: A foundation model for weather and climate. In Proceedings of the 40th International Conference on Machine Learning (ICML), Volume 202 of Proceedings of Machine Learning Research, pp. 25904–25938. PMLR
Norinder U, Lowry S, Müller H, Holzinger A (2026) Conformal prediction for counterfactual detection in concept learning from synthetic visual patterns. Electronics 15(7):1346. https://doi.org/10.3390/electronics15071346
Nowack P, Runge J, Eyring V, Haigh JD (2020) Causal networks for climate model evaluation and constrained projections. Nat Commun 11(1):1415. https://doi.org/10.1038/s41467-020-15195-y

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

Padovani G, Anantharaj V, Sacco L, Kurihana T, Bunino M, Tsolaki K, Girone M, Antonio F, Sopranzetti C, Fronza M, Fiore S (2024) A software ecosystem for multi-level provenance management in large-scale scientific workflows for AI applications. SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, pp 2024–2031

Padovani G, Anantharaj V, Fiore S (2025) Provenance tracking in large-scale machine learning systems. Workshop Proceedings of the 54th International Conference on Parallel Processing (ICPP Workshops '25). Association for Computing Machinery, New York, NY, pp 167–174

Pan, B., G.J. Anderson, A. Goncalves, D.D. Lucas, C.J.W. Bonfils, J. Lee, Y. Tian, and H.Y. Ma. 2021. Learning to correct climate projection biases. Journal of Advances in Modeling Earth Systems 13(10): e2021MS002509. https://doi.org/10.1029/2021MS002509

Pasche OC, 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

Pathak J, Subramanian S, Harrington P, Raja S, Chattopadhyay A, Mardani M, Kurth T, Hall D, Li Z, Azizzadenesheli K, Hassanzadeh P, Kashinath K, Anandkumar A 2022. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators arXiv:2202.11214

Petrie R, Denvil S, Ames S, Levavasseur G, Fiore S, Allen C, Antonio F, Berger K, Bretonnière PA, Cinquini L, Dart E, Dwarakanath P, Druken K, Evans B, Franchistéguy L, Gardoll S, Gerbier E, Greenslade M, Hassell D et al (2021) Coordinating an operational data distribution network for CMIP6 data. Geoscientific Model Development 14(1):629–644. https://doi.org/10.5194/gmd-14-629-2021

Pfeifer, B., A. Saranti, and A. Holzinger. 2022. Gnn-subnet: disease subnetwork detection with explainable graph neural networks. Oxford Bioinformatics 38(S-2): ii120–ii126. https://doi.org/10.1093/bioinformatics/btac478

Pfeifer B, Gevaert A, Loecher M, Holzinger A (2025) Tree smoothing: Post-hoc regularization of tree ensembles for interpretable machine learning. Inf Sci 690:121564. https://doi.org/10.1016/j.ins.2024.121564

Popper KR (1963) Science: Conjectures and refutations, Conjectures and Refutations: The Growth of Scientific Knowledge, 33–65. Routledge and Kegan Paul, London

Price I, Sanchez-Gonzalez A, Alet F, Andersson TR, El-Kadi A, Masters D, Ewalds T, Stott J, Mohamed S, Battaglia P, Lam R, Willson M (2025) Probabilistic weather forecasting with machine learning. Nature 637(8044):84–90. https://doi.org/10.1038/s41586-024-08252-9

Raissi M, Perdikaris P, Karniadakis GE (2019) Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J Comput Phys 378:686–707. https://doi.org/10.1016/j.jcp.2018.10.045

Rampal N, Hobeichi S, Gibson PB, Baño-Medina J, Abramowitz G, Beucler T, González-Abad J, Chapman W, Harder P, Gutiérrez JM (2024) Enhancing regional climate downscaling through advances in machine learning. Artificial Intelligence for the Earth Systems 3(2):e230066. https://doi.org/10.1175/AIES-D-23-0066.1

Rasp S, Pritchard MS, Gentine P (2018) Deep learning to represent subgrid processes in climate models. Proc Natl Acad Sci 115(39):9684–9689. https://doi.org/10.1073/pnas.1810286115

Rasp, S., P.D. Dueben, S. Scher, J.A. Weyn, S. Mouatadid, and N. Thuerey. 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

Rasp, S., S. Hoyer, A. Merose, I. Langmore, P. Battaglia, T. Russell, A. Sanchez-Gonzalez, V. Yang, R. Carver, S. Agrawal, M. Chantry, Z. Ben Bouallègue, P. Dueben, C. Bromberg, J. Sisk, L. Barrington, A. Bell, and F. Sha. 2024. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. Journal of Advances in Modeling Earth Systems 16(6): e2023MS004019. https://doi.org/10.1029/2023MS004019

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

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

Retzlaff CO, Angerschmid A, Saranti A, Schneeberger D, Röttger R, Müller H, Holzinger A (2024) Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. Cogn Syst Res 86:101243. https://doi.org/10.1016/j.cogsys.2024.101243

Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, Ross AS, Milojevic-Dupont N, Jaques N, Waldman-Brown A, Luccioni AS, Maharaj T, Sherwin ED, Mukkavilli SK, Kording KP, Gomes CP, Ng AY, Hassabis D, Platt JC, Creutzig F, Chayes J, Bengio Y (2023) Tackling climate change with machine learning. ACM Comput Surv 55(2):1–96. https://doi.org/10.1145/3485128

Runge, J., P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic. 2019. Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances 5(11): eaau4996. https://doi.org/10.1126/sciadv.aau4996
Runge J, Gerhardus A, Varando G, Eyring V, Camps-Valls G (2023) Causal inference for time series. Nature Reviews Earth & Environment 4(7):487–505. https://doi.org/10.1038/s43017-023-00431-y
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 (2024) Analysis, characterization, prediction, and attribution of extreme atmospheric events with machine learning and deep learning techniques: A review. Theoret Appl Climatol 155(1):1–44. https://doi.org/10.1007/s00704-023-04571-5
Salman AG, Kanigoro B, Heryadi Y (2015) Weather forecasting using deep learning techniques. 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, pp 281–285
Saranti A, Pfeifer B, Gollob C, Stampfer K, Holzinger A (2024) From 3D point-cloud data to explainable geometric deep learning: State-of-the-art and future challenges. WIREs Data Min Knowl Discovery 14(6):e1554. https://doi.org/10.1002/widm.1554
Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G (2009) The graph neural network model. IEEE Trans Neural Networks 20(1):61–80. https://doi.org/10.1109/TNN.2008.2005605
Scher S, Messori G (2018) Predicting weather forecast uncertainty with machine learning. Q J R Meteorol Soc 144(717):2830–2841. https://doi.org/10.1002/qj.3410
Schultz MG, Betancourt C, Gong B, Kleinert F, Langguth M, Leufen LH, Mozaffari A, Stadtler S (2021) Can deep learning beat numerical weather prediction? Phil Trans R Soc A 379(2194):20200097. https://doi.org/10.1098/rsta.2020.0097
Semmelrock H, Ross-Hellauer T, Kopeinik S, Theiler D, Haberl A, Thalmann S, Kowald D (2025) Reproducibility in machine-learning-based research: Overview, barriers, and drivers. AI Mag 46(2):e70002. https://doi.org/10.1002/aaai.70002
Shen, C., Z. Chen, D. Luo, D. Xu, H. Chen, and J. Ni 2025. Exploring multi-modal data with tool-augmented LLM agents for precise causal discovery. In Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, pp. 636–660. Association for Computational Linguistics
Shi X, Chen Z, Wang H, Yeung DY, Wong WK, Woo WC (2015) Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in Neural Information Processing Systems 28 (NIPS 2015). Curran Associates, pp 802–810
Shi, X., Z. Gao, L. Lausen, H. Wang, D.Y. Yeung, W.K. Wong, and W.c. Woo 2017. Deep learning for precipitation nowcasting: A benchmark and a new model. In Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 5622–5632. Curran Associates
Sønderby CK, Espeholt L, Heek J, Dehghani M, Oliver A, Salimans T, Agrawal S, Hickey J, Kalchbrenner N (2020) MetNet: A neural weather model for precipitation forecasting.arXiv:2003.12140
Thiyagalingam J, Shankar M, Fox G, Hey T (2022) Scientific machine learning benchmarks. Nature Reviews Physics 4(6):413–420. https://doi.org/10.1038/s42254-022-00441-7
Trok, J.T., E.A. Barnes, F.V. Davenport, and N.S. Diffenbaugh. 2024. Machine learning-based extreme event attribution. Science Advances 10(34): eadl3242. https://doi.org/10.1126/sciadv.adl3242
Ullrich, P.A., E.A. Barnes, W.D. Collins, K. Dagon, S. Duan, J. Elms, J. Lee, L.R. Leung, D. Lu, M.J. Molina, T.A. O'Brien, and F.O. Rebassoo. 2025. Recommendations for comprehensive and independent evaluation of machine learning-based Earth system models. Journal of Geophysical Research: Machine Learning and Computation 2(1): e2024JH000496. https://doi.org/10.1029/2024JH000496
Vandal, T., E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A.R. Ganguly 2017. DeepSD: Generating high resolution climate change projections through single image super-resolution. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, pp. 1663–1672. Association for Computing Machinery
Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin 2017. Attention is all you need. In Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 6000–6010. Curran Associates
Verma Y, Heinonen M, Garg V (2024) ClimODE: Climate and weather forecasting with physics-informed neural ODEs. International Conference on Learning Representations (ICLR
Von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, Kirsch B, Pfrommer J, Pick A, Ramamurthy R (2021) Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans Knowl Data Eng 35(1):614–633. https://doi.org/10.1109/TKDE.2021.3079836
Warren, M.S., S.P. Brumby, S.W. Skillman, T. Kelton, B. Wohlberg, M. Mathis, R. Chartrand, R. Keisler, and M. Johnson 2015. Seeing the Earth in the cloud: Processing one petabyte of satellite imagery in one day. In 2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1–12. IEEE
Watson-Parris, D., Y. Rao, D. Olivié, Ø. Seland, P. Nowack, G. Camps-Valls, P. Stier, S. Bouabid, M. Dewey, E. Fons, J. Gonzalez, P. Harder, K. Jeggle, J. Lenhardt, P. Manshausen, M. Novitasari, L. Ricard, and C. Roesch. 2022. ClimateBench v1.0: A benchmark for data-driven climate projections. Journal of Advances in Modeling Earth Systems 14(10): e2021MS002954. https://doi.org/10.1029/2021MS002954

Weyn JA, Durran DR, 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

Weyn, J.A., D.R. Durran, and R. Caruana. 2020. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. Journal of Advances in Modeling Earth Systems 12(9): e2020MS002109. https://doi.org/10.1029/2020MS002109

Willard J, Jia X, Xu S, Steinbach M, Kumar V (2023) Integrating scientific knowledge with machine learning for engineering and environmental systems. ACM Comput Surv 55(4):1–37. https://doi.org/10.1145/3514228

Wimmers A, Velden C, Cossuth JH (2019) Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. Mon Weather Rev 147(6):2261–2282. https://doi.org/10.1175/MWR-D-18-0391.1

Yuval, J., I. Langmore, D. Kochkov, and S. Hoyer. 2026. Neural general circulation models for modeling precipitation. Science Advances 12(2): eadv6891. https://doi.org/10.1126/sciadv.adv6891

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

Zscheischler J, Martius O, Westra S, Bevacqua E, Raymond C, Horton RM, van den Hurk B, AghaKouchak A, Jézéquel A, Mahecha MD, Maraun D, Ramos AM, Ridder NN, Thiery W, Vignotto E (2020) A typology of compound weather and climate events. Nature Reviews Earth & Environment 1(7):333–347. https://doi.org/10.1038/s43017-020-0060-z

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Andreas Holzinger[1,2,7] · Sandro Fiore[3] · Tullio Degiacomi[4] · Fabrizio Antonio[5] · Heimo Müller[6,7]

✉ Andreas Holzinger
andreas.holzinger@boku.ac.at

1 Human-Centered AI Lab, Institute of Forest Engineering, Department of Ecosystem Management, Climate and Biodiversity, BOKU University, Vienna, Austria

2 Institute of Human-Centered Computing, Faculty of Computer Science and Biomedical Engineering, Graz University of Technology, Graz, Austria

3 Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

4 Hypermeteo SRL, Padua, Italy

5 CMCC Foundation, Euro-Mediterranean Center on Climate Change, Lecce, Italy

6 AI Research, Human–Machine Mind Corporation, Graz, Austria

7 Machine Learning and Information Science Group, Medical University Graz, Graz, Austria