

Weather Flow Prediction Using Spatial-Temporal Graph Neural Networks: A Literature Review

Praise Jaravani

jrvpra001@myuct.ac.za

University of Cape Town

Computer Science Department

Cape Town, Western Cape, South Africa

Abstract

Spatial-Temporal Graph Neural Networks (ST-GNNs) have emerged as powerful tools for weather forecasting, with recent models outperforming traditional numerical weather prediction on standard benchmarks. However, these models systematically underestimate extreme weather events—heatwaves, cold snaps, typhoons—which constitute the greatest societal risks yet live at distribution tails where mean squared error (MSE) loss structurally smooths predictions. This literature review surveys 22 papers spanning ST-GNN architectures, extreme weather prediction, and chaos theory, with specific focus on the underexplored South African context. We identify the extreme weather gap as a training objective problem rooted in MSE’s double penalty effect: models learn to avoid predicting peaks because correctly forecasted extremes that are slightly mistimed incur disproportionate penalties. We propose a chaos-primed framework synthesizing three architectural properties essential for extreme prediction: (i) continuous-time dynamics, aligning with the differential equation nature of chaotic atmospheric flows; (ii) adaptive or dynamic graph structure, capturing state-dependent spatial connectivity; and (iii) distribution-sensitive training objectives, preserving fine-scale variance rather than smoothing tails. Comparative analysis reveals no existing architecture scores well across all three criteria. MTGODE achieves continuous-time and dynamic graphs but uses MAE loss (2/3). The most sophisticated regional model (Li et al., 2026) scores approximately 2.5/3 yet still underestimates heatwave and typhoon peaks, empirically proving partial alignment insufficient. The chaos-primed framework provides the missing analytical lens for evaluating whether ST-GNN architectures match the mathematical structure of the chaotic systems generating extremes, directly informing experimental design for South African weather forecasting where irregular station networks, complex topography, and growing extreme event exposure demand architectures where spatial structure, temporal continuity, and distribution sensitivity all matter simultaneously.

CCS Concepts

• **Computing methodologies** → **Neural networks; Machine learning**; Artificial intelligence; • **Applied computing** → *Physical sciences and engineering*; • **Mathematics of computing** → *Time series analysis*.

Keywords

Spatial-Temporal Graph Neural Networks, Weather Forecasting, Extreme Weather, Chaos Theory, South Africa, Deep Learning, Climate Prediction

ACM Reference Format:

Praise Jaravani. 2026. Weather Flow Prediction Using Spatial-Temporal Graph Neural Networks: A Literature Review. In *Proceedings of Research Methods in Computer Science (CSC4019Z’26)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Weather forecasting is fundamentally a spatial-temporal problem. Atmospheric conditions at any observation point are shaped by neighboring stations across time, creating a network where nodes (weather stations) exchange information through edges (atmospheric influence). This structure is irregular and dynamic: station spacing varies, dependencies shift with weather patterns, and influence propagates non-uniformly across geography. Spatial-Temporal Graph Neural Networks (ST-GNNs) are purpose-built for exactly this data structure, combining graph convolution (to model irregular spatial dependencies) with temporal modeling (to capture evolution over time) [1, 11].

Recent years have seen ST-GNNs achieve remarkable success in weather forecasting. GraphCast [12], a graph neural network trained on 40 years of reanalysis data, now outperforms ECMWF HRES—the leading operational numerical weather model—on 90% of 1380 forecast targets. Aurora [3], a 1.3-billion-parameter foundation model, demonstrates that deep learning approaches can match or exceed traditional physics-based models on standard benchmarks. These results represent a paradigm shift: weather forecasting is transitioning from solving partial differential equations on supercomputers to learning spatiotemporal patterns from historical data.

However, this success masks a critical limitation. Current ST-GNN implementations optimize for average performance across all weather conditions. Extreme weather events—heatwaves, cold snaps, typhoons, intense precipitation—live at the tails of the distribution and are systematically smoothed out during training. This is not incidental but structural: mean squared error (MSE) loss, the dominant training objective, penalizes all prediction errors symmetrically, causing models to minimize average error by predicting

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CSC4019Z’26, Cape Town, South Africa

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/26/03

<https://doi.org/XXXXXXX.XXXXXXX>

near the mean [15]. The result is that even state-of-the-art models underestimate extreme peaks. Li et al. [13], developing the most recent purpose-built regional ST-GNN with variable resolution and dynamic graph structure, explicitly report that their model underestimates heatwave temperatures and typhoon intensity beyond three-hour lead times. Hakim and Masanam [8], testing GraphCast, Pangu-Weather, and FourCastNet against real high-impact extreme events, find that all three underperform the numerical baseline despite superior benchmark performance. The extreme weather gap is real, well-documented, and unresolved.

This review argues that the gap is architectural. Extreme weather events are not statistical outliers—they are the signature of nonlinear amplification in chaotic dynamical systems. The atmosphere is a canonical chaotic system: deterministic yet locally unstable, where small perturbations grow exponentially and error growth peaks precisely when and where extreme events occur [2]. If extreme events emerge from the mathematical structure of chaos, then predicting them requires architectures aligned with that structure. Yet no existing work systematically evaluates ST-GNNs through this lens. Architectures are compared on benchmark datasets using aggregate metrics (RMSE, MAE), but not on whether their mathematical properties match the chaotic systems they model.

The South African context makes this work both novel and urgent. South Africa exhibits high climate variability, sparse and irregular weather station networks, and growing exposure to extreme events driven by climate change [5]. Only two prior studies [5, 6] have applied ST-GNNs to South African weather data, both establishing that spatial graph information adds predictive signal but neither targeting extremes or incorporating chaos-theoretic constraints. The irregular station topology—stations separated by hundreds of kilometers with spatial dependencies shaped by complex topography, coastal effects, and elevation gradients—makes South Africa an ideal testbed for evaluating whether ST-GNN architectures can adapt to chaotic, spatially heterogeneous dynamics.

This literature review proposes a **chaos-primed framework** as an analytical lens for evaluating ST-GNN architectures. We identify three properties that characterize structural alignment with chaotic dynamical systems: (i) continuous-time dynamics (chaotic systems evolve as continuous flows, not discrete steps), (ii) adaptive or dynamic graph structure (spatial dependencies shift with atmospheric state), and (iii) distribution-sensitive training objectives (extremes require loss functions that target the tails of the distribution, not the mean). These criteria are synthesized from recent advances in continuous-time neural ODEs [10, 22], dynamic graph structure learning [9], and work on training objectives for extreme events [19, 20]. Scoring existing architectures against these three criteria is the novel analytical contribution of this review.

The remainder of this paper is organized as follows. Section 2 surveys the foundations of graph neural networks and the emergence of ST-GNNs. Section 3 examines ST-GNN applications to weather forecasting, including recent advances and the South African baseline studies. Section 4 establishes the extreme weather gap as a training objective problem rooted in MSE loss and documents failed attempts to resolve it through architectural improvements alone. Section 5 introduces chaos theory as an analytical lens and identifies the three chaos-primed criteria. Section 6 presents a comparative analysis of reviewed architectures through this framework and

synthesizes what the literature reveals about the gap. Section 7 concludes with open questions and directions for future work.

2 Background: From GNNs to ST-GNNs

2.1 Graph Neural Networks

Standard deep learning architectures assume Euclidean data—images as regular 2D grids, text as sequential tokens. Weather station networks are irregular graphs: nodes (stations) separated by varying distances, edges (atmospheric influence) determined by complex factors including topography, elevation, and prevailing wind patterns. Graph Neural Networks (GNNs) solve this through message passing, where each node aggregates information from its neighbors to update its representation [1].

Two foundational approaches emerged. Graph Convolutional Networks (GCNs) take a spectral approach, transforming the graph into the frequency domain via the graph Laplacian to perform convolution—elegant mathematically but tied to a fixed graph structure. Graph Attention Networks (GATs) use a spatial approach where nodes aggregate directly from neighbors with learned attention weights determining which neighbors matter most. Spatial methods proved more scalable and now dominate the field, but both approaches share a critical limitation: GNNs handle spatial structure with no native temporal modeling. Weather forecasting requires both.

2.2 The Emergence of ST-GNNs

Spatial-Temporal Graph Neural Networks extend GNNs by adding temporal modeling layers. Two families emerged based on how they handle time. RNN-based architectures (e.g., DCRNN, Diffusion Convolutional RNN) embed graph convolution inside recurrent units, capturing temporal dependencies through hidden states that propagate across timesteps. CNN-based architectures (e.g., STGCN [24]) pair graph convolution with temporal convolutions, treating time as a local window rather than a sequential process. CNN-based approaches proved faster and parallelizable, becoming the dominant paradigm for large-scale forecasting.

GraphWaveNet represents a key architectural advance: it learns a self-adaptive adjacency matrix rather than relying on a predefined graph topology. This is critical for weather forecasting, where the true spatial dependency structure between stations is unknown and may not align with geographic distance. The model discovers hidden spatial relationships directly from data, allowing it to capture phenomena like distant stations that influence each other through atmospheric teleconnections or orographic effects.

MTGODE [10] represents a distinct architectural category. Rather than discretizing time into steps, it treats the system as a continuous dynamical system and solves coupled neural ODEs—one for spatial propagation, one for temporal aggregation. This continuous-time formulation eliminates the discontinuous latent state trajectories inherent to discrete architectures and allows deeper spatial propagation without over-smoothing. MTGODE's relevance to chaotic systems is direct: chaotic dynamics evolve as continuous flows described by differential equations, not discrete jumps. We return to this connection in Section 5.

2.3 The ST-GNN Family

Survey literature [1, 11] taxonomizes ST-GNNs along three axes. The *spatial approach* axis distinguishes spectral methods (operating in the graph frequency domain) from spatial methods (direct neighborhood aggregation). The *temporal approach* axis spans recurrent models (GRU, LSTM), convolutional models (temporal CNNs with dilation), attention-based models (Transformers), and continuous-time models (neural ODEs). The *graph type* axis categorizes graphs as static (fixed adjacency throughout training and inference), adaptive (learned during training but fixed at inference), or dynamic (recomputed at every timestep based on current features).

Notable variants include AGCRN (Adaptive Graph Convolutional Recurrent Network), which learns node-specific parameters via a node-adaptive parameter learning module; CLCRN (Conditional Local Convolution RNN), which uses distance-initialized graphs and conditional convolutions designed for meteorological forecasting; and ASTGCN (Attention-based ST-GCN), which applies attention mechanisms across both spatial and temporal dimensions. Each makes different tradeoffs between expressiveness, interpretability, and computational cost. Gaibie et al. [6], applying both AGCRN and CLCRN to South African weather data, find that AGCRN achieves better overall performance while CLCRN learns more plausible spatial dependencies—demonstrating that performance and interpretability can diverge.

The field expands rapidly. New architectures appear at major conferences (ICLR, NeurIPS, ICML, KDD) every few months, often targeting specific domains (traffic, energy, epidemics) with custom inductive biases. This makes comprehensive taxonomy a moving target and underscores the need for principled evaluation frameworks—precisely what the chaos-primed framework in Section 5 provides.

3 ST-GNNs Applied to Weather Forecasting

3.1 Problem Formulation

The spatial-temporal weather forecasting task can be formalized as follows. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ denote a weather station network where $\mathcal{V} = \{v_1, \dots, v_N\}$ is the set of N stations, $\mathcal{E}$ is the set of edges representing spatial dependencies, and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix encoding edge weights. For each station v_i , historical observations are represented as $\mathbf{X}^{(i)} = [\mathbf{x}_{t-T+1}^{(i)}, \dots, \mathbf{x}_t^{(i)}] \in \mathbb{R}^{T \times D}$, where T is the historical window length and D is the number of meteorological variables (e.g., temperature, pressure, humidity, wind speed). The complete input is $\mathbf{X} = [\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}] \in \mathbb{R}^{N \times T \times D}$.

Standard forecasting task. Given historical observations $\mathbf{X}$, predict future states $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_{t+1}, \dots, \hat{\mathbf{y}}_{t+\tau}] \in \mathbb{R}^{N \times \tau \times D}$ for lead times $1, \dots, \tau$. An ST-GNN learns a mapping $f_\theta : (\mathbf{X}, \mathcal{G}) \rightarrow \hat{\mathbf{Y}}$ parameterized by θ , optimizing a loss function $\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}})$ over the true future states $\mathbf{Y}$.

Extreme event prediction task. Extreme events are defined as states exceeding a threshold (e.g., 95th percentile of historical distribution) or exhibiting rapid intensification. Let $\mathcal{T}_{\text{extreme}} = \{t \mid \|\mathbf{y}_t^{(i)}\| > \theta_{\text{extreme}}\}$ denote the set of extreme event timesteps. The extreme forecasting task requires accurate prediction not just on

average but specifically on $\mathcal{T}_{\text{extreme}}$, which may constitute only 5–10% of timesteps. This requires loss functions $\mathcal{L}_{\text{extreme}}$ that explicitly target distribution tails rather than minimizing average error.

Evaluation metrics. Standard ST-GNN metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE). For extreme event evaluation, metrics stratify by event intensity: MAE and RMSE computed separately on extreme vs non-extreme timesteps, Structural Similarity Index (SSIM) for spatial pattern fidelity, and quantile skill scores for distribution tail accuracy. Chaos-specific metrics include Local Lyapunov Exponents (LLEs) [2], which quantify local predictability and can be estimated via supervised ML—high LLEs indicate low predictability precisely when and where extreme events occur. Spectral analysis via power spectral density (PSD) and coherence evaluates whether models preserve fine-scale variance [20], critical for capturing extreme event dynamics.

3.2 Early Applications and Baselines

Wilson et al. [23] provided the first demonstration of ST-GNNs for weather forecasting, applying a Low Rank Weighted GCN to the WeatherBench global dataset. The model combined weighted graph convolution (where edge weights decayed with distance) with temporal convolution, establishing that graph-based spatial modeling could capture atmospheric dependencies on global grids. This work proved viability but did not address irregular station networks or extreme events.

Davidson and Moodley [5] applied ST-GNNs to South African weather data for the first time, using 21 stations across the Western and Northern Cape provinces with eight years of hourly observations (2012–2019) covering six meteorological variables. They compared GraphWaveNet (GWN) and Low Rank Weighted GCN (WGN) against temporal-only baselines (LSTM, TCN). GWN achieved 8.30% SMAPE, outperforming WGN (9.02%), TCN (10.62%), and LSTM (16.25%). The key finding: GWN’s self-adaptive learned adjacency matrix discovered hidden spatial dependencies that distance-based initialization missed. Walk-forward validation with 27 splits demonstrated robustness across time. However, both models used MSE loss with no extreme event evaluation—the gap this review addresses.

3.3 ST-GNNs for South African Weather

Gaibie et al. [6] expanded the South African work to 45 stations across three provinces (Western Cape, Eastern Cape, Northern Cape) with 13 years of data (2010–2022) and four variables (temperature, pressure, humidity, wind speed). They compared AGCRN (Adaptive Graph Convolutional Recurrent Network) and CLCRN (Conditional Local Convolution RNN) against GWN and TCN. AGCRN achieved the best overall RMSE through node-adaptive parameter learning, which optimizes separate parameters for each station. CLCRN, designed specifically for meteorological forecasting with conditional local convolutions, learned more physically plausible spatial dependencies: learned paths aligned with coastal propagation along the Great Escarpment, and the model captured extreme temperature peaks more accurately than AGCRN despite slightly higher average error.

This divergence between performance and interpretability is critical. AGCRN optimizes for average accuracy via node-specific

tuning, smoothing extreme peaks in the process. CLCRN preserves local extreme structure through spatially-aware convolutions initialized with distance information. The finding that CLCRN better captures extremes while AGCRN has better average RMSE demonstrates that architectural choices affect extreme event prediction even when both models use identical MSE loss and training data. Both studies confirm spatial graph information adds predictive signal for South African weather, but both use standard MSE loss with no extreme-specific training or evaluation.

3.4 Recent Advances

GraphCast [12] represents a landmark result: a graph neural network trained on 40 years of ERA5 reanalysis data outperforms ECMWF HRES—the leading operational numerical weather model—on 90% of 1380 forecast targets at lead times up to 10 days. The architecture uses an icosahedral mesh for global coverage, encoding atmospheric state on 37 pressure levels with 6-hour autoregressive rollout. GraphCast's success demonstrated that purely data-driven deep learning could match decades of physics-based modeling. However, Hakim and Masanam [8], testing GraphCast against real high-impact extreme events (hurricanes, heatwaves, cold snaps), find that it underperforms ECMWF HRES despite superior benchmark performance—benchmark skill does not guarantee extreme event skill.

CloudNine-v2 [9] introduces fully dynamic graph structure learning via differentiable edge sampling (Gumbel-Softmax) and adaptive node degree estimation. Rather than learning a single global adjacency matrix, the model reconstructs spatial dependencies at every timestep based on current atmospheric state and distance regularization. CloudNine-v2 achieves 15% RMSE reduction over fixed-topology ST-GNNs (DCRNN, STGCN, GWN, AGCRN). Critically, the performance gap between high-variability and low-variability regions shrinks by a factor of 3-5 compared to fixed-graph models—high atmospheric variability occurs precisely where extreme weather events concentrate (coastal zones, mountainous regions). This suggests dynamic graphs capture time-varying spatial dependency structure during extreme conditions, though the study evaluates only average RMSE without separate extreme event metrics.

Li et al. [13] develop the most recent purpose-built regional ST-GNN with variable resolution (14km in the Pearl River Delta where extremes concentrate, 28km elsewhere) and dynamic graph structure learning. Trained on ERA5 reanalysis (2000-2022) for very short-range forecasting (up to 6 hours) in Guangdong, China, the model combines static geographic topology with learned time-varying adjacency matrices and uses Huber loss with spatial mean constraints and Laplacian regularization. This is the smoking gun paper for the extreme weather gap. In case studies on an extreme heatwave (Guangzhou, August 2024) and Super Typhoon Saola (September 2023), the model systematically underestimates peak temperatures and typhoon intensity beyond three-hour lead times. The authors explicitly acknowledge: "its ability to forecast the peaks (or troughs) remains limited." Despite incorporating dynamic graphs, variable resolution, and physics-aware loss components, extreme prediction fails—something more fundamental than architectural complexity is missing.

Graph-EFM [16], a probabilistic ensemble GNN, is explicitly motivated by the need to capture the full weather distribution including extremes. Rather than producing deterministic forecasts, it generates ensemble predictions with uncertainty estimates via hierarchical graph message passing. Aurora [3], a 1.3-billion-parameter foundation model, demonstrates the field's trajectory toward massive pre-trained models but is not graph-based and does not target extreme events specifically. Both represent where the broader field is heading but do not resolve the core gap this review identifies.

4 The Extreme Weather Gap

4.1 The Smoothing Problem

Mean Squared Error (MSE) loss penalizes all prediction errors symmetrically, optimizing models to minimize average error by predicting near the conditional mean. Extreme weather events—heatwaves, cold snaps, intense precipitation, tropical cyclones—live at the tails of the distribution and are systematically smoothed out during training. This is structural, not incidental. Olivetti and Messori [15] establish that this is a training objective problem, not a data volume problem: more data alone will not fix it. The issue is that MSE (and MAE, which optimizes for the median) implicitly assumes symmetric error distributions, but weather variables exhibit high autocorrelation and asymmetric truncated distributions where forecast errors are rarely symmetric.

Subich et al. [20] provide mathematical proof of why MSE fails. They decompose MSE in spherical harmonic space into two terms: spectral amplitude preservation and coherence maximization. These terms are coupled—when predictability is low (coherence < 1), the MSE optimum is to reduce spectral amplitude proportionally ($\sigma_x = \rho \cdot \sigma_y$). This creates the double penalty effect: a correctly predicted extreme event that is slightly mistimed gets penalized twice—once for appearing at the wrong location, once for not appearing at the correct location. Models learn to avoid predicting peaks entirely. Fine-scale atmospheric variability, where extremes emerge from nonlinear amplification, becomes unpredictable first (higher Lyapunov exponents at smaller scales), so MSE-trained models smooth precisely where extremes concentrate. Subich et al. demonstrate this empirically by fine-tuning GraphCast with Adjusted MSE (AMSE), a loss function that decouples amplitude from coherence. The result: effective resolution improves from 1,250km to 160km, hurricane intensity predictions match ECMWF HRES quality, and surface wind extreme distributions become realistic rather than systematically underpredicted.

4.2 What Has Been Attempted

Olivetti and Messori [15] survey leading AI weather models (GraphCast, Pangu-Weather, FourCastNet, FuXi) and find a consistent pattern: all exhibit superior average skill on standard benchmarks but underestimate extreme magnitudes. Pangu-Weather, despite using MAE loss, achieves high binary detection skill for temperature extremes at 6-day lead times yet systematically underestimates their magnitude. FourCastNet underestimates high quantiles of 10m zonal wind and total precipitation. GraphCast tracks cyclones and atmospheric rivers better than HRES on average but fails on peak intensities. The authors identify three structural failure modes:

sparsity of extremes in training data (rare events constitute only 5-10% of timesteps), symmetric loss functions that penalize over- and under-prediction equally despite asymmetric real-world impacts, and multi-step optimization that compounds smoothing across autoregressive rollout.

Hakim and Masanam [8] test GraphCast, Pangu-Weather, and FourCastNet against ECMWF HRES on five real high-impact extreme events (hurricanes, heatwaves, cold snaps). All three AI models underperform the numerical baseline despite superior benchmark performance. Benchmark skill does not guarantee extreme event skill—this is the empirical failure mode that motivates the chaos-theoretic lens. Sun et al. [21], examining the 2024 Dubai extreme rainfall event, find that GraphCast and AIFS forecast it successfully up to 8 days ahead, but the generalization mechanism is translocation (spatial similarity with events in other regions) rather than temporal extrapolation. Current models cannot extrapolate beyond their training range; they recognize spatial analogs of extremes seen elsewhere.

Attempts to address the gap have focused on post-processing or alternative training objectives. Bülte et al. [4] propose a GNN post-processing framework targeting rainfall distribution tails, demonstrating that graph structure improves extreme ensemble prediction, but this is not an end-to-end solution. Shi et al. [19] introduce ReFine, a model-agnostic approach framing extremes as out-of-distribution (OOD) events. Training involves reweighting the loss function to upweight tail events and fine-tuning with partial layer freezing to prevent catastrophic forgetting while adapting to extreme distributions. ReFine improves extreme predictions across multiple time series datasets (air quality, climate, electrical demand) but has not yet been applied to ST-GNNs for weather. These efforts validate that distribution-sensitive training matters but do not explain why architectural properties might interact with training objectives.

4.3 Articulating the Gap

Even the most recent purpose-built regional ST-GNN [13], incorporating variable resolution (14km in extreme-prone areas, 28km elsewhere), dynamic graph structure learning, and physics-aware loss components (Huber loss with spatial mean constraints and Laplacian regularization), systematically underestimates heatwave peaks and typhoon intensity beyond three-hour lead times. The authors explicitly acknowledge: “its ability to forecast the peaks (or troughs) remains limited.” This is the smoking gun—architectural sophistication alone does not resolve the extreme weather gap. Li et al.’s design choices reveal the tension: Laplacian regularization, intended to suppress high-frequency noise, structurally smooths local peaks (the signature of extremes). Huber loss is more robust to outliers than MSE but does not target distribution tails. The model learns time-varying spatial dependencies but optimizes for average performance.

No existing work systematically asks why certain architectures handle extreme dynamics better, grounded in the mathematical properties of the system generating those extremes. The literature treats extreme weather prediction as an empirical optimization problem—try different losses, add resolution, learn dynamic graphs—without a unifying theoretical framework. This review argues the

gap is architectural in a deeper sense: extreme weather events are not statistical outliers. They are the signature of nonlinear amplification in chaotic dynamical systems. The atmosphere is a canonical chaotic system where small perturbations grow exponentially, predictability is state-dependent, and error growth peaks precisely when and where extreme events occur [2]. If extremes emerge from the mathematical structure of chaos, then predicting them requires architectures aligned with that structure. This is why architectural fixes targeting spatial resolution or graph adaptivity alone have failed to solve the smoothing problem—they address symptoms, not the fundamental mismatch between discrete-time MSE-optimized architectures and the continuous-time chaotic dynamics that generate extremes.

5 Chaos Theory as an Analytical Lens

5.1 Chaos Theory and Weather

Chaos theory describes deterministic systems that are globally stable but locally unstable: governed by fixed physical laws yet exhibiting exponential sensitivity to initial conditions. Small perturbations grow exponentially along unstable directions, quantified by Lyapunov exponents—the rate at which nearby trajectories in phase space diverge. A positive Lyapunov exponent indicates chaos; the larger its value, the faster predictability degrades. The atmosphere is a canonical chaotic system, established by Lorenz in 1963 when he discovered that tiny rounding errors in initial conditions produced wildly divergent weather forecasts. This is foundational to weather prediction, not peripheral: operational ensemble forecasting exists precisely because initial condition uncertainty amplifies exponentially through chaotic dynamics.

Local Lyapunov Exponents (LLEs) extend this concept by measuring finite-time error growth rates along trajectory segments rather than long-term averages. LLEs fluctuate as the system explores different regions of its attractor—some atmospheric states are more predictable than others. Ayers et al. [2] demonstrate that supervised machine learning can estimate LLEs directly from system state observations without requiring computationally expensive tangent linear models. Crucially, their analysis reveals that error growth peaks precisely when and where extreme events occur: prediction accuracy degrades fastest in regions where LLE values are locally heterogeneous, fluctuating rapidly across nearby states. This establishes the direct connection between chaos and extremes—extreme weather events are not statistical outliers but the signature of nonlinear amplification in regions of phase space where the system is most unstable.

5.2 Continuous-Time Models and Chaotic Dynamics

Discrete-time ST-GNNs introduce an architectural mismatch with chaotic systems. Chaotic dynamics evolve as continuous flows described by differential equations; discretizing time into fixed steps creates discontinuous latent state trajectories and accumulates integration errors. Jin et al. [10] address this directly with MTGODE, which models fully continuous spatial-temporal dynamics via coupled neural ODEs. The Continuous Graph Propagation (CGP) module solves a spatial ODE for graph message passing at

each infinitesimal time step, while Continuous Temporal Aggregation (CTA) parameterizes the derivative of latent states rather than the states themselves. This formulation eliminates over-smoothing in deep graph propagation by decoupling aggregation depth from integration time, and it achieves 30% parameter reduction and 4× faster training than discrete alternatives while matching or exceeding their accuracy across five benchmark datasets. Critically, MTGODE’s continuous formulation produces smooth latent trajectories; discrete models (GRU, STGCN, MTGNN) exhibit discontinuities at every timestep.

Verma et al. [22] apply continuous-time modeling to real atmospheric data with ClimODE, a physics-informed neural ODE implementing the advection equation (continuity PDE from statistical mechanics). The architecture uses second-order flow velocity dynamics to enforce mass conservation—weather quantities are spatially transported but never lost or created. ClimODE demonstrates improved long-range forecast stability on ERA5 reanalysis data, and component ablation reveals that the physics-informed advection structure contributes more to performance than architectural complexity (attention mechanisms). This validates that continuous-time formulations aligned with governing physical laws outperform discrete architectures on atmospheric prediction.

Liu et al. [14] provide comprehensive survey evidence: across 100+ methods integrating GNNs with differential equations (ODEs, PDEs, SDEs), Graph Neural Differential Equations (Graph NDEs) consistently outperform discrete ST-GNNs on continuous dynamical systems. The survey establishes that GNN message passing is fundamentally a discretization of continuous diffusion flow on graphs ($\partial x / \partial t = \text{div}[G(x, t) \cdot \nabla x]$), and continuous formulations provide smoother feature propagation, natural handling of irregular time intervals, and principled connections to physics-based diffusion PDEs. For weather forecasting on irregular station networks, this evidence is compelling: continuous-time architectures are not merely empirically superior but structurally aligned with the mathematical nature of atmospheric dynamics.

5.3 Chaos-Primed Architectures

The extreme weather gap persists despite architectural sophistication because current ST-GNNs lack structural alignment with chaotic system properties. Platt et al. [17] demonstrate that enforcing dynamical invariants—specifically Lyapunov exponents and fractal dimension—as soft constraints during RNN training extends forecast horizons in chaotic systems. Testing on the Lorenz 1996 model and a spectral quasi-geostrophic atmospheric model, they show that constraining just the largest Lyapunov exponent raises valid prediction time from 2 to 7 Lyapunov timescales. In data-limited regimes, where standard training collapses to <10-day prediction horizons, Lyapunov-constrained training recovers 20-60 day stability. This establishes the principle: chaos-theoretic properties should be enforced during training, not left to emerge from MSE minimization.

Ribeiro et al. [18] test purely data-driven models (LSTMs, reservoir computing) against physics-guided approaches on canonical chaotic systems (Lorenz, Rössler). Purely data-driven models deteriorate systematically on long-horizon prediction without governing mechanism constraints. Physics-informed approaches that embed

conservation laws or dynamical structure into the training loss maintain forecast skill. This finding generalizes beyond toy systems: operational weather prediction requires models that respect the underlying physics, and for chaotic systems, that physics includes sensitivity to initial conditions, exponential error growth, and fractal attractor structure.

We synthesize these insights into three properties that characterize chaos-primed ST-GNNs: (i) **continuous-time dynamics**, aligning with the fact that chaotic systems evolve as continuous flows rather than discrete jumps; (ii) **adaptive or dynamic graph structure**, capturing state-dependent spatial connectivity as chaotic systems explore their attractors; and (iii) **distribution-sensitive training objectives**, preserving variance at all scales rather than smoothing extremes to minimize average error. These criteria emerge from the intersection of chaos theory’s mathematical requirements (continuous dynamics, sensitivity to initial conditions), the extreme weather gap literature (MSE smooths tails), and recent ST-GNN advances (dynamic graphs improve performance in high-variability regions). No existing architecture scores well across all three—MTGODE achieves continuous-time + dynamic graphs but uses MAE loss; Subich et al.’s AMSE provides distribution-sensitive training but applies to discrete-time GraphCast with static graphs; CloudNine-v2 has dynamic graphs but discrete time and symmetric loss. The gap is architectural: current models optimize one or two properties in isolation rather than aligning all three.

Scoring existing ST-GNNs against these chaos-primed criteria is the novel analytical contribution of this review. Section 6 presents the comparative analysis.

6 Discussion and Critical Comparison

6.1 What the Literature Reveals

Table 1 presents the comparative analysis of 22 papers through the chaos-primed framework. The patterns are stark. No single architecture scores well across all three criteria. MTGODE [10] achieves the highest score among ST-GNN implementations (2/3: continuous-time + dynamic graphs), validated at meta-level by the Graph ODEs survey [14] which synthesizes evidence across 100+ methods. Yet even MTGODE uses MAE loss, optimizing for central tendency rather than distribution tails. Li et al. [13], scoring approximately 2.5/3 with dynamic graphs, variable resolution, and Huber loss (more robust than MSE but not distribution-sensitive), still systematically underestimates heatwave peaks and typhoon intensity. This is the empirical proof that partial alignment is insufficient—extreme weather prediction requires all three properties in combination.

Distribution-sensitive training (criterion 3) is the rarest. Only three papers explicitly target extremes through training objectives: Bülte et al. [4] with GNN post-processing for rainfall tails, Shi et al. [19] with reweighted loss and fine-tuning (model-agnostic, not yet applied to weather ST-GNNs), and Subich et al.’s AMSE [20] which decouples spectral amplitude from coherence to preserve fine-scale variance. All three demonstrate that distribution-sensitive training improves extreme prediction, but none combine this with continuous-time dynamics or dynamic graphs. The gap is clear: architectures pursue spatial sophistication (adaptive graphs) or temporal sophistication (continuous-time ODEs) while training

objectives remain locked to symmetric loss functions that smooth tails.

The South African studies [5, 6] sit at one end of the spectrum: strong spatial modeling (adaptive graphs, 1/3 score), establishing that graph structure adds predictive signal to irregular station networks, but no extreme event targeting and no chaos-theoretic alignment. They represent the baseline from which this project builds. GraphCast [12] and Aurora [3]—the landmark SOTA models—score 0/3 or N/A: discrete-time, static graphs (GraphCast uses a fixed icosahedral mesh), MSE loss. Their benchmark dominance (90-94% of verification targets) does not translate to extreme event skill [8]. Hakim and Masanam’s validation on five real high-impact events exposes the disconnect: models optimized for average performance via symmetric loss systematically underperform on peaks.

CloudNine-v2 [9] offers indirect evidence for criterion 2. Its fully dynamic graph structure learning (reconstructing adjacency at every timestep via Gumbel-Softmax) reduces the performance gap between high-variability and low-variability regions by a factor of 3-5 compared to fixed-topology models. High atmospheric variability occurs precisely where extreme weather concentrates—coastal zones, mountainous terrain, frontal boundaries. This suggests dynamic graphs capture time-varying spatial dependency structure during extreme conditions, though CloudNine-v2 evaluates only average RMSE without stratifying by event intensity. The study validates that spatial structure matters for extremes but does not test whether dynamic graphs alone suffice without continuous-time dynamics or distribution-sensitive training.

Gupta et al. [7] provide critical context for the South African application. Comparing task-specific ST-GNNs against Time Series Foundation Models (Chronos, Moirai, TimesFM) on an urban temperature sensor network, they find ST-GNNs outperform when spatial dependencies are strong and node coverage is sparse. Foundation models’ general-purpose pretraining does not automatically capture domain-specific spatial structure. For South Africa’s irregular station network shaped by complex topography (Great Escarpment, coastal effects, elevation gradients), this validates the graph-based approach over foundation models. The chaos-primed framework extends this logic: if spatial structure matters, then *dynamic* spatial structure matters more during chaotic extremes when influence radii shift with atmospheric state.

The literature collectively demonstrates that architectural sophistication alone does not resolve the extreme weather gap. Li et al. [13], incorporating variable resolution, dynamic graphs, and Huber loss with spatial constraints, represents the current frontier—yet the authors explicitly acknowledge their model’s limited ability to forecast peaks. The gap is not computational (more parameters, higher resolution) but conceptual: current architectures and training objectives are not aligned with the mathematical structure of the chaotic systems generating extremes. Section 5 established that chaotic dynamics require continuous-time evolution, state-dependent connectivity, and preservation of fine-scale variance. Section 4 documented that MSE and MAE smooth distribution tails. Table 1 shows no existing work combines these insights. The chaos-primed framework is the missing analytical lens.

6.2 Comparative Table

The table reveals several additional patterns. First, explicit extreme event evaluation is rare—only 6 of 22 papers test performance on extremes separately from average metrics, and several of these (Hakim, Sun, Olivetti & Messori, Subich) are validation or analysis studies rather than architecture proposals. Second, temporal approaches cluster: RNN-based (GRU, LSTM) for adaptive learning, CNN-based for parallelization, autoregressive for long-range roll-out, and neural ODEs for continuous dynamics. The ODE cluster (MTGODE, ClimODE, Graph ODEs survey) consistently achieves criterion 1, suggesting continuous-time is an architectural choice that propagates across related work rather than emerging independently. Third, the chaos theory constraint papers (Ayers, Platt, Ribeiro) validate principles but do not implement full ST-GNN architectures, leaving a methodological gap: *how to integrate Lyapunov constraints, dynamic graphs, and distribution-sensitive training into a single operational forecasting system*.

This review synthesizes the landscape to inform experimental design. MTGODE serves as the natural baseline for continuous-time + dynamic graphs. AGCRN and GraphWaveNet represent discrete-time + adaptive graphs (the South African baseline). Subich et al.’s AMSE provides the distribution-sensitive training blueprint, though adapting spherical harmonic decomposition to irregular regional grids requires methodological innovation (wavelet lifting or local spectral analysis). The South African context—sparse stations, irregular topology, high climate variability, growing extreme event exposure—makes it an ideal testbed for chaos-primed architectures where spatial structure, temporal continuity, and distribution sensitivity all matter simultaneously.

7 Conclusions

This review establishes that predicting extreme weather events requires more than architectural sophistication—it requires structural alignment between ST-GNN architectures and the mathematical properties of the chaotic dynamical systems generating those extremes. We have demonstrated three core findings.

First, the extreme weather gap is real, well-documented, and unresolved by the latest architectures. GraphCast outperforms ECMWF HRES on 90% of 1380 benchmark targets yet systematically underestimates extreme event magnitudes [8]. Li et al.’s variable-resolution ST-GNN [13], incorporating dynamic graphs and physics-aware loss components, still underestimates heatwave peaks and typhoon intensity beyond three-hour lead times. The gap persists because current approaches treat extremes as an optimization problem (add resolution, learn dynamic graphs, try different losses) rather than recognizing that extremes are the signature of nonlinear amplification in chaotic systems. MSE and MAE loss structurally smooth distribution tails [15, 20]. Discrete-time architectures introduce discontinuous latent trajectories misaligned with continuous chaotic flows. Static graphs assume fixed spatial dependencies in systems where connectivity shifts with atmospheric state.

Second, chaos theory provides the analytical framework for understanding why partial solutions fail and which architectural properties matter. We synthesized three criteria that characterize chaos-primed ST-GNNs: continuous-time dynamics (aligning with the differential equation nature of chaotic flows), adaptive or

dynamic graph structure (capturing state-dependent spatial connectivity), and distribution-sensitive training objectives (preserving fine-scale variance rather than smoothing tails). These criteria emerge from the intersection of chaos theory's mathematical requirements [2, 17, 18], the extreme weather gap literature [15, 20], and recent ST-GNN advances [9, 10, 22]. The comparative analysis in Table 1 reveals that no existing architecture scores well across all three. MTGODE achieves 2/3 (continuous-time + dynamic graphs but MAE loss). Li et al. achieves approximately 2.5/3 yet still fails on extremes, empirically proving that partial alignment is insufficient.

Third, ST-GNNs remain the right paradigm for weather forecasting when spatial structure matters, and the South African context makes this work both novel and urgent. Gupta et al. [7] demonstrate that task-specific ST-GNNs outperform foundation models when spatial dependencies are strong and node coverage is sparse—precisely the characteristics of South Africa's irregular station network shaped by complex topography, coastal effects, and elevation gradients. Only two prior studies [5, 6] have applied ST-GNNs to South African weather, both establishing that adaptive graphs add predictive signal but neither targeting extremes or incorporating chaos-theoretic constraints. High climate variability, sparse observation networks, and growing extreme event exposure driven by climate change make South Africa an ideal testbed for evaluating whether chaos-primed architectures outperform standard approaches on regional extreme weather prediction.

Several open questions remain. Can chaos-theoretic constraints (Lyapunov exponents, fractal dimension) be incorporated into ST-GNN training without prohibitive computational cost? Platt et al. [17] show this is feasible for reservoir computing on low-dimensional chaotic systems; scaling to operational weather models with 10^9 variables is unresolved. Does the chaos-primed framework generalize beyond temperature to wind and precipitation, where distribution characteristics differ (non-negative, localized, non-Gaussian)? Subich et al. note that AMSE degrades on precipitation precisely because spherical harmonic decomposition assumes Gaussian statistics. Can the framework be validated experimentally, or does it remain a post-hoc analytical tool? This requires implementing a chaos-primed ST-GNN that scores 3/3 and comparing it against partial implementations in controlled ablations.

The path forward is clear. Benchmark MTGODE (continuous-time + dynamic graphs, MAE loss) against AGCRN and GraphWaveNet (discrete-time + adaptive graphs, MSE loss) as representatives of chaos-primed versus standard approaches on South African data. Implement distribution-sensitive training via AMSE [20] or quantile loss, adapting spherical harmonic techniques to irregular regional grids through wavelet lifting or local spectral analysis. Evaluate explicitly on extreme events (heatwaves, cold snaps) with metrics stratified by intensity: MAE and RMSE on extreme versus non-extreme timesteps, peak intensity error, timing accuracy for threshold exceedance. Estimate Local Lyapunov Exponents [2] from model predictions to verify whether chaos-primed architectures preserve the Lyapunov structure that MSE-trained models destroy through fine-scale smoothing. Apply the chaos-primed scoring framework as the novel analytical contribution, demonstrating that architectural alignment with chaotic system properties—continuous-time dynamics, dynamic graphs, distribution-sensitive training, all three

in combination—is not optional for extreme weather prediction. It is structurally necessary.

References

- [1] Zahraa Al Sahili and Mariette Awad. 2023. Spatio-Temporal Graph Neural Networks: A Survey. *arXiv preprint arXiv:2301.10569* (2023). Sections 2.1, 2.3.
- [2] John Ayers, William Lau, Javier Amezcua, and Alberto Carrassi. 2023. Supervised ML to Estimate Instabilities in Chaotic Systems: Local Lyapunov Exponents. *Quarterly Journal of the Royal Meteorological Society* (2023). Sections 5.1, 5.3.
- [3] Cristian Bodnar et al. 2025. A Foundation Model for the Earth System / Aurora. *Nature* (2025). Section 3.3.
- [4] Christopher Bulte, Sohir Maskey, Philipp Scholl, Jonas von Berg, and Gitta Kutnyok. 2025. Graph Neural Networks for Enhancing Ensemble Forecasts of Extreme Rainfall. *arXiv preprint arXiv:2504.05471* (2025). Sections 4.2, 4.3.
- [5] Mikhail Davidson and Deshen Moodley. 2022. ST-GNNs for Weather Prediction in South Africa. In *Southern African Conference for Artificial Intelligence Research (SACAIR)*. Sections 3.1, 3.2.
- [6] Adeeb Gaibie, Hamza Amir, Irene Nandutu, and Deshen Moodley. 2024. Predicting and Discovering Weather Patterns in South Africa Using Spatial-Temporal Graph Neural Networks. In *Southern African Conference for Artificial Intelligence Research (SACAIR)*. Section 3.2.
- [7] Abhiraj Gupta, Raghav Raina, Di Chen, Yuxun Chen, Dmitrii Danilov, Julius Eckhardt, Carlyn Bernard, and Klara Nahrstedt. 2025. Evaluating Spatio-Temporal Forecasting Trade-offs Between Graph Neural Networks and Foundation Models. In *ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys)*. Section 6.
- [8] Gregory Hakim and Ravi Masanam. 2025. Validating Deep Learning Weather Forecast Models on Recent High-Impact Extreme Events. In *AAAI/ACM Conference on AI, Ethics, and Society (AIIES)*. Section 4.2.
- [9] Hyunwook Jeon, Seungyeon Park, and Youngjoon Yoon. 2025. Discovering Spatial Correlations of Earth Observations for Weather Forecasting by Using Graph Structure Learning / CloudNine-v2. In *ACM International Conference on Relational and Algebraic Methods in Computer Science (RACS)*. Section 3.3.
- [10] Ming Jin, Yu Zheng, Yuan-Fang Li, et al. 2022. Multivariate Time Series Forecasting with Dynamic Graph Neural ODEs. *IEEE Transactions on Knowledge and Data Engineering* (2022). Sections 2.2, 5.2.
- [11] Ming Jin, Yu Zheng, Yuan-Fang Li, et al. 2023. Spatio-Temporal Graph Neural Networks for Predictive Learning in Urban Computing: A Survey. *IEEE Transactions on Knowledge and Data Engineering* (2023). Sections 2.1, 2.3.
- [12] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, et al. 2023. Learning Skillful Medium-Range Global Weather Forecasting. *Science* 382 (2023), 1416–1421. Section 3.3.
- [13] Qiang Li, Xuemei Chen, Yongfeng Huang, et al. 2026. A Variable-Resolution Unstructured STGNN for Very Short-Range Weather Forecasting in Guangdong, China. *Geoscience Letters* (2026). Sections 3.3, 4.3.
- [14] Zewen Liu, Xunqiang Wang, Zhiyuan Wang, Tongya Huang, Chaoze Yang, and Ming Jin. 2025. Graph ODEs and Beyond: A Comprehensive Survey on Integrating Differential Equations with Graph Neural Networks. *arXiv preprint arXiv:2503.23167* (2025). Section 5.2.
- [15] Leonardo Olivetti and Gabriele Messori. 2024. Advances and Prospects of Deep Learning for Medium-Range Extreme Weather Forecasting. *Geoscientific Model Development* 17 (2024), 2347–2358. Section 4.1.
- [16] Joel Oskarsson et al. 2024. Probabilistic Weather Forecasting with Hierarchical Graph Neural Networks / Graph-EFM. In *Advances in Neural Information Processing Systems (NeurIPS)*. Sections 3.3, 4.2.
- [17] John A. Platt et al. 2023. Constraining Chaos: Enforcing Dynamical Invariants in the Training of Recurrent Neural Networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science* (2023). Section 5.3.
- [18] Larissa Ribeiro et al. 2024. Toward a Physics-Guided ML Approach for Predicting Chaotic Systems Dynamics. *Frontiers in Big Data* (2024). Section 5.3.
- [19] Chen Shi, Amir Shirali, and Srinivasan Narasimhan. 2024. ReFine: Boosting Time Series Prediction of Extreme Events by Reweighting and Fine-tuning. In *IEEE International Conference on Big Data (BigData)*. Section 4.2.
- [20] Christopher Subich, Simon Chamberland, and Naveed Husain. 2025. Fixing the Double Penalty in Data-Driven Weather Forecasting Through a Modified Spherical Harmonic Loss Function. *arXiv preprint arXiv:2501.19374* (2025). Section 4.1.
- [21] Y. Qiang Sun, Pedram Hassanzadeh, Tiffany Shaw, and Hamid A. Pahlavan. 2025. Predicting Beyond Training Data via Extrapolation versus Translocation: AI Weather Models and Dubai's Unprecedented 2024 Rainfall. *arXiv preprint arXiv:2505.10241* (2025). Section 4.2.
- [22] Shashank Verma, Markus Heinonen, and Vikas Garg. 2024. ClimODE: Climate and Weather Forecasting with Physics-Informed Neural ODEs. In *International Conference on Learning Representations (ICLR)*. Section 5.2.
- [23] Andrew Wilson, Ken Tan, and Rui Luo. 2018. A Low Rank Weighted Graph Convolutional Approach to Weather Prediction. In *IEEE International Conference*

- on Data Mining (ICDM)*. Section 3.1.
- [24] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*

(IJCAI). Section 2.2.

Received 18 March 2026

Table 1: Comparative Analysis of ST-GNN Architectures and Related Work Through Chaos-Primed Framework

Paper	Domain / Dataset	Spatial	Temporal	Graph Type	Extremes?	Loss	Key Finding(s)	Chaos (CT/AG/DS)
ST-GNN Foundations & South African Baselines								
Yu et al. 2018 (STGCN) [24]	Traffic / PeMSD7	Spectral GCN	CNN (gated)	Static	No	MSE	First ST-GNN; established CNN-based temporal modeling as parallelizable alternative to RNNs	No / No / No
Wu et al. 2019 (GWN)	Traffic / METR-LA, PEMS-BAY	Adaptive GCN	CNN (dilated)	Adaptive	No	MAE	Self-adaptive adjacency matrix discovers hidden spatial dependencies beyond distance	No / Yes / No
Wilson et al. 2018 [23]	Global / WeatherBench	Low Rank WGN	CNN	Static	No	MSE	First ST-GNN for weather; established viability on global data	No / No / No
Davidson & Moodley 2022 [5]	SA / 21 stations	Adaptive GCN	CNN (TCN)	Adaptive	No	MSE	8.30% SMAPE baseline for SA; GWN beats WGN (9.02%); establishes SA as novel context	No / Yes / No
Gaibie et al. 2024 [6]	SA / 45 stations	Adaptive GCN	RNN (GRU)	Adaptive	No	MSE	AGCRN best performance; CLCRN learns plausible coastal paths; performance vs interpretability diverge	No / Yes / No
Recent SOTA Advances & Foundation Models								
Lam et al. 2023 (GraphCast) [12]	Global / ERA5	GNN (mesh)	AR (6h)	Static	Partial	MSE	Beats HRES on 90% of targets; underperforms on real extremes (Hakim 2025)	No / No / No
Jeon et al. 2025 (CloudNine-v2) [9]	Global / ERA5	GCN	CNN	Dynamic (GSL)	No	MSE	15% RMSE reduction vs fixed; 3-5x smaller gap in high-variability regions	No / Yes / No
Li et al. 2026 [13]	Regional / Guangdong 250 nodes	Variable-res + dynamic	CNN (AR)	Dynamic	Yes (HW, TC)	Huber + reg	Underestimates peaks 1°C, typhoon beyond 3h; dynamic + res insufficient	No / Yes / Partial
Bodnar et al. 2025 (Aurora) [3]	Global / multi	Transformer	AR	N/A	No	MSE	1.3B params; foundation model context; not graph-based	N/A
Oskarsson et al. 2024 (Graph-EFM) [16]	Global / ERA5	Hierarchical GNN	Ensemble	Static	Motivated	Ensemble	Probabilistic GNN for full distribution including extremes; uncertainty estimates	No / No / Partial
Continuous-Time & Physics-Informed								
Jin et al. 2022 (MTGODE) [10]	Traffic, energy	Neural ODE	Neural ODE	Dynamic	No	MAE	First continuous ST-GNN; 30% fewer params, 4x faster; no graph priors	Yes / Yes / No
Verma et al. 2024 (ClimODE) [22]	Global / ERA5	CNN + Attn	2nd-order ODE	N/A (grid)	No	MSE	Physics-informed ODE; advection PDE; improved long-horizon stability; not graph-based	Yes / N/A / No
Liu et al. 2025 (Survey) [14]	Survey	Graph NDEs	ODEs/PDEs	Dynamic	No	N/A	GNN-ODEs outperform discrete on continuous systems; handles irregular time	Yes / Yes / No
Extreme Weather Gap & Distribution-Sensitive Training								
Olivetti & Messori 2024 [15]	Review / Global	N/A	N/A	N/A	Yes	N/A	Training objective problem; 3 failure modes (sparsity, symmetric loss, multi-step)	N/A
Subich et al. 2025 [20]	Analysis	N/A	N/A	N/A	Yes	MSE vs AMSE	Double penalty: mistimed extreme penalized twice; models avoid peaks	N/A
Hakim & Masanam 2025 [8]	Validation	N/A	N/A	N/A	5 events	N/A	GraphCast, Pangu, FourCast all underperform HRES on real extremes despite benchmarks	N/A
Sun et al. 2025 [21]	Dubai case	N/A	N/A	N/A	Yes	N/A	GraphCast & AIFS forecast Dubai via translocation (spatial), not extrapolation	N/A
Bulte et al. 2025 [4]	Post-process	GNN	N/A	Static	Rainfall	Dist-target	GNN post-processing targets tails; graph structure matters for extremes	No / No / Yes
Shi et al. 2024 (ReFine) [19]	Time series	Agnostic	Agnostic	N/A	Yes (OOD)	Reweightd	Weighted loss + fine-tuning improves extremes; not yet for weather ST-GNNs	N/A / N/A / Yes
Chaos Theory & Physics-Informed ML								
Ayers et al. 2023 [2]	Lyapunov / Weather	N/A	N/A	N/A	Implicit	N/A	ML estimates LLEs; error growth highest at/when extremes occur; quantifies instability	N/A
Platt et al. 2023 [17]	RNN / Lorenz	RNN + constraints	RNN	N/A	Yes	MSE + constr	Enforcing Lyapunov + fractal dim as constraints → 7x longer forecasts	N/A / N/A / Partial
Ribeiro et al. 2024 [18]	Physics-ML / Lorenz	Hybrid	RNN, RC	N/A	Yes	N/A	Pure data-driven deteriorates on chaotic; physics-informed essential	N/A
Foundation Model Comparison								
Gupta et al. 2025 [7]	Urban IoT	ST-GNNs vs TSFMs	Multiple	Task-specific	No	MAE	ST-GNNs beat foundation models when spatial dependencies strong; validates graphs for sparse	No / Yes / No