

Evaluating Spatio-Temporal Graph Neural Networks and their Limits for Extreme Weather Forecasting

Génevieve Chikwanha
chkkar002@myuct.ac.za
University of Cape Town
Cape Town, South Africa

Abstract

Spatial-Temporal Graph Neural Networks (STGNNs) have emerged as a promising framework for weather forecasting. However, their efficacy in predicting extreme events has not extensively been explored. A review of the literature was conducted to evaluate the extent to which STGNNs can effectively model and predict extreme weather events. STGNNs extend deep neural network approaches by explicitly modelling both spatial and temporal dependencies in multivariate time series data, which makes them well-suited to weather forecasting tasks. Our analysis reveals that architectural components such as adaptive graph learning and higher-order relational modelling improve predictive performance under stable conditions. However, these models demonstrate consistent limitations when applied to extreme weather events. These limitations arise from data scarcity, bias inheritance from observational datasets, architectural constraints such as over-smoothing, and the inherent chaotic nature of weather systems. In particular, prediction error increases rapidly with lead time, which suggests a breakdown in model reliability over longer horizons. This suggests that the concept of a prediction horizon, which measures the temporal extent over which predictions remain valid, might reveal insights for understanding model performance. This review highlights the need for systematic evaluation of STGNN architectures under extreme weather conditions. In doing so, it aims to better characterise how accurately these models can predict weather as well as how long those predictions remain meaningful.

Keywords

Spatial-Temporal Graph-Based Neural Network (ST-GNN), Weather Prediction, Extreme Weather Prediction, Chaotic System

1 Introduction

Traditional forecasting methods do not explicitly encode both spatial and temporal dependencies present in weather systems [39]. This limits their ability to capture the complex interactions that govern atmospheric dynamics. In response, artificial intelligence (AI)-based approaches, particularly machine learning (ML) and deep learning (DL), have demonstrated strong capabilities for modelling and predicting weather events. These models can learn non-linear relationships directly from data without requiring explicit knowledge of the underlying physical processes [23]. Deep learning approaches are particularly effective in high-dimensional settings. Convolutional neural networks (CNNs) model spatial dependencies, while recurrent neural networks (RNNs) capture temporal

dynamics in time-series data [31]. However, these approaches often treat spatial and temporal dependencies separately and do not explicitly model the pairwise interdependencies between variables. This limits their ability to capture complex, structured interactions that are critical for detecting and predicting anomalous or extreme behaviour [14].

1.1 Spatio-Temporal Graph Neural Networks for Weather Forecasting

Spatial-Temporal Graph Neural Networks (STGNNs) address this limitation by representing weather systems as graphs, where nodes correspond to spatial locations and edges encode the relationships between them [3, 9]. However, weather prediction is not a uniform task. Standard weather forecasting focuses on modelling the continuous evolution of atmospheric variables, like temperature or humidity, whereas extreme event prediction requires identifying rare deviations from these patterns. These events are influenced by multiple interacting factors, including seasonality, geographical location, data quality, and forecast lead time [7]. Furthermore, they are often under-represented in observational datasets, which makes them difficult to learn and evaluate [14]. The difficulty is further compounded by the nature of the weather system itself. Weather is a non-linear, chaotic dynamical system, which means that small errors in initial conditions can lead to large deviations in future states. As a result, prediction accuracy degrades over time, which imposes an inherent limit on how far into the future reliable forecasts can be made [23, 24]. This raises an important questions about how accurate a model is and how long its predictions remain valid.

1.2 Research Gap and Objectives

Current STGNNs have been evaluated on a limited range of extreme weather scenarios, if at all [26]. This highlights a gap in the literature, which this review aims to address. This review evaluates the extent to which STGNNs can effectively model and predict extreme weather events. Specifically, we analyse these models in terms of their ability to capture complex spatial dependencies, model non-linear temporal dynamics, and maintain predictive performance under chaotic conditions and increasing forecast horizons. To do this, we examine recent STGNN architectures alongside temporal-only baselines, and analyse how design choices such as graph construction and adaptive learning impact performance. We also consider how these models behave under extreme conditions, where data is sparse and errors propagate more rapidly. The remainder of this

paper is structured as follows. Section 2 describes the weather system as a dynamical system and defines extreme weather events. Section 3 outlines the STGNN framework and the construction of spatio-temporal graphs from weather data. Section 4 presents and analyses recent STGNN architectures and their performance on weather prediction tasks. Section 5 examines the limitations of these models, with a focus on extreme events and predictability horizons. Finally, Section 6 concludes and outlines directions for future work.

2 Weather Systems and Extreme Events

Weather is a non-linear, dynamical, complex, and chaotic system. Each of these characteristics contributes to why weather prediction is fundamentally difficult, and why modelling it requires approaches capable of capturing uncertainty. Unlike linear systems, where outputs scale proportionally with inputs, non-linear systems do not exhibit such behaviour [11]. As a result, the scale of the input may differ significantly from the scale of the output. In a dynamical system, where variables evolve over time, this non-linearity means that the trajectory of the system is highly sensitive to both current conditions and how those conditions change [11]. Thus, small changes in input can produce disproportionately large or negligible changes in output. This often gives the impression that such systems are non-deterministic or inherently unpredictable, when in fact they may still be governed by deterministic rules. Weather is further characterised as a complex system that consists of multiple interacting subsystems operating across different spatial and temporal scales [8]. These interactions give rise to emergent behaviour that cannot be understood by analysing individual components in isolation [1, 8]. For example, rainfall can be described as the result of water evaporating from the ocean into the atmosphere, given sufficient heat from the sun, which is then released back down when the vapour condenses to form water droplets [1]. This process is not governed by a single timescale or linear mechanism. Instead, it emerges from the continuous interaction between atmospheric and oceanic conditions and solar energy input [1]. Finally, weather exhibits chaotic behaviour. While there is no universally agreed-upon definition of chaos, it is commonly understood as deterministic behaviour that is highly sensitive to initial conditions [5, 6]. Infinitesimal perturbations in the initial state of the system can lead to vastly different trajectories over time. This sensitivity limits the extent to which long-term predictions can be made reliably. The weather system can thus be understood as a complex collection of interacting chaotic subsystems evolving over multiple time scales [4].

2.1 Extreme Weather Events

Extreme events are typically defined relative to a threshold. A meteorological event is considered extreme when a measured variable, such as temperature, precipitation, or humidity, exceeds or falls below a predefined boundary [23]. These thresholds are often context-dependent and vary across regions and seasons, which makes the definition of “extreme” somewhat relative instead of absolute. There is currently no single, precise statistical definition that captures all

extreme weather events or their underlying mechanisms. Instead, events are often categorised based on their temporal scale. Sillmann et al. [30] distinguish between short-duration extremes (SDEs) and long-duration extremes (LDEs). SDEs persist for less than three days and include phenomena such as intense rainfall, storms, and heatwaves. In contrast, LDEs extend beyond three days and include events such as droughts, prolonged cold spells, and large-scale flooding [30]. Despite their differences in duration, both types of events arise from a combination of factors, including favourable initial atmospheric conditions, large-scale climatic drivers, local feedback mechanisms, and stochastic influences [2, 23, 30]. This dependence on multiple factors reinforces the complexity of extreme events, since their occurrence is not tied to a single causal mechanism but instead to the interaction of multiple processes across scales. It is also important to distinguish between extreme events and anomalous events. While both may be rare, anomalous events are defined primarily by their rarity, whereas extreme events are defined relative to some threshold [23]. Extreme events are often also rare, which introduces a significant challenge: limited data availability. Because these events don’t occur frequently, they are under-represented in observational datasets, which makes them difficult to model and evaluate [14].

2.2 Predictability Limits

Predicting weather events in the long term has been a long-held scientific goal [24, 30]. However, the nature of weather events, as we have described, makes them an inherently difficult prediction task. The weather is said to be predictable over some period of time if, given a perfect model with initial errors, it can be accurately forecasted [24, 33]. Traditionally, the predictability of a chaotic system is thought to be determined by the largest Lyapunov exponent, which measures the rate at which trajectories emanating from two infinitesimally-spaced points will exponentially separate over time [10]. A system can provide very short-term (up to 12 hours), short-term (from 12 to 72 hours), medium-term (up to 10-14 days), and long-term (monthly or seasonal) predictions [29]. The weather has proven to be predictable over short-medium periods [5]. Specifically, the Intrinsic Predictability Limit (IPL), which specifies a theoretical, estimated limit on the time range of daily weather forecasts, is 10-15 days in the Northern Hemisphere, after which the weather becomes deterministically unpredictable [23, 24]. This happens because any infinitesimal error in the initial state of a deterministic, non-linear system compounds over time, leading to a divergence in the prediction [4, 6, 21, 24]. So, accurate long-term weather prediction is inherently limited. Still, reliable forecasts, especially for extreme events, are crucial for disaster preparedness. Furthermore, even within the constraints of the IPL, determining how close STGNNs can get to theoretical limits remains a compelling and necessary challenge for both atmospheric and computational science.

2.3 Problem Formulation

Given a set of nodes V (where $v \in V$) and a set of edges E (where $e \in E$), we can construct a graph G (where $G = (V, E)$) that models weather system dynamics. v represents a weather station and

encapsulates the dynamics of that particular location, while E represents the relationships between stations in V [3, 9, 32]. Weather observations, and their corresponding meteorological features (e.g. temperature, rainfall, cloud cover, or humidity), can be represented as time-series data, which is a sequence of observations ordered by time:

$$X = \{X_1, X_2, X_3, \dots, X_T\} \in \mathbb{R}^{T \times n \times d} \quad (1)$$

where X_t is an observation at time $t \in T$ from a particular weather station, n is the number of locations (also given by $|V|$), and d is the number of meteorological features [14, 20, 31]. When $d = 1$, the time series is univariate [20]. For weather observations, often $d > 1$, and so they are multivariate [20]. The goal is to use a sequence of historical observations, X , from a network of n weather stations, along with the mathematical representation of the graph, the adjacency matrix $A \in \mathbb{R}^{n \times n}$, to predict some future sequence of values $y_1, y_2, y_3, \dots, y_H$ for a given feature at each v over some prediction horizon H that specifies the length of the sequence to be predicted [3, 9, 31, 36].

3 Designing and Evaluating Spatial-Temporal Graph Neural Networks

Deep Neural Networks (DNNs) have shown promising results for weather forecasting tasks, largely due to their ability to capture non-linear relationships in large, high-dimensional data [35, 36, 40]. However, earlier DNNs fail to fully capture the inherent spatial and temporal dependencies in weather data [3, 9]. Thus, Spatial-Temporal Graph Neural Networks (STGNNs) have emerged to address this shortcoming. STGNNs are able to capture spatial and temporal dependencies simultaneously, making them suitable for tasks with a large amount of interaction between spatial and temporal elements [22, 38]. Understanding how STGNNs are constructed, modelled, and evaluated allows for a more fine-grained comparison between models.

3.1 Spatio-Temporal Graph Neural Network Framework

Across the literature reviewed, forecasting models using STGNNs were developed using some approximation of the following steps. First, feed the model with some weather dataset and a pre-defined or learned adjacency matrix representing the graph. Next, pass the matrix on to the STGNN architecture, which includes some variation of a spatial and temporal module, to extract the relevant dependencies. Then, train the model until the loss function converges. Finally, evaluate the model's performance according to some evaluation metric and compare the results with existing STGNNs or other weather forecasting models. Pseudocode for this general pipeline is given in Algorithm 1.

3.2 Datasets for Weather Forecasting

Sun et al. [31] identify three categories of meteorological datasets.

Reanalysis Datasets. These datasets combine NWP models with historical weather observation data across a variety of meteorological features to reconstruct past atmospheric states [31]. ERA5 is one

Algorithm 1 General Spatio-Temporal Graph Neural Network (STGNN) Framework

Require: Time-series data X , optional adjacency matrix A , model parameters θ
Ensure: Predictions $\hat{Y}$

- 1: **for** each epoch **do**
- 2: **for** each batch (X, Y) **do**
- 3: **1. Input Preparation**
- 4: Extract temporal window $X_{t-k:t}$
- 5: Perform preprocessing (e.g., normalisation, embedding)
- 6: Initialise or provide adjacency matrix A
- 7: **2. Spatial Module**
- 8: Optionally refine or learn adjacency A
- 9: $Z_s \leftarrow \text{SpatialEncoder}(X_{t-k:t}, A)$
- 10: **3. Temporal Module**
- 11: $Z_t \leftarrow \text{TemporalEncoder}(Z_s)$
- 12: **4. Fusion**
- 13: $Z \leftarrow \text{Combine}(Z_s, Z_t)$ ▷ concatenation / residual / attention
- 14: **5. Prediction**
- 15: $\hat{Y} \leftarrow \text{Predictor}(Z)$
- 16: **6. Loss Computation**
- 17: $\mathcal{L}_{pred} \leftarrow \text{PredictionLoss}(\hat{Y}, Y)$
- 18: $\mathcal{L}_{reg} \leftarrow \text{RegularisationTerms}$
- 19: $\mathcal{L} \leftarrow \mathcal{L}_{pred} + \mathcal{L}_{reg}$
- 20: **7. Update**
- 21: Update θ using gradients of $\mathcal{L}$
- 22: **end for**
- 23: **end for**
- 24: **return** $\hat{Y}$

of the most widely used reanalysis datasets for global weather data in ML-based studies [17]. The dataset contains hourly observations from 1940 to present, with a spatial resolution of $0.25^\circ \times 0.25^\circ$ globally. This makes it suitable for medium-range forecasting tasks, but it may under-represent small-scale extreme events [17]. The relatively coarse spatial resolution limits the model's ability to capture fine-scale, localised meteorological dynamics [17]. Additionally, biases present in historical observations or introduced through NWP models can propagate into the dataset [31].

Surface Observation Datasets. These datasets are constructed from a global network of meteorological stations that directly record atmospheric conditions. NOAA's Integrated Surface Database (ISD) is one of the most widely used datasets in this category. It provides high-resolution, hourly observations from 1901 to present. While these datasets generally contain higher-quality measurements than reanalysis datasets, their spatial coverage is uneven, with denser observations in the Northern Hemisphere [31]. This imbalance may introduce geographical bias and limit generalisability.

Remote Sensing Datasets. Remote sensing datasets provide continuous, large-scale spatial coverage using satellite-based observations. For example, the GOES-R satellite series provides high-frequency observations of atmospheric and cloud dynamics over the

Table 1: Summary of Commonly-Used Meteorological Datasets in STGNN Studies

Category	Dataset	Time age	Coverage	Time Res.	Spatial Cov.	Spatial Res.	Key Variables
Reanalysis	ERA5	1940–present	Hourly	Global		0.25°	Geopotential (z), temperature (t), specific humidity (q), wind components (u, v), relative humidity (r)
	WeatherBench	1979–2018	Hourly	Global		1.4°–5.6°	z, t, q, r, u, v , vorticity (vo), potential vorticity (pv)
	WeatherBench2	1959–2023	Hourly	Global		0.25°	ERA5-derived variables + IFS ensemble (ENS) forecasts
Surface Observation	NOAA ISD	1901–present	Hourly	Global		Station-based	Temperature, wind speed/direction, pressure, precipitation
Remote Sensing	GOES-R	2017–present	5–15 min	Regional (Americas)		~0.5–2 km	Cloud properties, radiances, atmospheric motion vectors, moisture

Americas, and captures variables such as radiances and cloud properties. These datasets offer significantly higher spatial and temporal resolution compared to reanalysis and surface observations, making them valuable for capturing rapidly evolving weather systems. However, they introduce additional complexity for STGNN-based models due to their high dimensionality and indirect measurement of atmospheric variables [31].

3.2.1 Data Limitation. Extreme events are, by definition, rare, which leads to their under-representation in available datasets [39]. As a result, models are unable to train on sufficient examples of extreme events, which causes degraded performance during prediction [39]. This contributes to the generalisation problem in AI-based forecasting models, where performance declines when applied to unseen geographical regions or previously unobserved conditions [39]. These challenges are particularly pronounced for extreme weather events. Each event requires a distinct definition to capture its underlying physical mechanisms, which vary across temporal and spatial scales [23]. Trend analysis is difficult due to the rarity of such events and the limited availability of suitable observational data, which is often too short, too localised, or too homogeneous [23, 30]. The imbalance between extreme and non-extreme samples further limits model performance [23]. Current datasets do not contain sufficient representations of extreme events, nor are they always structured in a way that supports robust prediction. While individual case studies can provide insight, they are difficult to generalise without introducing bias, over-emphasising event-specific characteristics, or misidentifying underlying mechanisms [30]. As such, large-scale datasets that contain diverse samples of extreme events are required for reliable evaluation and forecasting [30].

Chaotic Datasets. There have also been recent attempts to construct datasets tailored to chaotic systems, which are relevant to atmospheric dynamics. Gilpin [10] introduces a dataset comprising 131 known chaotic dynamical systems, with precomputed uni- and multi-variate time series and annotated mathematical properties. These datasets focus on long time horizons and prediction tasks,

which makes them potentially useful for studying the behaviour of chaotic weather systems.

3.3 Graph Structure

The forecast accuracy of STGNNs is highly dependent on the graph structure [40]. This is partly because the graph structure specifies the relationships between pairs of nodes, and thus represents the spatial element of spatio-temporal graphs.

3.3.1 Nodal Relationships. Graphs model the pairwise relationships between nodes [31]. They can either be directed, where edges specify the direction of the relationship between nodes, or undirected, which denotes a reciprocal relationship between nodes [22]. The graph can also be homogeneous or heterogeneous. The former specifies a graph with the same type of node and edge, whereas the latter denotes a graph with diverse kinds of nodes and edges and depicts multi-relational data structures [22].

3.3.2 Predefined Graph Structure. Graph structures can be defined before training. The predefined adjacency matrix will then be given as input along with the weather observation data [36, 41]. In this case, the graph structure doesn’t need to be learned, and will remain fixed for the entire training duration. For some types of data, the graph structure is inherent or can be approximated relatively easily using prior knowledge [41]. This approach is effective when dealing with static graphs where the structure and node features are stationary [38]. A time series is stationary if its statistical properties (mean, variance, and covariance) remain constant over time [35]. However, most observations of natural phenomena, like weather events, are non-stationary [35]. Additionally, most multivariate time series, like weather observations, don’t have an explicit graph structure, so one will need to be generated using the input data [36, 38].

3.3.3 Automated Graph Structure Generation. Spatial-temporal graphs contain temporal node features that dynamically change over time [38]. The goal is to capture the spatial structure while analysing

the temporal changes. One way to construct this graph is to randomly initialise the adjacency matrix, and then refine it during training [38]. To accomplish this, some distance metric, such as Euclidean distance for spatial proximity or cosine similarity for nodes with similar attributes, can be used to measure the distance between node pairs, where a smaller distance between nodes denotes a strong relationship [14, 38]. Alternatively, the weight of connected nodes can be adjusted based on temporal inputs [36].

3.4 Graph Learning

Once the graph structure has been generated, a graph learning approach can be applied to effectively model the graph by learning the node embeddings. These embeddings capture both the structural relationships between nodes and their associated features. Representations are embedded into the spatial dimension and then recursively updated in the temporal dimension [31]. To achieve this, STGNNs generally include a spatial module and temporal module [14]. The spatial module captures dependencies between nodes based on the graph structure, while the temporal module models how node features evolve over time [14].

3.4.1 Graph Convolution Module.

Message Passing. STGNNs perform graph convolution to exchange information between nodes [14]. This is typically implemented through a message passing mechanism, where each node aggregates information from its local neighbourhood and updates its representation accordingly [15]. At each layer, a node updates its features by combining its own features with those of its neighbours, which allows the model to capture spatial dependencies in the graph. Through repeated application of this process, information propagates across the graph, allowing nodes to incorporate increasingly global contexts. This process is governed by the adjacency matrix A , which encodes the strength and structure of relationships between nodes. Each entry A_{ij} represents the influence of node j on node i , to determine how information flows through the graph.

Adjacency Matrix. In traditional approaches, A is predefined and remains fixed during training, assuming that the underlying spatial relationships are known [22]. However, this assumption does not always hold for spatio-temporal data such as weather, where relationships between locations are complex, dynamic, and not explicitly observable. To address this, more recent approaches allow the adjacency matrix to be learned directly from the data. This is typically achieved by parameterising A using learnable node embeddings and optimising it end-to-end via gradient descent [37]. In this setting, the model is able to uncover hidden spatial dependencies that are not captured by predefined structures, which allows for more flexible and data-driven representations of the underlying system. Within each graph convolution layer, the adjacency matrix determines how features are aggregated across different orders of neighbourhoods [37]. That is, information is not only exchanged between immediate neighbours, but can propagate across multiple hops in the graph through repeated aggregation. While this allows the model to capture increasingly global spatial context, it also introduces challenges such as over-smoothing.

Over-smoothing. Beyond the aforementioned components, STGNNs are kept relatively shallow, often limited to only two-three layers [34, 40]. This is due to the over-smoothing phenomenon, where, as the number of layers increase, the repeated aggregation causes node representations to become increasingly similar [40]. Over time, the features converge, but may do so to such an extent that the feature information becomes overly heterogenous and thus not informative [15, 40].

3.4.2 Temporal Module. The temporal module captures how node features evolve over time, modelling temporal dependencies in the data. This is commonly implemented using one-dimensional (1D) convolutions applied along the time axis, and allows the model to capture short- and long-range temporal patterns efficiently [22]. In practice, temporal convolution operates on sequences of node embeddings produced by the spatial module [22]. Alternative approaches may replace or augment temporal convolutions with recurrent neural networks or attention-based mechanisms, depending on the complexity of the temporal dynamics being modelled [22].

3.5 Evaluation Metrics

To evaluate the performance of forecasting models, appropriate quantitative metrics are required to measure how closely predicted values match true observations. In the context of spatio-temporal weather forecasting, this corresponds to assessing how accurately a model predicts future meteorological variables over a given horizon H across all nodes $v \in V$. Li et al. [17] define several commonly used regression-based metrics for this purpose. Let $y_t^{(i)}$ denote the true value of a target variable at node $i \in \{1, \dots, n\}$ and time step t , and $\hat{y}_t^{(i)}$ the corresponding prediction. Let N denote the total number of evaluated predictions across all nodes and forecast steps.

Mean Absolute Error. Mean Absolute Error (MAE) measures the average magnitude of prediction errors, without considering their direction. It provides a direct indication of how far predictions deviate from true values on average. Lower MAE indicates better performance, as predictions are, on average, closer to the true values.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (2)$$

Root Mean Squared Error. Root Mean Squared Error (RMSE) measures the dispersion of prediction errors. By squaring errors before averaging, it penalises larger deviations more heavily than MAE. Lower RMSE indicates better performance. Compared to MAE, RMSE is more sensitive to large errors. This makes it particularly relevant when evaluating performance on extreme events.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (3)$$

Coefficient of Determination. The Coefficient of Determination (R^2) measures how well the model explains the variance in the observed data relative to a mean-only baseline. A score closer to 1 indicates that the model captures a larger proportion of the variance,

and thus has stronger predictive performance.

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (4)$$

where $\bar{y}$ is the mean of the true values.

Mean Absolute Percentage Error. Mean Absolute Percentage Error (MAPE) measures the average relative error between predictions and true values over the forecast horizon. MAPE expresses error as a percentage, which makes it scale-independent and easier to interpret across different variables. Lower values indicate better performance.

$$MAPE = \frac{100}{H} \sum_{i=1}^H \frac{|y_{T+i} - \hat{y}_{T+i}|}{|y_{T+i}|} \quad (5)$$

Discussion. These metrics evaluate different aspects of model performance. MAE provides a general measure of average error, RMSE emphasises large deviations, R^2 captures how well the model explains variability, and MAPE provides a relative measure of error. For extreme weather forecasting, RMSE is particularly informative due to its sensitivity to large errors, while MAE provides a more stable estimate of overall performance. However, no single metric fully captures model performance, and multiple metrics are typically used together to provide a more comprehensive evaluation.

4 Analysis of STGNNs for Extreme Weather Prediction

The STGNN framework generally aligns well with the structure of meteorological systems. However, while this architectural alignment is theoretically appealing, the extent to which it translates into effective modelling of extreme events remains unclear. At a structural level, STGNNs operate by constructing a graph over spatial locations, learning node embeddings, and then iteratively propagating information through spatial and temporal modules. This pipeline assumes that (i) spatial dependencies can be captured through a graph structure and (ii) temporal evolution can be learned from historical observations. While this holds for smooth and continuous processes, extreme weather events challenge both of these assumptions.

4.1 Graph Learning

Each node has its own climate state, which in turn affects other nodes. This is understood as the principle of state waves, commonly used to describe the fluctuation characteristics of state variables evolving over time in dynamic systems [31]. Weather phenomena are governed by complex, overlapping “state waves” that propagate across space and time. These interactions are not confined to local neighbourhoods (pairwise modelling). Rather, distant regions with similar climatic patterns may exhibit correlated behaviour [36]. Thus, one of the key limitations that emerge across models is their reliance on local, pairwise spatial dependencies. Models such as MTGNN learn adaptive adjacency matrices but still fundamentally engage in pairwise interactions [36]. Although this is computationally efficient, it limits the model’s ability to capture higher-order

interactions that are intrinsic to meteorological systems [31]. As a result, purely local aggregation mechanisms may fail to capture the full structure of the system.

Hypergraph-based approaches, such as MORL, attempt to address this problem by modelling higher-order relationships between nodes [31]. A hypergraph is defined as $G_{hyper} = (V, E, I, X, H)$, where hyperedges connect multiple nodes simultaneously, thus allowing the model to capture non-pairwise interactions between variables [31]. For weather prediction, this means that correlations can be extracted from different weather stations with similar underlying climate patterns [31]. Empirically, this leads to improved performance across several meteorological variables, particularly temperature and humidity. However, this improvement is not uniform. During periods of sharp fluctuations, predictions exhibit systematic deviations. This suggests that increased structural complexity does not necessarily translate to robustness under extreme conditions [31]. Consequently, higher-order learning may introduce sensitivity to unstable patterns. In the context of extreme events, which are inherently rare and volatile, this could lead to over-fitting or failure to generalise.

4.2 Adaptive Graphs and Hidden Dependencies

Graph WaveNet provides a different perspective by learning spatial dependencies entirely from data through a self-adaptive adjacency matrix [37]. This removes the reliance on predefined graph structures and allows the model to uncover hidden spatial relationships. Empirically, Graph WaveNet outperforms both purely temporal models and earlier STGNNs across multiple datasets [3, 37]. This suggests that adaptive graph learning is critical for capturing non-obvious spatial dependencies. However, this also raises an important contradiction when compared to purely temporal models such as N-BEATS.

N-BEATS, despite lacking any explicit spatial modelling, achieves strong performance on chaotic time series through hierarchical decomposition and high-dimensional representations [10, 25]. This suggests that temporal structure alone may be sufficient under certain conditions. Consequently, this puts the value of graph-based spatial modelling in STGNNs into question. This tension motivates the need for comparative analysis, and must be evaluated empirically.

4.3 Data, Bias, and the Limits of Generalisation

Across all models, performance is heavily influenced by the underlying data. Weather datasets such as ERA5 and WeatherBench introduce biases through interpolation and resolution limitations, which are subsequently inherited by the models [18]. This is particularly evident in VRUSTGNN, where under-estimation of extreme temperatures and humidity in the data lead to an increase in prediction errors over time [18]. These biases fundamentally constrain the model’s ability to learn extreme behaviour, particularly over the long term. Since extreme events are under-represented in the data [39], models are trained predominantly on non-extreme conditions, which leads to poor generalisation when extremes occur.

Table 2: Critical comparison of models for weather forecasting with emphasis on extreme event prediction

Model	Key Idea / Architecture	Strengths	Limitations for Extreme Events
Spatio-Temporal Graph Neural Networks (STGNNs)			
Graph WaveNet [37]	Self-adaptive adjacency matrix learned via node embeddings; combines spatial graph convolution with gated temporal convolution	Captures hidden spatial dependencies without predefined graphs; strong performance across datasets; flexible when graph is unknown	Relies on learned correlations which may not generalise under rare/extreme conditions; robustness under chaotic deviations remains unclear
MTGNN [36]	Adaptive graph learning with sampled pairwise node interactions; dilated temporal convolutions for multi-scale temporal modelling	Scalable; strong baseline performance on weather datasets; captures local spatio-temporal dependencies effectively	Limited to pairwise relationships; struggles with non-local/global dependencies critical for large-scale extreme events
STNN [13]	Enhanced spatio-temporal modelling with improved relational learning over MTGNN	Consistent improvements over MTGNN across variables; strong performance on WeatherBench	Performance varies by variable; struggles with spatially sensitive features (e.g., cloud cover) and abrupt local changes
MORL [31]	Hypergraph-based modelling capturing higher-order relationships between nodes; dynamic + static graph learning	Captures complex multi-node interactions; improved performance for temperature and humidity; models multi-scale dependencies	Sensitive to sharp fluctuations; may overfit unstable patterns; reduced performance for highly localised/extreme phenomena
VRUSTGNN [18]	Variable-resolution unstructured graph (non-uniform spatial mesh); combines static and dynamic graphs	Better spatial representation in critical regions; transferable across domains; effective for short-range forecasts	Prediction error grows with lead time; struggles with peaks/troughs; inherits dataset biases (e.g., ERA5 underestimation of extremes)
ST-FiT [16]	STGNN with temporal data augmentation and spatial topology learning for limited training data	Handles missing temporal data; improves generalisation via synthetic temporal augmentation	Dependent on learned temporal structure; effectiveness under extreme/unseen conditions not well established
CloudNine-v2 [12]	Multi-source feature integration with adaptive graph learning; accounts for heterogeneity in meteorological systems	Strong performance across multiple variables; improved generalisation via richer inputs	Increased complexity reduces interpretability; impact on rare extreme events not explicitly evaluated
DNGNN [34]	Deep GNN (12 layers) with noise perturbation module and residual connections	Captures higher-order dependencies; improved robustness to noisy data; mitigates over-smoothing	Increased depth introduces stability challenges; performance on real-world weather extremes remains unclear
Temporal-Only / Non-Graph Models			
N-BEATS [25]	Purely temporal deep learning model using hierarchical decomposition of time series	Strong performance on chaotic time series; robust to noise; no reliance on graph structure	Ignores spatial dependencies; unclear whether temporal modelling alone suffices for spatially distributed extreme events

Additionally, the heterogeneity of meteorological variables further complicates prediction. For example, the cloud cover variable demonstrates significantly higher prediction error compared to humidity due to its spatial sensitivity and dependence on local conditions [13]. This indicates that model performance is not uniform across variables, and that certain phenomena, particularly those that are localised and abrupt like cloud coverage [13], are inherently more difficult to predict.

CloudNine-v2 addresses some of these challenges by incorporating multi-source data and adaptive graph learning, and achieves strong performance across variables [12]. This suggests that richer

input representations and more flexible graph structures may improve generalisation. However, the extent to which this translates to improved extreme event prediction remains to be evaluated.

4.4 Robustness, Noise, and Model Depth

Another critical limitation is model robustness. Real-world meteorological data is inherently noisy, and neural networks are highly sensitive to such noise [34]. Standard STGNNs, which are shallow due to over-smoothing, struggle to capture high-order dependencies while remaining robust. The Deep Noisy Graph Neural Network (DNGNNs) attempts to address this by increasing the depth of the model and introducing noise perturbation mechanisms [34]. This

allows the model to capture higher-order spatio-temporal interactions while improving robustness to noisy data. Empirical results demonstrate improved performance over baseline models. However, increasing depth introduces its own challenges. While DNGNN mitigates over-smoothing through architectural design, the trade-off between depth, stability, and generalisation is unresolved [34]. In the context of extreme events, where data is sparse and noise is significant, robustness is a central requirement.

4.5 Predictability Horizon and Chaotic Constraints

A fundamental limitation of all data-driven models arises from the chaotic nature of weather systems. Chaotic systems exhibit sensitive dependence on initial conditions, meaning that small errors grow exponentially over time [10]. This directly impacts the predictability horizon of any forecasting model. The concept of prediction horizon formalises this limitation by measuring the time until prediction error exceeds a threshold [42]. Empirical observations from models such as VRUSTGNN show that predictions remain accurate over short lead times (e.g. 1-3 hours), but degrade significantly beyond this range [18]. This aligns with theoretical results from chaos theory, where the largest Lyapunov exponent correlates with forecast error growth [10]. Additionally, the notion of shadowability highlights that even when numerical simulations approximate true trajectories locally, they may diverge globally over longer time spans [?]. This implies that long-term predictions of chaotic systems may not correspond to any true system trajectory, regardless of model accuracy.

Ensemble methods partially address this by quantifying uncertainty through multiple simulations [28]. However, the non-linear relationship between initial conditions and forecast error limits their effectiveness.

These results suggest that the limitations of STGNNs are not purely architectural, but are fundamentally tied to the physics of the system. As such, improving model design alone may not be sufficient to extend predictability horizons.

4.6 Characterisation of Extreme Events for STGNN Modelling

To forecast extreme events using STGNNs, the model must incorporate specific characteristics that define the "extremity" of weather phenomena. Ren et al. [27] describe three properties for this characterisation: magnitude, duration, and extent.

- **Magnitude:** This reflects the departure from a predefined threshold or baseline. This will allow the model successfully distinguish between a warm spell and a heatwave, as the latter is defined by the highest values observed during the year.
- **Duration:** This is defined as the time that has elapsed between the start and end of an event at some station. The overall duration is the time that has elapsed between the start of an event at its initial station until the end of the event at its final station.

- **Extent:** This represents the geographical area where the entire event was recorded.

Furthermore, Fathi et al. [7] and Li and Liu [19] identify four inherent weather prediction characteristics, namely that weather prediction is continuous, data-intensive, multidimensional, and chaotic, that must be considered when selecting or designing a weather forecasting model. These characteristics impose specific requirements on STGNN forecasting architectures. For instance, capturing large spatial extent requires the model to aggregate information across distant nodes, while modelling duration requires sufficiently large temporal receptive fields. However, standard STGNNs tend to be shallow due to over-smoothing constraints [34], which limits their ability to capture long-range spatial dependencies. Thus, there exists a tension between architectural constraints and the physical scale of extreme events. Furthermore, because chaotic nature of weather systems are highly sensitive to initial conditions, their predictive capacity may degrade rapidly over time even if STGNNs successfully learn spatio-temporal dependencies [7]. Future research could evaluate novel state-of-the-art STGNNs to see whether they incorporate these characteristics or can be adjusted to do so, and whether that increases forecast accuracy.

5 Conclusion

Taken together, the literature indicates that STGNNs are effective at modelling short-term spatio-temporal dependencies in weather systems, particularly under stable conditions. Adaptive graph learning, multi-source data integration, and higher-order relational modelling all contribute to improved performance. However, their effectiveness diminishes when applied to extreme events. This is due to a combination of factors: data scarcity, bias inheritance, architectural limitations, and the inherent unpredictability of chaotic systems. Models struggle to capture abrupt, localised phenomena and exhibit increasing error with longer prediction horizons.

This motivates two key directions for further investigation. First, a systematic evaluation of STGNN architectures is required to determine which design choices—such as adaptive graphs or higher-order modelling—are most effective for extreme event prediction. Second, the concept of predictability horizon should be explicitly incorporated into evaluation frameworks, allowing models to be assessed not only on accuracy, but on the temporal limits of their predictive capability. In this context, extending prediction horizon analysis using approaches analogous to Lyapunov exponents provides a promising direction. By quantifying when and how predictions diverge, it becomes possible to characterise the limits of model reliability, particularly for extreme events. Ultimately, the question is not only how accurately STGNNs can predict weather, but how far into the future those predictions remain meaningful.

References

- [1] Roberto Bove, Vinicio Pelino, and Luigi De Leonibus. 2006. Complexity in rainfall phenomena. *Communications in Nonlinear Science and Numerical Simulation* 11, 6 (2006), 678–684.
- [2] Gustau Camps-Valls, Miguel-Ángel Fernández-Torres, Kai-Hendrik Cohrs, Adrian Höhl, Andrea Castelletti, Aytac Pacal, Claire Robin, Francesco Martinuzzi, Ioannis Papoutsis, Ioannis Prapas, et al. 2025. Artificial intelligence for modeling and

- understanding extreme weather and climate events. *Nature Communications* 16, 1 (2025), 1919.
- [3] Mikhail Davidson and Deshendra Moodley. 2022. ST-GNNs for weather prediction in South Africa. In *Southern African Conference for Artificial Intelligence Research*. Springer, 93–107.
 - [4] Lin Ding, Yulong Bai, Donghai Zheng, Xiaoduo Pan, Manhong Fan, and Xin Li. 2025. Chaotic climate system forecasting using an improved echo state network with sparse observations. *Science China Earth Sciences* 68, 7 (2025), 2346–2360.
 - [5] William Ditto and Toshinori Munakata. 1995. Principles and applications of chaotic systems. *Commun. ACM* 38, 11 (1995), 96–102.
 - [6] Lorenzo Escot Mangas and Julio E Sandubete Galán. 2020. A brief methodological note on chaos theory and its recent applications based on new computer resources. *Revista ENERGIA Sociedad Ibero-Americana de Metodología Económica* (2020).
 - [7] Marzieh Fathi, Mostafa Haghi Kashani, Seyed Mahdi Jamei, and Ebrahim Mahdipour. 2022. Big data analytics in weather forecasting: A systematic review. *Archives of Computational Methods in Engineering* 29, 2 (2022), 1247–1275.
 - [8] Armin Fuchs. 2013. *Nonlinear dynamics in complex systems*. Springer.
 - [9] Adeb Gaibie, Hamza Amir, Irene Nandutu, and Deshendra Moodley. 2024. Predicting and Discovering Weather Patterns in South Africa Using Spatial-Temporal Graph Neural Networks. In *Southern African Conference for Artificial Intelligence Research*. Springer, 144–160.
 - [10] William Gilpin. 2021. Chaos as an interpretable benchmark for forecasting and data-driven modelling. *arXiv preprint arXiv:2110.05266* (2021).
 - [11] Stephen J Guastello, Matthijs Koopmans, and David Pincus. 2008. *Chaos and complexity in psychology: The theory of nonlinear dynamical systems*. Cambridge University Press.
 - [12] Hyeon-Ju Jeon, Jeon-Ho Kang, In-Hyuk Kwon, and O-Joun Lee. 2025. Discovering Spatial Correlations of Earth Observations for Weather Forecasting by using Graph Structure Learning. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*. 1–8.
 - [13] Junzhong Ji, Jing He, Minglong Lei, Muhua Wang, and Wei Tang. 2024. Spatio-temporal transformer network for weather forecasting. *IEEE Transactions on Big Data* 11, 2 (2024), 372–387.
 - [14] Ming Jin, Huan Yee Koh, Qingsong Wen, Daniele Zambon, Cesare Alippi, Geoffrey I Webb, Irwin King, and Shirui Pan. 2024. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. *IEEE transactions on pattern analysis and machine intelligence* 46, 12 (2024), 10466–10485.
 - [15] Nicolas Keriven. 2022. Not too little, not too much: a theoretical analysis of graph (over) smoothing. *Advances in Neural Information Processing Systems* 35 (2022), 2268–2281.
 - [16] Zhenyu Lei, Yushun Dong, Jundong Li, and Chen Chen. 2025. St-fit: Inductive spatial-temporal forecasting with limited training data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12031–12039.
 - [17] Delin Li, Zhiqiang Chen, Gang Huang, Jianjun Xu, and Yu Zhang. 2026. A variable-resolution unstructured spatio-temporal graph neural network: an application to very short-range weather forecasting in Guangdong, China. *Geoscience Letters* 13, 1 (2026), 11.
 - [18] Delin Li, Zhiqiang Chen, Gang Huang, Jianjun Xu, and Yu Zhang. 2026. A variable-resolution unstructured spatio-temporal graph neural network: an application to very short-range weather forecasting in Guangdong, China. *Geoscience Letters* 13 (2026). <https://api.semanticscholar.org/CorpusID:285403140>
 - [19] Kan Li and Yu-Shu Liu. 2005. A rough set based fuzzy neural network algorithm for weather prediction. In *2005 international conference on machine learning and cybernetics*, Vol. 3. IEEE, 1888–1892.
 - [20] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. 2024. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 6555–6565.
 - [21] Edward Lorenz. 1963. Deterministic non-periodic flow. *J. Atmos. Sci.* 20 (1963), 130–141.
 - [22] Sozan S Maghddid, Tarik A Rashid, and Shavan K Askar. 2025. Graph neural networks: historical backgrounds, present revolutions, and conventionalization for the future. *International Journal of Data Science and Analytics* 20, 6 (2025), 5237–5299.
 - [23] Stefano Materia, Lluís Palma García, Chiem van Straaten, Sungmin O, Antonios Mamelakis, Leone Cavicchia, Dim Coumou, Paolo de Luca, Marlene Kretschmer, and Markus Donat. 2024. Artificial intelligence for climate prediction of extremes: State of the art, challenges, and future perspectives. *Wiley Interdisciplinary Reviews: Climate Change* 15, 6 (2024), e914.
 - [24] Mu Mu, Wansuo Duan, and Xiaohao Qin. 2025. Disillusionment and rebirth of deterministic weather forecasting and climate prediction: A perspective on the Lorenz’s chaos theory. *Science China Earth Sciences* 68, 10 (2025), 3407–3417.
 - [25] Boris N Oreshkin, Dmitri Carpo, Nicolas Chapados, and Yoshua Bengio. 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437* (2019).
 - [26] Nian Ran, Peng Xiao, Yue Wang, Wesley Shi, Jianxin Lin, Qi Meng, and Richard Allmendinger. 2024. Hr-extreme: A high-resolution dataset for extreme weather forecasting. *arXiv preprint arXiv:2409.18885* (2024).
 - [27] Fumin Ren, Dshumanta Pattanaik, Randall Cervený, Blair Trewin, Andreas Walter, José Luis Stella, Boris Sherstyukov, Kouakou Bernard Dje, Panmao Zhai, Maria de los Milagros Skansi, Ahira Sánchez-Lugo, Ardhana Sopaheluwakan, Mirjana Ivanov, Manola Brunet, Fatima Driouech, Maya Körber, Kiyotoshi Takahashi, John Kennedy, Jessica Blunden, Freja Vamborg, Atsushi Goto, Stefan Rösner, and Peter Bissoli. 2023. Guidelines on the definition and characterization of extreme weather and climate events. *World Meteorological Organization, Geneva, Switzerland* (2023).
 - [28] Sebastian Scher and Gabriele Messori. 2018. Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society* 144, 717 (2018), 2830–2841.
 - [29] Marion Schroedter-Homscheidt and Stefan Wilbert. 2017. Methods to provide meteorological forecasts for optimum CSP system operations. In *The Performance of Concentrated Solar Power (CSP) Systems*. Elsevier, 253–281.
 - [30] Jana Sillmann, Thordis Thorarindottir, Noel Keenlyside, Nathalie Schaller, Lisa V Alexander, Gabriele Hegerl, Sonia I Seneviratne, Robert Vautard, Xuebin Zhang, and Francis W Zwiers. 2017. Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather and climate extremes* 18 (2017), 65–74.
 - [31] Yuming Su, Tinghui Ma, Huan Rong, Baobao Pan, Xuejian Huang, Li Jia, and Mohamed Magdy Abdel Wahab. 2026. Mixed-order relation learning spatio-temporal graph neural network for weather forecasting. *Expert Systems with Applications* 300 (2026), 130184.
 - [32] Xiaoni Sun, Jiming Li, Zhiqiang Zhao, Guodong Jing, Baojun Chen, Jinrong Hu, Fei Wang, and Yong Zhang. 2025. Utility of graph neural networks in short-to medium-range weather forecasting. *Computers, Materials, & Continua* 84, 2 (2025), 2121.
 - [33] Philip Duncan Thompson. 1957. Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus* 9, 3 (1957), 275–295.
 - [34] Songbo Wang, Yifan Chen, Haizhi Xu, Jianqiao Hu, Peiyu Zeng, and Ziming Hu. 2025. DNGNN: Efficient Deep Noisy Graph Neural Network for Spatio-Temporal Series Forecasting. In *2025 2nd International Conference on Algorithms, Software Engineering and Network Security (ASENS)*. IEEE, 15–18.
 - [35] Annette Witt, Jürgen Kurths, and A Pikovsky. 1998. Testing stationarity in time series. *physical Review E* 58, 2 (1998), 1800.
 - [36] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. 2020. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 753–763.
 - [37] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 2019. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121* (2019).
 - [38] Feng Xia, Ciyuan Peng, Jing Ren, Fahim Gozi Febrinanto, Renqiang Luo, Vidya Saikrishna, Shuo Yu, and Xiangjie Kong. 2026. Graph learning. *Foundations and Trends® in Signal Processing* 19, 4 (2026), 362–519.
 - [39] Manzhou Yu, Qunying Huang, and Zhenlong Li. 2024. Deep learning for spatiotemporal forecasting in Earth system science: a review. *International Journal of Digital Earth* 17, 1 (2024), 2391952.
 - [40] Huijun Zhang, Yaxin Liu, Chongyu Zhang, and Ningyun Li. 2025. Machine learning methods for weather forecasting: A survey. *Atmosphere* 16, 1 (2025), 82.
 - [41] Qi Zhang, Jianlong Chang, Gaofeng Meng, Shibiao Xu, Shiming Xiang, and Chunhong Pan. 2019. Learning graph structure via graph convolutional networks. *Pattern Recognition* 95 (2019), 308–318.
 - [42] Yong Zhuang, Matthew Almeida, Wei Ding, Shafiqul Islam, Zihan Li, and Ping Chen. 2025. Horizon Forcing: Improving the Recurrent Forecasting of Chaotic Systems. *ACM Transactions on Intelligent Systems and Technology* 16, 5 (2025), 1–22.

Received 23 March 2026