

Weather Flow Prediction using Spatial-Temporal Graph Neural Networks

Project Proposal

G enevieve Chikwanha*
CHKKAR002@myuct.ac.za
University of Cape Town
Cape Town, South Africa

Zoleka Fuyana*
FYNZOL001@myuct.ac.za
University of Cape Town
Cape Town, South Africa

Praise Jaravani*
JRVPR001@myuct.ac.za
University of Cape Town
Cape Town, South Africa

1 Introduction

Weather prediction involves using historical weather observations to predict some future weather condition [5]. Recent work [3, 5] has shown that Spatial-Temporal Graph Neural Networks (STGNNs) generally perform well on weather prediction tasks. In these models, weather prediction is formulated as a graph representation, where nodes represent the weather stations, and edges encode the local dynamics around a weather station and the global relationships between different stations [3, 5, 8, 14]. However, weather forecasting remains a fundamentally challenging problem, and current STGNNs for weather prediction face a number of limitations. STGNNs have been adapted to address these limitations. However, their effectiveness on normal and extreme weather prediction tasks has not been widely explored, if at all, in the literature. This project attempts to address this gap. To this end, this proposal serves to motivate this research direction and outline how the research will be conducted.

2 Project Motivation

2.1 Related Work

Recent work has increasingly explored STGNNs for weather prediction, motivated by their strong performance in related domains such as traffic flow forecasting and energy grid management. In particular, Davidson and Moodley [3] evaluated the performance of Graph WaveNet (GWN) and Weighted Graph Neural Network (WGN) for maximum temperature prediction across 21 weather stations in South Africa. By reformulating weather prediction as a spatial-temporal graph problem and benchmarking against deep neural network baselines, namely Long Short-Term Memory (LSTM) and Temporal Convolutional Networks (TCN), they showed that GWN outperformed the baselines across various prediction horizons. Building on this foundation, Gaibie et al. [5] extended this work by expanding the number of included South African weather stations and incorporating additional meteorological variables beyond temperature. Their study compared Adaptive Graph Convolutional Recurrent Network (AGCRN), Conditional Local Convolution Recurrent Network (CLCRN), and GWN against a TCN baseline. The results indicated that more recent architectures outperformed GWN, with AGCRN achieving the strongest predictive performance overall.

2.2 Current Limitations

A key gap in both studies, and in most STGNN studies, is that they don't evaluate model performance under extreme weather conditions. Additionally, they do not evaluate how long the models' predictions remain accurate for. This challenge is further formalised by the concept of the Intrinsic Predictability Limit (IPL), which limits reliable weather forecasting to approximately 10-15 days [15, 17]. Small perturbations in initial conditions can lead to disproportionately large deviations in future states [6], which results in prediction errors being amplified. Thus, beyond the IPL, predictions become increasingly unreliable [4, 13]. However, current STGNN-based approaches do not explicitly investigate how prediction accuracy degrades as the forecasting horizon increases. This leaves an important gap in understanding their temporal robustness. Additionally, existing STGNNs introduce modelling constraints. Most approaches adopt a deterministic framework, where future graph states are predicted directly from past observations [3, 5]. This fails to account for the inherent uncertainty in spatio-temporal evolution, which limits the model's ability to represent multiple plausible future trajectories [22]. Furthermore, STGNNs typically operate in discrete time by approximating system dynamics through stepwise updates [9]. This introduces artificial discontinuities that are particularly problematic for weather systems, which evolve as continuous physical processes [9, 25]. From a practical perspective, these challenges are compounded by the computational needs of STGNNs. These models often require a large number of parameters and long training times to achieve competitive performance [27]. This limits their scalability in resource-constrained or real-time forecasting settings. This issue becomes more pronounced as prediction horizons increase, due to the need to model longer-range dependencies [16]. Various STGNN techniques have been introduced to address these shortcomings; however, their ability to accurately forecast normal weather and extreme weather events is not widely explored in the literature.

3 Research Aims and Objectives

3.0.1 Research Aim. This project aims to evaluate and compare the effectiveness of recent STGNN techniques (lightweight, diffusion-based, and continuous-time) for predicting normal and extreme weather events in South Africa.

3.0.2 Research Questions. The research component of this study will be guided by the following research questions:

*All authors contributed equally to this research.

- (1) To what extent do diffusion-based, continuous-time, and lightweight STGNNs improve weather prediction performance compared to baseline STGNN models?
- (2) How effectively do diffusion-based, continuous-time, and lightweight STGNNs predict extreme weather events?
- (3) How long do predictions from different STGNN models remain valid, as measured by Valid Prediction Time (VPT)?
- (4) What trade-offs exist between accuracy, robustness, and computational efficiency across these approaches?

3.0.3 Research Objectives. In order to structure this investigation, we will compare the chosen STGNN techniques against state-of-the-art (SOTA) STGNN baselines. We will measure their predictive performance, using standard metrics, and their predictive limits, using VPT, on weather prediction and extreme weather prediction on South African datasets. Accordingly, the objectives are as follows:

- (1) Implement and standardise an experimental pipeline for training and evaluating baseline and chosen STGNN techniques on South African weather datasets.
- (2) Compare the predictive performance of the chosen STGNN techniques against baseline models using standard evaluation metrics.
- (3) Assess the ability of each STGNN approach to model and predict extreme weather events.
- (4) Evaluate the robustness of each model by measuring how long predictions remain accurate using VPT.
- (5) Analyse the trade-offs between predictive performance, uncertainty modelling, and computational efficiency across the different STGNN approaches.

4 Procedures and Methods

To address our research aims, this section presents the proposed experimental design and associated methodologies.

4.1 Datasets

Our data will be sourced from the South African Weather Services (SAWS), containing data from 70 stations across all nine provinces from the year 2011 to 2025. However, due to the teleconnection phenomenon, which suggests that weather patterns in one region can influence those in another [7, 10], understanding and predicting weather in South Africa requires considering patterns beyond its geographical boundaries. To account for this, we will introduce another dataset, namely the Weather Bench ERA5 dataset, which contains worldwide weather data since 1940. Both datasets contain hourly data on attributes such as temperature, wind speed, humidity, pressure, rainfall, and wind direction.

4.2 Problem Formulation

Given a set of nodes V (where $v \in V$) and a set of edges E (where $e \in E$), we construct a graph $G = (V, E)$ to model weather system dynamics. Each node v represents a weather station and encapsulates the dynamics of a specific location, while edges E capture spatial relationships between stations [3, 5, 20]. Weather observations, along with their associated meteorological features, are represented as multivariate time-series data:

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

where X_t denotes the observation at time step $t \in T$ across all stations, $n = |V|$ is the number of locations, and $d > 1$ is the number of meteorological features [8, 11, 19]. Given a sequence of historical observations X , the aim is to learn a mapping from past observations to future targets over a prediction horizon H , predicting a sequence $y_1, y_2, \dots, y_H$ for each node $v \in V$ [3, 5, 19, 24]. Within this formulation, we consider two prediction tasks.

The problem formulation includes two tasks: normal weather prediction and the more complex task of extreme weather event prediction.

4.2.1 Normal Weather Prediction. The normal weather prediction task is formulated as a regression problem [23]. The goal is to predict continuous-valued meteorological variables (e.g. temperature, rainfall, wind speed, wind direction) over the forecast horizon. Here, the model learns a mapping from historical spatio-temporal observations to future numerical values at each node.

4.2.2 Extreme Weather Event Prediction. In contrast, extreme weather event prediction is formulated as a classification problem where the aim is to determine whether an event is extreme or not [23]. This can be framed as either a binary or multi-class classification task. By using the same input structure (X, A) , the model learns to identify spatio-temporal patterns that precede or signal extreme conditions.

4.3 Preprocessing

Preprocessing is a critical step in preparing the datasets for model training, ensuring that the data is clean, consistent, and suitable for capturing both spatial and temporal dependencies. This involves handling missing values, structuring the data appropriately, and partitioning it for training and evaluation.

4.3.1 Handling missing values. Stations with more than 5% missing data will be excluded to avoid introducing excessive noise into the model [5]. To preserve spatial dependencies in the remaining station data, missing values will be imputed using the Inverse Distance Weighting (IDW) method [18], which estimates values based on neighbouring stations.

4.3.2 Data Segmentation. Following previous work [1, 12], the dataset will be split into training, validation and test sets using a 60:20:20 ratio. Importantly, this split will be performed chronologically rather than randomly.

4.3.3 Anomalous Data. Anomaly-aware preprocessing will be employed. Candidate anomalies will be identified using domain-relevant thresholds and temporal consistency checks. [23]

4.4 STGNN Techniques

This study investigates three categories of STGNNs, namely continuous-time, diffusion-based, and lightweight STGNNs, that have been proposed to address limitations regarding modelling temporal dynamics, uncertainty, and computational efficiency, respectively. Each approach introduces architectural modifications to standard STGNNs

Project Proposal

but still remains compatible with standard spatio-temporal forecasting frameworks. This allows for a consistent and controlled comparison across models.

4.4.1 Continuous-Time STGNNs. These STGNNs discrete updates with neural ordinary differential equations (NODEs) by parameterising the derivative of the latent state rather than the state itself. Although these STGNNs model dynamics continuously in the latent space, they accept standard sequences of discrete observations as input, which are first mapped to a latent representation via a lightweight convolution before the ODE solvers are applied [9]. These models are directly compatible with the regularly sampled hourly station data available from SAWS.

4.4.2 Diffusion-Based STGNNs. These incorporate denoising diffusion processes to model uncertainty in spatio-temporal predictions [22]. Instead of producing a single deterministic output, these models learn a distribution over future states [22, 25]. This allows them to represent multiple plausible outcomes [22, 25]. Although these properties make diffusion-based STGNNs particularly suitable for weather forecasting, they haven't, to the best of our knowledge, been applied to this task yet.

4.4.3 Lightweight STGNNs. These models aim to provide similar performance as standard "non-lightweight" STGNNs with a smaller number of parameters, shorter training times, lower memory requirements compared to standard STGNNs [21]. The models that will be evaluated in this family adopt a lightweight memory-based design [2]. Specifically, they incorporate a memory module with an attention mechanism that enables the model to capture both local and global spatio-temporal dependencies in a continuous manner [2].

4.5 Baseline Models

Our baseline STGNNs have achieved SOTA performance in previous experimentation of STGNN applications in weather prediction [5]. The two baselines that will be used are AGCRN and CLCRN. Since the selected baseline models have already been implemented in previous work, this study will reuse their publicly available source code where possible. Implementation will involve modifying the data-loading pipeline, input features, graph structure, forecasting horizon and output dimensions to match the weather dataset used in this study.

4.5.1 AGCRN. Adaptive Graph Convolutional Recurrent Network (AGCRN) captures fine-grained spatial and temporal correlations in time series data by integrating graph convolution with recurrent units [1]. AGCRN introduces a Data Adaptive Graph Generation (DAGG) module to infer inter-dependencies among nodes directly from data, and a Node Adaptive Parameter Learning (NAPL) module to learn node-specific patterns instead of shared global parameters. These adaptive mechanisms allow the model to automatically construct the graph structure and generate node-specific convolutional parameters, which allows for more expressive modelling of complex spatio-temporal dynamics.

4.5.2 CLCRN. Conditional Local Convolution Recurrent Network (CLCRN) is designed for spatio-temporal meteorological forecasting and models spatial dependencies through a conditional local

convolution kernel that varies across geographic locations [12]. The model constructs local coordinate systems on the sphere and applies a convolution kernel parameterised by a neural network to capture location-characterised meteorological patterns. This conditional convolution is embedded within a recurrent framework to model temporal dynamics, enabling the network to learn continuous, smooth, and spatially varying patterns of meteorological flows.

4.6 Experimental Design

Python and PyTorch will be used for the implementation of each model, following from original implementations of the selected STGNN architectures. Ensuring that the models are reproduced as accurately as possible.

Each model will take as input a sequence of historical spatio-temporal observations, X , represented as a multivariate time series over a graph of weather stations. Given an input window of the previous T time steps, the model will output y_H predictions over a forecast horizon H , either as continuous-valued weather variables (for regression) or as probabilities of extreme events (for classification). Each experiment will be repeated ten times with different initialisations as a way to ensure robustness. The results from these runs will be averaged, and that result will be used as the final prediction. Multiple experiments allow for reliability when performing a comparative study on the model performance. The predicted outputs will be compared against the ground truth values, and evaluation metrics will be computed accordingly. We propose a set of experiments that will be applied consistently across both prediction tasks and all models:

Experiment 1: Comparison against AGCRN. Each model will be evaluated against the AGCRN baseline to assess whether it can achieve improved predictive performance under the same experimental conditions.

Experiment 2: Comparison against CLCRN.

Each model will be evaluated against the CLCRN baseline to assess whether it can achieve improved predictive performance under the same experimental conditions.

4.7 Evaluation Metrics

This section presents the metrics that will be used to assess model performance. The metrics are divided according to regression and classification.

4.7.1 Normal Weather Event Prediction (Regression). To measure predictive accuracy for normal weather forecasting, we use Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE):

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

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

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

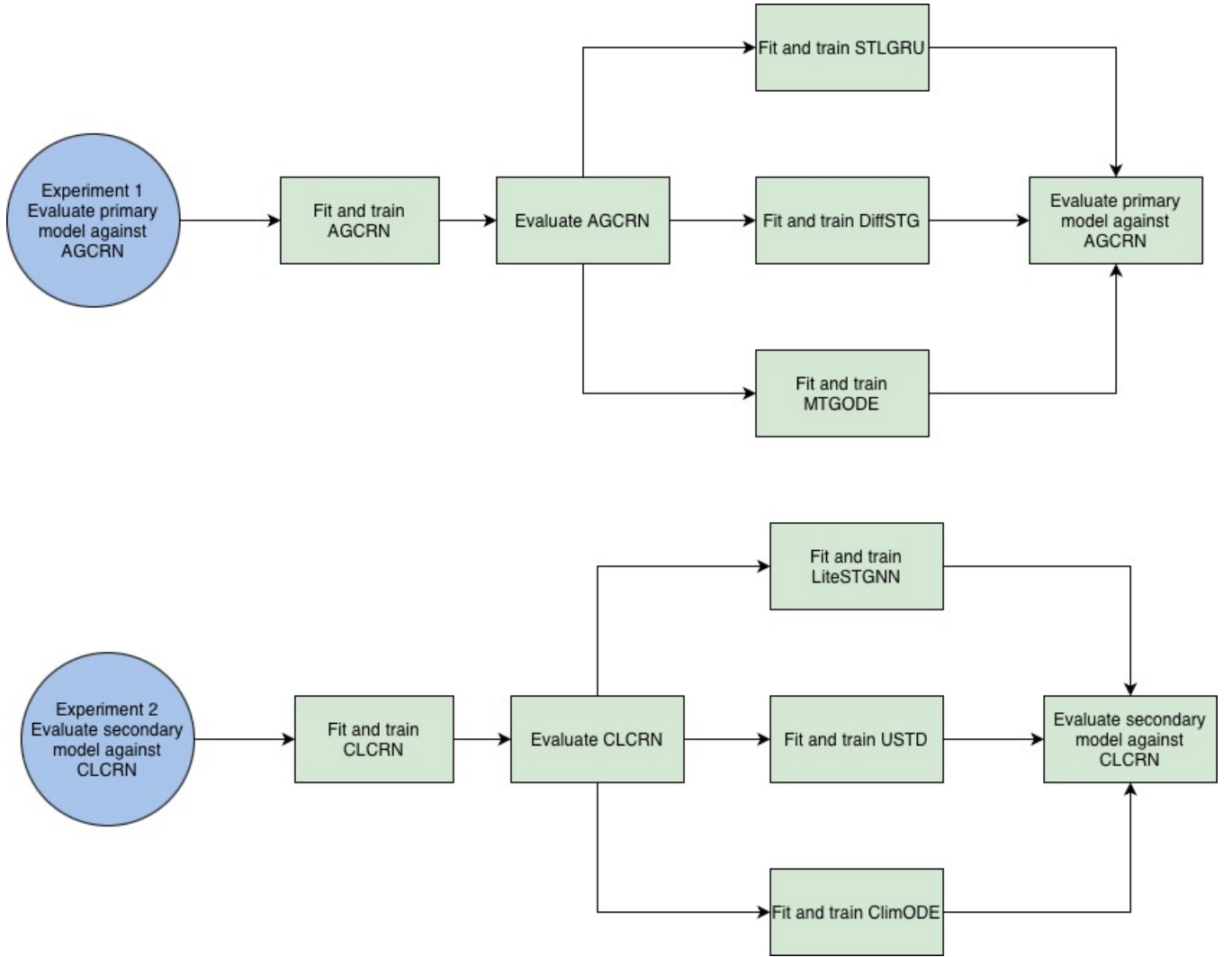


Figure 1: Diagram detailing the project experimental design.

$$\text{MAPE} = \frac{100}{N} \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{|y_i|} \quad (5)$$

$$\text{NRMSE}(t_f) = \sqrt{\frac{1}{N} \sum_{i=1}^N \frac{(y_{t_f,i} - \hat{y}_{t_f,i})^2}{\sigma_i^2}} \quad (6)$$

$$\text{VPT} = \arg \max_{t_f} \{ \text{NRMSE}(t_f) < \epsilon \} \quad (7)$$

where y_i represents the ground truth value, $\hat{y}_i$ the predicted value, and N the total number of observations.

4.7.2 Extreme Weather Event Prediction (Classification). For extreme weather event detection, predictive performance is evaluated using Accuracy, Precision, Recall, and F1-Score, with Weighted Binary Cross-Entropy (WBCE) as the training loss.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

$$\mathcal{L}_{\text{WBCE}} = -\frac{1}{N} \sum_{i=1}^N [w_1 y_i \log(\hat{y}_i) + w_0 (1 - y_i) \log(1 - \hat{y}_i)] \quad (12)$$

where TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively, and w_1 , w_0 are class weights used to address class imbalance.

5 Anticipated Outcomes

This section outlines the expected outcomes of the research project and identifies the key indicators that will be used to evaluate success across the different techniques.

5.1 Research Outcomes

5.1.1 Lightweight STGNNs. We expect lightweight STGNNs to achieve comparable predictive performance to the baseline models while reducing computational cost, memory usage, and training time. In particular, we anticipate that the integration of memory-based attention and gated information filtering will allow the model to retain essential spatio-temporal dependencies without the need for deep or highly parameterised architectures. However, slight performance trade-offs may occur in highly non-linear extreme weather scenarios due to reduced model complexity.

5.1.2 Continuous-Time STGNNs. We expect the continuous-time models to outperform the baselines in capturing fine-grained temporal dynamics. By modelling latent state evolution through the NODEs, these models are expected to produce smoother and more physically consistent forecasts over longer prediction horizons. Additionally, we anticipate improved robustness in modelling irregularly sampled or noisy station data. However, performance gains may depend on the stability of learned continuous trajectories in extreme event conditions.

5.1.3 Diffusion-Based STGNNs. For the diffusion-based models such, we expect improved uncertainty quantification compared to the baselines. In particular, we anticipate that these models will provide more reliable probabilistic forecasts for extreme weather events, where deterministic predictions tend to underperform. Based on prior literature, newer diffusion-based models is expected to outperform both deterministic baselines, while older models, specifically DiffSTG, may not consistently outperform the deterministic models in all forecasting scenarios.

5.1.4 Extreme Event Prediction and VPT Analysis. Across all model families, we expect performance differences to become more pronounced under extreme weather conditions. We hypothesise that diffusion-based and continuous-time models will demonstrate improved robustness in these settings compared to the baselines. Furthermore, we expect the VPT metric to reveal meaningful differences in model stability, with continuous-time models likely maintaining accuracy over longer horizons due to their smoother temporal formulation.

5.2 System Outcomes

The project is expected to produce a fully reproducible and modular experimental framework to evaluate STGNNs on South African weather data. This includes:

- A standardised preprocessing and training pipeline for SAWS station data
- Unified implementation of baseline and proposed STGNN architectures
- A comprehensive evaluation suite
- A benchmarking framework for fair comparison across models

5.3 Impact

5.3.1 Research Contribution. This project is expected to contribute empirical insights into the performance of STGNN architectures on weather prediction tasks. In particular, it will provide evidence on whether architectural adjustments in STGNNs, particularly continuous-time modelling, diffusion-based modelling, and lightweight approaches, lead to measurable improvements over established baselines. Additionally, the use of VPT as a stability metric is expected to provide a useful framework for assessing model robustness beyond standard error-based evaluation. Finally, given that predicting extreme weather events using STGNNs is an understudied topic in the current literature, a key expected contribution is a deeper understanding of how different STGNN techniques perform when predicting extreme weather.

5.4 Key Success Factors

The success of this project will be evaluated based on the following criteria:

- Successful implementation and training of all baseline and proposed STGNN models
- Consistent and reproducible evaluation across all STGNN techniques
- Production of statistically meaningful results across standard and extreme weather conditions
- Demonstrated comparative analysis of lightweight, continuous-time, and diffusion-based approaches
- Successful computation and interpretation of VPT across models

6 Ethical, Legal, and Professional Issues

This project does not involve human or animal subjects and therefore doesn't raise ethical concerns related to experimentation.

Licensing and Use. The software tools we use are open-source and distributed under permissive licenses for unrestricted use and redistribution. We will ensure that all referenced work is properly cited to comply with copyright and intellectual property requirements. Additionally, the use of weather data will align with the terms of use.

Project Release. We do not anticipate any legal barriers to the completion or dissemination of this project. All intellectual property generated through this work, including code and written outputs, will be owned by the University of Cape Town, with copyright in academic outputs vested in the authors, subject to university policies. Upon completion, we will release the project under the Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA 4.0) license, to allow for reuse and adaptation with appropriate attribution.

7 Project Plan

7.1 Work Allocation

- **Praise Jaravani:** Implementation and evaluation of continuous-time STGNNs, and paper writing.
- **Zoleka Fuyana:** Implementation and evaluation of lightweight STGNNs, and paper writing.

- **G enevieve Chikwanha:** Implementation and evaluation of diffusion-based probabilistic STGNNs, and paper writing.
- **Shared Responsibilities:** Data preprocessing pipeline, baseline model implementation, evaluation framework, statistical analysis, and project proposal.

7.2 Project Phases

The project is structured into five phases: a shared Setup phase followed by four sequential research phases of increasing depth and complexity. Phases 1 and 2 constitute the core experimental work, Phase 3 provides analytical depth through an extension task, and Phase 4 covers all writing and dissemination deliverables. A detailed Gantt chart is provided in Appendix A.3.

7.2.1 Setup: Shared Experimental Framework (~4 weeks, 5 May – 2 June 2026). Before any model training begins, the team will establish a shared experimental framework that serves as the foundation for all subsequent phases. This framework functions as a unified pipeline: models are passed in as inputs, and standardised results are returned as outputs, ensuring that all experimental models and baselines are evaluated under identical conditions and that results are directly comparable across all three workstreams.

The Setup phase encompasses the acquisition and preprocessing of both the SAWS station dataset and the ERA5/WeatherBench reanalysis dataset. Preprocessing includes missing value handling via IDW imputation, data normalisation, graph construction from the station network, and chronological data segmentation. The shared evaluation framework will also be implemented during this phase, covering all regression and classification metrics. Baseline models AGCRN and CLCRN will be trained and validated during Setup, establishing the performance benchmarks against which all experimental models will be compared.

By the end of Setup, the full pipeline is operational and validated in a regression setting across both datasets. Table 1 summarises the three technique families and their respective models.

Table 1: Experimental model allocation by technique family. Each team member implements their first model, followed by their second model within the same phase.

Technique Family	First Model	Second Model
Continuous-Time STGNNs	MTGODE	ClimODE
Lightweight STGNNs	STLGRU	LiteSTGNN
Diffusion-Based STGNNs	DiffSTG	USTD

7.2.2 Phase 1: Regression (~6 weeks, 2 June – 14 July 2026). Phase 1 begins immediately from the operational pipeline established in Setup. Each team member trains and evaluates their two models in a regression setting, starting with their first model before moving to their second. Regression predictions of continuous-valued meteorological variables are produced over the forecast horizon across both the SAWS station data and the ERA5/WeatherBench reanalysis data. At the end of Phase 1, a shared comparative analysis is conducted, consolidating regression results across all six models and both baselines, and identifying strengths and limitations of each technique family before moving to the classification reformulation in Phase 2.

7.2.3 Phase 2: Classification Reformulation (~3 weeks, 7 July – 27 July 2026). Phase 2 reformulates the prediction task as a binary classification problem for extreme event detection. No new data is introduced; all preprocessing and graph construction carried out in Setup is reused directly. The same six experimental models are retrained under the classification formulation, with the key methodological change being the adoption of a weighted binary cross-entropy loss function to address the severe class imbalance between extreme and non-extreme weather events. Extreme events are labelled using the 95th percentile threshold applied to the test set, consistent with standard practice in the literature. Classification performance is evaluated using precision, recall, F1-score, and AUC-ROC across all models and both datasets. At the close of Phase 2, a Wilcoxon signed-rank test will be conducted to assess the statistical significance of performance differences across technique families. Phase 2 results will be presented at the Project Progress Demonstration on 27 July 2026.

7.2.4 Phase 3: Extensions and Analysis (~2 weeks, 28 July – 11 August 2026). Phase 3 deepens the analytical scope of the project through a shared extension task and final cross-model analysis. The primary extension is custom loss function exploration for anomaly-sensitive prediction, which is contingent on related work by a PhD student being available in time. If this work is not available, the fallback extension is spatial-temporal dependency and path analysis following Gaibie et al. [5], which examines the spatial relationships learnt by each model through their self-adaptive adjacency matrices. Alongside the extension task, the team will consolidate all experimental results into final cross-model comparative visualisations and summary tables spanning both regression and classification tasks, both datasets, and all six models and two baselines. The outputs of Phase 3 feed directly into the paper writing process which begins at the end of this phase.

7.2.5 Phase 4: Writing and Dissemination (~3.5 weeks writing + dissemination, 10 August – 5 October 2026). Phase 4 covers all writing and dissemination deliverables. Paper writing begins with the scaffold due on 10 August 2026, overlapping with the tail end of Phase 3 to ensure continuity. Each team member is responsible for writing the sections corresponding to their workstream, with shared sections – including the introduction, related work, methodology, and conclusion – written collaboratively. The full draft is due 24 August 2026, followed by peer review, revisions, and final submission on 4 September 2026. Following paper submission, the team will prepare the code repository for final submission on 14 September 2026, deliver the final project demonstration between 16 and 23 September 2026, submit the project poster on 28 September 2026, and complete the project website by 5 October 2026.

7.3 Resources

7.3.1 Data Resources.

- SA Weather Service station network data (70 stations)
- WeatherBench global reanalysis dataset (ERA5)

7.3.2 Computational Resources.

- UCT High Performance Computing (HPC) cluster for model training

Project Proposal

- Personal computing resources for development and debugging
- GPU access for neural ODE integration and diffusion model denoising

7.3.3 Software and Tools.

- Python 3.8+ with PyTorch, PyTorch Geometric, torchdiffeq (for neural ODEs)
- Scikit-learn for preprocessing and baseline methods
- Matplotlib, Seaborn for visualisation
- Git/GitHub for version control and code sharing
- LaTeX for paper writing

7.4 Risks

Nine risk categories have been identified, spanning data quality, extreme event scarcity, model convergence, computational constraints, WeatherBench integration complexity, neural ODE instability, custom loss function dependency, baseline performance, and timeline delays. Each risk assessment is detailed in the Appendix A.1.

7.5 Deliverables

The project involves a number of deliverables, including:

- Three literature reviews (completed)
- A project proposal draft and presentation
- A project proposal final paper
- Three final paper drafts
- Three source code repositories with documentation
- Three final papers
- Three project visualisations (interactive dashboards or static plots)
- A project demonstration
- A project poster
- A project website

8 AI Usage Statement

Generative AI tools, specifically ChatGPT and Gemini Pro, were used in a limited capacity for this submission, in line with the guidelines provided in class. AI was primarily used for feedback on structure, clarity, and overall flow. No AI content was copied directly, and all suggestions were critically evaluated before being incorporated, if at all. At no point was AI used to generate original content for this proposal.

References

- [1] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems* 33 (2020), 17804–17815.
- [2] Kishor Kumar Bhaumik, Fahim Faisal Niloy, Saif Mahmud, and Simon S Woo. 2024. STLGRU: Spatio-temporal lightweight graph GRU for traffic flow prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 288–299.
- [3] Mikhail Davidson and Deshendran 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] Adeeb Gaibie, Hamza Amir, Irene Nandutu, and Deshendran 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.
- [6] Stephen J Guastello, Matthijs Koopmans, and David Pincus. 2008. *Chaos and complexity in psychology: The theory of nonlinear dynamical systems*. Cambridge University Press.
- [7] Andrew Hoell, Chris Funk, Tamuka Magadzire, Jens Zinke, and Greg Husak. 2015. El Niño–Southern Oscillation diversity and southern Africa teleconnections during austral summer. *Climate Dynamics* 45, 5 (2015), 1583–1599.
- [8] 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.
- [9] Ming Jin, Yu Zheng, Yuan-Fang Li, Siheng Chen, Bin Yang, and Shirui Pan. 2022. Multivariate time series forecasting with dynamic graph neural odes. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2022), 9168–9180.
- [10] Mark R Jury. 1996. Regional teleconnection patterns associated with summer rainfall over South Africa, Namibia and Zimbabwe. *International Journal of Climatology: A Journal of the Royal Meteorological Society* 16, 2 (1996), 135–153.
- [11] 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.
- [12] Haitao Lin, Zhangyang Gao, Yongjie Xu, Lirong Wu, Ling Li, and Stan Z Li. 2022. Conditional local convolution for spatio-temporal meteorological forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 7470–7478.
- [13] Edward Lorenz. 1963. Deterministic non-periodic flow. *J. Atmos. Sci.* 20 (1963), 130–141.
- [14] Sozan S Maghdid, 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.
- [15] 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.
- [16] Henok Tenaw Moges and Deshendran Moodley. 2025. A lightweight Spatial-Temporal Graph Neural Network for Long-term Time Series Forecasting. *arXiv preprint arXiv:2512.17453* (2025).
- [17] 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.
- [18] Mina Sadeghi, Mohammad Karimi, Hamidreza Rabiei-Dastjerdi, and Dipto Sarkar. 2026. Multi-scale dynamic population estimation: an Adaptive Inverse Distance Weighting (AIDW) model incorporating spatial characteristics. *Cartography and Geographic Information Science* 53, 1 (2026), 129–154.
- [19] Yuming Su, Tinghuai 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.
- [20] 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.
- [21] Dongjing Wang, Gangming Guo, Tianpei Ouyang, Dongjin Yu, Haiping Zhang, Bao Li, Rong Jiang, Guandong Xu, and Shuiguang Deng. 2025. A lightweight spatio-temporal neural network with sampling-based time series decomposition for traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems* (2025).
- [22] Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2023. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. In *Proceedings of the 31st ACM international conference on advances in geographic information systems*. 1–12.
- [23] E Williams. 2022. Predicting anomalous weather events using supervised machine learning... Faculty of Science, Department of Computer Science.
- [24] 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.
- [25] Chenchen Yang, Wenbing Zhang, and Zhou Yingjiang. 2025. An Overview of Spatiotemporal Network Forecasting: Current Research Status and Methodological Evolution. *Mathematics* 14, 1 (2025), 18.
- [26] Heeyong Yoon, Kang-Wook Chon, and Min-Soo Kim. 2025. SPATS: a practical system for comparative analysis of spatio-temporal graph neural networks. *Cluster Computing* 28, 13 (2025), 826.

- [27] Heeyong Yoon, Jinhong Jung, Kang-Wook Chon, and Min-Soo Kim. 2025. Search-Light: Neural Architecture Search for Lightweight Spatio-Temporal Graph Neural Networks. *IEEE Access* (2025).

A Appendix

A.1 Risk Assessment Matrix

Table 2 presents a comprehensive risk assessment for the project. STGNNs require long training times due to their large parameter counts [26], which is reflected in several of the risks below.

A.2 Evaluation Metrics

Table 3 provides a full summary of the evaluation metrics used across both task formulations.

A.3 Project Timeline

Figure 2 shows the detailed project timeline from May to October 2026. The schedule is structured around UCT Honours project weeks, with shared infrastructure tasks front-loaded before individual model workstreams begin. Key milestones are marked in red.

Project Proposal

Table 2: Risk Assessment Matrix

Risk	Prob.	Impact	Consequence	Mitigation	Monitoring	Management
Data quality issues (missing values, station dropouts)	Medium	High	Biased or incomplete training data; models may fail to generalise across stations	IDW imputation; exclude stations with >5% missing data; retain anomalous observations	Weekly data quality checks; verify station count before training	Re-run preprocessing with stricter thresholds; exclude affected stations and re-evaluate
Extreme event scarcity (class imbalance)	High	High	Models default to predicting normal conditions; poor recall and F1 on extreme class	95th percentile thresholding; weighted binary cross-entropy loss; stratified evaluation metrics	Track class distribution per variable; monitor F1 and recall on extreme class	Adjust class weights; re-threshold if extreme class remains underrepresented after initial training
Model convergence failures (neural ODEs, diffusion models)	High	Medium	Unreliable or uninterpretable results; inability to meet Phase 1 deliverables on time	Gradient clipping; alternative ODE solvers; cosine annealing learning rate schedule	Track training and validation loss; cross-check against published benchmarks [5]	Reduce model complexity; switch ODE solver; consult supervisor if convergence cannot be achieved
Diffusion model training time and sample quality	Medium	Medium	Phase 1 first model in each technique results delayed; blurry or oversmoothed probabilistic forecasts	Limit denoising steps during development; tune step count on validation set; apply early stopping	Monitor inference time per sample; evaluate forecast quality on held-out extreme windows	Reduce denoising steps; fall back to fewer diffusion iterations if HPC time is constrained
Computational resource limitations	Medium	High	Inability to train large models within project weeks; Phase 1 and 2 timelines at risk	Request UCT HPC access before June project weeks; optimise batch sizes and ODE step sizes	Monitor HPC queue times weekly; flag if baselines not complete by end of June	Fall back to cloud compute; reduce model depth or batch size to fit available resources
WeatherBench integration complexity (Phase 2)	Medium	Medium	Spatial misalignment between ERA5 grid and SAWS stations produces noisy or invalid results	Familiarise with ERA5 structure in Phase 1; validate spatial alignment before Phase 2 training	Confirm boundary overlap before Phase 2 begins; flag misalignment early	Adjust interpolation strategy; consult WeatherBench documentation and prior work
Station coverage bias affecting extreme event evaluation	Medium	Medium	Aggregated metrics mask poor inland performance; conclusions about extreme events may be geographically misleading	Conduct station-level stratified analysis; do not rely solely on aggregated metrics	Report per-station extreme event metrics; flag coastal vs inland divergence	Separate evaluation by region; discuss geographic limitations explicitly in results
Insufficient performance improvement over baselines	Medium	Low	Research questions remain answerable but findings are less impactful	Project designed to analyse how and why models differ, not only whether they improve; VPT and dependency analysis provide value regardless of result direction	Regular benchmarking against AGCRN and CLCRN from Phase 1 onward	Reframe analysis around interpretability and stability findings; treat negative results as valid contributions
Timeline delays during critical project weeks	Medium	Medium	Phase 1 deliverables slip past Demo #1; reduced time for Phase 2 and 3	Begin preprocessing and baseline setup in April–May so model training starts immediately when June project weeks open	Weekly progress check against Gantt chart in Appendix A.3; flag slip-page if baselines not training by end of June	Reprioritise core experiments; defer Phase 3 extensions if necessary; notify supervisor early
ERA5 data acquisition delays	High	High	Phase 1 start blocked if ERA5 is unavailable before end of May; both datasets are required from the outset of Phase 1	Initiate ERA5 access request immediately after proposal submission; identify alternative download sources (e.g. Copernicus CDS) as backup	Confirm ERA5 access by mid-May; flag immediately if download not initiated by 12 May	Fall back to SAWS-only experiments for Phase 1 if ERA5 is unavailable; reintroduce ERA5 in Phase 2 once acquired
Custom loss function extension dependency (Phase 3)	Medium	Medium	Primary Phase 3 extension cannot proceed if related PhD student work is not available in time; analytical depth of Phase 3 reduced	Identify fallback extension (spatial-temporal path analysis following Gaibie et al. [5]) early; confirm PhD student timeline by end of Phase 2	Check PhD student progress status at start of Phase 3 (28 July 2026); make go/no-go call immediately	Switch to path analysis fallback without delay; ensure fallback scope is scoped and ready to begin by 28 July regardless of primary extension outcome

Table 3: Evaluation Metrics by Task

Task	Metric	Purpose
Regression	MAE	Average prediction error
Regression	RMSE	Penalises large errors
Regression	MAPE	Scale-independent error
Regression	MSE	Training loss reference
Regression	Stratified MAE	MAE computed separately for extreme (>95th percentile) vs non-extreme timesteps
Regression	Peak Intensity Error	MAE on predicted maximum values during extreme event windows
Classification	Accuracy	Overall correctness
Classification	Precision	Correctness of extreme event predictions
Classification	Recall	Coverage of true extreme events
Classification	F1-Score	Harmonic mean of precision and recall; handles class imbalance
Both	VPT	Maximum lead time at which predictions remain reliable



Figure 2: WEATHER-FLOW project timeline, May–October 2026. Colour coding: blue = Setup, green = Phase 1, yellow = Phase 2, orange = Phase 3, red = Phase 4. Red diamonds mark UCT hard deadlines.