

Weather Flow Prediction using Diffusion-Based Spatial-Temporal Graph Neural Networks

Research Paper

Génevieve Chikwanha
CHKKAR002@myuct.ac.za
University of Cape Town
Cape Town, South Africa

Abstract

Spatio-Temporal Graph Neural Networks (STGNNs) have been successfully applied to weather forecasting problems, but the use of diffusion-based STGNNs for weather forecasting remains under explored. In this study, we evaluate whether diffusion-based STGNNs provide an advantage over deterministic approaches for weather prediction in South Africa. We compared two diffusion-based models, DiffSTG and Unified Spatio-Temporal Diffusion (USTD), with two deterministic models, Adaptive Graph Convolutional Recurrent Network (AGCRN) and Conditional Local Convolution Recurrent Network (CLCRN), using hourly observations from 32 South African Weather stations. We evaluated temperature, pressure, humidity, and wind speed across forecast horizons of up to 48 hours using MAE, RMSE, and CRPS. Additionally, we analysed predictive performance across forecast horizons and stations, as well as investigated spatial and elevation-related error patterns. CLCRN achieved the best overall point-forecast performance, while DiffSTG performed worst across the evaluated attributes. USTD performed competitively with AGCRN and achieved lower CRPS than DiffSTG for three of the four attributes. The results suggest that diffusion-based forecasting does not inherently improve weather prediction, furthermore, it highlights the importance of how spatial, temporal, and contextual information is incorporated into the forecasting architecture.

ACM Reference Format:

Génevieve Chikwanha. 2026. Weather Flow Prediction using Diffusion-Based Spatial-Temporal Graph Neural Networks [0.5 em] Research Paper. In *Proceedings of* . ACM, New York, NY, USA, 18 pages.

1 Introduction

Weather is a highly dynamic system in which observations at different locations are influenced by both temporal changes and spatial dependencies. These characteristics make weather forecasting a natural application for Spatio-Temporal Graph Neural Networks (STGNNs), which model temporal dynamics within individual nodes and spatial relationships between nodes [12, 29]. STGNNs have been widely applied to traffic and other flow prediction problems, but their application to weather forecasting remains comparatively under-explored. Previous studies have demonstrated the potential of STGNNs for weather prediction in South Africa. Davidson and Moodley [8] evaluated Graph WaveNet (GWN) and the Low Rank Weighted Graph Neural Network (WGN) for temperature prediction and found GWN to provide the best overall performance. Gaibie et al. [10] subsequently evaluated newer STGNNs, including the

Adaptive Graph Convolutional Recurrent Network (AGCRN) [1] and Conditional Local Convolution Recurrent Network (CLCRN) [16], and showed that these models could successfully predict multiple weather variables. These models provide strong deterministic baselines for evaluating newer approaches to weather forecasting. Recent work has introduced diffusion models for probabilistic spatio-temporal forecasting. DiffSTG [27] combines diffusion-based generation with graph neural networks to model uncertainty and spatio-temporal dependencies, while Unified Spatio-Temporal Diffusion (USTD) [11] extends diffusion-based forecasting with a deterministic spatio-temporal encoder and additional contextual information. Although diffusion models have shown promising results on spatio-temporal forecasting tasks, it remains unclear whether their probabilistic formulation provides an advantage for weather forecasting, particularly on irregularly distributed weather station networks. In this study, we compare DiffSTG and USTD with AGCRN and CLCRN for weather forecasting using hourly observations from 32 SAWS stations across South Africa. We evaluate temperature, pressure, humidity, and wind speed over forecast horizons of up to 48 hours using MAE, RMSE and, CRPS. Our results provide an empirical comparison of deterministic and diffusion-based STGNNs for South African weather forecasting and identify factors that may explain differences in their performance. In summary, the main contributions of this study are as follows:

- We provide an empirical comparison of two diffusion-based STGNNs, DiffSTG and USTD, with the deterministic AGCRN and CLCRN models for weather forecasting using South African weather station data.
- We evaluate both point and probabilistic forecasting performance across four weather variables and forecast horizons of up to 48 hours, using MAE, RMSE, and CRPS.
- We analyse forecasting errors across time and space, including node-level performance and the relationship between forecasting error, station elevation, and weather variability.

The remainder of this paper is structured as follows. Section 2 introduces the problem definition and reviews STGNNs, their application to weather forecasting, and diffusion-based spatio-temporal forecasting. Section 3 describes the dataset, data processing, evaluated models, experimental setup, evaluation metrics, and hyperparameter tuning procedure. Section 4 presents the predictive results and analyses performance across forecast horizons, spatial patterns, weather variability, elevation, and computational cost. Section 5 discusses the limitations of the study and directions for future

work. Finally, Section 6 summarises the main findings and their implications for diffusion-based weather forecasting.

2 Background and Related Work

This section first defines the weather forecasting problem as a spatio-temporal graph learning task. It then introduces the spatial and temporal components of STGNNs, reviews their application to weather forecasting, and concludes by describing diffusion-based approaches to spatio-temporal forecasting.

2.1 Problem Definition

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 $G = (V, E)$ that models the weather system. $v \in V$ represents a weather station and encapsulates the dynamics of that particular location, while E represents the set of relationships between stations in V [8, 10, 25]. Weather observations, and their corresponding meteorological features (such as temperature or humidity), can be represented as time-series data, which is a sequence of observations ordered by time:

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

where X_t is an observation at time $t \in T$ from a particular weather station, $|V|$ is the number of stations, and d is the number of meteorological features [12, 15, 24]. When $d = 1$, the series is univariate; when $d > 1$, it is multivariate [15]. Note that we investigate univariate weather prediction and leave multivariate prediction for future work. The goal is to use a sequence of historical observations, X , from a network of $|V|$ weather stations, the graph G , and an adjacency matrix $A \in \mathbb{R}^{|V| \times |V|}$ representing the strength and structure of the relationships between stations in V , 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 [8, 10, 24, 30]. Thus, the model must learn some function F that maps the true observed history to future signals [16].

2.2 Spatial Temporal Graph Neural Networks

STGNNs have been widely applied to forecasting problems such as traffic flow [14, 31], energy demand [18], air quality [11, 27], and increasingly, weather [8, 10]. They are effective at modelling data that is dependent on both its spatial relationships and its evolution over time [12, 29]. Traditional time-series approaches generally treat observations as independent sequences and therefore do not explicitly model spatial dependencies, while standard convolutional approaches assume a regular spatial structure, which many natural phenomena do not have [5, 29]. STGNNs address this by representing the data as a graph so that the model can learn both spatial dependencies between nodes and temporal dependencies within each node's sequence [5, 34]. STGNNs therefore consist of a spatial module and a temporal module, although the way these components are combined differs between architectures.

2.2.1 Spatial Module. Graph convolutional networks (GCNs) are commonly used as the spatial component [5, 12, 29]. GCNs aggregate information from a node's neighbours according to the strength and structure of its relationships with other nodes specified by the adjacency matrix so information can be propagated through

the graph [5, 29]. Other spatial modules include graph attention networks, which learn different weights for neighbouring nodes, diffusion-based graph convolutions, which model information propagation through directed graphs, and other more specialised graph convolutions designed for particular spatial structures [14, 16, 26].

2.2.2 Temporal Module. The temporal component is often implemented using recurrent neural networks (RNNs) and its variants, temporal convolutional networks (TCNs), or, more recently, attention-based architectures [5, 29]. RNN-based approaches maintain a hidden state that is updated as successive time steps are processed so that information from previous time steps influence later predictions. Gated recurrent units (GRUs) are a commonly-used RNN variant because their gating mechanism controls how information is retained and updated over time, while requiring less parameters than some other RNN variants [5, 29]. TCNs use one-dimensional convolutions over the temporal dimension. Dilated convolutions can increase the temporal receptive field without requiring the model to process each time step sequentially, which can make them more efficient for modelling longer temporal dependencies [31, 34].

2.3 STGNNs for Weather Prediction

STGNNs have increasingly been applied to weather forecasting, where observations are collected from irregularly-spaced locations and have complex spatial and temporal dependencies. Lin et al. [16] introduced CLCRN, which uses geographic coordinates to condition local graph convolutions on the relative distance and orientation between weather stations. CLCRN achieved state-of-the-art performance on the evaluated weather forecasting benchmarks. This showed the value of incorporating geographic information into the spatial modelling process. Davidson and Moodley [8] subsequently evaluated GWN and a Low-Rank Weighted Graph Neural Network for maximum-temperature forecasting across 21 South African weather stations, with GWN achieving the best average performance. Gaibie et al. [10] later compared AGCRN, CLCRN, TCN, and GWN across 45 South African stations and four weather variables. AGCRN achieved the best overall performance, although its advantage over the second-best model decreased from 24.23% at a 3-hour horizon to 4.66% at 24 hours. More recent work has continued to adapt the spatial structure of STGNNs to weather forecasting. Li et al. [13], for example, used an unstructured, variable-resolution graph with both static and dynamic spatial dependencies for very short-range forecasting which achieved high prediction accuracy for temperature and pressure but increasing errors for wind and humidity with lead time. These studies demonstrate that STGNNs can effectively model spatial and temporal dependencies in weather observations, but they primarily evaluate deterministic point forecasts. This is a limitation for weather forecasting because a single prediction does not represent the range of plausible future atmospheric states. The importance of representing forecast uncertainty is particularly evident for extreme events, where accurate prediction of the timing and spatial evolution of an event does not necessarily imply accurate prediction of its intensity. Li et al. [13], for example, found that their model reproduced the evolution and spatial structure of an extreme heatwave and Typhoon Saola but

underestimated peak temperatures, maximum wind speeds, and minimum sea-level pressures.

2.4 Diffusion-Based Spatio-Temporal Graph Forecasting

The deterministic nature of conventional STGNN forecasting motivates probabilistic approaches that model a distribution over possible future states rather than a single point prediction. Denoising diffusion models achieve this by defining a forward process that progressively adds noise to observed future data until it approaches a noise distribution, and then learning a corresponding reverse process that removes this noise to reconstruct plausible future samples [23, 27]. During forecasting, the reverse process starts from random noise and iteratively denoises it, conditioned on the available observations and other contextual information. Repeating this process allows the model to generate multiple possible future trajectories, which represents uncertainty in the forecast. TimeGrad introduced denoising diffusion for probabilistic multivariate time-series forecasting to generate future observations autoregressively from a distribution conditioned on previous observations [21]. Subsequent work explored conditional diffusion for time-series forecasting and imputation [19], including non-autoregressive forecasting approaches [23]. However, these methods do not explicitly model the spatial dependencies represented by STGNNs. DiffSTG extended denoising diffusion to spatio-temporal graphs, combining a diffusion model with an STGNN-based denoising network to generate multiple possible future trajectories non-autoregressively [27]. It reduced CRPS by 4 to 14% and RMSE by 2 to 7% relative to the evaluated probabilistic baselines across three real-world datasets. USTD subsequently proposed a unified diffusion framework for multiple spatio-temporal graph learning tasks, using a shared spatio-temporal encoder with task-specific probabilistic decoders for forecasting and kriging [11]. Other approaches have since explored alternative diffusion architectures, including dynamics-informed diffusion in DYffusion [4] and spectral diffusion in SpecSTG [17]. However, these methods have predominantly been evaluated on non-meteorological spatio-temporal datasets, while weather-specific diffusion-based STG forecasting remains comparatively limited. Recent work such as STGLDWeather demonstrates the potential of combining multi-scale STGNNs with latent diffusion for weather forecasting, but further evaluation is needed to establish how diffusion-based STGNNs compare with established deterministic weather STGNNs across different weather variables and forecast horizons [28].

3 Methodology and Experimental Design

This section describes the experimental design used to compare deterministic and diffusion-based STGNNs. It first describes the SAWS dataset and its preprocessing, then introduces the four evaluated models, followed by the implementation and evaluation procedures and the hyperparameter tuning strategy.

3.1 Dataset

We performed our experiment using a dataset from the South African Weather Service (SAWS) containing weather observations from 60 irregularly-spaced automatic weather stations distributed

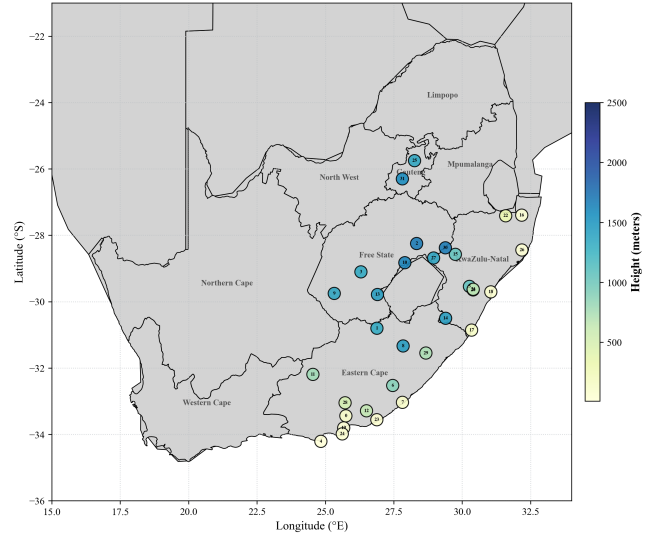


Figure 1: A map showing the region covered in the SAWS dataset, along with each station.

Table 1: Statistics of the SAWS (2016-2025) dataset.

Property	Temperature	Humidity	Pressure	Wind Speed
Unit	°C	%	hPa	m/s
Mean	17.35	67.07	927.05	2.68
Maximum	45.60	100.00	1050.00	22.80
Minimum	-10.20	1.00	800.00	0.00
Std. Dev.	6.83	1.00	63.41	2.21
Missing (%)	4.25	5.25	4.75	5.29

across four South African provinces, namely, Free State, Gauteng, KwaZulu-Natal, and the Eastern Cape. Each station contains hourly observations spanning 10 years, from 2016 to 2025, for six variables: humidity, pressure, rainfall, temperature, wind direction, and wind speed. The dataset contains additional metadata about each station, such as the station name, its longitude and latitude, and its height. The dataset does not contain personally identifiable information, was used in accordance with the SAWS data policy, and is publicly available upon request from the SAWS. A list of the stations requested is in the Appendix B.2 and statistics on the dataset can be found in Table 1.

3.2 Data Processing and Partitioning

To process this data for training and testing, we followed Davidson et al. [8] and Gaibie et al. [10].

3.2.1 Handling Missing Data. First, the 60 stations were filtered using a 10% missing data threshold [8, 33]. Because 28 of the 60 stations had more than 10% missing data across all variables, they were excluded from the study. The remaining 32 stations (5 from Free State, 2 from Gauteng, 12 from KwaZulu-Natal, and 13 from the Eastern Cape), shown in Figure 1, were then imputed using inverse distance weighting (IDW) interpolation to fill in the missing

data. This imputation method estimates missing values at a target station by calculating a distance-weighted average of concurrent observations from neighbouring stations [22]. This technique assumes that nearby stations are more related than distant stations, and thus stations closer to the target station influence its estimated data points more than those further away [22]. IDW is suitable for time-series data and is simple to implement [2], thus we found it suitable for this study. For this study, we only focus on four of the six features, excluding rainfall and wind direction, in line with Gaibie et al. [10].

3.2.2 Partitioning Data. After filtering and imputing, we had 87 672 hourly observations for each of the four features across the 32 stations, totalling 11 222 016 observations. These were used to construct an input matrix with dimensions (Timestamps, Stations, Features) = (87 672, 32, 4). We trained the models on each feature separately, as such, we sliced the matrix along the third dimension for each feature, so the dimensions became (Timestamps, Stations, Feature) = (87 672, 32, 1). We used the hold-out validation method to split the data into a 60% : 20% : 20% ratio for training, validating, and testing, respectively [9]. In line with other STGNN studies [8, 10, 11, 27], this split was applied chronologically to maintain the temporal dependencies of the data. The training set corresponded to the first six years (2016-2021), while the validation set was represented by the next two years (2022-2023) and the test set by the final two years (2024-2025).

3.2.3 Processing and Loading Data. Then, following Wen et al. and Hu et al. [11, 27], the input data was normalised using z-score normalisation. The mean and standard deviation were calculated using the training dataset to avoid data leakage from the test set. The processed data was then loaded by each model using the sliding window technique, where an input window of length T was used to predict the next window for a horizon of 48. Given an input window of historical data, the model predicts a non-overlapping but contiguous horizon. The window then slides forward by a single time step, such that the first sample uses observations from $t - T + 1$ to t to predict the subsequent observations from $t + 1$ to $t + h$. This process is repeated throughout the sequence, creating overlapping input samples while maintaining a contiguous prediction horizon. This allows the models to learn temporal patterns from different starting points in the sequence.

3.3 Models

We evaluated four STGNNs, consisting of two deterministic baselines, AGCRN [1] and CLCRN [16], and two diffusion-based models, DiffSTG [27] and USTD [11]. AGCRN and CLCRN were selected as baselines following Gaibie et al. [10], who evaluated these models for weather forecasting and found them to perform successfully in their implementation. This provides established deterministic baselines against which the diffusion-based models can be evaluated. DiffSTG and USTD were selected as the diffusion-based models because they explicitly combine diffusion-based probabilistic forecasting with spatio-temporal modelling, which allows us to investigate whether the benefits of diffusion models observed in other domains extend to weather forecasting.

3.3.1 AGCRN. This model combines graph convolution with recurrent modelling to capture spatial and temporal dependencies. Unlike models that rely on a predefined adjacency matrix, AGCRN learns the spatial relationships between nodes through its Data Adaptive Graph Generation (DAGG) module. Each node is assigned a learnable embedding, from which an adaptive adjacency matrix is generated. These same node embeddings are used by the Node Adaptive Parameter Learning (NAPL) module to generate node-specific graph convolution parameters, allowing different stations to learn different feature transformations. The graph-convolved features are then passed through GRU-style recurrent cells to model temporal dependencies. The final hidden representation is used to generate the multi-step forecast. In our implementation, the historical weather observations were provided as a sequence of node features. AGCRN therefore learned both the spatial graph and node-specific convolution parameters directly from the training data. No predefined adjacency matrix or additional covariates were provided to the model. This allows its learned spatial relationships to be analysed independently of the physical adjacency matrix used by the other models.

3.3.2 CLCRN. It is designed specifically for forecasting on irregularly distributed weather stations. Rather than treating the stations as points on a regular grid, CLCRN uses their geographic relationships to construct a local convolution around each centre node. The convolution is conditioned on the relative distance and orientation between stations, allowing nearby nodes to contribute differently depending on their spatial relationship to the centre node. The resulting spatial features are incorporated into a recurrent architecture, which models the temporal evolution of the weather observations. In our implementation, CLCRN received the historical weather observations together with the station coordinates. The spatial relationships between stations were used to construct the local graph and condition the convolution kernels, while the recurrent component processed the historical sequence to produce the forecasts. No additional temporal or elevation covariates were provided.

3.3.3 DiffSTG. It combines a spatio-temporal graph neural network with a denoising diffusion process to generate probabilistic forecasts. Its UGnet architecture uses temporal convolutional blocks to model temporal dependencies and graph convolutional blocks to propagate information between spatially connected nodes. During training, the future target sequence is progressively corrupted with noise, while the historical observations are used to condition the denoising process. The model learns to predict the noise added at each diffusion timestep. During inference, the process is reversed, starting from random noise and progressively denoising it to generate the future forecast. DiffSTG uses a predefined graph to perform spatial message passing. In our implementation, the historical weather observations were provided as the conditioning sequence and the future target variables were used as the denoising target. The same predefined adjacency matrix was used throughout the diffusion process, with no additional covariates provided. Temporal information was provided through the temporal position information generated by the dataset.

3.3.4 USTD. It combines a graph-based encoder with a diffusion Transformer to model conditional probabilistic forecasts. The encoder uses graph convolutions to extract a representation of the historical spatio-temporal context, while the diffusion Transformer uses self-attention to model relationships between stations and across time during denoising. Spatial positional information is derived from the graph using Laplacian positional encodings, while temporal positional information is provided separately. During training, noise is added to the future target variables, which are then progressively denoised by the Transformer. The historical representation, graph-based spatial information, temporal information, covariates, and diffusion timestep are used to condition this process. In our implementation, the historical target variables and covariates were first passed through the GWN encoder together with the predefined adjacency matrix. The resulting representation was then provided to the diffusion Transformer alongside the noisy future targets. We used time of day, day of year, and station height as covariates, with the same covariates provided for the corresponding historical and future time steps. The predefined adjacency matrix was also used to construct the Laplacian positional encoding supplied to the Transformer.

3.4 Model Implementation and Evaluation

All models were implemented with Python and PyTorch using publicly available repositories provided by the authors [1, 11, 16, 27]. Additional scripts created were also implemented using Python and PyTorch. Furthermore, the models were implemented as specified in their respective papers on the optimal hyperparameters we found by doing hyperparameter tuning. We used MAE as the target function and optimised using the Adam optimiser for a maximum of 50 epochs for the baselines and 300 epochs for the diffusion-based models. Early stopping was applied to all models with a patience of 15. Slight variations, detailed below, were implemented to ensure the models were suitable for the weather forecasting task. All experiments were run on the University of Cape Town’s High Performance Cluster (HPC) using the A100 GPU.

3.5 Evaluation Metrics

To measure the models’ average error over various horizons, specifically $H \in \{3, 6, 9, 12, 24, 48\}$, we use Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), in line with previous STGNN and meteorology studies [1, 11, 16, 27]. Using both allows for better analysis of the models’ error distribution [6]. Let y_i denote the true target value and $\hat{y}_i$ the corresponding prediction, with N denoting the total number of evaluated predictions across stations and forecast steps. Both metrics retain the units of the measured feature when calculated on the original scale. For the diffusion-based models, we additionally report the Continuous Ranked Probability Score (CRPS), which evaluates the quality of a predictive distribution.

Mean Absolute Error. MAE measures the average magnitude of prediction errors without considering their direction. Lower MAE indicates better performance:

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

Root Mean Squared Error. RMSE measures the magnitude of prediction errors while placing greater weight on larger errors:

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

RMSE is sensitive to outliers [6], which makes it informative when investigating extreme events.

Continuous Ranked Probability Score. CRPS evaluates the compatibility of a predictive cumulative distribution function F with an observed value y . Lower CRPS values indicate better probabilistic forecasts. It is defined as:

$$\int_{\mathbb{R}} (F(z) - \mathbb{I}\{y \leq z\})^2 dz, \quad (4)$$

where $\mathbb{I}\{y \leq z\}$ is an indicator function that equals 1 if $y \leq z$ and 0 otherwise.

3.6 Hyperparameter Tuning

Hyperparameter tuning was performed on all four models using the Tree-structured Parzen Estimator (TPE) algorithm, which is a variant of Bayesian Optimisation [3]. Rather than searching the entire search space using grid search, which would have been time-consuming for the diffusion-based models, TPE evaluates past configurations and splits them into those that performed well and those that did not. It then builds a density over the search space for each group and selects the next configuration by sampling candidates from the well-performing density and choosing whichever configuration is closer to other successful configurations and further from failed configurations [3]. Though it does not exhaustively evaluate the search space the way grid search does, it is cheaper to run since it scales linearly with the number of observations and dimensions, and has been shown empirically to outperform random and grid search on ML hyperparameter tuning tasks [3]. The models were tuned for 30 epochs (10% of the full training time for the diffusion-based models) for 30 trials using the validation dataset. The first 10 trials used a random search to explore configurations, after which the remaining 20 trials explored configurations based on those that performed well during the initial 10 trials. A 48-hour prediction horizon of was used to tune all models. For each model, the hyperparameters tuned aligned with those outlined and highlighted as impactful in their respective papers [1, 11, 16, 27]. The search space and optimal parameters found are shown in Appendix A in Table 3.

4 Results and Analysis

Each model was trained, validated, and tested on each of the four attributes three times using different seeds (2024, 2025, and 2026) to ensure model stability [11, 27]. The diffusion-based models each produced eight samples per run, which were averaged to enable comparison with the baselines. The mean and standard deviation across seeds for each metric are presented in Table 2, and a more comprehensive report of the results can be found in Appendix B.1. The results show that the diffusion-based models are able to predict weather with a similar degree of accuracy as the baselines, but do not consistently outperform them. We highlight our key findings below.

Table 2: Average forecasting performance (Mean and Standard Deviation) for all stations for MAE and RMSE across all attributes. The best performing model is highlighted in bold, and the second-best is underlined.

Model	Humidity		Pressure		Temperature		Wind Speed	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
AGCRN	9.923 $\pm$ 0.001	13.974 $\pm$ 0.010	4.340 $\pm$ 0.170	10.470 $\pm$ 0.350	2.194 $\pm$ 0.017	3.019 $\pm$ 0.025	1.125 $\pm$ 0.001	1.604 $\pm$ 0.001
CLCRN	9.519 $\pm$ 0.004	13.231 $\pm$ 0.005	3.883 $\pm$ 0.000	8.857 $\pm$ 0.049	2.146 $\pm$ 0.002	2.907 $\pm$ 0.003	1.122 $\pm$ 0.006	1.580 $\pm$ 0.001
DiffSTG	11.137 $\pm$ 0.017	14.883 $\pm$ 0.025	6.369 $\pm$ 0.245	13.904 $\pm$ 0.903	2.372 $\pm$ 0.004	3.183 $\pm$ 0.006	1.170 $\pm$ 0.000	1.650 $\pm$ 0.000
USTD	9.965 $\pm$ 0.233	<u>13.867 $\pm$ 0.310</u>	<u>4.329 $\pm$ 0.501</u>	<u>9.117 $\pm$ 1.365</u>	2.209 $\pm$ 0.077	<u>2.974 $\pm$ 0.096</u>	1.151 $\pm$ 0.013	1.647 $\pm$ 0.010

4.1 Predictive Performance

CLCRN performs the best across all metrics. Compared to AGCRN, CLCRN improves MAE by approximately 4% for humidity, 11% for pressure, and 2% for temperature, while its RMSE is lower by approximately 5%, 15%, and 4%, respectively. CLCRN also consistently outperforms USTD with a 3% to 10% reduction in error across both metrics for all attributes. Moreover, it achieves the most substantial improvement over DiffSTG, particularly for pressure, where it reduces error rates by nearly 40%. While our findings are in line with other studies comparing AGCRN and CLCRN [7, 16], these results differ from Gaibie et al. [10], where their AGCRN model outperformed their CLCRN model across all variables and metrics. This is most likely because our datasets covered different South African regions over a different time period.

Antithetically, DiffSTG achieves the highest average MAE and RMSE scores for every attribute. This is an expected result as the model was unable to surpass its deterministic STGNN baselines in the original study [27], and similarly was unable to surpass USTD in a subsequent study [11]. DiffSTG’s MAE score for pressure is roughly 47% higher than USTD’s and 64% higher than CLCRN’s. Similarly, across humidity and temperature forecasts, DiffSTG’s predictive accuracy is approximately 5% to 17% lower than the baselines and USTD. USTD is the more competitive of the two diffusion-based models. It performs similarly to AGCRN across both MAE and RMSE on most attributes, where the two models only outperform each other by relatively small margins. This was particularly noticeable for MAE, where the accuracy achieved between the two models was less than 3% across every attribute. The largest difference occurred for wind speed, where AGCRN achieved a 2.23% lower MAE than USTD. For RMSE, the models were similarly close for most attributes, with differences of less than 3% for three of the four attributes. The main exception was pressure, where USTD achieved a 12.93% lower RMSE than AGCRN.

This differs from the results reported in the original USTD study, where the model improved both MAE and RMSE over existing deterministic and probabilistic forecasting approaches on three benchmark datasets [11]. The difference is likely due to the characteristics of the present dataset and forecasting task as well as baselines chosen. The SAWS dataset contains a relatively small number of irregularly distributed weather stations compared with the large benchmark datasets commonly used in STG forecasting. Additionally, weather variables also have different spatial and temporal dependencies from the traffic and urban-sensor data on which

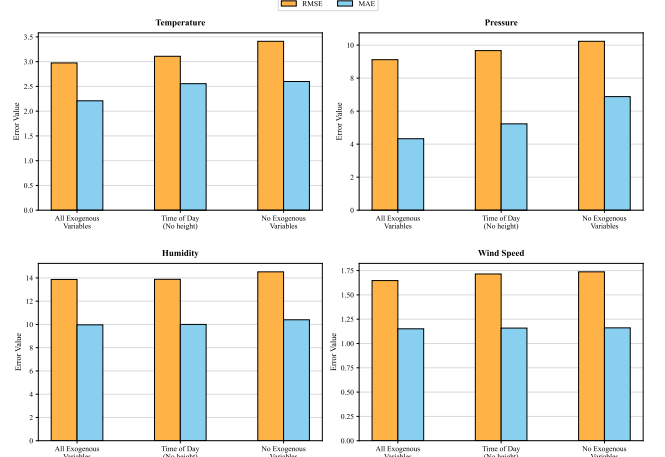


Figure 2: Ablation study showing the impact of exogenous variables on USTD’s predictive performance.

these models were primarily evaluated. Our results do not necessarily suggest that USTD’s advantages may not transfer uniformly across datasets.

4.1.1 The Role of Conditioning in Diffusion-Based Forecasting. The difference between USTD and DiffSTG suggests that the conditioning mechanism is an important contributor to the performance of the diffusion models. USTD combines a deterministic spatio-temporal encoder with probabilistic diffusion decoders, which ensures that the diffusion process can operate on representations that have already captured conditional spatio-temporal structure [11]. The ablation results, shown in Figure 2, provide supporting evidence for the importance of this conditioning. Removing all exogenous variables increased RMSE from 2.974 to 3.411 for temperature, from 9.117 to 13.230 for pressure, from 13.867 to 14.523 for humidity, and from 1.647 to 1.737 for wind speed, with the largest deterioration occurring for pressure. The ablation therefore suggests that contextual information contributes to USTD’s performance. Moreover, the difference in predictive accuracy between the two diffusion models suggests that the use of a diffusion-based probabilistic formulation alone does not guarantee improved point-forecast accuracy, and that the architecture used to condition and generate the forecasts might be important.

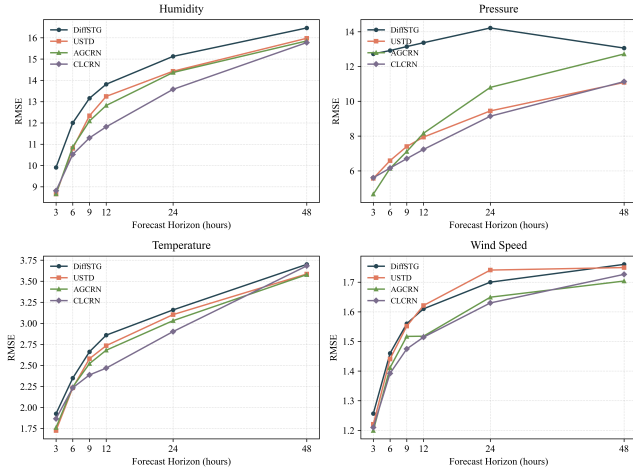


Figure 3: RMSE scores across 3, 6, 9, 12, 24, and 48 hour horizons for CLCRN, AGCRN, DiffSTG, and USTD.

This distinction may also help explain the difference between USTD and DiffSTG. USTD was explicitly designed around a deterministic encoder that captures conditional spatio-temporal patterns before probabilistic decoding, whereas DiffSTG directly uses its diffusion framework with UGnet to learn the spatio-temporal forecasting process [11, 27]. Hu et al. [11] argue that effectively learning the full shape of a probability distribution in order to provide reliable uncertainty ranges requires more information than learning a direct mapping from an input to a single output in the way deterministic STGNNs do. Thus, we suspect that a strong conditional representation provides a better foundation for probabilistic generation, particularly when the available observations alone do not fully describe the state of the atmosphere.

4.2 Prediction Horizon

To further analyse the performance of these models, we present the MAE and RMSE scores across six prediction horizons: 3, 6, 9, 12, 24, and 48. The RMSE results are visualised in Figure 3, while the MAE results are in Appendix B.5. The most noticeable finding is that, just like the baselines, the diffusion-based models’ error generally increases as the prediction horizon increases. For example, DiffSTG’s humidity MAE increases from 7.45 at horizon 3 to 12.43 at horizon 48, while its wind speed MAE increases from 0.91 to 1.25 over the same period. USTD shows the same general pattern. This is an expected result in weather forecasting since predicting further into the future is generally more difficult [20], and is consistent with trends reported in prior STGNN forecasting work [8, 10, 13]. However, the rate and magnitude of degradation differ between variables and models, as shown in Figure 3. Specifically, we see that the diffusion-based models degrade more than the baselines between horizons 6 to 24. Specifically, DiffSTG consistently produces the highest errors across all horizons for all attributes; USTD and AGCRN generally achieve the lowest errors at shorter horizons, after which CLCRN and AGCRN generally outperform the diffusion models between horizon 6 and 24. All models begin to converge at horizon 48.

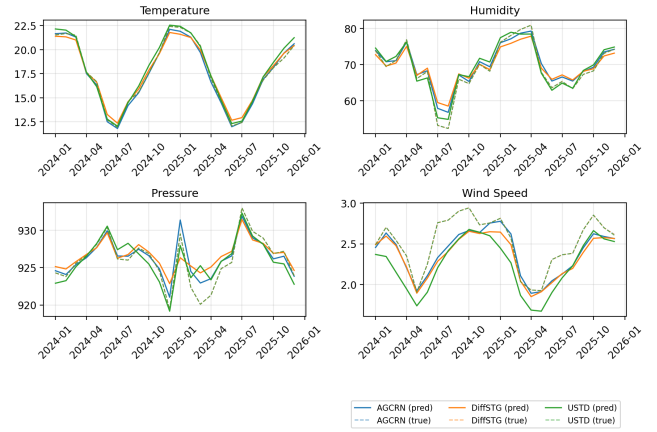


Figure 4: Each model’s predicted values versus the ground truth value for each attribute across the test set.

The comparatively poor performance of DiffSTG across the horizons may be related to how future temporal dependencies are represented. DiffSTG is explicitly non-autoregressive since it generates the complete prediction horizon through the diffusion process rather than recursively feeding previous predictions back into the model [23]. This is advantageous since predictions do not accumulate the recursive errors associated with repeatedly conditioning on previous forecasts [21, 23]. However, the absence of autoregressive feedback may also change how temporal dependencies across the forecast horizon are represented. In particular, the model must construct a coherent multi-step future without using its own intermediate predictions as inputs to subsequent steps [32]. However, USTD achieves competitive predictive performance despite being non-autoregressive as well. Another example is TimeDiff, which is explicitly non-autoregressive, but incorporates autoregressive initialisation and future-mixup conditioning to provide information about short-term temporal structure while retaining non-autoregressive generation [23]. The fact that non-autoregressive models can achieve strong predictive performance when given additional temporal conditioning suggests that non-autoregressive generation itself is unlikely to explain poor forecasting performance. Instead, the way temporal dependencies are encoded before or during generation may be more important. It is plausible that DiffSTG’s non-autoregressive nature may be putting greater demands on the model to represent relationships among multiple future time steps simultaneously. The monthly predictions, in Figure 4, show that DiffSTG generally captures the broad temporal structure of temperature and wind speed, but its predictions are more removed from the observed trajectory than the other three models. Thus, the issue may be a limitation in representing the conditional evolution of the forecast across the full prediction horizon.

Following Gaibie et al. and Davidson et al. [8, 8], we identified k edges, where $k=40$, with the highest weights for each model’s adjacency matrix, and plotted those edges on a map of the studied regions. We analyse the spatial correlation between nodes using their adjacency matrices, which is overlaid onto a map of Kwazulu Natal, Eastern Cape, Free State, and Gauteng. A station will have an

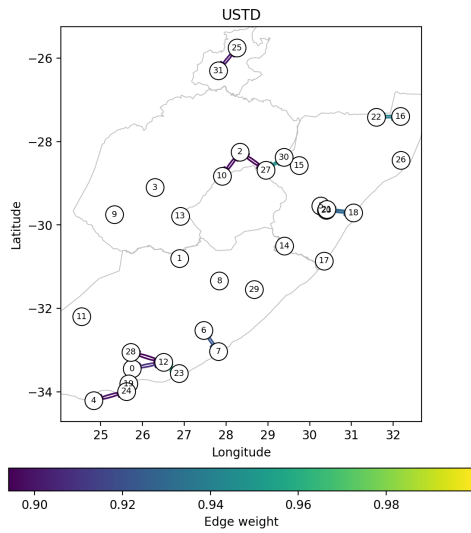


Figure 5: DiffSTG and USTD's fixed adjacency matrix showing the strongest 40 connections overlayed on a map of KZN, EC, Gauteng, and Free State

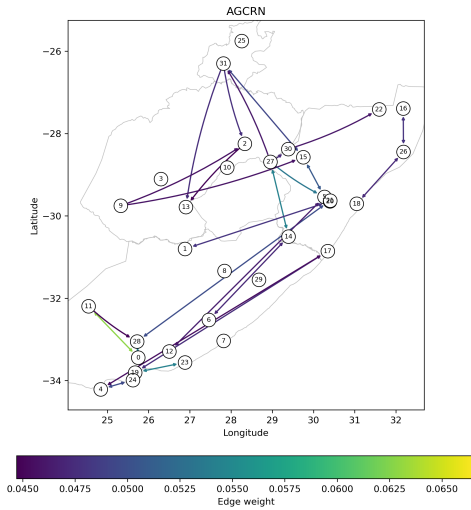


Figure 6: AGCRN's spatial node embeddings.

incoming arrow if it depends on the station from which the arrow originates. This will allow us to analyse the spatial dependencies that the models have learned. The predefined adjacency matrix was constructed by calculating the weighted distance between the stations, and thus places greater weight on stations closer together, as seen in Figure 5. It is also symmetric, making all edges bidirectional, so if station A depends on station B, then station B will also depend on station A.

4.3 Spatial Patterns

The predefined adjacency matrix focuses on local neighbourhoods, and this is reflected strong relationships between geographically

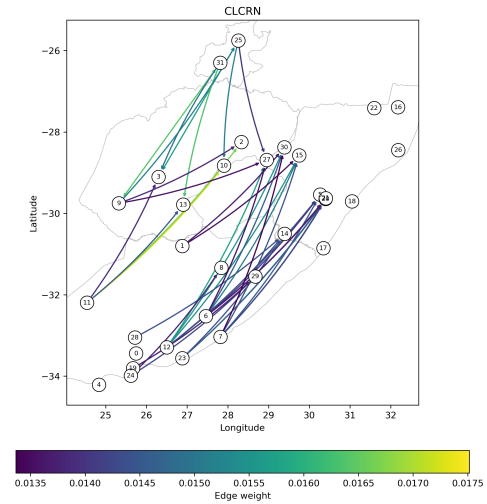


Figure 7: CLCRN's spatial node embeddings.

close stations, which results in the clustering seen in Figure 10 in the Appendix. However, geographical proximity does not necessarily capture all of the relationships that influence weather variables. This may be particularly important for the SAWS network because the stations are irregularly distributed across different geographic and climatic regions. Elevation, coastal proximity and local topography can cause geographically close stations to experience substantially different conditions. The node-level error analysis provides some evidence for this effect. For example, stations with similar geographic positions can exhibit different forecasting errors, while the higher errors observed at particular inland stations coincide with greater variability in some weather variables. The difference between nearby stations with different elevations is particularly relevant because elevation can introduce systematic differences that are not represented by distance alone. The baselines, which do not use pre-defined adjacency matrices, learn more complex spatial relationships as seen in Figures 6 and 7.

4.3.1 Weather Variability. The spatial patterns become more interesting when considering the underlying weather variability at each node, which can be found in B.3 in Table 6. For temperature, the coastal nodes are among the most stable stations in the dataset: nodes 16, 26, 18, 17, 7, 23, 24, and 48 are all within the ten most stable nodes (out of 32 nodes) based on standard deviation. In contrast, node 9, for which all four models report their highest MAE and RMSE scores, is the least stable node, with a standard deviation of 8.73 and a temperature range of 50.8°C (−9.9°C to 40.9°C). Other inland nodes show similar behaviour; for example, node 8 ranks 19th in stability and has a temperature range of 48°C. This suggests that, at least for temperature, the models may find nodes with greater weather variability more difficult to predict. Pressure shows a different pattern. Unlike temperature, humidity, and wind speed, there is greater variation between the models. DiffSTG performs particularly poorly along the KwaZulu-Natal and Eastern Cape coastlines, despite these nodes generally having relatively stable conditions. Nodes 17, 18, and 26 are particularly notable in

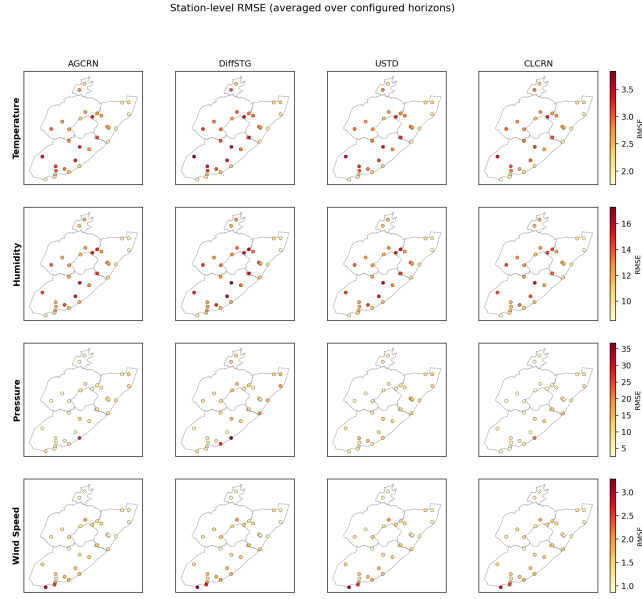


Figure 8: A grid of graphs containing node-specific RMSE errors which are overlaid on a map of the investigated regions.

this respect. This suggests that stability alone does not explain the errors observed for pressure and that the variable itself, or aspects of DiffSTG’s diffusion-based prediction process, may be contributing to this behaviour. There are also two notable pressure outliers, nodes 7 and 23. Both are highly stable, ranking third and seventh respectively in the stability ranking, yet they produce relatively high errors for some models. USTD appears to avoid this problem and achieves comparatively low MAE and RMSE values at these nodes. This is another case where the additional contextual information used by USTD may be helping it account for characteristics that are not captured by the spatial relationships alone. Overall, pressure therefore appears to reveal differences between the models that are less apparent for the other variables.

4.3.2 Elevation. Figure 8 shows that node 25 consistently reports higher temperature errors than its neighbouring node 31, despite the two being geographically close. This occurs across all four models, suggesting that geographical proximity alone does not explain the difference. USTD achieves the lowest errors at these nodes, while AGCRN also performs relatively well. Since AGCRN learns an adaptive adjacency matrix and USTD incorporates station height as contextual information, these results may indicate that both models are able to capture information beyond simple geographical proximity. However, the initial adjacency used by DiffSTG and USTD means that the learned spatial representation should be examined before drawing stronger conclusions. DiffSTG’s predictive performance is expected here since its adjacency matrix only uses fixed distance between stations to determine station relationships and how messages will be passed between stations. However, USTD, which also uses a fixed and predetermined adjacency matrix, and AGCRN report higher prediction accuracy than DiffSTG and CLCRN at this location. A possible reason for this could be that

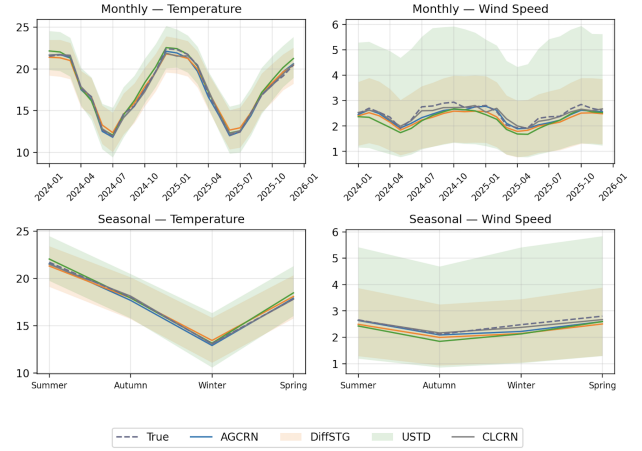


Figure 9: Predicted value versus true value for wind speed and temperature, including the eight samples generated by the diffusion models.

USTD is conditioned on station height as contextual information, which may have allowed it to better account for the relationship between elevation and temperature. We identify a similar pattern in the Free State, where three high-elevation nodes occur around the Free State–KwaZulu-Natal and Free State–Lesotho borders. In the middle of these three nodes, USTD and AGCRN again achieve lower errors than DiffSTG and CLCRN. This suggests that incorporating station height may help USTD distinguish between geographically close stations with different elevation characteristics. A similar pattern can be seen at the isolated node in the Eastern Cape. Although all four models report relatively high errors there, USTD and AGCRN perform better than DiffSTG and CLCRN. The isolation of the node could make it more difficult for the models to learn useful spatial relationships, although this would need to be investigated further using the models’ spatial representations. The fact that USTD achieves a lower RMSE score at this node may indicate that its use of station height provides useful information that is not captured by spatial proximity alone.

Uncertainty The probabilistic results provide a different perspective on the comparison between DiffSTG and USTD. Although USTD consistently produced lower MAE and RMSE than DiffSTG, its generated samples were visibly more dispersed, particularly for wind speed, as shown in Figure 9. This demonstrates that a narrower predictive distribution does not necessarily correspond to a more accurate point forecast. The central tendency of the generated distribution can be accurate even when individual samples span a relatively wide range. The CRPS results, in the Appendix B.6, strengthen this interpretation. USTD achieved lower CRPS than DiffSTG for humidity, pressure, and temperature, with the largest difference occurring for pressure. Its average CRPS decreased from 0.0060 for DiffSTG to 0.0043 for USTD. Temperature showed a smaller improvement, from 0.1076 to 0.1045, while humidity decreased from 0.1266 to 0.1219. These results indicate that USTD’s wider sample range does not simply correspond to poorer probabilistic forecasts. Instead, for these variables, USTD achieved both

lower point-forecast errors and lower distributional errors. However, the wind-speed results demonstrate why sample dispersion should not be interpreted as a direct measure of uncertainty quality. USTD produced a wider sample range and had slightly lower MAE and RMSE than DiffSTG, but its average CRPS was higher, increasing from 0.3745 for DiffSTG to 0.4073 for USTD. Thus, the greater spread of USTD’s samples does not universally improve probabilistic forecasting. For wind speed, the additional dispersion may instead indicate that the predictive distribution is less well matched to the observed distribution. This distinction is important because CRPS evaluates the predictive distribution as a whole, whereas MAE and RMSE evaluate a point prediction. Consequently, the sample plots should not be interpreted as showing that USTD explores the space better than DiffSTG. They show that USTD generates a more dispersed set of samples. Whether that dispersion is useful depends on how well the resulting distribution represents the observations. The CRPS results suggest that it is beneficial for pressure, humidity, and temperature, but not for wind speed. This variable-dependent behaviour is consistent with the fact that uncertainty is not necessarily distributed in the same way across different weather variables. This is also an important distinction from the deterministic models. AGCRN and CLCRN produce point predictions, whereas the diffusion models provide multiple possible future outcomes. The principal advantage of the diffusion models may therefore not be expected to appear entirely in MAE or RMSE. Diffusion-based STG forecasting was motivated in part by the inability of conventional STGNNs to represent intrinsic uncertainty, and DiffSTG explicitly evaluates this probabilistic capability using CRPS in addition to point-forecast metrics [27]. In the present results, the diffusion formulation did not provide a consistent point-accuracy advantage, but USTD did provide improved probabilistic performance for three of the four variables. This suggests that the value of probabilistic STGNNs should be assessed not only by whether they reduce point error, but also by whether their predictive distributions provide useful representations of uncertainty.

5 Limitations and Future Work

The study modelled each weather variable independently and used 32 stations across four provinces. Future work could therefore investigate multivariate forecasting and larger, more geographically diverse station networks. The evaluation also focused on overall forecasting error rather than performance during extreme weather events, which could provide a more practically relevant assessment of model robustness. The use of inverse distance weighting for missing observations may also introduce bias by relying primarily on spatial proximity rather than factors such as elevation or local topography. Alternative spatio-temporal imputation methods could be investigated in future work. Additionally, the four models differ in their architectural designs and available contextual inputs, making it difficult to attribute performance differences to individual components. More controlled ablation studies could isolate the effects of spatial structure, temporal conditioning, and exogenous variables. Finally, probabilistic evaluation was limited to the diffusion models. Future work could compare diffusion-based uncertainty estimates against probabilistic versions of deterministic STGNNs and assess their calibration, particularly during extreme

weather events. Extending the spatial analysis to relate learned representations to physical characteristics such as elevation, topography, and coastal proximity could also provide further insight into how these models represent weather dependencies.

6 Conclusion

In this study, we evaluated two diffusion-based STGNNs, DiffSTG and USTD, against the deterministic AGCRN and CLCRN models for weather forecasting in South Africa. Across temperature, pressure, humidity and wind speed, CLCRN achieved the best overall point-forecast performance, while DiffSTG produced the highest errors. USTD performed substantially better than DiffSTG and remained competitive with AGCRN, while also producing better probabilistic forecasts than DiffSTG for three of the four variables. Our analysis also showed that forecasting performance varied across both forecast horizons and weather stations. Errors generally increased with the prediction horizon, while the spatial analysis identified differences between coastal and inland stations and between stations with different levels of weather variability. The elevation analysis further showed that some neighbouring stations with similar spatial relationships exhibited different forecasting errors, suggesting that station-specific contextual information may be relevant. The USTD ablation study supports this, with the removal of exogenous covariates resulting in a notable deterioration in performance, particularly for pressure. Overall, our results show that diffusion-based STGNNs do not inherently provide more accurate weather forecasts than deterministic STGNNs. Rather, their performance depends on how spatial and temporal dependencies and contextual information are incorporated into the model. This provides a useful empirical baseline for future work on probabilistic STGNN weather forecasting. Future research should investigate multivariate forecasting, larger and more geographically diverse station networks, and improved conditioning mechanisms for diffusion-based models.

7 Acknowledgements

The author would like to thank Associate Professor Deshen Moodley for his supervision, guidance, and support throughout this research. The author also acknowledges the University of Cape Town High Performance Computing (HPC) facility for providing the computational resources required to conduct the experiments. We thank the South African Weather Service (SAWS) for providing the weather data used in this study. Finally, the author acknowledges the National Research Foundation (NRF) for providing the scholarship that made this research possible.

8 AI Usage Statement

All LLMs used were free-tier, and we only used them for guidance (as they don’t have the capacity and capability of doing anything beyond this to the level required). This includes help with scripts to do data processing (to process results) and data visualisation. These documents encompass the entirety of our AI usage for this project. GitHub Copilot was used to explore repositories for the four models in this study. For the writing of this paper, the free tier of Grammarly was used to correct spelling and grammar errors.

A Hyperparameter configurations

The hyperparameters and the search space we used are presented in Table 3.

B Additional Results

This appendix presents the full forecasting results, weather station metadata, and additional spatial analysis used in the experiments.

B.1 Full Forecasting Results

Table 4 presents the complete forecasting results for all evaluated attributes, models, metrics, and prediction horizons. Values are reported as mean $\pm$ standard deviation across the three experimental seeds.

B.2 Weather Station Metadata

Table 5 provides the geographical metadata for the 32 weather stations used in the experiments.

B.3 Outlier Station Analysis

Table 6 presents the stations identified as statistical outliers for each model and weather attribute.

B.4 Spatial Dependencies

The predefined adjacency matrix used by USTD and DiffSTG was examined using the 100 strongest off-diagonal edges to reveal the spatial dependencies represented in the graph.

B.5 Prediction Horizon

The MAE scores across all of our measured horizons for all four models.

B.6 Probability Ranking

The CRPS results show that probabilistic performance also changes with forecast horizon. For humidity, for example, DiffSTG’s CRPS increases from 0.0856 at three hours to 0.1422 at 48 hours. USTD follows a similar pattern, increasing from 0.0740 to 0.1438. This indicates that the predictive distribution becomes less accurate as the horizon increases, which is consistent with the increasing difficulty of the forecasting task.

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] Nir Benmoshe. 2025. A simple solution for the inverse distance weighting interpolation (IDW) clustering problem. *Sci* 7, 1 (2025), 30.
- [3] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems* 24 (2011).
- [4] Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. 2023. DYffusion: A Dynamics-informed Diffusion Model for Spatiotemporal Forecasting. In *Advances in Neural Information Processing Systems*, Vol. 36. doi:10.52202/075280-1961
- [5] Alessandro Capone, Carlo Casolaro, and Francesco Camastra. 2025. Spatio-temporal prediction using graph neural networks: A survey. *Neurocomputing* (2025). doi:10.1016/j.neucom.2025.130400
- [6] Tianfeng Chai and Roland R Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geoscientific model development* 7, 3 (2014), 1247–1250.
- [7] Yibi Chen, Kenli Li, Chai Kiat Yeo, and Keqin Li. 2024. Global-local feature learning via dynamic spatial-temporal graph neural network in meteorological prediction. *IEEE Transactions on Knowledge and Data Engineering* 36, 11 (2024), 6280–6292.
- [8] 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.
- [9] Hanlin Dong, Arian Prabowo, Hao Xue, Ao Shuang, Tianyi Zhou, Yuxuan Liang, and Flora D Salim. 2025. Double-Diffusion: Balancing Speed, Accuracy, and Uncertainty in Probabilistic Forecasting for Urban Sensor Networks. *arXiv preprint arXiv:2506.23053* (2025).
- [10] 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.
- [11] Junfeng Hu, Xu Liu, Zhencheng Fan, Yuxuan Liang, and Roger Zimmermann. 2024. Towards unifying diffusion models for probabilistic spatio-temporal graph learning. In *Proceedings of the 32nd ACM international conference on advances in geographic information systems*. 135–146.
- [12] Ming Jin, Huan Yee Koh, Qingsong Wen, Daniele Zamboni, 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.
- [13] 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.
- [14] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926* (2017).
- [15] 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.
- [16] 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.
- [17] Lequan Lin, Dai Shi, Andi Han, and Junbin Gao. 2025. SpecSTG: A Fast Spectral Diffusion Framework for Probabilistic Spatio-Temporal Traffic Forecasting. In *ICLR 2025 Workshops: DeLTA*.
- [18] Weixuan Lin, Di Wu, and Benoit Boulet. 2021. Spatial-Temporal Residential Short-Term Load Forecasting via Graph Neural Networks. *IEEE Transactions on Smart Grid* 12, 6 (2021), 5373–5384. doi:10.1109/TSG.2021.3093515
- [19] Xu Liu et al. 2023. PriSTI: A Conditional Diffusion Framework for Spatiotemporal Missing Data Imputation. In *Advances in Neural Information Processing Systems*.
- [20] Stefano Materia, Lluís Palma Garcia, Chiem van Straaten, Sungmin O, Antonios Mamalakis, 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.
- [21] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. 2021. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International conference on machine learning*. PMLR, 8857–8868.
- [22] Zakhele Phumlan Shabalala, Mokhele Edmond Moeletsi, Mphethe Isaac Tongwane, and Sabelo Marvin Mazibuko. 2019. Evaluation of infilling methods for time series of daily temperature data: Case study of Limpopo Province, South Africa. *Climate* 7, 7 (2019), 86.
- [23] Lifeng Shen and James Kwok. 2023. Non-autoregressive conditional diffusion models for time series prediction. In *International Conference on Machine Learning*. PMLR, 31016–31029.
- [24] 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.
- [25] 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.
- [26] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.
- [27] 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.
- [28] ZhiPeng Wu. 2026. Spatiotemporal weather forecasting via multi-scale graph neural networks and latent diffusion models. *PLOS ONE* 21, 6 (2026), e0348354. doi:10.1371/journal.pone.0348354

Table 3: Hyperparameter search spaces and selected configurations. A dash indicates that the parameter was not applicable or was not tuned for that model.

Parameter	Search Space	AGCRN	CLCRN	DiffSTG	USTD
Input Window Size	[12, 24, 48]	24	24	24	24
Hidden Size	[16, 32, 64]	-	32	32	-
N	[100, 200, 300]	-	-	300	-
beta end	[0.1, 0.2, 0.4, 0.5]	-	-	0.1	0.5
Number of Samples	[8, 16]	-	-	8	8
Mask Ratio	(0.0, 0.75)	-	-	0.2	0.70
Sample Steps	[50, 100, 200]	-	-	200	200
Batch Size	[32, 64, 128]	32	32	8	[64, 128]
Initial LR	(1e-4, 5e-2)	0.0002	0.005	0.002	0.001
Number of Layers	(1, 4)	2	2	-	-
RNN Units	[16, 32, 64]	64	-	-	-
Embedding Dimension	[1, 2, 3, 5, 8, 10, 15, 20, 30, 128]	-	8	30	128
End Dimension	[128]	-	-	-	128
Dropout	(0.0, 0.2)	-	-	-	0.025

- [29] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32, 1 (2020), 4–24.
- [30] 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.
- [31] 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).
- [32] Tianru Xie, Pingfu Chao, Weizhu Qian, Junhua Fang, and Jiajie Xu. 2026. Autoregressive STG-based Diffusion Model for Spatiotemporal Trajectory Generation. *ACM Transactions on Intelligent Systems and Technology* (2026).
- [33] Qidong Yang, Jonathan Giezendanner, Daniel Salles Civitarese, Johannes Jakubik, Eric Schmitt, Anirban Chandra, Jeremy Vila, Detlef Hohl, Chris Hill, Campbell Watson, et al. 2025. Local off-grid weather forecasting with multi-modal earth observation data. *Journal of Advances in Modeling Earth Systems* 17, 12 (2025), e2025MS005207.
- [34] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. 3634–3640. doi:10.24963/IJCAI.2018/505

Table 4: Normalised forecasting performance (mean $\pm$ standard deviation) across prediction horizons.

Attribute	Metric	Model	3 h	6 h	9 h	12 h	24 h	48 h	Average
Humidity	MAE	DiffSTG	7.4523 $\pm$ 0.2923	8.8872 $\pm$ 0.0274	9.8430 $\pm$ 0.0007	10.3925 $\pm$ 0.0293	11.3240 $\pm$ 0.0414	12.4260 $\pm$ 0.0000	11.1366 $\pm$ 0.0169
		USTD	6.0187 $\pm$ 0.1735	7.6550 $\pm$ 0.2775	8.8144 $\pm$ 0.6656	9.4845 $\pm$ 0.9474	10.3298 $\pm$ 0.1001	11.5171 $\pm$ 0.1293	9.9650 $\pm$ 0.2328
		AGCRN	6.02 $\pm$ 0.0127	7.7941 $\pm$ 0.0228	8.7036 $\pm$ 0.0207	9.2208 $\pm$ 0.0170	10.2618 $\pm$ 0.0324	11.5347 $\pm$ 0.0171	9.9225 $\pm$ 0.0016
		CLCRN	6.2305 $\pm$ 0.0036	7.6022 $\pm$ 0.0053	8.1992 $\pm$ 0.0110	8.5485 $\pm$ 0.0158	9.6965 $\pm$ 0.0038	11.3853 $\pm$ 0.0012	9.5186 $\pm$ 0.0035
	RMSE	DiffSTG	9.9107 $\pm$ 0.0721	12.0015 $\pm$ 0.2216	13.1583 $\pm$ 0.0960	13.8170 $\pm$ 0.0122	15.1260 $\pm$ 0.0239	16.4609 $\pm$ 0.0040	14.8829 $\pm$ 0.0252
		USTD	8.6544 $\pm$ 0.2078	10.8044 $\pm$ 0.4017	12.3378 $\pm$ 0.8154	13.2456 $\pm$ 1.1891	14.4274 $\pm$ 0.1598	15.9760 $\pm$ 0.2260	13.8674 $\pm$ 0.3102
		AGCRN	8.6684 $\pm$ 0.0256	10.8750 $\pm$ 0.0251	12.0887 $\pm$ 0.0422	12.8186 $\pm$ 0.0347	14.3686 $\pm$ 0.0353	15.8516 $\pm$ 0.0136	13.9744 $\pm$ 0.0104
		CLCRN	8.8117 $\pm$ 0.0037	10.5217 $\pm$ 0.0064	11.2966 $\pm$ 0.0191	11.8152 $\pm$ 0.0332	13.5773 $\pm$ 0.0070	15.7748 $\pm$ 0.0038	13.2307 $\pm$ 0.0046
	CRPS	DiffSTG	0.0856 $\pm$ 0.0000	0.1009 $\pm$ 0.0000	0.1109 $\pm$ 0.0000	0.1169 $\pm$ 0.0000	0.1291 $\pm$ 0.0000	0.1422 $\pm$ 0.0000	0.1266 $\pm$ 0.0000
		USTD	0.0740 $\pm$ 0.0017	0.0930 $\pm$ 0.0014	0.1038 $\pm$ 0.0023	0.1098 $\pm$ 0.0030	0.1302 $\pm$ 0.0055	0.1438 $\pm$ 0.0068	0.1219 $\pm$ 0.0040
Pressure	MAE	DiffSTG	5.0664 $\pm$ 0.4356	5.1940 $\pm$ 0.2575	5.3639 $\pm$ 0.2210	5.6226 $\pm$ 0.2941	6.5558 $\pm$ 0.3190	6.7969 $\pm$ 0.0252	6.3690 $\pm$ 0.2454
		USTD	2.3025 $\pm$ 0.2343	2.5136 $\pm$ 0.0890	3.4474 $\pm$ 0.7749	3.2817 $\pm$ 0.1044	4.5369 $\pm$ 0.7024	5.7832 $\pm$ 0.7361	4.3294 $\pm$ 0.5013
		AGCRN	1.55 $\pm$ 0.15	2.11 $\pm$ 0.19	2.54 $\pm$ 0.11	3.00 $\pm$ 0.11	4.52 $\pm$ 0.26	6.17 $\pm$ 0.25	4.34 $\pm$ 0.17
		CLCRN	1.8529 $\pm$ 0.0047	2.1880 $\pm$ 0.0023	2.4709 $\pm$ 0.0024	2.7762 $\pm$ 0.0042	4.0171 $\pm$ 0.0031	5.3907 $\pm$ 0.0022	3.8825 $\pm$ 0.0004
	RMSE	DiffSTG	12.7254 $\pm$ 0.8056	12.9169 $\pm$ 0.7886	13.1468 $\pm$ 0.8096	13.3672 $\pm$ 0.8454	14.2171 $\pm$ 1.0781	13.0607 $\pm$ 0.2915	13.9043 $\pm$ 0.9034
		USTD	5.5680 $\pm$ 0.1153	6.5861 $\pm$ 0.8811	7.3996 $\pm$ 1.1013	7.9397 $\pm$ 1.0027	9.4459 $\pm$ 1.3721	11.0816 $\pm$ 1.7371	9.1166 $\pm$ 1.3645
		AGCRN	4.66 $\pm$ 0.44	6.14 $\pm$ 0.53	7.11 $\pm$ 0.26	8.16 $\pm$ 0.28	10.80 $\pm$ 0.66	12.72 $\pm$ 0.34	10.47 $\pm$ 0.35
		CLCRN	5.6004 $\pm$ 0.1246	6.1583 $\pm$ 0.0867	6.7001 $\pm$ 0.1253	7.2307 $\pm$ 0.2261	9.1419 $\pm$ 0.1989	11.1308 $\pm$ 0.0750	8.8565 $\pm$ 0.0486
	CRPS	DiffSTG	0.0047 $\pm$ 0.0000	0.0048 $\pm$ 0.0000	0.0050 $\pm$ 0.0000	0.0052 $\pm$ 0.0000	0.0062 $\pm$ 0.0000	0.0065 $\pm$ 0.0000	0.0060 $\pm$ 0.0000
		USTD	0.0021 $\pm$ 0.0001	0.0024 $\pm$ 0.0002	0.0033 $\pm$ 0.0008	0.0029 $\pm$ 0.0001	0.0048 $\pm$ 0.0013	0.0057 $\pm$ 0.0006	0.0043 $\pm$ 0.0007
Temperature	MAE	DiffSTG	1.4439 $\pm$ 0.0012	1.7429 $\pm$ 0.0018	1.9826 $\pm$ 0.0024	2.1399 $\pm$ 0.0031	2.3764 $\pm$ 0.0045	2.8027 $\pm$ 0.0062	2.3721 $\pm$ 0.0035
		USTD	1.2562 $\pm$ 0.0768	1.6374 $\pm$ 0.1001	1.8939 $\pm$ 0.1200	2.0276 $\pm$ 0.1073	2.3013 $\pm$ 0.0923	2.7002 $\pm$ 0.0437	2.2090 $\pm$ 0.0773
		AGCRN	1.2653 $\pm$ 0.0037	1.6480 $\pm$ 0.0097	1.8571 $\pm$ 0.0137	1.9746 $\pm$ 0.0120	2.2296 $\pm$ 0.0162	2.6788 $\pm$ 0.0312	2.1944 $\pm$ 0.0173
		CLCRN	1.3459 $\pm$ 0.0002	1.6445 $\pm$ 0.0001	1.7674 $\pm$ 0.0005	1.8246 $\pm$ 0.0011	2.1351 $\pm$ 0.0027	2.7395 $\pm$ 0.0032	2.1459 $\pm$ 0.0018
	RMSE	DiffSTG	1.9264 $\pm$ 0.0021	2.3485 $\pm$ 0.0034	2.6616 $\pm$ 0.0042	2.8591 $\pm$ 0.0051	3.1590 $\pm$ 0.0068	3.7000 $\pm$ 0.0085	3.1827 $\pm$ 0.0055
		USTD	1.7257 $\pm$ 0.0822	2.2341 $\pm$ 0.1242	2.5795 $\pm$ 0.1734	2.7343 $\pm$ 0.1513	3.1031 $\pm$ 0.1043	3.5859 $\pm$ 0.0402	2.9739 $\pm$ 0.0956
		AGCRN	1.7609 $\pm$ 0.0078	2.2438 $\pm$ 0.0109	2.5212 $\pm$ 0.0184	2.6800 $\pm$ 0.0166	3.0314 $\pm$ 0.0155	3.5783 $\pm$ 0.0401	3.0185 $\pm$ 0.0250
		CLCRN	1.8667 $\pm$ 0.0003	2.2335 $\pm$ 0.0001	2.3869 $\pm$ 0.0004	2.4674 $\pm$ 0.0014	2.9016 $\pm$ 0.0062	3.6822 $\pm$ 0.0078	2.9071 $\pm$ 0.0034
	CRPS	DiffSTG	0.0652 $\pm$ 0.0001	0.0779 $\pm$ 0.0001	0.0887 $\pm$ 0.0001	0.0959 $\pm$ 0.0001	0.1070 $\pm$ 0.0002	0.1307 $\pm$ 0.0002	0.1076 $\pm$ 0.0001
		USTD	0.0629 $\pm$ 0.0057	0.0766 $\pm$ 0.0029	0.0888 $\pm$ 0.0049	0.0951 $\pm$ 0.0033	0.1076 $\pm$ 0.0041	0.1255 $\pm$ 0.0019	0.1045 $\pm$ 0.0028
Wind Speed	MAE	DiffSTG	0.9147 $\pm$ 0.0066	1.0500 $\pm$ 0.0141	1.1228 $\pm$ 0.0181	1.1593 $\pm$ 0.0132	1.2159 $\pm$ 0.0083	1.2580 $\pm$ 0.0112	1.1768 $\pm$ 0.0096
		USTD	0.8760 $\pm$ 0.0165	1.0234 $\pm$ 0.0127	1.0977 $\pm$ 0.0246	1.1377 $\pm$ 0.0225	1.1988 $\pm$ 0.0041	1.2297 $\pm$ 0.0340	1.1506 $\pm$ 0.0125
		AGCRN	0.8602 $\pm$ 0.0027	1.0097 $\pm$ 0.0017	1.0781 $\pm$ 0.0008	1.1148 $\pm$ 0.0007	1.1569 $\pm$ 0.0031	1.1992 $\pm$ 0.0020	1.1249 $\pm$ 0.0010
		CLCRN	0.8756 $\pm$ 0.0006	1.0020 $\pm$ 0.0055	1.0557 $\pm$ 0.0062	1.0785 $\pm$ 0.0051	1.1509 $\pm$ 0.0056	1.2183 $\pm$ 0.0068	1.1215 $\pm$ 0.0057
	RMSE	DiffSTG	1.2631 $\pm$ 0.0091	1.4721 $\pm$ 0.0170	1.5831 $\pm$ 0.0327	1.6344 $\pm$ 0.0345	1.7132 $\pm$ 0.0186	1.7733 $\pm$ 0.0188	1.6635 $\pm$ 0.0191
		USTD	1.2207 $\pm$ 0.0166	1.4408 $\pm$ 0.0213	1.5516 $\pm$ 0.0445	1.6212 $\pm$ 0.0505	1.7408 $\pm$ 0.0014	1.7493 $\pm$ 0.0275	1.6468 $\pm$ 0.0104
		AGCRN	1.1993 $\pm$ 0.0049	1.4114 $\pm$ 0.0031	1.5170 $\pm$ 0.0021	1.5764 $\pm$ 0.0045	1.6493 $\pm$ 0.0039	1.7038 $\pm$ 0.0010	1.6038 $\pm$ 0.0011
		CLCRN	1.2099 $\pm$ 0.0013	1.3922 $\pm$ 0.0031	1.4747 $\pm$ 0.0029	1.5143 $\pm$ 0.0028	1.6299 $\pm$ 0.0039	1.7263 $\pm$ 0.0025	1.5798 $\pm$ 0.0012
	CRPS	DiffSTG	0.2859 $\pm$ 0.00	0.3322 $\pm$ 0.00	0.3584 $\pm$ 0.00	0.3700 $\pm$ 0.00	0.3874 $\pm$ 0.00	0.4033 $\pm$ 0.00	0.3745 $\pm$ 0.00
		USTD	0.2950 $\pm$ 0.0082	0.3592 $\pm$ 0.0182	0.4037 $\pm$ 0.0243	0.4118 $\pm$ 0.0334	0.4038 $\pm$ 0.0281	0.4337 $\pm$ 0.0421	0.4073 $\pm$ 0.0245

Table 5: Geographical Description of the South African Weather Stations Used in the Experiments.

No.	Station Name	Latitude	Longitude	Province	Height (m)
0	Addo Elephant Park	-33.4420	25.7480	Eastern Cape	142.0
1	Aliwal-North Plaatkop	-30.8020	26.8830	Eastern Cape	1350.0
2	Bethlehem WO	-28.2490	28.3340	Free State	1689.0
3	Bloemfontein WO	-29.1030	26.2980	Free State	1354.0
4	Cape St Francis	-34.2110	24.8340	Eastern Cape	9.0
5	Cedara	-29.5410	30.2650	KwaZulu-Natal	1071.0
6	Dohne - AGR	-32.5270	27.4600	Eastern Cape	901.0
7	East London WO	-33.0350	27.8160	Eastern Cape	134.0
8	Elliot	-31.3360	27.8400	Eastern Cape	1464.0
9	Fauresmith	-29.7530	25.3230	Free State	1387.0
10	Ficksburg	-28.8270	27.9040	Free State	1628.0
11	Graaff - Reinet	-32.1930	24.5430	Eastern Cape	792.0
12	Grahamstown	-33.2900	26.5020	Eastern Cape	642.0
13	Knellpoortdam	-29.7860	26.9030	Free State	1468.0
14	Kokstad	-30.5020	29.3940	KwaZulu-Natal	1451.0
15	Ladysmith	-28.5750	29.7500	KwaZulu-Natal	1069.0
16	Makatini Research Centre	-27.3940	32.1760	KwaZulu-Natal	63.0
17	Margate	-30.8580	30.3430	KwaZulu-Natal	152.0
18	Mount Edgecombe	-29.7060	31.0460	KwaZulu-Natal	94.0
19	Ngqura - Coega	-33.8050	25.6690	Eastern Cape	47.0
20	Oribi Airport	-29.6470	30.3990	KwaZulu-Natal	721.0
21	Pietermaritzburg	-29.6270	30.4020	KwaZulu-Natal	673.0
22	Pongola	-27.4130	31.5900	KwaZulu-Natal	319.0
23	Port Alfred - Airport	-33.5590	26.8800	Eastern Cape	84.0
24	Port Elizabeth AWOS	-33.9860	25.6160	Eastern Cape	60.0
25	Pretoria University Proefplaas	-25.7520	28.2580	Gauteng	1380.0
26	Riverview	-28.4440	32.1820	KwaZulu-Natal	49.0
27	Royal National Park	-28.6850	28.9540	KwaZulu-Natal	1392.0
28	Somerset East	-33.0490	25.7190	Eastern Cape	513.0
29	Umthatha WO	-31.5490	28.6730	Eastern Cape	745.0
30	Van Reenen	-28.3780	29.3850	KwaZulu-Natal	1683.0
31	Zuurbekom AWS	-26.3000	27.8130	Gauteng	1581.0

Table 6: Outlier Stations and Comparative Model Performance (RMSE and MAE).

Station Name	Height (m)	Attribute	Metric	AGCRN	DiffSTG	USTD	CLCRN
Cape St Francis	9.0	Humidity	MAE	6.19	6.62	6.25	6.56
Elliot	1464.0	Humidity	MAE	11.75	13.48	11.47	12.73
Riverview	49.0	Humidity	MAE	6.04	6.52	6.02	6.47
Cape St Francis	9.0	Humidity	RMSE	9.07	9.36	9.08	9.28
Elliot	1464.0	Humidity	RMSE	16.32	17.27	16.34	17.49
Riverview	49.0	Humidity	RMSE	8.67	9.12	8.62	9.15
East London WO	134.0	Pressure	MAE	15.72	21.60	7.18	19.09
Elliot	1464.0	Pressure	MAE	5.11	7.19	4.90	5.70
Margate	152.0	Pressure	MAE	4.29	9.27	5.27	6.98
Port Alfred - Airport	84.0	Pressure	MAE	4.95	24.53	3.88	5.90
Riverview	49.0	Pressure	MAE	3.60	9.73	4.00	4.71
Umthatha WO	745.0	Pressure	MAE	4.97	6.28	5.32	6.89
East London WO	134.0	Pressure	RMSE	28.31	36.72	12.29	32.40
Port Alfred - Airport	84.0	Pressure	RMSE	9.21	27.54	6.70	8.91
Cape St Francis	9.0	Wind Speed	MAE	2.28	2.43	2.33	2.44
Port Elizabeth AWOS	60.0	Wind Speed	MAE	1.88	1.97	1.90	1.99
Cape St Francis	9.0	Wind Speed	RMSE	3.02	3.15	3.07	3.18
Port Elizabeth AWOS	60.0	Wind Speed	RMSE	2.53	2.61	2.55	2.62

Note: Bold values indicate that the error metric is a statistical outlier for that specific model and attribute across the weather station network.

Table 7: CRPS results for the diffusion-based models.

Attribute	Model	3 h	6 h	9 h	12 h	24 h	48 h	Average
Humidity	DiffSTG	0.0856	0.1009	0.1109	0.1169	0.1291	0.1422	0.1266
	USTD	0.0740	0.0930	0.1038	0.1098	0.1302	0.1438	0.1219
Pressure	DiffSTG	0.0047	0.0048	0.0050	0.0052	0.0062	0.0065	0.0060
	USTD	0.0021	0.0024	0.0033	0.0029	0.0048	0.0057	0.0043
Temperature	DiffSTG	0.0652	0.0779	0.0887	0.0959	0.1070	0.1307	0.1076
	USTD	0.0629	0.0766	0.0888	0.0951	0.1076	0.1255	0.1045
Wind Speed	DiffSTG	0.2859	0.3322	0.3584	0.3700	0.3874	0.4033	0.3745
	USTD	0.2950	0.3592	0.4037	0.4118	0.4038	0.4337	0.4073

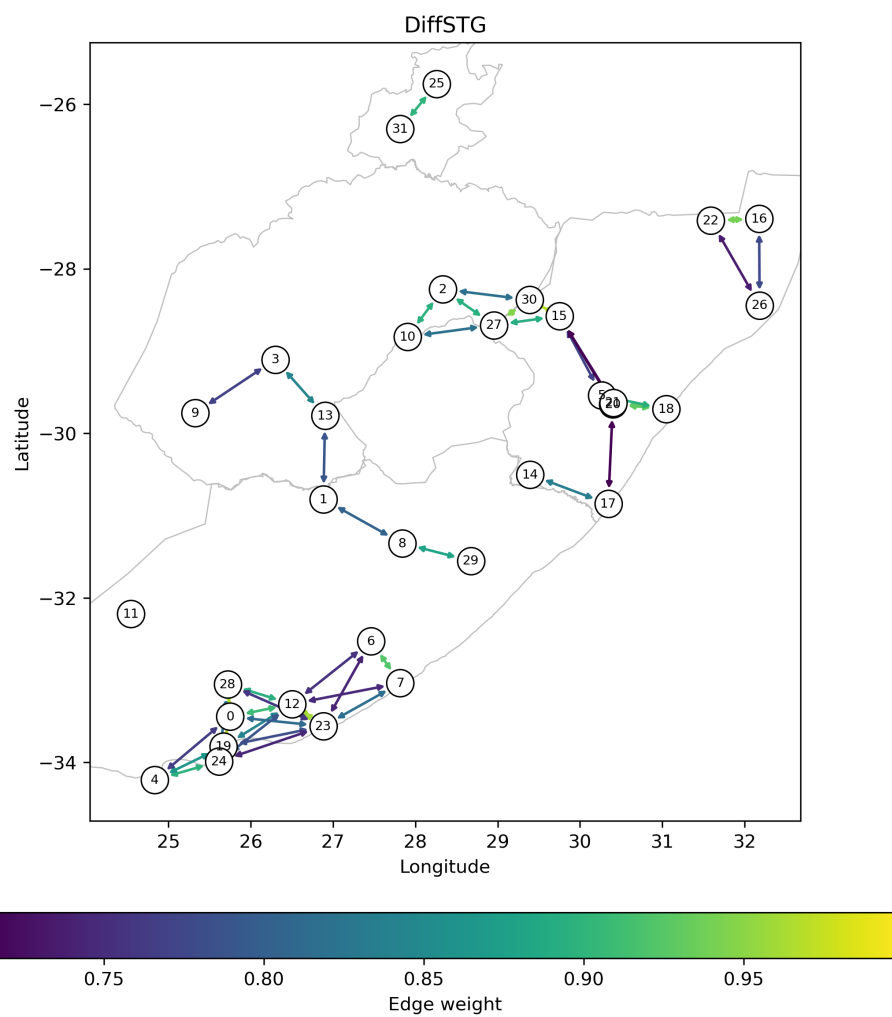


Figure 10: The 100 strongest off-diagonal edges in the predefined adjacency matrix used by USTD and DiffSTG.

