

Continuous-Time and Discrete Spatio-Temporal Graph Neural Networks for Normal and Extreme Weather Prediction on South African Station Data

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

Abstract

Spatio-temporal graph neural networks (STGNNs) achieve strong average-case accuracy for station-level weather forecasting. Two things stay untested: their behaviour on the extreme events that symmetric training losses smooth away, and how far ahead their forecasts remain skillful. We benchmark MTGODE—a continuous-time neural-ODE STGNN and, to our knowledge, the first used for weather forecasting—against STGODE, AGCRN, and CLCRN, crossing the spatial prior (learned versus geometric) with the temporal formulation (discrete versus continuous) in a 2×2 design over 32 South African Weather Service stations (2016–2025). Evaluation spans 48-hour regression, stratified by extremes, and extreme-event detection under calendar- and hour-relative percentile labels. MTGODE attains the lowest error on three of four variables and is inseparable from CLCRN on surface pressure, though the advantage is short-lived, the four architectures converging by two days. Neither axis helps alone: the interaction dominates—continuous time helps with a learned adaptive graph and hurts with a geometric one. All four models smooth the tails, degrading by up to a factor of 3.5 on extreme hours, and the model most accurate on average is not the most accurate on extremes—so mean error is the wrong criterion for selecting an extreme-event forecaster. The most accurate model is also among the cheapest, at 138,096 parameters against 1,663,104 for the weakest.

CCS Concepts

• **Computing methodologies** → **Machine learning; Neural networks**; • **Applied computing** → **Earth and atmospheric sciences**.

Keywords

spatio-temporal graph neural networks, weather forecasting, extreme events, neural ODEs, graph structure learning, South Africa

1 Introduction

Weather forecasting underpins planning in agriculture, water management, energy, and disaster preparedness, and it is the extremes that carry the cost: the April 2022 KwaZulu-Natal floods killed 459 people and were the most catastrophic natural disaster yet recorded in the province [12], while heatwaves recur over the northern interior at a frequency projected to rise [17]. Short-range forecasts are consequential for the region precisely because they are what warning decisions rest on. The system being forecast is dynamic and chaotic, and the spatial and temporal dependencies that drive

it are seldom directly observable [7]. Operational forecasting remains dominated by numerical weather prediction, while classical statistical approaches treat the problem as multivariate time-series prediction. Deep learning has challenged both over the past decade, but its most common architectures model temporal structure while neglecting the spatial dependencies between observation sites. Spatio-Temporal Graph Neural Networks (STGNNs) close this gap by representing each weather station as a node in a graph whose weighted edges express the strength of inter-station dependencies: a node captures the local dynamics at one location, and the graph structure captures the global dynamics of the system [7, 9].

Two prior studies form a direct lineage for STGNN weather prediction on South African station data: Davidson and Moodley [7] established STGNN viability at 21 stations, and Gaibie et al. [9] extended the comparison to AGCRN [4] and CLCRN [15] at 45 stations, contributing the hyperparameter optimisation protocol this study builds on (Section 2.2). The lineage shares three limits that motivate the present study: every model evaluated is discrete-time, evaluation stops at 24 hours, and extreme weather events are not a prediction target in either study.

Neural ordinary differential equations [6] replace discrete layers with a continuously integrated vector field, applied to spatio-temporal graphs by STGODE [8] and MTGODE [14] (Section 2.4). To the best of our knowledge, MTGODE has not previously been applied to meteorological forecasting, and no continuous-time graph model has been evaluated on an irregular meteorological station network.

We cross the two properties that separate the four models above: the spatial prior (a learned adaptive graph versus a geometric construction from station coordinates) and the temporal formulation (discrete recurrence versus a continuous-time neural ODE). AGCRN, CLCRN, MTGODE, and STGODE instantiate the four cells of this 2×2 design. All four are evaluated on a 32-station South African Weather Service (SAWS) network spanning 2016–2025 at hourly resolution, in two phases. Phase 1 is 48-hour regression on the lineage’s four target variables, evaluated at six horizons by truncation of the direct forecast, doubling the 24-hour ceiling of prior work. Phase 2 is extreme-event detection under calendar- and hour-relative 95th-percentile labels that extend the ETCCDI convention [37] to hourly resolution, with pre-registered naive baselines and decision rules. Every model runs through one benchmarking pipeline, WeatherFLOW, sharing preprocessing, splits, loss, training harness, seeds, and evaluation (Section 3.3). This paper covers the continuous-time workstream and the shared framework within a three-member project; the AGCRN and CLCRN baselines are common across the team’s papers.

This study evaluates whether continuous-time STGNNs improve weather prediction on South African station data relative to established discrete-time baselines, and how they behave outside average conditions. We expect a continuous-time formulation to reduce forecast error relative to discrete recurrence, but not independently of the spatial prior: because a learned adaptive graph and a continuous vector field impose comparable degrees of flexibility, the benefit of either may depend on the other. We further expect that accuracy on extreme observations, the horizon over which a forecast remains valid, and the cost in parameters and inference time of attaining that accuracy each rank the architectures differently from mean error, so that no single criterion selects the same model; and we treat the stability of any such ranking under reinitialisation as a quantity to be measured rather than assumed.

The contributions of this paper are as follows:

- (1) To the best of our knowledge, the first application of MT-GODE — and of learned-adaptive-graph neural ODEs generally — to meteorological forecasting, on an irregular station network.
- (2) A 2×2 comparison of spatial prior against temporal formulation on South African station data under a single pipeline, extending the Davidson–Gaibie lineage to 48-hour horizons.
- (3) An extreme-event detection formulation that extends ETC-CDI percentile indices to hourly resolution through calendar- and hour-relative thresholds, with pre-registered naive baselines and decision rules, removing by construction the calendar signal that static thresholds concede to any climatology-aware classifier.
- (4) WeatherFLOW, a reproducible benchmarking framework that integrates four STGNN families under a single training and evaluation harness, with implementation-fidelity audits reported in Section 3.

The remainder of this paper is organised as follows. Section 2 reviews the problem formulation, the four architectures, and the extreme-event detection literature; Section 3 describes the experimental design; Section 4 presents the results and analysis; Section 5 discusses the findings and concludes; and Section 6 addresses limitations and future work.

2 Background and Related Work

2.1 Problem Formulation

Following Davidson and Moodley [7], we represent the station network as a weighted graph $G = (V, E, A)$, where each of the $N = |V|$ nodes is a weather station, edges connect stations that influence one another, and the weight matrix $A \in \mathbb{R}^{N \times N}$ encodes the strength of these dependencies. At each hour t the network produces an observation matrix $X_t \in \mathbb{R}^{N \times D}$ over the $D = 6$ measured variables, and a length- w input window is written $X_{t-w+1:t}$.

Phase 1 is deterministic multi-step regression. For a designated target variable v , a model f_θ maps the input window and the graph to a joint forecast over all stations and the next H hours,

$$\hat{Y}_{t+1:t+H} = f_\theta(X_{t-w+1:t}, G), \quad \hat{Y} \in \mathbb{R}^{H \times N}, \quad (1)$$

with $w = 12$ and $H = 48$ in this study. Each experiment predicts one target variable while receiving all six variables as input, the univariate-target regime of Gaibie et al. [9].

Phase 2 is extreme-event detection. For label variable v , station n , and horizon step h , the ground truth is the exceedance indicator

$$E_{t+h,n,v} = \mathbb{1} [X_{t+h,n,v} > \tau_{n,v,d(t+h),k(t+h)}], \quad (2)$$

where the threshold tensor τ is conditioned on station, variable, day-of-year d , and hour-of-day k , and is estimated from training data only; its construction is specified in Section 3.2.2. The four architectures differ precisely in how they obtain and use A — two learn it, two construct it from station geometry — which is the first axis of our experimental design.

2.2 STGNNs for Weather Prediction in South Africa

Data-driven forecasting now rivals numerical weather prediction on global reanalysis grids, but skill on extreme weather continues to lag and requires dedicated evaluation [20]. Station-network forecasting is a distinct regime: observations are sparse, irregularly spaced, and locally noisy, and benchmarking in this regime has likewise been dominated by discrete-time models: Weather2K, a benchmark over 2,130 Chinese ground stations, evaluates discrete-time sequence and graph baselines exclusively [38].

For South Africa, two studies form the direct lineage. Davidson and Moodley [7] evaluated Graph WaveNet [34] and the low-rank weighted graph network of Wilson et al. [32] against LSTM and TCN baselines for maximum-temperature prediction at 21 stations. Graph WaveNet proved the most accurate, stable, and robust model under walk-forward validation, and the study introduced the practice of analysing the dependency structures the trained models learn. Gaibie et al. [9] extended the benchmark to AGCRN [4] and CLCRN [15] at 45 stations, predicting temperature, pressure, humidity, and wind speed at horizons up to 24 hours. Both newer models outperformed Graph WaveNet, with AGCRN achieving the best overall performance, and the study’s hyperparameter optima on SAWS data are inherited directly by our baselines (Section 3.4). The two studies differ in evidential strength: Davidson and Moodley validate under walk-forward resampling and report over ten independent runs per configuration, where Gaibie et al. average three runs on a single hold-out, so the later and more extensive benchmark rests on the weaker protocol.

A parallel strand at our institution benchmarks STGNNs on chaotic lattice dynamics and scores them by validity horizons rather than average error alone [18]; Section 3.5 shows how we adapt that perspective to station weather data.

2.3 Discrete-Time Recurrent Architectures

AGCRN [4] refines a GRU whose linear mappings are replaced by graph convolutions, and contributes two modules. Node Adaptive Parameter Learning factorises the convolution weights into a small per-node embedding multiplied by a shared weight pool, giving every station node-specific parameters without scaling the parameter count with the network size. Data Adaptive Graph Generation removes the need for a predefined graph entirely: the adjacency matrix is computed as the normalised similarity of learnable node embeddings and is trained end-to-end with the forecasting objective, extending the self-adaptive adjacency introduced by Graph WaveNet [34]. AGCRN was developed and evaluated on traffic

benchmarks; in our design it occupies the learned-adaptive, discrete-time cell.

CLCRN [15] is the only model of the four designed for meteorology. Its conditional local convolution builds a k -nearest-neighbour neighbourhood around each station on the sphere and generates the kernel weights as a smooth learned function of each neighbour’s local displacement, in geodesic distance and orientation, so the convolution is anisotropic and location-aware, approximating a continuous kernel over the Earth’s surface. The convolution is embedded in a recurrent encoder–decoder that decodes the forecast autoregressively, one step at a time; the consequences of this asymmetry relative to the direct multi-step models are addressed in Section 3.3. Evaluated on WeatherBench reanalysis fields [24] and transferred to SAWS stations by Gaibie et al. [9], CLCRN occupies the geometric, discrete-time cell.

2.4 Continuous-Time Graph ODE Architectures

Neural ordinary differential equations [6] replace a stack of residual layers with a vector field: the hidden state is defined as the solution of an initial-value problem $d\mathbf{H}(t)/dt = f_\theta(\mathbf{H}(t), t)$ computed by a numerical solver, so depth becomes integration time and the latent trajectory is continuous. On graphs, this construction permits deep spatial propagation without the over-smoothing that degrades stacked graph convolutions, and recent designs learn the graph itself within the ODE [16].

STGODE [8] introduced the formulation to spatio-temporal forecasting. It prescribes two static graphs: a spatial adjacency from a thresholded Gaussian kernel on inter-station distance, and a semantic adjacency that binarises dynamic-time-warping distances between node series, connecting stations whose histories share shape regardless of location. Features then propagate through a tensor-valued ODE that acts simultaneously on the node, temporal, and feature modes of the hidden state, with the block input re-entering as a restart term that counteracts the over-smoothing to which deep propagation is otherwise prone. Each block sandwiches the ODE between temporal convolutions, with spatial and semantic branches in parallel. STGODE was evaluated exclusively on road-traffic corpora; here it occupies the geometric, continuous-time cell.

MTGODE [14] targets three limitations its authors identify in prior work: discrete architectures produce discontinuous latent trajectories, stacked designs carry redundant parameters, and most methods rely on a graph prior. A graph learner in the style of MTGNN [33] constructs a sparse directed adjacency from trainable node embeddings. Temporal aggregation is itself an ODE whose derivative is parameterised by a single shared gated temporal-convolution block, and nested within each temporal solver step, spatial propagation runs as a second ODE,

$$\frac{d\mathbf{H}_G(s)}{ds} = (\hat{\mathbf{A}} - \mathbf{I}) \mathbf{H}_G(s), \quad (3)$$

the continuous generalisation of simplified graph convolution, with an attentive readout over intermediate states. Jin et al. position their design against STGODE directly: STGODE makes graph propagation continuous, but its temporal aggregation remains discrete [14], so the two models separate cleanly along our temporal axis as well as the spatial one. MTGODE was evaluated on

electricity-consumption, solar-power, and traffic datasets; it occupies the learned-adaptive, continuous-time cell.

Continuous-time methods have begun reaching atmospheric applications, with physics-informed ODEs for weather forecasting on regular reanalysis grids [30] and station-based air-quality modelling [13, 36], but not yet on station-network weather forecasting.

2.5 Extreme-Event Detection in Machine Learning Forecasting

Machine learning studies that target weather extremes divide by how they define an event. The dominant convention derives thresholds from climatological percentiles: the ETCCDI indices compute, for each location and calendar day, a percentile over a moving window of days pooled across the climatological baseline, so that an extreme is always extreme relative to what is normal for that place and time of year [37]. A recent scoping review of machine-learning heatwave prediction finds the same convention dominant, with the majority of its 23 reviewed studies using percentile thresholds, typically the 90th or 95th [3]. The same review cautions that inconsistent event definitions limit comparability across studies, and identifies both South Africa and spatio-temporal architectures such as GNNs as under-represented. The operational South African definition follows the same anomaly-relative principle: SAWS declares a heatwave when maximum temperature exceeds the hottest-month mean maximum by five degrees Celsius for at least three consecutive days [17, 26]. Our labels (Section 3.2.2) extend the ETCCDI construction to hourly resolution and adopt the anomaly-relative principle, relaxing the multi-day duration requirement, which a 48-hour horizon cannot express.

We formulate extremes as detection rather than regression because symmetric regression losses systematically underestimate the tails [35], whereas presence/absence classification permits imbalance-aware weighting; the marine-heatwave literature, whose percentile-based event definitions closely parallel ours, likewise emphasises imbalance-aware training and recommends rare-event metrics such as the critical success index [31]. Two alternatives do not fit this setting: extreme-value-theoretic block maxima [10] answer distributional questions rather than the per-timestep detection a forecast horizon requires, and expert-labelled event masks [23] do not exist for South African station data. Validating forecast systems on named historical extreme events has recent precedent [21], which our case study follows.

Taken together, the literature leaves a specific gap: the South African lineage validates discrete-time STGNNs up to 24 hours without extreme-event prediction, continuous-time graph models have never been evaluated on meteorological station networks, and, to our knowledge, no study crosses the spatial prior with the temporal formulation under one protocol. The design of Section 3 fills that gap.

3 Experimental Design

The design that follows is built to answer three questions: whether a continuous-time formulation transfers to an irregular station network, whether the spatial prior and the temporal formulation can be assessed independently of one another, and whether average-case accuracy predicts accuracy on extremes. Each is addressed by

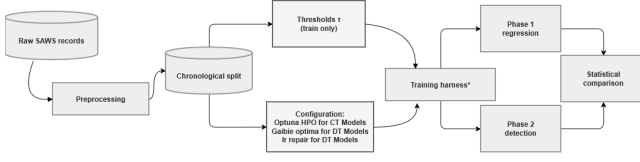


Figure 1: The WeatherFLOW experimental pipeline.

a controlled comparison across the four cells of Table 1, holding everything outside the two design axes fixed. This section describes the data and preprocessing pipeline, the two prediction tasks, the four model integrations, training and hyperparameter selection, the evaluation protocol, and the computing infrastructure. Values inherited from prior work are cited to their sources throughout, and every departure is disclosed where it occurs.

Figure 1 summarises the procedure end to end, and the subsections that follow expand each stage in turn.

3.1 Data and Preprocessing

The primary dataset is hourly surface-station data supplied by the South African Weather Service. The network comprises 32 stations spanning 1 January 2016 to 31 December 2025, giving 87,672 hourly timesteps of six variables per station: air temperature (t2m), precipitation (tp), relative humidity (rh), wind direction (wd), wind speed (ws), and surface pressure (sp). Stations were admitted under Davidson and Moodley’s criterion of at most 10% overall missingness [7]. The stations concentrate in the eastern half of the country (Figure 2), so results characterise South Africa’s predominantly summer-rainfall east rather than the winter-rainfall southwest. Eight stations pass overall while one variable, typically pressure or wind, exceeds 10% alone; the stricter 24-station all-variable subset is retained as a sensitivity option. Parsing follows the SAWS export convention that blank rainfall cells denote zero rather than missing, a convention that materially changes apparent precipitation coverage.

The preprocessing pipeline is ordered to be leakage-safe. Raw records are parsed to a time $\times$ station $\times$ variable array; missing values are imputed by inverse-distance weighting over the five nearest stations [27], with temporal fill as the fallback for regional outages; and the record is then split chronologically 60/20/20 — 52,603, 17,534, and 17,535 hours, the test partition covering the final fifth of the record (2024–2025) — before any statistic is computed. After windowing at $w = 12$ and $H = 48$ this yields 52,544, 17,475, and 17,476 samples. Extreme-event thresholds (Section 3.2.2) and per-station, per-variable z-score normalisation are estimated on the training partition only, and a $\pm 10\sigma$ clip guards against residual encoding artefacts.

3.2 Forecasting and Detection Tasks

3.2.1 Phase 1: multi-horizon regression. Phase 1 instantiates the mapping of Eq. (1); targets are {t2m, sp, rh, ws}, with precipitation and wind direction as inputs only. The input window is 12 hours, the optimum of Gaibie’s search over {12, 24, 48, 72, 168} hours [9] and identical to the MTGODE paper’s multi-step configuration [14]. The horizon is 48 hours, doubling the lineage’s ceiling, with results reported at six sub-horizons $h \in \{3, 6, 9, 12, 24, 48\}$ by truncation:

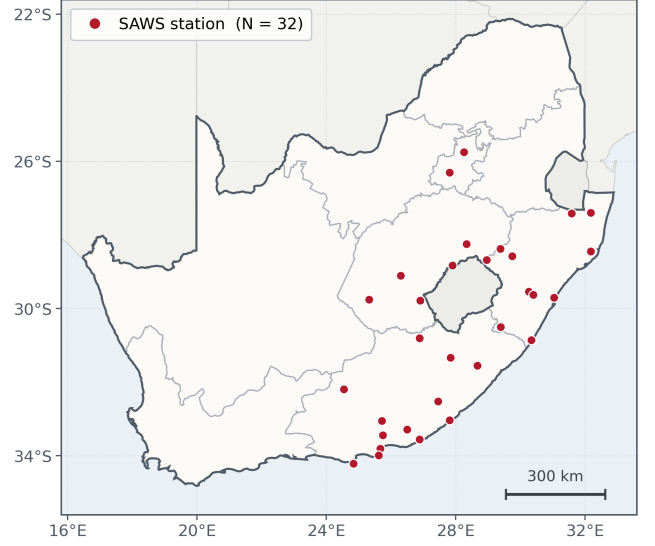


Figure 2: The 32-station SAWS network (2016–2025), concentrated in the eastern half of the country.

the model’s step- h prediction is the metric for horizon h . The first five are those of Gaibie et al.; the sixth is the endpoint of the extended horizon.

This design departs from the lineage in two respects. Davidson and Gaibie trained a separate model per horizon; doing so here would multiply the campaign roughly sixfold, and truncation reports the per-horizon skill contained in one 48-step forecast, at the cost that no horizon receives a dedicated training budget. Evaluation also uses a single chronological hold-out rather than walk-forward validation [7], with the test partition’s calendar coverage stated explicitly. The 48-hour horizon lies beyond the lineage’s validated regime and is read as an extrapolation finding rather than a tuned operating point. Phase 1 comprises 4 models $\times$ 4 targets $\times$ 10 seeds = 160 runs.

3.2.2 Phase 2: calendar-relative extreme-event detection. Phase 2 instantiates the detection task of Eq. (2) with a calendar- and hour-relative threshold tensor,

$$\tau_{n,v,d,k} = Q_{0.95} \{ X_{t,n,v} : t \in \mathcal{T}_{\text{train}}, |doy(t) - d|_{366} \leq 7, |hod(t) - k|_{24} \leq 1 \}, \quad (4)$$

the 95th percentile of training observations pooled over a ± 7 -day, ± 1 -hour window that wraps at year and day boundaries; exceedance is strict. This extends the ETCCDI moving calendar-window convention [37] to hourly resolution, with the day window widened from five days to seven so that our six-year baseline clears the roughly 200 samples a stable 95th-percentile estimate requires (median cell 270; ETCCDI’s 30-year, five-day pooling yields about 150). The window was fixed before any model training by a pre-registered sweep over twelve candidates. Labels cover {t2m, tp, rh, ws}: wind

Table 1: The 2×2 experimental design.

Spatial prior \ Temporal	Discrete time	Continuous time (neural ODE)
Learned adaptive graph	AGCRN [4]	MTGODE [14]
Geometric graph	CLCRN [15]	STGODE [8]

direction is circular and unsuited to exceedance labelling; pressure extremes are proxies for phenomena already expressed in the wind and humidity variables; and precipitation, unusable as an hourly regression target on a series dominated by zeros, is precisely the variable detection recovers. Sparse precipitation cells fall back through a pre-registered ladder of widened pools before a global wet-hour threshold.

Each model’s multivariate output is mapped in the adapter layer to one logit per station, horizon step, and label variable, and trained as a multi-label classifier with a weighted binary cross-entropy objective whose per-variable weight $\text{pos_weight}_v = (1 - p_v)/p_v$ is the ratio of negative to positive training examples. Training positive rates are 4–5% for temperature, humidity, and wind speed but $\approx 0.28\%$ for precipitation, giving a precipitation weight near 350, some fifteen times the other variables’. The classification decision threshold for each variable is selected on validation by maximum F1 and frozen for test. STGODE sits out Phase 2: its univariate output head is incompatible with the multi-label path, so the phase compares AGCRN, CLCRN, and MTGODE over 3 models $\times$ 10 seeds = 30 runs, for a campaign total of 190.

3.3 Models and Implementation

Table 1 summarises the design: the four architectures of Section 2 instantiate every combination of spatial prior and temporal formulation. The two baselines carry direct SAWS precedent [9]; the two continuous-time models are the studied family.

3.3.1 Benchmarking framework. WeatherFLOW follows the layered design of OpenSTL [29], separating shared data and evaluation utilities from the model integrations above them. Upstream repositories are vendored at pinned commits and never modified; each integration is an adapter that wraps the published code behind a common interface for model construction, forward computation, prediction, and parameter counting, so we run the code the original authors released. Every model therefore shares the preprocessed tensors, the splits, the loss, the optimiser, the seeds, and the evaluation suite. A static thresholded-Gaussian adjacency on haversine distance [7] is built for pipeline generality, but probe tests confirm that no evaluated architecture consumes it: AGCRN and MTGODE learn their graphs from embeddings, CLCRN builds spherical k-nearest-neighbour neighbourhoods from coordinates, and STGODE constructs its own spatial and semantic graphs, matching the original implementations and Gaibie et al.’s per-model handling [9].

3.3.2 Discrete-time baselines. AGCRN and CLCRN are configured to Gaibie et al.’s SAWS-45 optima, with architecture constants they did not tune held at their published defaults [4, 15]; every value and its source appears in Table 8. CLCRN’s autoregressive decoding (Section 2.3), against the direct multi-step output of the other three, is an uncontrolled protocol difference inherited from the original architectures that plausibly disadvantages CLCRN at long horizons.

3.3.3 Continuous-time models. MTGODE follows the paper’s multi-step configuration [14]: Euler solvers for both ODEs with integration time 1.0 and step size 0.25, the graph-learner embedding dimension pinned at the paper’s 40, and the remaining values set by hyperparameter search (Section 3.4). An external implementation audit preceded all reported runs and found three defects, each corrected before relaunch; they are itemised in Appendix A.3.

STGODE’s two graphs (Section 2.4) are built here with the paper-specified raw-distance kernels, with the semantic graph computed over stations’ 24-hour diurnal profiles. The integration departs from the released code in three respects: the raw-distance kernels replace the released z-scored construction, which collapses to a near-uniform graph on a fully connected haversine distance matrix at South African scale; the training loss is the uniform harness MAE rather than the released smooth-L1; and the optimiser is plain Adam. All other released behaviour is retained, including single-step explicit Euler integration over the configured span, a setting the audit exposed as tunable (Section 3.4).

3.4 Training and Hyperparameter Selection

All models train under one harness. The objective is mean absolute error — the training loss of the MTGODE paper and of the AGCRN and CLCRN reference implementations — rather than Davidson and Moodley’s MSE [7] or STGODE’s released smooth-L1, adopted for uniformity and cross-family comparability. Optimisation is Adam, and each model uses the learning-rate schedule from its original paper; early stopping has patience 15 and gradients are norm-clipped at 5.0. Every (model, target) cell runs ten seeds {42–51} with full seeding of Python, NumPy, and PyTorch, following Davidson and Moodley [7], who report over ten independent runs and argue that a configuration should be judged partly on the deviation of its test loss across runs; the MTGODE paper likewise repeats each experiment ten times [14].

Hyperparameter selection is asymmetric, following the three configuration paths of Figure 1, and Table 8 gives every production value with its source and, where searched, its space. The baselines inherit Gaibie et al.’s SAWS-45 optima unchanged, with one repair: their AGCRN learning rate of 0.01 exceeds the ceiling of Bai et al.’s published search range and diverged on our network (best validation at epoch 1 under both loss regimes); a pre-registered single-seed diagnostic over Bai’s range selected their published optimum of 0.003, a 42% reduction in validation MAE. The continuous-time models have no SAWS precedent and were each tuned with Optuna [2]: a tree-structured Parzen estimator (multivariate, sampler seed 42) with median pruning over 30 trials, target t2m, minimising validation MAE under Gaibie et al.’s reduced two-training-year protocol. Five knobs were searched per model; all winning values and spaces appear in Table 8. MTGODE’s winner is interior on four of five knobs; STGODE’s sits at the upper bound of its grid on three, a search-boundary observation revisited in the Discussion.

No separate confirmation stage was run between search and production. The fixed-seed search re-sampled each winning configuration multiple times and every genuine runner-up trailed by 0.002 validation MAE — two orders of magnitude below the seed spread — and the multi-seed production campaign itself supplies

the repeated validation. The full production configuration for every model, with provenance and search spaces, is given in Table 8 (Appendix A.4).

3.5 Evaluation Protocol

Phase 1 reports two naive baseline models: persistence, which holds the last observed value across the horizon, and climatology, the training-period mean of each station and variable. Phase 2 reports three naive baseline models: base rate — a constant predictor at each station–variable pair’s training positive rate — persistence, and seasonal climatology, a classifier that knows only the training exceedance frequency of the matching calendar–hour cell. The last is the matched opponent of the label design: under calendar-relative labels its skill should collapse toward base rate, a pre-registered rule treats a learned model failing to exceed it on PR-AUC as a reportable finding, and seasonal-climatology PR-AUC above 1.15× prevalence flags a label-pipeline defect.

Phase-1 metrics are reported per variable in physical units after de-normalisation: MAE and RMSE per horizon, NRMSE-per-horizon curves for the horizon analysis, and a stratified report of MAE over extreme versus non-extreme test samples under train-derived 95th-percentile thresholds. The valid forecast horizon (VFH) summarises predictability as the last horizon step at which NRMSE remains at or below $\epsilon = 0.3$, computed within the direct 48-step forecast, with the grid $\epsilon \in \{0.15, 0.2, 0.3, 0.5, 0.75, 1.0\}$ reported and censoring at both ends stated explicitly. All NRMSE is normalised by the frozen training-split standard deviation of each station and variable — a leakage-safe deviation from the source lineage’s test-trajectory normalisation — so $\epsilon = 1.0$ marks climatology-equivalent error. The construct follows the valid-prediction-time family [11, 18, 22]; VFH is a within-horizon metric for direct multi-step forecasts, not that lineage’s rollout VPT, in which predictions are fed back as inputs.

A run is recorded as divergent when its best validation loss improves on its first-epoch value by less than 3%; divergent runs are retained in all reported statistics and identified in Section 4.4.

Phase-2 reporting is led by per-variable PR-AUC, with F1 at the frozen validation-selected threshold, CSI, precision, recall, and ROC-AUC as secondary metrics; a macro mean over the four label variables is given as a summary only.

Prior work in this lineage reports the mean and standard deviation of a metric over repeated runs and argues from the size of the gap. We retain that reporting and add the probability of outperforming, $P(A > B)$ — the probability that a randomly drawn run of model A attains lower test error than a randomly drawn run of model B — estimated in its unpaired Mann–Whitney form over all 10×10 cross-pairs, with 95% confidence intervals from the percentile bootstrap over 20,000 resamples [5]. A comparison separates when the lower bound of the interval exceeds 0.5, and is meaningful when the upper bound exceeds the threshold $\gamma = 0.75$ of Bouthillier et al.; overlapping intervals are reported as inconclusive rather than as evidence of equivalence. Per-sample VFH values are summarised per seed before any comparison, since overlapping forecast windows are not independent. Computational cost is reported as trainable parameter count and inference time per sample.

3.6 Ethical Considerations

The study raises no human-subjects concerns: the dataset comprises physical measurements from automatic weather stations, with no personal or identifying information, and was supplied by the South African Weather Service for research use. Station identities and coordinates are reported because they are a matter of public record and are necessary to interpret the geographic results. Two considerations do apply. The campaign consumed substantial GPU time, and the computational profile of Appendix A.5 is reported partly so that the cost of each architecture is visible alongside its accuracy. More consequentially, forecasts of extreme weather carry asymmetric risk: a missed heatwave or storm is more costly than a false alarm, and the models evaluated here are not calibrated for operational warning. The results are reported as a benchmark comparison and should not be read as evidence of readiness for deployment in a public warning system.

3.7 Experiment Infrastructure

Production runs execute on the University of Cape Town HPC cluster on NVIDIA L40S GPUs, with reproducibility resting on pinned dependencies, full seeding, and a preprocessing cache fingerprinted over the raw files and configuration.

4 Results and Analysis

The results are presented in five parts: overall forecasting accuracy; behaviour on extreme events; horizon degradation and forecast validity; stability across random seeds; and the spatial prior read against the temporal formulation. Each (model, target) cell is evaluated over ten random seeds and reported in physical units; the single divergent run is treated in §4.4.

4.1 Overall Forecasting Accuracy

Table 2 reports test MAE by forecast horizon for the four architectures and the two naive baselines, in the format of Gaibie et al. [9] extended from their 24-hour ceiling to 48 hours. MTGODE attains the lowest horizon-averaged error on temperature, relative humidity and wind speed, and is inseparable from CLCRN on surface pressure, so neither is named lowest there. Below the leaders the ordering follows the pairwise probabilities of §4.4 rather than the mean column: MTGODE and CLCRN cannot be separated on temperature or pressure, while CLCRN separates from AGCRN, and AGCRN from STGODE, on all four variables. The mean column alone would place CLCRN last on temperature, an artefact of one divergent run addressed in §4.4.

Against the naive baselines the reduction in error divides sharply by variable. Every model reaches roughly half of the better baseline’s error on temperature and relative humidity, and about three quarters on wind speed; surface pressure is the exception, where the four barely improve on persistence and §4.2 takes the tail up. Set against that gap the architectures are close: their spread is 4–6% of the mean on the three variables where the divergent run does not inflate it, so the architecture accounts for only a small share of the achievable improvement over a non-learned predictor. Figure 4 shows this closeness widening at short lead and vanishing by two days.

Table 2: Test MAE by forecast lead (means over ten seeds); best model per lead in bold. Per-seed spread in §4.4.

Lead	AGCRN	CLCRN	MTGODE	STGODE	Pers.	Clim.
t2m (°C)						
3h	1.417	1.711	1.284	1.470	2.953	5.002
6h	1.730	1.983	1.640	1.864	5.168	5.002
9h	1.844	2.095	1.789	1.978	6.574	5.003
12h	1.886	2.143	1.848	1.991	7.064	5.005
24h	2.181	2.402	2.102	2.226	2.821	5.005
48h	2.713	2.910	2.623	2.702	3.462	5.007
Avg	2.197	2.410	2.108	2.256	5.318	5.005
sp (hPa)						
3h	2.049	1.845	1.252	1.587	1.295	7.749
6h	2.434	2.183	1.702	2.169	2.080	7.749
9h	2.753	2.467	2.219	2.557	2.493	7.749
12h	3.051	2.765	2.692	2.873	2.877	7.749
24h	4.207	3.954	4.059	4.158	4.133	7.750
48h	5.365	5.279	5.464	5.383	5.506	7.754
Avg	4.042	3.817	3.818	3.948	4.005	7.750
rh (%)						
3h	6.570	6.263	6.006	6.588	11.078	19.000
6h	8.003	7.648	7.560	8.328	18.983	19.000
9h	8.619	8.244	8.243	8.964	24.111	18.998
12h	8.931	8.587	8.579	9.230	26.044	18.996
24h	10.014	9.709	9.660	10.273	12.742	18.999
48h	11.493	11.337	11.233	11.673	14.673	18.995
Avg	9.842	9.527	9.454	10.075	20.361	18.997
ws (m s ⁻¹)						
3h	0.894	0.879	0.855	0.901	1.099	1.497
6h	1.037	1.004	0.989	1.061	1.550	1.497
9h	1.102	1.057	1.048	1.124	1.820	1.497
12h	1.124	1.083	1.076	1.142	1.935	1.497
24h	1.178	1.149	1.139	1.184	1.476	1.497
48h	1.227	1.213	1.203	1.227	1.558	1.497
Avg	1.149	1.120	1.110	1.157	1.721	1.497

Station difficulty is a property of the sites, not the architectures: the four models rank the 32 stations near-identically, with per-station error correlations of 0.98 to 1.00 on every variable. Figure 8 maps the geography. The coastal–inland contrast reverses by variable — temperature and relative humidity are forecast more accurately at the coast, wind speed and surface pressure less so — and all four differences are significant.

Capacity and accuracy are not aligned. CLCRN, the smallest model, separates from the far larger STGODE on all four variables (.90 on temperature, Table 7), while MTGODE is second-cheapest on both parameters and inference time. Parameter count and inference speed do not track one another, so no ranking of the four is bought with scale (Table ??).

These are average-case results; §4.2 asks whether the ordering holds when conditions are not average.

4.2 Extreme Events

4.2.1 Extreme-Event Detection. Table 3 reports detection performance per label variable as precision, recall, F1, PR-AUC and ROC-AUC; accuracy is not reported, because at test prevalences of 0.5 to 10.7% a model that never flags an extreme already scores near-perfect accuracy. MTGODE ranks first on PR-AUC and ROC-AUC on every variable, tied with CLCRN on precipitation, and all four models beat the persistence, seasonal-climatology and base-rate baselines. ROC-AUC separates the models cleanly from those baselines while PR-AUC stays low in absolute terms, as expected when the positive class is rare; relative to prevalence the models score 2.0 to 5.6 times, highest on the rarest label.

Table 3: Extreme-event detection per label variable, means over ten seeds; best on PR-AUC and ROC-AUC in bold.

Model	P	R	F1	PR-AUC	ROC-AUC
t2m (prevalence .065)					
AGCRN	.185	.299	.228	.167	.729
CLCRN	.215	.341	.264	.205	.769
MTGODE	.243	.313	.273	.218	.780
Persistence	—	—	.238	.106	—
Seasonal clim.	—	—	.092	.061	—
Base rate	—	—	—	.065	.500
tp (prevalence .005)					
AGCRN	.044	.072	.054	.020	.754
CLCRN	.057	.087	.068	.026	.794
MTGODE	.055	.091	.068	.026	.796
Persistence	—	—	.041	.006	—
Seasonal clim.	—	—	.000	.006	—
Base rate	—	—	—	.007	.500
rh (prevalence .107)					
AGCRN	.216	.334	.262	.215	.662
CLCRN	.232	.366	.284	.235	.696
MTGODE	.250	.373	.299	.254	.716
Persistence	—	—	.301	.165	—
Seasonal clim.	—	—	.093	.093	—
Base rate	—	—	—	.117	.500
ws (prevalence .039)					
AGCRN	.119	.193	.147	.096	.674
CLCRN	.137	.205	.164	.110	.690
MTGODE	.162	.226	.189	.132	.720
Persistence	—	—	.156	.057	—
Seasonal clim.	—	—	.061	.035	—
Base rate	—	—	—	.043	.500

Table 4: Seasonal-climatology PR-AUC relative to prevalence, under static and calendar/hour-relative thresholds; 1.0 is no skill.

Variable	Static	Calendar-relative	Δ
t2m	5.98	0.93	−5.05
tp	2.44	1.30	−1.15
rh	2.57	0.87	−1.70
ws	3.08	0.90	−2.18

The label design is verified before these results are read (Table 4). Under static per-station thresholds a classifier told only the calendar-hour exceedance frequency scores well above chance — up to 5.98 times prevalence on temperature, where the annual cycle dominates — but under the calendar- and hour-relative scheme it falls to no skill on three variables (0.87 to 0.93) and only just above on precipitation (1.30), whose 0.47% prevalence makes even that estimate unstable: its base-rate ROC-AUC is 0.614 rather than the 0.50 a rate-only predictor should return. The labels therefore concede no calendar signal, and the detection results are not an artefact of seasonality.

The thresholded metrics tell a different story from the ranking ones. At the validation-frozen decision threshold persistence matches or beats the models on F1 on temperature and humidity, though MTGODE leads it by 15% to 67% on the other two variables; the models run at recall well above precision, an asymmetry a single frozen threshold penalises. The gap is not a train-to-test prevalence shift on humidity, where the threshold was set — validation and test humidity prevalence are close (9.9% against 10.7%). The tail is nonetheless non-stationary: test positive rates run about 1.6 times

Table 5: Extreme-stratified test MAE: $MAE_{\text{extreme}}/MAE_{\text{non-extreme}}$ (degradation), means over ten seeds.

Model	t2m	sp	rh	ws
AGCRN	3.77 / 2.11 (1.79)	7.50 / 3.66 (2.05)	9.97 / 9.82 (1.02)	3.54 / 1.06 (3.35)
CLCRN	4.61 / 2.29 (2.02)	6.87 / 3.48 (1.97)	9.37 / 9.53 (0.98)	3.51 / 1.03 (3.42)
MTGODE	4.02 / 2.00 (2.01)	7.65 / 3.41 (2.24)	9.54 / 9.44 (1.01)	3.46 / 1.02 (3.40)
STGODE	4.01 / 2.16 (1.86)	7.63 / 3.55 (2.15)	10.24 / 10.04 (1.02)	3.69 / 1.06 (3.49)
Persistence	8.83 / 5.12 (1.72)	6.74 / 3.79 (1.78)	18.54 / 20.54 (0.90)	4.22 / 1.63 (2.58)
Climatology	13.16 / 4.55 (2.89)	18.30 / 6.58 (2.78)	31.78 / 17.58 (1.81)	4.69 / 1.37 (3.41)

the training rate on temperature and 2.8 times on humidity, which a fixed threshold cannot follow.

4.2.2 Regression Error Under Extreme Conditions. Table 5 stratifies test MAE by whether the observation exceeds its station’s training 95th percentile, and Figure 3 plots the degradation. How far error grows in the tail is set by the variable, not the architecture — steep on wind speed, moderate on temperature and pressure, and absent on relative humidity, whose upper tail is a saturation region bounded above, where variance compresses rather than expands. The harder baseline also changes: climatology is the harder of the two pooled on three variables, but persistence is harder in every tail, because a predictor fixed at the training mean is furthest from the observation exactly when it is extreme, and on humidity persistence alone improves in the tail.

The pooled ordering does not survive stratification on three of the four variables: a discrete-recurrence model holds the lowest extreme-regime error on temperature, pressure and humidity, and MTGODE — lowest pooled — leads the tail only on wind speed. Neither AGCRN nor CLCRN separates from the other on temperature, yet each beats both continuous-time models there: AGCRN attains the lower temperature extreme MAE in 90% of cross-pairs against MTGODE and 93% against STGODE, and CLCRN in 82% and 84%, over ten seeds. On pressure CLCRN separates from all three others. The divergent run lifts CLCRN’s temperature cell in Table 5, so Figure 3 plots the interquartile mean and says so. Average-case accuracy therefore does not predict accuracy in the tail.

A named case study of the December 2024 heatwave reproduces these findings on a single event — the tail ordering, the flattened amplitude, and a detection recall that collapses with lead — and is reported in full in the Supplementary Information (§A.1). That lead-dependence motivates the horizon analysis that follows.

4.3 Horizon Degradation and Forecast Validity

Figure 4 plots test MAE against lead time. Error growth is set by the target variable far more than by the architecture — shallow on wind speed, steep on surface pressure — and the spread between architectures narrows sharply with lead, from 47.3% of the mean near the first lead to 3.5% by two days on pressure. MTGODE has the steepest growth on every variable, so its advantage sits at short lead and shrinks across the horizon; on pressure it leads at the shortest leads and finishes last. Because the converged CLCRN curve dips below MTGODE at the nine- and twelve-hour steps on temperature, MTGODE is not lowest at every sub-horizon.

The models stay below persistence even at its diurnal minima, where the previous day’s value is free to persistence, so they capture

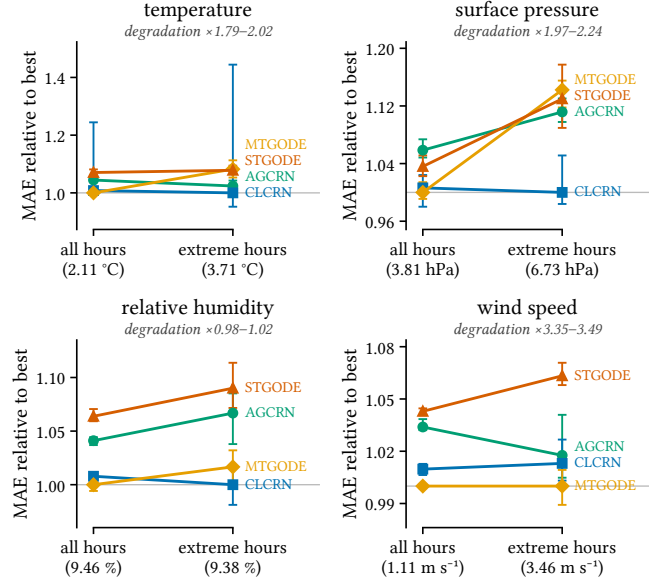


Figure 3: Extreme-hour against all-hour MAE per variable, relative to the best cell, one marker per model; subtitles give Table 5’s degradation range. Plotted on the interquartile mean, so CLCRN’s temperature point is not inflated by the divergent run.

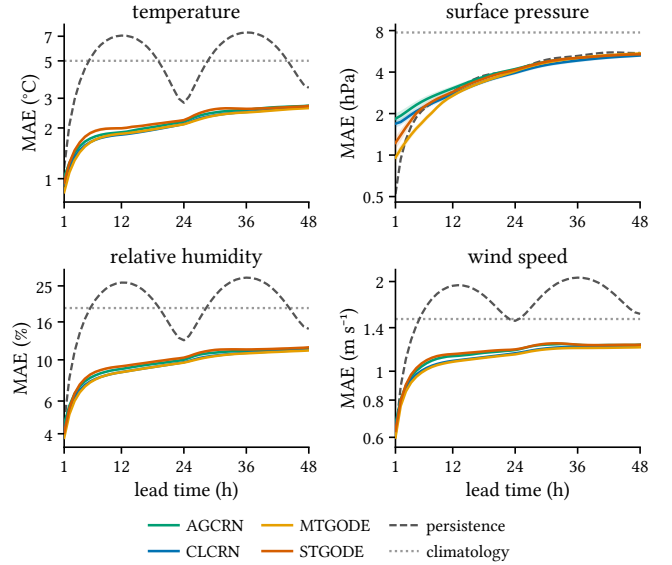


Figure 4: Test MAE against forecast lead for each target variable, on log axes, with the persistence and climatology baselines; means over ten seeds. Persistence’s diurnal sawtooth and the narrowing between architectures with lead are both visible.

Table 6: Valid forecast horizon (h): the last lead at which pooled NRMSE stays at or below ε , mean over ten seeds.

Model	$\varepsilon = 0.3$		$\varepsilon = 0.5$	
	t2m / sp	rh / ws	t2m / sp	rh / ws
AGCRN	2.0 / 0	1.0 / 0	26.2 / 0	5.2 / 0.9
CLCRN	2.6 / 0	1.0 / 0	25.6 / 0	6.6 / 1.0
MTGODE	3.0 / 0	1.0 / 0	29.2 / 2.6	7.0 / 1.0
STGODE	1.9 / 0	1.0 / 0	25.1 / 0.4	4.7 / 1.0
Persistence	1 / 0	0 / 0	2 / 3	1 / 1
Climatology	0 / 0	0 / 0	0 / 0	0 / 0

more than the diurnal cycle. Surface pressure is the exception at short lead: persistence is the stronger predictor below about eight hours, and at the first lead every model’s pressure error is 1.8 to 3.5 times persistence’s — yet all four overtake it by two days.

Table 6 reports the valid forecast horizon. At the pre-registered $\varepsilon = 0.3$ it is 0 h for every model on pressure and wind speed and 1 h on humidity — 8 of the 16 cells sit at 0 h, since first-lead NRMSE already exceeds the threshold there. At $\varepsilon = 0.5$ the horizon recovers the mean-error ordering on temperature and humidity but not on pressure, where CLCRN reaches 0 h yet ties for lowest MAE, and persistence outlasts every model.

The architectures therefore differ most at short lead and converge by two days, MTGODE’s lead is a short-lead phenomenon, and at the pre-registered threshold forecast validity does not separate the models. Whether the ordering is a property of the architectures or of their initialisation is the question of §4.4.

4.4 Stability Across Random Seeds

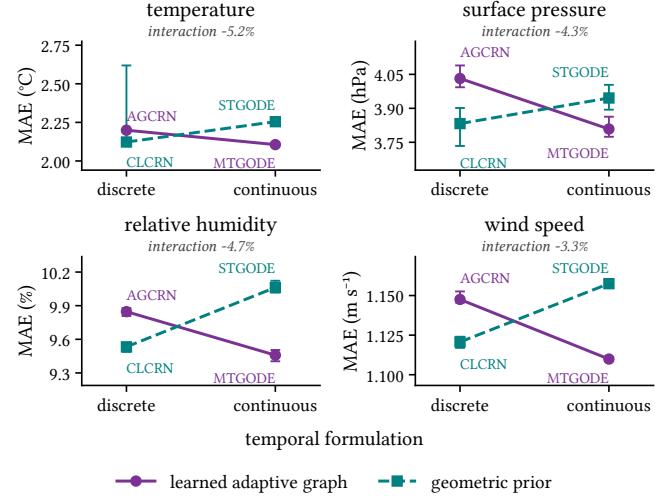
Fifteen of the sixteen (model, target) regression cells are stable under reinitialisation, with coefficients of variation of test MAE from 0.18% to 3.02%. Surface pressure carries the highest variation for every model, so a fixed seed budget buys the least precision there; where the signal is strongest, as on wind speed, independent runs converge to nearly the same error.

The sixteenth cell is the exception. On temperature CLCRN’s seed 50 converged to the training mean: its validation loss never fell below its first-epoch value and its test error equals the climatological forecast’s (5.003 against 5.005 °C). The failure reproduced under the same seed and is one run in 190; it is retained in every statistic, after Schmidt et al. and Mosbach et al., who keep and diagnose such runs rather than discard them [19, 25]. Its effect is not a blemish to remove but the reliability finding itself: it inflates the cell’s standard deviation to $\pm .912$ against $\pm .013$ to $\pm .029$ elsewhere and lifts its mean by 0.29 °C, about twice the widest gap between converged architectures (0.148 °C), and it struck the strongest-signal variable.

Read pairwise, 22 of the 24 comparisons separate at the 95% level (Table 7); the two that do not are MTGODE against CLCRN on temperature and pressure, so the leaders tie on those variables while MTGODE’s lead is decisive on humidity and wind speed. Across the 40 (target, seed) cells MTGODE takes first place in 31 and CLCRN in the other 9, and no other model leads anywhere. The one order that inverts is AGCRN against STGODE on pressure, which Figure 7 shows as the single point left of the reference line.

Table 7: Probability that a random run of the first model attains lower test MAE than the second; bold marks 95% intervals excluding 0.5.

Comparison	t2m	sp	rh	ws
MTGODE > CLCRN	.69 [.43,.91]	.56 [.28,.83]	.82 [.61,.97]	.99 [.94,1.0]
MTGODE > AGCRN	1.0 [1.0,1.0]	1.0 [1.0,1.0]	1.0 [1.0,1.0]	1.0 [1.0,1.0]
MTGODE > STGODE	1.0 [1.0,1.0]	.93 [.80,1.0]	1.0 [1.0,1.0]	1.0 [1.0,1.0]
CLCRN > AGCRN	.88 [.67,1.0]	.98 [.91,1.0]	1.0 [1.0,1.0]	1.0 [1.0,1.0]
CLCRN > STGODE	.90 [.70,1.0]	.83 [.63,.97]	1.0 [1.0,1.0]	1.0 [1.0,1.0]
AGCRN > STGODE	1.0 [1.0,1.0]	.20 [.04,.42]	1.0 [1.0,1.0]	.87 [.65,1.0]

**Figure 5: Factorial read of the design: each panel plots a variable’s four cell means against the temporal axis, one line for the learned adaptive graph (AGCRN→MTGODE) and one for the geometric prior (CLCRN→STGODE); the subtitle gives the interaction. Interquartile means over ten seeds.**

The bottom of the ranking is thus reproducible and the top is not on temperature and pressure, variability tracks predictability rather than architecture, and one run in 190 failed outright — so the factorial reading of §4.5 is made on a robust aggregate rather than the mean.

4.5 Spatial Prior versus Temporal Formulation

The four architectures instantiate the four cells of the design in Table 1, and Figure 5 reads them factorially. The lines cross in every panel: moving from a discrete to a continuous-time formulation lowers error under a learned adaptive graph and raises it under a geometric prior. On temperature the crossing holds under the interquartile mean and under the converged means, and is masked only by the divergent run in the ten-seed mean.

The interaction is negative on all four variables, from -3.28% to -5.19% of the four-cell mean, and each panel of Figure 5 carries its own value. It is computed from the interquartile mean [1], robust to the outlier run, rather than the raw mean, which the divergent run pushes positive on temperature (+1.46%); the interquartile mean departs from the raw mean by at most 0.40% in the other cells, so it rescales the estimator uniformly rather than excising a run. Both

marginal effects are small by comparison, so the interaction exceeds either axis alone on every variable.

The pattern is diagonal, not marginal: the cells pairing a learned graph with continuous time and a geometric prior with discrete recurrence fall below each variable’s mean, while the two crossed pairings fall above it. That the two design axes cannot be read independently is confounded here with differences the cells also carry — STGODE takes a single Euler step, and CLCRN decodes autoregressively (§3.3.3, §3.3.2) — so the effect of the temporal formulation depends on the spatial prior for this architecture set.

5 Discussion and Conclusions

Continuous-time formulations transfer to South African station data, but only with a learned graph. MTGODE leads on three variables and is inseparable from CLCRN on the fourth (§4.1), the first demonstration that a continuous-time graph model works on an irregular station network; yet STGODE is continuous-time too and is the weakest of the four, so continuity alone does not carry the result.

The factorial reading (§4.5) explains why: continuous time lowers error only under a learned graph, so a claim about either design axis is conditional on the other. This is consistent with Jin et al.’s positioning of MTGODE against STGODE on continuous temporal aggregation [14] without isolating it, since STGODE’s cell also differs in spatial prior, takes a single Euler step, and won its search at the grid bound on three of five knobs — a lower bound on that corner. And where Jin et al. report MTGODE’s edge as largest at long horizons, here it is largest at short lead (§4.3). Continuous-time graph formulations thus work when the graph is learned, and neither axis can be judged without the other.

No single criterion selects one architecture. On average error, on forecast validity at the looser threshold, and on detection ranking, MTGODE leads (§4.1, §4.3, §4.2.1); in the tail a discrete-recurrence model has the lower regression error on three of four variables (§4.2.2); and capacity buys none of it, the smallest model being competitive and the largest not (§4.1). The criteria fall into two families — average-case skill and tail amplitude — that do not agree, so the choice of model follows the intended use rather than a single ranking. The pre-registered validity threshold, already exceeded at the first lead on pressure and wind speed, is too strict to arbitrate between them.

Outside average conditions all four architectures smooth the tails. The stratified regression, the detection collapse at long lead and the named heatwave agree that no model reaches the extreme (§4.2.2, §A.1), and the case study shows the mechanism directly: each model follows the diurnal shape but flattens its amplitude, so error concentrates exactly where the observation is most extreme. This is the double-penalty behaviour of a smooth predictor under a squared-error objective [28], compounded by the scarcity of tail samples that leaves the high quantiles under-fit [20]. The deficit therefore sits in the training objective, not the architecture — a squared-error loss rewards hedging toward the mean — so the remedy is a tail-weighted or probabilistic objective rather than a larger model. Average error does not predict tail behaviour, and models should not be chosen on it when extremes are the target.

Static exceedance thresholds are not comparable across variables. Because a fixed per-station percentile absorbs a variable’s calendar cycle, a seasonal classifier scores far above chance under it, and by an amount that differs sharply by variable (§4.2.1) — the comparability problem Ashford raises for static extreme labels [3]. The calendar- and hour-relative scheme removes that signal, restoring no skill on three of the four variables; precipitation, an order of magnitude rarer, is the exception. Figure 6’s shaded bands are the scheme in operation — the exceedance region tracking the hour and the calendar rather than a fixed line. The design achieves comparability on three of four variables.

Set against the South African lineage, the accuracy results transfer in scale but not in order. Gaibie et al. rank AGCRN ahead of CLCRN on temperature, humidity and wind speed from three runs without dispersion [9]; here CLCRN separates from AGCRN on all four, the studies agreeing only on pressure. Several factors are confounded — the eastern rather than Cape network, the inherited optima, and three runs against ten seeds — and pressure is the sharpest contrast, most predictable in Gaibie’s normalised units but least predictable here in physical ones, because min–max normalisation over a wide altitude range shrinks pressure error by construction. Extended past the lineage’s 24-hour ceiling, the architectures converge: what transfers is the size of the differences, not their order (§4.1).

Replication separates a reproducible ranking from a single run’s. The bottom of the ordering is stable across seeds and the top is not (§4.4); the one divergence in 190 would have been a third of a cell under Gaibie’s three runs, and it flips the sign of the mean-based temperature interaction, which is why the factorial rests on a robust aggregate. That the failed run is a CLCRN run is consistent with CLCRN’s authors naming training instability an open problem of the method [15], though one run cannot establish a rate. Site, not graph construction, sets which stations are hard, consistent with the coastal and inland patterns reported by Davidson and Moodley and by Gaibie et al. [7, 9]. Reproducing a published architecture is also not the same as running its code — an audit found three defects in one integration (§3.3.3). A ranking should be read across seeds, not from a single run.

6 Limitations and Future Work

Several limits bound the comparison. Evaluation uses a single chronological hold-out rather than walk-forward validation, and hyperparameter selection is asymmetric — inherited optima for the baselines, a search for the continuous-time models. The network is confined to the summer-rainfall east, and the extreme-event picture is narrower still: it rests on one named event, on single-hour labels with no duration structure, and on a non-stationary tail that a frozen threshold cannot track.

Each points to a step: a single architecture whose spatial prior and temporal formulation switch independently would turn the interaction into a controlled effect (§4.5); a tail-weighted or probabilistic objective would address the smoothing every model shows (§5); and duration-aware labels with recalibrated thresholds would suit operational use.

Acknowledgments

The authors thank Dr. Deshendra Moodley for supervision; the WeatherFLOW team for shared baselines and data acquisition; UCT ICTS High-Performance Computing; and the South African Weather Service for station data.

References

- [1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. 2021. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, 29304–29320. arXiv:2108.13264. <https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html>
- [2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. ACM, New York, NY, 2623–2631. doi:10.1145/3292500.3330701
- [3] Adam Ashford, Fahad Ayaz, Muhammad Zeeshan Shakir, Naeem Ramzan, Michael Grebreslasie, Serestina Viriri, David Ndzi, Natalie Dickinson, Llinos Haf Spencer, Mary Lynch, and Saloshni Naidoo. 2026. Machine Learning for Heatwave Prediction: A Global Scoping Review of Environmental Predictors and Modelling Practices. *Forecasting* 8, 4 (2026), 63. doi:10.3390/forecast8040063
- [4] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. 2020. Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, 17804–17815. <https://proceedings.neurips.cc/paper/2020/hash/ce1aad92b939420fc17005e5461e6f48-Abstract.html>
- [5] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. 2021. Accounting for Variance in Machine Learning Benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, Vol. 3. 747–769. https://proceedings.mlsys.org/paper_files/paper/2021/hash/0184b0cd3cfb185989f858a1d9f5c1eb-Abstract.html
- [6] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. 2018. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, 6571–6583. <https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html>
- [7] Mikhail Davidson and Deshendra Moodley. 2022. ST-GNNs for Weather Prediction in South Africa. In *Artificial Intelligence Research (SACAIR 2022) (Communications in Computer and Information Science, Vol. 1734)*. Springer, 93–107. doi:10.1007/978-3-031-22321-1_7
- [8] Zheng Fang, Qingqing Long, Guojie Song, and Kunqing Xie. 2021. Spatial-Temporal Graph ODE Networks for Traffic Flow Forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, New York, NY, 364–373. doi:10.1145/3447548.3467430
- [9] Adeb Gaibie, Hamza Amir, Irene Nandutu, and Deshendra Moodley. 2024. Predicting and Discovering Weather Patterns in South Africa Using Spatial-Temporal Graph Neural Networks. In *Artificial Intelligence Research (SACAIR 2024) (Communications in Computer and Information Science, Vol. 2326)*. Springer, 144–160. doi:10.1007/978-3-031-78255-8_9
- [10] Asadullah Hill Galib, Andrew McDonald, Tyler Wilson, Lifeng Luo, and Pang-Ning Tan. 2022. DeepExtrema: A Deep Learning Approach for Forecasting Block Maxima in Time Series Data. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, 2980–2986. doi:10.24963/ijcai.2022/413
- [11] William Gilpin. 2023. Model Scale Versus Domain Knowledge in Statistical Forecasting of Chaotic Systems. *Physical Review Research* 5 (2023), 043252. doi:10.1103/PhysRevResearch.5.043252
- [12] Stefan W. Grab and David J. Nash. 2024. A New Flood Chronology for KwaZulu-Natal (1836–2022): The April 2022 Durban Floods in Historical Context. *South African Geographical Journal* 106, 4 (2024), 476–497. doi:10.1080/03736245.2023.2193758
- [13] Mohammad Amin Havaei, Vahid Shahhosseini, and Reza Maknoon. 2025. Continuous-Time Air Pollutant Forecasting Using Multi-Timescale Attention Neural Ordinary Differential Equations (MA-NODE). *Scientific Reports* 15, 1 (2025), 44178. doi:10.1038/s41598-025-27903-z
- [14] Ming Jin, Yu Zheng, Yuan-Fang Li, Siheng Chen, Bin Yang, and Shirui Pan. 2023. Multivariate Time Series Forecasting with Dynamic Graph Neural Ordinary Differential Equations. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2023), 9168–9180. arXiv:2202.08408. doi:10.1109/TKDE.2022.3221989
- [15] 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. AAAI Press, 7470–7478. doi:10.1609/aaai.v36i7.20711
- [16] Zewen Liu, Xiaoda Wang, Bohan Wang, Zijie Huang, Carl Yang, and Wei Jin. 2025. Graph ODEs and Beyond: A Comprehensive Survey on Integrating Differential Equations with Graph Neural Networks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*. ACM, 6118–6128. doi:10.1145/3711896.3736559
- [17] Innocent Mbokodo, Mary-Jane Bopape, Hector Chikooore, Francois Engelbrecht, and Nthaduleni Nethengwe. 2020. Heatwaves in the Future Warmer Climate of South Africa. *Atmosphere* 11, 7 (2020), 712. doi:10.3390/atmos11070712
- [18] Henok Tenaw Moges, Charalampos Skokos, and Deshendra Moodley. 2026. ChaosNetBench: Benchmarking Spatio-Temporal Graph Neural Networks on Chaotic Lattice Dynamics. *arXiv preprint arXiv:2605.09676* (2026). doi:10.48550/arXiv.2605.09676
- [19] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *International Conference on Learning Representations (ICLR)*. arXiv:2006.04884. <https://openreview.net/forum?id=nzplWnVAYah>
- [20] Leonardo Olivetti and Gabriele Messori. 2024. Advances and Prospects of Deep Learning for Medium-Range Extreme Weather Forecasting. *Geoscientific Model Development* 17, 6 (2024), 2347–2358. doi:10.5194/gmd-17-2347-2024
- [21] Olivier C. Pasche, Jonathan Wider, Zhongwei Zhang, Jakob Zscheischler, and Sebastian Engelke. 2025. Validating Deep Learning Weather Forecast Models on Recent High-Impact Extreme Events. *Artificial Intelligence for the Earth Systems* 4, 1 (2025), e240033. doi:10.1175/AIES-D-24-0033.1
- [22] Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott. 2018. Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach. *Physical Review Letters* 120, 2 (2018), 024102. doi:10.1103/PhysRevLett.120.024102
- [23] Prabhat, Karthik Kashinath, Mayur Mudigonda, Sol Kim, Lukas Kapp-Schwoerer, Andre Graubner, et al. 2021. ClimateNet: An Expert-Labeled Open Dataset and Deep Learning Architecture for Enabling High-Precision Analyses of Extreme Weather. *Geoscientific Model Development* 14 (2021), 107–124. doi:10.5194/gmd-14-107-2021
- [24] Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. 2020. WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems* 12, 11 (2020), e2020MS002203. doi:10.1029/2020MS002203
- [25] Robin M. Schmidt, Frank Schneider, and Philipp Hennig. 2021. Descending through a Crowded Valley – Benchmarking Deep Learning Optimizers. In *Proceedings of the 38th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 9367–9376. arXiv:2007.01547. <https://proceedings.mlr.press/v139/schmidt21a.html>
- [26] South African Weather Service. [n.d.]. What is a Heat Wave? <https://www.weathersa.co.za/home/weatherques>. Accessed August 2026.
- [27] A. T. Sree Dhevi. 2014. Imputing Missing Values Using Inverse Distance Weighted Interpolation for Time Series Data. In *2014 Sixth International Conference on Advanced Computing (ICoAC)*. IEEE, 255–259. doi:10.1109/ICoAC.2014.7229721
- [28] Christopher Subich, Syed Zahid Husain, Leo Separovic, and Jing Yang. 2025. Fixing the Double Penalty in Data-Driven Weather Forecasting Through a Modified Spherical Harmonic Loss Function. In *Proceedings of the 42nd International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 267)*. PMLR, 57191–57211. arXiv:2501.19374. <https://proceedings.mlr.press/v267/subich25a.html>
- [29] Cheng Tan, Siyuan Li, Zhangyang Gao, Wenfei Guan, Zedong Wang, Zicheng Liu, Lirong Wu, and Stan Z. Li. 2023. OpenSTL: A Comprehensive Benchmark of Spatio-Temporal Predictive Learning. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*. Curran Associates, arXiv:2306.11249. https://proceedings.neurips.cc/paper_files/paper/2023/hash/dcbff44d1130e75d09d3930411c23e1-Abstract.html
- [30] Yogesh Verma, Markus Heinonen, and Vikas Garg. 2024. ClimODE: Climate and Weather Forecasting with Physics-Informed Neural ODEs. In *International Conference on Learning Representations (ICLR)*. arXiv:2404.10024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/21f3c05cf5afcaebd8a71d7e98a1b171-Abstract-Conference.html
- [31] Suviní Welandawe, Y. H. P. P. Priyadarshana, Nipuna Senanayake, and E. Nishchitha S. Silva. 2025. Machine Learning Techniques for Marine Heatwave Prediction: A Comprehensive Review. *Intelligent Marine Technology and Systems* 3, 1 (2025), 28. doi:10.1007/s44295-025-00076-1
- [32] Tyler Wilson, Pang-Ning Tan, and Lifeng Luo. 2018. A Low Rank Weighted Graph Convolutional Approach to Weather Prediction. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 627–636. doi:10.1109/ICDM.2018.00078
- [33] 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 (KDD)*. ACM, New York, NY, 753–763. doi:10.1145/3394486.3403118
- [34] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 2019. Graph WaveNet for Deep Spatial-Temporal Graph Modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*. International

- Joint Conferences on Artificial Intelligence Organization, 1907–1913. doi:10.24963/ijcai.2019/264
- [35] Wanghan Xu, Kang Chen, Tao Han, Hao Chen, Wanli Ouyang, and Lei Bai. 2024. ExtremeCast: Boosting Extreme Value Prediction for Global Weather Forecast. *arXiv preprint arXiv:2402.01295* (2024). doi:10.48550/arXiv.2402.01295
- [36] Qingtian Zeng, Chao Wang, Geng Chen, Hua Duan, and Shuihua Wang. 2022. For the Aged: A Novel PM2.5 Concentration Forecasting Method Based on Spatial-Temporal Graph Ordinary Differential Equation Networks in Home-Based Care Parks. *Frontiers in Environmental Science* 10 (2022), 956020. doi:10.3389/fenvs.2022.956020
- [37] Xuebin Zhang, Lisa Alexander, Gabriele C. Hegerl, Philip Jones, Albert Klein Tank, Thomas C. Peterson, Blair Trewin, and Francis W. Zwiers. 2011. Indices for Monitoring Changes in Extremes Based on Daily Temperature and Precipitation Data. *Wiley Interdisciplinary Reviews: Climate Change* 2, 6 (2011), 851–870. doi:10.1002/wcc.147
- [38] Xun Zhu, Yutong Xiong, Ming Wu, Gaozhen Nie, Bin Zhang, and Ziheng Yang. 2023. Weather2K: A Multivariate Spatio-Temporal Benchmark Dataset for Meteorological Forecasting Based on Real-Time Observation Data from Ground Weather Stations. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) (Proceedings of Machine Learning Research, Vol. 206)*. PMLR, 2704–2722. <https://proceedings.mlr.press/v206/zhu23a.html>

A Supplementary Information

A.1 Case Study: A Named Historical Event

The most severe temperature extreme in the test partition is the December 2024 heatwave, for which the Weather Service issued warnings across six provinces. It spans 129 hourly steps from 8 to 13 December, with 19 of the 32 stations in simultaneous exceedance at its peak on 10 December. Figure 6 shows it at Makatini Research Centre, the network’s hottest reading of the episode at 41.2 °C, with forecasts drawn at a fixed 24-hour lead from each model’s median seed — by test MAE for the regression models, by temperature PR-AUC for the detection models.

Every model under-forecasts the peak, missing the Makatini maximum by 5.1 to 7.3 °C, and none reaches the observation at any hour; across the network they reproduce 74% to 87% of the observed daily temperature range. Event MAE inverts the pooled ordering — MTGODE falls from first at this horizon to third and CLCRN rises from last to second — and splits the models by temporal formulation: the discrete-recurrence AGCRN is more accurate during the event than on average (a ratio of 0.93) while the continuous-time pair is less accurate (1.16 and 1.15), with CLCRN between. Within the event, error on exceedance hours exceeds error on the rest by 1.66 to 2.20 times, matching the 1.79 to 2.02 measured over the whole test partition.

At its frozen threshold CLCRN recovers 76% of the network’s exceedance hours, against 63% for AGCRN and 55% for MTGODE, at precisions of 0.40 to 0.43, and all three rank exceedance hours above the rest. Recall falls sharply with lead — CLCRN from 86% at short lead to 14% at two days, AGCRN from 60% to 45%, MTGODE from 49% to 31% — the collapse that motivates §4.3. CLCRN’s network-wide advantage rests on the loosest decision bar and saturates at the hottest station, not on lower regression error, so ranking extremes and forecasting their magnitude are not the same task. Stratified error, detection and this named event agree: no model reaches the tail.

A.2 Figures

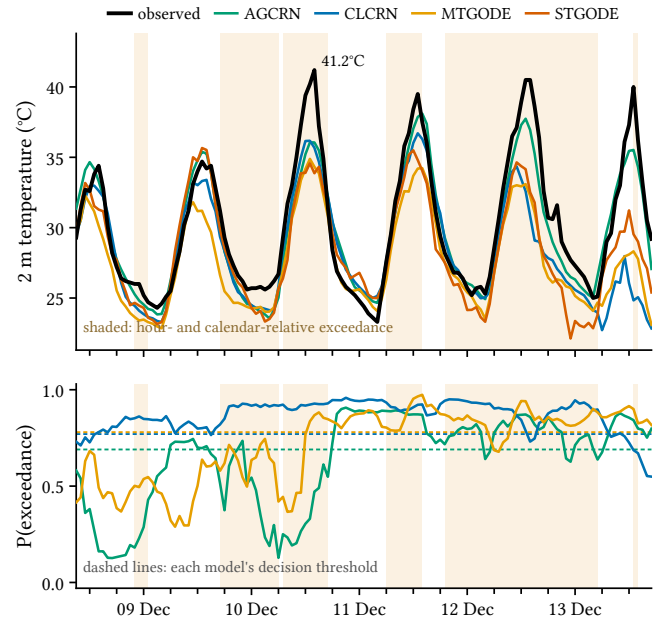


Figure 6: The December 2024 heatwave at Makatini, at 24-hour lead. Top: observed temperature and each model’s forecast, with hour- and calendar-relative exceedance shaded. Bottom: each detection model’s exceedance probability and its frozen decision threshold.

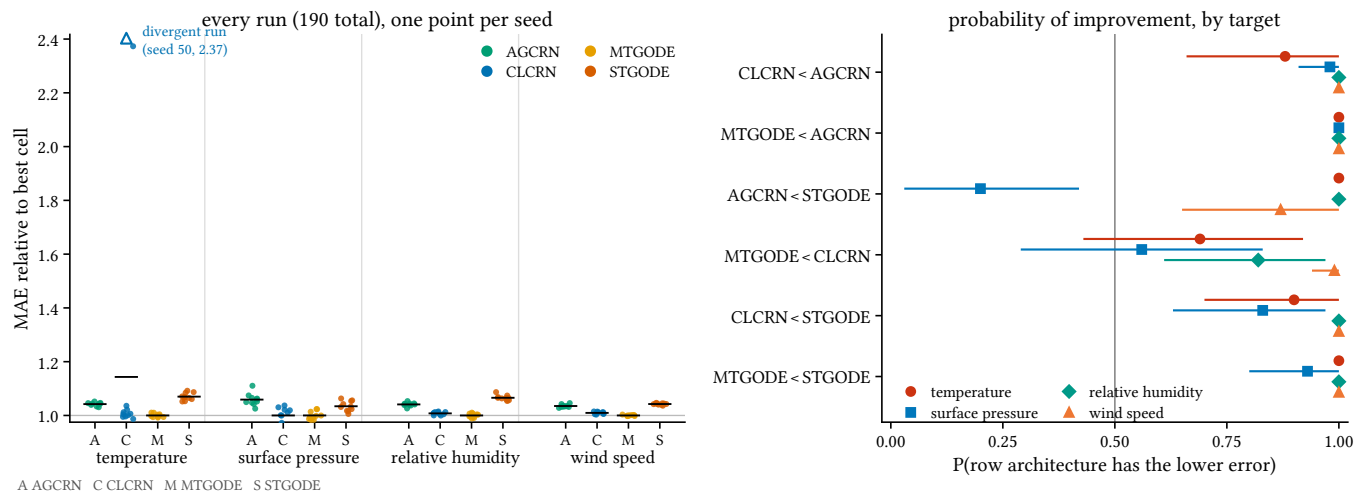


Figure 7: Left: every run's MAE relative to the best cell, one point per seed (190 in total), by variable and model, with the divergent CLCRN temperature run marked. Right: the probability that a randomly drawn run of one model beats another, by target, with 95% intervals.

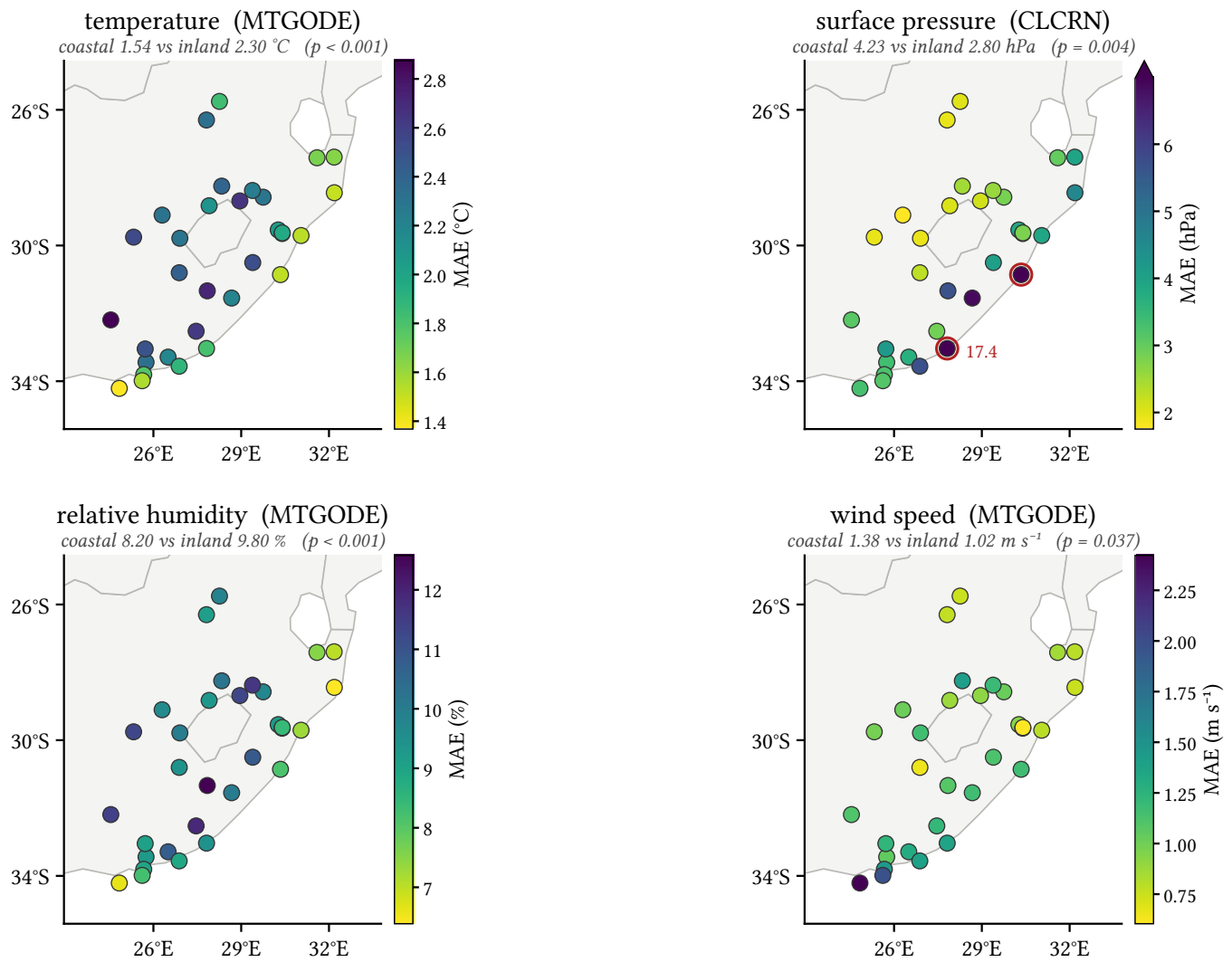


Figure 8: Per-station test MAE for one model per variable, named in each panel, one marker per station on the network map; subtitles give coastal against inland medians and a Mann–Whitney p . The four models' per-station error ranks correlate 0.98 to 1.00, so difficulty is shared.

A.3 MTGODE Implementation Audit

An external audit of the MTGODE integration preceded all reported runs and corrected three defects. First, the adjoint backward pass structurally excluded the graph learner’s parameters from gradient flow, so the learned adjacency remained at its random initialisation and the model trained as a graph-frozen variant. Second, the graph-learner embedding dimension had been conflated with the

convolutional width. Third, the spatial vector field was implemented at half the magnitude Eq. (3) requires. No result produced before these corrections is reported.

A.4 Production Configuration

A.5 Computational Profile

Table 8: Production configuration per model, with provenance and search spaces (braces). Sources: G24 = Gaibie et al.; B20 = Bai et al.; L22 = Lin et al.; J22 = Jin et al.; F21 = Fang et al.; HPO = Optuna winner.

Parameter	AGCRN	CLCRN	MTGODE	STGODE	Provenance
Hidden width / RNN units	64	32	64	—	G24 · G24 · HPO {32, 64, 128} · upstream defaults
Layers	2	2	2	—	G24 · G24 · J22 · upstream defaults
Node-embedding dim.	10	16	40	—	B20 · L22 · J22 (pinned per audit)
Graph construction	learned (DAGG)	spherical k -NN, $k = 25$; kernel layers [4, 8]; max view 2	learned, top- $k = 20$, $\alpha = 3$	Gaussian $\sigma = 500$ km, cutoff 500 km; DTW $\varepsilon = 0.8$	B20 · L22 · HPO {10, 13, 20} + J22 · HPO {100, 200, 500} / {0.3, 0.5, 0.8}
Chebyshev order	2	—	—	—	B20
Dropout	—	—	0.1	—	HPO {0.1, 0.2, 0.3, 0.4}
ODE solver	—	—	Euler, $T = 1.0$, $\Delta t = 0.25$ (both ODEs); CGP $\alpha = 2.0$	Euler, single step, span 12.0	J22 + audit · HPO {3, 6, 12}
Learning rate	0.003	0.05	6×10^{-3}	6×10^{-3}	B20 optimum (repair; §3.4) · G24 · HPO {1e-4, 3e-4, 1e-3, 3e-3, 6e-3, 1e-2} · HPO {1e-4, 3e-4, 1e-3, 2e-3, 3e-3, 6e-3}
Batch size	64	32	128	8	G24 · G24 · HPO {32, 64, 128} · HPO {8, 16, 32}
Epoch cap	40	40	200	200	G24 · G24 · J22 · F21

