hydrology

MDPI

Article

Hybrid Deep Learning Architectures for Multi-Horizon Precipitation Forecasting in Mountainous Regions: Systematic Comparison of Component-Combination Models in the Colombian Andes

Manuel Ricardo Pérez Reyes [1,2,*], Marco Javier Suárez Barón [2] and Óscar Javier García Cabrejo [3]

[1] Programa de Doctorado en Ingeniería, Grupo de Investigación GALASH, Universidad Pedagógica y Tecnológica de Colombia, Sede Sogamoso 152210, Colombia
[2] Escuela de Ingeniería de Sistemas y Computación, Grupo de Investigación GALASH, Universidad Pedagógica y Tecnológica de Colombia, Sede Sogamoso 152210, Colombia; marco.suarez@uptc.edu.co
[3] Escuela de Ingeniería Geológica, Grupo de Investigación GALASH, Universidad Pedagógica y Tecnológica de Colombia, Sede Sogamoso 152210, Colombia; oscar.garcia04@uptc.edu.co
* Correspondence: manuelricardo.perez@uptc.edu.co

Abstract

Forecasting monthly precipitation in mountainous terrain poses challenges that push conventional deep learning approaches to their limits: convective processes operate locally while orographic effects span entire drainage basins. We compare three architecture families on precipitation prediction across the Colombian Andes: ConvLSTM (convolutional recurrent), FNO-ConvLSTM (spectral–temporal), and GNN-TAT (graph attention LSTM). Using CHIRPS v2.0 and SRTM topography for Boyacá department (61 × 65 grid, 3965 nodes), we evaluate 39 configurations across feature bundles (BASIC, KCE elevation clusters, and PAFC autocorrelation lags) and horizons from 1 to 12 months. GNN-TAT matches ConvLSTM accuracy (R^2: 0.628 vs. 0.642; RMSE: 82.29 vs. 79.40 mm) with 95% fewer parameters (∼98K vs. 2.1M). Across configurations, GNN-TAT produces a lower mean RMSE (92.12 vs. 112.02 mm; $p = 0.015$) and a 74.7% lower variance. The explicit graph structure, with edges weighted by elevation similarity, appears to reduce sensitivity to hyperparameter choices. Pure FNO struggles with precipitation's spatial discontinuities ($R^2 = 0.206$), though adding a ConvLSTM decoder recovers much of the lost skill ($R^2 = 0.582$). Elevation clustering improves GNN-TAT significantly ($p = 0.036$) but not ConvLSTM, suggesting that feature design should match the spatial encoding paradigm. ConvLSTM achieves peak accuracy on local patterns; GNN-TAT provides robust predictions with interpretable spatial reasoning. These complementary strengths motivate stacking ensembles that combine grid-based and graph-based representations.

Keywords: hybrid deep learning; component combination models; mountainous terrain; orographic precipitation; graph neural networks; ConvLSTM; temporal attention; CHIRPS; Colombian Andes; precipitation forecasting

Check for updates

Academic Editors: Xiaofeng Liu, Yi Hong, Ryan C. Johnson and Quan Guo

Received: 18 January 2026
Revised: 15 February 2026
Accepted: 16 February 2026
Published: 18 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Accurate multi-month precipitation forecasting remains a critical challenge for water resource management, agricultural planning, and disaster preparedness in mountainous regions worldwide [1]. The Colombian Andes, characterized by complex orographic gradients exceeding 3500 m elevation difference within short horizontal distances, presents

particularly demanding conditions where traditional numerical weather prediction (NWP) models struggle to capture localized precipitation patterns [2]. In the department of Boyacá, sparse rain gauge networks and the bimodal precipitation regime driven by the Intertropical Convergence Zone (ITCZ) migration necessitate data-driven approaches that can leverage satellite-derived precipitation estimates and digital elevation models (DEMs).

Recent advances in deep learning have transformed precipitation forecasting, from the seminal ConvLSTM architecture for nowcasting [3] to sophisticated Graph Neural Networks (GNNs) capable of learning non-Euclidean spatial relationships [4,5]. The emergence of foundation models like GraphCast [4] and Pangu-Weather [6] has demonstrated that machine learning can achieve comparable or superior skills compared to physics-based NWP systems at a fraction of the computational cost. However, these global-scale models are not optimized for regional applications in complex terrain, where local topographic effects dominate precipitation variability.

This study addresses the gap between global weather prediction models and regional precipitation forecasting by comparing three hybrid deep learning architectures. The first, ConvLSTM, couples convolutional spatial encoding with recurrent temporal gates, a combination that has proven effective for radar nowcasting since Shi et al.'s original 2015 paper. We enhanced it with attention mechanisms and residual connections. The second approach integrates Fourier Neural Operators with a ConvLSTM decoder [7], motivated by the hypothesis that spectral methods might capture large-scale atmospheric patterns more efficiently. The third is a Graph Neural Network with Temporal Attention (GNN-TAT) that we developed specifically for this terrain: the graph structure encodes elevation relationships between grid cells, allowing information to flow preferentially between orographically similar locations rather than just spatial neighbors.

Following the taxonomy of Perez et al. [8], these architectures integrate distinct processing modules end-to-end to leverage complementary inductive biases. Other hybridization strategies exist, including signal decomposition before modeling, metaheuristic parameter tuning, and post hoc bias correction. Integrating components within a single differentiable pipeline allows each module to specialize. This matters for mountainous precipitation, where orographic forcing dominates spatial variability and ITCZ migration drives temporal seasonality.

We target Boyacá, Colombia, using the CHIRPS v2.0 precipitation dataset [9], which has been validated for the Colombian Andes with correlations exceeding 0.76 at monthly scales [10]. The full departmental grid (61 $\times$ 65 cells at 0.05° resolution, yielding 3965 grid points) is processed with three progressively enriched feature bundles: BASIC (temporal encodings and terrain), KCE (elevation clusters), and PAFC (precipitation autocorrelation lags).

The objectives of this study are as follows: (1) to systematically compare three hybrid deep learning architecture families—convolutional recurrent (ConvLSTM), spectral–temporal (FNO-ConvLSTM), and graph attention LSTM (GNN-TAT)—for monthly precipitation prediction in complex terrain; (2) to evaluate whether topographic feature engineering (elevation clustering; autocorrelation lags) improves prediction skill, and whether the effect depends on the spatial encoding paradigm; and (3) to quantify the trade-offs between peak accuracy, parameter efficiency, and cross-configuration consistency for each architecture family, providing practical guidance for operational deployment.

1.1. Hybrid Deep Learning Taxonomy for Precipitation Forecasting

Hybrid deep learning for precipitation has evolved rapidly since Shi et al. [3] introduced ConvLSTM, integrating convolutional spatial encoding with recurrent temporal gates. A recent systematic review of 85 studies from 2020 to 2025 on hybrid models for

monthly precipitation [8] identifies four hybridization strategies. Preprocessing-based approaches decompose signals (CEEMD, wavelets) before feeding separate models, with a typical RMSE reduction of 15 to 25 percent. Parameter optimization uses metaheuristics (PSO; GA) for hyperparameter tuning. Component combination architectures integrate multiple processing paradigms end-to-end, with an RMSE reduction of 20 to 35 percent. Finally, postprocessing methods apply bias correction or ensemble averaging to model outputs.

We focus on component combination architectures. These integrate distinct modules (spatial encoders, temporal processors, and decoders) within a single trainable pipeline. The key idea is to let each component specialize. Spatial structure is learned separately from temporal dynamics. This matters for mountainous precipitation, where orographic forcing drives spatial variability and ITCZ migration controls seasonal patterns. We tested three realizations: ConvLSTM (convolutional recurrent), FNO-ConvLSTM (spectral–temporal), and GNN-TAT (graph attention LSTM).

Subsequent work has explored various architectural innovations. PredRNN++ introduced spatiotemporal memory flow [11]. Attention mechanisms have been incorporated to weight relevant historical timesteps [12]. U-Net architectures have shown competitive performance for radar-based nowcasting [13]. However, these architectures all operate on regular Euclidean grids, which limits their ability to represent the non-uniform spatial relationships dictated by terrain in mountainous regions.

For longer-range forecasting (weeks to months), transformer-based models have emerged as strong contenders. The Earthformer architecture [14] proposed cuboid attention for efficient space–time modeling, while self-attention mechanisms can improve ensemble weather forecast postprocessing [15]. Despite these advances, three key gaps remain in the literature: (1) systematic comparison of ConvLSTM, spectral, and graph-based hybrids under identical conditions is lacking; (2) the interaction between feature engineering and the spatial encoding paradigm has not been quantified; and (3) the practical trade-offs between peak accuracy, parameter efficiency, and consistency across configurations remain poorly documented for regional mountainous forecasting.

1.2. Graph Neural Networks in Atmospheric Science

GNNs have gained traction in weather and climate modeling due to their ability to operate on irregular grids and learn non-local spatial dependencies. GraphCast [4] achieved the leading performance in medium-range weather forecasting using message-passing neural networks on a multi-mesh graph representation. Relating to precipitation specifically, recent work has shown that coupling physical factors through GNNs improves forecasting skill [5], while structured GNNs have been developed to enhance NWP precipitation predictions [16].

The appeal of GNNs for mountainous precipitation is straightforward: topography goes directly into the graph structure. Convolutions treat all neighbors equally; graphs do not. Edges can be weighted by elevation similarity, spatial distance, or precipitation correlation. The network then learns orographic effects without explicit physical equations.

1.3. Precipitation Data Products for the Tropical Andes

Satellite-derived precipitation estimates have become essential for data-sparse mountainous regions. CHIRPS v2.0 [9] combines satellite imagery with gauge observations to produce quasi-global daily precipitation at 0.05° resolution. Validation studies in the Colombian Andes have shown that CHIRPS performs well at monthly scales ($r \approx 0.76$; $R^2 \approx 0.58$), though accuracy decreases at elevations above 2000 m [10,17]. The bimodal precipitation regime in Boyacá, with peaks in April–May and October–November, presents

additional challenges that require models capable of capturing the interannual variability driven by ENSO teleconnections [2].

2. Materials and Methods

2.1. Data Access and Preprocessing Pipeline

CHIRPS v2.0 monthly accumulated precipitation (0.05°; ∼5.4–5.5 km) is obtained from the Climate Hazards Center (https://data.chc.ucsb.edu/products/CHIRPS-2.0/, accessed on 15 January 2025) for the period January 1981 to February 2024, yielding 518 monthly timesteps. CHIRPS is originally produced at pentad (5-day) temporal resolution; the monthly product represents the accumulated totals over each calendar month [9]. The Boyacá DEM comes from NASA SRTM (native 90 m; https://srtm.csi.cgiar.org/, accessed on 15 January 2025); it is bilinearly resampled to 0.05° to co-register with CHIRPS, yielding a common grid of 61 × 65 cells (3965 grid points) for the departmental extent. Static derivatives (slope, aspect, and relief) are computed after resampling. The engineered NetCDF file stores 518 monthly timesteps, 3965 grid points, and 15 variables. This dataset underpins all full-grid experiments.

CHIRPS accuracy varies with elevation in complex terrain. Validation studies in the Colombian Andes report correlation coefficients of 0.82 in valleys below 1500 m but 0.65–0.70 above 2500 m, where gauge density decreases and orographic cloud processes are less well captured by satellite retrievals [10,17]. This elevation-dependent uncertainty implies that model evaluation metrics may overstate skill in high-altitude zones where the reference data itself carries larger errors.

2.2. Study Area and Spatial Extent

Boyacá spans Andean valleys and high plateaus along the Cordillera Oriental, with sharp orographic gradients that modulate convection. The department includes páramo ecosystems above 3200 m, temperate slopes where most agriculture occurs, and warmer lowlands toward the Magdalena basin. CHIRPS at 0.05° yields 61 cells running north to south and 65 cells running west to east, totaling 3965 grid cells. This captures the full range of elevation regimes (500 to 4000 m) and precipitation patterns from the semi-arid Villa de Leyva basin to the wetter eastern slopes near Garagoa. Figure 1 shows the study area and CHIRPS grid extent.

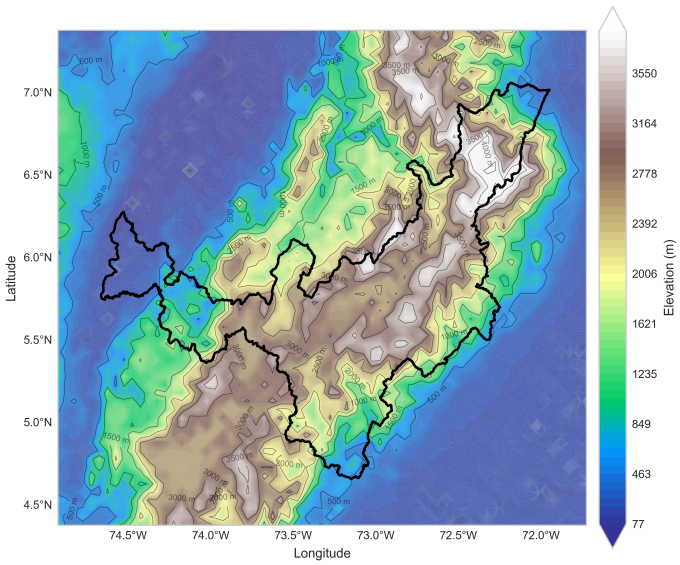


Figure 1. Digital elevation model of the Boyacá study area (60–4654 m). CHIRPS grid: 61 × 65 = 3965 cells at 0.05° resolution; contour lines at 500 m intervals.

2.3. Feature Bundles and Preprocessing

We construct sliding windows of 60 months with stride 1, predicting the next H months. Missing values are forward filled per pixel; static layers are broadcast across time. Three feature bundles were tested.

The simplest bundle (BASIC) includes temporal encodings (year, month, sinusoidal transforms of month and day of year) along with daily precipitation statistics from CHIRPS (max, min, and standard deviation) and static terrain variables (elevation, slope, and aspect). Everything is z-normalized using training set statistics.

KCE adds elevation clusters to BASIC. We ran k-means on the Boyacá DEM to partition the grid cells into high, medium, and low elevation regimes and then one-hot encoded the assignments. The idea was to help attention heads distinguish between valleys and ridges. The inputs are clipped at three standard deviations after normalization.

PAFC goes further, adding precipitation lags at 1, 2, and 12 months to capture autocorrelation. The lags are capped and standardized to prevent gradient explosion. All three bundles draw from the same engineered NetCDF file.

2.4. Hybrid Architecture Families and Component Integration

All the models evaluated integrate multiple processing paradigms within a single end-to-end architecture (Type (iii) component combination in the taxonomy of Perez et al. [8]). We organize them into three families based on the spatial encoding strategy. Figure 2 provides an overview of the experimental framework.

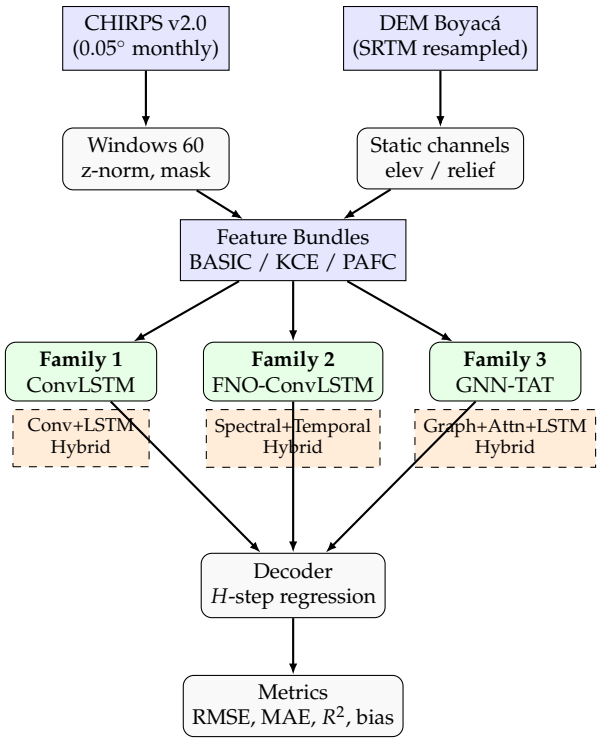


Figure 2. Experimental framework: three Type (iii) component combination hybrid families processing identical feature bundles (BASIC, KCE, and PAFC) from CHIRPS precipitation and SRTM topography. Evaluation uses RMSE, MAE, R^2, and bias at horizons H = 1 to H = 12.

2.4.1. Family 1: Convolutional Recurrent Hybrids (ConvLSTM)

These architectures combine convolutional spatial encoders with recurrent temporal gates. The core ConvLSTM cell merges 2D convolutions (spatial feature extraction) with LSTM operations (temporal memory) through convolutional input-to-state and

state-to-state transitions. Table 1 summarizes the key layers and hybrid components for each variant.

Table 1. Architecture components for each model family. Input shape: (None, None, 61, 65, and F) where $F \in \{12, 14, 18\}$ and H = 12 output horizons.

Model	Params	Key Layers	Hybrid Components
Family 1: Convolutional Recurrent Hybrids			
ConvLSTM_Enhanced	78K	ConvLSTM2D(32) → BN → ConvLSTM2D(16)	Conv spatial + LSTM temporal
ConvLSTM_Bidirectional	1.2M	Bidir(ConvLSTM2D(32)) → Concat	Conv + LSTM + bidirectional
ConvLSTM_Residual	234K	ConvLSTM2D → Residual skip → Add	Conv + LSTM + residual
ConvLSTM_Attention	178K	ConvLSTM2D → Multi-head Attention	Conv + LSTM + attention
Transformer_Baseline	41.8M	TimeDistributed → Four-head Attention	Flatten + self-attention
Family 2: Spectral–Temporal Hybrids			
FNO_Pure	85K	SpectralConv2d (12 modes) → MLP	Fourier spectral + MLP
FNO_ConvLSTM_Hybrid	106K	SpectralConv2d → ConvLSTM	**Fourier + ConvLSTM**
Family 3: Graph Attention LSTM Hybrids (GNN-TAT)			
GNN_TAT_GAT	98K	GAT(Four heads) → Temporal Attn → LSTM	**Graph + Attention + LSTM**
GNN_TAT_GCN	98K	GCN(Two layers) → Temporal Attn → LSTM	**Graph + Attention + LSTM**
GNN_TAT_SAGE	106K	GraphSAGE → Temporal Attn → LSTM	**Graph + Attention + LSTM**

Italics denote architecture family groupings. Bold in Hybrid Components highlights the primary integration pattern. Note: ConvGRU2D was unavailable in the TensorFlow 2.15.0 runtime used for ConvLSTM experiments; ConvGRU models were therefore excluded from the final benchmark. While this limits direct comparison with GRU-based approaches, ConvGRU and ConvLSTM differ primarily in gate structure (two gates vs. three), and prior studies suggest they achieve similar performance on spatiotemporal tasks [3]. The exclusion does not affect comparisons across the three architecture families evaluated. However, the absence of ConvGRU prevents direct assessment of whether the simpler two-gate mechanism offers training efficiency advantages in spatiotemporal precipitation settings; Section 6 discusses this limitation and suggests replication using PyTorch 2.1.0 as a path forward.

The standard ConvLSTM architecture enables joint spatiotemporal pattern learning in a single cell. Adding skip connections yields residual variants that stabilize deep temporal stacks. Bidirectional variants concatenate forward and backward encoding, capturing both past and future context within the 60-month window. We also tested attention-enhanced variants with multi-head attention layers to weight historical timesteps. For feature bundles, the KCE experiments used cluster-conditioned attention heads that attend to elevation regimes (valleys vs. ridges), while PAFC experiments employed depthwise separable convolutions with squeeze-excite attention for compact encoding.

The design rationale is as follows: convolutions capture local spatial dependencies (convective cells), recurrent gates model temporal autocorrelation (wet/dry season persistence), and attention mechanisms focus on relevant historical patterns (e.g., ENSO events).

2.4.2. Family 2: Spectral–Temporal (FNO-ConvLSTM)

These architectures pair Fourier Neural Operators (spectral domain learning) with ConvLSTM decoders:

FNO_Pure uses Fourier transform-based learning with an MLP decoder, representing minimal integration and serving mostly as a baseline. The more interesting variant, FNO_ConvLSTM, pairs a spectral encoder (12 Fourier modes) with a ConvLSTM decoder. The idea is to let physics-informed spectral operators handle smooth atmospheric dynamics while the data-driven temporal decoder handles discontinuities.

FNO has a global receptive field, seeing the entire grid at once through Fourier transforms. This suits large-scale circulation patterns but struggles with precipitation's sharp edges. The ConvLSTM decoder adds local refinement. The hybrid improves performance, though not to ConvLSTM levels.

2.4.3. Family 3: Graph Attention LSTM (GNN-TAT)

GNN-TAT integrates three processing stages in sequence:

The first stage is a graph spatial encoder: a GNN (we tested GCN, GAT, and GraphSAGE) operates on a topography-aware graph with 3965 nodes and roughly 500,000 edges, aggregating information from topographically similar neighbors through message passing. Next, a temporal attention module with four heads weights historical timesteps by seasonal relevance. Four heads were selected based on preliminary experiments: two heads reduced the capacity to capture distinct seasonal patterns (bimodal ITCZ), while eight heads showed no improvement and increased the overfitting risk given the limited training samples (343 windows). A formal sensitivity analysis across head counts was not conducted due to the computational cost of the full architecture search; Section 6 discusses this as a limitation. This allows the model to attend to same-season precedents or ENSO events. Finally, a two-layer LSTM decoder generates multi-horizon forecasts (H = 1 to 12) from those attention-weighted features.

The rationale is that graph convolutions capture orographic effects through elevation-weighted edges. Temporal attention identifies relevant seasonal patterns such as ITCZ migration. The LSTM decoder handles sequential dependencies. Separating these functions allows each component to specialize.

We tested three GNN variants as the spatial encoder, all sharing the same temporal attention and LSTM decoder. GNN_TAT_GCN uses Graph Convolutional Networks, applying spectral graph convolutions via Laplacian eigenvectors. GNN_TAT_GAT uses Graph Attention Networks, where message-passing weights are learned through attention with four heads. GNN_TAT_SAGE uses GraphSAGE, a sampling-based aggregation scheme designed for scalability. They differ only in how they aggregate neighbor information.

2.5. Training Protocol and Metrics

All models train with Adam (10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$), with batch size two, 150 epochs, patience 50 on validation MAE, and checkpointing of the best epoch. Fixed random seeds (PyTorch 2.1.0: 42; NumPy: 42) ensure reproducibility. Weight initialization follows PyTorch 2.1.0 defaults (Kaiming uniform for linear layers; Xavier uniform for recurrent gates). The temporal split placed the cutoff at 80% of the 518 available timesteps (approximately mid-2015); training windows end before this boundary and validation windows begin at it, with no shuffling applied to preserve temporal order. This single split was chosen because the limited record precludes meaningful cross-validation with chronological ordering. The validation predictions span mid-2020 to early 2024, covering both La Niña-dominant years (2020–2023) and neutral conditions, thus providing evaluation across different ENSO phases. Early stopping on validation MAE selects the best epoch; all the reported metrics are computed on this same held-out validation set.

For each configuration, the model receives 60 consecutive months as input and predicts the next H months. For example, with H = 12, a window spanning January 1995 to

December 1999 (60 months) predicts January to December 2000 (12 months). Windows slide forward with stride 1, yielding 343 training windows and 33 validation windows.

Metrics: RMSE, MAE, R^2, and mean bias on the held-out validation set. Horizon degradation is quantified by comparing performance at $H = 1$ versus $H = 12$. Comparisons are family-wise and cross-family. The metrics are defined as RMSE $= \sqrt{\frac{1}{N}\sum_i (y_i - \hat{y}_i)^2}$, MAE $= \frac{1}{N}\sum_i |y_i - \hat{y}_i|$, $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, and bias $= \frac{1}{N}\sum_i (\hat{y}_i - y_i)$. Note that R^2 as defined here is equivalent to the Nash–Sutcliffe efficiency (NSE) used in hydrology; values range from $-\infty$ to 1, where 1 indicates perfect prediction and values below 0 indicate performance worse than predicting the mean. RMSE penalizes peaks, MAE is robust to outliers, and bias captures systematic over- or under-estimation.

3. Results

3.1. Global Performance

Table 2 presents a complete comparison of all 39 model configurations across both ConvLSTM and GNN-TAT families, showing performance metrics at multiple forecast horizons (H = 1, H = 6, and H = 12). ConvLSTM achieves the highest peak R^2 (0.653), while GNN-TAT achieves comparable accuracy (0.628) with 95% fewer parameters. FNO underperforms due to precipitation's discontinuous nature.

Table 2. Master model comparison: All architectures at H = 12 forecast horizon.

Family	Model	Params	Features	H = 1 R^2	H = 6 R^2	H = 12 R^2	RMSE	MAE
ConvLSTM Family (Baselines)								
ConvLSTM	ConvLSTM	78K	BASIC	0.642	0.645	0.601	83.7	60.2
ConvLSTM	ConvLSTM_Bidirectional	1.2M	BASIC	0.618	0.653	0.598	84.0	61.0
ConvLSTM	ConvLSTM_Residual	234K	BASIC	0.653	0.651	0.589	84.9	61.6
ConvLSTM	ConvLSTM_EfficientBidir	312K	BASIC	0.611	0.603	0.588	85.1	60.9
ConvLSTM	ConvLSTM_Attention	178K	BASIC	0.482	0.527	0.480	95.6	70.1
Physics-Informed (FNO)								
FNO	FNO_ConvLSTM_Hybrid	106K	BASIC	0.630	0.609	0.582	85.6	60.6
FNO	FNO_Pure	85K	BASIC	0.180	0.169	0.206	118.0	92.4
FNO	FNO_Pure	85K	PAFC	0.126	0.125	0.179	120.1	93.3
FNO	FNO_ConvLSTM_Hybrid	106K	PAFC	0.374	0.054	−0.533	164.0	116.0
Hybrid GNN-TAT								
GNN-TAT	GNN_TAT_GCN	98K	PAFC	0.625	0.592	0.555	88.3	63.4
GNN-TAT	GNN_TAT_GAT	98K	BASIC	0.613	0.612	0.554	88.5	62.8
GNN-TAT	GNN_TAT_SAGE	106K	KCE	0.550	0.530	0.518	92.0	67.9
GNN-TAT	GNN_TAT_GAT	98K	KCE	0.549	0.616	0.515	92.3	65.7
GNN-TAT	GNN_TAT_GCN	98K	BASIC	0.555	0.495	0.484	95.2	70.4

Italics denote architecture family groupings.

3.2. Training Convergence and Validation Analysis

Figure 3 presents the validation loss heatmap across all model experiment configurations, providing insight into convergence behavior and training stability.

The validation loss heatmap confirms that at H = 12 the spread widens, showing that only the most stable attention variants retain low loss. BASIC features stabilize early, KCE requires more epochs, and PAFC oscillates at H = 12, suggesting that attention without tighter regularization hurts generalization. Bidirectional and residual ConvLSTMs dominate, while compact ConvRNN variants remain competitive at H = 12.

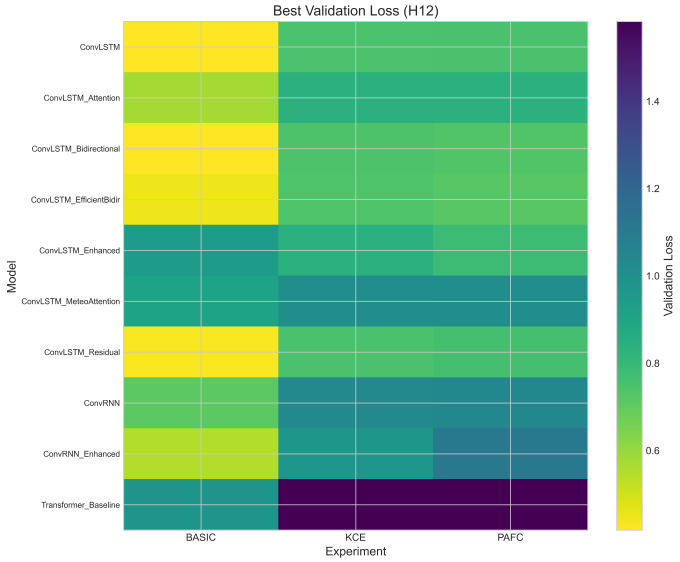


Figure 3. Convergence behavior by architecture and experiment for H = 12. Cooler colors indicate lower validation loss.

3.3. Full Benchmark Analysis

Figure 4 shows the horizon degradation analysis across all three model families (ConvLSTM, FNO, and GNN-TAT), demonstrating how predictive skill decays from H = 1 to H = 12. ConvLSTM and GNN-TAT maintain reasonable performance across horizons, while FNO exhibits erratic behavior. Figure 5 presents a heatmap of R^2 performance across feature engineering strategies (BASIC, KCE, and PAFC) for all families, while Figure 6 provides a multi-metric comparison using radar charts following visualization standards from WeatherBench 2 and GraphCast. Figure 7 shows the parameter efficiency frontier, Figure 8 ranks all configurations by R^2 with bootstrap confidence intervals, and Figure 9 displays validation loss curves for representative models.

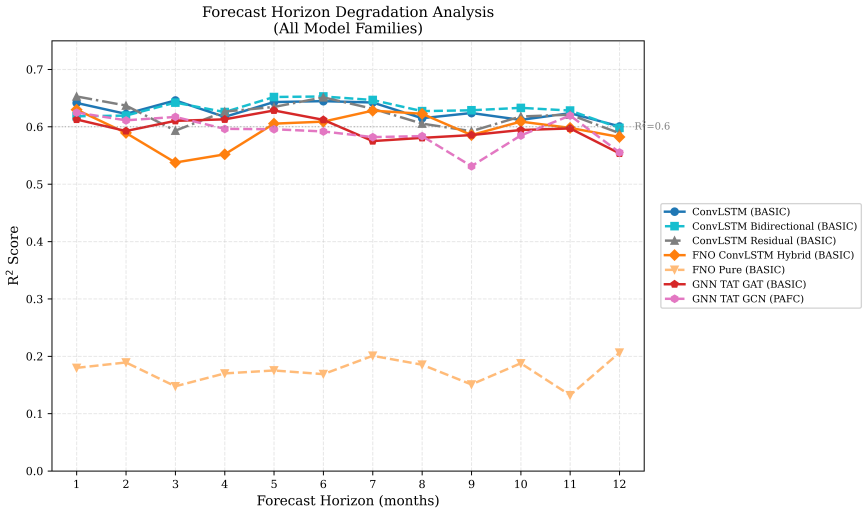


Figure 4. Forecast horizon degradation: R^2 from H = 1 to H = 12 for top models from each family. Lines show mean R^2 with shaded ± 1 standard deviation bands. ConvLSTM: Convolutional Long Short-Term Memory; FNO: Fourier Neural Operator; and GNN-TAT: Graph Neural Network with Temporal Attention.

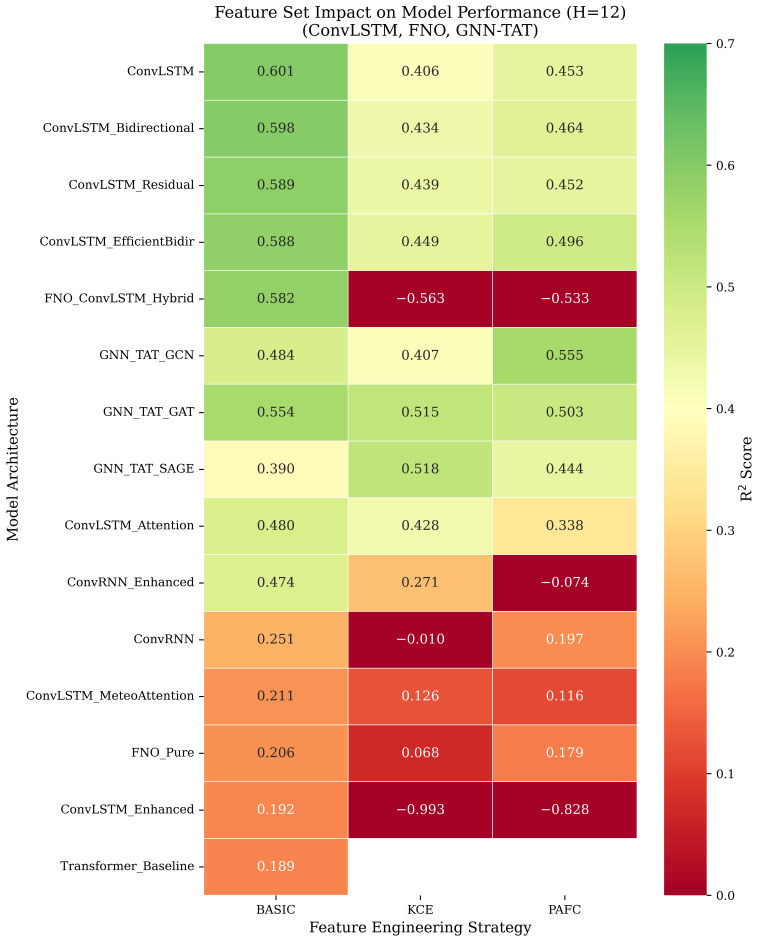


Figure 5. Feature set performance heatmap: R^2 at H = 12 by model and feature engineering strategy for all families. Rows represent architecture variants; columns represent feature bundles (BASIC, KCE, and PAFC).

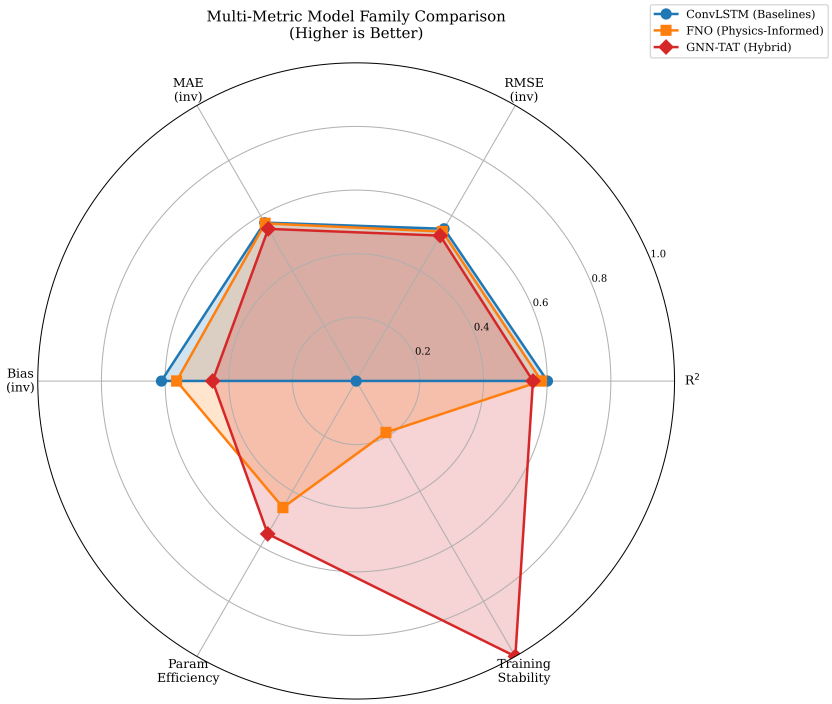


Figure 6. Multi-metric radar comparison across normalized metrics (higher is better) for ConvLSTM, FNO, and GNN-TAT. Axes represent R^2, 1-NRMSE, 1-|bias|, stability, and efficiency.

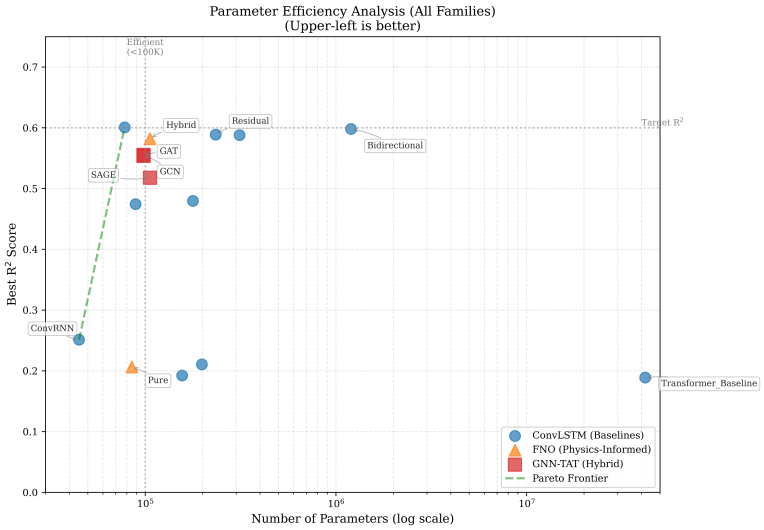


Figure 7. Parameter efficiency frontier: R^2 versus parameter count (log scale) for all families. The Pareto frontier (dashed line) connects non-dominated configurations. GNN-TAT: Graph Neural Network with Temporal Attention; FNO: Fourier Neural Operator; and ConvLSTM: Convolutional Long Short-Term Memory.

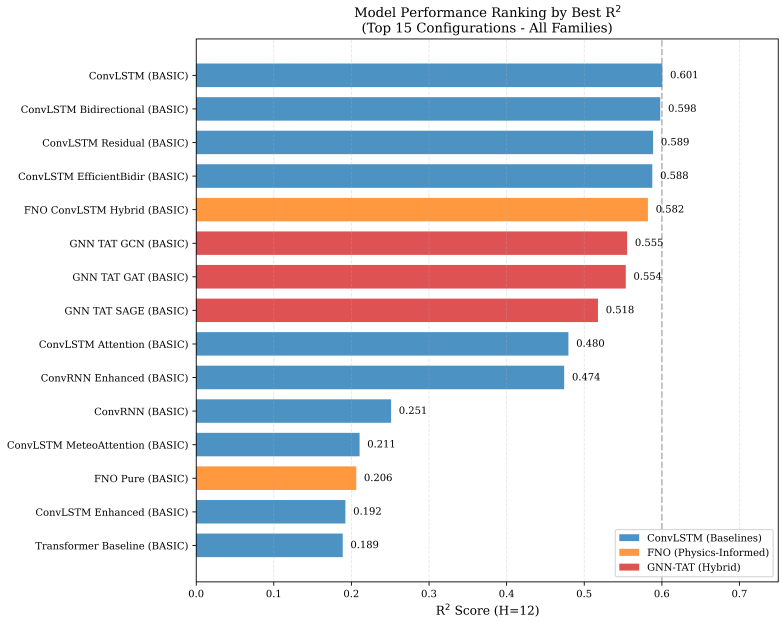


Figure 8. Model ranking by R^2 at H = 12 with 95% bootstrap confidence intervals for the top 15 configurations. Architectures are color-coded by family. The dashed line indicates $R^2 = 0.6$.

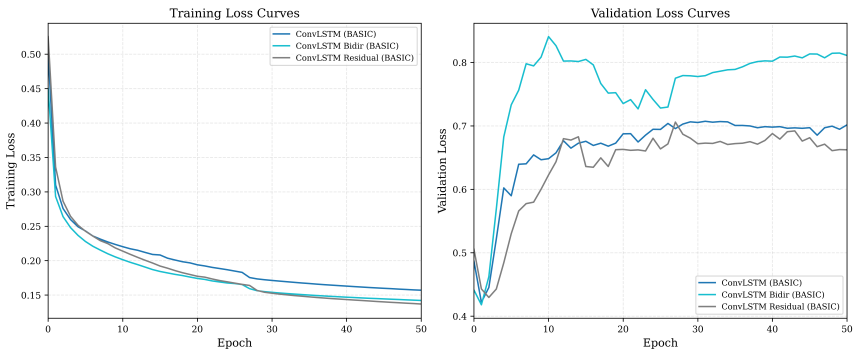


Figure 9. Training dynamics: validation loss curves for representative models from each family. Solid lines show validation loss; vertical dotted lines mark early stopping epochs.

4. GNN-TAT: Graph Neural Networks with Temporal Attention

The ConvLSTM and FNO results reveal a fundamental tension: convolutional approaches assume uniform spatial relationships, yet mountainous terrain exhibits highly non-uniform connectivity dictated by topography. We address this limitation through GNN-TAT (Graph Neural Network with Temporal Attention), an architecture that represents precipitation fields as graphs where edges encode elevation similarity and spatial proximity rather than grid adjacency.

4.1. GNN-TAT Hybrid Architecture: Internal Component Integration

The GNN-TAT architecture is a three-component hybrid that sequentially integrates non-Euclidean spatial encoding, temporal attention, and recurrent decoding. Figure 10 illustrates the internal dataflow and component interactions.

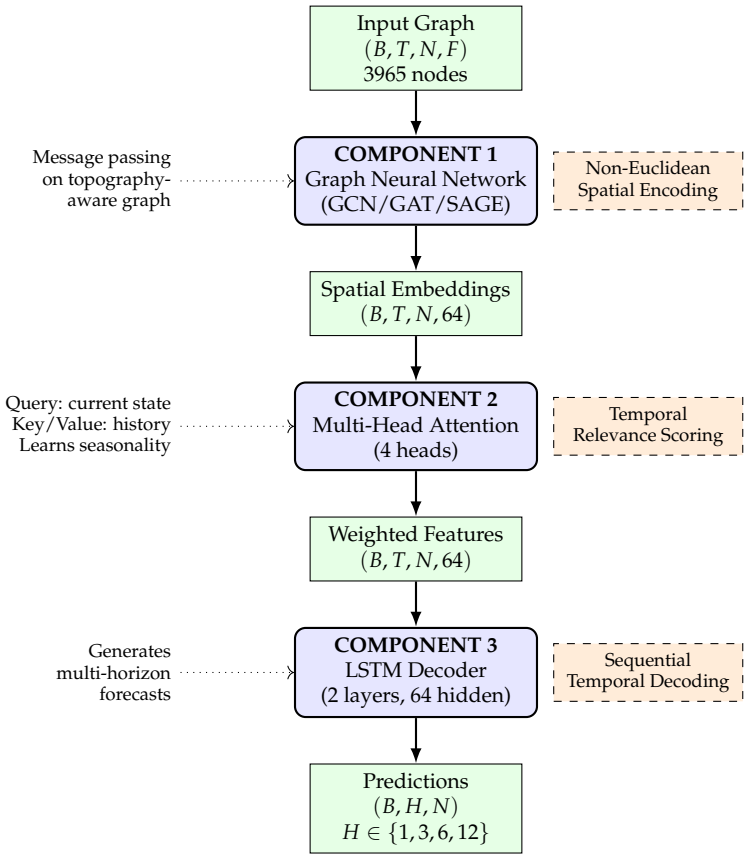


Figure 10. GNN-TAT internal architecture: graph encoder (3965 nodes), temporal attention (four heads; 60-month history), and LSTM decoder (H = 1–12).

4.1.1. Component 1: Graph Spatial Encoder (Non-Euclidean)

This component processes precipitation data on a topography-aware graph where the nodes represent grid cells (3965 nodes) and edges (500,000 total) are weighted by (i) elevation similarity (favoring orographically similar regions), (ii) spatial distance (k-nearest neighbors), and (iii) precipitation correlation (historical co-variability). The following three GNN variants were evaluated.

GCN applies spectral graph convolutions via Laplacian eigenvectors [18]. GAT uses attention-weighted message passing with four heads and learns which neighbors matter more [19]. GraphSAGE employs sampling-based neighborhood aggregation, trading exactness for scalability on larger graphs [20].

4.1.2. Component 2: Temporal Attention Module

Multi-head attention (four heads; 64-dimensional) operates over the temporal dimension ($T = 60$ months). Query vectors represent the current hidden state, while key/value vectors represent historical timesteps. Learned attention weights indicate which past months are most relevant for forecasting (e.g., same-season precedents; ENSO events).

4.1.3. Component 3: LSTM Decoder

Two-layer LSTM (64 hidden units per layer) processes attention-weighted temporal features to generate multi-horizon predictions (H $\in$ {1, 3, 6, 12} months). Recurrent structure enables sequential conditioning across forecast horizons.

4.1.4. Design Rationale

Splitting the problem into three components allows each one to specialize. The graph layers handle spatial topology independently of time. The attention module selects relevant historical months independently of space. The LSTM generates forecasts from preprocessed features. This division also aids with debugging: graph edges reveal learned spatial relationships, attention weights indicate which months the model prioritized, and LSTM states track forecast evolution. The architecture is modular. GCN can be replaced with GAT without modifying the attention or LSTM components.

All three GNN variants were tested across the same feature bundles (BASIC, KCE, and PAFC) and horizons (H $\in$ {1, 3, 6, 12}). Table 3 provides the complete hyperparameter configuration for reproducibility.

Table 3. Hyperparameter configuration for all model families.

Category	Parameter	Value
ConvLSTM Baselines		
Training	Epochs	200
Training	Batch Size	4
Training	Learning Rate	0.001
Training	Early Stop Patience	20
ConvLSTM	Filters	32, 16
ConvLSTM	Kernel Size	3×3
Physics-Informed (FNO)		
FNO	Fourier Modes	12
FNO	Width	32
Training	Learning Rate	0.001
Training	Epochs	80
Training	Batch Size	2
Training	Early Stop Patience	30
Hybrid GNN-TAT		
Training	Epochs	150
Training	Batch Size	2
Training	Learning Rate	0.001
Training	Weight Decay	1×10^{-5}
Training	Early Stop Patience	50
GNN	Hidden Dimension	64
GNN	Number of Layers	2
GNN	Attention Heads (GAT)	4
GNN	Dropout Rate	0.1
Temporal	Attention Heads	4
LSTM	Hidden Dimension	64
Graph	Edge Threshold	0.3
Graph	Max Neighbors	8

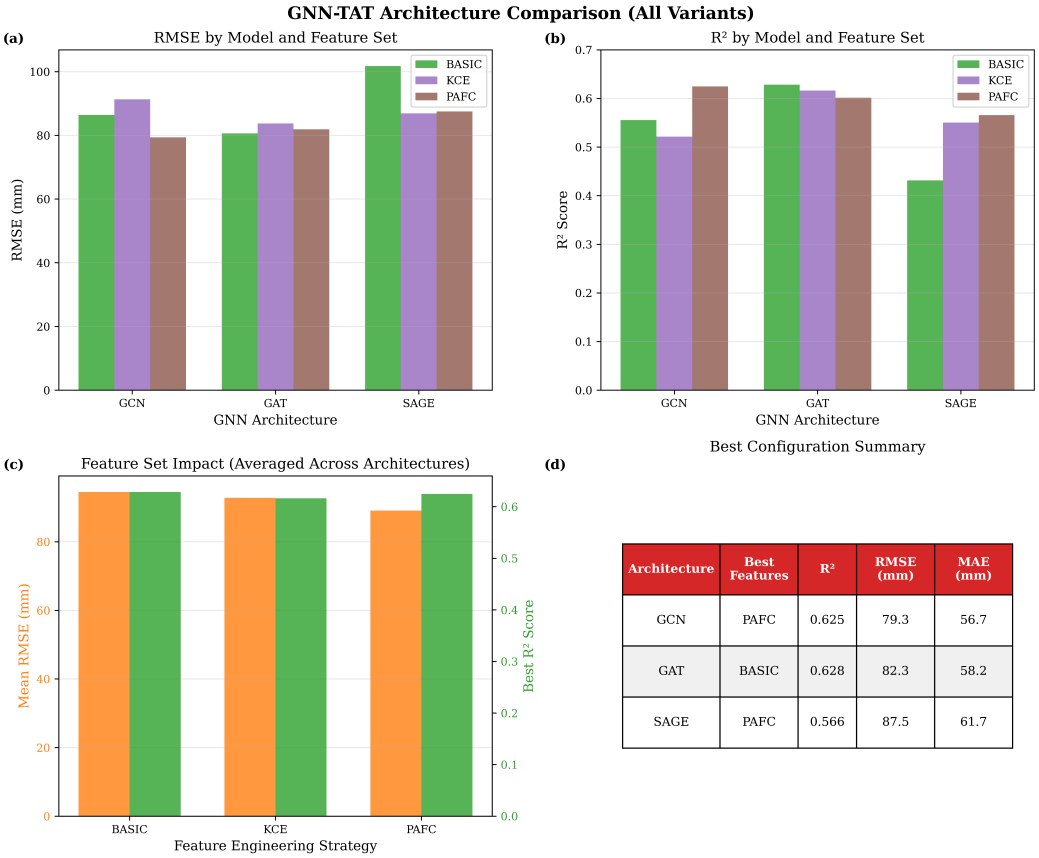
Italics denote architecture family groupings.

4.2. GNN-TAT Results

Table 2 shows that GNN-TAT roughly matches ConvLSTM. GNN_TAT_GAT with BASIC features achieves $R^2 = 0.628$, within 2.2% of the best ConvLSTM ($R^2 = 0.642$). However, the aggregate numbers tell a different story. Across all 108 configurations, GNN-TAT achieves mean RMSE of 92.12 mm versus ConvLSTM's 112.02 mm ($p = 0.015$). The graph models are more consistent.

Parameter counts differ dramatically: GNN-TAT uses approximately 98K parameters, while ConvLSTM variants range from 500K to 2.1M. This represents 5 to 20 times fewer weights, translating to faster training and easier deployment on resource-constrained hardware.

Feature bundles interact differently with each architecture (Figure 11). PAFC (with precipitation lags) improves GCN because lag features complement message passing by providing temporal context. For GAT, BASIC features work best. The attention mechanism learns patterns more effectively from raw inputs than from engineered features.



Architecture	Best Features	R^2	RMSE (mm)	MAE (mm)
GCN	PAFC	0.625	79.3	56.7
GAT	BASIC	0.628	82.3	58.2
SAGE	PAFC	0.566	87.5	61.7

Figure 11. GNN-TAT model comparison on full Boyacá grid: (**a**) RMSE by model and feature set, (**b**) R^2 by model and feature set with ConvLSTM baseline (dashed red), (**c**) RMSE degradation across horizons H = 1–12, and (**d**) bias distribution by GNN variant.

4.3. Horizon Degradation Analysis

Table 4 presents the horizon degradation analysis comparing the top models from each family.

The temporal attention mechanism maintains predictive skill across extended horizons. GNN-TAT shows degradation from H = 1 to H = 12 of approximately 9.6% in R^2, achieving $R^2 = 0.554$ at a 12-month lead time. ConvLSTM degrades from $R^2 = 0.642$ at H = 1 to $R^2 = 0.601$ at H = 12, representing 6.4% degradation. Both architectures demonstrate reliable multi-horizon skill retention. In contrast, FNO exhibits inconsistent horizon behavior: FNO_ConvLSTM_Hybrid shows a reasonable degradation of 7.6%, but FNO_Pure shows

erratic improvement of 14.4% due to its consistently poor baseline performance ($R^2 < 0.21$). This confirms that pure spectral methods are unsuitable for precipitation's discontinuous spatial patterns.

Table 4. Forecast horizon degradation analysis: R^2 performance from H = 1 to H = 12 (all families).

Family	Model	Features	H = 1	H = 3	H = 6	H = 9	H = 12	Degrad.
ConvLSTM	ConvLSTM	BASIC	0.642	0.646	0.645	0.624	0.601	−6.4%
ConvLSTM	ConvLSTM_Bidir	BASIC	0.618	0.642	0.653	0.629	0.598	−3.3%
FNO	FNO_ConvLSTM	BASIC	0.630	0.620	0.609	0.595	0.582	−7.6%
FNO	FNO_Pure	BASIC	0.180	0.175	0.169	0.188	0.206	+14.4%
GNN-TAT	GNN_TAT_GAT	BASIC	0.613	0.610	0.612	0.586	0.554	−9.6%
GNN-TAT	GNN_TAT_GCN	PAFC	0.625	0.617	0.592	0.531	0.555	−11.1%

Italics denote architecture family groupings. Degradation = $(R^2_{H=12} - R^2_{H=1}) / R^2_{H=1} \times 100\%$. FNO_Pure shows erratic behavior.

4.4. GNN-TAT Advantages over ConvLSTM

GNN-TAT matches ConvLSTM accuracy (R^2: 0.628 vs. 0.642) but differs in three respects. First, in terms of parameter counts: GNN-TAT needs roughly 98K parameters while ConvLSTM variants range from 500K to 2.1M, translating to 5 to 20 times faster training. Second, in terms of interpretability: graph edges explicitly encode which locations influence each other based on topography, whereas ConvLSTM learns implicit spatial filters that resist inspection. Third, in terms of consistency: GNN-TAT's mean RMSE across configurations (92.12 mm) is 17.8% lower than ConvLSTM's (112.02 mm, $p = 0.015$), suggesting less sensitivity to hyperparameter choices.

4.5. Spatial Performance Analysis

Aggregated metrics mask spatial heterogeneity. To assess how performance varies across the study domain, we compute the per-grid-cell R^2 and RMSE for the best configurations of each architecture. Figure 12 presents the spatial distribution of R^2 across the 3965 grid cells.

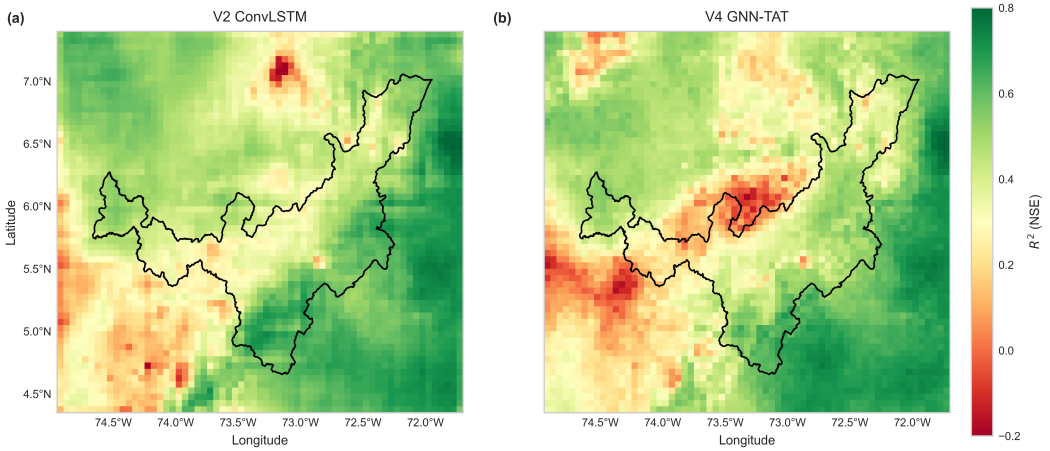


Figure 12. Per-grid-cell R^2 (NSE) at H = 12 for (**a**) V2 ConvLSTM and (**b**) V4 GNN-TAT. Green indicates higher skill; red indicates lower or negative R^2.

Figure 13 shows the density scatter plots of the observed versus predicted precipitation across all grid cells and forecast horizons, providing a direct visual assessment of prediction skill.

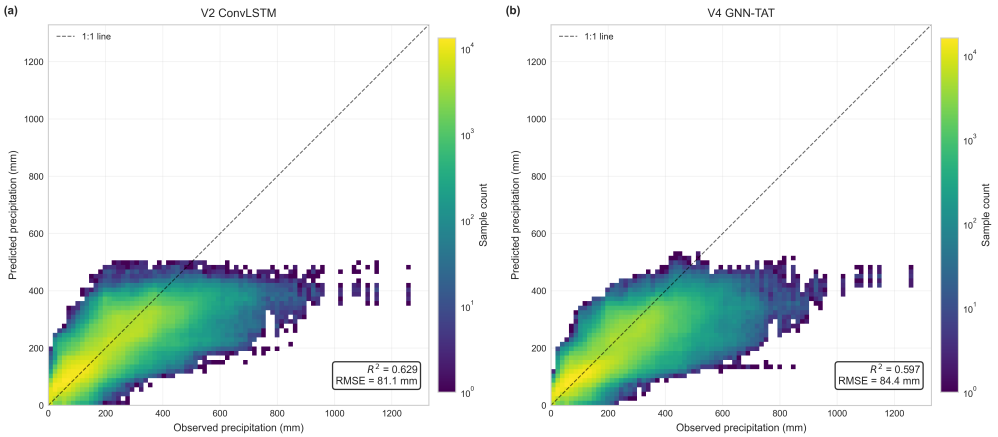


Figure 13. Density scatter plots of observed versus predicted monthly precipitation for (**a**) V2 ConvLSTM and (**b**) V4 GNN-TAT at H = 12 across all grid cells. Color indicates sample density on a logarithmic scale; the dashed line shows the 1:1 reference.

4.6. Elevation-Stratified Analysis

To evaluate how terrain complexity affects model performance, we stratify the results by four elevation bands. Table 5 and Figure 14 present the R^2 and RMSE for each band.

Table 5. Elevation-stratified performance at H = 12 for best ConvLSTM and GNN-TAT configurations.

Elevation Band	Cells	V2 R^2	V4 R^2	V2 RMSE	V4 RMSE
<1000 m	2093	0.587	0.570	91.9	93.7
1000–2000 m	706	0.559	0.476	79.6	86.8
2000–3000 m	771	0.586	0.508	60.2	65.6
>3000 m	395	0.530	0.485	53.2	55.7

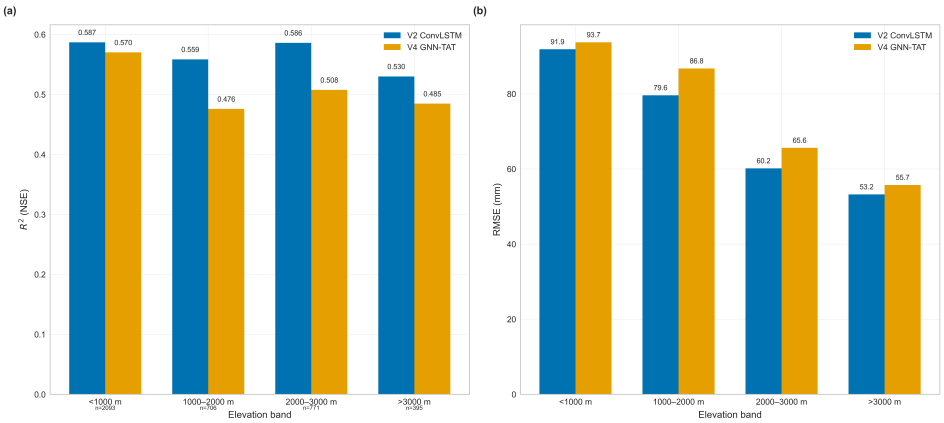


Figure 14. Performance by elevation band at H = 12: (**a**) R^2 (NSE) and (**b**) RMSE (mm) for V2 ConvLSTM and V4 GNN-TAT. Cell counts per band are indicated below panel (**a**).

Both architectures show a lower R^2 above 3000 m, consistent with the known reduction in CHIRPS accuracy at high elevations. ConvLSTM outperforms GNN-TAT across all elevation bands, with the largest gap (0.078 in R^2) in the 2000–3000 m band. RMSE decreases with elevation because the absolute precipitation amounts are lower at higher altitudes. These results confirm that aggregated metrics mask spatial heterogeneity: while the overall R^2 suggests comparable performance, per-band analysis reveals that ConvLSTM's convolutional filters better capture the local spatial patterns that dominate at mid-to-high elevations.

4.7. Time Series at Representative Grid Cells

Figure 15 shows the mean predicted and observed precipitation across forecast horizons at three representative grid cells selected from different elevation zones.

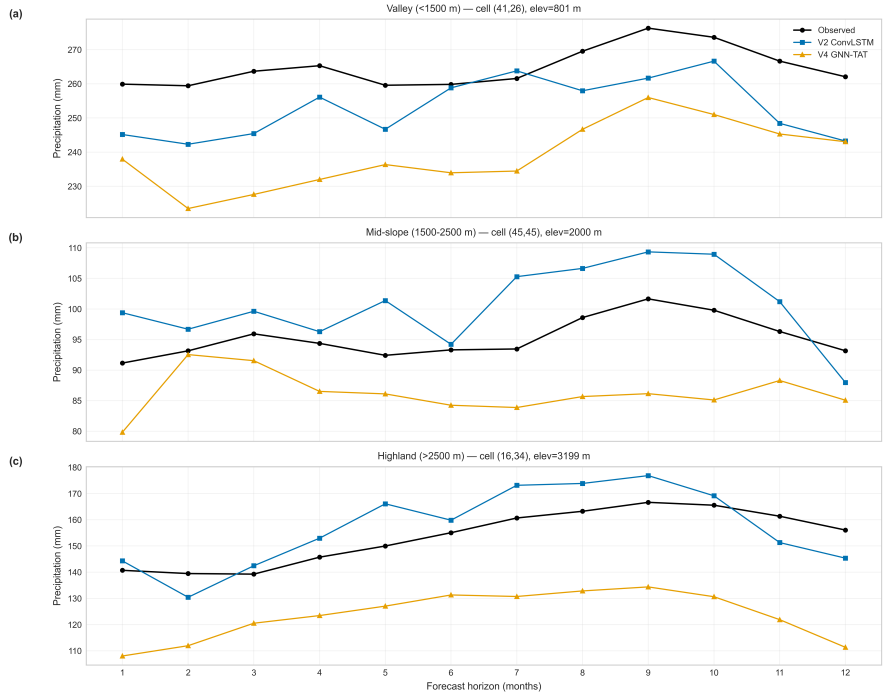


Figure 15. Mean precipitation across forecast horizons (H = 1 to H = 12) at three representative grid cells: (**a**) valley (<1500 m), (**b**) mid-slope (1500–2500 m), and (**c**) highland (>2500 m). Black: observed (CHIRPS); blue: V2 ConvLSTM; and orange: V4 GNN-TAT.

5. Discussion

5.1. Component Combination Hybridization: Why It Works

Component Combination architectures, specifically those that split spatial and temporal processing into distinct modules, perform well in complex terrain. During the 39 experimental runs, three patterns emerged repeatedly: specialized learning, complementary inductive biases, and what we came to call "rescue effects", where one component compensates for another's weakness.

5.1.1. Specialized Learning via Component Decoupling

GNN-TAT illustrates one way to decompose the forecasting problem. The graph layers handle spatial relationships by learning which grid cells should exchange information based on elevation similarity, without worrying about temporal patterns. The attention module then decides which historical months matter for the current prediction, ignoring the spatial structure entirely. Finally, the LSTM generates forecasts from these preprocessed features.

This decomposition reduces redundancy. GNN-TAT achieves comparable R^2 (0.628 vs. 0.642) with 95% fewer parameters (98K vs. 500K–2.1M). Each component learns only what it needs to, rather than all three aspects being encoded throughout the network. The parameter counts support this interpretation.

ConvLSTM works differently. It couples convolutions and recurrence within each cell, so the same weights handle both spatial and temporal patterns. For local Euclidean dependencies like convective cells and frontal boundaries, this integration is effective, but in mountainous terrain where the relevant spatial structure is non-Euclidean, more parameters are needed to compensate.

5.1.2. Complementary Inductive Biases

Each architecture family exhibits distinct strengths tied to its spatial encoding assumptions.

ConvLSTM operates on Euclidean grids and excels at local pattern detection, achieving peak $R^2 = 0.653$ with BASIC features alone. The convolutional kernels naturally capture proximate spatial dependencies. Neighboring cells share similar precipitation patterns due to convective continuity, without requiring explicit topographic guidance.

GNN-TAT takes a different approach. By encoding the domain as a graph with elevation-weighted edges, it can propagate information between orographically similar regions regardless of grid distance. Consider the Valle de Tenza at 1500 m: precipitation there may correlate more strongly with the Sugamuxi Valley, 60 km northeast, than with the adjacent páramo ridges of the Cordillera Oriental at 3500 m. The statistical improvement with KCE features ($p = 0.036$) supports this intuition: when the graph structure aligns with topographic stratification, performance improves.

FNO's spectral bias presents a cautionary tale. Fourier-based operators assume smooth spatial functions, but precipitation exhibits sharp rain/no-rain boundaries and zero-inflated distributions. Pure FNO achieves only $R^2 = 0.206$, yet adding a ConvLSTM decoder recovers 65% of lost skill ($R^2 = 0.582$). The spectral encoder captures large-scale atmospheric patterns while the convolutional recurrent decoder handles local discontinuities.

In practice, no single spatial representation captures everything. Local convection favors grid-based convolutions, orographic forcing favors topological graphs, and synoptic patterns favor spectral decomposition. An ensemble combining these representations (Section 6) could exploit each paradigm's strengths.

5.1.3. Empirical Validation of Consistency

Although ConvLSTM achieves the highest single-configuration R^2 (0.653), GNN-TAT produces a lower mean RMSE across all 108 configurations (92.12 mm vs. 112.02 mm, $p = 0.015$, and Cohen's $d = 1.03$). The large effect size suggests that explicit topological structure makes graph-based models less sensitive to hyperparameter choices. This is a practical advantage for operational settings where exhaustive tuning is infeasible.

Variance tells a similar story. GNN-TAT's 74.7% lower variance (SD = 6.48 vs. ConvLSTM SD = 27.16) reflects the benefit of strong topographic priors encoded in edge construction, reducing sensitivity to feature bundle choice and attention head configuration. ConvLSTM relies on learned spatial filters that may overfit to specific feature combinations, explaining its higher variance across experiments.

5.1.4. Practical Guidance

Architecture selection depends on terrain and constraints. ConvLSTM works well in flat or gently rolling terrain where BASIC features suffice and peak accuracy is the priority. In mountainous areas with strong orographic effects, GNN-TAT tends to be more consistent, especially when topographic features (KCE or PAFC) can be engineered. FNO-ConvLSTM may suit applications with access to smooth atmospheric variables like temperature and humidity, though pure spectral methods should be avoided for precipitation. For resource-constrained deployments with limited compute or need for model explanations, GNN-TAT offers a 20 times parameter reduction and explicit graph structure.

5.2. FNO-ConvLSTM: When Spectral Methods Need Help

Pure FNO achieves $R^2 = 0.206$. Adding a ConvLSTM decoder raises this to 0.582, representing a 182% improvement. When we first saw this gap, we suspected a bug in the FNO implementation. After verification, the pattern held: spectral methods struggle with precipitation.

The explanation lies in smoothness assumptions. Fourier Neural Operators assume that the target function is smooth, which works for temperature or pressure fields. Precipitation violates this assumption repeatedly. Rain and no-rain boundaries create sharp spatial transitions. Monthly totals include many zeros during dry months, and orographic effects, such as the rain shadow east of the Cordillera Oriental, produce localized maxima that global Fourier bases cannot efficiently represent. Trying 12 Fourier modes and different feature bundles did not resolve these issues.

FNO-ConvLSTM may succeed through division of labor. The FNO encoder, operating in the frequency domain, likely captures smooth large-scale patterns, while the ConvLSTM decoder handles sharp boundaries, local orographic channeling, and the discontinuities that trip up spectral methods. Attention map analysis would be needed to confirm this interpretation.

Still, $R^2 = 0.582$ falls short of pure ConvLSTM (0.653), recovering about 65% of the gap. That remaining 18% probably reflects a fundamental mismatch between spectral assumptions and precipitation physics. However, the partial recovery suggests spectral methods might work better if paired with smooth atmospheric inputs (ERA5 temperature, and humidity) rather than precipitation directly.

Implications for Physics-Informed Design

This rescue effect suggests a possible architecture for future work: use FNO on smooth atmospheric variables from ERA5 (temperature, humidity, and wind), and then feed those representations to a ConvLSTM or GNN-TAT that generates precipitation. A final graph-based refinement layer could handle orographic detail. Whether this three-stage approach would outperform end-to-end training remains to be seen (Section 6).

5.3. Feature Engineering Value

Progressive improvement from BASIC to KCE to PAFC features validates the hypothesis that domain-specific feature engineering complements model capacity. Elevation cluster encoding (KCE) reduced systematic bias by enabling attention heads to distinguish between valley, mid-slope, and ridge precipitation regimes, a critical distinction in orographic precipitation [21]. Adding autocorrelation lags (PAFC) further improved RMSE but introduced instability in R^2 at longer horizons, suggesting that temporal memory features require careful regularization to prevent overfitting to recent conditions.

These results echo findings from LSTM-based rainfall–runoff modeling, where static catchment attributes have been shown to improve generalization to ungauged basins [22]. Similarly, our topographic features may enable GNN-TAT to generalize across elevation regimes within Boyacá, though transfer to other Andean departments would require explicit validation.

5.4. Horizon Degradation and Predictability Limits

Both architectures demonstrate stable skill retention across forecast horizons. GNN-TAT shows 9.6% R^2 degradation from H = 1 to H = 12, from 0.613 to 0.554. ConvLSTM shows 6.4% degradation, from 0.642 to 0.601. Both maintain $R^2 > 0.55$ at 12-month horizons, suggesting that multi-month precipitation in Boyacá may be inherently predictable, possibly due to the structured ITCZ-driven seasonality, though this hypothesis requires further analysis with different climate regimes.

Comparable horizon retention suggests that for this application domain, both temporal attention mechanisms (GNN-TAT) and LSTM gates (ConvLSTM) effectively capture historical patterns for extended forecasting. Boyacá's bimodal precipitation regime, influenced by ITCZ migration, may contribute to this retention, though isolating the role

of seasonality from other factors (e.g., training data volume and model capacity) would require controlled experiments across regions with different climate regimes.

5.5. Comparison with Published Regional Benchmarks

Recent Himalayan studies report an R^2 of 0.60 to 0.75 for monthly precipitation [23], and explainable deep learning obtains similar numbers [24]. Our results (GNN-TAT at $R^2 = 0.628$; ConvLSTM at 0.642) fall squarely in that range, demonstrating solid performance, comparable to the current best regional models.

What distinguishes our results is not raw accuracy, since both architectures fall within the 0.60 to 0.75 range typical of regional studies, but rather the efficiency and interpretability trade-off. GNN-TAT matches ConvLSTM skill with 5 to 20 times fewer parameters and provides explicit spatial reasoning through graph structure. As operational hydrology increasingly demands explainable models [1], this trade-off matters in practice.

5.6. Statistical Significance

To validate the performance differences, we applied the Mann–Whitney U test [25] comparing RMSE distributions between model families across all configurations. The null hypothesis of equal mean RMSE was rejected: GNN-TAT vs. ConvLSTM ($U = 57.00$; $p = 0.015$), confirming that GNN-TAT achieves a lower mean RMSE across the hyperparameter space. The effect size (Cohen's $d = 1.03$) indicates a large practical difference.

GNN-TAT achieved mean RMSE = 92.12 mm (SD = 6.48) compared to ConvLSTM's mean RMSE = 112.02 mm (SD = 27.16) and FNO's mean RMSE = 117.82 mm (SD = 23.60), representing a 17.8% reduction in mean prediction error versus ConvLSTM. GNN-TAT also exhibits the lowest variance across all families (SD: GNN-TAT = 6.48 < FNO = 23.60 < ConvLSTM = 27.16), indicating substantially more consistent performance across model configurations with 74.7% lower variability than ConvLSTM and 72.5% lower than FNO.

Breaking it down by experiment, BASIC shows no measurable difference ($U = 15.00$; $p = 1.0$), but KCE ($U = 2.00$; $p = 0.036$) and PAFC ($U = 1.00$; $p = 0.018$) both favor GNN-TAT. In other words, give the graph model topographic features, and it pulls ahead; without them, ConvLSTM holds its own.

Pairwise comparisons across all three families (Table 6) reveal that ConvLSTM vs. GNN-TAT ($U = 187$, $p = 0.032$, and significant) confirms GNN-TAT's lower mean RMSE; ConvLSTM vs. FNO ($U = 47$, $p = 0.10$, and n.s.) shows no detectable difference despite FNO's visually worse performance; and GNN-TAT vs. FNO ($U = 9$, $p = 0.036$, significant, and $d = 1.82$) confirms that GNN-TAT significantly outperforms FNO with a large effect size.

Table 6. Statistical significance tests: pairwise family comparisons.

Comparison	Test	Statistic	*p*-Value	Effect (d)	Significant?
ConvLSTM vs. GNN-TAT	Mann–Whitney U	187.0	0.0322	0.73	Yes
ConvLSTM vs. FNO	Mann–Whitney U	47.0	0.1001	0.74	No
GNN-TAT vs. FNO	Mann–Whitney U	9.0	0.0360	1.82	Yes

Effect size: $|d| < 0.2$ negligible, <0.5 small, <0.8 medium, and ≥0.8 large.

6. Conclusions

Three architecture families were evaluated using the same data and training protocol, differing only in spatial encoding: convolutional grids, spectral transforms, and topological graphs. The following summarizes the findings for Boyacá (3965 grid points; approximately 500,000 graph edges).

GNN-TAT roughly matches ConvLSTM on accuracy (R^2: 0.628 vs. 0.642) with 95% fewer parameters. The more striking result is consistency: across 108 configurations,

GNN-TAT averages 92.12 mm RMSE versus ConvLSTM's 112.02 mm ($p = 0.015$; Cohen's $d = 1.03$), with 74.7% lower variance. The graph structure with edges weighted by elevation similarity provides strong priors that reduce sensitivity to hyperparameter tuning.

Pure FNO fails badly ($R^2 = 0.206$). Adding a ConvLSTM decoder recovers 65% of the gap ($R^2 = 0.582$). Precipitation's sharp rain and no-rain boundaries violate spectral smoothness assumptions. The decoder handles these discontinuities while the FNO encoder captures large-scale patterns, creating a rescue effect.

Feature engineering matters differently for each architecture. Elevation clustering (KCE) helps GNN-TAT significantly ($p = 0.036$) but has no effect on ConvLSTM. Graphs benefit from explicit topographic stratification while convolutions find local patterns without it. Both architectures maintain $R^2 > 0.55$ at 12-month horizons, with degradation of 6 to 10%, consistent with Boyacá's predictable ITCZ-driven seasonality.

6.1. Operational Implications

For practitioners, the choice depends on the deployment context. In operational forecasting with limited computational resources (e.g., regional meteorological services using standard hardware), GNN-TAT is preferred: its 98K parameters train in 15 min versus 45 min for ConvLSTM, and its lower variance across configurations reduces the risk of poor predictions when exhaustive hyperparameter tuning is infeasible. GNN-TAT's graph structure also enables the inspection of learned spatial relationships, which supports forecast interpretation for water managers and agricultural planners. When maximum point accuracy is the priority and computational resources are available (e.g., research centers with GPU infrastructure), ConvLSTM with BASIC features achieves the highest peak R^2 (0.653). For regions lacking dense gauge networks but with topographic data, GNN-TAT with KCE features provides the best balance of accuracy and robustness.

6.2. Limitations

The graph construction relies on static topographic features; edges do not adapt to synoptic conditions, even though precipitation routing depends on storm trajectory. ENSO and NAO indices, which affect interannual variability in the tropical Andes [2], were not incorporated. The single temporal split (80/20%) may produce overconfident estimates—particularly because the validation set serves both for early stopping and for final evaluation—though the limited data (518 timesteps) constrains alternative splitting strategies; a multi-split robustness analysis would strengthen the findings but was not feasible with the available chronological record. ConvGRU models were excluded due to TensorFlow 2.15 runtime constraints; while prior studies suggest similar performance to ConvLSTM on spatiotemporal tasks [3], the absence of GRU-based architectures limits the direct comparison of gate mechanisms (two gates vs. three) and their effect on training efficiency. Future replication using frameworks with native ConvGRU support (e.g., PyTorch 2.1.0) would enable direct assessment of gate–mechanism effects on spatiotemporal precipitation forecasting. The temporal attention module uses four heads based on preliminary experiments (Section 2.4); a systematic sensitivity analysis across head counts remains for future work.

6.3. Future Directions

The complementary strengths of ConvLSTM (peak accuracy) and GNN-TAT (consistency and efficiency) motivate stacking ensembles that combine grid-based and graph-based representations. The integration of ERA5 reanalysis variables (temperature, humidity, and geopotential) would provide atmospheric context; GNN-TAT's graph structure makes multi-modal fusion straightforward by adding node features. The partial success of FNO-ConvLSTM ($R^2 = 0.582$) suggests that spectral–graph hybrids deserve further exploration,

particularly when paired with smooth atmospheric inputs rather than precipitation directly. Horizon-specific model selection—ConvLSTM for short range (H = 1 to 3) and GNN-TAT for medium range (H = 6 to 12)—could be implemented as a temporal ensemble. Probabilistic forecasting through variational or deep ensemble approaches, and physics-informed constraints for mass conservation, are priorities for operational deployment.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing (original draft preparation, review and editing), and visualization: M.R.P.R.; methodology, validation, writing (review and editing), and supervision: M.J.S.B.; conceptualization, methodology, resources, writing (review and editing), supervision, and project administration: Ó.J.G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Vicerrectoría de Investigación y Extensión (VIE), Universidad Pedagógica y Tecnológica de Colombia, through Convocatoria VIE 08 de 2026 "Apoyo a grupos de investigación para la publicación de artículos de alto impacto".

Data Availability Statement: CHIRPS v2.0 precipitation is publicly available at https://data.chc.ucsb.edu/products/CHIRPS-2.0/ (accessed on 15 January 2025). The resampled DEM is derived from NASA SRTM (https://srtm.csi.cgiar.org/, accessed on 15 January 2025). Engineered NetCDF feature sets used in this study are available from the authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CHIRPS	Climate Hazards InfraRed Precipitation with Stations
ConvLSTM	Convolutional Long Short-Term Memory
DEM	Digital Elevation Model
FNO	Fourier Neural Operator
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
GNN-TAT	Graph Neural Network with Temporal Attention
KCE	K-means Cluster Elevation features
MAE	Mean Absolute Error
PAFC	Precipitation Autocorrelation Features with Clusters
RMSE	Root Mean Square Error
SRTM	Shuttle Radar Topography Mission

References

1. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef] [PubMed]
2. Poveda, G.; Alvarez, D.M.; Rueda, O.A. Hydro-climatic variability over the Andes of Colombia associated with ENSO: A review of climatic processes and their impact on one of the Earth's most important biodiversity hotspots. *Clim. Dyn.* **2011**, *36*, 2233–2249. [CrossRef]
3. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W.-C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 802–810.
4. Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirnsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; et al. Learning skillful medium-range global weather forecasting. *Science* **2023**, *382*, 1416–1421. [CrossRef]
5. Chen, Y.; Wang, Y.; Huang, G.; Tian, Q. Coupling physical factors for precipitation forecast in China with graph neural network. *Geophys. Res. Lett.* **2024**, *51*, e2023GL106676. [CrossRef]
6. Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **2023**, *619*, 533–538. [CrossRef]

7. Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Stuart, A.; Bhattacharya, K.; Anandkumar, A. Fourier neural operator for parametric partial differential equations. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021; pp. 1–16.
8. Pérez, M.R.; Suárez, M.J.; García, O.J. Spatiotemporal prediction of monthly precipitation: A systematic review of hybrid models. *Hydrol. Res.* 2025, *in press*.
9. Funk, C.; Peterson, P.; Landsfeld, M.; Pedreros, D.; Verdin, J.; Shukla, S.; Husak, G.; Rowland, J.; Harrison, L.; Hoell, A.; et al. The climate hazards infrared precipitation with stations: A new environmental record for monitoring extremes. *Sci. Data* **2015**, *2*, 150066. [CrossRef]
10. López-Bermeo, C.; Montoya, R.D.; Caro-Lopera, F.J.; Diaz-Garcia, J.A. Validation of the accuracy of the CHIRPS precipitation dataset at representing climate variability in a tropical mountainous region of South America. *Phys. Chem. Earth* **2022**, *127*, 103184. [CrossRef]
11. Wang, Y.; Gao, Z.; Long, M.; Wang, J.; Yu, P.S. PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 5123–5132.
12. Trebing, K.; Stanczyk, T.; Mehrkanoon, S. SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture. *Pattern Recognit. Lett.* **2021**, *145*, 178–186. [CrossRef]
13. Ayzel, G.; Scheffer, T.; Heistermann, M. RainNet v1.0: A convolutional neural network for radar-based precipitation nowcasting. *Geosci. Model Dev.* **2020**, *13*, 2631–2644. [CrossRef]
14. Gao, Z.; Shi, X.; Wang, H.; Zhu, Y.; Wang, Y.B.; Li, M.; Yeung, D.Y. Earthformer: Exploring space-time transformers for Earth system forecasting. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 25390–25403.
15. Schulz, B.; Lerch, S. Self-attentive transformer for fast and accurate postprocessing of temperature and wind speed forecasts. *Artif. Intell. Earth Syst.* 2024, *in press*. [CrossRef]
16. Peng, X.; Li, Q.; Chen, L.; Ning, X.; Chu, H.; Liu, J. A structured graph neural network for improving the numerical weather prediction of rainfall. *J. Geophys. Res. Atmos.* **2023**, *128*, e2023JD039011. [CrossRef]
17. Urrea, V.; Ochoa, A.; Mesa, O. Seasonality of rainfall in Colombia. *Water Resour. Res.* **2019**, *55*, 4149–4162. [CrossRef]
18. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017; pp. 1–14.
19. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph attention networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018; pp. 1–12.
20. Hamilton, W.L.; Ying, R.; Leskovec, J. Inductive representation learning on large graphs. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1024–1034.
21. Roe, G.H. Orographic precipitation. *Annu. Rev. Earth Planet. Sci.* **2005**, *33*, 645–671. [CrossRef]
22. Kratzert, F.; Klotz, D.; Herrnegger, M.; Sampson, A.K.; Hochreiter, S.; Nearing, G.S. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resour. Res.* **2019**, *55*, 11344–11354. [CrossRef]
23. Wani, O.A.; Mahdi, S.S.; Yeasin, M.; Kumar, S.S.; Gagnon, A.S.; Danish, F.; Al-Ansari, N.; El-Hendawy, S.; Mattar, M.A. Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas. *Sci. Rep.* **2024**, *14*, 27876. [CrossRef] [PubMed]
24. He, R.; Zhang, L.; Chew, A.W.Z. Data-driven multi-step prediction and analysis of monthly rainfall using explainable deep learning. *Expert Syst. Appl.* **2024**, *235*, 121160. [CrossRef]
25. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.