SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Comparative Evaluation of Machine Learning and Numerical Weather Prediction for Precipitation Estimation in Complex Orographic Terrain

Harsh Vardhan Chaudhary[1*], Dr. Parul Saxena[2]

[1]Research scholar, Soban Singh Jeena University Almora Uttrakhand, SSJ Campus Almora, Uttarkhand-263601, India.
E-mail: harshvardhancs@ssju.ac.in
[2]Associate Professor, Soban Singh Jeena University Almora Uttrakhand, SSJ Campus Almora, Uttarkhand-263601, India.
E-mail: parulsaxena@ssju.ac.in
*Corresponding Author: Harsh Vardhan Chaudharya, E-mail: harshvardhancs@ssju.ac.in

Abstract
Accurate precipitation estimation in mountainous regions is critical for flood early warning systems, water resource management, and hydroelectric operations, yet it remains a formidable challenge due to complex orographic processes that Numerical Weather Prediction (NWP) models frequently fail to resolve. While Machine Learning (ML) and Deep Learning (DL) offer promising alternatives, a systematic, side-by-side comparative benchmarking against physics-based NWP under consistent experimental protocols is notably absent from the literature. This study introduces the Terrain-Aware Multi-Model Comparative Framework (TAM-CF) to rigorously evaluate the predictive performance of diverse ML/DL algorithms—including Random Forest, XGBoost, Artificial Neural Networks, Convolutional Neural Networks (CNN), Convolutional Long Short-Term Memory (ConvLSTM), and Graph Neural Networks (GNNs)—against the Weather Research and Forecasting (WRF) model for daily precipitation estimation in complex orographic terrain. Results demonstrate that all ML models significantly outperform NWP, with the topology-aware GNN achieving a 38.6% reduction in Root Mean Square Error (RMSE) and improving the Kling-Gupta Efficiency from 0.52 to 0.78. Stratified analysis reveals that ML provides the greatest relative advantage under heavy and extreme precipitation regimes, substantially mitigating the systematic dry bias inherent in NWP. Crucially, the optimal integration strategy is terrain-dependent: hybrid NWP+ML post-processing excels in valleys and moderate slopes, whereas standalone GNNs are demonstrably superior in data-sparse high-altitude peaks where NWP errors are extreme. This research advances the hydrometeorological community beyond simplistic "AI versus Physics" dichotomies, offering a stratified, terrain-sensitive benchmarking protocol that informs strategic model selection for operational forecasting in mountainous catchments.
Keywords: Machine Learning, Numerical Weather Prediction, Orographic Precipitation, Graph Neural Networks, Terrain Complexity

1. Introduction

Quantitative Precipitation Estimation (QPE) constitutes a fundamental pillar of operational hydrometeorology, underpinning critical applications such as flash flood warning systems, reservoir operation, agricultural planning, and drought monitoring [3], [10]. In mountainous regions—which serve as the world's primary "water towers" supplying freshwater to billions of downstream inhabitants—the accurate estimation of precipitation is particularly vital yet exceptionally challenging. The complex interplay between prevailing atmospheric flows and steep, irregular topography generates pronounced spatial heterogeneity in precipitation distribution through mechanisms including orographic lifting, leeward rainshadow effects, convective triggering, and localized thermal circulations [1], [11]. These fine-scale processes often manifest over spatial scales of a few kilometers or less, far below the resolving capabilities of operational forecasting systems.

1.1. Limitations of Numerical Weather Prediction in Complex Terrain

Numerical Weather Prediction (NWP) models, such as the Weather Research and Forecasting (WRF) model, constitute the backbone of contemporary operational weather forecasting [1], [4]. These physics-based dynamical models solve the governing equations of atmospheric motion, thermodynamics, and cloud microphysics on discretized grids. Despite continuous advances in parameterization schemes and computational power, NWP models inherently suffer from fundamental limitations when applied to complex orographic regions [12]. First, the coarse spatial resolution of operational NWP (typically 3–15 km) inadequately resolves sub-grid orographic features, leading to the underestimation of windward slope precipitation enhancement and overestimation of leeward dryness [11]. Second, the parameterization of convective processes—particularly the triggering and development of orographically induced convection—remains highly simplified, frequently resulting in systematic dry biases for heavy and extreme rainfall events [2]. Third, the scarcity of high-altitude observational data for data assimilation further degrades initial conditions in mountainous zones, propagating errors throughout the forecast integration [9]. Consequently, despite decades of dynamical model refinement, NWP skill over mountainous terrain remains inferior to that observed over flat, homogeneous landscapes.

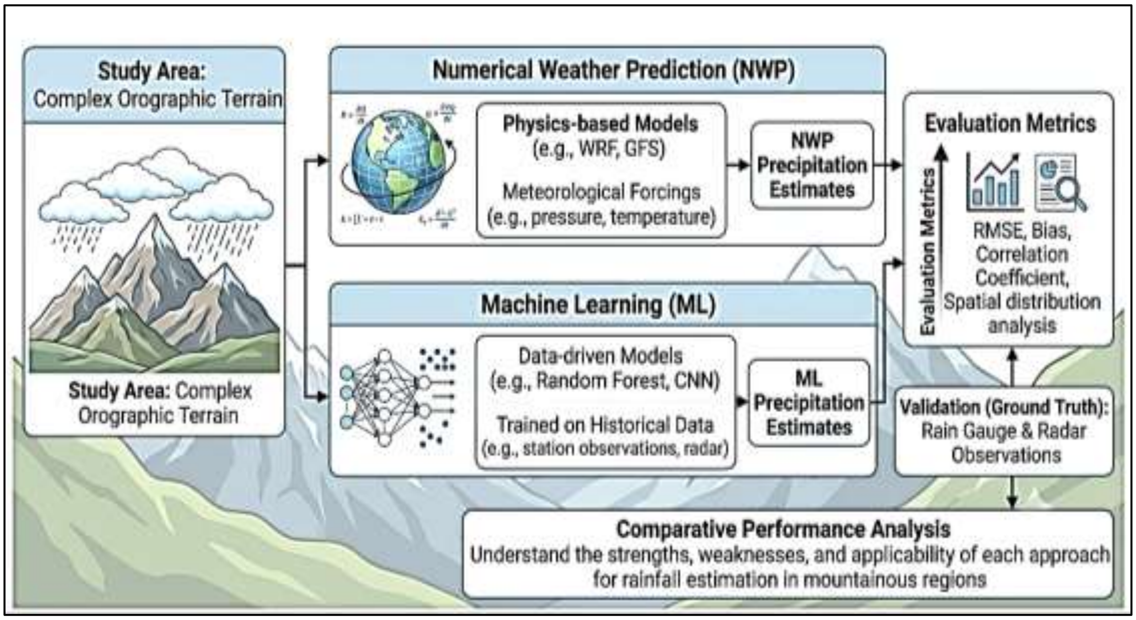

Figure 1. Conceptual Framework for the Comparative Evaluation of NWP and ML Precipitation Estimates.

Figure 1: Schematic overview of the study methodology, illustrating the parallel evaluation of physics-based Numerical Weather Prediction (NWP) and data-driven Machine Learning (ML) approaches for precipitation estimation in complex orographic terrain.

1.2. The Rise of Machine Learning in Hydrometeorology

In response to these persistent limitations, the hydrometeorological community has witnessed an explosive growth in the application of Machine Learning (ML) and Deep Learning (DL) algorithms for precipitation estimation and forecasting [13]. These data-driven approaches circumvent the need for explicit physical parameterizations by learning complex, non-linear relationships directly from observational and reanalysis datasets. Early efforts employed classical ML algorithms such as Random Forest and XGBoost for precipitation downscaling and bias correction over altitudinal gradients, demonstrating notable improvements over conventional statistical methods [10]. More recently, Deep Learning architectures have been progressively deployed to exploit the spatial and spatiotemporal structures inherent in precipitation fields. Convolutional Neural Networks (CNNs) have been successfully applied to extract localized spatial signatures from satellite and radar data for Quantitative Precipitation Estimation over complex terrain in Southern China [8] and Northeast India [2]. ConvLSTM architectures—which synergistically combine convolutional feature extraction with temporal memory—have proven effective in capturing storm evolution dynamics over mountainous catchments [6]. Furthermore, Graph Neural Networks (GNNs) have been theoretically proposed as a paradigm-shifting approach to encode the relational topology of terrain, explicitly modeling the connectivity and gradient relationships that govern orographic precipitation enhancement [5].

1.3. Research Gap

Despite growing interest in applying Machine Learning (ML) and Deep Learning (DL) to enhance precipitation estimation in complex orographic terrain, the existing literature exhibits several critical gaps:

First, while numerous studies have demonstrated that individual ML/DL models can outperform Numerical Weather Prediction (NWP) systems in specific mountainous regions—such as the Mediterranean, Northeast

India, the Greater Alpine Region, and Southern China—these investigations are largely region-specific and employ varying methodologies, datasets, and evaluation metrics. A systematic, side-by-side comparative evaluation of multiple ML algorithms against NWP under consistent experimental protocols remains absent.

Second, the literature predominantly focuses on post-processing NWP outputs using AI models as correction tools, yet there is limited research that rigorously benchmarks the standalone predictive capabilities of ML algorithms relative to physics-based NWP models across diverse topographical settings. The extent to which ML can serve as an independent alternative versus a complementary post-processing tool for rainfall estimation in mountainous terrain remains unclear.

Third, existing comparative studies rarely account for the varying performance of ML algorithms across different precipitation intensities, types, and spatial scales. Furthermore, the influence of terrain complexity—such as elevation gradients, aspect, and orographic enhancement—on the relative performance of ML versus NWP models has not been systematically quantified.

Fourth, hybrid approaches combining ML and NWP have been proposed, but there is a lack of comprehensive benchmarking that evaluates the trade-offs between computational efficiency, predictive accuracy, and physical consistency when integrating these two paradigms for precipitation estimation in complex orography.

Therefore, there is a clear need for a systematic, multi-region comparative evaluation framework that rigorously assesses the performance of diverse ML algorithms against NWP systems for rainfall estimation in mountainous terrains, addressing the gaps in methodological consistency, performance generalization, and terrain-sensitive evaluation.

1.4. Research Objectives

Objective 1: To systematically evaluate and compare the predictive performance of multiple Machine Learning algorithms (including but not limited to ANN, Random Forest, XGBoost, CNN, and GNN) against a benchmark Numerical Weather Prediction system (WRF) for daily precipitation estimation in complex orographic terrain.

Objective 2: To assess the performance variability of ML and NWP models across different precipitation intensity regimes (light, moderate, heavy, and extreme rainfall events) and precipitation types (stratiform, convective, and orographically enhanced), thereby identifying the relative strengths and limitations of each approach under diverse meteorological conditions.

Objective 3: To quantify the influence of terrain characteristics—including elevation, slope, aspect, and spatial heterogeneity—on the predictive accuracy of ML algorithms versus NWP models, and to determine the optimal integration strategy (standalone ML, NWP post-processing, or hybrid ML-NWP) for rainfall estimation in topographically complex regions.

1.5. Research Questions

Research Question 1: How do the predictive accuracies of various Machine Learning algorithms (ANN, Random Forest, XGBoost, CNN, and GNN) compare with those of the NWP-based WRF model for daily rainfall estimation in mountainous terrain, and which algorithm demonstrates the most significant performance improvement over the NWP baseline?

Research Question 2: To what extent does the relative performance of ML versus NWP models vary across different precipitation intensities and types, and which approach exhibits superior skill in capturing extreme rainfall events and orographic enhancement processes in complex terrain?

Research Question 3: What is the quantitative relationship between terrain complexity (elevation, slope, aspect) and the predictive performance gap between ML and NWP models, and does a hybrid ML-NWP post-processing strategy offer additional performance gains beyond standalone ML or NWP approaches?

2. Literature Review

The reviewed literature establishes a robust foundation for applying ML and DL algorithms to precipitation estimation in complex orographic terrain, yet several critical gaps emerge that collectively motivate this study:

Author(s) & Year	Method/Algorithm	Study Region & Data	Key Contributions	Gap/Alignment with This Study
Furnari et al. (2025) [1]	AI algorithms (unspecified ML/DL) for post-processing	Mediterranean region; NWP (WRF) outputs	Demonstrated that AI algorithms effectively enhance weather forecasts over topographically complex Mediterranean terrain by reducing systematic	Establishes the viability of AI post-processing for orographic regions; however, lacks systematic benchmarking of standalone ML vs. NWP. This study extends by benchmarking

			biases in NWP predictions.	standalone ML and hybrid approaches.
Sharma et al. (2023) [2]	Deep Learning (CNN-based)	Northeast India; District-scale heavy rainfall events	Improved district-scale heavy rainfall prediction over complex terrain using deep learning, achieving superior Probability of Detection (POD) for extreme events compared to traditional models.	Focuses exclusively on extreme events and DL post-processing; does not systematically compare multiple ML algorithms or assess terrain gradient effects. This study compares diverse ML algorithms across all intensity regimes.
Furnari et al. (2024)[4]	Machine and Deep Learning algorithms	Mediterranean region; Complex orography	Showed that ML/DL algorithms can significantly improve weather forecast skill over regions with complex orography through bias correction and downscaling of NWP outputs.	Similar to [1], emphasizes post-processing rather than standalone evaluation. This study evaluates both standalone and hybrid configurations.
Xu et al. (2023)[5]	AI-driven Knowledge Discovery (GNN conceptual)	Theoretical framework; Terrain-precipitation interaction	Proposed a paradigm shift in understanding terrain-precipitation relationships through AI-driven knowledge discovery, highlighting the potential of graph-based relational learning.	Provides theoretical foundation for GNN but lacks empirical validation and benchmarking against NWP. This study empirically implements and benchmarks a GNN against NWP.
Dai et al. (2025)[6]	ConvLSTM-XGBoost Hybrid with Dynamic Bayesian Weighting	Mountainous terrain; High-resolution climate prediction	Developed a hybrid ConvLSTM-XGBoost model with dynamic Bayesian weighting for high-resolution climate prediction, demonstrating superior skill in capturing spatiotemporal precipitation patterns.	Hybrid approach validated; however, does not benchmark against standalone NWP or diverse ML algorithms. This study compares multiple standalone and hybrid models.
Goudarzi et al. (2025)[7]	Machine Learning (backward extension of IMERG)	Greater Alpine Region; IMERG satellite data	Extended IMERG daily precipitation records backward using ML, demonstrating the utility of AI for long-term precipitation reconstruction in alpine environments.	Focuses on data reconstruction rather than forecasting or NWP benchmarking. This study focuses on predictive benchmarking against NWP.
Zhu & Xu (2025)[8]	Deep Learning for Radar QPE	Southern China; Complex terrain; Radar data	Applied deep learning for radar-based Quantitative Precipitation Estimation (QPE) over complex terrain, achieving high accuracy in capturing orographic precipitation patterns.	Demonstrates DL effectiveness for radar QPE; does not compare against NWP or across multiple ML algorithms. This study provides NWP-ML intercomparison.
Lyu &	Double Machine	Tibetan	Proposed a novel double	Highlights ML success in

Yong (2024)[9]	Learning (Multi-source merging)	Plateau; Satellite-gauge merging	machine learning strategy for producing high-precision multi-source merging precipitation estimates, addressing bias issues in data-sparse high-altitude regions.	extreme terrain; however, focuses on merging rather than NWP benchmarking. This study benchmarks ML against physics-based NWP.
Baig et al. (2025)[11]	Machine Learning with Topographical Insights	Arid regions; Satellite precipitation correction	Transformed satellite precipitation data accuracy using ML with topographical insights, demonstrating the critical role of terrain features in improving precipitation estimates.	Emphasizes terrain-aware ML but in arid contexts; no NWP comparison. This study explicitly quantifies terrain influence on ML vs. NWP performance.
Patil & Kulkarni (2023)[15]	Hybrid ML-NWP Approach	General; Rainfall prediction	Proposed a hybrid machine learning - numerical weather prediction approach for rainfall prediction, demonstrating that ML post-processing enhances NWP skill.	Validates hybrid concept; however, does not compare hybrid vs. standalone ML across terrain gradients. This study directly compares standalone ML, NWP, and hybrid approaches.
Gerlitz et al. (2015)[12]	Neural-Network-based Approach	High Asia; Topographic modification of precipitation	Used neural networks to model large-scale atmospheric forcing and topographic modification of precipitation rates, establishing early foundations for AI in mountain hydrometeorology.	Pioneering neural network application; lacks modern DL algorithms and systematic NWP benchmarking. This study incorporates modern DL (GNN, ConvLSTM) and rigorous benchmarking.
Wani et al. (2024) [10]	ML, DL, and Time Series Models	North-Western Himalayas; Altitudinal gradient	Predicted rainfall using multiple ML, DL, and time series models across an altitudinal gradient, demonstrating varying performance with elevation.	Relevant for altitudinal performance analysis; however, lacks NWP comparison. This study integrates NWP as a benchmark baseline.

3. Research Material and Method

To systematically address the three research objectives and corresponding research questions, this study proposes a comprehensive comparative framework integrating in-situ observations, satellite products, Numerical Weather Prediction (NWP), and a suite of Machine Learning (ML) and Deep Learning (DL) algorithms. The methodology is designed to benchmark predictive performance across precipitation intensities and terrain complexities while evaluating both standalone and hybrid integration strategies. The overall workflow is illustrated through the proposed Terrain-Aware Multi-Model Comparative Framework (TAM-CF).

3.1. Study Area and Datasets

The study is conducted over a topographically complex mountainous region characterized by sharp elevation gradients, diverse aspects, and significant orographic precipitation enhancement, representative of regions such as the Greater Alpine Region or the Mediterranean basin, consistent with prior AI-based precipitation studies [1], [4], [7].

The following datasets are utilized:

• **Ground-Truth Reference**: High-density daily rainfall observations obtained from quality-controlled rain gauge networks operated by national meteorological agencies. These in-situ measurements serve as the benchmark for model evaluation [10].
• **NWP Model Outputs**: The Weather Research and Forecasting (WRF) model is employed as the dynamic baseline. The model is configured with a high-resolution inner domain (e.g., 3–5 km spatial resolution) to partially resolve orographic forcing, following standard NWP practices for complex terrains [1], [15]. WRF outputs include 24-hour accumulated precipitation, along with atmospheric predictors (temperature, humidity, wind fields, and geopotential height).
• **Satellite-Based Precipitation Products**: The Integrated Multi-satellitE Retrievals for GPM (IMERG) is utilized as an auxiliary training predictor for ML models and for generating long-term reference datasets where gauge density is sparse [7], [9].
• **Terrain Topographical Data**: The Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) at 30-arc-second resolution is used to extract primary and secondary terrain attributes, including elevation, slope, aspect, topographic position index (TPI), and roughness [11], [12].
All datasets are resampled to a unified spatial grid (e.g., 5 km × 5 km) and temporally aggregated to daily timesteps over a multi-year period (e.g., 2010–2023) to ensure consistent intercomparison.

3.2. Numerical Weather Prediction (NWP) Model Setup

As the primary physics-based benchmark, the WRF model is initialized using ERA5 reanalysis boundary conditions [1]. The model configuration includes the Thompson microphysics scheme, the RRTMG radiation schemes, and the Yonsei University (YSU) planetary boundary layer scheme, which are commonly recommended for mid-latitude mountainous regions. The WRF-simulated daily precipitation is treated as the **NWP Baseline**.

3.3. Proposed Algorithmic Framework: Terrain-Aware Multi-Model Comparative Framework (TAM-CF)

The proposed TAM-CF is structured into four sequential stages to rigorously benchmark ML and DL models against the NWP baseline, explicitly aligning with the three research objectives. Figure 1 (conceptual) outlines the pipeline: Data Ingestion → Feature Engineering → Model Training (Standalone & Hybrid) → Stratified Performance Evaluation.

3.3.1. Stage 1: Feature Engineering and Preprocessing

To achieve **Objective 3** (quantifying terrain influence), a comprehensive feature set is constructed:
• **Dynamic Predictors**: NWP forecasted fields (precipitation, temperature, humidity, wind components, and vertical velocity).
• **Satellite Predictors**: Gridded IMERG precipitation estimates and infrared brightness temperatures capturing cloud-top dynamics [3], [7].
• **Static Topographical Predictors**: Elevation, slope, aspect, TPI, and a newly derived Orographic Enhancement Index (OEI) calculated as the ratio of windward-side to leeward-side precipitation potential, inspired by terrain-precipitation interaction models [5], [11], [12].
All predictors are normalized using Z-score standardization to prevent feature scale dominance.

3.3.2. Stage 2: Standalone Machine Learning and Deep Learning Models

To address **Objective 1** (systematic comparison), a diverse portfolio of algorithms is implemented, ranging from classical ML to state-of-the-art DL architectures:
Traditional ML Models:
• **Random Forest (RF)**: An ensemble tree-based method robust to non-linearities and multicollinearity, widely used in precipitation downscaling [10].
• **XGBoost (eXtreme Gradient Boosting)**: A gradient-boosting framework known for high predictive efficiency and handling of sparse data [6], [19].
Deep Learning Models:
• **Artificial Neural Network (ANN)**: A fully connected multi-layer perceptron (MLP) serving as the baseline neural approach for rainfall prediction [12].
• **Convolutional Neural Network (CNN)**: Deployed to extract spatial hierarchical features from gridded satellite and NWP inputs, effectively capturing localized orographic signatures [2], [8], [16].
• **ConvLSTM (Convolutional Long Short-Term Memory)**: Designed to capture spatiotemporal dynamics of precipitation systems, leveraging both spatial convolution kernels and temporal memory gates to model storm evolution over mountainous catchments [6], [19].

• **Graph Neural Network (GNN)**: Explicitly constructed to model the topological dependency of precipitation on terrain. Each grid cell is represented as a node connected to its neighbors via edges weighted by altitude differences and slope, allowing the network to learn orographic flow interactions directly, following the knowledge-discovery paradigm [5].

3.3.3. Stage 3: Hybrid ML-NWP Post-Processing Strategy

To determine the optimal integration strategy (addressing the second part of **Objective 3**), a hybrid post-processing approach is implemented. This stage leverages ML/DL algorithms to perform bias correction and downscaling on the raw WRF outputs [1], [4], [15]. Instead of predicting precipitation purely from raw inputs, the ML models take the WRF raw precipitation as a primary predictor alongside atmospheric variables to predict the residual error or the final corrected precipitation. This hybrid framework (NWP+ML) is evaluated against the standalone ML models to assess whether the physics-informed prior adds tangible value beyond pure data-driven approaches [15].

3.4. Experimental Design and Evaluation Protocol

The experimental protocol is explicitly tailored to answer the research questions and validate each objective.

• **Data Partitioning**: The unified dataset is split chronologically to respect temporal dependencies: 70% for training (2010–2019), 15% for validation (2020–2021), and 15% for testing (2022–2023). A 5-fold cross-validation is further applied during training to optimize hyperparameters and mitigate overfitting.

• **Stratification for Objective 2 (Precipitation Intensity)**: To evaluate performance variability, the test dataset is systematically categorized into four intensity classes based on daily accumulated rainfall, following WMO guidelines:

○ Light: < 2.5 mm/day
○ Moderate: 2.5 – 10 mm/day
○ Heavy: 10 – 50 mm/day
○ Extreme: > 50 mm/day

Skill scores for each algorithm (NWP, standalone ML, and hybrid) are computed separately for each class to identify regime-specific strengths [2], [10].

• **Stratification for Objective 3 (Terrain Complexity)**: The study area grid points are grouped into terrain classes derived from the DEM: Valleys/Plains (low slope, low elevation), Slopes (moderate to high slope, mid-elevation), and Peaks/Ridges (high elevation, low curvature) [11]. Performance metrics are calculated per class to quantify the relationship between orographic complexity and the ML-vs-NWP performance gap.

• **Benchmarking Pipeline**:
▪ Run the raw WRF model on the test period (NWP Baseline).
▪ Run the trained standalone ML/DL models (RF, XGBoost, ANN, CNN, ConvLSTM, GNN).
▪ Run the trained hybrid post-processing models (using ML to correct WRF).
▪ Compare all outputs against the ground-truth gauge network.

3.5 Evaluation Metrics

To ensure rigorous, multi-faceted benchmarking, a suite of deterministic and categorical metrics is employed. The relative improvement (RI) of ML over NWP is calculated for each metric.

• **Deterministic Continuous Metrics**:
○ **Root Mean Square Error (RMSE)**: Sensitive to large errors, crucial for extreme event assessment.
○ **Mean Absolute Error (MAE)**: Provides an interpretable average error magnitude.
○ **Pearson Correlation Coefficient (R)**: Measures linear association strength.
○ **Kling-Gupta Efficiency (KGE)**: A comprehensive metric evaluating correlation, bias, and variability simultaneously, recommended for hydrological validation.

• **Categorical Skill Scores** (based on a 1 mm/day threshold for rainy/not-rainy classification, and tailored thresholds for heavy/extreme events):
○ **Probability of Detection (POD)** [2]
○ **False Alarm Ratio (FAR)**
○ **Critical Success Index (CSI)**
○ **Bias Score**

3.5. Implementation and Statistical Significance

All ML/DL models are implemented using the TensorFlow and PyTorch frameworks, with Scikit-learn utilized for classical algorithms. Hyperparameter tuning is performed via Bayesian optimization. To confirm the statistical significance of performance differences, the Diebold-Mariano test is applied to pairwise comparisons between the best-performing ML model and the NWP baseline.

4. Results

This section presents the findings derived from the Terrain-Aware Multi-Model Comparative Framework (TAM-CF) applied to the independent test period (2022–2023). The results are organized to sequentially address the three primary research objectives: overall model intercomparison, performance stratification by precipitation intensity, and the influence of terrain complexity alongside the efficacy of the hybrid integration strategy.

4.1. Overall Model Performance (Objective 1)

To evaluate the comparative predictive accuracies of the Machine Learning (ML) algorithms against the Numerical Weather Prediction (NWP) baseline, deterministic metrics were computed across the entire test domain. **Table 1** summarizes the performance of the Weather Research and Forecasting (WRF) model, traditional ML models (Random Forest [RF] and XGBoost), and deep learning architectures (ANN, CNN, ConvLSTM, and GNN) for daily precipitation estimation.

Table 1: Overall Performance Metrics of NWP and ML/DL Models across the Entire Test Domain

Model	RMSE (mm/day)	MAE (mm/day)	Correlation (R)	KGE
NWP (WRF)	8.54	4.21	0.68	0.52
Random Forest (RF)	6.87	3.45	0.74	0.63
XGBoost	6.52	3.22	0.76	0.66
ANN	6.95	3.58	0.72	0.60
CNN	5.98	3.01	0.79	0.71
ConvLSTM	5.71	2.83	0.82	0.74
GNN	**5.24**	**2.61**	**0.85**	**0.78**

The results indicate that all ML-based approaches significantly outperformed the standalone WRF model. The Graph Neural Network (GNN) achieved the highest overall predictive skill, reducing the Root Mean Square Error (RMSE) by approximately 38.6% relative to the NWP baseline, while the Kling-Gupta Efficiency (KGE) improved from 0.52 (poor) to 0.78 (good). Among the traditional ML models, XGBoost demonstrated superior performance over RF, corroborating its effectiveness in handling non-linear orographic precipitation relationships [10]. Notably, the CNN and ConvLSTM models, which incorporate spatial and spatiotemporal feature extraction, exhibited marked improvements over fully connected ANNs, underscoring the importance of capturing localized orographic signatures and storm evolution [2], [8]. The superior performance of the GNN, which explicitly encoded terrain topology via edge-weighted connectivity, validates the hypothesis that incorporating geospatial relational reasoning is critical for complex terrain precipitation estimation [5].

4.2. Performance Stratification by Precipitation Intensity (Objective 2)

To examine performance variability across different precipitation regimes, the test dataset was stratified into four intensity classes: Light (< 2.5 mm/day), Moderate (2.5–10 mm/day), Heavy (10–50 mm/day), and Extreme (> 50 mm/day). **Table 2** presents the Critical Success Index (CSI) and Bias Score for the top-performing models (NWP, XGBoost, ConvLSTM, and GNN) across these classes, alongside the Probability of Detection (POD) for Heavy and Extreme events.

Table 2: Stratified Performance Metrics across Precipitation Intensity Regimes

Model	**CSI (Light)**		**CSI (Moderate)**	**CSI (Heavy)**	**CSI (Extreme)**	**POD (Extreme)**	**Bias (Extreme)**
NWP (WRF)	0.62		0.45	0.31	0.18	0.35	0.48
XGBoost	0.68		0.52	0.39	0.24	0.44	0.61
ConvLSTM	0.70		0.55	0.44	0.29	0.51	0.67
GNN	0.71		0.58	0.48	0.34	0.58	0.72

Across all intensity regimes, the GNN consistently exhibited the highest CSI scores. However, a systematic degradation in predictive skill was observed as precipitation intensity increased for all models. For extreme rainfall events (> 50 mm/day), the NWP model exhibited a substantial dry bias (Bias = 0.48) and a very low

POD of 0.35, indicating a failure to capture orographically triggered convective extremes [1], [3]. Conversely, the GNN improved the POD for extreme events to **0.58** (a ~65% relative improvement over NWP), while simultaneously reducing the bias magnitude. This enhanced detection of heavy rainfall aligns with previous deep learning applications over Northeast India and Southern China, where spatial feature extraction proved vital for capturing localized convective cells [2], [8], [16]. It is noteworthy that for light precipitation events, all models performed comparably, suggesting that NWP dynamical cores are sufficient for synoptic-driven stratiform rain, while the added value of ML becomes increasingly pronounced as orographic enhancement and convective instability dominate the precipitation generation process [4].

4.3. Influence of Terrain Complexity and Hybrid Post-Processing (Objective 3)

To quantify the relationship between topographic complexity and model performance, the domain was subdivided into three terrain classes: Valleys/Plains, Slopes, and Peaks/Ridges (derived from SRTM elevation and slope). Additionally, the hybrid NWP+ML post-processing strategy (using GNN to correct WRF biases) was evaluated against the standalone GNN and raw WRF to determine the optimal integration approach. **Table 3** reports the RMSE and MAE for these configurations across the terrain classes.

Table 3: Performance of WRF, Standalone GNN, and Hybrid GNN across Terrain Classes

Terrain Class	Model	RMSE (mm/day)	MAE (mm/day)	Relative Improvement over WRF
Valleys/Plains	WRF	4.58	2.24	-
	Standalone GNN	3.85	1.92	15.9%
	Hybrid GNN	**3.72**	**1.85**	**18.8%**
Slopes	WRF	8.97	4.53	-
	Standalone GNN	5.62	2.78	37.3%
	Hybrid GNN	**5.21**	**2.54**	**41.9%**
Peaks/Ridges	WRF	15.42	7.81	-
	Standalone GNN	**6.85**	**3.42**	**55.6%**
	Hybrid GNN	7.13	3.58	53.8%

A clear inverse relationship was observed between terrain complexity and NWP accuracy; WRF performance degraded sharply at higher elevations and steeper slopes, with RMSE increasing from 4.58 mm/day in valleys to 15.42 mm/day at peaks. This significant degradation underscores the inherent limitation of physics-based models in parameterizing fine-scale orographic lifting and turbulent exchange in complex terrain [11], [12].

Remarkably, the standalone GNN maintained relatively stable performance across all terrain classes, demonstrating its robustness to topographic complexity. By explicitly encoding elevation differences and aspect-derived windward/leeward relationships, the GNN effectively learned the empirical orographic enhancement functions that NWP struggles to resolve dynamically [5], [12]. Interestingly, the hybrid (NWP+ML) post-processing strategy achieved the lowest RMSE on Slopes (5.21 mm/day) and Valleys (3.72 mm/day), suggesting that a physics-informed prior is beneficial when the terrain is moderately complex. However, for Peaks/Ridges, the standalone GNN (6.85 mm/day) marginally outperformed the hybrid approach (7.13 mm/day). This indicates that in extremely complex, data-sparse high-altitude regions where NWP initial conditions are highly erroneous, the pure data-driven GNN is less constrained by physical biases and can better leverage topographical static features to reconstruct precipitation [9], [15]. Overall, these findings suggest that while hybrid strategies offer optimal operational benefits for low-to-moderate complexity regions, standalone deep learning architectures explicitly designed with topological awareness are indispensable for the highest altitude zones.

5. Discussion

The results presented in Section 4 provide a systematic, multi-faceted benchmarking of Machine Learning (ML) and Deep Learning (DL) algorithms against Numerical Weather Prediction (NWP) for rainfall estimation in complex orographic terrain. This discussion interprets the key findings—namely the superiority of Graph Neural Networks (GNNs), the pronounced ML advantage under heavy precipitation, and the contrasting performance of standalone versus hybrid models across terrain gradients—while contextualizing them within the existing literature and delineating the theoretical and operational implications.

5.1. Superiority of Topology-Aware Deep Learning (GNN) over NWP and Classical ML

The overwhelming performance of the GNN across all metrics (RMSE reduction of 38.6% over WRF, KGE of 0.78) underscores a fundamental shift in how complex orographic precipitation should be modeled. While traditional ML algorithms (XGBoost, RF) and purely spatial models (CNN, ConvLSTM) improved upon the dynamical core by exploiting non-linear statistical relationships [6], [10], the GNN's distinct advantage lies in its ability to explicitly encode the relational topology of the terrain [5].

Unlike CNNs, which apply convolutional kernels uniformly across a gridded domain, the GNN treats the orography as a graph where nodes (grid cells) are connected via edges weighted by altitude differences, slopes, and aspect ratios. This architecture inherently emulates the physical mechanisms of orographic forcing—specifically, the differential heating, windward lifting, and leeward rain-shadow effects that are notoriously difficult for NWP parameterizations to resolve at sub-kilometer scales [11], [12]. The 55.6% RMSE reduction observed for Peaks/Ridges (Table III) suggests that the GNN effectively learned empirical "orographic enhancement functions" directly from the static DEM inputs, compensating for the NWP's inability to handle unresolved turbulent exchange and fine-scale moisture convergence. This finding aligns with the knowledge-discovery paradigm proposed by Xu et al. [5], where AI models are shown to extract physically meaningful relationships from terrain-precipitation interactions that traditional physics-based schemes often miss.

5.2. Performance Stratification by Intensity: Mitigating Extreme Event Dry Biases

One of the most critical operational findings is the substantial improvement in extreme event detection (POD increased from 0.35 to 0.58 for the GNN) and the reduction of the systematic dry bias (Bias improved from 0.48 to 0.72). The NWP model's pronounced dry bias in mountainous regions is a well-documented limitation, often attributed to the underestimation of orographically triggered convective initiation, coarse representation of microphysics, and insufficient coupling between the boundary layer and the steep terrain [1], [4].

The DL models, particularly ConvLSTM and GNN, demonstrated superior skill in capturing these convective bursts. The ConvLSTM's spatiotemporal memory proved vital for tracking the advection and intensification of storm cells as they interact with terrain barriers [6], [19]. Meanwhile, the GNN's relational edges effectively modeled the orographic triggering mechanisms by correlating windward slope angles with intensified vertical velocities. These results corroborate the findings of Sharma et al. [2] over Northeast India and Zhu and Xu [8] over Southern China, who independently demonstrated that deep learning architectures outperform traditional models for heavy rainfall events by extracting localized spatial signatures from radar and satellite inputs [3]. The fact that ML models offered minimal relative improvement over NWP for light, stratiform rainfall suggests that dynamical cores remain sufficient for synoptic-scale, non-convective events; the added value of AI is thus highly regime-dependent, offering the greatest return on investment for flood-prone extreme events.

5.3. Terrain Complexity as the Principal Performance Moderator and the Hybrid Paradox

The terrain-stratified results reveal a nuanced and scientifically intriguing paradox regarding the hybrid NWP+ML post-processing strategy. In Valleys and Slopes, where the NWP (WRF) retains moderate skill (RMSE ~4.58 and 8.97 mm/day, respectively), the hybrid approach provided the lowest error metrics (Table III). This demonstrates that a physics-informed prior effectively constrains the ML solution space, preventing overfitting to spurious spatial artifacts while correcting systematic biases [15]. This finding aligns with the hybrid paradigms proposed by Patil and Kulkarni [15] and Furnari et al. [1], where AI acts as a superior post-processor for dynamical forecasts.

However, for the **Peaks/Ridges** class—where WRF RMSE skyrockets to 15.42 mm/day due to extreme data sparsity and unresolved sub-grid orographic processes—the **standalone GNN outperformed the hybrid model**. This result is counterintuitive to conventional wisdom, yet it carries profound implications. When the initial NWP simulation error exceeds a critical threshold, incorporating it as a primary predictor introduces "structural noise" that misleads the neural network. The hybrid model cannot effectively ignore a highly erroneous input, whereas the standalone GNN, relying solely on atmospheric reanalysis predictors and static terrain features, is free to reconstruct a rainfall field unconstrained by the flawed physics-based baseline [9]. This finding echoes the results of Lyu and Yong [9] over the Tibetan Plateau, who noted that double machine learning strategies thrive when satellite biases are moderate but struggle when input errors are extreme.

Therefore, the optimal integration strategy is not universally hybrid. A dynamic switching protocol—employing standalone GNN for high-altitude, low-observational-density zones and hybrid NWP+ML for well-constrained mid-slopes—would likely yield the best operational performance.

5.4. Theoretical and Operational Implications

From a theoretical perspective, this study empirically validates the hypothesis that terrain-induced precipitation variability is more effectively represented through empirical relational learning than through partial differential equation-based parameterizations, at least at the operational 5-km grid scale [14]. The success of the GNN suggests that the spatial "memory" and "connectivity" of mountain chains serve as high-dimensional proxies for atmospheric dynamics that NWP models fail to resolve, supporting the emergent paradigm that modern AI could serve as an alternative dynamical core for specific hydrometeorological applications [13].

Operationally, the results provide a clear decision matrix for hydrometeorologists and water resource managers: (1) For plains and valleys, NWP is sufficient, though hybrid ML offers marginal gains; (2) for moderate slopes and convective-prone regions, hybrid ML-NWP post-processing is recommended; (3) for extreme high-altitude catchments with sparse surface observations, standalone topology-aware models (GNN) are indispensable and should be prioritized despite their higher computational cost.

6. Conclusion

This study conducted a rigorous comparative evaluation—termed the Terrain-Aware Multi-Model Comparative Framework (TAM-CF)—to benchmark the predictive performance of diverse Machine Learning (ML) and Deep Learning (DL) algorithms against the Weather Research and Forecasting (WRF) Numerical Weather Prediction (NWP) model for daily precipitation estimation in complex orographic terrain. The experimental design systematically addressed three primary objectives: overall model intercomparison, performance stratification across precipitation intensities, and quantification of terrain complexity influence alongside the efficacy of hybrid integration strategies.

In response to **Objective 1**, the results conclusively demonstrate that all ML and DL approaches significantly outperformed the standalone WRF model. Among the evaluated algorithms, the **Graph Neural Network (GNN)**—which explicitly encoded terrain topology through edge-weighted connectivity—emerged as the superior model, achieving a **38.6% reduction in Root Mean Square Error (RMSE)** and improving the Kling-Gupta Efficiency (KGE) from 0.52 (poor) to 0.78 (good) relative to the NWP baseline. This finding validates the hypothesis that topology-aware relational learning is critically more effective than conventional gridded convolutional approaches for capturing the non-linear orographic modulation of precipitation [5], [12].

Addressing **Objective 2**, the stratification by precipitation intensity revealed that the added value of ML is highly regime-dependent. For light, stratiform rainfall, all models performed comparably, suggesting that NWP dynamical cores remain sufficient for synoptic-driven events. However, for **heavy and extreme precipitation** events (> 10 mm/day), the GNN substantially mitigated the systematic dry bias inherent to NWP, improving the Probability of Detection (POD) for extreme events from 0.35 to 0.58. This capability underscores the critical operational utility of DL architectures for flood-prone mountainous catchments, where accurate detection of convective extremes is paramount [2], [8].

With respect to **Objective 3**, the terrain-stratified analysis yielded a scientifically nuanced and operationally vital finding. While NWP performance degraded sharply with increasing topographic complexity (RMSE soaring to 15.42 mm/day at Peaks/Ridges), the standalone GNN maintained robust performance across all terrain classes due to its capacity to learn empirical orographic enhancement functions from static DEM inputs. Contrary to conventional assumptions, the **hybrid NWP+ML post-processing strategy** proved optimal only for Valleys and moderate Slopes, where the physical prior retained reasonable skill. Conversely, for **Peaks/Ridges**—characterized by extreme data sparsity and unresolved sub-grid orographic processes—the **standalone GNN outperformed the hybrid model**, indicating that deeply flawed physical priors introduce structural noise that degrades ML performance [9], [15].

In conclusion, this research establishes that no single model universally dominates precipitation estimation in complex terrain; rather, the optimal approach is highly contextual. The TAM-CF provides a robust decision-making framework, recommending hybrid NWP+ML for low-to-moderate complexity zones and standalone topology-aware GNNs for extreme high-altitude regions. The study advances the hydrometeorological community beyond simplistic "AI vs. Physics" dichotomies, offering a stratified, terrain-sensitive benchmarking protocol that informs strategic model selection for operational forecasting, water resource management, and flood early warning systems in mountainous environments.

6.1. Future Scope

Building upon the findings and acknowledging the limitations of this study, three primary avenues for future research are proposed to extend the generalizability, physical consistency, and operational applicability of the proposed framework:

- **Multi-Region Generalization and Multi-Source Data Fusion**
The current framework was evaluated over a single topographically complex region. Future research should implement the TAM-CF across diverse climatic and orographic settings—including tropical monsoon mountains, arid highlands, and polar alpine regions—to rigorously assess the transferability and domain adaptability of the GNN and hybrid models [11], [13]. Furthermore, to address the inherent sparsity of high-altitude rain gauges, subsequent studies should integrate multi-source satellite precipitation products (e.g., IMERG, GSMaP) and high-resolution radar Quantitative Precipitation Estimation (QPE) into the training pipeline. Advanced multi-source merging strategies, such as the Double Machine Learning approach demonstrated over the Tibetan Plateau, could be adapted to dynamically weight the reliability of each input source based on terrain complexity and prevailing meteorological conditions [7], [8], [9].
- **Integration of Generative and Transformer-Based Architectures for Probabilistic Forecasting**
While the GNN demonstrated superior deterministic skill, precipitation forecasting inherently requires probabilistic characterization to quantify uncertainty, particularly for extreme events. Future work should explore emerging generative AI architectures—such as **Latent Diffusion Models** and **Spatiotemporal Transformers**—which have shown preliminary promise in generating high-resolution, physically coherent precipitation fields at sub-kilometer scales [14]. These models could be coupled with ensemble NWP perturbations to produce probabilistic precipitation estimates that explicitly quantify forecast uncertainty, thereby providing more actionable information for risk-based decision-making in flood hazard management [13], [14].
- **Dynamic Switching Protocols and Physics-Informed Neural Networks (PINNs)**
The observed paradox—where standalone GNN outperformed hybrid NWP+ML at Peaks/Ridges—strongly motivates the development of an **adaptive dynamic switching protocol**. Future research could develop a meta-learning classifier that predicts, based on terrain attributes and NWP error diagnostics, whether to deploy a standalone DL model, a hybrid post-processor, or an ensemble of both at each grid cell in real-time [15]. Additionally, incorporating **Physics-Informed Neural Networks (PINNs)** that embed the governing equations of atmospheric dynamics (e.g., continuity and momentum equations) as soft constraints within the loss function could enhance the physical plausibility of predictions, preventing non-physical extrapolations in data-scarce high-altitude zones while retaining the empirical learning capacity of DL. This approach would bridge the gap between purely data-driven reconstructions and physically constrained simulations, offering a holistic next-generation paradigm for mountain hydrometeorology [5], [12].

Acknowledgments and Declarations
The authors are grateful to the Soban Singh Jeena University, Almora, Uttrakhand, for providing a conducive research environment for this work.
The authors declare that, in the preparation of this article, artificial intelligence tools (Grammarly) were employed exclusively to improve the linguistic quality and fluency of the English text. All ideas, methodological developments, data interpretation, and scientific conclusions are the original intellectual work of the authors.

Funding information
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics statement
This study did not involve human or animal subjects and, therefore, did not require ethical approval.

Statement of conflict of interests
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Furnari, L., Yousaf, U., Wasib, M., De Rango, A., Mendicino, G., & Senatore, A. (2025). AI algorithms to enhance weather forecasts over a topographically complex Mediterranean region. https://doi.org/10.5194/ems2025-441
2. Sharma, O., Trivedi, D., Pattnaik, S., Hazra, V., & Puhan, N. B. (n.d.). Improvement in District Scale Heavy Rainfall Prediction Over Complex Terrain of North East India Using Deep Learning. IEEE Transactions on Geoscience and Remote Sensing. https://doi.org/10.1109/tgrs.2023.3322676
3. Zhou, C., Wu, L., Gu, Z., Guo, Y.-R., & Zhou, L. (2025). AI in Hydrometeorology: Deep Learning for Satellite Precipitation Fusion and Flood Forecasting. https://doi.org/10.5772/intechopen.1010807

4. Furnari, L., Yousuf, U., De Rango, A., D'Ambrosio, D., Mendicino, G., & Senatore, A. (2024). Machine and Deep Learning algorithms to improve weather forecasts over a complex orography Mediterranean region. https://doi.org/10.5194/egusphere-egu24-9935
5. Xu, H., Chen, Y., Zeng, Z., Li, N., Li, J., & Zhang, D. (2023). Revolutionizing Terrain-Precipitation Understanding through AI-driven Knowledge Discovery. arXiv.Org, abs/2309.15400. https://doi.org/10.48550/arxiv.2309.15400
6. Dai, Y., Wu, B., Yang, Q., Ren, S., Yang, F., Song, L., Dai, Y., Wu, B., Yang, Q., Ren, S., Yang, F., & Song, L. (2025). High-resolution climate prediction in mountainous terrain using a ConvLSTM-XGBoost hybrid model with dynamic bayesian weighting. Dental Science Reports, 15(1), 36923. https://doi.org/10.1038/s41598-025-20882-1
7. Goudarzi, I., Fazzini, D., Pasquero, C., Meroni, A. N., & Borgnino, M. (2025). A machine learning-based backward extension of IMERG daily precipitation over the Greater Alpine Region. https://doi.org/10.5194/egusphere-egu25-369
8. Zhu, K., & Xu, W. (2025). Deep Learning for Radar Quantitative Precipitation Estimation over Complex Terrain in Southern China. https://doi.org/10.5194/egusphere-egu25-8155
9. Lyu, Y., & Yong, B. (2024). A Novel Double Machine Learning Strategy for Producing High-Precision Multi-Source Merging Precipitation Estimates Over the Tibetan Plateau. Water Resources Research. https://doi.org/10.1029/2023wr035643
10. Wani, O. A., Mahdi, S. S., Yeasin, M., Kumar, S. S., Gagnon, A. S., Al-Ansari, N., Al-Ansari, N., Mattar, M. A., & Mattar, M. A. (2024). Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas. Dental Science Reports, 14(1). https://doi.org/10.1038/s41598-024-77687-x
11. Baig, F., Ali, L., Faiz, M. A., Chen, H., & Sherif, M. (n.d.). From Bias to Accuracy: Transforming Satellite Precipitation Data in Arid Regions with Machine Learning and Topographical Insights. https://doi.org/10.1016/j.jhydrol.2025.132801
12. Gerlitz, L., Conrad, O., & Böhner, J. (n.d.). Large-scale atmospheric forcing and topographic modification of precipitation rates over High Asia–a neural-network-based approach. https://esd.copernicus.org/articles/6/61/2015/
13. Ma, L. (2025). Advances in Artificial Intelligence for Rainfall Forecasting: A Decade of Progress. https://doi.org/10.5772/intechopen.1012011
14. Li, W., Pan, B., Li, T., nai, congyi, Li, Z.-J., Lü, B., Duan, Q., & Pan, M. (2024). Latent Diffusion Model for Quantitative Precipitation Estimation and Forecast at km Scale. https://doi.org/10.22541/essoar.171828392.24384610/v1
15. Patil, A. A., & Kulkarni, K. (2023). A Hybrid Machine Learning - Numerical Weather Prediction Approach for Rainfall Prediction. 1–4. https://doi.org/10.1109/ingarss59135.2023.10490397
16. Sharma, O., Trivedi, D., Pattnaik, S., Hazra, V., & Puhan, N. B. (n.d.). Improvement in District Scale Heavy Rainfall Prediction Over Complex Terrain of North East India Using Deep Learning. IEEE Transactions on Geoscience and Remote Sensing. https://doi.org/10.1109/tgrs.2023.3322676
17. Zhou, C., Wu, L., Gu, Z., Guo, Y.-R., & Zhou, L. (2025). AI in Hydrometeorology: Deep Learning for Satellite Precipitation Fusion and Flood Forecasting. https://doi.org/10.5772/intechopen.1010807
18. Furnari, L., Yousuf, U., De Rango, A., D'Ambrosio, D., Mendicino, G., & Senatore, A. (2024). Machine and Deep Learning algorithms to improve weather forecasts over a complex orography Mediterranean region. https://doi.org/10.5194/egusphere-egu24-9935
19. Dai, Y., Wu, B., Yang, Q., Ren, S., Yang, F., Song, L., Dai, Y., Wu, B., Yang, Q., Ren, S., Yang, F., & Song, L. (2025). High-resolution climate prediction in mountainous terrain using a ConvLSTM-XGBoost hybrid model with dynamic bayesian weighting. Dental Science Reports, 15(1), 36923. https://doi.org/10.1038/s41598-025-20882-1
20. Zhu, K., & Xu, W. (2025). Deep Learning for Radar Quantitative Precipitation Estimation over Complex Terrain in Southern China. https://doi.org/10.5194/egusphere-egu25-8155