POLITECNICO DI TORINO
Repository ISTITUZIONALE

Uncertainty-aware methods for enhancing rainfall prediction with deep-learning based post-processing segmentation

Original
Uncertainty-aware methods for enhancing rainfall prediction with deep-learning based post-processing segmentation / Monaco, S., Monaco, L., Apiletti, D., Cremonini, R., Barbero, S.. - In: COMPUTERS & GEOSCIENCES. - ISSN 0098-3004. - ELETTRONICO. - 205:(2025). [10.1016/j.cageo.2025.105992]

Availability:
This version is available at: 11583/3002316 since: 2025-08-04T13:31:16Z

Publisher:
Elsevier

Published
DOI:10.1016/j.cageo.2025.105992

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

27 November 2026

Computers & Geosciences 205 (2025) 105992

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Uncertainty-aware methods for enhancing rainfall prediction with deep-learning based post-processing segmentation

Simone Monaco [a],*, Luca Monaco [b], Daniele Apiletti [a], Roberto Cremonini [c], Secondo Barbero [c]

[a] Department of Control and Computer Engineering, Politecnico di Torino, Italy
[b] Department of Environmental Engineering, Politecnico di Torino, Italy
[c] Environmental Protection and Research Agency of Piedmont (Arpa Piemonte), Italy

ARTICLE INFO

Keywords:
Rainfall prediction
Probabilistic deep learning
Deep learning post-processing

ABSTRACT

Precipitation forecast is critical in flood management, agricultural planning, water resource allocation, and weather warnings. Despite significant advancements in Numerical Weather Prediction (NWP) models, these systems often exhibit substantial biases and errors, particularly at high spatial and temporal resolutions. To address these limitations, we develop and evaluate uncertainty-aware deep learning ensemble architectures, focusing on characterizing forecast uncertainties while achieving high accuracy and an optimal balance between sharpness and reliability.

This study presents SDE U-Net, a novel adaptation of SDE-Net designed specifically for segmentation tasks in precipitation forecasting. We conduct a comprehensive evaluation of state-of-the-art ensemble architectures, including SDE U-Net, and compare their forecast uncertainty against that of a Poor Man's Ensemble (PME, i.e. NWPs forecast average) across diverse meteorological conditions, ranging from non-intense precipitation patterns to intense weather events. As an example case, we focus on predicting daily cumulative precipitation in northwest Italy, though our approach is broadly generalizable.

Our findings demonstrate that all the evaluated probabilistic deep learning models outperform the PME benchmark in terms of median RMSE for both non-intense and intense precipitation events. Among them, SDE U-Net achieves the best overall performance, delivering the lowest RMSE for intense events (2.637×10^{-2}) and demonstrating a more stable error distribution compared to other models. For non-intense events, SDE U-Net perform comparably to other deep learning models, still notably surpassing the baselines. Moreover, SDE U-Net effectively balances sharpness and reliability, making it particularly suitable for operational forecasting of extreme weather. Integrating uncertainty-aware models like SDE U-Net into forecasting workflows can enhance decision-making and preparedness for weather-related hazards.

1. Introduction

Within the broader domain of atmospheric science, precipitation forecast stands out as a critical focus area, attracting considerable attention due to its importance in addressing challenges such as flood risk and water resource management (Hernández et al., 2016; Huang et al., 2022; Shao et al., 2024). Despite notable advancements in numerical weather prediction (NWP) models, accurately representing the intricate variability of weather fields like precipitation remains challenging. Biases and uncertainties persist, especially at high spatial and temporal resolutions, due to atmospheric processes' non-linear and chaotic nature and the inherent approximations in NWPs (Bauer et al., 2015; Rasp et al., 2020). Furthermore, the direct model output (DMO) of NWPs is susceptible to factors such as initial and boundary conditions and model parameterizations. These sensitivities often lead to systematic errors, further compounding the difficulty of reliable precipitation forecasts. In addition, precipitation forecast is more challenging than other meteorological features due to precipitation events' highly imbalanced and sparse distribution. This imbalance makes predictions particularly difficult, especially for intense events, which are crucial for operational decision-making.

As a result, quantifying forecast uncertainty becomes essential. In Italy, for example, forecast uncertainty quantification plays a pivotal role for Civil Protection authorities in assessing the confidence of precipitation forecasts when determining weather alerts and emergency responses (Molini et al., 2009). On the other hand, a known shortcoming of traditional NWPs is their inability to quantify forecast uncertainty (Demargne et al., 2014) effectively.

* Corresponding author.
E-mail address: simone.monaco@polito.it (S. Monaco).

https://doi.org/10.1016/j.cageo.2025.105992
Received 20 February 2025; Received in revised form 26 May 2025; Accepted 19 June 2025
Available online 11 July 2025
0098-3004/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Addressing this need for uncertainty-aware forecasting methods is essential to improve the reliability and applicability of precipitation predictions.

Post-processing techniques address the limitations of NWPs and enhance the reliability of predictions. Traditional statistical approaches, such as Model Output Statistics (MOS) and Ensemble Model Output Statistics (EMOS), have achieved moderate success but often fall short in capturing the complexity of precipitation patterns and associated uncertainties (Gneiting et al., 2007; Scheuerer and Hamill, 2015). Despite its simplicity, the most widespread post-processing technique remains the average of all the available NWPs, sometimes referred as Poor Man's Ensemble (PME).

Recently, Machine Learning Weather Prediction methods (MLWPs) have shown significant potential in advancing geoscience and weather forecasting by leveraging large datasets and uncovering complex patterns that traditional methods struggle to model (Schultz et al., 2021; Vandal et al., 2018; Colomba et al., 2022). These techniques are designed to complement NWPs, including Global Circulation Models (GCMs) and Limited Area Models (LAMs), by expanding the range of available forecast information. They are also used to post-process the direct model output (DMO) from these models (Bi et al., 2023; Rasp et al., 2024).

While achieving high performance, Deep Learning (DL) models often suffer from opaque learning processes and can inadvertently capture biases, hindering generalization. Indeed, weather forecasts may also require more care when addressed using DLs (Ko et al., 2022). Explainability techniques, widely used in geosciences like climatology (Mamalakis et al., 2022a,b) and remote sensing (Monaco et al., 2021), address this by highlighting model priorities during predictions, fostering robustness and trustworthiness.

Alongside deterministic MLWP methods — including graph-based models (Lam et al., 2023; Lang et al., 2024), neural operators (Pathak et al., 2022; Bonev et al., 2023), and transformers (Bi et al., 2023; Nguyen et al., 2023b,a) — there is a growing interest in probabilistic models that quantify prediction uncertainty reliably. Embedding uncertainty from NWP forecasts can contribute to this effort. Therefore, Rather than relying on post hoc explainability, our work focuses on uncertainty quantification as a principled means to evaluate the trustworthiness of model predictions.

A prominent research direction for quantifying uncertainty in neural networks focuses on Bayesian Neural Networks (BNNs) (Denker and LeCun, 1990; MacKay, 1992). BNNs capture prediction uncertainty by assigning probability distributions to model parameters rather than relying on point estimates. While BNNs offer a principled framework for uncertainty quantification, deriving exact parameter posteriors is computationally challenging, particularly for large-scale datasets commonly encountered in tasks like computer vision.

Among non-Bayesian approaches, model ensembling (Lakshminarayanan et al., 2017) stands out as a widely recognized method. This technique involves training multiple Deep Neural Networks (DNNs) with different initializations and estimating uncertainty using the resulting prediction statistics. Some ensemble-based MLWPs enhance diversity by perturbing initial states and model parameters, mirroring physics-based methods. For example, several studies introduce random noise into the initial states to generate ensembles (Pathak et al., 2022; Bi et al., 2023). However, increasing the number of models raises computational costs, making scalability challenging for larger models.

An alternative approach to quantify uncertainty without training multiple models is Monte Carlo (MC) dropout. MC dropout provides a practical way to estimate uncertainty by enabling neural networks to produce variable outcomes. For example, Wang et al. (2019) analysed epistemic and aleatory uncertainties in CNN-based medical image segmentation, considering both pixel-level and structural-level uncertainties.

Other non-Bayesian methods (Geifman et al., 2018) often mix aleatory uncertainty with epistemic uncertainty. Separating these two sources of uncertainty is crucial for many tasks (Abdar et al., 2021). SDE-Net (Kong et al., 2020) addresses this problem by introducing a Brownian motion term into the network architecture to capture epistemic uncertainty (i.e. model uncertainty) and view DNN transformations as state evolution in a stochastic dynamical system. However, this architecture is demonstrated on simple classification and regression tasks with tabular data and does not directly apply to segmentation tasks and rainfall prediction without modifications.

A promising new research direction lies in generative and diffusion models. Diffusion models, such as GenCast (Price et al., 2023), inherently provide probabilistic results by learning the conditional probability distribution that transitions from one weather state to the next. GenCast produces global ensemble forecasts at 0.25° resolution, achieving competitive accuracy for up to 15 days. Other studies leverage diffusion models to expand the size of physics-based ensembles (Li et al., 2024) or stochastically downscale deterministic forecasts (Mardani et al., 2023). While diffusion models produce realistic samples, they typically require solving an ordinary differential equation involving multiple neural network passes for each time step, which can be computationally intensive. Addressing this limitation, Oskarsson et al. (2024) recently introduced a generative model for probabilistic weather forecasting based on Hierarchical Graph Neural Networks, enabling the generation of arbitrarily large ensembles with a single forward pass. However, generative models generally require vast amounts of data to achieve convergence, making these solutions impractical without well-curated large datasets and significant computational resources.

One limitation of the models discussed is their lack of explainability, a common issue in modern deep learning. In other words, while MLWPs can identify meaningful patterns from past weather conditions — and many studies demonstrate their effectiveness during extreme events (Lam et al., 2023; Price et al., 2023) — this does not guarantee their ability to generalize to future events. In contrast, NWPs are generally less accurate but grounded in solid mathematical and physical principles.

Our contribution focuses on deep learning to enhance the accuracy of NWP-provided quantitative precipitation forecasts while ensuring robust and reliable uncertainty quantification. We achieve this by post-processing QPF outputs from multiple NWPs and blending them onto a shared, regular grid. This multi-model approach exploits the complementary strengths of individual NWPs (Gagne et al., 2014), leading to improved predictive performance.

We frame precipitation estimation as an image segmentation task, and by using U-Net (Ronneberger et al., 2015) as our deterministic network, we investigate the use of advanced DL to generate ensembles and quantify uncertainty. Beyond leveraging state-of-the-art techniques, we propose SDE U-Net, an innovative variant of SDE-Net introduced by Kong et al. (2020), adapted explicitly for segmentation tasks in this domain. Our analysis emphasizes the critical sharpness-reliability tradeoff, striving to balance confidence in model predictions with the potential risks of forecasts failing to capture actual physical outcomes.

As a case study, we combine daily cumulative QPFs from multiple NWPs over Piedmont and Aosta Valley, two regions in northwestern Italy. The proposed post-processing framework is evaluated on this domain to assess model accuracy and uncertainty quantification in diverse precipitation regimes.

We design the post-processing frameworks in this study for seamless integration into operational forecasting systems. By adopting these frameworks, forecasters can enhance decision-making processes and strengthen preparedness for weather-related challenges.

Our main contributions can be summarized as follows:

- introduction of SDE U-Net, an enhanced version of SDE-Net (Kong et al., 2020) for dense regression in rainfall forecasting,
- a rigorous evaluation of uncertainty-aware deep learning models under common and intense weather conditions,

- the design of a testing framework that explicitly assesses generalization to out-of-distribution events.

The paper is structured as follows: Section 2 provides background on uncertainty in machine learning and describes the dataset used in this study. Section 3 presents the uncertainty-aware DL architectures, concluding with a formal definition of the novel SDE U-Net. Section 4 details the experimental design and evaluation metrics, followed by a comprehensive analysis of the results. Finally, Section 5 summarizes the key findings and discusses implications for operational weather forecasting.

2. Background

Machine Learning (ML) involves algorithms that identify patterns in data to make predictions or decisions. A key challenge is balancing bias and variance to address underfitting and overfitting. In supervised learning, models trained on labelled data handle tasks like regression, where loss functions like Mean Squared Error (MSE) and Mean Absolute Error (MAE) measure prediction errors. ML workflows typically divide data into training, validation, and test sets, with cross-validation enhancing robustness by averaging performance across multiple data splits.

Understanding the distinctions between aleatoric and epistemic uncertainties is crucial for managing the limitations and reliability of model predictions. Aleatoric uncertainty originates from inherent noise in the data, representing natural variability in the data generation process. This uncertainty does not decrease with additional data collection. In contrast, epistemic uncertainty arises from the model's lack of knowledge about its predictions, often due to insufficient training data or limited model capacity (Kendall and Gal, 2017).

2.1. Case study

This study estimates forecast uncertainties in daily cumulative QPF for an Area of Interest that includes Piedmont and Aosta Valley regions (Fig. 1), blending weather model forecasts on a unified grid. We focus on the first 24 h of daily precipitation forecasts as input to our machine learning framework, improving reliability and uncertainty characterization. In this work, a daily cumulative precipitation gridded observation is called an "event".

The Piedmont and Aosta Valley regions in northwestern Italy are shaped by the Alps, influencing their climate and hydrology. Piedmont consists of three main zones: the Alpine region, the pre-Alpine hills, and the flat Po Valley, a key drainage basin. Aosta Valley, Italy's smallest and most mountainous region, features glacial valleys and active glaciers that play a crucial role in seasonal hydrology. The Po Valley, an industrial hub, faces risks from intense rainfall and flash floods, emphasizing the need for accurate precipitation forecasts to mitigate impacts on infrastructure, agriculture, and industry.

2.2. Dataset building and description

The dataset in this study includes precipitation observations and forecasts from multiple Numerical Weather Prediction (NWP) models. ARPA Piemonte provides the observational data for the Area of Interest (AoI) through the North Western Italy Optimal Interpolation (NWIOI) dataset (ARPA Piemonte, 2022). This dataset offers daily cumulative gridded precipitation observations, derived from optimum interpolation (Gandin, 1963) of ground station measurements since 1958, with a spatial resolution of 0.125° (approximately 12 km horizontally).

For the forecast data, four NWPs operationally used by ARPA Piemonte were selected due to their extensive historical archives:

- BOLAM: a hydrostatic Limited Area Model (LAM) developed by the Institute of Atmospheric Sciences and Climate (ISAC-CNR). It operates at a grid resolution of 7 km over the Mediterranean domain and uses ECMWF-IFS for initialization. BOLAM has been used extensively for mesoscale weather forecasting in Europe and research purposes, including studies on extreme weather events (Buzzi et al., 1994).
- COSMO-5M: a non-hydrostatic LAM developed by the Consortium for Small-Scale Modelling (COSMO). It has a 5 km grid resolution, it uses ECMWF-IFS for initialization, and it focuses on high-impact weather phenomena, utilizing advanced parameterizations for convection and microphysics to improve predictions of precipitation and severe weather (Doms and Baldauf, 2018).
- COSMO-2I: LAM, it is a high-resolution (2.2 km) version of the COSMO model tailored for Italy (Baldauf et al., 2011). Although scheduled for decommissioning as ARPA Piemonte transitions to the ICON model framework (COSMO Consortium, 2024), COSMO-2I was included in this study for its capability to capture small-scale precipitation features critical for post-processing applications.
- ECMWF-IFS: Global Circulation Model (GCM), the high-resolution configuration of the European Centre for Medium-Range Weather Forecasts Integrated Forecasting System (IFS) operates at approximately 9 km horizontal resolution and 137 vertical levels. This model is widely recognized for its accuracy and global coverage, making it a benchmark in numerical weather prediction (ECMWF, 2016).

To focus on significant precipitation events, we include only days when the 99th percentile of observed precipitation across the grid exceeds 10 mm, removing fair-weather and low-precipitation days. We filter the dataset to retain only events with available forecasts from all four NWPs, resulting in 436 significant precipitation events between 2018 and 2024.

To evaluate the generalization capability of our machine learning architectures, we classify events as either intense or non-intense. The deep learning architectures are trained on non-intense events and tested on both categories, as described in Section 4.1. This classification was based on the distribution of spatial maxima of daily precipitation by season from the NWIOI dataset from 1958 to 2017, using the 99th percentile in each season as the threshold: this led to 64.58 mm for winter (DJF), 95.71 mm for spring (MAM), 93.26 mm for summer (JJA) and 140.40 mm for autumn (SON). Applying these thresholds, we classify 40 events as intense and 396 as non-intense.

Non-intense events are clustered based on precipitation nature — convective or stratiform — to ensure uniformity across training, validation, and test sets. Convective precipitation at mid-latitudes has high spatial variability and low spatial averages, while stratiform shows the opposite behaviour. Using k-means clustering (k = 3) in the Coefficient of Variation (CV) versus spatial average rainfall plane (Fig. 2), we identified convective, stratiform, and intermediate clusters and create nine non-overlapping training–validation–test subsets, ensuring an even distribution of cluster labels. CV is defined as the ratio between spatial standard deviation and average rainfall.

This process ensures that the model evaluation phase represents intense and non-intense events and various precipitation types.

3. Methods

We can define our task with a dual interpretation. From a *deterministic* interpretation, given a true precipitation map P for a specific event and a set of n *imperfect* predictions $\{P_i\}_{i=1,\dots,n}$, which are outputs from different NWPs, our deep learning algorithm — parameterized by weights θ — must generate an output $\hat{P}$ in the following form:

$$\hat{P} = f(\{P_i\}; \theta), \tag{1}$$

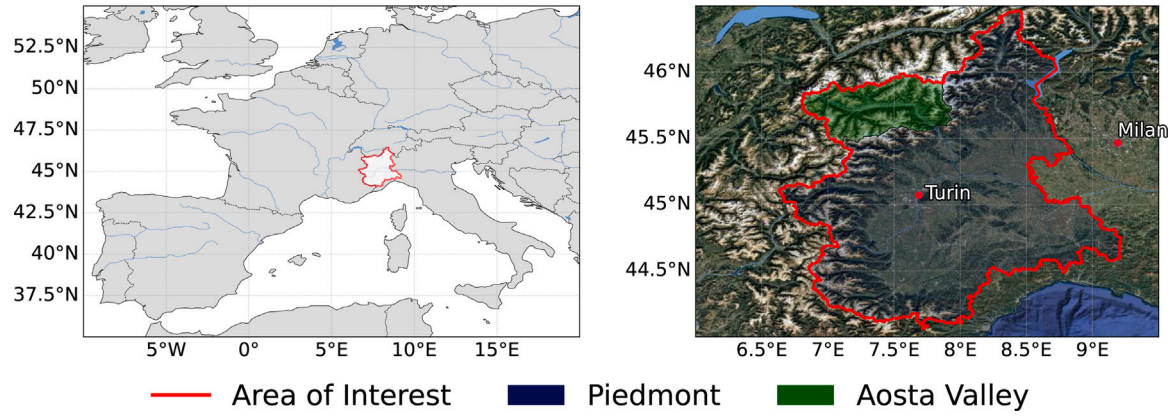

Fig. 1. The Area of Interest (AoI), consisting of Piedmont and Aosta Valley, is outlined in red using their unary union borders in the left panel, shown in the context of Western Europe. In the right panel, the focus shifts specifically to the AoI, still highlighted in red, with Piedmont and Aosta Valley individually shaded in blue and green, respectively, to emphasize their distinct regions.

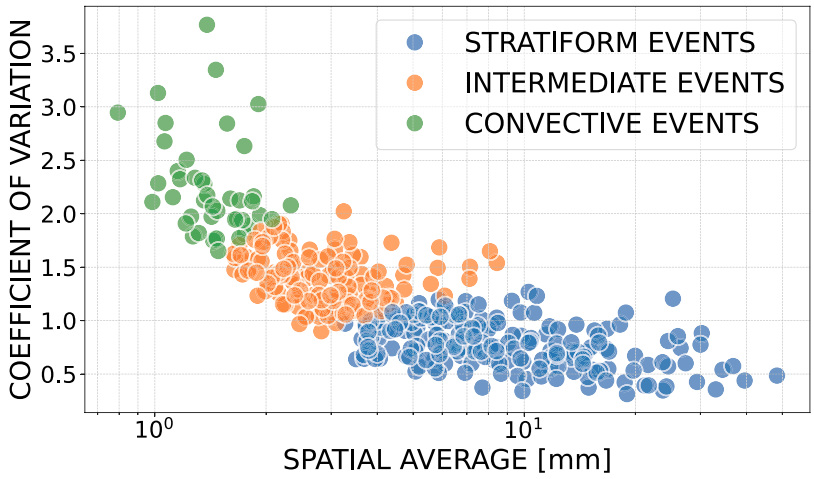

Fig. 2. The scatter plot visualizes non-intense precipitation events, clustered using the Coefficient of Variation (CV) to represent spatial variability and the spatial average rainfall to indicate mean precipitation across the study area. By applying k-means clustering, we categorize the dataset into three groups: stratiform, intermediate, and convective events. This classification relies on the observed tendency in mid-latitudes, where convective events typically exhibit greater spatial variability (higher CV), while stratiform events generally feature higher average precipitation.

such that a distance function (e.g., L_2 loss) is minimized.

Alternatively, from a *probabilistic* perspective, we interpret the NWP outcomes P_i as independent and identically distributed (i.i.d.) samples drawn from a distribution of a stochastic process, represented as:

$$P_i = P + \delta p_i \tag{2}$$

where δp_i denotes the epistemic error introduced by each numerical model. In this framework, we account for uncertainty in the model prediction $\hat{P}$ based on the characteristics of the training data. Additionally, we expect aleatoric uncertainty due to inherent measurement errors in the observational data. In this work, we will not explicitly separate these two types of uncertainty but provide overall forecast uncertainty estimates that encompass both sources of error.

Conventional deep learning models typically operate deterministically, generating only point predictions without indicating uncertainty. To overcome this limitation, we propose reformulating the problem by replacing the parametric model f with a variant capable of producing a distribution of outcomes rather than a single prediction. In this framework, the model prediction is represented as a sample from the predictive distribution:

$$\hat{P} \sim \tilde{f}(\{P_i\}; \theta), \tag{3}$$

where $\tilde{f}$ stands for the variational model. Given $\bar{Y} = \{\hat{P}_i\}_n$ and a set of n samples from the predictive distribution, we can introduce the Prediction Interval (PI) with a confidence level of $\gamma \in [0, 1)$ as the range $[l(\bar{Y}), u(\bar{Y})]$, such that the probability $\mathcal{P}\left(l(\bar{Y}) < \hat{P}_{n+1} < u(\bar{Y})\right) = \gamma$. This interval reflects the expected error between the prediction and the true target values. A wider PI indicates greater uncertainty in the model's predictions, suggesting higher variability in the input data or challenges in the prediction task.

Conversely, a narrower PI implies higher confidence in the model's predictions, with the actual precipitation likely closer to the predicted value. However, this increased confidence also raises the risk of the actual value falling outside the interval. Thus, the optimal PI range is highly context-dependent, representing a trade-off between prediction sharpness and reliability.

In precipitation forecasting, NWP simulations P_i typically yield broad PIs due to differing mathematical assumptions in the models. While this broad range is beneficial for capturing extreme meteorological events, it may introduce excessive uncertainty. An ideal model would refine these intervals, narrowing them without compromising the ability to capture significant weather phenomena.

3.1. Baseline deep learning architecture

We chose the U-Net architecture (Ronneberger et al., 2015) as our deterministic network to serve as the basis for generating forecasts probabilistically and quantifying uncertainty, effectively framing the

problem as a dense per-pixel regression, where the model predicts a continuous value at each pixel location, analogous to semantic segmentation but with continuous rather than categorical outputs. Although newer alternatives are available, U-Net remains a widely used and effective choice in domains such as medical imaging, remote sensing, and diffusion models (Siddique et al., 2021; Peebles and Xie, 2023). Its encoder–decoder structure with skip connections is particularly effective at capturing both local and global contextual information, making it well-suited to our task. Through experimentation, we found that more complex architectures did not significantly improve performance. However, it is important to note that choosing U-Net architecture is not critical to the subsequent analysis. The modifications we introduce can also be applied to other segmentation networks.

Building on these insights, we used several ensemble-generating models to manage uncertainty. These models incorporate well-established techniques from the literature to achieve optimal performance across various segmentation tasks. In the following sections, we will present these models and highlight our specific contributions to their development and refinement.

3.2. Monte Carlo dropout U-Net

This approach (referred to as MCD U-Net from now on) enhances the deterministic model by incorporating a Monte Carlo Dropout (MCD) strategy (Gal and Ghahramani, 2016). Dropout initially emerged as a regularization technique that randomly deactivates a subset of neurons during training to reduce overfitting and enhance generalization. In MCD, this concept is extended to the test phase to generate several of different outputs. By applying dropout during testing, we perform multiple forward passes, each with a different subset of neurons deactivated, thereby approximating a Bayesian approach. The resulting variance in the predictions serves as a measure of the model's uncertainty.

3.3. Deep Ensemble U-Net

The Deep Ensemble U-Net (Ens U-Net) employs an ensemble technique, independently training multiple U-Net models with different initial parameter values (Lakshminarayanan et al., 2017). Unlike MCD, which relies on stochasticity within a single model, the ensemble approach derives prediction variability from differences between independently trained models. This method creates diversity among the model outputs and improves the robustness. However, the number of independently trained U-Nets in the ensemble determines the extent of variability and uncertainty the model can capture.

3.4. SDE U-Net

Together with the baseline architectures, we also proposed an extension to the segmentation domain of the widespread SDE-Net, proposed by Kong et al. (2020) to integrate Stochastic Differential Equations (SDEs) into deep learning models for capturing uncertainty. This approach put his foundations into the broad field binding neural networks and dynamical systems, building upon neural ordinary differential equations (Chen et al., 2018), neural stochastic differential equations (Liu et al., 2019) and many other variations. The basic concept is to consider the learning task as a dynamical system by taking the limit for infinitesimal updates of residual networks (e.g. ResNet), in which the equation between one-layer output and the preceding one is

$$x_{t+1} = x_t + f(x_t, t). \tag{4}$$

This expression resembles an Euler integration step for a dynamical system. By substituting the unit update step with an infinitesimal Δt and rearranging the terms, we arrive at

$$\frac{x_{t+\Delta t} - x_t}{\Delta t} = f(x_t, t). \tag{5}$$

This formulation allows us to interpret the network as the following differential equation:

$$\frac{dx}{dt} = f(x, t), \tag{6}$$

that leverages an ODE solver to model continuous transformations of the hidden states. Neural ODEs allow for high-precision continuous-time evaluations of the hidden dynamics, providing memory and parameter efficiencies by conceptually transitioning from discrete to continuous layers.

Building on this foundation, SDE-Net captures epistemic uncertainty by incorporating a stochastic component through Brownian motion within its dynamic framework. This approach treats neural networks as continuous-time transformations, modelling epistemic uncertainty as a stochastic process governed by:

$$dx_t = f(x_t, t)dt + g(x_t, t)dW_t. \tag{7}$$

In this equation, the diffusion term $g(x_t, t)$ scales the Brownian motion dW_t, introducing randomness into the dynamics. Neural networks parameterize functions $f(\cdot;\theta_f)$ and $g(\cdot;\theta_g)$, with g specifically trained to model epistemic uncertainties. To maintain stability, we design the diffusion term g to depend only on the initial condition x_0 and time t, i.e., $g(x_0, t)$.

The network $g(x_t, t)$ should produce higher values when the model is uncertain, allowing the stochastic diffusion component to dominate, and lower values when uncertainty is limited, allowing the drift term to take precedence. Therefore, we train the model to predict an output x_{t_f} (the final state of the dynamics) from the initial input x_0 while maintaining this feature. The objective function thus combines the standard supervised loss on the true solution with a term that minimizes the diffusion strength for in-distribution data and a term that maximizes it for out-of-distribution inputs $\tilde{x}_0$, obtained by adding Gaussian noise to the actual inputs.

The complete objective function takes the form:

$$\mathbf{L} = \min_{\theta_f} \mathbb{E}\left[\mathcal{L}(x_{t_f}, y)\right] + \min_{\theta_g} \sum_t \mathbb{E}_{x_0}\left[g(x_0, t)\right] + \max_{\theta_g} \sum_t \mathbb{E}_{\tilde{x}_0}\left[g(\tilde{x}_0, t)\right], \tag{8}$$

where $\mathcal{L}$ is the task-specific loss function that encourages the terminal outcome x_{t_f} to match the target y. In our regression task, we opted for a MSE as loss function to emphasize larger errors. The other terms modulate the diffusion network g to differentiate between inputs in and out of distribution by minimizing diffusion in known data regions and amplifying it when encountering unfamiliar samples.

This objective allows the model to effectively capture epistemic uncertainty by allowing the stochastic component g to respond flexibly to different input distributions. For in-distribution data, where the model aims to minimize intrinsic uncertainty during training, the framework maintains low diffusion to stabilize predictions. In contrast, a larger diffusion term indicates greater uncertainty for out-of-distribution data, prompting the model to flag potentially unreliable predictions. This dual mechanism ensures that the SDE-Net can effectively manage predictive accuracy and uncertainty estimation under various and uncertain conditions. Therefore, it is particularly suitable for high-stakes segmentation tasks, where it is crucial to differentiate between reliable and uncertain predictions. Finally, SDE-Net provides theoretical guarantees on the existence and uniqueness of the solution x_t for $0 \le t \le t_f$ provided that f and g are both uniformly Lipschitz continuous and that the output of the diffusion term is bounded to ensure numerical stability (Kong et al., 2020).

SDE-Net original implementation defines the input–output system over the time interval $[0, t_f]$ using an Euler–Maruyama scheme, iteratively adding the two components of Eq. (7) with a fixed step size. This approach allows using the same networks at each time step, which helps reduce the overall number of weights.

Extending this strategy to the U-Net architecture presents challenges, as the primary strength of U-Net lies in its encoder–decoder structure, and each substructure in the encoder and the decoder has

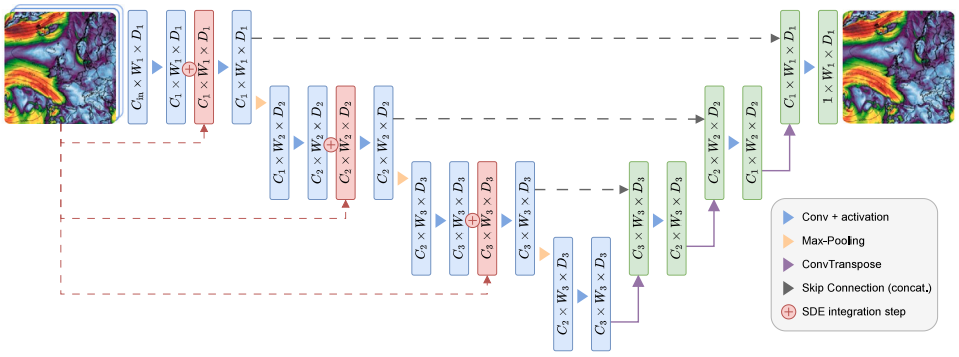

Fig. 3. Schema of the SDE U-Net architecture. The blue blocks represent the SDE's drift component within the encoder blocs, the red stays for the diffusion component, and the green ones are the decoder blocks.

different dimensions. To adapt the SDE-Net strategy, we align the number of time splits with the number of encoder blocks. For each encoder block, we introduce a diffusion block summing its signal at each skip connection and simulate an integration step at the encoder–decoder exchange. This approach allows us to incorporate the diffusion process into the U-Net framework.

Furthermore, this method allows for a more nuanced interaction between the encoded and decoded features, leveraging the strengths of both architectures. By integrating diffusion blocks, we introduce a mechanism that not only captures the hierarchical representations from the encoder but also enhances the reconstruction capabilities of the decoder through stochastic processes. The diffusion blocks function to robustify the output of the encoder before it passes to the decoder, potentially improving the quality of the reconstructed signals. Fig. 3 shows a 4-step SDE U-Net. The blue squares represent the input and output signals for the encoder blocks and their respective dimensions. These blocks employ the *drift* component of each SDE's update, while the red squares denote the output signals of the convolutional blocks reproducing the *diffusion* part. Finally, green squares indicate the input and output for the decoder blocks, resembling the original U-Net architecture. Each encoder block applies the SDE step using the following operation:

$$x_i = f_i(x_{i-1}) + g_i(x_0)\sqrt{t} \cdot \mathcal{N}(0,1) \tag{9}$$

where f_i is a convolutional block taking as input the output from the previous step and g_i is a combination of convolutional block and pooling operations taking as input x_0 and providing an output with the right dimension for the ith level. $\mathcal{N}(0,1)$ represents a Gaussian random variable with a mean of zero and a variance of one.

Finally, we trained this network with the strategy proposed by Kong et al. (2020) to assign higher uncertainty to out-of-distribution inputs, enabling effective quantification of uncertainty in segmentation tasks while preserving the U-Net structure.

4. Experiments

4.1. Experimental design and verification metrics

To assess the effectiveness of uncertainty-aware deep learning architectures in rainfall prediction and precipitation map reconstruction, we separate precipitation events into non-intense and intense categories as described in Section 2.2. Only non-intense events are used for training and validation, while both non-intense and intense events are used for testing. This separation enables a focused examination of model generalization: performance on non-intense events represents the expected baseline. In contrast, performance in intense events serves as a benchmark for robustness under challenging conditions.

This setup reflects a real-world scenario in which models are trained on abundant and moderate data, but are expected to make reliable predictions on even higher-impact events too.

Given the inherent differences between the two classes, we expect better model performances on non-intense events since they align more closely with the training data. However, the primary goal is to assess how well the model can extrapolate to unseen high-intensity precipitation events despite being trained solely on non-intense data. This design provides insights into the models' ability to capture complex patterns and uncertainties associated with extreme weather phenomena, which is critical to improving real-world rainfall prediction systems.

We compare the uncertainty estimates from the considered machine learning architectures with the forecast uncertainty provided by a Poor Man's Ensemble (PME, i.e. the average of NWP forecasts), which serves as our benchmark. We selected PME because it is a widely used operational baseline ensemble method that ensures a high proportion of target values fall within the Prediction Interval (PI). However, this comes at the cost of generating excessively wide PIs (Landberg et al., 2002). As a result, while it achieves high reliability, it lacks sharpness. The challenge is to develop an ensemble method that maintains a high proportion of correctly predicted targets while simultaneously reducing the length of PIs.

We begin the analysis by assessing statistical significance through cell-wise median p-values from the Wilcoxon test (Rosner et al., 2006) across the Area of Interest (AoI), comparing forecasts from the four input NWPs and the learning models against the PME baseline. For each split, the Wilcoxon test evaluates forecast differences at the grid-cell level using all events in that split, producing a distribution of p-values from which we extract the median. Near-zero p-values indicate a statistically significant difference from PME, while higher values suggest similar forecast distributions. To complement this analysis, we report the cell-wise median Root Mean Square Error (RMSE), which offers a basic estimate of uncertainty and indicates the direction and magnitude of distributional differences with respect to PME.

We extend the analysis beyond individual grid cells by computing summary metrics over the entire domain for each split and analysing their distributions. We include the Root Mean Square Error (RMSE) as a measure of forecast accuracy. Additionally, to study the trade-off between sharpness and reliability, we introduce a Coverage-Length-based Criterion (CLC), as defined in Khosravi et al. (2010). This metric allows us to assess how well the models balance prediction confidence and error coverage.

$$CLC = NMPIL \times \sigma(PICP, \eta, \gamma), \tag{10}$$

where the Normalized Mean Prediction Interval Length (NMPIL) represents the average length of the PI, and $\sigma(PICP, \eta, \gamma)$ is a sigmoid penalty function that models the trade-off between interval length and coverage and depends on PICP (Prediction Interval Coverage Probability), representing the proportion of target values that fall within the PI, a scaling parameter η and translation parameter γ:

$$\sigma(PICP, \eta, \gamma) = 1 + e^{-\eta(PICP-\gamma)} \tag{11}$$

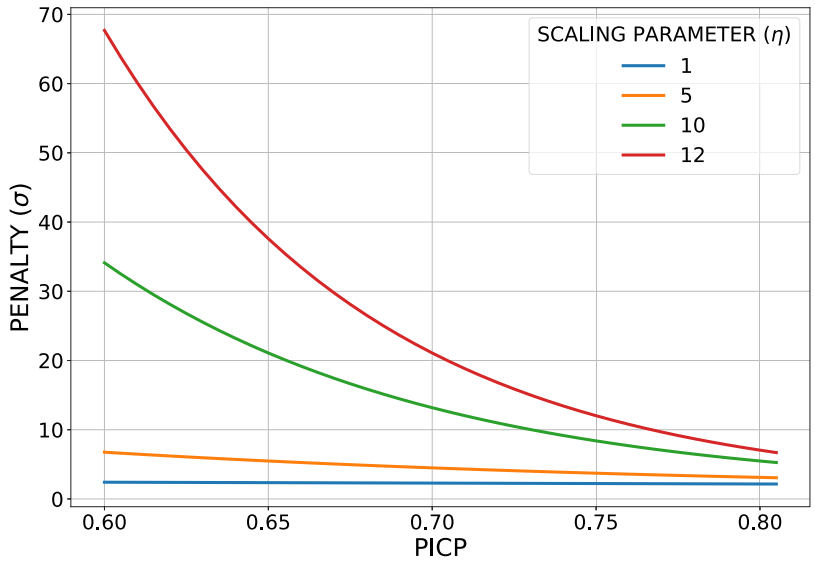

Fig. 4. Penalty function σ for different values of scaling parameter η at fixed $\gamma = 0.95$. The range of PICP is chosen looking from the expected results.

We aim to minimize the NMPIL, as smaller values indicate a narrower spread in ensemble predictions, resulting in more precise and valuable forecasts. However, reducing NMPIL can decrease the coverage PIs, causing an undesirable number of predictions to fall outside these intervals. To mitigate this, we aim for high PICP values. As a result, we strive to minimize the CLC, balancing sharpness and reliability. The parameter η regulates the penalty applied when PICP falls below the minimum acceptable threshold, γ.

The threshold value of acceptability γ should be as close as 1. In our experiments, we set $\gamma = 0.95$ to reflect a 95% PI. A significant penalty is applied when target values fall outside this 95% PI, rather than when the predictions themselves lie outside the interval. Consequently, we focus particularly on the CLC values for high values of the penalty parameter η where violations of the PI are penalized more heavily.

The choice of the considered values of η is as follows. Different works (Fan and Tang, 2013; Miyaguchi and Yamanishi, 2018) illustrates that while a higher penalty parameter can enforce stricter adherence to desired criteria, it may lead to preferentially conservative models. Fig. 4 shows the values of $\sigma(PICP, \eta, \gamma)$ for different η, with $\gamma = 0.95$ and within a PICP range of 0.6 to 0.8, ranging the whole spectra of the obtained PICP results.

Although there is not much difference between penalties at $\eta = 1$ and $\eta = 5$, the penalty essentially quadruples from $\eta = 5$ at $\eta = 10$ and doubles from $\eta = 10$ to $\eta = 12$, leading to an unnecessary tilt of the sharpness-reliability trade-off towards coverage. For this reason, $\eta = 10$ is considered an aggressive penalization value in a CLC metric. However, we have chosen $\eta = 12$ as the maximum penalization value to test our deep learning models further.

We compute all metrics using 20 sampled predictions for MCD U-Net (Section 3.2) and SDE U-Net (Section 3.4). We base the metrics on five independently trained models with different hyperparameters for Ens U-Net (Section 3.3). This approach allows us to estimate forecast uncertainty for each model, reflecting the epistemic error. We repeat the process in a 9-fold cross-validation to account for aleatoric uncertainty, ensuring statistical significance. Following the standard cross-validation scheme described in Bishop (2006), this approach involves training, validating and testing the selected learning models on nine different training–validation–test splits, ensuring that every event in the dataset at least appears in one of the nine training, validation, and testing phases. We derive the nine training–validation–test sets from weather physics considerations outlined in Section 2.2.

We normalize the precipitation grids (both observations and NWP forecasts) to the range $[0, 1]$ before using them in the learning models described in Section 3, ensuring all data remain adimensional. This normalization is then reversed back when computing the verification metrics. For a given data split, we compute the average metrics across all realizations for each learning model, accounting for epistemic uncertainty. Then, we obtain the distribution of these average metrics across the 9 splits, incorporating aleatoric uncertainty. Since PME is a deterministic model, it has only one realization, meaning we effectively consider only aleatoric uncertainty for its evaluation.

4.2. Results

Fig. 5 reports the median cell-wise p-value of the Wilcoxon test comparing each forecast with the PME baseline, across the four input NWPs (BOLAM, COSMO-2I, COSMO-5M, ECMWF-IFS) and the three learning-based models (ENS U-NET, MCD U-Net, SDE U-Net). A p-value close to zero indicates a statistically significant difference from PME, while higher values suggest similarity in the distribution.

Fig. 6 complements the p-value analysis by showing the median RMSE per grid cell. Both figures include orography and the Area of Interest (AoI) to highlight the spatial structure and orographic complexity affecting forecast performance.

Non-intense events show more regions with low p-values across both NWPs and learning-based models, suggesting more frequent departures from PME in these cases. Mountainous and pre-mountainous regions generally exhibit higher RMSE values compared to flat areas. Among NWPs, BOLAM stands out with high p-values even in the deeper parts of the flat Po Valley, indicating limited deviation from PME. This observation aligns with RMSE patterns, which remain similar to PME particularly in the mountainous and pre-mountainous regions.

Learning models, in contrast, produce lower p-values in these regions, indicating statistically significant differences from PME. They also achieve lower RMSE, especially in the northwestern portion of the AoI, where the reduction is substantial. SDE U-Net delivers the best performance in these areas. In the flat Po Valley, ENS U-NET yields consistently low p-values, while MCD U-Net appears closer to PME in terms of statistical similarity. RMSE differences remain limited in the flat region, with some degradation in the southeastern corner, caused by edge effects due to the lack of contextual data beyond the AoI boundaries.

For intense events, the four NWPs show no clear statistical divergence from PME across most of the AoI. COSMO-5M displays some localized differences in the flat region, though RMSE variations emerge more clearly in the mountainous and pre-mountainous zones. Learning

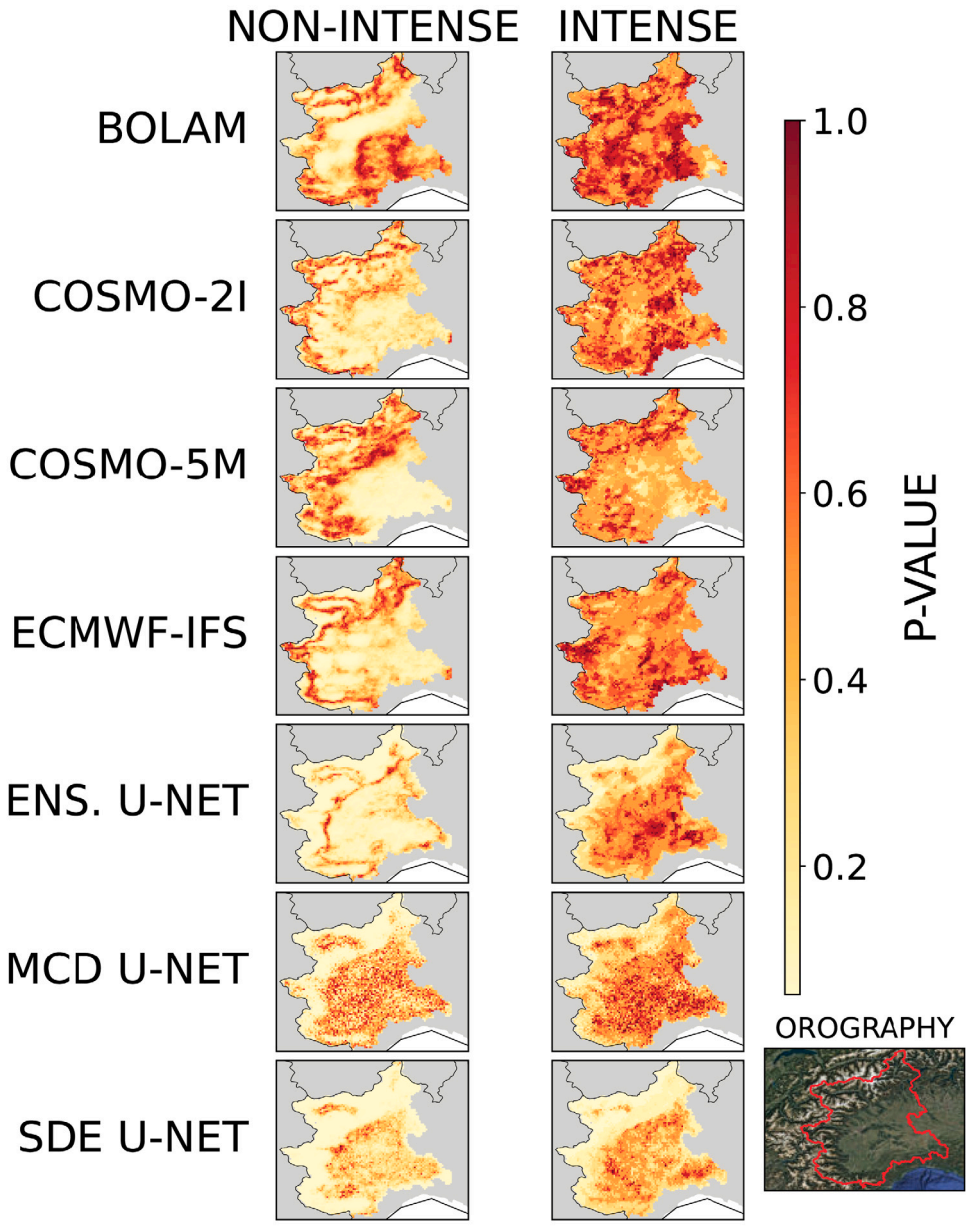

Fig. 5. Median spatial p-value from the Wilcoxon test computed across 9 data splits for each grid cell within the Area of Interest. The test compares forecast distributions from the four input NWPs (BOLAM, COSMO-2I, COSMO-5M, ECMWF-IFS), Deep Ensemble U-Net (ENS U-NET), Monte Carlo Dropout U-Net (MCD U-NET), and SDE U-NET against the Poor Man's Ensemble (PME). The Area of Interest and orography are shown to highlight terrain-related forecasting challenges.

models again produce lower p-values in these areas, suggesting consistent statistical differences from PME, while the flat region shows fewer significant changes. SDE U-Net exhibits the widest spread of low p-values in the flat area among the learning models. RMSE patterns reveal that all models struggle in the northeastern corner, though SDE U-Net achieves the lowest errors and most consistent improvements. The southeastern corner remains problematic for all learning models.

Because mountainous and pre-mountainous areas represent the most challenging regions for accurate forecasting, the improved performance in these zones — particularly from SDE U-Net — translates into a tangible statistical advantage.

Fig. 7 compares the performances of the selected learning models (Ens. U-Net, MCD U-Net, SDE U-Net) with that of the average of forecasts from four different Numerical Weather Prediction models (NWPs), referred to as PME. The comparison is based on a Coverage-Length-based Criterion (CLC) with a translation parameter $\gamma = 0.95$ and a scaling parameter $\eta = 9$, together with RMSE, Prediction Interval Coverage Probability (PICP), and Normalized Mean Prediction Interval Length (NMPIL). We provide boxplots representing the distribution forecast uncertainty taking into account both epistemic and aleatoric uncertainty, as described in Section 4. The boxplots are provided dividing them by non-intense and intense events. For each metric, an up or down arrow indicates whether the best values are the smallest or highest.

As expected, RMSE is generally higher for intense events than for non-intense ones. However, all deep learning models significantly outperform PME regarding median RMSE for both event types. Ens U-Net and SDE U-Net achieve the lowest median RMSE for non-intense events (7.79×10^{-3} and 8.15×10^{-3} on average). In comparison, SDE U-Net delivers the best RMSE for intense events (2.637×10^{-2}), demonstrating superior prediction accuracy. Additionally, SDE U-Net exhibits a narrower RMSE distribution for intense events compared to other

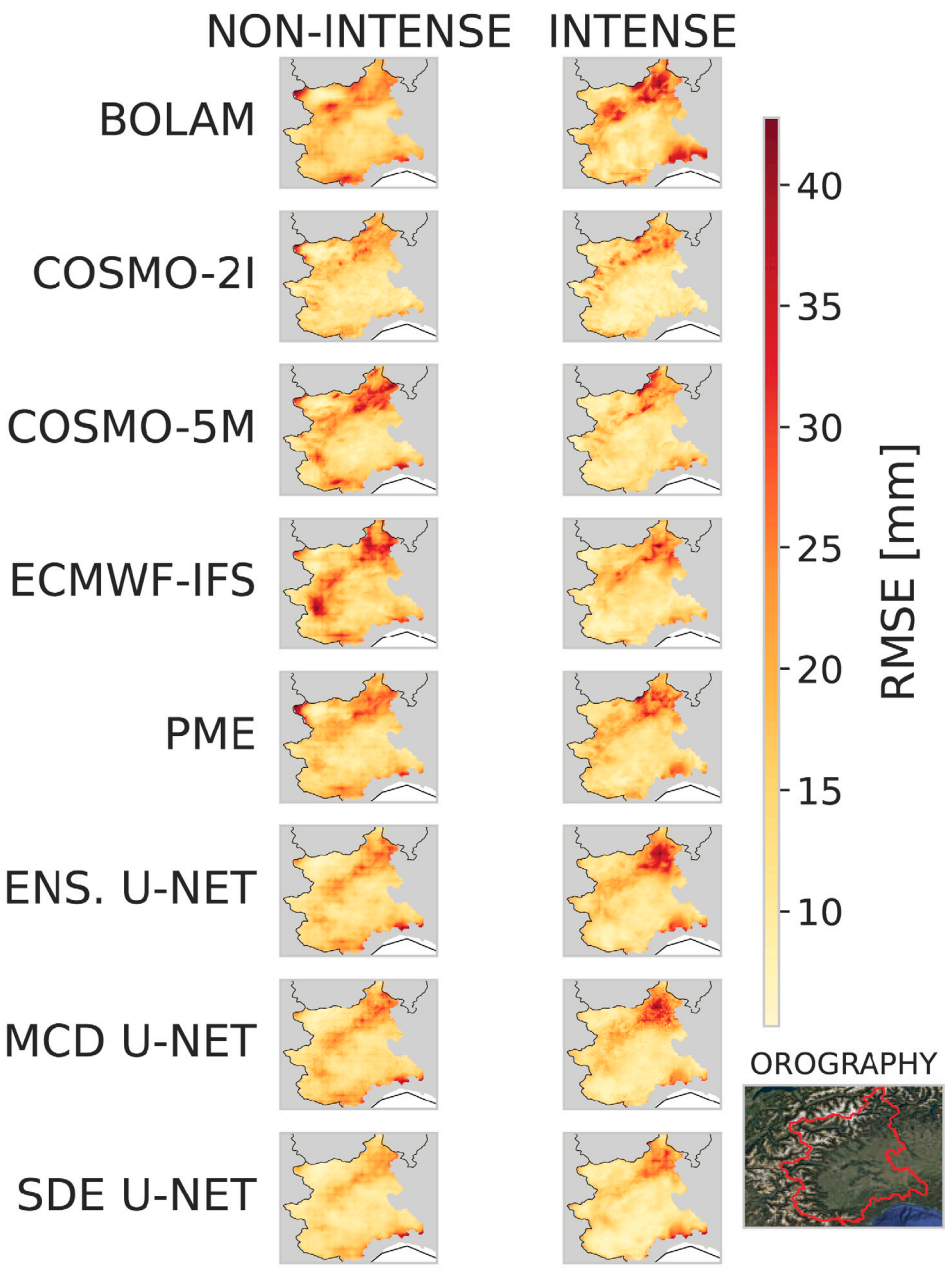

Fig. 6. Median spatial RMSE computed across 9 data splits for each grid cell within the Area of Interest. The analysis compares the forecasts from the four input NWPs (BOLAM, COSMO-2I, COSMO-5M, ECMWF-IFS), the Poor Man's Ensemble (PME), Deep Ensemble U-Net (ENS U-NET), Monte Carlo Dropout U-Net (MCD U-NET), and SDE U-NET. The Area of Interest and orography are shown to highlight terrain-related forecasting challenges.

deep learning models, indicating its effectiveness in balancing error minimization.

The PICP column indicates that the PME achieves 10%–15% better percentage coverage than the deep learning models considered in this work. Among them, MCD U-Net demonstrates the highest performance for this metric. However, this superior coverage comes at the expense of much wider prediction intervals, as shown in the NMPIL column, particularly for PME and both non-intense and intense events. As expected, PME predictions are highly reliable but lack sharpness. MCD U-Net follows a similar pattern, showing the worst NMPIL values among the deep learning models, with an even more significant disparity for non-intense events. The CLC column captures an aggregate of these trends. Here, we show the results for $\eta = 9$, which offers a reasonably high penalty and provides a more representative comparison; the discussion of results for other values will follow later. All deep learning models exhibit comparable performance for non-intense events and are consistently better than PME, with SDE U-Net showing a slightly higher median and a wider distribution range. All deep learning models outperform PME for intense events, and SDE U-Net shows the lowest CLC, indicating the best sharpness-reliability trade-off.

To delve deeper into the trade-off between sharpness and reliability, Fig. 8 illustrates the behaviour of CLC for $\gamma = 0.95$ and η values ranging from 0 to 12. We omit error bars for clarity.

We recall that smaller CLC values represent a better balance between sharpness and reliability, achieved through lower NMPIL and higher PICP values, especially at larger η. For non-intense events, CLC suggests that deep learning models offer substantial improvements in most of the analysed range, with MCD U-Net keeping stable results and outperforming the other neural models from $\eta \approx 9$. For η higher

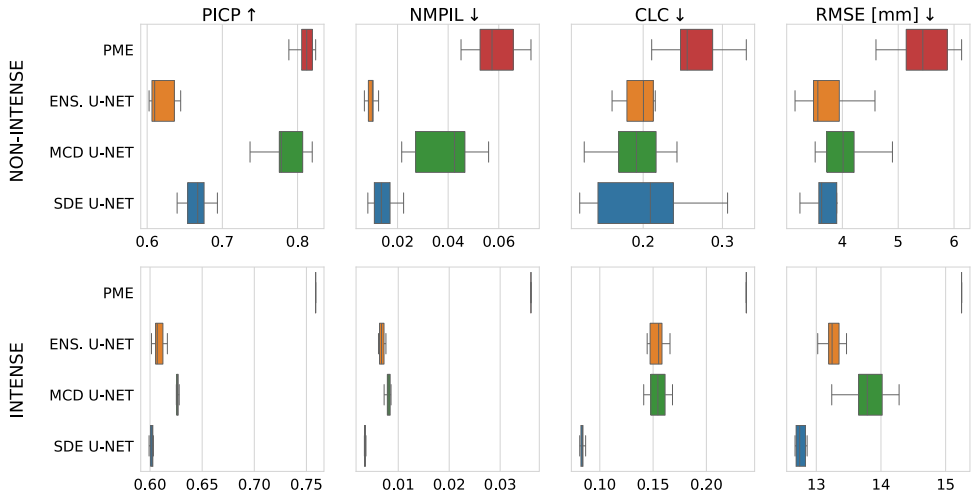

Fig. 7. Average PICP, NMPIL, CLC, and RMSE boxplots across the 9 splits, considering Deep Ensemble U-Net (ENS U-NET), Monte Carlo Dropout U-Net (MCD U-NET), and SDE U-NET against the Poor Man's Ensemble (PME). Up and down arrows indicate whether the best value is the higher or the lower, respectively.

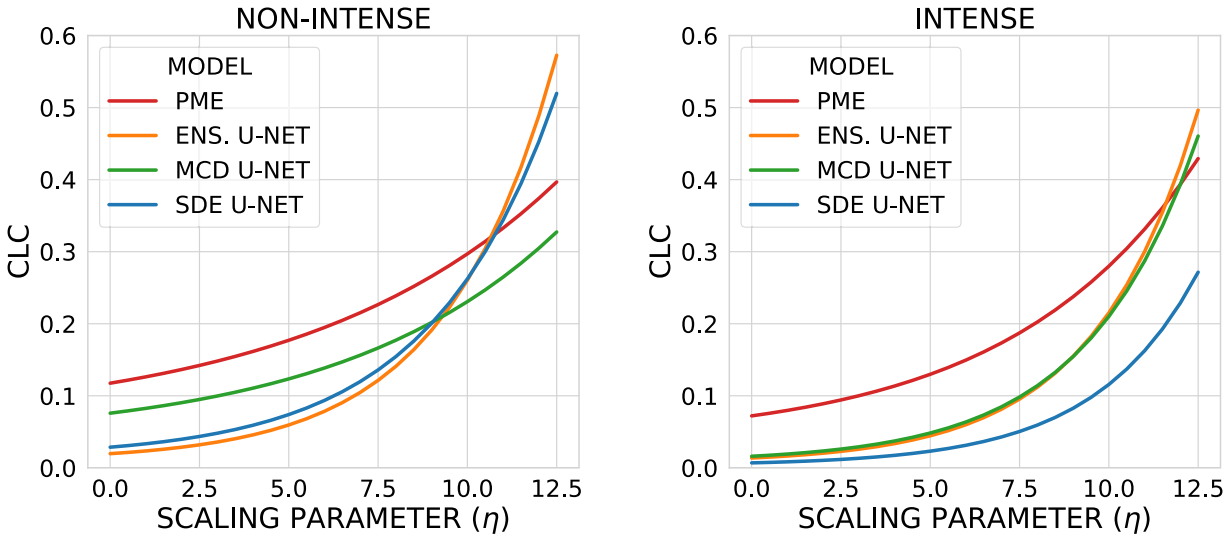

Fig. 8. CLC score of Deep Ensemble U-Net (ENS U-NET), Monte Carlo Dropout U-Net (MCD U-NET), and SDE U-NET against the Poor Man's Ensemble (PME), in function of the scaling parameter η, separated by intense and non-intense events.

than 11, both Ens. U-Net and SDE U-Net tend to give worse results than PME. Achieving the highest accuracy for non-intense events, reflected in the lowest RMSE as previously discussed, does not necessarily ensure a well-balanced trade-off between sharpness and reliability.

Operationally, MCD U-Net is the preferred choice when handling non-intense events, if the evaluation strongly penalizes forecasts falling outside a prediction interval with $\gamma = 0.95$.

SDE U-Net consistently outperforms the competitors for intense events, achieving the lowest CLC values for any η, while the differences among MCD U-Net and Ens U-Net are minimal. Finally, although SDE U-Net does not attain the best PICP for intense events in Fig. 7, its advantage lies in its narrow prediction intervals, as reflected by its NMPIL value. This narrow interval results in a highly favourable trade-off between sharpness and reliability, as evidenced by the CLC curve. This outcome, combined with the RMSE analysis, make SDE U-Net the preferred choice for intense events.

These results highlight the effectiveness of deep learning models, particularly the SDE U-Net, in providing accurate and precise rainfall predictions while ensuring reliable uncertainty quantification. These models' capability to balance predictive accuracy with uncertainty estimation makes them extremely valuable for operational forecasting systems.

5. Conclusions

Our study underscores the potential of probabilistic deep learning approaches to improve QPF's accuracy and sharpness-reliability trade-off. Through rigorous evaluation across both non-intense and intense precipitation scenarios, we demonstrated that all tested deep learning models significantly outperformed the benchmark PME solution especially on mountainous and pre-mountainous areas, which usually are the most challenging. For non-intense events MCD U-Net is the preferred operational choice while our customized SDE U-Net achieved the most relevant statistical advantage. Moreover, SDE U-Net model achieved the lowest RMSE and the most effective balance between sharpness and reliability for intense events, which are the most important operationally for emergency responses, establishing it as a leading choice for uncertainty-aware precipitation forecasting.

Integrating these models into operational forecasting systems can transform decision-making processes and improve preparedness for weather-related events like floods or agricultural disruptions. These models offer more robust and reliable predictions by explicitly accounting for uncertainty, enabling better resource allocation and risk

management. Furthermore, the associated computational cost is modest at inference time, making these models practical for real-time integration into operational forecasting workflows.

Future work will focus on enhancing the scalability and robustness of these models by refining their architectures and incorporating additional data sources, such as real-time observations or multi-resolution datasets. Further exploration of alternative approaches, including hybrid models that combine physical and data-driven methods, could provide more comprehensive solutions. Furthermore, expanding the application of these techniques to other regions and climatic conditions will help assess their generalizability and foster broader adoption in operational weather forecasting.

CRediT authorship contribution statement

Simone Monaco: Writing – original draft, Software, Methodology, Formal analysis, Validation, Resources, Investigation. **Luca Monaco:** Validation, Methodology, Data curation, Writing – original draft, Resources, Formal analysis, Conceptualization. **Daniele Apiletti:** Writing – review & editing, Funding acquisition, Supervision. **Roberto Cremonini:** Writing – review & editing, Supervision. **Secondo Barbero:** Funding acquisition.

Code availability section

The source codes with the model architecture and dataset to reproduce the results are available for download at the link: https://github.com/simone7monaco/probabilistic-rainprediction.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is part of the project NODES, funded by the Italian MUR (Ministry of University and Research) under M4C2 1.5 of the PNRR (National Plan for Recovery and Resilience) with grant agreement no. ECS00000036. This work is also part of Luca Monaco's PhD program, funded by the Italian Civil Protection Department through ARPA Piemonte. Luca Monaco extends his sincere gratitude to Renata Pelosini for initiating the project. The SmartData@PoliTO research centre of Politecnico di Torino, Italy has also partially funded this work. The authors acknowledge ARPA Piemonte for providing both historical observational data and numerical weather forecasts over the Piedmont and Aosta Valley regions.

Data availability

Data will be made available on request.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf. Fusion 76, 243–297.

ARPA Piemonte, 2022. Daily gridded climatic data: Optimal interpolation technique. MDPI Clim. 11, 120–140, URL: https://www.mdpi.com/10.3390/climate, Describes the NWIOI dataset used for daily precipitation and temperature data, covering Northwestern Italy from 1958 to the present.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., Reinhardt, T., 2011. Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. Mon. Weather Rev. 139 (12), 3887–3905.

Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical weather prediction. Nat. 525 (7567), 47–55.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q., 2023. Accurate medium-range global weather forecasting with 3D neural networks. Nat. 619 (7970), 533–538.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York, NY, USA.

Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., Anandkumar, A., 2023. Spherical fourier neural operators: Learning stable dynamics on the sphere. In: International Conference on Machine Learning. PMLR, pp. 2806–2823.

Buzzi, A., Fantini, M., Malguzzi, P., Nerozzi, F., 1994. Validation of a limited area model in cases of Mediterranean cyclogenesis: surface fields and precipitation scores. Meteorol. Atmos. Phys. 53 (3), 137–153.

Chen, R.T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.K., 2018. Neural ordinary differential equations. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 31, Curran Associates, Inc., p. 22, URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Colomba, L., Farasin, A., Monaco, S., Greco, S., Garza, P., Apiletti, D., Baralis, E., Cerquitelli, T., 2022. A dataset for burned area delineation and severity estimation from satellite imagery. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 3893–3897.

COSMO Consortium, 2024. COSMO-2I model decommissioning and transition to ICON framework. URL: https://www.cosmo-model.org/content/tasks/workgroups/wg6/icon_transition.htm, High-resolution model (2.2 km grid) for Italy, scheduled for replacement by ICON.

Demargne, J., Wu, L., Regonda, S.K., Brown, J.D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H.D., Fresch, M., et al., 2014. The science of NOAA's operational hydrologic ensemble forecast service. Bull. Am. Meteorol. Soc. 95 (1), 79–98.

Denker, J., LeCun, Y., 1990. Transforming neural-net output levels to probability distributions. In: Lippmann, R.P., Moody, J., Touretzky, D. (Eds.), Advances in Neural Information Processing Systems. Vol. 3, Morgan-Kaufmann, p. 7, URL: https://proceedings.neurips.cc/paper_files/paper/1990/file/7eacb532570ff6858afd2723755ff790-Paper.pdf.

Doms, G., Baldauf, M., 2018. A Description of the Nonhydrostatic Regional COSMO Model. Part I: Dynamics and Numerics. Technical Report, COSMO Technical Report, Deutscher Wetterdienst.

ECMWF, 2016. IFS Documentation CY43R1. Technical Report, ECMWF Technical Documentation.

Fan, Y., Tang, C., 2013. Tuning parameter selection in high dimensional penalized likelihood. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (3), 531–552.

Gagne, D.J., McGovern, A., Xue, M., 2014. Machine learning enhancement of storm-scale ensemble precipitation forecasts. Weather. Forecast. 29 (4), 1024–1043.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059.

Gandin, L.S., 1963. Objective Analysis of Meteorological Fields. Gidrometeorizdat, Leningrad.

Geifman, Y., Uziel, G., El-Yaniv, R., 2018. Bias-reduced uncertainty estimation for deep neural classifiers. In: Dy, J., et al. (Eds.), International Conference on Learning Representations. p. 14.

Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (2), 243–268.

Hernández, E., Sanchez-Anguix, V., Julian, V., Palanca, J., Duque, N., 2016. Rainfall prediction: A deep learning approach. In: Hybrid Artificial Intelligent Systems: 11th International Conference, HAIS 2016, Seville, Spain, April 18-20, 2016, Proceedings 11. Springer, pp. 151–162.

Huang, X., Luo, C., Ye, Y., Li, X., Zhang, B., 2022. Location-refining neural network: A new deep learning-based framework for heavy rainfall forecast. Comput. Geosci. 166, 105152.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS '17, Curran Associates Inc., Red Hook, NY, USA, pp. 5580–5590.

Khosravi, A., Nahavandi, S., Creighton, D., 2010. Construction of optimal prediction intervals for load forecasting problems. IEEE Trans. Power Syst. 25 (3), 1496–1503.

Ko, J., Lee, K., Hwang, H., Oh, S.-G., Son, S.-W., Shin, K., 2022. Effective training strategies for deep-learning-based precipitation nowcasting and estimation. Comput. Geosci. 161, 105072.

Kong, L., Sun, J., Zhang, C., 2020. SDE-Net: Equipping deep neural networks with uncertainty estimates. In: 37th International Conference on Machine Learning. ICML 2020, International Machine Learning Society (IMLS), pp. 5361–5371.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 30, Curran Associates, Inc., p. 12, URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al., 2023. Learning skillful medium-range global weather forecasting. Sci. 382 (6677), 1416–1421.

Landberg, L., Giebel, G., Myllerup, L., Badger, J., Madsen, H., Nielsen, T.S., 2002. Poor-man's ensemble forecasting for error estimation.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M.C., Lessig, C., Maier-Gerber, M., Magnusson, L., et al., 2024. AIFS-ECMWF's data-driven forecasting system. arXiv preprint arXiv:2406.01465.

Li, L., Carver, R., Lopez-Gomez, I., Sha, F., Anderson, J., 2024. Generative emulation of weather forecast ensembles with diffusion models. Sci. Adv. 10 (13), eadk4489.

Liu, X., Xiao, T., Si, S., Cao, Q., Kumar, S., Hsieh, C.-J., 2019. Neural sde: Stabilizing neural ode networks with stochastic noise. arXiv preprint arXiv:1906.02355.

MacKay, D.J., 1992. A practical Bayesian framework for backpropagation networks. Neural Comput. 4 (3), 448–472.

Mamalakis, A., Barnes, E.A., Ebert-Uphoff, I., 2022a. Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. Artif. Intell. Earth Syst. 1 (4), e220012.

Mamalakis, A., Ebert-Uphoff, I., Barnes, E.A., 2022b. Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. Environ. Data Sci. 1, e8.

Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Kashinath, K., Kautz, J., Pritchard, M., 2023. Generative residual diffusion modeling for km-scale atmospheric downscaling. arXiv preprint arXiv:2309.15214.

Miyaguchi, R., Yamanishi, K., 2018. A study of penalty function methods in multilevel optimization. Mach. Learn. 107 (2), 217–238.

Molini, L., Lanza, L.G., La Barbera, P., 2009. Improving the detection of heavy precipitation events by merging radar and rain gauge data: A neural network approach. Nat. Hazards Earth Syst. Sci. 9 (5), 1775–1786. http://dx.doi.org/10.5194/nhess-9-1775-2009, URL: https://nhess.copernicus.org/articles/9/1775/2009/.

Monaco, S., Greco, S., Farasin, A., Colomba, L., Apiletti, D., Garza, P., Cerquitelli, T., Baralis, E., 2021. Attention to fires: Multi-channel deep learning models for wildfire severity prediction. Appl. Sci. 11 (22), 11060.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J.K., Grover, A., 2023a. Climax: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343.

Nguyen, T., Shah, R., Bansal, H., Arcomano, T., Maulik, R., Kotamarthi, V., Foster, I., Madireddy, S., Grover, A., 2023b. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. arXiv preprint arXiv:2312.03876.

Oskarsson, J., Landelius, T., Deisenroth, M.P., Lindsten, F., 2024. Probabilistic weather forecasting with hierarchical graph neural networks. arXiv preprint arXiv:2406.04759.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al., 2022. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint arXiv:2202.11214.

Peebles, W., Xie, S., 2023. Scalable diffusion models with transformers. In: 2023 IEEE/CVF International Conference on Computer Vision. ICCV, IEEE Computer Society, Los Alamitos, CA, USA, pp. 4172–4182. http://dx.doi.org/10.1109/ICCV51070.2023.00387, URL: https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00387.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., et al., 2023. Gencast: Diffusion-based ensemble forecasting for medium-range weather. arXiv preprint arXiv:2312.15796.

Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S., Thuerey, N., 2020. WeatherBench: A benchmark dataset for data-driven weather forecasting. Geosci. Model. Dev. 13 (3), 1199–1210.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., et al., 2024. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. J. Adv. Model. Earth Syst. 16 (6), e2023MS004019.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.

Rosner, B., Glynn, R.J., Lee, M.-L.T., 2006. The Wilcoxon signed rank test for paired comparisons of clustered data. Biom. 62 (1), 185–192.

Scheuerer, M., Hamill, T.M., 2015. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. Mon. Weather Rev. 143 (4), 1321–1334.

Schultz, M.G., He, H., Kleinert, F., 2021. Can deep learning beat numerical weather prediction? Phil. Trans. R. Soc. A 379 (2194), 20200097.

Shao, P., Feng, J., Zhang, P., Lu, J., 2024. Interpretable spatial-temporal attention convolutional network for rainfall forecasting. Comput. Geosci. 185, 105535. http://dx.doi.org/10.1016/j.cageo.2024.105535, URL: https://www.sciencedirect.com/science/article/pii/S0098300424000189.

Siddique, N., Paheding, S., Elkin, C.P., Devabhaktuni, V., 2021. U-Net and its variants for medical image segmentation: A review of theory and applications. IEEE Access 9, 82031–82057. http://dx.doi.org/10.1109/ACCESS.2021.3086020.

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., Ganguly, A.R., 2018. Deepsd: Generating high resolution climate change projections through single image super-resolution. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, pp. 1663–1672.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34–45.