Ecological Informatics 90 (2025) 103255

Contents lists available at ScienceDirect

Ecological Informatics

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

Postprocessing of convection permitting precipitation forecast using UNets

Marcos Esquivel-González *, Albano González, Juan Carlos Pérez, Juan Pedro Díaz

Grupo de Observación de la Tierra y la Atmósfera (GOTA), Avenida Astrofísico Francisco Sánchez, s/n, La Laguna, 38200, Canary Islands, Spain

ARTICLE INFO

Keywords:
Convolution Neural Network (CNN)
Probabilistic Quantitative Precipitation Forecasting (PQPF)
Statistical postprocessing
Short-range prediction
UNet
WRF

ABSTRACT

Reliable precipitation forecasting is crucial in sectors like public safety, agriculture and water management. Although Numerical Weather Prediction (NWP) models form the backbone of modern forecasting, their inherent limitations and the chaotic behavior of atmospheric equations often result in errors, requiring postprocessing to improve accuracy and quantify uncertainties. This study assesses hourly precipitation probabilistic postprocessing models tailored for the Canary Islands, aiming to improve ensemble forecasting accuracy. UNet-based models were explored using two strategies: one incorporating all 25 km-scale convection-permitting ensemble forecast simulations as input, and another applying dimensionality reduction techniques to reduce input complexity. These models were compared to traditional benchmarks like the Censored Shifted Gamma Distribution (CSGD) with Ensemble Model Output Statistics (EMOS), the Analog Ensemble method and deep learning strategies through MLP architectures. In the analysis of the results, not only the reliability of the predictions for the set of available meteorological stations was considered, but also the generalization capacity of the UNet models to obtain precipitation predictions for the whole region.

UNet models generally improved the raw WRF ensemble and performed comparably or better than traditional approaches. Probabilistic (CRPSS, BSS) and deterministic (MAESS, RMSESS, Relative Bias) skill scores demonstrated improvements, particularly in forecasting light-to-moderate rainfall. The Integrated Gradients technique revealed that incorporating a height terrain map significantly influenced UNet's outputs, emphasizing the critical role of topographic data in rainfall forecasting. Furthermore, UNet models demonstrated strong spatial generalization, showcasing their potential for operational forecasting in areas with limited observational networks.

1. Introduction

Precipitation forecasting is a crucial aspect of meteorological prediction, with significant implications in different fields. Accurate and timely precipitation forecasts can support various sectors, including agriculture, flood control, and public safety. In the particular context of the Canary Islands, located in the North Atlantic, with a fragmented territory of high orographic complexity and subject to particular atmospheric circulation conditions, obtaining detailed and reliable precipitation forecasts is of particular interest (Suárez-Molina et al., 2021).

Currently, the most advanced method for forecasting precipitation events is Numerical Weather Prediction (NWP), based on state-of-the-art dynamic atmospheric models. Very important progress has been made in recent years with the increase in computational capacity, that allows the spatial resolution of these models to be increased, using fine enough grids to explicitly resolve deep convection (convection-permitting). However, the accuracy of these model results is limited by errors in initial conditions, numerical approximations, and incomplete representation of the underlying physical processes, in addition to the chaotic nature of the atmosphere (Arnold et al., 2012; Wang and Wu, 2022; Santanello et al., 2018). These errors in NWP models can propagate into subsequent hydrological simulations, affecting the quality and uncertainty of the final products (Lorenz, 1963). Due to this chaotic nature of the system and the impossibility of producing exact solutions, it is necessary to have a system that allows obtaining quantitative information on the uncertainties of the predictions, which cannot be easily determined from the results of a single deterministic simulation. To address this problem, in recent years, great effort is being made in the generation of probabilistic predictions from the results of a relatively large set of simulations (ensemble). As Palmer (2002) points out, ensemble forecasting provides a framework to represent and quantify forecast uncertainty, offering probabilistic information that is essential for decision-making in the face of inevitable model errors and chaotic variability. Probabilistic Quantitative Precipitation Forecasting (PQPF) has become a key approach within this framework, enabling

* Corresponding author.
E-mail address: marcos.esquivel.13@ull.edu.es (M. Esquivel-González).

https://doi.org/10.1016/j.ecoinf.2025.103255
Received 16 December 2024; Received in revised form 23 May 2025; Accepted 30 May 2025
Available online 19 June 2025
1574-9541/

the estimation of precipitation probabilities across different possible scenarios (Hamill and Scheuerer, 2018; Haiden et al., 2023).

However, these ensemble models usually have to be calibrated in order to improve the results of a single deterministic simulation. For this purpose, various postprocessing techniques have been developed. Among them, traditional methods such as Bayesian Model Averaging (BMA), Ensemble Model Output Statistics (EMOS) and Analog Ensemble (AnEn) have been widely used. EMOS (Gneiting et al., 2005) is a parametric postprocessing method that corrects biases and sharpness by fitting the parameters of the predictive distribution based on ensemble statistics, thereby improving forecast reliability and accuracy. Due to the pronounced skewness of rainfall distributions, employing the Censored Shifted Gamma Distribution (CSGD) has been an effective EMOS method for regression equation fitting. As alternatives to CSGD, recent studies have proposed the use of hurdle-type distributions, such as the hurdle gamma or hurdle Burr XII (EBXII) distributions, which explicitly model the probability of no precipitation and are particularly suited for handling zero-inflated and heavy-tailed rainfall data (Jin et al., 2024; Li et al., 2021). These models have shown strong performance in extreme event forecasting and provide additional flexibility, although they typically involve more complex estimation procedures. BMA (Raftery et al., 2005), on the other hand, combines the predictions of multiple models, assigning a weight to each simulation based on their historical performance, to produce a more robust joint predictive distribution. Regarding the AnEn method (Monache et al., 2013), it relies on historical observations that had historical predictions similar to the current one to create the probability distribution. It could be considered a lazy algorithm (Aha, 1997) as it does not build a general model but memorizes historical data.

Traditional methods have proven effective in reducing errors, but have limitations in terms of flexibility for including additional predictors and handling complex non-linear relationships compared to novel methods (Vannitsem et al., 2021). EMOS typically assumes linear relationships between ensemble summary statistics (e.g., mean, variance) and the parameters of a prespecified predictive distribution, which restricts its ability to incorporate additional predictors or complex interactions without increasing the risk of overfitting (Rasp and Lerch, 2018). BMA, while providing a full joint predictive distribution, also relies on linear projections (Ji et al., 2022). Similarly, the Analog Ensemble (AnEn) method is limited by the availability and representativeness of historical analogs, and it cannot explicitly model nonlinear interactions, making it less suitable for capturing rare events or complex spatiotemporal dependencies.

With the advent of machine learning (ML), new opportunities have emerged to overcome the limitations of these traditional methods. In this context, ML-based postprocessing techniques have been developed to correct some of the errors that affect the estimates of dynamic models (Zhong et al., 2023; Höhlein et al., 2024). Although NWP models excel at predicting and explaining large-scale phenomena, data-driven models efficiently estimate event characteristics using observed data and account for biases or anomalies within those data. By combining these approaches, hybrid forecasts offer enhanced predictive capabilities (Hu et al., 2023).

One of the challenges is to develop statistical postprocessing methods that enable forecasts at any location of the domain of interest, not just specific grid points or station sites. Possible approaches include incorporating location and surface data into regression models, removing spatial heterogeneity before modeling, and using spatial interpolation on well-calibrated station forecasts (Vannitsem et al., 2021; Yang et al., 2021) . Neural networks, and in particular Deep Learning (DL) such as Convolutional Neural Networks (CNNs), exhibit considerable promise the extraction of spatial features from high-resolution forecasts, as they can intrinsically embed spatial heterogeneity information within their architectural design. These networks have been successfully applied to NWP postprocessing (Scheuerer et al., 2020; Veldkamp et al., 2021) and to statistical downscaling of several variables (Rampal et al., 2022; Dupuy et al., 2023) in complex terrain, revealing their ability to discern fine spatial details. Recent studies have also explored the use of CNNs and other deep learning methods, such as Generative Adversarial Networks (GANs), for the postprocessing of ensemble precipitation forecasts (Jin et al., 2023a,b). These approaches demonstrate the potential of neural models to enhance spatial and temporal resolution and correct biases inherent in ensemble systems. In this context, UNet, a specific CNN architecture, has been shown to be very effective in solving pixel classification problems but it has been used to a lesser extent for pixel regression. However, some works, for example for climate regionalization, use UNets for this purpose (Doury et al., 2023, 2024). Compared to models based on basic convolutional networks, UNets can keep a better balance between location accuracy and feature values, presenting better spatial details in the results (Yao et al., 2018). Inspired by these previous works, the use of an UNet architecture is proposed in this study to fine-tune PQPF in the Canary Islands region.

In contrast to approaches that generate synthetic ensembles from deterministic forecasts (Shrestha et al., 2015), this work focuses on improving the calibration and spatial consistency of existing physically based ensemble forecasts, which already capture meaningful atmospheric variability. In this study, an ensemble of 25 different WRF (Weather Research and Forecasting) runs is used as input to the UNet network, with the aim of improving hourly precipitation predictions and providing probabilistic forecasting. Since the input to the network is the prediction data of each of the models, in a rectangular grid, and the output has the same spatial dimensions although fewer output channels, the UNet architecture is expected to be a suitable option because it considers, in addition to the precipitation values at each point of the grid of each simulation, the spatial relationships between them. Specifically, the network is trained to predict the three parameters of a Censored and Shifted Gamma Distribution (CSGD) (Baran and Nemoda, 2016) for each grid point, thus capturing the inherent uncertainty in the forecasts. The results obtained are compared with those obtained directly from the set of simulations through the parametric CSGD model (Scheuerer and Hamill, 2015) and non-parametric Analog Ensemble (AnEn) model, which are the benchmark postprocessing methods. The CSGD and AnEn are widely used due to their effectiveness and have shown good performance in related works (Hu et al., 2023; Baran and Nemoda, 2016) compared to BMA methods or other type of distributions. In addition, given the computational cost involved in performing 25 simulations for rainfall prediction, the possibility of choosing a subset of such simulations that provide results with performance equal or higher than the full set is studied.

The evaluation of our approach is performed using data from the meteorological stations available in the central islands of the archipelago, i.e. La Gomera, Tenerife and Gran Canaria for which high resolution (1 km) WRF simulations were carried out during the period analyzed (15 months), comparing the results with the benchmark methods to highlight the advantages and limitations of using UNet in precipitation prediction. Also, the UNet capability to provide a reliable PQPF throughout the entire territory is tested, instead of computing results only at weather station locations. Although the experiments focus on the Canary Islands, the proposed methodology is general and can be adapted to other regions through retraining with local data.

The remainder of the paper is organized as follows. Section 2 describes the data used and introduces both the proposed UNet architecture and the traditional statistical postprocessing methods; Section 3 presents the results, and finally, Section 4 summarizes the findings and provides additional discussion.

2. Materials and methods

2.1. Data

The data used in this work correspond to the results of an ensemble of WRF forecast simulations and observational data provided by meteorological stations distributed in the study region. The period used covers from July 2019 to September 2020.

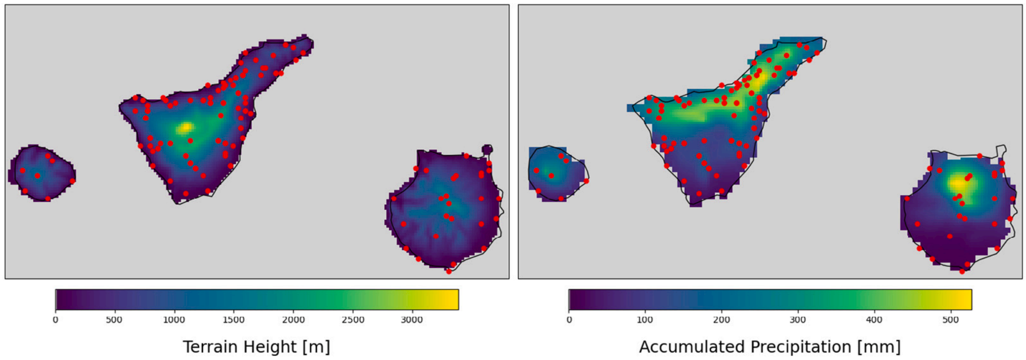


Fig. 1. Map of terrain height (HGT) in the innermost domain used in WRF simulations (left), and accumulated observed precipitation during the study period (right). In addition, the location of the weather stations is marked with red dots.

2.1.1. Observational data

The observational data used in this study consist of hourly precipitation values recorded by a dense network of 130 meteorological stations distributed across three islands: La Gomera, Tenerife and Gran Canaria (Fig. 1). These stations are located at different altitudes, from almost sea level to more than 2600 meters above sea level, and cover regions with highly diverse precipitation regimes, from areas recording only 50 mm during the study period to those exceeding 500 mm. This spatial distribution of the stations across diverse geographical and climatological areas provides robust data coverage, facilitating the analysis of precipitation patterns with high temporal and spatial resolution. The stations are managed by two primary institutions: the Spanish State Meteorological Agency (AEMET) and the Cabildo de Tenerife. Despite the moderate geographical extension of the Canary Islands, the region exhibits significant spatial inhomogeneity in precipitation, as clearly illustrated in Fig. 1 by the accumulated precipitation from the ROCIO+CAN,2.5 km gridded product supplied by AEMET. This inhomogeneity is primarily due to steep orographic gradients and complex meteorological influences. Given this pronounced spatial variability and the availability of such a dense network of weather stations, the application of spatial deep learning techniques, such as UNet, becomes particularly valuable. This station density not only enables the model to learn spatial dependencies but also provides suitable observational data for evaluating high-resolution simulations.

2.1.2. WRF ensemble forecast

To generate the ensemble of numerical precipitation predictions, the Weather Research and Forecasting (WRF) model has been used (Skamarock et al., 2008). WRF is a widely used non-hydrostatic numerical weather prediction system designed for both atmospheric research and operational forecasting. Developed by the National Center for Atmospheric Research (NCAR) and a collaborative partnership of several other research institutions, WRF offers high-resolution forecasts with flexible configurations. To generate the ensemble of numerical precipitation predictions, the WRF model has been used with different configurations and updated with different input and boundary conditions. In the present study, WRF version 3.9 was used with a configuration of 4 nested domains centered on the Canary Islands. The innermost domain (Fig. 1), from which the values presented in this paper are obtained, covers the central islands of La Gomera, Tenerife and Gran Canaria and is composed of 207 x 150 grid cells with a spatial resolution of 1 km and 34 vertical levels. The setup of the experiment consisted of performing 45 daily weather forecast simulations during the study period (July 2019– September 2020). The simulations for each day start at 00 UTC, which also corresponds to local time during the winter, and run for 30 h. Due to the high computational requirements of the simulations, they were executed in 3 batches of 15 runs on the Teide High-Performance Computing (TeideHPC) system, a state-of-the-art supercomputing facility managed by the Instituto Tecnológico y de Energías Renovables (ITER) in Tenerife, Spain ITER (2024), using a total of 64 computational cores for each run. The main two- and three-dimensional meteorological variables were stored every hour, although the first 6 h were not considered for subsequent analyses, as they are considered as a spin-up period.

In addition, WRF runs require the specification of some physical parameterizations to simulate various processes that are not resolved by its grid-scale equations. The available options for each type of parameterization are described in Skamarock et al. (2008), which also includes the corresponding references. In this study, the Noah LSM (Land-Surface Model) scheme has been used to provide heat and moisture fluxes over land points in all simulations. The Kain–Fritsch cumulus scheme, to account for subgrid-scale convection effects, was applied to the outer domains, being disabled in the domain of interest, since at 1 km these effects can be explicitly resolved, resulting in a convection-permitting simulation. Other parameterizations, as well as the forecasts of various global models used as boundary conditions for WRF, are different for the various simulations. The list of used choices, as well as the acronym used for each of them, is summarized as follows:

- Global forecast model (GFM): ARPEGE (Roehrig et al., 2020), GEM (Côté et al., 1998) and GFS (NCEP, 2015). The GFMs provide the initial and boundary conditions for the simulations, both for the atmospheric variables and the sea surface temperature (SST), which is essential for the study region, since the islands cover only a portion of the domain.
- Planetary boundary layer (PBL) estimates vertical sub-grid-scale fluxes due to eddy transports. PBL parameterizations are broadly classified into local and non-local schemes, based on how they handle turbulence and vertical mixing:
  - Local: Mellor-Yamada-Janjic (MYJ), Mellor–Yamada Nakanishi Niino (MYNN) Level 2.5, and University of Washington (UW).
  - Non-local: Yonsei University Scheme (YSU).
- Microphysics schemes represent water vapor, cloud and precipitation processes, including the formation, growth, and dissipation of different cloud hydrometeors. Used schemes: Thompson (T) and Morrison (M).
- Shortwave and longwave radiation parameterizations account for the absorption, reflection, and emission of solar and terrestrial radiation. Used options: Rapid Radiative Transfer Model for General Circulation Models Applications (RRTMG, labeled as R in this study), and NCAR Community Atmospheric Model (CAM, labeled as C).

When referring to a specific WRF configuration, the following nomenclature is used:

{GFM}_{PBL}{Microphysics}{Radiation}

Table 1
RMSE (mm) value for each configuration. Values are ordered from lowest to highest.

	Configuration	RMSE		Configuration	RMSE
0	ensemble	0.274	13	GFS_myjtr	0.312
1	GFS_uwtc	0.290	14	GFS_myjmc	0.317
2	GEM_uwtc	0.292	15	GEM_myjtr	0.317
3	GFS_uwtr	0.292	16	GFS_mynn2mc	0.320
4	ARPEGE_uwtc	0.300	17	ARPEGE_myjtc	0.321
5	GFS_uwmc	0.302	18	GEM_mynn2tr	0.325
6	GFS_myjtc	0.305	19	ARPEGE_mynn2tc	0.325
7	ARPEGE_uwmc	0.305	20	GEM_ysutr	0.326
8	GFS_mynn2tc	0.305	21	ARPEGE_myjtr	0.327
9	GFS_ysutc	0.308	22	ARPEGE_ysumc	0.329
10	GFS_mynn2tr	0.309	23	GEM_mynn2mc	0.332
11	GEM_myjtc	0.309	24	ARPEGE_ysutc	0.341
12	GFS_ysutr	0.311	25	ARPEGE_mynn2mc	0.347

So “GFS_ysutr” corresponds to simulations that take GFS forecasts as initial and boundary conditions, and use YSU, Thompsom (T) and RRTMG (R) parameterizations for PBL, microphysics, and radiation, respectively.

Due to the large number of simulations and instability problems in some of the selected configurations, this study focuses on 25 combinations for which data were available for the entire study period. These 25 combinations are presented in Table 1, where they are ordered according to their root mean square error (RMSE). Thus, for each station and each hour, the difference between the simulated and observed precipitation for the whole period is examined, and the average RMSE ordered according to the accuracy is shown; in addition, the average RMSE of the ensemble of simulations is also indicated, and, as can be observed, it provides a smaller error than any of the individual simulations.

2.2. Benchmark methods

The classical methods, whose results will be used as a reference for the comparison of those obtained with the proposed methods, will be presented below.

2.2.1. Censored shifted gamma distribution (CSGD) based EMOS

The gamma distribution is commonly used to represent variations in precipitation amounts in both meteorology and climatology (Wilks, 1990). The probability density function for this distribution is:

$$f_{k,\theta}(y) = \frac{y^{k-1}e^{-y/\theta}}{\theta^k \Gamma(k)} \tag{1}$$

with shape $k > 0$ and scale $\theta > 0$, and being y the total precipitation over the period of interest ($y \geq 0$). Since, in addition to the amount of rainfall, the occurrence/non-occurrence must also be modeled, a distribution that allows for negative values can be used, left-censoring it at zero (Scheuerer and Hamill, 2015). Censoring allows the probability of no precipitation to be taken into account, being the probability of the negative values of the uncensored distribution. As the gamma distribution is not defined for negative values, a shift ($\delta > 0$) is introduced when calculating the corresponding cumulative distribution function, which is directly related to the probability of zero precipitation

$$F^0_{k,\theta,\delta} = \begin{cases} F_{k,\theta}(y+\delta), & \text{if y} \geq 0 \\ 0 & \text{if y} < 0 \end{cases} \tag{2}$$

This Censored and Shifted Gamma Distribution (CSGD) captures the right-skewed and mixed nature of precipitation data, while addressing heteroscedasticity through nonlinear, nonhomogeneous regression.

Scheuerer and Hamill (2015) proposed a model to calculate relationships between ensemble statistics and CSGD parameters, using ensemble forecasts, that is, an EMOS method. In this study, simplified regression equations are used, which are based on the ensemble mean only:

$$\begin{aligned} &\mu = {}^{\mu_{cl}}/_{a_1} \log\left(1 + \left[\exp(a_1) - 1\right](a_2 + a_3 \bar{f}/\bar{f}_{cl})\right), \\ &\sigma = a_4 \sigma_{cl} \sqrt{\mu/\mu_{cl}}, \\ &\delta = \delta_{cl} \end{aligned} \tag{3}$$

where $\bar{f}$ and $\bar{f}_{cl}$ represent the raw ensemble mean forecast and its climatological mean in the training data, respectively. As in previous works (Hu et al., 2023; Ghazvinian et al., 2021), the shift parameter δ is not adjusted, keeping it identical to the climatological shift, ensuring a reversion to climatology when forecast skill is low.

The four regression coefficients (a_1, a_2, a_3, a_4) and climatological CSGD are optimized by minimizing the continuous ranked probability score (CRPS) over the training data, following the approach outlined by Scheuerer and Hamill (2015).

Previous studies (Scheuerer and Hamill, 2015; Baran and Nemoda, 2016; Lin and Wang, 2024) have demonstrated the effectiveness of the CSGD model and its variants, particularly in generating well-calibrated forecasts under various meteorological conditions. However, recent research (Ghazvinian et al., 2020) has revealed a tendency for the predictive CSGD to underestimate the probability of precipitation (PoP) due to an inflated climatological shift parameter. This bias is particularly pronounced at shorter lead times and during periods of high precipitation. In this work, the efficiency of the EMOS CSGD is tested for hourly precipitation, with each of the locations for which observations are available trained individually. This model is hereinafter referred to as PredCSGD.

2.2.2. ANN-CSGD

A traditional Artificial Neural Network (ANN) has also been used to postprocess WRF outputs in order to compare the results with those provided by the proposed UNet models. As in the case of EMOS discussed above, the CSGD will be used, so the network, called ANN-CSGD, is particularly designed to predict the three key parameters of the CSGD: mean (μ), standard deviation (σ), and shift (δ). By using ANNs, the ANN-CSGD can learn complex relationships between ensemble forecast inputs and the distribution parameters for each of the locations for which observations are available

The architecture of the ANN-CSGD model consists of multiple fully connected layers with Leaky ReLU activation functions. The network outputs three values corresponding to μ, σ, and δ, which are then transformed using a Softplus activation function to ensure positive values. The optimization process minimizes a custom loss function based on the Continuous Ranked Probability Score (CRPS), which directly evaluates the quality of probabilistic forecasts and will be explained in detail in Section 2.5.2.

The ensemble data, used as input to the network, are previously normalized with mean 0 and standard deviation 1 in order to transform data into a consistent and standard format. During training, the model iteratively updates its weights to minimize CRPS on the training

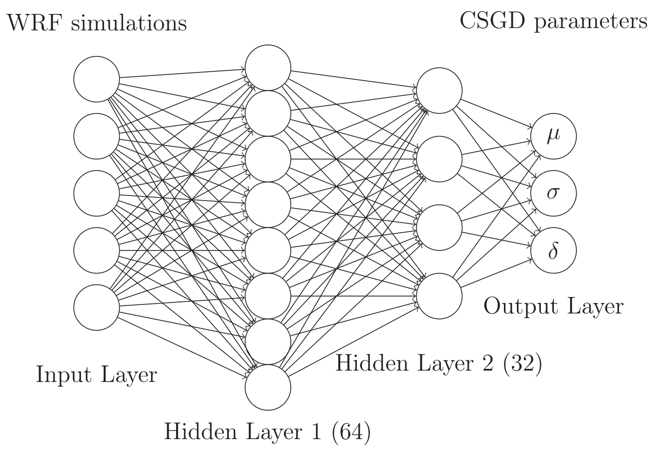


Fig. 2. Architecture of the ANN-CSGD model.

dataset. The validation process helps monitor performance and applies early stopping when the model does not improve for several epochs. Using validation set, extensive grid search of hidden layer dimensions was done to fine-tune the model for optimal performance. The final model architecture consists of two hidden layers with 64 and 32 neurons, respectively, and is presented in Fig. 2.

2.2.3. Analog ensemble

The Analog Ensemble (AnEn) technique is a method used to create forecast ensembles from deterministic weather predictions (Monache et al., 2013). Unlike earlier parametric methods, which first assume a predefined distribution for forecasting (e.g., precipitation) and then fit the distribution parameters to the data, AnEn works by leveraging similarities in past predictions and their corresponding observations to generate ensembles. The set of selected observations is used to construct the empirical CDF and the mean of the observations is used to obtain a deterministic value.

AnEn operates on the assumption that similar forecasts, generated by a consistent NWP model, will exhibit similar error patterns. By sampling from historical predictions that resemble the current forecast, the associated historical observation errors can be used to correct the forecast.

The distance between the multivariate deterministic forecast F_t at time t and the historical analog forecast $A_{t'}$ at time t' is calculated using a Euclidean distance:

$$\|F_t, A_{t'}\| = \sum_{i=1}^{N} w_i \sqrt{\sum_{j=-\tilde{t}}^{\tilde{t}} \Omega_j (F_{i,t+j} - A_{i,t'+j})^2} \tag{4}$$

where:

- N is the number of variables, which corresponds to the number of ensemble members.
- w_i is the weight for member i,
- $F_{i,t+j}$ is the target forecast for member i at time $t+j$,
- $A_{i,t'+j}$ is the historical analog forecast for variable i at time $t'+j$,
- $\tilde{t}$ is the half-size of the temporal window, used to find the most similar weather analogs, and
- Ω_j is the weight for time $t+j$.

This metric is calculated for all historical predictions. Those historical D times that have lower values of the metric will have their corresponding observations associated as members to form the analog ensemble. Once the analogs have been obtained, to calculate the CDF of the model, the empirical CDF is computed using the analog observations and the expected value is calculated as the average of the analog observations.

In this study, extensive grid search of parameters w_i, $\tilde{t}$, D, Ω_j were tested in the validation set for each location, finding $\tilde{t} = 1$, $D = 50$ and $\Omega = [0.8, 1, 0.8]$ as the best values for most cases.

The analog ensemble has demonstrated competitive performance when compared with other postprocessing algorithms (Hu et al., 2023), although it has difficulties in capturing extreme rainfall events (Jeworrek et al., 2023).

2.3. Proposed postprocessing model

To effectively process and analyze precipitation maps, this study proposes the use of a specific CNN. CNNs are deep learning models designed for processing grid-like data, such as precipitation maps. Its main feature is the use of convolution operations in the convolutional layers to extract spatial features. Additionally, CNNs may include other components to enhance learning, such as pooling layers to reduce spatial dimensions, batch normalization layers to stabilize training, and fully connected layers for classification or regression tasks. CNNs are particularly effective in tasks such as image recognition, segmentation and object detection.

2.3.1. Unet

In this study, a size-reduced version of the original UNet CNN (Ronneberger et al., 2015) is proposed for the probabilistic postprocessing of rainfall forecasting, due to its ability to capture fine spatial details thanks to its specific structure. The UNet architecture follows a U-shaped structure consisting of:

1. Contracting Path (Encoder): A series of convolutional and pooling layers that capture high-level features while reducing spatial dimensions.
2. Bottleneck: The transition between encoding and decoding, where the deepest feature representations are learned.
3. Expanding Path (Decoder): A series of upsampling and convolutional layers that restore the image resolution while preserving essential features.
4. Skip Connections: Direct links between corresponding encoder and decoder layers to retain fine-grained spatial information.

In Fig. 3 a detailed scheme of the proposed UNet model is shown. In this diagram, rectangles represent tensors, and arrows indicate different operations, as detailed in the figure legend. In addition, above each rectangle the corresponding number of channels (equal to the number of convolution filters) is indicated and next to each tensor block the height and the width are also displayed. The input to the network consists of N channels: N-1 channels corresponding to precipitation-related variables predicted by the ensemble members (detailed in Section 2.3.2), along with a terrain height map (HGT).

At the bottleneck, additional convolutional blocks are applied to capture more abstract features in the low resolution space. Next, the decoder where the reconstruction stage takes place is applied. In this stage, the network performs upsampling using deconvolution operations to gradually recover the spatial resolution, followed by two convolution blocks. During this process, skip connections are used, where feature maps from corresponding encoder layers are concatenated with the upsampled features, helping to retain spatial information that might otherwise be lost.

When the feature map height and width dimensions match the input size, an additional convolutional block is applied to obtain the three channels of the final output. These three channels pass through a softplus activation function to get a positive output corresponding to the three parameters of CSGD for each grid point.

To optimize the distribution parameters in the training process, the CRPS metric is minimized, which is related to minimizing the area between the Heaviside function shifted by the ground truth (empirical CDF) and the CDF of predicted CSGD.

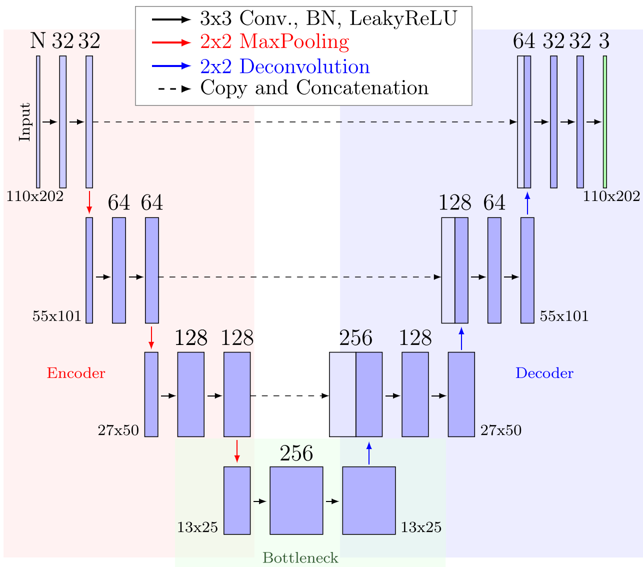


Fig. 3. Proposed UNet model. In the legend, BN stands for Batch Normalization, and Conv. represents Convolution. Multiplication notation indicates kernel size; for example, 2 × 2 MaxPooling corresponds to max pooling with a kernel size of 2. Additionally, N represents the number of input channels.

Optimizing the CRPS function explicitly can entail a large computational cost. For this reason, a numerical approach previously presented in Scheuerer and Hamill (2015) is used:

$$\begin{aligned}\mathrm{CRPS}(F_{k_i,\theta_i,\delta_i}, y_i) = {} & (y_i - \delta_i)[2F_{k_i,\theta_i}(y_i - \delta_i) - 1] \\ & - {}^{\theta_i k_i}/_{\pi}\,\beta\left(\frac{1}{2}, k_i + \frac{1}{2}\right)[1 - F_{2k,\theta_i}(-2\delta_i)] \\ & + \theta_i k_i[1 + 2F_{k_i,\theta_i}(-\delta)F_{k_i+1,\theta_i}(-\delta_i) - F_{k_i,\theta_i}(-\delta_i)^2 \\ & - 2F_{k_i+1,\theta_i}(y_i - \delta_i)] + \delta F_{k_i,\theta_i}(-\delta_i)^2\end{aligned} \tag{5}$$

where β is the beta function and i indicates each grid point. Then, the loss function is defined for all those grid points that coincide with the locations of the weather stations (N):

$$\mathcal{L} := \overline{\mathrm{CRPS}}(k, \theta, \delta, y) = {}^1/_N \sum_{i=1}^{N} \mathrm{CRPS}(F_{k_i,\theta_i,\delta_i}, y_i) \tag{6}$$

Previous studies have shown numerical instabilities with the use of this analytical form and propose regularizations (Hu et al., 2023). A new regularization in the predicted standard deviation is proposed in this study:

$$\mathcal{L} = \overline{\mathrm{CRPS}}(\mu, \sigma_{reg}, \delta, y) = \overline{\mathrm{CRPS}}(\mu, \sigma + \sqrt{\mu}, \delta, y) \tag{7}$$

where the mean μ and the standard deviation σ of the predicted CSGD are obtained from the relationships $k = {}^{\mu^2}/_{\sigma^2}$ and $\theta = {}^{\sigma^2}/_{\mu}$. This regularization ensures that the standard deviation always has an uncertainty related to the intensity of the rainfall event, facilitating the convergence of the UNet and also improving its performance.

2.3.2. Dimensionality reduction

The precipitation obtained from the 25 different WRF simulations can provide a valuable diversity of rainfall information to the UNet network allowing it to have a good inference performance, but it also has the disadvantage of possibly having a considerable amount of redundant information. To check this, different dimensionality reduction techniques are investigated: Feature Selection (FS) based on the correlation between the ensemble members, Principal Component Analysis (PCA), and Down Selection (DS) based on Integrated Gradients applied to the UNet trained with the full set of simulations. Although the latter methodology is also actually a feature selection method, which is performed from the trained model instead of directly using the results of the WRF simulations, it has been decided to use a common term in these cases, Down Selection, to distinguish it from the former one. Although newer methods for feature number reduction have appeared in recent years, the PCA technique, which has been widely used in different application fields, has also been used in this work. The input channels were previously standardized before using these dimensionality reduction techniques.

For feature selection, the D_{FS} members of the ensemble set that were least correlated with each other were selected. This not only helps to minimize data redundancy, allowing the model to capture a greater diversity of precipitation patterns with less redundant data, but also brings the added benefit of discarding the execution of unused WRF models, thus achieving a significant computational saving. Since the variables represent the same magnitude (precipitation), linear correlation is a reliable measure of similarity between them. Selecting the least linearly correlated variables is, therefore, a simple yet effective approach to reduce dimensionality without losing diversity of relevant information. In Fig. 4 the correlation matrix of all the variables used in this study is shown.

In the case of PCA, the D_{PCA} principal components with the highest explained variance were chosen. As the principal components consist of linear combinations of the precipitation predicted by the ensemble members, these components can retain some semantic meaning, allowing the UNet model to capture important patterns from the original data without requiring all channels.

The third approach, DS, uses Integrated Gradients (IG) (Sundararajan et al., 2017), a model interpretation method that allows assessing the contribution of each input feature to the prediction made by a model. IG measures how much the model output varies as the input changes, based on the idea that the sensitivity of a feature can be estimated by integrating the gradient of the prediction with respect to that feature along a path from a baseline point to the current feature value. This baseline input is used as a reference point representing the "absence" of any significant input features. Applying IG to the UNet trained with all simulations, the 25 ensemble members were ranked by their attribution scores, selecting the D_{DS} channels with the highest scores. Three cases from the validation set were used to compute IG scores, prioritizing cases where the UNet model demonstrated a marked improvement over the ensemble mean in precipitation estimation.

For D_{PCA} and D_{FS}, values of 5, 10, and 15 were tested to train the network. From the results in the validation set, the value 10 turned out to be the most favorable using the CRPS metric, that will be explained in Section 2.5. This corresponds to having a 66% total variance of the ensemble explained by the 10 PCA components and a maximum ensemble member correlation of 0.37 for the feature selection model. Following the previous findings from PCA and Feature Selection, the same number ($D_{DS} = 10$) was used for Integrated Gradients Down Selection method. To these 10 selected channels the HGT map is added to get the 11 final input channels.

Regarding the training details of the Unet, several configurations were tested, and the optimal parameters were found to be a batch size of 32, with the AdamW optimizer and a learning rate of $1e-3$. The model was trained for up to 40 epochs, with early stopping applied, if validation performance did not improve for 10 consecutive epochs, to prevent overfitting.

The UNet model with feature selection preprocessing will be referred to as UNet-FS, the model with PCA as UNet-PCA, and the model with Integrated Gradients Down Selection as UNet-DS.

To ensure a fair comparison across all methodologies, the feature selection approach applied to UNet-FS was also implemented for the other models, including the original WRF ensemble. This guarantees that all models are tested with the same reduced set of relevant input features, allowing a more controlled evaluation of their effectiveness across models. To facilitate the identification of all the models to be tested, their acronyms and the type of approach used, a summary is presented in Table 2.

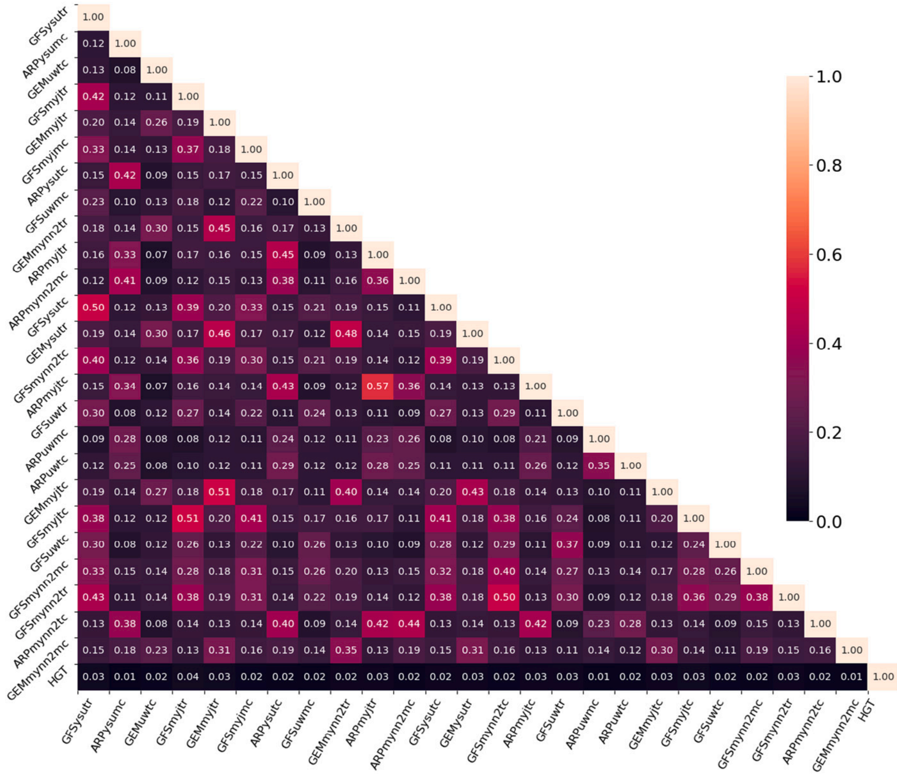


Fig. 4. Lower triangle of the correlation matrix using the 25 WRF ensemble members and HGT map.

Table 2
Summary of models used in this study.

Model type	Model name	Dimensionality Reduction	Distribution type
UNet	UNet-All	None	CSGD
	UNet-PCA	PCA	CSGD
	UNet-FS	FS	CSGD
	UNet-DS	DS	CSGD
Analog ensemble	AnEn	None	Empirical CDF
	AnEn-FS	FS	Empirical CDF
CSGD based EMOS	PredCSGD	None	CSGD
	PredCSGD-FS	FS	CSGD
ANN	ANN-CSGD	None	CSGD
	ANN-CSGD-FS	FS	CSGD
Ensemble	Ensemble	None	Empirical CDF
	Ensemble-FS	FS	Empirical CDF

2.4. Experimental setup

The experiments are divided into two groups. Firstly, the prediction skills at the weather stations are compared and, in a second stage, the spatial generalization capabilities of the UNet network are tested leaving stations out of the training.

For all the experiments the hourly precipitation data are divided by day into training (70%), validation (15%) and test (15%) sets. As the observational data are not balanced, with most days having no rainfall, in order to maintain a similar ratio in the different sets, the splitting is done in a stratified way by daily rainfall amount. Specifically, the daily amount of rainfall averaged by the weather stations was divided into four bins given by these limits: (0, 0.1, 4, 10, ∞). These four bins are the ones that maintained their proportionality in the three data sets. As the division is made by daily rainfall, all hours associated with each day are included in the same set.

Additionally, for the spatial generalization experiment, the station data are divided into 5 random folds. So the network is trained with the training data and stations from four of the folds and in the validation and test set the remaining fold is used. The folds are then rotated until all folds have been once in the separated set (a total of 5 runs). Regarding the benchmark methods, since a different model was trained for each station, it is not possible to assess the model's generalization to other geographical locations.

2.5. Evaluation metrics

Forecast quality is assessed through a suite of deterministic and probabilistic metrics, along with diagnostic tools that evaluate calibration and bias. Deterministic metrics, such as the Mean Absolute Error (MAE), Relative Bias, and Root Mean Square Error (RMSE), assess direct deviations between forecasts and observations. Probabilistic metrics, including the Continuous Ranked Probability Score (CRPS) and Brier Score (BS), evaluate the quality of uncertainty estimation in forecasts. An additional diagnostic tool, the Reliability Diagram, helps to illustrate the calibration of probabilistic forecasts methods.

2.5.1. Deterministic metrics

The MAE and RMSE measure the accuracy of deterministic forecasts by quantifying the error between observed and predicted values while the Relative Bias assesses the systematic tendency to over- or underpredict:

$$\begin{aligned} \text{MAE} &= \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \\ \text{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \\ \text{Relative Bias} &= \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{y}_i - y_i}{y_i + \epsilon} \end{aligned} \tag{8}$$

where y_i represents observations, $\hat{y}_i$ the predictions, N the number of observations, and ϵ is a small constant added to prevent division by zero (0.01). MAE provides a straightforward measure of average absolute error, while RMSE penalizes larger errors more heavily due to squaring, making it more sensitive to extreme discrepancies. The Relative Bias indicates whether the forecast tends to overestimate (positive values) or underestimate (negative values) the observations. Lower values, particularly lower absolute values for Relative Bias, indicate better forecast accuracy.

2.5.2. Continuous ranked probability score (CRPS)

The Continuous Ranked Probability Score (CRPS) is used to evaluate probabilistic forecasts by comparing the cumulative distribution function (CDF) of the forecast with the observed value. For a forecast with cumulative distribution function $F(y)$ and observation y, CRPS is defined as:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty}\left(F(x) - H_{\{x\geq y\}}\right)^2 dx \tag{9}$$

where $H_{\{x\geq y\}}$ is the Heaviside step function, which equals 1 when $x \geq y$ and 0 otherwise. The CRPS penalizes deviations between the forecast CDF and the actual empirical CDF (with the Heaviside function), effectively summarizing the forecast reliability and sharpness. Lower CRPS values indicate better probabilistic forecasts, as they represent smaller discrepancies between predicted and observed distributions. For deterministic forecasts, the CRPS simplifies to the MAE when evaluating against observed values, as both metrics effectively capture the magnitude of forecast errors.

2.5.3. Brier score (BS)

The Brier Score (BS) measures the accuracy of probabilistic forecasts for binary events, such as whether a precipitation threshold occurs (yes/no). It is defined as:

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{10}$$

where p_i is the predicted probability of the event, o_i is the observed outcome (1 if the event occurs, 0 otherwise), and N is the number of instances. The Brier Score ranges from 0 to 1, with values closer to 0 indicating more accurate probability forecasts. BS evaluates both the reliability and resolution of probabilistic forecasts, making it a valuable metric for assessing forecast quality in categorical prediction tasks.

2.5.4. Reliability diagram and calibration errors

The reliability diagram is a graphical tool to assess the calibration of probabilistic forecasts. It plots the observed relative frequency of an event against the predicted probability (Wilks, 2011). In this case, the event consists in exceeding a certain precipitation threshold. Perfect calibration corresponds to points lying along the diagonal. Deviations indicate whether forecasts are underconfident (below the diagonal) or overconfident (above the diagonal).

To quantify the degree of miscalibration observed in reliability diagrams, two metrics are commonly used (Guo et al., 2017):

- Expected Calibration Error (ECE), defined as:

$$\text{ECE} = \sum_{b=1}^{B}\frac{n_b}{N}\,|\text{acc}(b) - \text{conf}(b)| \tag{11}$$

  where B is the number of bins used to create the reliability diagram, n_b is the number of samples in bin b, N is the total number of samples, $\text{acc}(b)$ is the observed frequency in bin b, and $\text{conf}(b)$ is the mean forecast probability in the bin. ECE reflects the average discrepancy between confidence and accuracy across bins.
- Maximum Calibration Error (MCE), which measures the worst-case deviation:

$$\text{MCE} = \max_{b=1,\ldots,B}|\text{acc}(b) - \text{conf}(b)| \tag{12}$$

  This metric emphasizes the largest error across all probability bins and is useful to detect extreme miscalibration even if the overall average is low.

Together, the reliability diagram, ECE, and MCE provide a comprehensive view of how well the predicted probabilities align with observed event frequencies, helping to identify whether forecasts are trustworthy in a probabilistic sense.

2.5.5. Skill score

To enable fair comparison between models, skill scores (SS) are computed for each metric (if applicable). The skill score compares a forecast to a reference model (e.g., the raw ensemble):

$$\text{SS} = 1 - \frac{M_{\text{forecast}}}{M_{\text{reference}}} \tag{13}$$

where M denotes the value of the metric (e.g., CRPS, RMSE, BS). A skill score of 1 indicates a perfect forecast, 0 indicates parity with the reference, and negative values denote worse performance. By normalizing each metric in this way, the skill score allows for a consistent and interpretable comparison of forecast accuracy across RMSE, MAE, CRPS and BS.

3. Results

This section presents a comprehensive evaluation of the proposed models. First, the prediction skills are compared in the test set at the weather station locations. Then, an input sensitivity analysis is carried out to assess the relative influence of each ensemble member on the UNet predictions. Next, the ability of the models to generalize spatially is assessed through cross-validation on unseen locations. Finally, it is explored how the model performance is affected when the percentage of training stations is reduced, evaluating robustness under more constrained data scenarios.

Besides Relative Bias and Reliability Diagrams, the analysis evaluates MAE, RMSE, CRPS, BS($>$ 0.1 mm) and BS($>$ 99th percentile) skill scores using the ensemble mean as the reference, as it yields lower errors than individual simulations (Table 1). These scores are calculated for each observation site independently, computing also their confidence interval using bootstrapping (Ferro, 2007) by taking 1000 random samples with replacement. In particular, a block bootstrap has been used to resample the data in a way that respects the underlying serial dependence, since the assumption of independent random samples may provide narrower intervals than the real ones when serial correlation is present in the time series (Ferro, 2007; Wilks, 2010). The method proposed by Politis et al. (2004), and subsequently corrected by Patton et al. (2009), was used to estimate the size of the blocks. The values of the scores will be presented using box plots, one for each model considered (Table 2), in order to represent the dispersion of the results across the 130 observation sites. Specifically, the shaded boxes correspond to the 25–75 percentile interval, while the whiskers cover the 5%–95% interval. In addition, the values averaged over all locations are presented, with the corresponding confidence intervals, calculated by propagating the intervals of the individual scores, assuming that they are independent.

In order to calculate RMSE, MAE and Relative Bias for the analysis, the expected value of the distribution is taken as the deterministic value of the models.

3.1. Forecasting performance

3.1.1. Overall metrics

Figs. 5 and 6 present the skill score results of the probabilistic and deterministic metrics, respectively. The CRPSS results (Fig. 5(a)) show that all tested approaches (except Ensemble-FS) achieve positive values, confirming their improvement over the ensemble reference. All UNet-based architectures exhibit slightly higher median CRPSS values in the boxplots and higher mean CRPSS values (red markers), indicating their ability to enhance probabilistic forecasting. Furthermore, as the box plots illustrate, the improvement with respect to the ensemble occurs in most of the locations. However, differences among the UNet models (and other tested approaches) are not statistically significant within the 95% confidence level, showing the difficult task of processing hourly rainfall. The PredCSGD models also show competitive results, while ensemble-based methods like AnEn and AnEn-FS display slightly lower but still positive skill scores. Notably, the ANN-CSGD models present a wider spread, suggesting higher variability in their probabilistic forecasts.

For the BSS metric at the 0.1 mm threshold (Fig. 5(b)), the models exhibit significant variability in their ability to predict rainfall occurrence. UNet models achieve the highest skill scores and median values, but the overlapping confidence intervals prevent establishing a definitive advantage. The ANN-CSGD models, both with and without feature selection, do not perform as well as the other models, although they are still better than the ensemble mean.

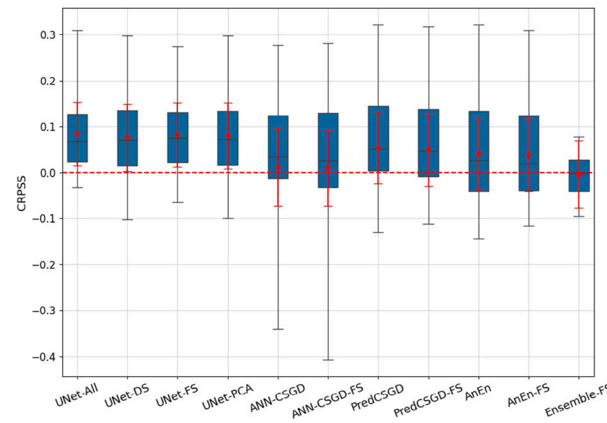
To further evaluate model performance under rare-event conditions, Fig. 5(c) shows the BSS results for the 99th percentile precipitation threshold. As expected, skill scores are generally lower for all models in this more challenging context. UNet models retain slightly positive median and averaged values, whereas Ensemble-FS and ANN-CSGD models exhibit scores closer to or below zero. Conversely, the results of the PredCSGD variants are centered near zero; although they may appear lower, they are not statistically distinguishable from the UNet models at the 95% confidence level. These results, where no model shows a clear improvement over the reference ensemble, underline the difficulty of predicting extreme rainfall events.

Figures S1 and S2 in the supplementary material, show the spatial distribution of the CRPSS and BSS(> 0.1 mm) skill scores. These maps do not show clear spatial patterns. In general, post-processing methods tend to achieve greater skill on the western part of the central island (Tenerife) and on the easternmost island (Gran Canaria).

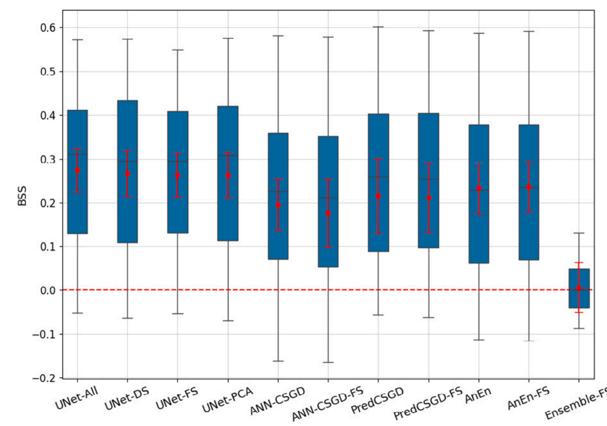
Furthermore, to explore model behavior over lead time, Figures S3 and S4 in the supplementary material show the hourly evolution of the CRPSS and BSS(> 0.1 mm) across the forecast horizon. UNet-based models maintain the highest average skill throughout nearly all hours, showing robustness across the lead forecast period. Since the skill scores compare each method to the reference one, the absence of a clear trend may mean that both corresponding scores behave over time in a similar manner.

Fig. 6 presents the results for the deterministic metrics (RMSESS, MAESS and Relative Bias), following the same structure as the probabilistic metrics. These deterministic metrics, while losing the probabilistic information of the models, provide more information on the models' ability to reduce biases and absolute and squared errors relative to the ensemble mean.

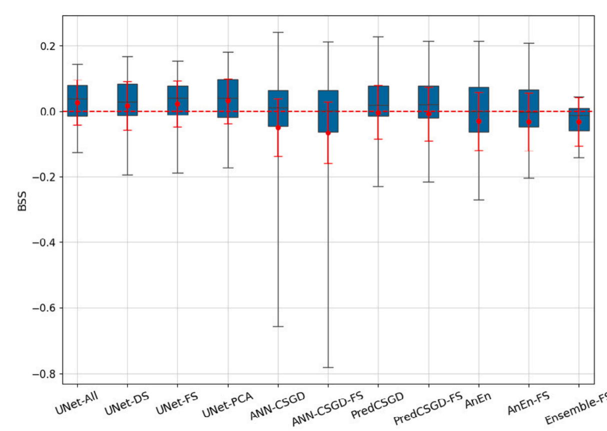
The RMSESS results (Fig. 6(a)) suggest that UNet-based and PredCSGD models are the only ones showing moderate improvements in deterministic forecasting. The feature selection variants (UNet-FS and ANN-CSGD-FS) demonstrates competitive performance while requiring fewer inputs, indicating that reducing the number of ensemble members does not strongly impact deterministic forecast accuracy. On the other hand, the spread in RMSESS is more pronounced for AnEn-FS and ANN-CSGD-FS, reflecting the model inconsistencies across stations when reducing ensemble members. Similarly, the MAESS metric (Fig. 6(b)) suggests that all models, except analog-based methods and Ensemble-FS, show a robust improve upon the ensemble mean, with no single approach displaying a dominant advantage.



(a) CRPSS



(b) BSS (> 0.1 mm)



(c) BSS (> 99^{th} percentile)

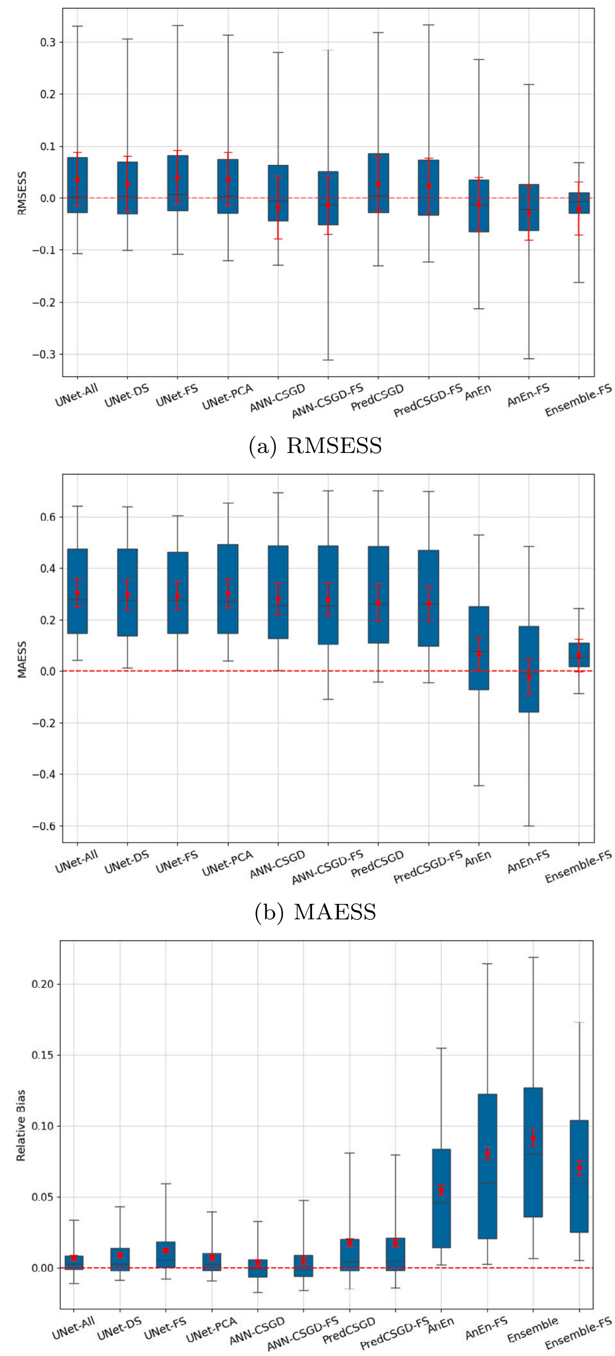
Fig. 5. Skill score results of probabilistic metrics with the ensemble of WRF simulations as reference. Red error bars indicate 95% confidence intervals of the mean across locations, while boxplots display individual skill scores for each station.

An important point to note is that RMSE penalizes large errors more heavily due to the squared differences, making it particularly sensitive to deviations in extreme precipitation events. MAE, by taking the absolute differences rather than squared differences, is less sensitive to large errors. The higher MAESS values compared to RMSESS further confirm this behavior, suggesting that while post-processing generally reduces absolute error, their ability to mitigate large overpredictions or underpredictions is more limited.

To complement the analysis, Fig. 6(c) shows the Relative Bias of each method. All UNet-based architectures display low bias, with median values close to zero and narrow interquartile ranges, indicating a stable central tendency in their predictions. ANN-CSGD models also maintain low and consistent bias values. In contrast, ensemble-based methods and AnEn variants show a systematic tendency toward overestimation, with wider dispersion and significantly higher mean bias, especially for Ensemble-FS and AnEn-FS. PredCSGD models present intermediate behavior, with low medians but slightly higher variability.

Overall, feature selection does not seem to deteriorate the quality of the models, indicating that reducing the number of simulations does not significantly diminish ensemble variability. Evidence of this is that the Ensemble-FS model achieves very similar results in the metrics to the raw ensemble, as shown by skill score error bars centered at zero. Additionally, no significant differences are observed between the use of principal components (UNet-PCA) and feature selection (UNet-FS and UNet-DS), suggesting that both methods effectively preserve the key variability of the ensemble while reducing input dimensionality, with no statistically significant advantage of one approach over the other in terms of forecast skill. However, PCA requires an extra preprocessing step and the use of all ensemble members, while feature selection can be directly applied based on correlation analysis, potentially simplifying implementation and reducing computational overhead.

In addition to skill scores, forecast reliability is evaluated in Fig. 7, which presents reliability diagrams for a 0.1 mm threshold. Each plot compares predicted probabilities to observed frequencies, with Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) also reported in each panel. UNet models, particularly UNet-FS and UNet-PCA, are the best calibrated, showing smooth curves close to the diagonal and lower ECE/MCE values. This seems to indicate the benefits of dimensionality reduction with PCA for the results but especially of the FS feature selection methodology, which has reduced the number of ensemble members that are used. PredCSGD models closely follow results of UNets, and raw ensemble methods (both with all channels and with FS approach) tend to overpredict rainfall occurrence, evidenced by a calibration function entirely located to the right of the 1:1 line, and by the highest calibration errors. ANN-CSGD models present more erratic reliability curves and large MCEs, consistent with their wider spread in skill scores. AnEn models show reasonable calibration but with a tendency toward underconfidence, especially in low-probability bins. These results reinforce the advantage of convolutional approaches, not only in improving skill scores but also in delivering reliable probabilistic forecasts.



(a) RMSESS

(b) MAESS

(c) Relative Bias

Fig. 6. Same as Fig. 5 but for deterministic metrics.

3.1.2. Input sensitivities

To assess the relative attribution of each input feature to the UNet model's accuracy in precipitation forecast, a comprehensive sensitivity analysis was conducted in order to identify the most influential factors driving the UNet model's rainfall forecasting capabilities. The purpose is to gain insight into how the network prioritizes features such as specific ensemble members or spatial patterns, thus contributing to the understanding of its predictions.

To this end, Integrated Gradients (IG) technique is used to study the sensitivities of the final UNet models. The IG attribution technique was conceptually presented in Section 2.3.2 to explain the Down Selection method applied to the UNet trained with all input variables. Readers can also refer to Sundararajan et al. (2017) for more details on how IG works.

To implement this technique, the following procedure is used:

1. Three of the rainiest samples of the ensemble were chosen in the test set. Also, zero values were used as baseline inputs.
2. Attributions were calculated using IG at the locations of the weather stations in the three channels of the UNet output maps, these attributions were summed for each input channel.
3. Additionally, for the PCA model case, the principal component matrix was used to reverse the change of basis and thus express

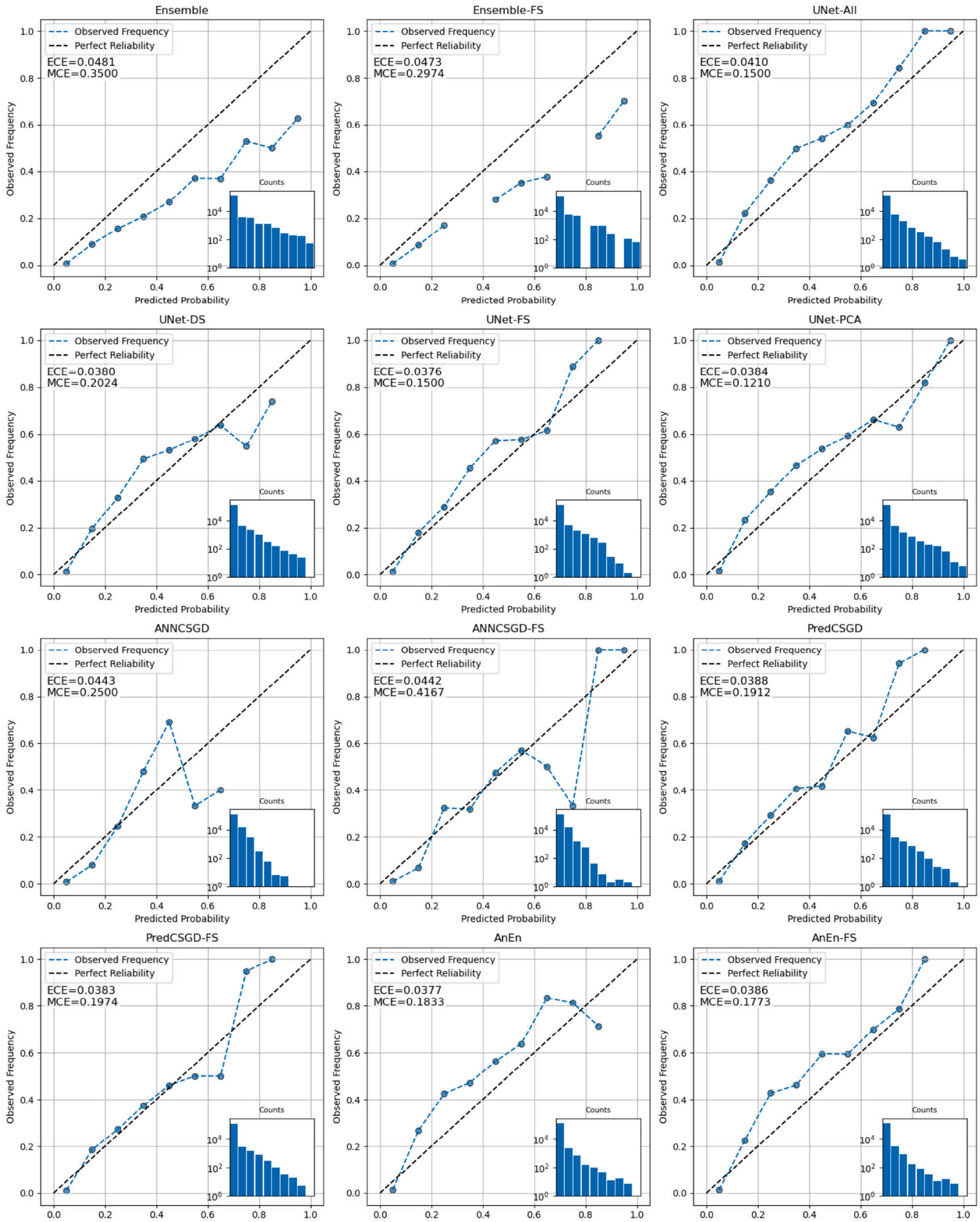


Fig. 7. Reliability diagrams for all models at a 0.1 mm precipitation threshold. Inset histograms show the count of samples per bin in logarithmic scale.

the contributions in the original space to make them comparable with the feature selection results. All final sensitivities are expressed in absolute value.

Fig. 8 shows the sensitivities of the 10 PCA components plus the HGT map for the UNet-PCA model. Looking at the principal components, it can be seen that the model is, as expected, highly sensitive to the first component, while for the rest of the components the model is equally sensitive with a very low sensitivity. The analysis of the PCA transformation revealed that the first principal component is close to the average of the ensemble members (although not normalized). This makes sense since the ensemble average is usually taken to give the final results of the ensemble. In addition, since all the original variables contribute a significant weight in the formation of the first principal component, this also explains why the sensitivity of this component is high.

On the other hand, a very high sensitivity of the model to the HGT map can be seen, even higher than the first principal component. It is important to note that the high sensitivity of the HGT map does not directly imply that this channel is the most important for the prediction in all instances, but rather indicates that the precipitation outputs are highly dependent on how the inputs vary with respect to the fixed topography of the area. Since the value of HGT does not change between instances of the dataset, its high sensitivity reflects how the

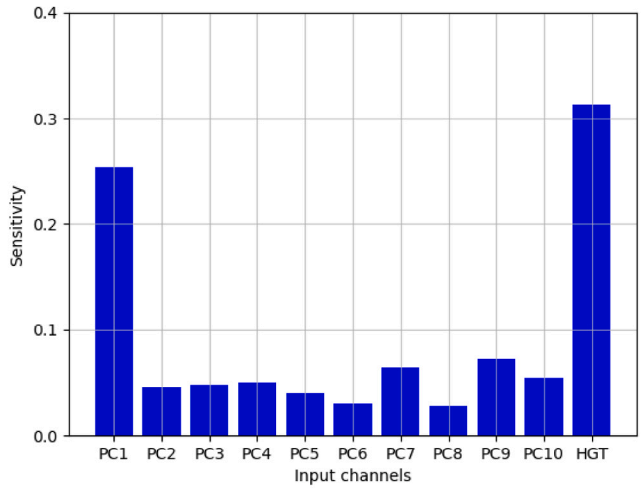


Fig. 8. Channel sensitivities with respect to the output in the location of the weather stations for UNet-PCA model.

model adjusts its rainfall prediction for different meteorological inputs depending on the underlying topographic structure. This suggests that the variations of the other channels are modulated by the terrain, which is crucial to improve precipitation predictions in orographically complex regions such as the Canary Islands.

Continuing with the analysis, Fig. 9 shows the sensitivities of all Unet-based models, with the UNet-PCA broken down into the original variables by undoing the PCA transformation. In addition, the bars corresponding to the 11 channels with the highest sensitivity are colored in a darker red. For the feature selection and down selection models this corresponds to all input channels. In all models, it is observed a clear trend again: the HGT map presents the highest sensitivity by far compared to other channels. This reinforces the idea of the dependence of the final results of the UNet models on the orography.

Regarding the channels corresponding to the ensemble members, for UNet-PCA and UNet-All a similar sensitivity can be observed for all members, none of them exceeding 0.1. So these models seem to spread the sensitivity over the whole ensemble benefiting from their variety. Especially for UNet-PCA this behavior is clear, where the attribution to the HGT map is not as high as in UNet-All and all ensemble members have a similar contribution to each other. This is consistent with Fig. 8, as the first principal component had high sensitivity, undoing the decomposition spreads its sensitivity across all features. For UNet-FS the sensitivity of the inputs is not as equally distributed, which could be understandable due to the differences in the dimensionality reduction approach. The model seems to give important attributions to individual members such as GEM_mynn2mc and GFS_uwtc, the latter being the member with the best RMSE metric obtained individually for rainfall prediction (see Table 1). By reducing the number of members at the UNet input, the model seems to start paying attention to individual advantages of each member. In this sense, it makes sense to look at a member with competent results such as GFS_uwtc. In the case of UNet-DS, which had not reached as competitive results as the rest in the test set as shown in the previous Section 3.1.1, the attribution to ensemble members is not as localized as in UNet-FS, giving larger importance to the HGT map. Even so, model attribution is still observed not only to GFS_uwtc but also to GEM_uwtc, which is the second best performing RMSE member (Table 1).

3.2. Spatial generalization performance

As previously mentioned in Section 1, the benchmark methods are mainly focused on obtaining results only at the grid points corresponding to weather station locations. However, since the predictions of mesoscale numerical models, such as WRF, provide results on a grid covering the entire area of interest, postprocessing methods that could directly improve the outputs of such models not only at those points where observations are available, but throughout the entire simulation domain, would be desirable (Yang et al., 2021).

To rigorously assess the network's ability to generalize spatially, that is, to predict precipitation at locations not included in training, a five-fold cross-validation strategy is employed. In each fold, one-fifth of the weather stations are set aside for evaluation, ensuring that the model's performance is evaluated on unseen grid points in every fold. This methodology provides insight into the model's capacity to interpolate precipitation patterns across the entire domain based on limited observation points.

Figs. 10 and 11 illustrate the probabilistic and deterministic metrics, respectively. In the case of skill scores, they are computed using the ensemble mean as the reference forecast. These figures follow the same plotting scheme as Figs. 5 and 6, described in the initial paragraph of Section 3. The results are concatenated across all test splits (one for each fold) to provide a comprehensive evaluation. Additionally, Ensemble-FS results are also shown as it directly provides precipitation maps with the selected WRF simulations.

The results confirm that the UNet-based models are able to generalize well to unseen grid points, maintaining positive skill scores in both probabilistic and deterministic verification metrics. Specifically, BSS (> 0.1 mm), MAESS and Relative Bias show statistically significant improvement (with 95% level) to the ensemble and ensemble with feature selection (Ensemble-FS) suggesting that the proposed UNet models specially enhance light-to-moderate rainfall predictions and improve the calibration of overall bias even at locations not seen during training.

Regarding extreme precipitation events, the BSS (99th percentile) metric (Fig. 10(c)) reveals no statistically significant improvement over the ensemble. Although the UNet models do not underperform in these cases, their advantage over the ensemble diminishes when it comes to rare, high-intensity rainfall. This highlights that, while spatial generalization is preserved for common rainfall intensities, capturing extremes at unseen locations remains a challenging task.

Further analysis of the different UNet variants reveals that models leveraging dimensionality reduction techniques (UNet-PCA, UNet-DS and UNet-FS) do not exhibit a degradation or improvement in performance compared to UNet-All, which directly uses all ensemble members. This reinforces the idea that reducing input dimensionality, whether through principal component analysis or feature selection, does not discard crucial information for spatial generalization.

Overall, the results support the applicability of UNet-based architectures for high-resolution precipitation mapping across the Canary Islands. Furthermore, they highlight the potential for reducing the high computational cost of WRF simulations by using fewer ensemble members (as in UNet-FS and UNet-DS), while maintaining predictive performance.

3.3. Effect of training station density

To assess the importance of station density for training spatial models, we conducted an additional experiment in which the number of training stations was progressively reduced, while keeping the test stations set fixed. The UNet-FS model was retrained using 80%, 60%, 40%, and 20% of the original training stations, corresponding to 104, 78, 52, and 26 weather stations, respectively. Fig. 12 shows the resulting bootstrap confidence intervals of the CRPSS, RMSESS, BS(0.1), and BS(99th percentile) skill scores. Boxplots show the dispersion of skill scores across the test stations.

The results of this station reduction experiment show a logical but not very pronounced degradation of skill scores in all metrics as the number of training stations is reduced. This trend is evident in CRPSS, RMSESS, and BS metrics. While the median values remain

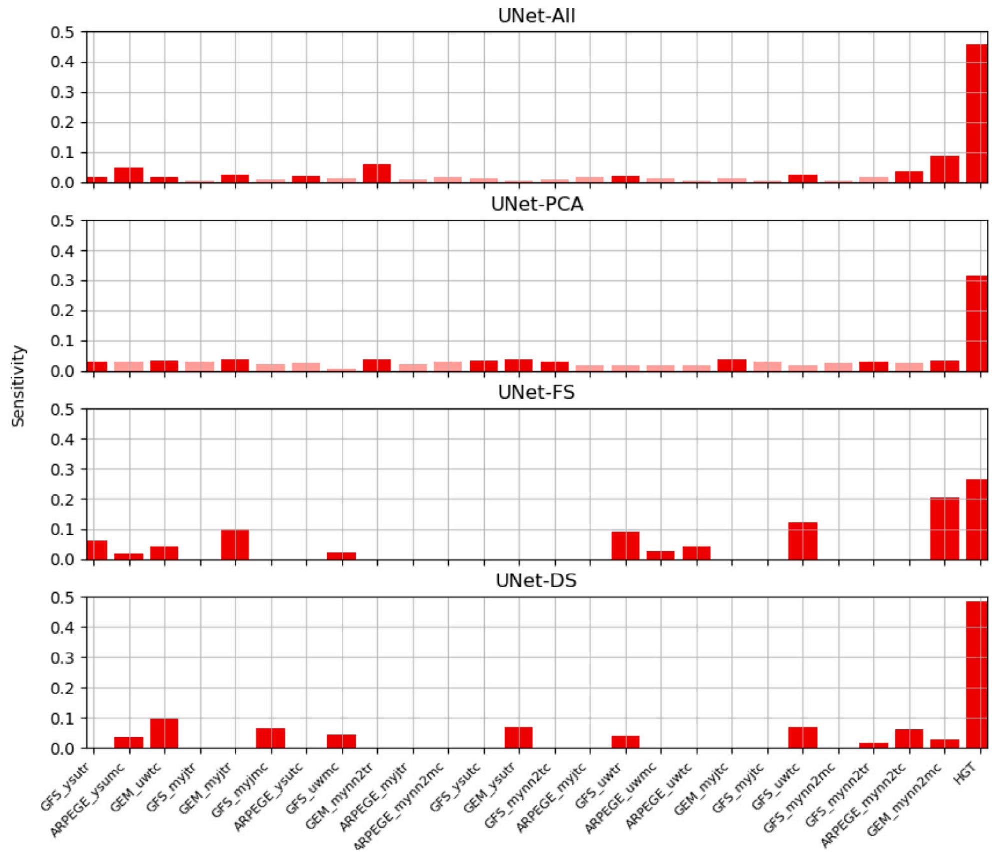


Fig. 9. Channel sensitivities with respect to the output in the location of the weather stations. The 11 channels with higher sensitivities are colored in a darker red.

relatively stable, the spread of the scores, specially shown in the boxplots, increases noticeably with fewer training stations. This suggests a moderate deterioration in prediction precision and spatial generalization when less spatial information is available during training, though the proposed method is still valuable, improving some of the metrics with respect to the ensemble even when a low number of stations are available for training. However, the impact is particularly significant in the prediction of extreme events, as reflected in the confidence intervals of BSS at 99th percentile threshold metric (Fig. 12d). This indicates that with fewer training stations, the network may resort to learning more global corrections of the input WRF ensemble rather than capturing localized variability, which negatively affects its ability to properly correct for high rainfall values. This loss of resolution in the extremes underscores the importance of a dense and geographically diverse observational network when training deep learning models for spatial postprocessing.

4. Summary and conclusions

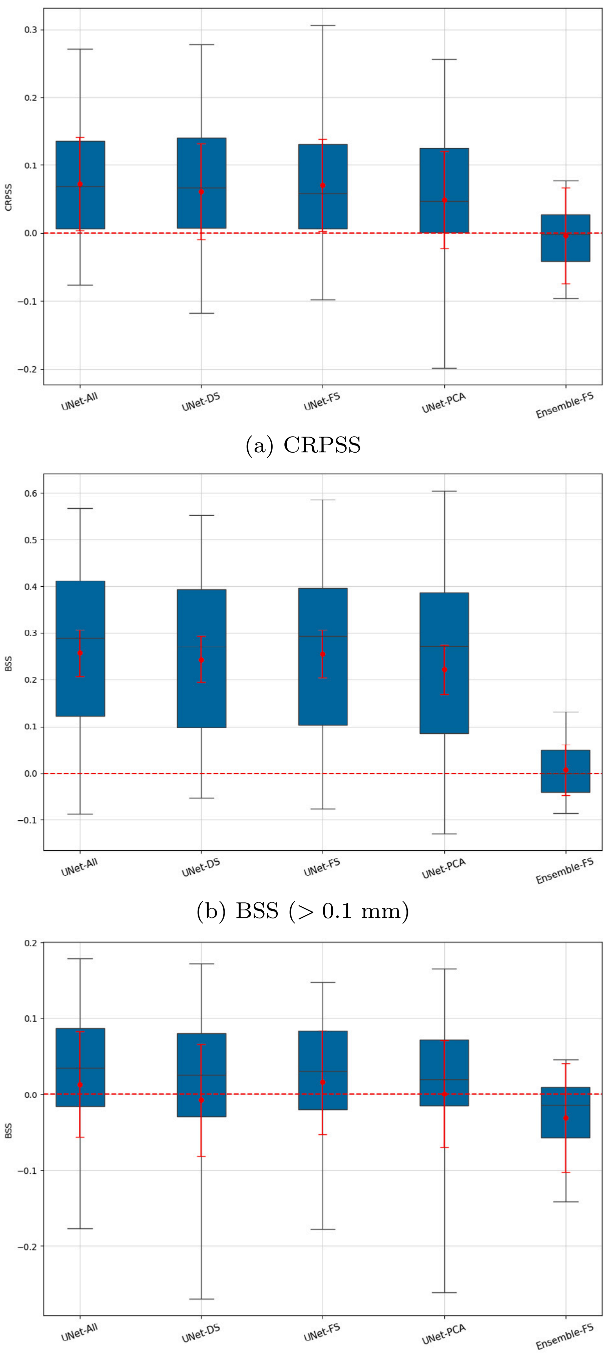
This study presented a comprehensive analysis of different data-driven postprocessing models applied to the Canary Islands, in order to improve ensemble numerical WRF precipitation forecasting outputs. The results demonstrated that the proposed models based on the UNet architecture, without dimensionality reduction and with it through PCA, feature selection and down selection, performed as well as or better than traditional benchmarks such as Censored Shifted Gamma Distribution fitting with Ensemble Model Output Statistics (EMOS), Analog Ensemble and ANN in all analyzed metrics. The proposed UNet models efficiently learned the parameters of a CSGD to capture forecast uncertainty effectively.

The results demonstrate that UNet-based models consistently outperform the WRF simulations ensemble mean, achieving positive skill scores across all tested probabilistic (CRPSS, BSS) and deterministic (MAESS, RMSESS) metrics, and reducing Relative Bias. Specially, the models showed an enhance in the prediction of light-to-moderate rainfall events. These improvements suggest that deep learning architectures can effectively learn spatial precipitation patterns from sparse observations, leveraging convolutional operations to infer high-resolution output maps.

Despite the overall positive results, significant variability was observed in the skill scores across different locations, as reflected by the large interquartile ranges and extended whiskers in the boxplots. This variability highlights the challenging nature of spatial generalization, indicating that forecast performance depends in some way on the station locations. Nevertheless, the fact that the majority of skill score distributions lie above zero confirms that UNet models consistently outperform the raw ensemble forecast, even in unseen locations during UNets training. These findings suggest that the deep learning approach effectively captures spatial precipitation patterns, even when trained with sparse observational data.

The integration of the HGT variable as an additional input further contributed to the success of the UNet models, particularly in orographically complex regions. The sensitivity analysis using Integrated Gradients highlighted the significance of the terrain's topography, showing that the incorporation of the static variable HGT in the UNet models adequately modulates rainfall forecasts. This underlines the necessity of incorporating terrain information in precipitation models for regions with complex geographical features.

This study demonstrates that convolutional neural networks like UNet, combined with dimensionality reduction techniques such as PCA, feature selection, and down selection, are highly effective for hourly precipitation forecasting in complex geographical regions. The proposed feature selection methodology significantly reduces the need

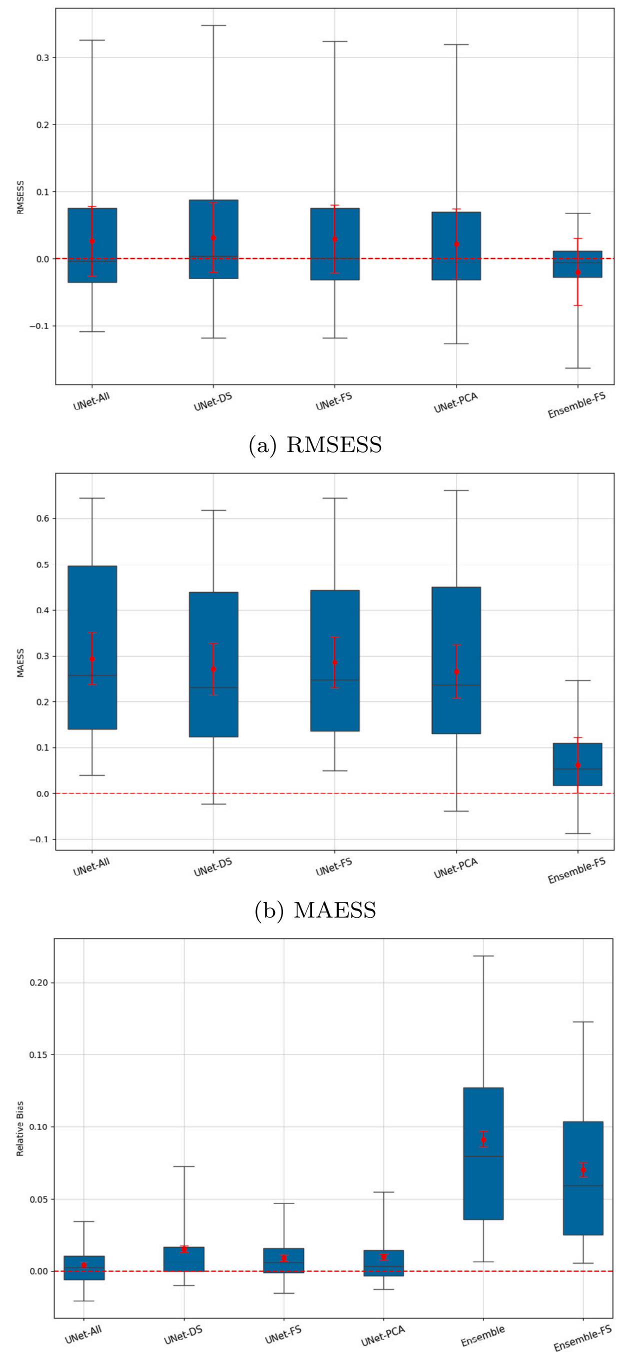


(a) CRPSS

(b) BSS (> 0.1 mm)

(c) BSS (> 99^{th} percentile)

Fig. 10. Skill score results of probabilistic metrics with the ensemble as reference. Red error bars show 95% confidence intervals from bootstrapping and blue boxplots show individual skill scores across stations. The results are constructed by concatenating the test splits from each of the 5 folds.



(a) RMSESS

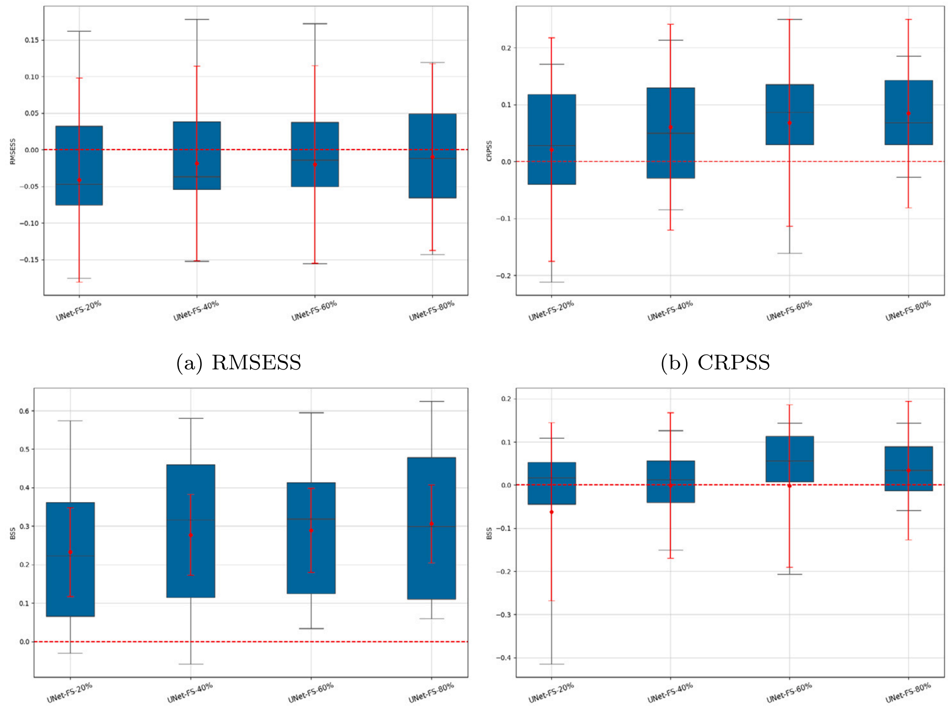
(b) MAESS

(c) Relative Bias

Fig. 11. Same as Fig. 10 but for deterministic metrics.

for computationally expensive numerical simulations while maintaining predictive accuracy. Results show that models utilizing principal component analysis (UNet-PCA) and feature selection (UNet-FS, UNet-DS) perform comparably to the full-input model (UNet-All). Notably, UNet-FS and UNet-DS achieve this efficiency by discarding 15 WRF simulation members without compromising predictive quality.

Overall, the results suggest that convolutional-based architectures like UNet offer a promising framework for ensemble post-processing. Their ability to generalize spatially, learn from limited observational data, and incorporate orographic influences makes them well-suited for high-resolution precipitation forecasting in complex terrain regions. Further research could explore more sophisticated dimensionality reduction techniques like auto-encoders (Michelucci, 2022) and/or the use of additional predictors from each ensemble member, such as the Integrated Vapor Transport (IVT) which has also been shown to be possibly relevant for rainfall forecasting (Hu et al., 2023). Additionally, it would be interesting to compare the performance of the CSGD against other advanced postprocessing distributions such as the hurdle EBXII distributions((Jin et al., 2024).



(a) RMSESS (b) CRPSS

(c) BSS (>0.1 mm) (d) BSS (> 99^{th} percentile)

Fig. 12. Skill metrics and their associated uncertainty (95% CI) for different proportions of training stations.

5. Software and data availability

The codebase for this study is built using PyTorch, a Python-based deep learning library. All code used in this work is publicly available in a GitHub repository, along with principal documentation provided in the README.md file. This README file also contains guidance for downloading and using the data and the pre-trained models used in this study. The repository can be accessed at https://github.com/marcosesgonz/ProbabilisticPostprocessingPrecipitationUNets.

CRediT authorship contribution statement

Marcos Esquivel-González: Writing – original draft, Software, Methodology. **Albano González:** Writing – review & editing, Methodology. **Juan Carlos Pérez:** Writing – review & editing, Conceptualization. **Juan Pedro Díaz:** Writing – review & editing, Data curation.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve readability and language. After using this service, the authors reviewed and edited the content as needed and assume full responsibility for the content of the published article.

Funding sources

This research was supported by the PLANCLIMAC2 Project (1/MAC/2/2.4/0006), financed through the European Union's INTERREG MAC 2021–2027 Program. Additional support was provided by the Government of the Canary Islands, through the Consejería de Transición Ecológica *y* Energía, as part of the cooperation agreement signed on November 8, 2024.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Juan Pedro Diaz reports financial support was provided by European Union INTERREG MAC 2021–2027 Program. Juan Pedro Diaz reports financial support was provided by Government of the Canary Islands, Consejeria de Transicion Ecologica y Energia. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge the support provided by the Teide High-Performance Computing (TeideHPC) facilities. These resources are made available by the Instituto Tecnológico *y* de Energías Renovables (ITER), S.A. (URL: teidehpc.iter.es)

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103255.

References

Aha, D.W., 1997. Lazy Learning. Springer Dordrecht, USA, http://dx.doi.org/10.1007/978-94-017-2053-3.

Arnold, D., Morton, D., Schicker, I., Seibert, P., Rotach, M., Horvath, K., Dudhia, T., Satomura, T., Müller, M., Zängl, G., et al., 2012. Issues in high-resolution atmospheric modeling in complex topography–The HiRCoT Workshop. Hrvatski meteorološki časopis 47, 3–11.

Baran, S., Nemoda, D., 2016. Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. Environmetrics 27, 280–292, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2391, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.2391.

Côté, J., Gravel, S., Méthot, A., Patoine, A., Roch, M., Staniforth, A., 1998. The operational CMC–MRB global environmental multiscale (GEM) model. Part I: Design considerations and formulation. Mon. Weather Rev. 126, 1373–1395. http://dx.doi.org/10.1175/1520-0493(1998)126{<}1373:TOCMGE{>}2.0.CO;2, URL: https://journals.ametsoc.org/view/journals/mwre/126/6/1520-0493_1998_126_1373_tocmge_2.0.co_2.xml.

Doury, A., Somot, S., Gadat, S., 2024. On the suitability of a convolutional neural network based RCM-emulator for fine spatio-temporal precipitation. Clim. Dyn. 62, 8587–8613. http://dx.doi.org/10.1007/s00382-024-07350-8.

Doury, A., Somot, S., Gadat, S., Ribes, A., Corre, L., 2023. Regional climate model emulator based on deep learning: concept and first evaluation of a novel hybrid downscaling approach. Clim. Dyn. 60, 1751–1779. http://dx.doi.org/10.1007/s00382-022-06343-9.

Dupuy, F., Durand, P., Hedde, T., 2023. Downscaling of surface wind forecasts using convolutional neural networks. Nonlinear Process. Geophys. 30, 553–570. http://dx.doi.org/10.5194/npg-30-553-2023, URL: https://npg.copernicus.org/articles/30/553/2023/.

Ferro, C.A.T., 2007. Comparing probabilistic forecasting systems with the Brier score. Weather. Forecast. 22, 1076–1088. http://dx.doi.org/10.1175/WAF1034.1, URL: https://journals.ametsoc.org/view/journals/wefo/22/5/waf1034_1.xml.

Ghazvinian, M., Zhang, Y., Seo, D.J., 2020. A nonhomogeneous regression-based statistical postprocessing scheme for generating probabilistic quantitative precipitation forecast. J. Hydrometeorol. 21, 2275–2291. http://dx.doi.org/10.1175/JHM-D-20-0019.1, URL: https://journals.ametsoc.org/view/journals/hydr/21/10/jhmD200019.xml.

Ghazvinian, M., Zhang, Y., Seo, D.J., He, M., Fernando, N., 2021. A novel hybrid artificial neural network - parametric scheme for postprocessing medium-range precipitation forecasts. Adv. Water Resour. 151, 103907. http://dx.doi.org/10.1016/j.advwatres.2021.103907, URL: https://www.sciencedirect.com/science/article/pii/S0309170821000622.

Gneiting, T., Raftery, A.E., Westveld, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133, 1098–1118. http://dx.doi.org/10.1175/MWR2904.1, URL: https://journals.ametsoc.org/view/journals/mwre/133/5/mwr2904.1.xml.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning. PMLR, pp. 1321–1330, URL: https://proceedings.mlr.press/v70/guo17a.html.

Haiden, T., Janousek, M., Vitart, F., Ben-Bouallegue, Z., Prates, F., 2023. Evaluation of ECMWF forecasts, including the 2023 upgrade. http://dx.doi.org/10.21957/d47ba5263c.

Hamill, T.M., Scheuerer, M., 2018. Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. Mon. Weather Rev. 146, 4079–4098. http://dx.doi.org/10.1175/MWR-D-18-0147.1, URL: https://journals.ametsoc.org/view/journals/mwre/146/12/mwr-d-18-0147.1.xml.

Höhlein, K., Schulz, B., Westermann, R., Lerch, S., 2024. Postprocessing of ensemble weather forecasts using permutation-invariant neural networks. Artif. Intell. the Earth Syst. 3, e230070. http://dx.doi.org/10.1175/AIES-D-23-0070.1, URL: https://journals.ametsoc.org/view/journals/aies/3/1/AIES-D-23-0070.1.xml.

Hu, W., Ghazvinian, M., Chapman, W.E., Sengupta, A., Ralph, F.M., Monache, L.D., 2023. Deep learning forecast uncertainty for precipitation over the Western United States. Mon. Weather Rev. 151, 1367–1385. http://dx.doi.org/10.1175/MWR-D-22-0268.1, URL: https://journals.ametsoc.org/view/journals/mwre/151/6/MWR-D-22-0268.1.xml.

ITER, 2024. TeideHPC. URL: https://www.iter.es/. (Online; Accessed 4 July 2024).

Jeworrek, J., West, G., Stull, R., 2023. Improved analog ensemble formulation for 3-hourly precipitation forecasts. Weather. Forecast. 38, 1323–1339. http://dx.doi.org/10.1175/WAF-D-23-0018.1, URL: https://journals.ametsoc.org/view/journals/wefo/38/8/WAF-D-23-0018.1.xml.

Ji, Y., Zhi, X., Ji, L., Zhang, Y., Hao, C., Peng, T., 2022. Deep-learning-based post-processing for probabilistic precipitation forecasting. Front. Earth Sci. http://dx.doi.org/10.3389/feart.2022.978041, Volume 10-2022. URL: https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2022.978041.

Jin, H., Jiang, W., Chen, M., Li, M., Bakar, K.S., Shao, Q., 2023a. Downscaling long lead time daily rainfall ensemble forecasts through deep learning. Stoch. Environ. Res. Risk Assess. 37, 3185–3203. http://dx.doi.org/10.1007/s00477-023-02444-x.

Jin, H., Liu, Y., Shao, Q., Li, M., 2023b. Postprocessing ensemble forecasts with generative adversarial networks for daily precipitation. In: 2023 IEEE International Conference on Knowledge Graph. ICKG, pp. 152–159. http://dx.doi.org/10.1109/ICKG59574.2023.00024.

Jin, H., Mahani, M.E., Li, M., Shao, Q., Crimp, S., 2024. Probabilistic seasonal precipitation forecasts using quantiles of ensemble forecasts. Stoch. Environ. Res. Risk Assess. 38, 2041–2063. http://dx.doi.org/10.1007/s00477-024-02668-5.

Li, M., Jin, H., Shao, Q., 2021. Improvements in subseasonal forecasts of rainfall extremes by statistical postprocessing methods. Weather. Clim. Extrem. 34, 100384. http://dx.doi.org/10.1016/j.wace.2021.100384, URL: https://www.sciencedirect.com/science/article/pii/S2212094721000748.

Lin, H.M., Wang, L.P., 2024. Enhancing radar-based ensemble nowcasting with CSGD: A UK study. In: EGU General Assembly Conference Abstracts. p. 2799. http://dx.doi.org/10.5194/egusphere-egu24-2799.

Lorenz, E.N., 1963. Deterministic nonperiodic flow. J. Atmospheric Sci. 20, 130–141. http://dx.doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2, URL: https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml.

Michelucci, U., 2022. An introduction to autoencoders. CoRR, abs/2201.03898. URL: https://arxiv.org/abs/2201.03898, arXiv:2201.03898.

Monache, L.D., Eckel, F.A., Rife, D.L., Nagarajan, B., Searight, K., 2013. Probabilistic weather prediction with an analog ensemble. Mon. Weather Rev. 141, 3498–3516. http://dx.doi.org/10.1175/MWR-D-12-00281.1, URL: https://journals.ametsoc.org/view/journals/mwre/141/10/mwr-d-12-00281.1.xml.

NCEP, 2015. NCEP GFS 0.25 degree global forecast grids historical archive. http://dx.doi.org/10.5065/D65D8PWK, URL: https://rda.ucar.edu/datasets/dsd084001/.

Palmer, T.N., 2002. The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. Q. J. R. Meteorol. Soc. 128, 747–774. http://dx.doi.org/10.1256/0035900021643593, URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/0035900021643593.

Patton, A., Politis, D.N., and, H.W., 2009. Correction to automatic block-length selection for the dependent bootstrap by D. Politis and H. White. Econometric Rev. 28, 372–375. http://dx.doi.org/10.1080/07474930802459016.

Politis, D.N., H.W., 2004. Automatic block-length selection for the dependent bootstrap. Econometric Rev. 23, 53–70. http://dx.doi.org/10.1081/ETC-120028836.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155–1174. http://dx.doi.org/10.1175/MWR2906.1, URL: https://journals.ametsoc.org/view/journals/mwre/133/5/mwr2906.1.xml.

Rampal, N., Gibson, P.B., Sood, A., Stuart, S., Fauchereau, N.C., Brandolino, C., Noll, B., Meyers, T., 2022. High-resolution downscaling with interpretable deep learning: Rainfall extremes over New Zealand. Weather. Clim. Extrem. 38, 100525. http://dx.doi.org/10.1016/j.wace.2022.100525, URL: https://www.sciencedirect.com/science/article/pii/S2212094722001049.

Rasp, S., Lerch, S., 2018. Neural networks for postprocessing ensemble weather forecasts. Mon. Weather Rev. 146, 3885–3900. http://dx.doi.org/10.1175/MWR-D-18-0187.1, URL: https://journals.ametsoc.org/view/journals/mwre/146/11/mwr-d-18-0187.1.xml.

Roehrig, R., Beau, I., Saint-Martin, D., Alias, A., Decharme, B., Guérémy, J.F., Voldoire, A., Abdel-Lathif, A.Y., Bazile, E., Belamari, S., Blein, S., Bouniol, D., Bouteloup, Y., Cattiaux, J., Chauvin, F., Chevallier, M., Colin, J., Douville, H., Marquet, P., Michou, M., Nabat, P., Oudar, T., Peyrillé, P., Piriou, J.M., Salas y Mélia, D., Séférian, R., Sénési, S., 2020. The CNRM global atmosphere model ARPEGE-Climat 6.3: Description and evaluation. J. Adv. Model. Earth Syst. 12, e2020MS002075. http://dx.doi.org/10.1029/2020MS002075, URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002075.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. URL: https://arxiv.org/abs/1505.04597, arXiv:1505.04597.

Santanello, Jr., J.A., Dirmeyer, P.A., Ferguson, C.R., Findell, K.L., Tawfik, A.B., Berg, A., Ek, M., Gentine, P., Guillod, B.P., Van Heerwaarden, C., et al., 2018. Land–atmosphere interactions: The LoCo perspective. Bull. Am. Meteorol. Soc. 99, 1253–1272.

Scheuerer, M., Hamill, T.M., 2015. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. Mon. Weather Rev. 143, 4578–4596. http://dx.doi.org/10.1175/MWR-D-15-0061.1, URL: https://journals.ametsoc.org/view/journals/mwre/143/11/mwr-d-15-0061.1.xml.

Scheuerer, M., Switanek, M.B., Worsnop, R.P., Hamill, T.M., 2020. Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. Mon. Weather Rev. 148, 3489–3506. http://dx.doi.org/10.1175/MWR-D-20-0096.1, URL: https://journals.ametsoc.org/view/journals/mwre/148/8/mwrD200096.xml.

Shrestha, D.L., Robertson, D.E., Bennett, J.C., Wang, Q.J., 2015. Improving precipitation forecasts by generating ensembles through postprocessing. Mon. Weather Rev. 143, 3642–3663. http://dx.doi.org/10.1175/MWR-D-14-00329.1, URL: https://journals.ametsoc.org/view/journals/mwre/143/9/mwr-d-14-00329.1.xml.

Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Wang, W., Huang, X.Y., Duda, M., 2008. A Description of the Advanced Research WRF Version 3. Technical Report.

Suárez-Molina, D., Fernández-González, S., Montero, G., Oliver, A., González, J.C.S., 2021. Sensitivity analysis of the WRF model: Assessment of performance in high resolution simulations in complex terrain in the Canary Islands. Atmos. Res. 247, 105157. http://dx.doi.org/10.1016/j.atmosres.2020.105157, URL: https://www.sciencedirect.com/science/article/pii/S0169809520310930.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. URL: https://arxiv.org/abs/1703.01365, arXiv:1703.01365.

Vannitsem, S., Bremnes, J.B., Demaeyer, J., Evans, G.R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Bouallègue, Z.B., Bhend, J., Dabernig, M., Cruz, L.D., Hieta, L., Mestre, O., Moret, L., Plenković, I.O., Schmeits, M., Taillardat, M., den Bergh, J.V., Schaeybroeck, B.V., Whan, K., Ylhaisi, J., 2021. Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. Bull. Am. Meteorol. Soc. 102, E681 – E699. http://dx.doi.org/10.1175/BAMS-D-19-0308.1, URL: https://journals.ametsoc.org/view/journals/bams/102/3/BAMS-D-19-0308.1.xml.

Veldkamp, S., Whan, K., Dirksen, S., Schmeits, M., 2021. Statistical postprocessing of wind speed forecasts using convolutional neural networks. Mon. Weather Rev. 149, 1141–1152. http://dx.doi.org/10.1175/MWR-D-20-0219.1, URL: https://journals.ametsoc.org/view/journals/mwre/149/4/MWR-D-20-0219.1.xml.

Wang, Y., Wu, J., 2022. Overview of the application of orographic data in numerical weather prediction in complex orographic areas. Adv. Meteorol. 2022, 1279625.

Wilks, D.S., 1990. Maximum likelihood estimation for the gamma distribution using data containing zeros. J. Clim. 3, 1495–1501. http://dx.doi.org/10.1175/1520-0442(1990)003<1495:MLEFTG>2.0.CO;2.

Wilks, D., 2010. Sampling distributions of the Brier score and Brier skill score under serial dependence. Q. J. R. Meteorol. Soc. 136, 2109–2118. http://dx.doi.org/10.1002/qj.709.

Wilks, D., 2011. Chapter 8 - forecast verification. In: Wilks, D.S. (Ed.), Statistical Methods in the Atmospheric Sciences. In: International Geophysics, vol. 100, Academic Press, pp. 301–394. http://dx.doi.org/10.1016/B978-0-12-385022-5.00008-7.

Yang, Q., Wang, Q.J., Hakala, K., 2021. Achieving effective calibration of precipitation forecasts over a continental scale. J. Hydrol.: Reg. Stud. 35, 100818. http://dx.doi.org/10.1016/j.ejrh.2021.100818, URL: https://www.sciencedirect.com/science/article/pii/S2214581821000471.

Yao, W., Zeng, Z., Lian, C., Tang, H., 2018. Pixel-wise regression using u-net and its application on pansharpening. Neurocomputing 312, 364–371. http://dx.doi.org/10.1016/j.neucom.2018.05.103.

Zhong, X., Ma, Z., Yao, Y., Xu, L., Wu, Y., Wang, Z., 2023. WRF–ML v1.0: a bridge between WRF v4.3 and machine learning parameterizations and its application to atmospheric radiative transfer. Geosci. Model. Dev. 16, 199–209. http://dx.doi.org/10.5194/gmd-16-199-2023, URL: https://gmd.copernicus.org/articles/16/199/2023/.