Distributional Regression U-Nets for the Postprocessing of Precipitation Ensemble Forecasts

Romain Pic, Clément Dombry, Philippe Naveau, Maxime Taillardat

Distributional Regression U-Nets for the Postprocessing of Precipitation Ensemble Forecasts

Romain Pic[1], Clément Dombry[1], Philippe Naveau[2], and Maxime Taillardat[3]

[1]Université de Franche Comté, CNRS, LmB (UMR 6623), F-25000 Besançon, France
[2]Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, CEA-CNRS-UVSQ, EstimR, IPSL & U Paris-Saclay, Gif-sur-Yvette, France
[3]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

July 2, 2024

Abstract

Accurate precipitation forecasts have a high socio-economic value due to their role in decision-making in various fields such as transport networks and farming. We propose a global statistical postprocessing method for grid-based precipitation ensemble forecasts. This U-Net-based distributional regression method predicts marginal distributions in the form of parametric distributions inferred by scoring rule minimization. Distributional regression U-Nets are compared to state-of-the-art postprocessing methods for daily 21-h forecasts of 3-h accumulated precipitation over the South of France. Training data comes from the Météo-France weather model AROME-EPS and spans 3 years. A practical challenge appears when consistent data or reforecasts are not available.

Distributional regression U-Nets compete favorably with the raw ensemble. In terms of continuous ranked probability score, they reach a performance comparable to quantile regression forests (QRF). However, they are unable to provide calibrated forecasts in areas associated with high climatological precipitation. In terms of predictive power for heavy precipitation events, they outperform both QRF and semi-parametric QRF with tail extensions.

1 Introduction

Correctly forecasting precipitation is crucial for decision-making in various fields such as flood levels, transport networks, water resources and farming, among others (see, e.g., Olson et al. 1995). Moreover, high-impact events are expected to intensify in the future as a consequence of climate change (Planton et al., 2008). Numerical weather prediction (NWP) systems have been continuously improving to take into account uncertainty of the atmosphere and the limitations of their physical modeling (Bauer et al., 2015). NWP systems produce ensemble forecasts, consisting of multiple runs of deterministic scenarios with different parameters. Nonetheless, raw ensemble forecasts suffer from bias and underdispersion (see, e.g., Hamill and Colucci 1997; Bauer et al. 2015; Ben Bouallègue et al. 2016; Baran and Nemoda 2016). This phenomenon affects all NWP systems regardless of the weather service and of the variable of interest. Furthermore, the limited number of ensemble members coupled with underdispersion implies that raw ensemble forecasts may have a limited predictive power regarding extremes (Williams et al., 2013). In order to correct these systematic errors, it has become standard practice to use statistical postprocessing of ensemble prediction systems (EPS) in both research and operations.

A popular spatial statistical postprocessing strategy consists of separately postprocessing marginal distributions at each location and the spatial dependence structure. Numerous methods for postprocessing univariate marginals have been developed over the past two decades. There has been a rise in the number of machine learning based statistical postprocessing techniques as they provide a flexible framework enabling the modeling of complex relationships between the output of NWP models and the target variable. Moreover, they facilitate the use of a large number of predictors. These methods range from well-established statistical learning techniques, such as random forests (Taillardat et al., 2016) or gradient boosting (Messner et al., 2017), to neural networks or deep learning techniques, such as fully connected neural networks (Rasp and Lerch, 2018) and transformers (Ben Bouallègue et al., 2024). For

a thorough review of the existing statistical postprocessing techniques, readers may refer to Vannitsem et al. (2021) and Schulz and Lerch (2022a). Once calibrated univariate marginals are obtained, the spatial dependence structure may be needed by downstream applications. The spatial dependence structure can be obtained from the raw ensemble as done by ensemble copula coupling (ECC; Schefzik et al. 2013) and its variants (e.g., Ben Bouallègue et al. 2016) or from historical observations as done by Schaake shuffle (ScS; Clark et al. 2004). Alternatively, if raw ensembles or historical data do not model the spatial dependence sufficiently well, it can be postprocessed using adapted techniques (see, e.g., Schefzik and Möller 2018).

An alternative postprocessing strategy consists of direct postprocessing of raw ensemble members to obtain calibrated members. This can be achieved by postprocessing each member individually (Van Schaeybroeck and Vannitsem, 2014) or by using ensemble-agnostic methods (Ben Bouallègue et al., 2024).

In order to circumvent (potential) data scarcity, it is common to use parametric methods as they are usually less affected by smaller training datasets. The choice of a specific parametric distribution can be motivated by prior knowledge (or assumption) on the distribution of the variable of interest. Parametric methods can enable extrapolation beyond the range available in the training data, which is of interest to consider extreme events (see, e.g., Friederichs et al. 2018 and Taillardat et al. 2019). In particular, certain meteorological variables have a heavy-tailed distribution; thus, a parametric method can be used to ensure that postprocessed distributions will have an appropriate tail behavior (e.g., Lerch and Thorarinsdottir 2013).

Previous studies, such as Hemri et al. (2014) and Taillardat and Mestre (2020), have highlighted that all meteorological quantities do not represent the same difficulty in terms of postprocessing. Variables with heavy-tailed climatological distributions or variables with short-scale spatio-temporal dependence (e.g., rainfall or wind gusts) are more difficult to treat than light-tailed variables or spatially smooth variables (e.g., surface temperature or sea level pressure). In the same vein, Schulz and Lerch (2022a) states that "wind gusts are a challenging meteorological target variable as they are driven by small-scale processes and local occurrence, so that their predictability is limited even for numerical weather prediction (NWP) models run at convection-permitting resolutions."

NWP models produce forecasts on a grid that are of interest to downstream applications (Hamill, 2018, Section 7.3.2). However, consistent gridded data suited to postprocessing is computationally costly since reanalyses and reforecasts of gridded products are demanding in terms of both storage and computation. Numerous observation networks are station-based (e.g., temperature, wind speed, or pressure), but they vary in coverage and quality. When forecasts are required at nearby locations, spatial modeling procedures are required. Both station-based and grid-based approaches present benefits and drawbacks (Hamill, 2018, Section 7.3.2). No preference has reached a consensus for any variable, but Feldmann et al. (2019) shows that the relative improvement is greater for station-based 2-m temperature postprocessing when station-based observations are used. In the case of precipitation, observations can be measured by hybrid observations (gauge-adjusted radar images), allowing for improvement in the quality of gridded postprocessing.

As mentioned by Schulz and Lerch (2022a), one of the main challenges of postprocessing is to preserve the spatio-temporal information while optimally utilizing the whole available input data. This motivates the use of global statistical postprocessing models (e.g., a single model for multiple locations). Distributional regression networks (DRN; Rasp and Lerch 2018) use an embedding module to learn a representation of stations, allowing the model to learn from nearby and similar stations in order to preserve the spatial information of the data. When working with gridded data, a postprocessing method could benefit from taking into account this spatial structure of the data within its architecture. Convolutional neural networks (CNN) rely on the image-like structure of their input. Numerous CNN-based methods have been developed to perform postprocessing (see, e.g., Dai and Hemri 2021 and Lerch and Polsterer 2022). Here, we want the output of the statistical postprocessing method to be grid-based. U-Net (Ronneberger et al., 2015) architectures appear to be a natural solution to preserve the spatial structure of the data. U-Nets use a sequence of convolutional blocks to learn complex features and upscaling blocks to retrieve parameters of interest at the desired resolution. We propose a U-Net-based method to postprocess marginals at each grid point using predictors at nearby grid points for high-resolution precipitation ensemble forecasts.

The paper is organized as follows. Section 2 presents the dataset used in this study. In Section 3, three state-of-the-art methods composing the reference methods of this study, namely quantile regression

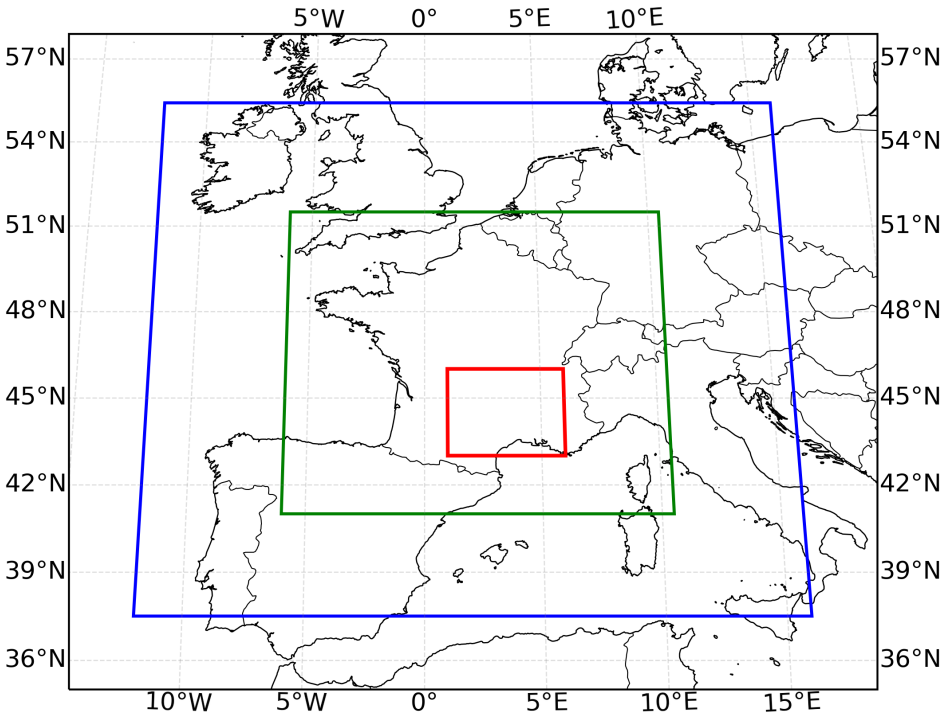


Figure 1: Domains covered by AROME-EPS (blue), ANTILOPE (green) and the region of interest (red).

forests (QRF), QRF with tail extension (TQRF) and DRN, are presented and compared based on their known benefits and limitations. A U-Net-based method, called distributional regression U-Nets (DRU), is introduced and compared with U-Net-based postprocessing methods in the literature. The predictive performance of the models is compared in terms of multiple univariate metrics in Section 4. An emphasis is put on the predictive performance of extremes. Finally, Section 5 sums up the performance of DRU and offers possible perspectives.

The code used to implement the different methods and their verification is publicly available[1].

2 Data

In this study, we focus on 3-h accumulated precipitation over the South of France (see Fig. 1) at a forecast lead time of 21-h initialized at 15:00UTC daily. Ensemble forecasts are taken from the 17-member limited area ensemble forecasting system AROME-EPS (Bouttier et al., 2015) driven by a subsampling of the global[2] PEARP ensemble. AROME-EPS produces ensembles with one control member and 16 perturbed members for forecasts up to 51 hours on four different initialization times. It produces a gridded ensemble over Western Europe with a horizontal resolution of 0.025° based on a model run at 1.3 km resolution. The probabilistic forecasts are compared to 3-h accumulated precipitation data obtained from the gauge-adjusted radar product ANTILOPE (Champeaux et al., 2009), which has a spatial resolution of 0.001° over Western Europe. We project observations of ANTILOPE onto the AROME-EPS grid using bilinear interpolation.

The region of interest in this study covers areas, such as the *Cévennes*, prone to heavy precipitation events (HPEs) (Ricard et al., 2012). HPEs affect Mediterranean coastal regions regularly causing flash floods. Mediterranean HPEs are typically characterized by quasi-stationary convective precipitation and

[1] https://github.com/pic-romain/unet-pp
[2] in the sense of globe-wide

may have limited predictability due to their intensity and being very local (Caumont et al., 2021). Statistical postprocessing methods can help improve forecasting such events.

Our period of interest spans 4 years from November 2019 to October 2023. The period from November 2019 to October 2022 is used as a training/validation dataset using 7-fold cross-validation to tune hyperparameters of the models. The folds are based on the day of the week. The period from November 2022 to October 2023 is used as a hold-out test set. All the results of Section 4 are provided for models trained on the entirety of the training/validation dataset and evaluated on the test dataset. The dataset is composed of forecasts and reforecasts from two different cycles of AROME-EPS. Consistency of both raw ensembles and observations is important since independent and identically distributed (i.i.d.) data is assumed. The two cycles of AROME-EPS used, namely 43t2 and 46t1, only have minor differences, making the i.i.d. assumption reasonable.

We use summary statistics of the AROME-EPS ensemble as predictors. The following variables were selected based on experts' opinions: precipitation, convective available potential energy, maximal reflectivity, pseudo wet-bulb potential temperature, relative humidity and AROME convection index. For each of these variables, the mean, the minimum, the maximum and the standard deviation of the raw ensemble were computed at each grid point and used as predictors.

In addition to summary statistics from AROME-EPS, distributional regression U-Nets (DRU) use constant fields carrying information about the topography and the type of terrain as predictors. The constant fields used are the altitude, a land-sea mask, the distance to sea and the first four components of a principal component analysis decomposition called AURHELY (Bénichou, 1994). Lerch and Polsterer (2022) showcased that the use of constant fields, such as altitude or orography, improves the performance of DRN. The first four components of AURHELY can be interpreted as local peak/depression, Northern/Southern slope, Eastern/Western slope and saddle effects, respectively. Figure 2 shows the seven constant fields used as predictors in DRU. Table 1 summarizes the predictors issued from both the raw ensemble and constant fields. Table 2 lists the dimensions of the dataset.

Type	Variable
Raw ensemble (mean, min, max, sd)	Precipitation
	Convective available potential energy
	Maximal reflectivity
	Pseudo wet-bulb potential temperature
	Relative humidity
	AROME convection index
Constant fields	Altitude
	Land-sea mask
	AURHELY components (1-4)
	Distance to sea

Table 1: List of weather and topographic variables used as predictors.

Variable	Value	Description
d	31	number of predictors
H	112	height (in grid points) of the region of interest (latitude)
W	192	width (in grid points) of the region of interest (longitude)
$n_{trainval}$	1091	# of days in the training/validation dataset
n_{test}	365	# of days in the test dataset

Table 2: Dimensions of the dataset used in this study.

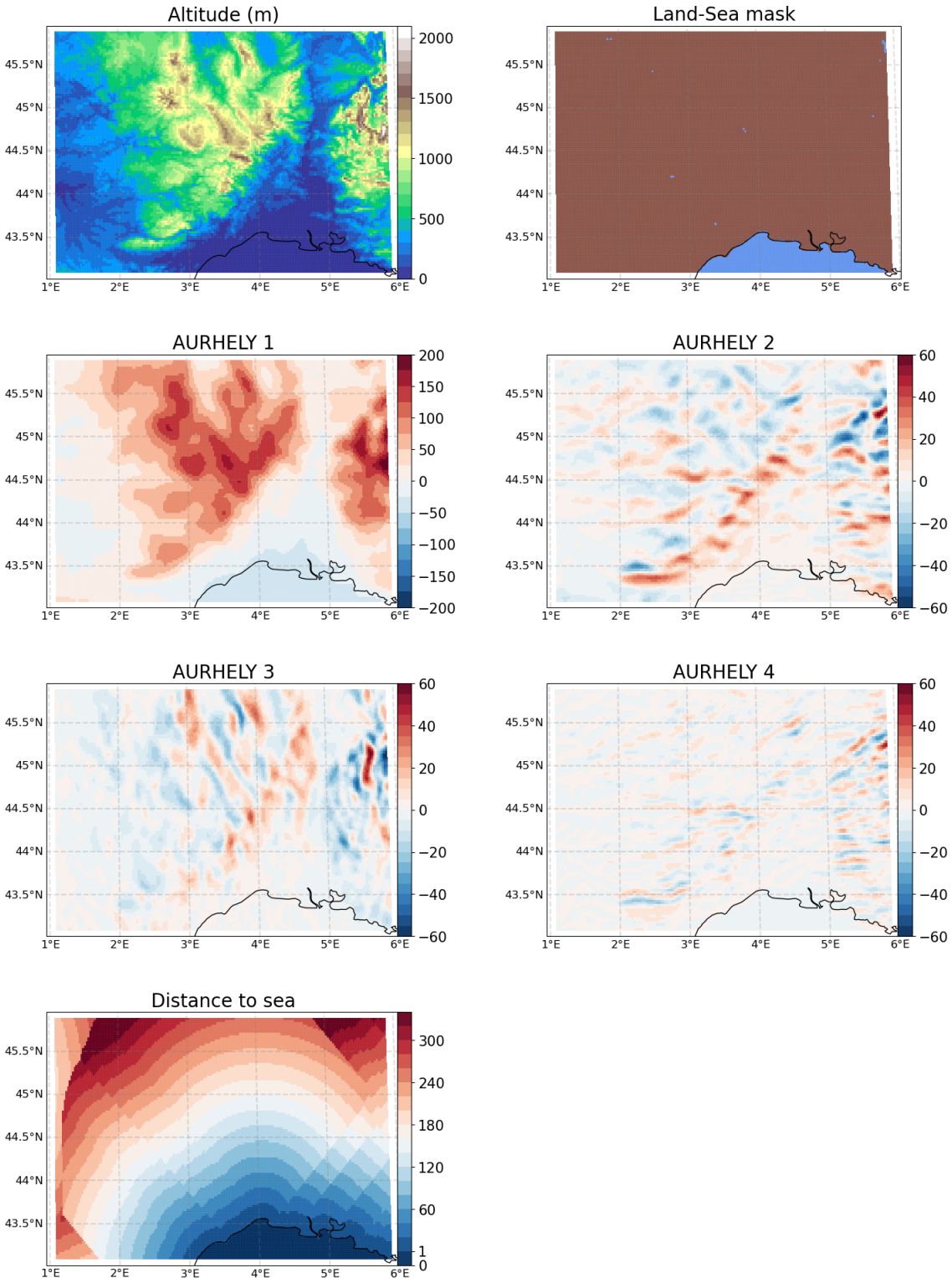


Figure 2: Constants fields used as predictors in distributional regression U-Nets: altitude, land-sea mask, four first components of AURHELY procedure, and distance to the sea.

3 Methods

We compare several postprocessing methods for the marginal distributions of gridded spatial ensemble forecasts of 3-h accumulated precipitation over the South of France. In a complete postprocessing scheme used operationally, the multivariate dependencies can then be retrieved using ECC or ScS, for example. We compare our U-Net-based distributional regression method to two benchmark methods: quantile

regression forest (QRF; Taillardat et al. 2016) and QRF with tail extension (TQRF; Taillardat et al. 2019). The performance of postprocessed forecasts using these different methods will be compared to the performance of the raw ensemble. Additionally, we recall distributional regression networks (DRN; Rasp and Lerch 2018) since our method can be seen as an extension of this approach.

These methods differ in their degree of reliance on parametric distributions (nonparametric, semi-parametric and parametric), in the fact of being local (i.e., a different model for each grid point) or global (i.e., a single model for the whole grid). Among global methods, differences lie in the representation of the spatial structure of the data. We briefly present the benchmark techniques and their limitations.

3.1 Quantile regression forests (QRF)

Quantile regression forests (QRF; Meinshausen 2006) is a nonparametric method able to predict conditional quantiles or, more generally, a conditional distribution. The method is based on random forests (Breiman, 2001). Similarly, it uses the data in terminal nodes (i.e., leaves) to compute a weighted average of empirical distributions. QRFs have proven their performance for postprocessing of wind speed and temperature forecasts (Taillardat et al., 2016) and for precipitation forecasts (Whan and Schmeits, 2018; van Straaten et al., 2018). QRFs can outperform complex postprocessing methods, such as neural network (NN-)based methods, at specific locations due to their local adaptability (Rasp and Lerch, 2018; Schulz and Lerch, 2022a). Moreover, QRF is used operationally as a postprocessing method at Météo-France (Taillardat and Mestre, 2020). This, as well as its overall performance, makes it a relevant benchmark method for this study.

QRFs are known to have three main limitations: potential spatial inconsistency, storage memory voracity (Taillardat and Mestre, 2020) and inability to extrapolate. The fact that QRF is a local model (i.e., a different model is used for each location, lead time, and variable) may cause problems. There is no guarantee that the output of the models is consistent spatially or temporally. Additionally, QRFs need to store the construction parameters (such as variables and thresholds of splits) of each tree of the forest and the samples used for training. This latter limitation results in the need to store a large number of parameters (especially when working with gridded data) to perform postprocessing. Lastly, QRF is incapable of extrapolating as its output is a weighted average of the training samples and does not provide a model for the distribution tail.

3.2 Quantile regression forest with tail extension (TQRF)

In order to circumvent the extrapolation inability of QRF, semi-parametric methods based on a combination of parametric modeling and random forest were proposed. Schlosser et al. (2019) introduced distributional regression forests using maximum likelihood to infer the parameters of a censored Gaussian distribution. Taillardat et al. (2019) proposed a method using probability-weighted moments (Diebolt et al., 2007) on the output of QRF to infer the parameters of an extended generalized Pareto distribution (EGPD; Naveau et al. 2016). The EGPD is a flexible parametric class of distributions able to jointly model the whole range of the distribution while in alignment with extreme value theory, without the requirement of threshold selection. The methods proposed in Schlosser et al. (2019), Taillardat et al. (2019) and, more recently, Muschinski et al. (2023) can all be adapted to any suitable parametric distribution. We choose to use the semi-parametric method of Taillardat et al. (2019) based on probability-weighted moments inference.

Our implementation of TQRF differs from the original method described in Taillardat et al. (2019). It uses refinements that have proven to be useful in operational settings: the tail extension is only activated if the QRF forecast assigns a large enough probability of exceedance of certain levels of interest, and in that case, only the quantiles that are higher for the fitted distribution than in the output of the QRF are updated. Moreover, we did not use EGPD because, while the QRF+EGPD is robust and efficient, the minimization of its continuous ranked probability score (CRPS; Matheson and Winkler 1976) for parameter inference is not direct due to its complex form (Taillardat et al., 2019, 2022). These implementation issues could, for example, be circumvented by using Monte-Carlo sampling to estimate the CRPS or by fixing the tail parameter to its climatological value.

Instead of the EGPD, the generalized truncated/censored normal distribution (GTCND; Jordan et al. 2019) and the censored-shifted gamma distribution (CSGD; Scheuerer and Hamill 2015) are used as tail

extensions of the QRF and as parametric distributions for DRU. The GTCND used here has a lower endpoint equal to 0 and no upper endpoint and its cumulative distribution function (cdf) is defined as

$$F^{\text{gtcnd}}_{L,\mu,\sigma}(z) = \begin{cases} L + \frac{1-L}{1-\Phi(-\mu/\sigma)}\left(\Phi(\frac{z-\mu}{\sigma}) - \Phi(-\mu/\sigma)\right) & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases},$$

where $0 \leq L \leq 1$ is the probability of a dry event (i.e., absence of precipitation), Φ is the cdf of the standard normal distribution, $\mu \in \mathbb{R}$ is the location parameter of the truncated normal distribution and $\sigma > 0$ is its scale parameter. The cdf of the CSGD is defined as

$$F^{\text{csgd}}_{k,\theta,\delta}(z) = \begin{cases} G_k(\frac{z-\delta}{\theta}) & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases},$$

where G_k is the cdf of the gamma distribution of shape $k > 0$, θ is the scale parameter and $\delta < 0$ is a shift parameter. The probability of dry events has a point mass of $G_k(-\delta/\theta)$. These distributions are both suited to the forecast of precipitation since they have point masses in 0 and take positive values. Moreover, the CSGD can reflect the variations of skewness observed in precipitation distributions (Scheuerer and Hamill, 2015). Details on the moments method for GTCND and CSGD, as well as CRPS formulas, are provided in Appendix A and Appendix B.

We denote QRF+*distrib* the TQRF method where *distrib* is the name of the parametric distribution family. The QRF+EGPD method is used operationally for rainfall postprocessing at Météo-France (Taillardat and Mestre, 2020). Nonetheless, this semi-parametric method remains local and thus also suffers from both potential spatial inconsistency and memory voracity (Taillardat and Mestre, 2020). To bypass these limitations, methods need to be global (i.e., use one model for all locations) while staying efficient locally.

3.3 Distributional regression networks (DRN)

Rasp and Lerch (2018) proposed distributional regression networks (DRN), a NN-based approach to postprocess 2-m temperature forecasts. DRN is a global model predicting the parameters of a distribution of interest. It leverages the flexibility of NN to model the dependency of parameters on the covariables (used as input of DRN). DRN can be seen as an extension of EMOS (Gneiting et al., 2005), which itself fits a parametric distribution where the parameters linearly depend on summary statistics of the raw ensemble. DRN is a global model thanks to the presence of an embedding module within its architecture, allowing the network to learn location-specific parameters and to benefit from data at similar locations. DRN learns the embedding and parameters of a dense NN by minimizing a strictly proper scoring rule (Gneiting and Katzfuss, 2014) such as the CRPS.

Rasp and Lerch (2018) and Schulz and Lerch (2022a) have shown that DRN outperforms other state-of-the-art methods in most stations over Germany for the postprocessing of temperature and wind gusts, respectively. Moreover, Schulz and Lerch (2022a) studied other NN-based postprocessing techniques, namely Bernstein quantile network (BQN; Bremnes 2020) and histogram estimation network (HEN; see, e.g., Scheuerer et al. 2020 and Veldkamp et al. 2021). BQN and HEN are nonparametric approaches where NNs learn the coefficient of Bernstein polynomials to predict a quantile function and probabilities of bins to predict a probability density function (pdf), respectively. At particular stations, BQN outperforms other postprocessing techniques, including DRN, for wind gust forecasts.

In spite of being a global model, the architecture of DRN makes it ill-suited to gridded data. Its architecture does not use knowledge of the spatial structure of the points and thus has to try to learn it through its embedding module. Moreover, DRN only uses information available at the location of interest as predictors. Convolutional neural network (CNN)-based architectures make use of the gridded structure of the data and can use the information at neighboring locations as a predictor. Lerch and Polsterer (2022) studied a modified DRN architecture using the representation of global fields from a convolutional auto-encoder as predictors and showed an improvement in skill compared to regular DRN.

DRNs' architecture makes their implementation on gridded data very costly. They need to flatten the data across locations (i.e., reshape it into a 1D vector), and they cannot benefit from GPU computing. For these reasons and their impact on the search for optimal hyperparameters, DRNs are not used as a benchmark method in this study.

3.4 Distributional regression U-Nets (DRU)

Convolutional blocks are the main ingredient of CNN-based architectures. The simplest convolutional blocks are composed of a convolutional layer and a max-pooling layer. The role of the convolutional layer is to learn kernels able to extract useful features from the input of the convolutional block. The max-pooling layer reduces the resolution of the features, allowing the following layers to work at broader scales. The succession of convolutional blocks allows CNNs to learn patterns at different spatial scales and to learn complex patterns (see, e.g., Simonyan and Zisserman 2014). CNN-based architectures have been used in numerous postprocessing studies (e.g., Dai and Hemri 2021; Veldkamp et al. 2021; Li et al. 2022; Chapman et al. 2022; Lerch and Polsterer 2022).

Since we are interested in global models using the data's gridded structure and want the output to be the distributional parameters of marginals on the same grid, we use a U-Net architecture (Ronneberger et al., 2015). The U-Net architecture was initially designed for images but is compatible with gridded data to obtain a grid-based output. It has been used for various postprocessing applications. Grönquist et al. (2021) used it in a bias/uncertainty postprocessing scheme of temperature and geopotential forecasts. Dai and Hemri (2021) used a U-Net as a generator within a conditional generative adversarial network (cGAN) for cloud cover postprocessing. Hu et al. (2023) used U-Nets to predict the parameters of a CSGD corresponding to the postprocessed daily precipitation given a deterministic forecast. Horat and Lerch (2023) used U-Nets to perform postprocessing of temperature and precipitation at the sub-seasonal to seasonal scale. The task is a three-level classification problem with below-normal, near-normal and above-normal conditions as classes. Ben Bouallègue et al. (2024) used transformers within a U-Net architecture to postprocess ensemble members directly with temperature and precipitation as variables of interest.

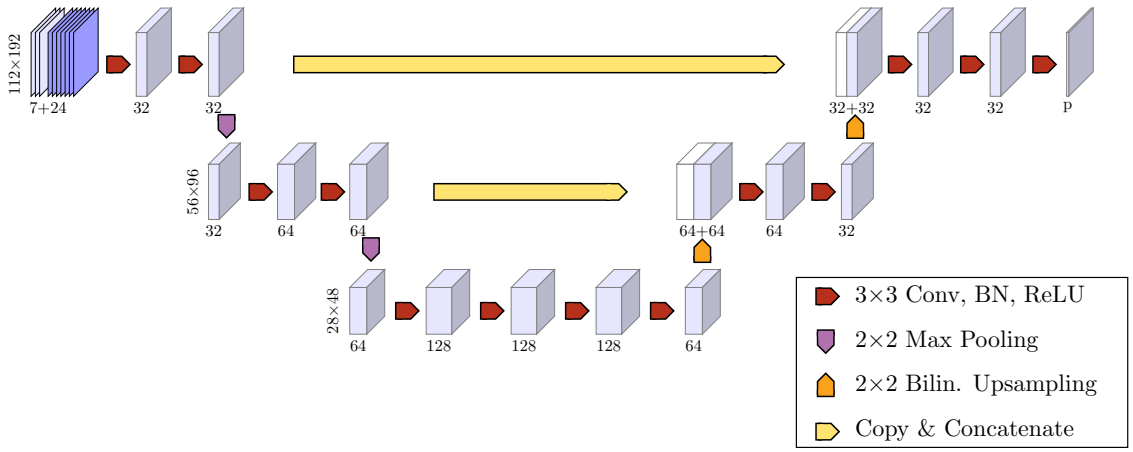


Figure 3: Architecture of distributional regression U-Nets. *Conv* stands for convolution, *BN* stands for batch normalization, *ReLU* stands for rectified linear unit and *Bilin. Upsampling* stands for bilinear upsampling. p is the number of distribution parameters: for GTCND and CSGD, $p = 3$.

The U-Net architecture used in this work is presented in Figure 3. The U-Net input is a concatenation of constant fields and summary statistics of the ensemble members. The output is the parameters of the postprocessed marginal distribution at each grid point (i.e., parameters of a GTCND or a CSGD). The architecture can be decomposed into two parts. On the left part, the succession of specific convolutional blocks (red and purple arrows) leads to an increase in the number of features and a reduction of the spatial dimension (i.e., a coarsening of the spatial resolution) as the data progresses through the network. As explained above, the convolutional blocks are constructed in order to learn useful representations of the features of the fields at various spatial scales. On the right part, upscaling blocks (red and orange arrows), based on bilinear upsampling, use the features learned in the central part of the architecture to predict features at finer resolutions and finally learn the parameters of the distribution selected. Additionally, we use skip-connections (yellow arrows), consisting of copying and concatenating features, as bridges between the left and right parts of the U-Net. Skip-connections have proven to improve

the stability of the convergence of NN (see, e.g., Li et al. 2018). This U-Net-based method is a global model enabling extrapolation through a parametric distribution (e.g., GTCND or CSGD). We denote U-Net+*distrib* the distributional regression U-Net (DRU) where *distrib* is the parametric distribution.

DRU learns to predict the parameters of a distribution by minimizing the CRPS at each grid point. Both the parameterized distribution and the scoring rule to minimize can be chosen to be suited to the variable of interest or to facilitate computations, thus making the architecture flexible. The convolution blocks allow the parameters of marginal distribution to be learned from neighboring grid points, potentially accounting for dependencies between grid points (Schefzik and Möller, 2018, Section 4.5). Moreover, the use of constant fields as input enables the convolutional layers to learn representations of these fields that are relevant to the postprocessing task at hand. This can be seen as a natural extension of the embedding module in DRN (Rasp and Lerch, 2018).

DRNs are built to bypass the limitations of the methods presented above. The model is global and uses the predictor fields of the whole grid, this construction enables the predicted marginals to be spatially consistent. Moreover, the use of convolutional layers facilitates the learning of relevant spatial features compared to DRN. Memory voracity is not an issue as the model is global and the number of parameters is contained. Finally, as highlighted previously, any parameterized distribution can be used as the output of DRU accounting for extrapolation and relevance to the target variable at hand. Table 3 summarizes the characteristics of the postprocessing methods studied in this article.

The U-Net-based method of this article is related to the one of Hu et al. (2023) in the sense that both approaches use U-Nets to predict the parameters of a distribution corresponding to the marginals of the variable of interest. The main differences between the approaches are the following: they studied daily precipitation accumulations, where we are interested in 3-h accumulated precipitation; they postprocess deterministic forecasts, where we postprocess ensemble forecasts; and finally, we use constant fields as additional predictors. Moreover, in terms of the number of years in the training data, our work (with only 3 years of training data) falls in a "gray area" where their U-Net-based method is outperformed by analog ensemble (Delle Monache et al., 2013), which is a simpler approach (Hu et al., 2023, Figure 11). Table 4 summarizes the characteristics of the different U-Net-based postprocessing methods available.

	QRF	TQRF	DRN	DRU
Local/Global	local	local	global	global
Principles	grid point per grid point	grid point per grid point	embedding to learn from similar stations	constant fields and architecture aware of the gridded structure
Ability to extrapolate	✗	✓	✓	✓
Number of parameters	∼15.3 B	∼15.3 B	∼450,000	∼1,000,000
Storage necessary for prediction	splits of each tree and training data	splits of each tree and training data	parameters and architecture	parameters and architecture

Table 3: Comparison of the postprocessing methods mentioned in this study. The number of parameters is provided for hyperparameters selected by cross-validation on the training/validation data set and for the setup described in Section 2 (e.g., a 112×192 grid). In the case of DRN, an architecture similar to the one in Rasp and Lerch (2018) has been considered. *B* stands for billion.

The following hyperparameters of the U-Net architecture have been selected using the training/validation dataset: the learning rate, the batch size and the number of epochs. The optimizer is Adam with default parameters (except for the learning rate) from its `Keras` implementation. In order to limit the number of parameters and prevent overfitting, the depth of the U-Net is kept at two levels (as shown in Fig. 3) and

	Grönquist et al. (2021)	Dai and Hemri (2021)	Horat and Lerch (2023)	Ben Bouallègue et al. (2024)	Hu et al. (2023)	Pic et al. (2024)
Variable of interest	temperature, geopotential	cloud cover	temperature, 2-w precip.	temperature, 6-h precip.	24-h precip.	3-h precip.
Output	bias	samples from a cGAN	probability of classes	postprocessed ensemble members	parameters of a CSGD	parameters of a GTCND/CSGD
Lead times	48h	1-120h	2-4w	6-96h	0-4d	21h
Dataset: raw forecast	ECMWF-ENS10 *ensemble*	COSMO-E, ECMWF-IFS *ensemble*	ECMWF-IFS (S2S) *ensemble*	ECMWF-IFS *ensemble*	West-WRF *deterministic*	AROME-EPS *ensemble*
Dataset: obs.	ERA5	EUMETSAT	NOAA-CPC	ERA5	PRISM	ANTILOPE
Resolution	0.5°	0.02°	1.5°	1°	0.04°	0.025°
Training data range	17 years	3 years	20 years	19 years	2-30 years	3 years

Table 4: Comparison of the postprocessing methods relying on U-Nets.

separable convolutions were used instead of standard ones. Moreover, in order to contain the variability due to random initialization, we aggregate forecast distributions of 10 models as recommended in Schulz and Lerch (2022b).

Most of the implementation was conducted in `Python` and the implementation of DRU is based on `Tensorflow` (Abadi et al., 2015) and `Keras` (Chollet et al., 2015). QRF and TQRF are implemented in `R` (R Core Team, 2023) using the `ranger` package (Wright and Ziegler, 2017).

4 Results

We provide a comparison of DRU to QRF, TQRF and the raw ensemble using verification tools targeting three different aspects of forecasts: verification of the overall performance with the CRPS, calibration and extreme events. First, we compare the performance of the postprocessing techniques in terms of their relative improvement compared to the raw ensemble and among themselves. This improvement is quantified in terms of continuous ranked probability skill score (CRPSS). Second, we assess the calibration of the postprocessed forecasts using rank histograms. Finally, the improvement of the postprocessing methods in terms of extreme forecasting is evaluated using receiver operating characteristic (ROC) curves for events corresponding to the exceedance of various thresholds.

4.1 Continuous ranked probability score

Since the postprocessing techniques considered act on the 1-dimensional marginals, the improvement and comparison of the postprocessing techniques can be done with univariate scoring rules. The continuous ranked probability score (CRPS; Matheson and Winkler 1976) is one of the most popular univariate scoring rules in weather forecasting and is defined as

$$\begin{aligned}\mathrm{CRPS}(F, y) &= \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y \leq z})^2 \mathrm{d}z; && (1)\\ &= 2 \int_0^1 (\mathbb{1}_{y \leq F^{-1}(\alpha)} - \alpha)(F^{-1}(\alpha) - y) \mathrm{d}\alpha; && (2)\\ &= \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'|, && (3)\end{aligned}$$

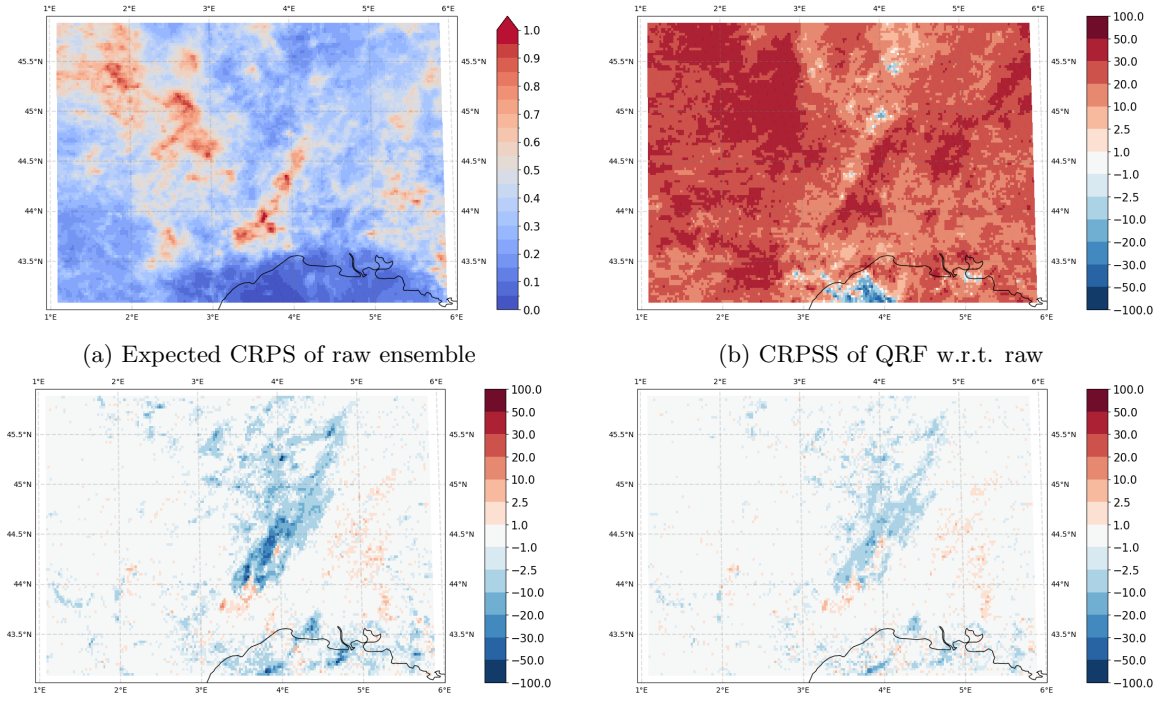
where the forecast F is assimilated to its cdf, F^{-1} is its quantile function and X and X' follow the distribution F. The CRPS is strictly proper on the set of measures with a finite first moment. Moreover,

it benefits from multiple representations that help both its computation and interpretation. Equation (1) is the threshold or Brier score (Brier, 1950) representation and expresses the CRPS as the integrated squared error between the cdf of the forecast and the empirical cdf associated with observation y over all thresholds z. Equation (2) is the quantile representation and shows that the CRPS is expressed as the pinball loss over all quantile levels α. Equation (3) is the kernel representation and is particularly useful to compute the score of ensemble forecasts. The CRPS formulas for the parametric distributions of this article are available in the Appendix A and B. For the raw ensemble, QRF and TQRF forecasts, the CRPS has been estimated using the fair estimator (Ferro, 2013).

When working with (strictly) proper scoring rules to compare forecasts, the comparison of the scoring rules of two forecasts can be summarized by the skill score. For a proper scoring rule S, the skill score of a forecast F with respect to (w.r.t.) a reference forecast F_{ref} is defined as

$$\text{SS}(F, F_{\text{ref}}) = \frac{\mathbb{E}_G[\text{S}(F_{\text{ref}}, Y)] - \mathbb{E}_G[\text{S}(F, Y)]}{\mathbb{E}_G[\text{S}(F_{\text{ref}}, Y)]}, \tag{4}$$

where G is the distribution of the observations and $\mathbb{E}_G[\cdots]$ is the expectation with respect to $Y \sim G$. The skill score is positive if the forecast F improves the expected score w.r.t. the reference forecast F_{ref} and negative otherwise. The skill score can be expressed in percentage. In the context of postprocessing, a reference of choice is the raw ensemble that the postprocessing procedure aims to improve upon.



(a) Expected CRPS of raw ensemble

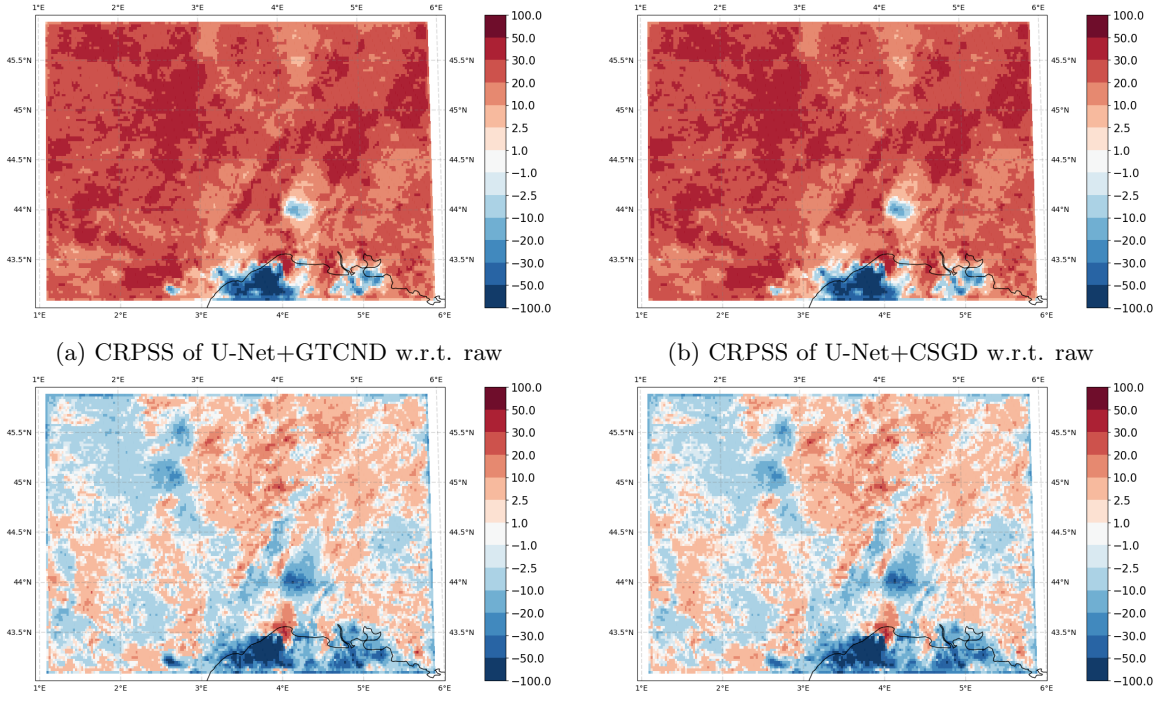
(b) CRPSS of QRF w.r.t. raw

(c) CRPSS of QRF+GTCND w.r.t. QRF

(d) CRPSS of QRF+CSGD w.r.t. QRF

Figure 4: Predictive performance of the benchmark methods in terms of CRPS. (a) Expected CRPS of the raw ensemble, (b) CRPSS of QRF w.r.t. the raw ensemble and CRPSS w.r.t. QRF of (c) QRF+GTCND and (d) QRF+CSGD.

We compared the continuous ranked probability skill score (CRPSS) for the different postprocessing methods studied w.r.t. other benchmark methods. Figure 4 shows the expected CRPS of the raw ensemble, the CRPSS of QRF w.r.t. the raw ensemble and the CRPSS of QRF+GTCND and QRF+CSGD w.r.t. QRF. The raw ensemble has an expected CRPS of 0.3725 mm when averaged over the whole region of interest. However, the expected CRPS greatly fluctuates over the whole grid and most grid points of higher altitude have larger expected CRPS since they correspond to higher precipitation accumulations (see Fig. 4a). The lowest expected CRPS values are located over the Mediterranean Sea corresponding to an area of low precipitation as discussed further (see Fig. 6). Moreover, observations in this area are of lower quality since it is far from the nearest radar and cannot be corrected by gauges.



(a) CRPSS of U-Net+GTCND w.r.t. raw

(b) CRPSS of U-Net+CSGD w.r.t. raw

(c) CRPSS of U-Net+GTCND w.r.t. QRF

(d) CRPSS of U-Net+CSGD w.r.t. QRF

Figure 5: Predictive performance of the distributional regression U-Nets in terms of CRPS. CRPSS w.r.t. the raw ensemble of (a) U-Net+GTCND and (b) U-Net+CSGD and CRPSS w.r.t. QRF of (c) U-Net+GTCND and (d) U-Net+CSGD.

Figure 4b confirms that QRF is able to improve the predictive performance in terms of CRPSS compared to the raw ensemble (23.51% after averaging over the region of interest). The CRPSS of QRF w.r.t. the raw ensemble is positive (i.e., improvement of skill) over the whole domain except for some localized regions. In particular, over the area that has the lowest expected CRPS for the raw ensemble, QRF is not able to improve compared to raw ensemble in terms of expected CRPS. This may be caused by the fact that this area is already well-predicted by the raw ensemble and the QRF is not able to improve its CRPS. Figures 4c and 4d show the CRPSS w.r.t. QRF of QRF+GTCND and QRF+CSGD, respectively. Overall, QRF+GTCND and QRF+CSGD have a close but slightly smaller expected CRPS than that of QRF (average CRPSS w.r.t. QRF of -1.04% and -0.33%, respectively). For both GTCND and CSGD tail extensions, the areas of lower skill (in blue) are located in a mountainous region (the Eastern part of Massif Central) and near the Mediterranean coast. Nonetheless, the areas are wider and have lower CRPSS values for QRF+GTCND compared to QRF+CSGD. Both methods also present areas of improvement of CRPSS (in orange/red) that are sparser and smaller than the areas of negative CRPSS.

Figure 5 provides the CRPSS of U-Net+GTCND and U-Net+CSGD w.r.t. the raw ensemble and QRF. Figures 5a and 5b show the CRPSS of DRU w.r.t. the raw ensemble. Both GTCND and CSGD lead to methods improving CRPSS w.r.t. the raw ensemble with 22.28% and 22.36%, respectively, when averaged over the region of interest. As the QRF, DRU leads to improvement in terms of CRPSS over the vast majority of grid points. Nonetheless, there are areas where they have a poorer predictive performance compared to raw ensemble. These areas are also located over the Mediterranean Sea or near the coast, and one patch is located in the Rhône River valley. When censoring grid points located over the sea and at the border, the average CRPSS w.r.t. the raw ensemble is 24.34% and 24.48% for U-Net+GTCND and U-Net+CSGD, respectively.

Figures 5c and 5d show the CRPSS of U-Net+GTCND and U-Net+CSGD w.r.t. QRF. Overall, DRU has a higher expected CRPS than QRF (CRPSS of -1.52% for the U-Net+GTCND and -1.37% for the U-Net+CSGD), but it has an improved predictive performance (in terms of CRPS) over a non-negligible part of the region of interest. Due to their architecture, DRUs are affected by a border effect, leading to a less predictive performance on the grid points located at the boundaries of the grid (see Fig. 5c and Fig. 5d). Using the censoring mentioned above, U-Net+GTCND and U-Net+CSGD have an average

		Reference			
		Full region		Censored region	
		Raw ensemble	QRF	Raw ensemble	QRF
Postprocessing methods	QRF	**23.51%**	–	23.56%	–
	QRF+GTCND	22.67%	-1.04%	22.72%	-1.05%
	QRF+CSGD	23.23%	-0.33%	23.29%	-0.34%
	U-Net+GTCND	22.25%	-1.52%	24.34%	0.05%
	U-Net+CSGD	22.36%	-1.37%	**24.48%**	**0.26%**

Table 5: Summary of the performance in terms of CRPSS averaged over the full region of interest and over the censored one.

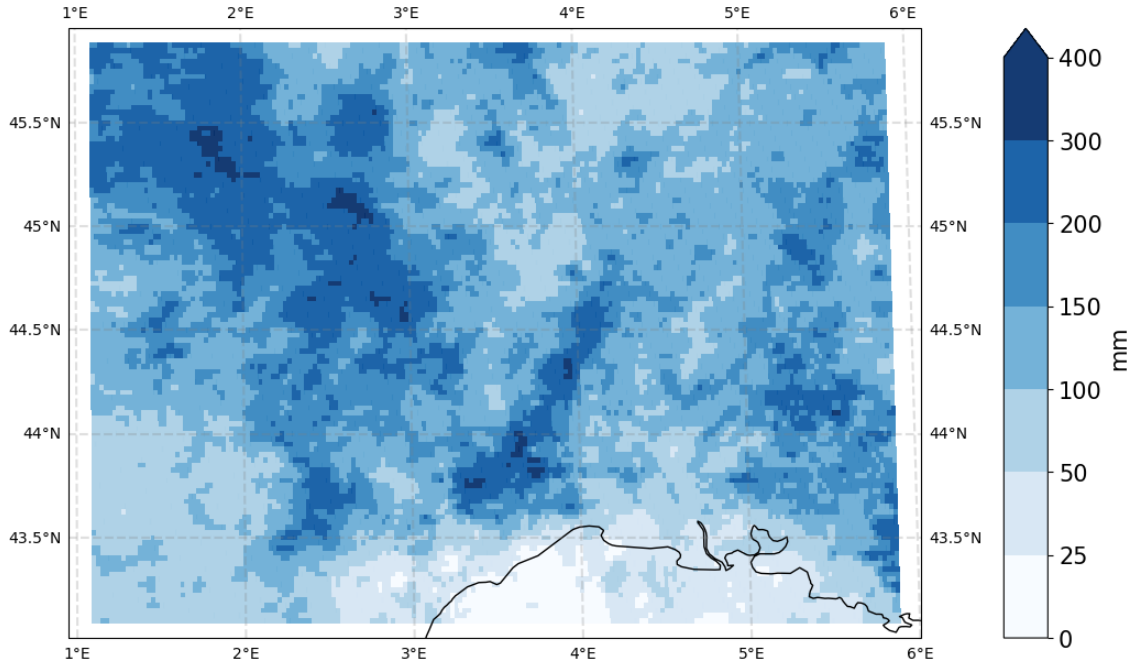
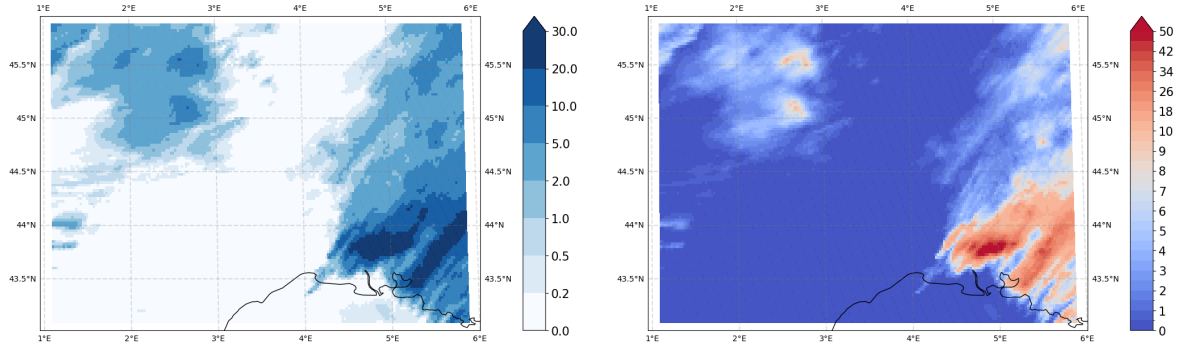


Figure 6: Total precipitation over the test set. Due to the initial time and the lead time considered, only precipitation between 12:00UTC and 15:00UTC are taken into account.

CRPSS w.r.t. QRF of 0.05% and 0.26%, respectively. Table 5 summarizes the comparisons of methods in terms of CRPSS.

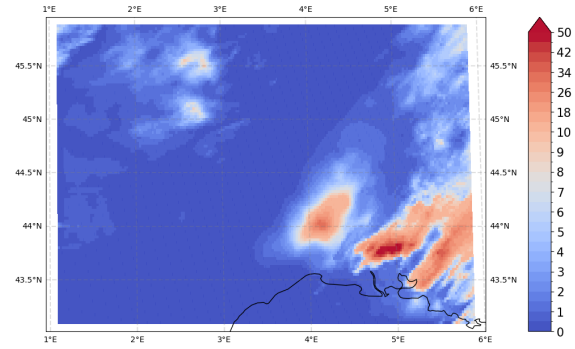
For the training/validation dataset, DRUs are prone to numerical instabilities. This led to areas of negative CRPSS w.r.t. the raw ensemble caused by the divergence of predicted parameters (σ is the case of U-Net+GTCND and θ in the case of U-Net+CSGD) (not shown). In addition to standard numerical stabilizing tricks, we have tried to constrain the range of diverging parameters using the value of the climatological fits since higher values would lead to forecasts less informative than the climatological forecasts. This solved the divergence issues over both the training/validation and test datasets for U-Net+CSGD but not for U-Net+GTCND (not shown). However, it increased the border effects causing deteriorating performance for both models. Hence, the constraining of the range of the parameters for DRU method is not used and the numerical stability of the methods needs to be understood and prevented.

Despite being prone to numerical instabilities, the areas of negative CRPSS w.r.t. the raw ensemble for the test dataset are not all caused by numerical instabilities. The largest area of negative CRPSS

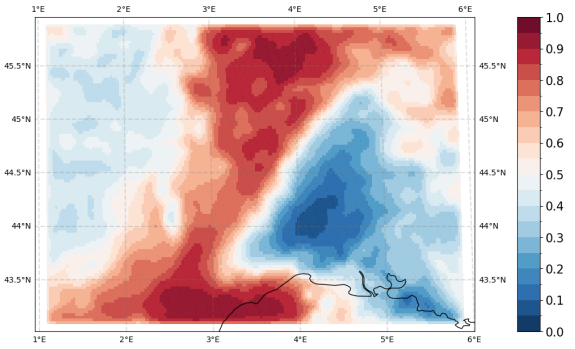


(a) Precipitation (mm)

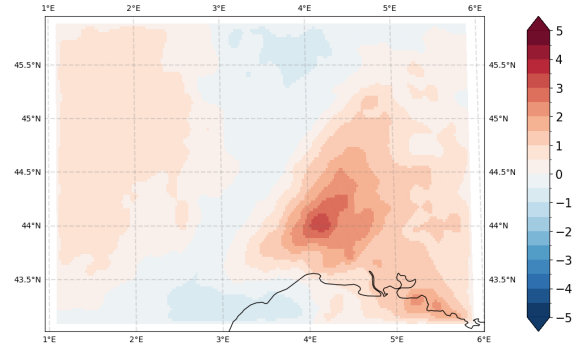
(b) CRPS of raw ensemble



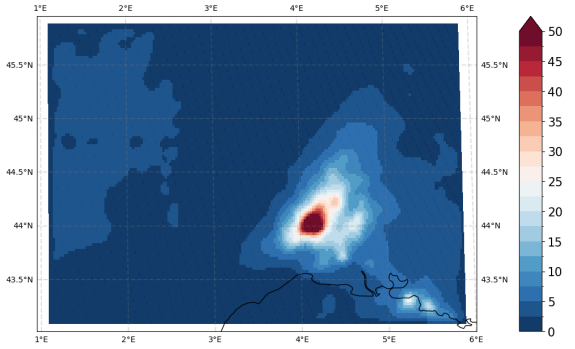
(c) CRPS of U-Net+GTCND



(d) Predicted L



(e) Predicted μ



(f) Predicted σ

Figure 7: Example of a numerical instability of U-Net+GTCND for a forecast valid on November 3, 2022 at 15 :00UTC. Note the different scales for CRPS below and above 10 mm.

w.r.t. raw (see Fig. 5a and 5b) coincide with the area with the lowest total precipitation over the test period (see Fig. 6). This area matches the area of the lowest expected CRPS for the raw ensemble (see Fig. 4a). Numerous dry events occur at this location and are perfectly predicted by the raw ensemble (i.e., all members predict 0 mm of precipitation). However, in order to perfectly predict a dry event, U-Net+GTCND and U-Net+CSGD need to predict $L = 1$ and $-\delta/\theta = \infty$, respectively, which is never the case in practice. This may explain why the CRPSS w.r.t. the raw ensemble of this area is highly negative for DRU. The CRPS of QRF (and TQRF) has been computed using 107 quantiles, rendering perfect prediction of dry events harder and resulting in a deterioration in terms of CRPS over the aforementioned area (see Fig. 4b).

The other smaller areas of negative CRPSS w.r.t. the raw ensemble for DRU seem to be caused by numerical instabilities. For example, Figure 7 presents a numerical instability for a U-Net+GTCND forecast valid on November 3, 2022 at 12:00UTC. It corresponds to heavy precipitation over the Easter part of the region of interest (see Fig. 7a). Both raw ensemble and U-Net+GTCND seem not able to correctly predict heavy precipitation, as reflected in the high values of their CRPS (see Fig. 7b and 7c). However, the CRPS of U-Net+GTCND presents an additional area of high CRPS that is caused by the prediction of precipitation where no precipitation has been observed. This incorrect prediction

is characterized by a low value of L (i.e., low probability of dry event), a positive value of μ and a very high value of σ (see Fig. 7d, 7e and 7f). The abnormally large value of σ seems to be caused by a numerical instability and gives a larger probability to large precipitation. The high CRPS over this region associated with a low value of CRPS for raw ensemble causes the CRPSS for U-Net+GTCND w.r.t. the raw ensemble over the test set to be negative (see Fig. 5a).

DRUs are able to reach a predictive performance slightly lower but comparable to the QRF. U-Net+CSGD has a slightly better expected CRPS than U-Net+GTCND. In order to be deemed worthy postprocessing methods, U-Net+GTCND and U-Net+CSGD need to be calibrated.

4.2 Calibration

Since the ideal forecast (i.e., the true conditional distribution) is unknown, it is impossible to know if a postprocessed forecast has reached the minimum expected CRPS. In order to decompose the contribution of calibration and sharpness to scoring rules (Winkler, 1977; Winkler et al., 1996), rank histograms are used to evaluate the calibration of the different postprocessing techniques.

Multiple definitions of calibration exist with different levels of hypotheses (see, e.g., Tsyplakov 2013, 2020). The most used definition is probabilistic calibration which, broadly speaking, consists of computing the rank of observations among samples of the forecast and checking for uniformity with respect to observations. If the forecast is calibrated, observations should not be distinguishable from forecast samples, and thus, the distribution of their ranks should be uniform, leading to a flat histogram. The shape of the rank histogram gives information about the type of (potential) miscalibration: a triangular-shaped histogram suggests that the probabilistic forecast has a systematic bias, a $\cup$-shaped histogram suggests that the probabilistic forecast is underdispersed and a $\cap$-shaped histogram suggests that the probabilistic forecast is overdispersed. Jolliffe and Primo (2008) proposed a statistical test to assess the uniformity (i.e., flatness) of rank histograms. Moreover, slopes in the rank histograms can be accounted for. Zamo (2016) proposed a test accounting for the presence of a wave in rank histograms. This test is called the Jolliffe-Primo-Zamo (JPZ) test in the following.

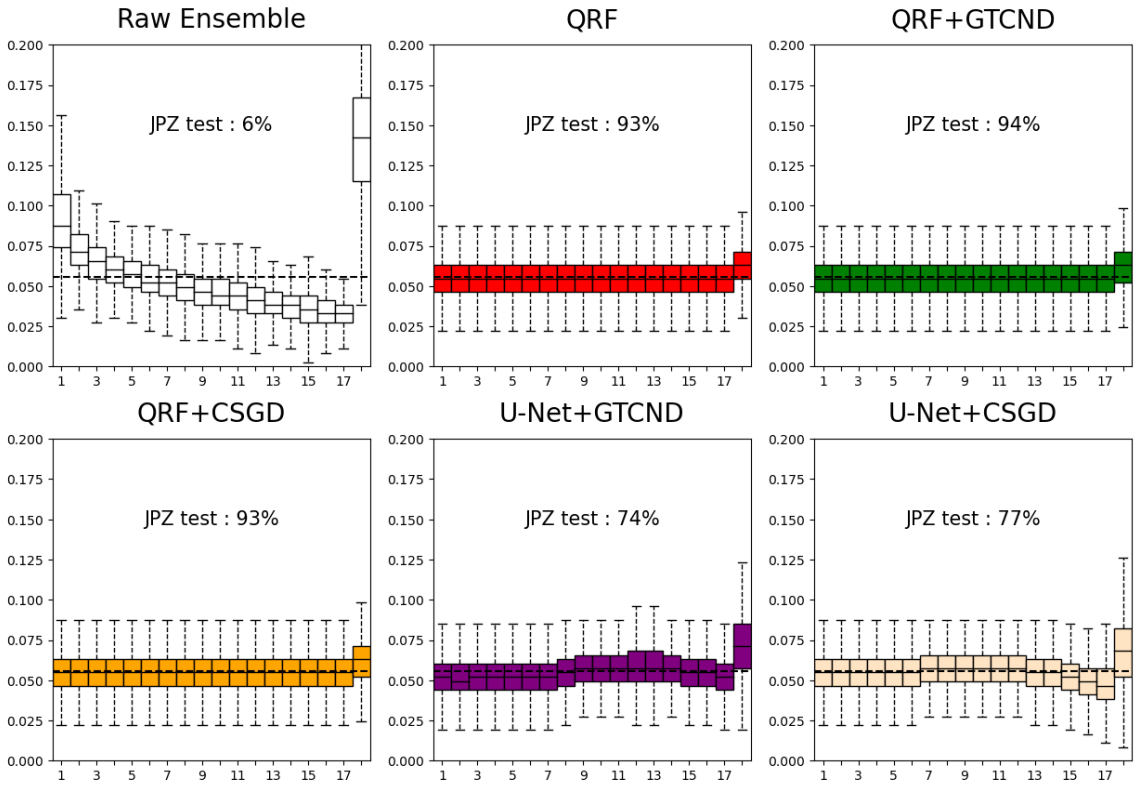


Figure 8: Rank histogram for raw ensemble, QRF, TQRF (namely, QRF+GTCND and QRF+CSGD) and distributional regression U-Nets associated with the GTCND and the CSGD. The hyperparameters are selected as the best performing by cross-validation on the training dataset.

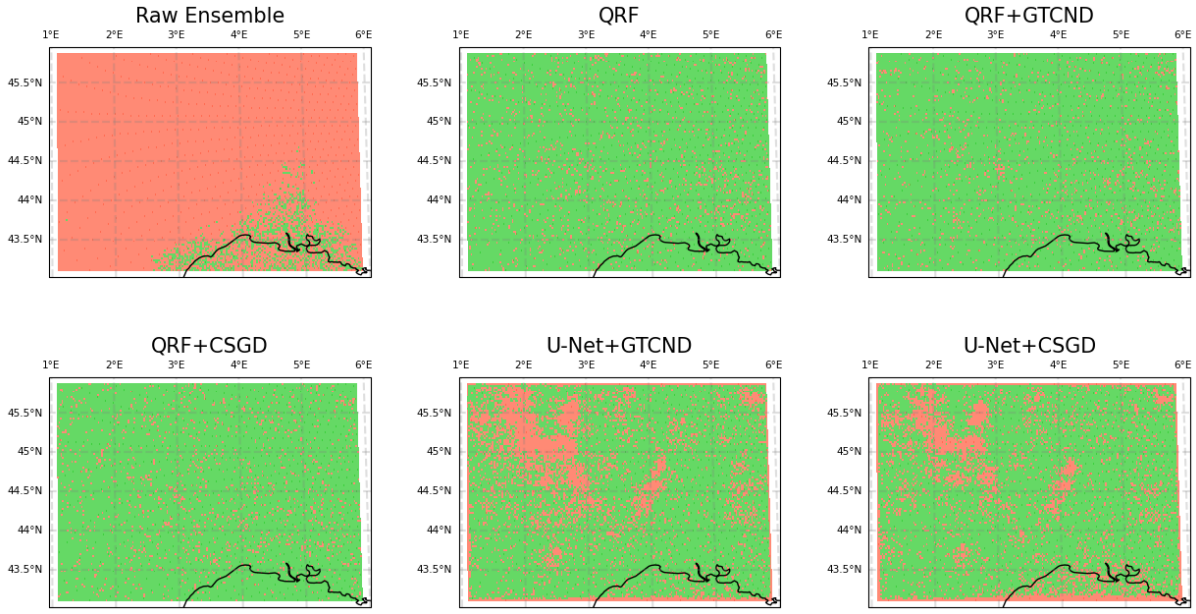


Figure 9: Map of rejection (red) and non-rejection (green) of the flatness of the rank histogram for the forecasting methods considered: raw ensemble, QRF, QRF+GTCND, QRF+CSGD, U-Net+GTCND and U-Net+CSGD.

To conciliate with the AROME-EPS raw ensemble composed of 17 members, the rank histograms can take 18 different classes and 107 quantiles of the forecasts were produced for the QRF, TQRF and DRU methods (each group of 6 consecutive ranks are gathered as a single rank).

Figure 8 shows the rank histograms of each forecast over the whole grid and the JPZ tests for flatness of rank histograms. As is often the case, the raw ensemble is biased and underdispersed, which is visible by the triangular shape of the rank histograms and the fact that the lowest and highest ranks are over-represented. Its JPZ test confirms that the raw ensemble forecast is not calibrated (only 6% of grid points do not reject the flatness of the rank histogram). QRF, QRF+GTCND and QRF+CSGD all show very high calibration with JPZ tests not rejecting flatness at 93%, 94% and 93% of grid points. Contrary to what was observed in Taillardat et al. (2019), no noticeable difference in calibration seems to be present between the QRF and its tail extension. This may be caused by the operational refinement used in the implementation, the fact that different parametric distributions are used and the smaller precipitation accumulations compared to the original article (i.e., 3-h vs. 6-h). DRUs present a lower calibration level compared to QRF-based methods, but their calibration is still significant. The JPZ tests do not reject the flatness hypothesis at 74% and 77% of the grid points for the U-Net+GTCND and U-Net+CSGD, respectively. Both DRU forecasts present a slight underdispersion in the right tail revealed by the higher representation of the largest rank in their histograms.

Figure 9 shows a map of the rejection and non-rejection of the flatness of the rank histogram given by JPZ tests. Calibrated grid points for the raw ensemble are sparsely located over the Mediterranean Sea, the coast and the South of the Rhône valley. QRF, QRF+GTCND and QRF+CSGD are able to calibrate the marginals homogeneously across the region of interest. The areas explaining the lower rate of calibrated grid point for DRU compared to QRF-based methods correspond to high climatological precipitation (see Fig. 6). The lack of calibration over these areas may be caused by the small depth of the training/validation data (only 3 years) resulting in not enough high precipitation observed. Moreover, the lower performance due to border effects affects the calibration of the DRU forecasts. DRU leads to spatially inconsistent forecasts in terms of calibration whereas the QRF-based methods are homogeneously calibrated over the whole domain.

4.3 Extreme events

Extreme events are of particular interest. They may lead to the highest socio-economic impacts. However, if verification were to focus only on cases of extreme events, forecasters might be encouraged to propose

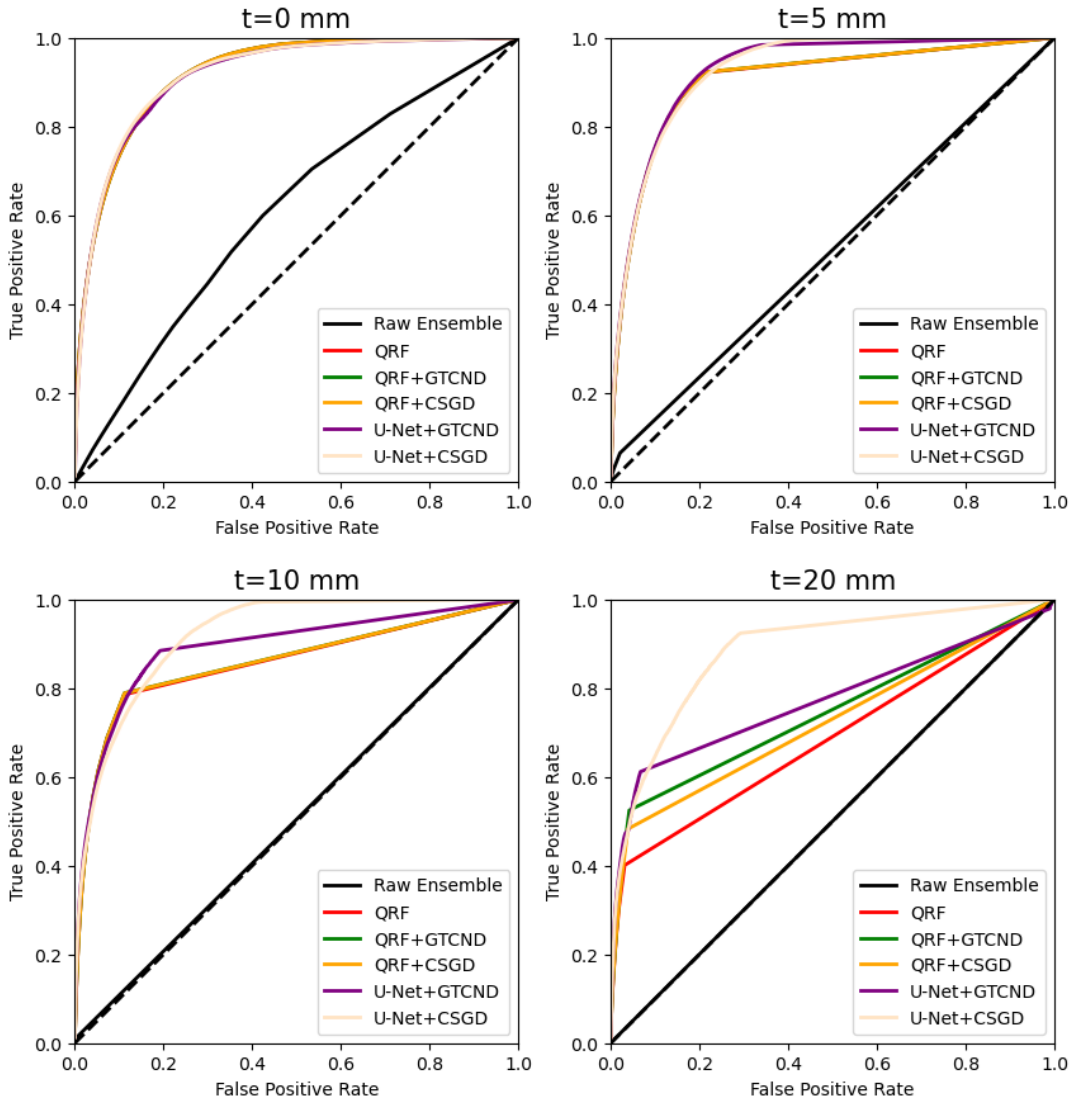


Figure 10: Receiver operating characteristic (ROC) curves of binary events corresponding to the exceedance of a threshold $t \in \{0, 5, 10, 20\}$ (in mm of precipitation). As for Figure 8, the hyperparameters are selected as the best performing by cross-validation on the training dataset.

forecasts that are overly alarming and, thus, of lower general predictive performance. Lerch et al. (2017) pinpointed this phenomenon and named it the *forecaster's dilemma*. Since we have compared the general predictive performance of postprocessing techniques, we can conduct verification focused on extreme events and not be affected by the forecaster's dilemma.

To focus on forecasts' predictive performance regarding extreme events, we are interested in predicting binary events in the form of the exceedance of a high threshold t. We use ROC (receiver operating characteristic) curves to evaluate the discriminant power of forecasts in terms of binary decisions. In particular, ROC curves can inform on the risk of missing an extreme event. Given the binary event $\mathbb{1}_{y>t}$ (i.e., exceedance of the threshold t), the ROC curve is the plot of the rate of predicted events (i.e., true positive), also called hit rate, versus the rate of false alarms (i.e., false positive). A good forecast should maximize the rate of events detected while minimizing false alarms. In practice, the compromise between the highest hit rate of the method and its lowest false alarm rate depends on the application. In the case of high-impact events, forecasts with a non-negligible false alarm rate may be tolerated if it is accompanied by a better hit rate. In addition to thresholds associated with extreme events, lower thresholds corresponding to lower precipitation events are investigated. Note that grid point by grid point computation of ROC curves does not prevent potential double-penalty effects (Ebert, 2008).

In Figure 10, ROC curves for the exceedance of various thresholds are represented for the raw ensem-

ble, QRF, TQRF and DRU. The lowest threshold is $t = 0$mm, which characterizes the prediction of dry events (i.e., absence of precipitation). Raw ensemble has a poor performance regarding the prediction of the presence of precipitation. All the postprocessing methods have comparable performances, as seen in the overlap of their ROC curves. During the cross-validation over the training/validation dataset, the raw ensemble had a better predictive power regarding the prediction of dry events but was still lower than the postprocessing methods (not shown). The threshold $t = 5$ mm corresponds to intermediate precipitations. The performance of the raw ensemble already decreases and a difference between DRU and QRF-based methods appears. DRUs have a slightly higher predictive performance compared to QRF-based methods. The raw ensemble lacks resolution because of the nature of its miscalibration (i.e., bias and underdispersion).

For the highest thresholds $t = 10$ mm and $t = 20$ mm (corresponding to the quantile of level 0.995 and 0.999, respectively, of the climatology over the region of interest), the ROC curves of the different postprocessing methods can be distinguished. For $t = 10$ mm, the performance of the raw ensemble continues to deteriorate and is close to the random guess (dashed line). All the postprocessing techniques are able to maintain a good predictive power but start to noticeably lack resolution, which can be seen in the sudden change of slope. U-Net+GTCND and U-Net+CSGD have a better performance compared to QRF-based techniques which continue to have overlapping ROC curves. U-Net+CSGD has the overall best performance. For $t = 20$ mm, the raw ensemble has a performance indistinguishable from a random guess. DRUs are better than QRF-based methods. QRF+GTCND and QRF+CSGD denote from QRF as the tail extension improves predictive performance. QRF+GTCND seems to have a slightly better performance than QRF+CSGD. The gap in performance between U-Net+CSGD and U-Net+GTCND continues to grow and U-Net+CSGD clearly has to the best predictive power w.r.t. the exceedance of the threshold $t = 20$ mm.

All postprocessing methods compete favorably with the raw ensemble, which has the same predictive performance as a random guess for the highest thresholds ($t = 10$ mm and $t = 20$ mm). All postprocessing methods have comparable predictive performances for dry events. For heavy precipitation events corresponding to quantiles of levels 0.995 and 0.999, DRUs, and in particular U-Net+CSGD, have a distinctly better predictive power. Moreover, as already observed in Taillardat et al. (2016), TQRF is able to improve the prediction of heavy precipitation with respect to QRF (even for a light-tailed extension as the GTCND).

5 Discussion

We proposed a U-Net-based method, namely distributional regression U-Nets, to postprocess marginal distributions for gridded precipitation data. This approach extends DRN to gridded data by substituting the fully connected NN and embedding module for a U-Net architecture aware of the gridded structure of the data. Simultaneously predicting marginal distributions at each grid point using information from nearby grid points represents a means to account for dependencies between grid points. Both U-Net+GTCND and U-Net+CSGD have predictive performances comparable to the QRF and TQRF in terms of CRPS. DRUs are (probabilistically) calibrated over a large part of the domain studied except for areas associated with the highest precipitation over the test set (see Fig. 6). This may result from the relatively small training/validation set and could improve with a larger training/validation set. Future studies could try to limit this by emphasizing the learning of high precipitation events using weighted scoring rules for inference. In terms of heavy precipitation, U-Net+CSGD outperforms QRF-based methods.

One of the challenges of the dataset used is the small amount of available training data. This is encountered in practice where consistent data is required, but large reforecast and reanalysis are too computationally expensive. In a more general context, the lack of consistency can be induced at larger time scales by climate change or in specific regions of the world by El Niño forcing.

We focused on distributional regression U-Nets where outputs are distribution parameters based on CRPS minimization. DRU can rely on the minimization of other (strictly) proper scoring rules. Moreover, DRU can directly be extended to learn nonparametric distributions such as BQN (Bremnes, 2020) where the quantile function is a combination of Bernstein polynomials or as HEN (e.g., Scheuerer et al. 2020) where the pdf is modeled by the probability of bins.

As U-Net architecture is aware of the spatial gridded structure of the data, specific architectures can also be used for common data structures. We present architectures related to temporal and graph-based

structures that are currently used in probabilistic forecasting settings. Their application to postprocessing provides an interesting for future works. For example, if the temporal structure of the data is of interest, recurrent neural networks can be used to predict a parametric distribution. Pasche and Engelke (2024) proposed to forecast flood risk using high-quantile prediction based on fitting a generalized Pareto distribution via logarithmic score (i.e., negative log-likelihood) minimization. In the case of spatial structure relying on an irregular or more abstract grid (e.g., station network), graph neural networks (GNNs) are able to predict graph-based quantities (Battaglia et al., 2018). Cisneros et al. (2024) used graph convolutional neural networks to learn the parameters of a mixture of a logistic distribution and EGPD via logarithm score minimization to predict wildfire spread. Using the 3D spatial graph-based structures, GNNs are already able to produce deterministic forecasts reaching performance comparable to ECMWF deterministic high-resolution forecasts in performance (Keisler, 2022; Pathak et al., 2022; Bi et al., 2023; Lam et al., 2022; Chen et al., 2023).

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project) and the Energy-oriented Centre of Excellence II (EoCoE-II), Grant Agreement 824158, funded within the Horizon2020 framework of the European Union. Part of this work was also supported by the ExtremesLearning grant from 80 PRIME CNRS-INSU and this study has received funding from Agence Nationale de la Recherche - France 2030 as part of the PEPR TRACCS program under grant number ANR-22-EXTR-0005 and the ANR EXSTA.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Sándor Baran and Dóra Nemoda. Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, 27(5):280–292, May 2016. ISSN 1099-095X. https://doi.org/10.1002/env.2391.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. June 2018. https://doi.org/10.48550/ARXIV.1806.01261.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, sep 2015. https://doi.org/10.1038/nature14956.

Zied Ben Bouallègue, Tobias Heppelmann, Susanne E. Theis, and Pierre Pinson. Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Monthly Weather Review*, 144(12):4737–4750, November 2016. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-15-0403.1.

Zied Ben Bouallègue, Jonathan A. Weyn, Mariana C. A. Clare, Jesper Dramsch, Peter Dueben, and Matthew Chantry. Improving medium-range ensemble weather forecasts with hierarchical ensemble transformers. *Artificial Intelligence for the Earth Systems*, 3(1), January 2024. ISSN 2769-7525. https://doi.org/10.1175/aies-d-23-0027.1.

Patrick Bénichou. Cartography of statistical pluviometric fields with an automatic allowance for regional topography. In *Global Precipitations and Climate Change*, pages 187–199. Springer Berlin Heidelberg, 1994. https://doi.org/10.1007/978-3-642-79268-7_11.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, July 2023. ISSN 1476-4687. https://doi.org/10.1038/s41586-023-06185-3.

François Bouttier, Laure Raynaud, Olivier Nuissier, and Benjamin Ménétrier. Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. *Quarterly Journal of the Royal Meteorological Society*, 142(S1):390–403, sep 2015. https://doi.org/10.1002/qj.2622.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. https://doi.org/10.1023/a:1010933404324.

John Bjørnar Bremnes. Ensemble postprocessing using quantile function regression based on neural networks and bernstein polynomials. *Monthly Weather Review*, 148(1):403–414, jan 2020. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-19-0227.1.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78 (1):1–3, 1950. ISSN 1520-0493. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

Olivier Caumont, Marc Mandement, François Bouttier, Judith Eeckman, Cindy Lebeaupin Brossier, Alexane Lovat, Olivier Nuissier, and Olivier Laurantin. The heavy precipitation event of 14–15 october 2018 in the aude catchment: a meteorological study based on operational numerical weather prediction systems and standard and personal observations. *Natural Hazards and Earth System Sciences*, 21(3): 1135–1157, March 2021. ISSN 1684-9981. https://doi.org/10.5194/nhess-21-1135-2021.

Jean-Louis Champeaux, Pascale Dupuy, Olivier Laurantin, Isabelle Soulan, Pierre Tabary, and Jean-Michel Soubeyroux. Les mesures de précipitations et l'estimation des lames d'eau à Météo-France : état de l'art et perspectives. *La Houille Blanche*, 95(5):28–34, October 2009. ISSN 1958-5551. https://doi.org/10.1051/lhb/2009052.

William E. Chapman, Luca Delle Monache, Stefano Alessandrini, Aneesh C. Subramanian, F. Martin Ralph, Shang-Ping Xie, Sebastian Lerch, and Negin Hayatbini. Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Monthly Weather Review*, 150(1):215–234, jan 2022. https://doi.org/10.1175/mwr-d-21-0106.1.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, Yuanzheng Ci, Bin Li, Xiaokang Yang, and Wanli Ouyang. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. April 2023. https://doi.org/10.48550/ARXIV.2304.02948.

François Chollet et al. Keras. https://keras.io, 2015. URL https://keras.io.

Daniela Cisneros, Jordan Richards, Ashok Dahal, Luigi Lombardo, and Raphaël Huser. Deep graphical regression for jointly moderate and extreme australian wildfires. *Spatial Statistics*, 59:100811, March 2024. ISSN 2211-6753. https://doi.org/10.1016/j.spasta.2024.100811.

Martyn Clark, Subhrendu Gangopadhyay, Lauren Hay, Balaji Rajagopalan, and Robert Wilby. The schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5(1):243–262, feb 2004. https://doi.org/10.1175/1525-7541(2004)005<0243:tssamf>2.0.co;2.

Y. Dai and S. Hemri. Spatially coherent postprocessing of cloud cover ensemble forecasts. *Monthly Weather Review*, 149(12):3923–3937, dec 2021. https://doi.org/10.1175/mwr-d-21-0046.1.

Luca Delle Monache, F. Anthony Eckel, Daran L. Rife, Badrinath Nagarajan, and Keith Searight. Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, 141(10):3498–3516, September 2013. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-12-00281.1.

Jean Diebolt, Armelle Guillou, and Imen Rached. Approximation of the distribution of excesses through a generalized probability-weighted moments method. *Journal of Statistical Planning and Inference*, 137(3):841–857, mar 2007. https://doi.org/10.1016/j.jspi.2006.06.012.

Elizabeth E. Ebert. Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, 15(1):51–64, 2008. https://doi.org/10.1002/met.25.

Kira Feldmann, David S. Richardson, and Tilmann Gneiting. Grid- versus station-based postprocessing of ensemble temperature forecasts. *Geophysical Research Letters*, 46(13):7744–7751, July 2019. ISSN 1944-8007. https://doi.org/10.1029/2019gl083189.

C. A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, December 2013. ISSN 0035-9009. https://doi.org/10.1002/qj.2270.

Petra Friederichs, Sabrina Wahl, and Sebastian Buschow. Postprocessing for extreme events. In *Statistical Postprocessing of Ensemble Forecasts*, pages 127–154. Elsevier, 2018. https://doi.org/10.1016/b978-0-12-812372-0.00005-4.

Tilmann Gneiting and Matthiass Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and its Applications*, 2014. https://doi.org/10.1146/annurev-statistics-062713-085831.

Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118, May 2005. ISSN 0027-0644. https://doi.org/10.1175/mwr2904.1.

Peter Grönquist, Chengyuan Yao, Tal Ben-Nun, Nikoli Dryden, Peter Dueben, Shigang Li, and Torsten Hoefler. Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200092, 2021. https://doi.org/10.1098/rsta.2020.0092.

Thomas M. Hamill. *Practical Aspects of Statistical Postprocessing*, pages 187–217. Elsevier, 2018. ISBN 9780128123720. https://doi.org/10.1016/b978-0-12-812372-0.00007-8.

Thomas M. Hamill and Stephen J. Colucci. Verification of eta–rsm short-range ensemble forecasts. *Monthly Weather Review*, 125(6):1312–1327, June 1997. ISSN 1520-0493. https://doi.org/10.1175/1520-0493(1997)125<1312:voersr>2.0.co;2.

S. Hemri, M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden. Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41(24):9197–9205, dec 2014. https://doi.org/10.1002/2014gl062472.

Nina Horat and Sebastian Lerch. Deep learning for post-processing global probabilistic forecasts on sub-seasonal time scales. June 2023.

Weiming Hu, Mohammadvaghef Ghazvinian, William E. Chapman, Agniv Sengupta, Fred Martin Ralph, and Luca Delle Monache. Deep learning forecast uncertainty for precipitation over the western united states. *Monthly Weather Review*, 151(6):1367–1385, jun 2023. https://doi.org/10.1175/mwr-d-22-0268.1.

Ian T. Jolliffe and Cristina Primo. Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review*, 136(6):2133–2139, jun 2008. https://doi.org/10.1175/2007mwr2219.1.

Alexander Jordan, Fabian Krüger, and Sebastian Lerch. Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*, 90(12), 2019. ISSN 1548-7660. https://doi.org/10.18637/jss.v090.i12.

Ryan Keisler. Forecasting global weather with graph neural networks. February 2022. https://doi.org/10.48550/ARXIV.2202.07575.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Graphcast: Learning skillful medium-range global weather forecasting. December 2022. https://doi.org/10.48550/ARXIV.2212.12794.

Sebastian Lerch and Kai L. Polsterer. Convolutional autoencoders for spatially-informed ensemble post-processing. April 2022.

Sebastian Lerch and Thordis L. Thorarinsdottir. Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, 65(1):21206, December 2013. ISSN 1600-0870. https://doi.org/10.3402/tellusa.v65i0.21206.

Sebastian Lerch, Thordis L. Thorarinsdottir, Francesco Ravazzolo, and Tilmann Gneiting. Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32(1), February 2017. ISSN 0883-4237. https://doi.org/10.1214/16-sts588.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.

Wentao Li, Baoxiang Pan, Jiangjiang Xia, and Qingyun Duan. Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *Journal of Hydrology*, 605:127301, feb 2022. https://doi.org/10.1016/j.jhydrol.2021.127301.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22, 1976. https://doi.org/10.2307/2629907.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

Jakob W. Messner, Georg J. Mayr, and Achim Zeileis. Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, 145(1):137–147, 2017. https://doi.org/10.1175/mwr-d-16-0088.1.

Thomas Muschinski, Georg J. Mayr, Achim Zeileis, and Thorsten Simon. Robust weather-adaptive post-processing using model output statistics random forests. *Nonlinear Processes in Geophysics*, 30(4): 503–514, November 2023. ISSN 1607-7946. https://doi.org/10.5194/npg-30-503-2023.

Philippe Naveau, Raphael Huser, Pierre Ribereau, and Alexis Hannart. Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769, apr 2016. https://doi.org/10.1002/2015wr018552.

David A. Olson, Norman W. Junker, and Brian Korty. Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Weather and Forecasting*, 10(3):498–511, sep 1995. https://doi.org/10.1175/1520-0434(1995)010<0498:eoyoqp>2.0.co;2.

Olivier C. Pasche and Sebastian Engelke. Neural networks for extreme quantile regression with an application to forecasting of flood risk. 2024. https://doi.org/10.48550/ARXIV.2208.07590.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. February 2022.

Serge Planton, Michel Déqué, Fabrice Chauvin, and Laurent Terray. Expected impacts of climate change on extreme climate events. *Comptes Rendus Geoscience*, 340(9-10):564–574, sep 2008. https://doi.org/10.1016/j.crte.2008.07.009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

Stephan Rasp and Sebastian Lerch. Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900, 2018. https://doi.org/10.1175/mwr-d-18-0187.1.

Didier Ricard, Véronique Ducrocq, and Ludovic Auger. A climatology of the mesoscale environment associated with heavily precipitating events over a northwestern mediterranean area. *Journal of Applied Meteorology and Climatology*, 51(3):468–488, March 2012. ISSN 1558-8432. https://doi.org/10.1175/jamc-d-11-017.1.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science*, pages 234–241. Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

Roman Schefzik and Annette Möller. *Ensemble Postprocessing Methods Incorporating Dependence Structures*, pages 91–125. Elsevier, 2018. ISBN 9780128123720. https://doi.org/10.1016/b978-0-12-812372-0.00004-2.

Roman Schefzik, Thordis L. Thorarinsdottir, and Tilmann Gneiting. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28(4), nov 2013. https://doi.org/10.1214/13-sts443.

Michael Scheuerer and Thomas M. Hamill. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions*. *Monthly Weather Review*, 143(11):4578–4596, October 2015. ISSN 1520-0493. https://doi.org/10.1175/mwr-d-15-0061.1.

Michael Scheuerer, Matthew B. Switanek, Rochelle P. Worsnop, and Thomas M. Hamill. Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over california. *Monthly Weather Review*, 148(8):3489–3506, aug 2020. https://doi.org/10.1175/mwr-d-20-0096.1.

Lisa Schlosser, Torsten Hothorn, Reto Stauffer, and Achim Zeileis. Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *The Annals of Applied Statistics*, 13(3), sep 2019. https://doi.org/10.1214/19-aoas1247.

Benedikt Schulz and Sebastian Lerch. Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, 150(1):235–257, 2022a. https://doi.org/10.1175/mwr-d-21-0150.1.

Benedikt Schulz and Sebastian Lerch. Aggregating distribution forecasts from deep ensembles. April 2022b. https://doi.org/10.48550/ARXIV.2204.02291.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. September 2014. https://doi.org/10.48550/ARXIV.1409.1556.

Maxime Taillardat and Olivier Mestre. From research to applications – examples of operational ensemble post-processing in france using machine learning. *Nonlinear Processes in Geophysics*, 27(2):329–347, may 2020. https://doi.org/10.5194/npg-27-329-2020.

Maxime Taillardat, Olivier Mestre, Michaël Zamo, and Philippe Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144 (6):2375–2393, 2016. https://doi.org/10.1175/mwr-d-15-0260.1.

Maxime Taillardat, Anne-Laure Fougères, Philippe Naveau, and Olivier Mestre. Forest-based and semi-parametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34(3):617–634, 2019. https://doi.org/10.1175/waf-d-18-0149.1.

Maxime Taillardat, Anne-Laure Fougères, Philippe Naveau, and Olivier Mestre. Corrigendum. *Weather and Forecasting*, 37(7):1305, July 2022. ISSN 1520-0434. https://doi.org/10.1175/waf-d-22-0057.1.

Alexander Tsyplakov. Evaluation of probabilistic forecasts: Proper scoring rules and moments. *SSRN Electronic Journal*, 2013. ISSN 1556-5068. https://doi.org/10.2139/ssrn.2236605.

Alexander Tsyplakov. Evaluation of probabilistic forecasts: Conditional auto-calibration, 2020. URL https://www.sas.upenn.edu/~fdiebold/papers2/Tsyplakov_Auto_calibration_sent_eswc2020.pdf.

Bert Van Schaeybroeck and Stéphane Vannitsem. Ensemble post-processing using member-by-member approaches: theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*, 141(688): 807–818, jun 2014. https://doi.org/10.1002/qj.2397.

Chiem van Straaten, Kirien Whan, and Maurice Schmeits. Statistical postprocessing and multivariate structuring of high-resolution ensemble precipitation forecasts. *Journal of Hydrometeorology*, 19(11): 1815–1833, nov 2018. https://doi.org/10.1175/jhm-d-18-0105.1.

Stéphane Vannitsem, John Bjørnar Bremnes, Jonathan Demaeyer, Gavin R. Evans, Jonathan Flowerdew, Stephan Hemri, Sebastian Lerch, Nigel Roberts, Susanne Theis, Aitor Atencia, Zied Ben Bouallègue, Jonas Bhend, Markus Dabernig, Lesley De Cruz, Leila Hieta, Olivier Mestre, Lionel Moret, Iris Odak Plenković, Maurice Schmeits, Maxime Taillardat, Joris Van den Bergh, Bert Van Schaeybroeck, Kirien Whan, and Jussi Ylhaisi. Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102(3):E681–E699, mar 2021. https://doi.org/10.1175/bams-d-19-0308.1.

Simon Veldkamp, Kirien Whan, Sjoerd Dirksen, and Maurice Schmeits. Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Monthly Weather Review*, 149(4):1141–1152, apr 2021. https://doi.org/10.1175/mwr-d-20-0219.1.

Kirien Whan and Maurice Schmeits. Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods. *Monthly Weather Review*, 146(11):3651–3673, oct 2018. https://doi.org/10.1175/mwr-d-17-0290.1.

R. M. Williams, C. A. T. Ferro, and F. Kwasniok. A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1112–1120, jul 2013. https://doi.org/10.1002/qj.2198.

R. L. Winkler, Javier Muñoz, José L. Cervera, José M. Bernardo, Gail Blattenberger, Joseph B. Kadane, Dennis V. Lindley, Allan H. Murphy, Robert M Oliver, and David Ríos-Insua. Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60, June 1996. ISSN 1863-8260. https://doi.org/10.1007/bf02562681.

Robert L. Winkler. *Rewarding Expertise in Probability Assessment*, pages 127–140. Springer Netherlands, 1977. ISBN 9789401012768. https://doi.org/10.1007/978-94-010-1276-8_10.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, 77(1), 2017. ISSN 1548-7660. https://doi.org/10.18637/jss.v077.i01.

Michaël Zamo. *Statistical Post-processing of Deterministic and Ensemble Wind Speed Forecasts on a Grid*. Theses, Université Paris Saclay (COmUE), December 2016. URL https://theses.hal.science/tel-01598119.

A Generalized Truncated/Censored Normal Distribution

We recall quantities related to the generalized truncated/censored normal distribution (GTCND). Denote l and u the lower and upper boundaries, L and U are the point masses at these boundaries. Since we are working with precipitation, we are interested in the case where $u = \infty$ (implying that $U = 0$) and $l = 0$, leaving L a parameter to determine along μ and σ. Formulas for the general case are available in Jordan et al. (2019).

The cumulative distribution function (cdf) of the GTCND is

$$F^{gtcnd}_{L,\mu,\sigma}(z) = \begin{cases} \dfrac{1-L}{1-\Phi(-\mu/\sigma)}\big(\Phi(\frac{z-\mu}{\sigma}) - \Phi(-\mu/\sigma)\big) + L & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

where Φ is the cdf of the standard normal distribution. Its quantile function is expressed as

$$F^{\text{gtcnd}^{-1}}_{L,\mu,\sigma}(p) = \begin{cases} 0 & \text{if } p \leq L \\ \mu + \sigma\Phi^{-1}\left(\frac{(p-L)(1-\Phi(-\mu/\sigma)}{1-L} + \Phi(-\mu/\sigma)\right) & \text{if } p > L \end{cases}$$

for $p \in (0, 1)$. The special case of GTCND used here can be expressed using the truncated normal distribution :

$$F^{\text{gtcnd}}_{L,\mu,\sigma}(z) = L\mathbb{1}_{z\geq 0} + (1-L)N^0_{\mu,\sigma}(z),$$

where $N^0_{\mu,\sigma}$ is the cdf of the zero-truncated normal distribution.

Moments methods

$$\mathbb{E}[\mathbb{1}_{X=0}] = L$$

$$\mathbb{E}[X] = \mu + \frac{\phi(-\mu/\sigma)\sigma}{1-\Phi(-\mu/\sigma)}$$

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sigma^2\left\{1 - \frac{\mu}{\sigma}\frac{\phi(\mu/\sigma)}{1-\Phi(-\mu/\sigma)} - \left(\frac{\phi(-\mu/\sigma)}{1-\Phi(-\mu/\sigma)}\right)^2\right\}$$

Continuous Ranked Probability Score

$$\begin{aligned}\mathrm{CRPS}(F^{\mathrm{gtcnd}}_{L,\mu,\sigma}, y) = {} & |y - y_+| + \mu L^2 \\ & + \frac{1-L}{1-\Phi(-\frac{\mu}{\sigma})}(y_+ - \mu)\left\{2\Phi\left(\frac{y_+-\mu}{\sigma}\right) - \frac{1-2L+\Phi\left(-\frac{\mu}{\sigma}\right)}{1-L}\right\} \\ & + 2\sigma\frac{1-L}{1-\Phi(-\frac{\mu}{\sigma})}\left(\phi\left(\frac{y_+-\mu}{\sigma}\right) - \phi\left(-\frac{\mu}{\sigma}\right)L\right) \\ & - \left(\frac{1-L}{1-\Phi(-\frac{\mu}{\sigma})}\right)^2\frac{\sigma}{\sqrt{\pi}}\Phi\left(\frac{\mu\sqrt{2}}{\sigma}\right)\end{aligned}$$

with $y_+ = \max(0, y)$ and ϕ the probability density function of the standard normal distribution.

B Censored-Shifted Gamma Distribution

We recall quantities related to the censored-shifted gamma distribution (CSGD). The expressions can be found in Scheuerer and Hamill (2015) and Baran and Nemoda (2016). The cumulative distribution function (cdf) of the CSGD is

$$F^{\mathrm{csgd}}_{k,\theta,\delta}(z) = \begin{cases} G_k(\frac{z-\delta}{\theta}) & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases},$$

with G_k the cdf of the gamma distribution of shape k. Its quantile function is expressed as

$$F^{\mathrm{csgd}\,-1}_{k,\theta,\delta}(p) = \delta + \theta\gamma^{-1}(k, p\Gamma(k)),$$

where γ is the lower incomplete gamma function, Γ is the gamma function and $p \in (0, 1)$.

Moments method

Let $\tilde{c} = -\delta/\theta$.

$$\mathbb{E}[X] = (1 - G_k(\tilde{c}))\Big\{\theta k(1 - G_{k+1}(\tilde{c})) - \delta(1 - G_k((\tilde{c}))\Big\}$$

$$\begin{aligned}\mathbb{E}[X^2] = (1 - G_k((\tilde{c}))\Big\{ & k(k+1)\theta^2(1 - G_{k+2}(\tilde{c})) \\ & - 2\delta k\theta(1 - G_{k+1}(\tilde{c})) \\ & + \delta^2(1 - G_k(\tilde{c}))\Big\}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X^3] = (1 - G_k(\tilde{c})\Big\{ & k(k+1)(k+2)\theta^3(1 - G_{k+3}(\tilde{c})) \\ & - 3\delta k(k+1)\theta^2(1 - G_{k+2}(\tilde{c})) \\ & + 3\delta^2 k\theta(1 - G_{k+1}(\tilde{c})) \\ & - \delta^3(1 - G_k(\tilde{c}))\Big\}\end{aligned}$$

Continuous Ranked Probability Score

The continuous ranked probability score (CRPS) of the CSGD is

$$\begin{aligned}\mathrm{CRPS}(F^{\mathrm{csgd}}_{k,\theta,\delta}, y) = \theta\Big\{ &\tilde{y}\,(2G_k(\tilde{y}) - 1) - \tilde{c}G_k^2(\tilde{c}) + \theta k\left(1 + 2G_k(\tilde{c})G_{k+1}(\tilde{c}) - G_k^2(\tilde{c}) - 2G_{k+1}(\tilde{y})\right) \\ &- \frac{\theta k}{\pi} B(1/2, k+1/2)\,(1 - G_{2k}(2\tilde{c}))\Big\},\end{aligned}$$

where $\tilde{y} = \frac{y-\delta}{\theta}$, $\tilde{c} = -\delta/\theta$ and B is the beta function.