Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR035088

Intercomparison of Deep Learning Architectures for the Prediction of Precipitation Fields With a Focus on Extremes

Noelia Otero[1] **and Pascal Horton**[1]

[1]Institute of Geography and Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

Key Points:

- We present an intercomparison of deep learning models to assess the ability of different architectures to predict precipitation events
- U-Net-based architectures outperformed the rest of the models
- A layer-wise relevance propagation explainability method quantify the most important feature to predict precipitation extreme

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

N. Otero,
noelia.otero@unibe.ch

Citation:

Otero, N., & Horton, P. (2023). Intercomparison of deep learning architectures for the prediction of precipitation fields with a focus on extremes. *Water Resources Research*, *59*, e2023WR035088. https://doi.org/10.1029/2023WR035088

Received 20 APR 2023
Accepted 15 SEP 2023

Abstract In recent years, the use of deep learning methods has rapidly increased in many research fields. Similarly, they have become a powerful tool within the climate scientific community. Deep learning methods have been successfully applied for different tasks, such as the identification of atmospheric patterns, weather extreme classification, or weather forecasting. However, due to the inherent complexity of atmospheric processes, the ability of deep learning models to simulate natural processes, particularly in the case of weather extremes, is still challenging. Therefore, a thorough evaluation of their performance and robustness in predicting precipitation fields is still needed, especially for extreme precipitation events, which can have devastating consequences in terms of infrastructure damage, economic losses, and even loss of life. In this study, we present a comprehensive evaluation of a set of deep learning architectures to simulate precipitation, including heavy precipitation events (>95th percentile) and extreme events (>99th percentile) over the European domain. Among the architectures analyzed here, the U-Net network was found to be superior and outperformed the other networks in simulating precipitation events. In particular, we found that a simplified version of the original U-Net with two encoder-decoder levels generally achieved similar skill scores than deeper versions for predicting precipitation extremes, while significantly reducing the overall complexity and computing resources. We further assess how the model predicts through the attribution heatmaps from a layer-wise relevance propagation explainability method.

Plain Language Summary With the increasing success of machine learning methods in Earth Sciences, deep learning is becoming a promising tool for building data-driven models for meteorological applications. Yet, predicting extreme events, such as heavy rainfall, is still challenging. Here, we present an intercomparison of deep learning models to assess the ability of different architectures to predict precipitation events. While most of the models perform relatively well, we show that U-Net-based architectures outperformed the rest of the models. We additionally applied explainability methods to quantify which input features are the most important to predict precipitation events. Overall, the relative humidity showed the highest relevance values, followed by both wind components, particularly in western and southern Europe.

1. Introduction

Predicting precipitation is challenging for numerical weather prediction (NWP) models. NWP models solve numerically coupled partial differential equations subject to dynamic and thermodynamic laws that describe the atmospheric state (Schultz et al., 2021). However, precipitation generation involves complex microphysical processes that occur on very small scales, and that cannot be explicitly resolved in most NWP due to inadequate grid resolution, which might lead to large uncertainties. Such processes are inferred from parametrization schemes, which are generally sources of parametric uncertainty (Bauer et al., 2015). Moreover, NWP models are computationally expensive, particularly at a high resolution.

A major concern relates to extreme precipitation events that are expected to change in intensity and frequency under a changing climate, leading to higher socio-economic impacts (Donat et al., 2016; Trenberth et al., 2003). The skill of climate models, or more specifically general circulation models, to predict extreme events is rather limited due to their lack of ability to represent mesoscale processes that require higher spatio-temporal resolutions (Gao & Adam Schlosser, 2019). Regional climate models can better represent topography and small-scale microphysical processes thanks to a higher spatial resolution (2–25 km) but are computationally expensive (Adewoyin et al., 2021).

With the rapid development of machine learning (ML) techniques, sophisticated deep learning (DL) models, and the availability of large data sets, there is an increasing interest in the weather and climate research community to

tackle climate-related problems using ML. ML models can extract feature representations from observed patterns and relate them to general meteorological situations. Moreover, ML models are computationally much cheaper than physically based modeling of precipitation. Recent studies have proposed different ML methods and DL architectures to predict precipitation at several time scales, including nowcasting, subseasonal, and seasonal forecast (Civitarese et al., 2021; Hwang et al., 2019; Vandal et al., 2019). These ML applications have shown promising results for predicting precipitation (Adewoyin et al., 2021).

Data-driven approaches have become very popular in many fields of natural sciences due to their ability to learn and efficiently represent underlying physical processes (Rasp et al., 2020). Several studies have shown the great potential of convolutional neuronal network (CNN) architectures to reproduce synoptic patterns (Chattopadhyay et al., 2020), weather extreme events (Liu et al., 2016), and provide weather forecasting (Scher, 2018; Weyn et al., 2019). In particular, precipitation forecasting has been the subject of DL studies that have proposed advanced network architectures that can outperform conventional forecast models (Rasp et al., 2020).

The use of DL is exponentially growing in climate and weather applications, as it is in other fields. Some state-of-the-art architectures, such as generative adversarial networks (GANs) have shown promising results for nowcasting precipitation (Ravuri et al., 2021), downscaling tasks (Harris et al., 2022; Leinonen et al., 2021), and bias correction (Pan et al., 2021). Moreover, the success of transformers in computer vision tasks has sparked interest in adapting these models to climate and weather applications, as shown in Dabrowski et al. (2020) and Nguyen et al. (2023). For example, researchers have leveraged transformers to improve seasonal forecasts (e.g., Civitarese et al., 2021). Due to the wide range of DL architectures, the effort needed to adapt the models to a specific case study and to train them, and the pace at which new models emerge, it is almost infeasible to perform an exhaustive and up-to-date comparison.

Previous works have used DL to predict extreme precipitation for spatially aggregated time series (Davenport & Diffenbaugh, 2021; Huang, 2022) or to predict high-resolution precipitation locally (i.e., statistical downscaling) (Adewoyin et al., 2021; Pan et al., 2019). However, the extreme values in the predicted precipitation fields over a larger domain have not yet been investigated enough nor improved. Therefore, this work aims to fill this gap by assessing the performance of existing DL models in predicting extremes in precipitation fields over a large domain. Building upon recent works, we present an intercomparison of DL architectures and assess their ability to predict extreme precipitation events over Europe.

The context of this work is the comparison of models for the prediction of daily precipitation for the same day as the inputs, that is, without temporal extrapolation of the evolution of the weather situation. The objective is, therefore, not to forecast precipitation for a given lead time, but to establish a statistical relationship between meteorological variables and the concomitant precipitation. Within this framework, we assess the ability of the different models to infer precipitation extremes resulting from different synoptic-scale situations. Although not providing temporal extrapolation, such an approach is interesting for bypassing or improving the prediction of precipitation by NWP, which is uncertain and computationally intensive, or by climate models, whose resolutions are most often inappropriate for explicitly modeling precipitation.

While our primary focus is to test the model performance to capture precipitation extremes, we also examine the DL performance for precipitation estimates. Contrasting with most of the existing literature where the domain of interest focused on precipitation over the U.S. (e.g., Davenport & Diffenbaugh, 2021; Pan et al., 2019), here we present a model comparison over the European domain. The skills of the models are compared for the prediction of the precipitation amount over the entire domain as well as for the probability of exceedance of the 95th (i.e., heavy precipitation) and 99th (i.e., extreme precipitation) percentiles. In addition, a baseline model was used to benchmark the performance of the selected DL architectures. The baseline consists in a random forest (RF) model (Breiman, 2001), which is applied point-wise. RF is a common and robust algorithm that has been previously applied to predict precipitation (e.g., A. J. Hill & Schumacher, 2022; G. R. Hill et al., 2022; Wolfensberger et al., 2021).

In the second step, we select the best architecture based on the skill scores, and we apply an eXplainable artificial intelligence (XAI) (Montavon et al., 2018) method to provide further physical insights about the connections between the inputs and the predictions obtained from the model. Specifically, we apply a layer-wise relevance propagation (LRP) (Bach et al., 2015) method to interpret how the network predicts precipitation events, that is, we assess the most important variables in the inputs that helped the model to make a specific prediction. Based on such relevance, we additionally test the effect of the number of input features on the network performance.

Table 1
Meteorological Variables Used by the Selected Studies

Study	Nb	SLP	Z	T	SH	RH	U/V	TCW/PW
Davenport and Diffenbaugh (2021)	2	1×	500	–	–	–	–	–
Huang (2022)	2	1×	500	–	–	–	–	–
Pan et al. (2019)	4	–	500	–	–	–	–	1×
			850					
			1,000					
Shi (2020)	30	–	300	300	–	300	300	–
			500	500		500	500	
			700	700		700	700	
			850	850		850	850	
			925	925		925	925	
			1000	1,000		1,000	1,000	

Note. The variables are: sea-level pressure (SLP), geopotential height (Z), air temperature (T), specific humidity (SH), relative humidity (RH), the zonal and meridional wind components (U/V), and the total column water vapor (TCW). The column "Nb" contains the number of variables used. The table values for Z, T, SH, RH, and U/V represent the pressure levels selected (hPa).

The rest of the paper is organized as follows: Section 2 discusses previous related work; the data and methods are introduced in Section 3; Section 4 shows the results, and the main conclusions are summarized in Section 5.

2. Related Works

Recently, many studies have proposed using sophisticated ML methods to improve precipitation estimates in various contexts, such as precipitation nowcasting (Ayzel et al., 2019) and post-processing of NWP precipitation output (Hess & Boers, 2022). This section reviews the most relevant studies closely related to our objectives and methodology.

Davenport and Diffenbaugh (2021) analyzed extreme precipitation days (above 95th percentile) over the U.S. Midwest and their links to large-scale atmospheric circulation patterns using a CNN with daily sea level pressure and geopotential height anomalies as input fields (Table 1). The model architecture consisted of two convolutional layers, each followed by a max-pooling layer, a dense 16-neuron layer, and a final classification layer of extreme and nonextreme precipitation days. The CNN showed high accuracy (91%) for the identification of extreme precipitation days, although some extreme events were not captured. The authors suggested that additional variables representing smaller-scale processes might improve the model performance. Moreover, due to the differences in the seasonal distribution of precipitation during extreme events, they pointed out the relevance of incorporating temporal information.

Building upon the work of Davenport and Diffenbaugh (2021), Huang (2022) proposed a self-attention augmented convolution mechanism for short-term extreme precipitation forecasting over the U.S. Midwest. The network consisted of two attention-augmented convolutional layers, a max-pooling, and a dropout layer. The proposed model outperformed classical convolutional models by 12%. However, a limitation to capturing some extreme events was acknowledged, likely due to localized processes for which additional information (e.g., variables) might be required.

Focusing on precipitation downscaling to point locations, Pan et al. (2019) proposed a CNN model as an alternative to parameterization schemes for numerical precipitation estimation. They built a CNN model based on convolutional and pooling layers using the geopotential height at several pressure levels and the total column water (TCW) as inputs (at a 3-hourly time step; see Table 1). The extracted features were flattened and processed by two final dense layers. The authors tested the CNN in different locations across the U.S. and showed that the CNN outperformed the reanalysis precipitation products and classical statistical methods. However, the model underestimated large precipitation values.

Similarly, Shi (2020) evaluated the performance of ML methods, including CNNs, for statistical downscaling of extreme precipitation in three Asian regions. They compared two DL architectures: RaNet, with three

convolutional layers and five fully connected layers, and RxNet, a more complex model with 58 layers, including residual connections similar to the original Xception model (Chollet, 2017). The results showed that deep CNN with an intermediate-level complexity structure (e.g., RaNet) generally performed better than a more complex architecture (e.g., RxNet). Moreover, while the CNNs well captured the precipitation extremes in the subtropical regions, they performed poorly in the tropical regions, illustrating the challenge of representing extreme precipitation in certain regions.

Adewoyin et al. (2021) developed TRU-NET (Temporal Recurrent U-Net), a DL model based on a U-Net (Section 3.2.1) architecture and featuring a novel 2D cross-attention mechanism to account for the spatio-temporal nature of weather processes. It relies on Convolutional Long Short-Term Memory cells, more specifically Convolutional Gated Recurrent Units. Their objective is to improve the prediction of high-resolution precipitation for climate models, which provide low-resolution outputs. They propose a fused temporal cross attention as a better aggregation strategy than averaging the six-hourly data to a daily time step. They show that TRU-NET outperforms other models, including U-Net, but notice that it under-predicts high precipitation events (>20 mm/day).

Recently, Hess and Boers (2022) showed that a U-Net-based network, using NWP ensemble simulations as input features, captures well heavy rainfall events. They applied DL as a post-processing step to correct biases in the NWP-predicted rainfall. They proposed a frequency-based weighting of the loss function that combines a continuously weighted mean squared error (MSE) with a multiscale structural similarity measure, which improved the training for high values when using both metrics separately.

3. Data and Methods

3.1. Data

The input variables and the precipitation fields were retrieved from the ERA5 (Hersbach et al., 2020) reanalysis. Reanalyses are produced using a single version of a data assimilation system coupled with a forecast model constrained to follow observations over a long period. They provide multivariate outputs that are physically consistent, also for variables that are not directly observed (Gelaro et al., 2017). ERA5 is the state-of-the-art reanalysis at the time of writing and was shown to outperform other reanalyses for predicting precipitation using a simpler statistical downscaling method (Horton, 2021). ERA5 provides data with high temporal (hourly) and spatial (0.25°) resolutions.

The weather variables used as input to the DL model should be robust, that is, not depend too much on the NWP or climate model, for the DL model to be transferable to other model outputs (Adewoyin et al., 2021). We thus selected frequently used variables: geopotential height (Z), air temperature (T), relative humidity (RH), TCW, and both wind components (U, V). All variables were selected at six pressure levels, that is, 300, 500, 700, 850, 925, and 1,000 hPa, except the TCW, which has a single vertical dimension. To reduce the computational costs of training all the networks (see Section 3.2), the spatial resolution of ERA5 data was degraded to 1°. Additionally, the variables were temporally aggregated at a daily time step. The domain on which these variables are used is: latitude = (30, 75) and longitude = (−25, 30) (Figure 1).

The precipitation data were also extracted from ERA5 over the same domain and spatial resolution (1°) and aggregated to a daily time step. Our study period is from 1979 to 2021. In this work, heavy precipitation events are identified based on the 95th percentile of the total distribution (1979–2021; the thresholds would be on average 0.1 mm lower on the training period) for each grid cell (i.e., pixel-wise definition). Similarly, extreme precipitation events are defined as those days exceeding the 99th percentile (Figure S1 in Supporting Information S1).

3.2. Methods

3.2.1. Deep Convolutional Neural Networks: Selected Architectures

CNNs have proven successful in different applications in climate science, including extreme weather forecasting (Liu et al., 2016; Racah et al., 2016), clustered weather patterns prediction (Chattopadhyay et al., 2020), precipitation nowcasting (Shi et al., 2015, 2017), or extreme precipitation (Davenport & Diffenbaugh, 2021; Shi, 2020). They are a type of neural network designed to process high-dimensional data, such as images or geospatial data (LeCun & Bengio, 1995). They have become tremendously popular due to their ability to automatically learn

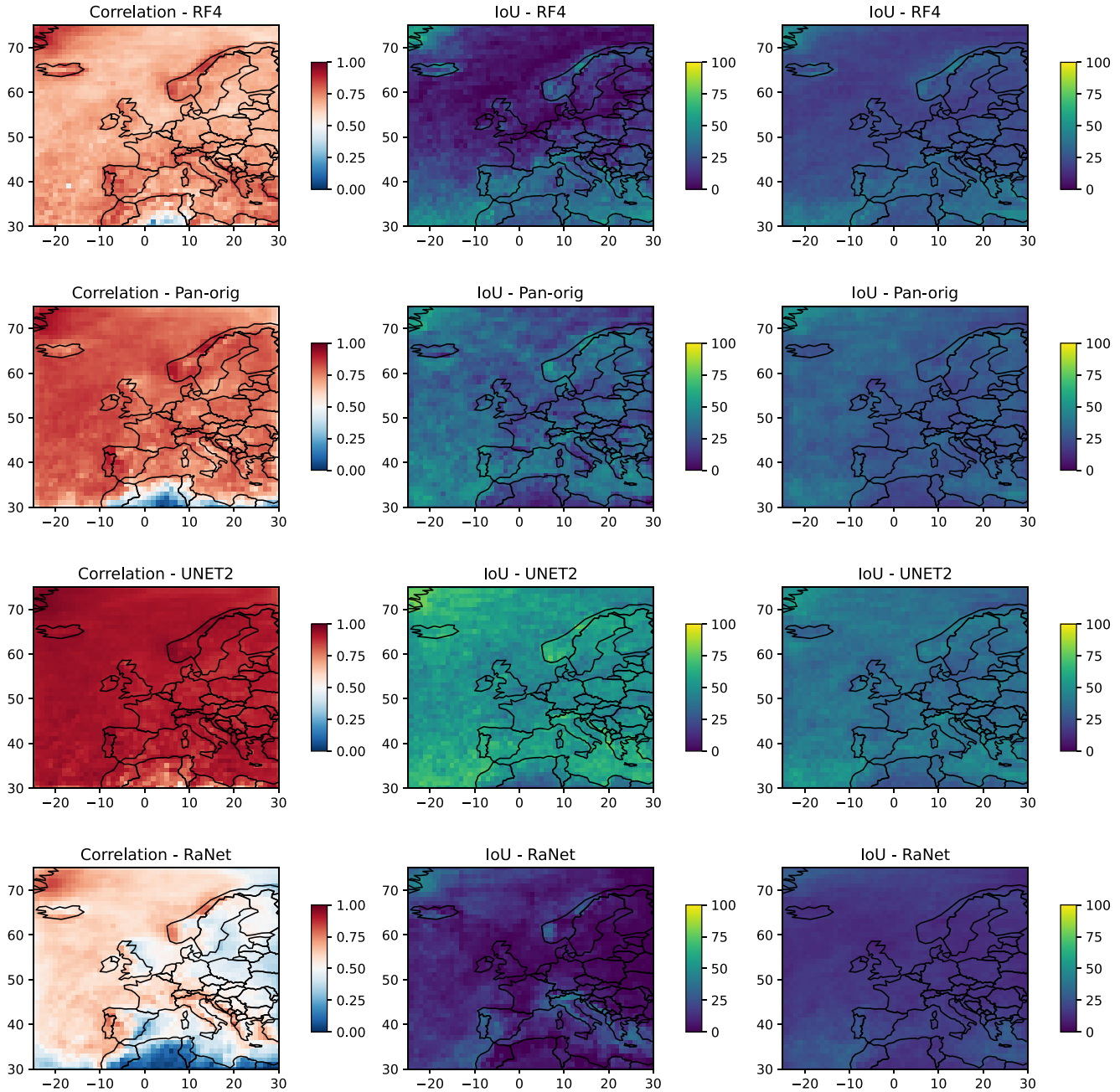

Figure 1. Correlation patterns between the ground truth and the predictions during the testing period (2016–2021) for the best architectures (left), intersection over union (IoU) for the occurrence of heavy precipitation events (from the regressor model) (middle), and IoU for heavy precipitation events (from the classifier model) with an exceeding probability >0.5.

spatial hierarchies of features, from low to high-level patterns (Goodfellow et al., 2016). The principle of CNN relies on a mathematical operation called *convolution*, a specialized linear operation used for feature extraction (Goodfellow et al., 2016). CNNs usually consist of three types of layers: (a) convolutional layers that perform the convolution operation, (b) pooling layers that reduce the dimensionality of the inputs, and (c) fully connected layers. The first two types of layers extract and condense the feature information used by dense layers. A typical CNN architecture is often composed of successive convolutional and pooling layers.

Building on CNNs, the popular U-Net, which was originally introduced by Ronneberger et al. (2015) for biomedical image segmentation, has shown good performance in climate applications, such as post-processing

Table 2
The Different Model Structures Assessed in This Work With the Number of Input Fields and Parameters

Type	Variant	Nb input fields	Nb parameters	Latent dimension
2D-CNN	Dav-orig	31	51,721	16
	Dav-64	31	178,105	64
	Pan-orig	31	238,054	60
	CNN-2l	31	746,345	64
U-Net	U-Net4	31	31,071,169	1,024
	U-Net3	31	7,724,481	512
	U-Net2	31	1,883,841	256
	U-Net1	31	421,697	128
	U-Net-att	31	37,514,533	1,024
3D-CNN	RaNet	30	1,859,627	256

weather forecasts (Grönquist et al., 2021; Hess & Boers, 2022), downscaling (e.g., Adewoyin et al., 2021) and precipitation nowcasting (e.g., Trebing et al., 2021). Larraondo et al. (2019) tested several encoder-decoder configurations and found the best results with U-Net-based architectures to forecast total precipitation using geopotential height as input. In Weyn et al. (2020), the authors used a U-Net architecture and mapped the input grid values to a cubed-sphere achieving a good performance to forecast complex surface temperature patterns from a few input atmospheric state variables. The U-Net architecture consists of two parts: a contracting path to capture the context (encoder) and a symmetric expanding path that enables precise localization (decoder). The encoder part is composed of stacked convolutions and pooling operations to extract the features, while the decoder part combines these features (through skip connections) with the upscaled output to reconstruct the spatial information. The encoder-decoder network enables propagating high-resolution features from the contracting path that are combined with the upscaled output (Ronneberger et al., 2015).

Recently, Trebing et al. (2021) proposed an adapted U-Net with a combination of attention modules and depthwise-separable convolutions for precipitation nowcasting. Introducing an attention mechanism into the convolutional neural network structure has also become popular in image segmentation processes (Oktay et al., 2018). In particular, the Attention U-Net proposed by Oktay et al. (2018) exploits the use of attention gates added to the encoder-decoder structure. This soft-attention mechanism is implemented for the skip connections. The attention gates actively suppress activations in irrelevant regions and, thus, reduce the number of redundant features. The authors showed that the use of attention gates improved the prediction performance of U-Net as the model learned to focus on useful features information, enhancing the accuracy of the network in locating tissues and organs, in the medical context.

Among the DL models presented in the literature for predicting precipitation, we have selected a number of representative studies closely related to our objectives. Given that our approach and model domain differ from the selected original studies, we have adapted the original architectures to our purpose. We classified the methods into three classes: 2D CNN models, U-Net architectures, and 3D CNN models. Table 1 summarizes the inputs originally used in the selected studies, and Table 2 lists the different structures assessed in this work and their corresponding number of parameters. Below, we briefly describe the models considered in our study.

In the 2D CNN class, we considered the model from Davenport and Diffenbaugh (2021) (hereafter named Dav-orig), which includes two convolutional layers with 16 3×3 filters, followed by two 2×2 max-pooling with a stride of 2. In the original configuration, a dense 16-neuron layer follows the convolution and max-pooling layers, followed by a final classification layer providing the probability of the outcomes. To predict a spatial precipitation field over the European domain, we added a decoder part made of a dense layer, two deconvolution layers, and a final convolution layer, symmetrically to the original model. Based on this architecture, we tested the use of a latent space of dimension 64 instead of 16 (Dav-64). Then, we considered the architecture from Pan et al. (2019, here named Pan-orig), which consisted of two convolutional and pooling layers followed by two consecutive dense layers. As in the previous model configurations, a symmetrical decoder part was added to keep the spatial dimensions. Following the architectures described above, we additionally tested a convolutional encoder-decoder made of two layers, with a latent space of dimension 64 (CNN-2l). Further experiments with additional layers were conducted but were not successful.

Then, based on the success shown by U-Net in diverse applications, we added this architecture to our analyses. We considered the original U-Net structure as proposed by Ronneberger et al. (2015), which consists of four levels (U-Net4), but also assessed shallower versions with three (U-Net3), two (U-Net2) and one (U-Net1) levels. The shallower versions have substantially fewer parameters to calibrate (Table 2). We also tested whether the inclusion of attention gates improves the accuracy of simulating extreme precipitation events (U-Net-att).

Finally, we included a 3D CNN, namely the RaNet architecture as proposed by Shi (2020). This model consists of three 3D CNN layers (using three-dimensional filters) and four fully connected layers, followed by a symmetric decoder part of upscaling layers that allows reconstructing the output into its original size. The use of this architecture allows us to further assess the performance of 3D CNN against 2D CNN.

3.2.2. Models Implementation

While our primary goal is to assess the model performance to reproduce precipitation extremes, we also tested the prediction of precipitation amounts. Therefore, the implemented models were assessed for different objectives: (a) for the prediction of the precipitation estimates, (b) for the occurrence of heavy precipitation (i.e., >95th percentile), and (c) for the occurrence of extremes (>99th percentile). The configuration of the models is the same in all cases, the only difference being the activation function of the last layer. A rectified linear unit (ReLU) that ensures nonnegative output values is used for predicting the precipitation amount, and a sigmoid is applied for predicting the probability of heavy/extreme events. It is important to note that the models were trained independently for the three different targets. The loss function used was the MSE for the prediction of precipitation estimates and the weighted binary cross-entropy for the occurrence of extremes (with weights computed to balance both classes). These scores were computed pixel-wise and averaged over the domain. An early stopping strategy has been used, with a maximum of 200 epochs. For all models, dropout and spatial dropout for the convolutional layers have been used.

A class was written in Python to generate the different model architectures with multiple options and handle common tasks, such as an eventual initial zero-padding when necessary and output cropping. The code is publicity available at https://github.com/ML-precip/precip-predict. It also sets the final activation layer to ReLU for the precipitation estimates or sigmoid for the probability of extremes. The models were implemented using Keras (Chollet et al., 2015) and designed according to the description in the related paper.

The input data is a tensor of shape $46 \times 56 \times 31$; 31 represents the number of atmospheric fields (i.e., channels): six fields for Z, RH, T, U, V, and one for TCW; 46×56 represents the spatial dimensions (latitude $\times$ longitude) of the domain considered. All models use the same number of channels (i.e., 31), except the RaNet model, for which TCW was excluded as 3D variables are required. The training period ranges from 1979 to 2005, and validation from 2006 to 2015. The testing period runs from 2016 to 2021. The model performance is assessed through standard metrics (see Section 4.1). It is important to note that there is no temporal lag between the predictors and the predictand, as our primary goal is to assess the ability of DL architectures to extract synoptic and mesoscale information and establish relationships to estimate precipitation.

3.2.3. Baseline Model

To compare the performance of the DL models with more traditional methods, a RF model (Breiman, 2001) was used as a baseline. The RF was fed with the same input data and trained/tested on the same periods as the DL architectures. As RF models do not predict spatial fields by nature, one model was here trained per pixel of the domain and then used to predict for that same pixel. Then, all predicted pixel-wise time series were aggregated into maps to provide daily fields.

As with the DL models, two different kinds of contexts were considered: the prediction of (a) the precipitation amount using a regressor RF and (b) the occurrence of heavy/extreme events (95/99th percentile) using a classifier. In the last case, the weights between event occurrence and non-occurrence were balanced. Different values of the maximum depth of RFs, which is an important parameter to avoid overfitting, have been tested, and the optimal one (4) was further used. Additional parameters, such as the number of trees (number of estimators) and the number of predictors (features) were set by default.

3.2.4. Feature Importance: Layer-Wise Relevance Propagation

We used a LRP approach, an XAI method applicable to ML models (e.g., Montavon et al., 2018), to better understand the importance of the input variables for heavy precipitation events, that is, which variables are more important for the network to make a prediction. Among the existing DL interpretation methods, LRP is a backward propagation technique used for explaining complex network outputs. The LRP creates heatmaps, which in our case help identify the most relevant regions of the input for predicting a heavy precipitation event (Barnes et al., 2022). Different LRP rules have been presented in recent literature for geoscience applications (e.g., Davenport & Diffenbaugh, 2021; Toms et al., 2020). Here, we tested three popular LRP rules, namely: (a) the LRP_z that redistributes the contributions of each input to the neuron activation as they occur (Montavon et al., 2018) (b) the $\text{LRP}_{\alpha-\beta}$ rule with $\alpha = 1$ and $\beta = 0$, which identifies locations for which higher activation values positively contribute to a likely output (i.e., predicted class); thus, with this formulation, only positive contributions to the neural network output are tracked; and (c) LRP composite (LRP_{comp}) (Kohlbrenner et al., 2020) that combines the LRP_z and the $\text{LRP}_{\alpha-\beta}$.

Preliminary analysis showed that in our case, the LRP_z rule lead to generally noisy results, which might happen when using deep networks (Mamalakis et al., 2022). While we observed similar heatmaps when comparing $LRP_{\alpha-\beta}$ and the LRP_{comp} rules, the LRP_{comp} exhibited the highest correlation with the ground truth (not shown). Thus, unless noted, the results presented hereinafter refer to the LRP_{comp} rule.

The LRP produces a map with the same dimensions as the input, where the pixel values indicate the importance of the predicted class. A total of 31 maps (i.e., 31 input variables) are obtained for each day. Then, we computed composite maps (for each input feature separately) by calculating for every pixel the average value of the relevance of a specific input feature for all days with an extreme event within the training period. To this end, we considered a larger area of influence for each pixel by calculating the averages of the maximum relevance within a small spatial domain for each feature when an event occurred:

$$\overline{R} = \frac{1}{N_i} \sum \max(R \pm z)$$

where z represents the number of the closest pixels to calculate the relevances at each grid cell. We performed additional sensitivity analyses for different values of z and decided to use $z = 3$ as a good compromise to account for local processes that might be relevant for pixel-wise precipitation events. It is important to note that the averages of the relevances were calculated for the *true* extremes.

As detailed below (see Section 4.3), after selecting our best model for predicting precipitation, we apply the most suitable LRP rule to examine the importance of the input features for simulating heavy precipitation events. Based on the relevance values obtained for the training sample, we ranked the predictors by their average relevance. These values were obtained by averaging the composite maps produced for each input feature. Then, we conducted a number of experiments for differing subsets of predictors to examine the role of the number of features in the network performance.

4. Results

4.1. Networks Performance

We trained the models separately for predicting precipitation amounts (i.e., as a regression task) and precipitation events (i.e., as a classification task). In the first case, we assessed the prediction of the precipitation amount with the RMSE, and we further estimated the predicted threshold exceedances (95th and 99th percentile for each pixel) to compute the precision and recall scores (Table 3 for the 95th percentile and Table 4 for the 99th percentile). The U-Net outperformed the rest of the models for predicting precipitation amounts (lower RMSE) and provided the highest precision and recall scores when assessing the threshold exceedances. Its optimization also stops earlier than other models (not shown). However, while using an attention gate in U-Net showed an improvement for medical image data sets (Oktay et al., 2018), this was not the case in our application, as the results showed similar performances with or without the attention gates, with slightly improved precision observed in extreme precipitation case (see Table 4). Therefore, the attention gates were not further used in the following analyses.

The prediction skills for heavy and extreme precipitation events were evaluated in terms of the AUC (ROC under curve area), precision, and recall scores based on a probability threshold of 0.5. Tables 5 and 6 show the score values obtained for classifying both heavy (>95th) and extreme (>99th) precipitation events.

Similarly to the regression case, the results show clearly that U-Net, which has significantly more trainable parameters, is the best to predict precipitation extremes. However, a difference between both settings becomes obvious: when trained for the prediction of extremes, the model's outputs result in a much higher recall than when trained for the precipitation amount while presenting a lower precision. The models trained for the extremes predict them better than when trained for the whole precipitation distribution (Table 4), but overestimate the number of extreme events (i.e., Table 6). It can be expected that balancing the weights differently in the weighted binary cross-entropy will result in other recall and precision scores.

When comparing the variants per type, Pan-orig performs better than other 2D CNN variants, and different U-Net variants show similar scores, depending on the metric used. We have selected U-Net2 due to its ability to provide similar performance with fewer parameters.

We further analyze the ability of the DL models to represent the spatial distribution of precipitation events realistically. To do so, we first select the best network from each type (see Table 2) according to the scores obtained above.

Table 3
Scores of the Tested Models When Trained to Predict the Precipitation Amount

Type	Variant id	RMSE train	RMSE test	Precision train	Precision test	Recall train	Recall test
Random forest	RF4	2.66	2.93	0.73	0.66	0.27	0.23
2D-CNN	Dav-orig	3.19	3.33	0.55	0.51	0.21	0.19
	Dav-64	2.73	2.91	0.60	0.58	0.42	0.38
	Pan-orig	2.44	2.59	0.73	0.71	0.40	0.37
	CNN-2l	2.30	2.65	0.69	0.61	0.53	0.45
U-Net	U-Net4	1.43	1.72	0.79	0.76	0.70	0.65
	U-Net3	**1.40**	1.74	0.79	0.75	**0.74**	**0.67**
	U-Net2	1.41	**1.69**	**0.83**	0.80	0.66	0.62
	U-Net1	1.58	1.75	0.82	**0.81**	0.63	0.60
	U-Net-att	1.42	1.73	0.80	0.77	0.71	0.65
3D-CNN	RaNet	3.01	3.30	0.60	0.53	0.16	0.14

Note. Precision and recall are computed for the exceedance of the 95th percentile. The best scores are highlighted in bold.

Then, we calculated the correlation patterns between the model outputs and the truth to quantify the ability of the networks to capture the spatial distribution of precipitation amounts. Additionally, the metric intersection over union (IoU) to evaluate the model's performance to classify heavy events (Prabhat et al., 2021; Wang et al., 2023) is calculated. The IoU was also quantified by considering an extreme event to be occurring for any probability higher than 0.5. The U-Net architecture shows the highest correlation values, followed by the Pan-orig and the RF models (see Figure 1). The RaNet architecture shows the lowest correlation values, pointing out a lower ability of the network to represent here the spatial distribution of precipitation. Similar conclusions can be derived from the IoU (Figure 1): the highest values correspond to the U-Net architecture. Table 7 shows the spatial averages of the metrics illustrated in Figure 1. In agreement, the spatial distribution of the metrics calculated for the extreme precipitation case showed the highest correlation and IoU values for the U-Net architecture (Figure S2 in Supporting Information S1).

Next, we also examine the predictions of the different models for the day with the highest amount of observed precipitation exceeding the 95th and the 99th percentiles during the test period and over the considered domain, that is, we selected the day (13 October 2018) with the highest number of grid cells exceeding the respective percentile. The meteorological context during that day was characterized by a large long-lived extratropical cyclone in the Atlantic that became a powerful posttropical system resulting in heavy precipitation in several regions in Western Europe. As in Figure 1, this analysis is performed for the best DL model.

Table 4
Scores of the Tested Models When Trained to Predict the Precipitation Amount

Type	Variant id	RMSE train	RMSE test	Precision train	Precision test	Recall train	Recall test
Random forest	RF4	2.66	2.93	0.67	0.28	0.09	0.05
2D-CNN	Dav-orig	3.19	3.33	0.33	0.14	0.02	0.02
	Dav-64	2.73	2.91	0.58	0.45	0.14	0.10
	Pan-orig	2.44	2.59	0.65	0.61	0.29	0.25
	CNN-2l	2.30	2.65	0.69	0.57	0.31	0.21
U-Net	U-Net4	1.43	1.72	0.83	0.77	0.56	0.46
	U-Net3	1.40	1.74	0.81	0.76	0.56	0.47
	U-Net2	1.41	**1.69**	0.83	0.78	**0.58**	**0.48**
	U-Net1	1.58	1.75	0.81	**0.79**	0.47	0.42
	U-Net-att	**1.37**	1.73	**0.86**	**0.79**	0.55	0.41
3D-CNN	RaNet	3.01	3.30	0.57	0.31	0.07	0.03

Note. Precision and recall are computed for the exceedance of the 99th percentile. The best scores are highlighted in bold.

Table 5
Scores of the Tested Models When Trained to Predict Precipitation Extremes

Type	Variant id	AUC train	AUC test	Precision train	Precision test	Recall train	Recall test
Random forest	RF4	0.90	0.86	0.27	0.27	0.93	0.85
2D-CNN	Dav-orig	0.90	0.89	0.18	0.18	0.86	0.83
	Dav-64	0.95	0.93	0.24	0.24	0.92	0.88
	Pan-orig	0.96	0.95	0.30	0.30	0.92	0.89
	CNN-2l	0.97	0.94	0.26	0.25	0.96	0.89
U-Net	U-Net4	**0.99**	**0.98**	0.36	0.36	**0.99**	**0.96**
	U-Net3	**0.99**	**0.98**	**0.39**	**0.39**	**0.99**	0.95
	U-Net2	**0.99**	**0.98**	0.38	0.38	**0.99**	**0.96**
	U-Net1	**0.99**	**0.98**	**0.39**	**0.39**	0.98	**0.96**
	U-Net-att	**0.99**	**0.98**	0.38	0.38	0.98	0.95
3D-CNN	RaNet	0.90	0.88	0.18	0.18	0.89	0.84

Note. Precision and recall are computed for the exceedance of the 95th percentile. The best scores are highlighted in bold.

Figures 2 and 3 show the results of the models trained for the prediction of the precipitation amount (two first columns) and the results of the models trained for the prediction of the occurrence of extremes (last column). From Figure 2 it can be seen that, in general, most of the models simulate fairly well heavy precipitation events, and most of them show consistent patterns when compared with the truth (i.e., ERA5). The differences between the models become larger when comparing their performance in capturing extreme precipitation events (Figure 3). While the overall scores obtained for the baseline RF model show a close performance to some of the best DL architectures (e.g., Pan-orig), the RF represents poorly the spatial distribution of the selected precipitation event, compared to the DL models. This highlights the ability of CNN to extract the spatial information, being more efficient in treating complex spatial features. In that case, it can be observed that a U-Net-based architecture (UNET2) is superior and reproduces the closest pattern to the *truth*. In agreement with the skill scores in Tables 3–6, the UNET2 outperforms the rest of the models for both the amount of precipitation and the threshold exceedances.

4.2. Assessment of U-Net Variants

Motivated by the good performance of U-Net in simulating precipitation events, we conducted further experiments to assess the predictive capabilities of several U-Net-based architectures for precipitation events.

Table 6
Scores of the Tested Models When Trained to Predict Precipitation Extremes

Type	Variant id train	AUC test	AUC train	Precision test	Precision train	Recall test	Recall
Random forest	RF4	0.90	0.89	0.05	0.06	0.98	0.95
2D-CNN	Dav-orig	0.94	0.92	0.05	0.05	0.90	0.85
	Dav-64	0.97	0.96	0.08	0.08	0.94	0.88
	Pan-orig	0.98	0.97	0.09	0.09	0.98	0.94
	CNN-2l	0.97	0.94	0.08	0.07	0.96	0.88
U-Net	U-Net4	**0.99**	**0.99**	**0.17**	**0.18**	**0.99**	0.97
	U-Net3	**0.99**	**0.99**	**0.17**	0.17	**0.99**	0.97
	U-Net2	**0.99**	**0.99**	0.16	0.16	**0.99**	**0.98**
	U-Net1	**0.99**	**0.99**	**0.17**	0.17	**0.99**	0.97
	U-Net-att	0.99	0.98	0.16	0.16	0.99	0.97
3D-CNN	RaNet	0.93	0.89	0.05	0.05	0.92	0.80

Note. Precision and recall are computed for the exceedance of the 99th percentile. The best scores are highlighted in bold.

Table 7
Average Values of the Metrics Illustrated in Figure 5: Correlation Coefficients, Intersection Over Union (IoU) for the Occurrence of Heavy Precipitation Events (From the Regressor Model) and IoU for Heavy Precipitation Events (From the Classifier Model) Exceeding a Probability >0.5

Model	Correlation	IoU	IoU
Random forest	0.70	19.52	31.62
Pan-orig	0.76	36.20	28.78
U-Net2	**0.90**	**51.51**	**39.27**
RaNet	0.52	10.56	17.05

Note. The best scores are highlighted in bold.

4.2.1. Sensitivity to the Number of Input Features

In this section, we further explore the role of the number of inputs. Inputs that may bring redundant or unnecessary information can increase the complexity of the model while not improving its skill. To do so, we conducted several sensitivity analyses to test whether reducing the number of inputs would lead to comparable results to the full model that uses 31 input features with a considerable number of trainable parameters (see Table 2). One of the greatest potentials of XAI methods is that they highlight the most important features that help the model make accurate predictions. Therefore, we rely on the relevance values obtained from the LRP_{comp} to apply a forward selection procedure. We can expect the model performance to increase as the relevant features identified by the LRP_{comp} are added. Our aim is to analyze the necessary number of inputs that would lead to a reasonably good model performance while reducing computational efforts.

A typical forward/backward stepwise selection procedure where the predictors are included/removed one at a time would be computationally expensive. Thus, the predictors were included in the models five at a time according to the ranking provided by the LRP. We first explore the sensitivity of the XAI method to the depth of the U-Net networks (see Figure 4). Overall, we found a similar ranking of the most important features according to the XAI method when applied to the different U-Net depths.

Next, we assess the model performance in terms of number of inputs. To keep this analysis concise and to limit computational costs, the forward selection procedure is only applied to the selected network (U-Net2), which is characterized by good performances and a lower number of parameters. The U-Net2 is then trained separately for six predictor subsets.

As stated in the previous section, we evaluate the prediction skill of heavy precipitation events through the categorical skill scores commonly used for classification problems and can be obtained from the contingency table. The AUC, precision, and recall scores were calculated for both training and test data sets. Figure 5 illustrates the results corresponding to the U-Net architectures used in the experiments for different subsets of predictors. It can be observed that the performance of the network is considerably lower for a small group of features and improves when increasing the number of predictors. Overall, we observed a steady improvement in the model performance when increasing the number of input features up to an optimal around 20. Additional inputs are found less relevant as they do not contribute to improving the prediction. It should be noted that these optimums likely depend on the random seed, and some variability is expected between different random seeds. These results show anyway that more data does not always mean better performance. Based on Figures 4 and 5, the predictors that are less informative for the precipitation prediction are the air temperature (T) and the geopotential height (Z). While this last one is often considered as key in weather forecasting, it is here superseded by the wind components that do also contain information on atmospheric circulation.

4.3. Interpretability: LRP

The LRP_{comp} previously used was also mapped to visualize which features are important for the U-Net network to predict heavy precipitation events in the different geographical regions. Note that for the interpretability analysis, we focus on the network that showed the best performance (i.e., UNET2). We first examined the composite LRP maps (Section 3.2.4) for all heavy events occurring during the training period (1979–2005). These maps highlight the relevant features at a pixel scale for predicting heavy precipitation at that same pixel (Figure 6). Note that a zero-input baseline predicts a very small precipitation amount (0.3 mm max; 0.086 mm on average), which indicates no spatial structure except some noise and border effects and a probability of occurrence of extreme events equal to zero (not shown).

From Figure 6, it can be observed that some features show an extended area of relevance inland (e.g., RH fields). Overall, the RH shows the highest values, followed by both wind components, particularly in western and southern Europe. The high relevance of the wind components in some areas reflects the dependence between extreme precipitation events and strong wind conditions due to the same mesoscale and/or synoptic features, as shown by previous studies (Martius et al., 2016). For example, one can observe the higher relevance values of the zonal and meridional wind in the Iberian Peninsula, which often experiences concurrent extreme precipitation and winds

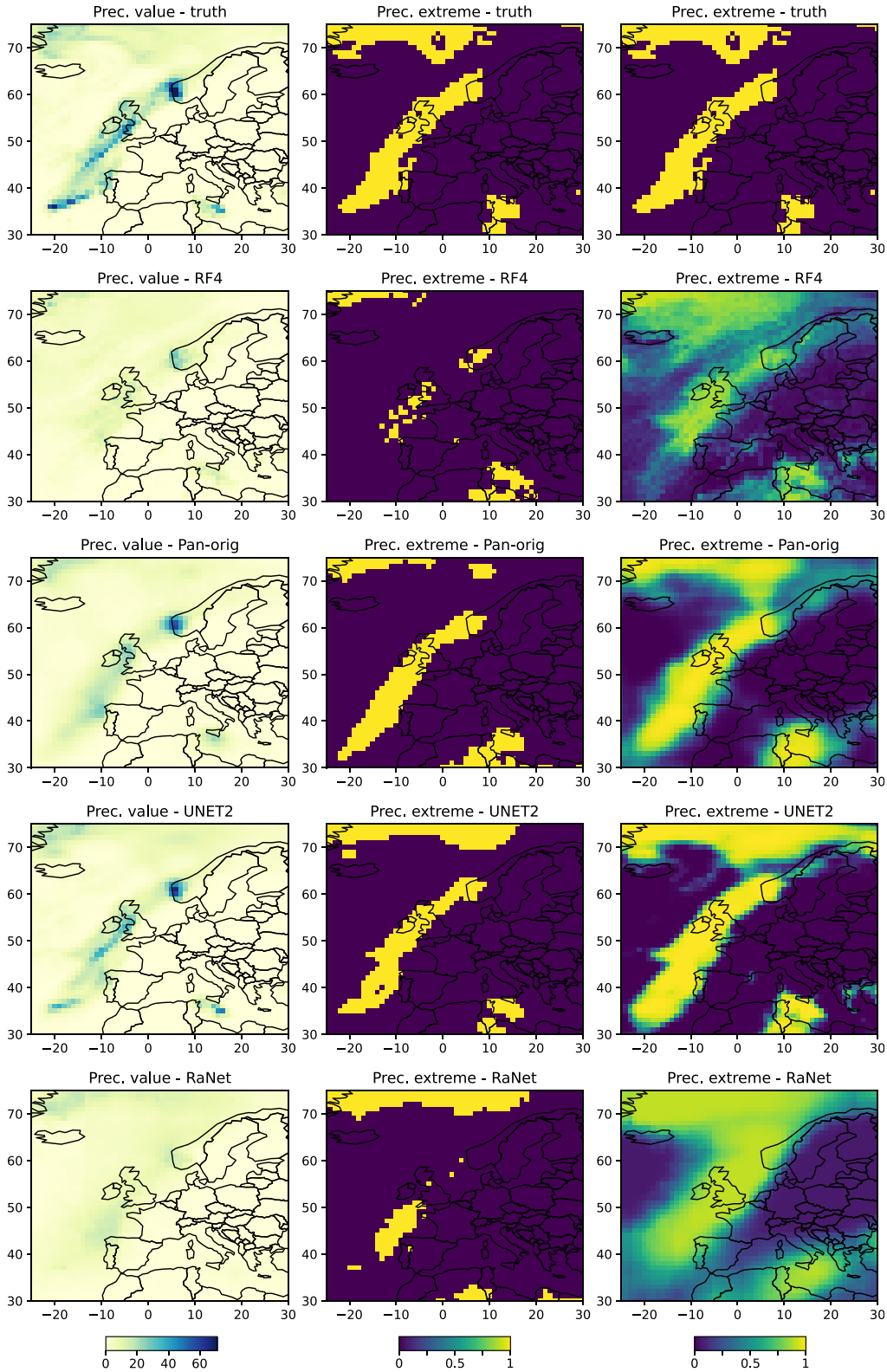

Figure 2. First row: true values of the precipitation amount and the corresponding threshold exceedance for the 95th percentile. Next rows: the prediction of each model for the same date, in terms of precipitation amount (first column), the corresponding threshold exceedance (second column), and the probability of the occurrence of heavy precipitation (third column).

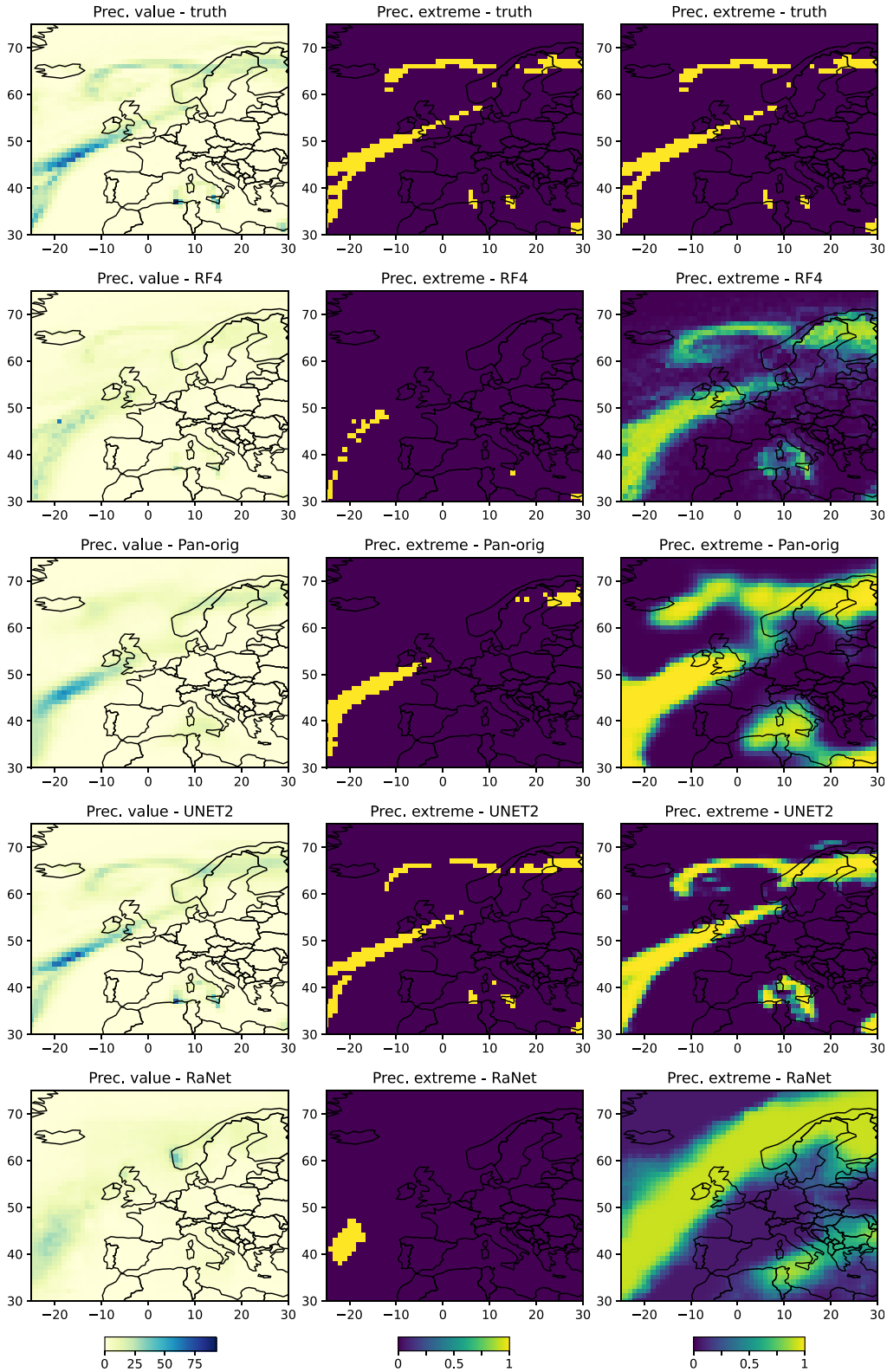

Figure 3. First row: true values of the precipitation amount and the corresponding threshold exceedance for the 99th percentile. Next rows: the prediction of each model for the same date, in terms of precipitation amount (first column), the corresponding threshold exceedance (second column), and the probability of the occurrence of extreme precipitation (third column).

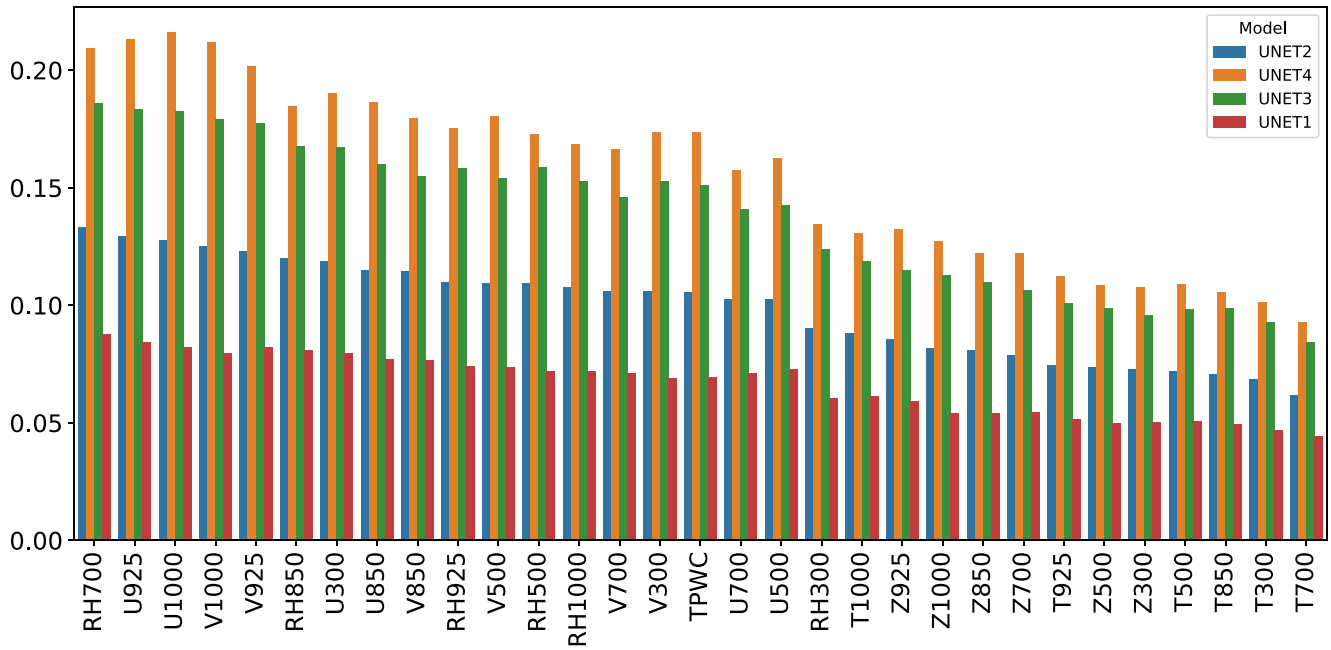

Figure 4. Ranked relevances (averages) obtained for heavy precipitation events in the training sample (1979–2005) for each U-Net architecture (Table 2).

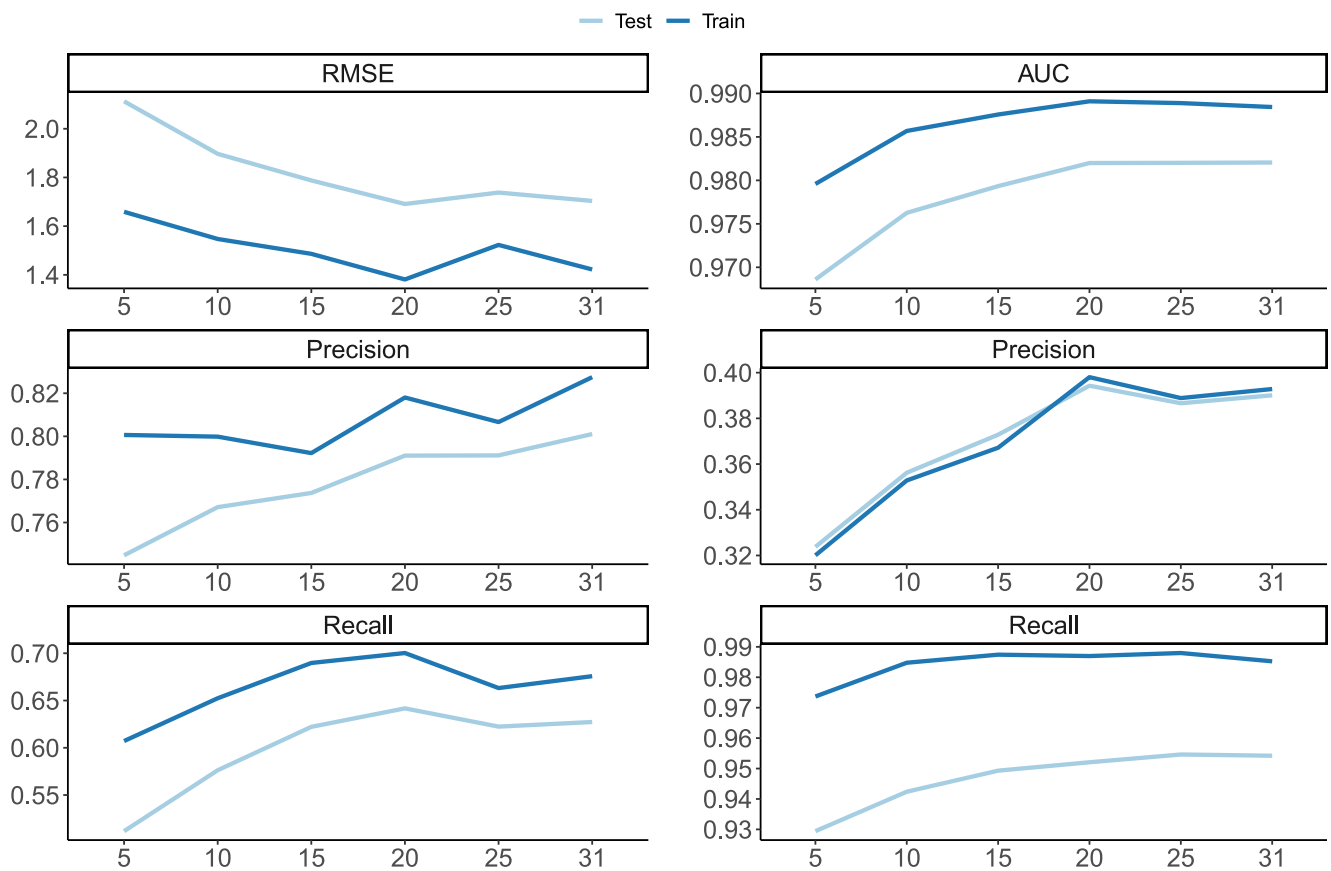

Figure 5. Scores obtained using different input subsets for the selected best U-Net, UNET2. Left plots show the scores corresponding to the model performance to predict precipitation amounts. Right plots show the scores of the model performance to predict heavy precipitation.

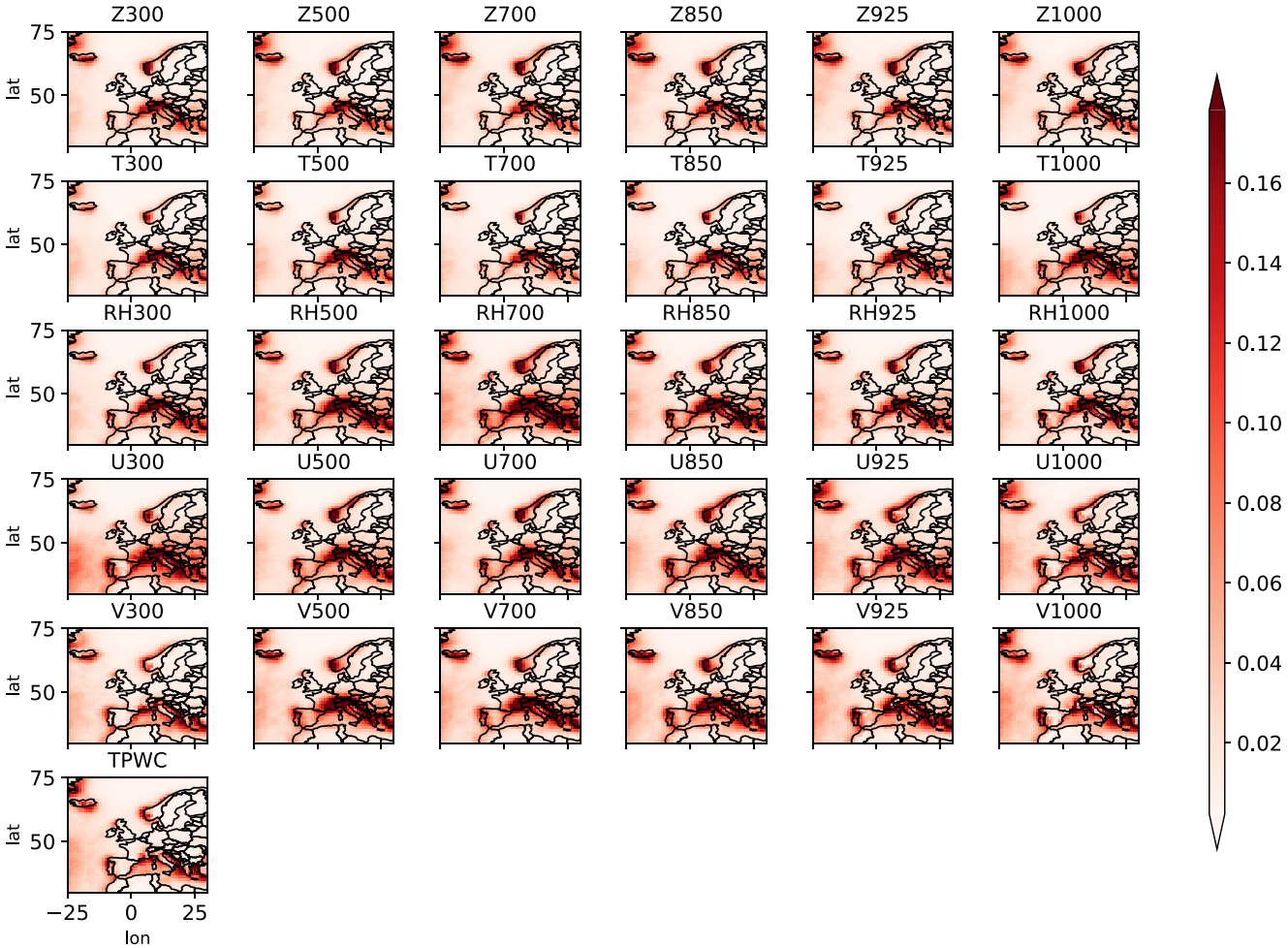

Figure 6. Composite relevance maps for heavy precipitation events (>95th percentile) derived from the best U-Net architecture during the training period (1979–2005) for each feature.

conditions, mostly related to extratropical cyclones and their atmospheric fronts (Hénin et al., 2021). We can also distinguish the relevance of the wind components for the alpine region, which is known to be related to heavy precipitation events due to the orographic forcing of air masses that transport moisture from the Mediterranean. This influence of the atmospheric circulation comes in pair with the moisture information, heavily represented by the RH variable at 700 hPa (RH700).

Additionally, we examined the predictors relevance for specific extreme precipitation events (see Figures S3 and S4 in Supporting Information S1). In these examples, the network only highlights one area of relevance where the inputs have major weights. Such areas of relevance appear to be associated with the location of the highest relative precipitation amount (e.g., Figure S4 in Supporting Information S1), that is, the network finds the most relevant geographical regions at the same location as the heavy precipitation event evolves. This is also reasonable as a shallower U-Net (two encode-decoder levels) was used, and local relevant areas are used. Moreover, the LRP analysis indicates that the local predictors contain enough information to predict the event and that no remote information is needed.

5. Conclusions

The use of ML has exponentially grown in the past years in a wide number of fields. In particular, DL methods have shown enormous potential to address complex Earth Science problems, which might be useful in tackling climate change-related issues. Here, we have presented an intercomparison of existing architectures used to predict precipitation, either for aggregated precipitation or spatial precipitation fields. Due to the wide range of DL architectures, we here focused on a selected set of networks. The architectures used were classified into

three classes: 2D CNN models, U-Net architectures, and 3D CNN models. To benchmark the performance of the selected DL architectures, we applied a point-wise Random Forest as a baseline model. We examined the prediction skill not only to simulate heavy and extreme precipitation events but also to predict the amount of precipitation over the European domain. For the interpretability of the networks, we applied a layer-wise propagation technique, which was further used as a tool for feature selection to test the importance of the number of input parameters on the model performance. It is important to note that while some of these DL topologies have been previously presented in the literature, the original application slightly differs from ours, and more importantly, the original configuration was adapted to our purposes (e.g., in each case, we added a decoder part to preserve the spatial dimensions of the input data).

In general, most of the analyzed DL were able to reproduce reasonably well the occurrence of precipitation events. However, we found that U-Net-based networks generally outperformed the rest of the tested architectures by a large margin, which is in line with previous studies (Hess & Boers, 2022; Larraondo et al., 2019) that used a U-Net architecture to simulate precipitation. In general, the skill scores that measure the precision to classify heavy precipitation events (i.e., >95th percentile) were higher than those obtained for extreme precipitation events (i.e., >99th percentile), due to the unbalanced number of classes where the number of extremes is significantly reduced in the training data.

While the original U-Net already showed a good performance, we found that a shallower network, in terms of number levels compared to the original architecture, would be sufficient to classify heavy precipitation events correctly. This likely means that, for this spatial resolution and with no temporal extrapolation, most of the information needed to predict precipitation is available at the location where the precipitation occurs. Our results showed that in such a context, a shallower U-Net, which significantly reduces the number of trainable parameters and the computation time, is able to predict precipitation events fairly well.

We additionally conducted a number of experiments on U-Net-based configurations to examine how the number of inputs plays a role in the performance of the model. As expected, the network showed the poorest performance when using only a few input variables. The optimal number of inputs for this application seems to be around 20, after which no substantial improvement was found.

We hope that our comparison analysis can help guide future studies on precipitation prediction in the choice of the DL architecture and can highlight the good performance of U-Net networks in predicting precipitation fields and the probability of extremes. Additionally, a shallower U-Net was found to perform as well as the original architecture while having substantially fewer parameters. The use of LRP for the selection of the most relevant predictors is an approach that can also be used in other contexts, and that is less computationally intensive than typical forward/backward stepwise selection procedures. We plan to extend this work by focusing on downscaling tasks and exploring alternative and state-of-the-art architectures, including the utilization of GANs.

Data Availability Statement

The ERA5 data is available for download at the Copernicus Climate Change Service (Hersbach et al., 2020; Store, 2023). The code used for the analysis is available in Horton and Otero (2023).

Acknowledgments
We thank the editor and the anonymous reviewers for their constructive comments and insightful suggestions that led to the improvement of the manuscript. We acknowledge the support of the HPC cluster at the University of Bern, UBELIX (http://www.id.unibe.ch/hpc), where all the calculations were performed.

References

Adewoyin, R. A., Dueben, P., Watson, P., He, Y., & Dutta, R. (2021). TRU-NET: A deep learning approach to high resolution prediction of rainfall. *Machine Learning*, *110*(8), 2035–2062. https://doi.org/10.1007/s10994-021-06022-6

Ayzel, G., Heistermann, M., Sorokin, A., Nikitin, O., & Lukyanova, O. (2019). All convolutional neural networks for radar-based precipitation nowcasting. *Procedia Computer Science*, *150*, 186–192. https://doi.org/10.1016/j.procs.2019.02.036

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Barnes, E. A., Barnes, R. J., Martin, Z. K., & Rader, J. K. (2022). This looks like that there: Interpretable neural networks for image tasks when location matters. *Earth and Space Science Open Archive*, *34*. https://doi.org/10.1002/essoar.10509984.2

Bauer, P., Thorpe, A., & Bruner, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*, 47–55. https://doi.org/10.1038/nature14956

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

Chattopadhyay, A., Hassanzadeh, P., & Pasha, S. (2020). Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. *Scientific Reports*, *10*, 1317. https://doi.org/10.1038/s41598-020-57897-9

Chollet, F. (2015). Keras. Retrieved from https://github.com/fchollet/keras.GitHub

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1800–1807). https://doi.org/10.1109/CVPR.2017.195

Civitarese, D. S., Szwarcman, D., Zadrozny, B., & Watson, C. (2021). Extreme precipitation seasonal forecast using a transformer neural network. *arXiv:2107.06846*. Retrieved from http://arxiv.org/abs/2107.06846

Dabrowski, J. J., Zhang, Y., & Rahman, A. (2020). ForecastNet: A time-variant deep feed-forward neural network architecture for multi-step-ahead time-series forecasting. *arXiv:2002.04155*. Retrieved from https://arxiv.org/abs/2002.04155

Davenport, F. V., & Diffenbaugh, N. S. (2021). Using machine learning to analyze physical causes of climate change: A case study of U.S. Midwest extreme precipitation. *Geophysical Research Letters*, *48*(15). e2021GL093787. https://doi.org/10.1029/2021GL093787

Donat, M., Lowry, A., Alexander, L., O'Gorman, P. A., & Maher, N. (2016). More extreme precipitation in the world's dry and wet regions. *Nature Climate Change*, *6*, 508–513. https://doi.org/10.1038/nclimate2941

Gao, X., & Adam Schlosser, C. (2019). Mid-western US heavy summer-precipitation in regional and global climate models: The impact on model skill and consensus through an analogue lens. *Climate Dynamics*, *6*, 1569–1582. https://doi.org/10.1007/s00382-018-4209-0

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. Retrieved from http://www.deeplearningbook.org

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *379*(2194), 20200092. https://doi.org/10.1098/rsta.2020.0092

Harris, L., McRae, A. T. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS003120. https://doi.org/10.1029/2022MS003120

Hénin, R., Ramos, A. M., Pinto, J. G., & Liberato, M. L. R. (2021). A ranking of concurrent precipitation and wind events for the Iberian Peninsula. *International Journal of Climatology*, *41*(2), 1421–1437. https://doi.org/10.1002/joc.6829

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis [Dataset]. Quarterly Journal of the Royal Meteorological Society, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hess, P., & Boers, N. (2022). Deep learning for improving numerical weather prediction of heavy rainfall. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002765. https://doi.org/10.1029/2021MS002765

Hill, A. J., & Schumacher, R. S. (2022). Forecasting excessive rainfall with random forests and a deterministic convection-allowing model, weather and forecasting. *Weather and Forecasting*, *36*(5), 1693–1711. https://doi.org/10.1175/WAF-D-21-0026.1

Hill, G. R., Herman, A. J., & Schumacher, R. S. (2022). Forecasting severe weather with random forests. *Monthly Weather Review*, *148*(5), 2135–2161. https://doi.org/10.1175/MWR-D-19-0344.1

Horton, P. (2021). Analogue methods and ERA5: Benefits and pitfalls. *International Journal of Climatology*, *42*, 4078–4096. https://doi.org/10.1002/joc.7484

Horton, P., & Otero, N. (2023). Ml-precip/precip-predict: V1.0.0 (v1.0.0) [Software]. Zenodo. https://doi.org/10.5281/zenodo.7847311

Huang, W. (2022). Extreme precipitation forecasting using attention augmented convolutions. *arXiv:2201.13408*. Retrieved from https://arxiv.org/abs/2201.13408

Hwang, J., Orenstein, P., Cohen, J., Pfeiffer, K., & Mackey, L. (2019). Improving subseasonal forecasting in the western U.S. with machine learning. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2325–2335). https://doi.org/10.1145/3292500.3330674

Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., & Lapuschkin, S. (2020). Towards best practice in explaining neural network decisions with LRP. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–7). https://doi.org/10.1109/IJCNN48605.2020.9206975

Larraondo, P. R., Renzullo, L. J., Inza, I., & Lozano, J. A. (2019). A data-driven approach to precipitation parameterizations using convolutional encoder-decoder neural networks. *arXiv:1903.10274*. https://doi.org/10.48550/ARXIV.1903.10274

LeCun, Y., & Bengio, Y. (1995). *Convolutional networks for images, speech, and time series, the handbook of brain theory and neural networks* (pp. 255–258). MIT Press.

Leinonen, J., Nerini, D., & Berne, A. (2021). Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(9), 7211–7223. https://doi.org/10.1109/TGRS.2020.3032790

Liu, Y., Racah, E., Prabhat, Correa, J., Khosrowshahi, A., Lavers, D., et al. (2016). Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv:1605.01156*.

Mamalakis, A., Barnes, E. A., & Ebert-Uphoff, I. (2022). Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems*, *1*(4), e220012. https://doi.org/10.1175/AIES-D-22-0012.1

Martius, O., Pfahl, S., & Chevalier, C. (2016). A global quantification of compound precipitation and wind extremes. *Geophysical Research Letters*, *43*(14), 7709–7717. https://doi.org/10.1002/2016GL070017

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). Climax: A foundation model for weather and climate.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-net: Learning where to look for the pancreas. *arXiv:1804.03999*. https://doi.org/10.48550/ARXIV.1804.03999

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J. W., Lee, J., et al. (2021). Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002509. https://doi.org/10.1029/2021MS002509

Pan, B., Hsu, K., AghaKouchak, A., & Sorooshian, S. (2019). Improving precipitation estimation using convolutional neural network. *Water Resources Research*, *55*(3), 2301–2321. https://doi.org/10.1029/2018WR024090

Prabhat, Kashinath, K., Mudigonda, M., Kim, S., Kapp-Schwoerer, L., Graubner, A., et al. (2021). ClimateNet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development*, *14*(1), 107–124. https://doi.org/10.5194/gmd-14-107-2021

Racah, E., Beckham, C., Maharaj, T., Prabhat, & Pal, C. J. (2016). Semi-supervised detection of extreme weather events in large climate datasets. *arXiv:1612.02095*. Retrieved from http://arxiv.org/abs/1612.02095

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002203. https://doi.org/10.1029/2020ms002203

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*, 672–677. https://doi.org/10.1038/s41586-021-03854-z

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015* (pp. 234–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28

Scher, S. (2018). Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, *45*(22), 12616–12622. https://doi.org/10.1029/2018GL080704

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., et al. (2021). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *379*(2194). https://doi.org/10.1098/rsta.2020.0097

Shi, X. (2020). Enabling smart dynamical downscaling of extreme precipitation events with machine learning. *Geophysical Research Letters*, *47*(19), e2020GL090309. https://doi.org/10.1029/2020GL090309

Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., & Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *arXiv:1506.04214*. Retrieved from http://arxiv.org/abs/1506.04214

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D., Wong, W., & Woo, W. (2017). Deep learning for precipitation nowcasting: A benchmark and a new model. *arXiv:1706.03458*. Retrieved from http://arxiv.org/abs/1706.03458

Store, C. D. (2023). Climate data store. Retrieved from https://cds.climate.copernicus.eu/cdsapp#!/search?type=dataset

Toms, B. A., Barnes, E. A., & Imme, E.-U. (2020). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. https://doi.org/10.1029/2019MS002002

Trebing, K., Stańczyk, T., & Mehrkanoon, S. (2021). SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture. *Pattern Recognition Letters*, *145*, 178–186. https://doi.org/10.1016/j.patrec.2021.01.036

Trenberth, K., Dai, A., Rasmussen, R. M., & Parsons, D. B. (2003). The changing character of precipitation. *Bulletin of the American Meteorological Society*, *84*, 1205–1218. https://doi.org/10.1175/BAMS-84-9-1205

Vandal, T., Kodra, E., & Ganguly, A. R. (2019). Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation. *Theoretical and Applied Climatology*, *137*(1–2), 557–570. https://doi.org/10.1007/s00704-018-2613-3

Wang, F., Tian, D., & Carroll, M. (2023). Customized deep learning for precipitation bias correction and downscaling. *Geoscientific Model Development*, *16*(2), 535–556. https://doi.org/10.5194/gmd-16-535-2023

Weyn, J. A., Durran, D. R., & Caruana, R. (2019). Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2680–2693. https://doi.org/10.1029/2019MS001705

Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002109. https://doi.org/10.1029/2020MS002109

Wolfensberger, D., Gabella, M., Boscacci, M., Germann, U., & Berne, A. (2021). Rainforest: A random forest algorithm for quantitative precipitation estimation over Switzerland. *Atmospheric Measurement Techniques*, *14*, 3169–3193. https://doi.org/10.5194/amt-14-3169-2021