https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

Spatially explicit assessment of machine learning post-processing of daily precipitation forecasts – a multimodel approach

Gabriel Stachura[1,2,3], Bogdan Bochenek[2], Piotr Sekuła[2], Anna Jurczyk[2], Zbigniew Ustrnul[2,3]

[1]Doctoral School of Exact and Natural Sciences, Jagiellonian University, Krakow, 31-007, Poland
[2]Institute of Meteorology and Water Management – National Research Institute, Krakow, 30-215, Poland
[3]Department of Climatology, Institute of Geography and Spatial Management, Jagiellonian University, Krakow, 31-007, Poland

Correspondence to: Gabriel Stachura (gabriel.stachura@imgw.pl)

Abstract. Accurate precipitation forecasting is crucial from both meteorological and hydrological points of view. Numerical weather prediction (NWP) models, however, still suffer from biases in this regard. In this paper, a machine-learning-based post-processing method is proposed and thoroughly evaluated using the object-based Structure-Amplitude-Location (SAL) method. For that purpose, a U-Net model is trained to combine 24-h precipitation forecasts from three deterministic NWP models (IFS, ALARO and AROME) over south-central Poland for the period 2018–2024. A dataset combining rain-gauge, radar and satellite measurements is used as the reference. Overall, U-Net slightly outperforms the input models, mainly due to improvements in the amplitude and location components. The structure component of U-Net forecasts exhibits a significant seasonal cycle, with the greatest bias during summer due to oversmoothing of precipitation objects, likely associated with convective events. The structural bias of the NWP models for that season corresponds to their horizontal resolution with IFS being the most biased and AROME – the least. In winter, however, all the NWP models suffer from overestimation of the structure component, while U-Net is barely biased. Substantial positive bias for the input models in winter also occurs in the amplitude component, whereas U-Net shows only slight underestimation for this season. However, considerable underestimation of the domain-maximum precipitation for warning-relevant precipitation sums severely limits the applicability of U-Net in supporting forecasters in predicting heavy precipitation events. At the same time, U-Net substantially reduces the number of false alarms, however, at the expense of an increased number of misses. Analysis of the relative importance of the NWP models indicates a dominant contribution from IFS, suggesting that high-resolution information from AROME might not be fully utilized by U-Net.

1 Introduction

Precipitation is certainly one of the key meteorological quantities forecasted by numerical weather prediction (NWP) models. However, its accurate prediction is still commonly considered to be challenging (Eden et al., 2012; Tapiador et al., 2019).

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

Errors stem from, among other factors, spatial discretization, the quality of initial conditions and shortcomings in physical parametrisations of convection, cloud microphysics, radiation and advection. These sources of error were thoroughly discussed in review work of Lopez (2007). In addition to limitations of NWP models themselves, the spatiotemporal characteristics of precipitation also affect its predictability. In contrast to continuous fields such as temperature or wind, precipitation exhibits strong spatiotemporal intermittency and sharp gradients (Tapiador et al., 2019). Its spatial heterogeneity can occur at hectometric scale, which is still beyond the resolution provided by operational NWP models in most European countries (ACCORD All Staff Workshop, 2025). It is therefore not surprising that numerous approaches to improve quantitative precipitation forecast (QPF) have already been proposed in the literature. As reviewed by Vannitsem et al. (2021), various statistical methods were developed to post-process QPF already before machine learning (ML) methods emerged. However, due to the growing popularity of ML in recent years, research on this topic has increased significantly. Most of this research uses ensemble forecasts from a single NWP model as input. Hess and Boers developed a frequency-weighted loss function to enhance the skill of the Integrated Forecasting System (IFS) ensemble model over the contiguous United States, particularly for heavy rain (Hess and Boers, 2022). Ensemble forecasts of the Consortium for Small-scale Modelling Ensemble Prediction System (COSMO EPS) were post-processed and downscaled to 1 km horizontal resolution using several ML algorithms over Germany (Rojas-Campos et al., 2023). Li et al. (2022) corrected the parameters of a probabilistic distribution of IFS ensemble over eastern China, producing an ensemble on output. A similar approach was taken by Pic et al. (2025) in southern France when refining AROME-EPS forecasts. Valuable contributions from an operational perspective come from Australian and French national weather services, who post-processed ensembles of IFS and AROME, respectively (Taillardat and Mestre, 2020; Trotta et al., 2024).

One of the earliest applications of ML to reduce precipitation bias in a multi-model approach is the work of Krasnopolsky and Lin (2012). They combined ensemble forecasts from as many as eight different sources to create a neural-network ensemble which mitigated severe dry bias present in input forecasts over the contiguous US. The concept of a superensemble was also deployed by Quintero Plaza and Garcia-Moya Zapata (2019), who calibrated QPF for Spanish airports (point locations). They stated that the outcome for precipitation was positive, but the gain was less substantial than in the case of 2-m air temperature or 10-m wind speed.

Deterministic forecasts were explored less often in that subject. Several studies concerned bias correction of the Weather Research and Forecasting (WRF) model forecasts only at station points, disregarding the spatial aspect (Sayeed et al., 2023; Zhang et al., 2024). Sayeed et al. (2023) reported that the improvement for precipitation was minor compared to other meteorological variables. Hu et al. (2021) employed two-dimensional reforecasts of Yin-He global spectral model over the Sichuan Basin in China. Their root mean square error reduction was substantial, however, the resolution of input and output fields was quite coarse (0.5°). A comprehensive study involving as much as 34 years of WRF forecasts over the western US was conducted by Badrinath et al. (2023). They managed to lower the forecast error by around 15% in total (for extreme precipitation it was even more). The major limitation relevant to our study is that their baseline verification scores were pixel-based.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

Particularly interesting, yet relatively rare, is research where bias correction is accomplished by combining information from several deterministic NWP models (a deterministic multi-model). An example of this is the study conducted by Zhong et al. (2025) who blended outputs of models of different scales (global and mesoscale) to improve QPF. Despite an overall improvement, they pointed out overestimation of precipitation area as a major weakness of the ML model. However, their evaluation was limited primarily to aggregate scores and single case studies, lacking deeper insight how the enhancement varied across meteorological conditions. Other valuable research combining deterministic forecasts of NWP models was carried out over Italy by Monaco et al. (2024). They reported improvement in all seasons except spring. During training, they ensured the balance between convective, intermediate and large-scale precipitation types in a training dataset. However, the evaluation is based on pixel-based metrics with very limited stratification of forecast performance.

Collectively, each of the studies reviewed in this section has some limitations relevant to the present work. First, many of them rely solely on pixel-wise verification metrics, which are prone to the double penalty problem. Another major limitation is a lack of detailed evaluation. What is mostly reported is the overall score, often accompanied by individual case studies. The main emphasis is placed on achieving an overall improvement with little information on the circumstances under which it occurs. The most commonly addressed aspect of verification is precipitation intensity, which is particularly relevant in the context of operational warning systems. However, variability with respect to weather regimes, months or dominant precipitation type remains insufficiently explored, limiting our understanding of the strengths and weaknesses of a given post-processing method. Another aspect relevant from an operational perspective is the form of a post-processed forecast. While some studies produce probabilistic output, a single deterministic output could provide more readily interpretable information to a forecaster.

In light of these limitations, the aim of this study is to assess the potential of an ML model to correct bias in 24-h precipitation (hereafter: daily precipitation) forecasts by blending information from three deterministic NWP models available operationally at the Polish national weather service (IMGW). The resulting post-processed field is designed to exploit complementary information from the input NWP models and improve the spatial representation of daily precipitation. Beyond evaluating the overall performance of the ML model, a detailed analysis is conducted how the performance varies across atmospheric circulation types, months and precipitation amount. The results are related to the variability observed in the input NWP models. This allows us to identify the conditions under which the proposed method provides the greatest benefits, as well as those under which its performance deteriorates.

The paper is divided into five sections. The first section describes the study domain and the different data sources used to train and evaluate the ML model. The second section introduces the ML method and verification scores used to assess its performance. It also describes the construction of the loss function and the adopted training strategy. Additionally, the section explains how the data were preprocessed and how predictors were selected. The Results section provides a detailed assessment of forecast accuracy of both the ML model and the input NWP models from several perspectives. This is followed by a discussion that addresses conditions associated with poor performance of the ML model and assesses the informativeness of the input models. Finally, the main conclusions are summarized in the last section.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

2 Study area and data

In the study, spatial precipitation forecasts from three NWP models are combined with gridded reference data over the area of south-central Poland. The study period spans over 6 years, from 1st August 2018 to 31st December 2024. Daily precipitation sums were aggregated from 6:00 UTC to 6:00 UTC on the following day, which corresponds to a 6–30h forecast range. A detailed description of the study area can be found in Sect. 2.1. The NWP models are described in Sect. 2.2. Section 2.3 addresses the Rain Gauge-Radar-Satellite (RainGRS) - a system producing a spatially consistent estimated precipitation field. A weather patterns classification used as an additional aspect of forecast verification is briefly described in Sect. 2.5.

2.1 Study area

The study domain is a 256 km × 256 km square covering south-central Poland (Fig. 1). Its northern and western parts are dominated by lowlands belonging to the Central European Lowlands with annual precipitation around 550 mm. Annual precipitation increases to approximately 800–900 mm over the highlands in the central and southern parts of the domain and exceeds 1000 mm in the mountainous area along the southern edge of the domain. Nearly 40% of the annual precipitation typically falls in summer months and only 18% in winter (Kirschenstein and Baranowski, 2005). Liquid phase accounts for 82% of the annual precipitation. January is the only month in which solid precipitation accounts for a greater amount than liquid precipitation (Łupikasza and Małarzewski, 2023). Concerning storm-related convective precipitation, thunderstorms occur on average on 25 days per year (Bielec-Bakowska and Lupikasza, 2009). The thunderstorm season lasts from mid-April to late September and peaks in June. Around 80% of cloud-to-ground flashes occur between June and August (Poręba et al., 2022).

Several factors determined the exact location and shape of the study area. The domain dimensions were set such that, at a horizontal resolution of 1 km, the number of grid cells along each dimension is a power of two, which ensures compatibility with the multiscale encoder-decoder architecture of the CNN used in the study. During training, the network performs repeated operations of down- and upsampling by factors of two (Ronneberger et al., 2015). Choosing dimensions to be powers of two ensures retention of the full-domain spatial context at all scales without the need of any additional cropping. The second factor determining the choice of the study area was a marked deterioration of reference data quality outside Poland due to a lack of data from ground-based stations as input. Hence, the domain needed to be located predominantly over Poland. Finally, under certain synoptic conditions, large daily precipitation can occur in areas of complex orography in the southern part of the country. Inclusion of such events could be particularly informative for training an ML model (Dewettinck et al., 2026a) . Based on the considerations discussed above, the study area was defined as shown in Fig. 1.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

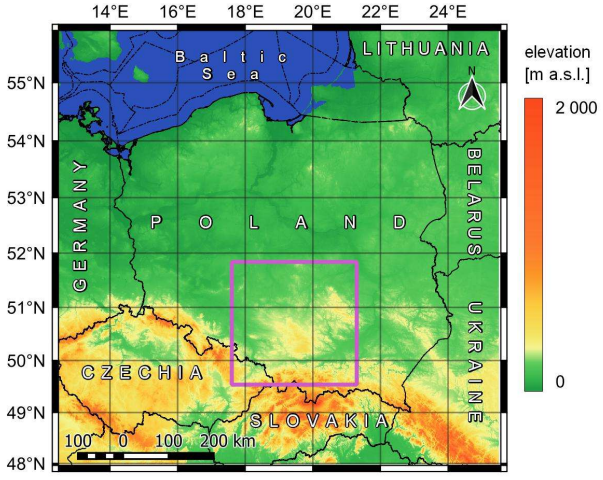

Figure 1: Orography map of the study area with country borders shown in black and the study area highlighted with purple.

2.2 Forecast data

Out of several NWP models available operationally at IMGW, three of them were selected to provide input forecasts: IFS, ALARO and AROME. The selection of ALARO and AROME was driven by their long record of archived forecasts in the local archive (since August 2018) and their known skill in precipitation forecasting (particularly AROME). IFS was chosen because its forecasts are commonly regarded as more accurate than those from GFS (Mass, 2023). Based on the availability of archived forecasts of ALARO and AROME, the study period was set to 1 August 2018 – 31 December 2024. Because of occasional operational breakdowns, twenty days were missing. In total, complete data were available for 2324 days. Only forecasts initialized at 00UTC were used in the study. As AROME forecasts were limited to 30 h for most of the study period, we consider daily precipitation only for the period from 06 UTC to 06 UTC the following day (6–30 h forecast lead times). In addition to precipitation fields, the downloaded data also included other meteorological and physiographic fields which are physically related to precipitation and therefore might be informative for an ML model. While the range of available fields was broad for IFS, for ALARO and AROME it was limited to what was stored in the local archive. The selected fields are listed in Table A1 in the Appendix.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.



2.2.1 ALARO and AROME

ALARO and AROME are limited-area numerical weather prediction (NWP) models developed within the A Consortium for Convection-scale Modelling Research and Development (ACCORD), formerly the ALADIN-HIRLAM consortium (Termonia et al., 2018). Both models are run operationally at IMGW in a dynamical adaptation mode (Kolonko et al., 2023).

ALARO is a non-hydrostatic model with a horizontal resolution of 4 km × 4 km and 70 vertical levels. The lateral and lower boundary conditions for the model are provided by the global ARPEGE model. ALARO uses a single-moment microphysics scheme with five prognostic water phases (water vapour, cloud ice and droplets, snow and rain) plus a diagnostic pseudo-graupel (Lopez, 2002). This scheme relies on the 3MT (Modular Multiscale Microphysics and Transport) framework to handle both stratiform (layered) and convective (storm-driven) precipitation consistently (Giot et al., 2016). Deep convection is partly resolved, while shallow convection is handled within a TOUCANS turbulence scheme (Dewettinck et al., 2026b).

During the study period, the model underwent two major updates. On 5 June 2019, the model version was upgraded from CY40T1 to CY43T2, including the introduction of a non-hydrostatic dynamical core. On 18 November 2019, the number of vertical levels was increased from 60 to 70. Seasonal verification of the ALARO forecasts showed that the updated configuration provided results consistent with those of the previous version (not shown). Several changes were introduced also to the lateral boundary conditions, including an increase in the horizontal resolution of ARPEGE from 15.7 km to 9.4 km on 15 November 2019, as well as changes in the model version from CY43T2 to CY46T1 in June 2022 and from CY46T1 to CY48T1 in October 2024 (Roehrig, 2022).

Output from the ALARO model is used to provide initial and lateral boundary conditions for AROME. The operational AROME configuration has a horizontal resolution of 2 km × 2 km and 70 vertical levels, with a forecast range of 30 h. Its computational domain consists of 799 × 799 grid points and is centered at 19.3°E, 52.3°N. Deep convection is assumed to be fully resolved by the model's dynamics, while shallow convection is parameterized. The model uses a single-moment microphysics scheme with three-class ice parametrisation (ICE3) involving five prognostic quantities of water condensates (Seity et al., 2011). During the analyzed period, the AROME model was upgraded from CY40T1 to CY43T2 on 11 February 2020. Detailed forecast verification demonstrated consistency between the new and previous model versions (not shown).

2.2.2 IFS

Operational, high-resolution deterministic forecasts of the global IFS model were retrieved from the European Centre for Medium-Range Weather Forecasts (ECMWF) Meteorological Archival and Retrieval System via dedicated Web API service. Their horizontal resolution was around 9 km and has not changed throughout the study period. The microphysics scheme used in the model concerns grid-scale clouds and precipitation and includes following prognostic variables: cloud liquid water, cloud ice, rain, snow and cloud fraction. Subgrid-scale convection (including deep, shallow and mid-level convection) is represented using a mass-flux convection parametrisation. Land surface processes are parametrized by the Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land (HTESSEL) land surface model. Several updates of model versions were

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.



175 introduced during the study period (Table 1). While the above-described parametrisations and schemes remained conceptually the same during this period, some of their components did change. Detailed scientific documentation of individual model versions is available at the ECMWF webpage (IFS documentation, 2022).

Table 1 Upgrades of the IFS model versions during the study period (Integrated Forecasting System, 2022)

model version	operational period
cy45r1	1 August 2018 – 10 June 2019
cy46t1	11 June 2019 – 29 June 2020
cy47t1	30 June 2020 – 10 May 2021
cy47r2	11 May 2021 – 11 October 2021
cy47r3	12 October 2021 – 26 June 2023
cy48t1	27 June 2023 – 11 November 2024
cy49r1	12 November 2024 – 31 December 2024

180 **2.3 Reference data**

The RainGRS system combines precipitation data from rain gauges, weather radars and satellite observations. It has been working operationally at IMGW since 2015. The system consists of several components responsible for the quality control of radar measurements, the quality-based compositing of data from individual weather radars, the quality control and spatial interpolation of rain gauge observations, the derivation of satellite precipitation estimates from visible and infrared channels
185 and the merging of data obtained using different measurement techniques (Jurczyk et al., 2020). Operational precipitation data are obtained from a network of telemetric rain gauges, the Polish POLRAD radar network consisting of 10 C-band Doppler polarimetric radars and satellite products developed within the European Organisation for the Exploitation of Meteorological Satellites Satellite Application Facilities on Support to Nowcasting and Very Short Range Forecasting (EUMETSAT NWC SAF) programme. The final product - a multi-source estimate of precipitation field - is generated every 10 minutes with a high
190 spatial resolution of 1 km x 1 km. It is then temporally aggregated into longer accumulation intervals. Temporal interpolation is applied to ensure the completeness of the accumulations when data gaps occur.

The combination algorithm is an extension of the conditional merging technique proposed by Sinclair and Pegram (2005). In this approach, rain gauges are considered to be the most accurate direct measurements of precipitation but only at discrete points, whereas weather radar provides trustworthy information about the spatial variability of rainfall.

195 Verification of the multi-source combination has shown that its accuracy strongly depends on the quality of the input data and the type of precipitation. For heavy rain associated with convective events, the reliability of the GRS estimates is higher than that of any individual input and tends to improve as the quality of the interpolated rain gauge field increases (Jurczyk et al., 2020). This conclusion was confirmed by an analysis conducted during the large flood in Poland in September 2024, caused by extremely high widespread precipitation in an orographically diversified basin. It showed that multi-source GRS estimates

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

were the most reliable when evaluated against manual rain gauge observations used as reference data (Szturc et al., 2025). In the case of stratiform precipitation, the accuracy of multi-source combination is strongly related to the distance from the nearest radar site. This is caused by geometry of the radar beam propagation and the fact that weather radars are more likely to sample above the tops of stratiform clouds, which are lower than convective clouds (Jurczyk et al., 2020).

In addition to the operationally generated precipitation fields, reanalyses of daily precipitation using more complete input data than are available in real time are also produced since March 2021. The algorithm takes advantage of several additional possibilities, including the incorporation of delayed data, access to high-quality observations from manual rain gauges that are not available operationally, and the improved performance of quality-control and merging algorithms when applied to longer accumulation periods. The results have shown that the reanalysed estimates are more reliable than those generated operationally. Adjusting the reanalysed data using observations from manual rain gauges resulted in a further substantial improvement in quality (Jurczyk et al., 2023). Therefore, reference data used in this work are a combined dataset consisting of operational RainGRS (from 1 August 2018 to 1 March 2021) and the refined RainGRS (from 2 March 2021 to 31 December 2024).

2.4 Atmospheric circulation types

Since precipitation in the study area is strongly related to the direction of the incoming air masses (Ustrnul and Czekierda, 2001), the Grosswetterlagen (GWL) classification of circulation types over Europe was used as an additional dimension in verification of precipitation forecasts. This is a well-known subjective classification proposed by Hess and Brezowsky (Gerstengarbe and Werner, 2005) and, as reported by James (2007), updated several times later. It classifies large-scale circulation patterns over Europe and the North Atlantic with a primary focus on synoptic-scale circulation conditions affecting weather in Central Europe (Gerstengarbe and Werner, 2005). A circulation type is assigned to particular conditions manually by a meteorologist based on, e.g., spatial distribution of mean sea-level pressure as well as 500-hPa geopotential height. The classification comprises 29 defined circulation types (4 zonal, 7 mixed and 18 meridional) plus an undefined (transitional) category. They are listed in Table A2 in the Appendix. Despite daily temporal resolution, one circulation type generally lasts for at least 3 days, thus representing a persistent circulation pattern rather than an instantaneous one. The classification is currently maintained and updated by the German Weather Service (DWD).

3 Methods

3.1 Machine Learning

The post-processing model employed in the study is based on a U-Net – a type of convolutional neural network with encoder-decoder architecture originally developed for biomedical image segmentation (Ronneberger et al., 2015). Owing to its ability to capture spatial patterns at multiple scales simultaneously, the U-Net has become one of the most widely adopted deep learning networks in precipitation-related applications, including nowcasting (Ayzel et al., 2020), statistical downscaling

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

(Böker et al., 2023; Sha et al., 2020) and post-processing of numerical QPF (Hess and Boers, 2022; Taillardat and Mestre, 2020; Zhang et al., 2023). Two structural properties of the U-Net make it particularly well-suited to the task at hand. First, the contracting (encoder) path restores the original spatial resolution. Second, skip connections transfer feature maps directly from the encoder to the corresponding decoder levels, preserving small-scale, high-frequency information that would otherwise be lost during downsampling (Hess and Boers, 2022). The combination of both mechanisms enables the network to blend information from the three input NWP models across a range of spatial scales – from the location of precipitation systems down to sharp local gradients – which is essential given the strong spatial intermittency of daily precipitation fields.

The network architecture used here follows the canonical U-Net design. The input consists of a number of predictor channels given on a common 256x256 grid at 1-km horizontal resolution (Sect. 2.1). The encoder comprises four stages, each consisting of a block of two 3x3 convolutions followed by a rectified linear unit activation and a 2x2 max-pooling operation that halves the spatial dimensions. The number of feature maps doubles at each stage – from 32 in the first block to 512 in the bottleneck, where a field is represented at a 16-fold reduced resolution (16x16 grid points). At the stage of decoding, the resolution is restored in four steps by 2x2 transposed convolutions with a stride of 2. After each upsampling step, the resulting feature maps are concatenated with the corresponding encoder feature maps (so-called skip connections) and processed by another double-convolution block. The final 1x1 convolution transforms the 32 feature maps of the last decoder stage onto a single output channel representing the corrected daily precipitation field. No activation is applied to the output layer; instead, after the inverse transformation, negative values are truncated to zero. A Python implementation of the U-Net model within the PyTorch library was used in the study (Paszke et al., 2019).

3.2 Verification scores

To evaluate the performance of precipitation forecasts, the object-based Structure–Amplitude–Location (SAL) verification method was employed (Wernli et al., 2008). Unlike pixel-based verification metrics, SAL avoids the double-penalty problem commonly encountered in the evaluation of high-resolution precipitation forecasts and provides quantitative information on the spatial characteristics of precipitation events (Gilleland et al., 2010). The method first identifies precipitation objects within the forecast and observed fields based on their spatial distribution. It then evaluates forecast quality using three independent components describing the Structure (S), Amplitude (A), and Location (L) of the precipitation field (Wernli et al., 2008).

The A component ranges from −2 to 2 and quantifies the difference in the domain-averaged precipitation between a forecast and an observation field. Positive values indicate an overestimation of the overall precipitation amount, whereas negative values indicate an underestimation. A value equal to 0 corresponds to a perfect agreement in the domain-averaged precipitation (Wernli et al., 2008). The component is undefined when neither the forecast nor the observation contains any detected precipitation objects (Pulkkinen et al., 2019).

The S component also ranges from −2 to 2 and measures differences in the spatial structure of precipitation objects. Positive values indicate that the forecast precipitation objects are too large and/or too flat compared with the observations, whereas negative values indicate that they are too small and/or too peaked. A value of S equal to 0 represents a

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

perfect match in the structure of precipitation objects (Wernli et al., 2008). If no objects are detected in either field (because the precipitation fields are entirely zero or all values are below the object detection threshold), the structure cannot be evaluated and S is not a number (NaN) (Pulkkinen et al., 2019).

The L component ranges from 0 to 2 and quantifies the spatial displacement between the forecast and observed precipitation fields. It consists of two terms describing the displacement of the precipitation centroid and the difference in the spatial distribution of individual precipitation objects. Lower values indicate better agreement, while L=0 corresponds to a perfect spatial match (Wernli et al., 2008). An undefined value occurs when no objects are identified in one or both fields (Pulkkinen et al., 2019).

A special case occurs when all three SAL components are returned as NaN. This indicates that neither the forecast nor the observation contains any precipitation objects, making the SAL verification undefined. Such cases were excluded from the evaluation.

The SAL method has been applied in numerous studies of QPF verification (Gofa et al., 2018; Li et al., 2023; Peng et al., 2023; Radanovics et al., 2018). It is also included in the internal verification toolbox of the ACCORD consortium (ACCORD, 2020). In the present study, SAL was computed using the *pySTEPS* Python library (Pulkkinen et al., 2019). The implementation applies a relative object-detection threshold by default. We modified the threshold to exclude precipitation lower than 1 mm, so that it is calculated as $\max(1, 0.067p_{95})$, where p_{95} is the 95th percentile of a precipitation field. The threshold defined in this way allows to neglect small precipitation, which is irrelevant within a 24h accumulation period, particularly from an operational perspective. At the same time, setting it higher could result in an artificially increased number of false-negative (miss) cases, particularly if forecasts tend to underestimate precipitation.

To additionally focus on misses and false alarms (which correspond to extreme values of S and A and NaNs in case of the L component), a contingency table (also known as a confusion matrix) was constructed.

To summarize the overall SAL performance for the models, an aggregated SAL-based metric SAL_{agg} was introduced in this study. The metric is defined as the sum of the medians of the absolute values of the three SAL components. Additionally, to account for cases where L is undefined, penalty terms for misses and false alarms were added. The quantity is calculated as follows:

$$SAL_{agg} = med(|A|) + med(|S|) + med(|L|) + miss_{frac} + \frac{FA_{frac}}{2} \quad (1)$$

where *med* denotes the the median, | | – the absolute value, $miss_{frac}$ – fraction of misses, FA_{frac} – fraction of false alarms.

3.3 Data preprocessing

To begin with, IFS forecasts were regridded to a regular longitude-latitude grid. Next, forecast files of all NWP models were converted to the NetCDF format and temporally aggregated to daily values using the mean, sum or maximum. RainGRS data were reprojected to the WGS84 coordinate system and converted to the NetCDF format. Finally, the data were cropped to the

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.



study domain. The processing was done using the Climate Data Operators software and Python (Schulzweida, 2023). To filter out precipitation noise (very small amounts of daily precipitation), a threshold of 0.05 mm was applied. The value represents a reasonable trade-off between including data of negligible relevance and losing some meaningful data (as the lowest precipitation measurable by a rain gauge is 0.1 mm). Temporal fields were encoded as sine and cosine.

3.4 Training strategy and feature selection

As precipitation distribution is usually zero-inflated and heavy-tailed, neural networks have problems to efficiently learn the patterns if default loss functions are applied, which are usually quadratic (Bishop, 1995; Wilson et al., 2022). Hence, the mean squared error (MSE) is dominated by the most frequent low values which drives the network towards overly smooth fields and a systematic underestimation of heavy precipitation (Hess and Boers, 2022; Rojas-Campos et al., 2023). To mitigate the imbalance, the target data have been log-transformed, and a composite loss function L was introduced (Eq. (2)). The function consists of three components. The first one is a weighted MSE computed in the log-transformed space, where grids with reference daily precipitation of at least 1 mm receive a weight multiplied by 1+α, where α=0.5, thereby emphasizing wet areas relative to the dry background. The second term penalizes the absolute difference between the 99th percentile of the predicted and reference precipitation fields (computed per sample in physical units) and is weighted by the γ parameter (γ=0.01). While the weighted MSE component controls the overall accuracy of the field, the quantile term constrains the upper tail of the predicted distribution and prevents the damping of heavy precipitation. The design is conceptually similar to the frequency-weighted loss proposed by, e.g., Hess and Boers (2022). Additionally, to prevent the model from underestimating precipitation particularly below the verification threshold, a penalty term has been added as well (with a weight β=0.1).

$$L = \frac{1}{N}\sum_{i=1}^{N}[1 + \alpha \mathbb{1}(o_i \geq o_{thr})]\,(y_i - o_i)^2 + \gamma[p_{99}(y) - p_{99}(o)]^2 + \beta\frac{\sum_{i=1}^{N}\mathbb{1}(o_i \geq o_{thr})\max(o_i - y_i, 0)^2}{\sum_{i=1}^{N}\mathbb{1}(o_i \geq o_{thr})} \quad (2)$$

where:

N – number of grid points

α, β, γ – tunable weights of different components

o_i – reference precipitation at the i-th grid point

y_i – forecasted precipitation at the i-th grid point

o_{thr} - minimum precipitation threshold

$\mathbb{1}$ – an indicator function, $\mathbb{1}(o_i \geq o_{thr}) = \begin{cases} 1, & o_i \geq o_{thr} \\ 0, & \text{otherwise} \end{cases}$

p_{99} – 99th percentile of a field.

Training of the U-Net model was performed with the Adam optimizer using a learning rate of 10^{-4} and a batch size of 16 for up to 60 epochs. The learning rate was halved whenever the validation loss did not improve for 5 consecutive epochs. The data were split into training, validation, and testsets by year. The final model was selected based on the performance on the validation dataset (the lowest value of validation loss). To ensure reproducibility, experiments were run in deterministic mode

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

with a fixed random seed. Experiments were repeated using three random seeds to account for sensitivity to stochastic initialization of model weights and the order of training batches. Training was performed on NVIDIA H100 GPUs using *bfloat16* mixed precision with a single training run taking approximately 1.5 hours.

A three-phase predictor selection procedure was applied to a set of around 100 predictors (obtained by temporal aggregation of the initial predictors listed in Table A1 in the Appendix, as described in Sect. 3.3). At first, features were divided into several groups (e.g., instability indices, humidity profiles, precipitation components) and ablation experiments were performed. The goal was to identify groups that caused the largest deterioration of the score on the validation set (relative to the model using the full set of features). Then, within the most informative groups, more detailed ablation experiments were conducted to assess the added value of individual predictors. In the final phase, candidate predictors were gradually added to the core features (daily precipitation forecasts from each NWP model). After each addition, model performance was assessed. Eventually, the lowest validation error was obtained for the following set of predictors: daily precipitation from each of the NWP models, daily mean specific humidity from IFS at all selected pressure levels, daily accumulated lightning density from IFS, elevation and a temporal field accounting for the annual cycle (Fig. 2).

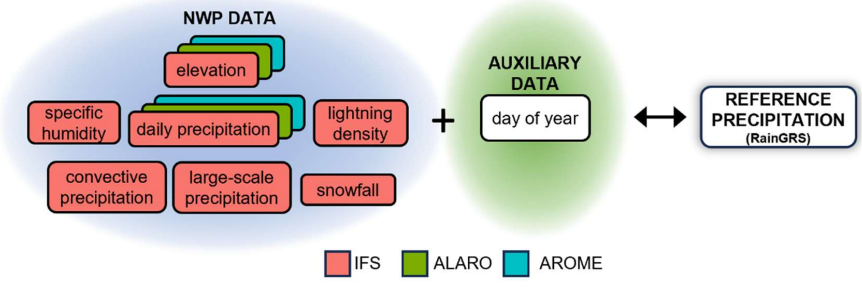

Figure 2: A schematic diagram of data used to train the U-Net model. The input predictors (left side of the diagram) are matched with the target daily precipitation according to the RainGRS system.

Predictor selection was performed using a fixed temporal training-validation split, with 2024 reserved for validation and the remaining dataset used for training. Upon completion of the process, the selected predictor set was kept fixed and used to produce forecasts for each year of the dataset through cross-year evaluation. To this end, a series of models were trained and evaluated across different temporal splits. In each rotation, one year was assigned to the test set, another one to the validation set while the rest of the data comprised the training set. As data from 2018 cover only 5 months, the year was treated jointly with 2019 in all rotations. All considered configurations of the cross-year evaluation are explicitly listed in Table A3 in the Appendix.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

4 Results

Unless stated otherwise, all results presented below are based on predictions from the cross-year evaluation and therefore cover the entire study period.

4.1 Overall performance

Overall model performance was assessed using the SAL_{agg} metric and its components (Table 2). On average, U-Net provides the most accurate precipitation forecasts, however, outperforming the best input model (AROME) by only 0.1. The improvement does not occur for every SAL component. It mostly derives from better estimation of amplitude as well as more accurate location of a precipitation field. However, U-Net clearly underperforms AROME regarding the structure component.

Table 2 Overall assessment of QPF accuracy across models for the whole study period. The lowest value in each column is highlighted in bold.

	SAL_{agg}	med(|S|)	med(|A|)	med(|L|)	penalty term
IFS	1.81	0.77	0.79	0.17	0.10
ALARO	1.70	0.62	0.81	0.18	0.07
AROME	1.49	**0.50**	0.71	0.19	0.07
U-Net	**1.39**	0.61	**0.54**	**0.14**	**0.04**

A qualitative view of the overall scores is provided by the so-called salplot (Fig. 3) – a scatterplot showing SAL components for each day (Wernli et al., 2008). A diagonal pattern is evident for the input models – underestimation of structure usually coincides with underestimation of amplitude, while overestimation of structure – with overestimation of amplitude. The largest values of the location error usually accompany extreme values of A and S. The input models vary with respect to the area of the main concentration of points (represented approximately by the median). The most biased model is IFS. Most of the points lie in the first quadrant indicating overestimation of both the structure and the amplitude of precipitation. Especially the S component is overestimated by this model. A similar tendency occurs for ALARO, however, overestimation of the S component is less severe. The most balanced scores are produced by AROME with most points concentrated very near the origin. U-Net scores feature a very dense cloud of points near the origin, the median however is discernibly shifted towards negative values of the A component. The number of cases in the first quadrant is much smaller than for the input models. On the other hand, there is a large number of points with overestimated structure and underestimated amplitude (fourth quadrant), which is rare for the input models.

However, one limitation of a salplot is overlapping of points, particularly at the extreme values of S and L, which can obscure the frequency of such cases. Therefore, the distributions of individual SAL components should also be inspected (Fig. 4). The figure reveals that U-Net suffers from a large number of cases with extreme underestimation of S and A, which corresponds to misses (no precipitation is forecasted while precipitation is observed). Although the input models also exhibit an increased

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

frequency of such cases (particularly for the A component), but less so than U-Net. On the other hand, the input models feature a substantial number of cases at the opposite extreme (S and A=2), corresponding to false alarms. Such cases are less frequent for U-Net, with a moderate reduction for S and much more pronounced for A.

As far as the L component is considered (Fig. 4b), it is interesting to note that IFS performs slightly better than ALARO and AROME. Both the median and the mode of the U-Net distribution are closer to zero than those of the input models. Moreover, the violin is narrower at higher L values indicating that such values occur less often. However, for extreme location errors, the improvement is negligible. In fact, U-Net is inferior to IFS in this respect.

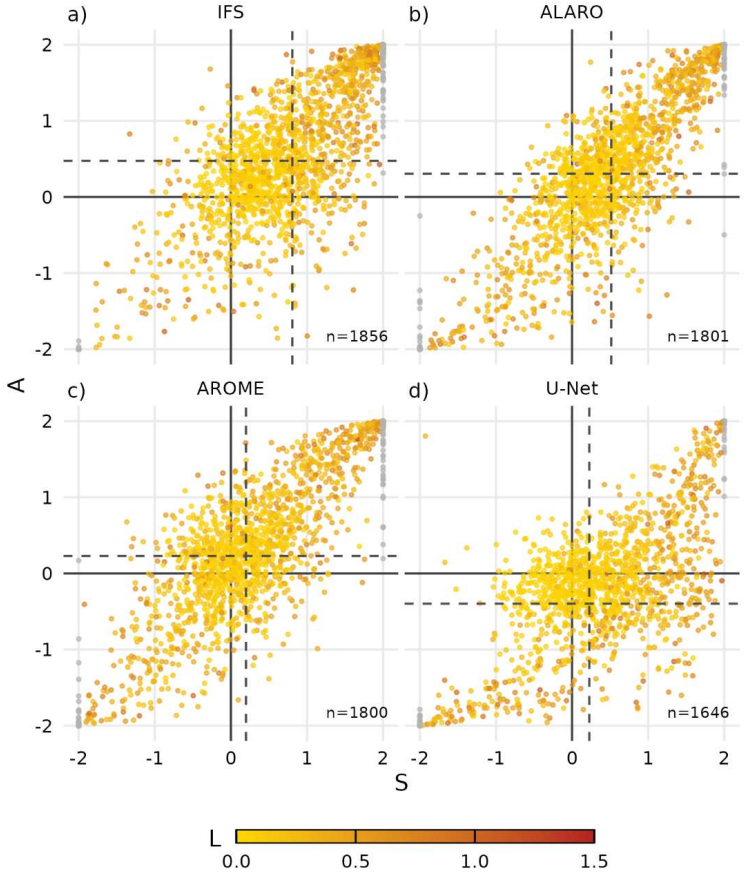

Figure 3: Scatterplots of SAL components for the different models. The S and A components are shown on the x- and y-axes, respectively, while the L component is indicated by colour. Grey denotes cases with undefined location. If either S or A is undefined, the point is not shown. The total number of depicted points is given in the bottom-right corner of each plot. The medians of S and A components are represented by dashed lines

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

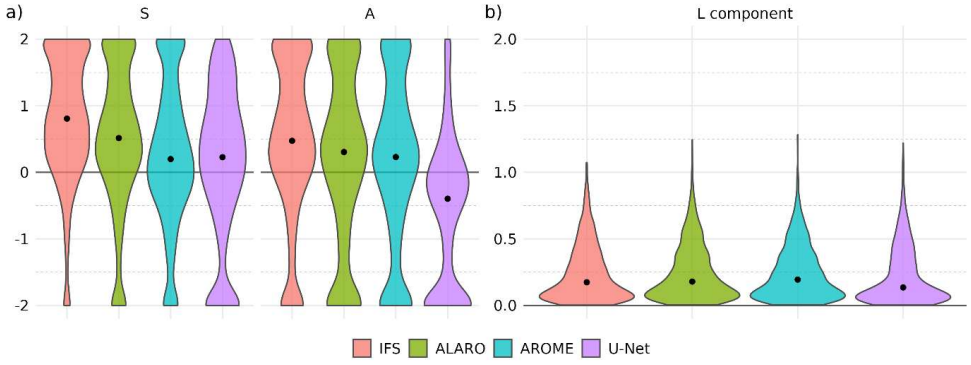

Figure 4: Distributions of the S and A (a) and the L component (b) for the different models. The median is represented by a black dot.

From an operational perspective, a desirable feature of a post-processing method is the ability to accurately predict the maximum precipitation within an area, so that appropriate warnings could be issued. Therefore, the skill of the models in capturing the magnitude of the domain-maximum is examined in Fig. 5. At IMGW, warnings are issued on a three-level scale, with the first degree representing the lowest severity and the third degree – the highest. The corresponding warning-relevant intervals are indicated in the figure with colours. From the figure it can be seen that the input models tend to overestimate precipitation just below the threshold of the first-degree warning (20 mm). For IFS and AROME, the overestimation persists within the first degree warning interval (20–30 mm). For higher precipitation, however, IFS and ALARO rapidly shift towards underestimation, while AROME is neutral. A similar tendency could be observed for the heaviest precipitation events – IFS and ALARO exhibit dry bias there while AROME underpredicts only slightly. In contrast, U-Net shows systematic underestimation of precipitation sums above 7 mm. Within the warning intervals, it underestimates precipitation by as much as 50–60%. As a result, U-Net is in fact the least suitable model for predicting precipitation that exceeds warning thresholds.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

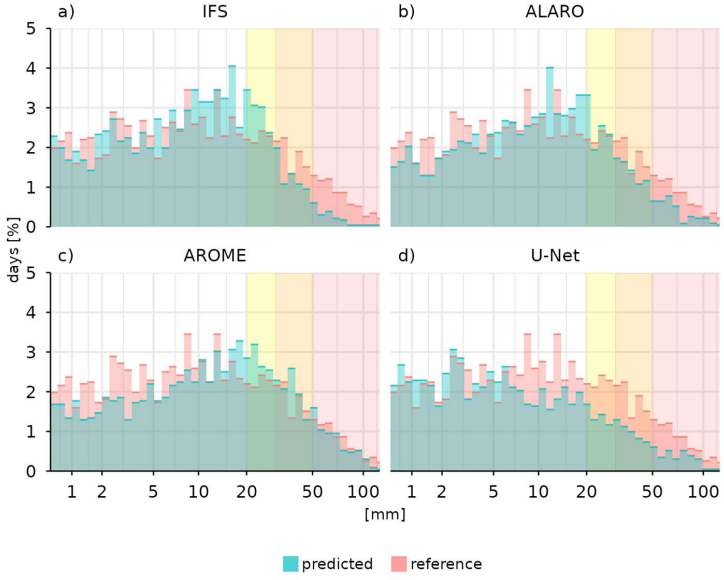

Figure 5: Histogram of predicted versus reference domain-maximum precipitation. Warning intervals are shown in yellow (lowest degree), orange (moderate degree) and red (highest degree). Note that the x-axis is transformed to the log(1+x) scale. Precipitation below 0.5 mm is omitted.

4.2 Variability across months

As seasons change, the dominant precipitation mechanisms shift between large-scale and convective processes, along with corresponding changes in precipitation type. Consequently, QPF accuracy varies across months and seasons. The dependence is particularly marked for the structure component (Fig. 6). IFS exhibits substantial positive bias for this component throughout the year with highest values in summer as well as winter. Similar bias in winter could be observed for ALARO, however, in summer it is much lower compared to IFS. AROME is typically the most accurate input model across the months with the lowest bias in April, August and September and the highest – in winter. As far as U-Net is considered, severe overestimation in warm season is striking. For the rest of the months, the bias is relatively small, sometimes even falling slightly negative. Compared to the performance of the input models, it is winter when relative improvement of U-Net is the most evident.

A single-mode pattern is to be seen when variability of the amplitude component is analysed. The input models tend to overestimate precipitation amounts in the coldest part of the year (November – March). In summer and autumn, the mean bias is very small and for ALARO and AROME, it even switches to negative values. U-Net underestimates daily precipitation persistently across the months, with largest bias in September and October and the lowest – in February.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

The location component varies across months the least. The results for all the models are pretty much alike. The largest values of the error occur in summer months. Additionally, IFS exhibits strong bias also in October. For the NWP models, most accurate forecasts in terms of locations occur in transitional months (April, May and September). U-Net outperforms the input models in this aspect in every month except May. The gain is however usually slight.

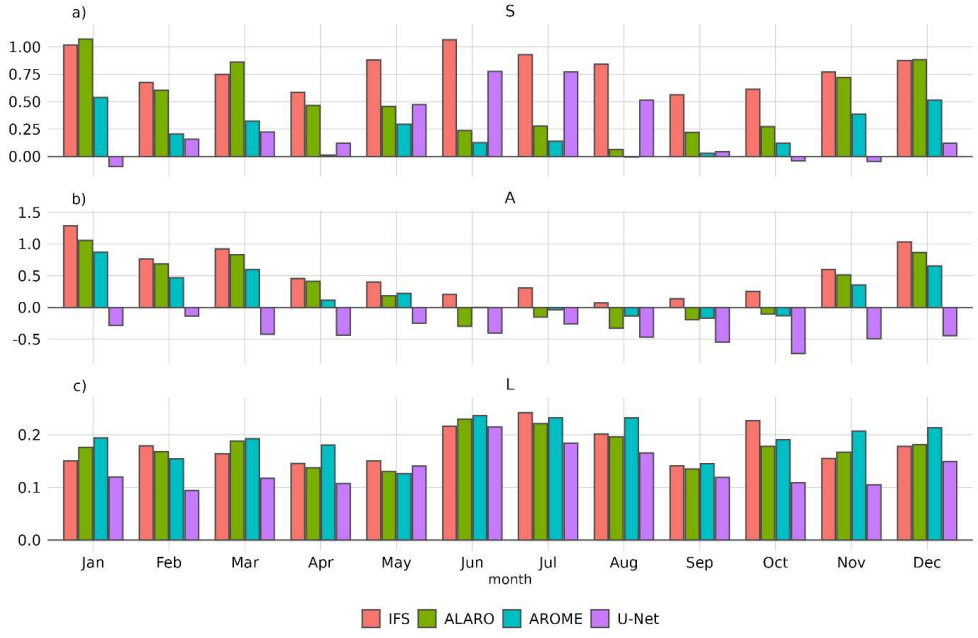

Figure 6: Monthly variability of the median values of the S (a), A (b) and L (c) components for different models. Undefined cases were excluded from the analysis. Note that y-axis scale range differs across panels.

To address specifically cases with extreme or undefined values of SAL components, a confusion matrix across the months was calculated and summarized in Fig. 7. As far as false alarms occurrence is concerned, the input models produce such cases at most in the colder period (November – February). Secondary maximum occurs for July, but it is less pronounced. Interestingly, no clear added value of AROME with respect to ALARO is visible in this aspect. U-Net considerably reduced the frequency of such cases as it correctly forecasted no precipitation then. The improvement is particularly notable in winter, when the input models can have up to 15% of such errors, while for U-Net it is around 1%. On the other hand, U-Net suffers from an increased ratio of miss cases. No clear seasonality is visible in this aspect, however, the colder part of the year seems to favour such cases. In contrast, the input models produce misses at most for August and September. IFS is the model that misses precipitation least often – only 1–3% per month, while for ALARO and AROME it is 1–8%.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

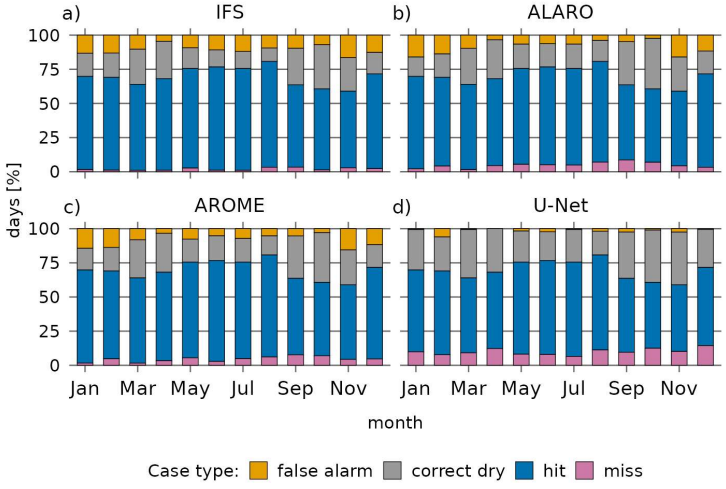

Figure 7: Monthly variability of the relative frequency of days classified according to the confusion matrix for different models.

4.3 Variability across GWL circulation types

To enhance the readability of charts, the GWL circulation types were grouped by representative airflow direction. The grouping was based on the cardinal and intercardinal directions indicated by the type names. If the name did not specify any direction, the cardinal direction indicated by the Grosswettertypen was assigned. The types with no clear direction of advection (BM, HM, TM and U) were analysed separately. The full list of circulation types and their assigned groups is provided in Table A3 in the Appendix. The SAL scores are presented in Fig. 8. For the S and A components, the largest errors are associated with southern circulation types (mainly SW and SE). For some of these circulation types the median reaches the maximum possible value, indicating substantial underestimation or overestimation in most cases. These types have generally moderate frequency (1–3% in the study period) and, as reported by Ustrnul and Czekierda (2001), typically bring little precipitation. Only the south-western ones (SWz and SWa) occur more frequently (5% each) and are associated with greater precipitation, especially in winter (Ustrnul and Czekierda, 2001). For the S component, performance of the models corresponds roughly to their horizontal resolution. U-Net does not improve accuracy for S and A for most of these types, with errors generally comparable to or larger than those of the NWP models. Among the 8 types for which U-Net improves the S component, the largest relative gain occurs for the type featuring a low over the British Isles (TB) and for the south-eastern types. Regarding the A component, the scores of the models are much more similar, especially regarding the most problematic sectors: SW and SE. However, taking into consideration all circulation types, U-Net improves the scores for most of them, with the largest improvement for the type involving a Scandinavia-Iceland high with a trough across Central Europe (HNFz) and the largest deterioration for

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

the undefined or transitional type (U). As far as the L component is considered, the scores vary little across the circulation sectors. U-Net is more accurate than the NWP models for all sectors except the south-eastern one.

The most frequent circulation type during the study period was a zonal ridge across Central Europe (BM), which typically brings dry weather with rare precipitation (Ustrnul and Czekierda, 2001). U-Net outperformed the input models in terms of amplitude and location for that type. Additionally, as seen in Fig. 9d, the NWP models tend to forecast spurious rain in such situations. The number of false alarms was greatly reduced by U-Net. On the other hand, U-Net exhibits missed events systematically across all circulation types (Fig. 9a). This is particularly evident for the western and northern types as well as for the BM type.

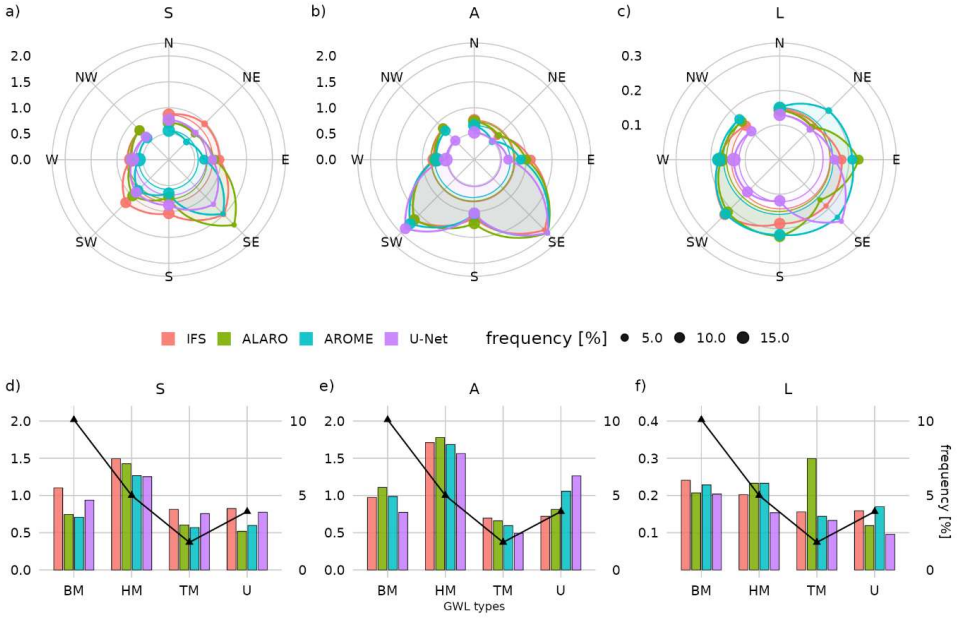

Figure 8: Variability of the median of absolute values of the SAL components across the GWL circulation types grouped by representative airflow direction (a-c). Types without a specific airflow direction are depicted separately in panels d-f. The relative frequency of each directional group is represented by the size of a dot (panels a-c), while the relative frequency of the remaining GWL types is shown on the secondary axis (panels d-f).

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

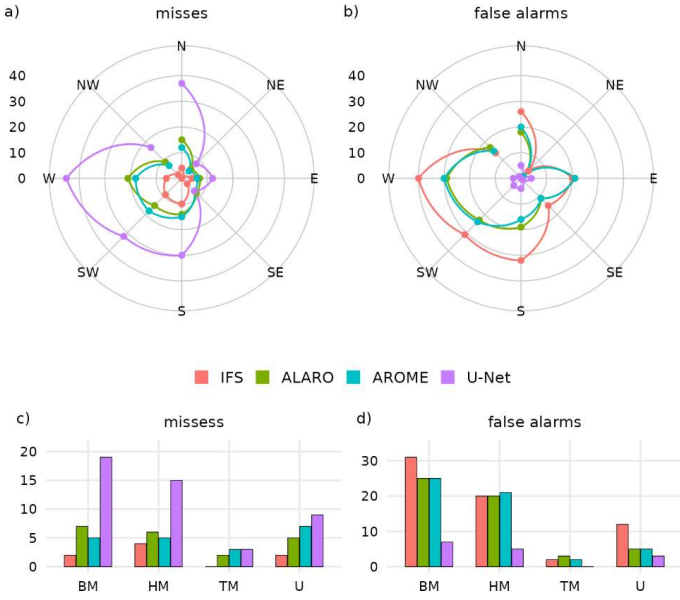

Figure 9: Variability of the number of misses and false alarms across the GWL circulation types grouped by representative airflow direction (a-b). Types without a specific airflow direction are depicted separately in panels c-d.

4.4 Importance assessment of the input models

Ablation experiments were performed to quantify the impact of each input NWP model on the performance of U-Net. To this end, all features derived from a given NWP model were removed. Then, U-Net was retrained using the same cross-year procedure as in the main model but with a single seed (see Sect. 3.4). This resulted in 12 runs for every NWP model. Finally, the mean of validation scores from the runs was calculated and expressed as a relative change (in %) with respect to the reference model, i.e. the one including all NWP models. The validation scores involve both validation loss as well as the aggregated SAL metric SAL_{agg} (calculated on the validation set). Estimation uncertainty was quantified using paired non-parametric bootstrap confidence intervals over temporal split pairs (Efron and Tibshirani, 1986). For each NWP model, a set of 12 values of relative change was bootstrapped with replacement 10000 times to create a distribution of the mean relative change. The 2.5^{th} and 97.5^{th} percentiles of the distribution were used as the 95% confidence interval. The validation scores as well as their uncertainty are shown in Fig. 10. IFS is clearly the most informative input for the U-Net model. Excluding its features increases both the validation loss and SAL_{agg} by around 10%. For ALARO and AROME, the impact is much smaller and differs between the two evaluation metrics. If the validation loss is considered, both models have small but positive contributions to the overall performance. However, results for SAL_{agg} suggest that ALARO is neutral while AROME makes

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

485 the U-Net performance even worse. Although both metrics suggest that ALARO provides more informative input than AROME, their confidence intervals partly overlap.

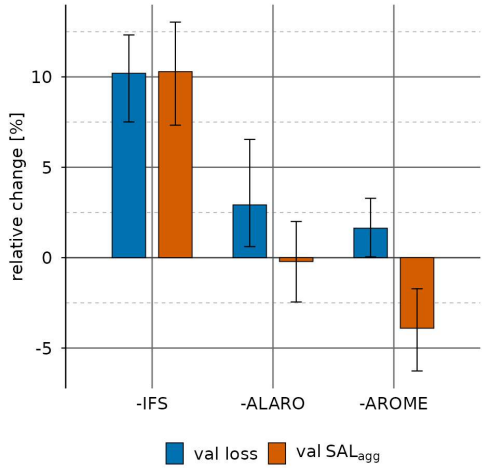

Figure 10: Mean relative changes [%] in the validation loss and the SAL_{agg} metric after the exclusion of inividual NWP models relative to the reference model. The 95% confidence intervals are indicated by whiskers.

490 **5 Discussion**

5.1 Circumstances favouring poor performance of U-Net

Based on the analyses presented above, the following major deficiencies of the U-Net could be identified: 1) overestimation of the structure component in summer; 2) underestimation of domain-maximum precipitation for warning-relevant accumulations; 3) an increased frequency of miss cases. Regarding the first issue, oversmoothing of precipitation fields is a
495 common problem in CNN-based models, as observed and discussed, e.g., by Hess et al. (2023), Rojas-Campos et al. (2023), Zhong et al. (2025) and Ayzel et al. (2020). The cited studies associate also spatial smoothing with the attenuation of small-scale heavy precipitation events, which relates to the second issue indicated above. Our attempts to alleviate it by introducing a dedicated component into the loss function yielded little improvement. One possible issue here concerns the tuning of the component weight – if no satisfactory effect was observed, the weight could be increased. However, as shown by our auxiliary
500 experiments, the resulting improvement in the S component was accompanied by deterioration in the A component (and vice versa). A similar relationship was also observed when tuning the miss weight. Therefore, optimizing the loss function component weights effectively involved a trade-off between structure and amplitude.

Although the tendency of U-Net models to excessively smooth precipitation fields has been widely reported, the analysis of its monthly variability casts a new light on the problem. First of all, the issue occurs mostly during the warm part of the year

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

(May – August), while it is marginal in autumn and winter. This was not reported by Rojas-Campos et al. (2023) or Ayzel et al. (2020) since they did not examine the monthly variability of the scores. To our knowledge, this seasonal pattern has rarely been reported in previous studies. A few studies conducted over regions of China have revealed the largest errors of ML methods in summer. However, their study areas experience the monsoon during this time of year, so non-normalized pixel-based error metrics may be affected by the seasonal variability of precipitation (Zhang et al., 2023; Zhang and Ye, 2021). Additionally, although CNN-based architectures were considered, neither study employed a U-Net specifically. Another factor limiting a direct comparison is the fact that the observed seasonal dependence may vary across climatic regions. Ha and Park (2024) found the highest U-Net skill in late summer, however, large-scale, advected rainfall typically prevails in their study area during this period.

Although the linkage between oversmoothing of precipitation objects and underestimation of the amplitude for heavy precipitation as reported by Rojas-Campos (2023) and Ayzel (2020) was not directly investigated in this study, this could be inferred by the abundance of points in the fourth quadrant in Fig. 3. However, examining monthly variability in Fig. 6 revealed that the period of the largest overestimation of structure does not correspond to the largest underestimation of amplitude. As a matter of fact, the highest amplitude error occurs for autumn months, when the structure error is close to zero. These findings seem to suggest that negative bias in precipitation amplitude does not necessarily coexists with positive bias in structure. As convection in autumn is typically infrequent in Poland (Poręba et al., 2022), the underestimation might also occur for large-scale precipitation events with considerable daily sums (hence reflected in warning-relevant intervals in Fig. 5). However, it is important to be reminded that monthly variability in Fig. 6 is represented by the median value. Therefore, the median close to zero can be either a sign of great prediction skill or information that the frequency of cases with overestimation is roughly the same as underestimated cases. Additionally, warning-relevant precipitation constitutes only a tiny fraction of the dataset and consequently, might be barely reflected by measures of central tendency. To fully address this issue, monthly histograms of observed versus forecasted precipitation should be analysed.

Despite an inherent tendency for oversmoothing, several studies reported U-Net yielding considerable improvement in representing summer extreme precipitation. A particularly relevant example is the work of Badrinath et al. (2023). To address the observed underestimation of heavy precipitation in their initially developed U-Net model, they designed an additional model specifically to handle extreme precipitation. As a result, the performance for the heaviest 10% precipitation events improved by around 20%. However, they link extreme events with atmospheric river patterns, which represent rather large-scale phenomena. Chen et al. (2023) introduced an intensity-weighted loss function which led to improved predictions of heavy summer rainfall over China compared to a standard U-Net. A similar approach was adopted by Hess and Boers (2022). These studies, however, lack spatial verification metrics which would allow for a more direct comparison with the scores obtained in this work. Nevertheless, they indicate potential ways to improve the model, which could be explored in future work.

An alternative direction for future work is also to investigate different ML methods. For example, Random Forests have been used to skilfully forecast the probability of storm-scale convection by Gagne et al. (2014) and Loken et al. (2019). A more

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

recent alternative involves a conditional generative adversarial network, which is reportedly able to reflect small-scale, sparse convection events more accurately than U-Net (Ravuri et al., 2021; Rojas-Campos et al., 2023; Xu et al., 2026).

5.2 Importance of the input models

Ablation experiments carried out to assess the relative importance of the input models yielded an unexpected result - the contribution of the NWP model with the highest spatial resolution and the highest forecast accuracy (AROME) turned out to be marginal or even negative. Conversely, the model with the coarsest resolution (IFS) appeared to be the most informative. The comparison of the models might not, however, be entirely fair due to the unequal number of features provided by the individual models. There are five extra features from IFS (four precipitation components and lightning density) which were added to the model during feature selection because their inclusion yielded better validation scores. Analogous fields from ALARO and AROME turned out to be either irrelevant (precipitation components) or not available (i.e., not stored in the local archive). A similar pattern of relative importance was reported by Zhong et al. (2025), who combined forecasts from IFS and two local models with comparable or finer horizontal resolution. Even though each NWP model provided the same set of features to an ML model, IFS turned out to be the most informative. However, the authors examined the variability of importance for different precipitation thresholds and found that for extremely heavy precipitation (>50 mm $3h^{-1}$), the high-resolution model was as informative as IFS.

A second factor that might favour IFS over ALARO and AROME is information redundancy. Since ALARO provides initial and lateral boundary conditions to AROME, their forecasts are correlated to some extent. Hence, some information contained in AROME might already be represented by ALARO. It is commonly known that correlation among features in machine learning can lead to underestimation of their importance (Nicodemus et al., 2010; Scikit-learn developers, 2026). IFS forecasts, in turn, provide a more independent source of information what may increase its importance score.

A third possible explanation is that U-Net was unable to extract fine-scale information from AROME and therefore lost its advantage with respect to other input models. This seems to be supported by the significant overestimation of precipitation structure observed mostly in summer, when convective precipitation prevails (see Fig. 6). In such situations, the added value of fine-scale forecasts is generally expected to be particularly pronounced (Clark et al., 2016; Plas et al., 2017). This is consistent with our analysis – the lowest median of the structure component error in summer is obtained by AROME. Nevertheless, U-Net does not appear to have effectively utilized this advantage. A possible way to tackle this issue is to add more fine-scale fields with high spatial variability, e.g., convection-related fields. Unfortunately, such fields are not stored in IMGW repositories and could only be obtained by rerunning the forecasts. Although technically feasible, this approach would require considerable time and computational resources.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

6 Conclusions

In this study, a U-Net model was developed to combine information from three NWP models and increase the accuracy of daily precipitation forecasts. Its performance was thoroughly assessed using SAL to indicate its strengths and weaknesses in relation to the input models. A major deficiency of the model identified in the study is severe overestimation of the structure component, an issue commonly reported for U-Net models. However, our results revealed that it is confined to the period from May to August, largely corresponding to the thunderstorm season in the study area. During this period, the performance of the NWP models generally improves with increasing horizontal resolution. In winter, all NWP models exhibit a positive bias for this component, whereas U-Net is generally unbiased.

Considerable underestimation of domain-maximum precipitation for warning-relevant sums makes U-Net the least useful model for predicting heavy precipitation events. To some extent, this is likely related to oversmoothing of precipitation fields, however, no distinct rise in amplitude error for U-Net is visible in summer months. This suggests that some cases of amplitude underestimation might be related to heavy large-scale precipitation events occurring outside the thunderstorm season. Conversely, during some summer convective events, the amplitude might be estimated correctly despite forecasting excessively large objects. On average, U-Net outperforms the input models for the amplitude component, mostly by reducing considerable positive bias in winter present in all input models. Another benefit of U-Net concerns the location component. For most months and circulation types, precipitation objects are located more accurately by U-Net. The improvement is, however, relatively small. Analysis of the confusion matrix reveals that U-Net significantly reduces the number of false alarms, which were particularly frequent for the input models in winter (up to 15% of cases). On the other hand, U-Net produces the highest frequency of misses. IFS shows the opposite tendency, with the fewest misses but the most false alarms. As for the variability across the GWL circulation types, southern circulations turned out to be the most problematic for both the NWP models and U-Net. As these circulation types typically produce little precipitation, their poor performance may be related to the predominance of sparse convection under these conditions.

Despite its coarse resolution and moderate verification scores, IFS is the most informative input NWP model for U-Net. This is likely caused by the greater number of IFS predictors used for training. The limited ability of U-Net to extract fine-scale information from high-resolution models (particularly AROME) represents another plausible explanation. To some extent, redundancy between ALARO and AROME may also play a role. Considering U-Net's tendency to oversmooth precipitation fields, it might be inferred that U-Net overly relies on IFS. Including convective parameters from high-resolution models therefore appears to be crucial for deriving greater benefit from ALARO and AROME models. Further research should include examining methods for mitigating data imbalance, such as tailoring the loss function or augmenting the training data, and employing different ML approaches, e.g., generative adversarial networks.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

Appendix A

Table A1 List of model fields retrieved from each NWP model. Precipitation components are shaded in grey.

variable name	unit	atmospheric levels [hPa]	IFS	ALARO	AROME
geopotential	$m^2 s^{-2}$	500, 700, 850, 925	x	x	x
air temperature	K		x	x	x
u wind component	$m s^{-1}$		x	x	x
v wind component	$m s^{-1}$		x	x	x
specific humidity	$kg kg^{-1}$		x		
relative humidity	%			x	x
vertical velocity	$Pa s^{-1}$		x		
CAPE	$J kg^{-1}$		x		
surface geopotential	$m^2 s^{-2}$	-	x	x	x
total column water	$m^2 s^{-2}$		x		
total-column vertically-integrated water vapour	$m^2 s^{-2}$		x		
mean sea level pressure	Pa		x	x	x
low cloud cover	(0-1)		x	x	x
averaged total lightning flash density in the last hour	$km^{-2} day^{-1}$		x		
K index	K		x		
total totals index	K		x		
precipitation type	-		x		
large scale precipitation	m		x		
convective precipitation	m		x		
snowfall	m, $m^2 s^{-2}$		x		x
convective rainfall	$m^2 s^{-2}$			x	
convective snowfall	$m^2 s^{-2}$			x	
large scale rainfall	$m^2 s^{-2}$			x	
large scale snowfall	$m^2 s^{-2}$			x	
rainfall	$m^2 s^{-2}$				x
graupel	$m^2 s^{-2}$				x

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

Table A2 List of circulation types in the Hess-Brezowsky Grosswetterlagen classification (as given by James (2007)).

pattern symbol	description	assigned direction
Wa	westerly anticyclonic	W
Wz	westerly cyclonic	W
WS	south-shifted westerly	W
WW	maritime westerly (block over Eastern Europe)	W
SWa	westerly anticyclonic	SW
SWz	south-westerly cyclonic	SW
NWa	north-westerly anticyclonic	NW
NWz	north-westerly cyclonic	NW
HM	high over Central Europe	-
BM	zonal ridge across Central Europe	-
TM	low over central Europe	-
Na	northerly anticyclonic	N
Nz	northerly cyclonic	N
HNa	Icelandic high anticyclonic	N
HNz	Icelandic high cyclonic	N
HB	high over British Isles	N
TRM	trough over the British Isles	N
NEa	north-easterly anticyclonic	NE
NEz	north-easterly cyclonic	NE
HFa	Scandinavian high, ridge across C. Europe	E
HFz	Scandinavian high, trough across C. Europe	E
HNFa	Scandinavia-Iceland high, ridge across C. Europe	E
HNFz	Scandinavia-Iceland high, trough across C. Europe	E
SEa	south-easterly anticyclonic	SE
SEz	south-easterly cyclonic	SE
Sa	southern anticyclonic	S
Sz	southern cyclonic	S
TB	low over the British Isles	S
TRW	trough over Western Europe	S
U	undefined or transitional pattern	-

605

Table A3 Temporal split configurations used in the cross-year evaluation of the U-Net model. Each split was repeated for three seeds.

No.	training set	validation set	test set

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

1	2018-2022	2023	2024
2	2018–2019,2021–2023	2020	2024
3	2018–2022	2024	2023
4	2018–2019,2021–2022,2024	2020	2023
5	2018–2021,2023	2024	2022
6	2018–2021,2023,2024	2020	2022
7	2018–2020,2022–2023	2024	2021
8	2018–2019,2022–2024	2020	2021
9	2018–2019,2021–2023	2024	2020
10	2021–2024	2018–2019	2020
11	2020–2023	2024	2018–2019
12	2021–2024	2020	2018–2019

Code and data availability

Forecasts of ALARO and AROME as well as the reference data (RainGRS estimations) are publicly available at https://danepubliczne.imgw.pl/ (last accessed: 10 August 2026). This does not include reference data from 2 March 2021 onwards, as these data come from additional reanalyses available only internally at IMGW. IFS forecasts are available at the ECMWF Meteorological Archival and Retrieval System (https://apps.ecmwf.int/mars-catalogue, last accessed: 10 July 2026), subject to the applicable access conditions.

The source code developed to train and evaluate the U-Net model is available on Zenodo at https://doi.org/10.5281/zenodo.23214572 (Stachura, 2026).

Author contributions

GS and ZU conceptualized the goals of the research; GS downloaded and preprocessed the data. The development of methodology and training of ML models was performed by GS and BB. PS delivered descriptions of the NWP models as well as the SAL method. AJ described the reference data and supported their postprocessing. ZU prepared climatological description of the study area. Investigations and validation were done by GS under the supervision of ZU and BB. All co-authors reviewed and proposed corrections to the final version of the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.

Acknowledgements

The authors acknowledge the use of OpenAI ChatGPT (version 5.5) during the preparation of this study. ChatGPT was used for correcting grammar and spelling errors and logic structure of the text. The OpenAI Codex was used to develop a set of Python scripts responsible for training, validation and testing of the U-Net model based on prompts and specifications provided by the authors. The authors reviewed and verified the AI-assisted code and take full responsibility for its implementation and use in the study.

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018763. The results were produced with partial support of the DS-6 projects run in IMGW-PIB.

Financial support

The research has been supported by a grant from the Faculty Geography and Geology of Jagiellonian University in Kraków as well as from statutory fundings of the Institute of Meteorology and Water Management – National Research Institute. More detailed information will be given upon the acceptance of the manuscript.

References

ACCORD All Staff Workshop, https://www.umr-cnrm.fr/accord/?5th-All-Staff-Workshop-31-March-4-April-2025-Zalakaros-hybrid, last access: 24 July 2026, 2025.

ACCORD: ACCORD Strategy 2021-2025, https://www.umr-cnrm.fr/accord/IMG/pdf/strategy.pdf, last access: 27 August 2026, 2020., 2020.

Ayzel, G., Scheffer, T., and Heistermann, M.: RainNet v1.0: a convolutional neural network for radar-based precipitation nowcasting, Geosci. Model Dev., 13, 2631–2644, https://doi.org/10.5194/gmd-13-2631-2020, 2020.

Badrinath, A., Monache, L. D., Hayatbini, N., Chapman, W., Cannon, F., and Ralph, M.: Improving Precipitation Forecasts with Convolutional Neural Networks, Weather and Forecasting, 38, 291–306, https://doi.org/10.1175/WAF-D-22-0002.1, 2023.

Bielec-Bakowska, Z. and Lupikasza, E.: Long-term precipitation variability on thunderstorm days in Poland (1951–2000), Atmospheric Research, 93, 506–515, https://doi.org/10.1016/j.atmosres.2008.09.018, 2009.

Bishop, C. M.: Neural Networks for Pattern Recognition, Oxford University PressOxford, https://doi.org/10.1093/oso/9780198538493.001.0001, 1995.

Böker, B., Laux, P., Olschewski, P., and Kunstmann, H.: Added value of an atmospheric circulation pattern-based statistical downscaling approach for daily precipitation distributions in complex terrain, International Journal of Climatology, 43, 5130–5153, https://doi.org/10.1002/joc.8136, 2023.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.



Chen, Y., Huang, G., Wang, Y., Tao, W., Tian, Q., Yang, K., Zheng, J., and He, H.: Improving the heavy rainfall forecasting using a weighted deep learning model, Front. Environ. Sci., 11, https://doi.org/10.3389/fenvs.2023.1116672, 2023.

Clark, P., Roberts, N., Lean, H., Ballard, S. P., and Charlton-Perez, C.: Convection-permitting models: a step-change in rainfall forecasting, Meteorological Applications, 23, 165–181, https://doi.org/10.1002/met.1538, 2016.

Dewettinck, W., Van Den Bleeken, D., Van Ginderachter, M., Adriaensen, L., Van De Vyver, H., Caluwaerts, S., and Termonia, P.: Can AI Weather Models Extrapolate Extremes? Evaluating Generalisation to Unseen Precipitation Events, EMS Annual Meeting 2026, Utrecht, Netherlands, https://doi.org/10.5194/ems2026-487, 2026a.

Dewettinck, W., Van De Vyver, H., Degrauwe, D., Hamdi, R., Van Ginderachter, M., Van Schaeybroeck, B., Van Weverberg, K., Vandelanotte, K., Caluwaerts, S., and Termonia, P.: Evaluation of the ALARO1-SFX (CY43T2) regional climate model over Belgium across different resolutions, Geosci. Model Dev., 19, 7911–7937, https://doi.org/10.5194/gmd-19-7911-2026, 2026b.

Eden, J. M., Widmann, M., Grawe, D., and Rast, S.: Skill, Correction, and Downscaling of GCM-Simulated Precipitation, Journal of Climate, 25, 3970–3984, https://doi.org/10.1175/JCLI-D-11-00254.1, 2012.

Efron, B. and Tibshirani, R.: Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, Statistical Science, 1, 54–75, https://doi.org/10.1214/ss/1177013815, 1986.

Gagne, D. J., McGovern, A., and Xue, M.: Machine Learning Enhancement of Storm-Scale Ensemble Probabilistic Quantitative Precipitation Forecasts, Wea. Forecasting, 29, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1, 2014.

Gerstengarbe, F.-W. and Werner, P. C.: Katalog der Großwetterlagen Europas (1881-2004) nach Paul Hess und Helmut Brezowsky. 6. verbesserte und ergänzte Auflage, 100, Potsdam-Institut für Klimafolgenforschung, Potsdam, http://www.pik-potsdam.de/pik_web/publications/pik_reports/reports/pr.100/pr100.pdf (last access: 16 September 2026), 2005.

Gilleland, E., Ahijevych, D. A., Brown, B. G., and Ebert, E. E.: Verifying Forecasts Spatially, Bull. Amer. Meteor. Soc., 91, 1365–1376, https://doi.org/10.1175/2010BAMS2819.1, 2010.

Giot, O., Termonia, P., Degrauwe, D., De Troch, R., Caluwaerts, S., Smet, G., Berckmans, J., Deckmyn, A., De Cruz, L., De Meutter, P., Duerinckx, A., Gerard, L., Hamdi, R., Van Den Bergh, J., Van Ginderachter, M., and Van Schaeybroeck, B.: Validation of the ALARO-0 model within the EURO-CORDEX framework, Geosci. Model Dev., 9, 1143–1152, https://doi.org/10.5194/gmd-9-1143-2016, 2016.

Gofa, F., Boucouvala, D., Louka, P., and Flocas, H. A.: Spatial verification approaches as a tool to evaluate the performance of high resolution precipitation forecasts, Atmospheric Research, 208, 78–87, https://doi.org/10.1016/j.atmosres.2017.09.021, 2018.

Ha, J.-H. and Park, J.: Advancing very short-term rainfall prediction with blended U-Net and partial differential approaches, Front. Earth Sci., 11, https://doi.org/10.3389/feart.2023.1301523, 2024.

Hess, P. and Boers, N.: Deep Learning for Improving Numerical Weather Prediction of Heavy Rainfall, Journal of Advances in Modeling Earth Systems, 14, e2021MS002765, https://doi.org/10.1029/2021MS002765, 2022.

Hess, P., Lange, S., Schötz, C., and Boers, N.: Deep Learning for Bias-Correcting CMIP6-Class Earth System Models, Earth's Future, 11, e2023EF004002, https://doi.org/10.1029/2023EF004002, 2023.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.



Hu, Y., Yin, F., and Zhang, W.: Deep learning-based precipitation bias correction approach for Yin–He global spectral model, Meteorological Applications, 28, e2032, https://doi.org/10.1002/met.2032, 2021.

IFS documentation, https://www.ecmwf.int/en/publications/ifs-documentation, last access: 19 August 2026, 2022.

Integrated Forecasting System, https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model, last access: 19 August 2026, 2022.

James, P. M.: An objective classification method for Hess and Brezowsky Grosswetterlagen over Europe, Theor. Appl. Climatol., 88, 17–42, https://doi.org/10.1007/s00704-006-0239-3, 2007.

Jurczyk, A., Szturc, J., Otop, I., Ośródka, K., and Struzik, P.: Quality-Based Combination of Multi-Source Precipitation Data, Remote Sensing, 12, 1709, https://doi.org/10.3390/rs12111709, 2020.

Kirschenstein, M. and Baranowski, D.: Sumy opadów atmosferycznych w Polsce w latach 1951-1995, Badania Fizjograficzne, 56, 55–72, 2005.

Kolonko, M., Szczęch-Gajewska, M., Bochenek, B., Stachura, G., and Sekuła, P.: Using ALARO and AROME numerical weather prediction models for the derecho case on 11 August 2017, Meteorol. Hydrol. Water Manage., 10, 88–105, https://doi.org/10.26491/mhwm/156260, 2023.

Krasnopolsky, V. M. and Lin, Y.: A Neural Network Nonlinear Multimodel Ensemble to Improve Precipitation Forecasts over Continental US, Advances in Meteorology, 2012, 649450, https://doi.org/10.1155/2012/649450, 2012.

Li, W., Pan, B., Xia, J., and Duan, Q.: Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts, Journal of Hydrology, 605, 127301, https://doi.org/10.1016/j.jhydrol.2021.127301, 2022.

Li, W., Zhang, J., Sun, R., and Duan, Q.: Evaluation of Tianji and ECMWF high-resolution precipitation forecasts for extreme rainfall event in Henan in July 2021, Water Science and Engineering, 16, 122–131, https://doi.org/10.1016/j.wse.2022.11.002, 2023.

Loken, E. D., Clark, A. J., McGovern, A., Flora, M., and Knopfmeier, K.: Postprocessing Next-Day Ensemble Probabilistic Precipitation Forecasts Using Random Forests, Weather and Forecasting, 34, 2017–2044, https://doi.org/10.1175/WAF-D-19-0109.1, 2019.

Lopez, P.: Implementation and validation of a new prognostic large-scale cloud and precipitation scheme for climate and data-assimilation purposes, Quart J Royal Meteoro Soc, 128, 229–257, https://doi.org/10.1256/00359000260498879, 2002.

Lopez, P.: Cloud and Precipitation Parameterizations in Modeling and Variational Data Assimilation: A Review, Journal of the Atmospheric Sciences, 64, 3766–3784, https://doi.org/10.1175/2006JAS2030.1, 2007.

Łupikasza, E. B. and Małarzewski, Ł.: Trends in the indices of precipitation phases under current warming in Poland, 1966–2020, Advances in Climate Change Research, 14, 97–115, https://doi.org/10.1016/j.accre.2022.11.012, 2023.

Mass, C.: The Uncoordinated Giant II: Why U.S. Operational Numerical Weather Prediction Is Still Lagging and How to Fix It, Bulletin of the American Meteorological Society, 104, E851–E871, https://doi.org/10.1175/BAMS-D-22-0037.1, 2023.

Monaco, L., Cremonini, R., and Laio, F.: Precipitation forecast post-processing: blending deterministic NWPs with machine learning, EGU24-2361, Copernicus Meetings, https://doi.org/10.5194/egusphere-egu24-2361, 2024.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.



Nicodemus, K. K., Malley, J. D., Strobl, C., and Ziegler, A.: The behaviour of random forest permutation-based variable importance measures under predictor correlation, BMC Bioinformatics, 11, 110, https://doi.org/10.1186/1471-2105-11-110, 2010.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, https://doi.org/10.48550/arXiv.1912.01703, 3 December 2019., 2019.

Peng, X., Li, Q., Chen, L., Ning, X., Chu, H., and Liu, J.: A Structured Graph Neural Network for Improving the Numerical Weather Prediction of Rainfall, JGR Atmospheres, 128, e2023JD039011, https://doi.org/10.1029/2023JD039011, 2023.

Pic, R., Dombry, C., Naveau, P., and Taillardat, M.: Distributional Regression U-Nets for the Postprocessing of Precipitation Ensemble Forecasts, Artificial Intelligence for the Earth Systems, 4, https://doi.org/10.1175/AIES-D-24-0067.1, 2025.

Plas, E. van der, Schmeits, M., Hooijman, N., and Kok, K.: A Comparative Verification of High-Resolution Precipitation Forecasts Using Model Output Statistics, Monthly Weather Review, 145, 4037–4054, https://doi.org/10.1175/MWR-D-16-0256.1, 2017.

Poręba, S., Taszarek, M., and Ustrnul, Z.: Diurnal and Seasonal Variability of ERA5 Convective Parameters in Relation to Lightning Flash Rates in Poland, Weather and Forecasting, 37, 1447–1470, https://doi.org/10.1175/WAF-D-21-0099.1, 2022.

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., and Foresti, L.: Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0), Geosci. Model Dev., 12, 4185–4219, https://doi.org/10.5194/gmd-12-4185-2019, 2019.

Quintero Plaza, D. and García-Moya Zapata, J. A.: Statistical Postprocessing of Different Variables for Airports in Spain Using Machine Learning, Advances in Meteorology, 2019, https://doi.org/10.1155/2019/3181037, 2019.

Radanovics, S., Vidal, J.-P., and Sauquet, E.: Spatial Verification of Ensemble Precipitation: An Ensemble Version of SAL, Weather and Forecasting, 33, 1001–1020, https://doi.org/10.1175/WAF-D-17-0162.1, 2018.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., and Mohamed, S.: Skilful precipitation nowcasting using deep generative models of radar, Nature, 597, 672–677, https://doi.org/10.1038/s41586-021-03854-z, 2021.

Roehrig, R.: Recent Advances of the Météo-France global NWP system, 37th Session of the Working Group on Numerical Experimentation, Boulder, USA, 2022.

Rojas-Campos, A., Langguth, M., Wittenbrink, M., and Pipa, G.: Deep learning models for generation of precipitation maps based on numerical weather prediction, Geoscientific Model Development, 16, 1467–1480, https://doi.org/10.5194/gmd-16-1467-2023, 2023.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, vol. 9351, edited by: Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., Springer International Publishing, Cham, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.



Sayeed, A., Choi, Y., Jung, J., Lops, Y., Eslami, E., and Salman, A. K.: A Deep Convolutional Neural Network Model for improving WRF Forecasts, IEEE Trans. Neural Netw. Learning Syst., 34, 750–760, https://doi.org/10.1109/TNNLS.2021.3100902, 2023.

Schulzweida, U.: CDO User Guide, https://doi.org/10.5281/ZENODO.10020800, 2023.

Scikit-learn developers: Scikit-learn documentation, https://scikit-learn/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html, last access: 23 September 2026., 2026.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France Convective-Scale Operational Model, Monthly Weather Review, 139, 976–991, https://doi.org/10.1175/2010MWR3425.1, 2011.

Sha, Y., Gagne Ii, D. J., West, G., and Stull, R.: Deep-Learning-Based Gridded Downscaling of Surface Meteorological Variables in Complex Terrain. Part II: Daily Precipitation, Journal of Applied Meteorology and Climatology, 59, 2075–2092, https://doi.org/10.1175/JAMC-D-20-0058.1, 2020.

Sinclair, S. and Pegram, G.: Combining radar and rain gauge rainfall estimates using conditional merging, Atmospheric Science Letters, 6, 19–22, https://doi.org/10.1002/asl.85, 2005.

Stachura, G.: Training and validation of U-Net to reduce precipitation bias, https://doi.org/10.5281/ZENODO.23214572, 2026.

Szturc, J., Jurczyk, A., Ośródka, K., Kurcz, A., Szaton, M., Figurski, M., and Pyrc, R.: Can we reliably estimate precipitation with high resolution during disastrously large floods?, Hydrol. Earth Syst. Sci., 29, 5405–5427, https://doi.org/10.5194/hess-29-5405-2025, 2025.

Taillardat, M. and Mestre, O.: From research to applications – examples of operational ensemble post-processing in France using machine learning, Nonlinear Processes in Geophysics, 27, 329–347, https://doi.org/10.5194/npg-27-329-2020, 2020.

Tapiador, F. J., Roca, R., Del Genio, A., Dewitte, B., Petersen, W., and Zhang, F.: Is precipitation a good metric for model performance?, Bull Am Meteorol Soc, 100, 223–233, https://doi.org/10.1175/BAMS-D-17-0218.1, 2019.

Termonia, P., Fischer, C., Bazile, E., Bouyssel, F., Brožková, R., Bénard, P., Bochenek, B., Degrauwe, D., Derková, M., El Khatib, R., Hamdi, R., Mašek, J., Pottier, P., Pristov, N., Seity, Y., Smolíková, P., Španiel, O., Tudor, M., Wang, Y., Wittmann, C., and Joly, A.: The ALADIN System and its canonical model configurations AROME CY41T1 and ALARO CY40T1, Geosci. Model Dev., 11, 257–281, https://doi.org/10.5194/gmd-11-257-2018, 2018.

Trotta, B., Owen, B., Liu, J., Weymouth, G., Gale, T., Hume, T., Schubert, A., Canvin, J., Mentiplay, D., Whelan, J., and Johnson, R.: RainForests: A Machine Learning Approach to Calibrating NWP Precipitation Forecasts, Weather and Forecasting, 39, 1715–1732, https://doi.org/10.1175/WAF-D-23-0211.1, 2024.

Ustrnul, Z. and Czekierda, D.: Circulation background of the atmospheric precipitation in Central Europe (based on the Polish example), metz, 10, 103–111, https://doi.org/10.1127/0941-2948/2001/0010-0103, 2001.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Bouallègue, Z. B., Bhend, J., Dabernig, M., Cruz, L. D., Hieta, L., Mestre, O., Moret, L., Plenković, I. O., Schmeits, M., Taillardat, M., Bergh, J. V. den, Schaeybroeck, B. V., Whan, K., and Ylhaisi, J.: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World, Bulletin of the American Meteorological Society, 102, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1, 2021.

https://doi.org/10.5194/egusphere-2026-7158
Preprint. Discussion started: 24 November 2026
© Author(s) 2026. CC BY 4.0 License.



Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL - A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts, Monthly Weather Review, 136, 4470–4487, https://doi.org/10.1175/2008MWR2415.1, 2008.

Wilson, T., McDonald, A., Galib, A. H., Tan, P.-N., and Luo, L.: Beyond Point Prediction: Capturing Zero-Inflated & Heavy-Tailed Spatiotemporal Data with Deep Extreme Mixture Models, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2020–2028, https://doi.org/10.1145/3534678.3539464, 2022.

Xu, J., Dai, K., Ma, J., Zhang, Q., Chen, Y., Zhang, F., and Ng, C.-P.: Postprocessing for 24-Hour Advanced Forecasting of Extreme Precipitation Using Deep Learning Generative Models, Weather and Forecasting, 41, 381–401, https://doi.org/10.1175/WAF-D-24-0199.1, 2026.

Zhang, J., Gao, Z., and Li, Y.: Deep-Learning Correction Methods for Weather Research and Forecasting (WRF) Model Precipitation Forecasting: A Case Study over Zhengzhou, China, Atmosphere, 15, 631, https://doi.org/10.3390/atmos15060631, 2024.

Zhang, T., Liang, Z., Li, W., Wang, J., Hu, Y., and Li, B.: Statistical post-processing of precipitation forecasts using circulation classifications and spatiotemporal deep neural networks, Hydrology and Earth System Sciences, 27, 1945–1960, https://doi.org/10.5194/hess-27-1945-2023, 2023.

Zhang, Y. and Ye, A.: Machine Learning for Precipitation Forecasts Postprocessing: Multimodel Comparison and Experimental Investigation, J. Hydrometeor., 22, 3065–3085, https://doi.org/10.1175/JHM-D-21-0096.1, 2021.

Zhong, Q., Fang, Z., Chen, H., Shen, L., Zhang, Y., and Hou, S.: Fusing Multi-Scale Numerical Model Forecasts to Improve Short-Term Intense Rainfall Forecast with a Deep Learning Rain Network, Journal of Meteorological Research, 39, 1441–1462, https://doi.org/10.1007/s13351-025-4151-0, 2025.