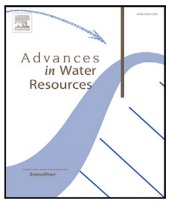
Advances in Water Resources 204 (2025) 105061

Contents lists available at ScienceDirect

Advances in Water Resources

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



Performance assessment of neural network models for seasonal weather forecast postprocessing in the Alpine region

Sameer Balaji Uttarwar [a,*], Sebastian Lerch [b,c,d], Diego Avesani [a], Bruno Majone [a]

[a] Department of Civil, Environmental and Mechanical Engineering, University of Trento, via Mesiano, 77, Trento, 38123, Trentino-South Tyrol, Italy
[b] Department of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Straße 6, 35043, Marburg, 38098, Hessen, Germany
[c] Institute of Statistics, Karlsruhe Institute of Technology, Blücherstraße 17, Karlsruhe, 76185, Baden-Württemberg, Germany
[d] Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, Heidelberg, 69118, Baden-Württemberg, Germany

ARTICLE INFO

Dataset link: https://github.com/Sam-Uttarwar/NN_arch.git, https://doi.pangaea.de/10.1594/PANGAEA.924502, https://www.ecmwf.int/en/forecasts/dataset/operational-archive

Keywords:
ECMWF-SEAS5
Statistical postprocessing
Neural networks
Forecast skill evaluation
Complex alpine region
Hydrometeorological applications

ABSTRACT

Seasonal weather forecasts are crucial for water-related sectors. However, the presence of systematic biases limits the usefulness of global seasonal weather forecasts produced by numerical weather prediction models. Although statistical postprocessing approaches, such as empirical quantile mapping, are widely used to improve accuracy and reliability, they have limitations in the accuracy of forecast values outside the training period and difficulties in incorporating multiple static and dynamic environmental variables to capture non-linear dependencies. This study seeks to overcome these limitations by implementing a neural network-based distributional regression method as a postprocessing tool. The study investigates the performance of these methods using seasonal forecasts of total precipitation and 2-meter temperatures for a one-month lead time over the Trentino-South Tyrol region in the northeastern Italian Alps. The forecast dataset is the fifth-generation seasonal weather forecast system (SEAS5) generated by the European Centre for Medium-Range Weather Forecasts (ECMWF), which has a 0.125° x 0.125° horizontal grid resolution with 25 ensemble members over the period from 1981 to 2016. The reference dataset is a high-resolution (250 m x 250 m) gridded observational data over the region. The performance of both methods is evaluated with a focus on the effects of forecast lead times, location, and seasonal variability. Results show that the neural network-based approach consistently outperforms empirical quantile mapping, especially during short lead times and at higher elevations.

1. Introduction

One of the most important objectives of using seasonal weather forecasting is to assist water management strategies with the provision of streamflow forecasts generated by a hydrometeorological modeling chain (Emerton et al., 2018; Vogel et al., 2021; Robertson et al., 2024). In particular, hydrological models driven by weather forecasts can play a crucial role in the water–energy–food nexus (Kumar et al., 2023), allowing decision-makers to optimize planning and resource allocation in critical sectors such as water and energy, thereby ensuring that community needs are more effectively addressed (Klemm and McPherson, 2017; Alexander and Block, 2022). Furthermore, accurate streamflow forecasts are also essential for mitigating risks associated with extreme weather events, such as droughts and floods, thereby facilitating the optimal operation and management of hydraulic infrastructure (Tripathy et al., 2020).

These challenges are particularly acute in the Alpine region, often referred to as Europe's "water tower", due to its unique topographic and climatic characteristics (Hohenwallner et al., 2011; Brouwer et al., 2013). The unique topography and climatic characteristics of the Alpine region not only play a pivotal role in shaping precipitation patterns, hydrological processes, and regional water balances (Viviroli and Weingartner, 2004) but also are considered a hot-spot for the likely effects of climate change (Viviroli et al., 2007; Gobiet et al., 2014; de Jong, 2015; Arnoux et al., 2020).

In response to these challenges, considerable efforts have been made to enhance the performance of numerical weather prediction (NWP) models, including seasonal models (Johnson et al., 2019; Brotzge et al., 2023), by applying postprocessing methods designed to correct errors in raw ensemble forecasts (Hemri et al., 2014; Vannitsem et al., 2018). Furthermore, Crochemore et al. (2016) illustrated that post-processed forecasts could improve the skill of streamflow forecasts for lead times of up to 3 months. Golian and Murphy (2022) compared different statistical postprocessing methods in improving the accuracy of seasonal precipitation forecasts from ECMWF-SEAS5 over

* Corresponding author.
E-mail address: sameer.uttarwar@unitn.it (S.B. Uttarwar).

https://doi.org/10.1016/j.advwatres.2025.105061
Received 7 April 2025; Received in revised form 21 July 2025; Accepted 21 July 2025
Available online 29 July 2025

lead times of 1–6 months. Over the past years, the use of machine learning (ML) techniques has been gaining significant attention, often outperforming traditional state-of-the-art postprocessing methods such as ensemble model output statistics (EMOS; Gneiting et al., 2005) and Bayesian model averaging (BMA; Raftery et al., 2005). ML-based approaches offer the distinct advantage of straightforwardly enabling the incorporation of additional predictor variables beyond the ensemble forecasts of the target variable, while effectively capturing non-linear dependencies in a data-driven manner (Vannitsem et al., 2021).

In a pioneering paper, Rasp and Lerch (2018) employed neural networks (NNs) to estimate the parameters of a specified probability distribution for 2-m temperature at a lead time of 48 h, allowing for the modeling of non-linear relationships between arbitrary predictor variables (i.e., ensemble predictors and station-specific information) and forecast distribution parameters. This parametric distributional regression approach and related methods have been extended towards other weather variables and use cases, including wind gusts (Schulz and Lerch, 2022) and solar irradiance (Horat et al., 2024). Ghazvinian et al. (2021) utilized an NN-based approach that outperforms EMOS schemes in postprocessing ensemble forecasts of precipitation, specifically focusing on spatially averaged precipitation at the sub-basin level with a seven-day lead time.

Motivated by the successes of ML methods on shorter time scales and the broader need to enhance seasonal forecasting skills in alpine regions, this study aims to evaluate the performance of these techniques for probabilistic postprocessing of two key hydrometeorological variables – precipitation and temperature – in regions characterized by complex terrain. Specifically, the evaluation focuses on the Trentino-South Tyrol region, situated in the northeastern Italian Alps, and utilizes the reforecasts from the European Centre for Medium-Range Weather Forecasts - System 5 (ECMWF-SEAS5; Johnson et al., 2019) for the period 1981–2016. The standard Empirical Quantile Mapping (E-QM) approach is employed here as a benchmark for comparison. Given the topographic and climatic diversity, the region is a challenging case study for further developing advanced postprocessing techniques for seasonal weather forecasting. Understanding the dependencies of different static and dynamic environmental processes with respect to complex topography is crucial for improving forecast accuracy. The static processes include elevation that remains constant over time and affects the atmospheric circulation, moisture content, and energy fluxes. On the other hand, dynamic processes, such as precipitation, temperature, wind patterns, cloud cover, and humidity, are time-varying and dependent on terrain characteristics.

A key novelty of this work is its departure from the site-specific training and evaluation of NN schemes used in previous studies. Instead, this study utilizes the gridded observational product developed by Crespi et al. (2021a), which provides accurate data at a 250 m × 250 m resolution at a daily time scale for the period of 1981–2018. To the best of our knowledge, this study is the first application of an NN-based probabilistic postprocessing approach for seasonal weather forecasts using a high-resolution gridded observational dataset over the topographically complex Alpine region, also targeting forecasts at a daily timestep, an aspect which is particularly valuable for hydrological applications in the region. While we acknowledge that the use of a gridded observation dataset is subject to an additional source of uncertainty related to the interpolation procedure, which sums up to the inherent bias present in any measurement data, the benefit of producing high-resolution postprocessed forecast data is of pivotal importance for hydrological applications in a complex topographic domain like the Alpine region.

The remainder of the paper is organized as follows: Section 2 describes the study area and data. Section 3 provides a detailed overview of the empirical quantile mapping (E-QM) and neural network (NN) statistical postprocessing techniques. Section 4 evaluates and discusses the forecast performance of both these methods, and finally, in Section 5, concluding remarks are drawn.

2. Study area and datasets

2.1. Study area

The Trentino-South Tyrol region is geographically located between latitude 45.5° to 47.5°N and longitude 10.5° to 12.5°E with an area of approximately 13,000 km^2. The region is located in the northeastern Italian Alps (as depicted in the inset of Fig. 1), and is characterized by a highly diverse and complex topography. The elevation ranges from 61 m (m a.s.l.) in the valleys to 4000 m (m a.s.l.) in the mountain peaks, with an average elevation of 1600 m, making it a challenging location for accurate weather forecasting. The region has only 4% gently sloped terrain (slope steepness ≤5%) of the total area, with the remaining 96% being characterized by steep slopes (slope steepness ≥5%). These terrain characteristics enhance substantial variations of local climatic processes and, thus, weather patterns across short distances. Hydrometeorological weather variables, such as temperature and precipitation, are influenced by temperature inversion and orographic uplifting. The valley region experiences relatively moderate winters and hot summers, while the mountains experience cold winters with heavy snowfall and cool summers around the year. However, the average annual precipitation ranges from around 800 mm in the valleys to over 1600 mm in the mountainous regions (Laiti et al., 2018; Mallucci et al., 2019).

The complex terrain characteristics cause substantial seasonal variations influencing the temperature, snow cover, and precipitation between summer and winter within the region (Morlot et al., 2023; Bertoldi et al., 2023). These factors make Trentino-South Tyrol a particularly challenging region for weather forecasting, particularly for precipitation and temperature.

2.2. Observational dataset

The study utilizes a 250 m × 250 m gridded daily precipitation and mean temperature dataset from 1980 to 2018 as a reference for postprocessing. The gridded reference dataset has been developed by using a climatological interpolation technique (Crespi et al., 2021a) based on observations with daily precipitation and temperature records from 243 and 311 ground stations, respectively. The analysis presented in this work is based on a horizontal grid resolution of 1 km × 1 km, therefore, the observational grid dataset was rescaled from 250 m to 1 km using the average sampling method. The dataset is available from: https://doi.pangaea.de/10.1594/PANGAEA.924502.

2.3. Forecast data

The study utilizes the ECMWF Integrated Forecast System (IFS) version 43r1 global seasonal reforecast dataset (SEAS5), with monthly releases forecasting up to 7 months, including 25 ensemble members for the period of 1981 to 2016 at a 6-h time step (Johnson et al., 2019). The dataset has a horizontal grid resolution of 0.125° × 0.125° (i.e., approximately 14 km). In particular, this study focuses on postprocessing daily forecasts of precipitation and 2-m temperature for lead times of up to one month. The dataset, extracted from ECMWF operational archives, is compiled using the daily one-month lead forecasts from each monthly release, ensuring that the first month of forecast data is used. The ECMWF-SEAS5 dataset significantly differs from its predecessor, ECMWF-SEAS4, with a larger ensemble size (25 in the reforecast, compared to 15 in SEAS4) and a longer reforecast period (1981–2016, compared to 1981–2010 in SEAS4), along with an improved ocean and sea-ice coupling model (Johnson et al., 2019). This extensive dataset, in terms of reforecast years and the number of ensemble members, thus enables more robust training of statistical postprocessing techniques. That said, the monthly releases pose additional challenges for model development, as typical medium-range datasets will often include substantially more (re-)forecast cases due to the more frequent model runs, see, e.g., the EUPPBench dataset (Demaeyer et al., 2023). The forecast variables are resampled to a 1 km × 1 km horizontal grid resolution using the nearest neighbor technique, resulting in 13,289 spatially consistent data grid points across the area.

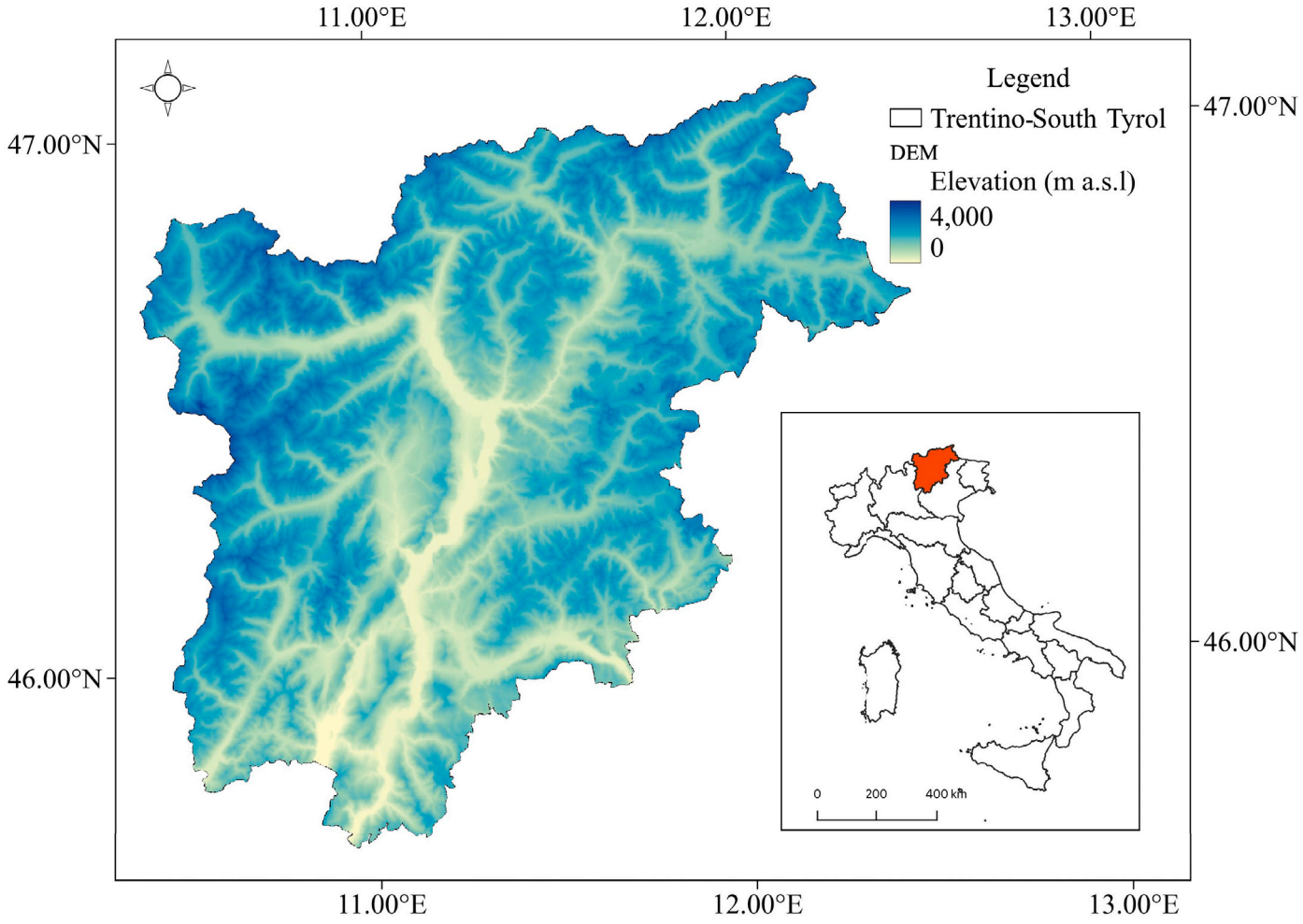


Fig. 1. Study area map with the inset showing the location within the Italian territory. The study area map also displays the digital elevation model of the region.

3. Methods

This section briefly introduces two state-of-the-art univariate post-processing methods for daily precipitation and 2-m temperature. Empirical quantile mapping (E-QM) is employed as the benchmark method against which the neural network approach is evaluated. The available forecast and observation data are divided into 2 parts: the training period (1981–2010, 30 years) used for training and validating the postprocessing methods and the test period (2011–2014, 4 years) used to evaluate the forecast accuracy of both methods. Furthermore, to tune the hyperparameters of the neural network model, a validation period of 3 years (2008–2010) is extracted from the training period, as detailed in the ensuing section. The training period consists of 145,607,573 samples (i.e., 13,289 grid points × 10,957 days), the testing period contains 19,415,229 samples (i.e., 13,289 grid points × 1461 days), and the validation period contains 14,564,744 samples (i.e., 13,289 grid points × 1096 days).

3.1. Empirical Quantile Mapping (E-QM)

The E-QM method adjusts forecast data by aligning the empirical cumulative distribution functions (ECDFs) of observed and forecasted datasets (i.e., considering all 25 ensemble members) over a training period. This process creates a statistical transfer function that is afterwards used to adjust biases during the test period. Once established, the transfer function can be applied to new forecasts ensemble-by-ensemble, improving their accuracy by ensuring they match the statistical characteristics of the observations at each grid point and target day, respectively (Themeßl et al., 2011; Cannon et al., 2015). Following Monhart et al. (2018), the bias correction for a target day was derived from a sample of reforecasts and corresponding observations collected within a 2-month window (i.e., 61 days) centered on the target day. As shown in Fig. 2, the postprocessing function for a specific lead time in a year is derived from a sample of reforecasts and corresponding observations within a 61-day window centered around the target day, thereby capturing essential temporal dynamics and enhancing forecast accuracy. Notice that in this approach, the target day coincides with the definition of lead time, given that we are considering releases occurring always on the first day of each month. The approach is synthetically outlined below:

Following the approach proposed by Monhart et al. (2018), the empirical quantile mapping (E-QM) method adjusts the forecast data by aligning the empirical cumulative distribution functions (ECDFs) of the observed and forecast datasets, considering the 25 members of the ensemble, over a designated training period. This alignment defines a statistical transfer function, which is subsequently applied during the test period to correct forecast biases. The correction is performed thus independently for each ensemble member, improving forecast accuracy by ensuring consistency with the statistical characteristics of the observational reference at each grid point and target day (Themeßl et al., 2011; Cannon et al., 2015).

Specifically, the method applies a quantile-based transformation defined as follows:

$$z_{i,t}^{n'} = ECDF_{i,d}^{o,-1}(ECDF_{i,d}^{f}(x_{i,t}^{n'})) \quad (1)$$

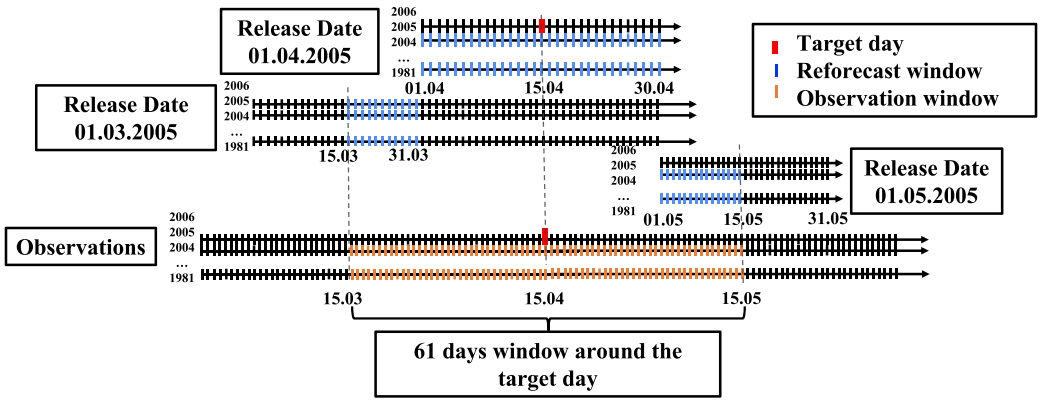


Fig. 2. Schematic illustration of the data used to train the E-QM procedure. For any forecast target day (red mark), the model builds an empirical cumulative distribution function using all reforecast dates within a 61-day window (data from multiple release dates are used for this purpose as outlined by the blue marks in the top rows) and compares them with corresponding observations (orange marks in the bottom row). This results in a sample size of 1769 daily observations (61 days × 29 years) and 44,225 daily reforecast values (61 days × 29 years × 25 ensemble members), which are used to estimate the statistical transfer function for the E-QM technique associated to a given forecast target day.

where $z_{i,t}^{n'}$ denotes the bias corrected forecast for the n'th member of the ensemble at time t and grid point i, and $x_{i,t}^{n'}$ is the corresponding raw forecast; the variable d refers to the calendar day associated with time t; $ECDF_{i,d}^{f}$ represents the empirical cumulative distribution function estimated from the raw forecast; and $ECDF_{i,d}^{o,-1}$ is the inverse of the empirical CDF constructed from observational data.

It is important to note that both $ECDF_{i,d}^{f}$ and $ECDF_{i,d}^{o}$ are individually computed for each grid point i and for each calendar day d of the year. This requires mapping both the forecast times and the observation timestamps to their corresponding calendar day d, ensuring that the data from both sources are consistently aligned. The empirical distributions are then estimated using values collected within a centered 61-day moving window around day d (i.e., ±30 days), aggregated over the historical reference period. Forecast ECDFs include all ensemble members.

For the sake of clarity, Fig. 2 illustrates the construction of the E-QM transfer function for a specific lead time. The schematic shows how reforecast–observation pairs are selected within a 61-day window around the target day and used to derive the empirical CDFs of model and observed data over the training period.

3.2. Neural networks

A neural network consists of interconnected nodes and were originally designed to imitate the structure and functionality of the biological human brain (Maier et al., 2023). These nodes are categorized into layers to process and transfer information. NNs generally include an input layer, multiple hidden layers, and an output layer. Each node in a layer receives input from the nodes in the previous layer, processes the data using an activation function that allows it to introduce non-linearity into the model, and then transmits the output to the nodes in the next layer (Goodfellow et al., 2016). This interconnected network system allows NNs to learn complex relations in the data.

In this study, NN-based parametric distributional regression network (DRN) approaches (Rasp and Lerch, 2018) are implemented for both precipitation and 2-m temperature forecasts over a one-month lead time using a gridded observational dataset as a reference. The DRN method extends the general ensemble model output statistics (EMOS) approach by including ensemble predictors of additional meteorological variables and location-specific environmental inputs. The DRN technique links the input predictor variables to distribution parameters by learning the non-linear relationships in a data-driven manner rather than relying on a predetermined link function. The traditional feed-forward neural network models are typically used for single-value prediction, and they usually do not account for uncertainty. In contrast, DRNs are specifically designed for distributional regression, making them more suitable for probabilistic forecasting applications where both the mean and spread of the forecast are important. For this reason, DRN was selected as the most appropriate model for our objectives. Indeed, there are other architectures that can be used and have been used, such as CNNs (Horat and Lerch, 2024), or GNNs (Feik et al., 2024). That said, our aim here is mainly to first investigate simple architectures that have shown success in related studies, while exploring more advanced approaches will be part of future work. The application of the DRN method for precipitation and 2-m temperature forecast variables differs in the assumed univariate parametric distribution, as it will be explicitly presented in the corresponding subsections.

To enhance the NN model's training, we considered ensemble forecasts of 10 additional meteorological variables as auxiliary features that may particularly relate to the target variables, i.e., precipitation and temperature. Additional features further include the daily observed values of the North Atlantic and Atlantic Oscillation (NAO & AO) indices provided by the NOAA National Weather Service (https://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml), the day of the year (sine and cosine transformed), and grid point latitude and longitude. We also included observed and model grid point elevations and the difference between observed and model grid point elevations as further location-specific information. The additional features were selected based on known physical relationships of each feature to the target variable. A summary of these additional variables is provided in Table 1.

Following Rasp and Lerch (2018), the ensemble forecasts of all variables are transformed into their respective mean and standard deviations and processed in the input layer. Fig. 3 depicts the schematic structure of the NN-based postprocessing model implemented for precipitation (blue color) and 2-m temperature (orange color). The input layer also processes the additional features (cyan color) presented in Table 1. Following Xu et al. (2019), the input features are normalized with respect to their maximum value. The rectangular blue box represents the input layer of the NN model for precipitation and the orange box represents the input layer of the NN model for temperature forecasts, respectively. Additionally, an embedding layer connects to the input layer (see purple nodes in Fig. 3), which allows the network to learn the information about the grid identifier by mapping the discrete grid identifier value into a vector of real numbers (see Table 1). This embedding layer provides a way to transform categorical inputs into a numeric format that captures semantic relationships that are fed into the network (Guo and Berkhahn, 2016). The remaining elements of the NN model are distinct, particularly based on the assumption of a corresponding parametric distribution to postprocess the forecast variable of interest. This is distinguished in Fig. 3 with orange dashed lines (i.e., upper part) for temperature and blue dashed lines (i.e., lower part) for precipitation, respectively.

Table 1
List of adopted features and their abbreviations. Note that the meteorological features are introduced in the NN models as the mean and standard deviation of the ensemble.

Features (Mean & Std. dev.)	
Precipitation	prec
2-m temperature	t2m
2-m dewpoint temperature	d2m
10-m horizontal wind component	u10
10-m vertical wind component	v10
100 hPa horizontal wind component	u100
100 hPa vertical wind component	v100
Total cloud cover	tcc
Surface latent heat flux	slhf
Surface sensible heat flux	sshf
Convective precipitation	cp
Surface net solar radiation	ssr
Other features	
North Atlantic oscillation	nao
Atlantic oscillation	ao
Day of year (sine/cosine transformed)	doy
Elevation of observation data grid	elv
Elevation of model grid	geo
Difference in elevation	elv-geo
Latitude of grid	lat
Longitude of grid	lon
Lead time (days)	lead
Embeddings	
Grid identifier	Grid ID

The NN model is trained using the adaptive moment estimation (Adam) algorithm (Kingma and Ba, 2015), and the weights are optimized over the 1981–2010 (30 years) training period by minimizing the loss function (i.e., continuous ranked probability score, see Sections 3.3 3.3.1).

The following hyperparameters (i.e., adjustable parameters that govern the training process of machine learning models) were tested while training the model, with the values adopted being presented in Table 2:

- Number of nodes in hidden layer: $\{5, 25, 125, 512, 1025\}$
- Batch size (i.e., number of random training samples used in one iteration): $\{1250, 4096\}$
- Epochs (i.e., an epoch corresponds to one complete iteration over the entire training dataset during which the model updates its parameters to minimize the loss function): 5, 10, 30
- Learning rate: $\{1e{-}1, 1e{-}2, 1e{-}3, 1e{-}4, 1e{-}5\}$
- Embedding (i.e., discrete grid point ID mapped to a continuous vector space dimension): 2, 4

To evaluate the model's performance and prevent overfitting, we use 10% of the training period (i.e., the last 3 years: 2008–2010) as a validation period. This procedure ensures that the model's performance accurately reflects its ability to make predictions on unseen data when evaluated with validation period data that was not included in the training set. This subdivision of the training period into the validation period ensures an independent dataset for early stopping and hyperparameter tuning. A single hidden layer is used for both NN-based approaches, as the additional layers with different parameter configurations tested increased the network complexity and computational demand but did not provide significant performance improvements. Ablation studies testing deeper networks with two and three hidden layers, which indicated that a single hidden layer worked best, are presented in Table S1 in the supplementary material (SM). For precipitation, as the number of hidden layers increases, the training loss is slightly worse, and validation loss increases, indicating model performance degradation. Furthermore, for temperature, the training loss slightly improves as the depth of the network increases, but the validation loss is slightly worse, indicating possible overfitting. This observation is well in line with findings from other studies, where different hyperparameter configurations were evaluated and simpler models with only one or two hidden layers were found to work best, and with deeper networks often deteriorating predictive performance (e.g., Rasp and Lerch (2018), Bremnes (2020), Schulz and Lerch (2022) and others). One potential explanation for this observation might be that in the context of postprocessing, more complex neural network models rarely provide any benefits due to the availability of the generally highly informative predictors in the form of NWP ensemble forecasts. During model training, we ensured that the distribution of input features remained consistent across the training and validation datasets. The total number of parameters in the NN model for precipitation and temperature forecasts are 27,231 and 27,205, respectively. Additionally, to account for variability caused by random fluctuations between the model runs, we followed the approach of Rasp and Lerch (2018) by training an ensemble of 10 networks, each one initialized with different random NN weights for each model variant. The loss curves of the NN model for both precipitation and temperature forecast variables during the training and validation periods are presented in Figs. S1 and S2 of the SM, respectively. In particular, the curves show the mean losses, along with the ± 1 standard deviation band for each epoch, obtained by running an ensemble of 10 networks to account for random fluctuations between model runs. For precipitation (see Fig. S1 in SM), the validation loss is slightly above the training loss, which can be attributed to the complexity and high variability of precipitation forecasts. For temperature (see Fig. S2 in SM), the training and validation loss converge in the initial epochs and remain very close to each other, indicating the ability to apply what the model has learned during training and make accurate predictions on unseen data with no overfitting. These ensemble members are then aggregated to produce the final output by averaging the distribution parameters obtained as the output, ensuring more robust and reliable predictions (Schulz et al., 2024). The Python code for postprocessing both precipitation and temperature forecast variables using the NN model is made available through a GitHub repository link: https://github.com/Sam-Uttarwar/NN_arch.git.

Table 2
Hyperparameters used in the NN model for the respective forecast variable of interest.

Variable	Epochs	Learning rate	Batch size	Hidden nodes	Embedding size
Precipitation	10	1e−5	4096	25	2
Temperature	10	1e−4	4096	25	2

3.2.1. Precipitation

As argued by Scheuerer and Hamill (2015), precipitation is a challenging variable for statistical modeling as it exhibits both discrete and continuous characteristics (i.e., dry and wet days), with a non-zero probability of precipitation being zero. Therefore, Scheuerer and Hamill (2015) proposed a censored shifted gamma (CSG) distribution where point masses at zero precipitation are enabled by allowing for probability mass at negative values and left-censoring the obtained distribution at zero. Furthermore, we use the formulation of Scheuerer and Hamill (2015) and Baran and Nemoda (2016) for the cumulative distribution function of CSG (denoted as $\mathcal{G}^0$) which is characterized by the parameters shape (α), scale (β), and shift (δ). The CSG-DRN model for precipitation (see the blue box in Fig. 3) is here implemented over each grid point i and lead time t considering a single hidden layer and a single output layer. In our model, let $y_{i,t}$ represents the precipitation observation data and $X_{i,t}$ denote the vector of mean and standard deviation of the ensemble features (see Table 1), at each grid point i and lead time t. Additionally, the vector of $X_{i,t}$ also includes other features (see Table 1), including location-specific environmental variables, grid identifier, and lead time, to further refine the forecast

and capture spatial variability. The conditional distribution of $y_{i,t}$ given $X_{i,t}$ is modeled by the parametric CSG distribution,

$$y_{i,t} \mid X_{i,t} \sim \mathcal{G}^{0}{}_{(\alpha_{i,t},\beta_{i,t},\delta_{i,t})}, \tag{2}$$

where the shape parameter $\alpha_{i,t}$ and the scale parameter $\beta_{i,t}$ are the distribution parameters predicted by the model. The shape and scale parameters can be computed analytically from the predicted mean $\mu_{i,t}$ and standard deviation $\sigma_{i,t}$ estimated as $\alpha_{i,t} = \mu^2_{i,t}/\sigma^2_{i,t}$ and $\beta_{i,t} = \sigma^2_{i,t}/\mu_{i,t}$ (Scheuerer and Hamill, 2015), respectively. The three parameters $\mu_{i,t}$, $\sigma_{i,t}$ and $\delta_{i,t}$ are the outputs of the CSG-DRN at each grid point i and lead time t. For the CSG-DRN model, we utilize a single hidden layer utilizing the exponential linear unit (ELU) activation function with a parameter $\theta = 1$ to introduce the non-linear relationships to the network, where

$$\mathrm{ELU}(x) = \begin{cases} x, & x > 0 \\ \theta(e^x - 1), & x \le 0. \end{cases}$$

We apply a softplus activation function ranging from [0, +∞] to the output layer to ensure that the NN output is always positive, with

$$\mathrm{softplus}(x) = \log(1 + e^x).$$

3.2.2. Temperature

In this model, let $y_{i,t}$ represent the temperature observation data and $X_{i,t}$ denote the vector of mean and standard deviation of ensemble features (see Table 1), at each grid point i and lead time t. Similarly to precipitation, the vector of $X_{i,t}$ also includes the additional features presented in Table 1. The conditional distribution of $y_{i,t}$ given $X_{i,t}$ for 2-m temperature is modeled by a parametric Gaussian distribution and can be expressed as (Gneiting et al., 2005):

$$y_{i,t} \mid X_{i,t} \sim \mathcal{N}_{(\mu_{i,t},\sigma_{i,t})} \tag{3}$$

The two parameters $\mu_{i,t}$ and $\sigma_{i,t}$ of the Gaussian forecast distribution at each grid point i and lead time t are obtained as the outputs of the DRN (shown in Fig. 3). We utilize a single hidden layer with a rectified linear unit (ReLU) activation function to introduce the non-linear relationships to the network, where

$$\mathrm{ReLU}(x) = \max(0, x).$$

3.3. Forecast verification

3.3.1. Continuous ranked probability score (CRPS)

To compare the performance of postprocessing methods in improving the weather forecast accuracy for each grid point (i) and lead time (t) in the test period from 2011–2014, we use the CRPS (Matheson and Winkler, 1976). The CRPS measures the difference between the forecast cumulative distribution function (CDF), $F(z_{i,t})$, and the step function at the observed value, $\mathbb{I}(z_{i,t} \ge y_{i,t})$, i.e.,

$$\mathrm{CRPS}_{i,t} = \int_{-\infty}^{\infty} \left[F(z_{i,t}) - \mathbb{I}(z_{i,t} \ge y_{i,t})\right]^2 dz, \tag{4}$$

where $F(z_{i,t})$ is the forecast cumulative distribution function and $\mathbb{I}(z_{i,t} \ge y_{i,t})$ is a step function that equals 1 when $z_{i,t} \ge y_{i,t}$ and 0 otherwise. Therefore, the CRPS values range between [0, +∞], and values closer to zero indicate better forecasts. The CRPS value is influenced by the bias in the ensemble mean and the ensemble spread. The bias in the ensemble mean illustrates a systematic error, and the too-small or too-large ensemble spread represents forecast uncertainty. These characteristics reflect the reliability and accuracy of the forecast, capturing how well the forecast represents both the average outcome and the spread of possible outcomes (Monhart et al., 2018). Analytical expression of the CRPS can be derived for many families of probability distributions (Jordan et al., 2019). In this study, the DRN model was trained using an analytical CRPS equation for the Gaussian distribution as a loss function with parameters $\mu_{i,t}$ and $\sigma_{i,t}$, and the CSG-DRN model was trained using an analytical CRPS equation for CSG distribution as a loss function with the parameters being constrained to $(\alpha_{i,t} = \mu^2_{i,t}/\sigma^2_{i,t}) > 0$, $(\beta_{i,t} = \sigma^2_{i,t}/\mu_{i,t}) > 0$ and $\delta_{i,t} < 0$ (see Eq. 10 in Scheuerer and Hamill (2015)).

3.3.2. Continuous ranked probability skill score (CRPSS)

The CRPSS has been adopted to evaluate the relative gain achieved by postprocessing methods over the raw ensemble forecasts across different lead times and variables in the test period. This metric provides a comprehensive measure of the effectiveness of postprocessing methods in improving forecast accuracy relative to the raw forecast (F^{raw}) and is defined as

$$\mathrm{CRPSS}(F, y) = 1 - \frac{\overline{\mathrm{CRPS}}(F, y)}{\overline{\mathrm{CRPS}}(F^{raw}, y)}, \tag{5}$$

where $\overline{\mathrm{CRPS}}(F, y)$ and $\overline{\mathrm{CRPS}}(F^{raw}, y)$ represent the mean CRPS values of the postprocessed forecast variable and the raw forecast variable, respectively, computed over the spatial dimension. We will additionally calculate the CRPSS based on the mean CRPS values for a single grid point over the temporal dimension for both the postprocessed and raw forecast variables to better highlight spatial aspects of the differences in predictive performance.

CRPSS values range from $-\infty$ to 1. A value of 0 indicates that the postprocessing method performs not better than the raw forecast, while values closer to 1 indicate greater improvement in forecast accuracy and negative values indicate a worse performance than the raw ensemble.

4. Results and discussions

4.1. Raw forecast skill

The accuracy and reliability of raw ensemble weather forecasts are assessed by computation of the CRPS at each grid point and for different lead times within the specified region. During the reforecast period (1981–2016), the temporal average CRPS for precipitation ranges from a minimum of 1.63 in the valley to a maximum of 3.94 in the mountains, with an overall spatial mean of 2.46, as shown in Fig. 4a. In contrast, the temporal average CRPS exhibits the opposite pattern for temperature, ranging from a maximum of 10.13 in the valley to a minimum of 1.56 on the mountain peaks, with an overall spatial mean of 3.57, as shown in Fig. 4b. The high CRPS values for both raw weather forecast variables reveal that the forecasts predicted probability distribution differs substantially from the given observed value.

Substantial biases in SEAS5 raw forecasts compared to various reference datasets have been reported in multiple studies. For example, Crespi et al. (2021b) analyzed the performance of seasonal weather forecasts over Europe, Ehsan et al. (2021) studied Ethiopia, Ratri et al. (2019a) explored Java, Indonesia, and Gubler et al. (2020) analyzed South America. These studies used different forecast verification metrics and reference observation datasets, but reveal the elevation-dependence of the biases and the presence of significant spatial variations in the biases for both precipitation and 2-m temperature over the respective study regions. Our results confirm these findings, showing similar spatial patterns of the discrepancies, particularly in relation to the presence of elevation-dependent biases. This spatial heterogeneity indicates that traditional statistical postprocessing methods are likely insufficient and motivates the use of more flexible, location-aware postprocessing approaches to improve forecast accuracy.

4.2. Postprocessed forecast skill

Figs. 5 and 6 present the boxplots of spatial averaged CRPS values during the test period (2011–2014) for precipitation and temperature

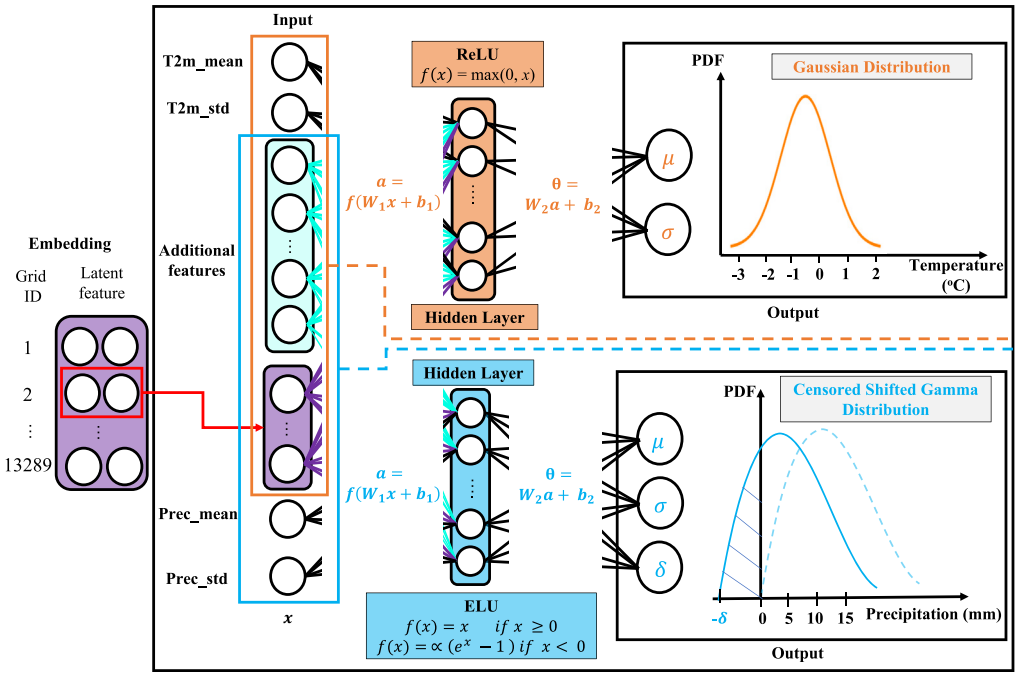


Fig. 3. Schematic illustration of the NN-based postprocessing models. The orange dashed part is the DRN architecture for 2-m temperature forecasting. The blue dashed part is the CSG-DRN architecture for precipitation. Purple colored nodes and connections are the grid point ID embedding, and cyan colored nodes are additional features. Orange and blue nodes are hidden layers for DRN and CSG-DRN architecture, respectively.

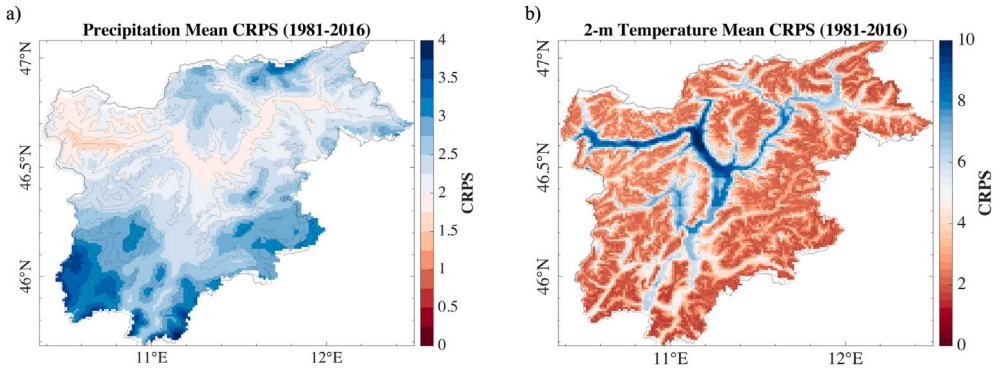


Fig. 4. Temporal mean CPRS values of the raw ensemble forecasts during the reforecast period 1981–2016: mean over all the lead times (a) for precipitation; and (b) for temperature.

forecasts, respectively, comparing the postprocessing methods E-QM, CSG-DRN, and DRN, with the raw ensemble forecasts. In particular, all these methods are evaluated across multiple lead times (from 1 to 31 days), highlighting differences in variability and central tendency.

Concerning precipitation (see Fig. 5), CSG-DRN achieves the best performance across most of the lead times, with lower median CRPS values, reduced variability, and fewer extreme values compared to both the raw forecasts and E-QM. Notably, it reaches a minimum mean CRPS value (white dot) of 1.33 at lead time three. Over the first seven days of lead time, CSG-DRN consistently outperforms E-QM, as indicated by lower mean CRPS values and narrower boxplots. Between 8 and 29 days lead time, the two postprocessing methods exhibit comparable performance, with similar mean CRPS values and variability, though on some days, CSG-DRN is preferable. While the boxplot spreads remain stable across most lead times, they substantially increase for both post-processing methods at the longest lead times tested (days 30 and 31). Both methods, however, outperform the raw simulations, particularly until day 7 lead time, with the latter generally showing higher CRPS values, greater variability, and more extreme outliers across all lead times.

Concerning temperature (see Fig. 6), more pronounced improvements over the raw forecasts and be observed, and the DRN method shows a clear advantage over E-QM, particularly during shorter lead times. It achieves a minimum averaged CRPS value (white dot) of 0.73 on day 1. In particular, over the first ten days of lead time, DRN consistently outperforms E-QM, as indicated by lower mean CRPS values and narrower boxplots. At longer lead times, the performance of DRN and E-QM becomes more comparable, with overlapping boxplots and similar medians for both methods. Across all lead times, both post-processing methods significantly outperform the raw forecast, which exhibits higher median CRPS values, larger variability, and a greater number of extreme values. Unlike precipitation, the performance of both postprocessing methods for temperature shows a clear decline with increasing lead times, as indicated by higher mean CRPS values. Notably, on day ten of lead time, there is a marked jump in the

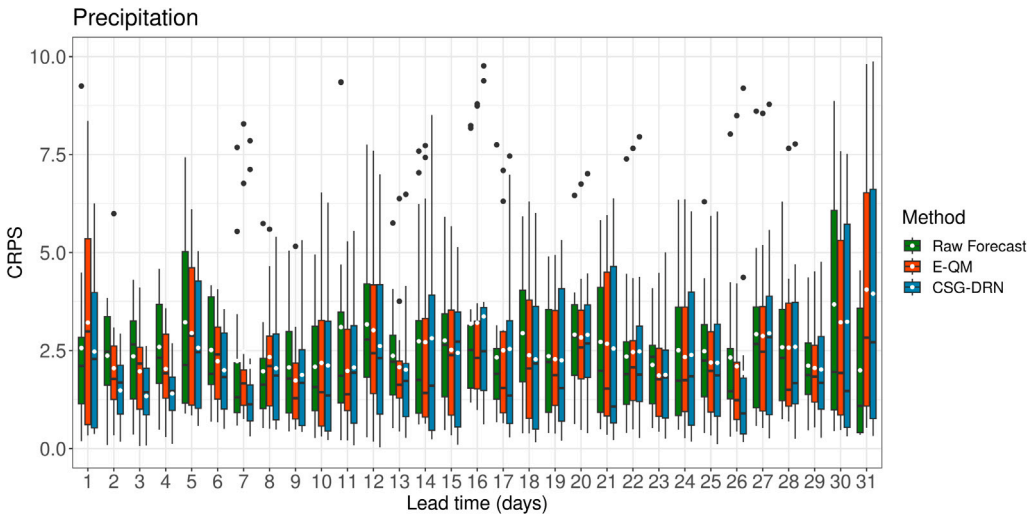


Fig. 5. Boxplot of spatial averaged CRPS values at each day of lead time for raw precipitation forecast (in green), E-QM (in orange), and CSG-DRN (in blue). The white dot and the horizontal bar within the boxplot represent the mean and median values for each box, respectively.

reduction of performance for both methods, highlighting a critical transition point in forecast skill.

To gain deeper insights into the relative performance of the post-processing methods, Figs. 7 and 8 show the CRPSS values for both precipitation and temperature, respectively. For precipitation, CSG-DRN exhibits the highest median values, particularly for short lead times (e.g., days 1–7), indicating greater improvement over the raw forecast compared to E-QM. As lead times increase, the median CRPSS values for both methods gradually decline, reflecting the typical decrease in forecast skills associated with long prediction horizons with respect to the raw forecast (Buizza and Leutbecher, 2015; Ratri et al., 2019b). However, CSG-DRN consistently exhibits higher CPRSS values than E-QM, even at longer lead times (e.g., days 20–31).

For temperature, DRN exhibits higher CRPSS values than E-QM during short lead times (e.g., days 1–7), where it also exhibits narrower boxplots This is in line with the minimum average CRPS values for DRN, as shown in Fig. 6. As lead times increase (e.g., beyond day 10), the magnitude and range of CRPSS values for both E-QM and DRN increase in a similar manner. After 17 days, E-QM exhibits narrower boxplots, indicating that its performance is more stable over longer lead times and highlighting a trade-off between consistency and overall performance. This suggests that while DRN maintains higher median CRPSS values than E-QM, the performance varies with longer lead times. Overall, DRN achieves higher median CRPSS values during most lead times, underlining its generally superior skill in improving temperature forecast accuracy with respect to the raw forecast.

The ability of CSG-DRN and DRN to capture more complex relationships between predictors and forecast variables and to incorporate location-specific environmental variable information highlights their potential to improve the forecast skill for precipitation and temperature in mountainous regions, such as the Trentino-South Tyrol region. A key insight is the significant short lead-time gain shown by both NN-based methods, which is essential in applications where weekly or biweekly forecast accuracy is important, such as in water management issues in small alpine catchments, including, e.g., hydropower operations (Anghileri et al., 2019; Dasgupta et al., 2023). The inclusion of location-specific environmental variables and non-linear dependencies, coupled with the flexibility of NNs, helps to explain the better performance of these methods compared to E-QM. However, the gradual convergence in performance between the methods as lead times increase indicates limitations in addressing predictive uncertainty for lead times larger than 10–15 days, which challenges both NN-based and traditional statistical postprocessing approaches (Hou et al., 2022; Ghazvinian et al., 2021).

The decreasing accuracy with increasing lead times suggests that while CSG-DRN and DRN are effective for short to intermediate lead-time day improvements, they may need further advancement to improve forecast accuracy above about 10–15 days. This highlights the need for continuous exploration of multivariate (Lakatos et al., 2023), hybrid (Reichstein et al., 2019), or more flexible postprocessing techniques to address the inherent challenges of weather forecasting in complex regions.

4.3. Elevation dependence of postprocessed forecast skill

Postprocessing methods aim to reduce systematic biases, which are often strongly affected by topographic variations, particularly in complex terrain (Velasquez et al., 2020; J.J. et al., 2018). To provide a more detailed understanding of the performance of the investigated postprocessing methods across different ranges of elevation differences, Figs. 9 and 10 present boxplots of temporally averaged CRPSS for precipitation and temperature forecasts, respectively, over the 2011–2014 test period. The elevation difference values (i.e., the difference between the elevation of the observational gridded dataset and SEAS5 model elevation at the same location) range from −1000 m to 1000 m, where a negative elevation difference indicates that the elevation value represented in the model is higher than the observed one (occurring in the valleys), and a positive elevation difference indicates that the elevation values represented in the model are lower than observed (at high elevations).

Concerning precipitation, the boxplots shown in Fig. 9 reveal that the CSG-DRN method consistently outperforms E-QM across all elevation difference ranges, demonstrating the CSG-DRN's ability to better capture key features influencing precipitation forecasts. This is in line with the results presented in Fig. 7, where the CSG-DRN model showed consistently higher CRPSS computed based on spatially averaged CRPS values. On the other hand, Fig. 10 reveals that for temperature forecasts, both DRN (blue) and E-QM (orange) perform similarly when the elevation difference is negative (i.e., in valley regions). For positive elevation difference ranges (i.e., at high elevations), DRN exhibits a minor advantage over E-QM, thus suggesting that DRN is more effective in capturing topographic effects, leading to better performance in complex terrain.

The analysis of postprocessing performance across different elevation ranges provides valuable insights into the robustness of the CSG-DRN and DRN methods in complex topographic regions. The superior performance of CSG-DRN over E-QM for precipitation forecasts across all elevation different ranges can be attributed to the ability

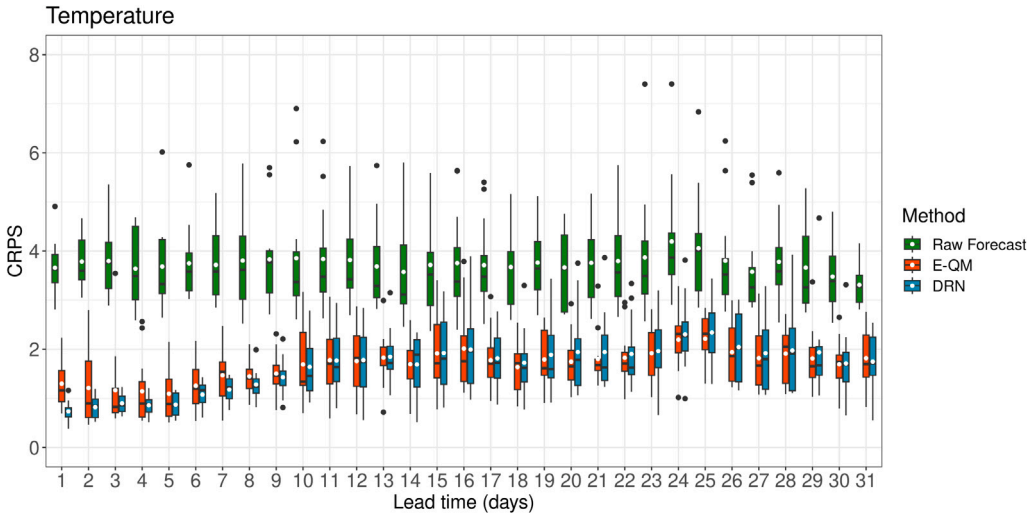


Fig. 6. Boxplot of spatial averaged CRPS values at each day of lead time for raw 2-m temperature forecast (in green), E-QM (in orange), and DRN (in blue). The white dot and the horizontal bar within the boxplot represent the mean and median values for each box, respectively.

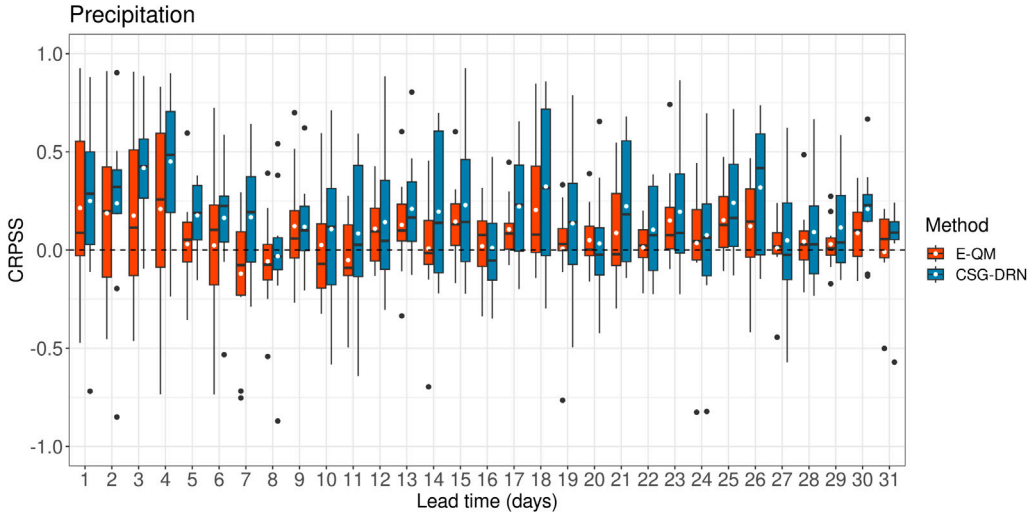


Fig. 7. Boxplot of spatial averaged CRPSS values at each day of lead time for precipitation forecast for both E-QM (in orange) and DRN (in blue). The white dot and the horizontal bar within the boxplot represent the mean and median values for each box, respectively.

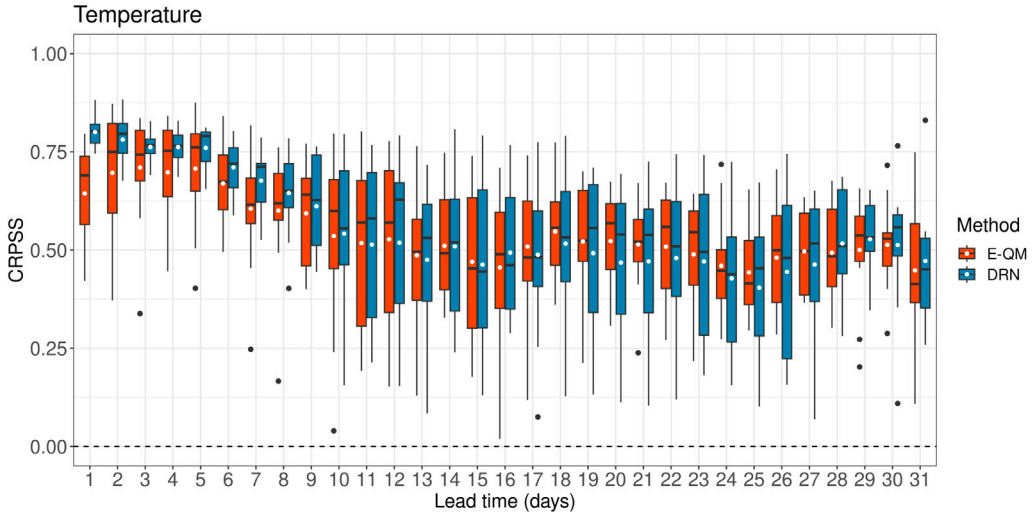


Fig. 8. Boxplot of spatial averaged CRPSS values at each day of lead time for 2-m temperature forecast for both E-QM (in orange) and DRN (in blue). The white dot and the horizontal bar within the boxplot represent the mean and median values for each box, respectively.

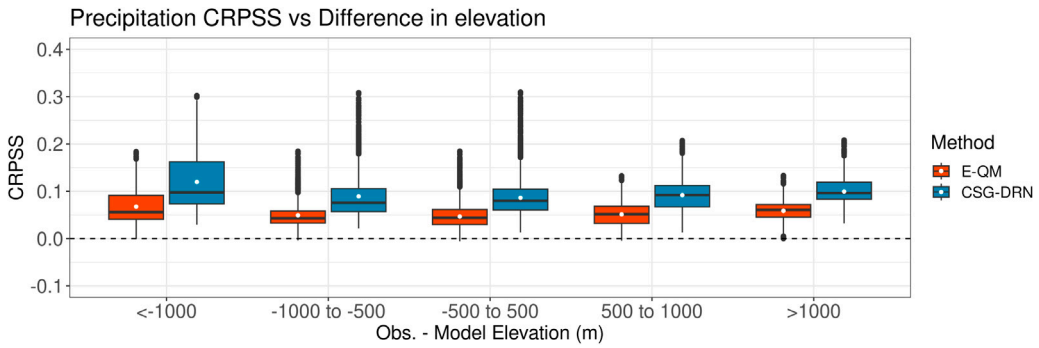


Fig. 9. Boxplots of temporally averaged CRPSS values for precipitation aggregated over grid points with different elevation difference ranges.

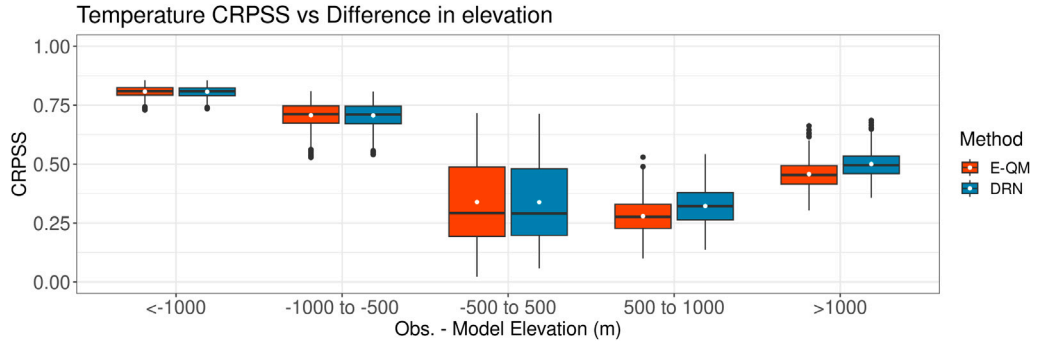


Fig. 10. Boxplots of temporally averaged CRPSS values for 2-m temperature aggregated over grid points with different elevation difference ranges.

of NNs to capture non-linear relationships between topography and forecast variables. In mountainous regions, where weather patterns are influenced by localized topographic effects such as orographic precipitation (Napoli et al., 2019), the flexibility of CSG-DRN in incorporating additional predictors and spatial information indeed enhances its ability to correct systematic forecast biases.

The results also suggest a more subtle effect of elevation dependence on temperature forecasts. In valley regions (i.e., negative elevation differences), both DRN and E-QM perform similarly, which can be attributed to the presence of a more stable environmental lapse rate, making statistical methods like E-QM sufficient to correct forecast biases. However, as elevation increases (i.e., positive elevation differences), DRN outperforms E-QM, indicating that in higher elevations, where temperature is more influenced by elevation, slope orientation, and thermal inversions (Mahrt, 2006; Dimri et al., 2022), the DRN provides a better framework for capturing these complex interactions.

The variation in performance at higher elevations may also be influenced by E-QM's inherent limitations, which apply the same transformation function across grid points without fully accounting for local geographical features (Velasquez et al., 2020). In contrast, DRN's inclusion of location-specific environmental variables and additional ensemble predictors (see Table 1) allows it to adjust forecasts to reflect the unique climatological dynamics at higher elevations, such as the cooling effect of elevation and the increased variability of temperature and precipitation in alpine regions (Pepin et al., 2022).

These results overall indicate that NN-based methods can be particularly valuable in regions where topographic complexity substantially affects weather forecast variables and that these methods can model the complex relationships between topography and atmospheric variables more effectively than traditional postprocessing methods like E-QM.

4.4. Spatial characteristics of postprocessed forecast skill

To provide a more explicit perspective on spatial aspects of the improvements achieved by the NN-based approach, we calculated the CRPSS of the NN-based techniques relative to the E-QM method for each forecast variable (i.e., the denominator in Eq. (5) refers to CRPS of E-QM instead of raw forecast). In addition, the CRPSS is calculated using the CRPS values averaged over 10-day intervals for each grid point. The results are further averaged temporally on annual and seasonal scales. The spatial distribution of CRPSS for precipitation and temperature forecasts, as depicted in Figs. 11 and 12 highlights, in particular, the locations where NN-based techniques achieve substantial improvements with respect to the traditional E-QM, providing a comprehensive overview of their performance across different seasons, lead times, and climatic conditions. Positive CRPSS values here show a superior forecast accuracy of the NN-based approach with respect to E-QM, whereas negative values indicate the contrary.

For precipitation (see Fig. 11), the CSG-DRN method demonstrates substantial improvements with respect to E-QM over almost all the grid points for the annual average, the winter (DJF) and spring (MAM) seasons in the first ten lead-time days. The improvements are less evident in summer (JJA) and autumn (SON). As lead times extend from day ten to day twenty (see Fig. 11 middle row), the CRPSS for the annual, DJF, and SON seasons indicates that E-QM is preferable to CSG-DRN at most of the grid points, except for the northern and northwestern valleys. During the spring and summer seasons, CSG-DRN performs slightly better than E-QM at these intermediate lead time scales. Starting from day twenty lead time onward, the performance of CSG-DRN and E-QM becomes broadly similar, with each method occasionally outperforming the other. This convergence indicates that while CSG-DRN provides better forecasts initially, its advantage decreases as lead times increase, resulting in both methods delivering comparable performance over longer lead times. As expected, forecast skill is substantially influenced by seasonality, lead time, and regional factors, in line with the findings of Ehsan et al. (2021), which demonstrated that seasonality and lead time play a dominant role in forecast accuracy. The performance of CSG-DRN in complex terrain, especially in northern and northwestern valleys (see Fig. 11), indicates that NN-based methods are performing better in regions where traditional methods face challenges. This is primarily due to the complex interactions between surface characteristics

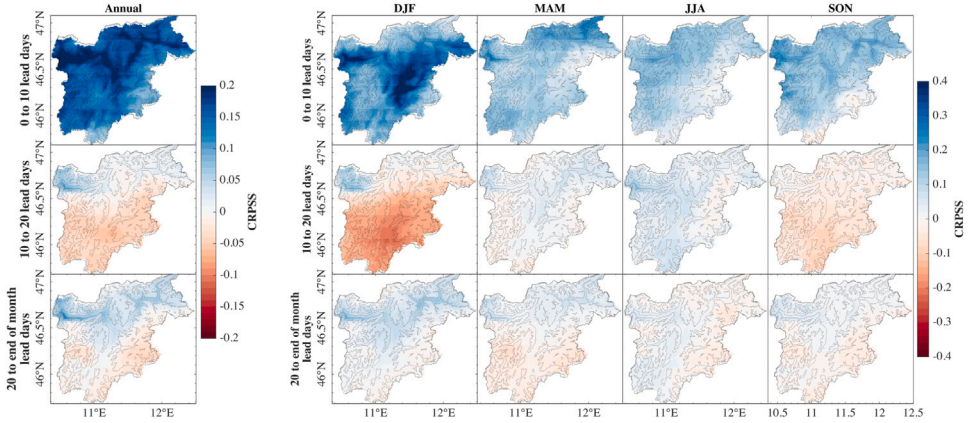


Fig. 11. Average CRPSS of CSG-DRN relative to E-QM for precipitation forecasts, calculated at annual and seasonal scales with ten-day intervals.

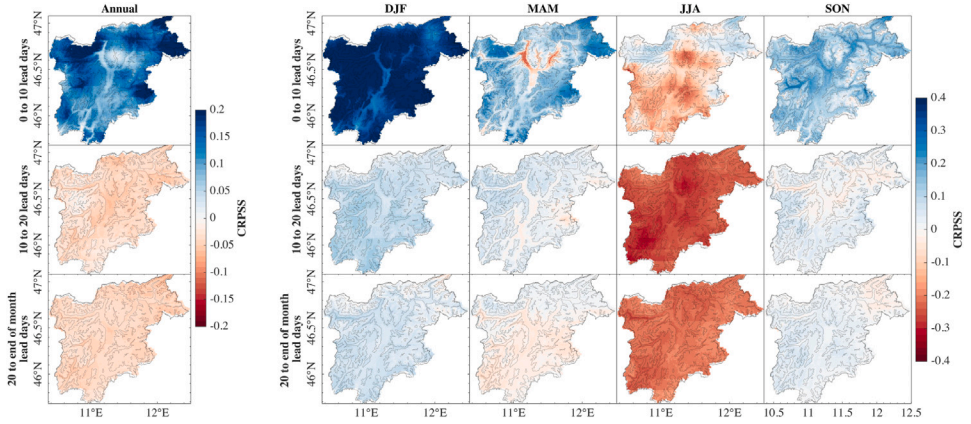


Fig. 12. Average CRPSS of DRN relative to E-QM for 2-m temperature forecasts, calculated at annual and seasonal scales with ten-day intervals.

and atmospheric conditions, interactions that are captured by the NN-based techniques, thus allowing for a substantial enhancement of the weather forecasts in regions with complex topography (Rasp and Lerch, 2018; Schulz and Lerch, 2022).

For temperature (see Fig. 12), DRN demonstrates substantial skill improvements with respect to E-QM over most grid points for all three lead time intervals during all seasons, with the exception of summer. In the first ten days of lead time (see Fig. 12 top row), DRN substantially outperforms E-QM during the winter season, especially at high elevation regions. E-QM performs slightly better or similarly to DRN only in the valley regions during spring. In contrast, during the summer season, E-QM generally outperforms DRN, except in certain northern or northwestern valleys where DRN maintains an advantage. Overall, the DRN is performing much better in the first ten days of lead time. The performance of NN-based methods in complex terrain, especially in northern and northwestern valleys (see Figs. 11 and 12), indicates that NN-based methods are better performing in regions where traditional methods face challenges. This is primarily due to the complex interactions between surface characteristics and atmospheric conditions. As the lead time extends beyond day ten, E-QM slightly surpasses the performance of DRN for the annual average and summer season forecasts, while DRN maintains a slight edge for the other seasons. The same behavior is confirmed during lead times between 20 and 30 (bottom row in Fig. 12). This trend suggests that DRN's strength is most evident for short lead times, but its performance may gradually align with or is surpassed by E-QM in some seasons as the lead times increase. Overall, both methods offer comparable skill levels at extended lead times, with fluctuations in performance depending on the region, season, and lead time. Finally, the performance comparison of DRN and E-QM for temperature forecasts suggests that while NNs perform better in general, they may not consistently outperform traditional methods like E-QM, especially over longer lead times. This highlights the need for adaptive methods to postprocessing, where the strengths of different methods are leveraged depending on the forecast lead time, season, or regional factors (Chen et al., 2024; Wessel et al., 2024).

In summary, the CSG-DRN and DRN methods significantly enhance short lead-time forecasts for both precipitation and temperature, especially at high elevations and during the winter season. However, as lead times increase, the performance of both methods tends to be comparable, with the E-QM method providing better performance for precipitation during 10 to 20 lead-time days in the winter (see Fig. 11) and for temperature during summer across all lead times (see Fig. 12). Our results also indicate that the effectiveness of the CSG-DRN and DRN methods varies notably with the seasons. These methods perform better in the winter season (DJF) and at high elevations, likely because of their ability to capture complex interactions between atmospheric processes and topography. In contrast, the less evident improvements in the JJA season can be attributed to increased atmospheric variability, which is challenging to reproduce accurately. An important aspect to consider is how sensitive these methods are to lead time. The better performance of NN-based models during the first ten days of lead time is consistent with their ability to capture short-term atmospheric dynamics (see Figs. 11 and 12). However, as the lead time increases, their performance decreases. This performance decrease can be attributed to growing uncertainty in the forecast data and the limitations of NNs in inferring complex relationships beyond short-term patterns.

Figs. 13 and 14 further present boxplots of spatially averaged CRPSS for each season for precipitation and temperature, respectively. The

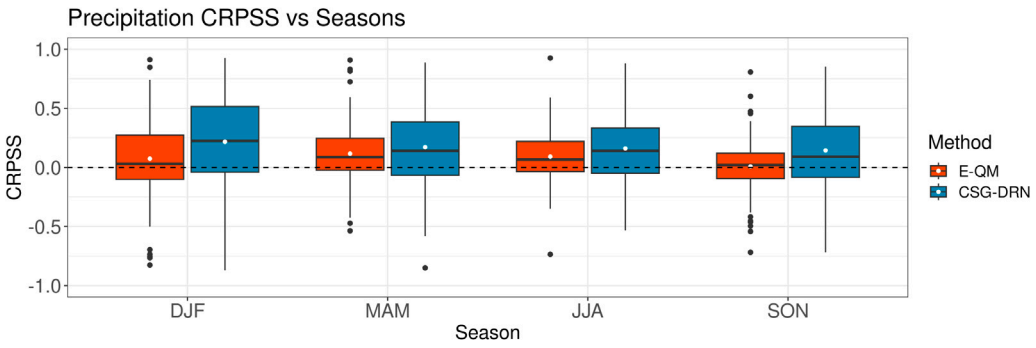


Fig. 13. Boxplot of spatial mean CRPSS for CSG-DRN and E-QM relative to raw precipitation forecasts across different seasons.

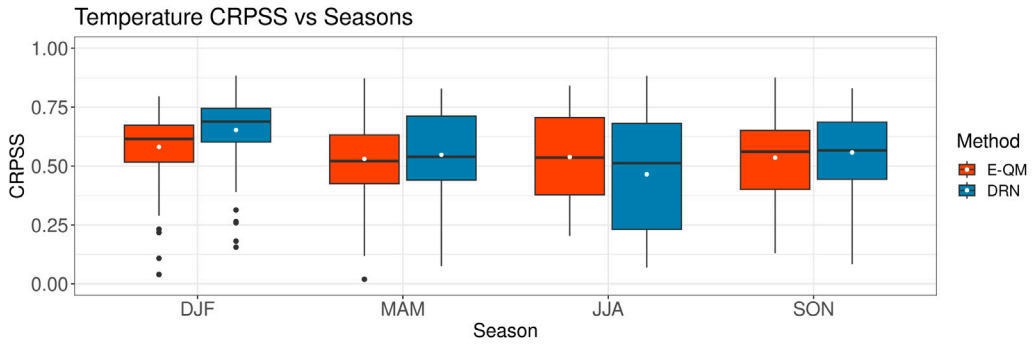


Fig. 14. Boxplot of spatial mean CRPSS for DRN and E-QM relative to raw 2-m temperature forecasts across different seasons.

CRPSS of the NN-based approaches and E-QM is here computed with respect to the raw ensemble forecast (see Eq. (5)) over the test period (2011–2014). In the case of precipitation, Fig. 13 highlights that CSG-DRN (in blue) consistently performs better than E-QM (in orange) in improving forecast skill across all seasons. For temperature, Fig. 14 reveals that DRN (in blue) performs better than E-QM except in the summer, in line with results presented in Fig. 12. This suggests that on a spatial average, the NN-based approach is more effective at correcting systematic biases and improving accuracy regardless of seasonal variability, with the only exception of temperature during the summer season.

4.5. Feature importance

To understand the importance of the input features utilized in the CSG-DRN and DRN methods, we performed a permutation-based feature importance analysis, as first proposed by Breiman (2001). Importance scores calculated for individual input features thereby reflect the relative importance of that each feature in shaping the outputs of the NN.

We follow the approach proposed by Rasp and Lerch (2018) and compute feature's importance in influencing the predictive performance of the NN-based models. The values of a given feature are randomly shuffled across the grid points and time steps while keeping all other features fixed. This disrupts the relationship between the shuffled feature and the target variable, allowing us to measure the contribution of the shuffled feature in predictive performance. The importance of each input feature is assessed by comparing the mean CRPS of the model for the shuffled features against the non-shuffled features in the test period.

Fig. 15a clearly demonstrates that convective precipitation (cp) is an essential feature after precipitation (prec) itself in the CSG-DRN method. This highlights the crucial role of convective processes in shaping precipitation patterns, specifically in complex terrain (Houze, 2012). Other important features include the vertical wind component at the 100 hPa pressure level (v100), which captures the northward motion of atmospheric winds; the latent and sensible heat fluxes between the surface and atmosphere (sshf and slhf); and the surface net solar radiation (ssr). Each of these features provides essential insights into atmospheric dynamics. In addition to these physical atmospheric features, spatiotemporal information, represented by lead times and embeddings, is also found to be important. These features allow capturing the transition of weather patterns over time and across different grid points, substantially improving the forecast accuracy. However, features with near-zero importance scores have an insignificant effect on the performance of NN-based techniques. This suggests that these features could potentially be removed in future applications to reduce computational costs without compromising forecast accuracy.

The feature importance analysis applied to the DRN technique for temperature forecast is shown in Fig. 15b. The plot reveals the important role of the bias in elevation (elv-geo), which is recognized as a feature that influences temperature forecasts in complex terrain. The importance of the 2-m dewpoint temperature (d2m) is also highlighted, as it reflects the moisture content in the air, which can directly influence temperature. Additionally, sensible and latent heat fluxes (sshf and slhf) contribute to the DRN's accuracy by explaining the energy exchanges between the Earth's surface and atmosphere. The cosine-transformed days of the year (cos-doy) and embeddings are important features that are crucial to capturing seasonal cycles and location-specific information, respectively. Furthermore, for the DRN technique, Atlantic Oscillation (AO) also emerges as an important feature that can be attributed to influencing temporal weather patterns and temperature variability. However, in the operational forecast postprocessing context, neither the AO nor the NAO indices are available and can therefore be excluded, especially given their relatively low importance compared to other features.

Overall, this feature importance analysis highlights the importance of using suitable features that help effectively capture complex atmospheric interactions, topographical influences, and temporal dynamics by integrating insights from precipitation and temperature weather

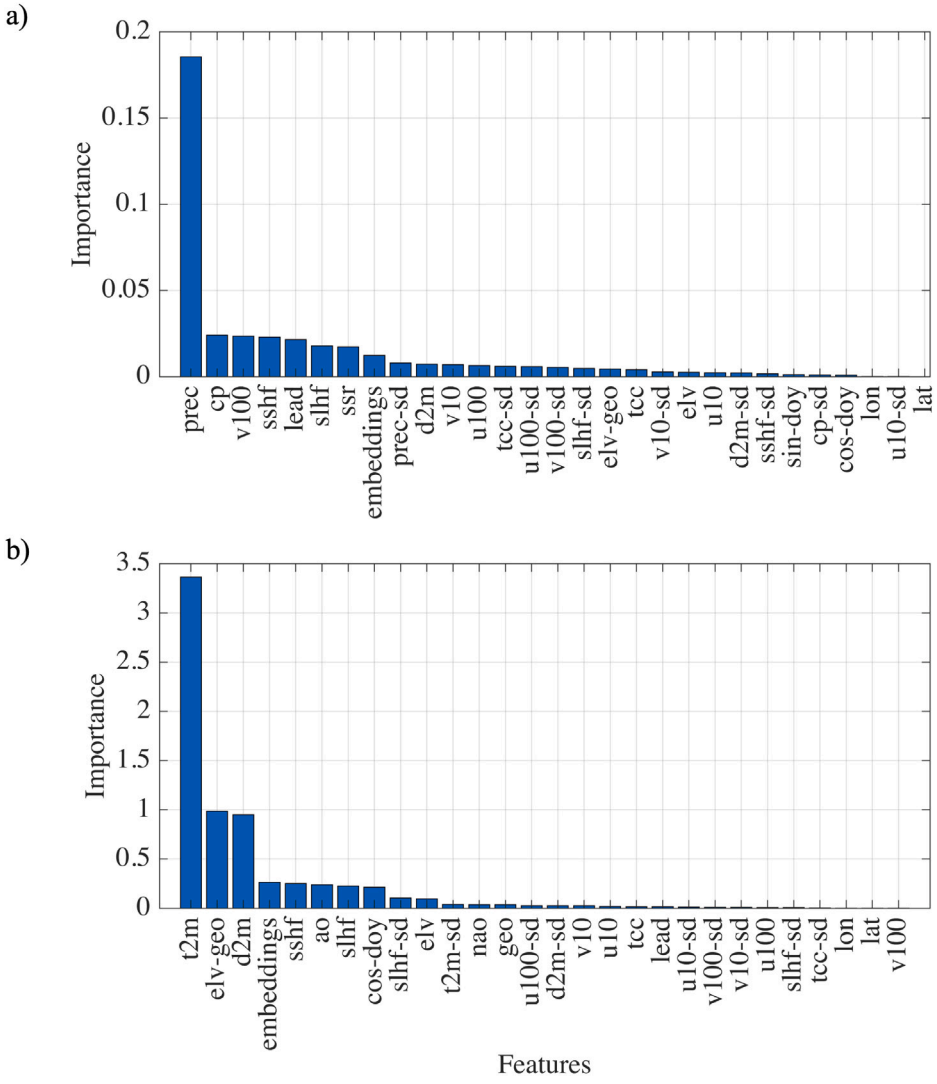


Fig. 15. Feature importance plot of (a) the CSG-DRN method for precipitation (b) the DRN method for 2-m temperature.

forecast postprocessing methods. For both target variables, the inclusion of uncertainty information from the ensemble forecasts is only of negligible importance, in line with corresponding results from other studies (Rasp and Lerch, 2018; Schulz and Lerch, 2022; Höhlein et al., 2024). In a possible extension of this work, we will consider whether to use only the most important input features to train the NN model while obtaining comparable forecast accuracy. In particular, this reduction in the input features can reduce the model complexity and computational demand, making it more suitable for real-time forecasting contexts.

5. Conclusion

This study presents a novel approach for improving the accuracy and reliability of seasonal weather forecasts by addressing the inherent limitations of traditional postprocessing methods. We aim to overcome constraints associated with the E-QM method, which struggles to correct values that are outside the training dataset. Additionally, it is challenging for E-QM to include additional environmental variables to account for linear or non-linear relationships, as it only adjusts the marginal distribution of the forecast variable at each grid point and lead time. To address these challenges, we implemented an NN-based distributional regression method. The study was focused on the Trentino-South Tyrol region in the northeastern Italian Alps, a region with complex topography where accurate forecasting is particularly challenging.

The NN-based method and E-QM demonstrated clear improvements in forecast skill compared to the raw ensemble forecasts; however, the NN-based method performed better than E-QM, particularly in shorter lead times for precipitation and temperature variables. The CSG-DRN for precipitation consistently outperforms E-QM across all ranges of elevation differences and over the first ten days of lead time, indicating its ability to capture non-linear relationships. Similarly, DRN outperformed E-QM for temperature forecasts, particularly at shorter lead times (within the first 10 days) and at positive elevation differences. However, as the lead time extended beyond 10 days, the performance of NN-based methods and E-QM became comparable, with E-QM occasionally outperforming in certain seasons and regions. Furthermore, the NN-based methods showed substantial improvements in forecast skills based on seasons, with superior performance in winter, but E-QM outperforming DRN in summer. Indeed, integrating multiple meteorological features, such as convective precipitation, vertical wind components, heat and energy fluxes, and solar radiation, contributed to the ability of the NN-based method to capture the complex dynamics of atmospheric conditions in this complex topography region.

Feature importance analysis also revealed that key meteorological variables, such as convective precipitation, wind components, surface fluxes, and elevation, provide relevant information to improve the accuracy of the NN-based methods. Furthermore, lead times and embeddings also played a significant role in improving forecast accuracy.

In conclusion, this study demonstrates the potential of NN-based postprocessing to significantly improve seasonal weather forecasts, particularly in complex terrain regions where traditional methods like E-QM struggle. By leveraging the non-linear relationships between meteorological variables and incorporating additional features, the NN-based model offers a more dynamic and flexible solution. Future work could explore combining the strengths of both methods for further improvement, especially at longer lead times. Overall, this approach represents a promising advancement in postprocessing techniques, with the potential for broader application in improving forecast accuracy across various regions and seasons.

Evidently, there are multiple promising opportunities for extensions and future work. One limitation of the NN-based methods considered here is that they process each grid point individually without taking spatial relationships explicitly into account. Over the past years, various ML-based spatial postprocessing methods building on advances in convolutional neural networks have been proposed (e.g., Scheuerer et al., 2020; Veldkamp et al., 2021; Chapman et al., 2022; Lerch and Polsterer, 2022; Li et al., 2022; Hu et al., 2023; Horat and Lerch, 2024). Those might enable further improvements by leveraging spatial structures in the input features, but the large dimension of the study area might require adaptations such as meteorologically-informed subdivisions into smaller regions. Further, our study only considered univariate aspects of forecast performance. However, many practical applications require accurate models of spatial, temporal, and inter-variable dependencies. A variety of multivariate postprocessing methods has been proposed over the past years, see Lerch et al. (2020) and Lakatos et al. (2023) for comparisons, including modern ML-based approaches (Chen et al., 2024), the adaptation of which to the setting of our study would be of interest. Connecting these advances to the downstream applications discussed in the introduction – particularly those related to hydrological modeling, water resource management, and the operation of hydraulic infrastructure in complex terrain – could further enhance the value and impact of seasonal forecast systems (Avesani et al., 2021, 2022).

CRediT authorship contribution statement

Sameer Balaji Uttarwar: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sebastian Lerch:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Diego Avesani:** Writing – review & editing, Writing – original draft, Validation, Supervision. **Bruno Majone:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

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

During the preparation of this work the author(s) used Grammarly and QuillBot in order to improve language and readability of the original draft, with caution. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

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

Acknowledgments

This research has been funded by the European Union - NextGenerationEU program, under the PRIN 2022 PNRR project "SAHARA - StorAge enHAnced droughts management for Resilient river bAsins" (Prot. no. P20227NPLW, CUP E53D23021860001). It acknowledges the Italian Ministry of Education, Universities and Research (MUR), in the framework of the project DICAM-EXC Departments of Excellence 2023–2027, grant L232/2016. Diego Avesani acknowledges support from the European Union – FSE-REACT-EU, PON Research and Innovation 2014–2020 DM1062/2021. Bruno Majone also acknowledges support from "iNEST (Interconnected Nord-Est Innovation Ecosystem)" project funded by the European Union under NextGenerationEU (PNRR, Mission 4.2, Investment 1.5, Project ID: ECS 00000043). Sebastian Lerch gratefully acknowledges support by the Vector Stiftung through the Young Investigator Group "Artificial Intelligence for Probabilistic Weather Forecasting".

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.advwatres.2025.105061.

Data availability

The code used for the analysis is available on GitHub repository https://github.com/Sam-Uttarwar/NN_arch.git. We acknowledge that the observational dataset used is available at https://doi.pangaea.de/10.1594/PANGAEA.924502. The ECMWF seasonal forecast dataset is downloaded from the operational archives of the MARS catalog https://www.ecmwf.int/en/forecasts/dataset/operational-archive.

References

Alexander, S., Block, P., 2022. Integration of seasonal precipitation forecast information into local-level agricultural decision-making using an agent-based model to support community adaptation. Clim. Risk Manag. 36, 100417. http://dx.doi.org/10.1016/j.crm.2022.100417.

Anghileri, D., Monhart, S., Zhou, C., Bogner, K., Castelletti, A., Burlando, P., Zappa, M., 2019. The value of subseasonal hydrometeorological forecasts to hydropower operations: How much does preprocessing matter? Water Resour. Res. 55, 10159–10178. http://dx.doi.org/10.1029/2019WR025280, URL: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2019WR025280.

Arnoux, M., Halloran, L.J., Berdat, E., Hunkeler, D., 2020. Characterizing seasonal groundwater storage in alpine catchments using time-lapse gravimetry, water stable isotopes and water balance methods. Hydrol. Process. 34, 4319–4333. http://dx.doi.org/10.1002/hyp.13884.

Avesani, D., Galletti, A., Piccolroaz, S., Bellin, A., Majone, B., 2021. A dual-layer mpi continuous large-scale hydrological model including human systems. Environ. Model. Softw. 139, 105003. http://dx.doi.org/10.1016/j.envsoft.2021.105003, URL: https://www.sciencedirect.com/science/article/pii/S1364815221000463.

Avesani, D., Zanfei, A., Di Marco, A., Ravazzolo, F., Righetti, M., Majone, B., 2022. Short-term hydropower optimization driven by innovative time-adapting econometric model. Appl. Energy 310, 118510. http://dx.doi.org/10.1016/j.apenergy.2021.118510, URL: https://www.sciencedirect.com/science/article/pii/S0306261921017244.

Baran, S., Nemoda, D., 2016. Censored and shifted gamma distribution based EMOS model for probabilistic quantitative. pp. 280–292. http://dx.doi.org/10.1002/env.2391.

Bertoldi, G., Bozzoli, M., Crespi, A., Matiu, M., Giovannini, L., Zardi, D., Majone, B., et al., 2023. Diverging snowfall trends across months and elevation in the northeastern italian alps. Int. J. Climatol. 2023, 2794–2819.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Bremnes, J.B., 2020. Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. Mon. Weather Rev. 148, 403–414. http://dx.doi.org/10.1175/MWR-D-19-0227.1.

Brotzge, J.A., Berchoff, D., Carlis, D.L., Carr, F.H., Carr, R.H., Gerth, J.J., Gross, B.D., Hamill, T.M., Haupt, S.E., Jacobs, N., et al., 2023. Challenges and opportunities in numerical weather prediction. Bull. Am. Meteorol. Soc. 104, E698–E705.

Brouwer, S., Rayner, T., Huitema, D., 2013. Mainstreaming climate policy: The case of climate adaptation and the implementation of eu water policy. Environ. Plan. C Gov. Policy 31, http://dx.doi.org/10.1068/c11134.

Buizza, R., Leutbecher, M., 2015. The forecast skill horizon. Q. J. R. Meteorol. Soc. 141, 3366–3382. http://dx.doi.org/10.1002/qj.2619.

Cannon, A.J., Sobie, S.R., Murdock, T.Q., 2015. Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? J. Clim. 28, 6938–6959. http://dx.doi.org/10.1175/JCLI-D-14-00754.1.

Chapman, W.E., Monache, L.D., Alessandrini, S., Subramanian, A.C., Ralph, F.M., Xie, S.P., Lerch, S., Hayatbini, N., 2022. Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. Mon. Weather Rev. 150, 215–234. http://dx.doi.org/10.1175/MWR-D-21-0106.1.

Chen, J., Janke, T., Steinke, F., Lerch, S., 2024. Generative machine learning methods for multivariate ensemble postprocessing. Ann. Appl. Stat. 18, 159–183. http://dx.doi.org/10.1214/23-AOAS1784, URL: http://arxiv.org/abs/2211.01345, arXiv:2211.01345.

Crespi, A., Matiu, M., Bertoldi, G., Petitta, M., Zebisch, M., 2021a. A high-resolution gridded dataset of daily temperature and precipitation records (1980–2018) for Trentino-South tyrol (north-eastern Italian alps). Earth Syst. Sci. Data 13, 2801–2818. http://dx.doi.org/10.5194/essd-13-2801-2021.

Crespi, A., Petitta, M., Marson, P., Viel, C., Grigis, L., 2021b. Verification and bias adjustment of ecmwf seas5 seasonal forecasts over europe for climate service applications. Climate 9, 1–17. http://dx.doi.org/10.3390/cli9120181.

Crochemore, L., Ramos, M.H., Pappenberger, F., 2016. Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. Hydrol. Earth Syst. Sci. 20, 3601–3618. http://dx.doi.org/10.5194/hess-20-3601-2016.

Dasgupta, A., Arnal, L., Emerton, R., Harrigan, S., Matthews, G., Muhammad, A., O'Regan, K., Pérez-Ciria, E., van Osnabrugge, B., Werner, M., Buontempo, C., Cloke, H., Pappenberger, F., Pechlivanidis, I.G., Prudhomme, C., Ramos, M.H., Salamon, P., 2023. Connecting hydrological modelling and forecasting from global to local scales: Perspectives from an international joint virtual workshop. J. Flood Risk Manag. 1–44. http://dx.doi.org/10.1111/jfr3.12880.

Demaeyer, J., Bhend, J., Lerch, S., Primo, C., Van Schaeybroeck, B., Atencia, A., Ben Bouallègue, Z., Chen, J., Dabernig, M., Evans, G., Faganeli Pucer, J., Hooper, B., Horat, N., Jobst, D., Merše, P., Möller, A., Mestre, O., Taillardat, M., Vannitsem, S., 2023. The euppbench postprocessing benchmark dataset v1.0. Earth Syst. Sci. Data 15, 2635–2653. http://dx.doi.org/10.5194/essd-15-2635-2023.

Dimri, A.P., Palazzi, E., Daloz, A.S., 2022. Elevation dependent precipitation and temperature changes over Indian Himalayan region. Clim. Dyn. 59, 1–21. http://dx.doi.org/10.1007/s00382-021-06113-z.

Ehsan, M.A., Tippett, M.K., Robertson, A.W., Almazroui, M., Ismail, M., Dinku, T., Acharya, N., Siebert, A., Ahmed, J.S., Teshome, A., 2021. Seasonal predictability of ethiopian kiremt rainfall and forecast skill of ECMWF's SEAS5 model. Clim. Dyn. 57, 3075–3091. http://dx.doi.org/10.1007/s00382-021-05855-0.

Emerton, R., Zsoter, E., Arnal, L., Cloke, H.L., Muraro, D., Prudhomme, C., Stephens, E.M., Salamon, P., Pappenberger, F., 2018. Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS-Seasonal v1.0. Geosci. Model. Dev. 11, 3327–3346. http://dx.doi.org/10.5194/gmd-11-3327-2018.

Feik, M., Lerch, S., Stühmer, J., 2024. Graph neural networks and spatial information learning for post-processing ensemble weather forecasts. arXiv preprint arXiv:2407.11050, arXiv:2407.11050.

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

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

Gobiet, A., Kotlarski, S., Beniston, M., Heinrich, G., Rajczak, J., Stoffel, M., 2014. 21St century climate change in the European Alps-A review. Sci. Total Environ. 493, 1138–1151. http://dx.doi.org/10.1016/j.scitotenv.2013.07.050.

Golian, S., Murphy, C., 2022. Evaluating bias-correction methods for seasonal dynamical precipitation forecasts. J. Hydrometeorol. 23, 1350–1363. http://dx.doi.org/10.1175/jhm-d-22-0049.1.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, http://www.deeplearningbook.org.

Gubler, S., Sedlmeier, K., Bhend, J., Avalos, G., Coelho, C.A., Escajadillo, Y., Jacques-Coper, M., Martinez, R., Schwierz, C., de Skansi, M., Spirig, C., 2020. Assessment of ECMWF SEAS5 seasonal forecast performance over south america. Weather. Forecast. 35, 561–584. http://dx.doi.org/10.1175/WAF-D-19-0106.1.

Guo, C., Berkhahn, F., 2016. Entity embeddings of categorical variables. pp. 1–9, URL: http://arxiv.org/abs/1604.06737, arXiv:1604.06737.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., Haiden, T., 2014. Trends in the predictive performance of raw ensemble weather forecasts. Geophys. Res. Lett. 41, 9197–9205. http://dx.doi.org/10.1002/2014GL062472, URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GL062472.

Hohenwallner, D., Saulnier, G.M., Anton, B., Bertoncelj, I., Brenčič, M., Bruno, M.C., Cadoux-Rivollet, M., Calvi, C., Carolli, M., Castaings, W., Chenut, J., Bona, A., Defrancesco, C., Doering, M., Dutto, E., Freundl, G., Harum, T., Jamsek, A., Zolezzi, G., 2011. Water resources management and water scarcity in the alps: Recommendations for water resources managers and policy-makers.

Höhlein, K., Schulz, B., Westermann, R., Lerch, S., 2024. Postprocessing of ensemble weather forecasts using permutation-invariant neural networks. Artif. Intell. Earth Syst. http://dx.doi.org/10.1175/AIES-D-23-0070.1.

Horat, N., Klerings, S., Lerch, S., 2024. Improving model chain approaches for probabilistic solar energy forecasting through post-processing and machine learning. Adv. Atmos. Sci. http://dx.doi.org/10.1007/s00376-024-4219-2.

Horat, N., Lerch, S., 2024. Deep learning for postprocessing global probabilistic forecasts on subseasonal time scales. Mon. Weather Rev. 152, 667–687. http://dx.doi.org/10.1175/MWR-D-23-0150.1.

Hou, Z., Li, J., Wang, L., Zhang, Y., Liu, T., 2022. Improving the forecast accuracy of ECMWF 2-m air temperature using a historical dataset. Atmos. Res. 273, 106177. http://dx.doi.org/10.1016/j.atmosres.2022.106177.

Houze, R.A., 2012. Orographic effects on precipitating clouds. Rev. Geophys. 50, 1–47. http://dx.doi.org/10.1029/2011RG000365.

Hu, W., Ghazvinian, M., Chapman, W.E., Sengupta, A., Ralph, F.M., Monache, L.Delle., 2023. Deep learning forecast uncertainty for precipitation over Western US. Mon. Weather Rev. 151, 1367–1385. http://dx.doi.org/10.1175/MWR-D-22-0268.1.

J.J., Gómez-Navarro, Raible, C.C., Bozhinova, D., Martius, O., Valero, J.A.G., Montávez, J.P., 2018. A new region-aware bias-correction method for simulated precipitation in areas of complex orography. Geosci. Model. Dev. 11, 2231–2247. http://dx.doi.org/10.5194/gmd-11-2231-2018.

Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S.P., Mogensen, K., Zuo, H., Monge-Sanz, B.M., 2019. SEAS5: The new ECMWF seasonal forecast system. Geosci. Model. Dev. 12, 1087–1117. http://dx.doi.org/10.5194/gmd-12-1087-2019.

de Jong, C., 2015. Challenges for mountain hydrology in the third millennium. Front. Environ. Sci. 3, 1–13. http://dx.doi.org/10.3389/fenvs.2015.00038.

Jordan, A., Krüger, F., Lerch, S., 2019. Evaluating probabilistic forecasts with scoringrules. J. Stat. Softw. 90, http://dx.doi.org/10.18637/jss.v090.i12, arXiv:1709.04743.

Kingma, D.P., Ba, J.L., 2015. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. pp. 1–15, arXiv:1412.6980.

Klemm, T., McPherson, R.A., 2017. The development of seasonal climate forecasting for agricultural producers. Agricult. Forest. Meterol. 232, 384–399. http://dx.doi.org/10.1016/j.agrformet.2016.09.005.

Kumar, H., Zhu, T., Sankarasubramanian, A., 2023. Understanding the food-energy-water nexus in mixed irrigation regimes using a regional hydroeconomic optimization modeling framework. Water Resour. Res. 59, 1–24. http://dx.doi.org/10.1029/2022WR033691.

Laiti, L., Mallucci, S., Piccolroaz, S., Bellin, A., Zardi, D., Fiori, A., Nikulin, G., Majone, B., 2018. Testing the hydrological coherence of high-resolution gridded precipitation and temperature data sets. Water Resour. Res. 54, 1999–2016.

Lakatos, M., Lerch, S., Hemri, S., Baran, S., 2023. Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts. Q. J. R. Meteorol. Soc. 149, 856–877. http://dx.doi.org/10.1002/qj.4436, arXiv:2206.10237.

Lerch, S., Baran, S., Möller, A., Groß, R., Hemri, S., Graeter, M., 2020. Simulation-based comparison of multivariate ensemble post-processing methods. Nonlinear Process. Geophys. 27, 349–371. http://dx.doi.org/10.5194/npg-27-349-2020.

Lerch, S., Polsterer, K.L., 2022. Convolutional autoencoders for spatially-informed ensemble post-processing. In: International Conference on Learning Representations (ICLR) 2022 - AI for Earth and Space Science Workshop. arXiv. URL: https://arxiv.org/abs/2204.05102.

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

Mahrt, L., 2006. Variation of surface air temperature in complex terrain. J. Appl. Meteorol. Clim. 45, 1481–1493. http://dx.doi.org/10.1175/JAM2419.1.

Maier, H.R., Galelli, S., Razavi, S., Castelletti, A., Rizzoli, A., Athanasiadis, I.N., Sànchez-Marrè, M., Wu, W., Humphrey, G.B., 2023. Exploding the myths: An introduction to artificial neural networks for prediction and forecasting. Environ. Model. Softw. 167, http://dx.doi.org/10.1016/j.envsoft.2023.105776.

Mallucci, S., Majone, B., Bellin, A., 2019. Detection and attribution of hydrological changes in a large alpine river basin. J. Hydrol. 575, 1214–1229.

Matheson, J.E., Winkler, R.L., 1976. Scoring rules for continuous probability distributions.. Manag. Sci. 22, 1087–1096. http://dx.doi.org/10.1287/mnsc.22.10.1087.

Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C., Liniger, M.A., 2018. Skill of subseasonal forecasts in europe: Effect of bias correction and downscaling using surface observations. J. Geophys. Res.: Atmospheres 123, 7999–8016. http://dx.doi.org/10.1029/2017JD027923, URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2017JD027923, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2017JD027923.

Morlot, M., Russo, S., Feyen, L., Formetta, G., 2023. Trends in heat and cold wave risks for the italian trentino-alto adige region from 1980 to 2018. Nat. Hazards Earth Syst. Sci. 23, 2593–2606.

Napoli, A., Crespi, A., Ragone, F., Maugeri, M., Pasquero, C., 2019. Variability of orographic enhancement of precipitation in the alpine region. Sci. Rep. 9, 1–8. http://dx.doi.org/10.1038/s41598-019-49974-5.

Pepin, N.C., Arnone, E., Gobiet, A., Haslinger, K., Kotlarski, S., Notarnicola, C., Palazzi, E., Seibert, P., Serafin, S., Schöner, S., Thornton, J.M., Vuille, M., Adler, C., 2022. Climate changes and their elevational patterns in the mountains of the world. Rev. Geophys. 60, 1–40. http://dx.doi.org/10.1029/2020RG000730.

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

Rasp, S., Lerch, S., 2018. Neural networks for postprocessing ensemble weather forecasts. Mon. Weather Rev. 146, 3885–3900, arXiv:1805.09091.

Ratri, D.N., Whan, K., Schmeits, M., 2019a. A comparative verification of raw and bias-corrected ECMWF seasonal ensemble precipitation reforecasts in Java (Indonesia). J. Appl. Meteorol. Clim. 58, 1709–1723. http://dx.doi.org/10.1175/JAMC-D-18-0210.1.

Ratri, D.N., Whan, K., Schmeits, M., 2019b. A comparative verification of raw and bias-corrected ecmwf seasonal ensemble precipitation reforecasts in java (indonesia). J. Appl. Meteorol. Clim. 58, 1709–1723. http://dx.doi.org/10.1175/JAMC-D-18-0210.1, URL: https://journals.ametsoc.org/view/journals/apme/58/8/jamc-d-18-0210.1.xml.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven earth system science. Nature 566, 195–204. http://dx.doi.org/10.1038/s41586-019-0912-1.

Robertson, D.E., Fu, G., Barron, O., Hodgson, G., Schepen, A., 2024. A new approach of coupled long-range forecasts for streamflow and groundwater level. J. Hydrol. 631, 130837. http://dx.doi.org/10.1016/j.jhydrol.2024.130837.

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

Scheuerer, M., Switanek, M.B., Worsnop, R.P., Hamill, T.M., 2020. Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over california. Mon. Weather Rev. 148, 3489–3506. http://dx.doi.org/10.1175/MWR-D-20-0096.1.

Schulz, B., Köhler, L., Lerch, S., 2024. Aggregating distribution forecasts from deep ensembles. http://dx.doi.org/10.48550/arXiv.2204.02291, arXiv:2204.02291.

Schulz, B., Lerch, S., 2022. Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. Mon. Weather Rev. 150, 235–237. http://dx.doi.org/10.1175/MWR-D-21-0150.1, arXiv:2106.09512.

Themeßl, M., Gobiet, A., Leuprecht, A., 2011. Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. Int. J. Climatol. 31, 1530–1544. http://dx.doi.org/10.1002/joc.2168.

Tripathy, S.S., Vittal, H., Karmakar, S., Ghosh, S., 2020. Flood risk forecasting at weather to medium range incorporating weather model, topography, socio-economic information and land use exposure. Adv. Water Resour. 146, 103785. http://dx.doi.org/10.1016/j.advwatres.2020.103785.

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

Vannitsem, S., Wilks, D.S., Messner, J.W., 2018. Statistical Postprocessing of Ensemble Forecasts. Elsevier.

Velasquez, P., Messmer, M., Raible, C.C., 2020. A new bias-correction method for precipitation over complex terrain suitable for different climate states: A case study using WRF (version 3.8.1). Geosci. Model. Dev. 13, 5007–5027. http://dx.doi.org/10.5194/gmd-13-5007-2020.

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

Viviroli, D., Dürr, B., Meybeck, M., Weingartner, R., 2007. Mountains of the world, water towers for humanity: Typology, mapping, and global significance. Water Resour. Res. 43, 1–13. http://dx.doi.org/10.1029/2006WR005653.

Viviroli, D., Weingartner, R., 2004. The hydrological significance of mountains: from regional to global scale. Hydrol. Earth Syst. Sci. 8, 1017–1030. http://dx.doi.org/10.5194/hess-8-1017-2004.

Vogel, E., Lerat, J., Pipunic, R., Frost, A.J., Donnelly, C., Griffiths, M., Hudson, D., Loh, S., 2021. Seasonal ensemble forecasts for soil moisture, evapotranspiration and runoff across Australia. J. Hydrol. 601, 126620. http://dx.doi.org/10.1016/j.jhydrol.2021.126620.

Wessel, J.B., Ferro, C.A., Kwasniok, F., 2024. Lead-time-continuous statistical postprocessing of ensemble weather forecasts. Q. J. R. Meteorol. Soc. 150, 2147–2167. http://dx.doi.org/10.1002/qj.4701.

Xu, J., Sun, X., Zhang, Z., Zhao, G., Lin, J., 2019. Understanding and improving layer normalization. Adv. Neural Inf. Process. Syst. 32, 1–19, arXiv:1911.07013.