The EUPPBench postprocessing benchmark dataset v1.0

Jonathan Demaeyer[1,2], Jonas Bhend[3], Sebastian Lerch[4], Cristina Primo[5], Bert Van Schaeybroeck[1], Aitor Atencia[6], Zied Ben Bouallègue[7], Jieyu Chen[4], Markus Dabernig[6], Gavin Evans[8], Jana Faganeli Pucer[9], Ben Hooper[8], Nina Horat[4], David Jobst[10], Janko Merše[11], Peter Mlakar[9,11], Annette Möller[12], Olivier Mestre[13,14], Maxime Taillardat[13,14], and Stéphane Vannitsem[1,2]

[1]Royal Meteorological Institute of Belgium, Brussels, Belgium
[2]European Meteorological Network (EUMETNET), Brussels, Belgium
[3]Federal Office of Meteorology and Climatology MeteoSwiss, Zurich, Switzerland
[4]Karlsruhe Institute of Technology, Karlsruhe, Germany
[5]Deutscher Wetterdienst, Offenbach, Germany
[6]GeoSphere Austria, Vienna, Austria
[7]European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom
[8]Met Office, Exeter, United Kingdom
[9]University of Ljubljana, Faculty of Computer and Information Science, Slovenia
[10]University of Hildesheim, Hildesheim, Germany
[11]Slovenian Environment Agency, Ljubljana, Slovenia
[12]Bielefeld University, Bielefeld, Germany
[13]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France
[14]Météo-France, Toulouse, France

Correspondence: Jonathan Demaeyer (Jonathan.Demaeyer@meteo.be)

Abstract. Statistical Postprocessing of medium-range weather forecasts is an important component of modern forecasting systems. Since the beginning of modern data science, numerous new postprocessing methods have been proposed, complementing an already very diverse field. However, one of the questions that frequently arises when considering different methods in the framework of implementing operational postprocessing is the relative performance of the methods for a given specific task. It is particularly challenging to find or construct a common comprehensive dataset that can be used to perform such comparisons. Here, we introduce the first version of *EUPPBench*, a dataset of time-aligned forecasts and observations, with the aim to facilitate and standardize this process. This dataset is publicly available at https://github.com/EUPP-benchmark/climetlab-eumetnet-postprocessing-benchmark. We provide examples on how to download and use the data, propose a set of evaluation methods, and perform a first benchmark of several methods for the correction of 2-meter temperature forecasts.

1 Introduction

Since the advent of numerical weather prediction, statistical postprocessing techniques have been used to correct ~~forecasts~~ forecast biases and errors. The term "postprocessing techniques" here refers to methods which use past forecasts and observations ~~combined together to learn about the models weather~~ to learn information about forecast deficiencies, ~~aiming to use that knowledge to correct their~~ that can then be used to correct future forecasts. Nowadays, postprocessing of weather forecasts is an important part of the forecasting chain in modern prediction systems at national and international meteorological services.

Many postprocessing approaches have been proposed during the last half century, ranging from the so-called *Perfect Prog* method (Klein et al., 1959; Klein and Lewis, 1970) to *Bayesian Model Averaging* (BMA) techniques (Raftery et al., 2005), and including the emblematic *Model Output Statistics* (MOS) approach (Glahn and Lowry, 1972). Some of these methods have been adapted to deal with ensemble forecasts and also calibrate the associated forecast probabilities, like the ~~EMOS~~ Ensemble MOS (EMOS) method (Gneiting et al., 2005). Recently, machine learning-based methods ~~were~~ have been proposed (Taillardat et al., 2016; Rasp and Lerch, 2018; Bremnes, 2020), which were shown to improve upon the conventional methods (Schulz and Lerch, 2022).

Systematic intercomparison exercises of both univariate ~~(e.g., Rasp and Lerch, 2018; Schulz and Lerch, 2022; Chapman et al., 2022; Ch~~ (e.g., Rasp and Lerch, 2018; Schulz and Lerch, 2022; Chapman et al., 2022) and multivariate ~~(e.g., Wilks, 2015; Perrone et al., 2020; Lerc~~ (e.g., Wilks, 2015; Perrone et al., 2020; Lerch et al., 2020; Chen et al., 2022; Lakatos et al., 2023) postprocessing methods exist, often based on artificial simulated datasets mimicking properties of real-world ensemble forecasting systems, or based on real-world datasets consisting of ensemble forecasts and observations for specific use-cases. However, there ~~currently~~ is no comprehensive, widely applicable benchmark dataset available for station- and grid-based postprocessing that facilitates re-use and comparisons~~by serving multiple purposes, e.g. by~~ , including a large set of potential input predictors and several target variables relevant to operational weather forecasting at meteorological services. The aim of the present work is to pave the way towards achieving these aims, with the publication of an extensive - *analysis-ready* - forecast and observation dataset, ~~both gridded and~~ consisting of both gridded data and data at station locations. By an analysis-ready dataset, we mean that the dataset formatting is tailored to obtain the most optimal match between observations and forecasts. In practice, this means that the observations are not provided as conventional time series but rather at the times and locations that match the forecasts.

Recently, the need for a common platform based on which different postprocessing techniques of weather forecasts can be compared, was highlighted (Vannitsem et al., 2021), and extensively discussed ~~in the context~~ between several members of the expert team of the postprocessing module running within the programs of the European Meteorological Network (EUMETNET)~~working group on postprocessing, called EUPP~~. Here, we introduce the first step in the development of such a platform, in the form of an easily-accessible dataset that can be used by a large community of users interested in the design of efficient postprocessing algorithms of weather forecasts for different applications. As stated in Dueben et al. (2022), comprehensive benchmark datasets are needed to enable a fair quantitative comparison between different tools and methods, while reducing the need to design and build them, a task which requires domain-specific knowledge. In this view, common benchmark datasets facilitate the collaboration of different communities with different expertise, by lowering the "energy barrier" required to embark on specific problems which would have otherwise required an excessive and discouraging amount of resources.

Many datasets related to weather and climate prediction were released during the last 3 years, emphasizing the need and appetite of the field for ever more data. For instance, datasets have been published related to sea ice drift (Rabault et al., 2022), to hydrology (Han et al., 2022), to learning of cloud classes (Zantedeschi et al., 2019), to sub-seasonal and seasonal weather forecasting (Rasp et al., 2020; Garg et al., 2022; Lenkoski et al., 2022; Wang et al., 2022), to data-driven climate projections (Watson-Parris et al., 2022), and - most relevant to the present work - to the benchmarking of postprocessing methods (Haupt et al., 2021; Ashkboos et al., 2022; Kim et al., 2022). Haupt et al. (2021) distribute a collection of (partly pre-

existing) different datasets for specific postprocessing tasks, including ensemble forecasts of the Madden-Julian Oscillation, integrated vapour transport over the Eastern Pacific and Western United States, temperature over Germany and surface road conditions in the United Kingdom. By contrast, Ashkboos et al. (2022) provide a reduced set of global gridded 10-member ECMWF ensemble forecasts for selected target variables.

Providing weather- or climate-related datasets to the scientific community in a standardized and persistent way remains a challenge, which was recently simplified by the introduction of efficient tools to store and provide data to the users. We can mention for example *xarray* (Hoyer and Joseph, 2017), *Zarr* (Miles et al., 2020), *dask* and the package *climetlab* recently developed by the European Center for Medium-range Weather Forecasts (ECMWF). The dataset introduced in the present article is for instance provided by a *climetlab plugin*, but also accessible through other means and programming languages (see the Supplementary Information).

The EUPPBench dataset consists of gridded and point ECMWF sub-daily forecasts of ~~all~~ different kinds (deterministic high-resolution, ensemble forecasts and reforecasts) over central Europe (see Figure 1). EUPPBench encompasses both station- and grid-based forecasts for many different variables, enabling a large variety of applications. This - ~~complement~~ complemented by the inclusion of reforecasts - enables a realistic representation of operational ~~posptroccessing~~ postprocessing situations, allowing users and institutions to learn and improve their skills on this crucial process. These operational aspects are, to our knowledge, missing in the currently available postprocessing benchmark datasets.

The forecasts and reforecasts of EUPPBench are paired with station observations and gridded reanalysis for the purpose of training and verifying postprocessing methods. To demonstrate how this dataset can be used, a benchmark of state-of-the-art postprocessing methods has been conducted to improve medium-range temperature forecasts. Although limited in scope, the outcome of this benchmark already emphasizes the potential of the dataset to provide meaningful results and provides useful insights in the potential, diversity and limitations of postprocessing over the study domain. Additionally, performing the benchmark for the first time with a large community also allows to address the usefulness of the established guidelines and protocols and to draw lessons-learned which are important assets for the delivery of many more benchmarks to come.

This article is structured as follow: The dataset structure and metadata ~~is~~ are introduced in Section 2. The design and the verification setup of the benchmark which was carried out ~~on the occasion of the~~ upon publication of this dataset is explained in Section 3, while in Section 4, the benchmarked methods are detailed. The results of the benchmark are presented in Section 5. We draw some interesting conclusions in Section 6 and ~~give some prospects on~~ outline plans for the future development of the dataset and of other benchmarks to come. Finally, the code of the benchmark and the data availability of the dataset are provided in Section 7.

2 EUPPBench v1.0 dataset

The EUPPBench dataset is available on a portion of Europe covering ~~47.75~~45.75° to 53.5° in latitude, and 2.5° to 10.5° in longitude. Therefore, this domain includes mainly Belgium, France, Germany, Switzerland, Austria and the Netherlands. It is stored in Zarr format, a CF compatible format (Gregory, 2003; Eaton et al., 2003) which provides easy access and allows users

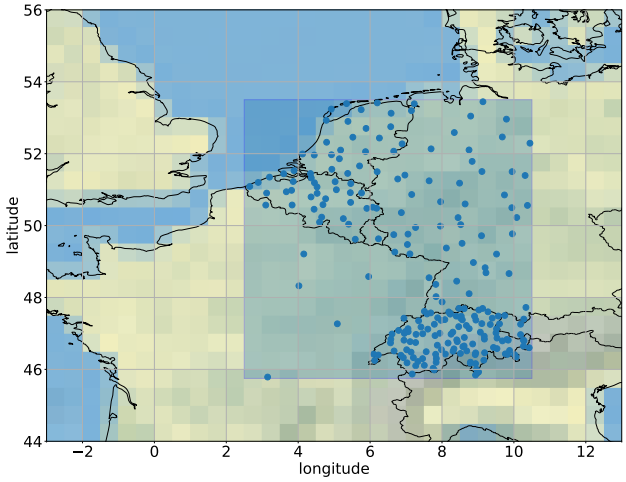

Figure 1. ~~The~~ Spatial coverage of the dataset. Blue rectangle: spatial domain of the gridded dataset~~and the~~ . Blue dots: position of the stations included in the dataset. Grey lines depict the latitude and longitude grid.

to "slice" the data along various dimensions in an effortless and efficient manner. In addition, the forecast and observation data are already paired together along corresponding dimensions, providing therefore an analysis-ready dataset for postprocessing benchmarking purposes.

EUPPBench includes both the 00Z (midnight) sub-daily ensemble forecasts and reforecasts (Hagedorn et al., 2008; Hamill et al., 2008) produced by the Integrated Forecasting System (IFS) of ECMWF during the years 2017 and 2018, which are released by the forecasting center under the CC-BY-4.0 license. Therefore there are 730 forecast dates and 209 reforecast dates over the ~~2-years~~ 2-year span, reforecasts being produced twice a week (Monday and Thursday). Apart from the ensemble forecasts and reforecasts, the high-resolution deterministic forecasts is also included. Each reforecast date, however, consists of 20 past forecasts computed with the model version valid at the reforecast date, and initialized from 1 to 20 years in the past at the same date of the year, thereby covering the period 1997 - 2017. In total, there are 4180 reforecasts. The number of ensemble members is 51 and 11 for the forecasts and reforecasts, respectively. This includes the forecast control run which is assumed to have the "closest" initial conditions to reality. The choice of the years 2017 and 2018 was motivated by the relative small number of model changes of the ECMWF forecast system during that period, and most importantly, the absence of model resolution modifications, as shown by Table 1. This implementation constraint is crucial to ensure that no supplementary model error biases are introduced in the datasets, as those biases can lead to a more-or-less severe degradation of the postprocessing performances (Lang et al., 2020; Demaeyer and Vannitsem, 2020).

Table 1. ECMWF IFS model changes during the 2017-2018 time span

Implementation date	Summary of changes	Resolution	Full IFS documentation
05-Jun-2018	Cycle 45r1	Unchanged	Cycle 45r1 full documentation
11-Jul-17	Cycle 43r3	Unchanged	Cycle 43r3 full documentation

Source: https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model

The forecasts and reforecasts time steps are 6-hourly (including the 0-th analysis time steps) up to a lead time of 120 hours (5 days). The variables considered are mainly surface variables and can be classified in two main categories: instantaneous or processed. Table 2 details these two different kinds of variable. Here, a "processed variable" means that the corresponding variable has either been accumulated, averaged or filtered over the past 6 hours. In addition to these surface variables, the Extreme Forecast Indices[1] (Lalaurette, 2003; Zsótér, 2006) and some pressure-level variables are also available (see respectively ~~the~~ Tables 3 and 4).

2.1 General data structure

The EUPPBench dataset consists of observations and forecasts in two types: a gridded dataset and a dataset at station locations. All forecasts are based on the ECMWF IFS forecasts. However, while the observational dataset at the station locations is based on ground measurements, the reanalysis ERA5 is taken as the gridded observational dataset. All forecast and reforecast datasets are provided for 31 variables, and additionally, the forecast dataset includes 9 EFI variables. The observations, on the other hand, include only 5 and 21 variables for the station-location and gridded datasets, respectively. Additionally metadata on the model and observations ~~is~~ are provided.

How to access the datasets is documented in the Section 7. We now detail in the following subsections the sources and properties of both dataset formats.

2.2 Gridded data

All gridded EUPPBench data is provided on a regular grid of $0.25° \times 0.25°$ corresponding roughly to a 25 km horizontal resolution at mid-latitude. As ~~aforementioned~~mentioned before, the forecasts and reforecasts are provided by the ECMWF forecasting model in operation at the moment of their issuance. They have both been re-gridded from the ECMWF original ensemble forecasts *O640* (or *O1280* for the deterministic forecasts)[2] grid to the regular grid using the ECMWF MIR interpolation package (Maciel et al., 2017), provided automatically by the MARS archive system. This re-gridding was done to be in line with the resolution of the ERA5 reanalysis (Hersbach et al., 2020) which provides the gridded observations of the EUPPBench dataset.

[1] Commonly abbreviated as "EFI".

[2] The ensemble forecasts grid *O640* has a horizontal resolution of 18 km while the deterministic forecasts grid *O1280* has a 9 km horizontal resolution.

Table 2. List of instantaneous and processed forecast variables on the surface level available in EUPPBench, all available in the EUPPBench gridded and station-location forecast datasets, and the availability of the corresponding gridded and station-location observations.

Parameter name	Short name	Units	Gridded obs.	Station obs.
2 metre temperature	t2m	K	x	x
10 metre U wind component	10u	$\mathrm{m\ s^{-1}}$	x	
10 metre V wind component	10v	$\mathrm{m\ s^{-1}}$	x	
Total cloud cover	tcc	$\in [0,1]$	x	x
100 metre U wind component	100u	$\mathrm{m\ s^{-1}}$		
100 metre V wind component	100v	$\mathrm{m\ s^{-1}}$		
Convective available potential energy	cape	$\mathrm{J\ kg^{-1}}$	x	
Soil temperature level 1	stl1	K	x	
Total column water	tcw	$\mathrm{kg\ m^{-2}}$	x	
Total column water vapour	tcwv	$\mathrm{kg\ m^{-2}}$	x	
Volumetric soil water layer 1	swvl1	$\mathrm{m^3\ m^{-3}}$	x	
Snow depth	sd	m	x	
Convective inhibition	cin	$\mathrm{J\ kg^{-1}}$		
Visibility	vis	m		x
Total precipitation	tp6	m	x	x
Surface sensible heat flux	sshf6	$\mathrm{J\ m^{-2}}$	x	
Surface latent heat flux	slhf6	$\mathrm{J\ m^{-2}}$	x	
Surface net solar radiation	ssr6	$\mathrm{J\ m^{-2}}$	x	
Surface net thermal radiation	str6	$\mathrm{J\ m^{-2}}$	x	
Convective precipitation	cp6	m	x	
Maximum temperature at 2 metres	mx2t6	K	x	
Minimum temperature at 2 metres	mn2t6	K	x	
Surface solar radiation downwards	ssrd6	$\mathrm{J\ m^{-2}}$	x	
Surface thermal radiation downwards	strd6	$\mathrm{J\ m^{-2}}$	x	
10 metre wind gust	10fg6	$\mathrm{m\ s^{-1}}$	x	x

Remark: A '6' was added to the usual ECMWF short names to indicate the span (in hours) of the accumulation or filtering.

We recognize that gridded observational datasets over the study domain exist for specific variables that are more accurate than ERA5. For instance, in the case of precipitation-related variables (like the *total precipitation* contained in the dataset at hand), ERA5 has been shown to provide - compared to other datasets - a poor agreement with stations observations (Zandler et al., 2019), mixed performances when used to derive hydrological products (Hafizi and Sorman, 2022), yet, good results when using Perfect prog downscaling methods (Horton, 2022). Notwithstanding, we emphasize that the goal of this gridded dataset

Table 3. List of available Extreme Forecast Indices, all available in the EUPPBench gridded and station-location forecast datasets.

Parameter name	Short name
2 metre temperature EFI	2ti
10 metre wind speed EFI	10wsi
10 metre wind gust EFI	10fgi
CAPE EFI	capei
CAPE shear EFI	capesi
Maximum temperature at 2m EFI	mx2ti
Minimum temperature at 2m EFI	mn2ti
Snowfall EFI	sfi
Total precipitation EFI	tpi

Remark: By definition, observations are not available for the EFI. The EFI are available for the model step ranges (in hours) 0-24, 24-48, 48-72, 72-96, 96-120, 120-144 and 144-168. The range of values of EFI goes from -1 to +1.

Table 4. List of variables on pressure levels, all available in the EUPPBench gridded and station-location forecast datasets.

Parameter name	Level	Short name	Units
Temperature	850	t	K
U component of wind	700	u	$m\ s^{-1}$
V component of wind	700	v	$m\ s^{-1}$
Geopotential	500	z	$m^2\ s^{-2}$
Specific humidity	700	q	$kg\ kg^{-1}$
Relative humidity	850	r	%

Remark: Only gridded observations (reanalysis) are available for these variables.

is to provide a representative "truth" for the purpose of benchmarking of postprocessing methods. Additionally, the availability of a wide range of variables in ERA5 and the spatio-temporal consistency among different meteorological variables ~~are very important in this context~~(a very important aspect in the present context) cannot be provided by gridded observational datasets.

2.3 EUPPBench data at station locations

Subdaily station observations have been provided by many National Meteorological Services (NMS) participating in the construction of this dataset, and a big part of the station data can be considered as open data (see section 7). The observations of 234 stations cover the entire 22-year time period 1997-2018 necessary to match the reforecasts and forecasts. The elevation of these stations varies from a few meters below the sea level up to 3562 meters for the Jungfraujoch station in Switzerland. These

Table 5. List of available constant fields

Parameter name	Short name	Units
Land use	landu	1,2,3,...,44,48
Model terrain height	mterh	m
Surface Geopotential	z	$m^2\ s^{-2}$

The land usage is extracted from the CORINE2018 dataset (Copernicus Land Monitoring Service, 2018). More details are provided in the "legend" entry of the metadata within each file.
The model terrain height is extracted from the EU-DEM v1.1 data elevation model dataset (Copernicus Land Monitoring Service).
Finally, the model orography can be obtained by dividing the surface geopotential by $g = 9.80665\ m\ s^{-2}$.

stations constitute the most authoritative sources of information about weather and climate provided in each of the involved countries, being constantly monitored and the quality of the data being checked.

The EUPPBench dataset at station locations consists of the ECMWF forecasts and reforecasts at the grid point closest to the station locations and the associated observations, matched for each lead time. As shown in the Table 2, there are ~~mainly~~ 5 variables currently available: 2 metre temperature (*t2m*), total cloud cover (*tcc*), visibility (*vis*), total precipitation (*tp6*) and 10 metre wind gust (*10fg6*). More observation variables will be added in subsequent versions of the dataset.

2.4 Static data and metadata

In addition to the forecasts and reforecasts, auxiliary fields are provided, such as the land usage ~~or~~ and the surface geopotential which is proportional to the model orography (see Figure 2). Table 5 synthesizes this part of the dataset. These constant fields have been extracted and are also provided in the ~~stations~~ station metadata.

Depending on the kind of dataset, dimensions and different information are embedded in the data: For gridded data, the metadata available in the forecast, reforecast and observation datasets are detailed in Table 6. For station data, the forecast and reforecast metadata are detailed ~~are detailed~~ in Table 7, while the observation metadata are detailed in Table 8. For all data, attributes specifying the sources and the license are always ~~present~~provided.

3 Postprocessing benchmark

To illustrate the usefulness of the EUPPBench dataset, a benchmark of several state-of-the-art postprocessing methods - ~~and~~ many of which are currently in operation in NMSs - was performed, ~~along with~~ including some more recent and more advanced methods. This first exercise was based on a small subset of the dataset. Along the same line, the verification process of this benchmark also focused on some general aspects typically considered for operational postprocessing. In this section,

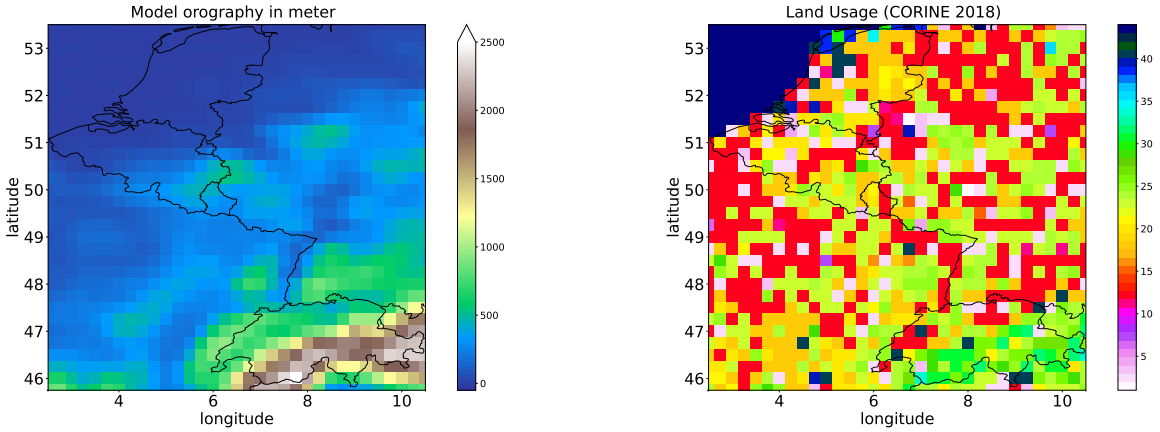

Figure 2. Static fields in the gridded dataset. Left panel: The model orography obtained by dividing the model surface geopotential by $g = 9.80665$ m s^{-2}. Right panel: Grid point land usage provided by the CORINE 2018 dataset (Copernicus Land Monitoring Service, 2018). Numerical codes indicating the usage categories ~~is~~ are included in the dataset metadata.

Table 6. The metadata provided in the files of the gridded forecasts, reforecasts and observational datasets.

Metadata	Description
latitude	Latitude of the grid points.
longitude	Longitude of the grid points.
depthBelowLandLayer	Layer below the surface (valid for some variables only, here there is only the upper surface level).
number	Number of the ensemble member. The 0-th member is the control run. Also present in observation for compatibility reasons, but set to 0.
time	Forecast or reforecast date (reforecasts are only issued on Mondays and Thursdays).
year	Dimension to identify the year in the past, year=1 means a forecast valid 20 years ago at the reforecast day and month, year=20 means a forecast valid one year before the reforecast date. Only valid for reforecasts.
step	Step of the forecast (the lead time).
surface	Layer of the variable considered (here there is just one, at the surface).
isobaricInhPa	Pressure level in hectopascal (or millibar).
valid_time	Actual time and date of the corresponding forecast data.

Remark: **Bold** metadata denote dimensions indexing the datasets.

we describe the general framework that we used to conduct this benchmark. The following sections will be devoted to the methods and the results obtained. We begin by detailing the design of this experiment.

Table 7. The metadata provided in the files of the forecast, reforecast at the station locations.

Metadata	Description
station_latitude	Latitude of the station.
station_longitude	Longitude of the station.
station_altitude	Elevation of the station (in ~~meter~~meters).
station_id	Unique identifier ~~of~~ for the station.
depthBelowLandLayer	Layer below the surface (valid for some variables only, here there is only the upper surface level).
number	Number of the ensemble member. The 0-th member is the control run. Also present in observation for compatibility reasons, but set to 0.
time	Forecast or reforecast date (reforecasts are only issued on Mondays and Thursdays).
year	Dimension to identify the year in the past, year=1 means a forecast valid 20 years ago at the reforecast day and month, year=20 means a forecast valid one year before the reforecast date. Only valid for reforecasts.
step	Step of the forecast (the lead time).
surface	Layer of the variable considered (here there is just one, at the surface).
isobaricInhPa	Pressure level in hectopascal (or millibar).
station_land_usage	Land usage at the station location, extracted from the CORINE 2018 dataset.
station_name	Name of the station.
model_latitude	Latitude of the model grid point.
model_longitude	Longitude of the model grid point.
model_altitude	True elevation (in ~~meter~~meters) of the model grid point, extracted from the EU-DEMv1.1 data elevation model dataset.
model_orography	Surface height (in ~~meter~~meters) in the model at the model grid point.
model_land_usage	Land usage at the model grid point, extracted from the CORINE 2018 dataset.
valid_time	Actual time and date of the corresponding forecast data.

Remark 1: **Bold** metadata denote dimensions indexing the datasets.
Remark 2: The metadata with 'model' in their name indicate properties of the closest model grid point to the station location, and at which the forecasts corresponding to the station observations was extracted from the gridded dataset.

3.1 Experiment design

The postprocessing benchmark at hand considers the correction of the ensemble forecasts of the 2 ~~meter~~ metre temperature at the nearest forecast grid point from every station available in the dataset, spanning several European countries and the whole EUPPBench area. We note that this area includes orographically difficult regions, nearly-flat plains, and also stations close to the sea or located on islands. Discrepancies may therefore occur between ~~forecast~~ forecasts and observations due to poor observational representativity at the scale of the model or due to challenges in the model representation of a wide range of

Table 8. Station observations metadata

Metadata	Description
altitude	Elevation of the station (in meter).
land_usage	Land usage at the station location, extracted from the CORINE 2018 dataset.
latitude	Latitude of the station.
longitude	Longitude of the station.
station_id	Unique identifier of the station.
station_name	Name of the station.
step	Step of the forecast (the lead time).
time	Forecast or reforecast date (reforecasts are only issued on Mondays and Thursdays).

Remark: **Bold** metadata denote dimensions indexing the datasets.

physical processes. We note also that due to the coarse nature of the gridded forecast dataset, the forecast grid points are not evenly situated with respect to the stations they represent, with sometimes huge differences in elevation or situation (e.g. forecast point at sea), that may induce large temperature biases.

Within this simplified benchmark exercise, the only predictor that could be used to perform the postprocessing is the temperature at 2 ~~meter~~ meters itself. Additionally the use of the (static) metadata was allowed and some methods used latitude and longitude, elevation, land use, model orography, lead time and also the day of the year. The 11-member reforecasts produced during the 2017-2018 period were considered as training data while the 51-member forecasts for the same period was used as test data for verification. This setup introduced some challenges for the implementation of some of the postprocessing methods described below.

To avoid a potential overlap between the reforecasts and the forecasts, the forecast from 2017 that were included in the reforecasts of 2018 have been removed from the training dataset. Since the ECMWF reforecasts date are ~~being~~ produced each Monday and Thursday, the reforecasts (per lead time) for 2017 do not overlap with those of 2018.

However, one notable difference between the training data and the test data is the number of ensemble members: ensemble forecasts contain 51 members while ensemble reforecasts include 11 members. Also, note that in both cases, the ECMWF control run forecast is included in the ensemble (as the 0-th ensemble member). The high-resolution deterministic forecast runs were not used nor postprocessed in the current benchmark.

3.2 Verification setup and methodology

Forecasts ~~are given at a particular run time for a particular hour and~~ from the various methods analyzed here are available for particular forecast initialization times, lead times, and for a specific place of interest ~~. Differences between forecast and observation provide an idea about the forecast error. When the study extends to time periods and areas, these errors have to be aggregated , and according~~ (here the locations of measurement stations). For those distinct pairs of ensemble forecasts

and verifying observations, we compute a range of forecast verification measures to quantify forecast performance. These verification measures are then aggregated in time or space in order to extract summary information on forecast performance. According to how the aggregation is done, the ~~analyses~~ analysis will focus on different aspects of the forecasts: ~~Can we distinguish spatial patterns or dependence on the elevation in skill~~Do the forecast exhibit systematic temporal or spatial errors? How does the forecast quality decrease with the lead time? ~~Are there systematic deviations between the forecast members and the observations. In addition, when dealing with an ensemble of forecasts, is the ensemble well calibrated, i.e. is the forecast spread a good measure of the uncertainty?~~

Can we distinguish spatial patterns or does the performance depend on the elevation? The verification study ~~answers~~ addresses these questions by comparing the performance of the different methods ~~showing aggregated statistical scores using temporal series, maps and rank histograms. More specifically, the post-processed~~ using these aggregated verification measures. In particular, the postprocessed forecasts at the station locations are compared with the station observations within the test dataset (2017-2018)~~are compared here with observations. We use two metrics that quantify the forecast quality, or in other words, how well the ensemble forecast matches the observations~~.

Forecast quality is multi-faceted (Murphy, 1993) and no single score can capture all aspects. Here we use four metrics to address different aspects of forecast quality: the Bias ~~and~~ to diagnose forecast *bias*, the Continuous Ranked Probability Score (CRPS~~) Hersbach (2000). Bias addresses the skill of the ensemble average and corresponds to the average differences between forecast and observation. CRPS, on the other hand, addresses the probabilistic skill by comparing~~ , Hersbach 2000) to quantify forecast *accuracy*, the forecast spread to quantify *sharpness*, and the spread-error-ratio as an indication of forecast *reliability*. The Bias is defined as the average difference between the ensemble mean and observation, and points out if an ensemble has positive or negative systematic errors. The CRPS compares the Cumulative Distribution Functions (CDF) of the ~~forecast against the corresponding observation. The calibration of the ensemble is analyzed via a~~ forecasts with the corresponding observations. The CRPS generalizes the mean absolute error to probabilistic forecasts and is sensitive to both forecast *reliability* and *sharpness*.

For calibrated forecasts, the ensemble standard deviation (commonly referred to as forecast spread) corresponds to the magnitude of the forecast error. A sharper ensemble forecast (i.e. an ensemble forecast with low spread) is therefore more informative and skillful. In this study we analyse the spread/skill~~score equal to~~ relationship by comparing the ratio of the average ensemble standard deviation divided by the root-mean squared error of the ensemble mean. ~~As a reference forecasting dataset~~ A spread-error-ratio smaller than one indicates a lack of forecast spread (forecast under-dispersion), whereas values larger than one indicate over-dispersion. It should be noted that spread-error-ratio equal to one is only a necessary but not sufficient condition for forecast reliability and care should be taken when interpreting the spread-error-ratio in particular in the presence of remaining systematic biases. To complement, we also analyse rank histograms. These histograms show where the observation places within the ensemble when it is sorted from the lowest to the highest value. A reliable ensemble would lead to a flat rank histogram. The shape of the rank histogram can help to detect deficiencies in ensemble calibration, e.g. a U-shaped rank histogram indicates under-dispersion or conditional biases.

The verification using the different measures allows us to detect if the compared postprocessing methods have systematic errors or biases and if the postprocessed ensembles are well calibrated, over-, or under-confident.

The reference forecast dataset will be the raw IFS ensemble at the nearest forecast grid point from every station~~is used with an additional~~. The difference between IFS orography and station elevation is taken into account by applying a constant lapse-rate correction of 6.5°C/km~~to account for the difference between the IFS orography and the station elevation. Some~~. In this study, some results are also presented conditioned on the station elevation to detect remnant orographic influences.

The verification results for the different postprocessing methods are obtained after performing quality-control tests on the initial data to detect for possible inconsistencies, unrealistic values and missing data. Missing postprocessed predictions for individual time steps and locations in the test set are replaced by the direct model output (DMO). Postprocessing methods with missing values are therefore intentionally penalized. The rationale behind this is that EUPP aims for improving *operational* forecasting systems in which forecasts need to be provided in any case. Additionally, this approach discourages hedging, i.e. artificially increasing the performance of a postprocessing method, by replacing known cases with underperforming skill by a missing value. ~~Moreover, significant tests are run to assess if the score differences were significant~~Additionally, score differences are tested for statistical significance (see Appendix A).

4 Postprocessing methods

Along with the dataset and verification framework described above, the present work further includes a collection of forecasts of exemplary postprocessing methods along with corresponding code for their implementation. Note that with providing forecasts of a selected set of methods, we do not intend to provide a comprehensive or systematic comparison to establish the "best" approach, but rather aim to present an overview of both commonly used and more advanced methods ranging from approaches from statistics to machine learning. Those can be used in subsequent research for developing extensions to existing approaches and for comparing novel methods to established baselines. Short descriptions of the methods available in the present benchmark are provided below, verification results are presented in Section 5. Specific details regarding the adaptation and implementation of the different methods, as well as code are available from the corresponding Github repositories[3]. For a general overview of recent developments in postprocessing methodology, we refer to Vannitsem et al. (2018, 2021). Note that a direct comparison of computational costs is challenging because of the differences in terms of the utilized hardware infrastructure, software packages and parallelization capabilities, and might be considered in future work, ideally within a fully automated procedure (see Section 6). That said, the computational costs of all considered postprocessing methods are by several orders of magnitude lower than those required for the generation of the raw ensemble forecasts.

Within the present section, we use the following notation. For a specific forecast instance t (at a specific location and for a specific initialization and lead time), we denote the ensemble forecasts by $x_m(t), m = 1, ..., M$, their mean value by $\mu^{\text{ens}}(t)$, and their standard deviation by $\sigma^{\text{ens}}(t)$. The corresponding observation is denoted by $y(t)$.

[3]See a detailed list of the methods Github repositories at https://github.com/EUPP-benchmark/ESSD-benchmark.

4.1 Accounting for systematic and representativeness errors (ASRE)

ASRE postprocessing tackles systematic and representativeness errors in two independent steps. A local bias correction approach is applied to correct for systematic errors. For each station and each lead time, the averaged difference between reforecasts and observations in the training dataset is computed and removed from the forecast in the validation dataset. The difference averaging is performed over all training dates centered around the forecast ~~validity~~ valid date within a window of ± 30 days.

Representativeness errors are accounted for separately using a universal method inspired by the *Perfect Prog* approach (Klein et al., 1959; Klein and Lewis, 1970). A normal distribution is used to represent the diversity of temperature values that can be observed at a point within an area given the average temperature of that area. For an area of a given size (i.e., a model grid box), the variance of the distribution is expressed as a function of the difference between station elevation and model orography only (see Eq. 4 in Ben Bouallègue, 2020). Random draws from this probability distribution is added to each ensemble member to simulate representativeness uncertainty.

4.2 Reliability Calibration (RC)

Reliability Calibration is a simple, non-parametric technique that specifically targets improving the forecast reliability without degrading forecast resolution. Two additional steps are applied prior to Reliability Calibration, targeted at correcting forecast bias; initially a lapse rate correction of 6.5°C/km between the station elevation and model orography is applied, followed by a simple bias correction calculated independently at each station. Following bias correction, probabilistic forecasts are derived from the bias corrected ensemble member forecasts by calculating the proportion of ensemble members which exceed thresholds at 0.5°C intervals. At each threshold, the exceedance probabilities are calibrated separately. The Reliability Calibration implementation largely follows Flowerdew (2014), although in this study, all sites are aggregated into a single reliability table which is used to calibrate forecasts across all sites. As in Flowerdew (2014), a set of equally spaced percentiles are extracted using linear interpolation between the thresholds, which are treated as pseudo-ensemble members for verification. The non-parametric nature of Reliability Calibration makes it attractive for a range of diagnostics, including temperature, if combined with other simple calibration techniques such as those applied here. Reliability Calibration was implemented using IMPROVER (Roberts et al., 2022), an open-source codebase developed by the Met Office and collaborators.

4.3 Member-By-Member postprocessing (MBM)

The Member-By-Member approach calibrates the ensemble forecasts by correcting the systematic biases in the ensemble mean with a linear regression-based MOS technique and rescaling the ensemble members around the corrected ensemble mean (Van Schaeybroeck and Vannitsem, 2015). This procedure estimates the coefficients α^{MBM}, β^{MBM} and γ^{MBM} in the formula providing the corrected ensemble:

$$T^{C}\tilde{x}_m(t) = \alpha^{\mathrm{MBM}}(t) + \beta^{\mathrm{MBM}}(t)\,\mu^{\mathrm{ens}}(t) + \gamma^{\mathrm{MBM}}(t)\,{}^{\mathrm{ens}}x'_m(t), \tag{1}$$

by optimizing the CRPS ~~,~~separately for each station and for each lead time. ~~$\bar{T}_m^{\text{ens}}(t) = T_m^{\text{ens}}(t) - \mu^{\text{ens}}(t)$~~ $x'_m(t) = x_m(t) - \mu^{\text{ens}}(t)$ here denotes the deviation of the member m from the ensemble mean. The results were obtained with the Pythie package (Demaeyer, 2022a), training on the 11 members of the training dataset to obtain the coefficients α^{MBM}, β^{MBM} and γ^{MBM}, and then using them to correct the 51 ~~members~~ member forecasts of the test dataset. One of the main advantages of MBM postprocessing is that - by design - it preserves simultaneously spatial, temporal, and inter-variable correlations in the forecasts.

4.4 Ensemble model output statistics (EMOS)

EMOS is a parametric postprocessing method introduced in Gneiting et al. (2005). The temperature observations are modelled by a Gaussian distribution. The location (μ) and scale (σ) parameters of the forecast distribution can be described by two linear regression equations via

$$y(t) \sim \mathcal{N}(\mu, \sigma) \begin{cases} \mu(t) = \beta_0^{\text{EMOS}} + f_1^{\text{EMOS}}(\text{doy}) + \beta_1^{\text{EMOS}} \mu^{\text{ens}}(t) \\ \log(\sigma) = \gamma_0^{\text{EMOS}} + g_1^{\text{EMOS}}(\text{doy}) + \gamma_1^{\text{EMOS}} \log(\sigma^{\text{ens}}(t)), \end{cases} \tag{2}$$

with β_0^{EMOS}, γ_0^{EMOS}, β_1^{EMOS} and γ_1^{EMOS} as regression coefficients, and $f_1^{\text{EMOS}}(\text{doy})$ and $g_1^{\text{EMOS}}(\text{doy})$ as seasonal smoothing functions to capture a seasonal bias of location and scale. The seasonal smoothing function is a combination of annual and bi-annual base functions ($\sin(2\pi\,\text{doy}/365)$, $\cos(2\pi\,\text{doy}/365)$, $\sin(4\pi\,\text{doy}/365)$, and, $\cos(4\pi\,\text{doy}/365)$) as presented in Dabernig et al. (2017). The implemented EMOS version is based on the R-package `crch` (Messner et al., 2016) with maximum likelihood estimation. 51 equidistant quantiles between 1 and 99 % of the distribution are drawn to match the amount of members from the raw ECMWF forecasts, which were needed for verification. EMOS is applied separately to every station and lead time.

4.5 EMOS with heteroscedastic autoregressive error adjustments (AR-EMOS)

AR-EMOS extends the EMOS approach by estimating parameters of the predictive distribution based on ensemble forecasts adjusted for autoregressive behavior (Möller and Groß, 2016). For each ensemble forecast $x_m(t)$, the respective error series $z_m(t) := y(t) - x_m(t)$ is defined and an autoregressive (AR) process of order p_m is fitted to each $z_m(t)$ individually. Based on the estimated parameters of the AR(p_m) processes an AR-adjusted forecast ensemble is obtained via

$$\widetilde{x}_m(t) = x_m(t) + \alpha_m^{\text{AR}} + \sum_{j=1}^{p_m} \beta_{m,j}^{\text{AR}} [y(t-j) - x_m(t-j) - \alpha_m^{\text{AR}}], \tag{3}$$

where α_m^{AR} and $\beta_{m,j}^{\text{AR}}$, $j = 1, \ldots, p_m$ are the coefficients of the respective AR(p_m) process. The adjusted ensemble forecasts are employed to estimate the mean and variance parameter of the predictive Gaussian distribution. Estimation of the predictive variance was further refined in Möller and Groß (2020). The method is implemented in the R package `ensAR` (Groß and Möller, 2019). However, some adaptations had to be made to the method and implementation in order to accommodate the benchmark data, see code documentation in the corresponding Github repository.

4.6 D-vine copula based postprocessing (DVQR)

In the D-vine (drawable vine) copula based postprocessing, a multivariate conditional copula C is estimated using a pair-copula construction for the graphical D-vine structure according to Kraus and Czado (2017). D-vine copulas enable a flexible modelling of the dependence structure between the observation y and the ensemble forecast $x_1, \ldots, x_m$ (see, e.g., Möller et al., 2018). The covariates $x_1, \ldots, x_m$ are selected by their predictive strength based on the conditional log-likelihood. Afterwards, D-vine copula quantile regression (DVQR) allows to predict quantiles $\alpha \in (0, 1)$ that represent the postprocessed forecasts via

$$F^{-1}_{y|x_1,\ldots,x_m}(\alpha|x_1(t),\ldots,x_m(t)) := F_y^{-1}\left(C^{-1}(\alpha|F_{x_1}(x_1(t)),\ldots,F_{x_m}(x_m(t)))\right), \tag{4}$$

where F_{x_i} denotes the marginal distributions of x_i for all $i = 1, \ldots, m$, F_y^{-1} the inverse marginal distribution of y and C^{-1} is the conditional copula quantile function. The implementation of this method is mainly based on the R package `vinereg` by Nagler (2020), where the marginal distributions are kernel density estimates. DVQR is estimated separately for every station and lead time using a seasonal adaptive training period.

4.7 Distributional regression network (DRN)

Rasp and Lerch (2018) first proposed the use of neural networks (NNs) for probabilistic ensemble postprocessing. In a nutshell, their DRN approach extends the EMOS framework by replacing pre-specified link functions with a NN connecting inputs and distribution parameters, enabling flexible nonlinear dependencies to be learned in a data-driven way. The parameters of a suitable parametric distribution are obtained as the output of the NN, which may utilize arbitrary predictors as inputs, including additional meteorological variables from the NWP system and station information. In our implementation for EUPPBench, we closely follow Rasp and Lerch (2018), and assume a Gaussian predictive distribution. We fit a single DRN model per lead time jointly for all stations, and encode the station identifier and land-use via embedding layers to make the model locally adaptive. Since the use of additional input information has been a key aspect in the substantial improvements of DRN and subsequent extensions in other NN-based methods over EMOS, similar benefits are less likely here due to the limitation to ensemble predictions of the target variable only in the experimental setup, see Rasp and Lerch (2018) for more detailed comparisons.

4.8 ANET

ANET is a NN approach, similar to DRN, for postprocessing ensembles with variable member counts. ANET estimates the parameters of a predictive Gaussian distribution jointly for all lead times and over all stations. ANET processes individual ensemble members first, and combines them into a single output inside the architecture later. A dynamic attention mechanism facilitates focusing on important sample members, enabling ANET to retain more information about individual members in cases where the ensemble describes a more complex distribution. Likewise, we take advantage of the fact that we are predicting the parameters of a Gaussian distribution by computing the mean and spread of the residuals μ^i_Δ and σ^i_Δ rather than the direct distribution parameter values. ANET thus computes the distribution parameters for a lead time i as $\mu_i^{\text{ANET}}(t) = \mu_i^{\text{ens}}(t) + \mu_{\Delta,i}$, $\sigma_i^{\text{ANET}}(t) = S(\sigma_i^{\text{ens}}(t) + \sigma_\Delta, i)$, where S denotes the softplus activation function $S(x) = \ln(1 + \mathrm{e}^x)$, ensuring that the standard

deviation remains positive. ~~To further increase the stability of ANET we randomly varied the number of ensemble members passed to the network during training.~~ The model is trained by minimizing the negative log-likelihood function. For more details about the method, see Mlakar et al. (2023a).

5 Results

Here we present the results from the verification of the submission to the benchmark. The CRPS (Fig. 3a) as a measure of forecast accuracy clearly demonstrates the benefit of postprocessing. The elevation-corrected ECMWF DMO exhibits pronounced diurnal variability in CRPS with forecast errors at night being considerably more pronounced than during the day. Postprocessing achieves a reduction of these forecast errors by up to 50% early in the forecast lead time and by 10-40% on day 5. Most postprocessing methods perform similarly with the notable exception of ANET that achieves the lowest CRPS and exhibits ~~much~~ less diurnal variability in forecast errors.

Postprocessing improves forecast performance by reducing systematic biases (Fig. 3b) and by increasing ensemble spread (Fig. 3c) to account for sources of variability not included in the NWP system. ~~Again the~~ The ensemble spread of most postprocessing methods is similar with the notable exception of RC that generates much more dispersed forecasts in particular early in the forecast lead time.

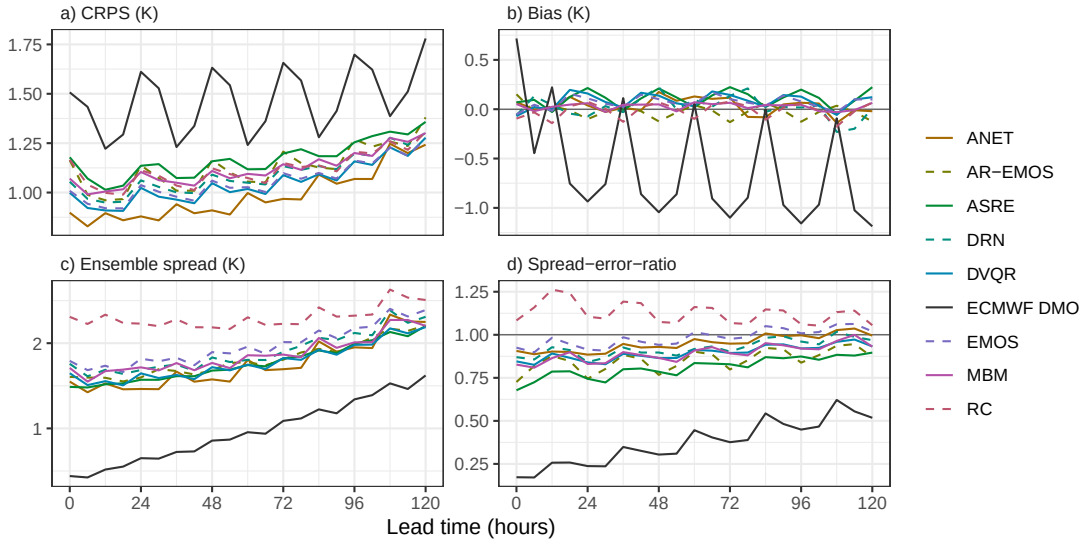

Figure 3. Average scores ~~in dependence of~~ at each lead time, aggregated across all stations and forecasts.

Forecast calibration is assessed with the spread-error-ratio (Fig. 3d) and the rank histogram (Fig. 4). ECMWF DMO is heavily over-confident ~~i.e.~~ resulting in a spread-error-ratio smaller than 1 and a U-shaped rank histogram. Postprocessed forecasts are much better calibrated with indication of some remaining forecast over-confidence for all methods but RC (Fig. 3d). The

rank histogram in Fig. 4 allows for a different perspective on forecast calibration with indication of forecast over-dispersion (inverse U-shape) for many of the postprocessed forecasts.

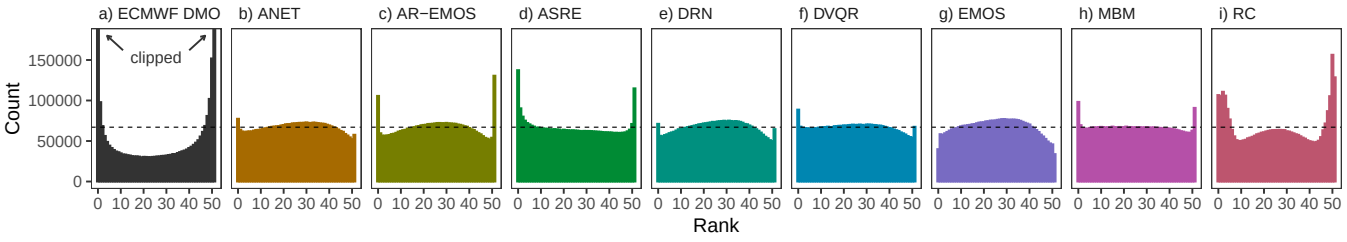

Figure 4. Rank histogram of forecasts submitted to the benchmark experiment, aggregated across all stations, forecasts and lead times. ~~Please note~~ Note that the visualization for ECMWF DMO is clipped for better comparison with the rank histograms of the postprocessed forecasts.

Postprocessed forecasts have been produced for a number of stations in central Western Europe. With very few exceptions, postprocessing improves forecast quality everywhere as illustrated by the positive values of CRPSS in Figure 5. These findings are corroborated by the significance testing presented in Appendix A. Most of the postprocessing methods perform similarly with more pronounced improvements in complex topography and less pronounced improvements in the northern and predominantly flat part of the domain. As a notable exception, ANET forecasts perform better in particular for high-altitude stations in Switzerland ~~and for coastal stations in Belgium and the Netherlands~~(see also Fig. 6).

In Fig. 6 we present the relationship with station altitude for a range of scores ~~in dependence of the station altitude~~ to further explore the specifics of the postprocessing methods. For example forecasts with the elevation-corrected ECMWF DMO for high-altitude stations are systematically too cold, indicating that the constant lapse rate correction applied to ECMWF DMO is an approximation at best. The AR-EMOS methods appears to produce the smallest biases overall, whereas there is some remaining negative bias at altitude in many of the methods and positive biases in RC. The remaining large biases in the AR-EMOS ~~and EMOS methods~~ method are from missing predictions that have been filled with ECMWF DMO.

The CRPS at each station in Fig. 6 shows that the reduction in forecast errors and correspondingly increased forecast skill is generally more pronounced at altitude. Compared with the other postprocessing approaches, ANET achieves lower CRPS for high-altitude stations (above 1000 m)~~and for a cluster of stations below 100 m predominantly located in the Netherlands (see also Fig. 5)~~.

The spread-error-ratio as a necessary condition for forecast calibration also reveals considerable differences between postprocessing approaches. The ASRE and RC methods in particular exhibit large variations in spread-error-ratio from station to station, whereas the other methods exhibit much more uniform spread-error-ratios. Please note that the strong under-dispersion of ECMWF DMO as indicated by the spread-error-ratio in Fig. 3 and Fig. 6 is slightly reduced when systematic biases are removed (not shown). For the postprocessed forecasts, the effect of remaining systematic biases on the spread-error-ratio is negligible. More detailed analyses of the results would of course be possible, but are beyond the scope of this publication on the benchmark dataset, and will be the subject of a dedicated work.

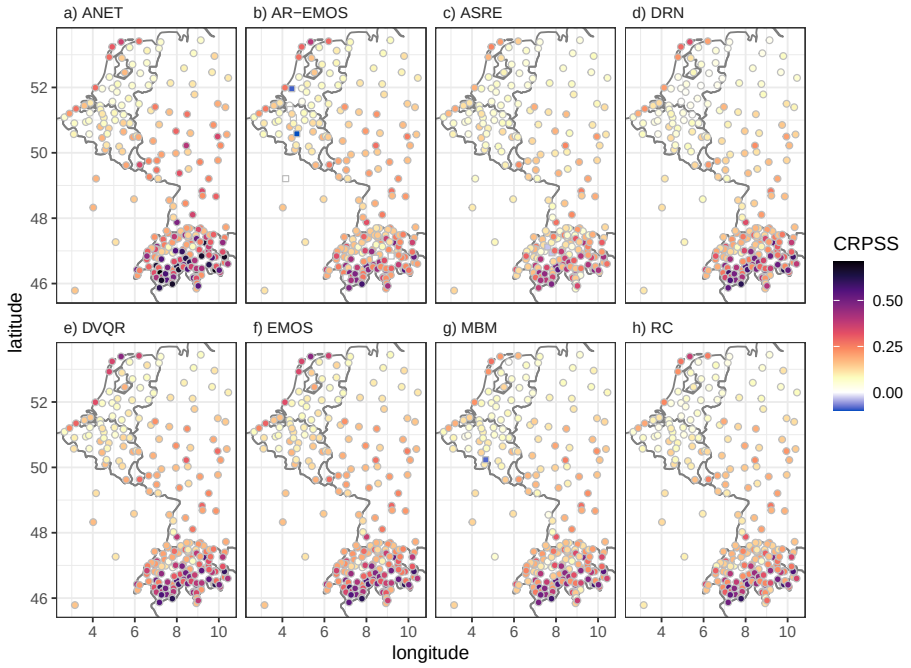

Figure 5. Continuous ranked probability skill score (CRPSS) per station, averaged across all forecasts and all lead times. CRPSS is computed using the ECMWF DMO as the reference forecast and positive values indicate that the postprocessed forecasts outperform ECMWF DMO. Stations at which forecast skill is negative are marked by square symbols.

6 Conclusions and prospects

A benchmark dataset is proposed in the context of the EUMETNET PostProcessing (EUPP) program for comparing statistical postprocessing techniques that are nowadays an integral part of many operational weather-forecasting suites. This dataset includes ensemble forecasts and reforecasts of the ECMWF for the period 2017-2018, and the corresponding gridded and station observations, over a region covering a small portion of western Europe. This region covers a variety of topographies including ~~as~~ coastal, flat and mountainous areas. To illustrate the usefulness of this dataset, a standardized ~~excercise~~ exercise is established in order to allow an objective and rigorous intercomparison of postprocessing methods. This exercise included the contribution of many well-established state-of-the-art postprocessing techniques. Despite the limited scope of the presented exercise, this collaborative effort will serve as a reference framework and will be strongly extended. The whole process includes: (i) the download of the data or their access on the European Weather Cloud (where the dataset is stored, see Section 7); (ii) the application of the different techniques by the contributors; and (iii) the verification of the results by the verification team. This proof-of-concept proved to be very successful.

While the authors constructed and performed this benchmark, some lessons were learned along the way:

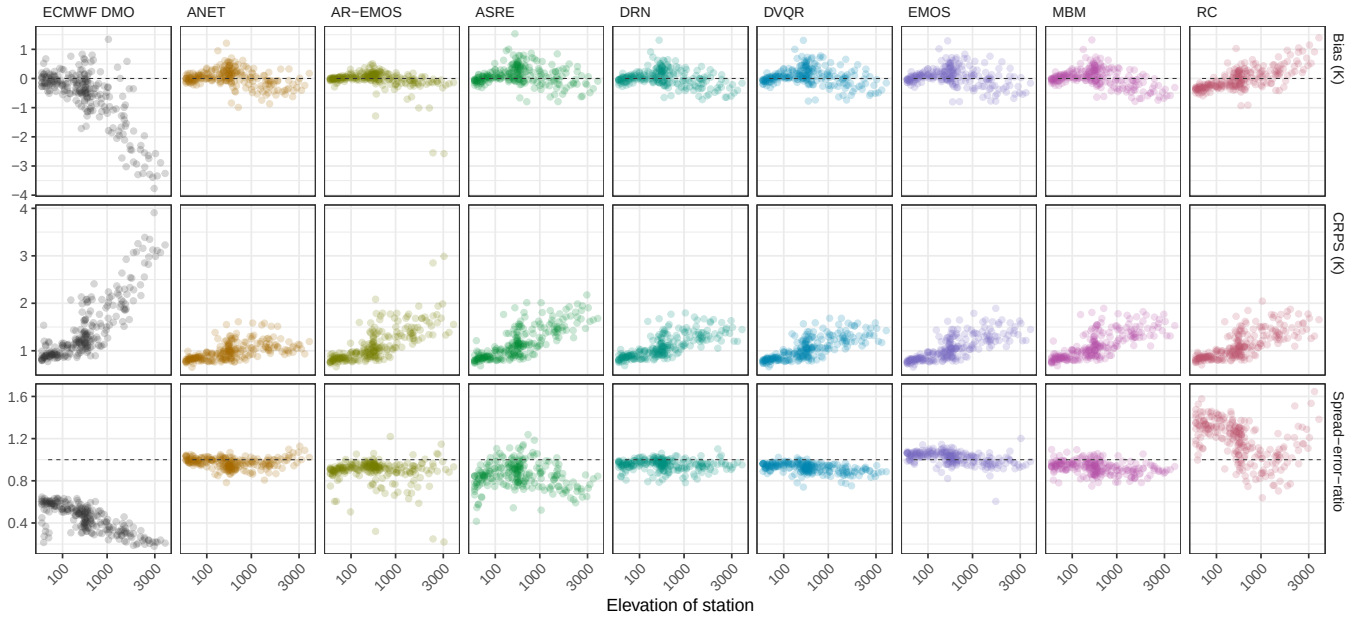

Figure 6. Average scores ordered by station ~~in dependence of station~~ elevation, the elevation-corrected ECMWF DMO is shown alongside the results from the postprocessing methods submitted to the benchmark experiment. Aggregated across all forecasts and all lead times.

- As much as possible, avoid ~~to maintain~~ maintaining an archive or a database of scores for the experiments. Instead compute verification results for the experiments on the fly and only store the summary results. This has the advantage that you can easily add (or remove) scores, summaries of scores, without going through the complex process necessary to update an archive (with new submissions and additional scores).

- Either be very strict about the format of submitted predictions, or use software that is aware of the NetCDF data model and that can handle slight inconsistencies (e.g. re-ordering of dimensions or dimension values)

- Quality control is imperative: while the verification results generally quickly indicate whether there are any major issues with the submitted predictions, issues may already arise earlier than that (making ~~a~~ verification impossible). Catching these errors and establishing a feedback loop with the submitters is important. One way to solve this with NetCDF format is to check the NetCDF header of the submission for format compliance.

These points are important for the next EUPP projects, which will aim to harness the full potential of this dataset, by postprocessing other, less predictable variables (e.g. rainfall, radiation), on station and gridded data, and by allowing many predictors instead of only the target variable itself, ~~as it is the case in the setup of the methods presented here~~which can be expected to yield substantial improvements in predictive performance, in particular for the more advanced machine learning approaches (Rasp and Lerch, 2018). By considering broader aspects of forecasting (e.g. spatial and temporal aspects) as well as more specific scores, the verification task for these forthcoming studies will allow us to use more advanced and cutting-edge

concepts in the field. The lessons learned from these experiments will also be valuable to other groups engaging with the design and operation of such benchmarking experiments. Ultimately, one of the long-term ~~goal of the EUPP module~~ goals of the current benchmark is to provide an automated procedure to upload and compare new approaches to the existing pool of methods available. It is an ambitious goal, with many challenges ahead, but the benefits it will bring make it worth pursuing.

7 Code and data availability

The most straightforward way to access the dataset is through the climetlab EUMETNET postprocessing benchmark plugin at https://github.com/EUPP-benchmark/climetlab-eumetnet-postprocessing-benchmark. This plugin provides easy access to the dataset stored on the ECMWF European Weather Cloud. An example on how to use the plugin is documented in the Supplementary Information, along with other unofficial ways to access the data.

In addition, the dataset has been uploaded in Zarr format on Zenodo for long-term storage. See Demaeyer (2022b) for the gridded data and Bhend et al. (2023) for the station data.

However, the Switzerland station data which are part of the dataset are not presently freely available. These station data may be obtained from IDAWEB (https://gate.meteoswiss.ch/idaweb/) at MeteoSwiss and we are not entitled to provide it online. Registration with IDAWEB can be initiated here: https://gate.meteoswiss.ch/idaweb/prepareRegistration.do. For more information, please also read https://gate.meteoswiss.ch/idaweb/more.do?language=en.

The documentation of the dataset is available at https://eupp-benchmark.github.io/EUPPBench-doc/files/EUPPBench_datasets.html and is also provided in the supplementary information.

The code and scripts used to perform the benchmark are available on GitHub and have been centralized in a single repository: https://github.com/EUPP-benchmark/ESSD-benchmark-codes. This repository contains links to the scripts sub-repositories along with a detailed description of each method. In addition, these codes have been also uploaded to Zenodo: verification code (Primo-Ramos et al., 2023), MBM method (Demaeyer, 2023), Reliability Calibration method (Evans and Hooper, 2023), ASRE method (Ben Bouallègue, 2023), EMOS method (Dabernig, 2023), AR-EMOS method (Möller, 2023), DRN method (Chen et al., 2023b), DVQR method (Jobst, 2023), ANET method (Mlakar et al., 2023b).

Finally, to allow further studies and a better reproducibility, the output data (the corrected forecasts) provided by each methods have also been uploaded to Zenodo (See Chen et al. (2023a)).

Appendix A: Significance assessment

To assess the significance of the CRPS differences observed between each pair of postprocessing methods in the benchmark results, we compute the percentage of station and lead time combinations for which a standard t test of the null hypothesis of equal predictive performance indicates a significant difference at a level of 5%. The p-values of these tests have been adjusted for multiple testing by controlling the false discovery rate using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) . The results are shown in Figure A1, where each cell in the table shows for what percentage of the station and lead time

combinations the method denoted in the row performs significantly better than the method denoted in the column. From this, additional conclusions can be drawn. For instance, all the methods produce a large fraction of significantly better forecasts (i.e. with a lower CRPS) than the ECMWF DMO, while ANET, EMOS, DVQR and AR-EMOS outperform the other methods.

	ANET	AR-EMOS	ASRE	DRN	RC	MBM	DVQR	EMOS	ECMWF DMO
ANET	NA	49.4	74.3	69.7	65.4	62.1	38.0	42.1	89.1
AR-EMOS	7.6	NA	76.0	46.6	39.1	36.1	7.6	7.7	91.9
ASRE	0.9	6.1	NA	13.4	10.5	4.9	1.1	0.3	89.7
DRN	0.9	21.4	52.9	NA	36.1	38.8	3.4	5.9	84.1
RC	1.4	11.9	44.6	26.8	NA	15.9	0.5	0.3	87.2
MBM	3.4	14.3	59.5	32.5	31.5	NA	1.9	0.8	92.6
DVQR	7.8	42.7	75.5	65.3	67.8	58.8	NA	22.1	91.0
EMOS	14.2	44.1	84.9	66.4	81.1	73.4	7.5	NA	94.9
ECMWF DMO	0.2	0.5	1.2	2.1	1.2	2.1	0.5	0.2	NA

Figure A1. Percentage of station and lead time combinations for which the forecast denoted in the row performs significantly (at 5% level) better in terms of the CRPS than the forecast denoted in the column. The p-values have been adjusted for multiple testing using the Benjamini-Hochberg correction.

Author contributions. **Jonathan Demaeyer** led the overall coordination of the benchmark, as well as the collection and dissemination of the dataset. He further contributed to the verification setup, implemented the MBM method (Section 4.3), and coordinated the writing of the manuscript.

Jieyu Chen and Nina Horat implemented the DRN method (Section 4.7).

Sebastian Lerch coordinated the writing of Section 4, and contributed to the implementation of DRN (Section 4.7).

Annette Möller implemented the AR-EMOS method (Section 4.5).

Cristina Primo coordinated the verification work and the writing of Section 3.2.

Gavin Evans and Ben Hooper utilized the IMPROVER codebase to implement the Reliability Calibration method, along with bias correction approaches (Section 4.2).

Markus Dabernig participated in the collection of the data, and provided EMOS corrected forecasts (Section 4.4).

Bert Van Schaeybroeck contributed to the verification results (Section 3.2).

David Jobst implemented the DVQR method (Section 4.6).

Aitor Atencia contributed to the verification work.

Jonas Bhend participated in the collection of the data, contributed to the verification work by producing Figs. 3-6 and drafted the results Section 5.

Zied Ben Bouallègue implemented the ASRE postprocessing method (Section 4.1).

Stéphane Vannitsem leads the PP module of EUMETNET within which the benchmark has been conceptualized.

Peter Mlakar developed and implemented the ANET method (Section 4.8).

Janko Merše and Jana Faganeli Pucer contributed to the development and implementation of the ANET method (Section 4.8).

Olivier Mestre and Maxime Taillardat participated in the collection of the data.

All the authors participated to the writing and the review of the manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. Jonathan Demaeyer thanks Florian Pinault and Baudoin Raoult from ECMWF for their help on the setup of the climetlab plugin, and Francesco Ragone, Lesley De Cruz and David Docquier from RMIB for their support to gather the gridded ~~the~~ data. He also thanks Veerle De Bock and Joffrey Schmitz for their help with the RMIB data. The authors thank Tom Hamill for his guidance on the selection of variables during the dataset design phase. ~~The EUMETNET PP module~~ They also thank Thomas Muschinski and Reto Stauffer for raising an important issue about the station data which has since been corrected.

This benchmark activity organized within the postprocessing module of the 2019-2023 EUMETNET program phase is largely supported financially by its members, constituted by a large number of European National Meteorological and Hydrological Services. The EUMETNET Postprocessing module also thanks the ECMWF for its support on the European Weather Cloud. Jieyu Chen, Nina Horat, and Sebastian Lerch gratefully acknowledge support by the Vector Stiftung through the Young Investigator Group "Artificial Intelligence for Probabilistic Weather Forecasting". Annette Möller acknowledges support by Deutsche Forschungsgemeinschaft (DFG) Grant Number MO 3394/1-1, by the Hungarian National Research, Development and Innovation Office under Grant Number NN125679, and by the Helmholtz Association's pilot project "Uncertainty Quantification". Finally, the authors would like to thank the 3 anonymous reviewers who helped us to increase the value of this manuscript.

References

Ashkboos, S., Huang, L., Dryden, N., Ben-Nun, T., Dueben, P., Gianinazzi, L., Kummer, L., and Hoefler, T.: Ens-10: A dataset for post-processing ensemble weather forecast, arXiv preprint arXiv:2206.14786, 2022.

Ben Bouallègue, Z.: Accounting for representativeness in the verification of ensemble forecasts, ECMWF Technical Memorandum, 865, 2020.

Ben Bouallègue, Z.: EUPP-benchmark/ESSD-ASRE: version 1.0 release, https://doi.org/10.5281/zenodo.7477735, 2023.

Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society. Series B (Methodological), 57, 289–300, http://www.jstor.org/stable/2346101, 1995.

Bhend, J., Dabernig, M., Demaeyer, J., Mestre, O., and Taillardat, M.: EUPPBench postprocessing benchmark dataset - station data, https://doi.org/10.5281/zenodo.7708362, 2023.

Bremnes, J. B.: Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials, Monthly Weather Review, 148, 403–414, 2020.

Chapman, W. E., Monache, L. D., Alessandrini, S., Subramanian, A. C., Ralph, F. M., Xie, S.-P., Lerch, S., and Hayatbini, N.: Probabilistic Predictions from Deterministic Atmospheric River Forecasts with Deep Learning, Monthly Weather Review, 150, 215–234, https://doi.org/10.1175/MWR-D-21-0106.1, 2022.

Chen, J., Janke, T., Steinke, F., and Lerch, S.: Generative machine learning methods for multivariate ensemble post-processing, arXiv preprint arXiv:2211.01345, 2022.

Chen, J., Dabernig, M., Demaeyer, J., Evans, G., Faganeli Pucer, J., Hooper, B., Horat, N., Jobst, D., Lerch, S., Mlakar, P., Möller, A., Merše, J., and Bouallègue, Z. B.: ESSD benchmark output data, https://doi.org/10.5281/zenodo.7798350, 2023a.

Chen, J., Horat, N., and Lerch, S.: EUPP-benchmark/ESSD-DRN: version 1.0 release, https://doi.org/10.5281/zenodo.7477698, 2023b.

Copernicus Land Monitoring Service, E. U.: EU-DEM, https://land.copernicus.eu/imagery-in-situ/eu-dem, European Environment Agency, CLC.

Copernicus Land Monitoring Service, E. U.: CORINE Land Cover, https://land.copernicus.eu/pan-european/corine-land-cover, European Environment Agency, CLC, 2018.

Dabernig, M.: EUPP-benchmark/ESSD-EMOS: version 1.0 release, https://doi.org/10.5281/zenodo.7477749, 2023.

Dabernig, M., Mayr, G. J., Messner, J. W., and Zeileis, A.: Spatial Ensemble Post-Processing with Standardized Anomalies, Quarterly Journal of the Royal Meteorological Society, 143, 909–916, https://doi.org/https://doi.org/10.1002/qj.2975, 2017.

Demaeyer, J.: Climdyn/pythie: Version 0.1.0 alpha release, https://doi.org/10.5281/zenodo.7233538, 2022a.

Demaeyer, J.: EUPPBench postprocessing benchmark dataset - gridded data - Part I, https://doi.org/10.5281/zenodo.7429236, 2022b.

Demaeyer, J.: EUPP-benchmark/ESSD-mbm: version 1.0 release, https://doi.org/10.5281/zenodo.7476673, 2023.

Demaeyer, J. and Vannitsem, S.: Correcting for model changes in statistical postprocessing – an approach based on response theory, Nonlinear Processes in Geophysics, 27, 307–327, https://doi.org/10.5194/npg-27-307-2020, 2020.

Dueben, P. D., Schultz, M. G., Chantry, M., Gagne, D. J., Hall, D. M., and McGovern, A.: Challenges and Benchmark Datasets for Machine Learning in the Atmospheric Sciences: Definition, Status, and Outlook, Artificial Intelligence for the Earth Systems, 1, e210 002, https://doi.org/10.1175/AIES-D-21-0002.1, 2022.

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H., et al.: NetCDF Climate and Forecast (CF) metadata conventions, http://cfconventions.org/Data/cf-conventions/cf-conventions-1.8/cf-conventions.pdf, 2003.

Evans, G. and Hooper, B.: EUPP-benchmark/ESSD-reliability-calibration: version 1.0 release, https://doi.org/10.5281/zenodo.7476590, 2023.

Flowerdew, J.: Calibrating ensemble reliability whilst preserving spatial structure, Tellus A: Dynamic Meteorology and Oceanography, 66, 22 662, https://doi.org/10.3402/tellusa.v66.22662, 2014.

Garg, S., Rasp, S., and Thuerey, N.: WeatherBench Probability: A benchmark dataset for probabilistic medium-range weather forecasting along with deep learning baseline models, arXiv preprint arXiv:2205.00865, 2022.

Glahn, H. R. and Lowry, D. A.: The use of model output statistics (MOS) in objective weather forecasting, Journal of Applied Meteorology and Climatology, 11, 1203–1211, 1972.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Monthly Weather Review, 133, 1098–1118, 2005.

Gregory, J.: The CF metadata standard, CLIVAR Exchanges, 8, 4, 2003.

Groß, J. and Möller, A.: ensAR: Autoregressive postprocessing methods for ensemble forecasts, https://github.com/JuGross/ensAR, R package version 0.2.0, 2019.

Hafizi, H. and Sorman, A. A.: Assessment of 13 Gridded Precipitation Datasets for Hydrological Modeling in a Mountainous Basin, Atmosphere, 13, https://doi.org/10.3390/atmos13010143, 2022.

Hagedorn, R., Hamill, T. M., and Whitaker, J. S.: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures, Monthly Weather Review, 136, 2608 – 2619, https://doi.org/10.1175/2007MWR2410.1, 2008.

Hamill, T. M., Hagedorn, R., and Whitaker, J. S.: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation, Monthly Weather Review, 136, 2620 – 2632, https://doi.org/10.1175/2007MWR2411.1, 2008.

Han, J., Miao, C., Gou, J., Zheng, H., Zhang, Q., and Guo, X.: A new daily gridded precipitation dataset based on gauge observations across mainland China, Earth System Science Data Discussions, 2022, 1–33, https://doi.org/10.5194/essd-2022-373, 2022.

Haupt, S. E., Chapman, W., Adams, S. V., Kirkwood, C., Hosking, J. S., Robinson, N. H., Lerch, S., and Subramanian, A. C.: Towards implementing artificial intelligence post-processing in weather and climate: proposed actions from the Oxford 2019 workshop, Philosophical Transactions of the Royal Society A, 379, 20200 091, 2021.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather and Forecasting, 15, 559–570, 2000.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Horton, P.: Analogue methods and ERA5: Benefits and pitfalls, International Journal of Climatology, 42, 4078–4096, https://doi.org/https://doi.org/10.1002/joc.7484, 2022.

Hoyer, S. and Joseph, H.: xarray: N-D labeled Arrays and Datasets in Python, Journal of Open Research Software, 5, https://doi.org/10.5334/jors.148, 2017.

Jobst, D.: EUPP-benchmark/ESSD-DVQR: version 1.0 release, https://doi.org/10.5281/zenodo.7477640, 2023.

Kim, T., Ho, N., Kim, D., and Yun, S.-Y.: Benchmark Dataset for Precipitation Forecasting by Post-Processing the Numerical Weather Prediction, arXiv preprint arXiv:2206.15241, 2022.

Klein, W. H. and Lewis, F.: Computer forecasts of maximum and minimum temperatures, Journal of Applied Meteorology (1962-1982), pp. 350–359, 1970.

Klein, W. H., Lewis, B. M., and Enger, I.: Objective prediction of five-day mean temperatures during winter, Journal of Atmospheric Sciences, 16, 672–682, 1959.

Kraus, D. and Czado, C.: D-vine copula based quantile regression, Computational Statistics & Data Analysis, 110, 1–18, https://doi.org/10.1016/j.csda.2016.12.009, 2017.

Lakatos, M., Lerch, S., Hemri, S., and Baran, S.: Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts, Quarterly Journal of the Royal Meteorological Society, https://doi.org/10.1002/qj.4436, 2023.

Lalaurette, F.: Early detection of abnormal weather conditions using a probabilistic extreme forecast index, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 129, 3037–3057, 2003.

Lang, M. N., Lerch, S., Mayr, G. J., Simon, T., Stauffer, R., and Zeileis, A.: Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression, Nonlinear Processes in Geophysics, 27, 23–34, https://doi.org/10.5194/npg-27-23-2020, 2020.

Lenkoski, A., Kolstad, E. W., and Thorarinsdottir, T. L.: A Benchmarking Dataset for Seasonal Weather Forecasts, NR-notat, https://nr.brage.unit.no/nr-xmlui/bitstream/handle/11250/2976154/manual.pdf, 2022.

Lerch, S., Baran, S., Möller, A., Groß, J., Schefzik, R., Hemri, S., and Graeter, M.: Simulation-based comparison of multivariate ensemble post-processing methods, Nonlinear Processes in Geophysics, 27, 349–371, https://doi.org/10.5194/npg-27-349-2020, 2020.

Maciel, P., Quintino, T., Modigliani, U., Dando, P., Raoult, B., Deconinck, W., Rathgeber, F., and Simarro, C.: The new ECMWF interpolation package MIR, pp. 36–39, https://doi.org/10.21957/h20rz8, 2017.

Messner, J. W., Mayr, G. J., and Zeileis, A.: Heteroscedastic Censored and Truncated Regression with crch, The R Journal, 8, 173–181, https://journal.r-project.org/archive/2016-1/messner-mayr-zeileis.pdf, 2016.

Miles, A., Kirkham, J., Durant, M., Bourbeau, J., Onalan, T., Hamman, J., Patel, Z., shikharsg, Rocklin, M., raphael dussin, Schut, V., de Andrade, E. S., Abernathey, R., Noyes, C., sbalmer, pyup.io bot, Tran, T., Saalfeld, S., Swaney, J., Moore, J., Jevnik, J., Kelleher, J., Funke, J., Sakkis, G., Barnes, C., and Banihirwe, A.: zarr-developers/zarr-python: v2.4.0, https://doi.org/10.5281/zenodo.3773450, 2020.

Mlakar, P., Merše, J., and Pucer, J. F.: Ensemble weather forecast post-processing with a flexible probabilistic neural network approach, https://doi.org/10.48550/arXiv.2303.17610, 2023a.

Mlakar, P., Pucer, J. F., and Merše, J.: EUPP-benchmark/ESSD-ANET: version 1.0 release, https://doi.org/10.5281/zenodo.7479333, 2023b.

Möller, A. and Groß, J.: Probabilistic temperature forecasting based on an ensemble autoregressive modification, Quarterly Journal of the Royal Meteorological Society, 142, 1385–1394, https://doi.org/https://doi.org/10.1002/qj.2741, 2016.

Möller, A. and Groß, J.: Probabilistic Temperature Forecasting with a Heteroscedastic Autoregressive Ensemble Postprocessing model, Quarterly Journal of the Royal Meteorological Society, 146, 211–224, https://doi.org/https://doi.org/10.1002/qj.3667, 2020.

Möller, A., Spazzini, L., Kraus, D., Nagler, T., and Czado, C.: Vine copula based post-processing of ensemble forecasts for temperature, 2018.

Murphy, A. H.: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, Weather and Forecasting, 8, 281–293, https://doi.org/https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2, 1993.

Möller, A.: EUPP-benchmark/ESSD-AR-EMOS: version 1.0 release, https://doi.org/10.5281/zenodo.7477633, 2023.

Nagler, T.: vinereg: D-Vine Quantile Regression, https://CRAN.R-project.org/package=vinereg, r package version 0.7.2, 2020.

Perrone, E., Schicker, I., and Lang, M. N.: A case study of empirical copula methods for the statistical correction of forecasts of the ALADIN-LAEF system, Meteorologische Zeitschrift, 29, 277–288, https://doi.org/10.1127/metz/2020/1034, 2020.

Primo-Ramos, C., Bhend, J., Atencia, A., Van schaeybroeck, B., and Demaeyer, J.: EUPP-benchmark/ESSD-Verification: version 1.0 release, https://doi.org/10.5281/zenodo.7484371, 2023.

Rabault, J., Müller, M., Voermans, J., Brazhnikov, D., Turnbull, I., Marchenko, A., Biuw, M., Nose, T., Waseda, T., Johansson, M., et al.: A dataset of direct observations of sea ice drift and waves in ice, arXiv preprint arXiv:2211.03565, 2022.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Monthly weather review, 133, 1155–1174, 2005.

Rasp, S. and Lerch, S.: Neural networks for postprocessing ensemble weather forecasts, Monthly Weather Review, 146, 3885–3900, 2018.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 203, https://doi.org/https://doi.org/10.1029/2020MS002203, e2020MS002203 10.1029/2020MS002203, 2020.

Roberts, N., Ayliffe, B., Evans, G., Moseley, S., Rust, F., Sandford, C., Trzeciak, T., Abernethy, P., Beard, L., Crosswaite, N., Fitzpatrick, B., Flowerdew, J., Gale, T., Holly, L., Hopkinson, A., Hurst, K., Jackson, S., Jones, C., Mylne, K., Sampson, C., Sharpe, M., Wright, B., Backhouse, S., Baker, M., Brierley, D., Booton, A., Bysouth, C., Coulson, R., Coultas, S., Crocker, R., Harbord, R., Howard, K., Hughes, T., Mittermaier, M., Petch, J., Pillinger, T., Smart, V., Smith, E., and Worsfold, M.: IMPROVER: the new probabilistic post processing system at the UK Met Office, Bulletin of the American Meteorological Society, 2022.

Schulz, B. and Lerch, S.: Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison, Monthly Weather Review, 150, 235–257, 2022.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P.: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics, Monthly Weather Review, 144, 2375–2393, 2016.

Van Schaeybroeck, B. and Vannitsem, S.: Ensemble post-processing using member-by-member approaches: theoretical aspects, Quarterly Journal of the Royal Meteorological Society, 141, 807–818, https://doi.org/https://doi.org/10.1002/qj.2397, 2015.

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

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., and Ylhaisi, J.: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World, Bulletin of the American Meteorological Society, 102, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1, 2021.

Wang, W., Yang, D., Hong, T., and Kleissl, J.: An archived dataset from the ECMWF Ensemble Prediction System for probabilistic solar power forecasting, Solar Energy, 248, 64–75, https://doi.org/https://doi.org/10.1016/j.solener.2022.10.062, 2022.

Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., Stier, P., Bouabid, S., Dewey, M., Fons, E., Gonzalez, J., Harder, P., Jeggle, K., Lenhardt, J., Manshausen, P., Novitasari, M., Ricard, L., and Roesch, C.: ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections, Journal of Advances in Modeling Earth Systems, 14, e2021MS002 954, https://doi.org/https://doi.org/10.1029/2021MS002954, 2022.

Wilks, D. S.: Multivariate ensemble Model Output Statistics using empirical copulas, Quarterly Journal of the Royal Meteorological Society, 141, 945–952, https://doi.org/https://doi.org/10.1002/qj.2414, 2015.

Zandler, H., Haag, I., and Samimi, C.: Evaluation needs and temporal performance differences of gridded precipitation products in peripheral mountain regions, Scientific reports, 9, 1–15, 2019.

Zantedeschi, V., Falasca, F., Douglas, A., Strange, R., Kusner, M. J., and Watson-Parris, D.: Cumulo: A dataset for learning cloud classes, arXiv preprint arXiv:1911.04227, 2019.

Zsótér, E.: Recent developments in extreme weather forecasting, ECMWF newsletter, 107, 8–17, 2006.