Politecnico di Torino

Master’s degree Environmental and Land Engineering
Climate Change track
A.y. 2025/2026
Degree Session July 2026

Weather forecast assessment at the daily time scale in northwestern Italy

Supervisors:
Prof. Stefania Tamea
Prof. Alberto Viglione

Candidate:
Chiara Rondana

A mio nonno Silvano,

che mi ha insegnato l’ingegneria senza formule,

senza nemmeno saperlo.

Abstract

This Master's thesis analyses the forecasts of meteorological variables, precipitation and temperature, in the northwestern Italian regions of Piedmont and Aosta Valley, with the aim of evaluating the accuracy along the year through a statistical analysis of forecast errors.

In the preliminary phase of the analysis a systematic classification was performed of available forecast and observational datasets, based on criteria such as lead time, spatial and temporal resolution and extent, and data accessibility. This step allows the identification of most suitable datasets for the study purposes, that are: the forecast dataset derived from the ICON-2I numerical model, including its 20 ensemble members, made available through the MeteoHub platform by Meteo Italia, and the observational dataset NWIOI provided by ARPA Piemonte.

For comparison purposes, both deterministic and ensemble forecast datasets are aligned with the observational dataset through regridding and temporal aggregation procedures. In this context, model performance is evaluated using statistical indicators of accuracy, bias, and dispersion, aimed at providing a comprehensive description of forecast quality. The analysis also includes the quantification of performance dependence on the main variability factors, namely lead time, seasonality, and orographic complexity of the territory. Verification tools for probabilistic forecasts of continuous hydrological variables are also adopted to further assess ensemble simulations.

The results indicate that temperature predictions exhibit a lower dependence on seasonality compared to precipitation forecasts, which are instead more strongly influenced by the time of year and show greater difficulty in accurately capturing both event intensity and spatial localization. Regarding lead time, temperature forecasts remain relatively stable, showing similar predictive behaviour at both short and long lead times, whereas precipitation forecasts demonstrate an improvement in performance as the lead time decreases. Furthermore, ensemble probabilistic forecasts present distinct systematic behaviours depending on the variable: precipitation ensemble forecasts are underdispersed, reflecting an underestimation of uncertainty, whereas temperature ensemble forecasts show a persistent negative bias in the ensemble mean, indicating systematic underprediction.

This study provides a quantitative assessment of the reliability of meteorological forecasts in the investigated area, supporting their potential use in decision-support tools tailored to agriculture and water management, contributing to improved irrigation efficiency and water resources planning.

Summary

Introduction

Nowadays, climate change is an increasingly significant driver of environmental and socio-economic impacts, with water resources representing one of the most affected components of natural systems. In the agricultural sector, this translates into increasing pressure on water availability and management, highlighting the importance of weather forecasts as valuable sources of information to support decision-making processes, particularly in irrigation management.

The agricultural sector is particularly affected by these changes due to its strong dependence on water resources. In this context, improving irrigation management is essential to promote sustainable water use, and reliable meteorological forecasts can provide valuable information to support decision-making processes and optimize irrigation planning.

Numerical weather prediction models have become fundamental tools for providing weather forecasts at different spatial and temporal scales. Nevertheless, forecast accuracy depends on several factors, including lead time, seasonality, terrain complexity, and the meteorological variable considered. While temperature forecasts generally show higher predictability, precipitation remains more challenging to forecast, particularly over mountainous regions. To better represent forecast uncertainty, ensemble prediction systems complement deterministic forecasts by providing probabilistic information, making forecast verification an essential step in assessing their reliability for operational applications.

In this framework, this master's thesis evaluates the performance of deterministic and ensemble forecasts of precipitation and temperature over the northwestern Italian regions of Piedmont and Aosta Valley. The analysis investigates forecast reliability through statistical verification methods, examining the influence of lead time, seasonality, and orographic complexity on forecast performance, as a possible future development for the use of meteorological forecasts as decision-support tools for agricultural water management.

The thesis is organized as follows. Chapter 1 presents the context of the available datasets. Chapter 2 describes the methodology used in the analysis. Chapter 3 presents the deterministic and probabilistic results of the forecast verification. Finally, Chapter 4 discusses the main findings and draws the conclusions of the study.

1. Datasets

This thesis focuses on the analysis of meteorological forecast and observational datasets, with particular attention to temperature and precipitation variables as the primary elements describing the state and evolution of atmospheric conditions within a weather forecasting context. These variables constitute the core information used to evaluate forecast skill and to compare model outputs against observed atmospheric behaviour.

To begin, it is important to frame the two types of datasets involved, highlighting their role within the forecasting chain, as well as the broader context of data availability.

1.1. Weather observational datasets

The first type of weather dataset considered in this thesis is the observational dataset, defined as a structured collection of measured atmospheric data recorded by weather stations, in situ sensors, or remote sensing systems (Dati osservati, s.d.). It represents the actual state of the atmosphere at specific times and locations and typically includes variables such as temperature, precipitation, humidity, wind speed and direction, and atmospheric pressure. In this context, observational data are assumed to provide the closest available representation of the true atmospheric state, although they remain affected by measurement errors, sampling limitations, and instrument-specific uncertainties.

Several observational products are available, differing in spatial and temporal resolution, measurement techniques, and spatial coverage. These differences are crucial when selecting a dataset for a specific application, as no single product simultaneously provides optimal performance at both high spatial and temporal scales. In practice, dataset selection requires a compromise between spatial representativeness and temporal continuity, depending on the variable of interest and the characteristic spatial and temporal variability of atmospheric processes within the study domain. Such considerations are particularly relevant in heterogeneous environments, where topography and local climatic effects can strongly influence the representativeness of observations.

Observational data are widely used in climate and environmental studies, particularly for monitoring climate variability and validating numerical weather prediction models. They are essential for applications such as trend analysis and environmental monitoring, and in this thesis, they provide a reference dataset for forecast validation. For this reason, they constitute the benchmark against which both deterministic and ensemble forecast performance is assessed throughout the analysis.

1.1.1. Overview of available observational datasets

To provide a structured overview of the available products, Table 1 summarizes the main observational datasets considered in this study, together with their key characteristics and data accessibility.

In general, these datasets span a range of spatial scales, from station-based point observations to gridded products with resolutions ranging approximately from 24 km to 1 km, and temporal resolutions from monthly to sub-daily as hourly or 3-hourly data. Their spatial coverage is highly heterogeneous, ranging from global coverage to continental-scale domains such as Europe, and further down to sub-national and regional scales. This diversity reflects the different methodologies used in data production, including interpolation techniques and reanalysis systems.

From a spatial resolution perspective, the datasets exhibit significant variability. Among the products considered, EMO-1 and CERRA represent the highest spatial resolution datasets over the European domain, with grid spacings of approximately 1.8 km and 5.5 km respectively. These datasets are particularly suitable for analyses requiring a fine representation of spatial variability, especially at regional to sub-regional scales. In addition, CERRA provides a relatively high temporal resolution, 3-

hourly, whereas EMO-1 is available at daily temporal resolution, making CERRA more suitable for applications requiring a more detailed representation of temporal variability.

From a temporal perspective, ERA5 stands out as the most comprehensive dataset. Although its spatial resolution of 0.25°, approximately 24 km, is coarser than other products, it provides hourly temporal resolution and a long and consistent global coverage starting from 1940. This makes it particularly suitable for time-resolved analyses and studies requiring long-term temporal consistency.

Table 1. Main weather observational datasets for the study area and their characteristics.

WEATHER OBSERVATIONAL DATASETS								
Name	Source	Spatial resolution (lat. 45°)	Temporal resolution	Spatial extent	Temporal extent	Update frequency	Accessibility	Data access provider
NWIOI (North-Western Italy Optimal Interpolation) (+ specific dataset for precipitation)	Piedmont's ARPA	0.125° ≈ 12 km (0.01° ≈ 1 km)	Daily	Piedmont and Aosta Valley	From December 1957	Each day after 18:00 local time	Open access under CC BY 2.5 Italy (Open access under request)	Piedmont's ARPA
E-OBS	KNMI	0.1° ≈ 10 km and 0.25° ≈ 24 km	Daily	Europe (land only)	From January 1950	Every 6 months with a latency of 3 months	Open access under E-OBS product licence	Copernicus Climate Data Store (CDS)
ERA5 (European Reanalysis)	ECMWF	0.25° ≈ 24 km (atmosphere)	Hourly and monthly	Global	From January 1940	Daily with a latency of about 5 days	Open access under the CC-BY licence	Copernicus Climate Data Store (CDS)
• EMO-5 • EMO-1 (European Meteorological Observation)	Copernicus Emergency Management Service (CEMS)	• 0.05° ≈ 5 km (only land) • 0.02° ≈ 1.82 km	6-hourly and daily	Europe	• 1990 – 2019 • 1990 - 2023	• Not supported anymore • Annually	Open access • under CC-BY 4.0 licence • under European Commission Reuse and Copyright Notice	Joint Research Centre (JRC) Data Catalogue
CERRA	Copernicus Climate Data Store (CDS)	0.06° ≈ 5.5 km	3-hourly	Europe	From 1984	Monthly	CC-BY licence	Copernicus Climate Data Store (CDS)
WWR (World Weather Records)	NOAA / NCEI (National Centers for Environmental Information)	Point data for off-center stations, not grid	Monthly	Global (Italy is not available)	From 1951 (From January 1755)	Annually	Open access	NOAA / NCEI (National Centers for Environmental Information)

Note: Datasets references are available in the bibliography.

1.1.2. Selected observational dataset

The observational dataset selected for this study is the NWIOI (North-Western Italy Optimal Interpolation) product. NWIOI is a gridded meteorological dataset defined on a regular latitude-longitude grid with a spatial resolution of 0.125°, covering north-western Italy within the domain bounded by 6.5°-9.5° E longitude and 44.0°-46.5° N latitude. It is provided by ARPA Piemonte in NetCDF (Network Common Data Form) format, a widely used standard in geoscientific applications. The dataset is released as open data under the Creative Commons Attribution 2.5 License (CC BY 2.5), allowing its use for research and analysis provided that appropriate attribution is given.

The dataset provides spatially continuous fields of three meteorological variables, namely minimum temperature, maximum temperature, and accumulated precipitation, all available at daily temporal resolution from 1 December 1957 onward. Data are made available daily, typically after 18:00 local time on the day following the observations (Dataset su griglia NWIOI, n.d.).

NWIOI is generated through an optimal interpolation procedure that merges in situ station observations to produce spatially continuous fields. The current observational network used in the interpolation consists of approximately 480 stations, which form the observational basis for the gridded reconstruction. This approach ensures a consistent spatial representation of meteorological variables even in the presence of an irregular or sparse monitoring network, thereby providing spatially homogeneous fields across the domain. This gridded structure is a key advantage for the present study, as it enables a more straightforward and consistent analysis compared to point-based observational data, avoiding additional preprocessing steps related to spatial aggregation and station representativeness.

The figure 1 illustrates a Voronoi tessellation derived from the spatial distribution of temperature stations, shown in red, for the years 1958, 1990, 2000, and 2009, highlighting the temporal evolution of the observational network used in the interpolation procedure for the generation of the NWIOI gridded fields. This evolution ultimately defines the effective domain of validity of the dataset, which is representative of Piedmont and Aosta Valley. In addition, the evolution of the mean station radii over time provides an indication of the effective resolution of the optimal interpolation scheme, progressively converging towards the current nominal resolution of 0.125° (Metodologia interpolazione NWIOI).

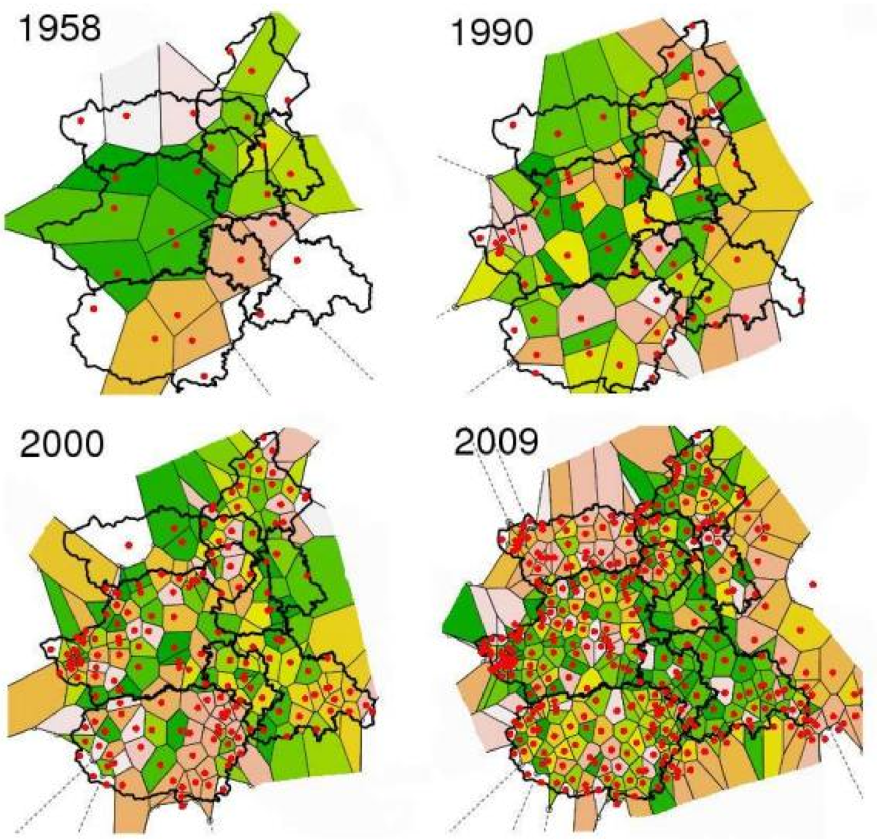

Figure 1. Voronoi tessellation derived from the spatial distribution of temperature stations (in red) for the years 1958, 1990, 2000, and 2009 (Metodologia interpolazione NWIOI).

1.2. Weather forecast datasets

The second type of dataset considered in this thesis is the weather forecast dataset. Weather forecast datasets consist of structured representations of the future state of the atmosphere, generated through numerical weather prediction (NWP) models. These models are based on fundamental physical principles and simulate the evolution of atmospheric processes by numerically solving a coupled set of governing equations describing the conservation of momentum, mass, and energy under the framework of fluid dynamics and thermodynamics (Numerical weather prediction, n.d.).

Starting from an initial atmospheric state derived from observational data through data assimilation procedures, NWP systems integrate the governing equations forward in time on a discretised spatial grid, producing forecasts of key meteorological variables such as temperature, precipitation, wind components, and surface pressure (Numerical weather prediction, n.d.). The quality of the forecast is therefore strongly dependent on both the accuracy of the initial conditions and the ability of the model to represent sub-grid scale physical processes through parameterisation schemes.

Given the chaotic nature of the atmosphere and its sensitivity to initial conditions, small perturbations in the starting state can lead to significantly divergent outcomes over time. For this reason, forecasts are often produced not only in deterministic form, but also in probabilistic ensemble configurations. Ensemble forecasting systems consist of multiple simulations generated from slightly perturbed initial conditions and/or perturbed model physics, allowing for a more complete representation of forecast uncertainty and providing a quantitative framework for assessing predictability and confidence in the predicted atmospheric state.

1.2.1. Overview of available observational datasets

Several operational forecasting systems are available over the study area, ranging from global to European and regional-scale models. These systems differ in terms of spatial resolution, temporal resolution, forecast horizon, ensemble configuration, and update frequency reflecting the different design purposes of global and limited-area numerical weather prediction systems.

An overview of the main operational weather forecasts over the study area, together with their main characteristics, is provided in Table 2.

In terms of spatial coverage, global-scale models, such as ECMWF-HRES and GFS, prioritize forecast range and spatial coverage over fine spatial detail. With resolutions ranging from about 0.1° to 0.3°, approximately from 10 to 28 km, they provide global forecasts extending up to 16 days ahead and are primarily designed to represent large-scale atmospheric processes. While their ability to resolve local-scale features is limited, they offer reliable medium-range predictions, frequent updates, and comprehensive spatial coverage. In contrast, regional and convection-permitting models such as ICON-2I, HARMONIE-AROME, and MOLOCH operate at significantly higher spatial resolutions, reaching values close to 2-3 km or even finer in the case of MOLOCH. These models are specifically designed to better resolve local atmospheric processes, including orographic effects and convective precipitation, which are particularly relevant in complex terrain such as northwestern Italy. However, this increase in spatial detail comes at the expense of a reduced forecast horizon, typically limited to 48-72 hours, that correspond to 2-3 days.

Closely related to the modelling scale is also the configuration of ensemble systems, which is strongly constrained by computational limitations. Ensemble sizes differ between modelling systems: global systems such as ECMWF provide relatively large ensembles up to 51 members, whereas high-resolution regional models such as HARMONIE-AROME Europe and ICON-2I typically operate with smaller number of members like 16 and 20 respectively.

The temporal resolution also varies consistently across datasets. While global models generally provide outputs at 3-6 hours intervals, high-resolution limited-area models offer hourly forecasts, which are more suitable for hydrological applications and for capturing fast-evolving meteorological phenomena.

Overall, Table 2 illustrates an unavoidable trade-off between spatial resolution, temporal resolution, forecast horizon, and computational constraints, which ultimately shape the design of numerical weather prediction systems across different modelling scales.

Table 2. Main weather forecast datasets for the study area and their characteristics.

WEATHER FORECAST DATASET											
Name	Provider	Ensemble	Spatial resolution (lat. 45°)	Temporal resolution	Forecast Horizon	Spatial extent	Forecast Frequency	Historical data availability	Update available/ delay	Accessibility	Data access provider
HRES	ECMWF	ENS 51 members	0.1° ≈ 10 km	3-hourly up to 6 days and 6-hourly up to 15 days	16 days	Global	4 times daily	August 2023 - present	1 hour	Open access for 0.25° spatial resolution under the CC-BY 4.0 Licence	ECMWF
ICON-2I (ICON-EU)	ItaliaMeteo ARPAE (DWD)	ICON-EPS 20 members	0.02° ≈ 2.2 km (0.07° ≈ 7 km)	1-hourly (1-hourly up to 78h and 3-hourly up to 120h)	3 days (5 days)	Italy (Europe)	2 times daily (4 times daily)	June 2025 - present	Every 12 hours (every 3 hours)	Open access under CC BY 4.0 Licence	MeteoHub portal
HARMONIE-AROME Europe	Météo-France	HarmonEPS 16 members	0.03 ≈ 2.5 km	Every 6 hours	-	Europe	4 times daily	They only retain 72 hours	Every 1 hours	Open access under CC-BY 4.0 Licence	KNMI data platform
NEMSGLOBAL (NEMS12)	MeteoBlue	No	0.3° ≈ 28 km (0.125° ≈ 12 km)	Hourly	7 days	Global (Europe)	Daily	1984 - present	< 24 hours	Paid license - scientific collaborations available	MeteoBlue
ALARO-LAEF (successor of the former ALADIN-LAEF)	ECMWF	A-LAEF 17 members	0.05° ≈ 4.8 km	-	3 days	Central Europe (not the entire Italy is included)	2 times daily	-	-	-	-
GFS	National Centers for Environmental Prediction (NCEP)	GEFS	0.3° ≈ 28 km up to 10 days and 0.7° ≈ 70 km over 10 days	1-hourly up to 120 hours and 3-hourly up to 16 days	16 days	Global	4 times daily	January 2015 - present	Every 6 hours	Open Data with attribution request	NOAA
• MOLOCH • BOLAM • GLOBO	CNR-ISAC	• No • No • 41 members	• 0.075° ≈ 8.3 km • 0.0113° ≈ 1.25 km • 0.2° ≈ 19 km	• 1-hourly • 3-hourly • Daily	• 2 days • 3 days • 7 days	• Italy • Mediterranean-European • Global	Daily	(In maps) • Since 2012 • Since 2012 • Since 2014	Daily	Open access under CC BY-NC-ND 4.0 Licence	CNR-ISAC

Note: Datasets references are available in the bibliography.

1.2.2. Selected forecast dataset

The forecast dataset selected for this study is derived from the high-resolution numerical weather prediction system ICON-2I. ICON is a global model originally developed by the German national meteorological service Deutscher Wetterdienst (DWD), while ICON-2I represents its Italian configuration, specifically adapted for operational use over the Italian domain.

For the analysis two ICON-2I configurations are employed: a deterministic configuration and an ensemble configuration consisting of 20 members. Both configurations provide a comprehensive set of atmospheric variables; however, the analysis is restricted to the variables of interest, namely 2 m air temperature and accumulated precipitation. Both systems operate at a horizontal resolution of 0.02°, corresponding to approximately 2.2 km, over a domain extending from 33°N to 49°N in latitude and from 3°E to 22°E in longitude. Forecasts are available at hourly temporal resolution and are produced up to 72 hours ahead, with the data subsequently aggregated to ensure consistency in the comparison between forecast and observational datasets.

The deterministic system is run twice daily, with model initializations at 00 UTC and 12 UTC. In contrast, the ensemble system is initialized once per day at 21 UTC, providing multiple realizations of the atmospheric state through perturbed initial conditions. In this study, ensemble forecasts are limited to the first local day of the forecast period.

Both deterministic and ensemble outputs are obtained through MeteoHub, the official operational platform for national meteorological data distribution developed by ItaliaMeteo. The dataset is provided in GRIB format, a widely used standard in numerical weather prediction for the efficient storage and exchange of gridded meteorological data. Ath the time of this thesis, the ICON-2I archive available through MeteoHub covers the period starting from June 2025. Consequently, the analysis is based on approximately one year of forecast data, corresponding to the full period currently available from the operational archive.

2. Methods

2.1. Preprocessing workflow

To ensure physical consistency and mutual comparability prior to analysis, both observational and forecast datasets undergo a preliminary preprocessing workflow. The two datasets are handled separately due to their different native formats, NetCDF for observations and GRIB for forecasts, requiring dedicated reading and decoding procedures for each source. The preprocessing is organized by variable, with temperature and precipitation treated independently in order to preserve their distinct physical meanings and to ensure consistency in subsequent analyses. This variable-wise approach also facilitates a more structured alignment of the datasets during the regridding and temporal aggregation steps.

2.1.1. Unit conversion and variable adjustment

The preprocessing workflow begins with a unit conversion and variable adjustment stage.

Starting from temperature variables, values originally expressed in Kelvin are converted into Celsius degrees to ensure physical interpretability across datasets. Within the observational dataset, daily mean temperature is derived as the arithmetic average of daily maximum and minimum values, since only these two components are available, ensuring coherence with the treatment of forecast temperature at later stages.

Regarding the precipitation variable, the model output is provided as accumulated surface mass flux in kg/m^2, a quantity that is equivalent to rainfall depth in mm, under standard hydrological assumptions. In the forecast dataset, accumulated precipitation is provided over the forecast lead time up to 72 hours. Hourly values are obtained by computing successive differences between accumulation steps, and daily totals are derived by temporal aggregation of the resulting hourly series.

2.1.2. Temporal handling

The temporal handling is designed to ensure consistency between observational and forecast datasets in terms of temporal resolution and reference time system.

Observations are available at daily resolution, whereas forecast outputs are provided at hourly resolution. A temporal aggregation of the forecast data is therefore required to align the two datasets and enable a consistent comparison. For precipitation, daily values correspond to the accumulated precipitation over the day, whereas for temperature they are obtained by averaging the hourly forecast values. This aggregation is performed after the forecast data have been organized into hourly-based structures, as described in Section 2.3.

In addition, a temporal alignment is applied to ensure consistency in the reference time system. Forecast outputs are provided in Coordinated Universal Time (UTC), while observations are referenced to local time. Accordingly, all forecast timestamps are converted to local time by applying 1 hour offset during standard time from late October to late March and by applying 2 hours offset during daylight saving time from late March to late October.

2.1.3. Spatial processing

From a spatial perspective, the processing ensures full correspondence between the two datasets in terms of spatial domain and resolution, so that corresponding grid cells represent the same geographical locations.

The spatial resolution of the NWIOI dataset represents the limiting factor of the analysis and is therefore adopted as the reference grid for all subsequent processing steps. The NWIOI grid is defined as a regular grid of 20 × 24 cells, characterized by:

- latitude ranging from 46.4380°N to 44.0630°N
- longitude ranging from 6.5630°E to 9.4380°E
- constant spatial resolution of 0.125°, approximately 12 km.

Since the NWIOI dataset covers a broader north-western Italy domain, a geographic mask is applied to explicitly isolate the regions of interest, namely Piedmont and the Aosta Valley, which define the actual valid domain of the dataset. This operation ensures a precise and unambiguous definition of the valid grid cells to use in the analysis.

Forecast fields are then processed to ensure full spatial consistency with observational reference. The native ICON-2I domain is first restricted through a latitude-longitude cut, applied by selecting only grid points falling within a predefined bounding box defined by the minimum and maximum latitudinal and longitudinal coordinates of the NWIOI domain. The resulting subset is then regridded onto the NWIOI observational grid. These two steps are applied consistently to both temperature and precipitation fields, ensuring that all forecast variables are defined over the same geographical extent prior to further analysis.

The regridding procedure is implemented in MATLAB through a spatial aggregation approach. Given the native and fine spatial resolution of ICON-2I, 0.02° corresponding to about 1.2 arcminutes or about 2.2 km, each NWIOI grid cell contains multiple forecast grid points, typically on the order of 25 to 30 depending on latitude and local grid geometry. For each NWIOI grid cell, all ICON-2I grid points falling within its spatial extent are identified and aggregated using a simple arithmetic mean.

Since the ICON-2I grid is substantially finer than the observational grid and exhibits relatively smooth spatial variability at the aggregation scale, the use of an unweighted average provides a computationally efficient and physically consistent representation of the mean signal within each coarser cell. This approach allows the higher-resolution ICON-2I output to be consistently mapped onto the NWIOI grid, retaining the large-scale spatial structure of the original forecast fields while filtering out sub-grid variability.

A MATLAB code snippet illustrating the domain extraction and spatial regridding procedure is shown in the following figure.

```matlab
%% -------------------------
% SPATIAL CUTTING
ilat_min = find(lat_orig >= lat_max, 1, 'last');        % considering decreasing lat
ilat_max = find(lat_orig <= lat_min, 1, 'first');
ilon_min = find(lon_orig <= lon_min, 1, 'last');
ilon_max = find(lon_orig >= lon_max, 1, 'first');

lat_sub  = lat_orig(ilat_min:ilat_max);
lon_sub  = lon_orig(ilon_min:ilon_max);
data_sub = data_orig(ilat_min:ilat_max, ilon_min:ilon_max, 1:n_bande);
[LON_SUB, LAT_SUB] = meshgrid(lon_sub, lat_sub);

%% -------------------------
% REGRIDDING ONTO THE OBSERVATIONAL GRID
nlat_obs = length(lat_obs);
nlon_obs = length(lon_obs);
fore_on_obs = NaN(nlat_obs, nlon_obs, n_bande);

lat_edges = [lat_obs + 0.125/2; lat_obs(end) - 0.125/2];    % considering decreasing lat
lon_edges = [lon_obs - 0.125/2; lon_obs(end) + 0.125/2];

for b = 1:n_bande
    slice_b = data_sub(:,:,b);
    for i = 1:nlat_obs
        for j = 1:nlon_obs
            mask_sub = LAT_SUB <= lat_edges(i) & LAT_SUB > lat_edges(i+1) & ...
                       LON_SUB >= lon_edges(j) & LON_SUB < lon_edges(j+1);
            vals = slice_b(mask_sub);
            if ~isempty(vals) && any(~isnan(vals))
                fore_on_obs(i,j,b) = mean(vals);        % created considering decreasing lat
            end
        end
    end
end
```

Figure 2. MATLAB code snippet for domain cut and spatial regridding procedure.

Finally, the same geographic mask corresponding to the effective area of interest is also applied to the regridded forecast fields.

The spatial processing workflow is illustrated in the following figures, which refer to an arbitrary example day and are provided simply for illustrative purposes. They show the domain reduction, the subsequent regridding of both precipitation and temperature fields, and the final application of the spatial mask.

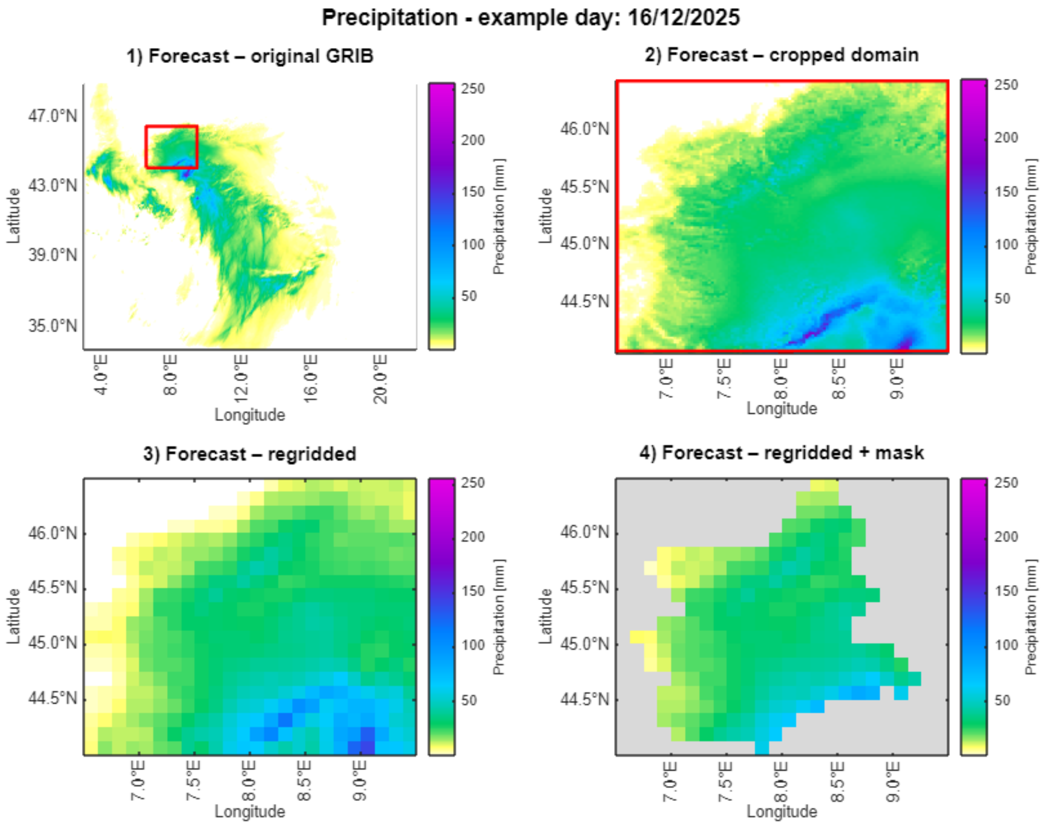

Figure 3. ICON-2I domain cut and regridding onto NWIOI grid and application of the spatial mask. (precipitation example).

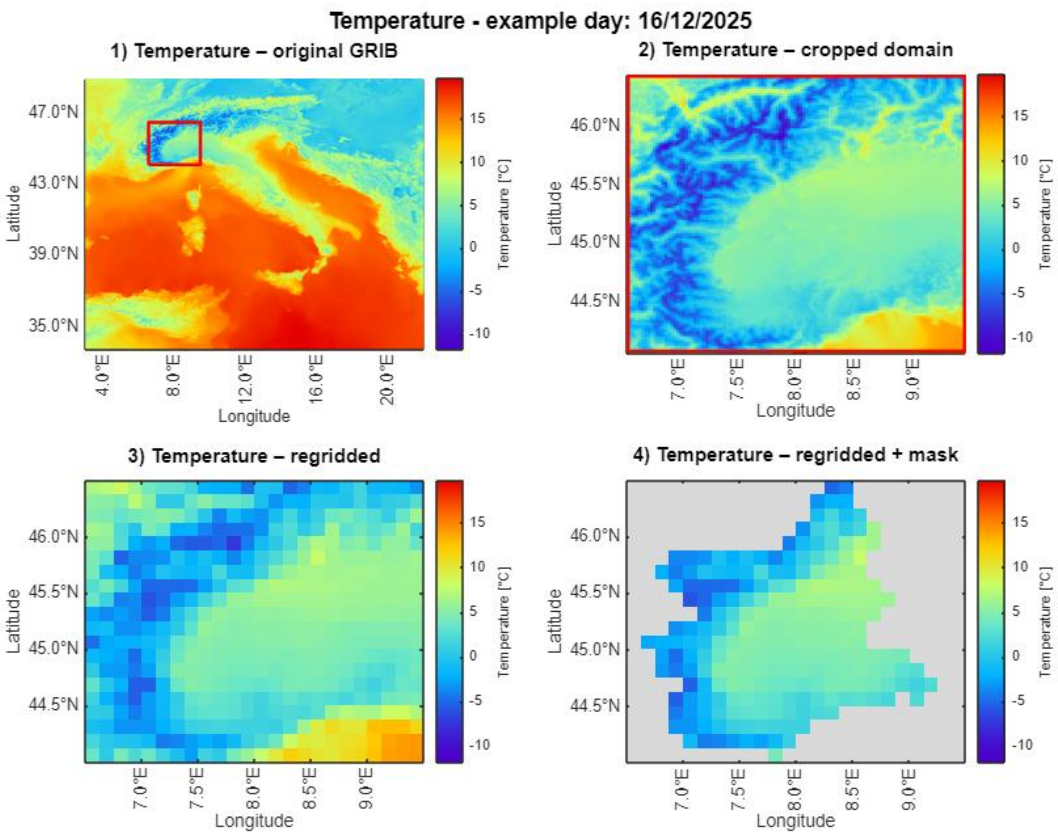

Figure 4. ICON-2I domain cut and regridding onto NWIOI grid and application of the spatial mask. (temperature example).

2.2. Orographic classification

To account for terrain effects on forecast performance, the territory of Piedmont and Aosta Valley is classified into four orographic classes using a high-resolution Digital Terrain Model (DTM). The classification is based on methodology from Mombrini et al. (2023), in which topographic variability is represented by the standard deviation of elevation within each grid cell. For each grid cell, the standard deviation of elevation is computed from all DTM values contained within the spatial extent of the cell, providing a measure of local topographic variability.

The study domain, as shown in the following figure, is then divided into four orographic classes:

- Plains
- Hilly region
- Low mountains
- High mountains.

Classification thresholds are set to 40 m, 160 m, and 410 m, selected to ensure a balanced distribution of grid cells across the classes. The classification is applied only within the study domain using the geographic mask and provides a static descriptor of terrain complexity used in the subsequent analysis.

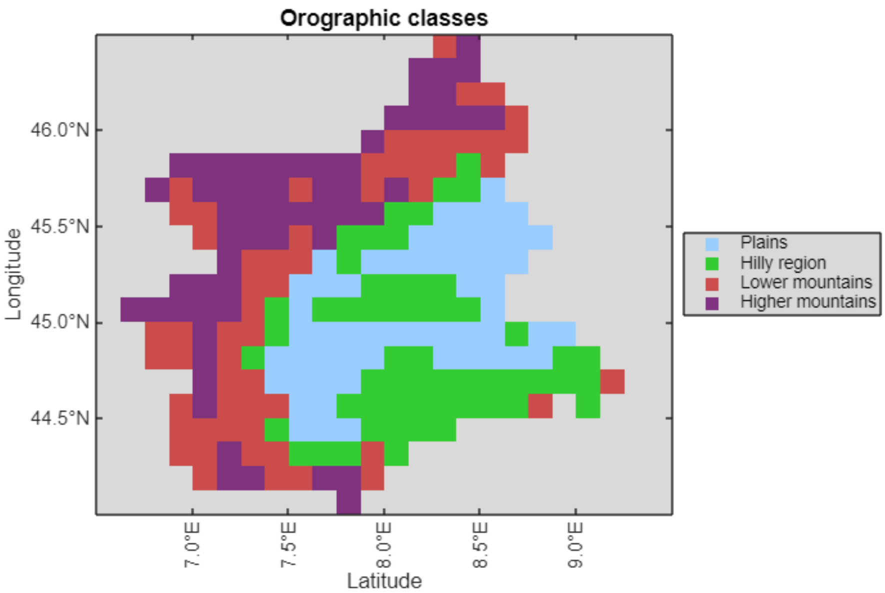

Figure 5. Orographic classification map over study domain.

2.3. Data organization

After the acquisition and preliminary processing stages, the datasets are organized into dedicated MATLAB data structures to facilitate their use in the subsequent analyses. Observational and forecast data are stored separately for temperature and precipitation variables, with each structure containing both measured and modeled components.

Since observational data are already available as single daily aggregated values, no further processing steps are required. In contrast, the forecast dataset requires more structured organization due to the presence of two daily deterministic model initializations, namely run 1 at 00 UTC and run 2 at 12 UTC, each extending up to 72 forecast hours.

To enable a consistent comparison across initialization times and prediction horizons, hourly forecasts are reorganized into overlapping lead-time windows associated with the same valid local day, ensuring that all forecasts are evaluated against a common observational reference framework. For each valid day, five forecast windows are considered, corresponding to forecasts derived from run 1 of the same day and from both initialization runs of the two preceding days. These windows are separated by 12-hour lead-time intervals and collectively provide full coverage of the local day reference.

Because forecasts are issued in UTC, a systematic offset of one to two hours with respect to local time arises depending on the time regime. This issue affects the reconstruction of the full 24-hour local period for run 1 of the same day, which does not fully cover the first one to two hours of the local day, resulting in a temporal gap of one or two missing hours for this specific initialization.

The temporal arrangement of the forecast windows, including the UTC-based offset and the resulting alignment between lead-time segments and the local day, is schematically represented in the following figure.

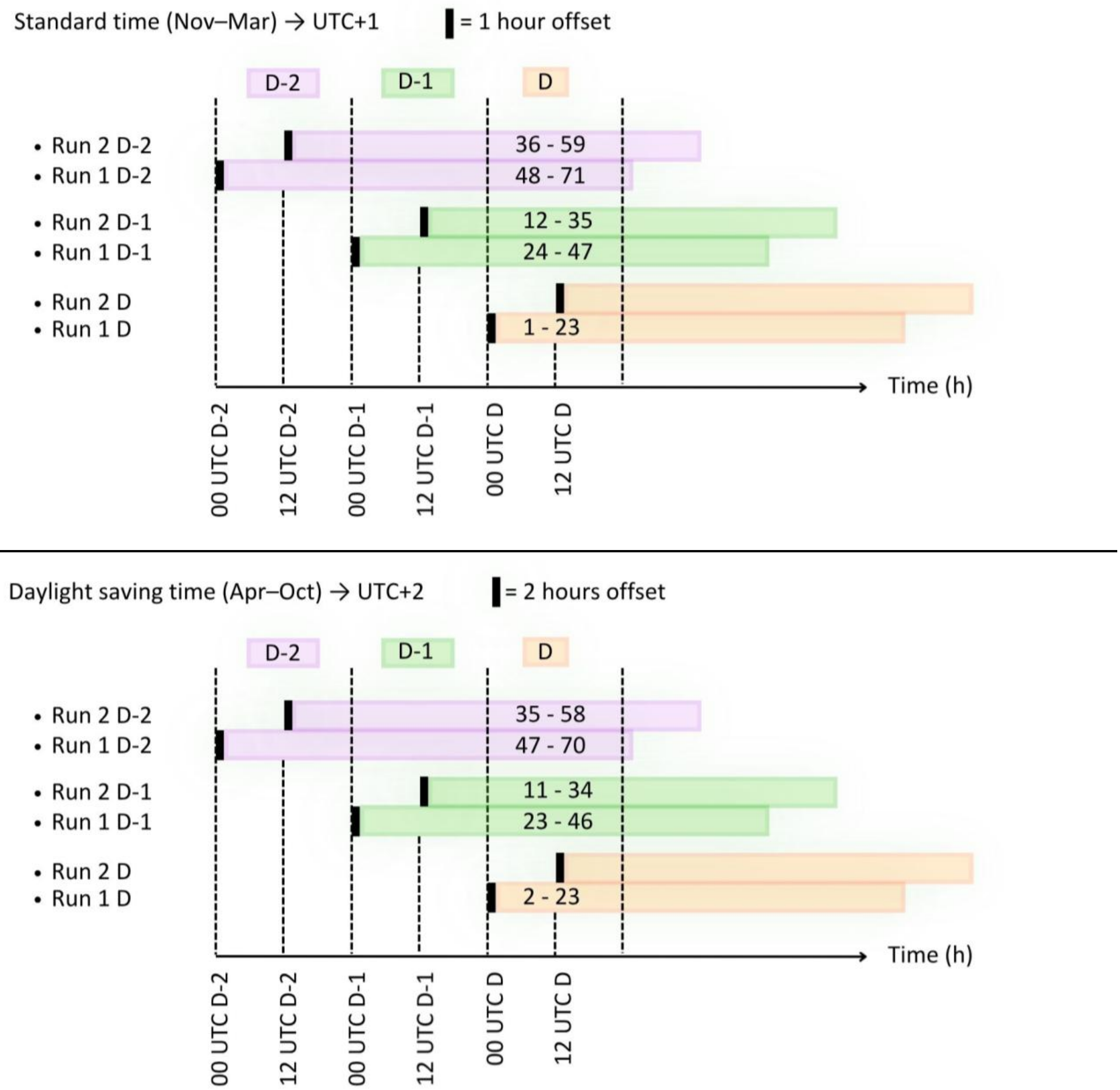

Figure 6. Forecast run structure and definition of overlapping daily validation windows under standard and daylight-saving time regimes.

The mismatch between UTC-based forecasts and local-day reconstruction introduces a systematic effect: for precipitation, the missing initial hour(s) may lead to a slight underestimation of the accumulated daily totals in the corresponding forecast window. For temperature, the gap affects the daily means by excluding one to two hours that correspond to the first hours of the diurnal cycle, which occur during nighttime and are typically the coldest part of the day. As a result, a small systematic effect may occur, potentially inducing a slight increase in the bias of the forecasted daily mean temperature for the first initialization of the same day.

In this way, each valid day is associated with a complete and temporally coherent set of forecast segments spanning progressively increasing prediction horizons, allowing a consistent comparison with the corresponding observational daily aggregates and enabling a systematic analysis of forecast performance as a function of lead time.

2.4. Deterministic verification metrics

Following the preprocessing phase, the classification of the study area into different orographic classes, and the organization of the datasets into dedicated MATLAB structures, the deterministic performance of the ICON-2I model is evaluated. The verification framework relies on three standard deterministic skill metrics: Bias, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), which provide complementary information on forecast accuracy and systematic errors. These metrics are designed to capture complementary aspects of forecast quality, specifically the systematic model deviation, the average error magnitude, and the sensitivity to large deviations.

The verification is performed at daily temporal resolution over a gridded domain defined by the study-area mask and subsequently aggregated to monthly statistics. For each day, forecast and observational fields are compared at grid-cell level, and the resulting error field is defined as:

$$E(x, y, t) = F(x, y, t) - O(x, y, t)\,,$$

where F is the forecast value and O is the observed value, both at a specific grid position (x, y) and time t. This error field constitutes the fundamental quantity for the computation of Bias, MAE, and RMSE.

The deterministic error metrics are computed through a two-step aggregation procedure:

1. Spatial aggregation is first performed for each day by averaging over all grid cells within the domain, equal contribution of each spatial unit to the daily error representation.
2. Temporal aggregation is then applied over all days within a given month, ensuring equal weighting of each day while preserving the temporal variability of forecast errors.

Since the dataset covers only a single hydrological year, each monthly estimate is based on a limited number of daily samples and should therefore be interpreted in the context of the specific meteorological conditions occurring within the considered period, rather than as long-term climatological statistics.

The analysis is performed separately for precipitation and temperature and for each forecast lead time, ordered from longer to shorter forecast horizons to evaluate the evolution of forecast error metrics with lead time. All metrics retain the physical units of the corresponding variable, namely millimetres (mm) for precipitation and degrees Celsius (°C) for temperature.

Within this framework, all grid cells and all days within each month are grouped according to observed intensity classes, defined separately for precipitation and temperature. For each class, the Mean Absolute Error (MAE) is computed as the average absolute difference between forecast and observed values over all grid cell–day pairs belonging to that class. Consequently, each MAE value represents the characteristic forecast error associated with a given range of observed conditions across the entire monthly dataset, allowing a consistent comparison of model performance across different regimes.

Furthermore, relative formulations of MAE and RMSE are also applied but exclusively to the precipitation field to provide a dimensionless assessment of forecast performance across different precipitation intensity classes. For each precipitation class, the MAE and RMSE are first computed by considering all grid cells and all days belonging to that class. The resulting error metric is then normalised by the mean observed precipitation of the same class, yielding percentage-based relative error measures. This approach expresses the average forecast error relative to the characteristic precipitation intensity of each class, enabling a consistent comparison of forecast performance across different precipitation regimes.

2.4.1. Bias

The Bias represents the first-order statistical moment of the error distribution and "measures the correspondence between the average forecast and the average observed value of the predictand" (Wilks, 2011). It quantifies the systematic deviation of the forecasting system and preserves information on the sign of the error, indicating whether the model tends to systematically overestimate or underestimate the observed variable (Jolliffe & Stephenson, 2003). A direct consequence of this formulation is that opposite-signed errors may compensate each other, leading to a potential underestimation of the overall error magnitude. For this reason, Bias should not be interpreted as a measure of accuracy but rather as an indicator of systematic model behaviour, and it is therefore analysed in combination with MAE and RMSE (Wilks, 2011).

The operational formulation used in this study follows the standard definition of the metric while incorporating the two-step spatial and temporal aggregation procedure described above, and is expressed as:

$$\mathrm{Bias}_{\mathrm{month}} = \frac{1}{\mathrm{N}_{\mathrm{days\ of\ month}}} \sum_{\mathrm{t}=1}^{\mathrm{N}_{\mathrm{days\ of\ month}}} \left(\frac{1}{\mathrm{N}_{\mathrm{cells}}} \sum_{\mathrm{x,y}} \left(\mathrm{F}(\mathrm{x},\mathrm{y},\mathrm{t}) - \mathrm{O}(\mathrm{x},\mathrm{y},\mathrm{t}) \right) \right).$$

In this context, the Bias is used to identify persistent model tendencies across different months and orographic settings. Given the limited temporal coverage of a single hydrological year, it may be influenced by event sequencing effects and should therefore be interpreted as an event-conditioned estimate of systematic behaviour.

2.4.2. Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a first-moment measure of the magnitude of forecast errors, defined as the mean of the absolute deviations between forecasts and observations. By removing the sign of the error, it avoids compensation between positive and negative deviations, providing a robust estimate of typical error magnitude (Jolliffe & Stephenson, 2003). This property makes MAE suitable for heterogeneous spatial domains, although it does not retain information on the direction of the error, and therefore does not distinguish between systematic overestimation and underestimation.

The operational formulation used in this study is expressed as:

$$\mathrm{MAE}_{\mathrm{month}} = \frac{1}{\mathrm{N}_{\mathrm{days\ of\ month}}} \sum_{\mathrm{t}=1}^{\mathrm{N}_{\mathrm{days\ of\ month}}} \left(\frac{1}{\mathrm{N}_{\mathrm{cells}}} \sum_{\mathrm{x,y}} \left| \mathrm{F}(\mathrm{x},\mathrm{y},\mathrm{t}) - \mathrm{O}(\mathrm{x},\mathrm{y},\mathrm{t}) \right| \right).$$

MAE provides a stable and physically interpretable measure of typical forecast error across lead times, intensity classes, orographic conditions, and months, and is relatively robust to isolated extreme events.

2.4.3. Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) is a second-order metric that quantifies forecast error magnitude through the quadratic mean of deviations. It is defined as the square root of the mean of squared differences between forecasts and observations, which increases the contribution of larger errors (Jolliffe & Stephenson, 2003). This property makes RMSE particularly sensitive to extreme values and rare high-impact events (Matsuura & Willmott, 2005), which can be useful for identifying significant forecast failures.

The operational formulation used in this study is expressed as:

$$RMSE_{month} = \sqrt{\frac{1}{\mathrm{N_{days\ of\ month}}}\sum_{\mathrm{t=1}}^{\mathrm{N_{days\ of\ month}}}\left(\frac{1}{\mathrm{N_{cells}}}\sum_{\mathrm{x,y}}(\mathrm{F(x,y,t)} - \mathrm{O(x,y,t)})^2\right)}\ .$$

Due to its quadratic formulation, RMSE is more influenced by extreme events than MAE and is therefore interpreted jointly with it to distinguish between typical error levels and high-impact deviations.

2.5. Ensemble verification method

Following the deterministic verification, the analysis focuses on the evaluation of the probabilistic ensemble configuration of the ICON-2I forecasting system. While deterministic verification evaluates the performance of a single forecast realization against observations, ensemble verification assesses the model's ability to represent forecast uncertainty through the ensemble forecast distribution. To this end, a set of probabilistic verification techniques is employed to evaluate the statistical consistency and reliability of the forecast distribution.

In addition to diagnostics based on the full ensemble distribution, the ensemble mean is also used as a representative forecast value for comparison with observations and with deterministic forecasts. It is important to note that this quantity differs from the deterministic forecast, which corresponds to run 1 (00 UTC) of the same analysed day and is generated as an independent model realization rather than as an aggregation of ensemble members. As a result, the deterministic forecast and the ensemble mean provide complementary representations of model behaviour, corresponding respectively to a single model realization and to the mean state of the ensemble.

Because meteorological variables and forecast errors often exhibit complex statistical behaviour that cannot be adequately represented by simple theoretical distributions, ensemble verification is based on non-parametric diagnostic tools that do not rely on specific distributional assumptions and are therefore well suited to real atmospheric data.

Within this context, the verification framework is designed to investigate three main aspects of ensemble performance: statistical calibration, reliability of the probabilistic representation, and dependence structure. Statistical calibration refers to the consistency between the ensemble forecast distribution and the observations, while reliability concerns the internal statistical behaviour of the ensemble-based representation. Dependence structure refers to the presence of residual temporal dependence in the PIT time series, which should not be present in a statistically consistent system.

These aspects are assessed through three complementary diagnostic tools: statistical calibration using the Probability Integral Transform (PIT), reliability through ensemble consistency, and dependence structure via the Kendall's Tau test applied to the time series of domain-averaged PIT values.

2.5.1. Probability Integral Transform (PIT)

The Probability Integral Transform (PIT) is used to assess the statistical calibration of the ensemble forecasts. For each forecast–observation pair, the PIT value is defined as:

$$z_i = P_i(x_i),$$

where P_i is the predictive cumulative distribution function derived from the ensemble and $x_i$is the corresponding observation. (Berkovitz, 2012)

Under the null hypothesis of a well-calibrated probabilistic system, PIT values are independent and identically distributed according to a uniform distribution $U(0,1)$. Deviations from uniformity indicate deficiencies in calibration, such as bias or dispersion errors (Laio & Tamea, 2007).

PIT diagnostics therefore provides a direct assessment of whether the ensemble distribution is statistically consistent with the observed data.

2.5.2. Spread and reliability diagnostics

Ensemble spread is defined as the variability among ensemble members and provides a first-order estimate of forecast uncertainty.

Reliability is assessed by examining the relationship between ensemble spread and the corresponding forecast error, under the expectation that larger spread should be associated with larger errors in a well-calibrated system. This relationship is used to evaluate whether the ensemble appropriately represents observed variability.

The diagnostic interpretation is based on the probability plot representation of PIT values following Laio and Tamea (2007). PIT values $z_i$are plotted against their empirical cumulative distribution function, computed as R_i/n, where $R_i$denotes the rank of $z_i$in the ordered sample and $n$is the total number of cases.

The shape of the resulting curve provides a direct visual assessment of ensemble calibration. Curves close to the bisector indicate well-calibrated forecasts, while systematic deviations highlight deficiencies in ensemble spread. Underdispersion and overdispersion can be identified through characteristic distortions of the curve, reflecting an excess of observations in the tails or in the central part of the distribution, corresponding respectively to narrow and large ensemble forecasts.

This representation provides a direct diagnostic of ensemble reliability and is summarized in Figure 7.

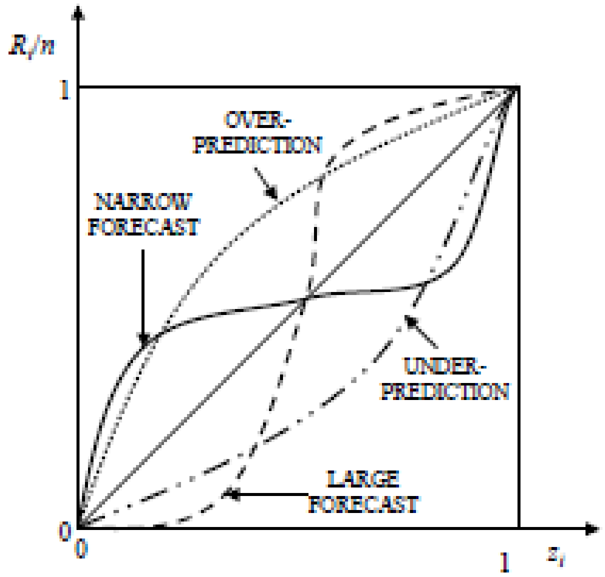

Figure 7. PIT probability plot used for ensemble forecast calibration and dispersion diagnostics. (Laio & Tamea, 2007)

2.5.3. Kendall Tau independence test

The Kendall Tau test is a non-parametric method used to assess dependence based on rank ordering. It does not rely on the numerical magnitude of the data, making it suitable for non-linear and non-Gaussian behaviour typical of meteorological variables.

In this study, following Kendall and Stuart (1977), the test is used as a diagnostic tool to detect residual dependence in ensemble-derived sequences from the ICON-2I forecasting system. The objective is to verify whether transformed forecast quantities behave as statistically independent realisations or whether they retain systematic structure.

Given a sequence of length n:

$$R_1, R_2, \dots, R_n,$$

we define the set of adjacent pairs:

$$(R_1, R_2), (R_2, R_3), \dots, (R_{n-1}, R_n).$$

Each pair represents a local transition of the sequence. The test evaluates all possible combinations of these adjacent transitions, introducing two indices $i$and $j (i \neq j)$. Each pair (R_i, R_{i+1}) is compared with every other pair (R_j, R_{j+1}), providing a global assessment of consistency across all local changes.

A discordance occurs when two transitions exhibit opposite behaviour, i.e.:

- $R_i < R_{i+1}$and $R_j > R_{j+1}$, or
- $R_i > R_{i+1}$and $R_j < R_{j+1}$.

The total number of discordances is denoted by N_d, and the Kendall Tau statistic is defined as:

$$\tau = 1 - \frac{4N_d}{(n-1)(n-2)}\,.$$

The denominator reflects the total number of possible comparisons between adjacent transitions, which increases quadratically with the sample size.

Under the null hypothesis of independence, the standardized statistic is approximately normally distributed, allowing formal hypothesis testing. Within this framework, the test is applied to time series derived from ensemble diagnostics. For the Probability Integral Transform (PIT), the sequence $z_i = P_i(x_i)$ is expected to behave as an independent and uniformly distributed sample in [0,1] if the ensemble is statistically consistent, implying the absence of residual temporal structure.

Accordingly, the Kendall Tau test is used to detect any remaining temporal dependence in the PIT time series. Values close to zero indicate that no systematic ordering is present, while significant deviations from zero suggest residual temporal structure not fully captured by the ensemble system.

3. Results

This chapter presents the results of the verification analysis conducted for both deterministic and ensemble forecasts. The evaluation of deterministic forecasts investigates forecast performance based on forecast lead time, seasonal variability, terrain complexity, and the intensity of the analysed variables. The assessment of ensemble forecasts complements this analysis by examining the probabilistic characteristics of the predictions based on forecast uncertainty, ensemble spread, statistical calibration, and the ability of the ensemble to represent the observed variability. Together, these results provide a comprehensive assessment of forecast performance and its sensitivity to the main meteorological and geographical factors influencing the study domain.

3.1. Deterministic forecast verification

3.1.1. Daily time series of deterministic forecasts and observations

The daily time series of deterministic forecasts and observations provide a direct visual comparison between forecasts at different lead times and the corresponding observations. Daily values are computed as spatial averages over the study domain.

The lead times considered in the analysis arise from two daily model runs designed to provide overlapping coverage of the same validation day through a structured forecast dataset, as previously described in Section 3.2. The runs are issued at different initialization times and account for both standard time (UTC+1) and daylight-saving time (UTC+2), as summarized in Table 3.

Table 3. Forecast run structure and overlapping validation windows under UTC+1 and UTC+2, with corresponding lead-time ranges and colour coding used in the time-series plots.

Forecast run	Lead time (UTC+1) [h]	Lead time (UTC+2) [h]	Colour in figures
Run 1 D	1 - 23	2 - 23	Red
Run 2 D – 1	12 - 35	11 - 34	Orange
Run 1 D – 1	24 - 47	23 - 46	Yellow
Run 2 D – 2	36 - 59	35 - 58	Light green
Run 1 D – 2	48 - 71	47 - 70	Light blue

The forecast series are displayed using a colour scale ranging from red to light blue, representing increasing forecast lead time. This colour scheme allows for an immediate visual interpretation of forecast performance as the verification time approaches, as shown in Figures 8 and 9 for precipitation and temperature, respectively. Furthermore, to facilitate comparisons across months, all monthly plots share the same vertical axis scale.

Given that the deterministic verification is based on a single-year dataset, all available daily samples are considered relevant for the analysis. However, some deterministic forecast runs are missing within the study period, specifically: 16 July; 4-5 and 12 October; and 5 December. No missing days are present in August, February, and May. Due to the 72-hour forecast length, these gaps are partially compensated by overlapping runs from the preceding and following days, which still provide partial temporal coverage of the affected periods. As a result, the overall temporal continuity of the dataset is preserved and the impact of the missing runs on the verification analysis is limited.

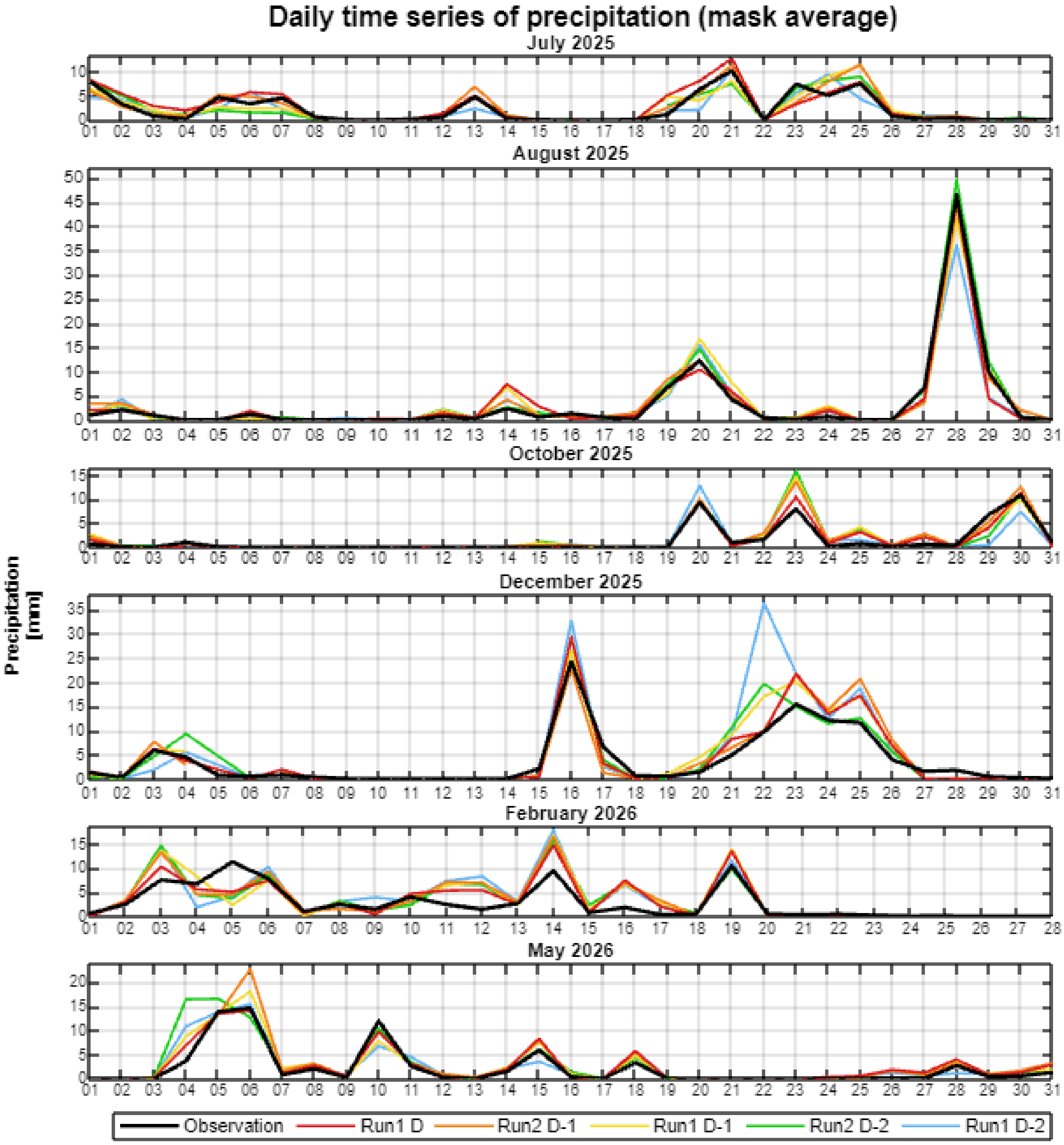

Figure 8. Precipitation time series for the selected months of the study period, showing observed and forecasted values across different lead-times.

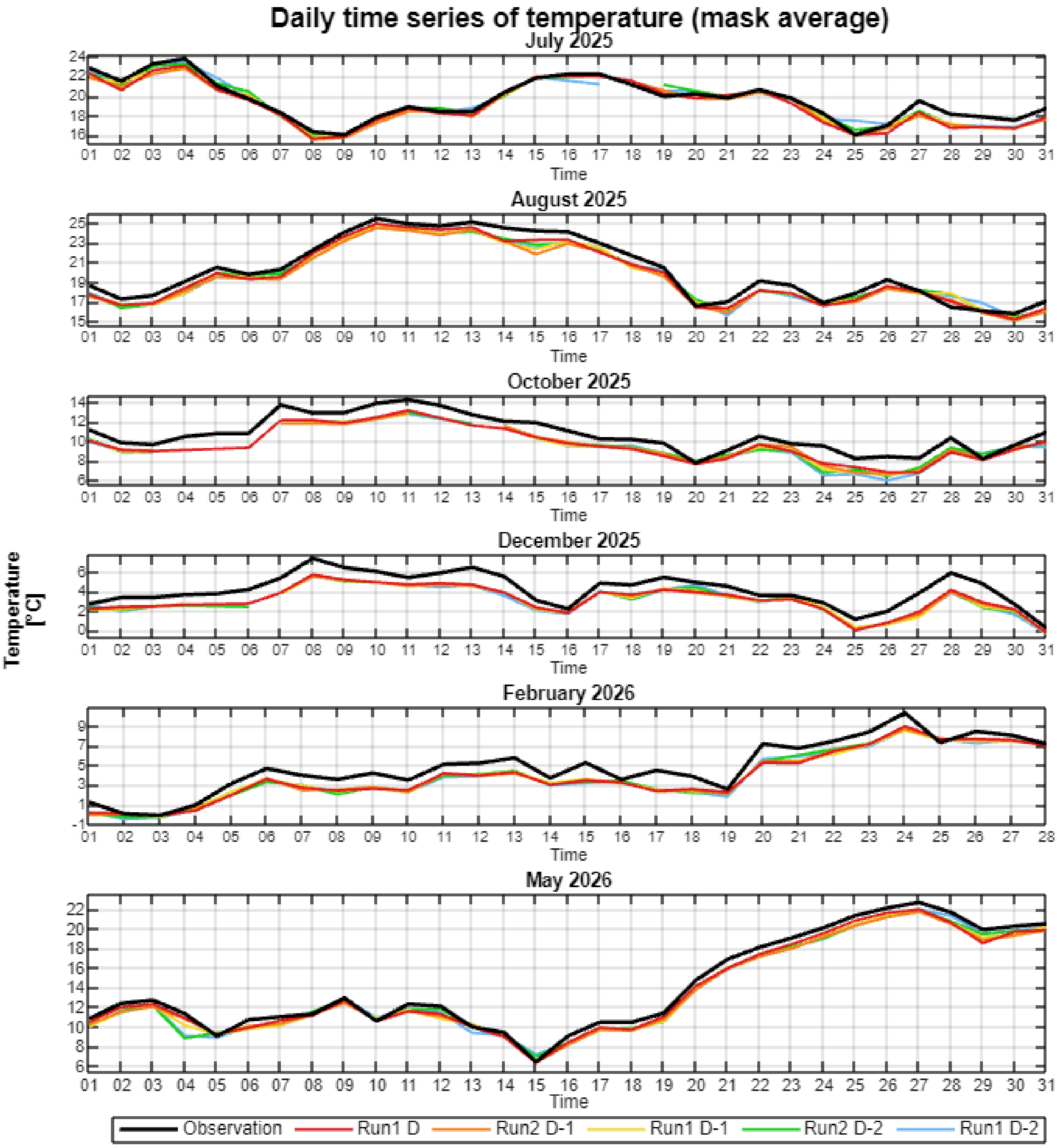

Figure 9. Temperature time series for the selected months of the study period, showing observed and forecasted values across different lead-times.

3.1.2. Magnitude and evolution of error metrics with forecast lead time

The results obtained from the analysis of the magnitude and temporal evolution of the main deterministic verification metrics, namely Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Bias, as a function of forecast lead time, are presented below.

The temporal evolution of MAE, RMSE, and Bias for the precipitation variable as a function of forecast lead time is shown in Figure 10. The results are analysed monthly to highlight variability in model performance.

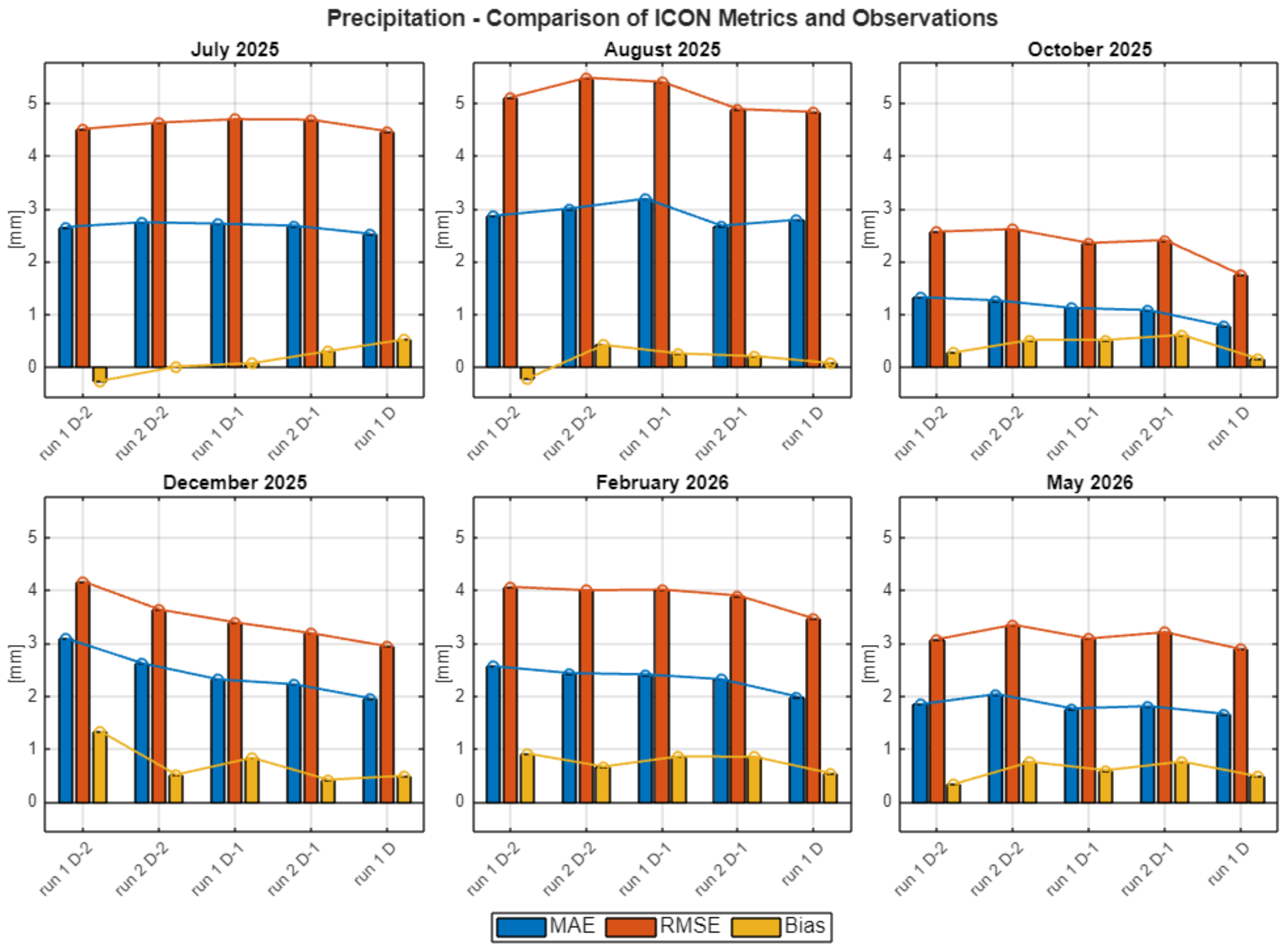

Figure 10. Temporal evolution of MAE, RMSE, and Bias for precipitation as a function of forecast lead time, aggregated monthly over the selected study period.

MAE shows an overall magnitude of approximately 2-3 mm/day when averaged over the selected months, computed as a spatial mean over the study domain. This represents the typical absolute deviation between modelled and observed precipitation. Overall, MAE exhibits a general decreasing trend with decreasing forecast lead time, indicating a progressive reduction in forecast error as the verification time approaches. This behaviour suggests an overall improvement in model performance at shorter lead times, although the intensity of this reduction is not uniform throughout the study period. From a monthly perspective, July and May display a relatively stable MAE across lead times, with no marked improvements at shorter horizons. In contrast, December shows the most pronounced improvement, with reductions exceeding 1 mm/day, while intermediate behaviour is observed in October and February, where improvements are more moderate, of the order of approximately 0.5 mm/day.

RMSE exhibits a similar temporal behaviour, but with consistently higher values compared to MAE, ranging between 2 and 5 mm/day on average. This reflects the greater sensitivity of RMSE to large errors and extreme precipitation events, which tend to exert a stronger influence on this metric. The month of August exhibits the highest overall error levels, a behaviour strongly influenced by a single major precipitation event that significantly affects the statistics.

The bias remains generally low across all months, with a mean positive value below 1 mm/day, indicating a slight systematic tendency of the model to overestimate precipitation when averaged over the domain. Unlike the other two metrics, it does not exhibit a clear dependence on forecast lead time, remaining predominantly positive across most months without a consistent monotonic trend, with only limited cases of sign changes. A notable exception is July, where the bias shifts from slightly negative values at longer lead times to slightly positive values at shorter lead times, indicating a transition from a slight initial underestimation to a subsequent overestimation. Overall, the low bias values should be interpreted in relation to the MAE magnitude, which is larger by approximately 1-2 mm/day. This discrepancy suggests that the small bias values are mainly the result of compensation between positive and negative errors across the domain rather than an intrinsically accurate spatial representation. In other words, while the model can reproduce the overall precipitation amount, it may still exhibit errors in the spatial distribution of precipitation events, leading to partial error cancellation in the spatial average.

After the analysis of the precipitation variable, the same methodology is applied to the temperature field to assess whether similar lead-time dependencies and seasonal behaviours are observed across different surface variables. In this context, Figure 11 reports the temporal evolution of error metrics for temperature as a function of forecast lead time.

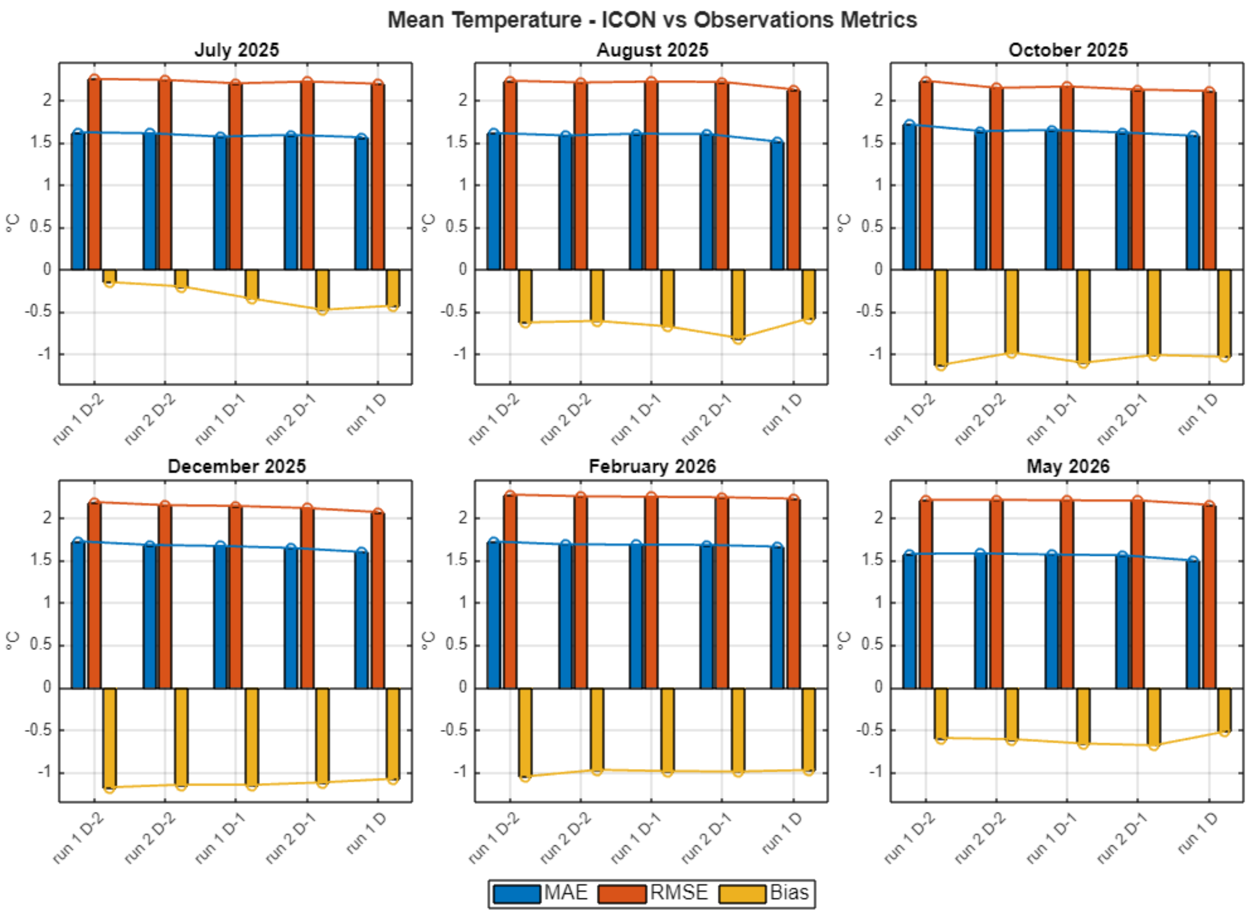

Figure 11. Temporal evolution of MAE, RMSE, and Bias for temperature as a function of forecast lead time, aggregated monthly over the selected study period.

The temporal evolution of MAE, RMSE, and Bias for temperature is characterized by a generally stable behaviour, with little variation across the forecast horizon and comparable performance already at longer lead times. Overall, temperature forecasts show no meaningful dependence on lead time, and results remain highly consistent across the selected months, with similar error magnitudes and temporal patterns.

MAE shows an average magnitude of approximately 1.7 °C over the selected months, computed as a spatial mean over the study domain. This represents the average absolute deviation between modelled and observed daily temperature. Overall, MAE remains stable across forecast lead times, indicating that the model maintains a comparable level of performance even at longer horizons.

RMSE displays similar behaviour, with an average value of about 2.3 °C. As expected, RMSE is consistently higher than MAE due to its greater sensitivity to larger deviations and extreme temperature errors, although no marked variation with forecast lead time is observed.

The bias shows a mean negative value of approximately −1 °C, indicating a general tendency of the model to underestimate temperature. This behaviour remains largely stable across the forecast horizon, with no clear dependence on lead time. The only exception is observed in July, where the bias becomes less negative at shorter lead times, indicating a slight reduction in the underestimation as the forecast approaches the verification time.

3.1.3. Error metrics as a function of orographic class

The dependence of MAE, RMSE, and Bias on orographic classes is analysed in Figure 13 for precipitation and Figure 14 for temperature, to assess how forecast performance varies with terrain complexity. The results are also shown for the different forecast lead times, which are distinguished by the colour coding introduced in the previous sections.

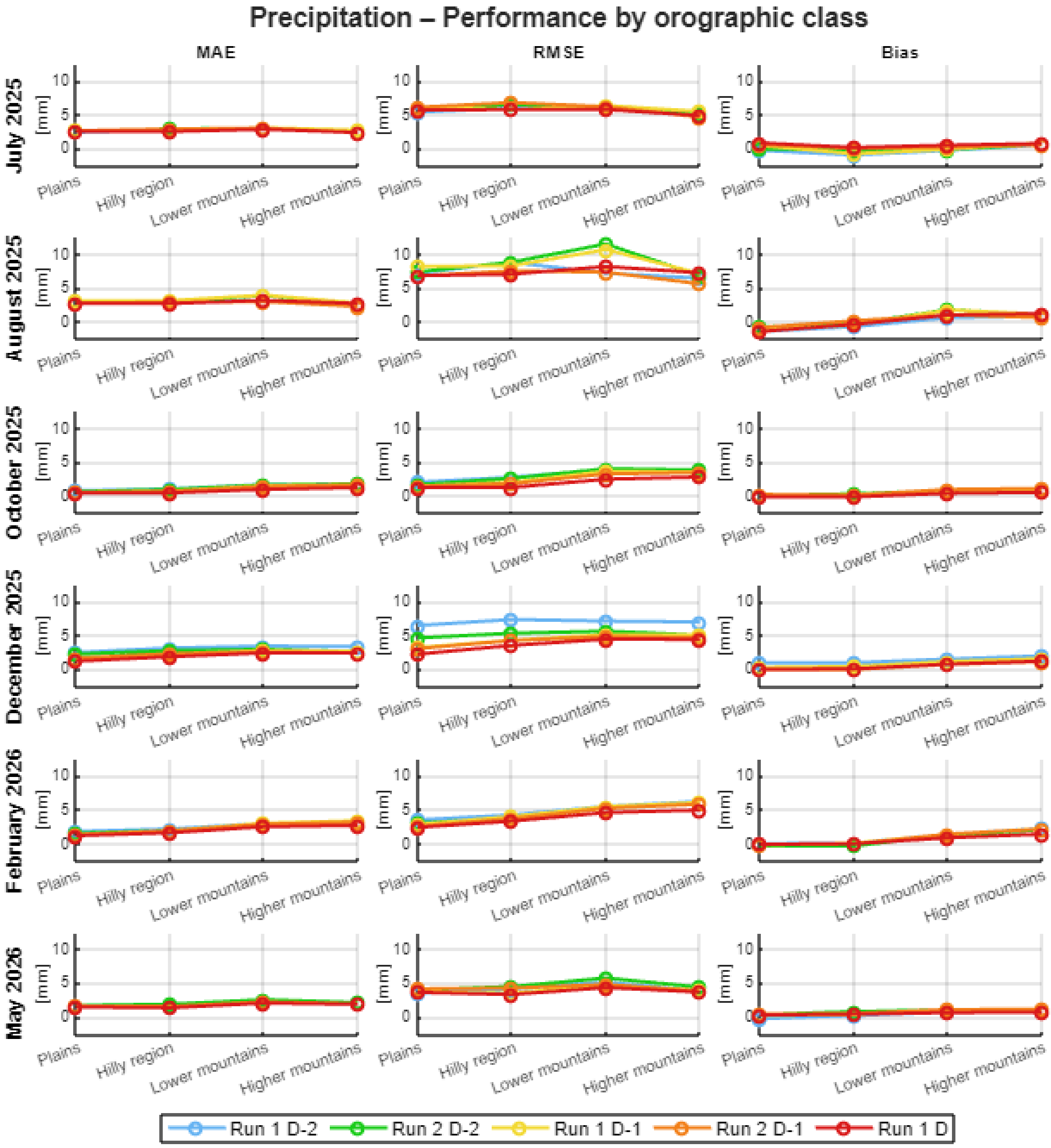

Figure 12. MAE, RMSE, and Bias by orographic class for precipitation, with results shown for different forecast lead times.

For precipitation, both MAE and RMSE exhibit a clear increasing trend from plains to hills and further towards mountainous areas. Higher error values are generally observed in mountainous regions, while plains and hills show comparatively lower and more homogeneous error levels. Within the mountainous class, low and high mountain areas appear overall consistent, without substantial differences in error magnitude. The Bias remains close to zero in both plains and hills, indicating no strong systematic deviation in these areas. In contrast, in mountainous regions the Bias becomes positive, with increases of the order of 1-2 mm depending on the month, suggesting a tendency towards precipitation overestimation at regions with higher orographic complexity.

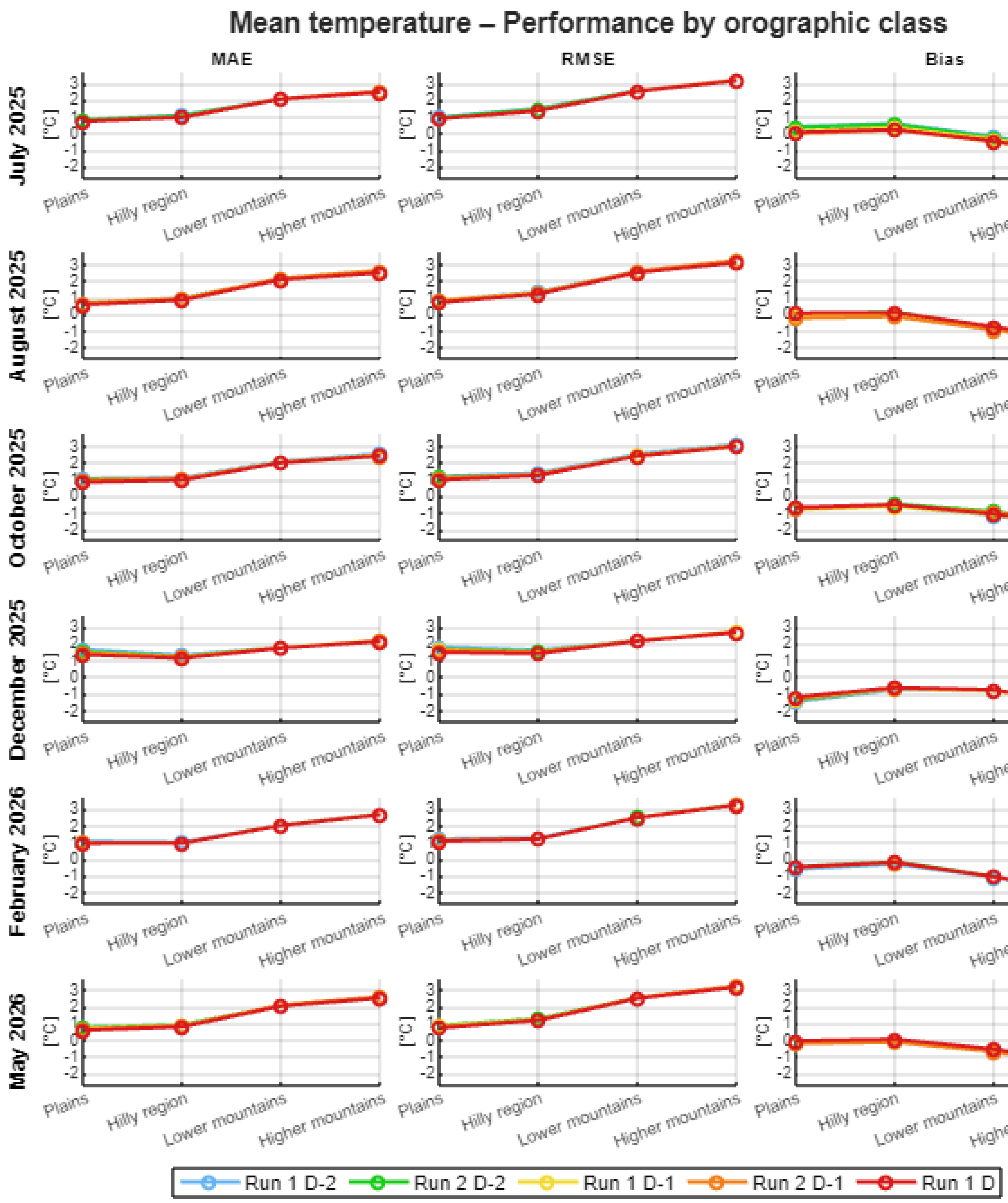

Figure 13. MAE, RMSE, and Bias by orographic class for temperature, with results shown for different forecast lead times.

For temperature, a similar orographic dependency is observed. Both MAE and RMSE increase progressively from plains to hills and then to mountainous regions, with an additional increase when moving from low to high mountains, indicating a gradual degradation of performance with orographic complexity. The Bias shows values close to zero or slightly negative in plains and hills, while it becomes increasingly negative with altitude, reaching values of approximately -2 °C in high mountain areas. This indicates a systematic tendency to underestimate temperature at higher orographic complexity.

3.1.4. Monthly spatial error distribution and forecast-observation comparison

This subsection presents the spatial distribution of error metrics, MAE, RMSE and Bias, at monthly scale, together with a direct comparison between forecasts and observations. The analysis is carried out using the deterministic forecast with lead time of the run 1 of the same day.

For precipitation, the spatial maps of MAE, RMSE and Bias, shown in Figure 14, highlight how forecast performance varies monthly across the domain and provide a direct visual representation of the spatial coherence of the error structure, complementing the previously discussed aggregated and class-based statistics.

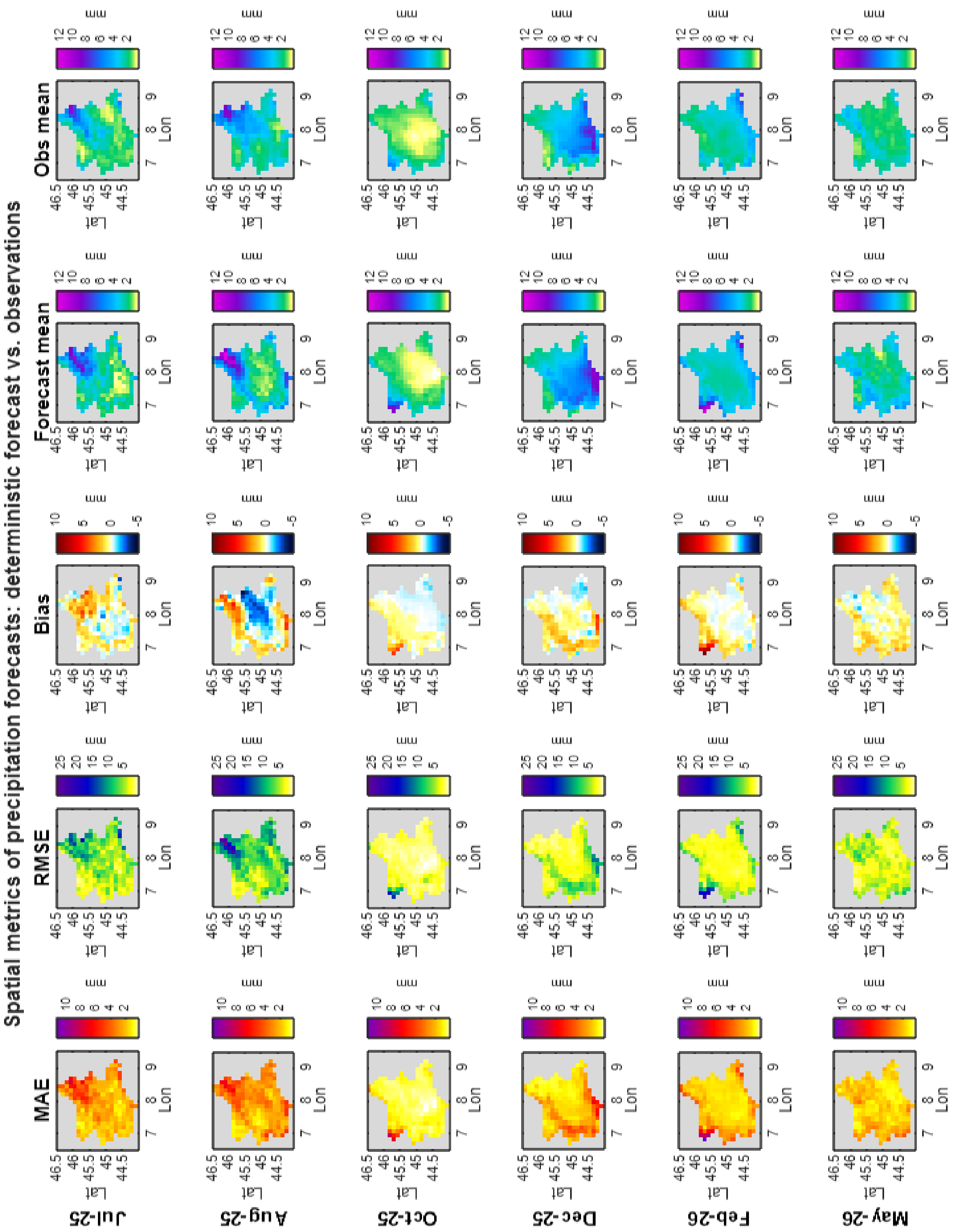

Figure 14. Monthly spatial patterns of forecast–observation errors (MAE, RMSE and Bias) for precipitation.

For temperature, complementary spatial maps are presented in Figure 15, showing the monthly mean forecast and observed temperature fields together with the corresponding MAE, RMSE, and Bias distributions. This combined representation provides spatial context for the error metrics by linking them directly to the spatial structure of both the forecast and the observations, facilitating the interpretation of forecast performance across the domain.

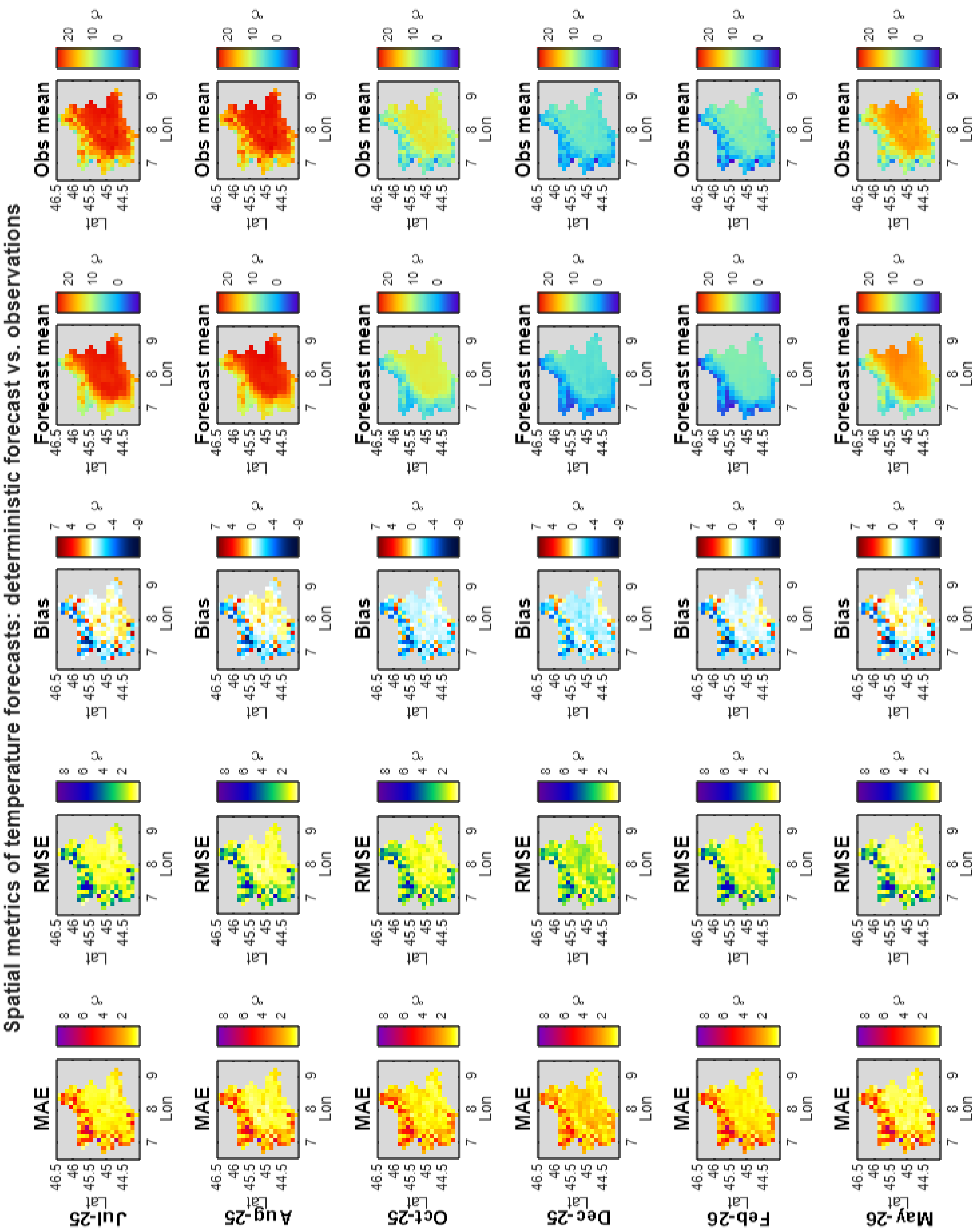

Figure 15. Monthly spatial patterns of forecast–observation errors (MAE, RMSE and Bias) for temperature.

Overall, the combined forecast-observation comparison and error metric maps provide a consistent spatially distributed assessment of model performance, highlighting areas of persistent overestimation or underestimation and supporting the interpretation of the statistical results presented in the previous subsections. The detailed relationships between forecast errors and the analysed environmental and intensity-dependent factors are further investigated in the subsequent subsections.

3.1.5. Spatial and seasonal variability of variable-error relationships

The analysis of precipitation-error relationships allows the identification of several consistent patterns describing the dependence structure between forecast performance and key hydro-meteorological drivers. Rather than being limited to a graphical comparison, this section synthesizes the main behavioural relationships emerging from the joint analysis of precipitation magnitude, error metrics, seasonality, and orographic effects.

The following key relationships have been identified:

- Precipitation intensity vs. error metrics (Figure 16):

The following figure illustrates the relationships between forecast errors and observed precipitation intensity classes, providing both absolute and relative formulations of MAE and RMSE to support a comprehensive interpretation of how error behaviour varies with precipitation intensity.

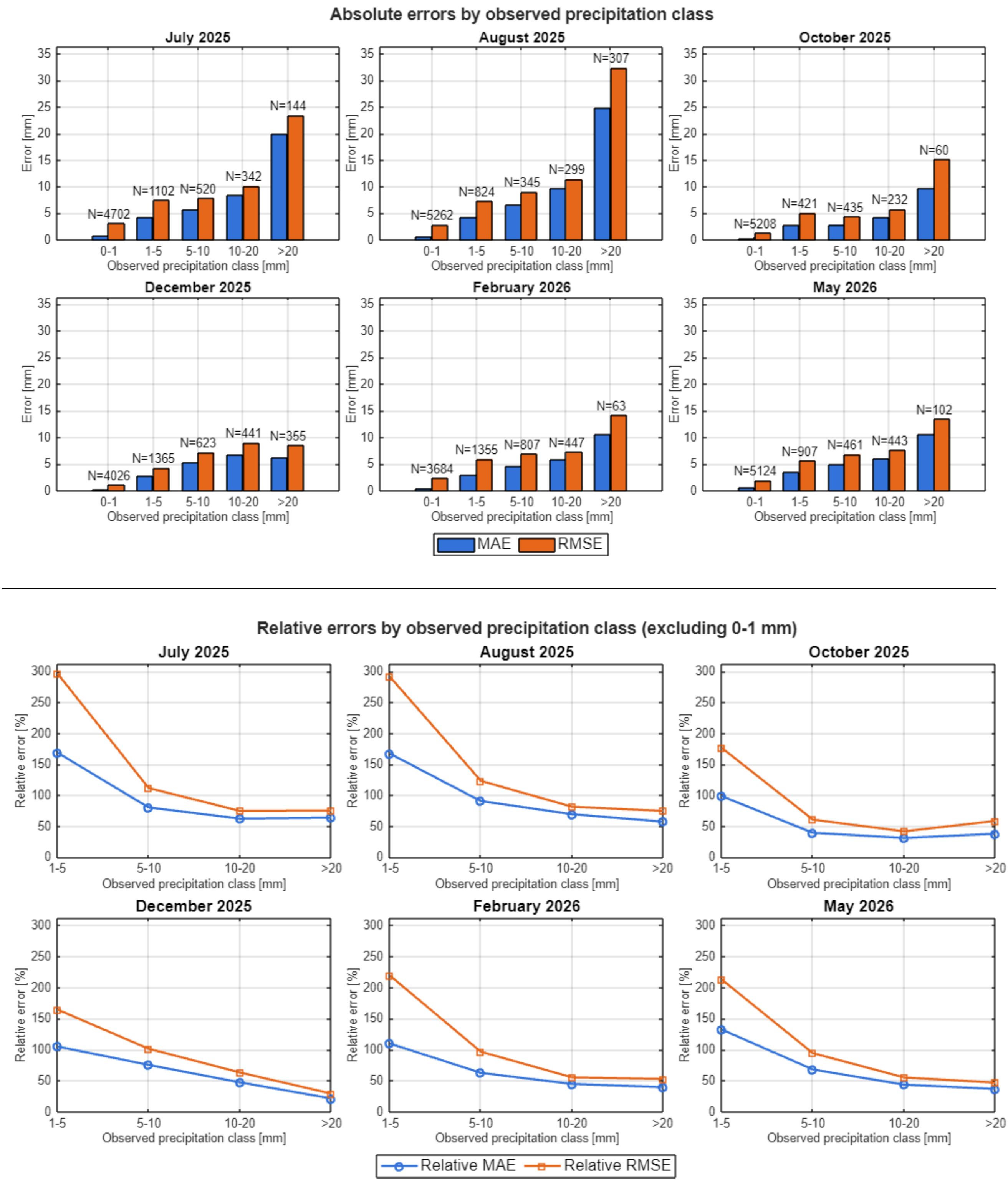

Figure 16. Dependence of absolute and relative forecast errors on precipitation intensity.

Above the bars in the absolute-error panels, it's reported the number of samples (N) associated with each precipitation class. Each sample represents a single grid cell for a single day within the month considered whose observed precipitation falls within the class considered. This information provides an indication of the statistical robustness of the estimated metrics and helps assess the extent to which the results may be influenced by the available sample size. For the relative metrics, the first precipitation class (0-1 mm) is excluded because very small observed values in the denominator can disproportionately large relative errors, reducing the interpretability of the results.

Overall, a clear contrast emerges between the behaviour of absolute and relative error metrics across precipitation classes. Both MAE and RMSE increase with precipitation intensity, indicating that larger precipitation amounts are associated with greater absolute forecast deviations and reflecting the increasing difficulty of accurately reproducing intense precipitation events. In contrast, relative MAE and RMSE generally decrease as precipitation increases, as a given absolute error represents a progressively smaller fraction of the observed amount. This behaviour suggests that, although forecasts tend to exhibit larger absolute errors during intense events, their performance improves when evaluated relative to the magnitude of the precipitation being predicted.

- Orographic sensitivity on forecast errors for comparable precipitation intensities (Figure 17):

When forecast errors are analysed across orographic classes for comparable precipitation intensities, a clear seasonal dependence emerges. During October, December, and February, the largest errors are generally observed in mountainous areas, whereas in May, July, and August higher errors tend to occur over plains and hilly regions. These results suggest that the relationship between forecast error and orography is not constant throughout the year but is strongly modulated by seasonal atmospheric conditions and the dominant precipitation-generating mechanisms characterising each month.

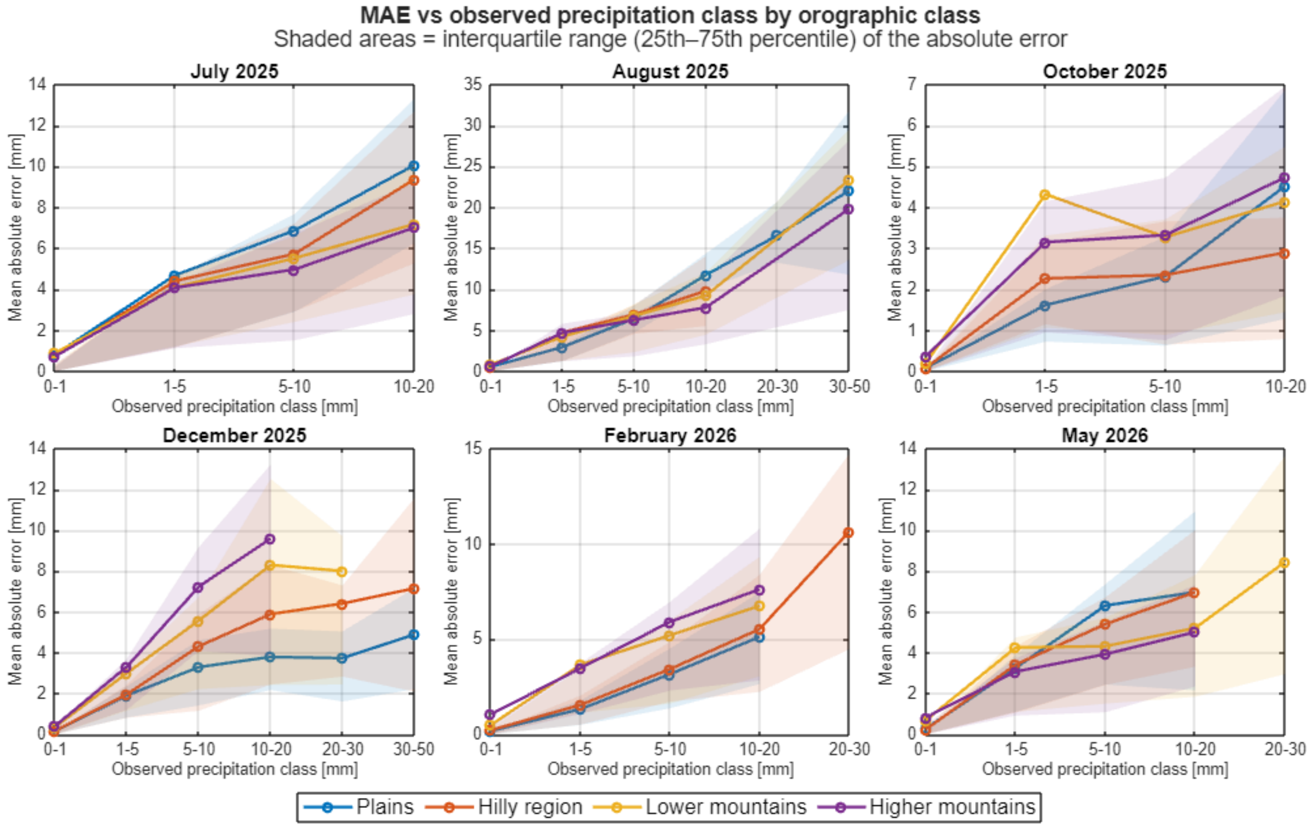

Figure 17. Dependence of orography on monthly forecast errors for comparable precipitation intensities.

- Monthly variability of MAE, in both absolute and relative terms, across comparable precipitation intensity classes (Figure 18):

The MAE, in both absolute and relative terms, is computed for each precipitation intensity class and month to characterise the seasonal dependence of forecast errors under comparable rainfall conditions. Both metrics are explicitly conditioned on the precipitation class, ensuring a consistent comparison of error behaviour across intensity ranges. The inter-month comparison shows a coherent pattern in which, for the same precipitation class, July and August consistently exhibit the highest error levels. This behaviour suggests a potentially stronger difficulty in capturing warm-season precipitation processes compared to other months; however, this interpretation should be made with caution, as the analysis is based on a single year of available data.

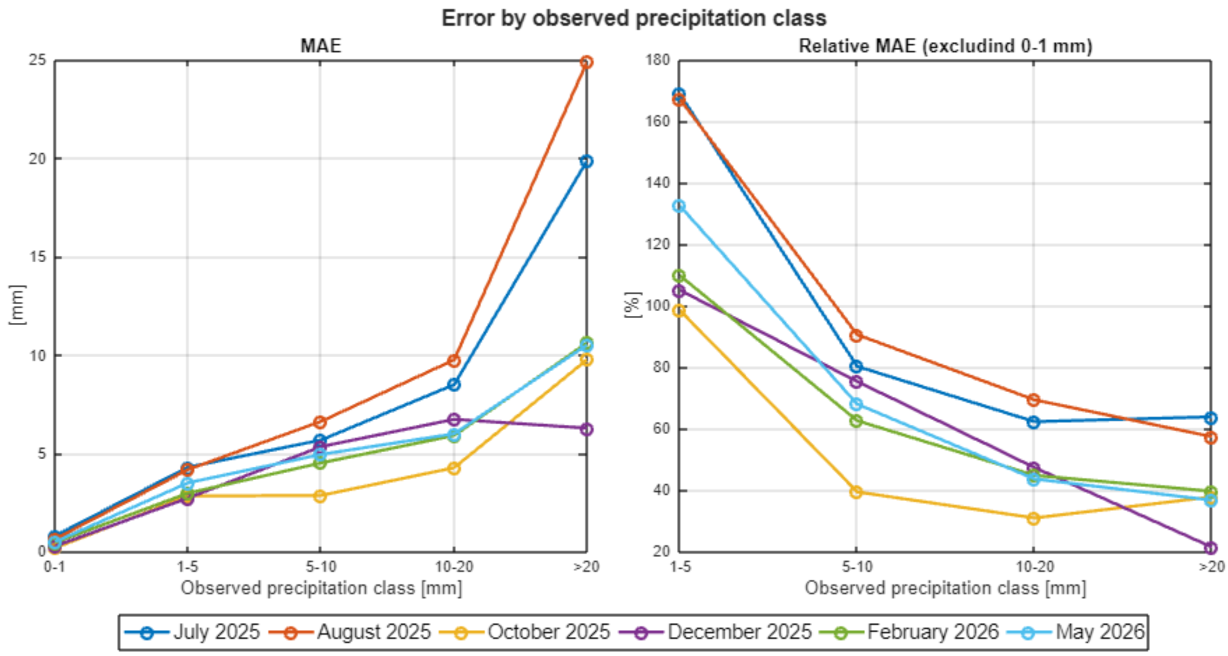

Figure 18. Monthly variability of MAE and RMSE for comparable precipitation intensities.

- Joint dependence between precipitation distribution, orography and error patterns (Figure 19):

A joint analysis is performed using monthly panels in which, for each orographic class, the mean precipitation, MAE, and bias are jointly evaluated. This structure allows a direct comparison between the climatological distribution of precipitation and the associated forecast error characteristics within each orographic class. In July, August, May, and October, classes characterised by higher mean precipitation generally also exhibit higher MAE within the same classes, indicating a co-variation between rainfall intensity and error magnitude. Conversely, in December and February, precipitation is mainly concentrated over plains and hills, whereas mountainous classes show lower mean precipitation but higher MAE and more systematic bias. Therefore, even though in these two months the largest precipitation amounts are observed in plains and hilly regions, the largest errors are consistently found in mountainous regions. This pattern may be consistent with winter cyclonic precipitation events, which are generally better represented in regions of low orographic complexity, whereas in mountainous areas the error may be partly related to limitations in the representation of snowfall processes.

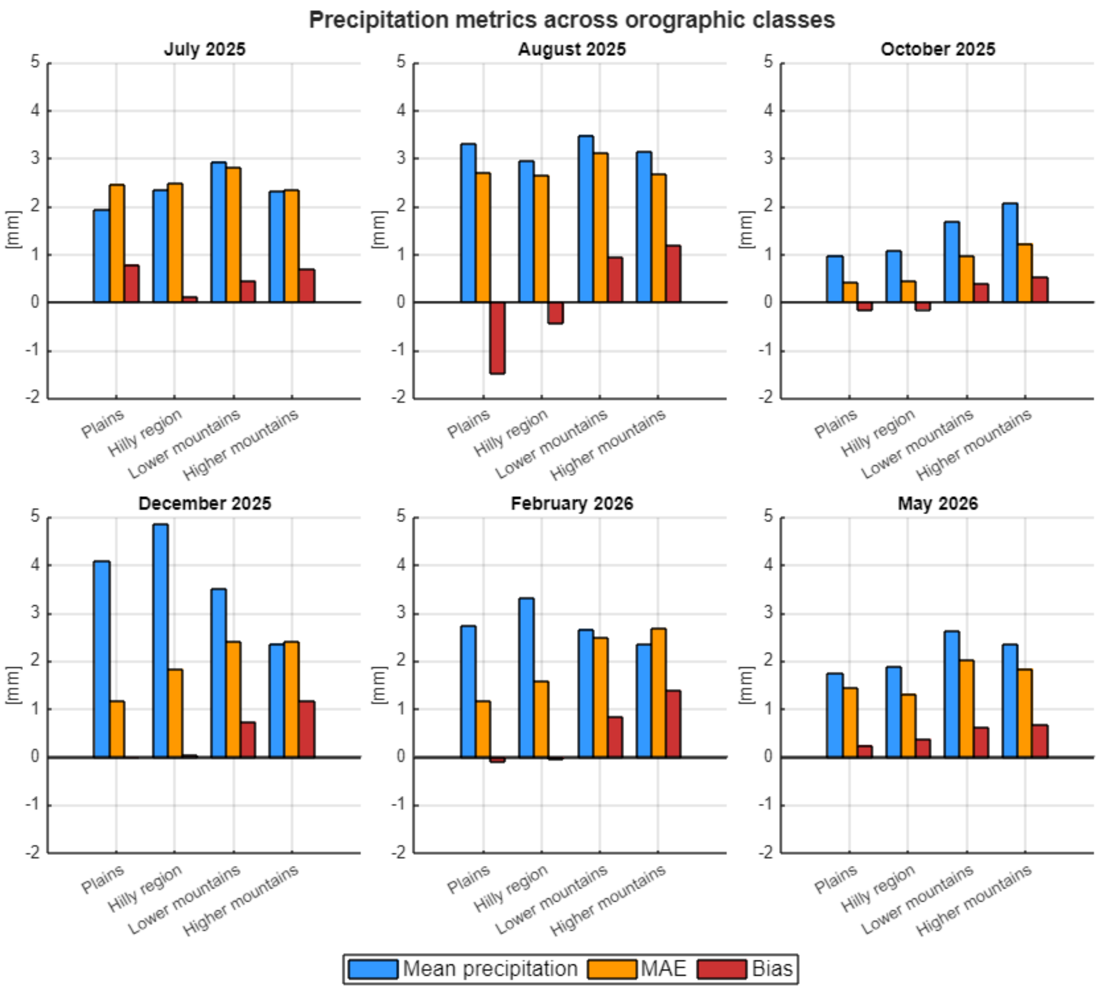

Figure 19. Joint dependence between precipitation distribution, orography and error patterns.

Following the analysis of relationships for precipitation, the analysis is extended to temperature to investigate whether the systematic error behaviours observed so far also emerge when considering a different atmospheric variable and associated physical controls. The analysis of temperature-error relationships highlights systematic dependencies between forecast performance, temperature magnitude, and orographic setting. The following key behaviours are identified:

- Temperature intensity vs. error metrics (Figure 20):

The following figure shows the relationship between forecast errors and observed temperature through monthly scatter plots of MAE and bias.

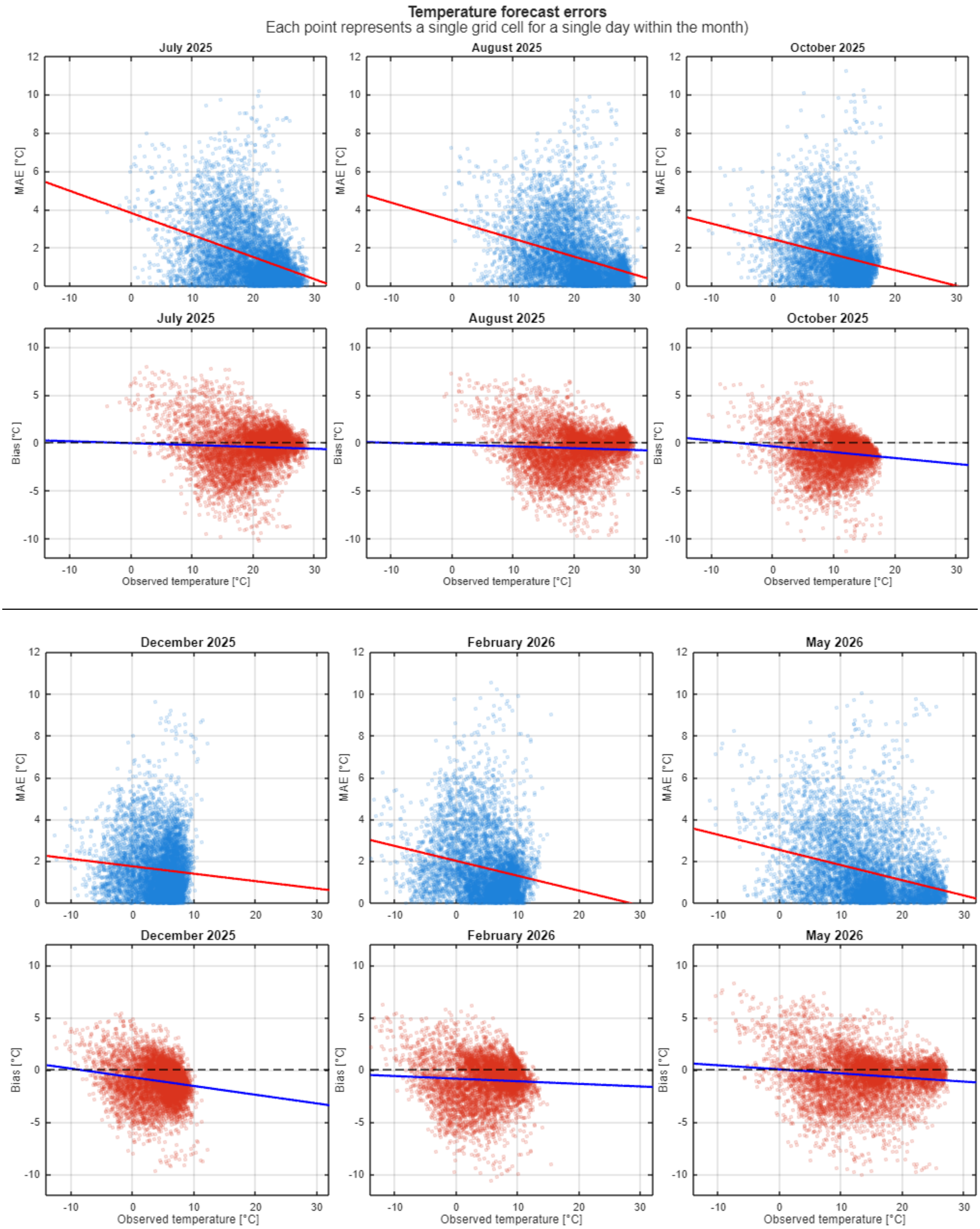

Figure 20. Monthly scatter plots of MAE and bias as a function of observed temperature.

Scatter plots of MAE and bias are constructed as a function of observed temperature, where each point represents a single grid cell for a single day within the corresponding month. The MAE exhibits a decreasing trend with increasing temperature, as indicated by the negative slope of the fitted regression line. This behaviour is partly related to the higher density of observations at warmer temperatures, where the point cloud becomes more concentrated and the dispersion of errors is reduced.

The structure of the bias distribution provides additional insight into the model error characteristics. Although the regression line indicates an overall negative bias across the temperature range, the scatter reveals a clear dependence of the error sign on temperature, with predominantly positive bias values at lower temperatures and negative values at higher temperatures. This sign reversal indicates a systematic regression toward the mean state, whereby the model overestimates low temperatures and underestimates high temperatures. This behaviour leads to a contraction of the simulated temperature range, with forecasted values being systematically pulled toward intermediate conditions compared to observations.

The approximately linear trend of the bias shows a slight negative slope, indicating that the magnitude of the cold bias increases with increasing temperature. In other words, higher temperatures are increasingly underestimated, while at lower temperatures the bias exhibits a broader spread, including both overestimation and underestimation.

As a result, the reduction in MAE at higher temperatures is mainly associated with a decrease in error dispersion. In contrast, the bias highlights a persistent and temperature-dependent systematic error, characterized by an increasing underestimation under warmer conditions.

- Orographic sensitivity at comparable temperature levels (Figure 21):

The following figure presents the relationship between forecast performance and orographic complexity within comparable temperature classes.

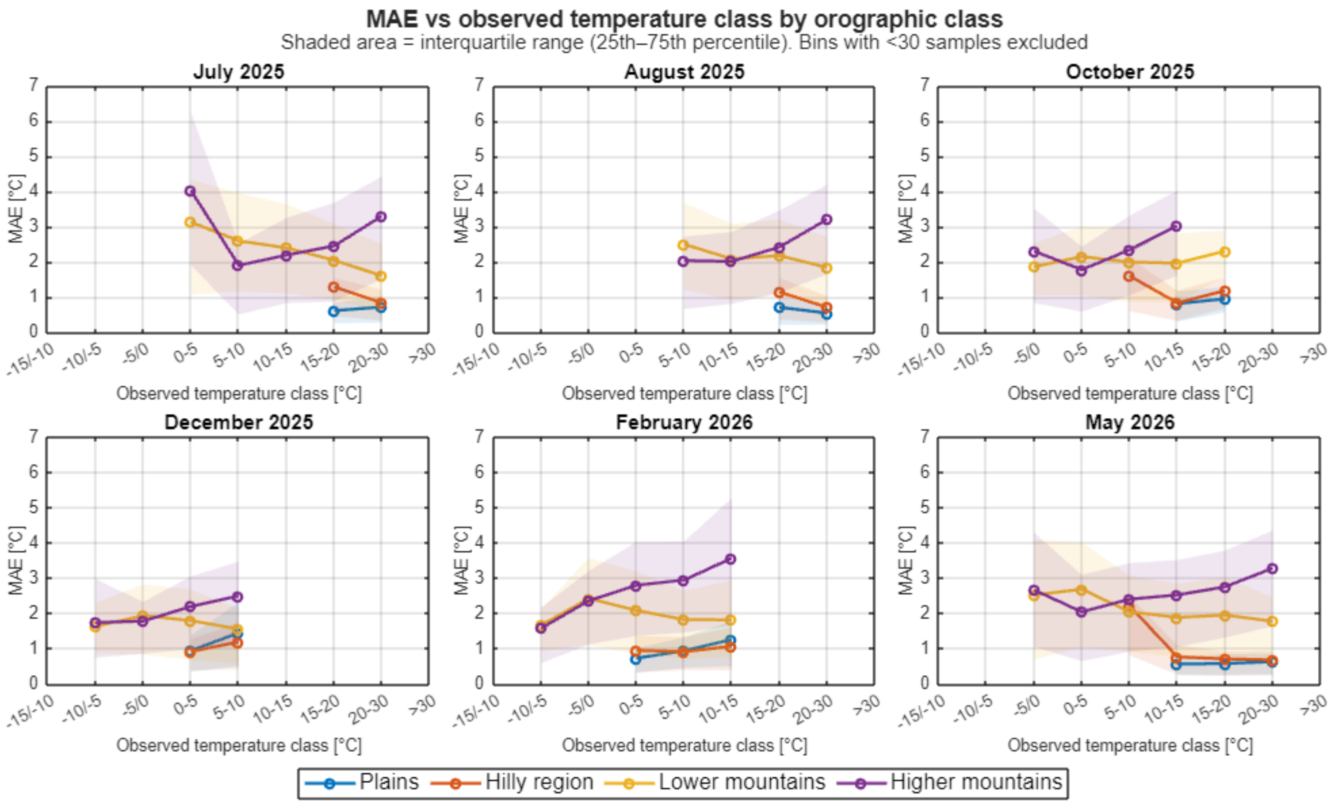

Figure 21. Dependence of orography on forecast errors for comparable temperature levels month by month.

Within the same temperature class, mountains consistently exhibit lower forecast performance compared to plains and hilly regions, which show more favourable error characteristics. This behaviour highlights a persistent sensitivity of forecast errors to terrain complexity, which is consistently observed across all temperature classes and remains evident throughout all months analysed.

To ensure the robustness and interpretability of the results, bins containing fewer than 30 samples were excluded from the analysis, as low-sample conditions introduced excessive variability and could significantly distort the estimated error statistics.

- Seasonal behaviour at comparable temperature levels (Figure 22):

The following figure illustrates the seasonal dependence of forecast errors across comparable temperature classes.

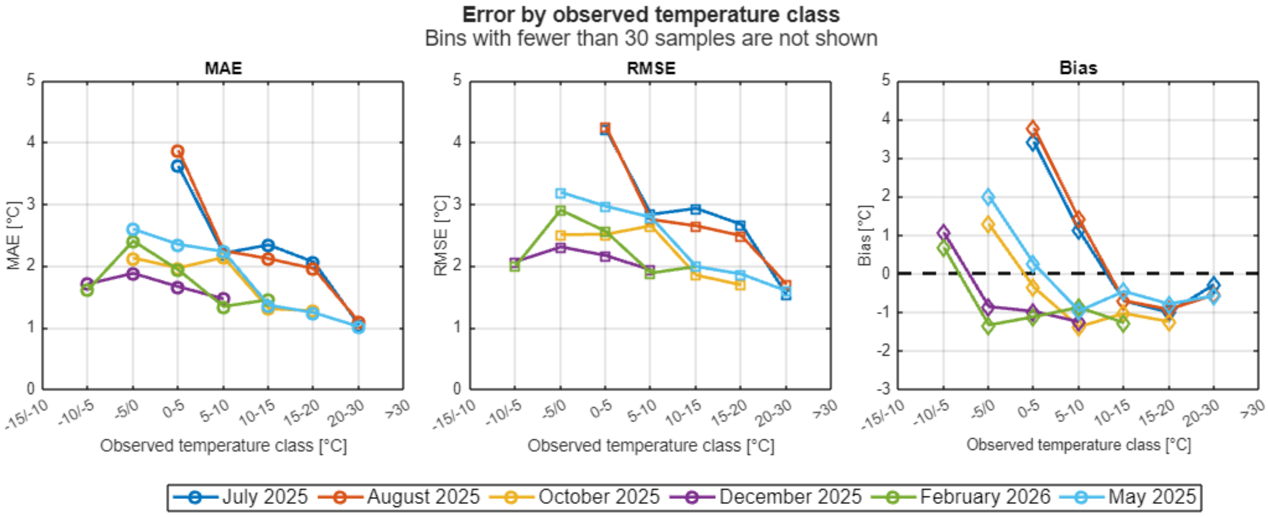

Figure 22. Monthly variability of MAE and RMSE for comparable temperature levels.

When comparing the same temperature ranges across months, July and August consistently exhibit the highest error levels, followed by May and October, while February and December show the lowest errors overall. This behaviour is derived from a single year of available data and should therefore be interpreted with caution, as it does not fully capture interannual variability.

Overall, the pattern suggests a clear seasonal signal, with higher errors during the warm season and reduced errors in the cold season.

To ensure robustness and reliability of the analysis, bins containing fewer than 30 samples were excluded, as low-sample conditions introduced excessive variability and could distort the estimated error statistics.

3.2. Ensemble analysis results

This section presents the results of the verification analysis conducted for the ensemble forecasting system. In addition to the evaluation of deterministic forecasts, the ensemble assessment focuses on the probabilistic representation of atmospheric variability, with particular attention to forecast performance, ensemble spread and statistical consistency.

Given that also the ensemble verification is based on a single-year dataset, all available daily samples contribute to the statistical analysis. However, a few ensemble forecast runs are missing during the study period, specifically: 6 July; 3 August; from 3 to 6 and from 9 to 12 October; and 30 May. No missing days are present in December and February. Unlike the deterministic configuration, the ensemble forecasts are evaluated only for the first forecast day, without exploiting overlapping lead-time windows from adjacent initializations. In particular, the extended sequence of missing runs in October reduces the number of available forecast samples for that month, leading to a less robust statistical basis. Consequently, the statistical analyses for October should be interpreted with greater caution than those for the other months.

3.2.1. Daily time series of ensemble and deterministic forecast and observations

This subsection presents the daily time series of ensemble forecasts, deterministic forecasts, and observations, providing a direct comparison between deterministic and probabilistic model outputs. Ensemble forecasts are represented through their distributional characteristics, while the ensemble mean is used as a single representative metric to compare with deterministic predictions and observed values.

Daily values are computed as spatial averages over the study domain, ensuring consistency with the deterministic framework and allowing a direct comparison of the temporal evolution across the different forecast products.

The time series provide a first qualitative assessment of the ensemble behaviour, including its ability to reproduce the observed variability and the relationship between ensemble spread and forecast uncertainty under different atmospheric conditions.

Figures 23 and 24 illustrate the ensemble behaviour for precipitation and temperature, respectively.

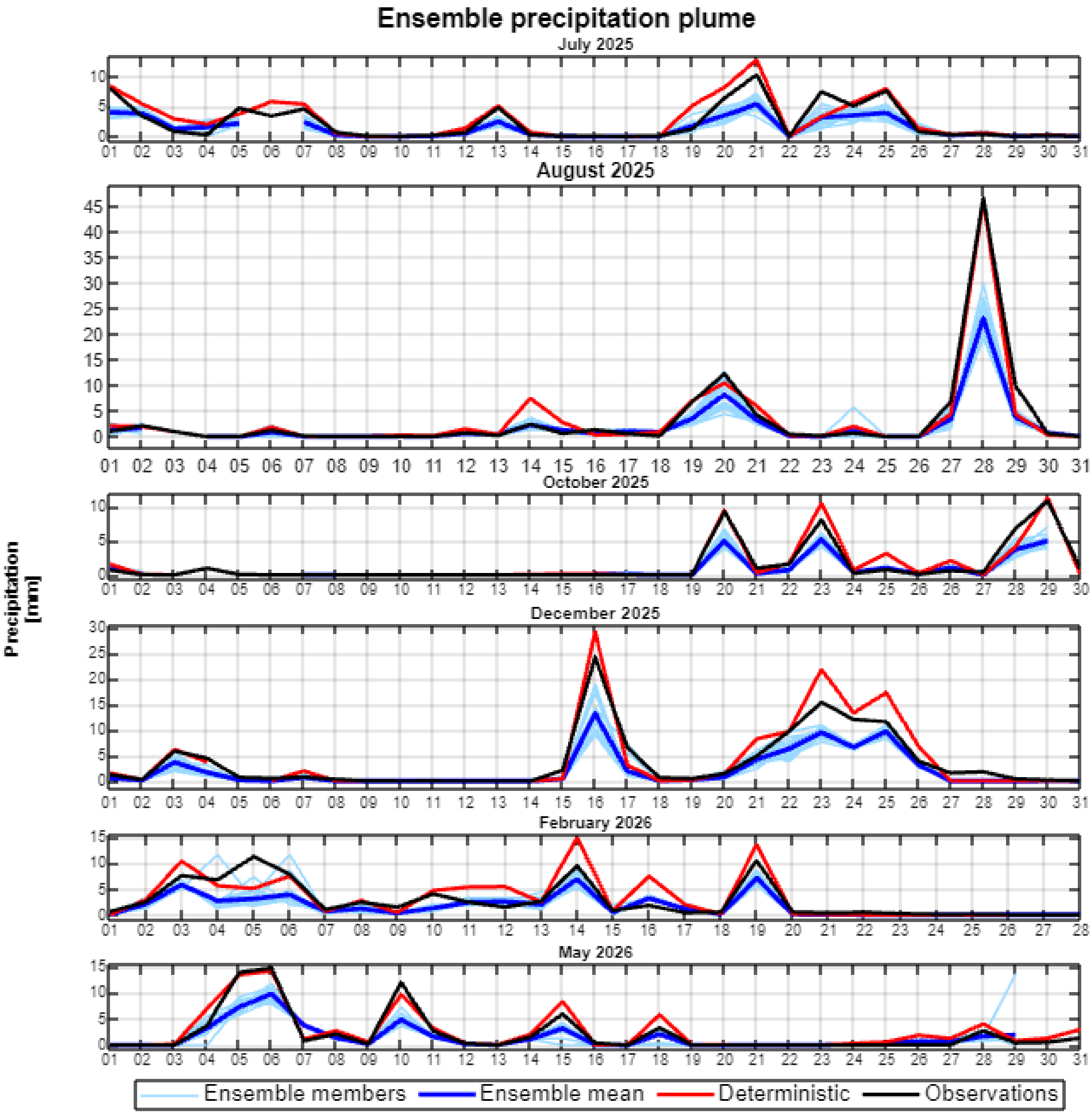

Figure 23. Precipitation time series for the selected months of the study period, showing observed values, ensemble forecasts, ensemble mean and deterministic forecast.

In Figure 23, the ensemble distribution for precipitation shows a limited ability to fully encompass the observed values within its spread, indicating that observations frequently lie outside the ensemble range.

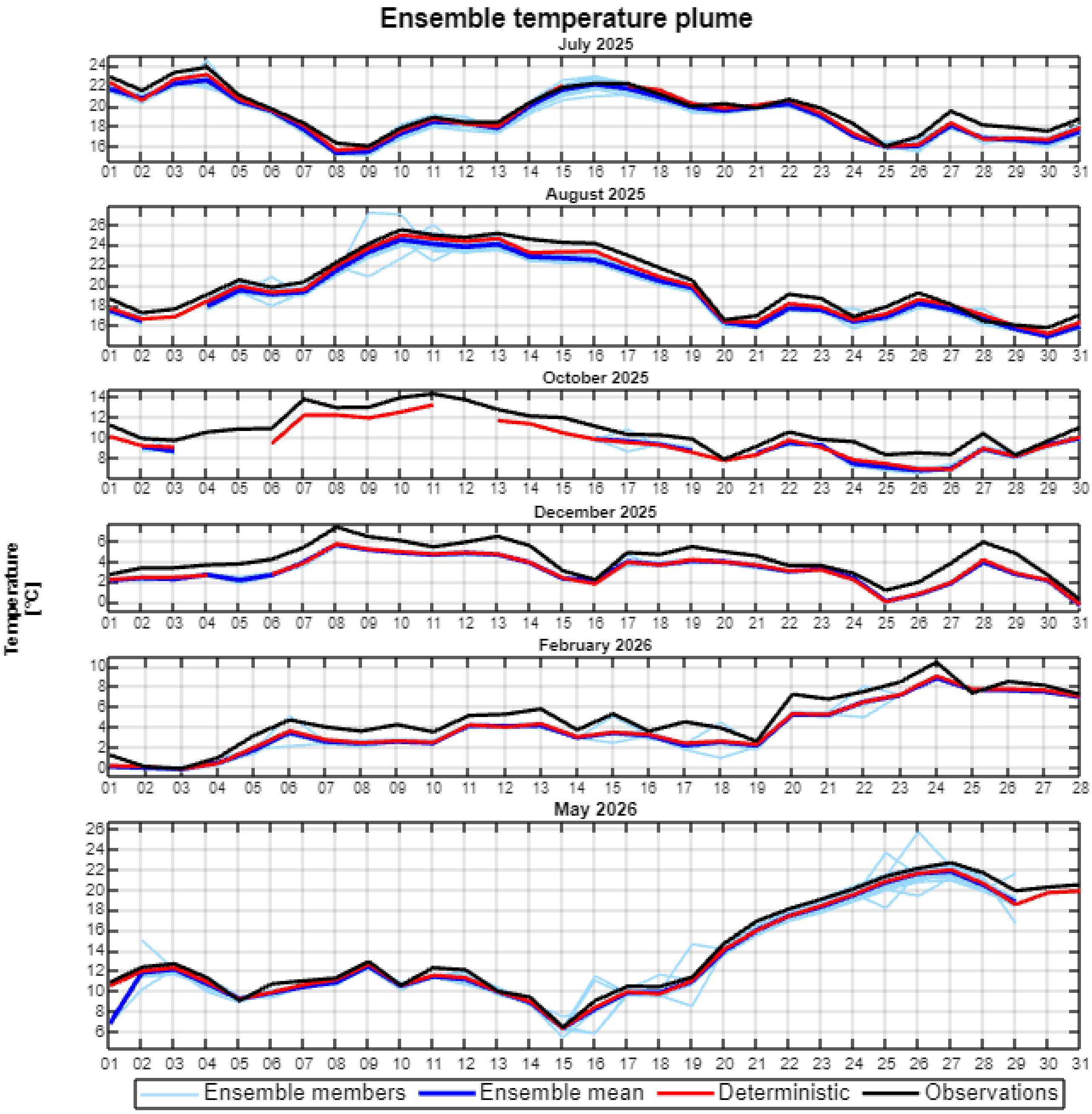

Figure 24. Temperature time series for the selected months of the study period, showing observed values, ensemble forecasts, ensemble mean and deterministic forecast.

In contrast, Figure 24 shows that for temperature the ensemble generally captures the observed variability more consistently, although a tendency for the ensemble to lie below the observations can be observed, suggesting a systematic underestimation.

3.2.2. Comparison with deterministic predictions

Based on a preliminary visual inspection of the ensemble forecasts, a quantitative comparison of the error metrics, MAE, RMSE, and bias, is performed for the precipitation variable between the ensemble and deterministic configurations, as shown in Figure 25. This comparison provides a concise overview of the relative performance of the two forecasting approaches across the analysed months. Overall, the ensemble system generally exhibits slightly higher error values across most metrics, while maintaining a seasonal behaviour similar to that observed for the deterministic forecasts.

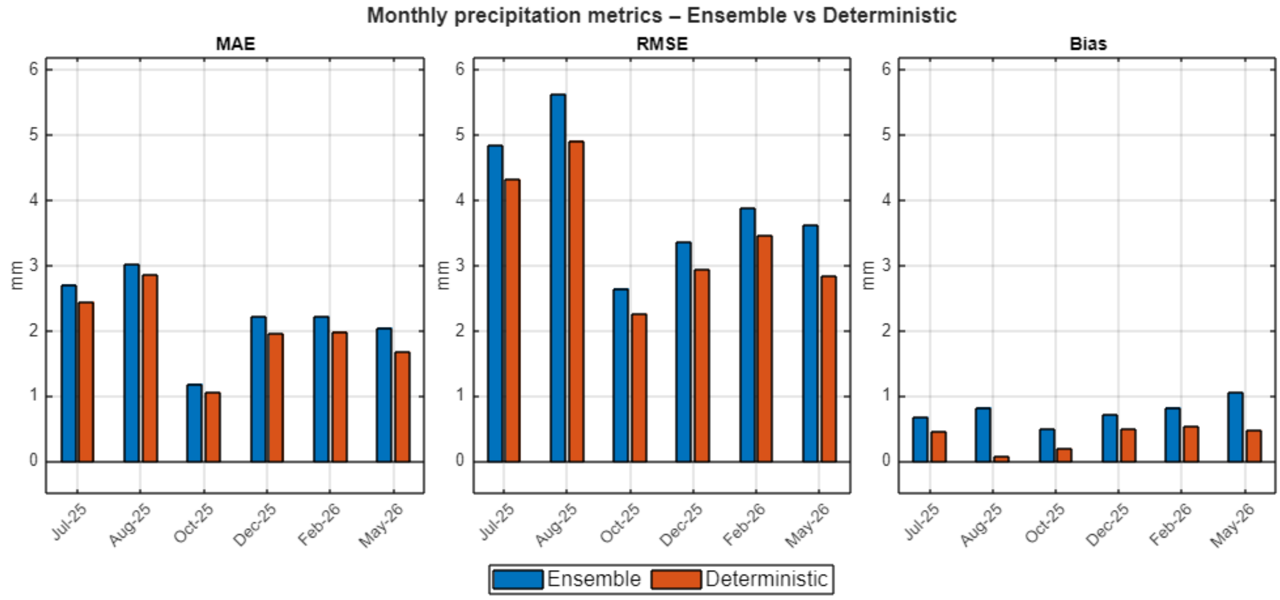

Figure 25. Comparison of ensemble and deterministic error metrics for precipitation.

A similar analysis is then performed for the temperature variable, using the same graphical structure and error metrics, to allow a consistent comparison of ensemble and deterministic performance across different meteorological variables, as shown in Figure 26. Also in this case, the ensemble configuration generally shows slightly higher error values across all three metrics. However, the differences between ensemble and deterministic forecasts are consistently very small, remaining almost negligible in magnitude, indicating a very similar performance between the two configurations for the temperature variable. In more detail, temperature error metrics show very similar values across ensemble and deterministic forecasts for MAE and RMSE, while the bias is the only metric exhibiting a slightly more pronounced separation between the two configurations, although this difference does not exceed approximately 0.5 °C.

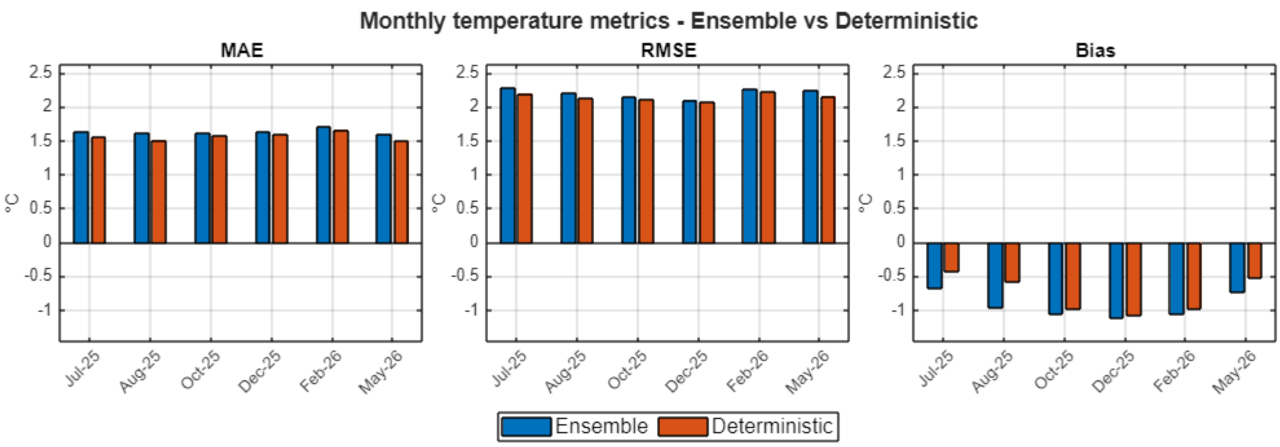

Figure 26. Comparison of ensemble and deterministic error metrics for temperature.

3.2.3. PIT and its time representation

Following the evaluation of deterministic and ensemble error metrics, appropriate diagnostic tools are employed to further assess the ensemble forecasting system. Among these, the Probability Integral Transform (PIT) provides a direct measure of probabilistic calibration by comparing the forecast distribution with the observed outcomes. All spatial grid cells and all available days contribute to the computation of the PIT, ensuring that the diagnostic represents the full spatio-temporal sampling of the dataset.

Figure 27 shows the PIT histogram for precipitation. The distribution exhibits a pronounced U-shaped behaviour, with a higher frequency of values close to 0 and 1 compared to intermediate values. PIT values near 1 are slightly more frequent than those near 0, indicating a mild asymmetry. This pattern suggests that the ensemble spread is somewhat limited, indicating a tendency toward underdispersion, although the overall behaviour remains within an acceptable range.

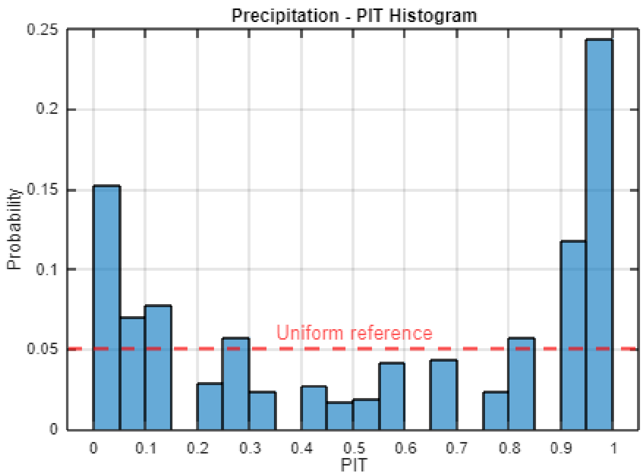

Figure 27. Probability Integral Transform (PIT) histogram for precipitation.

Deviations from an ideal calibration are reflected in the reduced frequency of intermediate PIT values compared to what would be expected under perfect consistency between forecasts and observations. From a statistical perspective, a perfectly calibrated ensemble would produce a uniform PIT distribution, meaning that all probability intervals are equally populated and observations are indistinguishable from ensemble realisations in a probabilistic sense. In this case, perfect uniformity would imply a probability of 0.05 for each PIT bin of width 0.5.

From an interpretative point of view, PIT values can also be related to systematic behaviour: values below 0.5 indicate a tendency to overestimate precipitation, values around 0.5 correspond to good calibration, and values above 0.5 indicate underestimation. The observed structure therefore indicates a generally reasonable probabilistic behaviour, with some residual underdispersion and a non-uniform representation of uncertainty, but without indicating severe miscalibration.

Continuing the analysis of precipitation, Figure 28 shows the temporal evolution of PIT values. Each daily value is obtained by spatially averaging the PIT over the entire study domain. Therefore, the resulting time series does not represent the pointwise PIT distribution shown in the histogram but rather describes the daily average ensemble calibration over the territory with respect to observations. Overall, the values are distributed across the full PIT range, although a slight predominance of values above 0.5 is observed, indicating a systematic tendency towards higher PIT values.

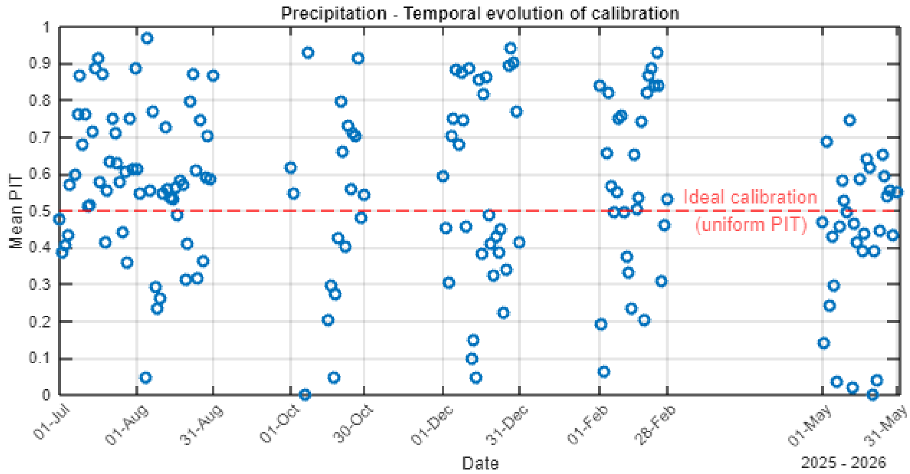

Figure 28. Temporal evolution of spatially averaged PIT values for precipitation, representing the daily domain-mean calibration of the ensemble forecasts.

A similar analysis is performed for the temperature variable to assess the probabilistic calibration of the ensemble forecasts.

Figure 29 shows the PIT histogram for temperature. The distribution departs from the ideal uniform shape and exhibits a broad dome-like pattern, with frequencies increasing progressively from the central PIT values toward higher probabilities. This increase culminates in a pronounced peak at PIT near to 1, while PIT values near 0 also show a local peak, but remain below the uniform reference level and do not constitute a dominant mode. The resulting structure indicates a greater concentration of observations in the upper portion of the ensemble forecast distribution.

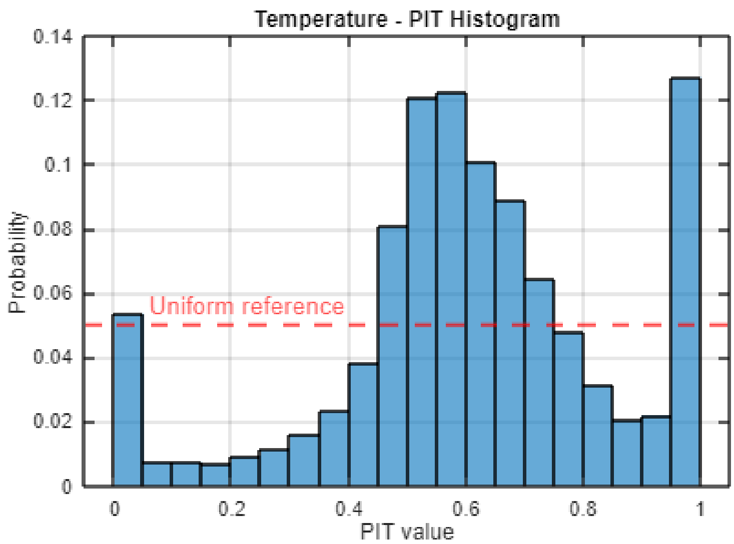

Figure 29. Probability Integral Transform (PIT) histogram for temperature.

Since a perfectly calibrated ensemble would produce a uniform PIT histogram, with all probability intervals equally populated and frequencies close to the reference value, the observed departure from this behaviour indicates that the ensemble does not fully represent the observed temperature variability, particularly in the upper part of the distribution.

From an interpretative perspective, PIT values below 0.5 indicate that the ensemble tends to overestimate temperature, with forecast values generally warmer than the observations. Values around 0.5 correspond to good calibration, while values above 0.5 indicate underestimation, with ensemble forecasts tending to be colder than observed conditions. The predominance of PIT values above 0.5, together with the pronounce PIT peak near to 1, therefore indicates a systematic underestimation and an asymmetric representation of temperature uncertainty. Overall, while the histogram departs from the ideal uniform distribution, the deviations are moderate and mainly concentrated in the upper PIT range.

Continuing the analysis of temperature, Figure 30 shows the temporal evolution of PIT values. As for precipitation, each daily value is obtained by spatially averaging the PIT over the entire study domain, therefore the time series does not represent the pointwise distribution shown in the histogram, but rather the daily average ensemble calibration over the territory with respect to observations.

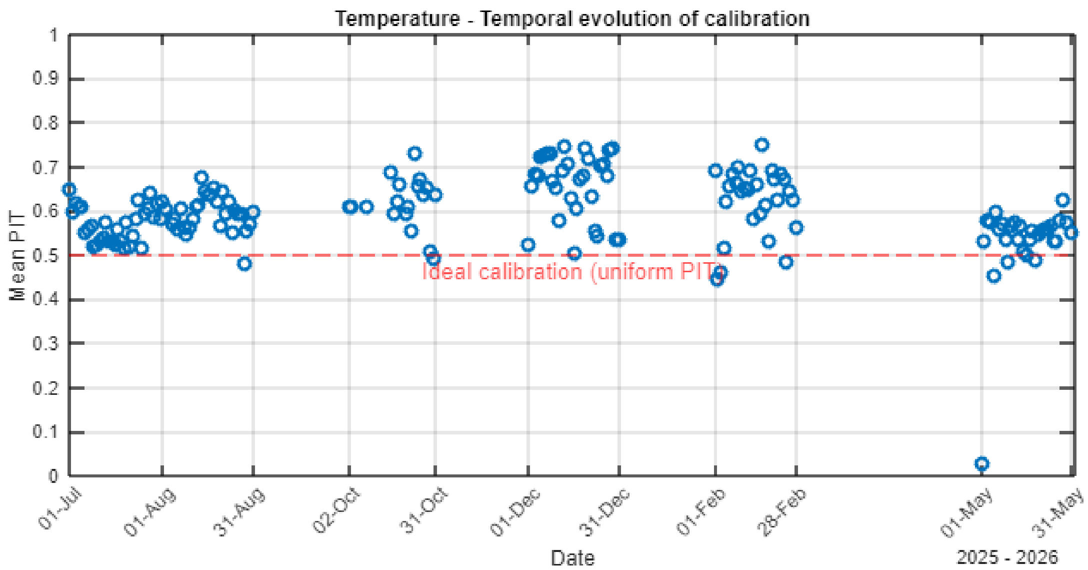

Figure 30. Temporal evolution of spatially averaged PIT values for temperature, representing the daily domain-mean calibration of the ensemble forecasts.

The time series exhibits consistent behaviour across all months, with values predominantly above 0.5 and generally confined within the range between approximately 0.5 and 0.8. No significant seasonal differences are observed, and the same pattern is maintained throughout the analysed period. Values above 0.5 indicate a tendency toward underestimation, meaning that the ensemble forecasts are generally colder than the observations on average. However, this behaviour remains relatively stable in time, without strong temporal variability or pronounced differences between months.

3.2.4. Diagnosis of the forecast

In this section, the diagnostic interpretation of the ensemble forecasts is performed. The analyses presented so far have provided a qualitative and distributional understanding of ensemble behaviour through the ensemble plume and the PIT distribution. In contrast, the probability plot offers a more direct diagnostic tool, where deviations from the reference line can be immediately interpreted in terms of ensemble dispersion and systematic forecast errors.

As previously shown by the uniformity plot of the precipitation variable, the frequency of PIT values is concentrated near the extremes, with an increased occurrence of values close to 0 and 1. According to the diagnostic framework of Laio and Tamea (2007), as visualised in the precipitation PIT probability plot in Figure 31, the precipitation behaviour indicates a narrow forecast, indicating that the ensemble spread is generally insufficient to encompass the observed precipitation variability. This leads to an underdispersed forecast in which observations frequently fall outside the range represented by the ensemble members.

The same diagnostic analysis is applied at both global and monthly scales. The monthly results don’t show substantial deviations from the global results, indicating that the identified patterns are consistent across the different months of the study period.

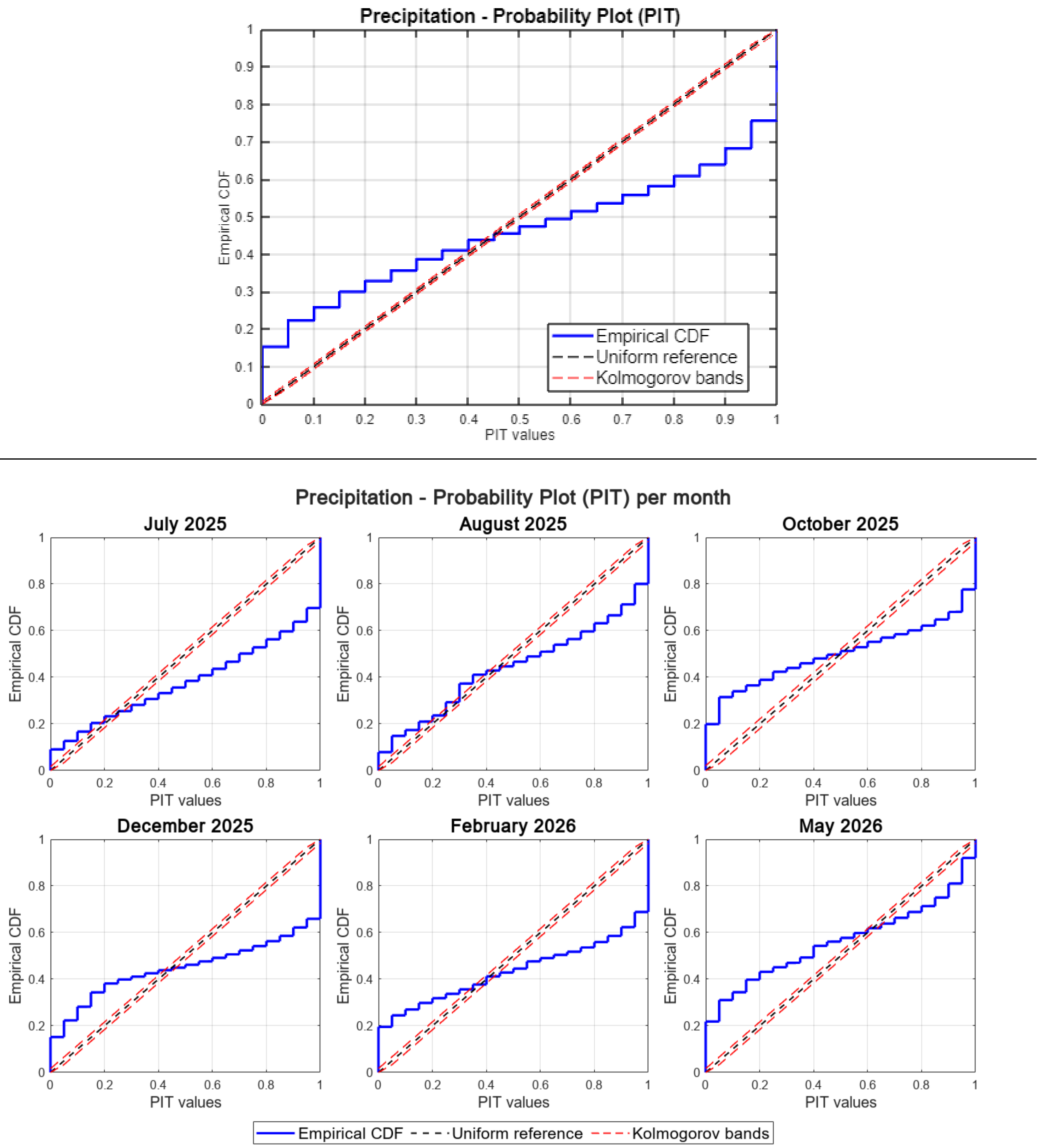

Figure 31. Global and monthly probability plots of precipitation PIT values.

A similar diagnostic approach is then applied to the temperature variable, for which the probability plot exhibits a systematic structure characterized by a displacement of the empirical curve with respect to the reference bisector. The curve remains predominantly above the bisector over a large portion of the distribution, which, according to the diagnostic scheme, is consistent with an underprediction behaviour. This indicates that observations are more frequently located in the upper part of the ensemble distribution and tend to lie above the central tendency of the ensemble forecasts, as shown in Figure 32.

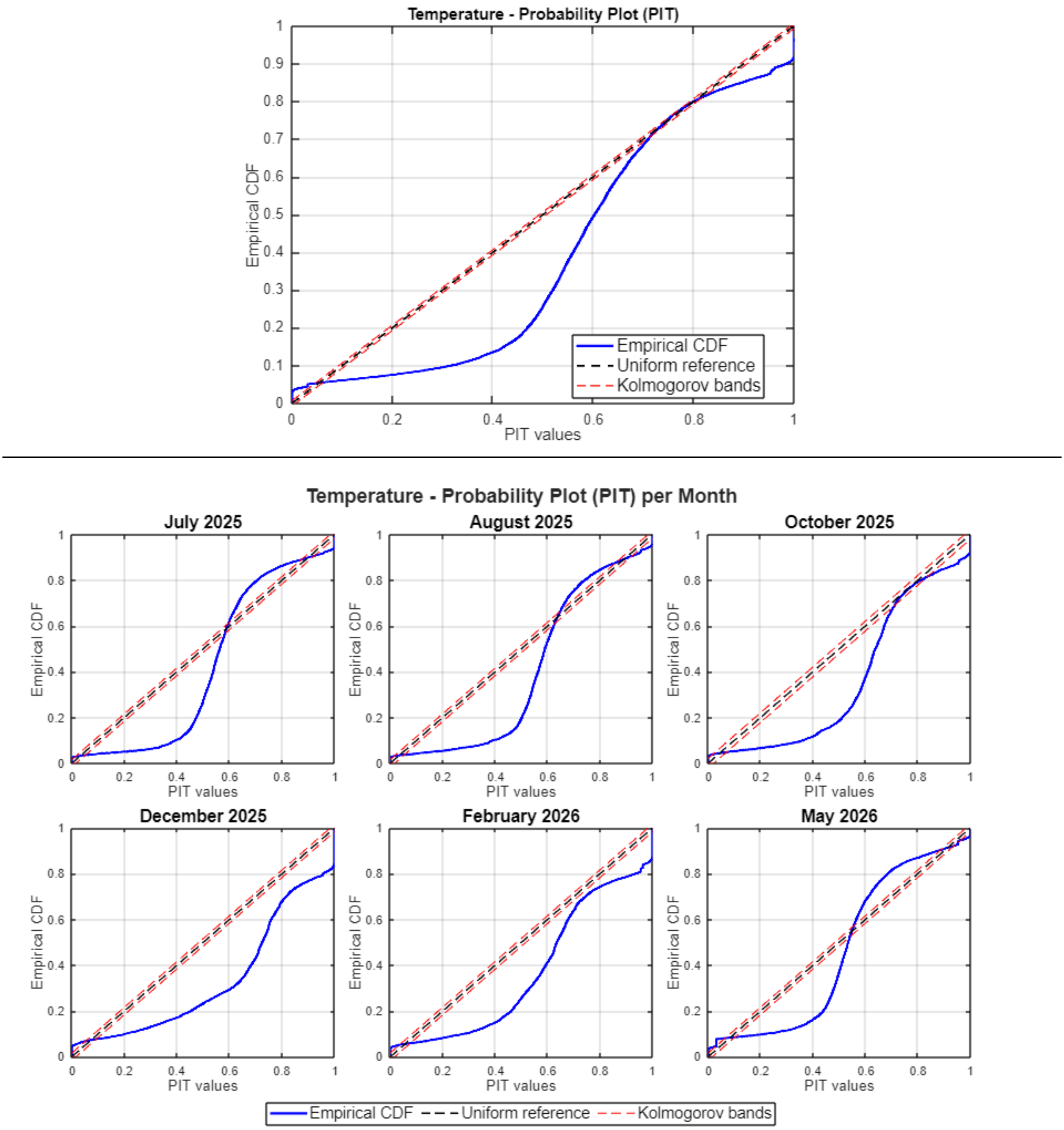

Figure 32. Global and monthly probability plots of temperature PIT values.

An aspect emerging from the comparison between the two variables is the different level of discretisation of the PIT distributions: precipitation exhibits a substantially lower number of unique PIT values compared to temperature, resulting in a more clustered empirical distribution. This behaviour is consistent with the intermittent nature of precipitation and the higher frequency of repeated forecast-observation combinations.

Overall, the shape of the probability plots provides a direct interpretation of the diagnostic characteristics of the ensemble forecasts. Curvature and deviations from the reference line highlight limitations in the representation of forecast variability, while systematic displacements of the curve indicate persistent deviations of the forecast distribution with respect to the observations. In this framework, precipitation is mainly characterized by underdispersion, whereas temperature exhibits a systematic negative displacement of the forecast distribution.

3.2.5. Kendall Tau independence test

The Kendall Tau test is applied to the PIT time series to assess the presence of residual temporal dependence in the ensemble-derived diagnostics. The test is first performed on the complete dataset and subsequently monthly to evaluate the consistency of the results over time. The assessment of independence is based on the standardized Kendall Tau statistic, with values exceeding the critical threshold of ±1.645 indicating rejection of the null hypothesis at the 95% confidence level.

For the precipitation variable, the global test indicates a statistically significant departure from the hypothesis of independence, with a standardized Kendall Tau value equal to 2.8. This result suggests the presence of residual temporal structure in the PIT sequence when considered over the full period.

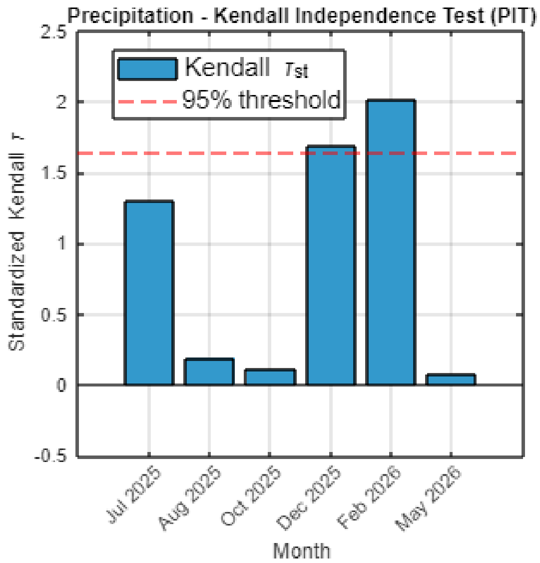

Figure 33. Monthly standardized Kendall Tau statistics for precipitation with independence threshold.

Figure 33 presents the monthly analysis of the Kendall Tau test for precipitation and shows a more variable behaviour. More in detail, the hypothesis of independence is not rejected for July, August, October, and May, indicating no statistically significant temporal dependence in these periods. Conversely, for December and February, the test indicates rejection of independence, suggesting the presence of temporal structure within these specific months.

For the temperature variable, the global test also indicates a statistically significant departure from independence, with a standardized Kendall Tau value equal to 6.9. The monthly results shown in the Figure 34 show a more intermittent behaviour, with statistically significant dependence detected in July and December, while for the remaining months the hypothesis of independence cannot be rejected. October is the only month showing a negative value; however this result should be interpreted with caution due to the reduced sample size associated with missing days in that month, which makes October the weakest period in terms of data availability.

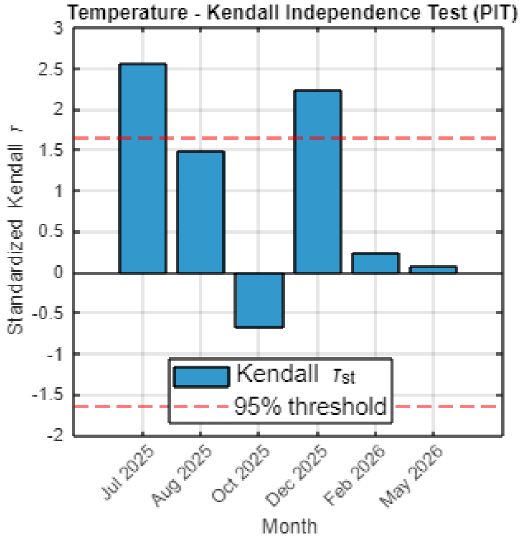

Figure 34. Monthly standardized Kendall Tau statistics for temperature with independence threshold.

Overall, both variables exhibit evidence of residual temporal dependence at the global scale, while the monthly analysis highlights a non-uniform behaviour over time, with periods of independence alternating with periods where temporal structure is detected. This suggests that residual dependence in the PIT series is present at the global scale throughout the entire study period, while showing more variability at the monthly scale.

4. Discussion and conclusions

4.1. Discussion of the results

This study provides a comprehensive verification of deterministic and ensemble forecasts of daily precipitation and temperature over the study domain. By analysing forecast performance with respect to forecast lead time, seasonality, terrain complexity, and the magnitude of the observed variables, the results provide a detailed characterization of the forecast errors and of the conditions under which they become more pronounced.

The analyses reveal distinct error characteristics for precipitation and temperature. Forecast performance is primarily influenced by forecast lead time, seasonality, terrain complexity, and the magnitude of the observed variables, although the relative importance of these factors differs between the two forecast variables.

The deterministic verification reveals markedly different behaviours for precipitation and temperature. Precipitation forecasts exhibit a clear dependence on forecast lead time, with forecast accuracy generally improving as the verification time approaches. This behaviour is particularly evident during the cold season, whereas summer months show comparatively limited improvement and consistently larger forecast errors. Such differences likely reflect the greater predictability of large-scale winter synoptic systems compared with the highly localized and rapidly evolving convective processes that dominate summer precipitation. Conversely, temperature forecasts remain remarkably stable across the entire forecast horizon. Both MAE and RMSE vary only slightly with lead time, indicating that forecast errors remain relatively constant even several days before verification. This behaviour suggests that near-surface temperature is represented more consistently than precipitation, owing to its stronger dependence on large-scale atmospheric conditions and lower spatial variability.

Terrain complexity emerges as one of the dominant factors controlling forecast performance. For both variables, forecast errors increase progressively from plains to mountainous areas, confirming the greater difficulty of representing atmospheric processes over complex topography. However, the nature of these errors differs between precipitation and temperature. While precipitation forecasts tend to exhibit a systematic positive bias over mountainous regions, indicating an overestimation of accumulated rainfall, temperature forecasts display an increasingly pronounced cold bias with increasing orographic complexity.

The conditional analyses further demonstrate that forecast performance depends strongly on the magnitude of the observed variables. For precipitation, absolute errors increase with rainfall intensity, whereas relative errors decrease. This behaviour indicates that intense precipitation events remain difficult to predict accurately in absolute terms, although forecast quality improves when evaluated relative to the magnitude of the observed rainfall. In contrast, temperature forecasts exhibit a systematic regression toward the mean, characterized by the overestimation of lower temperatures and underestimation of higher temperatures. Consequently, the forecasted temperature range is narrower than the observed one, suggesting that extreme thermal conditions remain more challenging to reproduce than average conditions.

The analyses also reveal a clear seasonal modulation of forecast performance. When comparable precipitation or temperature classes are considered, July and August consistently exhibit the largest forecast errors, whereas winter months generally produce more accurate predictions. This seasonal behaviour is likely associated with the transition from predominantly stratiform precipitation during winter to convective precipitation during summer, together with the increased spatial and temporal variability characteristic of warm-season atmospheric processes. Moreover, the interaction between seasonality and terrain demonstrates that forecast errors cannot be attributed solely to orographic complexity but instead arise from the combined influence of topography and the prevailing meteorological conditions.

The ensemble verification complements the deterministic analysis by providing insight into the probabilistic behaviour of the forecasting system. Overall, the deterministic forecasts achieve slightly lower MAE, RMSE, and bias values than the ensemble mean, although the differences remain relatively small. This indicates that the ensemble mean reproduces deterministic forecast performance while additionally providing information on forecast uncertainty. The probabilistic diagnostics reveal that the ensemble is not perfectly calibrated. For precipitation, the U-shaped PIT distribution indicates moderate underdispersion, suggesting that the ensemble spread is frequently insufficient to encompass the observed variability. This behaviour is particularly relevant for high-impact precipitation events, where forecast uncertainty is expected to be largest. Temperature forecasts also depart from ideal calibration, with PIT diagnostics confirming the presence of a systematic cold bias that is consistent with the deterministic verification.

Finally, the Kendall's Tau independence analysis demonstrates that residual temporal dependence exists when the entire verification period is considered. However, the monthly analyses reveal that this dependence is not persistent throughout the study period, suggesting that the statistical characteristics of the ensemble forecasts vary according to the prevailing atmospheric conditions rather than reflecting a systematic weakness of the forecasting system.

Overall, the verification highlights that forecast errors are primarily controlled by the interaction of seasonality, terrain complexity, forecast lead time, and the magnitude of the observed variables. Precipitation errors are mainly associated with warm-season convection, complex terrain, and high-intensity events, whereas temperature errors are characterized by a persistent cold bias over mountainous regions and a tendency to underestimate the observed temperature variability. The ensemble analyses further show that the probabilistic forecasts reproduce the main deterministic error patterns while providing additional information on forecast uncertainty, although moderate underdispersion for precipitation and residual calibration deficiencies for temperature remain evident.

4.2. Conclusions

This study presented a comprehensive verification of deterministic and ensemble forecasts of daily precipitation and temperature, identifying the main factors controlling forecast performance across different environmental conditions.

The results highlight clear and consistent differences between the two variables. Precipitation forecasts are strongly influenced by forecast lead time, seasonality, terrain complexity, and event intensity, with larger errors occurring under convective summer conditions, in mountainous regions, and for high-intensity events. Temperature forecasts instead exhibit a more stable behaviour over lead time but are affected by systematic biases related to terrain complexity and a reduced variability compared with observations.

Overall, forecast errors are primarily structured by the interaction between topography and prevailing atmospheric conditions, which governs the spatial and temporal variability of forecast performance.

The ensemble analysis confirms the main deterministic error patterns while providing additional information on forecast uncertainty. However, residual issues remain in terms of underdispersion for precipitation and systematic cold bias for temperature, indicating limitations in the calibration of the probabilistic forecasts.

4.3. Limitations of the research

The results of this study should be interpreted considering some limitations that affect the robustness and generality of the conclusions.

The main limitation is the relatively short verification period. At the time this thesis was conducted, only one year of forecasts and observations was available, limiting the ability of the assessment to capture interannual variability. Consequently, the results may be influenced by the specific meteorological conditions that occurred during the selected period. Individual high-impact precipitation events may have a non-negligible impact on the computed statistics. Furthermore, the analyses presented in this thesis were based on only six months of data, which further limits the statistical representativeness of the findings. Nevertheless, the study provides an initial assessment of the forecasting system and establishes a foundation for future analyses based on longer verification periods.

A further aspect related to the available datasets concerns the temporal consistency between forecasts and observations. Forecast data are provided in Coordinated Universal Time (UTC), whereas observations are referenced to local time, UTC+1 or UTC+2 depending on daylight saving time. Consequently, the first daily accumulation period cannot always be perfectly aligned, and the closest available time window was adopted. This results in a small temporal mismatch, on the order of one to two hours, in the first verification period, which is however limited in magnitude.

Regarding temperature, only daily minimum and maximum values are directly available, while daily mean temperature is derived as their arithmetic average. Although this is a common and consistent approximation, it does not fully represent the true diurnal temperature cycle and may introduce small deviations under strongly asymmetric daily temperature conditions.

Finally, the verification is based on daily aggregated variables, namely daily precipitation totals and daily mean temperature. While this temporal resolution is suitable for assessing overall forecast performance, it does not allow the investigation of sub-daily variability or the internal structure of precipitation events. Moreover, the spatial aggregation applied in parts of the analysis may reduce the representation of local-scale variability.

4.4. Possible future developments

Future developments of this work primarily point towards the operational application of weather forecasts in water resource and irrigation management. A key objective is the development of decision-support tools for institutions responsible for irrigation management, based on the quantitative use of meteorological data and existing forecasting systems. The aim is to support decision-making processes in water resource management, promoting more informed planning of irrigation practices and reducing inefficiencies associated with partially empirical water management strategies.

In this context, meteorological forecasts can be integrated into irrigation requirement models capable of describing the hydrological balance of the system. Such an approach would allow observed and forecasted precipitation and temperature data to be directly used within operational frameworks, improving the link between meteorological information and agricultural water demand.

In parallel, continuous developments in numerical weather prediction systems, particularly in model physics, data assimilation, and ensemble design, are expected to progressively improve forecast quality over time. These improvements may contribute to reducing systematic errors and enhancing the representation of forecast uncertainty, thereby increasing the potential applicability of meteorological forecasts in operational contexts. Moreover, the availability of longer forecast datasets in the future will allow more comprehensive verification analyses, improving the assessment of forecast performance over a wider range of meteorological conditions.

Overall, the methodology developed in this study provides a foundation for linking meteorological forecast verification with practical applications in sustainable agriculture. The final objective is to translate weather forecasts into operational indicators that support irrigation planning, contributing to more efficient and resilient water resource management under increasing climate variability.

Bibliography

Berkovitz, J. (2012). Testing Density Forecasts, With Applications to Risk Management.

CERRA sub-daily regional reanalysis data for Europe on single levels from 1984 to present. (n.d.). Retrieved from Copernicus Climate Data Store (CDS): https://cds.climate.copernicus.eu/datasets/reanalysis-cerra-single-levels?tab=overview

CNR-ISAC. (n.d.). *GLOBO - BOLAM - MOLOCH forecasts*. Retrieved from Previsioni meteorologiche CNR-ISAC: https://www.isac.cnr.it/dinamica/projects/forecasts/

Copernicus. (n.d.). *Limited Area Ensemble Forecasting (ALADIN-LAEF)*. Retrieved from Copernicus: https://asr.copernicus.org/articles/16/63/2019/

Dataset su griglia NWIOI. (n.d.). Retrieved from Arpa Piemonte: https://www.arpa.piemonte.it/scheda-informativa/dataset-griglia-nwioi

Dati osservati. (n.d.). Retrieved from MeteoItalia: https://www.agenziaitaliameteo.it/meteo/

ECMWF. (n.d.). *IFS Medium-range Control forecast (formerly HRES) and Analysis Data (15-days) (Set I)*. Retrieved from ECMWF forecast Datasets: https://www.ecmwf.int/en/forecasts/datasets/set-i

ECMWF. (n.d.). *The mesoscale ensemble prediction system A-LAEF*. Retrieved from ECMWF: https://www.ecmwf.int/en/newsletter/172/earth-system-science/mesoscale-ensemble-prediction-system-laef

EMO: A high-resolution multi-variable gridded meteorological data set for Europe. (n.d.). Retrieved from data.europa.eu: https://data.europa.eu/data/datasets/0bd84be4-cec8-4180-97a6-8b3adaac4d26?locale=en

E-OBS daily gridded meteorological data for Europe from 1950 to present derived from in-situ observations. (n.d.). Retrieved from Climate Data Store: https://cds.climate.copernicus.eu/datasets/insitu-gridded-observations-europe?tab=overview

ERA5 hourly data on single levels from 1940 to present. (2018). Retrieved from Climate Data Store: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview

European Commission Joint Research Centre. (n.d.). *EMO-1arcmin dataset*. Retrieved from JRC Data Catalogue: https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/CEMS-EFAS/meteorological_forcings/README-EMO-1arcmin.txt

Jolliffe, I., & Stephenson, D. (2003). *Forecast Verification: A Practitioner's Guide in Atmospheric Science.*

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics.*

Laio, F., & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous. *Hydrology and Earth System Sciences*.

Matsuura, K., & Willmott, C. (2005). Advantages of the mean absolute error (MAE) over.

meteoblue. (n.d.). *Datasets* . Retrieved from meteoblue Technical Documentation: https://docs.meteoblue.com/en/meteo/data-sources/datasets

Metodologia interpolazione NWIOI. (n.d.). Retrieved from Arpa Piemonte: https://www.arpa.piemonte.it/export/sites/default/tematismi/clima/confronti-storici/dati/metodologia.pdf

MISTRAL. (n.d.). Retrieved from About Mistral: https://www.mistralportal.it/about-mistral/#

Mombrini, E., Tamea, S., Viglione, A., & Revelli, R. (2023). A 60-year drought analysis of meteorological data in the western Po River basin. *Hydrology and Earth System Sciences*.

Monaco, S., Monaco, L., Apiletti , D., Cremonini , R., & Barbero, S. (2025). Uncertainty-aware methods for enhancing rainfall prediction with deep-learning based post-processing segmentation. *ScienceDirect*. Retrieved from Uncertainty-aware methods for enhancing rainfall prediction with deep-learning based post-processing segmentation - ScienceDirect

National Centers for Environmental Information (NCEI), N. (n.d.). *World Weather Records*. Retrieved from National Centers for Environmental Information (NCEI), NOAA: https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc%3AC00160

National Centers for Environmental Information (NCEI), NOAA. (n.d.). *World Weather Records*. Retrieved from National Centers for Environmental Information (NCEI), NOAA: https://www.ncei.noaa.gov/products/world-weather-records

NOAA. (n.d.). *NOAA Global Forecast System (GFS)*. Retrieved from Registry of Open Data on AWS: https://registry.opendata.aws/noaa-gfs-bdp-pds/

NOAA/NCEP/EMC. (n.d.). *GFS: Global Forecast System*. Retrieved from Earth Engine Data Catalog: https://developers.google.com/earth-engine/datasets/catalog/NOAA_GFS0P25?hl=it

Numerical weather prediction. (n.d.). Retrieved from Cordulus: https://www.cordulus.com/glossary/numerical-weather-prediction

Open-meteo. (n.d.). *Model Updates and Data Availability*. Retrieved from Open-meteo: https://open-meteo.com/en/docs/model-updates

Publications Office of the European Union. (n.d.). *Meteorological forecasting model (AROME) - historical analysis data*. Retrieved from European data: https://data.europa.eu/data/datasets/9d8a1894-9f6c-4d74-9d2e-d481f8737ced?locale=en

Publications Office of the European Union. (n.d.). *Weather forecast for HARMONIE-AROME*. Retrieved from European data: https://data.europa.eu/data/datasets/56eac114-43e9-4cae-922d-d4f486728fb1?locale=en

Wetterdienst, D. (n.d.). *NWP forecast data* . Retrieved from Deutscher Wetterdienst: https://www.dwd.de/EN/ourservices/nwp_forecast_data/nwp_forecast_data.html

Wilks, D. (2011). *Statistical Methods in the Atmospheric Sciences.* 3rd Edition, Elvisier.

Acknowledgements

Raggiunto questo importante traguardo del mio percorso universitario, desidero dedicare alcune parole di gratitudine alle persone che, in modi diversi, ne hanno fatto parte.

Prima di tutto, ringrazio i miei professori Stefania Tamea e Alberto Viglione per aver accolto la mia idea di svolgere una tesi su uno degli argomenti che più mi hanno appassionata durante i miei studi. Vi ringrazio per avermi dato questa opportunità e per avermi guidata nello sviluppo di un lavoro di cui vado molto fiera.

A questo punto, ovviamente, un grande grazie va ai miei genitori, mamma e papà. In modi diversi mi avete accompagnata lungo questo percorso, tortuoso ma anche il più bello della mia vita. Mi avete dato fiducia e avete avuto pazienza e adesso non ci resta che festeggiare insieme per un traguardo che, per molto tempo, non mi credevo capace di raggiungere.

Vitto, a te semplicemente grazie, perché sei la prima persona che mi viene in mente quando ho una notizia da condividere. So che con te ogni mia emozione viene ascoltata, vissuta e condivisa. Hai sempre saputo vivere i miei progressi e traguardi come se fossero anche un po' tuoi. Anche questa laurea, quindi, non poteva che essere condivisa con te, e io non posso che essere grata alla vita per avermi regalato un'amica e una sorella come te.

Un grazie di cuore ad Elia, che già dal giorno prima del test di ingresso è stato una spalla su cui piangere e la persona su cui trovare sostegno e conforto, ma soprattutto con cui condividere la felicità e le emozioni che hanno accompagnato questo percorso.

Un pensiero speciale va alle mie amiche Cali, Ias e Sonia, per avermi sempre sostenuta con affetto e per aver saputo essere presenti con sincerità in ogni momento, trovando sempre una parola di incoraggiamento.

Grazie alla mia grande famiglia, che è sempre stata orgogliosa di vedermi raggiungere i miei obiettivi, pur continuando a chiedermi che cosa studiassi esattamente e quale lavoro potrò fare in futuro. Un pensiero affettuoso va a nonno Silvano, a cui dedico questa tesi e che non mi ha mai fatto mancare le sue chicche di ingegneria senza formule.

Un abbraccio va anche alla mia seconda famiglia, quella che mi ha fatto conoscere Elia e che mi ha accolta facendomi sentire sempre a casa.

Ai miei compagni di università e, in particolare, ad Eriola e Lorenzo un grazie perché, anche senza vederci spesso, ho sempre saputo di poter contare su di voi e per aver condiviso con me le avventure e le vicissitudini che l'università ci ha dato l'opportunità di vivere.

Infine, grazie a me, per aver avuto la volontà e il coraggio di mettermi sempre alla prova con nuove sfide e avventure che mi hanno arricchita e permesso di crescere.