Parametric post-processing of ensemble forecasts across multiple weather variables and resolutions

Thesis for the Degree of Doctor of Philosophy (PhD)

by Marianna Lakatos-Szabó

Supervisor:
Prof. Dr. Sándor Baran

UNIVERSITY OF DEBRECEN
Doctoral Council of Natural Sciences and Information Technology
Doctoral School of Informatics
Debrecen, 2024

Hereby I declare that I prepared this thesis within the Doctoral Council of Natural Sciences and Information Technology, Doctoral School of Informatics, University of Debrecen in order to obtain a PhD Degree in Informatics at the University of Debrecen.

The results published in the thesis are not reported in any other PhD theses.

Debrecen, 202. . . . . .

signature of the candidate

Hereby I confirm that Marianna Lakatos-Szabó candidate conducted her studies with my supervision within the Theoretical Foundation and Applications of Information Technology and Stochastic Systems Doctoral Program of the Doctoral School of Informatics between 2019 and 2023. The independent studies and research work of the candidate significantly contributed to the results published in the thesis.
I also declare that the results published in the thesis are not reported in any other PhD theses.
I support the acceptance of the thesis.

Debrecen, 202. . . . . .

signature of the supervisor

Parametric post-processing of ensemble forecasts across multiple weather variables and resolutions

Dissertation submitted in partial fulfilment of the requirements
for the doctoral (PhD) degree in Informatics

Written by Marianna Lakatos-Szabó, certified computer scientist

Prepared in the framework of the Doctoral School of Informatics,
University of Debrecen
(Theoretical Foundation and Applications of Information Technology and
Stochastic Systems programme)

Thesis advisor: Prof. Dr. Sándor Baran

The official opponents of the dissertation:

Dr. ....
Dr. ....
Dr. ....

The evaluation committee:

chairperson: Dr. ....
members: Dr. ....
Dr. ....
Dr. ....
Dr. ....

The date of the dissertation defence: ..... 202...

List of abbreviations

ALADIN-HUNEPS	Aire Limitée Adaptation Dynamique Développement International-Hungary Ensemble Prediction System
BFGS	Broyden–Fletcher–Goldfarb–Shanno
BMA	Bayesian model averaging
BS	Brier score
BSS	Brier skill score
CDF	Cumulative distribution function
CRPS	Continuous ranked probability score
CRPSS	Continuous ranked probability skill score
CSG	Censored and shifted gamma
DM	Diebold-Mariano
ECMWF	European Centre for Medium-Range Weather Forecasts
EFAS	European Flood Awareness System
EMOS	Ensemble model output statistics
EPS	Ensemble prediction system
GEV	Generalized extreme value
HPC	High-performance computing

IFS	Integrated Forecast System
JJA	June-July-August
LHPC	Large high-performance computing
LN	Log-normal
MAE	Mean absolute error
ML	Maximum likelihood
MOS	Model output statistics
NWP	Numerical weather prediction
PDF	Probability density function
PIT	Probability integral transform
QM	Quantile-mapped
QM+W	Weighted quantile-mapped
QS	Quantile score
QSS	Quantile skill score
RMSE	Root mean squared error
SHPC	Small high-performance computing
SYNOP	Synoptic observations
TGEV	Truncated generalized extreme value
TN	Truncated normal
twCRPS	Threshold-weighted continuous ranked probability score
twCRPS	Threshold-weighted continuous ranked probability skill score
UWME	University of Washington mesoscale ensemble
VRH	Verification rank histogram

Contents

List of Figures

List of Tables

Introduction

Ensemble weather forecasting has emerged as a transformative approach in the field of meteorology. By combining multiple runs of numerical weather prediction models, ensemble forecasts provide valuable probabilistic predictions that help capture the uncertainty inherent in weather forecasting. These forecasts have proven instrumental in enhancing the accuracy of weather predictions and supporting decision-making processes in various sectors. However, the raw output from ensemble forecasts often exhibits certain limitations, such as underdispersion and bias, which can adversely affect their reliability and usability.

The spatial resolution and ensemble size are two critical factors that significantly impact the performance of ensemble forecasts. Spatial resolution refers to the level of detail at which weather models represent atmospheric phenomena, while ensemble size refers to the number of members within an ensemble. The choice of spatial resolution affects the level of computational resources required for running the models and the forecast accuracy. Similarly, the ensemble size influences the representativeness and spread of the ensemble members, thereby affecting the reliability of the forecast. Striking the right balance between spatial resolution and ensemble size is crucial to optimize forecast performance while managing computational costs.

In the context of ensemble weather forecasting, the concept of dual-resolution ensembles has gained attention in recent years. Dual-resolution ensembles combine members with different spatial resolutions, enabling a trade-off between computational cost and forecast skill. This approach has the potential to enhance the accuracy of ensemble forecasts while maintaining reasonable computational requirements. However, there is a need to systematically assess the impact of dual-resolution ensembles on forecast performance and determine the optimal configurations that yield the best results.

Statistical post-processing techniques play a vital role in improving the skill and reliability of ensemble forecasts. These techniques involve the calibration

of raw ensemble outputs to correct for biases and enhance their predictive capabilities. While there has been extensive research on statistical post-processing for single-resolution ensembles, the application and evaluation of different calibration methodologies in the context of dual-resolution ensemble predictions are relatively limited. Understanding how statistical post-processing can effectively calibrate dual-resolution ensemble forecasts is crucial to harnessing the full potential of this approach.

One of the primary objectives of this thesis is to investigate the impact of statistical post-processing on dual-resolution ensemble forecasts. Specifically, the research assesses the effectiveness of different calibration methodologies in improving forecast accuracy and reliability. By comparing and evaluating various post-processing techniques, the purpose of this study is to contribute to a better understanding of the calibration requirements and challenges associated with dual-resolution ensemble forecasts. Moreover, this research seeks to explore the optimal balance between spatial resolution and ensemble size in dual-resolution ensembles. The study aims to provide insights into the trade-offs between forecast skill, computational cost, and the configurations of dual-resolution ensemble predictions by conducting comprehensive experiments and performance evaluations.

The thesis also provides multiple case studies on the application and validation of a novel truncated version of the generalized extreme value distribution-based nonhomogeneous regression model for the purpose of calibrating wind speed forecasts in order to improve their forecast skill. The aim is to correct the deficiency of the otherwise efficient generalized extreme value distribution-based method of Lerch and Thorarinsdottir (2013) of occasionally predicting negative wind speed.

This thesis is structured as follows to address the research objectives outlined above. Chapter 1 provides a comprehensive review of the literature on ensemble weather forecasting, dual-resolution ensembles, and statistical post-processing techniques. It establishes the theoretical foundation and contextualizes the research within the existing body of knowledge.

Chapter 2 presents the methodology employed in these studies, including the different modelling techniques applied to different weather variables, such as temperature, wind speed and precipitation accumulation. This chapter also gives details about the parameter estimation process, the spatial considerations for choosing the training data, and the evaluation metrics to validate the results. It describes the framework used for statistical post-processing and the calibration methodologies applied to the dual-resolution ensemble forecasts.

Chapter 3 explores the selection of an appropriate statistical model for the

calibration of wind speed forecasts. Drawing on the non-homogeneous regression approach introduced by Gneiting et al. (2005), various parametric models for wind speed are compared. To provide a thorough comparison to our novel truncated generalized extreme value distribution-based model, we assess the conventional model, which relies on truncated normal distribution (Thorarinsdottir and Gneiting, 2010), and other models based on log-normal and generalized extreme value distributions. All of the findings can be seen in Baran et al. (2021).

We investigate the case studies of dual-resolution temperature and precipitation accumulation forecast calibrations in Chapter 4 and 5, respectively. In the articles by Baran et al. (2019) and Szabó et al. (2023), we provide the full comprehensive analysis of these two studies conducted. Apart from the calibration of the ensemble forecasts, we also address the questions regarding the balance between spatial resolution and ensemble size in dual-resolution ensembles within the limitations of the available data.

Finally, Chapter 6 summarizes the key findings of the study, discusses their implications, and provides recommendations for future research. This chapter serves as a concluding section of the thesis, highlighting the importance of advancing research in two key areas: dual-resolution ensemble prediction systems and the development and evaluation of distribution-based models for challenging weather variables. By delving deeper into these subjects, a more comprehensive understanding of the underlying characteristics can be attained, leading to improved predictive performance in future forecasts. The chapter emphasises the need for further exploration and investigation in order to maximize the benefits of these approaches and enhance the accuracy of weather predictions.

Chapter 1

Literature review

Weather forecasting is essential for many areas of society, from agriculture to transportation, energy, and disaster response. Numerical weather prediction (NWP) models, which simulate the behaviour of the atmosphere using mathematical equations, have made significant advances in recent decades, but they still have limitations and uncertainties that affect their accuracy. Post-processing techniques aim to improve forecast skill by correcting biases, reducing errors, and enhancing the information content of the forecast. The development of post-processing methods has been fuelled by the increasing availability of observations, the growth of computational resources, and the need for more reliable and informative forecasts. This review of the literature provides an overview of ensemble forecasting methods and the main approaches to weather forecast post-processing, with the main focus being on statistical methods, as well as their applications and challenges. By synthesizing and analyzing the existing research, this review aims to identify the current state of the field, the gaps in knowledge, and the future directions for research and applications in the post-processing of weather forecasts.

1.1 Ensemble forecasting

The numerical prediction of weather variables is a critical tool in modern meteorology, providing forecasts of atmospheric variables for various time frames and geographical areas. NWP models are based on the physical laws that govern the atmosphere, the land surface, and the ocean, including the laws of thermo-

dynamics, fluid dynamics, and radiation (Kalnay, 2003). These models divide the atmosphere into a grid of discrete points, each representing a set of atmospheric variables such as temperature, pressure, and humidity. The equations that describe the evolution of these variables are solved numerically at each grid point, with the results being combined to create a forecast.

According to the comprehensive first chapter of Vannitsem et al. (2018) and the eighth chapter of Wilks (2019), one of the most significant advances in NWP has been the development of ensemble forecasting, which involves running multiple models with slightly different initial conditions or model configurations. These various runs produce a range of forecasts from which statistical information can be derived, such as the probability of certain weather events occurring. In Figure 1.1, it becomes clear how the level of uncertainty increases with the increase of the lead time.

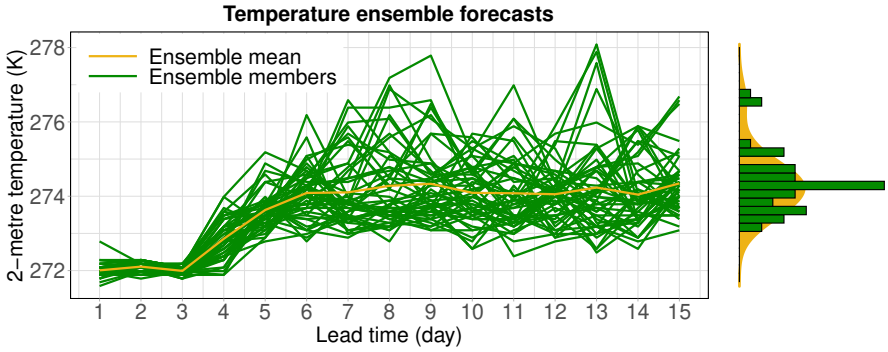

Figure 1.1: Flame diagram of the 50-member ECMWF ensemble forecast of 2-metre temperature and the corresponding ensemble mean for 1 June 2016 initialised at 0000 UTC for lead times from 1 through 15 days.

The fundamental paper by Lorenz (1963) demonstrated that solutions to systems of deterministic nonlinear differential equations could exhibit sensitive dependence on initial conditions. Despite the deterministic nature of the equations, the computed solutions can diverge strongly from each other when initiated from slightly different initial conditions. This phenomenon of sensitive dependence on initial conditions was later coined "chaotic dynamics" by Li and Yorke (1975). The sensitivity to initial conditions in atmospheric modelling

presents a fundamental limitation in forecasting the weather accurately beyond a few days. As proposed by Edward Lorenz, it is impossible for long-range forecasts—those made more than two weeks in advance—to predict the state of the atmosphere with any degree of skill owing to the chaotic nature of the fluid dynamics equations involved. Eady (1951) expressed that dynamical forecasts would unavoidably be uncertain due to the amplifying effect of initial-state errors and that these uncertainties should be described using probabilistic terms. Following this school of thought, Epstein (1969) began to experiment with the idea of generating multiple forecasts from slightly different initial conditions and model configurations. He conducted independent random draws from the uncertainty distribution of the initial conditions to select the initial ensemble members. With his stochastic-dynamic predictions, Epstein was able to expand the forecasts with means and variances. Reflecting on these ideas, Leith (1974) provided a Monte Carlo forecasting procedure to represent a practical, computable approximation to the stochastic dynamic forecasts.

The first ensemble prediction systems (EPSs) that produce operational global medium-range ensemble weather forecasts were developed in 1992 at the European Centre for Medium-Range Weather Forecasts (ECMWF, Palmer et al., 1993; Buizza et al., 1999) and at the National Centers for Environmental Prediction (NCEP Toth and Kalnay, 1993, 1997). Shortly after, in 1995, they were followed by the Meteorological Service of Canada (MSC Houtekamer et al., 1996). Other weather centres with operational global ensemble prediction systems include the Australian Bureau of Meteorology (BMRC), the Chinese Meteorological Administration (CMA), the Brazilian Center for Weather Prediction and Climate Studies (CPTEC), the Japanese Meteorological Administration (JMA), the Korean Meteorological Administration (KMA), Meteo-France, and the UK Met Office (UKMO).

Since then, probabilistic or ensemble forecasting (Gneiting and Raftery, 2005) has become a standard tool in the field of weather forecasting, and it has been widely used for a variety of applications, including short-term weather forecasting (Stensrud et al., 1999), seasonal forecasting (Goddard et al., 2001; Palmer et al., 2004), and climate modelling (Palmer et al., 2005). Ensemble forecasting has also been shown to be useful in predicting extreme weather events (Stensrud, 2001) such as hurricanes, tornadoes, and severe thunderstorms.

Ensemble forecasting has several advantages over traditional single-model forecasting. One of the main advantages is its ability to provide more accurate and reliable predictions, as ensemble-mean forecasts are expected to outperform traditional high-resolution single-integration dynamical forecasts. More importantly, ensemble forecasts also provide a measure of uncertainty, which is

essential for decision-making in weather-sensitive industries such as agriculture, transportation, and energy.

Despite its many advantages, ensemble forecasting also has several challenges, one of which is the high computational cost of generating and processing multiple forecasts. Another concern is obtaining precise and reliable data, which can be challenging in some parts of the world.

1.2 Dual-resolution

The increasing need for accurate forecasts infers the urgency of computing on a higher-resolution grid, to provide more detailed forecasts and be able to represent complex topography and coastal regions better, which can have a strong influence on local weather patterns. As computing power continues to advance, higher-resolution models are becoming more feasible, and are expected to become increasingly important for improving the accuracy of weather forecasts. Advances in supercomputing and data storage technology have enabled the use of higher-resolution models, but the challenge of balancing accuracy with computational resources remains an ongoing concern. The cost of computations intensely depends on the spatial resolution, and the ensemble size chosen for a forecast system. As proven by Machete and Smith (2016) and Leutbecher (2018), the more ensemble members an EPS has, the better it is able to estimate forecast uncertainty. However, it remains a perpetual question of whether to invest resources in higher-resolution numerical models or in larger EPS (Ferro et al., 2012). The answer should always depend on the specific needs that arise. For instance, according to Richardson (2001) and Mullen and Buizza (2002) forecasts of extreme events benefit from larger ensembles at the expense of their resolution. In contrast, shorter lead times benefit more from improved resolution (this could be due to the already low overall uncertainty). At medium and longer lead times, the tendency shifts back in favour of larger ensemble sizes (Ma et al., 2012). It is also argued that there is a fundamental limit on the possible improvements gained by increasing the grid resolution (Lorenz, 1969; Palmer et al., 2014). After all, meteorological centres only have limited computational resources, and since computational costs rise with both ensemble size and resolution, a fair trade-off on these crucial variables should be made before implementing a new operational EPS configuration. Finding the right balance between increasing spatial resolution and increasing ensemble size is, therefore, essential.

The studied global medium-range ECMWF EPS had only 51 ensemble members with a resolution of approximately 18 km grid spacing (Haiden et al., 2018). However, since June 2023, ECMWF operationally produces a 51-member medium-range 9 km spatial resolution EPS while supplementing the forecasts by generating a 101-member extended-range ensemble on a 36 km grid (ECMWF, 2023). This new setup enables the use of dual-resolution ensemble forecasts operationally, while also benefiting from both the higher resolution of the medium-range and the larger ensemble size of the extended-range ensemble. The advantages expected from this setup are justified by preliminary analysis conducted by Leutbecher and Ben Bouallègue (2020) and Gascón et al. (2019) on 2-metre temperature and precipitation accumulation, respectively. Evidently, the 9 km and 36 km resolution was not available, thus instead, the 18 km and 45 km grid resolution was tested in both studies. To expand on these initial findings, we have conducted further investigation of the differences between mixture and single-resolution systems, summarised in Baran et al. (2019) and Szabó et al. (2023) and in Chapters 4 and 5, respectively.

In accordance with the strategic plans of the ECMWF for the upcoming years (ECMWF, 2021), "ECMWF will continue to investigate a mixture of a larger ensemble and increased vertical and horizontal resolution and a blend of variational and ensemble methods across the Earth system components."

1.3 Post-processing methods

As emphasised by Buizza (2018), ensemble forecasts indicate systematic errors in both dispersion and magnitude, and these must be accounted for before the results can be interpreted probabilistically. They can either under- or over-represent the forecast uncertainty, but often, operational ensemble forecasts exhibit too little dispersion (e.g. Buizza, 1997; Hamill, 2001; Toth et al., 2001; Buizza et al., 2005; Wang and Bishop, 2005). For example, this problem is also present in the case studies of Chapter 3, where it can be seen in Figure 3.2. This evidently results in overconfidence in assessing probability if ensemble relative frequencies are interpreted directly as estimating probabilities, as stated by Wilks (2019). Generally, we say that ensembles are probabilistically calibrated if the proportion of ensemble members predicting a given weather event aligns with the corresponding observed relative frequency when evaluated over a large sample.

The methods of calibrating the ensemble forecasts are rooted in the long-established approaches for statistical post-processing of dynamical weather fore-

casts. One of the simplest examples is the model output statistics (MOS, Glahn and Lowry, 1972) method, which relates past predictions from a forecast model to future weather quantities. Extending its principles to ensemble forecasts, Gneiting et al. (2005) suggests the ensemble model output statistics (EMOS), which is also referred to as nonhomogeneous regression or parametric distributional regression model. To improve calibration, the EMOS approach applies a single probability distribution to the ensemble forecast, thus resulting in a full predictive distribution. The parameters of the fitted distribution model depend on the ensemble members with the option of using differing link functions of the members. The importance of choosing an appropriate distribution for the different weather variables at hand should be emphasised. In Section 2 various weather variables and the suggested parametric distributions are detailed. These are based on numerous studies conducted to assess the skill of post-processed forecasts. In addition to widely used distribution-based models for wind speed calibration, our novel truncated generalised extreme value distribution-based approach is also presented (see Section 2.2.4).

An alternative to the EMOS post-processing technique is the Bayesian model averaging (BMA, Raftery et al., 2005) approach, where the forecast distribution is provided by a weighted mixture of parametric densities, each of which depends on a single member of the ensemble, with the weights of the mixture being determined by the performances of the ensemble members across the training data.

Taking the EMOS approach one step further, there are numerous studies that utilize a mix of distributions to base the models on, thus alleviating the limited flexibility of single-parametric distribution models. For wind speed forecasts, Lerch and Thorarinsdottir (2013) suggested a regime-switching combination model and Baran and Lerch (2016) provided a mixture EMOS model for calibration. The article by Gneiting and Ranjan (2013) proposes combination methods based on prediction spaces and cumulative distribution functions and assesses different aggregation methods, while Baran and Lerch (2018) provides an empirical assessment of the merits of combining forecast distributions from post-processing models for wind speed and precipitation accumulation forecasts of two datasets.

Furthermore, there are ensemble post-processing methods that are able to represent multivariate dependencies, which is key in accounting for intervariable, spatial and temporal dependencies. E.g. Lerch et al. (2020) and Lakatos et al. (2023) give a comprehensive review and comparison of state-of-the-art methods for multivariate ensemble post-processing, including the ensemble copula coupling (Schefzik et al., 2013), the Schaake shuffle (Clark et al., 2004), and

the Gaussian copula approach (Möller et al., 2013).

In recent years, approaches using machine learning techniques gained rapid popularity as they provide more flexibility in modelling and newly developed methodologies can be adapted from other research fields.Vannitsem et al. (2021) presents a general overview on how the rapidly advancing methods of machine learning and in particular neural networks fit into the context of post-processing techniques. To elaborate on the separate weather variables, Valdivia-Bautista et al. (2023) gives an analysis of artificial intelligence approaches in wind speed forecasting, Schulz and Lerch (2022) provides a systematic review of deep learning methods for wind gust forecast calibration, for forecasting solar irradiance with neural networks see Gneiting et al. (2023), for precipitation forecasts see Scheuerer et al. (2020) and Ghazvinian et al. (2022) and for temperature forecasts see Rasp and Lerch (2018). For a non-parametric neural network approach for wind speed forecast calibration we refer to Bremnes (2020).

There is also a wide range of non-parametric post-processing methods that have been developed to improve the predictive skill of ensemble forecasts, including quantile regression forest (Taillardat et al., 2016), censored quantile regression (Friederichs and Hense, 2007) and constrained quantile regression splines (Bremnes, 2019). However, these methods require sufficiently long training periods and generally lead to high computational costs, but have the benefit of eliminating the non-trivial choice of a specific distribution. However, this dissertation has its main focus on the parametric EMOS post-processing approach, which is one of the most widely used methods.

1.4 Overview and key challenges

Statistical post-processing techniques have become integral components of forecasting suites used by many meteorological services. The objective of these methods is to counter the different types of errors in the predictions in order to provide better forecasts overall. However, there are numerous challenges to consider when transferring these methods of calibration into operational use, and as a result, the number of post-processing methods for weather forecasts is rapidly growing. Despite the significant advances in post-processing methods related to weather forecasts, there is no single method that can overcome every challenge, particularly for every type of weather variable. This indicates the importance of developing and utilizing a range of post-processing techniques that can address the specific challenges associated with different types of weather quantities.

Some of the challenges are presented in the paper by Vannitsem et al. (2021), supplemented by a comprehensive list of the different state-of-the-art post-processing methods. These approaches are grouped by their need for distributional assumptions and are assessed by their flexibility and implementation complexity. The gap between the implementation of these methods in research projects and the operational use can be quite vast, and so before transferring these post-processing techniques into operations, one must handle the findings as preliminary results. The key problems include the management of training data, and having suitable quality control as these methods are usually retrained for every forecast. These reruns can take up a lot of computational resources and need to be optimised for fast real-time calculations through parallel implementations.

As suggested by Vannitsem et al. (2021): "To really be able to benchmark the value of new methods, a common platform on which the different techniques can be compared on a set of appropriately chosen meteorological forecasts is highly desirable." Since then, the EUPPBench was developed and published by Demaeyer et al. (2023), a dataset of time-aligned forecasts and observations, with the aim to facilitate and standardize the process of comparing different methodologies for various weather variables. Vannitsem et al. (2021) also states that there is a noticeable change in approach occurring in the field of weather forecasting, where there is a transition from physical modelling methods towards data-driven approaches. This shift is a result of the abundance of new datasets, emerging technologies, and advancements in computing power resources and data science techniques. These resources provide new means of improving forecast accuracy beyond the traditional method of refining NWP models.

In this dissertation, the focus will be on three continuous weather variables, namely temperature, wind speed and precipitation accumulation. Furthermore, extreme weather events are not analysed, as the emphasis is on general weather patterns. Climate, hydrological and atmospheric modelling are not included in the scope of this dissertation. Although it is evident that multivariate post-processing and neural networks are increasingly popular in weather forecasting, the primary focus is on univariate post-processing techniques. In addition, the relatively new field of dual-resolution ensemble forecasts is analysed in great detail.

Chapter 2

Statistical post-processing with EMOS

Post-processing is a key factor in correcting bias and dispersion errors in ensemble forecasts, thus providing more skilful predictions, as discussed in Section 1.3. One of the most efficient and most widely used distribution-based approaches is the EMOS framework introduced by Jewson et al. (2004) and Gneiting et al. (2005), which specifies a parametric model for the forecast distribution by selecting a suitable probability law to match the characteristics of the weather variable at hand. This approach is based on the idea that the forecast distribution can be modeled using a parametric probability distribution, such as a Gaussian distribution for temperature forecasts, and that the parameters of this distribution can be estimated using historical data. The parametric distributional regression model is fitted with the help of the ensemble predictions and the corresponding observations of a training period by linking the distribution parameters to the ensemble members appropriately. As was introduced by Gneiting et al. (2005) the Gaussian forecast mean is a corrected weighted average of the ensemble member predictions, with coefficients that represent the relative contributions of the member models to the ensemble. On the other hand the variance of the predictive distribution is obtained by an affine function of the ensemble variance. The regression coefficients are estimated by optimizing a suitable loss function, and then the constructed model goes through a validation process on a separate dataset. The following sections provide the specific model structures for three different weather variables: temperature, wind speed and

precipitation accumulation. Naturally, for all of these quantities a multitude of parametric predictive distributions exists, however we only provide the detailed models for those, that were used in the studies. These models were implemented in order to provide the results presented in Chapter 3, 4 and 5.

As suggested by Gneiting et al. (2005), when fitting the EMOS model one should consider parameters (e.g. location and scale) of the predictive probability density function to be affine functions of the ensemble forecasts and the ensemble variance, respectively. In what follows, let us denote by $f_1, f_2, \ldots, f_K$ the ensemble forecast for a given location, time and forecast horizon and let $\overline{f}$ denote the ensemble mean and S^2 denote the ensemble empirical variance:

$$S^2 := \frac{1}{K-1} \sum_{k=1}^{K} \left(f_k - \overline{f}\right)^2.$$

As defined by Vannitsem et al. (2018), ensemble members are considered non-exchangeable, if they have distinct statistical characteristics e.g. derived from single integrations of models, whereas ensemble members are considered exchangeable if they have the same characteristics e.g. produced by the same model with slight perturbations. If an EPS has M ensemble members divided into K exchangeable groups, where the kth group contains $M_k \geq 1$ ensemble members ($\sum_{k=1}^{K} M_k = M$), then let us denote its mean as $\overline{f}_k$.

To give examples of the varying configurations adopted, the ECMWF currently employs in its operational use an EPS with 1 control member and 50 exchangeable ensemble members produced by perturbed initial conditions, whereas 40 non-exchangeable members are offered by the ICON EPS German operational small-scale modelling system.

2.1 Temperature

Temperature is a fundamental variable in weather forecasting, and the accuracy of temperature forecasts is essential across all aspects of society, e.g. the transportation, healthcare and energy sectors. According to Harmel et al. (2001) and Gneiting et al. (2005), the assumption of normality is prevalent in models which deal with temperature forecasts, therefore in the next section, a normal distribution-based EMOS model is given in detail.

2.1.1 Normal EMOS model

Let $\mathcal{N}(\mu, \sigma^2)$ denote a normal distribution with mean μ and standard deviation $\sigma > 0$. The associated predictive distribution of temperature, suggested by Gneiting et al. (2005) is given as

$$\mathcal{N}\big(a_0 + a_1 f_1 + \cdots + a_K f_K, b_0 + b_1 S^2\big), \tag{2.1}$$

where $a_0 \in \mathbb{R}$ and $a_1, \ldots, a_K, b_0, b_1 \geq 0$.

If we have exchangeable ensemble groups, Gneiting (2014) and Wilks (2018) suggest using the same coefficients within a given group, as also shown in the case of a multimodel setup by Fraley et al. (2010). Consequently, the EMOS predictive distribution will be given by

$$\mathcal{N}\big(a_0 + a_1 \overline{f}_1 + \cdots + a_K \overline{f}_K, b_0 + b_1 S^2\big). \tag{2.2}$$

2.2 Wind speed

With the ever-growing prominence of wind power as a renewable energy source, the calibration of wind speed forecasts has gained even more importance in recent years. To model wind speed data one must take into account its non-negative nature and the possibility of frequent high wind speed values. These characteristics make the choice of an appropriate parametric distribution harder than in case of e.g. temperature forecasts. The recommendation is a skewed distribution with non-negative support, such as the Weibull and log-normal distributions (Justus et al., 1978; Garcia et al., 1998), others have assessed the truncated normal distribution (TN; Thorarinsdottir and Gneiting, 2010) and the gamma distribution (Scheuerer and Möller, 2015). In order to provide a better fit to high wind speed values, Lerch and Thorarinsdottir (2013) and Baran and Lerch (2015) suggest models based on generalized extreme value (GEV) and also log-normal (LN) distributions, respectively. In the latter the authors compared the predictive performance of TN, LN, GEV and mixtures of TN-LN and TN-GEV based models as well. In the following sections, we explore the exact formulations of those distribution-based parametric regression models that we have applied in our studies.

2.2.1 Truncated normal EMOS model

Let us denote by $\mathcal{N}_0(\mu, \sigma^2)$ a TN distribution with location μ, scale $\sigma > 0$, and lower truncation at 0, having probability density function (PDF)

$$g(x|\mu,\sigma) := \begin{cases} \frac{1}{\sigma}\varphi\big((x-\mu)/\sigma\big)/\Phi(\mu/\sigma), & \text{if } x \geq 0; \\ 0, & \text{otherwise,} \end{cases}$$

where φ is the PDF, while Φ denotes the cumulative distribution function (CDF) of the standard normal distribution. For the TN EMOS predictive distribution, the location and scale are linked to the ensemble members via the following equations:

$$\mu = a_0 + a_1 f_1 + \cdots + a_K f_K \qquad \text{and} \qquad \sigma^2 = b_0 + b_1 S^2. \tag{2.3}$$

where $a_0 \in \mathbb{R}$ and $a_1, \ldots, a_K, b_0, b_1 \geq 0$.

If the ensemble can be split into K groups of exchangeable members, then forecasts within a given group will share the same location parameter (Gneiting, 2014; Wilks, 2018) resulting in link functions

$$\mu = a_0 + a_1 \overline{f}_1 + \cdots + a_K \overline{f}_K \qquad \text{and} \qquad \sigma^2 = b_0 + b_1 S^2. \tag{2.4}$$

2.2.2 Log-normal EMOS model

A LN distribution has a heavier upper tail than a TN distribution and is, therefore, more appropriate to model high wind speed values (Baran and Lerch, 2015). The PDF of a LN distribution $\mathcal{LN}(\mu, \sigma)$ with location μ and scale $\sigma > 0$ is

$$h(x|\mu,\sigma) := \begin{cases} \frac{1}{x\sigma}\varphi\big((\log x - \mu)/\sigma\big), & \text{if } x \geq 0; \\ 0, & \text{otherwise,} \end{cases}$$

while the mean m and variance v are

$$m = \mathrm{e}^{\mu+\sigma^2/2} \qquad \text{and} \qquad v = \mathrm{e}^{2\mu+\sigma^2}\big(\mathrm{e}^{\sigma^2} - 1\big),$$

respectively. Obviously, a LN distribution can also be parametrized by these latter two quantities via equations

$$\mu = \log\left(\frac{m^2}{\sqrt{v+m^2}}\right) \qquad \text{and} \qquad \sigma = \sqrt{\log\left(1 + \frac{v}{m^2}\right)},$$

and in the LN EMOS model of Baran and Lerch (2015) m and v are affine functions of the ensemble and the ensemble variance, respectively:

$$m = \alpha_0 + \alpha_1 f_1 + \cdots + \alpha_K f_K \qquad \text{and} \qquad v = \beta_0 + \beta_1 S^2. \tag{2.5}$$

In the case of the existence of groups of exchangeable ensemble members, similar to (2.4), the equation for the mean in (2.5) is replaced by

$$m = \alpha_0 + \alpha_1 \overline{f}_1 + \cdots + \alpha_K \overline{f}_K. \tag{2.6}$$

2.2.3 Generalized extreme value EMOS model

In a similar manner, one can consider a GEV distribution-based model, which also has a heavier upper tail (Lerch and Thorarinsdottir, 2013) as an alternative to the TN EMOS approach. Let $\mathcal{GEV}(\mu, \sigma, \xi)$ denote a GEV distribution with location μ, scale $\sigma > 0$ and shape ξ, defined by its CDF

$$G(x|\mu, \sigma, \xi) := \begin{cases} \exp\Big(- \big[1 + \xi(\frac{x-\mu}{\sigma})\big]^{-1/\xi}\Big), & \text{if } \xi \neq 0; \\ \exp\Big(- \exp\big(- \frac{x-\mu}{\sigma}\big)\Big), & \text{if } \xi = 0, \end{cases} \tag{2.7}$$

for $1 + \xi(\frac{x-\mu}{\sigma}) > 0$ and $G(x|\mu, \sigma, \xi) := 0$, otherwise.

However, this model has the disadvantage of assigning positive probabilities to negative wind speed values (see e.g., Baran and Lerch, 2015; Baran et al., 2021).

The model proposed by Lerch and Thorarinsdottir (2013) uses location and scale parameters

$$\mu = \gamma_0 + \gamma_1 f_1 + \cdots + \gamma_K f_K \qquad \text{and} \qquad \sigma = \sigma_0 + \sigma_1 \overline{f}, \tag{2.8}$$

with $\sigma_0, \sigma_1 \geq 0$, while the shape parameter ξ does not depend on the ensemble members.

2.2.4 Truncated generalized extreme value EMOS model

To mitigate the issue that was mentioned in the preceding section, let us consider a truncated version of a GEV distribution, as this does not forecast negative wind speed with a positive probability. The solution proposes a novel EMOS model where the predictive GEV distribution is truncated from below at 0. Let

$\mathcal{TGEV}(\mu, \sigma, \xi)$ denote a truncated GEV (TGEV) distribution with location μ, scale $\sigma > 0$ and shape ξ. For $x \geq 0$ the CDF of a TGEV distribution is

$$G_0(x|\mu, \sigma, \xi) = \begin{cases} \frac{G(x|\mu,\sigma,\xi) - G(0|\mu,\sigma,\xi)}{1 - G(0|\mu,\sigma,\xi)}, & \text{if } G(0|\mu, \sigma, \xi) < 1; \\ 1, & \text{if } G(0|\mu, \sigma, \xi) = 1, \end{cases} \tag{2.9}$$

whereas negative values are obviously excluded from the support set. For $\xi < 1$ (and $G(0|\mu, \sigma, \xi) < 1$) the $\mathcal{TGEV}(\mu, \sigma, \xi)$ distribution has a finite mean of

$$\begin{cases} \mu + (\Gamma(1-\xi) - 1))\frac{\sigma}{\xi}, & \text{if } \xi > 0 \text{ and } \xi\mu - \sigma > 0; \\ \mu - \frac{\sigma}{\xi} + \frac{\sigma(\Gamma_\ell(1-\xi, [1-\xi\mu/\sigma]^{-1/\xi}))/\xi}{1-\exp(-[1-\xi\mu/\sigma]^{-1/\xi})}, & \text{if } \xi \neq 0 \text{ and } \xi\mu - \sigma \leq 0; \\ \frac{\mu+\sigma(C-\mathrm{Ei}(-\exp[\mu/\sigma]))}{1-\exp(-\exp[\mu/\sigma])}, & \text{if } \xi = 0, \end{cases} \tag{2.10}$$

where Γ and Γ_ℓ denote the gamma and the lower incomplete gamma function, respectively, defined as

$$\Gamma(a) = \int_0^\infty t^{a-1}\mathrm{e}^{-t}\mathrm{d}t \qquad \text{and} \qquad \Gamma_\ell(a, x) = \int_0^x t^{a-1}\mathrm{e}^{-t}\mathrm{d}t,$$

and $\mathrm{Ei}(x)$ is the exponential integral

$$\mathrm{Ei}(x) = \int_{-\infty}^x \frac{e^t}{t}\mathrm{d}t = C + \ln|x| + \sum_{k=1}^\infty \frac{x^k}{k!k}$$

with C being the Euler–Mascheroni constant. It is important to emphasize, that the case $\xi < 0$ and $\xi\mu - \sigma > 0$, does not appear in the formula (2.10), since in that case the PDF of a $\mathcal{GEV}(\mu, \sigma, \xi)$ is positive only on $]-\infty, \mu - \sigma/\xi] \subset \mathbb{R}_-$. Further, as for $\xi > 0$ and $\xi\mu - \sigma > 0$ the support of $\mathcal{GEV}(\mu, \sigma, \xi)$ is $[\mu - \sigma/\xi, \infty[\subset \mathbb{R}_+$, truncation does not change the distribution and the means of $\mathcal{GEV}(\mu, \sigma, \xi)$ and $\mathcal{TGEV}(\mu, \sigma, \xi)$ distributions coincide. For the proof of the remaining cases of (2.10) see Appendix A.1.

The parameters of the TGEV EMOS model are also linked to the ensemble members according to (2.8), which is replaced by

$$\mu = \gamma_0 + \gamma_1\overline{f}_1 + \cdots + \gamma_K\overline{f}_K \qquad \text{and} \qquad \sigma = \sigma_0 + \sigma_1\overline{f}, \tag{2.11}$$

in the exchangeable case. Note that alternative expressions

$$\sigma = \sigma_0 + \sigma_1 S, \qquad \sigma = \sqrt{\sigma_0 + \sigma_1 S^2} \qquad \text{and} \qquad \sigma = \sigma_0 + \sigma_1 \mathrm{MD}$$

of the scale have also been tested, where

$$\mathrm{MD} := \frac{1}{K^2} \sum_{k,\ell=1}^{K} \left| f_k - f_\ell \right|$$

is the ensemble mean absolute difference (see e.g. Scheuerer, 2014; Baran et al., 2020). However, in our case studies TGEV EMOS models with link functions (2.8) and (2.11) show the best predictive performance.

2.3 Precipitation accumulation

Calibrating precipitation forecasts is a far more difficult task compared to the calibration of temperature or wind speed forecasts, as precipitation accumulation has a unique discrete-continuous nature that makes it challenging to model accurately. Firstly, one must address its non-negative characteristic, which indicates the need for a distribution that has a non-negative support. Secondly, zero precipitation values are very common in everyday observations, thus only those predictive distributions are advised to be used, that are able to assign positive mass to the zero precipitation event. A popular choice is to consider a continuous distribution that can take both negative and positive values and left censor it at zero, such as the GEV (Scheuerer, 2014) or the censored, shifted gamma distribution (CSG; Scheuerer and Hamill, 2015; Baran and Nemoda, 2016). In what follows, we explore the detailed formulations of only the CSG distribution-based parametric regression model that we have applied in our studies.

2.3.1 Censored and shifted gamma EMOS model

Let $G_{\kappa,\theta}$ denote the CDF of a gamma distribution $\Gamma(\kappa, \theta)$ with shape $\kappa > 0$ and scale $\theta > 0$ defined by PDF

$$g_{\kappa,\theta}(x) := \begin{cases} \frac{x^{\kappa-1}\mathrm{e}^{-x/\theta}}{\theta^{\kappa}\Gamma(\kappa)}, & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The distribution $\Gamma(\kappa, \theta)$ can be characterized by its mean $\mu > 0$ and standard deviation $\sigma > 0$ as well, since there are direct relations between these parameters, and the shape κ and scale θ parameters of the corresponding gamma

distribution. This correspondence can be expressed through the following equations:

$$\kappa = \frac{\mu^2}{\sigma^2} \qquad \text{and} \qquad \theta = \frac{\sigma^2}{\mu}.$$

After extending the support of the gamma distribution to negative values with the help of a shift parameter $\delta > 0$, one can introduce $\Gamma^0(\kappa, \theta, \delta)$ denoting a shifted gamma distribution, left censored at zero with shape κ, scale θ and shift δ, given by its CDF

$$G^0_{\kappa,\theta,\delta}(x) \coloneqq \begin{cases} G_{\kappa,\theta}(x+\delta), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Following the notations given at the beginning of Chapter 2, let us consider the parameters of the CSG distribution to be affine functions of the ensemble members. In the CSG EMOS model mean μ and variance σ^2 of the underlying gamma distribution are linked to the ensemble members as

$$\mu = a_0 + a_1 f_1 + \cdots + a_K f_K \qquad \text{and} \qquad \sigma^2 = b_0 + b_1 \overline{f}, \tag{2.12}$$

where $a_0, \ldots, a_K, b_0, b_1 \geq 0$, and the shift parameter $\delta > 0$ is independent of the ensemble forecast. The variance is dependent only on the ensemble mean which choice is due to the extensive tests Baran and Nemoda (2016) has conducted with the same model. Among the various ensemble statistics used, the ensemble mean proved to be the best in terms of the predictive performance of the calibrated forecasts.

However, in the cases where the ensemble members are exchangeable, the link functions in (2.12) should be replaced by

$$\mu = a_0 + a_1 \overline{f}_1 + \cdots + a_K \overline{f}_K \qquad \text{and} \qquad \sigma^2 = b_0 + b_1 \overline{f}. \tag{2.13}$$

2.4 Parameter estimation

According to the optimal score estimation approach of Gneiting and Raftery (2007), model parameters should be estimated by optimizing the mean value of a proper scoring rule as a function of the parameters over appropriately chosen training data. Scoring rules can evaluate the accuracy of probabilistic forecasts by assigning a numerical score that is based on both the predictive distribution and the observed value. For predictive distributions, one of the most widely used strictly proper scoring rule is the continuous ranked probability score (CRPS;

Gneiting and Raftery, 2007; Wilks, 2019). Given a (predictive) CDF $F(y)$ and real value (observation) x, the CRPS is defined as

$$\mathrm{CRPS}\big(F,x\big) := \int_{-\infty}^{\infty} \big(F(y) - \mathbb{I}_{\{y\geq x\}}\big)^2 \mathrm{d}y = \mathsf{E}|X-x| - \frac{1}{2}\mathsf{E}|X-X'|, \quad (2.14)$$

where $\mathbb{I}_H$ denotes the indicator of a set H, whereas X and X' are independent random variables with CDF F and finite first moment. Note that the CRPS is a negatively oriented score, and the right-hand side of (2.14) shows that the CRPS has the same unit as the observation. For example, for the normal distribution family, a closed form was given by Gneiting and Raftery (2007). Closed-form expressions are essential for the efficient computation of the score for a large number of forecasts, to avoid the need for computationally expensive numerical integration, which can be time-consuming. For the TN, LN, GEV and CSG distributions the closed form of the CRPS has already been derived, see Thorarinsdottir and Gneiting (2010); Baran and Lerch (2015); Friederichs and Thorarinsdottir (2012); Scheuerer and Möller (2015), respectively. The closed-form of the CRPS for a TGEV distribution $\mathcal{TGEV}(\mu,\sigma,\xi)$ defined in Section 2.2.4 with CDF $G_0(x)$ derived from a GEV CDF $G(x)$ is given by

$$\begin{aligned}\mathrm{CRPS}(G_0,x) =& \big(2G_0(x)-1\big)\Big(x-\mu+\frac{\sigma}{\xi}\Big) \\ &+ \frac{\sigma}{\xi(1-G(0))^2}\Big[-2^{\xi}\Gamma_\ell\big(1-\xi,-2\ln G(0)\big) \\ &+ 2G(0)\Gamma_\ell\big(1-\xi,-\ln G(0)\big) + 2\big(1-G(0)\big)\Gamma_\ell\big(1-\xi,-\ln G(x)\big)\Big]\end{aligned} \quad (2.15)$$

for $\xi \neq 0$, whereas for $\xi = 0$ we have

$$\begin{aligned}\mathrm{CRPS}(G_0,x) =& (x-\mu)\big(2G_0(x)-1\big) \\ &+ \frac{\sigma}{(1-G(0))^2}\Big(C - \ln 2 + \mathrm{Ei}\big(2\ln G(0)\big) \\ &- (G(0))^2 \ln\big[-\ln G(0)\big] - 2G(0)\mathrm{Ei}\big(\ln G(0)\big)\Big) \\ &+ \frac{2\sigma}{1-G(0)}\Big[G(x)\ln\big[-\ln G(x)\big] - \mathrm{Ei}\big(\ln G(x)\big)\Big].\end{aligned} \quad (2.16)$$

For the proof of (2.15) and (2.16) see Appendix A.2.

Another aspect of the parameter estimation process that needs addressing is the numerical challenges presented by the optimisation techniques applied. During optimisation the most important factor to consider is the ratio between

the number of parameters to estimate and the data available, which challenge has been addressed in the previous Section regarding the choice of training period length. More implementation details can be found in Section 3.2, 4.2 and 5.2.

2.5 Spatial and temporal selection of training data

The use of any statistical post-processing model for weather forecasting requires estimating the model parameters based on observed data. One of the key challenges is determining how to choose the training data to achieve accurate parameter estimation. There are three main approaches to this task concerning spatial selection: regional, local, and semi-local.

The regional approach uses data from a wide ensemble domain to estimate the parameters. If the dataset spans an area with small variability in terms of the weather variable of interest, then all of the forecast values are used as training data to train one model, resulting in the same parameters across all locations. In contrast, the local approach focuses on a single location, thus using only the data from that specific location to train its distinct model, resulting in differing models for all locations. The semi-local approach (Lerch and Baran, 2017) is a good compromise between the two, providing a good alternative for highly variable areas where running a local model training is not effective. The semi-local approach is based on clustering the locations based on different features, consequently providing the model with a smaller dataset, from relatively similar locations. Evidently, the locations that are in the same cluster will have the same parameters as well, however, the clustering is usually rerun for every forecast horizon and every forecast day. Specifically, the clustering applied in Section 4.3 is implemented similarly to Lerch and Baran (2017), where the k-means clustering of stations is based on 24-dimensional feature vectors consisting of 12 equidistant quantiles of the climatological CDF and 12 equidistant quantiles of the empirical CDF of forecast errors of the ensemble mean over the training period.

The other challenge is regarding the temporal selection of the training data, which is explored in the comparative analysis of Lang et al. (2020). Different time-adaptive training strategies have been developed for non-homogeneous regression in order to adjust for seasonally varying error characteristics between ensemble forecasts and observations. These schemes include a smooth model ap-

proach using data from multiple years, as well as the standard sliding-window approach. In the latter case, to estimate the EMOS model parameters, a rolling training period is applied and the estimates are obtained using ensemble forecasts and corresponding validating observations for the preceding n calendar days. This means that given a verification day, the utilised training period consists of the preceding n days, and this sliding window shifts with the verification period day by day. It is also important to take the lead time into account, which is the time between making the forecast and when it becomes available, as we do not want to use forecasts that are not available in reality. So one must shift the training period backwards with the lead time as well. We refer to Sections 3.2, 4.2 and 5.2 for specific details on the choice of n for the different data sets, and for some further examples of choosing an appropriate n see Hemri et al. (2014).

2.6 Validation metrics

Forecast validation (or verification) is the process of assessing the quality of forecasts, in which matched pairs of forecasts are compared to the observations to which they pertain. Wilks (2019) provides a detailed list of the different attributes that characterize the forecast quality, such as accuracy, bias, reliability and sharpness, etc. The standard tool for quantifying the predictive performance of probabilistic forecasts both in terms of calibration and sharpness is calculating the mean of the CRPS over the verification data, defined in Section 2.4.

Similarly, one can consider the Brier score (BS; Wilks, 2019, Section 9.4.2) for the dichotomous event that the observed continuous weather variable x exceeds a given threshold y. For a predictive CDF $F(y)$, the Brier score is defined as

$$\mathrm{BS}\left(F, x; y\right) := \left(F(y) - \mathbb{I}_{\{y \geq x\}}\right)^2,$$

(see e.g. Gneiting and Ranjan, 2011), and note that the CRPS is the integral of the BS over all possible thresholds. In the results provided in Section 4.3 we consider the thresholds of 5, 10, ..., 90, and 95 percentiles of the corresponding station climatology for the verification period.

Further, let $q_\tau(F)$ denote the τ-quantile $(0 \leq \tau \leq 1)$ of a CDF $F(y)$, that is

$$q_\tau(F) := F^{-1}(\tau) := \inf\{y : F(y) \geq \tau\},$$

and consider the loss function

$$\rho_\tau(x) := \begin{cases} \tau|x|, & \text{if } x \geq 0, \\ (1-\tau)|x|, & \text{if } x < 0. \end{cases}$$

Then for a given value x the quantile score (QS; see e.g. Bentzien and Friederichs, 2014) is defined as

$$\mathrm{QS}_\tau(F, x) := \rho_\tau\big(x - q_\tau(F)\big).$$

In order to evaluate the CRPS, BS, and QS of the raw ensemble, it is necessary to replace the predictive CDF with the empirical one.

To be able to study the predictive performance of the predictions for higher forecast values (e.g. in the case of wind speed) we also consider the threshold-weighted continuous ranked probability score (twCRPS; Gneiting and Ranjan, 2011)

$$\mathrm{twCRPS}(F, x) := \int_{-\infty}^{\infty} \Big[F(y) - \mathbb{I}_{\{y \geq x\}}\Big]^2 \omega(y)\mathrm{d}y, \tag{2.17}$$

where $\omega(y) \geq 0$ is a weight function. Setting $\omega(y) \equiv 1$ results in the traditional CRPS (2.14), whereas with the help of $\omega(y) = \mathbb{I}_{\{y \geq r\}}$ one can address weather variable values above a given threshold r. Note that in the case studies of Chapter 3 the thresholds correspond approximately to the 90th, 95th and 98th percentiles of the wind speed observations.

Predictive performance is presented as a skill score in a lot of case studies, as it can even express small differences between competing forecasts. Firstly, to define its general formula, let us denote a particular measure of accuracy by $\mathcal{S}_F$ for a given forecast F and the accuracy measure of a reference forecast F_{ref} by $\mathcal{S}_{F_{ref}}$. Secondly, let $\overline{\mathcal{S}}_F$ and $\overline{\mathcal{S}}_{F_{ref}}$ represent the mean score values over the verification data for F and F_{ref}, respectively. The skill score $\mathcal{S}^{skill}$ of $\mathcal{S}$ is given by

$$\mathcal{S}^{skill}\big(F, F_{ref}\big) := 1 - \frac{\overline{\mathcal{S}}_F}{\overline{\mathcal{S}}_{F_{ref}}}.$$

Employing this definition for CRPS, BS, QS and twCRPS one can introduce the continuous ranked probability skill score (CRPSS; see e.g. Gneiting and Raftery, 2007), the Brier skill score (BSS), the quantile skill score (QSS; Friederichs and Thorarinsdottir, 2012) and the threshold-weighted continuous ranked probability skill score (twCRPSS) quantifying improvement in a forecast

F over a reference forecast F_{ref}. Obviously, in contrast to the original scoring measures, the corresponding skill scores are positively oriented, that is the larger the better.

In the case of point forecasts, such as ensemble and EMOS medians and means, a good evaluation method uses mean absolute errors (MAEs) and root mean squared errors (RMSEs), where the former is optimal for the median and the latter for the mean (Gneiting and Ranjan, 2011), although for the normal EMOS model, these quantities – the median and the mean – coincide.

To assess the calibration and sharpness of a predictive distribution, it is advised to investigate the coverage and average width of the $(1-\alpha)100\,\%$, $\alpha \in]0,1[$, central prediction interval, respectively. The coverage refers to the proportion of validating observations situated within the lower and upper $\alpha/2$ quantiles of the predictive CDF, and the level α should be selected to match the nominal coverage of the raw ensemble, that is, $(K-1)/(K+1)100\,\%$, where K represents the size of the ensemble. Choosing α in this manner allows for a direct comparison with the ensemble coverage, as the coverage of a properly calibrated predictive distribution is expected to be around $(1-\alpha)100\,\%$.

To assess the statistical significance of the differences between the verification scores we make use of the Diebold-Mariano (DM; Diebold and Mariano, 1995) test of equal predictive performance, as it allows accounting for the temporal dependencies in the forecast errors. Adhering to the notations used by Gneiting and Ranjan (2011) in their detailed description of the Diebold-Mariano test and making use of the already given notations in the preceding paragraphs, let $\overline{\mathcal{S}}_F$ and $\overline{\mathcal{S}}_G$ denote the mean values of a particular scoring rule over the verification data, corresponding to the competing F and G forecasts, respectively. Then the test statistic of the DM test is given by

$$t_N = \sqrt{N}\frac{\overline{\mathcal{S}}_F - \overline{\mathcal{S}}_G}{\hat{\sigma}_N},$$

where $\hat{\sigma}_N$ is a suitable estimator of the asymptotic standard deviation of the sequence of score differences between $\overline{\mathcal{S}}_F$ and $\overline{\mathcal{S}}_G$ over the verification data of size N. Under certain weak regularity assumptions, t_N asymptotically follows a standard normal distribution when the null hypothesis of equal predictive performance is true. If t_N has negative values, it suggests a better predictive performance of F, while positive values favour G.

The assessment of uncertainty in the verification scores involves the utilization of confidence intervals for both mean score values and skill scores. These intervals are obtained through 2,000 block bootstrap samples, which are gener-

ated using the stationary bootstrap scheme with mean block length computed based on the method developed by Politis and Romano (1994).

However, as formal tests may indicate an inadequate fit, they may not provide enough information on the specific nature of the dissimilarities. The evaluation of the adequacy of the parametric representation can also be executed through a graphical comparison between the data and the fitted distribution. This comparison serves to identify areas where the parametric model may not be suitable and to assess the degree of its inadequacy. One of the graphical tools suggested by Wilks (2019) is the Quantile–Quantile (Q–Q) plot, which compares empirical data and fitted CDFs in terms of the dimensional values of the variable (the empirical quantiles). Another commonly used alternative is the verification rank histogram (VRH) or Talagrand diagram of ensemble predictions and its continuous counterpart, the probability integral transform (PIT) histogram. The verification rank is the rank of the verifying observation with respect to the corresponding ensemble forecast (see e.g., Wilks, 2019, Section 9.7.1), whereas the PIT is the value of the predictive CDF evaluated at the verifying observation (Dawid, 1984; Raftery et al., 2005) with a possible randomisation at the points of discontinuity. In the scenario of a K-member ensemble that has been correctly calibrated, it can be observed that the verification ranks adhere to a uniform distribution on the set $\{1, 2, \ldots, K + 1\}$. On the other hand, the PIT values, representing the transformed CDF of calibrated predictive distributions, display uniformity across the $[0, 1]$ interval.

Chapter 3

Calibration of wind speed forecasts

In this chapter, we present the results of our investigation into the performance and effectiveness of the TGEV EMOS method defined in Section 2.2.4 for wind speed forecasting. The primary objective is to assess the impact of TGEV EMOS calibration on improving the accuracy and reliability of wind speed predictions. Through comprehensive evaluation and analysis, we examine the extent to which TGEV EMOS calibration improves the forecast skill, providing comparable results for various benchmark EMOS models based on other more common distributions, such as TN, LN, and GEV, thus providing valuable insights into the effectiveness of this novel method for enhancing wind speed forecasting capabilities. The direct comparison is made possible by using the same datasets containing the ensemble forecasts and observations of wind speed that were studied by Baran and Lerch (2015, 2016, 2018).

3.1 Data

Within this section, we provide an overview of the datasets used in the study for the evaluation of the TGEV wind speed EMOS model. To assess the predictive performance of the TGEV distribution-based model, we used three sets of data that had already been used to test existing EMOS models to provide a fair comparison. Each data set contains ensemble predictions for a single forecast horizon ranging from 24 to 48 hours, which are called short-range forecasts.

Each differs in the observed wind quantity, in the forecast lead time and in the stochastic properties of the ensemble. Further, we expand the comparison of the different EMOS models on a much more extensive database, providing ensemble forecasts with varying lead times ranging from 24 hours to 360 hours.

3.1.1 Short-range ensemble forecasts

UWME forecasts

The eight members of the University of Washington mesoscale ensemble (UWME) are generated by separate runs of the fifth-generation Pennsylvania State University-National Center for Atmospheric Research mesoscale model (PSU-NCAR MM5) with different initial conditions (Grell et al., 1995). The EPS domain covers the Pacific Northwest region of North America with a 12-km grid, and the dataset at hand contains forecasts for 48 hours ahead with the corresponding observations of 10-metre maximal wind speed (maximum of the hourly instantaneous wind speeds over the previous 12 hours, given in m/s; see, e.g., Sloughter et al. (2010)) for 152 stations in the Automated Surface Observing Network (National Weather Service, 1998) in the US states of Washington, Oregon, Idaho, California, and Nevada for the two years of 2007–2008. The forecasts are initialised at 0000 UTC, and the ensemble generation ensures that its members are clearly distinguishable. Our analysis focuses on 2008 with additional data from December 2007 used for model training. Removing days and locations with missing data and stations where data is only available on very few days results in 101 stations (see Figure 3.1) with a total of 27,481 individual forecast cases.

ALADIN-HUNEPS ensemble

The Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System (ALADIN-HUNEPS) of the Hungarian Meteorological Service covers a large part of continental Europe with a horizontal resolution of 8 km. The forecasts are obtained by dynamic downscaling of the global ARPEGE1-based PEARP2 system of Metéo-France (Horányi et al., 2006; Descamps et al., 2015). The EPS provides an unperturbed analysis-initiated control member and 10 members calculated with perturbed initial conditions. These members are statistically indistinguishable and therefore can be considered interchangeable, which should be considered when formulating post-processing models. We use ensembles of 42-hour ahead forecasts (initialised at

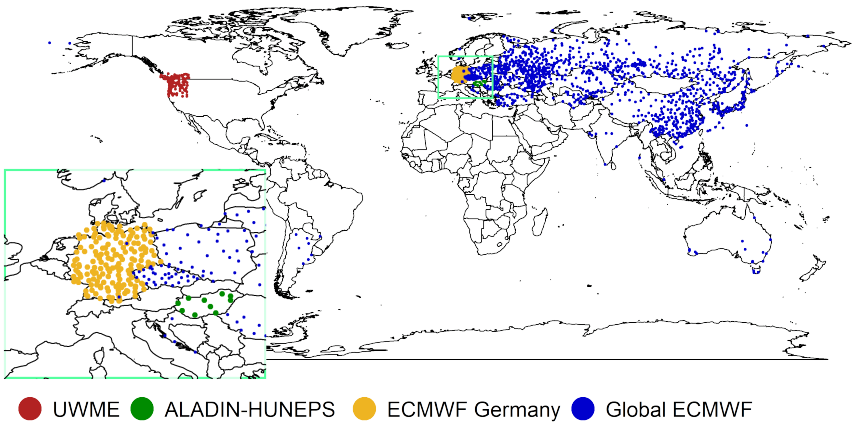

Figure 3.1: Map of the stations from the four datasets.

1800 UTC) of the 10-metre instantaneous wind speed (m/s) issued for 10 major cities in Hungary (see Figure 3.1) for the 1-year period from 1 April 2012 to 31 March 2013, together with the corresponding validation observations. The 6 days with missing forecasts and/or observations are excluded from the analysis.

ECMWF ensemble forecasts for Germany

The operational EPS of the ECMWF comprises 50 perturbed (thus exchangeable) members and operates on a global 18 km grid (Molteni et al., 1996; Leutbecher and Palmer, 2008). First, we consider 24-hour ahead ECMWF forecasts of 10-metre daily maximum wind speed initialised at 0000 UTC between 1 February 2010 and 30 April 2011, calculated for 228 stations (see Figure 3.1). We also consider corresponding verifying observations from the same 228 synoptic observation (SYNOP) stations over Germany. This dataset is identical to the one studied by Lerch and Thorarinsdottir (2013) and Baran and Lerch (2015, 2016). Post-processed forecasts are verified for the 1-year period between 1 May 2010 and 30 April 2011, containing 83,220 individual forecast cases. Forecast-observation pairs from April 2010 are used for training purposes.

3.1.2 Global ECMWF forecasts with different forecast horizons

To compare the predictive performance of the various EMOS models for different prediction horizons, we also investigate a global dataset of the ECMWF ensemble forecasts of 10-metre daily maximal wind speed with lead times from 1 day up to 15 days initialised at 1200 UTC between 1 January 2014 and 24 June 2018, and the corresponding validating SYNOP observations. Thus, one has observations and corresponding ensemble forecasts with 15 different lead times for the period from 16 January 2014 to 25 June 2018, with the exception of 2 days in between with missing forecast data. For consistency, our analysis is restricted to SYNOP stations with complete data, meaning 1059 stations mostly located in Europe and Asia. The stations considered and depicted on Figure 3.1 have only two overlaps in Germany.

3.2 Implementation details

Since the 8 members of the UWME are non-exchangeable, we employ TN (2.3) and LN (2.5) EMOS models for post-processing. For the GEV and TGEV EMOS models, we utilize the parametrization of (2.8), where $K = 8$. Ensemble forecasts for the calendar year 2008 are calibrated regionally using a 30-day rolling training period, which choice is a result of a detailed preliminary analysis, see Baran and Lerch (2015). The ALADIN-HUNEPS ensemble is structured in a way that naturally divides the ensemble members into two exchangeable groups. The first group includes only the control member, while the second group comprises members derived from random perturbations of the initial conditions ($M = 11,\ K = 2,\ M_1 = 1,\ M_2 = 10$). Hence, regional calibration is performed using EMOS models with distribution locations/means linked to the ensemble members via (2.4), (2.6) and (2.11). The detailed data analysis of Baran et al. (2014) suggests a 43-day rolling training period for EMOS post-processing of ALADIN-HUNEPS ensemble forecasts, leaving 315 calendar days (3 150 forecast cases) between 15 May 2012 and 31 March 2013 for forecast verification.

The 50 members of operational ECMWF EPS are regarded as exchangeable, so in the link functions (2.4), (2.6) and (2.11) we have $K = 1$ and $\overline{f}_1$ equals the ensemble mean. Following the suggestions of Baran and Lerch (2015), the parameters of the EMOS models for calibrating the ECMWF ensemble forecast for the period 1 May 2010 – 30 April 2011 are estimated regionally using a rolling

training period of 20 days. The large ensemble domain does not allow global modelling in the case of the global ECMWF forecasts. Thus, local estimation with a rolling training period of 100 days is applied, which ensures a reasonably stable parameter estimation for all investigated EMOS approaches and leaves the period of 10 May 2014 – 25 June 2018 (1508 calendar days after excluding the two days with missing data) for validation purposes.

In the four case studies presented, the estimated parameters of the TN and LN EMOS models minimise the mean CRPS of the forecast-observation pairs of the training data. The objective functions are optimized using the popular Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (see e.g. Press et al., 2007, Section 10.9). However, the methods of estimation for more complex GEV and TGEV models differ. In the short-range cases in Section 3.3.1, we calculate the maximum likelihood (ML) estimates of the GEV parameters as suggested by Lerch and Thorarinsdottir (2013). For the TGEV model, we achieve positive skewness and a finite mean for the distribution by keeping the shape parameter ξ in the same interval as the ML estimate of the shape of the GEV model, namely $]-0.278, 1/3[$. To implement this, we use the box-constrained version of BFGS (L-BFGS-B; Byrd et al., 1995). For the case of the global forecast in Section 3.3.2, we minimise the mean CRPS to estimate the parameters of both the GEV and the TGEV EMOS. The optimisation algorithm is also the BFGS with a maximum allowed iterations of 200, and the constraints on the scale and shape parameters are done by appropriate transformations. The TN and LN models utilize a linear regression of the observations on the corresponding forecasts to determine the starting parameters for location/mean. The scale parameters have fixed starting points. On the other hand, the GEV and TGEV models begin all iterations from predetermined initial points.

3.3 Results

The forecast skill of the novel TGEV EMOS model proposed in Section 2.2.4 is tested on both the short-range (24–48 hours) wind speed forecasts and on more recent global surface wind forecasts of the operational EPS of the ECMWF with lead times $1, 2, \ldots, 15$ days. We use the TN, LN and GEV EMOS approaches described in Section 2.2.1, 2.2.2 and 2.2.3 as reference models, as well as the raw ensemble and climatological forecasts (observations from the training period are considered as an ensemble).

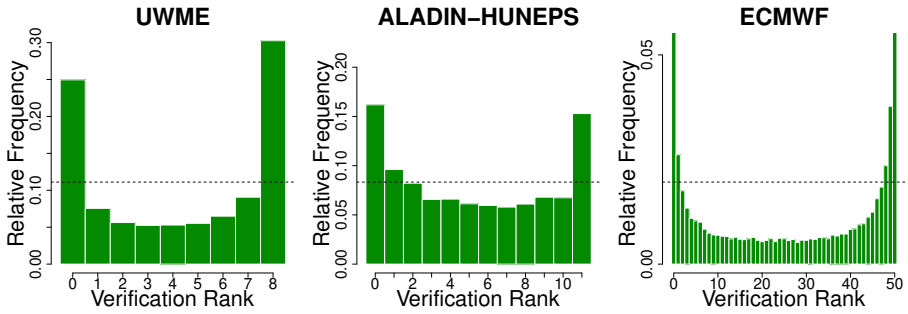

Figure 3.2: Verification rank histograms of raw ensemble forecasts: (*left*) UWME for the calendar year 2008; (*middle*) ALADIN-HUNEPS ensemble for the period 1 April 2012 – 31 March 2013; (*right*) ECMWF ensemble for the period 1 May 2010 – 30 April 2011.

3.3.1 Short-range ensemble forecasts

This section uses the three wind speed data sets examined by Baran and Lerch (2015, 2016). We utilize identical training and verification data for the TGEV modelling, following the same approach as the previous studies. This enables us to directly compare the performance of the TGEV EMOS model with the previously examined TN, LN, and GEV EMOS models.

UWME forecasts

A close examination of Figure 3.2 (*left*) shows that the verification rank histogram of the UWME wind speed forecasts for the calendar year 2008 is strongly U-shaped, indicating that the forecasts are underdispersive. In only 45.24 % of cases does the ensemble range contain the validating observation, which is far below the nominal coverage of 77.78 %, requiring some form of calibration.

Table 3.1 presents a summary of verification scores, coverage, and average width of nominal 77.78 % central prediction intervals for the different EMOS models, as well as the raw and climatological UWME forecasts. Meanwhile, Table 3.2 shows the mean twCRPS values for various thresholds. Climatological forecasts exhibit worse mean CRPS, MAE, and RMSE compared to the raw ensemble, but have better skill on the tails which is quantified in lower mean twCRPS values. However, the raw forecasts suffer from underdispersion,

Forecast	CRPS (m/s)	MAE (m/s)	RMSE (m/s)	Cover. (%)	Av. w. (m/s)
TN	1.114 (1.052,1.188)	**1.550** (1.466,1.655)	2.048	78.65	4.67
LN	1.114 (1.052,1.188)	1.554 (1.465,1.658)	2.052	77.29	4.69
GEV	1.100 (1.041,1.174)	1.554 (1.463,1.656)	2.047	77.20	4.69
TGEV	**1.099** (1.038,1.173)	1.551 (1.464,1.656)	**2.046**	76.69	4.62
Ensemble	1.353 (1.274,1.460)	1.655 (1.554,1.775)	2.169	45.24	2.53
Climatology	1.412 (1.291,1.539)	1.987 (1.820,2.170)	2.629	81.10	5.90

Table 3.1: Mean CRPS and MAE of median forecasts together with 95 % confidence intervals, RMSE of mean forecasts, coverage and the average width of 77.78 % central prediction intervals for the UWME. Mean and maximal probability of predicting negative wind speed by the GEV model: 0.05 % and 4 %.

Forecast	twCRPS (m/s)		
	$r = 9$	$r = 10.5$	$r = 14$
TN	0.150 (0.116,0.189)	0.074 (0.054,0.099)	**0.010** (0.005,0.016)
LN	0.149 (0.115,0.186)	0.073 (0.053,0.098)	**0.010** (0.005,0.017)
GEV	**0.145** (0.112,0.183)	**0.072** (0.052,0.095)	**0.010** (0.005,0.018)
TGEV	**0.145** (0.112,0.180)	**0.072** (0.052,0.096)	**0.010** (0.005,0.017)
Ensemble	0.175 (0.134,0.226)	0.085 (0.061,0.115)	0.011 (0.005,0.019)
Climatology	0.173 (0.132,0.220)	0.081 (0.058,0.111)	**0.010** (0.005,0.017)

Table 3.2: Mean twCRPS for various thresholds r together with 95 % confidence intervals for the UWME.

resulting in poor coverage and overly narrow central prediction intervals. On the other hand, the wider climatological prediction intervals offer improved coverage. EMOS post-processing significantly enhances the calibration and forecast skill of the raw ensemble, as evidenced by lower score values (except for the mean twCRPS at extreme wind speeds) compared to the raw and climatological forecasts. Notably, the mean CRPS shows a significant improvement. The calibrated forecasts achieve coverage close to the nominal value, although the central prediction intervals are less sharp than those derived from the raw ensemble. Among the competing EMOS methods, the novel TGEV model performs best in terms of mean CRPS, RMSE, and twCRPS (comparable to GEV EMOS scores), while slightly trailing behind the TN EMOS method in MAE. Furthermore, the TGEV model produces the narrowest central prediction intervals, albeit with a slight decrease in coverage, which can be expected.

Beyond comparing the twCRPS values reported in Table 3.1, one can get a deeper insight into the tail behaviour of the different EMOS approaches by examining Figure 3.3 showing the twCRPSS with respect to the TN EMOS

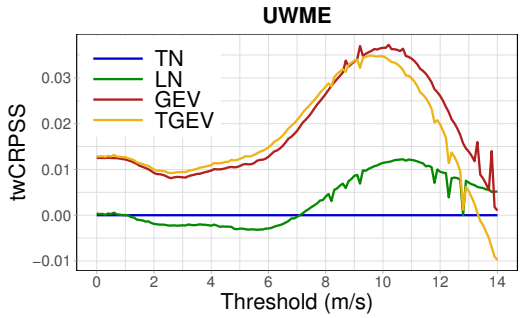

Figure 3.3: twCRPSS values with respect to the TN EMOS model for the UWME.

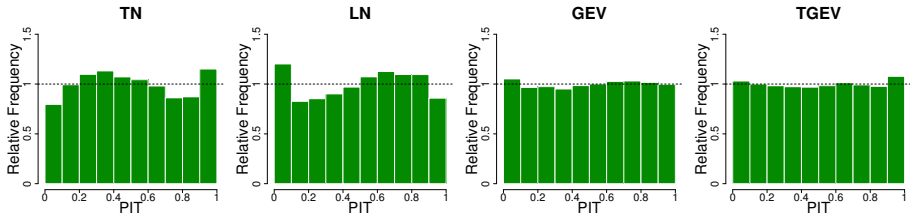

Figure 3.4: PIT histograms of the EMOS-calibrated UWME forecasts.

as a function of the threshold. GEV and TGEV models exhibit very similar behaviour and up to 13 m/s both approaches outperform the TN and LN EMOS methods. For lower threshold values TGEV EMOS results in the highest skill score, but after 8 m/s GEV demonstrates the best predictive performance.

In contrast to the verification rank histogram of the raw UWME forecasts (Figure 3.2 (*left*)), the PIT histograms of the various EMOS models depicted in Figure 3.4 exhibit a much closer resemblance to the desired uniform distribution, suggesting enhanced calibration. The PIT histograms for TN and LN EMOS show slight biases and a hump-shaped pattern, while the histograms for GEV and TGEV approaches appear nearly flat. These observed shapes align well

with the corresponding CRPS values reported in Table 3.1.

Considering the results detailed above, one can conclude that among the competing EMOS approaches for the UWME forecasts, the novel TGEV model demonstrates the highest level of forecast skill, with the GEV EMOS model closely trailing behind. However, it is important to note that the GEV model carries the possibility of predicting negative wind speed values. For the examined UWME forecasts, the average and maximum probabilities associated with these negative values are 0.05 % and 4 %, respectively (Baran and Lerch, 2015).

ALADIN–HUNEPS ensemble

In comparison to the previously discussed UWME, the ALADIN-HUNEPS ensemble exhibits better calibration. Despite some overconfidence, as indicated by the verification rank histogram shown in Figure 3.2 *(middle)* and the presence of larger bins at the edges, it is much closer to a uniform distribution than the histogram in Figure 3.2 *(left)*. Furthermore, the ensemble coverage of 61.21 % is in closer proximity to the nominal value of 83.33 %.

The verification scores and characteristics of the central prediction intervals presented in Table 3.3 provide compelling evidence for the effectiveness of statistical post-processing. All EMOS models generate well-calibrated forecasts with sharp intervals, closely matching the nominal coverage and surpassing the raw and climatological forecasts across all evaluated scores. The improvement brought by statistical calibration is also evident in the mean twCRPS values shown in Table 3.4, although it is important to acknowledge the inherent forecast uncertainty.

Among the different post-processing approaches, the TGEV EMOS stands out by exhibiting the lowest mean CRPS and MAE, along with the sharpest central prediction interval and coverage that is the second closest to the nominal value. However, when considering the twCRPS metric, which assesses predictive performance at high wind speeds, the GEV EMOS demonstrates superior forecast skill. This distinction is further illustrated in Figure 3.5, where the twCRPSS values relative to the TN EMOS are plotted against the threshold. The GEV EMOS consistently outperforms its competitors in this regard. Nevertheless, it is worth noting that the ALADIN-HUNEPS ensemble forecasts have 9.46 % as the maximal probability of predicting negative wind speeds, and the mean probability of such predictions is 0.33 %, which adds a nuanced consideration to the interpretation.

The improved calibration of post-processed ALADIN-HUNEPS forecasts is

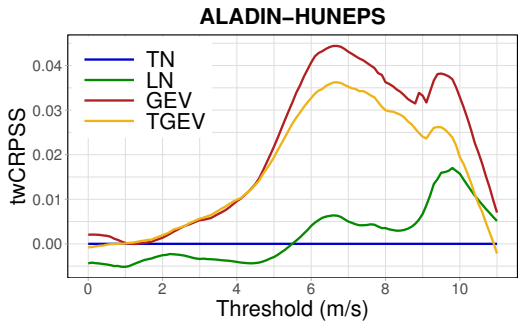

Figure 3.5: twCRPSS values with respect to the TN EMOS model for the ALADIN-HUNEPS ensemble.

evident from the PIT histograms depicted in Figure 3.6, which exhibit a much closer resemblance to a uniform distribution compared to the corresponding verification rank histogram shown in Figure 3.2 *(middle)*. Notably, the TGEV model yields the flattest PIT histogram, while the histograms of the TN, LN, and GEV models display slight biases and hump-shaped patterns. Consequently, among the four presented EMOS approaches for the ALADIN-HUNEPS ensemble forecasts, the TGEV model demonstrates the most favourable overall performance.

Forecast	CRPS (m/s)	MAE (m/s)	RMSE (m/s)	Cover. (%)	Av.w. (m/s)
TN	0.738 (0.689,0.793)	**1.037** (0.966,1.112)	1.357	83.59	3.53
LN	0.741 (0.690,0.799)	1.038 (0.960,1.125)	1.362	80.44	3.57
GEV	0.737 (0.685,0.793)	1.041 (0.970,1.117)	**1.355**	81.21	3.54
TGEV	**0.736** (0.685,0.793)	**1.037** (0.969,1.114)	1.356	82.13	3.53
Ensemble	0.803 (0.749,0.865)	1.069 (1.001,1.136)	1.373	68.22	2.88
Climatology	1.046 (0.944,1.149)	1.481 (1.333,1.627)	1.922	82.54	4.92

Table 3.3: Mean CRPS and MAE of median forecasts together with 95 % confidence intervals, RMSE of mean forecasts and coverage and average width of 83.33 % central prediction intervals for the ALADIN-HUNEPS ensemble. Mean and maximal probability of predicting negative wind speed by the GEV model: 0.33 % and 9.46 %.

Forecast	twCRPS (m/s)		
	$r=6$	$r=7$	$r=9$
TN	0.102 (0.062,0.147)	0.054 (0.027,0.085)	0.012 (0.003,0.022)
LN	0.102 (0.062,0.145)	0.054 (0.028,0.084)	**0.011** (0.004,0.022)
GEV	**0.098** (0.062,0.143)	**0.052** (0.026,0.081)	**0.011** (0.003,0.021)
TGEV	0.099 (0.058,0.145)	**0.052** (0.026,0.082)	**0.011** (0.003,0.022)
Ensemble	0.112 (0.069,0.163)	0.059 (0.030,0.093)	0.013 (0.004,0.026)
Climatology	0.127 (0.076,0.190)	0.064 (0.031,0.102)	0.012 (0.003,0.023)

Table 3.4: Mean twCRPS for various thresholds r together with 95 % confidence intervals for the ALADIN-HUNEPS ensemble.

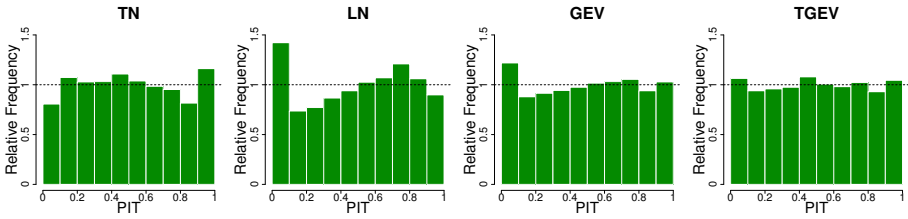

Figure 3.6: PIT histograms of the EMOS-calibrated ALADIN-HUNEPS ensemble forecasts.

ECMWF ensemble forecasts for Germany

Among the three investigated EPSs discussed in Section 3.1.1, it is observed that the ECMWF ensemble displays the greatest lack of calibration. In a majority of cases, the ensemble forecasts either underestimate or overestimate the validating observation, leading to a coverage of 43.40 %, significantly deviating from the nominal coverage of 96.08 %. The underdispersive nature of the forecasts is also evident from the verification rank histogram depicted in Figure 3.2 *(right)*.

Similar to the previous two short-range forecast case studies, in Table 3.5 the mean CRPS, MAE and RMSE of post-processed, raw and climatological forecasts are reported together with the corresponding coverage and average width of 96.08 % (nominal) central prediction intervals. Furthermore, Table 3.6 provides the mean twCRPS scores for three different thresholds. Upon analyzing these values, a similar pattern emerges as observed in previous cases: post-processing leads to enhanced predictive performance and improved calibration. The lowest CRPS, MAE and twCRPS values belong to the TGEV EMOS model,

Forecast	CRPS (m/s)	MAE (m/s)	RMSE (m/s)	Cover. (%)	Av.w. (m/s)
TN	1.045 (0.974,1.125)	1.388 (1.298,1.488)	2.148	92.19	6.39
LN	1.037 (0.970,1.112)	1.386 (1.298,1.482)	2.138	93.16	6.91
GEV	1.034 (0.960,1.114)	1.388 (1.300,1.488)	**2.134**	94.84	8.22
TGEV	**1.031** (0.962,1.112)	**1.385** (1.298,1.480)	2.135	92.89	7.37
Ensemble	1.263 (1.194,1.345)	1.441 (1.373,1.523)	2.232	45.00	1.80
Climatology	1.550 (1.406,1.700)	2.144 (1.948,2.340)	2.986	95.84	11.91

Table 3.5: Mean CRPS and MAE of median forecasts together with 95 % confidence intervals, RMSE of mean forecasts and coverage and average width of 96.08 % central prediction intervals for the ECMWF ensemble forecasts for Germany. Mean and maximal probability of predicting negative wind speed by the GEV model: 0.01 % and 5 %.

Forecast	twCRPS (m/s)		
	$r = 10$	$r = 12$	$r = 15$
TN	0.200 (0.150,0.255)	0.110 (0.075,0.147)	0.042 (0.024,0.062)
LN	0.198 (0.146,0.254)	0.109 (0.075,0.149)	0.042 (0.024,0.062)
GEV	0.195 (0.145,0.250)	**0.106** (0.072,0.145)	**0.041** (0.024,0.059)
TGEV	**0.194** (0.143,0.248)	**0.106** (0.072,0.143)	**0.041** (0.024,0.060)
Ensemble	0.211 (0.155,0.272)	0.113 (0.077,0.152)	0.043 (0.025,0.061)
Climatology	0.251 (0.182,0.326)	0.128 (0.087,0.172)	0.045 (0.026,0.066)

Table 3.6: Mean twCRPS for various thresholds r together with 95 % confidence intervals for the ECMWF ensemble forecasts for Germany.

which has a fair coverage but is slightly less sharp than the TN and LN EMOS.

The mean twCRPS values and their corresponding 95 % confidence intervals for the GEV and TGEV models, as presented in Table 3.6, show minimal differences. However, a closer examination of Figure 3.7, which illustrates the twCRPSS in relation to TN EMOS, highlights the contrasting behaviour in the tails of the two methods, thereby indicating the superiority of the novel TGEV EMOS approach. Additionally, it is worth noting that the GEV model's mean and maximum probabilities of predicting negative wind speed stand at 0.01 % and 5 %, respectively.

Finally, the comparison of the PIT histograms presented in Figure 3.8 with the verification rank histogram of the raw ECMWF ensemble (refer to Figure 3.2 *(right)*) highlights the substantial improvement in forecast calibration through post-processing. However, it is important to note that none of the competing EMOS methods results in perfectly uniform PIT values. For example, the GEV EMOS model exhibits a slight overdispersion with heavy tails, which aligns

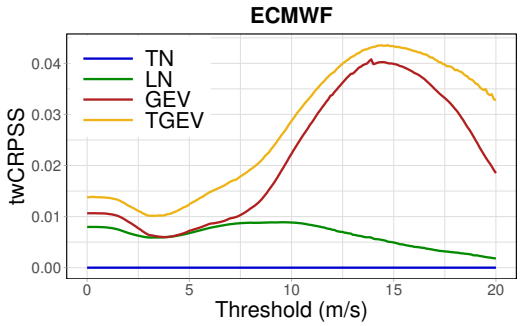

Figure 3.7: twCRPSS values with respect to the TN EMOS model for the ECMWF forecasts for Germany.

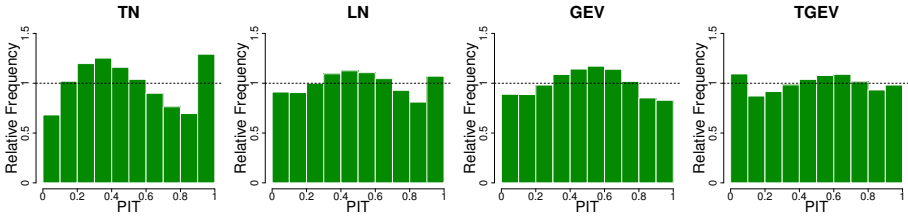

Figure 3.8: PIT histograms of the EMOS-calibrated ECMWF forecasts for Germany.

with the wider nominal central prediction intervals reported in Table 3.5. On the other hand, the TN EMOS model displays slightly lighter tails. Among the competing methods, the TGEV and LN EMOS models demonstrate the smallest deviation from uniformity. Therefore, for the ECMWF forecasts under investigation, the TGEV EMOS model shows the best overall performance.

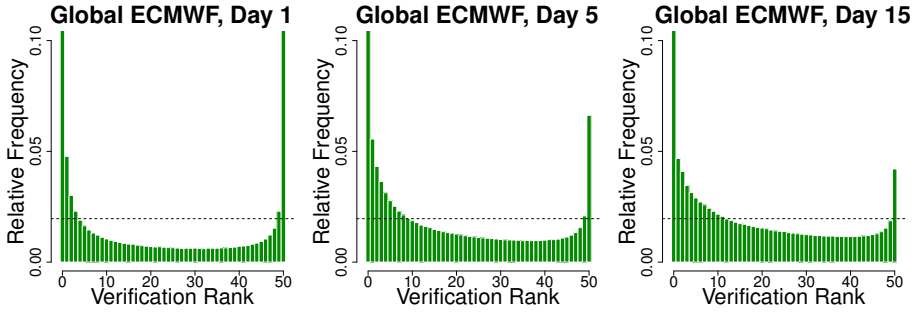

Figure 3.9: Verification rank histograms of the global ECMWF ensemble forecasts for the period 16 January 2014 – 25 June 2018.

3.3.2 Global ECMWF forecasts with different forecast horizons

The case studies of Section 3.3.1 verify the positive effect of EMOS post-processing on the calibration of short-term wind speed ensemble forecasts in general, and the superiority of the TGEV EMOS approach as well. However, as argued in the discussion of Feldmann et al. (2019), the longer the lead time, the more training data is needed for post-processing to outperform the raw ensemble and a similar conclusion can be derived from the results of Baran et al. (2020), too. This motivates the case study presented in this section, where calibration of global ECMWF wind speed ensemble forecasts with lead times $1, 2, \ldots, 15$ days covering a very long time period of almost four and a half years is considered.

The verification rank histograms of Figure 3.9 show that the global ECMWF forecasts are strongly U-shaped for all lead times; however, the increase of the forecast horizon reduces underdispersion. This could be due to the increase of forecast uncertainty resulting in a wider ensemble range and better coverage, which improves from 52.05 % of day 1 to 85.74 % of day 15 (see also Figure 3.12).

In contrast to previous studies on ECMWF temperature forecasts (Feldmann et al., 2019; Baran et al., 2020), the mean CRPS analysis reveals a significant improvement in forecast performance for all lead times when considering EMOS models compared to raw wind speed ensemble forecasts (Figure 3.10 *(right)*). It is worth noting that the non-monotonic shape of the mean CRPS for the raw

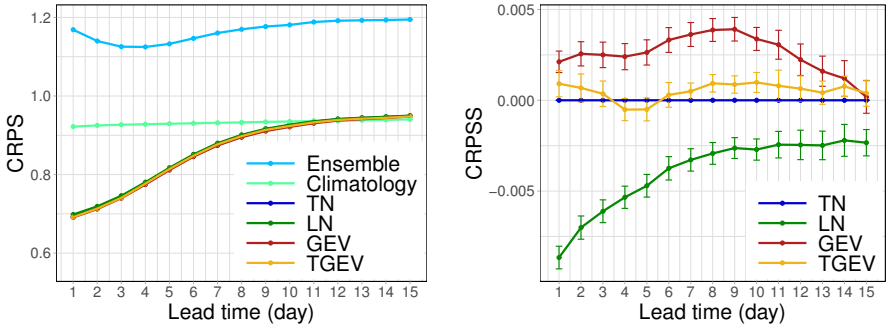

Figure 3.10: *(left)* CRPS of the raw, climatological and calibrated ECMWF global forecasts; *(right)* CRPSS with respect to the TN EMOS model together with 95 % confidence intervals.

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Mean	2.48	2.48	2.48	2.49	2.51	2.53	2.54	2.56	2.57	2.59	2.60	2.61	2.63	2.63	2.65
Q90	7.36	7.30	7.28	7.32	7.32	7.30	7.30	7.30	7.37	7.42	7.48	7.51	7.58	7.59	7.61
Q95	14.20	13.95	13.76	13.59	13.38	13.14	13.02	12.92	13.00	13.02	12.99	13.12	13.22	13.25	13.30
Q99	32.95	32.32	31.65	30.82	29.79	29.23	28.53	28.18	27.90	27.83	27.63	27.84	27.84	27.77	27.94

Table 3.7: Mean and the 90th, 95th and 99th quantiles of probabilities (in %) of predicting negative wind speed by the GEV model.

ensemble is attributed to representativeness error in the verification process, which can be partially mitigated by incorporating observation uncertainty into the ensemble spread (Ben Bouallègue, 2020). EMOS models also outperform climatology for shorter lead times, although this advantage diminishes as the lead time increases and disappears after day 11. To highlight the differences between the various EMOS approaches in terms of the mean CRPS, Figure 3.10 *(middle)* presents the CRPSS values relative to the TN EMOS model. LN EMOS exhibits the lowest forecast skill, but this disadvantage decreases for longer forecast horizons. GEV EMOS demonstrates superior performance compared to its competitors, followed by TGEV EMOS, which consistently achieves significantly positive skill scores across most lead times. However, it is important to note that the issue of predicting negative wind speed values is more prominent in the GEV EMOS approach in this case compared to the

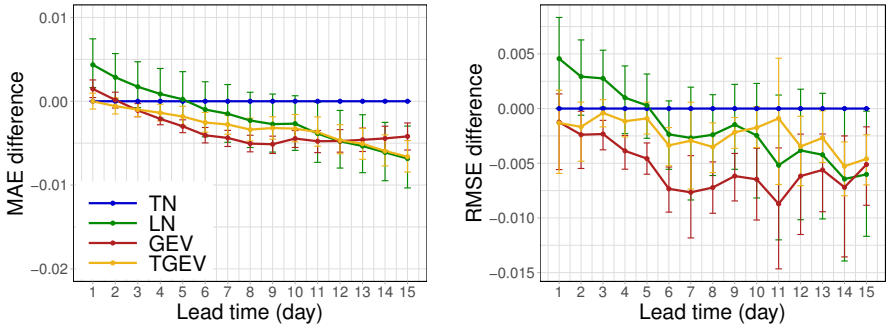

Figure 3.11: Difference in MAE *(left)* and RMSE *(right)* values from the reference TN EMOS model together with 95 % confidence intervals.

studies discussed in Section 3.3.1. According to Table 3.7, the mean probability of predicting negative wind speeds is approximately 2.5 %, with 99th quantiles ranging from 27.63 % to 32.95 %, posing challenges for potential operational applications.

Figures 3.11 *(left)* and 3.11 *(right)* present the differences in MAE and RMSE compared to the reference TN EMOS model, respectively, where smaller values indicate better performance. For short and very long lead times the TGEV EMOS results in the lowest MAE values, whereas between 4 and 10 days the GEV EMOS significantly outperforms its competitors. After day 11 the performance of the LN EMOS is similar to that of the TGEV EMOS; however, the uncertainty of the former is much higher. In terms of RMSE, a different ranking is observed in Figure 3.11 *(right)*. The GEV EMOS yields the lowest scores, followed by the TGEV EMOS model, which for medium lead times behaves very similarly to the LN EMOS. It is important to note that the calculation of EMOS means may occasionally encounter numerical issues for all models, leading to unrealistic squared errors. To address this, forecast cases with absolute errors exceeding 100 m/s (less than 0.5 % of the total cases) are excluded from the analysis.

As anticipated, climatological forecasts exhibit the highest coverage (Figure 3.12 *(left)*, closely followed by the GEV EMOS approach. The coverage values for the TGEV, TN, and LN EMOS models are slightly below 90 % across all lead times, with relatively flat curves close to each other. In terms of sharp-

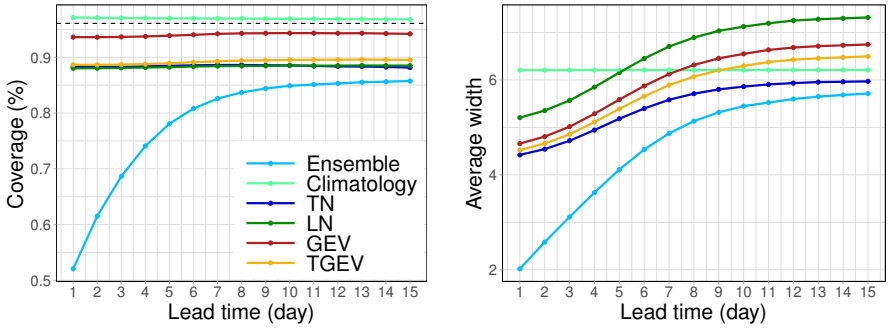

Figure 3.12: Coverage *(left)* and average width *(right)* of nominal 96.08 % central prediction intervals. In the *(left)* panel the ideal coverage is indicated by the horizontal dotted line.

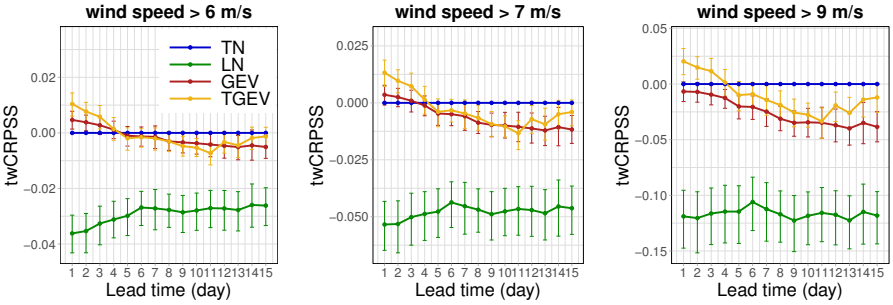

Figure 3.13: twCRPSS values with respect to the TN EMOS model for thresholds 6 m/s, 7 m/s and 9 m/s together with 95 % confidence intervals.

ness, Figure 3.12 *(right)* shows a clear ranking of the competing post-processing methods. TN EMOS has the narrowest central prediction intervals followed by the TGEV, the GEV and the LN EMOS models.

To assess the tail behaviour of the different EMOS models, we examine the twCRPSS values relative to the TN EMOS approach at thresholds corresponding to the 90th, 95th, and 98th quantiles of the wind speed observations (see

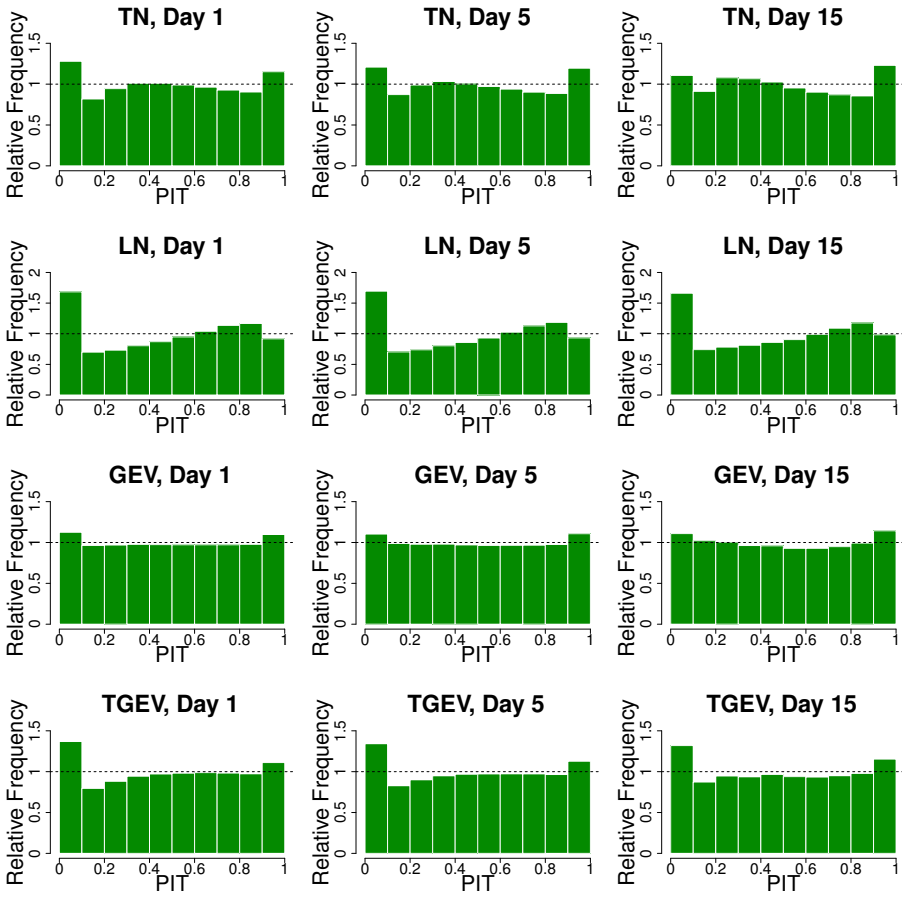

Figure 3.14: PIT histograms of the EMOS post-processed ECMWF global forecasts for days 1, 5 and 15.

Figure 3.13). The ranking of the different EMOS models remains consistent across all three thresholds, with variations primarily observed in their relationship to the reference TN EMOS. After day 3 TN EMOS results in the best forecast skill, whereas the LN EMOS approach, similar to Figure 3.10 *(right)*,

is far behind its competitors.

The PIT histograms of the EMOS post-processed forecasts for lead times of 1, 5, and 15 days, as shown in Figure 3.14, highlight the positive impact of post-processing. These histograms exhibit a much closer resemblance to uniformity compared to the verification rank histograms of the raw ECMWF ensemble forecasts depicted in Figure 3.9. Furthermore, the shapes of the presented PIT histograms align well with the corresponding CRPS scores illustrated in Figure 3.10, as well as the coverage and average widths of nominal central prediction intervals shown in Figure 3.12. The PIT histograms of the LN EMOS approach display the largest deviation from uniformity, while the histograms of the GEV model appear nearly flat with a slight underdispersion, particularly for longer lead times. The TGEV EMOS also yields relatively flat PIT histograms with slightly lighter lower tails for all lead times.

For the ECMWF data set at hand across all lead times, the GEV EMOS model shows the best overall predictive performance, followed by the TGEV EMOS. Considering the mean probabilities of predicting negative wind speed by the GEV model presented in Table 3.7, it is advisable to opt for the slightly less skilful yet more reliable TGEV EMOS approach.

3.4 Conclusions

We proposed a novel distribution-based approach, called TGEV EMOS, to calibrate wind speed ensemble forecasts and address the issue of occasionally predicting negative wind speed encountered in the GEV EMOS method by Lerch and Thorarinsdottir (2013). We evaluate the performance of the TGEV EMOS model on short-range forecasts from three different ensemble prediction systems and a large dataset of global ECMWF forecasts spanning several years. Verification is done using CRPS, MAE, and RMSE scores for probabilistic and point forecasts, along with the analysis of the coverage and the average width of central prediction intervals. Additionally, we assess the model's performance in predicting high wind speed values using twCRPS.

Comparing the TGEV EMOS model with TN, LN, and GEV EMOS approaches, as well as raw and climatological forecasts, we consistently find that post-processing improves calibration and accuracy. The TGEV EMOS model exhibits the best overall performance among the four methods considered, closely followed by the GEV EMOS model. However, it should be noted that the GEV model has a mean probability of predicting negative wind speed values around 2.5 % for all lead times, as observed in the case study of Section 3.3.2.

Regarding the performance of extreme events evaluated with the twCRPS and twCRPSS, the four case studies show a slight difference between the GEV and TGEV models, half of them favouring the non-truncated version, the other half the truncated one. Moreover, the lighter-tailed TN model even outperformed the competing approaches in the global case. This suggests that there might be room for further investigation in this area (see e.g.), but one should keep the challenges that forecast evaluation of extreme events poses in mind.

It is also important to note that in most of the case studies regarding many of the verification metrics, the confidence intervals of the compared approaches have a substantial overlap. This indicates that the differences in predictive performance between the distributional models tend to be small. Note that Baran and Baran (2021) also compared the skill of TN, LN and TGEV EMOS models for calibrating the short-term 100-metre wind speed forecasts of the AROME-EPS of the Hungarian Meteorological Service and found a slightly different ranking of the competing approaches. In their case study, the TN EMOS model outperformed the TGEV EMOS approach, but all EMOS models were surpassed by a distributional regression network approach based on a TN predictive distribution.

Chapter 4

Calibration of dual-resolution temperature forecasts

This chapter of the dissertation presents the findings of a comprehensive investigation into the probabilistic skill of dual-resolution ensemble forecasts, with a specific focus on the connection between spatial resolution and ensemble size. It is widely recognized that balancing between these two factors can significantly impact forecast performance (Mullen and Buizza, 2002; Raynaud and Bouttier, 2017; Leutbecher and Ben Bouallègue, 2020). Recently, dual-resolution ensembles have emerged as a promising approach to explore this balance by utilising ensemble members with different spatial resolutions to generate probabilistic forecasts.

By leveraging the same datasets, we can build upon the findings of Leutbecher and Ben Bouallègue (2020) and further examine the impact of statistical post-processing on the optimal dual-resolution configuration. This allows us to assess the generalisability and robustness of their conclusions in the context of calibrated dual-resolution ensemble forecasts. The investigation focuses on medium-range dual-resolution ensemble forecasts of 2-metre temperature (K), employing the Integrated Forecast System (IFS) of ECMWF with horizontal resolutions ranging from 18 to 45 km and ensemble sizes varying from 8 to 254 members.

TCo399 - TCo639 1:4				TCo255 - TCo639 1:16			
LHPC		SHPC		LHPC		SHPC	
M_L	M_H	M_L	M_H	M_L	M_H	M_L	M_H
0	50	0	8	0	16	0	8
40	40	8	6	16	15	16	7
120	20	16	4	32	14	32	6
160	10	24	2	64	12	64	4
200	0	28	1	128	8	128	0
		32	0	254	0		

Table 4.1: The investigated dual-resolution mixtures.

4.1 Data

By considering the same datasets as Leutbecher and Ben Bouallègue (2020), we establish a strong basis for the comparison and build on the existing knowledge in the field of dual-resolution ensemble forecasting. Both ensemble forecasts and observation data of 2-metre temperature were provided by the ECMWF for this study. We calibrated the global medium-range ensemble forecasts of the IFS of ECMWF with three horizontal resolutions, namely:

- 50 members at TCo639 (grid resolution ∼18 km),
- 200 members at TCo399 (grid resolution ∼29 km),
- 254 members at TCo255 (grid resolution ∼45 km).

The TCo639 ensemble with the highest resolution was the operational ECMWF medium-range ensemble at the time of the study, while the lower-resolution TCo399 and TCo255 ensembles were generated with the same model version as the operational ensemble at the time (cycle 41r2). The investigation period is the boreal summer of 2016, with each ensemble forecast initialised once daily between 1 June and 31 August 2016, resulting in a total of 92 days.

The aspect of interest, besides the horizontal resolution of the ensembles, is the cost ratio between the different forecasts. Table 4.1 shows the different combinations of lower- (M_L) and higher-resolution (M_H) ensemble members we considered, and their corresponding cost ratios in the second row. This means that when constructing different dual-resolution configurations, one TCo639

member can be traded against four TCo399 members or against sixteen TCo255 members. The configurations are separated into a mixture of TCo399–TCo639 ensembles and TCo255–TCo639 ensembles. The mixtures were composed with the first n members of the different ensembles. Additionally, two scenarios have been considered, corresponding to different assumptions concerning the availability of high-performance computing (HPC) resources. The first scenario, referred to as the large supercomputer (LHPC) scenario, assumes the availability of HPC resources, which were utilised in 2018 at the ECMWF. The second scenario, referred to as the small supercomputer (SHPC) scenario, is based on the assumption of one-sixth of these resources. Notably, TCo399–TCo639 and TCo255–TCo639 combinations are based on different ensemble sizes of TCo639 members, leading to two different LHPC scenarios. This technicality arises from the fact that the cost of 50 TCo639 members is equivalent to that of 800 TCo255 members, and the largest possible ensemble size that can be handled with the current GRIB settings is 255, storing the ensemble size in a single byte.

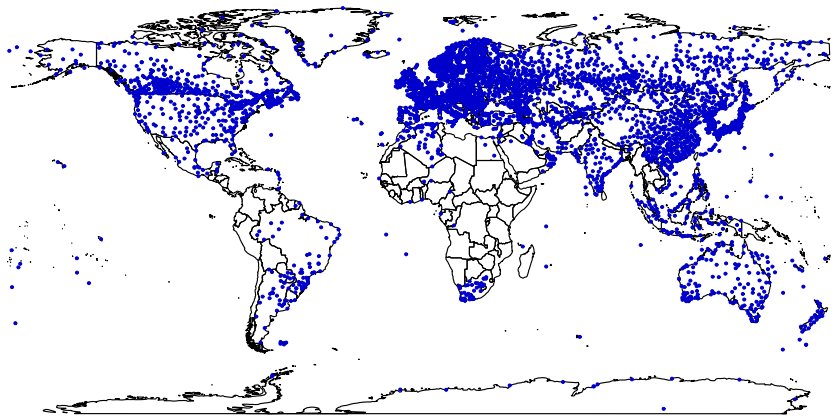

Figure 4.1: Map of the 4560 SYNOP stations across the globe.

To supplement the forecasts and to be able to apply the models and also verify them, we used measurements from surface SYNOP stations. These measurements are reported from various locations around the globe with high densities in Europe and North America and low observation densities in tropical and subtropical regions. Filtering was applied to these observations as the number of measurements stored can vary from day to day. After it, we were able to use a subset of 4,560 stations with full availability over the verification period.

Forecasts are evaluated by comparing them to observations at the nearest grid point of the native forecast grid. However, due to the coarse representation of the orography in the model, systematic representativity errors may arise. To address this issue, an orographic correction is applied to the raw temperature forecasts. This correction involves adjusting the forecasts linearly, based on the Δz height difference between the station and the model representation. The following formula based on the temperature lapse rate was utilised: $\Delta T/\Delta z = -0.0065$ K/m, where ΔT denotes the correction calculated for the temperature forecast. By incorporating this orographic correction, we aim to mitigate the potential biases introduced by the coarse orography description and improve the accuracy of the forecasted temperatures at each station.

4.2 Implementation details

The choice of the training data is important for statistical post-processing. Here, we focus on the clustering-based semi-local estimation (see Section 2.5), where the observation sites are grouped into clusters using k-means clustering of the stations with 24-dimensional feature vectors comprising 12 equidistant quantiles of the climatological CDF and 12 equidistant quantiles of the empirical CDF of forecast errors for the ensemble mean during the training period. Regional parameter estimation is then performed within each cluster. With the help of this method one can get reliable parameter estimates even for short training periods and the obtained models may outperform the local EMOS approach (Lerch and Baran, 2017), hence we mainly focus on the semi-local results in Section 4.3.

Due to the limited time period of the available dataset, which encompasses only the boreal summer of 2016 spanning 92 calendar days, careful consideration is required to strike a balance between a sufficiently reliable parameter estimation and an adequate amount of data for model verification. To address this trade-off, a rolling 30-day training period is employed for calibration. Consequently, verification scores are computed for ensemble forecasts initiated between 1 July and 31 August 2016, along with the corresponding validating observations. It should be noted that the forecast periods are shifted by 1–15 days depending on the lead times of the ensemble predictions.

A total of 200 clusters are considered, resulting in a comparable mean number of stations per cluster as in Lerch and Baran (2017). Local EMOS estimates 4–5 parameters based on 30 forecast-observation pairs, while semi-local EMOS estimates the same parameters using around 600 forecast-observation

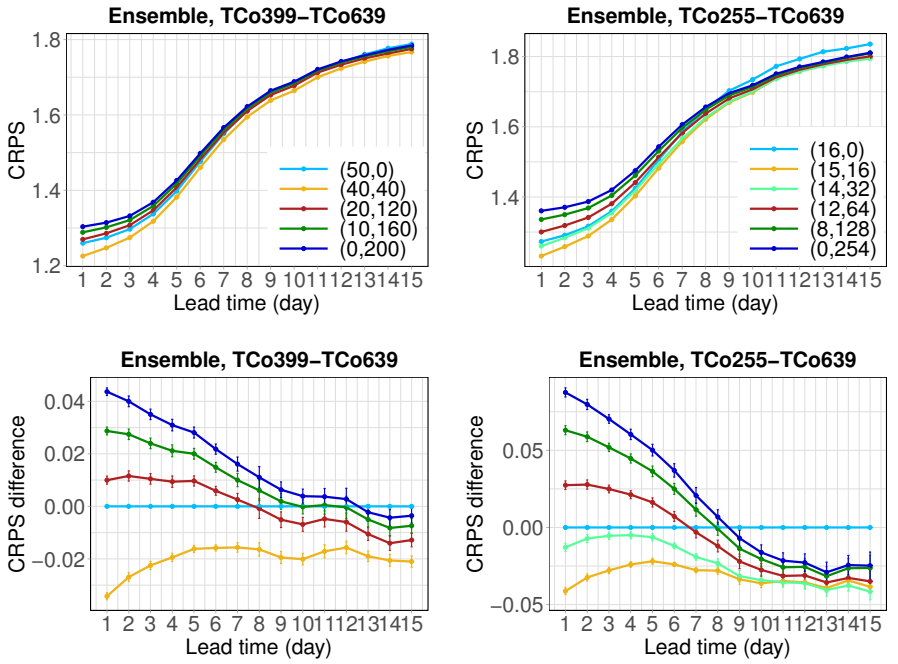

Figure 4.2: Mean CRPS values (the lower the better) of global dual-resolution ensemble forecasts for 2-metre temperature (*top*) and the difference in mean CRPS (the lower the better) from the reference pure high-resolution ensemble (*bottom*) with 95 % confidence intervals, LHPC scenario.

pairs. Consequently, the latter approach is expected to provide more constrained parameter estimates. In order to highlight the distinctions between local and semi-local approaches, a very short 10-day training period is also examined.

In our case, the ensemble members of a given resolution can be considered exchangeable, thus the normal distribution-based model defined in (2.2) is implemented with the following link functions:

$$\mu = a + b_H \overline{f}_H + b_L \overline{f}_L \qquad \text{and} \qquad \sigma^2 = c + d\overline{f}, \tag{4.1}$$

where $\overline{f}_H$ and $\overline{f}_L$ denote the mean of high- and low-resolution members,

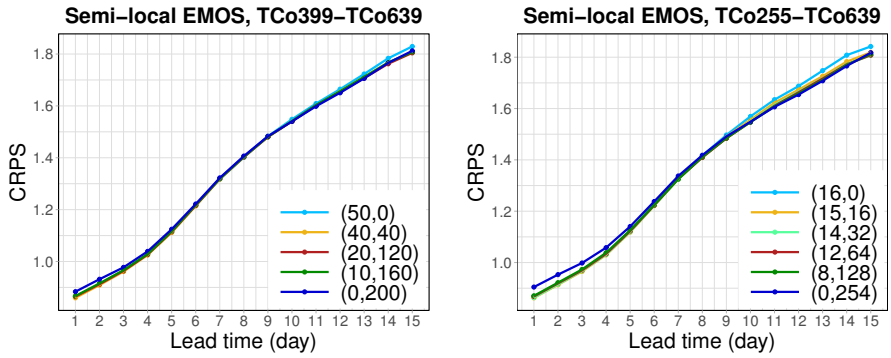

Figure 4.3: Mean CRPS values (the lower the better) of semi-local EMOS post-processed global dual-resolution ensemble forecasts for 2-metre temperature, LHPC scenario.

respectively. Model parameters are estimated by minimizing the mean CRPS over the training data where we fix $b_L = 0$ in the pure high-resolution $(M_L = 0)$ and $b_H = 0$ in the pure low-resolution $(M_H = 0)$ case.

To ensure consistency with the results obtained from the raw ensemble, EMOS predictive distributions are derived using orographically corrected ensemble forecasts. While local EMOS does not require this preliminary bias correction, employing the orographic correction for semi-local EMOS leads to improved skill in terms of verification scores. This is due to the fact that the orographic correction allows for the representation of local variability within each cluster, thus enhancing the overall performance of the semi-local EMOS approach.

4.3 Results

4.3.1 Calibration of mixtures for large supercomputer

The analysis of raw ensemble forecasts reveals that both resolution combinations consistently favour balanced mixtures across all lead times. Specifically, the combination (40,40) for TCo399 - TCo639 and (15,16) for TCo255 - TCo639 demonstrate a preference for balanced ensembles, which aligns with the findings

of Leutbecher and Ben Bouallègue (2020). This pattern is clearly depicted in Figure 4.2, which illustrates the mean CRPS values of dual-resolution ensemble forecasts for 2-metre temperature, as well as the difference in mean CRPS compared to the pure high-resolution case, as a function of lead time.

The application of statistical post-processing significantly alters these findings. Figure 4.3 demonstrates that the implementation of semi-local EMOS leads to a notable decrease in the mean CRPS across all lead times (but day 15). The disparities in predictive performance among the different mixture combinations are greatly diminished, particularly for TCo399 – TCo639. Although not shown here, local EMOS produces similar outcomes.

Figure 4.4 illustrates the CRPSS compared to the pure high-resolution case. In the semi-local case, from day 8 onwards, the pure low-resolution ensemble exhibits superior performance to the pure high-resolution ensemble for both TCo399 – TCo639 and TCo255 - TCo639 resolutions. However, for shorter lead times, semi-local EMOS still favours balanced combinations, which align with the optimal choices identified in the raw ensemble. It is worth noting that the score differences associated with local EMOS become more variable for longer lead times, resulting in wider confidence intervals. As a result, the subsequent section primarily focuses on presenting the results obtained through semi-local EMOS calibration.

The analysis of Brier skill scores using thresholds corresponding to the $5, 10, \ldots, 95$ percentiles of the station sample climatology during the verification period yielded similar findings. BSS values with respect to the pure high-resolution case of semi-local EMOS post-processed forecasts displayed in Figure 4.5 are fully consistent with the graphs in the top row of Figure 4.4. Specifically, on days 1 and 5, the balanced combinations demonstrate the highest forecast skill across all thresholds, while on day 10, the other combinations equal or even outperform the pure high-resolution case.

Figure 4.6 presents the quantile skill scores for ensemble configurations post-processed using semi-local EMOS, with respect to the pure higher-resolution configuration. The results for the median (50% quantile) align with the CRPS differences, as seen in the middle row. For the more extreme quantiles (2% and 98%), shown in the top and bottom rows, the score differences are comparable to those of the median, but the confidence intervals tend to be wider. It is important to note that, for longer lead times, the quantile score differences between configurations are statistically not strongly significant.

We explored the question, of whether stations exhibiting a significant difference in mean CRPS between the optimal combination and the reference pure high-resolution case display any clear spatial patterns. However, visualizing the

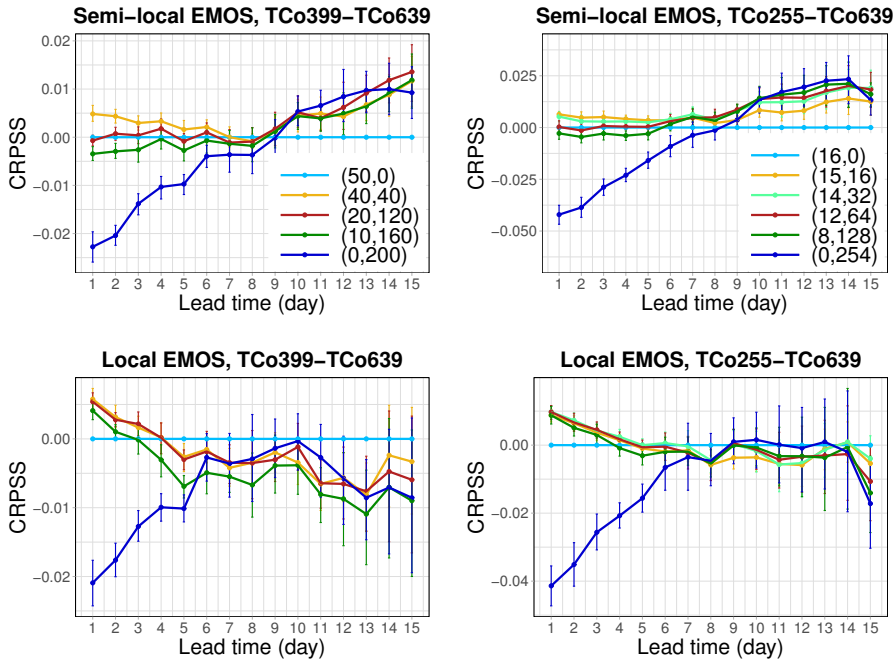

Figure 4.4: CRPSS from the reference pure high-resolution case with 95 % confidence intervals of semi-local (*top*) and local (*bottom*) EMOS post-processed global dual-resolution ensemble forecasts for 2-metre temperature, LHPC scenario.

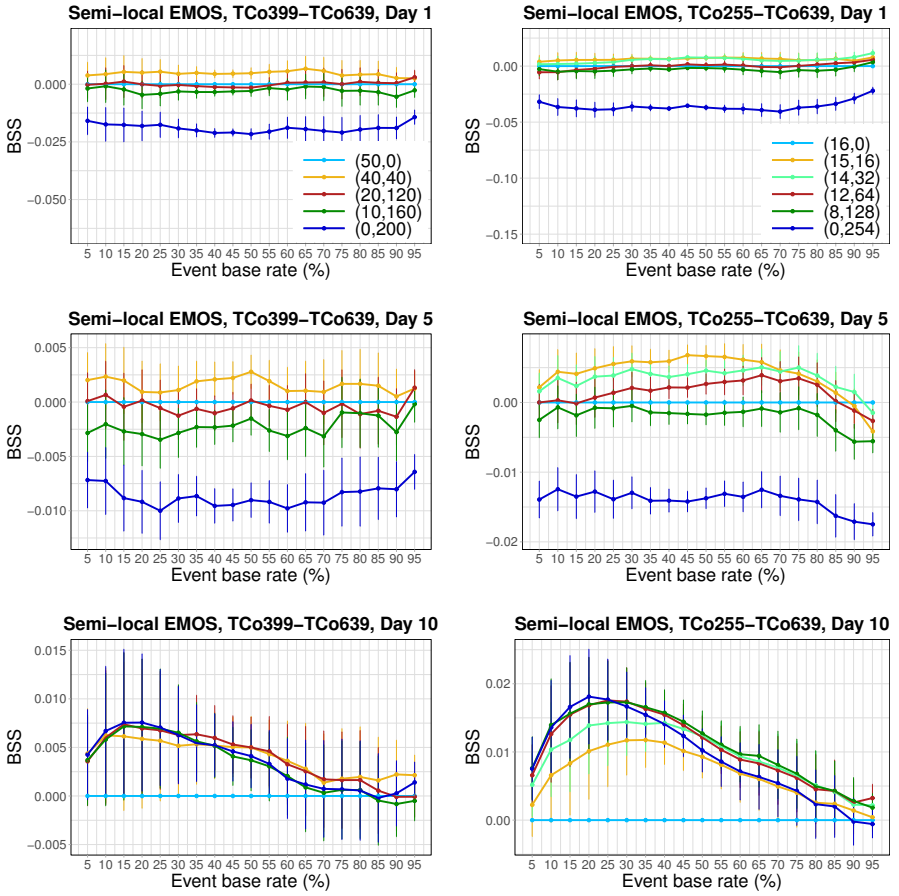

Figure 4.5: Brier skill scores (the higher the better) with respect to the reference pure high-resolution case with 95 % confidence intervals of semi-local EMOS post-processed global dual-resolution ensemble forecasts for 2-metre temperature, LHPC scenario.

stations with a significant difference at a 5 % level on maps did not reveal any apparent connection to their geographical locations.

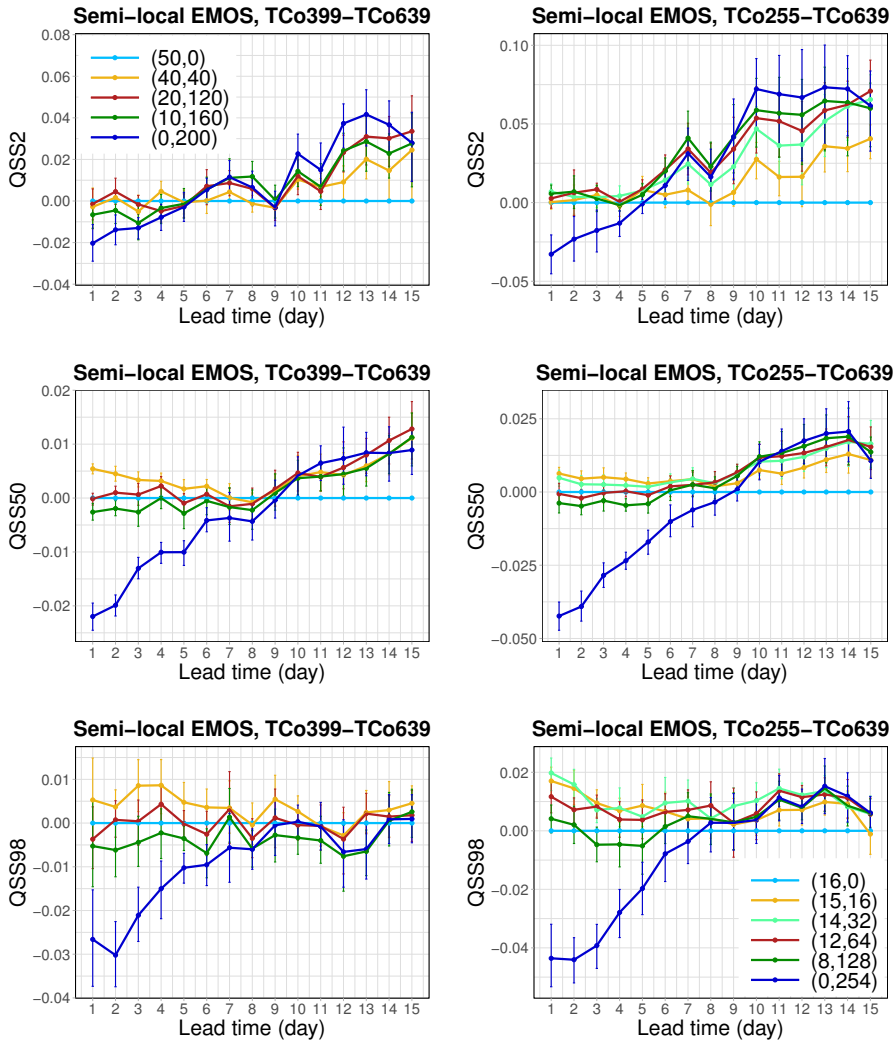

Figure 4.6: Quantile skill scores (the higher the better) for percentiles 2 (*top*), 50 (*middle*) and 98 (*bottom*) with respect to the reference pure high-resolution case with 95 % confidence intervals of semi-local EMOS post-processed global dual-resolution ensemble forecasts for 2-metre temperature, LHPC scenario.

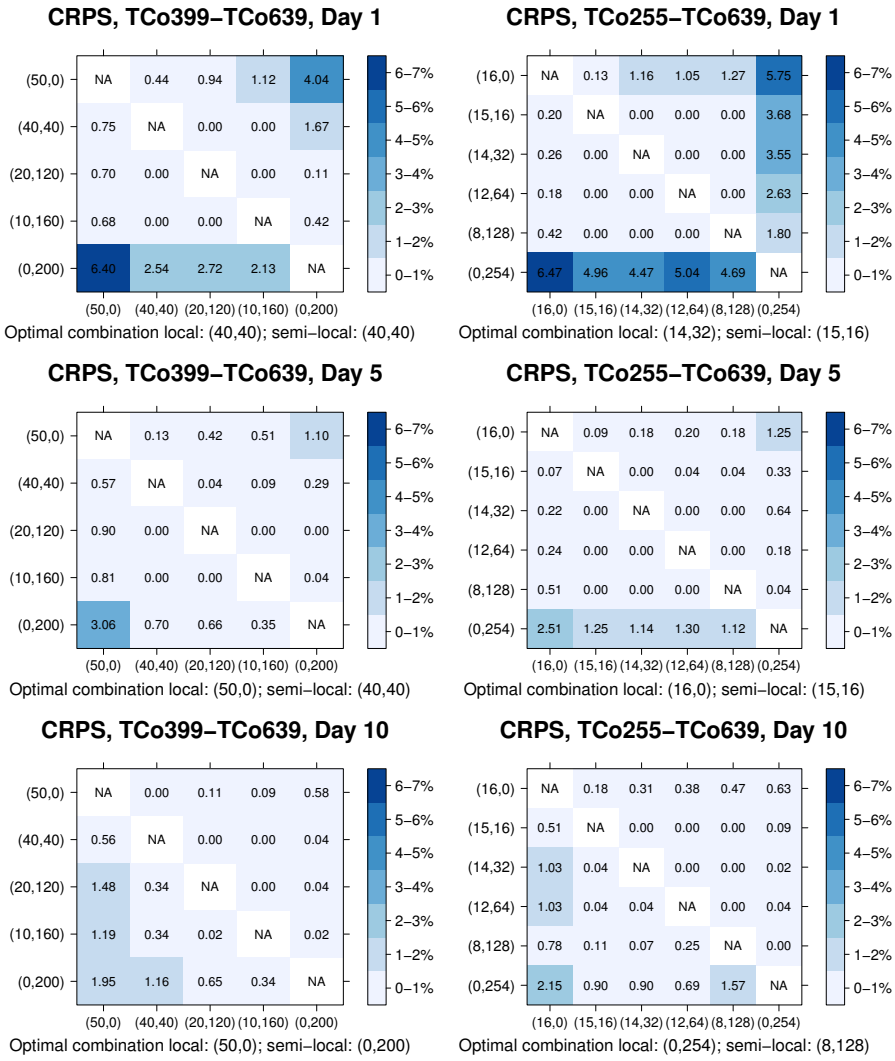

Figure 4.7: Proportion of stations with significantly different mean CRPS at a 5 % level for different lead times for local (lower triangle) and semi-local (upper triangle) parameter estimation approaches, LHPC scenario.

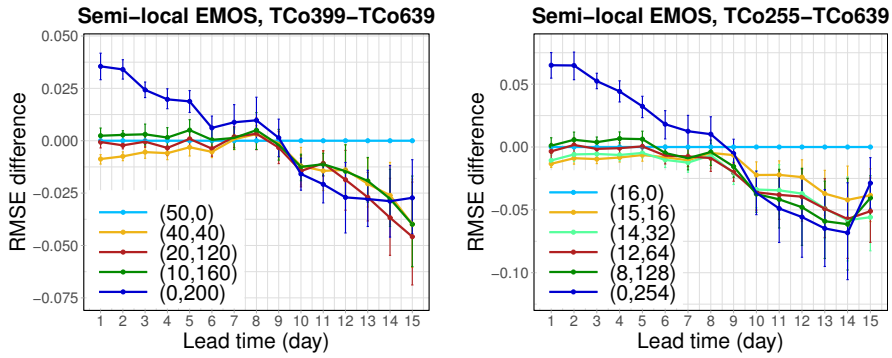

Figure 4.8: Difference in RMSE values (the lower the better) from the reference pure high-resolution case with 95 % confidence intervals of semi-local EMOS post-processed global dual-resolution ensemble forecasts for 2-metre temperature, LHPC scenario.

To assess the statistical significance of differences in mean CRPS values across various combinations, station-wise DM tests were conducted. Figure 4.7 presents the proportions of stations with a significant difference in mean CRPS at a 5% significance level for different lead times. The lower triangle corresponds to results obtained using the local parameter estimation approach, while the upper triangle represents the findings from the semi-local approach. Generally, as lead times increase, the proportion of stations with a significant difference decreases for both local and semi-local EMOS post-processing. The values in the first column and row of each matrix align with the observations presented in Figure 4.4.

Finally, let us consider the root mean squared errors of EMOS mean forecasts. Figure 4.8 illustrates the differences in RMSE values for semi-local EMOS post-processed combinations compared to the reference pure high-resolution case. It is worth noting that the graphs for both mixtures closely resemble the ones in the top row of Figure 4.4. This similarity in the observed values for CRPSS (Figure 4.4, top) and RMSE (not shown) can also be observed for local EMOS.

In the LHPC scenario, the examined verification scores consistently show that the balanced combination (40,40) for TCo399 - TCo639 performs the best

up to day 9, while all combinations outperform the pure high-resolution case from day 10 onward, with small differences observed for combinations involving TCo399 forecasts. This contrasts with the results for the raw ensemble where the balanced combination consistently exhibits the best forecast skills for all lead times (Figure 4.2, left column). For TCo255 – TCo639, the balanced combination (15,16) is clearly preferred up to day 5, and again all combinations outperform the pure high-resolution case from day 10 onward, with more pronounced differences compared to the other mixture. However, for long lead times, the pure low-resolution ensemble performs well, and the ordering of combinations deviates from that based on the raw ensemble forecasts, which consistently identify the balanced combination as the best for all lead times (Figure 4.2, right column).

4.3.2 Calibration of mixtures for small supercomputer

In the small HPC scenario, when considering the raw ensemble, the balanced combinations ((6,8) for TCo399 - TCo639 and (7,16) for TCo255 - TCo639) show the highest skill for short lead times. However, for longer lead times, larger ensemble sizes perform better, with the pure low-resolution ensemble exhibiting the best forecast skill (Figure 4.9).

Semi-local EMOS post-processing significantly improves calibration in terms of mean CRPS and reduces differences between combinations, aligning with LHPC results (Figure 4.10). Figure 4.11 displays skillscores for local and semi-local EMOS, both showing a consistent ranking. Balanced combinations perform better for shorter lead times, while larger ensemble sizes are preferred for longer lead times.

The statistical significance of score differences between configurations has been computed station-wise (not shown). The proportion of stations with significant differences decreases with longer lead times. Compared to the LHPC scenario, the proportions on day 1 are similar to the top line of Figure 4.7. At day 5, the proportions are smaller for both TCo399 - TCo639 and TCo255 - TCo639 mixtures in the SHPC scenario. However, at day 10, more stations in SHPC show significant differences in mean CRPS compared to LHPC, likely due to the greater importance of ensemble size at longer lead times in the SHPC scenario.

For TCo399 - TCo639 mixtures (Figure 4.12, left column), the Brier skill scores with respect to the post-processed pure high-resolution case show that the most balanced combination (6,8) performs best at day 1. At day 5, combinations that include low-resolution members outperform the post-processed pure

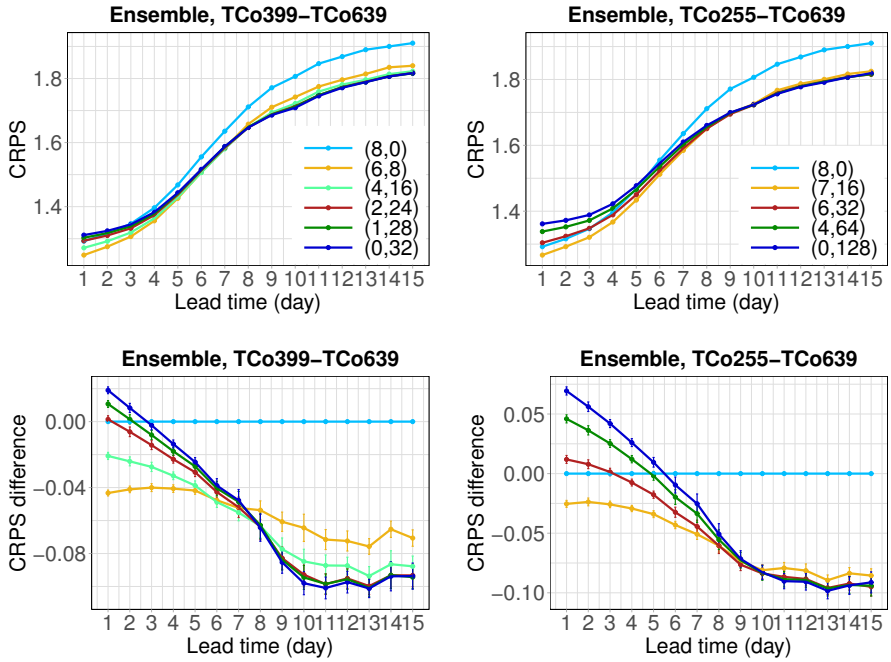

Figure 4.9: Mean CRPS values (the lower the better) of global dual-resolution ensemble forecasts for 2-metre temperature (*top*) and the difference in mean CRPS (the lower the better) from the reference pure high-resolution ensemble (*bottom*) with 95 % confidence intervals, SHPC scenario.

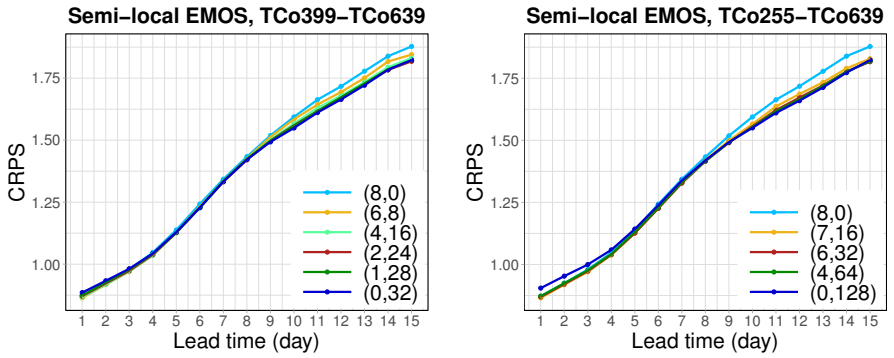

Figure 4.10: Mean CRPS values (the lower the better) of semi-local EMOS post-processed global dual-resolution ensemble forecasts for 2-metre temperature, SHPC scenario.

high-resolution ensemble. On day 10, the ordering clearly reflects the ensemble size, with larger ensembles performing better. For TCo255 - TCo639 mixtures (Figure 4.12, right column) in general, larger ensemble sizes are considered. On days 1 and 5, the 128-member post-processed pure low-resolution ensemble performs worse than the 8-member post-processed pure high-resolution ensemble. The effect of ensemble size dominates only at day 10. The results for Brier scores are consistent with those for CRPSS (see Figures 4.12 and 4.11, respectively).

Similarly to the LHPC scenario, the differences in 50 % quantile skill scores with respect to the post-processed pure high-resolution case align with the CRPS results. For the tails, the performance improvement of the different combinations compared to the pure high-resolution case is similar to the LHPC scenario, and the differences are often not significant, especially for the TCo399 - TCo639 mixture.

Furthermore, the evaluation of the mean accuracy of post-processed dual-resolution ensemble forecasts using root mean squared error (RMSE) yields a consistent pattern with the verification scores for probabilistic forecasts. The differences in RMSE values from the post-processed pure high-resolution case for semi-local EMOS align closely with the CRPS results.

Similar to the LHPC scenario, all verification scores support the same conclusions. The balanced combinations (6,8) and (7,16) are favoured up to 4 days,

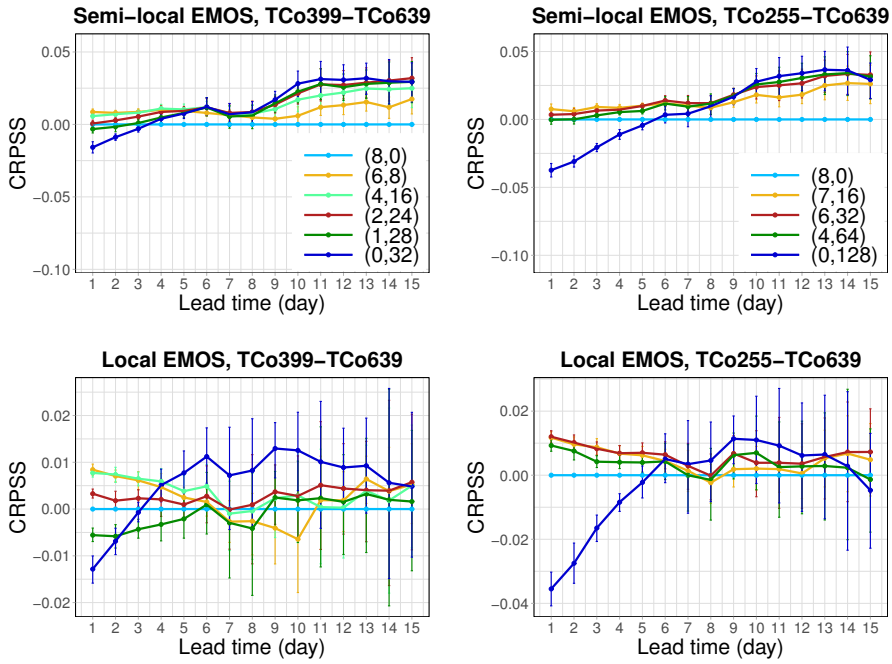

Figure 4.11: CRPSS from the reference pure high-resolution case with 95 % confidence intervals of semi-local (*top*) and local (*bottom*) EMOS post-processed global dual-resolution ensemble forecasts for 2-metre temperature, SHPC scenario.

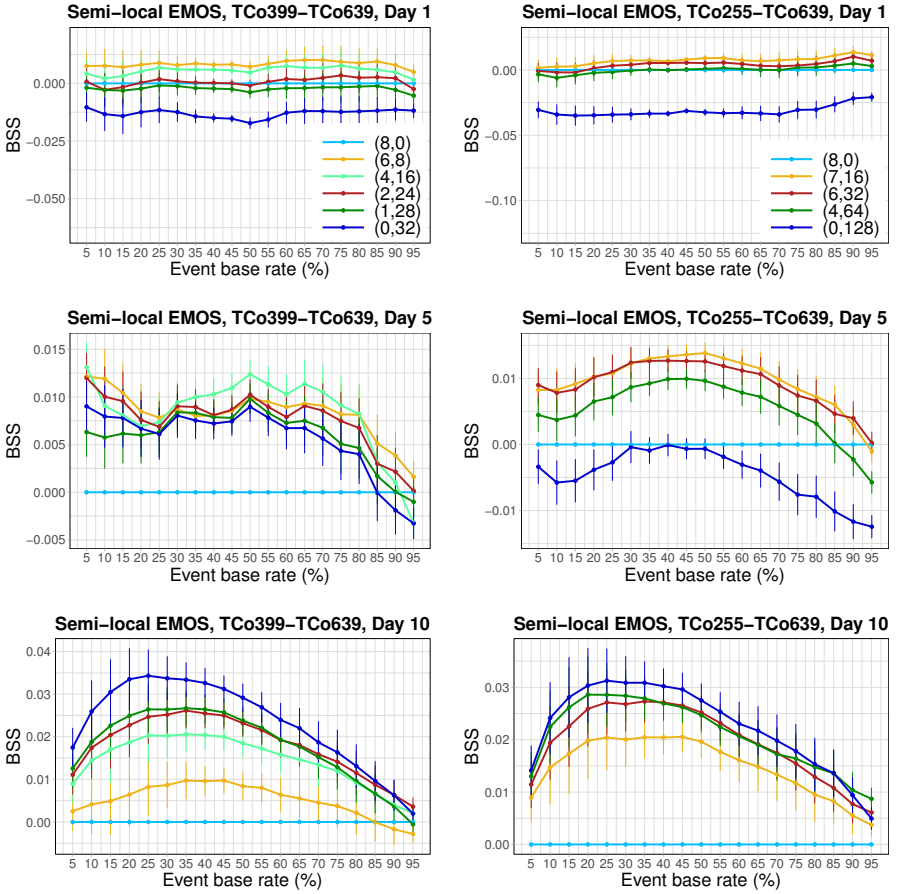

Figure 4.12: Brier skill scores (the higher the better) with respect to the reference pure high-resolution case with 95 % confidence intervals of semi-local EMOS post-processed global dual-resolution ensemble forecasts for 2-metre temperature, SHPC scenario.

although the differences compared to combinations (4,16) and (6,32) are minimal. For TCo399 - TCo639, after day 7, and for TCo255 - TCo639, after day 9, the post-processed pure low-resolution ensemble exhibits the best predictive performance. For longer lead times, the forecast skill is influenced by the ensemble size. This behaviour aligns closely with the patterns observed in the raw ensemble (see Figures 4.9 and 4.11).

4.3.3 Calibration using a very short training period

Additional results were obtained using a 10-day training period to compare the effectiveness of local and semi-local EMOS post-processing. Specifically, the TCo399 - TCo639 ensemble configurations of the LHPC scenario were calibrated for lead times up to 10 days. Local EMOS utilized 10 forecast-observation pairs to estimate 4–5 parameters, while semi-local EMOS employed an average of around 225 forecast-observation pairs within each cluster to estimate the same parameters. The verification was performed using data from the same number of calendar days as in Sections 4.3.1 and 4.3.2, which encompassed ensemble forecasts initialized between 1 July and 31 August 2016 and the corresponding validating observations.

Figure 4.13 displays the verification scores for lead times of 1, 5 and 10 days. Notably, all post-processing approaches for all combinations yielded substantial improvements over the raw ensemble for a shorter lead time. However, this effect diminishes at day 10 for both local and semi-local EMOS with 10 day training period, while the longer, 30 day variation is still performing well. With the 30-day training period, there was no clear preference for local EMOS or semi-local EMOS. In contrast, for the 10-day training period, semi-local EMOS clearly outperformed local EMOS. Therefore, the clustering-based semi-local estimation of EMOS parameters offers a viable alternative to the local approach, especially when the ensemble data cover a relatively short time period.

4.4 Conclusions

EMOS calibration significantly improves the skill of single and dual-resolution ensemble forecasts, reducing the CRPS values from around 1.3 K to just under 1.0 K at day 3 using semi-local EMOS. While the improvements are substantial, they are not as large as those reported by Hemri et al. (2014). Our raw forecasts, which include an orographic correction, have lower CRPS values compared to their uncorrected forecasts. The clustering-based semi-local estimation

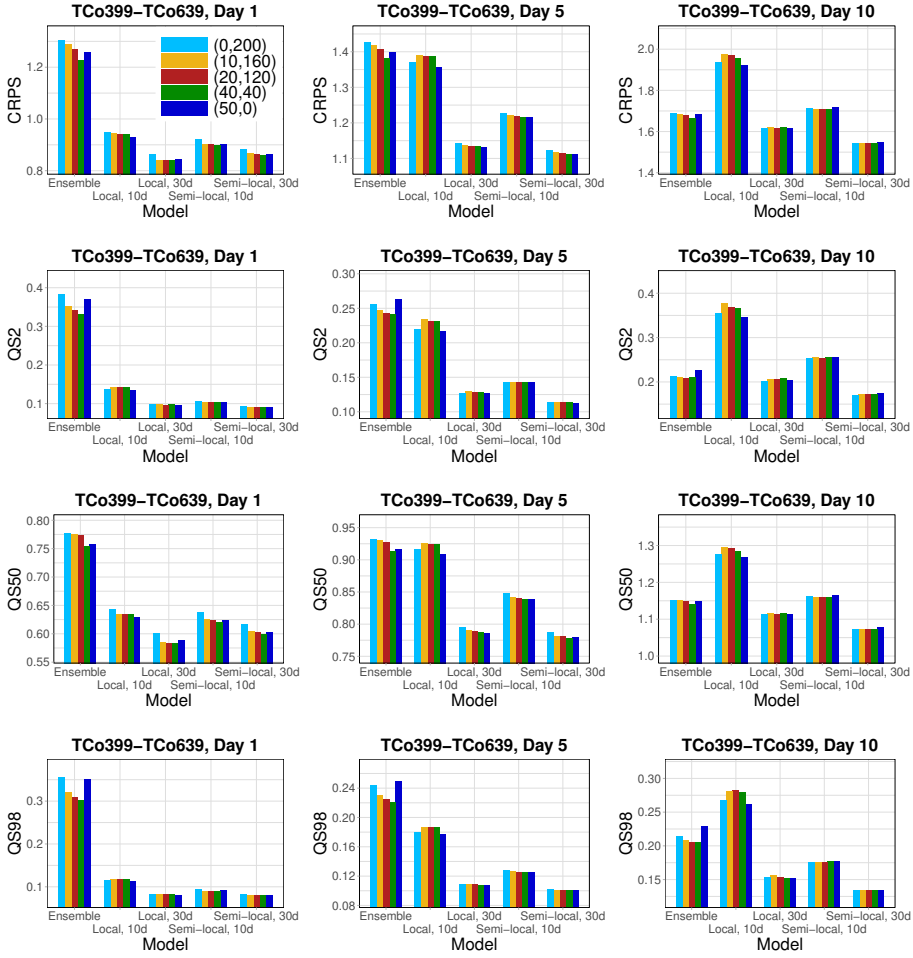

Figure 4.13: Verification scores of 2-metre temperature (K) for the local and semi-local EMOS post-processed forecasts using 10-day and 30-day training periods, TCo399 - TCo639 mixture, LHPC scenario.

of EMOS parameters offers a reasonable alternative to the local approach, particularly when ensemble data cover a short time period, which is consistent with findings by Lerch and Baran (2017) for wind speed forecasts.

EMOS calibration optimizes skill in terms of CRPS and improves the accuracy of point forecasts such as the ensemble mean or median. Brier scores and quantile scores exhibit consistent rankings among calibrated ensemble configurations for different event thresholds and probability levels, respectively.

EMOS calibration reduces differences in skill among equal-cost single and dual-resolution ensemble configurations to a larger extent than the reduction seen from raw to calibrated forecasts. This implies that the selection of the best resolution/ensemble size configuration is less critical for users relying on EMOS-calibrated forecasts rather than raw forecasts.

The optimal single or dual-resolution configuration can change after EMOS calibration. For example, in the large supercomputer scenario, the (40,40) configuration is best for raw forecasts at all lead times but remains superior only until about day 7 after calibration. At longer lead times, configurations with at least 140 members exhibit comparable skills. Similarly, in the small supercomputer scenario, the overall ranking remains similar before and after EMOS calibration. Beyond day 7, skill is mainly determined by the ensemble size, with the pure low-resolution ensemble showing the best performance.

The study by Leutbecher and Ben Bouallègue (2020) identified situations where a dual-resolution ensemble of 2-metre temperature is considerably more skilful than a single-resolution configuration with the same computational cost. After EMOS calibration, the benefit of dual-resolution configurations becomes more marginal compared to before calibration.

The results shown in Section 4.3 demonstrate that statistical post-processing substantially reduces score differences between different single- and dual-resolution configurations. Consequently, the advantages of certain dual-resolution setups over single-resolution configurations are less pronounced when considering post-processed forecasts. Furthermore, the statistical post-processing can alter the ranking of ensemble configurations, highlighting the influence of this calibration technique on the evaluation of optimal dual-resolution approaches. Through a comprehensive analysis of these results, this study contributes to a deeper understanding of the interplay between resolution, ensemble size, and statistical post-processing in the context of dual-resolution ensemble forecasting.

Chapter 5

Calibration of dual-resolution precipitation forecasts

In this chapter, we focus on evaluating the predictive performance of the censored shifted gamma EMOS approach described in Section 2.3.1 for statistical post-processing of dual-resolution precipitation accumulation ensemble forecasts over Europe with various forecast horizons. Our methodology is based on the -at the time of the study- operational 50-member ECMWF ensemble as the high-resolution component which is augmented by a low-resolution (29-km grid) 200-member experimental forecast. The combinations of these two forecast ensembles, with an equal computational cost equivalent to that of the then operational ensemble, form the basis of our investigation.

The primary objective of this case study is to assess the impact of EMOS post-processing on forecast skill, comparing it to raw ensemble combinations. We also aim to determine whether there are statistically significant differences in skill among the various mixtures of dual-resolution combinations. Additionally, we explore the effectiveness of the semi-locally trained CSG EMOS as a powerful alternative to the quantile mapping technique, which typically relies on historical data. For more details on the specifics of the quantile mapping approach see the studies by Hamill and Scheuerer (2018); Gascón et al. (2019) and also Appendix B.

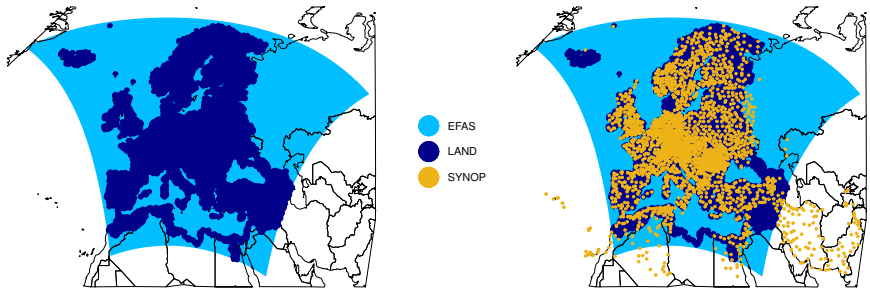

Figure 5.1: (*left*) Map of the domain of the EFAS gridded data and the land subset; (*right*) SYNOP stations in the land subset of the EFAS gridded data.

5.1 Data

In this study we focus on 24-hour precipitation accumulation(from 0600 UTC to the same time of the next day) and our datasets are identical to the ones considered by Gascón et al. (2019). The dual-resolution system consists of different combinations between ensemble forecasts of

- 50 members at TCo639 (grid resolution ∼18 km),
- 200 members at TCo399 (grid resolution ∼29 km).

Note that the cost ratio between these two resolutions is the same as mentioned in section 4.1, so in the different dual-resolution configurations four TCo399 members can be traded against a single TCo639 forecast.

The first dataset consists of 24-hour gridded accumulated precipitation analyses of the European Flood Awareness System (EFAS: Ntegeka et al., 2013) for 1996–2016 covering Europe and some of the surrounding countries (see Figure 5.1 (*left*).

The validation of the post-processing methods under investigation is based on data from 2016, while analyses from the period 1996-2015 are utilized for training the quantile mapping-based techniques. It is important to note that the training process involves all EFAS grid points that correspond to the land subset (363 534 grid points with a 5 km grid spacing). However, for the purposes of validation, only data from 2 370 grid points corresponding to SYNOP stations are taken into account (Figure 5.1 (*right*)).

Post-processing techniques are applied to 24-hour precipitation accumulation forecasts from the ECMWF's Integrated Forecast System for the June-July-August (JJA) period of 2016. The forecasts have lead times ranging from 6 hours to 246 hours, with initialization at 0000 UTC. Following the methodologies employed in previous studies such as Gascón et al. (2019), Baran et al. (2019) and Leutbecher and Ben Bouallègue (2020), we analyze 50 perturbed members of the operational TCo639 ensemble and forecasts from the 200-member TCo399 experiment. The investigated dual-resolution mixtures are the following:

Low resolution	0	40	120	160	200
High resolution	50	40	20	10	0

All of them have the same computational cost corresponding to the available HPC resources of the ECMWF at the moment the forecasts were generated. The mixtures were composed with the first n members of the different ensembles.

To support the training of quantile mapping-based approaches, we utilize 11-member gridded reforecasts for the JJA period from 1996 to 2016. These reforecasts have forecast horizons that match the lead times of the dual-resolution combinations generated at both TCo639 and TCo399 resolutions. For a more comprehensive description of the datasets used in this analysis, we refer interested readers to Gascón et al. (2019) and the relevant references therein.

5.2 Implementation details

In terms of computational costs and model complexity, EMOS is one of the most efficient post-processing approaches (see e.g. Vannitsem et al., 2021, Fig. 1) showing excellent performance for a large variety of weather quantities.

Following the optimum score estimation principle of Gneiting and Raftery (2007), mean parameters $a, b_1, \ldots, b_K$, variance parameters c, d, and shift parameter $\delta > 0$ of the CSG EMOS model specified either by (2.12) or by (2.13) are estimated by optimizing the mean CRPS over an appropriate set of training data. As for both resolutions we consider only forecasts obtained using perturbed initial conditions, ensemble members at a given resolution can be considered as exchangeable. Hence, link functions (2.13) of the CSG EMOS model reduce to

$$\mu = a^2 + b_H^2 \overline{f}_H + b_L^2 \overline{f}_L \qquad \text{and} \qquad \sigma^2 = c^2 + d^2 \overline{f}, \tag{5.1}$$

where $\overline{f}_H$ and $\overline{f}_L$ denote the mean of high- and low-resolution members, respectively. Model parameters are estimated by minimizing the mean CRPS

over the training data where we fix $b_L = 0$ in the pure high-resolution ($M_L = 0$) and $b_H = 0$ in the pure low-resolution ($M_H = 0$) case. The data augmentation technique of Hamill and Scheuerer (2018) is applied in the reference quantile mapping approaches, whereas in EMOS modelling we consider the clustering-based semi-local method of Lerch and Baran (2017).

Due to the high frequency of zero values in both predicted and observed precipitation accumulations, statistical post-processing of this weather variable requires a larger amount of training data compared to variables such as temperature or wind speed. In the case of local EMOS models, Hemri et al. (2014) recommend using almost 5 years (1816 calendar days) of data. However, the dual-resolution forecasts from ECMWF cover a much shorter time period, specifically the 97-day interval from June 1, 2016, to September 5, 2016. Moreover, the heterogeneity and extension of the EFAS domain (as shown in Figure 5.1a) make regional modelling unreliable, warranting the use of a clustering-based semi-local approach. After conducting thorough data analysis and testing various combinations of training period length and number of clusters, we choose to estimate the parameters of the CSG EMOS model using a rolling 30-day training period and 8000 clusters. Similar to the previous case study by Baran et al. (2019), the clustering is performed using 24-dimensional feature vectors. Half of these features are obtained by taking equidistant quantiles of the climatological CDF over the training period, while the other half consists of quantiles of the empirical distribution of the forecast error of the ensemble mean. This configuration provides an average of 1363 forecast-observation pairs for each estimation task, with 5 or 6 parameters to be estimated, and allows for a 52-day verification period (11 July 2016, to 31 August 2016). Remember, that the CSG EMOS post-processed predictions are validated using data from 2370 SYNOP stations (Figure 5.1 (*right*)), enabling a direct comparison with the quantile-mapped (QM) and weighted quantile-mapped (QM+W) forecasts of Gascón et al. (2019). To align with their work, in Section 5.3 we report the various verification scores only for forecast horizons of 1, 3, 5, 7, and 10 days.

5.3 Results

In Figure 5.2, we present the average CRPS for both raw and post-processed forecasts of the dual-resolution combinations under investigation. It is observed that, up to day 5, all post-processed forecast combinations exhibit superior performance compared to the raw dual-resolution ensemble. The largest improvement is observed on day 1, while the smallest gain is observed on day 5.

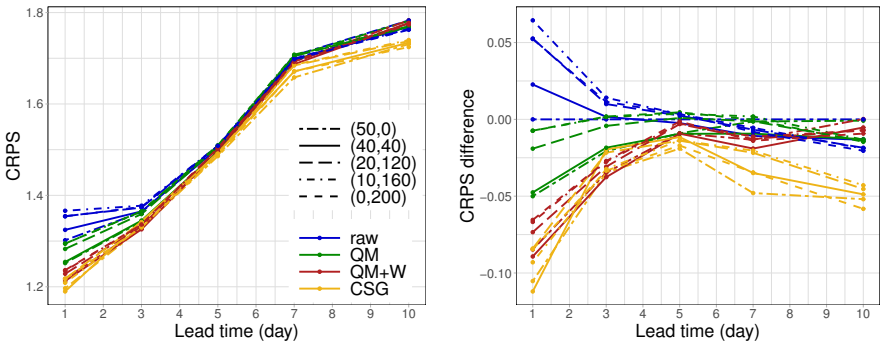

Figure 5.2: (*left*) CRPS of raw and post-processed forecasts; (*right*) difference in CRPS from the raw (50,0) combination as a function of the forecast horizon.

However, for longer lead times, the advantage of QM and QM+W forecasts diminishes, and CSG EMOS models achieve the lowest mean CRPS.

Figure 5.3 provides additional insights into the forecast differences by showing the skill scores of the dual-resolution CSG EMOS models compared to the raw pure high-resolution (50,0) forecast (Figure 5.3 (*left*)) and the corresponding QM+W forecast (Figure 5.3 (*right*)). It is observed that the skill differences among the CSG EMOS models for different dual-resolution combinations are minimal, which aligns with the findings of Baran et al. (2019). These models demonstrate significant improvements over the raw high-resolution forecasts for all lead times except day 5. Furthermore, Figure 5.3 (*right*) shows that the relatively simpler CSG EMOS approach is able to perform on par with the QM+W forecasts for all dual-resolution configurations and lead times.

The analysis of Brier skill scores using thresholds of 0.1 mm, 5 mm, and 10 mm leads to similar conclusions. The CSG EMOS forecasts consistently outperform the ECMWF high-resolution (50,0) precipitation accumulation forecast for all lead times, dual-resolution combinations, and threshold values (Figure 5.4). However, for the 0.1 mm threshold at day 10, the difference is only significant at a 5 % level for the EMOS models based on either the pure high-resolution or pure low-resolution ensemble. Furthermore, Figure 5.5 confirms that there is no dual-resolution combination or forecast horizon where the difference in skill between the matching CSG EMOS and QM+W forecasts is significant at a 5 % level.

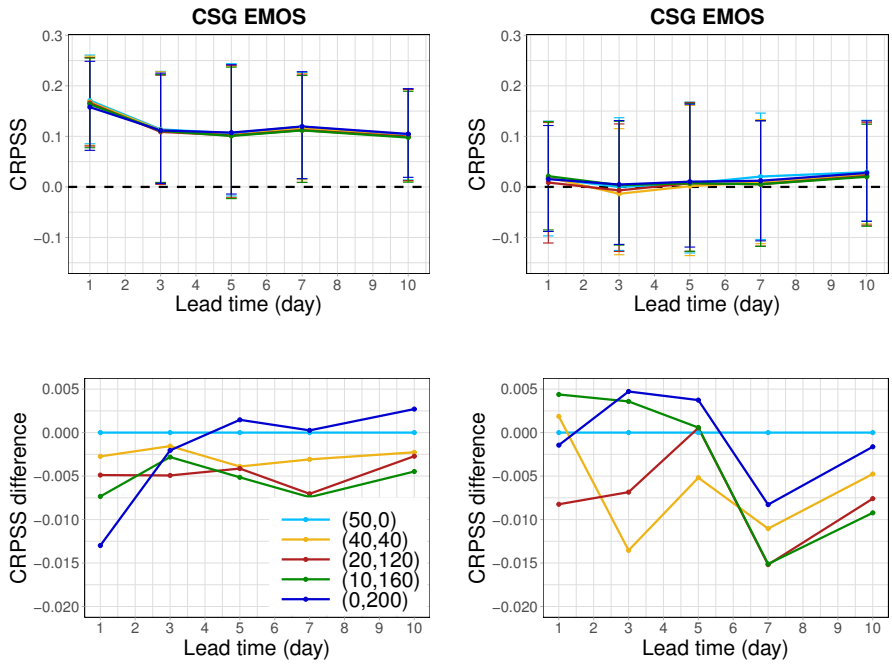

Figure 5.3: CRPSS of the CSG EMOS model for different dual-resolution configurations (*left*) with respect to the raw (50,0) combination; (*right*) with respect to the corresponding QM+W forecast with 95% confidence intervals. The (*bottom*) panels provide the differences in CRPSS from the curves on (*top*), corresponding to the mixture (50,0).

The simultaneous DM tests - using the Benjamini-Hochberg algorithm (Benjamini and Hochberg, 1995) to control the false discovery rate at a 5% level of significance (e.g., Wilks, 2016) - conducted for all considered stations confirm that there are no substantial differences in skill between the CSG EMOS models corresponding to different dual-resolution mixtures. On days 1, 3, 5, and 10, there are practically no stations where the difference in mean CRPS between any pairs of combinations is significant at a 5 % level. Only at day 7, did the mixtures (50,0) and (40,40) show a significant difference in mean CRPS at 1.47 % of the locations. Additionally, up to day 7, these mixtures are the only ones

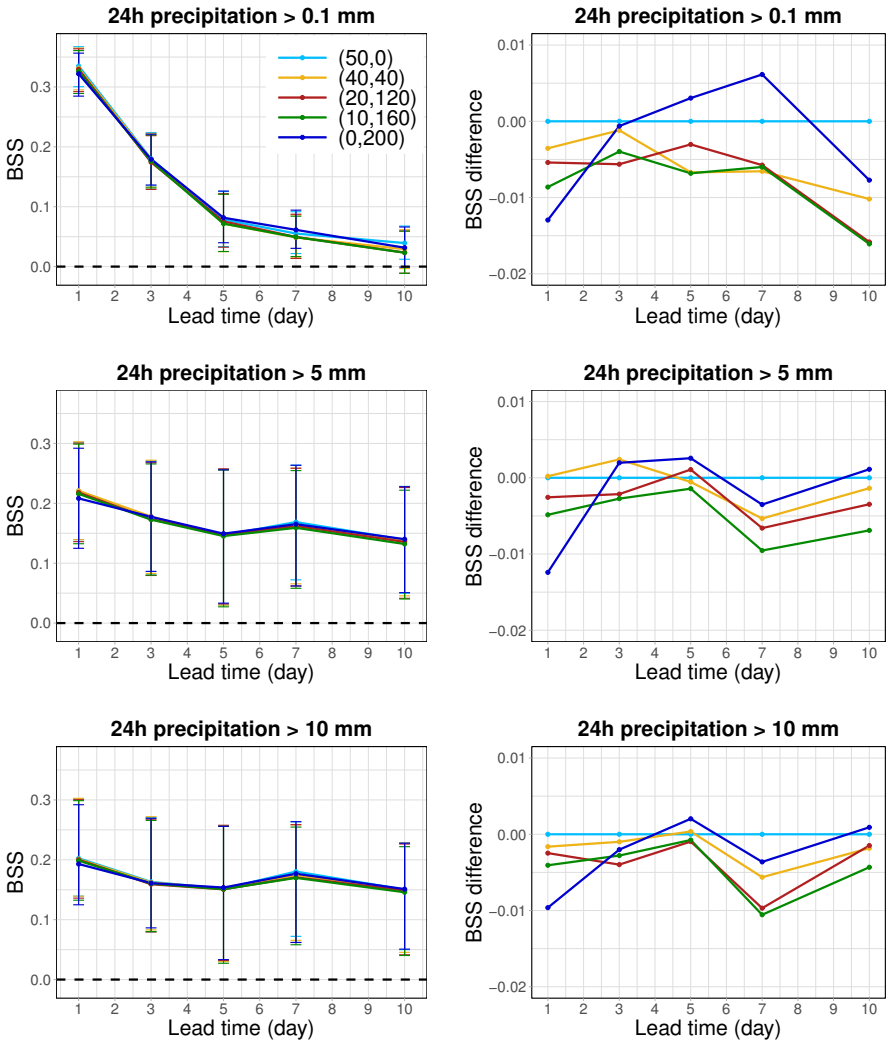

Figure 5.4: BSS of the CSG EMOS model for different dual-resolution configurations with respect to the raw (50,0) configuration with 95% confidence intervals for thresholds (*top*) 0.1 mm; (*middle*) 5 mm; (*bottom*) 10 mm. Panels on the (*right*) provide the differences in BSS from the curves on the (*left*), respectively, corresponding to the mixture (50,0).

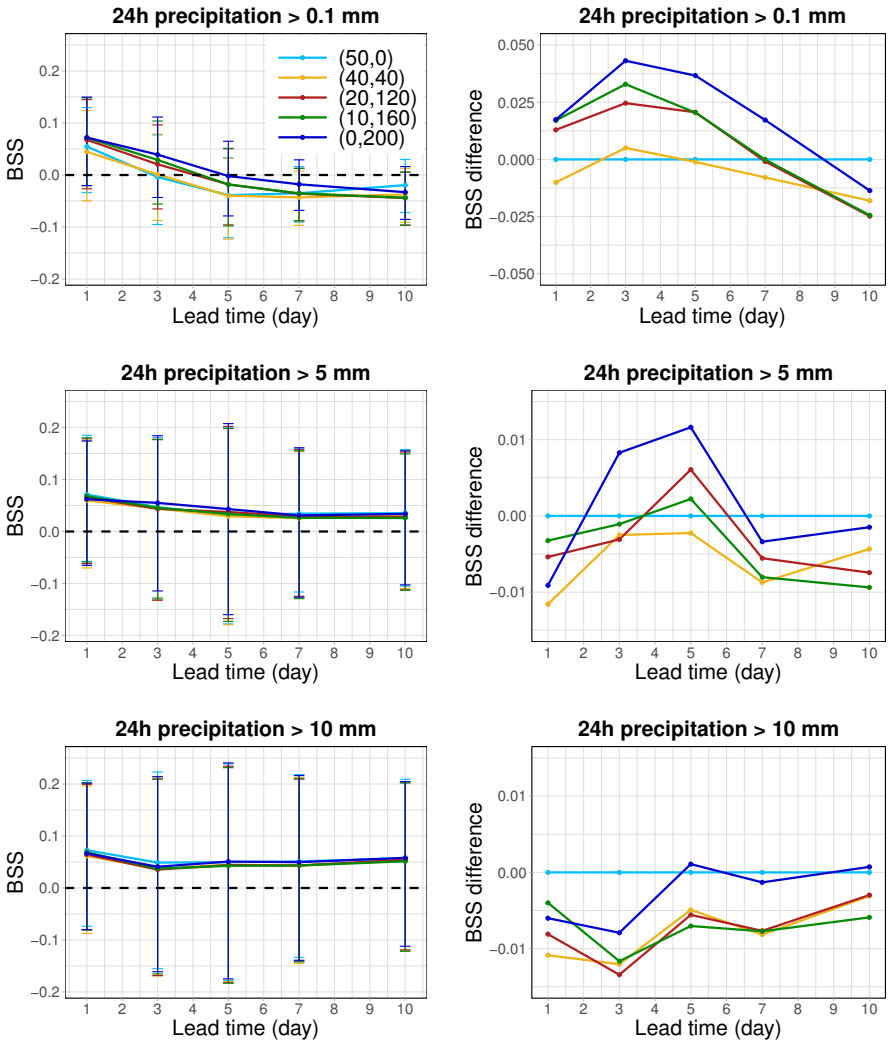

Figure 5.5: Brier Skill scores with respect to the corresponding QM+W forecasts for each dual-resolution combination with 95% confidence intervals for all 3 thresholds. Panels on the (*right*) provide the differences in BSS from the curves on the (*left*), respectively, corresponding to the mixture (50,0).

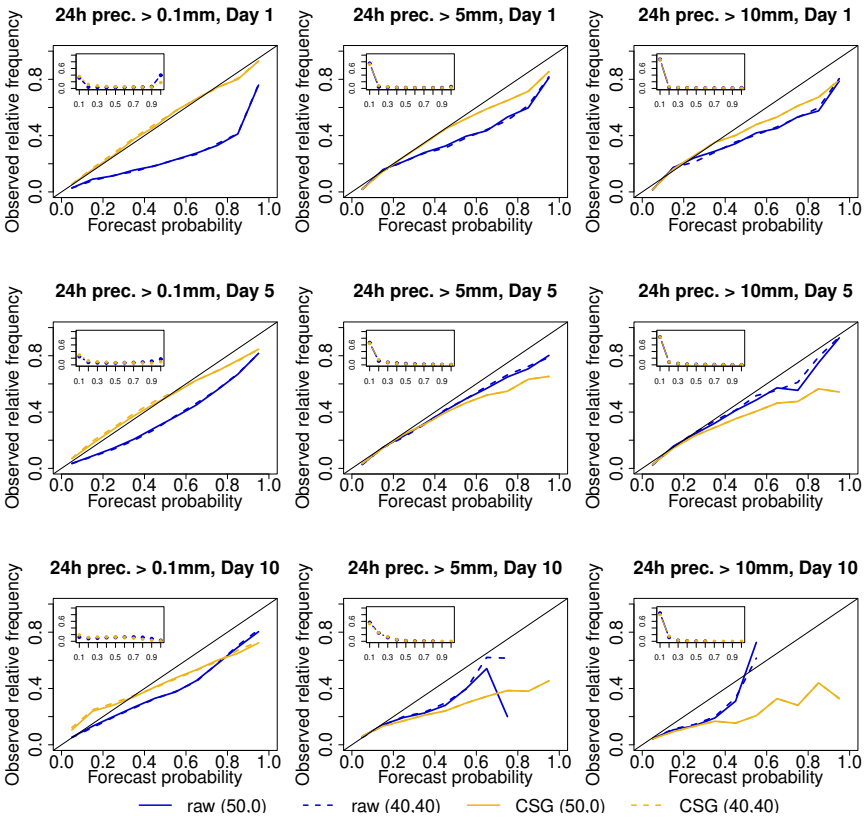

Figure 5.6: Reliability diagrams for 0.1, 5 and 10 mm thresholds of raw (50,0) and (40,40) combinations and corresponding CSG EMOS forecasts for days 1, 5 and 10. The inset curves display the relative frequency of cases within the respective bins for the (50,0) mixture.

that exhibit significantly different mean BS values at some stations for all three thresholds, with proportions ranging from 4.41 % to 6.85 % for 0.1 mm, 3.4 % to 10.72 % for 5 mm, and 2.13 % to 11.68 % for 10 mm. However, on day 10, the situation changes as there are locations with significantly different mean BS values for all pairs of mixtures and thresholds, although the proportions of such locations only slightly exceed 9 %.

Figure 5.6 provides the reliability diagrams for 0.1 mm, 5 mm, and 10 mm thresholds of the raw (50,0) and (40,40) combinations, as well as the corresponding CSG EMOS forecasts for days 1, 5, and 10. At day 1, the CSG EMOS models clearly outperform the raw forecasts, particularly for the 0.1 mm threshold, where the fit to the reference line is nearly perfect. However, for longer lead times, the advantage of post-processing is primarily observed at the lowest threshold, where the reliability diagrams are based on 36.5 % (day 5) and 34.9 % (day 10) of the observations. For the 5 mm and 10 mm thresholds, these proportions decrease to 11.6 % and 10.8 %, and 5.5 % and 6.2 %, respectively. Additionally, the inset histograms illustrate that the distribution of forecast cases is biased, with very low frequencies in the upper bins. This scarcity of data may explain the erratic behaviour of the reliability diagrams for the 5 mm and 10 mm thresholds on day 10.

5.4 Conclusions

We investigate the performance of the censored shifted gamma EMOS approach for statistical post-processing of dual-resolution 24-hour precipitation accumulation ensemble forecasts over Europe. All dual-resolution combinations have equal computational costs and are compared with the raw ensemble combinations. Reference post-processing methods, such as quantile mapping and weighted quantile mapping, are also considered. The calibration methods are trained using forecast-analysis pairs at EFAS grid points and validated using data from SYNOP stations.

The results show that semi-local EMOS post-processing significantly improves forecast skill compared to the raw ensemble combinations for various lead times and thresholds. Among the dual-resolution combinations, there are no significant differences in the skill of CSG EMOS forecasts. Furthermore, CSG EMOS forecasts outperform the reference QM and QM+W predictions in terms of mean CRPS, but the differences are not significant. The same holds for the Brier scores. These findings suggest that the semi-local CSG EMOS method, trained on a 30-day rolling training period, can match the performance of the more complex quantile mapping based on 20 years of historical data.

The introduction of the new ECMWF 48r1 cycle in 2023, which includes dual-resolution forecasts, opens up new research avenues in investigating statistical calibration. Further investigation can explore the skill of machine learning-based parametric post-processing approaches in the context of dual-resolution predictions, focusing on methods that require short training data.

Chapter 6

Summary

In conclusion, this thesis presents a comprehensive exploration of the topic of statistical post-processing of single- and dual-resolution forecasts, analysing the efficacy of parametrised EMOS models based on various distributions and testing them across multiple weather variables. In light of the results from the case studies, this section offers a concise synthesis of the addressed topics and the obtained results and proposes avenues for future research in the field of post-processing of single- and dual-resolution weather forecasts.

In Chapter 3 we introduced a novel approach for calibrating wind speed ensemble forecasts, based on a truncated GEV distribution (TGEV). We addressed the limitations of the otherwise efficient GEV EMOS method proposed by Lerch and Thorarinsdottir (2013), which occasionally predicts negative wind speed. The TGEV EMOS model is tested on short-range (24–48-hour) wind speed forecasts of three completely different EPSs (8-member UWME, 11-member ALADIN–HUNEPS and 50-member ECMWF) covering different and relatively small geographical regions. Furthermore, we compared the EMOS models on a much larger medium-range dataset of global ECMWF forecasts with lead times from 1 to 15 days. To verify the models, we utilised several metrics including the CRPS for assessing the accuracy of probabilistic forecasts, and the MAE and the RMSE for evaluating the median and the mean forecasts, respectively. Additionally, we examined the calibration of the forecasts through the coverage and average width of nominal central prediction intervals, and we assessed the predictive performance at high wind speed values using the twCRPS corresponding to the 90th, 95th, and 98th percentiles of the observed wind speed. The forecast skill of the TGEV EMOS model is compared to that of the TN,

LN, and GEV EMOS approaches, as well as the raw and climatological forecasts. The results of the four case studies demonstrate that post-processing consistently improves the calibration of probabilistic forecasts and the accuracy of point forecasts. Moreover, all EMOS models outperform both the raw ensemble and climatology. Among the four competing methods, the TGEV EMOS approach exhibits the best overall performance, closely followed by the GEV EMOS model. However, it should be noted that the GEV EMOS model occasionally predicts negative wind speed values with a mean probability of approximately 2.5% for the case study of the global ECMWF forecasts in Section 3.3.2 for all considered lead times.

Throughout these case studies, our scope was limited to univariate forecasts for a single location and lead time. However, many practical applications, such as wind energy forecasting (Pinson and Messner, 2018), necessitate accurate modelling of spatial and temporal dependencies. Therefore, an intriguing avenue for future research would involve extending the proposed TGEV EMOS model to encompass multivariate forecasts, enabling the provision of spatially and temporally consistent calibrated wind speed forecasts. For an extensive overview of potential approaches in this area, Lerch et al. (2020); Lakatos et al. (2023) provide valuable insights.

Chapter 4 focuses on the case study of the calibration of dual resolution 2-metre temperature forecasts. With the help of various validation metrics (see Section 2.6) we showed that the EMOS calibration leads to substantial improvements in skill for all examined single- and dual-resolution ensemble forecasts. For example, when employing the semi-local EMOS, we observed a decrease in the CRPS from approximately 1.3K to slightly below 1.0K at day 3. Although the improvements were notable, they were not as substantial as those reported by Hemri et al. (2014). In comparison, our raw ensemble forecasts exhibited significantly smaller CRPS values, which difference can be attributed to the application of an orographic correction to our forecasts. Our raw forecasts were adjusted using a simple correction method that accounts for the altitude disparity between the model's orography and the station's height.

In terms of spatial considerations, the clustering-based semi-local estimation of EMOS parameters provides a reasonable alternative to the local approach, especially in situations where ensemble data cover only a rather short time period. This is fully in line with results reported by Lerch and Baran (2017), where multi-model ensemble forecasts of wind speed over Europe and North Africa were calibrated. The EMOS calibration parameters were obtained by optimizing skill in terms of the CRPS. EMOS demonstrates its effectiveness in enhancing both probabilistic and point forecasts, such as the ensemble mean

and median. Additionally, when comparing calibrated ensemble configurations, we observe consistent rankings in terms of score differences for metrics like the Brier score and quantile score across various event thresholds and probability levels, respectively. EMOS calibration can alter which single- or dual-resolution configuration is optimal. For example, in the large supercomputer scenario, the TCo399-TCo639 (40, 40) configuration is initially the best for raw forecasts at all lead times. After calibration, it remains the best until around day 7, but for longer lead times, configurations with at least 140 members show equal skill. After calibration, the 200 lower-resolution members show slightly higher skill than the 50 higher-resolution members, even if initially, for the raw forecasts, 50 members at TCo639 resolution perform as well as 200 members at TCo399 resolution. Similarly, in the case of the small supercomputer scenario, the overall ranking remains similar before and after EMOS calibration. Beyond day 7, the predictive performance is primarily determined by ensemble size, with the pure low-resolution ensemble exhibiting the best skill. EMOS calibration significantly reduces the skill differences between equal-cost configurations of single- and dual-resolution ensembles. This means that selecting the "best resolution/ensemble size configuration" becomes less crucial for users relying on EMOS-calibrated forecasts instead of raw forecasts. In terms of direct model output, dual-resolution ensemble forecasts for 2-metre temperature show greater skill compared to a single-resolution configuration with the same computational cost. However, the advantage of dual-resolution configurations becomes marginal when EMOS calibration is applied. The question of whether more sophisticated post-processing approaches provide the same answer arises, thus providing a possible direction for further research.

In Chapter 5 we addressed the third case study, which analysed the calibration of dual-resolution precipitation accumulation forecasts. The predictive performance of the censored shifted gamma EMOS approach by Baran and Nemoda (2016) is studied using various dual-resolution 24-hour precipitation accumulation ensemble forecasts across Europe. The computational costs of all dual-resolution combinations are equivalent to that of the then operational 50-member ECMWF ensemble. Reference post-processing methods, such as quantile mapping and weighted quantile mapping of Hamill and Scheuerer (2018), are used for direct comparison. Compared to raw ensemble combinations, semi-local EMOS post-processing significantly improved the mean CRPS and mean BS for different thresholds at all lead times. The mixture of 40 high- and 40 low-resolution forecasts outperforms other combinations until day 5 in the case of the raw ensemble. However, there are no significant skill differences among the various mixtures in CSG EMOS forecasts. CSG EMOS

forecasts outperform QM and QM+W predictions in terms of mean CRPS, but the differences are not significant. The same is true for the Brier scores between CSG EMOS and QM+W. These results suggest that the semi-local CSG EMOS method, trained using a 30-day rolling training period, can achieve similar performance to the more complex quantile mapping based on 20 years of historical data.

Concerning the two case studies of Chapter 4 and 5 it is important to emphasise the differing characteristics of the datasets to have a better understanding of the results in comparison to each other. Pertaining to the datasets at hand, the temperature forecasts were station-based, while the precipitation forecasts has a gridded structure. Further distinction should be made that - due to the nature of the target weather variables - the latter is relatively much harder to model than the other. But, despite all of these differences the key findings of the dual-resolution models persist. On the one hand, the uncalibrated ensemble forecasts show the optimality of the balanced combination, however this significance diminishes as post-processing is applied. On the other hand, both case studies show the increasing importance of the size of the ensemble over the resolution of it. These factors all contribute to the considerations that the operational use of these dual-resolution ensembles needs to address. The decreasing weight of the choice of mixture configurations when calibration is applied poses further need for analysing the current setup in future studies. This highlights the importance of exploring the calibration methods for dual-resolution ensemble forecasts, as well as considering the impact of ensemble size on forecast performance. The introduction of the new ECMWF 48r1 cycle in 2023, with 51 forecasts at TCo1279 resolution and 101 forecasts at TCo319 resolution, opens up new research possibilities for calibrating these predictions.

Additionally, exploring the skill of machine learning-based parametric post-processing approaches in the dual-resolution context, specifically methods that require short training data similar to EMOS, could be a potential direction for further investigation (Baran and Baran, 2021, 2023; Ghazvinian et al., 2022). The recent findings of Höhlein et al. (2023) shed light to the potential that relying only on summary statistics of the ensemble rather than specifically tailoring a model to the ensemble structure can yield just as good results. The question that arises is whether this finding applies to dual-resolution data as well. It could be interesting to test various configurations, utilising separate or overall summary statistics, or including other, more detailed information about the structure of the mixtures. Of course, all of these analysis will require substantially more training data than what was available for the case studies in Chapter 4 and 5.

Chapter 7

Összefoglalás

A disszertáció átfogóan vizsgálja az egy- és kétfelbontású előrejelzések statisztikai utófeldolgozásának témáját, elemezve a különböző eloszlásokon alapuló, parametrizált EMOS modellek hatékonyságát, és tesztelve azokat több időjárási változóra vonatkozóan. Az esettanulmányok eredményeinek fényében ez a szakasz tömören összefoglalja a tárgyalt témákat és a kapott eredményeket, és javaslatokat tesz az egyféle és duális felbontású időjárás-előrejelzések utófeldolgozásának lehetséges jövőbeli kutatási irányaira.

A 3. fejezetben egy újszerű, a csonkolt általánosított extrém érték (TGEV) eloszláson alapuló megközelítést mutattunk be a szélsebesség ensemble előrejelzések kalibrálására. Megvizsgáltuk a Lerch and Thorarinsdottir (2013) által javasolt, egyébként hatékony általánosított extrém érték (GEV) eloszláson alapuló EMOS módszer korlátait, amely esetenként negatív szélsebességet jelez előre. A TGEV EMOS modell hatékonyságát először három rövidtávú (24–48 óra) ensemble predikciós rendeszer szélsebesség előrejelzésein teszteltük (8 tagú UWME, 11 tagú ALADIN–HUNEPS és 50 tagú ECMWF), melyek földrajzilag viszonylag kisebb régiókat fednek le. Ezenkívül összehasonlítottuk az EMOS modelleket az ECMWF középtávú (1-15 nap) ensemble predikciós rendszerének lényegesen nagyobb, globális adathalmazán. A modellek ellenőrzéséhez számos alkalmas metrikát használtunk, köztük a CRPS-t a valószínűségi előrejelzések pontosságának, míg az átlagos abszolút hibát (MAE) és az átlagos négyzetes hiba négyzetgyökét (RMSE) az előrejelzések átlagának és mediánjának a kiértékelésére. Ezenkívül megvizsgáltuk az előrejelzések kalibrálását a nominális központi előrejelzési intervallumok lefedettségén és átlagos szélességén keresztül, illetve megfigyeltük a modellek nagy szélsebesség-

értékeknél nyújtott prediktív teljesítményét a megfigyelt szélsebesség 90., 95. és 98. percentilisének megfelelő twCRPS kiszámításának segítségével. A TGEV EMOS modell előrejelzési képessége összehasonlításra került a TN, LN és GEV EMOS megközelítésekkel, valamint a kalibrálatlan és klimatológiai előrejelzésekkel. A négy esettanulmány eredményei azt mutatják, hogy az utófeldolgozás konzisztensen javítja a valószínűségi előrejelzések kalibráltságát és a kategorikus előrejelzések pontosságát, sőt, minden EMOS modell felülmúlja mind a nyers ensemble előrejelzés, mind a klimatológia eredményeit. A négy versengő módszer közül a TGEV EMOS mutatja a legjobb teljesítményt, amelyet szorosan követ a GEV EMOS modell. Meg kell azonban jegyezni, hogy a GEV EMOS modell alkalmanként negatív szélsebesség-értékeket jelez előre, körülbelül 2,5%-os átlagos valószínűséggel a 3.3.2 szakaszban leírt globális ECMWF előrejelzések esetén az összes figyelembe vett előrejelzési horizontra vonatkozóan.

Az esettanulmányaink olyan egyváltozós előrejelzésekre korlátozódtak, melyek egyetlen helyszínre és előrejelzési horizont adataira vonatkoznak, ezáltal figyelmen kívül hagyva a fennálló térbeli és időbeli összefüggéseket. Számos gyakorlati alkalmazás azonban, mint például a szélenergia előrejelzése (Pinson and Messner, 2018), szükségessé teszi ezen térbeli és időbeli függőségek pontos modellezését. Ezért a jövőbeli kutatásaink során érdekes lehetőség lenne a javasolt TGEV EMOS modell kiterjesztése többváltozós előrejelzésekre, lehetővé téve a térben és időben konzisztens, kalibrált szélsebesség-előrejelzések biztosítását. Lerch et al. (2020); Lakatos et al. (2023) átfogó áttekintést nyújt a lehetséges megközelítésekhez ezen a területen.

A 4. fejezet a duális felbontású 2 méter magasságra készített hőmérséklet előrejelzések kalibrálásának esettanulmányára fókuszál. Különféle validációs metrikák segítségével (lásd a 2.6 részt) megmutattuk, hogy az EMOS-alapú utófeldolgozás lényeges javulást eredményez az összes vizsgált egyféle és duális felbontású ensemble előrejelzéshez. Például a modell paraméterek klaszterezésével kiegészített szemi-lokális EMOS alkalmazásakor megfigyeltük az átlagos CRPS értékek csökkenését körülbelül $1,3K$-ról valamivel $1,0K$ alá a 3. napon. Bár ez a javulás figyelemre méltó, nem olyan jelentős, mint amiről Hemri et al. (2014) számolt be. Ehhez képest a kalibrálatlan ensemble előrejelzések lényegesen kisebb CRPS értékeket mutattak, amely különbség annak tulajdonítható, hogy az előrejelzéseinkre orográfiai korrekciót alkalmaztunk. A nyers előrejelzéseket egy egyszerű korrekciós módszerrel igazítottuk ki, amely figyelembe veszi a modell orográfiája és az állomás magassága közötti magassági eltérést.

A térbeli szempontok szerint a szemi-lokális becslés ésszerű alternatívát

nyújt a lokális megközelítéssel szemben, különösen olyan helyzetekben, amikor az ensemble adatok csak viszonylag rövid időszakot fednek le. Ez teljes mértékben összhangban van a Lerch and Baran (2017) által közölt eredményekkel, ahol az Európa és Észak-Afrika feletti szélsebességre vonatkozó több modellből álló ensemble előrejelzéseket kalibrálták. Az EMOS kalibrációs paramétereit a CRPS metrikára optimalizálással kaptuk. Alátámasztottuk az EMOS hatékonyságát mind a valószínűségi, mind a pontszerű előrejelzésekkel szemben, például az ensemble átlag és a medián javításában. Továbbá, amikor a kalibrált ensemble-konfigurációkat összehasonlítjuk, konzisztens rangsort figyelhetünk meg az olyan metrikák esetén, mint a CRPS, a Brier-score és a quantile score. Az EMOS modell alkalmazása megváltoztathatja, hogy melyik egyféle vagy duális felbontású konfiguráció az optimális. Például a nagy szuperszámítógépes környezet esetén a TCo399-TCo639 (40, 40) konfiguráció kezdetben a legjobb a nyers előrejelzésekhez minden előrejelzési horizonton. A kalibrálás után körülbelül a 7. napig marad ez a kiegyensúlyozott kombináció a legjobb, de távolabbi előrejelzési horizont esetén a legalább 140 tagú konfigurációk azonosan jó képességet mutatnak. A kalibrálás után a 200 alacsonyabb felbontású tag valamivel nagyobb előrejelzési képességet mutat, mint az 50 nagyobb felbontású tag, még akkor is, ha kezdetben a nyers előrejelzések esetében az 50 TCo639 felbontású tag ugyanolyan jól teljesít, mint a 200 TCo399 felbontású tag. Hasonlóképpen, a kis szuperszámítógépes környezet esetében a teljes vizsgálatot átfedő rangsor az EMOS-kalibrálás előtt és után is hasonló marad. A 7. napon túl az előrejelzési teljesítményt elsősorban az ensemble mérete határozza meg, és a legjobb képességet a csupán alacsony felbontású tagokat tartalmazó ensemble mutatja. Az EMOS alkalmazása jelentősen csökkenti az egyféle és duális felbontású ensemble előrejelzések azonos költségű konfigurációi között mérhető különbségeket.

Ez azt jelenti, hogy a "legjobb felbontás/ensemble méret" konfigurációjának kiválasztása kevésbé lesz kulcsfontosságú azon felhasználók számára, akik a nyers előrejelzések helyett az kalibrált előrejelzésekre alapoznak. A modell közvetlen kimenete szempontjából a duális felbontású ensemble előrejelzések a hőmérséklet előrejelzésekre vonatkozóan jobb predikciós képességet mutatnak, mint az azonos számítási költséggel rendelkező egyféle felbontású konfigurációk. A duális felbontású konfigurációk előnye azonban marginális lesz, ha EMOS modellt alkalmaznak az ensemble kalibrálására. Felmerül a kérdés, hogy a kifinomultabb utófeldolgozási megközelítések ugyanezt a választ adják-e, ami a további kutatások lehetséges irányát adja.

A 5. fejezetben a duális felbontású csapadékösszeg előrejelzések kalibrációjának esettanulmányával foglalkoztunk. A Baran and Nemoda (2016)

által alkalmazott cenzorált eltolt gamma (CSG) eloszláson alapuló EMOS megközelítés előrejelzési teljesítményét különböző duális felbontású 24 órás csapadékösszeg ensemble előrejelzések segítségével vizsgáltuk egy Európai adathalmazon. Az összes duális felbontású kombináció számítási költségei megegyeznek az ECMWF korábbi, operatívan használt 50 tagú ensemble költségével. Az összehasonlításhoz referencia utófeldolgozási módszereket is használtunk, úgy mint Hamill and Scheuerer (2018) által vizsgált kvantilis regressziót (QM) és a súlyozott kvantilis regressziót (QM+W). A nyers ensemble kombinációkhoz képest a szemi-lokális EMOS utófeldolgozó módszer jelentősen javította az átlagos CRPS és az átlagos Brier score értékeket a különböző küszöbértékek esetében minden előjelzési horizontra. A 40 magasabb és 40 alacsonyabb felbontású tag keveréke az 5. napig felülmúlja a többi kombinációt a nyers ensemble esetében. A CSG EMOS eredményeit tekintve azonban nincsenek jelentős különbségek a keverékek között. A CSG EMOS kalibrált előrejelzései az átlagos CRPS tekintetében felülmúlják a QM és a QM+W kalibrált előrejelzéseit, de a különbségek nem szignifikánsak. Ugyanez igaz a CSG EMOS és a QM+W közötti Brier-score értékeire is. Ezek az eredmények arra utalnak, hogy a szemi-lokális CSG EMOS módszer, amelyet 30 napos gördülő tanuló periódus segítségével tanítottunk, legalább olyan jó teljesítményt tud elérni, mint a 20 éves historikus adatokon alapuló, összetettebb kvantilis regressziós módszer.

Végezetül, kitekintésként ismertetjük az ECMWF duális felbontású ensemble előrejelző rendszerének aktuális és jövőbeli fejlődési kilátásait. Az új ECMWF 48r1 ciklus 2023-as bevezetése, amely 51 előrejelzést tartalmaz TCo1279 felbontásban és 101 előrejelzést TCo319 felbontásban, új kutatási lehetőségeket nyit a duális felbontású előrejelzések kalibrálására. Mivel a dolgozatban ismertetett vizsgálatok elvégézésekor még nem álltak rendelkezésre ezek a felbontású adatok, így a korábban levont következtetések csak egy sejtést adhatnak a duális kombinációk nyújtotta lehetőségekről. Ezen túlmenően további utófeldolgozó módszerek, mint például a gépi tanuláson alapuló parametrikus utófeldolgozási módszerek vizsgálata szintén egy új irányt adhat a duális felbontású előrejelzések témakörében (Baran and Baran, 2021, 2023; Ghazvinian et al., 2022).

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my PhD supervisor, Dr. Sándor Baran, for his invaluable guidance and mentorship throughout my doctoral journey. I am indebted to him for his constant availability and willingness to provide constructive feedback not just on my dissertation but also on any other project, fostering an environment of collaborative work from the start. Furthermore, I am sincerely thankful to him for his confidence in my research potential as well as his generosity in providing me with opportunities to present my work at conferences and publish in reputable journals, which have been instrumental in creating this dissertation.

I am beyond grateful to my beloved husband, István Lakatos, whose unwavering support of me and my academic goals has been a constant source of strength and encouragement. His belief in my abilities, even during moments of self-doubt, have propelled me forward and given me the confidence to overcome any challenge. His love has been my greatest source of peace on this stressful and demanding journey. His selfless dedication remains an anchor, grounding me in moments of uncertainty and elevating every achievement to a shared triumph.

I would also like to thank my parents for igniting the spark of my interest in science from a very early age and fostering this throughout my school years with love and care. I would like to thank all my friends and family for cheering me on this journey.

I would like to thank all my teachers at the University of Debrecen as well, especially at the Department of Applied Mathematics and Probability Theory, who have guided and helped me through my academic years.

I would also like to extend my appreciation to Martin Leutbecher, Zied Ben-Bouallègue and Estíbaliz Gascón for many helpful discussions, and insightful comments, and for sharing data, and ideas through collaboration.

I also acknowledge that the studies included in this dissertation were made possible with the support of the project EFOP-3.6.3-VEKOP-16-2017-00002, co-

financed by the Hungarian Government and the European Social Fund and the support by the ÚNKP-19-3 New National Excellence Program of The Ministry for Innovation and Technology, as well as the support by the Hungarian National Research, Development and Innovation Office under Grant No. K142849.

Bibliography

Baran, S. and Baran, Á. (2021) Calibration of wind speed ensemble forecasts for power generation. *Időjárás*, **125**, 609–624.

Baran, S. and Baran, Á. (2023) A two-step machine learning approach to statistical post-processing of weather forecasts for power generation. *Q. J. R. Meteorol. Soc.*, doi:10.1002/qj.4635.

Baran, S., Baran, Á., Pappenberger, F. and Ben Bouallègue, Z. (2020) Statistical post-processing of heat index ensemble forecasts: Is there a royal road? *Q. J. R. Meteorol. Soc.*, **146 (732)**, 3416–3434.

Baran, S., Horányi, A. and Nemoda, D. (2014) Comparison of the BMA and EMOS statistical methods in calibrating temperature and wind speed forecast ensembles. *Időjárás*, **118 (3)**, 217–241.

Baran, S. and Lerch, S. (2015) Log-normal distribution based ensemble model output statistics models for probabilistic wind speed forecasting. *Q. J. R. Meteorol. Soc.*, **141**, 2289–2299.

Baran, S. and Lerch, S. (2016) Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, **27**, 116–130.

Baran, S. and Lerch, S. (2018) Combining predictive distributions for the statistical post-processing of ensemble forecasts. *Int. J. Forecast.*, **34**, 477–496.

Baran, S., Leutbecher, M., Szabó, M. and Ben Bouallègue, Z. (2019) Statistical post-processing of dual-resolution ensemble forecasts. *Q. J. R. Meteorol. Soc.*, **145**, 1705–1720.

Baran, S. and Nemoda, D. (2016) Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, **27**, 280–292.

Baran, S., Szokol, P. and Szabó, M. (2021) Truncated generalized extreme value distribution-based ensemble model output statistics model for calibration of wind speed ensemble forecasts. *Environmetrics*, **32 (6)**, e2678.

Ben Bouallègue, Z. (2020) *Accounting for representativeness in the verification of ensemble forecasts*. Technical Memorandum No. 865, ECMWF, Reading, UK, doi:10.21957/5z6esc7wr.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57**, 289–300.

Bentzien, S. and Friederichs, P. (2014) Decomposition and graphical portrayal of the quantile score. *Q. J. R. Meteorol. Soc.*, **140**, 1924–1934.

Bremnes, J. B. (2019) Constrained quantile regression splines for ensemble postprocessing. *Mon. Weather Rev.*, **147**, 1769–1780.

Bremnes, J. B. (2020) Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Mon. Weather Rev.*, **148**, 403–414.

Buizza, R. (1997) Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Weather Rev.*, **125 (1)**, 99–119.

Buizza, R. (2018) Ensemble forecasting and the need for calibration. In Vannitsem, S., Wilks, D. S., Messner, J. W. (Eds.). *Statistical Postprocessing of Ensemble Forecasts*, Elsevier, 15–48.

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y. and Wei, M. (2005) A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Weather Rev.*, **133 (5)**, 1076–1097.

Buizza, R., Milleer, M. and Palmer, T. N. (1999) Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **125 (560)**, 2887–2908.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, **16 (5)**, 1190–1208.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. and Wilby, R. (2004) The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields. *J. Hydrometeorol.*, **5 (1)**, 243–262.

Dawid, A. P. (1984) Present Position and Potential Developments: Some Personal Views Statistical Theory the Prequential Approach. *J. R. Stat. Soc. A*, **147 (2)**, 278–290.

Demaeyer, J., Bhend, J., Lerch, S., Primo, C., Van Schaeybroeck, B., Atencia, A., Ben Bouallègue, Z., Chen, J., Dabernig, M., Evans, G., Faganeli Pucer, J., Hooper, B., Horat, N., Jobst, D., Merše, J., Mlakar, P., Möller, A., Mestre, O., Taillardat, M. and Vannitsem, S. (2023) The EUPPBench postprocessing benchmark dataset v1.0. *Earth Syst. Sci. Data*, **15**, 2635–2653.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P. and Cébron, P. (2015) PEARP, the Météo-France short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **141 (690)**, 1671–1685.

Diebold, F. X. and Mariano, R. S. (1995) Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263.

Eady, E. T. (1951) The quantitative theory of cyclone development. In T. Malone (Ed.). *Compendium of Meteorology*, American Meteorological Society, Boston, MA, 464–469.

ECMWF (2021) ECMWF Strategy 2021-2030. ECMWF, Reading, UK, doi: 10.21957/s21ec694kd.

ECMWF (2023) IFS Documentation CY48R1 - Part V: Ensemble Prediction System. ECMWF, Reading, UK, doi:10.21957/e529074162.

Epstein, E. S. (1969) Stochastic dynamic prediction. *Tellus*, **21 (6)**, 739–759.

Feldmann, K., Richardson, D. S. and Gneiting, T. (2019) Grid-versus station-based postprocessing of ensemble temperature forecasts. *Geophys. Res. Lett.*, **46 (13)**, 7744–7751.

Ferro, C. A. T., Jupp, T. E., Lambert, F. H., Huntingford, C. and Cox, P. M. (2012) Model complexity versus ensemble size: allocating resources for climate prediction. *Philos. Trans. R. Soc. A*, **370 (1962)**, 1087–1099.

Fraley, C., Raftery, A. E. and Gneiting, T. (2010) Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Weather Rev.*, **138 (1)**, 190–202.

Friederichs, P. and Hense, A. (2007) Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Weather Rev.*, **135**, 2365–2378.

Friederichs, P. and Thorarinsdottir, T. L. (2012) Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, **23**, 579–594.

Garcia, A., Torres, J., Prieto, E. and De Francisco, A. (1998) Fitting wind speed distributions: A case study. *Sol. Energy*, **62 (2)**, 139–144.

Gascón, E., Lavers, D., Hamill, T. M., Richardson, D. S., Ben Bouallègue, Z., Leutbecher, M. and Pappenberger, F. (2019) Statistical postprocessing of dual-resolution ensemble precipitation forecasts across Europe. *Q. J. R. Meteorol. Soc.*, **145**, 3218–3235.

Ghazvinian, M., Zhang, Y., Hamill, T. M., Seo, D.-J. and Fernando, N. (2022) Improving probabilistic quantitative precipitation forecasts using short training data through artificial neural networks. *J. Hydrometeor.*, **23**, 1365–1382.

Glahn, H. R. and Lowry, D. A. (1972) The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol. Climatol.*, **11 (8)**, 1203–1211.

Gneiting, T. (2014) *Calibration of medium-range weather forecasts*. Technical Memorandum No. 719. ECMWF, Reading, UK.

Gneiting, T., Lerch, S. and Schulz, B. (2023) Probabilistic solar forecasting: Benchmarks, post-processing, verification. *Sol. Energy*, **252**, 72–80.

Gneiting, T. and Raftery, A. E. (2005) Weather forecasting with ensemble methods. *Science*, **310**, 248–249.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction and estimation. *J. Am. Stat. Assoc.*, **102**, 359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.*, **133**, 1098–1118.

Gneiting, T. and Ranjan, R. (2011) Comparing density forecasts using threshold-and quantile-weighted scoring rules. *J. Bus. Econ. Stat.*, **29 (3)**, 411–422.

Gneiting, T. and Ranjan, R. (2013) Combining predictive distributions. *Electron. J. Stat.*, **7**, 1747–1782.

Goddard, L., Mason, S. J., Zebiak, S. E., Ropelewski, C. F., Basher, R. and Cane, M. A. (2001) Current approaches to seasonal to interannual climate predictions. *Int. J. Climatol.*, **21 (9)**, 1111–1152.

Grell, G. A., Dudhia, J. and Stauffer, D. R. (1995) A description of the fifth-generation Penn state/NCAR mesoscale model (MM5). *Boulder*, NCAR Technical Note, NCAR/TN-398+STR. 122.

Haiden, T., Janousek, M., Bidlot, J., Buizza, R., Ferranti, L., Prates, F. and Vitart, F. (2018) *Evaluation of ECMWF forecasts, including the 2018 upgrade.* ECMWF Technical Memorandum No. 831, doi:10.21957/ldw15ckqi.

Hamill, T. M. (2001) Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.*, **129 (3)**, 550–560.

Hamill, T. M., Engle, E., Myrick, D., Peroutka, M., Finan, C. and Scheuerer, M. (2017) The U.S. national blend of models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Mon. Weather Rev.*, **145 (9)**, 3441–3463.

Hamill, T. M. and Scheuerer, M. (2018) Probabilistic precipitation forecast post-processing using quantile mapping and rank-weighted best-member dressing. *Mon. Weather Rev.*, **146**, 4079—4098.

Harmel, R. D., Richardson, C. W., Hanson, C. L. and Johnson, G. L. (2001) Simulating maximum and minimum daily temperature with the normal distribution. *ASABE Paper No.: 012240.*, ASABE: St. Joseph, Michigan.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K. and Haiden, T. (2014) Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.*, **41**, 9197–9205.

Horányi, A., Kertész, S., Kullmann, L. and Radnóti, G. (2006) The ARPEGE/ALADIN mesoscale numerical modeling system and its application at the Hungarian Meteorological Service. *Időjárás*, **110 (3-4)**, 203–227.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H. and Mitchell, H. L. (1996) A system simulation approach to ensemble prediction. *Mon. Weather Rev.*, **124 (6)**, 1225–1242.

Höhlein, K., Schulz, B., Westermann, R. and Lerch, S. (2023) Postprocessing of Ensemble Weather Forecasts Using Permutation-invariant Neural Networks. *Artif. Intell. Earth Syst.*, doi:10.1175/AIES–D–23–0070.1.

Jewson, S., Brix, A. and Ziehmann, C. (2004) A new parametric model for the assessment and calibration of medium-range ensemble temperature forecasts. *Atmos. Sci. Lett.*, **5 (5)**, 96–102.

Justus, C. G., Hargraves, W. R., Mikhail, A. and Graber, D. (1978) Methods for estimating wind speed frequency distributions. *J. Appl. Meteorol.*, **17 (3)**, 350–353.

Kalnay, E. (2003) *Atmospheric modeling, data assimilation and predictability.* Cambridge University Press.

Lakatos, M., Lerch, S., Hemri, S. and Baran, S. (2023) Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts. *Q. J. R. Meteorol. Soc.*, **149 (752)**, 856–877.

Lang, M. N., Lerch, S., Mayr, G. J., Simon, T., Stauffer, R. and Zeileis, A. (2020) Remember the past: A comparison of time-adaptive training schemes for non-homogeneous regression. *Nonlinear Process. Geophys.*, **27**, 23–34.

Leith, C. E. (1974) Theoretical skill of Monte-Carlo forecasts. *Mon. Weather Rev.*, **102 (6)**, 409–418.

Lerch, S. and Baran, S. (2017) Similarity-based semi-local estimation of EMOS models. *J. R. Stat. Soc. C*, **66**, 29–51.

Lerch, S., Baran, S., Möller, A., Groß, J., Schefzik, R., Hemri, S. and Graeter, M. (2020) Simulation-based comparison of multivariate ensemble post-processing methods. *Nonlinear Process. Geophys.*, **27 (2)**, 349–371.

Lerch, S. and Thorarinsdottir, T. L. (2013) Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A*, **65 (1)**, 21 206.

Leutbecher, M. (2018) Ensemble size: How suboptimal is less than infinity? *Q. J. R. Meteorol. Soc.*, **145**, 107–128.

Leutbecher, M. and Ben Bouallègue, Z. (2020) On the probabilistic skill of dual-resolution ensemble forecasts. *Q. J. R. Meteorol. Soc.*, **146**, 707–723.

Leutbecher, M. and Palmer, T. N. (2008) Ensemble forecasting. *J. Comput. Phys.*, **227 (7)**, 3515–3539.

Li, T. Y. and Yorke, J. A. (1975) Period three implies chaos. *Am. Math. Mon.*, **82 (1)**, 985–992.

Lorenz, E. N. (1963) Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20 (2)**, 130–141.

Lorenz, E. N. (1969) The predictability of a flow which possesses many scales of motion. *Tellus*, **21 (3)**, 289–307.

Ma, J., Zhu, Y., Wobus, R. and Wang, P. (2012) An effective configuration of ensemble size and horizontal resolution for the NCEP GEFS. *Adv. Atmos. Sci.*, **29**, 782–794.

Machete, R. L. and Smith, L. A. (2016) Demonstrating the value of larger ensembles in forecasting physical systems. *Tellus A*, **68 (1)**, 28 393.

Möller, A., Lenkoski, A. and Thorarinsdottir, T. L. (2013) Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Q. J. R. Meteorol. Soc.*, **139 (673)**, 982–991.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. (1996) The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.*, **122 (529)**, 73–119.

Mullen, S. L. and Buizza, R. (2002) The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Weather Forecast.*, **17 (2)**, 173–191.

National Weather Service (1998) Automated Surface Observing System (ASOS) users guide. National Weather Service: Silver Spring, MD.

Ntegeka, V., Salomon, P., Gomes, G., Sint, H., Lorini, V., Zambrano-Bigiarini, M. and Thielen, J. (2013) *EFAS-Meteo: a European daily high-resolution gridded meteorological data set for 1990—2011.*, EU, Ispra: Joint Research Centre, Technical Report JRC86388.

Palmer, T. N., Doblas-Reyes, F. J., Hagedorn, R. and Weisheimer, A. (2005) Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philos. Trans. R. Soc. B*, **360 (1463)**, 1991–1998.

Palmer, T. N., Döring, A. and Seregin, G. (2014) The real butterfly effect. *Nonlinearity*, **27 (9)**, R123.

Palmer, T. N., Molteni, F., Mureau, R., Buizza, R., Chapelet, P. and Tribbia, J. (1993) Ensemble prediction. *Proc. ECMWF Seminar on Validation of models over Europe*, European Centre for Medium-Range Weather Forecasts Reading, United Kingdom, Vol. 1, 21–66.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Delécluse, P., Déqué, M., Diez, E., Doblas-Reyes, F. J., Feddersen, H. et al. (2004) Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Am. Meteorol. Soc.*, **85 (6)**, 853–872.

Pinson, P. and Messner, J. W. (2018) Application of postprocessing for renewable energy. In Vannitsem, S., Wilks, D. S., Messner, J. W. (Eds.). *Statistical Postprocessing of Ensemble Forecasts*, Elsevier, 241–266.

Politis, D. N. and Romano, J. P. (1994) The stationary bootstrap. *J. Am. Stat. Assoc.*, **89**, 1303–1313.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007) *Numerical recipes in C++: The art of scientific computing (3rd ed.)*. Cambridge university press.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.*, **133**, 1155–1174.

Rasp, S. and Lerch, S. (2018) Neural networks for postprocessing ensemble weather forecasts. *Mon. Weather Rev.*, **146**, 3885–3900.

Raynaud, L. and Bouttier, F. (2017) The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Q. J. R. Meteorol. Soc.*, **143 (11)**, 3037–3047.

Richardson, D. S. (2001) Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteorol. Soc.*, **127 (577)**, 2473–2489.

Schefzik, R., Thorarinsdottir, T. L. and Gneiting, T. (2013) Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28 (4)**, 616–640.

Scheuerer, M. (2014) Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Q. J. R. Meteorol. Soc.*, **140**, 1086–1096.

Scheuerer, M. and Hamill, T. M. (2015) Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Weather Rev.*, **143**, 4578–4596.

Scheuerer, M. and Möller, D. (2015) Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Ann. Appl. Stat*, **9 (3)**, 1328–1349.

Scheuerer, M., Switanek, M. B., Worsnop, R. P. and Hamill, T. M. (2020) Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Weather Rev.*, **148**, 3489–3506.

Schulz, B. and Lerch, S. (2022) Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Mon. Weather Rev.*, **150 (1)**, 235–257.

Sloughter, J. M., Gneiting, T. and Raftery, A. E. (2010) Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Am. Stat. Assoc.*, **105**, 25–37.

Stensrud, D. J. (2001) Using short-range ensemble forecasts for predicting severe weather events. *Atmos. res.*, **56 (1-4)**, 3–17.

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S. and Rogers, E. (1999) Using ensembles for short-range forecasting. *Mon. Weather Rev.*, **127 (4)**, 433–446.

Szabó, M., Gascón, E. and Baran, S. (2023) Parametric post-processing of dual-resolution precipitation forecasts. *Weather Forecast.*, **38 (8)**, 1313–1322.

Taillardat, M., Mestre, O., Zamo, M. and Naveau, P. (2016) Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Weather Rev.*, **144 (6)**, 2375–2393.

Thorarinsdottir, T. L. and Gneiting, T. (2010) Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. R. Stat. Soc. A*, **173**, 371–388.

Toth, Z. and Kalnay, E. (1993) Ensemble forecasting at NMC: The generation of perturbations. *Bull. Am. Meteorol. Soc.*, **74 (12)**, 2317–2330.

Toth, Z. and Kalnay, E. (1997) Ensemble forecasting at NCEP and the breeding method. *Mon. Weather Rev.*, **125 (12)**, 3297–3319.

Toth, Z., Zhu, Y. and Marchok, T. (2001) The use of ensembles to identify forecasts with small and large uncertainty. *Weather Forecast.*, **16 (4)**, 463–477.

Valdivia-Bautista, S. M., Domínguez-Navarro, J. A., Pérez-Cisneros, M., Vega-Gómez, C. J. and Castillo-Téllez, B. (2023) Artificial intelligence in wind speed forecasting: A review. *Energies*, **16 (5)**, 2457.

Vannitsem, S., Wilks, D. S. and Messner, J. (2018) *Statistical postprocessing of ensemble forecasts*. Elsevier, Amsterdam.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Ben Boualègue, Z., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Odak Plenkovič, I., Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K. and Ylhaisi, J. (2021) Statistical postprocessing for weather forecasts – review, challenges and avenues in a big data world. *Bull. Am. Meteorol. Soc.*, **102**, E681–E699.

Wang, X. and Bishop, C. H. (2005) Improvement of ensemble reliability with a new dressing kernel. *Q. J. R. Meteorol. Soc.*, **131 (607)**, 965–986.

Wilks, D. S. (2016) "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Amer. Meteor. Soc.*, **97**, 2263–2273.

Wilks, D. S. (2018) Univariate ensemble forecasting. In Vannitsem, S., Wilks, D. S., Messner, J. W. (Eds.). *Statistical Postprocessing of Ensemble Forecasts*, Elsevier, 49–89.

Wilks, D. S. (2019) *Statistical Methods in the Atmospheric Sciences. 4th ed.* Elsevier, Amsterdam.

Appendix A

A.1 Mean of a TGEV distribution

To simplify the formulation of the results, similar to the notations of Section 2.2.4, in what follows we set aside the indication of the parameters of the GEV and TGEV CDFs G and G_0 defined by (2.7) and (2.9), respectively.

The present section is devoted to the verification of the formula (2.10) for the TGEV mean. Let $\xi < 1$ and $G(0) < 1$. The PDF $g_0(x)$ of a $\mathcal{TGEV}(\mu, \sigma, \xi)$ distribution defined by (2.9) equals

$$g_0(x) = \begin{cases} \frac{\left[1+\xi(\frac{x-\mu}{\sigma})\right]^{-1/\xi-1} \exp\left(-\left[1+\xi(\frac{x-\mu}{\sigma})\right]^{-1/\xi}\right)}{\sigma(1-G(0))}, & \text{if } \xi \neq 0; \\ \frac{\exp\left(\frac{x-\mu}{\sigma}\right)\exp\left(-\exp\left[-\frac{x-\mu}{\sigma}\right]\right)}{\sigma(1-G(0))}, & \text{if } \xi = 0, \end{cases} \tag{A.1}$$

for $x \geq 0$ and $x\xi \geq \mu\xi - \sigma$, and $g_0(x) = 0$ otherwise, where

$$G(0) = \begin{cases} \exp(-[1-\xi\mu/\sigma]^{-1/\xi}), & \text{if } \xi \neq 0, \\ \exp(-\exp[\mu/\sigma]), & \text{if } \xi = 0. \end{cases}$$

Let X be a TGEV random variable and assume first $1 > \xi > 0$ and $\xi\mu - \sigma > 0$. Then

$$\mathsf{E}X = \frac{1}{\sigma(1-G(0))} \int_{\mu-\sigma/\xi}^{\infty} x \left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi-1} \times \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right) \mathrm{d}x. \tag{A.2}$$

After setting $t = \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}$ and applying the change of variables, a short straightforward calculation shows that for $\xi > 0$ and $\xi\mu - \sigma > 0$ one has

$$\mathsf{E}X = \frac{1}{(1 - G(0))} \int_0^\infty \left[\frac{(t^{-\xi} - 1)\sigma}{\xi} + \mu\right] \exp(-t)\mathrm{d}t = \frac{\mu + \sigma(\Gamma(1-\xi) - 1)/\xi}{1 - \exp(-[1 - \xi\mu/\sigma]^{-1/\xi})}.$$

Now, let $\xi \neq 0$ and $\xi\mu - \sigma \leq 0$. If $\xi > 0$, then the support of $g_0(x)$ is $[0, \infty[$, so the integral in (A.2) should be taken over this particular interval. For $\xi < 0$ the support of $g_0(x)$ changes to $[0, \mu - \sigma/\xi]$; however, in both cases the change of integral leads to

$$\begin{aligned}\mathsf{E}X &= \frac{1}{1-G(0)} \int_0^{\left(1-\frac{\xi\mu}{\sigma}\right)^{-1/\xi}} \left[\frac{(t^{-\xi} - 1)\sigma}{\xi} + \mu\right] \exp(-t)\mathrm{d}t \\ &= \mu - \frac{\sigma}{\xi} + \frac{\sigma(\Gamma_\ell(1-\xi, [1 - \xi\mu/\sigma]^{-1/\xi}))/\xi}{1 - \exp(-[1 - \xi\mu/\sigma]^{-1/\xi})}.\end{aligned}$$

Finally, let $\xi = 0$. In this case

$$\mathsf{E}X = \frac{1}{\sigma(1 - G(0))} \int_0^\infty x \exp\left(\frac{x-\mu}{\sigma}\right) \exp\left(-\exp\left[-\frac{x-\mu}{\sigma}\right]\right) \mathrm{d}x,$$

where the change of variables with respect to $t = \exp\left(-\frac{x-\mu}{\sigma}\right)$ results in

$$\mathsf{E}X = \frac{1}{\sigma(1 - G(0))} \int_0^{\exp(\mu/\sigma)} (\mu - \sigma \ln t) \exp(-t)\mathrm{d}t = \frac{\mu + \sigma(C - \mathrm{Ei}(-\exp[\mu/\sigma]))}{1 - \exp(-\exp[\mu/\sigma])}.$$

□

A.2 CRPS of a TGEV distribution

Following the ideas of Friederichs and Thorarinsdottir (2012), the CRPS of a TGEV distribution is derived using representation

$$\mathrm{CRPS}(G_0, x) = x\big(2G_0(x) - 1\big) - 2\int_0^1 tG_0^{-1}(t)\mathrm{d}t + 2\int_{G_0(x)}^1 G_0^{-1}(t)\mathrm{d}t, \quad \text{(A.3)}$$

where G_0^{-1} denotes the quantile function corresponding to G_0. Short calculation shows that for $0 < y < 1$

$$G_0^{-1}(y) = \begin{cases} \mu + \frac{\sigma}{\xi}\Big(-1 + \big[-\ln\tau(y)\big]^{-\xi}\Big), & \text{if } \xi \neq 0, \\ \mu - \sigma\Big(\ln\big[-\ln\tau(y)\big]\Big), & \text{if } \xi = 0, \end{cases}$$

where $\tau(y) := \big(1 - G(0)\big)y + G(0)$.

First, let us assume $\xi \neq 0$. Then the first integral of (A.3) equals

$$\begin{aligned} 2\int_0^1 tG_0^{-1}(t)\mathrm{d}t &= \mu - \frac{\sigma}{\xi} + \frac{2\sigma}{\xi}\int_0^1 t\big[-\ln\tau(t)\big]^{-\xi}\mathrm{d}t \\ &= \mu - \frac{\sigma}{\xi} + \frac{2\sigma}{\xi}\int_{G(0)}^1 \frac{\tau - G(0)}{(1-G(0))^2}[-\ln\tau]^{-\xi}\mathrm{d}\tau \\ &= \mu - \frac{\sigma}{\xi} + \frac{2\sigma}{\xi}\frac{1}{(1-G(0))^2} \\ &\qquad \times \left[\int_{G(0)}^1 \tau[-\ln\tau]^{-\xi}\mathrm{d}\tau - G(0)\int_{G(0)}^1 [-\ln\tau]^{-\xi}\mathrm{d}\tau\right]. \end{aligned}$$

Now, let Γ_u denote the upper incomplete gamma functions, defined as

$$\Gamma_u(a, x) = \int_x^\infty t^{a-1}\mathrm{e}^{-t}\mathrm{d}t.$$

Using $\Gamma(a) = \Gamma_\ell(a, x) + \Gamma_u(a, x)$, short calculations involving appropriate changes of variables show

$$\begin{aligned} \int_{G(0)}^1 \tau[-\ln\tau]^{-\xi}\mathrm{d}\tau &= 2^{\xi-1}\Big[\Gamma(1-\xi) - \Gamma_u\big(1-\xi, -2\ln G(0)\big)\Big] \\ &= 2^{\xi-1}\Gamma_\ell\big(1-\xi, -2\ln G(0)\big), \\ \int_{G(0)}^1 [-\ln\tau]^{-\xi}\mathrm{d}\tau &= \Gamma(1-\xi) - \Gamma_u\big(1-\xi, -\ln G(0)\big) = \Gamma_\ell\big(1-\xi, -\ln G(0)\big). \end{aligned}$$

Hence,

$$2\int_0^1 tG_0^{-1}(t)\mathrm{d}t = \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi(1-G(0))^2}\Big[2^{\xi}\Gamma_\ell\big(1-\xi,-2\ln G(0)\big) - G(0)\Gamma_\ell\big(1-\xi,-\ln G(0)\big)\Big]. \tag{A.4}$$

The second integral of (A.3) can be evaluated in a similar way, resulting in

$$\int_{G_0(x)}^1 G_0^{-1}(t)\mathrm{d}t = \big(1-G_0(x)\big)\Big(\mu-\frac{\sigma}{\xi}\Big) + \frac{\sigma}{\xi(1-G(0))}\Gamma_\ell\big(1-\xi,-\ln G(x)\big)\big). \tag{A.5}$$

Finally, the combination of equations (A.3), (A.4) and (A.5) gives

$$\begin{aligned}\mathrm{CRPS}(G_0,x) =& \big(2G_0(x)\ -1\big)\Big(x-\mu+\frac{\sigma}{\xi}\Big) + \frac{\sigma}{\xi(1-G(0))^2}\\ &\times\Big[-2^{\xi}\Gamma_\ell\big(1-\xi,-2\ln G(0)\big) + 2G(0)\Gamma_\ell\big(1-\xi,-\ln G(0)\big)\\ &+2\big(1-G(0)\big)\Gamma_\ell\big(1-\xi,-\ln G(x)\big)\Big].\end{aligned}$$

Now, let $\xi = 0$. In this case for the integrals in (A.3) we have

$$\begin{aligned}2\int_0^1 tG_0^{-1}(t)\mathrm{d}t &= \mu - 2\sigma\int_0^1 t\ln\big[-\ln\tau(t)\big]\mathrm{d}t\\ &= \mu - 2\sigma\int_{G(0)}^1 \frac{\tau-G(0)}{(1-G(0))^2}\ln[-\ln\tau]\mathrm{d}\tau\\ &= \mu - \frac{2\sigma}{(1-G(0))^2}\left[\int_{G(0)}^1 \tau\ln[-\ln\tau]\mathrm{d}\tau - G(0)\int_{G(0)}^1 \ln[-\ln\tau]\mathrm{d}\tau\right],\\ \int_{G_0(x)}^1 G_0^{-1}(t)\mathrm{d}t &= \mu\big(1-G_0(x)\big) - \sigma\int_{G_0(x)}^1 \ln\big[-\ln\tau(t)\big]\mathrm{d}t\\ &= \mu\big(1-G_0(x)\big) - \frac{\sigma}{1-G(0)}\int_{G(x)}^1 \ln\big[-\ln\tau\big]\mathrm{d}\tau.\end{aligned}$$

Hence, keeping in mind that

$$\int\tau\ln\big[-\ln\tau\big]\mathrm{d}\tau = \frac{\tau^2}{2}\ln\big[-\ln\tau\big] - \frac{1}{2}\mathrm{Ei}(2\ln\tau)\big] \quad \text{and}$$

$$\int\ln\big[-\ln\tau\big]\mathrm{d}\tau = \tau\ln\big[-\ln\tau\big] - \mathrm{Ei}(\ln\tau)\big],$$

we obtain

$$\begin{aligned}\mathrm{CRPS}(G_0,x) = {} & x(2G_0(x)-1)+\mu-2\mu G_0(x) \\ & + \frac{2\sigma}{(1-G(0))^2}\left\{\left[\frac{s^2}{2}\ln\big[-\ln s\big]-\frac{1}{2}\mathrm{Ei}(2\ln s)\right]_{s=G(0)}^{s=1}\right. \\ & - G(0)\Big[\big(s\ln\big[-\ln s\big]-\mathrm{Ei}(\ln s)\big]_{s=G(0)}^{s=1} \\ & \left. - \big(1-G(0)\big)\Big[s\ln\big[-\ln s\big]-\mathrm{Ei}(\ln s)\Big]_{s=G(x)}^{s=1}\right\}.\end{aligned}$$

Finally, since

$$\begin{aligned}& s^2\ln\big[-\ln s\big] - \mathrm{Ei}(2\ln s) - 2G(0)\big(s\ln\big[-\ln s\big]-\mathrm{Ei}(\ln s)\big) \\ & \qquad -2\big(1-G(0)\big)\big(s\ln\big[-\ln s\big]-\mathrm{Ei}(\ln s)\big) \\ & = s^2\ln\big[-\ln s\big] - 2s\ln\big[-\ln s\big] - \mathrm{Ei}(2\ln s) + 2\mathrm{Ei}(\ln s) \\ & = C - \ln 2 + (s-1)^2\ln\big[-\ln s\big] + \sum_{k=1}^{\infty}\frac{-(2\ln s)^k + 2(\ln s)^k}{k!k} \to C-\ln 2 \quad \text{as} \quad s\uparrow 1,\end{aligned}$$

the CRPS of a TGEV distribution with $\xi = 0$ equals

$$\begin{aligned}\mathrm{CRPS}(G_0,x) = {} & (x-\mu)\big(2G_0(x)-1\big) + \frac{\sigma}{(1-G(0))^2} \\ & \times\Big(C - \ln 2 + \mathrm{Ei}\big(2\ln G(0)\big) - (G(0))^2\ln\big[-\ln G(0)\big] \\ & - 2G(0)\mathrm{Ei}\big(\ln G(0)\big)\Big) \\ & + \frac{2\sigma}{1-G(0)}\Big[G(x)\ln\big[-\ln G(x)\big] - \mathrm{Ei}\big(\ln G(x)\big)\Big].\end{aligned}$$

□

Appendix B

Quantile mapping

Quantile mapping, detailed by Hamill et al. (2017) is a nonparametric statistical post-processing method used to adjust an uncalibrated forecast to match the distribution of the corresponding observations. This approach relies on climatological CDFs of forecasts and observations to make the adjustments. In the case of an ensemble forecast, quantile mapping is applied to each member separately.

Let $F_f(y)$ and $F_o(y)$ denote the climatological CDF of forecasts and observations, respectively, representing the probability that a random variable (e.g., precipitation amount) is less than or equal to a particular threshold value y.

Furthermore, let $F_o^{-1}(p)$, where $p \in [0, 1]$, refer to the quantile function as the inverse distribution function, which maps a cumulative probability p to a threshold value y. And thus, the adjusted forecast $\widetilde{f}$ can be given by

$$\widetilde{f} \coloneqq F_o^{-1}\big(F_f(f)\big).$$

In the case study discussed in Chapter 5, $F_f(y)$ and $F_o(y)$ are estimated from historical data of 19 years using the control of the 11-member ECMWF reforecasts and the EFAS analysis, respectively. For each grid point and each date of the verification period, climatological CDFs are developed from 9000 sample values, that are derived from 20 years $\times$ 1 member $\times$ 9 closest dates to the given Julian date $\times$ 50 supplemental similar locations chosen according to suggestions of Hamill et al. (2017).

In the case of the weighted quantile mapping approach, defined by Hamill and Scheuerer (2018), first, quantile mapping was applied to each member of

the ensemble separately. For a given calendar year, climatological CDFs are calculated in the same way as before utilizing the matching reforecasts and corresponding analyses of 9 neighbouring dates from the remaining 19 years for 50 similar supplemental locations. Second, the 11-member quantile-mapped reforecasts for the 19 years are then applied to derive the 11-bin closest-member histograms, which are histograms of ranks of the adjusted reforecast members closest to the analysed precipitation amount, for various quantile-mapped ensemble mean values. Third, weights can be produced for the operational TCo639 or experimental TCo399 ensemble forecasts (controls included, resulting in 51 and 201 bins, respectively) by fitting a beta distribution to an 11-bin nearest member histogram. For even more details we refer to Gascón et al. (2019) and Hamill and Scheuerer (2018).

Appendix C

List of publications and conferences

Journal papers

Baran, S., Leutbecher, M., **Szabó, M.** and Ben Bouallègue, Z. (2019) Statistical post-processing of dual-resolution ensemble forecasts. *Q. J. R. Meteorol. Soc.*, **145**, 1705–1720. doi:10.1002/qj.3521

(**IF: 3.471, SJR: D1**) [Ind. cit.: 1]

Baran, S., Szokol, P. and **Szabó, M.** (2021) Truncated generalized extreme value distribution-based ensemble model output statistics model for calibration of wind speed ensemble forecasts. *Environmetrics*, **32 (6)**, e2678. doi:10.1002/env.2678

(**IF: 1.527, SJR: Q1**) [Ind. cit.: 7]

Szabó, M., Gascón, E. and Baran, S. (2023) Parametric post-processing of dual-resolution precipitation forecasts. *Weather Forecast.*, **38(8)**, 1313–1322. doi:10.1175/WAF-D-23-0003.1.

(**IF: 2.9, SJR: Q1**) [Ind. cit.: 0]

Unrelated journal papers

Bogacsovics, G., Hajdu, A., Harangi, B., Lakatos, I., Lakatos, R., **Szabó, M.**, Tiba, A. and Tóth, J. (2021) Napelemfarmok Magyarország területén történő elhelyezését segítő döntéstámogató rendszer fejlesztése. *KözigazgatásTudomány*, **1(2)**, 134–145. doi:10.54200/kt.v1i2.23

Bogacsovics, G., Hajdu, A., Harangi, B., Lakatos, I., Lakatos, R., **Szabó, M.**, Tiba, A., Tóth, J. and Tarcsi, Á. (2021) Adatelemzési folyamat és keretrendszer a közigazgatás számára. *KözigazgatásTudomány*, **1(2)**, 146–158. doi:10.54200/kt.v1i2.24

International conferences

Baran, S., Leutbecher, M., **Szabó, M.** and Ben Bouallègue, Z. Parametric post-processing of dual resolution precipitation forecasts. *12th International Conference of the ERCIM WG on Computational and Methodological Statistics and 13th International Conference on Computational and Financial Econometrics (CFE-CMStatistics)*, London, UK, 18 – 21 December 2019.

Baran, S., Szokol, P. and **Szabó, M.** Calibration of wind speed ensemble forecasts using truncated GEV-based EMOS approach. *Bernoulli – IMS One World Symposium*, Online, 24 – 28 August 2020.

Baran, S., Szokol, P. and **Szabó, M.** Calibration of wind speed ensemble forecasts using truncated GEV-based EMOS approach. *The 1st Conference on Information Technology and Data Science (CITDS)*, Online, 6 – 8 November 2020.

Szabó, M., Gascón, E. and Baran, S. Distribution-based statistical post-processing methods for dual-resolution precipitation forecasts. *The sixth conference of the Deutsche Arbeitsgemeinschaft Statistik (DAGStat)*, Hamburg, Germany, 28 March - 1 April 2022.

Baran, S., Szokol, P. and **Szabó, M.** Calibration of wind speed ensemble forecasts using truncated GEV-based EMOS approach. *15th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics)*, London, UK, 17 – 19 December 2022.

Szabó, M., Gascón, E. and Baran, S. Distribution-based statistical post-processing methods for dual-resolution precipitation forecasts. *The 23rd European Young Statisticians Meeting (EYSM)*, Online, 11-15 September 2023.

Poster presentations

Baran, S., Ben Bouallègue, Z., Leutbecher, M. and **Szabó, M.** Statistical post-processing of dual resolution ensemble forecasts. *The 9th International Workshop on Applied Probability (IWAP)*, Budapest, Hungary, 17 – 21 June 2018.

Szabó, M., Gascón, E. and Baran, S. Parametric post-processing of dual-resolution precipitation ensemble forecasts. *The 16th German Probability and Statistics Days (GPSD)*, Essen, Germany, 7 - 10 March 2023.