ORIGINAL ARTICLE

# Statistical post-processing of dual-resolution ensemble precipitation forecasts across Europe

**Estíbaliz Gascón[1]* | David Lavers[1] | Thomas M. Hamill[2] | David S. Richardson[1] | Zied Ben Bouallègue[1] | Martin Leutbecher[1] | Florian Pappenberger[1]**

[1]European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

[2]NOAA/Earth System Research Laboratory, Physical Sciences Division, Boulder, Colorado, USA

**Correspondence**
Estíbaliz Gascón, European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom.
Email: estibaliz.gascon@ecmwf.int

**Funding information**
The authors gratefully acknowledge financial support from the Horizon 2020 IMPREX project (Grant Agreement No. 641811)

This article verifies 1 to 10-day probabilistic precipitation forecasts in June, July, and August 2016 from an experimental dual-resolution version of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble prediction system. Five different ensemble combinations were tested. These comprised subsets of the 51-member operational ECMWF configuration (18 km grid) and an experimental 201-member lower-resolution configuration (29 km grid). The motivation of the dual-resolution ensemble forecast is to trade some higher resolution members against a larger number of lower resolution members to increase the overall ensemble size at constant overall computational cost. Forecasts were verified against precipitation analyses over Europe. Given substantial systematic errors of precipitation forecasts, both raw and post-processed dual-resolution ensemble predictions were evaluated. Post-processing consisted of quantile mapping, tested with and without an objective weighting of sorted ensemble members using closest-member histogram statistics. Reforecasts and retrospective precipitation analyses were used as training data. However, the reforecast ensemble size and the dual-resolution ensemble sizes differed, which motivated the development of a

* Corresponding author.

novel approach for developing closest-member histogram statistics for the larger real-time ensemble from the smaller reforecast ensemble. Results show that the most skillful combination was generally 40 ensemble members from the operational configuration and 40 from the lower-resolution ensemble, evaluated by continuous ranked probability score, Brier Scores at various thresholds, and reliability diagrams. This conclusion was generally valid with and without post-processing. Reliability was improved by post-processing, though the improvement of resolution component is not so clear. The advantages of many members at higher resolution was diminished at longer lead times; predictability of smaller scale features was lost, and there is more benefit from increasing ensemble size to reduce sampling uncertainty. This paper evaluates only one aspect in deciding on any future ensemble configuration and other skill related considerations need to be taken into account.

**KEYWORDS**
Dual-resolution ensemble, closest-member weighting, precipitation, post-processing, quantile mapping, verification

## 1 | INTRODUCTION

When deciding on the future configuration of an operational ensemble prediction system, it is common to assume that some fixed amount of computational resources and wall time will be available for real-time production. Evaluations may then be performed on potential tradeoffs of ensemble size versus resolution to determine a final configuration meeting these constraints. Currently, the European Centre for Medium-Range Weather Forecasts (ECMWF) generates 51 real-time ensemble predictions twice daily at TCo639 resolution, of approximately 18 km grid spacing, with 91 vertical levels (Haiden et al., 2018). The optimal configuration depends in part on the fidelity of the ensemble predictions, which generally improves with each upgrade. Fewer members might be computed at higher resolution, improving each member's forecast (Buizza, 2010) but with increased sampling variability due to the fewer members (Buizza and Palmer,

1998; Richardson, 2001). Alternatively, a larger ensemble with potentially greater biases for each member but with reduced sampling variability could be generated. From previous testing, the meteorological community has learned that the relative tradeoff of ensemble size and resolution may have complex dependencies, changing with the variable of interest, the metric used for evaluation (Lei and Whitaker, 2017), the forecast lead time (Ma et al., 2012), and whether statistical post-processing was applied or not (Baran et al., 2019).

The ECMWF 2016–2025 Roadmap (https://www.ecmwf.int/en/about/what-we-do/strategy) describes the organization's goal to produce some operational ensemble forecasts at 5-km grid spacing by 2025. This has motivated ECMWF to conduct investigations with *dual − resolution* ensemble prediction with some members at higher resolution (eventually 5 km) to exploit the value of high resolution and additional members at lower resolution, to decrease the sampling error. The main motivation of this dual-resolution configuration is to trade some higher resolution members against a larger number of lower resolution members to increase the overall ensemble size at constant overall computational cost. Other research is ongoing at ECMWF to determine the potential tradeoffs of such a dual-resolution system. See Leutbecher and Ben Bouallègue (in preparation); Baran et al. (2019).

The research question to be addressed in this article is the relative skill of probabilistic forecasts of precipitation in various configurations of a dual-resolution version of the ECMWF ensemble prediction system. Accurate probabilistic precipitation forecasts are important to many customers, including hydrologists. For example, improved precipitation guidance for hydrological models can improve flood prediction. This is one of the goals of the European Union (EU) 2020 IMPREX (Improving PRedictions and management of hydrological EXtremes; Van den Hurk et al. (2016) project. Hence, probabilistic precipitation forecast skill and reliability should be carefully evaluated when making decisions on future ensemble prediction system configurations.

Despite the many improvements in numerical weather predictions (NWP) over the last two decades (Buizza and Leutbecher, 2015), probabilistic precipitation forecasts are still typically unreliable in part because of limitations in the underlying prediction system (Hamill et al., 2017). These limitations include simple sampling variability, but also as a lack of spread (Hamill and Colucci, 1998; Buizza, R., 2018) and biases, both location and state dependent. For example, (Hamill, 2012) found that light precipitation in operational global ensemble predictions was commonly over-forecasted and heavy precipitation under-forecasted. Such biases in precipitation may also change from one

season to the next (Hamill, T. M., 2018). For these reasons, statistical post-processing of the output of deterministic and ensemble prediction systems is commonly an integral part of the numerical weather prediction process. With statistical post-processing, the statistician develops relationships between past model forecasts and observations which are then used to adjust the real-time forecast. These commonly improve the skill and reliability of the probabilistic quantitative precipitation forecasts, or PQPF; (Wilks and Hamill, 2007; Hamill et al., 2008, 2006, 2013; Hamill and Whitaker, 2007; Baran and Nemoda, 2016; Ben Bouallègue, 2013). Since careful statistical post-processing can add greatly to PQPF skill and reliability and might change the resolution/ensemble size tradeoff, an evaluation of possible ensemble configurations would be more informative if the skill of post-processed PQPFs were also considered.

The article will thus evaluate PQPF skill and reliability from a dual-resolution ensemble in different configurations, both raw and after post-processing. The evaluation will include 24-h PQPFs over Europe from five different dual-resolution ensemble combinations and lead times from +1 to +10 days. In this study, each ensemble is calibrated separately. Readers interested in optimal combination of multi-model ensembles can refer to Ben Bouallègue et al. (2019) and references therein.

While many precipitation post-processing methods have been proposed in the literature, we choose to use one that has been recently demonstrated to perform well in a US-based application (Hamill and Scheuerer, 2018). This approach sequentially applies two commonly used post-processing components, *quantile mapping* (Hopson and Webster, 2010) and an approach inspired by *best − member dressing*. (Roulston and Smith, 2003; Fortin et al., 2006; Hamill and Scheuerer, 2018). Quantile mapping leverages cumulative distribution functions (CDFs) of forecasts and observations in a training dataset. The quantile in the CDF associated with a particular forecast amount is determined. The forecast amount is then replaced with the amount associated with the same quantile in the observed/analysed CDF, thereby ameliorating amount-dependent bias. Subsequently, each quantile-mapped member is objectively weighted, and the final event probabilities are estimated from the weighted relative frequency. The statistical characteristics of the weights are determined from past ensemble forecasts, specifically the frequency of a given sorted, quantile-mapped member closest to the observed.

There are particular challenges associated with the statistical post-processing of precipitation. Post-processing of other variables such as short-lead temperature forecasts may yield improved probabilistic forecasts when trained

with shorter training data sets (Stensrud and Yussouf, 2003; Yussouf and Stensrud, 2007; Hagedorn et al., 2008; Hamill, 2012). Unfortunately, the successful calibration of heavier precipitation amounts typically requires larger training sample sizes. Also, precipitation forecast errors may be strongly location dependent, and because heavier precipitation amounts are uncommon, there may be an insufficient number of similarly heavy precipitation forecasts at a given location in a short training data set to properly estimate the location-dependent forecast-error characteristics. Two possible approaches to help address this problem are the use of supplemental locations (Hamill et al., 2008, 2017; Lerch and Baran, 2017) and the use of a longer, more complete time series of reforecasts. With the former approach, at every location where calibration is desired, other locations are identified that have similar precipitation climatologies and geographic characteristics. The assumption is that the systematic errors will be similar at the original location and at the supplemental locations, and thus the training data for the original location can be bolstered by training data at the supplemental locations. With the latter approach, the limitations of a short time series of past forecasts is explicitly acknowledged. The prediction centre thus generates as many retrospective forecasts as is practical with the same model version and ideally the same data assimilation system used to generate the real-time forecasts. This "reforecast" procedure has been applied for several model versions of the US National Weather Service Global Ensemble Forecast System (GEFS) (Hamill et al., 2004, 2006, 2013; Hamill and Whitaker, 2007) and reforecasts are now regenerated for each model version in the ECMWF ensemble (Hagedorn, 2008; Vitart et al., 2019). A complication to the use of reforecasts with the proposed objective dressing approach is that the ECMWF reforecast ensemble has only 11 members while the real-time configuration has 51 members. Previously, the quantification of dressing statistics (Hamill and Scheuerer, 2018) assumed that training ensemble size and real-time ensemble size were the same. In this application, we will thus discuss a novel algorithmic modification that permits objective estimation of dressing statistics for a larger, real-time ensemble to be estimated from training data comprised of a smaller ensemble. That is a secondary goal of this paper.

The article now provides more detail on the specifics of the post-processing technique and the results of an evaluation with a dual-resolution ensemble. Section 2 contains the description of the data to be used in the study (2.1 - 2.3), the calibration methodology (2.4), and the verification methodologies (2.5). Section 3 provides the evaluation of the different dual-resolution ensemble tests with several verification scores and reliability diagrams. Finally, section 4

contains the discussion and conclusions of this study.

# 2 | DATA, CALIBRATION, AND VERIFICATION METHODOLOGY

## 2.1 | Reforecast training data

For the calibration process, we utilise the 11-member reforecasts (1 control forecast and 10 perturbed forecasts) that were computed twice weekly (Mondays and Thursdays) covering the JJA 1996-2016 period. Data to +246-h lead were utilised here. The 11-member reforecasts were computed for both resolutions of the dual-resolution system, TCo639 and TCo399 simulating the availability of dual-resolution reforecast training data in a hypothetical future operational prediction system.

## 2.2 | EFAS gridded precipitation analyses

The European Flood Awareness System (EFAS; Ntegeka et al. 2013) provided the 24-h gridded accumulated precipitation validation and training data . The EFAS analysis extended domain database covers Europe and some surrounding countries (Figure 1). The data set used contained 24-h accumulated daily precipitation analyses from 06 UTC of a given day to 06 UTC of the following day. Data were archived on a Lambert Azimuthal Equal Area projection grid (5 km grid spacing). The interpolation algorithm from the station observations to the EFAS extended domain grid was SPHEREMAP (Willmott et al., 1985), with the spherical adaptation of the interpolation scheme developed by Shepard (1968). It is based on a combined distance and angular weighting plus a correction using the gradient of the observations. EFAS data were available for years from 1996 to 2016, covering both the training (1996-2015) and validation (2016) periods. This data set is used as the observation analysis input to initialise the EFAS hydrological model. The IMPREX project has as main goal to improve meteorological and hydrological predictions for a better forecast of floods. As a final step to achieve this objective, this calibration and the different dual-resolution ensemble configurations will be tested as forcings in the EFAS hydrological model and for this reason we decided to use the same analysis database. This test

will be developed in another scientific article.

**FIGURE 1** Extended EFAS domain, encompassing the verification region (gray shaded area).

For training of the post-processing algorithm, all EFAS grid points were considered, a practical necessity given the use of the supplemental locations algorithm. For verification, forecast characteristics were validated only at a smaller number (2400) of more trustworthy analysis grid points corresponding to the locations of European SYNOP stations. These points are usually used in ECMWF forecast verification (Haiden et al., 2018). Tests using all the grid points were also performed, with quite similar verification results (not presented).

## 2.3 | Dual-resolution ensemble configurations

In this examination of dual-resolution ensemble forecast characteristics, two horizontal resolutions of the ECMWF Integrated Forecast System (IFS) were examined: TCo639 (~18 km resolution) and TCo399 (~29 km resolution). 51 members were produced at TCo639, and 201 members at TCo399, but in the dual-resolution ensemble investigation we will only use the perturbed ensembles (50 and 200 ensembles, respectively) and not the control members. Each ensemble system (higher-resolution and lower-resolution) is calibrated separately before setting up the different dual-resolution ensemble combinations. Five different dual-ensemble combinations with *HH* higher-resolution and *LL* lower-resolution perturbed members were tested with the structure *HH/LL*: 50/0, 40/40, 20/120, 10/160, and

0/200, all with similar computational expense. To choose the subsample of each ensemble forecast system to create the different dual-ensemble combinations, we select the first HH (from high-resolution) or LL (from low-resolution) from the original ensemble members (not from the sorted ones when we apply the weighting step) and we give to them the same weight in the dual combination. It means that for the combination 40/40, we will select the first 40 ensemble members from the high-resolution system and the first 40 from the low-resolution one and all 80 members will contribute equally to the combined dual-resolution ensemble forecast. These ensemble forecast systems will be referred to as the "real-time" ensembles hereafter. Both higher-resolution and lower-resolution ensembles use IFS model cycle 41r2, the operational model version during the verification period. All the initial conditions and stochastic representation of model uncertainties were the same for both ensemble resolutions; Leutbecher (2018) describes further details.

The real-time dual-resolution ensemble forecasts were generated once daily during the June, July, and August (JJA) 2016 period to 246-h lead time (10 days), with all forecasts initialised at 00 UTC. To match the validation data periods, discussed below, 24-h accumulated precipitation were calculated from 06 UTC of the corresponding study day to the 06 UTC of the following day, for example from +6 to +30 h lead time (day +1). This was chosen to coincide with the accumulated period for the EFAS precipitation analyses. Both the 2016 dual-resolution simulated operational forecast data and the reforecast training data were interpolated to the EFAS horizontal grid, discussed below, before the calibration and verification processes, using a nearest-neighbour technique. That is, the forecast value at the EFAS grid point is simply obtained by taking the value from the nearest model grid point.

## 2.4 | Calibration

A schematic providing high-level details of the calibration method is presented in Figure 2. Each ensemble system (higher-resolution and lower-resolution) is calibrated separately before setting up the different dual-resolution ensemble combinations. The procedure will be explained as applied to the single-resolution, 51-member real-time ensemble forecast system. An identical procedure was applied to calibrate the lower-resolution ensemble with 201 ensemble members. The dashed line on the diagram separates the processing of reforecast data (above the line) from the real-time

processing (below the line). We first outline the calibration procedure at a high level of abstraction, followed by a detailed description of each component. Much of the algorithmic detail follows that outlined in Hamill and Scheuerer (2018).

The reforecast processing begins with generation of the cumulative distribution functions (CDFs) for the reforecasts and EFAS analyses. These will leverage a precomputed set of supplemental locations that indicate what other grid points are suitable for increasing the sample size used to estimate the CDFs. With reforecast and analyzed CDFs generated, the reforecasts are quantile mapped. These quantile-mapped reforecasts are then compared to the analysed data to determine the closest-member histograms.

The rest of the processing in Figure 2 is performed on the real-time ensemble, and both the perturbed and control were calibrated. The 51-member real-time ensemble is quantile mapped using CDFs developed from reforecast and EFAS data using supplemental locations. We will apply the post-processing to the 51-members ensemble system as might be applied in current operations, however only the 50 perturbed members will be used to create the experimental dual-resolution ensemble combinations. Given the differing sizes of the reforecast and real-time ensembles, the 11-dimensional closest-member histograms are unsuitable for determining weights to apply to a sorted 51-member ensemble and determination of probabilities for 51 intervals between 0 and 1. Estimated 51-dimensional closest-member histograms are thus determined through a procedure that involves fitting a Beta distribution. The resulting raw, quantile-mapped, and quantile-mapped and weighted ensembles can then each be verified using standard methods.

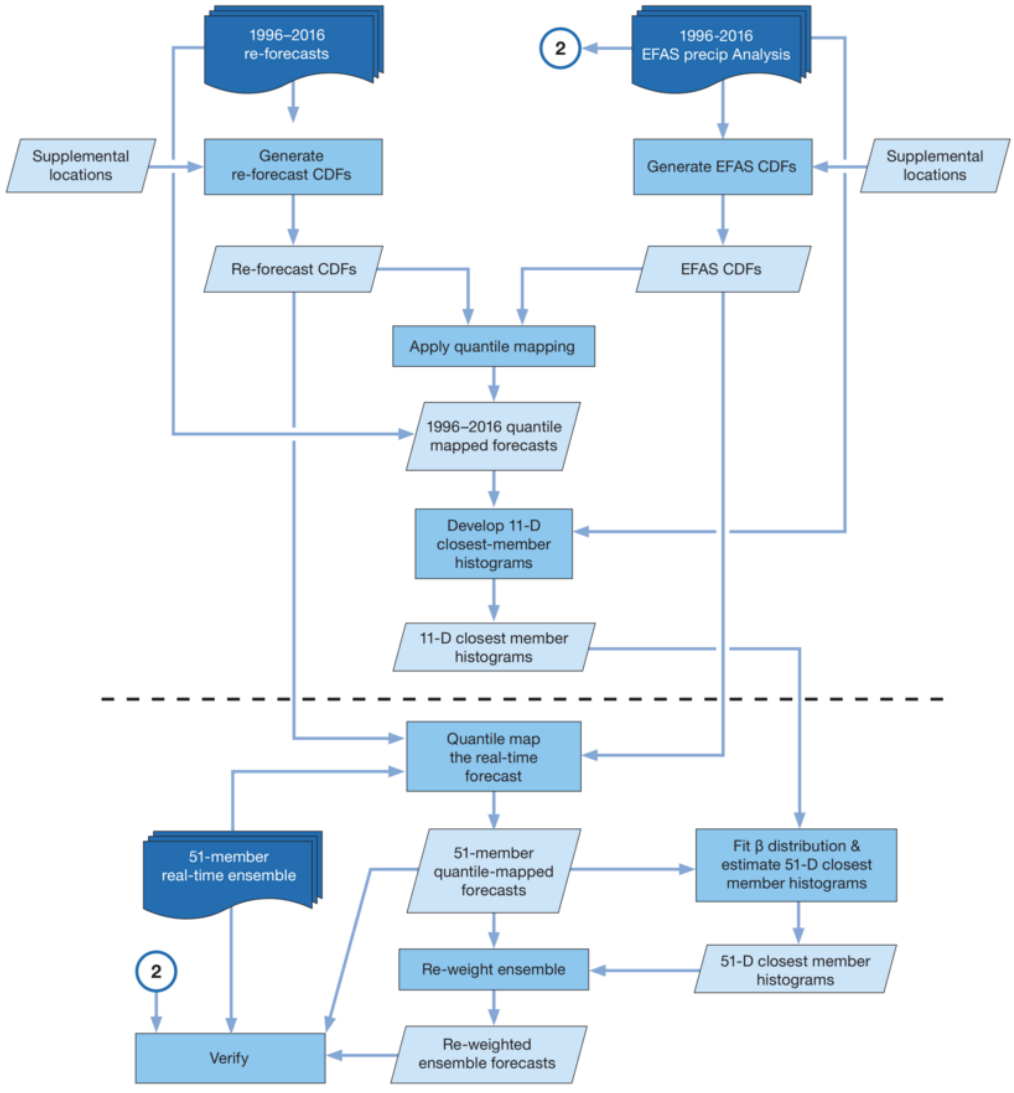


**FIGURE 2** Data and process flow diagram for the quantile mapping and closest-member weighting calibration procedures. Rectangles denote processing steps; parallelograms represent data stores.

### 2.4.1 | Quantile mapping

The statistical adjustment of ensemble forecasts begins with quantile mapping. Assume we have a raw forecast amount $\bar{x}$ which provides an estimate of the true (unknown) precipitation amount $x$. Assume we have climatological CDFs

$\Phi_f(x)$ and $\Phi_a(x)$ for the forecasts and analyses, respectively. Given a precipitation amount, the CDFs return the non-exceedance probability $q$. The inverse function, $\Phi_a^{-1}(q)$ is the analysis quantile function, which here returns the corresponding analysed amount associated with that quantile. Quantile mapping thus adjusts the forecast to be consistent with the analysed CDF:

$$\tilde{y} = \Phi_a^{-1}\left[\Phi_f(\tilde{x})\right]. \quad (1)$$

CDFs were needed to perform the quantile mapping. The reforecasts are split into 19 years of training data to populate the CDFs (using the nine reforecast dates of the year closest to the Julian day of the reforecast). Then, the remaining year of training data is quantile mapped. The procedure is repeated to provide quantile-mapped precipitation amounts spanning the 20 years and we repeat the procedure for each one of the 11 reforecast members separately. These data are then used in a second step of the training process, as input for developing closest-member histograms, discussed below. Reforecasts and corresponding analyses at 50 unique supplemental locations were used at each EFAS grid point to provide extra training data. The CDFs were estimated with a fraction zero, i.e., a fraction of samples with zero precipitation, and with the shape $\alpha$ and scale $\beta$ parameters of a fitted Gamma distribution for non-zero amounts. At each grid point there were thus 20 years $\times$ 1 member (each member is calibrated separately) $\times$ 9 dates $\times$ 50 supplemental locations = 9000 samples used to populate the forecast CDFs for the quantile mapping of each the 11 ensemble members, and the same 9 dates to populate the EFAS analysis CDFs. The 50 supplemental locations were selected based on the similarities of analysed climatologies and terrain characteristics and different for each month of the year, directly following the Hamill and Scheuerer (2018) methodology. Figure 3 provides an example of chosen supplemental locations for Madrid (Spain) illustrating how the supplemental locations are chosen to match the underlying precipitation climatology characteristics.

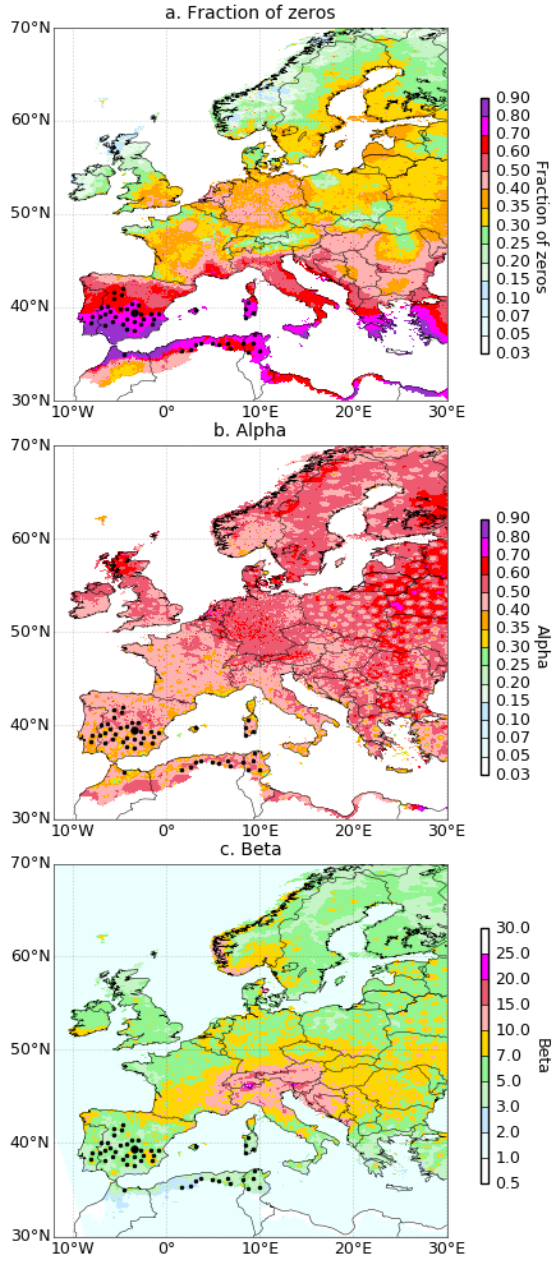


**FIGURE 3** Illustration of supplemental locations (black dots) of a point near Madrid (Spain, large dot) for the month of June. The location for which supplemental locations are desired is indicated by the large black dot. Chosen supplemental locations are identified by the smaller black dots. Climatological precipitation distribution parameters are comprised of a fraction zero, i.e., the fraction of samples with zero precipitation, and with the shape ($\alpha$) and scale ($\beta$) parameters of a fitted Gamma distribution for non-zero amounts. Colours on the maps denote the underlying EFAS 24-h precipitation analysis climatology of (a) fraction zero (b) $\alpha$, and (c) $\beta$.

Quantile mapping was also applied to the real-time ensemble as the first step in the correction of systematic error. In this case, the CDFs for the quantile mapping were developed from the full 20 years × 1-member (control forecast) × 9 cases × 50 supplemental locations, thus providing 9000 total samples from real-time forecast and EFAS analysis precipitation to generate the empirical CDF. This step is only applied to the verification period that corresponds to JJA 2016 (3 months).

Because of the model's tendency to over-forecast light precipitation, quantile mapping sometimes adjusted a forecast light precipitation amount to zero. Suppose the CDFs indicated an under-forecasting of light precipitation. In this case there were multiple quantiles of $\Phi_f(x)$ that were likely associated with zero, and we face a non-uniqueness problem when zero precipitation is forecast: Is this representing the 0th percentile of the forecast CDF, or perhaps the 5th percentile? This problem was avoided by implementing an ad-hoc rule, such that zero raw forecast amounts were retained without quantile mapping.

### 2.4.2 | Generating weights to apply to sorted ensemble members

The second corrective step during the training process was applied after the quantile mapping of ensemble members. Suppose for the moment there was a rational basis to believe that the analysed state was more likely to be near to one sorted member than the others. Let's assume we have a vector of weights $\mathbf{w} = [w_{(1)}, \ldots, w_{(m)}]$ associated with the sorted members that reflect this likelihood, where $_{(i)}$ denotes the $i$th rank. Weighted probabilities can then be generated in a straightforward manner. When considering the probability of exceeding the threshold amount $t$, we define an indicator function for the $i$th sorted member:

$$I(i) = \begin{cases} 0 & \text{if } \bar{y}_{(i)} < t \\ 1 & \text{if } \bar{y}_{(i)} \geq t\,. \end{cases} \tag{2}$$

Weighted probabilities of exceeding the amount $t$ are then generated as follows:

$$P(x > t) = \sum_{i=1}^{m} I(i)\, w_{(i)}\,. \tag{3}$$

The question then turns to how to objectively generate weights associated with each sorted member. A procedure for doing so was described in Hamill and Scheuerer (2018) using the previously mentioned closest-member histograms. To generate closest-member histograms from reforecasts, after a set of cases of ensemble training data for a particular lead time has been quantile mapped, we have an 11-dimensional vector $\tilde{\mathbf{y}} = [\tilde{y}_1, \ldots, \tilde{y}_{11}]$ of estimates of the unknown precipitation amount. For the training sample (each date and each grid point), these quantile-mapped ensemble data were sorted, $\tilde{\mathbf{y}}^s = [\tilde{y}_{(1)}, \ldots, \tilde{y}_{(11)}]$ and then compared to the analysed precipitation amount. The rank of the nearest sorted member was determined, and the histogram count associated with that was incremented by one. Closest-member histograms were thus generated tallying over these many samples which sorted member was closest to the analyzed amount. Following Hamill and Scheuerer (2018), separate closest-member histograms were generated in this application for different quantile-mapped ensemble-mean amounts. However, separate histograms were *not* estimated separately for each grid point; it was assumed that the previous quantile mapping removed any location-dependent biases.

How can one use the 11-dimensional closest-member histograms from reforecasts to estimate weights in a sorted, 51-member ensemble? Closest-member histograms for a 51-member ensemble can be estimated through the fitting of Beta distributions (Wilks, 2011, section 4.4.4). A Beta distribution provides a continuous probability density function associated with a quantile $q$ in the range of (0,1). The pdf $f(q, \alpha, \beta)$ of the Beta distribution is

$$f(q, \alpha, \beta) = \left(\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) q^{\alpha-1}(1-q)^{\beta-1}; \quad 0 < q < 1; \quad \alpha, \beta > 0. \tag{4}$$

Here $\alpha$ and $\beta$ are the parameters of the Beta distribution, and $\Gamma(\cdot)$ is the Gamma function. Parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ are commonly generated from the method of moments as

$$\hat{\alpha} = \frac{\bar{q}^2(1-\bar{q})}{s^2} - \bar{q} \quad \text{and} \tag{5}$$

$$\hat{\beta} = \frac{\hat{\alpha}(1-\bar{q})}{\bar{q}}, \tag{6}$$

where $\bar{q}$ and $s^2$ are the sample mean and standard deviation, respectively. Beta distributions have flexible shapes and can be fit to resemble the closest-member histograms.

The procedure for generating closest-member histograms for the real-time, 51-member ensemble was thus as follows: (a) fit a Beta distribution to the 11-dimensional closest-member histogram based on the ECMWF reforecast training data. (b) Create closest-member histogram weights associated with the larger $HH$=51-member ensemble by integrating the Beta distribution into 51 equally spaced regions spanning 0 to 1. For step (a), sample means and variances were needed to apply the method of moments to estimate the Beta distribution parameters. Let $\mathbf{w}^{11}$ represent the appropriate 11-dimensional closest-histogram vector of weights from the reforecast ensemble based on the quantile-mapped ensemble mean. Let's also denote a vector $\mathbf{a}$ that provides the corresponding central value associated with each rank in the closest-member histogram when mapped to the interval (0,1):

$$\mathbf{a} = \left(a_1, \ldots, a_{11}\right) = \left(\frac{0}{11} + \frac{1}{2\times 11}, \ldots, \frac{10}{11} + \frac{1}{2\times 11}\right) \tag{7}$$

The sample mean $\bar{q}$ is

$$\bar{q} = \sum_{i=1}^{11} a_i w_i^{11} \quad , \tag{8}$$

and the sample variance is calculated from a closest-histogram weighted sum of squared differences from the sample mean

$$s^2 = \frac{10}{11} \sum_{i=1}^{11} (a_i - \bar{q})^2 \times w_i^{11} \quad . \tag{9}$$

In the second step, the closest-member histogram weights are computed through integration of the fitted Beta distribution. Let $j$ indicate the rank in the sorted, 51-member ensemble and the index in the closest-member histogram vector $\mathbf{w}^{51}$. The closest-member histogram weight for this rank was calculated as

$$w_j^{51} = \int_{(j-1)/m}^{j/m} f(q, \hat{\alpha}, \hat{\beta}) dx \, . \tag{10}$$

An example of closest-member histograms and Beta-distribution fits are provided in Figure 4. Panels (a), (b), and (c) provide the closest-member histograms for light precipitation, moderate precipitation, and heavier precipitation, respectively. Light precipitation was defined as 0.01 mm $\leq \bar{\mathbf{x}} <$ 2.0 mm; moderate precipitation was defined as 2.0 mm $\leq \bar{\mathbf{x}} <$ 6.0 mm, and heavy precipitation was defined as 6.0 mm $\geq \bar{\mathbf{x}}$, where $\bar{\mathbf{x}}$ was the raw ensemble-mean precipitation amount. When ensemble-mean precipitation was less than 0.01 mm, a uniform closest-member histogram was assumed. From Figure 4, we see that when light-mean precipitation was forecast, there was a U-shaped histogram that indicated some under-dispersion of the bias-corrected forecasts. When heavier precipitation was forecast, the lower ranks were

more heavily weighted; this would have the effect of decreasing heavy-precipitation event probabilities relative to an equally weighted ensemble. Panels (a), (b), and (c) also show histograms generated from the integration of the fitted Beta distributions into 11 bins. These appear to provide a reasonable estimate of the shape of the original closest-member histograms. Figures 4 (d), (e) and (f) then provide an example of the estimated 51-dimensional closest-member histograms, illustrating the similar histogram shapes but with finer discretization.

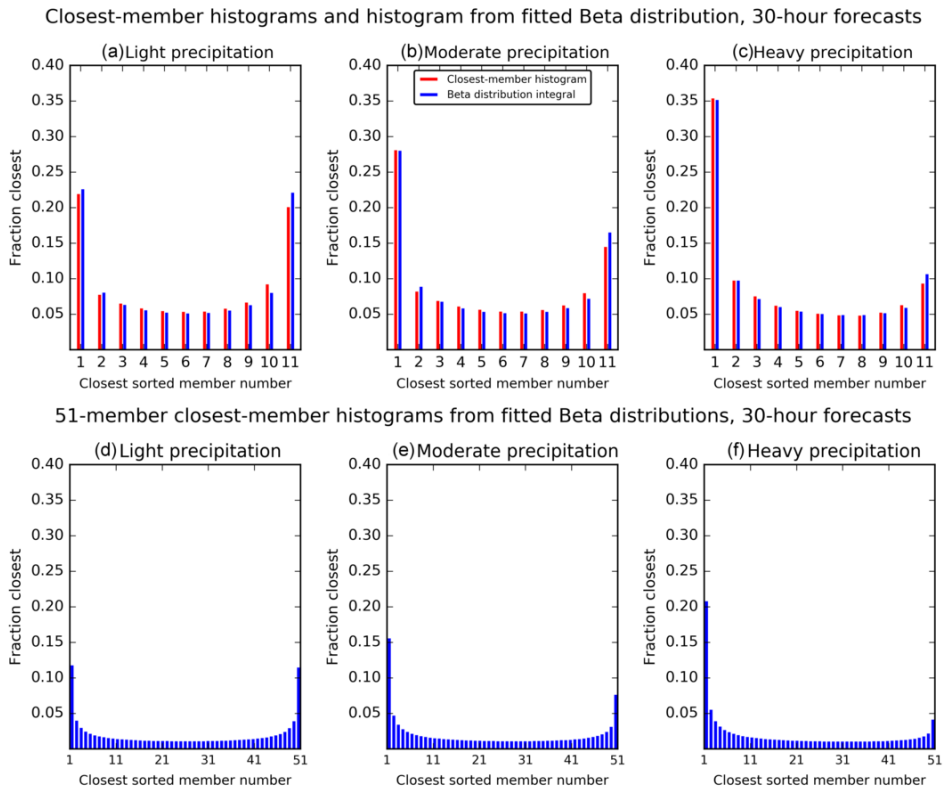


**FIGURE 4** Illustration of estimated 11-dimensional closest-member histograms (red) and histograms from fitted Beta distributions (blue) for the 14 July 2016 ECMWF reforecast ensemble, +6 to +30 h forecasts. (a) Histograms for light ensemble-mean precipitation, (b) moderate, and (c) heavy, as defined in the text. Panels (d), (e), and (f) provide estimates of the closest-member histograms for a 51-member ensemble and for light, moderate, and heavier precipitation respectively. These histograms were generated through integration of the fitted Beta distributions into 51 equally spaced bins.

With the 51-dimensional closest-member histograms generated from the training data, the statistical adjustment of the real-time forecasts proceeded. Note that the calculation of the closest-member histogram was developed for

the 51-member ensemble (high-resolution) and 201-member ensemble (low resolution) system, thus before the dual-resolution combinations were built. Then, each real-time ensemble member will be weighted based on its corresponding ensemble system closest-member histogram (51 or 201), before extracting the number of ensembles that we need to create each dual combination. The weights could be different if the closest-member histograms are built for each dual-resolution combination (for instance, considering 40 members from HH and 40 from LL), and this could be a topic for further research.

Real-time forecasts were quantile-mapped using reforecast-based CDFs (equation 1). Based on the ensemble-mean precipitation amount, the appropriate closest-member histogram was selected. Now, for the re-weight ensemble, we will use the the quantile-mapped ensemble for a particular lead time and grid point and the associated 51-dimensional closest-member histogram. The procedure to be applied to adjust the quantile-mapped members again leverages the machinery of quantile mapping, using it to perform a stretching of the original ensemble so that members are more equally likely in their statistical character.

For the procedure here to adjust the quantile-mapped members to have characteristics more like equally likely members, $\Phi_f(x)$ and will no longer represent a CDF of past forecasts. Instead, it now depicts a distribution for a particular grid point fitted to today's quantile-mapped under the assumption that all members are given equal weight. Similarly, $\Phi_a(x)$ now depicts a distribution for a particular grid point fitted to today's quantile-mapped and closest-histogram weighted ensemble.

The procedure for estimating the fitted distributions for the prior (quantile-mapped) and posterior (quantile mapped and weighted) are functionally equivalent. In the latter case, weights in the procedure are supplied by the closest-member histograms. In the former, weights are constant, 1/51. Fraction zero and positive precipitation will be separately processed, following Hamill and Scheuerer (2018) procedure.

With the fraction zero and gamma-distribution parameters separately estimated for quantile-mapped unweighted and weighted ensembles, we have fitted $\Phi_f(x)$ and $\Phi_a(x)$ and the original ensemble of quantile-mapped values. The second mapping procedure is now applied.

In the limit of infinite training data, the quantile mapping should produce a climatological distribution of the forecasts that is identical to the climatological distribution of the analyses provided the real-time forecasts are consistent

with the reforecasts. The weighting introduced in this second step discussed in this section can in principle deteriorate the climatological distribution of the quantile mapped forecasts. It will be an important question for future research to examine whether this is a limitation of the method in practical applications.

## 2.5 | Verification methods

Verification procedures were applied to the predefined dual-resolution ensemble combinations and considering 3 types of calibration: raw (no calibration), quantile mapped (QM), and quantile mapped combined with a weighting using the closest-member histogram methodology (QM+W). The verification period covers 3 months (JJA) in 2016 and we focus on 24-h precipitation forecasts. All the verification scores are computed from lead time day +1 (+6 to +30 h) to day +10 (+222 to +246 h) lead times with 2-days step, but only relevant results will be shown in the next section.

The Continuous Ranked Probability Score (CRPS; Matheson and Winkler 1976; Unger 1985) is the first measure used to evaluate the overall quality of PQPFs. The CRPS measures the integrated squared difference between the CDF of the ensemble forecasts and the corresponding CDF of the observations. The CRPS is sensitive to calibration and sharpness (Gneiting et al., 2014). We plot the results as raw CRPS values and CRPS differences between all the dual-resolution combinations (raw and calibrated) and the reference current ensemble operational configuration without applying any calibration (raw 50/0 combination). This shows how much improvement or degradation is obtained from different combinations of dual-resolution ensembles and post-processing.

Similar results were also calculated using the Brier Score (BS; Brier 1950; Wilks 2011). This score is the mean-squared error of the probability forecasts over the verification sample (binary) for a specific threshold of a specific variable (in our case, 24-h accumulated precipitation). We evaluated three different precipitation thresholds: $\geq$0.1 mm, $\geq$5 mm and $\geq$10 mm. The BS can be decomposed into reliability, resolution, and uncertainty components (Murphy, 1973). We will examine the BS resolution component of the forecast for the different precipitation thresholds. The CRPS corresponds to the integral of the BS over all possible thresholds. Additionally, reliability diagrams are provided for selected thresholds and different lead times.

Looking at the aggregated verification scores over Europe, one can conclude whether, on average, one ensemble

combination or a type of calibration improves probabilistic forecast performance. However, this does not take into account the potential spatial distribution of these improvements, and the user might wonder if in some areas there may be a degradation. Hence, the differences of the mean CRPS at each verification point was also determined for different ensemble dual combinations and calibration methods, compared to the raw 50/0 as a reference. Following Baran et al. (2019), a Diebold-Mariano test (DM; Diebold and Mariano 2002) was also applied at each verification point separately. This test of equal predictive performance compares the errors of the different ensemble forecasts and takes into account their temporal dependencies. We apply the test in its factor one version (the factor applied to each forecast error) but we acknowledge that other versions of the test can lead to different results. Moreover, confidence intervals associated with CRPS and Brier Score differences are obtained with the help of 2000 block bootstrap samples using the stationary bootstrap scheme with mean block length according to Politis and Romano (1994) and following the same approach as Baran et al. (2019).

# 3 | VERIFICATION OF DUAL-RESOLUTION ENSEMBLES

## 3.1 | Case study

We start with a case study to visually illustrate the typical effect of statistical post-processing on precipitation ensemble forecasts. Figure 5a shows the EFAS precipitation analysis for July 2016, while panels (b), (c), and (d) present the +126 h (day +5) probability of precipitation greater than 0.1 mm derived from the raw, QM, and QM+W ensembles, respectively. High-intensity precipitation is visible over part of Central Europe in Figure 5(a) . In the case of the raw probabilistic forecast (Figure 5b), a large red area indicating probabilities near 100 % covers most of Central and Northern Europe. After QM (Figure 5c), a decrease in the area covered by high probabilities is observed. QM+W further reduced the geographic extent of high probabilities (Figure 5d). The reduction of high probabilities, except in areas of consistently high precipitation across ensemble members, is a characteristic of the QM+W post-processing method.

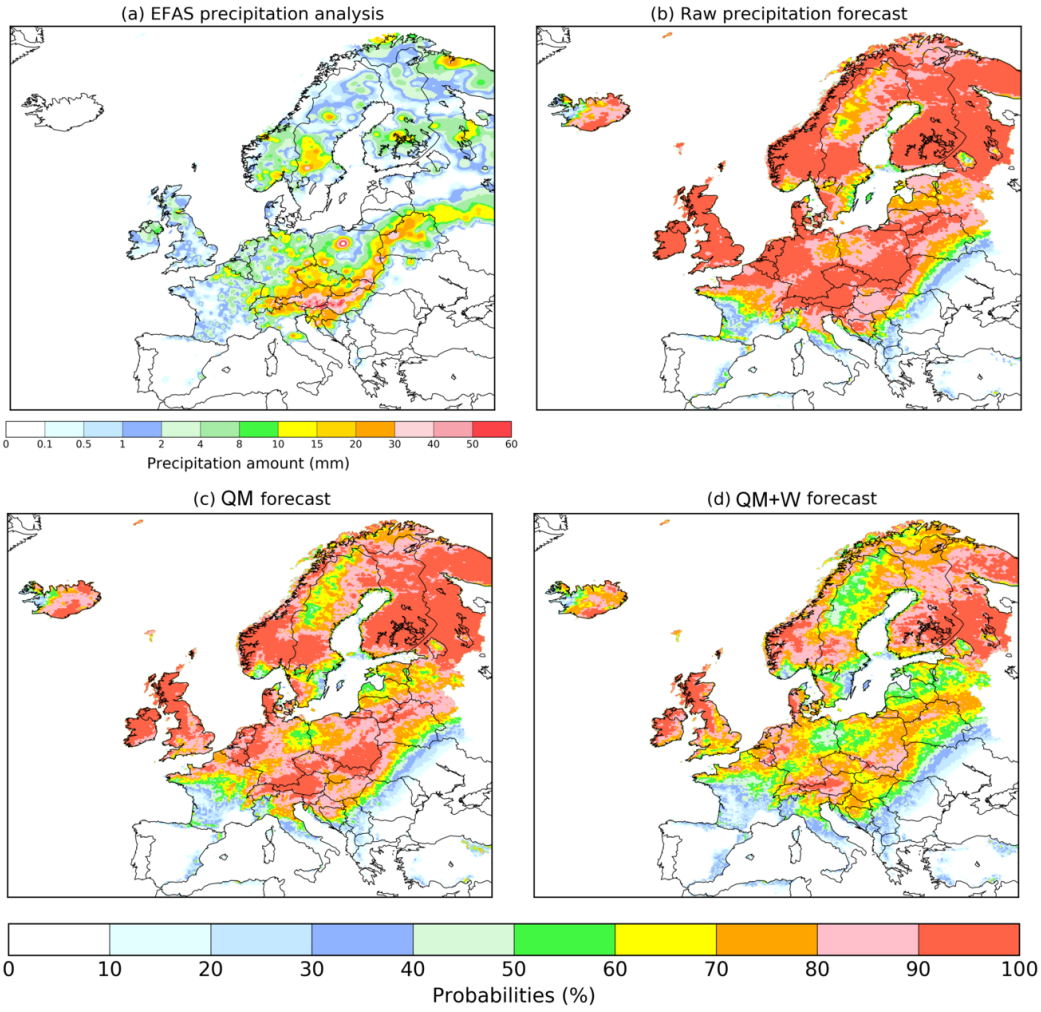


**FIGURE 5** Example of verifying precipitation analysis and associated probabilistic forecasts. (a) Verifying EFAS precipitation analysis for the 06 UTC 14 July 2016. Corresponding day +5 probability forecast of precipitation greater 0.1 mm from (b) the raw 50-member ECMWF ensemble, (c) the QM 50-member ensemble, and (d) the QM+W ensemble.

## 3.2 | Domain-averaged verification

We now consider CRPS results for different configurations of the dual-resolution ensemble ( Figure 6). Results are presented in terms of CRPS for all investigated dual-resolution combinations (Figure 6a, b, and c) and in terms of CRPS

differences with respect to the raw 50/0 ensemble, (CRPS raw 50/0 - CRPS HH/LL), where again HH is the number of higher-resolution members and LL is the number of lower-resolution members (Figure 6d, e, and f). In the former case, the lower the better, while in the later case, the higher the better.

At short lead times (up to day +5), the 40/40 ensemble is the most skillful combination, followed by the 50/0 ensemble (Figure 6a and d). At longer lead times, CRPS for the raw forecasts have similar values for all combinations except 50/0. Figures 6b and 6c show that all ensemble combinations benefit from post-processing (QM or QM+W), in particular at short lead times, and with more positive significant changes with the second technique (QM+W).

The skill improvement when applying QM with respect to the raw 50/0 forecasts is shown in Figure 6e. Note the difference in the scale of the y-axis with respect to panel (d). In this configuration, the 40/40 ensemble is also the most skillful combination across all lead times. The difference is significant up to day 5 but at longer lead times all QM calibrated combinations have comparable skill.

Figure 6(f) shows differences in skill with respect to the raw 50/0 ensemble when applying QM+W. Comparing panels (e) and (f) (note different scaling of the y-axis), we see that QM+W further improves the forecast performance, though differences are small at long lead times. In that case, the mean CRPS differences between all the combinations are small, which is consistent with results in Baran et al. (2019).

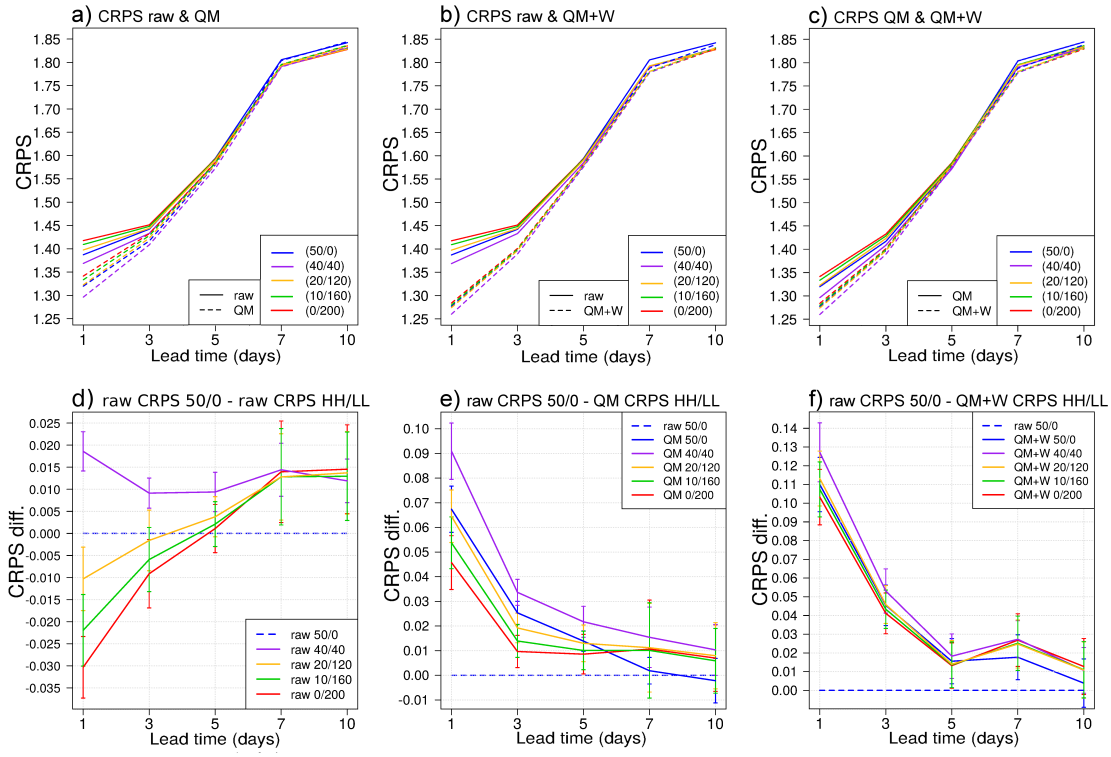


**FIGURE 6** Top panels: CRPS as a function of the forecast lead time for all investigated dual-resolution combinations for (a) raw and QM, (b) raw and QM+W, and (c) QM and QM+W forecasts. Bottom panels: CRPS differences with respect to the raw 50/0 forecasts (the higher the better) for (d) the raw ensemble combination, (e) QM dual-resolution ensembles, and (f) QM+W ensembles. 95 % confidence intervals are indicated by vertical bars.

Figure 7 shows Brier score differences for all the investigated ensemble configurations and different post-processing approaches. For a given configuration HH/LL, results are presented in the form (BS raw 50/0 - BS HH/LL), with positive differences indicating a forecast improvement with respect to the 50-member higher-resolution ensemble. As in the results for the CRPS (Fig. 6), the 40/40 combination appears to be either the best or among the best dual-resolution configurations. The 40/40 ensemble clearly outperforms the other configurations when focusing on high-intensity events and short lead times. Similarly to the CRPS results, the differences between the different dual-resolution combinations decrease with QM calibration and even more so with the QM+W calibration.

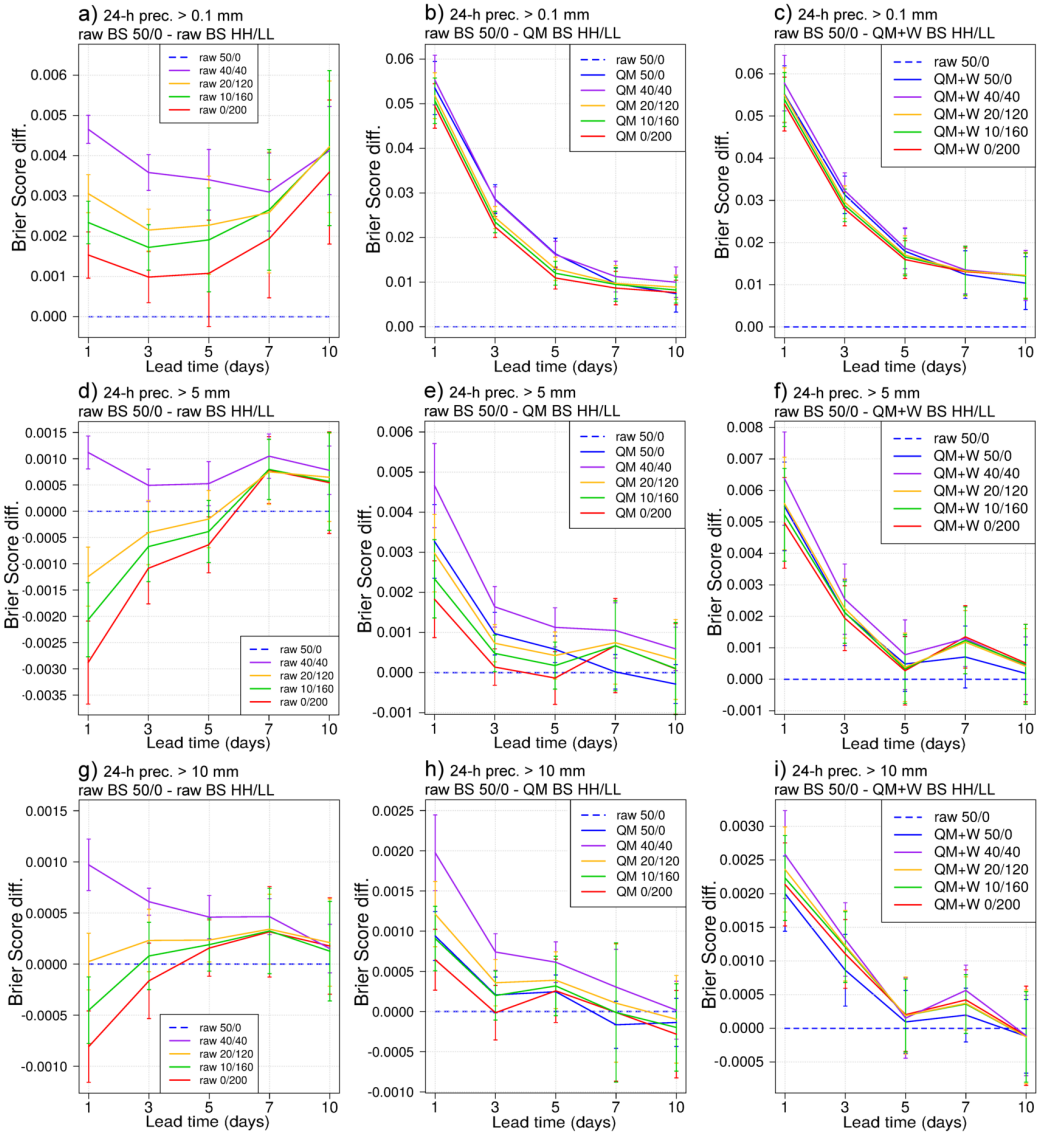


**FIGURE 7** Brier score differences for the dual-resolution ensemble configurations as a function of the forecast lead time, presented in the form BS raw 50/0 - BS HH/LL (the higher the value the better). Rows indicate the event threshold (0.1 mm, 5 mm, and 10 mm, from top to bottom) and vertical bars indicate 95 % confidence intervals. Columns indicate the type of calibration (raw ensembles, QM ensembles, and QM+W ensembles, from left to right).

Figure 8 provides reliability diagrams for the 50/0 and 40/40 combinations focusing on 3 different lead times (rows) and 3 different thresholds (columns). Similar results are obtained with other combinations (not show). Indeed,

346 we see that changing the dual resolution configuration from 50/0 to 40/40 has little impact on the reliability curves.
347 Reliability is more strongly affected by post-processing. The lack of reliability of the raw ensemble is evident at short
348 lead times and very low precipitation thresholds (Figure 8a). Figures 8 (a) and (d) show the substantial impact of QM
349 on light precipitation, with especially pronounced positive effect on the reliability at day +1. QM+W provides further
350 improvement in terms of reliability wich is consistent with results in Hamill and Scheuerer 2018. At longer lead time,
351 the raw ensemble are much better calibrated and post-processing has therefore less of an impact. The limited sample
352 size due to the short verification period explains the increased noise of the reliability curves at long lead times and for
353 high precipitation thresholds (Figures 8 h and i).

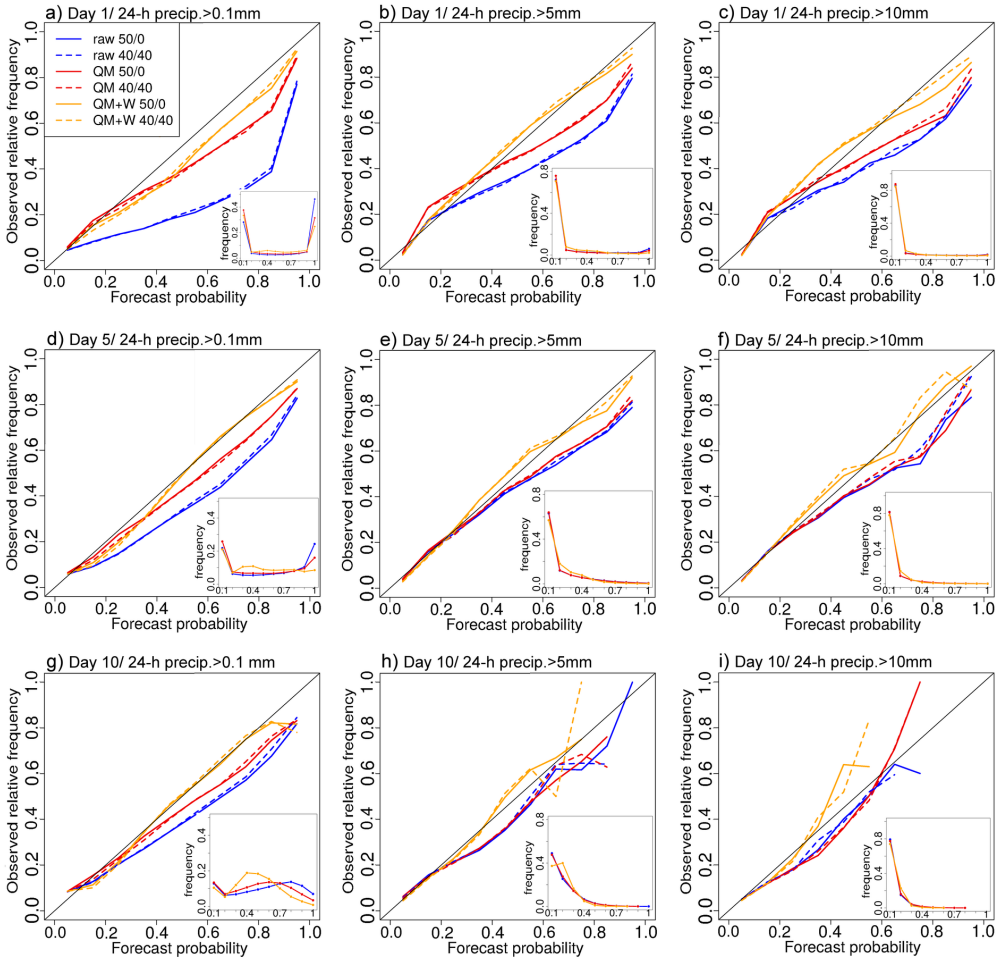


**FIGURE 8** Reliability diagrams for different lead times (rows) and for different event thresholds (columns): day +1, +5, and +10 forecasts (from top to bottom) and threshodls ≥0.1 mm, ≥5 mm, and ≥10 mm (from left to right). Blue colour corresponds to results for the raw ensemble forecast, red is for the QM forecast, and orange for the QM+W forecast. Continuous lines indicate results for the 50/0 dual-resolution ensemble combination while dashed lines indicate results for the 40/40 ensemble. Bottom right subplots show the frequency of forecast falling in each of the probability category for the 50/0 combination only.

To complement the reliability diagrams, we present results in terms of BS resolution component. Figure 9 shows BS resolution (the higher the better) for lead times of day +1 (first row), day +5 (second row), and day +10 (third row), and for different precipitation thresholds: ≥0.1 mm (first column), ≥5 mm (second column) and ≥10 mm (third column). While reliability is improved after post-processing for all investigated event thresholds and lead times (Figure 8 ), the

impact of post-processing on the forecast resolution is less univocal: we see a large improvement with QM and QM+W at day 1 for low-intensity events, some improvement with QM at day 5, but a degradation with both QM and QM+W at day 10. We would expect that post-processing would retain (or improve) the resolution of the raw forecast. This is not achieved here which suggests that there may be room for improvement of the weighting post-processing method.

It is also interesting to note that BS resolution as a function of the ensemble configuration shows a peak (maximum) for the 40/40 combination for nearly all thresholds and lead times. This is an indication that the superiority of the 40/40 combination (seen in Figures 6 and 7) originates from an increased forecast information content.

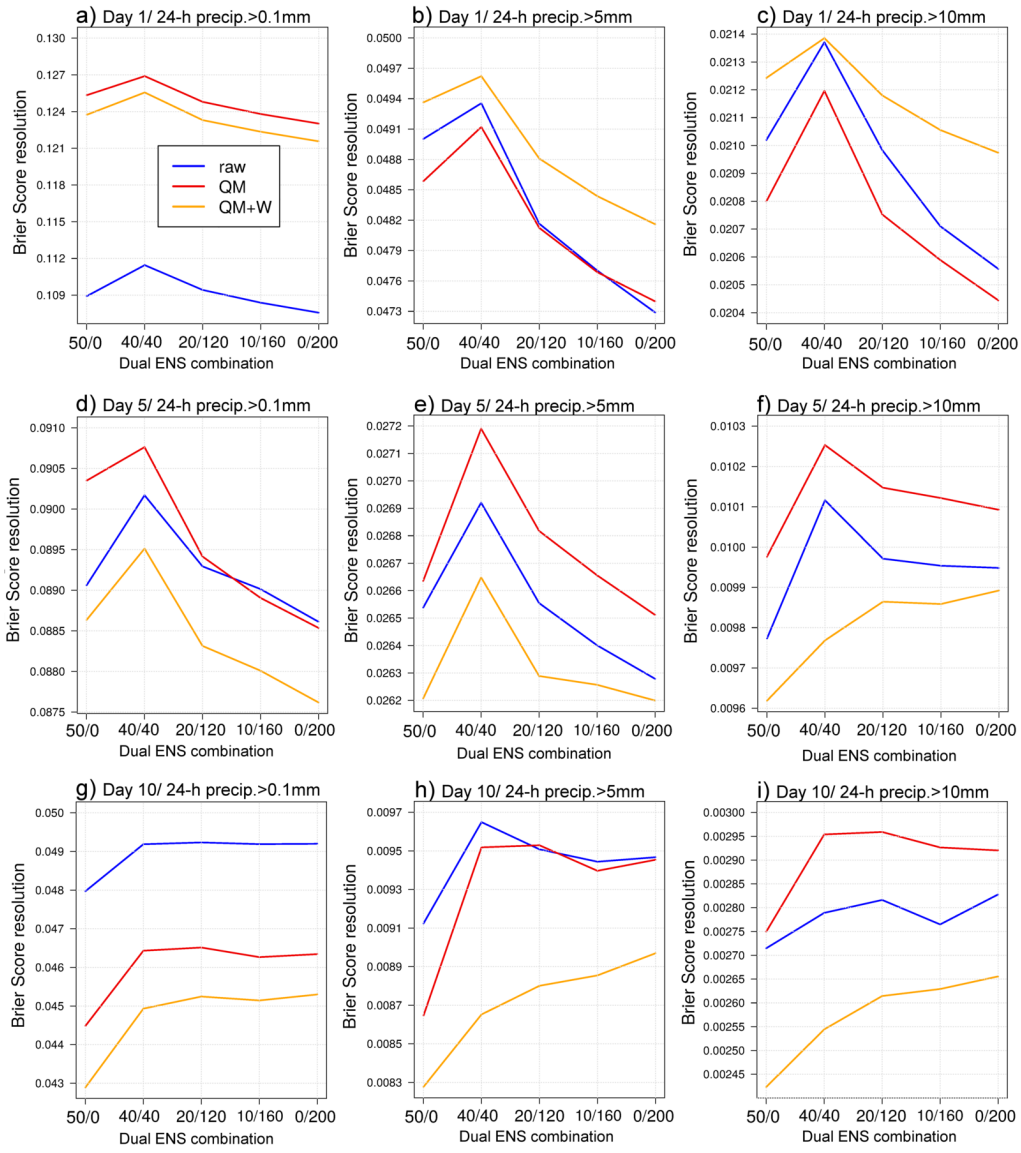


**FIGURE 9** Brier score resolution component (the higher the better) as a function of the dual-resolution ensemble combination for different lead times. Rows indicate the forecast lead time, with day +1, day +5, and day +10 (from top to bottom). Each column corresponds to a different threshold: ≥0.1 mm, ≥5 mm, and ≥10 mm (from left to right). Blue colour corresponds to the raw ensemble forecast, red to the QM forecast, and orange to the QM+W forecast.

## 3.3 | Spatial variation of CRPS

We now investigate the following question: are there any spatial pattern of improvement or degradation associated with the results presented so far? To answer this question, we consider the geographical distribution of the PQPF performance differences and their significance for different ensemble configurations and post-processing approaches. We evaluate the statistical significance of CRPS differences at each verification point using the DM test with a p-value threshold of 0.05.

Figure 10 shows a spatial representation of local CRPS differences and their significance for day +1 (left) and day +10 (right). The area shown on the map is a zoom of the study area, covering 90 % of the verification points. Stations with statistically not-significant differences are represented with small dots, while stations with statistically significant differences with larger triangles. The symbol colour indicates the value of the mean CRPS differences at the station level: bluish colours indicate an improvement while reddish colours indicate a degradation with respect to the reference (the raw 50/0 operational ensemble). Large CRPS differences (deep blue or deep red) are not always associated with statistical significance, because the mean difference may be affected by outliers.

Top panels in Figure 10 show the differences between the raw 50/0 and 40/40 configurations. From visual inspection, we do not see any specific pattern in the differences between both raw combinations at day +1. At day +10, local improvements and degradations of the performance are more balanced and some pattern is observed: the raw 50/0 ensemble outperforms the 40/40 combination over coastal and mountainous areas (as for example along the Atlantic coast and over the Alps) while the 40/40 combination improves the forecast significantly over continental areas over Northern and Central Europe.

Middle and bottom panels in Figures 10 show the differences between the raw operational forecast and both types of calibration, QM and QM+W, respectively. The dominant blue colours indicate that post-processing improves the skill of the forecast. At day 1, an improvement with statistical significance is registered over the whole area of study. At day +10 large areas with degradation are observed mostly over Eastern Europe with QM (Figure 10d) and mostly over Western Europe with QM+W (Figure 10f). At day +10, the positive impact of post-processing is identified over coastal areas in Northern Europe and mountainous areas over Central Europe, matching the areas where the 40/40

combination shows less skill than the operational ensemble.

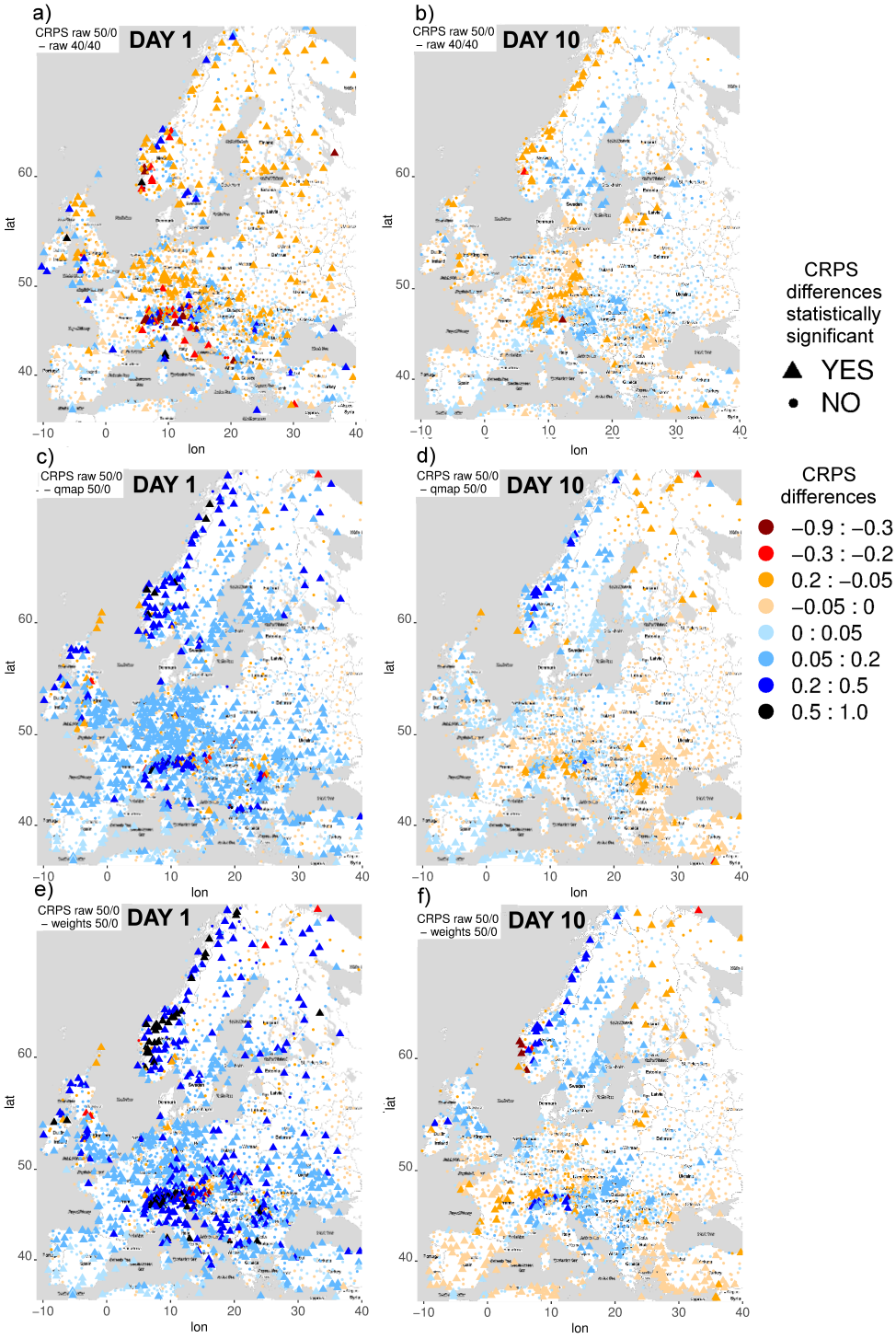


**FIGURE 10** Spatial distribution of mean CRPS differences for lead times of day +1 (left) and day +10 (right). The first row presents the mean CRPS of raw 50/0 minus the CRPS of the raw 40/40 combination. The second row presents the mean CRPS of the raw 50/0 compared to QM 50/0. The third row presents the mean CRPS of the raw 50/0 compared to QM+W 50/0. Large triangles indicate stations that are statistically significant different and smaller dots indicate stations that are not.

# 4 | CONCLUSION

This paper explores the skill and reliability of probabilistic quantitative precipitation forecasts (PQPFs) over Europe for various dual-resolution ensemble combinations. The evaluation is performed for raw ensemble forecasts but also for statistically post-processed forecasts with (a) quantile mapping and (b) quantile mapping combined with an objective weighting of the sorted ensemble members. Five different combinations of *HH* higher-resolution members and *LL* lower-resolution, which have equal computational cost, are tested. The intent is to determine: (a) whether combinations of lower- and higher-resolution ensemble provide improved PQPFs with respect to a single-resolution ensemble, and (b) whether the optimal combination was notably different after post-processing. This paper, which focuses on 24 h precipitation, complements other studies on the probabilistic skill of dual-resolution ensemble forecasts (Leutbecher and Ben Bouallègue, in preparation) and the statistical post-processing of 2 m temperature dual-resolution ensemble forecasts (Baran et al., 2019).

The post-processing methodology applied here follows Hamill and Scheuerer (2018): a quantile mapping with the use of supplemental locations to increase the training sample size. In addition, closest-member histogram statistics is used for an objective re-weighting of a sorted ensemble members PQPF. The methodology as applied here provided some novel aspects. In particular, training data are supplied by reforecasts. Since the reforecasts have a different ensemble size (11 members) than the real-time forecasts considered, e.g., 51 higher-resolution members, closest-member histogram statistics from the 11-member reforecasts cannot be directly used for the objective re-weighting of the 51-member real-time ensemble. This problem is addressed by fitting a Beta distribution to the closest-member histogram.

Regarding the impact of post-processing, verification results reveal similar conclusions to previous studies. As Hamill and Scheuerer (2018) concluded, the primary ensemble forecast deficiency corrected by quantile mapping is the over-prediction of light precipitation amounts, especially at very short lead times. On the other hand, the primary deficiency of forecasts of heavier amounts is overconfidence and it is addressed through the closest-histogram rank weighting. Reliability is improved at all lead times and precipitation thresholds, in particular at short lead times and low thresholds. However, and this is a new result, forecast resolution is decreased by the calibration process at longer

lead times. This could be explained by a suboptimal choice of supplemental locations in some mountainous or low precipitation areas because of the absence of grid points with similar orographic characteristics and/or precipitation climatology. In addition, the weighting step might undo some of the benefits of the quantile mapping. Whether this is an actual issue for the longer lead times remains to be investigated in future work.

Moreover, post-processing is applied to higher and the lower resolution ensembles separately. Dual-resolution ensemble would further benefit from an optimal combination of (calibrated) ensembles, which can be achieved following for example Ben Bouallègue et al. (2019).

Regarding the dual-resolution ensemble performance, the best dual-resolution ensemble, among the ones tested in this study, is a balance between both resolution ensembles, namely the 40/40 combination. At short lead times, the second best is the ensemble with 50 higher-resolution members which corresponds to the current operational configuration. At longer lead times, the difference in performance is small between ensemble with large number of members. These results are in line with the findings in Leutbecher and Ben Bouallègue (in preparation) but illustrate how the positive impact of higher-resolution members vanishes more rapidly in the case of precipitation forecasts than in the case of 2 m temperature forecasts. The interpretation of these results is that higher-resolution forecasts are more valuable at short lead times where predictable features are better resolved with the higher-resolution system. At longer lead times, the predictability of the small scale features is lost and sampling error dominates, which favour larger ensembles.

Regarding the dual-resolution ensemble performance after post-processing, results presented in this paper confirm the conclusion of Baran et al. (2019). Post-processing techniques, in particular quantile mapping combined with a member weighting, strongly reduces the differences in the skill between all the dual ensemble configurations. This could imply that the choice of the ensemble configuration, that is the balance between horizontal resolution/ensemble size, might be less important for users making decision based on calibrated forecasts than for the ones using raw forecasts.

The evaluation presented in this paper provides some guidance on the skill of different ensemble configurations. However, a multitude of applications and other skill-related considerations, together with technical and practical aspects all need to be taken into account when deciding on any operational configuration. Future work will address these other considerations. As we discussed at the beginning of the article, this will include the evaluation of the

different dual-resolution ensemble configurations in the EFAS hydrological model and exploring the benefits of each calibration process in small and large catchments for different seasons (summer and autumn)

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support from the European Union Research and Innovation Programme Horizon 2020 IMPREX project (Grant Agreement No. 641811). Funding for T. Hamill was provided both by ESRL Physical Sciences Division base funding and by a funding from the US NWS Office of Science and Technology Integration through the Meteorological Development Lab, project number T8MWQML.P00. Thanks to Sándor Baran for his valuable help in the application and interpretation of the Diebold-Mariano test.

## REFERENCES

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

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

Ben Bouallègue, Z. (2013) Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Wea. Forecasting*, **28**, 515–524.

Ben Bouallègue, Z., Ferro, C. A. T., Leutbecher, M. and Richardson, D. S. (2019) Predictive verification for the design of multi-model ensembles. *Tellus A. (in review)*.

Brier, G. W. (1950) Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Buizza, R. (2010) Horizontal resolution impact on short-and long-range forecast error. *Q. J. R. Meteorol. Soc.*, **136**, 1020–1035.

Buizza, R. and Leutbecher, M. (2015) The forecast skill horizon. *Q. J. R. Meteor. Soc.*, **141**, 3366–3382.

Buizza, R. and Palmer, T. N. (1998) Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503–2518.

Buizza, R. (2018) Chapter 2 - Ensemble Forecasting and the Need for Calibration. In *Statistical Postprocessing of Ensemble Forecasts* (eds. S. Vannitsem, D. S. Wilks and J. W. Messner), 15–48. Elsevier.

Diebold, F. X. and Mariano, R. S. (2002) Comparing predictive accuracy. *Journal of Business & economic statistics*, **20**, 134–144.

Fortin, V., Favre, A.-C. and Saïd, M. (2006) Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Q. J. R. Meteorol. Soc*, **132**, 1349–1369.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2014) Probabilistic forecast, calibration and sharpness. *J. Roy. Stat. Soc. B*, **69**, 243–268.

Hagedorn, R. (2008) Using the ECMWF reforecast dataset to calibrate EPS forecasts. In *ECMWF Newsletter 117*, 8–13.

Hagedorn, R., Hamill, T. M. and Whitaker, J. S. (2008) Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619.

Haiden, T., Janousek, M., Bidlot, J.-R., Buizza, R., Ferranti, L., Prates, F. and Vitart, F. (2018) Evaluation of ecmwf forecasts, including the 2018 upgrade. *ECMWF Technical Memorandum 831*, 243–268.

Hamill, T. M. (2012) Verification of TIGGE Multimodel and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous United States. *Mon. Wea. Rev*, **140**, 2232–2252.

Hamill, T. M., Bates, G., Whitaker, J. S., Murray, D. R., Fiorino, M., Jr., T. J. G., Zhu, Y. and Lapenta, W. (2013) NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565.

Hamill, T. M. and Colucci, S. J. (1998) Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724.

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. Wea. Rev.*, **145**, 3441–3463.

Hamill, T. M., Hagedorn, R. and Whitaker, J. S. (2008) Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation. *Mon. Wea Review*, **136**, 2620–2632.

Hamill, T. M. and Scheuerer, M. (2018) Probabilistic Precipitation Forecast Postprocessing Using Quantile Mapping and Rank-Weighted Best-Member Dressing. *Mon. Wea. Rev.*, **146**, 4079–4098.

Hamill, T. M. and Whitaker, J. S. (2007) Ensemble Calibration of 500-hPa Geopotential Height and 850-hPa and 2-m Temperatures Using Reforecasts. *Mon. Wea. Rev.*, **135**, 3273–3280.

Hamill, T. M., Whitaker, J. S. and Mullen, S. L. (2006) Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33–46.

Hamill, T. M., Whitaker, J. S. and Wei, X. (2004) Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447.

Hamill, T. M. (2018) Practical Aspects of Statistical Postprocessing. In *Statistical Postprocessing of Ensemble Forecasts* (eds. S. Vannitsem, D. S. Wilks and J. W. Messner). Elsevier.

Hopson, T. M. and Webster, P. J. (2010) A 1–10-day ensemble forecasting scheme for the major river basins of bangladesh: Forecasting severe floods of 2003–07. *J. Hydrometeor.*, **11**, 618–641.

Lei, L. and Whitaker, J. S. (2017) Evaluating the trade-offs between ensemble size and ensemble resolution in an ensemble-variational data assimilation system. *Journal of Advances in Modeling Earth Systems*, **9**, 781–789.

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

Leutbecher, M. (2018) Ensemble size: How suboptimal is less than infinity? *Q. J. R. Meteorol. Soc.* doi:10.1002/qj.3387.

Leutbecher, M. and Ben Bouallègue, Z. (in preparation) On the probabilistic skill of dual-resolution ensemble forecasts.

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

Matheson, J. E. and Winkler, R. (1976) Scoring rules for continuous probability distributions. *Manage. Sci.*, **22**, 1087–1095.

Murphy, A. H. (1973) A new vector partition of the probability score. *J. Appl. Meteor*, **12**, 595–600.

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. *Tech. Rep. JRC86388*, Joint Research Centre, EU, Ispra (VA), Italy.

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

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**, 2473–2489.

514 Roulston, M. S. and Smith, L. A. (2003) Combining dynamical and statistical ensembles. *Tellus A: Dynamic Meteorology and*
515 *Oceanography*, **55**, 16–30.

516 Shepard, D. (1968) A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM*
517 *national conference*, 517–524. ACM, New York.

518 Stensrud, D. J. and Yussouf, N. (2003) Short-Range Ensemble Predictions of 2-m Temperature and Dewpoint Temperature over
519 New England. *Mon. Wea. Rev.*, **131**, 2510–2524.

520 Unger, D. A. (1985) A method to estimate the continuous ranked probability score. Preprints. In *Proceedings of the Ninth Con-*
521 *ference on Probability and Statistics in Atmospheric Sciences*, 206–213. American Meteorological Society, Boston, USA.

522 Van den Hurk, B., Bouwer, L. M., Buontempo, C., Döscher, R., Ercin, E., Hananel, C., Hunink, J. E., Kjellström, E., Klein, B., Manez,
523 M., Pappenberger, F., Pouget, L., Ramos, M. H., Ward, P. J., Weerts, A. H. and Wijngaard, J. B. (2016) Improving predictions
524 and management of hydrological extremes through climate services. climate services. *Climate Services*, 6–11.

525 Vitart, F., Balsamo, G., Bidlot, J.-R., Lang, S., Tsonevsky, I., Richardson, D. and Alonso-Balmaseda, M. (2019) Use of era5 to
526 initialize ensemble re-forecasts. *ECMWF Technical Memorandum*.

527 Wilks, D. S. (2011) *Statistical methods in the atmospheric sciences (3rd Ed)*. Academic Press.

528 Wilks, D. S. and Hamill, T. M. (2007) Comparison of Ensemble-MOS Methods Using GFS Reforecasts. *Mon. Wea. Rev.*, **135**,
529 2379–2390.

530 Willmott, C., Rowe, C. and Philpot, W. (1985) Small-scale climate maps: A sensitivity analysis of some common assumptions
531 associated with grid-point interpolation and contouring. *The American Carthographer*, **12**, 5–16.

532 Yussouf, N. and Stensrud, D. J. (2007) Bias-corrected short-range ensemble forecasts of near-surface variables during the
533 2005/06 cool season. *Weather and Forecasting*, **22**, 1274–1286.