How to create an operational multi-model of seasonal forecasts?

Stephan Hemri[1] · Jonas Bhend[1] · Mark A. Liniger[1] · Rodrigo Manzanas[2] · Stefan Siegert[3] · David B. Stephenson[3] · José M. Gutiérrez[4] · Anca Brookshaw[5] · Francisco J. Doblas-Reyes[6,7]

Abstract

Seasonal forecasts of variables like near-surface temperature or precipitation are becoming increasingly important for a wide range of stakeholders. Due to the many possibilities of recalibrating, combining, and verifying ensemble forecasts, there are ambiguities of which methods are most suitable. To address this we compare approaches how to process and verify multi-model seasonal forecasts based on a scientific assessment performed within the framework of the EU Copernicus Climate Change Service (C3S) Quality Assurance for Multi-model Seasonal Forecast Products (QA4Seas) contract C3S 51 lot 3. Our results underpin the importance of processing raw ensemble forecasts differently depending on the final forecast product needed. While ensemble forecasts benefit a lot from bias correction using climate conserving recalibration, this is not the case for the intrinsically bias adjusted multi-category probability forecasts. The same applies for multi-model combination. In this paper, we apply simple, but effective, approaches for multi-model combination of both forecast formats. Further, based on existing literature we recommend to use proper scoring rules like a sample version of the continuous ranked probability score and the ranked probability score for the verification of ensemble forecasts and multi-category probability forecasts, respectively. For a detailed global visualization of calibration as well as bias and dispersion errors, using the Chi-square decomposition of rank histograms proved to be appropriate for the analysis performed within QA4Seas.

✉ Stephan Hemri
stephan.hemri@meteoswiss.ch

[1] Federal Office of Meteorology and Climatology MeteoSwiss, Zurich-Airport, Switzerland

[2] Meteorology Group, Dpto. de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Santander, Spain

[3] University of Exeter, Exeter, UK

[4] Meteorology Group, Instituto de Física de Cantabria (CSIC-Universidad de Cantabria), Santander, Spain

[5] European Centre for Medium-Range Weather Forecasts, Reading, UK

[6] ICREA, Pg. Lluis Companys, Barcelona, Spain

[7] Earth Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain

1 Introduction

Seasonal forecasts of atmospheric variables like near-surface temperature or precipitation are becoming increasingly important for a wide range of stakeholders in fields like agriculture (Ouédraogo et al. 2015; Ramírez-Rodrigues et al. 2016; Roudier et al. 2016; Rodriguez et al. 2018), hydrology (Demirel et al. 2015; Yuan et al. 2015a, b), or wind energy production (Alonzo et al. 2017; Clark et al. 2017; Torralba et al. 2017). The EU Copernicus Climate Change Service (C3S) Quality Assurance for Multi-model Seasonal Forecast Products (QA4Seas) contract C3S 51 lot 3 aimed at *developing a strategy for the evaluation and quality control (EQC) of the multi-model seasonal forecasts provided by the C3S* (see also Barcelona Supercomputing Center 2018). An in-depth scientific assessment of the seasonal forecast products was one of the core activities performed in QA4Seas. The first part of this assessment focused on the comparison of different bias adjustment and ensemble recalibration approaches (Manzanas et al. 2019, 2020). Following up

these results, we discuss strategies for multi-model combination and verification.

Typically, multi-model combination of seasonal ensemble forecasts leads to a forecast skill, which is greater than the one of the best single forecast system. Besides error compensation, multi-model combination also improves consistency and reliability (Hagedorn et al. 2005). The effects of multi-model combination and single model recalibration on forecast skill are comparable (Doblas-Reyes et al. 2005; Weigel et al. 2009). However, multi-models tend to benefit less from additional recalibration than single models. Further, the effects of recalibration and multi-model combination vary strongly among geographical areas, variables, and the forecast models considered. Accordingly, each multi-model forecast system needs to be assessed separately (Doblas-Reyes et al. 2005). A recent study by Mishra et al. (2018) analyses different multi-model combination approaches for the European Multimodel Seasonal to Interannual Prediction (EUROSIP, Vitart et al. 2007; Stockdale 2013) system, which is composed of the Met Office GloSea5 model, the European Centre for Medium Range Weather Forecasts (ECMWF) SEAS4 model, the National Centers for Environmental Prediction (NCEP) System 2 model, and the Météo France System 5 model. Focusing on seasonal temperature and precipitation predictions for the European region and a verification period from 1992 to 2012 their results indicate that a simple equally weighted multi-model on average outperforms two different unequally weighted multi-models. But, as the best multi-model does not always outperform all single model predictions, they recommend to assess predictions provided by both the single models and the multi-model combination. The high geographical variability in relative forecast skill is in line with Kharin et al. (2017) who performed Monte Carlo analyses which indicated that sophisticated processing of seasonal forecasts is impaired by the small sample size of less than 30 years of hindcasts. Further, Baker et al. (2018) assessed both the skill and the dispersion errors, i.e. too small or too large ensemble spread, for the wintertime North Atlantic Oscillation (NAO) of the EUROSIP multi-model predictions. While they could reveal substantial forecast skill for the NAO, they emphasize also the dispersion errors of the seasonal forecast models.

Keeping in mind that there is no general consensus on how to assess and post-process seasonal multi-model ensemble predictions, we compare and discuss different post-processing procedures based on the three forecast systems that have been available through C3S at the time we have prepared this study. From Manzanas et al. (2019) we know that both simple bias adjustment approaches like mean variance rescaling (MVA, Doblas-Reyes et al. 2005; Torralba et al. 2017) and more sophisticated recalibration approaches like climate conserving recalibration (CCR, Doblas-Reyes et al. 2005; Weigel et al. 2009) are able to correct the biases of

the single models and lead to an increase in forecast skill. In terms of the continuous ranked probability score (CRPS, Hersbach 2000; Gneiting and Raftery 2007), recalibration approaches are able to put their ability to specifically improve reliability to good use, but outperform bias adjustment methods only in a few particular regions and seasons with high predictability. However, both bias adjustment and recalibration degrade measures that focus on multi-category probability forecasts like the ranked probability score (RPS, Epstein 1969; Murphy 1969, 1971). While the different recalibration approaches proved to perform similarly, it turned out that MVA outperformed the other bias adjustment methods. Accordingly, we will focus on MVA and CCR bias adjusted predictions for the following assessment of multi-model combination. For the sake of brevity we will use the term bias adjustment also for CCR, hereafter.

In the following, we will make suggestions on how to apply multi-model combination to both ensemble and multi-category probability seasonal predictions. To this end different multi-model combination approaches are tested. The different forecast (products) are then verified based on the skill metrics CRPS and RPS, respectively. Further, we follow the paradigm of probabilistic forecasting to maximize sharpness subject to calibration (Gneiting et al. 2005; Gneiting and Raftery 2007), which states that a probabilistic forecast should be as focused as possible, i.e. the narrower the prediction intervals the better, as long as it is well calibrated in the sense that the observations cannot be distinguished from random samples from the predictive distributions. Here, we assess calibration and sharpness in more detail based on rank histograms and the widths of prediction intervals, respectively. For further insights into the effects of bias correction and multi-model combination we apply χ^2 decompositions of the rank histograms (Jollife and Primo 2008) of the different forecast models.

Section 2 presents the data and methods used for this study. This comprises also suggestions on how to generate forecast products and how to verify them. The plausibility of the suggestions is underpinned by the exemplary results from the quality assessment of near-surface temperature seasonal predictions performed within QA4Seas in Sect. 3 and are discussed in Sect. 4. Conclusions are provided in Sect. 5.

2 Data and methods

2.1 Seasonal forecasts and reference observations

We analyse seasonal forecasts from three forecasting systems available in the C3S Climate Data Store (CDS) at the time of the QA4Seas project: ECMWF SEAS5 (Johnson et al. 2019), Météo France System 5 (Batté et al. 2018), and Met Office GloSea5 (MacLachlan et al. 2015). Hindcasts for

monthly mean quantities are available on a common global $1° \times 1°$ grid. The common period for which hindcasts are available from all three forecasting systems is 1993–2014 (1994–2014 for the January initialization). While additional hindcasts are available for the individual models, we use the common hindcast period for calibration and verification of all forecasting systems. Note that such a data homogenization comes at the cost of potentially misinterpreting the skill of (recalibrated) forecasts due to the short training and verification periods. ECMWF ERA-Interim analysis data are used as reference for both training and verification.

The analyses presented here focus on seasonal averages of near-surface temperature on the global grid for the four typical boreal seasons DJF (winter), MAM (spring), JJA (summer), SON (autumn). We use forecasts available one month before the seasons of interest, i.e. those available at the beginning of February, May, August, and November, respectively. This corresponds to a lead time of one month. Within the framework of QA4Seas, the analyses have been complemented by analyses of ensemble forecasts of sea surface temperature and precipitation. In order to keep this paper short and because the results do not differ much among the different variables, results for these two variables are available as supplemental material.

2.2 Multi-model combination

Prior to summarizing suggestions on how to generate and verify seasonal forecast products (see Sect. 2.5), we introduce the multi-model combination methods used for this study. Most of these methods can be applied independently from bias adjustment. In general we have tested both MVA and CCR for bias adjustment. However, one of the multi-model combination approaches cannot work with MVA by construction. Refer to Manzanas et al. (2019) for details on MVA and CCR. While there is convincing evidence that multi-model forecasts outperform single-model forecasts on average (Doblas-Reyes et al. 2005; Hagedorn et al. 2005; Weigel et al. 2009), how to best form the multi-model forecast is still a matter of debate. If large training samples were available, training skill based weighting of model systems

would produce the best-performing multi-model forecasts. Methods that apply such a weighting like ensemble model output statistics (EMOS; Gneiting et al. 2005) or Bayesian model averaging (BMA; Raftery et al. 2005) are used in medium-range forecasting. In seasonal forecasting, however, it is not clear if skill based weighting increases forecast quality of multi-model forecasts (Weigel et al. 2010; DelSole et al. 2013) due to the relatively small sample size to estimate model weights. Despite its expected poor performance for seasonal forecasts, we have tested EMOS within the framework of QA4Seas, since it is still a rather simple method that requires only a few coefficients to be estimated. Alternatively, weighting based on similarity of model errors is being used for climate change projections (Knutti et al. 2017; Sanderson et al. 2017). Whether such weighting or a combination with the above is beneficial in the case of seasonal forecasting has yet to be explored. Here we assess a cascade of combination methods, which are summarized in Table 1, with increasing complexity:

– Pooling of first n-members (MFN): This approach consists of pooling the first n-members of calibrated single-model forecasts. In our case, we use the first 15 members of each model to give each individual forecasting system equal weight in the multi-model forecast. Technically, the following steps are performed:

1. Let n be the ensemble size of the smallest single model ensemble.
2. Select members $1, \ldots, n$ from each ensemble, which for the lagged GloSea5 Météo France System 5 ensembles should correspond to the most recent model runs. For burst ensembles, which consist of ensemble members that have been initialized simultaneously and are run in parallel, statistical exchangeability between the individuals members is typically fulfilled. Hence, in principle any subset of size n could be selected.
3. Pool the just selected members to a multi-model ensemble of size $r \times n$, where r is the number of ensemble models considered.

Table 1 Overview over the different multi-model combination approaches

Method	forecast type	parametric?	weighting?	bias adjustment
MFN	ensemble	no	no	MVA or CCR
MDE	ensemble	yes	no	MVA or CCR
MDR	ensemble	yes	yes	MVA or CCR
MDM	ensemble	yes	yes	CCR
EMOS	ensemble	yes	yes	EMOS
MPE	categorical	no	no	implicit
MPR	categorical	no	yes	implicit

The cell colors refer to the corresponding steps in the processing chain as defined by the flowchart in Fig. 1. Note that EMOS performs both multi-model combination and bias adjustment at the same time

– Equidistant quantiles from multi-model PDF (MDE): Multi-model ensemble members are drawn as equidistant quantiles from the parametric multi-model probability density function (PDF). Technically:

1. Obtain the first two moments (mean and variance) of each single model ensemble.
2. Use these moments to generate a normal mixture distribution. For this study this has been done using the R package `nor1mix` (Mächler 2017).
3. Obtain the multi-model ensemble by selecting equidistant quantiles, i.e. the $\frac{0.5}{m}, \frac{1.5}{m}, \ldots, \frac{m-0.5}{m}$ quantiles, where m is the desired size of the multi-model ensemble, from the normal mixture distribution.

– Equidistant quantiles from multi-model PDF with skill-based weighting (MDR): As in the above approach, but instead the multi-model PDF is a weighted average of the single-model PDFs with weights proportional to the inverse of the cross-validated RMSE of the individual calibrated single-model forecasts. Technically:

1. Obtain the first two moments (mean and variance) of each single model ensemble.
2. Obtain weights for each model that are proportional to the inverse of the cross-validated RMSE of the single models' forecast means.
3. Use the estimates of the moments and model weights to generate a weighted normal mixture distribution (also done using the R package `nor1mix`).
4. Select equidistant quantiles from the weighted normal mixture distribution as in step 3 of MDE.

– Equidistant quantiles from multi-model PDF with ensemble mean difference based weighting (MDM): As in the above approach, but instead the multi-model PDF is a weighted average of the single-model PDFs with weights proportional to the inverse ensemble mean difference of the individual calibrated single-model forecasts. Ensemble mean difference, which has been introduced by Scheuerer (2014) in an ensemble forecasting context, is a more robust measure of ensemble spread than the variance and is computed as

$$MD(\boldsymbol{x}) = \frac{1}{Q^2} \sum_{q,q'} |x_q - x_q'|, \qquad (1)$$

where x_q, $q = 1, \ldots, Q$, denote the ensemble members of an ensemble of size Q. Note that this method works only with input ensembles that have been corrected such that a strict spread-skill dependence can be ensured. Otherwise, it cannot generally be assumed that ensemble forecasts with a smaller spread perform better than those with a larger spread. In the scientific assessment at hand, strict spread-skill dependence can only be ensured for CCR recalibrated single-model forecasts. The direct link between ensemble spread and weight in the multi-model allows for a dynamical weighting that changes from initialization date to initialization date. This is generally not the case for MDR.

– EMOS, which simultaneously performs bias adjustment and multi-model combination: For EMOS, we use non-homogeneous Gaussian regression as introduced in Gneiting et al. (2005). The mean parameter depends linearly on the raw ensemble means $\bar{x}_1, \ldots, \bar{x}_r$ via

$$\mu_{\mathrm{emos}} = a_0 + a_1 \bar{x}_1 + \cdots + a_r \bar{x}_r, \qquad (2)$$

and the variance depends on the variance s^2 of all members of a pooled raw ensemble, which in a multi-model combination context consists of the ensemble members from all forecast systems considered, through

$$\sigma_{\mathrm{emos}}^2 = b_0 + b_1 s^2. \qquad (3)$$

The coefficients of the EMOS model have been obtained by minimum CRPS estimation over cross-validated leave-1-year-out training periods.

Further, there are two multi-models that provide only multi-category probability forecasts:

– Unweighted average of forecast category probabilities from the raw ensembles (MPE).
– Weighted average of forecast category probabilities (MPR). The weights are proportional to the inverse of the ranked probability scores (RPS) of tercile forecasts over the training period.

2.3 Metrics of forecast skill

As stated above, for ensemble forecasts of continuous variables like near-surface temperature we use the CRPS to measure forecast skill. For (multi-)category probability forecasts such as tercile forecasts, we use the RPS. We made this choice, since we want to focus on the verification of probabilistic forecasts and not deterministic ensemble statistics thereof like the ensemble mean. Hence, we omit deterministic verification measures like correlation or RMSE. CRPS and RPS values are calculated using the R package `SpecsVerification` (Siegert 2017a). A sample version of the CRPS, i.e. a version that can be used to approximate the CRPS from random samples like an ensemble forecast, can be derived from the more general

formula for the energy score provided by Gneiting et al. (2008). Accordingly, the CRPS can be written as

$$crps(\mathbf{x}, y) = \frac{1}{m} \sum_{i=1}^{m} |x_i - y| - \frac{1}{2m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} |x_i - x_j|, \quad (4)$$

where x_i denotes member i of an ensemble of size m and y is the verifying observation. Similarly, the RPS of multi-category probability forecasts is given by

$$rps(\mathbf{f}, \mathbf{o}) = \frac{1}{K-1} \sum_{k=1}^{K} (f_k - o_k)^2, \quad (5)$$

where f_k and o_k denote the kth component of the cumulative forecast and observation probability vectors.

We use climatology as reference forecast to compute skill scores CRPSS and RPSS. Climatology refers to the grid point wise leave-one-year-out cross-validated ERA-Interim analysis near-surface temperature values for the season of interest. The (C)RPSS relative to climatology is computed as

$$(C)RPSS = 1 - \frac{(C)RPS_{forc}}{(C)RPS_{clim}}, \quad (6)$$

where $(C)RPS_{forc}$ and $(C)RPS_{clim}$ denote the grid point wise mean (C)RPS of the forecasts of interest and the reference climatology, respectively, for a given season and initialization month over the entire verification period.

Significance of the difference in area-weighted global mean CRPS and RPS between the different hindcast variants and the corresponding reference predictions (climatological forecasts for single ensemble hindcasts and ECMWF SEAS5 hindcasts for multi-model combinations, respectively) are computed using a paired t test, where each sample corresponds to the difference of the area-weighted global mean scores averaged over the considered initializations of the two predictions for a particular year. As we are averaging over a large number of grid points, these differences follow approximately a normal distribution. Further, differences in yearly averages of scores are not expected to exhibit a strong serial correlation. Hence, it is appropriate to apply a paired t test. In order to control for multiple testing, we apply a Bonferroni correction for each group of similar tests. Such a group consists of all tests with the same reference model, the same bias adjustment method, and the same lead time. Significance of the difference in CRPS or RPS compared to climatological forecasts on a grid point level is computed using two-sided Diebold–Mariano tests (Diebold and Mariano 1995) with a significance level of 0.05 for all forecast variants. In order to avoid erroneous interpretation of random clusters of locally significant grid points in the CRPSS and RPSS maps under spatial correlation, we have applied false discovery rate (FDR) correction that provides an upper boundary to the fraction of type 1 errors as described in

Wilks (2016). For further details on FDR we refer to Benjamini and Hochberg (1995) and Ventura et al. (2004). Note that statistical significance could also be computed using bootstrapping approaches. However, for the sake of a computationally efficient analysis on the entire global grid, we prefer applying Diebold–Mariano tests with FDR correction. Similarly, we apply FDR correction in the same way to the χ^2 goodness of fit (GOF) tests which we use to evaluate reliability, bias, and dispersion errors as described in Sect. 2.4.

2.4 Evaluating reliability and sharpness

For an in-depth analysis of the reliability we rely on the decomposition of the χ^2 GOF test statistic for uniform rank histograms that has been introduced by Jollife and Primo (2008). While the χ^2 GOF test assesses only whether the rank histogram resembles that of a random sample from a discrete uniform distribution (Elmore 2005), the decomposition of the χ^2 test statistic allows to test for particular violations of uniformity that arise, for instance, from biases or dispersion errors. For this study, we select a linear contrast to test for additive biases and a u-shape contrast to test for under- or overdispersion. Hence, the test statistic T_{full} is decomposed as follows

$$T_{full} = T_{linear} + T_{u\_shape} + T_{resid}, \quad (7)$$

where T_{linear}, T_{u_shape}, and T_{resid} denote the contributions of the linear, the u-shape, and a residual term to T_{full}.

Following Jollife and Primo (2008) the χ^2 GOF test statistics can be written as

$$T_{full} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} = \sum_{r=1}^{k} \left(\sum_{i=1}^{k} l_{ri} \frac{n_i - e_i}{\sqrt{e_i}} \right)^2 = \sum_{r=1}^{k} u_r^2, \quad (8)$$

where n_i are the observed numbers in class i and e_i are the corresponding expected numbers, while l_{ri} are the elements of a $k \times k$ matrix L. Each row of L can be interpreted as a contrast vector, which should be orthonormal to each other. Technically, the linear contrast is a linearly increasing vector. The u-shape, or dispersion contrast, is represented by a vector with minimum values in the center and quadratically increasing values towards both ends. The contrast vectors are adjusted such that their mean is zero. Details on the construction of contrast vectors can be found in Jollife and Primo (2008). In practice, one needs to specify only the $j < k$ first contrasts of interest and calculates only the u_r^2 for $r = 1, \ldots, j$ test statistics and T_{resid} is obtained by subtraction according to Eq. (7). Further, the sign of u_r reflects the direction of a contrast. For the linear contrast, a negative value of u_r reflects that most observations are rather low compared to the ensemble, which, for instance, corresponds to a warm bias for temperature or to a wet bias

for precipitation. A positive value would indicate a cold or dry bias, respectively. In case of the u-shaped contrast, a negative value would reflect an overdispersive ensemble. An underdispersive ensemble would lead to a positive value.

In order to obtain a sample size that is large enough to assess reliability, we have pooled the hindcasts from the four initializations considered. Having 22 years of data, this leads to a sample size of almost 90. As the (multi-)model ensembles are too large for direct computation of the ranks, i.e. 45 members, the ranks of the observations are mapped to eight categories leading to a rank histogram with 8 bins. In a small simulation study, using 8 bins has proved to be optimal for detection of biases and/or dispersion errors. For details on the statistical tests we refer again to Jollife and Primo (2008).

Sharpness is evaluated based on the relative mean width of centered 90 % prediction intervals compared to the mean width of the corresponding 90 % prediction intervals of the climatological forecasts. The narrower the mean width, the sharper is the forecast on average. Note that we have chosen quite wide prediction intervals, despite the rather small multi-model ensemble size of 45. The advantage of such an approach is a higher probability to detect regions with particularly heavy tailed multi-model forecasts that may arise from combining ensembles with considerable differences in the forecast mean.

2.5 Choices to be made for forecast product generation and verification

Prior to any analysis or processing of the seasonal forecasts available through the CDS, we want to embed this study into the broader context of choices that have to be made by the user. First, any user has to define the goal of processing

Fig. 1 Flowchart illustrating the generation of seasonal forecast products within the framework of QA4Seas. The box colors refer to the processing steps defined in Sect. 4: **a** aggregation in red; **b** bias correction in cyan; **c** combination in orange
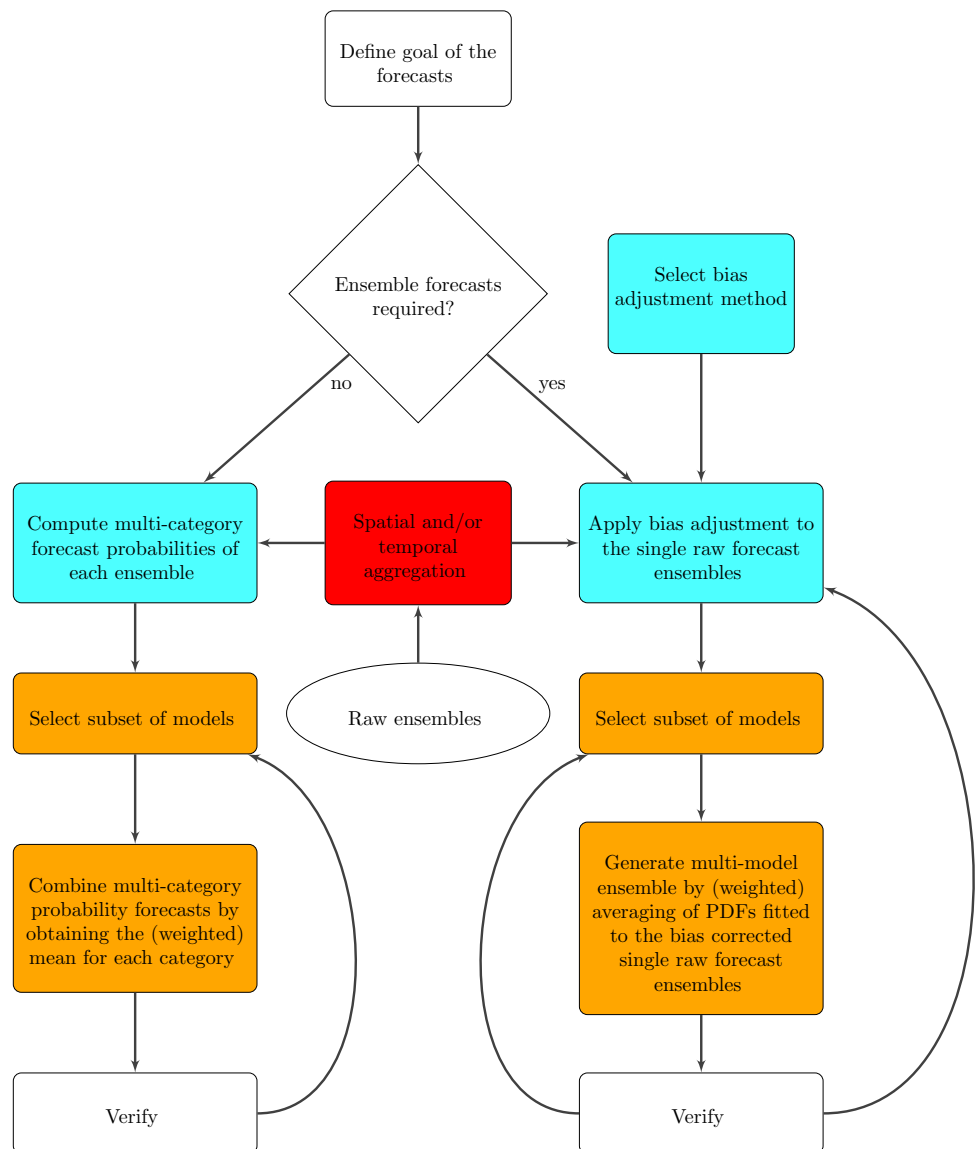
the seasonal raw ensembles. In other words: what kind of forecast products do they need? As shown in Fig. 1, a key question is whether ensemble or multi-category probability forecasts are requested. Depending on the forecast product the requirements on and the effects of bias adjustment may be quite different. While some users might just be happy with forecast anomalies others may need absolute forecasts, and hence, more sophisticated bias adjustment. However, all users will be affected by a lack of reliability, which in our opinion should always be bias adjusted, which refers to step (b) in the discussion in Sect. 4 and Fig. 1.

The discerning reader may realize that the processing chain for multi-category probability forecasts illustrated by Fig. 1 generally does not include any explicit correction of reliability. However, for the seasonal data at hand it turned out that an implicit bias adjustment by computing multi-category forecast probabilities based on thresholds, which correspond to quantiles of the forecast climatologies, leads to reasonably reliable predictions.

A second early choice is related to spatial and/or temporal aggregation or smoothing, which refers to step (a) in the discussion in Sect. 4 and Fig. 1. Usually, product skill is enhanced by aggregating, or smoothing the forecasts as early as possible in the product generation chain or by smoothing the estimates of bias adjustment coefficients (see e.g. Gong et al. 2003; Kharin et al. 2017). However, depending on the specific use case this can be done at any point in the product generation chain.

If seasonal forecasts from different models are available, a further choice suggests itself. Should all models be considered for the generation of a multi-model based product? After looking at the outcome of a verification exercise of a number of forecast systems, a user may favor a multi-model that combines data from just a subset of the forecast systems available. In our opinion, the benchmark multi-model should always be the unweighted multi-model ensemble of bias adjusted forecasts. However, depending on the forecast product needed, more sophisticated multi-model combinations, as presented in this study, or applying bias adjustment after multi-model combination may be favored by the user. Multi-model combination refers to step (c) in the discussion in Sect. 4 and Fig. 1.

Finding verification measures that are suitable for the targeted forecast product is the last, yet very important, choice to be made by the user. The choice of verification measures eventually drives the selection of bias adjustment and multi-model combination approaches.

In the specific setting of QA4Seas, the forecast products comprise either multi-category probability multi-model forecasts or ensemble multi-model forecasts. In this particular case, for the multi-category probability forecasts we suggest to:

1. Compute directly the multi-category forecasts probabilities for each raw ensemble relative to its forecast climatology. This goes in hand with an implicit bias adjustment.
2. Combine the single model multi-category probability forecasts to a multi-model by (weighted) averaging of the forecast category probabilities. Details on multi-model combination can be found in Sect. 2.2.
3. Forecast verification needs then to be performed using measures for categorical forecasts like the RPS.

However, for ensemble forecasts we suggest to:

1. Bias adjust the single raw forecast ensembles.
2. As elaborated in Sect. 2.2, generate a multi-model ensemble either by pooling of the ensemble members or by sampling from a (weighted) mixture distribution.
3. Verify the ensemble forecasts based on measures suitable for ensembles like the CRPS.
4. If the user is not only interested in a comparison of forecast skill, but also in what causes the differences in skill, e.g. differences in reliability, bias, or dispersion, we suggest to use the χ^2 decomposition by Jollife and Primo (2008) summarized in Sect. 2.4.

3 Results

3.1 Skill on common grid

We get a first overview on the predictive performance of the different forecast models and products by calculating mean scores over all grid points. Table 2 depicts the CRPS values for near-surface temperature forecasts averaged over the February, May, August, and November initializations for MVA and CCR bias adjusted hindcasts. In most of the cases the simpler MVA bias adjustment method leads to a slightly better CRPS, while for the Météo France System 5 model, which tends to be particularly underdispersed, CCR considerably outperforms MVA. The results for the raw ensembles are not shown here, because they perform considerably worse than climatology just because of a simple additive bias. For both bias adjustment approaches, the multi-model combination can improve the global CRPS for all forecast months, while MDR and its equally weighted counterpart MDE perform best, as reflected by the CRPS values in Table 2.

The global distribution of CRPSS relative to climatology is very similar among the different post-processed forecasts as depicted by Fig. 2. As a benchmark model we use here ECMWF SEAS5 forecasts with a simple leave-one-year-out cross-validated additive bias correction (A_BC ECMWF SEAS5). Differences in forecast skill between

Table 2 Global mean CRPS of ensemble forecasts for all land grid points averaged over the February, May, August, and November initializations for different debiased (MVA on the left and CCR corrected on the right side) near-surface temperature (T2M) (multi-model) hindcasts verified against ERA-Interim

Bias adjustment	MVA				CCR			
Forecast months	1-3	2-4	3-5	4-6	1-3	2-4	3-5	4-6
Climatology	0.63	0.64	0.63	0.63	0.63	0.64	0.63	0.63
ECMWF SEAS5	**0.56**	**0.61**	0.61	0.61	**0.56**	**0.62**	**0.62**	**0.62**
GloSea5	**0.57**	**0.61**	0.62	0.62	**0.58**	**0.62**	0.62	0.63
Météo France System 5	**0.66**	**0.66**	**0.66**	**0.67**	0.62	**0.65**	0.64	**0.64**
MFN	0.55	0.60	0.60	0.60	**0.55**	**0.60**	**0.61**	**0.61**
MDE	**0.54**	**0.59**	**0.59**	0.60	**0.55**	**0.60**	**0.60**	**0.60**
MDR	**0.54**	**0.59**	**0.59**	**0.59**	**0.55**	**0.60**	**0.60**	**0.60**
MDM	NA	NA	NA	NA	**0.55**	**0.60**	**0.60**	**0.60**
MM EMOS	**0.57**	**0.63**	**0.63**	**0.63**	NA	NA	NA	NA

The cell color scale runs from blue (high CRPS) to red (low CRPS). Note that the EMOS multi-model forecast has been obtained by applying EMOS on the raw ensembles. The global mean scores have been obtained by area weighted averaging. Bold numbers refer to scores that are significantly different from the corresponding scores of the climatological forecasts or ECMWF SEAS5 hindcasts according to a two-sided, Bonferroni corrected, paired t test at a significance level of 0.05, for single ensemble models and multi-model combinations, respectively
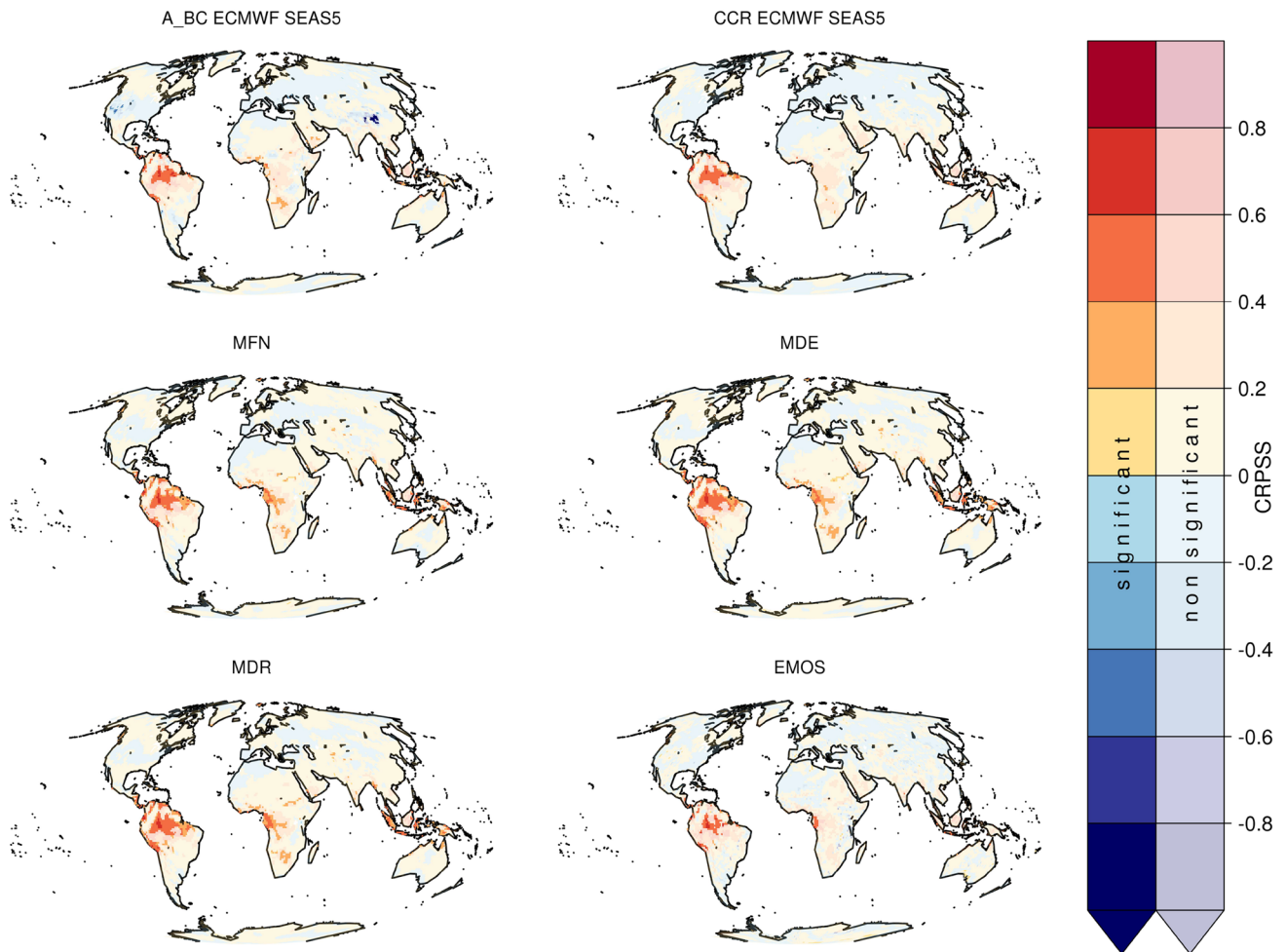

Fig. 2 Maps showing the skill in terms of CRPSS relative to a reference climatology of seasonal T2M ensemble hindcasts initialized in November and valid for DJF. The multi-model combinations are all based on ECMWF SEAS5, GloSea5, and Météo France System 5. The hindcasts are verified over the period from 1993 to 2014 against ERA Interim. Except from A_BC ECMWF SEAS5 all hindcasts shown here are CCR recalibrated. The significance level is 0.05, and false discovery rate (FDR) correction as described in Wilks (2016) has been applied to avoid multiple testing issues
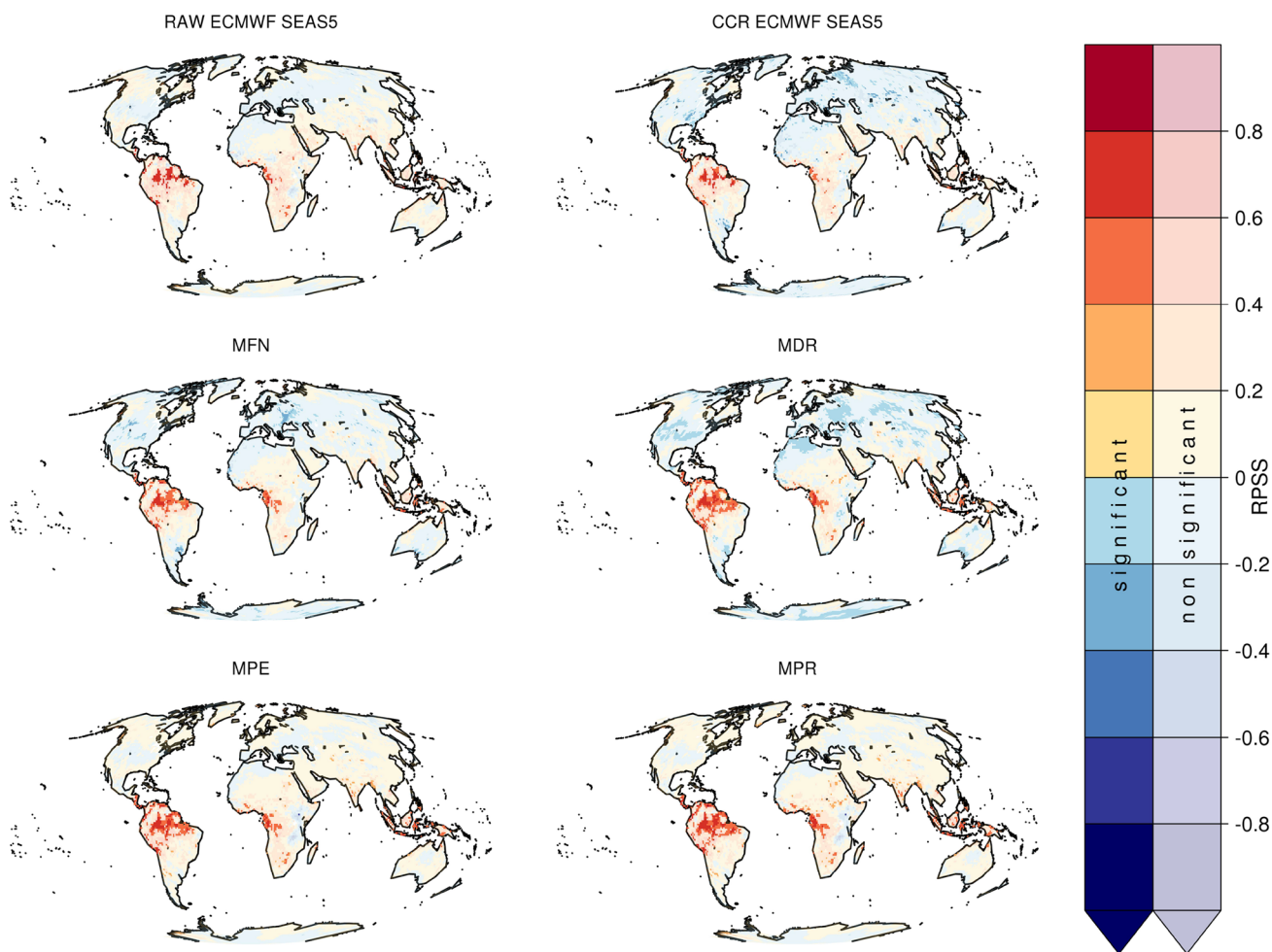
Fig. 3 Maps showing the skill in terms of RPSS relative to a reference climatology of seasonal T2M multi-category (here terciles) probability hindcasts initialized in November and valid for DJF. The multi-model combinations are all based on ECMWF SEAS5, Glo-Sea5, and Météo France System 5. The hindcasts are verified over the period from 1993 to 2014 against ERA Interim. For CCR ECMWF SEAS5, MFN, and MDR CCR recalibration has been applied. Significance level 0.05 and FDR corrected

A_BC ECMWF SEAS5 and CCR ECMWF SEAS5 seem to be marginal. Comparing CCR ECMWF SEAS5 with the multi-model combinations MFN, MDE, and MDR reveals a considerable increase in forecast skill by multi-model combination over West Africa. Additional analyses based on the mean sea surface temperature in NINO regions, which are not shown here, indicated a very small, but significant increase in skill in terms of CRPS when using MDR compared to MDE. The more sophisticated EMOS post-processing method, which applies bias adjustment and multi-model combination simultaneously, performs worse than all other multi-model combination approaches and even worse than the bias adjusted ECMWF SEAS5 or GloSea5 single models. Relative to MDE, EMOS performs worse in many regions. Though the difference in performance is rather small at most of the grid points, the areas in the tropics,

where climatology can be outperformed significantly, are considerably smaller for EMOS compared to MDE.

Let us now assess the effects of multi-model combination on the skill of multi-category probability forecasts. Here, we show the results for the tercile probability hindcasts derived from the ensembles. Figure 3 reveals the effect of the implicit bias adjustment [cf. step (b) in the processing flowchart in Fig. 1] that follows from calculating multi-category probabilities. The uncorrected tercile probability ECMWF SEAS5 hindcasts in the top-left panel outperform their bias adjusted counterpart, i.e. CCR recalibrated ECMWF SEAS5 hindcasts, in terms of RPS. This may be in part due to the use of cross-validation for CCR, which is expected to have a negative effect on the skill of a forecast product, which is implicitly bias adjusted, like tercile forecasts. However, this effect is expected to be rather small, as the tercile

Table 3 Global mean RPS of multi-category probability forecasts (here terciles) for all land grid points averaged over the February, May, August, and November initializations for different T2M (multi-model) hindcasts without any bias correction verified against ERA-Interim

Forecast months	1-3	2-4	3-5	4-6
Climatology	0.43	0.43	0.43	0.43
ECMWF SEAS5	**0.37**	**0.40**	**0.41**	**0.41**
GloSea5	**0.39**	**0.41**	**0.42**	0.42
Météo France System 5	**0.46**	0.46	**0.46**	**0.47**
MPE	0.37	**0.39**	**0.40**	**0.40**
MPR	0.36	**0.39**	**0.40**	**0.40**
MDR	**0.39**	**0.42**	**0.43**	**0.43**

MDR has been added for comparison only and has been bias adjusted using CCR prior to computing the multi-model and then from the multi-model the probability forecast. The cell color scale runs from blue (high RPS) to red (low RPS). The global mean scores have been obtained by area weighted averaging. Bold numbers refer to scores that are significantly different from the corresponding scores of the climatological forecasts or ECMWF SEAS5 hindcasts according to a two-sided, Bonferroni corrected, paired t test at a significance level of 0.05, for single ensemble models and multi-model combinations, respectively

boundaries have also been calculated in cross-validation mode. As for the ensembles, we look now at average scores. As shown in Table 3 the performance of multi-category probability forecasts can hardly be improved by multi-model combination. The good performance of MPE and MPR compared to MDR confirms that direct combination of the single model categorical probability forecasts is to be preferred over first obtaining a bias adjusted ensemble multi-model, here MDR, and then calculating categorical probability forecasts from that ensemble. Comparing the spatial distribution of RPSS relative to climatology of MPR and MDR in Fig. 3 reveals that MPR does not outperform MDR in areas with considerable forecasts skill, whereas areas with poor forecast skill in the extratropics clearly benefit from applying MPR instead of MDR.

Forecast months	1-3	2-4	3-5	4-6
Climatology	0.43	0.43	0.43	0.43
ECMWF SEAS5	**0.37**	**0.40**	**0.41**	**0.41**
GloSea5	**0.39**	**0.41**	**0.42**	0.42
Météo France System 5	**0.46**	0.46	**0.46**	**0.47**
MPE	0.37	**0.39**	**0.40**	0.40
MPR	0.36	**0.39**	**0.40**	0.40
MDR	**0.39**	**0.42**	**0.43**	**0.43**

3.2 Assessment of reliability, bias, and dispersion errors

However, for a sound assessment of the quality of the forecast products the causes of the differences in (skill) scores need to be revealed. To this end, we apply the χ^2 decomposition of the GOF test for flat rank histograms described in Sect. 2.4 to each grid point individually.

As depicted by Fig. 4 the additive bias corrected A_BC ECMWF SEAS5 model, which serves as a benchmark, shows dispersion errors in the rank histograms in some regions, in particular in the tropics. This implies that simple cross-validated additive bias correction is able to remove mean bias, whereas ensemble spread is inaccurate in some regions.

For the corresponding benchmark multi-model with additive bias correction, A_BC MFN, there are still some regions in which the null hypothesis of a flat rank histogram is rejected at a significance level of 0.05. But overall, the number of grid points at which a significant dispersion error can be detected decreases considerably by multi-model combination of the additive bias corrected ensembles, indicating that bias can be reduced by multi-model combination. Like for A_BC ECMWF SEAS5, there are no grid points with a significant mean bias.

Likewise MVA bias adjustment of raw ECMWF SEAS5 reduces the number of grid points with a significant dispersion error compared to A_BC ECMWF SEAS5. However, there are still grid points with a significant u-shape term indicating a dispersion error, in particular in the tropics. This is much less the case for the corresponding multi-model MVA MFN. The more sophisticated bias adjustment by CCR seems to be able to correct the dispersion error successfully, as there are almost no grid points with a significant u-shape term for ECMWF SEAS5 after applying CCR, labelled as CCR ECMWF SEAS5, and even fewer grid points for the corresponding multi-model, labelled as CCR MFN.

As bias adjustment, i.e. step (b) as illustrated in Fig. 1, is achieved by any of the forecast processing variants we do not show any detailed results of the bias (linear) term of the χ^2 decomposition of the GOF test statistics described in Sect. 2.4. In contrast, it is worth to assess dispersion in more detail.

Here, dispersion is computed based on the dispersion (u-shape) term of the χ^2 decomposition. A_BC ECMWF SEAS5 hindcasts are in particular underdispersive in some tropical regions (Fig. 5). Overdispersion, on the other hand, is quite rare and is confined to a few grid points in central Asia, North America, and southern Africa. Note that there are slightly more grid points with significant dispersion errors in Fig. 5 than in Fig. 4. As more separate statistical tests have been performed for the latter, FDR correction has a stronger effect on significances.

While additive bias corrected GloSea5 performs equally well as A_BC ECMWF SEAS5 in terms of dispersion, Météo France System 5 seems to be much more underdispersed. Météo France System 5 sea surface temperature forecasts are particularly underdispersed as well, while its precipitation forecasts are comparable to the other models in terms of dispersion errors (see supplemental material for the corresponding figures). Simple multi-model
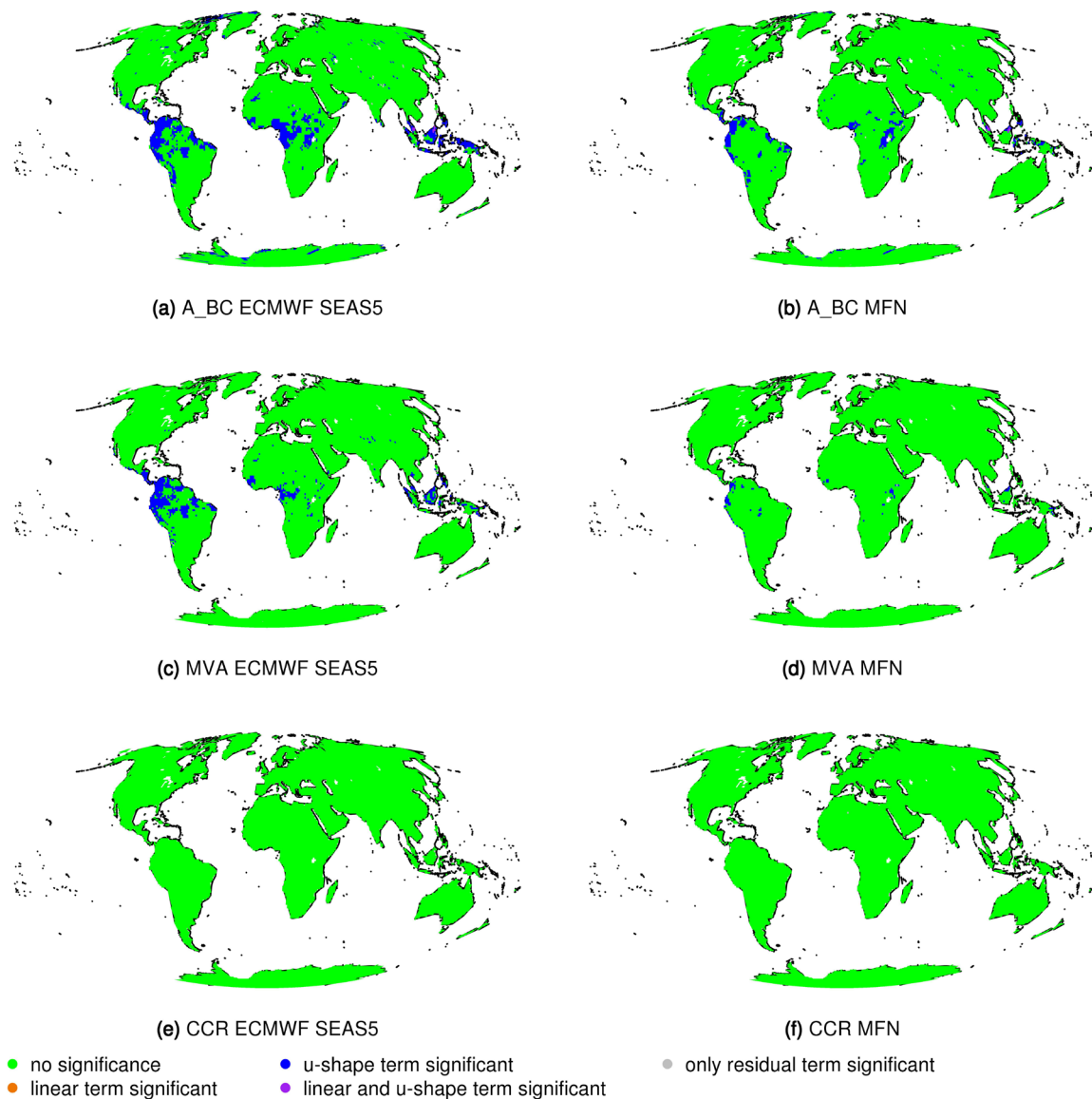
(a) A_BC ECMWF SEAS5

(b) A_BC MFN

(c) MVA ECMWF SEAS5

(d) MVA MFN

(e) CCR ECMWF SEAS5

(f) CCR MFN

- no significance
- linear term significant
- u-shape term significant
- linear and u-shape term significant
- only residual term significant

Fig. 4 Assessment of the "flat" rank histogram assumption using the χ^2 decomposition. Note: significance level 0.05 and FDR corrected

combination of the additive bias adjusted benchmarks, i.e. applying only a version of step (b), which does not affect variance, and step (c) depicted in Fig. 1, attenuates underdispersion in many regions.

The MVA bias adjustment approach can correct for the dispersion errors at most of the grid points, while there remain some regions, in particular the northern half of South America and parts of Africa where forecasts are still underdispersed. CCR leads to an additional reduction in the number of underdispersive grid points compared to MVA. In a nutshell, combining bias adjustment with multi-model combination, leads to a substantial reduction in the number of grid points with significant underdispersion.

3.3 Assessment of sharpness

Having assessed reliability in detail, let us now have a look at the sharpness, which is a measure of how focused a forecast is (e.g. Gneiting and Raftery 2007). The forecasts to be compared are selected following the same rationale as in the previous Sect. 3.2.

As depicted by Fig. 6 applying bias adjustment methods [step (b) in the flowchart in Fig. 1], on average leads to a loss of sharpness for ECMWF SEAS5. In contrast to mean-adjustment, bias adjustment methods allow for changes in the variance. The accompanying loss of sharpness is rather small when using MVA, but more pronounced when using
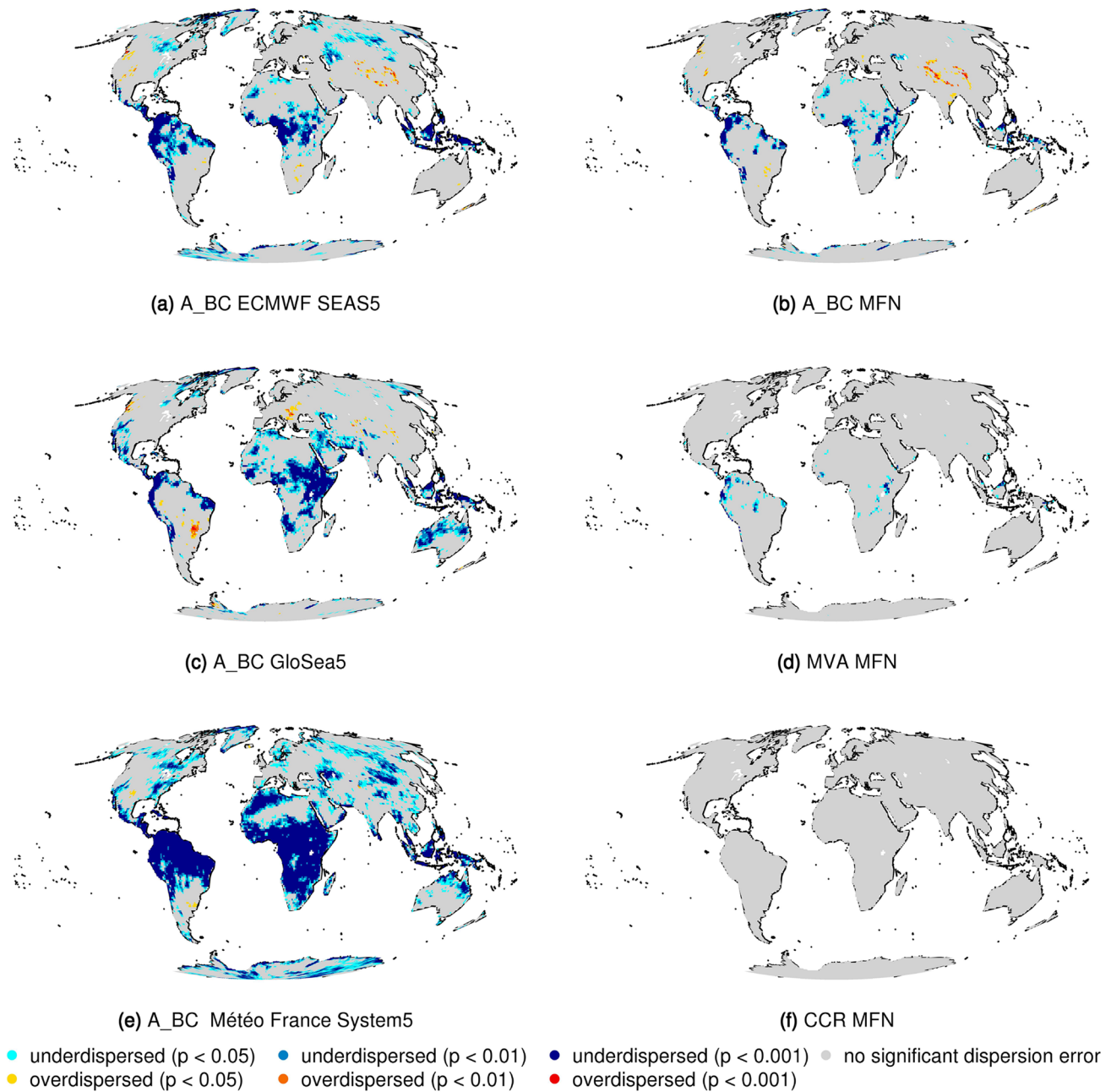
(a) A_BC ECMWF SEAS5

(b) A_BC MFN

(c) A_BC GloSea5

(d) MVA MFN

(e) A_BC Météo France System5

(f) CCR MFN

● underdispersed (p < 0.05) ● underdispersed (p < 0.01) ● underdispersed (p < 0.001) ● no significant dispersion error
● overdispersed (p < 0.05) ● overdispersed (p < 0.01) ● overdispersed (p < 0.001)

Fig. 5 Assessment of the dispersion term of the χ^2 decomposition. Significance level 0.05 and FDR corrected

CCR. It is difficult, however, to find systematic spatial patterns of changes in sharpness.

Multi-model combination, step (c) in the flowchart of the mean-adjusted raw ensemble hindcasts leads also to a decrease in sharpness in most parts of the world as revealed by a comparison of A_BC ECMWF SEAS5 with A_BC MFN. Multi-model combination of the MVA or CCR bias adjusted ensemble forecasts, however, does not affect sharpness much as can be seen from comparing panels (c) with (d) and panels (e) with (f) in Fig. 6, respectively. Together

with the increased reliability of bias adjusted multi-model ensembles, this emphasizes the gain in forecast quality by multi-model combination of calibrated forecasts.

4 Discussion

Stepping back to the flowchart shown in Fig. 1, there are three main tasks related to the generation of multi-model seasonal forecast products:
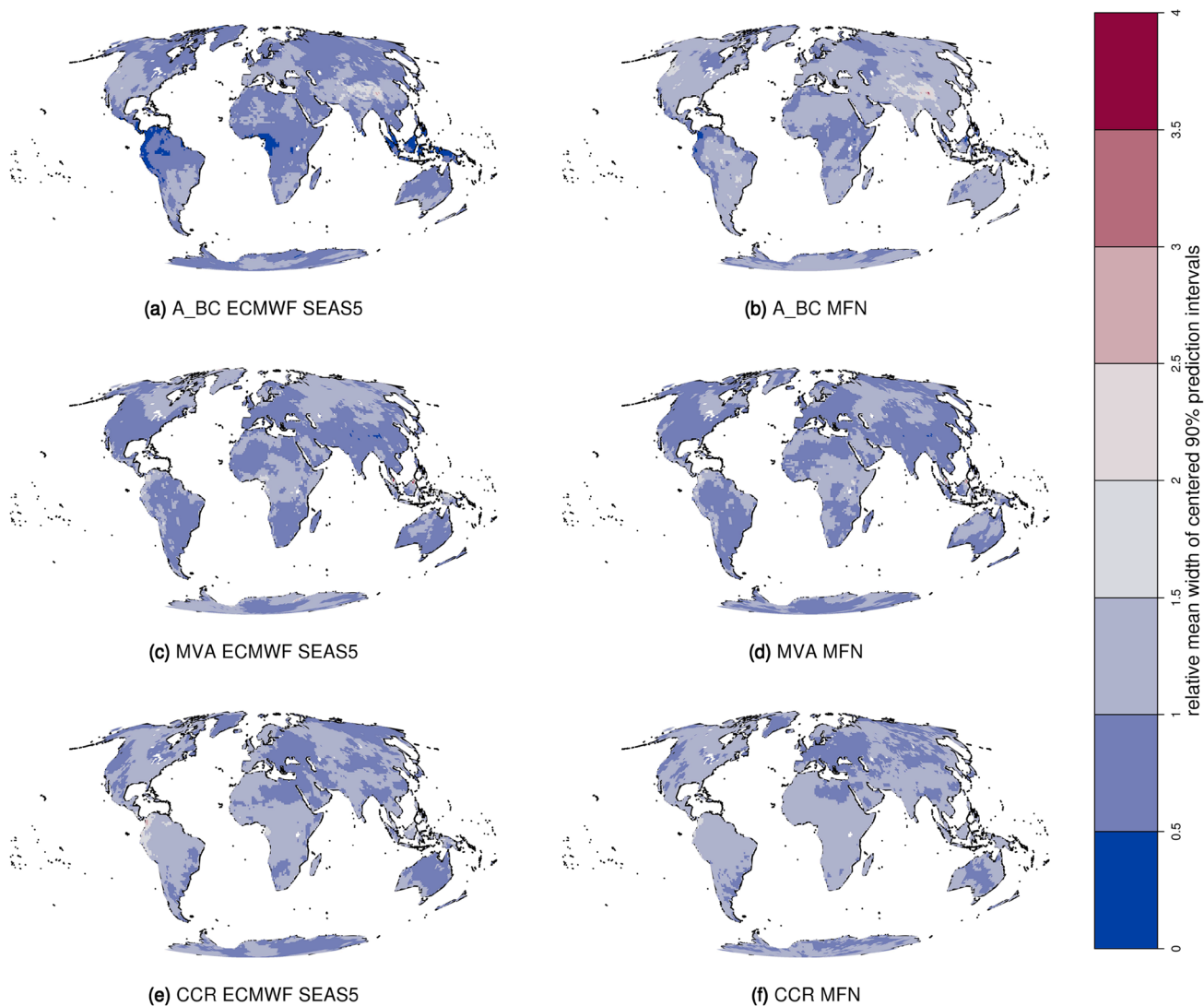
Fig. 6 Maps of the relative mean sharpness of pooled seasonal T2M forecasts for MAM, JJA, SON, and DJF, initialized in February, May, August, and November, respectively, in terms of the mean width of 90% prediction intervals verified over the period 1993–2014 relative to the mean width of the corresponding 90% prediction intervals of cross-validated climatological forecasts. The panels on the left show the sharpness of the single ECMWF SEAS5 hindcasts with additive, MVA, and CCR bias adjustment, respectively. The panels on the right show the sharpness of pooled multi-models (ECMWF SEAS5, GloSea5, and Météo France System 5) with the corresponding bias adjustments

(a) data aggregation in space and time,
(b) forecast bias adjustment or statistical post-processing in general, and
(c) multi-model combination.

In principle, all 6 permutations of (abc, cab, ...) are possible, but some are more suitable and practical than others.

Usually it makes sense to start with data aggregation (a), because the aggregation level is often predetermined by the specific application, which is the case for the analyses performed within QA4Seas as shown in the flowchart (cf. Fig. 1). However, if there is freedom in the choice of the aggregation level, an appropriate aggregation level should be defined carefully.

The moment we aggregate forecasts and observations in space and time, we lose spatio-temporal information about biases, signal-to-noise, trends etc., that might be useful for post-processing. But by aggregating we also average out high-frequency noise that might increase the estimation variance of bias correction parameters and combination weights. In seasonal forecasting we typically have low signal to noise ratio (i.e. the variance of the ensemble mean forecast is small compared to the ensemble variance (e.g. Scaife and Smith 2018), and we don't tend to use spatio-temporal information

in post-processing, so it usually makes sense to aggregate the data first even when this is not dictated by a specific forecast application.

For instance Gong et al. (2003) showed that seasonal forecasts of precipitation exhibit increasing skill with increasing spatial aggregation as long as precipitation in the entire aggregation area is forced by a common signal, typically up to areas of about 15° in latitude and/ or longitude. However, aggregating over geographically distant regions may be detrimental. Kharin et al. (2017), who considered several variables, reported that spatial smoothing did not generally improve skill. We hypothesize here that spatial aggregation, which can be understood as very strong smoothing, would also rather deteriorate forecast skill in the setting of Kharin et al. (2017). Hence, if the user is interested in regional forecast products, spatial aggregation should be done as a first step, whereas for forecast products spanning large areas up to the globe it is probably beneficial to do spatial aggregation after bias correction and multi-model combination.

While Kharin et al. (2017) confirm that temporal aggregation of seasonal forecasts typically also improves forecast skill, for instance Salles et al. (2016) report benefits from temporal aggregation only if the variable to be forecast can be represented as a stationary time series, whereas temporal aggregation should not be applied for variables showing non-stationary temporal patterns. Hence, the exact position of temporal aggregation among the steps shown in the flowchart strongly depends on statistical properties of the variable of interest that may depend on geographical region and season.

The sequence of (b) and (c) depends on the similarity of the forecast systems considered. If the selected forecast models have similar errors (in terms of sign and magnitude) it makes little sense to bias-correct them individually. Instead, one can post-process the forecasts jointly and assume strong dependency between the post-processing parameters (as in, e.g., Siegert and Stephenson 2019). In extreme cases, when the assumption of completely exchangeable forecasts holds, one would average the raw forecasts first and post-process the combined forecast. The latter approach would favour the sequence (a) → (c) → (b), unlike the sequence in the flowchart in Fig. 1, but under specific bias-correction methods, the sequence (a) → (b) → (c) might also be practical. In principle, (b) and (c) can be done in one step, e.g. by applying EMOS. However, at least for the study performed within QA4Seas, EMOS is outperformed by methods that apply (b) and (c) separately.

Returning to the analyses performed within QA4Seas it becomes obvious that there are quite a few regions where the different raw models show errors of opposite sign (bias maps of the raw ensemble hindcasts are not shown). Hence, if applied globally, the sequence (a) → (b) → (c), as shown

in the flowchart in Fig. 1, is the safe option to use knowing that it may be outperformed by (a) → (c) → (b) in regions, where all models show similar errors. In practice, often a single forecast product is requested, which requires the generally applicable sequence (a) → (b) → (c). Further, it may be difficult to determine if the forecasts are similar. Should similarity be assessed grid point by grid point? How should inconsistencies at the boundaries between areas of similar and dissimilar models be handled?

The above discussion on the sequence of processing steps applies to both ensemble and multi-category probability forecasts. For the latter, bias adjustment is done implicitly when computing probabilities. While the sequence (a) → (b) → (c) is straightforward, the sequence (a) → (c) → (b) would imply multi-model combination by computing probabilities from the pooled ensemble consisting of all ensemble models to be combined. This in turn necessitates a forecast climatology of the pooled multi-model ensemble, which is not straightforward to generate.

Note also that some implicit, but important, assumptions have been made for this study: first, we have used ECMWF ERA-Interim analysis data for training and verification tacitly assuming that it represents the true observations. This is obviously not the case. However, we did also some analyses using CRU-TS (Harris et al. 2014) as reference, which led to comparable results. Second, using ECMWF ERA-Interim analysis as observation data set potentially leads to biased verification results in that ECMWF SEAS5 may be favored over the other forecasting systems just by the fact that it is based on the same analysis as the observational data set. Third, the post-processing approaches presented in this study only considered forecasts for the variable of interest as predictor at a specific location, e.g. for sea surface temperature we only use ensemble forecasts of sea surface temperature at that specific location. Obviously, our results may look different when including also forecasts for other grid points and other forecast variables. Such a teleconnections based post-processing approach, however, probably needs longer training periods in order to outperform the basic approaches applied in this study.

In the meantime, additional seasonal forecasting systems have been made available through the CDS. Also, the ECMWF ERA-Interim analysis, which we have used as reference, has been replaced by ERA5. At the time of writing, these datasets have not been available and therefore could not be included. A follow-up analysis would definitely need to include these more up-to-date datasets. Furthermore, the assessment of reliability, bias, and dispersion errors using PIT GOF tests may be improved by including recent results on GOF testing under serial correlation (Bröcker and Ben Bouallègue 2020).

5 Conclusions

The results stress the importance of bias adjustment of seasonal ensemble forecasts. CCR proved to be the optimal approach for the study at hand, since the simpler MVA approach sometimes fails to correct underdispersion and—not surprisingly—the more sophisticated EMOS post-processing approach does not perform well in terms of forecast skill. The poor performance of EMOS is most likely due to the low predictability of seasonal forecasts in combination with the limited time sample of only 22 years, which leads to only 21 training data points in the cross-validated model estimation setting applied. This issue underpins also the importance of cross-validation for any type of processing of seasonal forecast ensembles.

We emphasize again that the provision of skill optimized ensemble and multi-category probability forecast products need different processing steps. Though it depends on the target application, our results suggest that in general using an unweighted multi-model combination of CCR bias adjusted single forecast systems leads to well calibrated seasonal forecast with a relatively good predictive skill. Unless a much larger training set is available or the skill of the systems improves substantially, more sophisticated multi-model combination approaches are unlikely to perform better.

As stated in Sect. 2.5 any analysis of seasonal forecasts is driven by the type of forecast product requested by the user. We have chosen to verify forecast products formulated in the form of multi-category probabilities using the RPS. Forecast products formulated as continuous probability functions are preferably verified using the CRPS. On the one hand RPS and CRPS are proper scoring rules, which prevent the forecaster from hedging (Gneiting and Raftery 2007; Weigel et al. 2007) and on the other hand they allow for calculation of skill scores, significance testing, and aggregation in space and time. Additionally, these scores can be decomposed into a reliability, a resolution, and a uncertainty term using the score decomposition proposed by Siegert (2017b).

Furthermore, the χ^2 decomposition of the rank histogram proved to be a suitable tool for visualization of miscalibration, bias, and dispersion errors on the global grid. In combination with an assessment of sharpness, a detailed picture of forecast quality can be obtained.

Compliance with ethical standards

References

Alonzo B, Ringkjob HK, Jourdier B, Drobinski P, Plougonven R, Tankov P (2017) Modelling the variability of the wind energy resource on monthly and seasonal timescales. Renew Energy 113:1434–1446. https://doi.org/10.1016/j.renene.2017.07.019

Baker LH, Shaffrey LC, Sutton RT, Weisheimer A, Scaife AA (2018) An intercomparison of skill and overconfidence/underconfidence of the wintertime North Atlantic Oscillation in multimodel seasonal forecasts. Geophys Res Lett 45(15):7808–7817. https://doi.org/10.1029/2018GL078838

Barcelona Supercomputing Center (2018) Quality assurance for multi-model seasonal forecast products. https://climate.copernicus.eu/quality-assurance-multi-model-seasonal-forecast-products. Accessed 9 June 2020

Batté L, Ardilouze C, Déqué M (2018) Forecasting West African heat waves at subseasonal and seasonal time scales. Mon Weather Rev 146(3):889–907. https://doi.org/10.1175/MWR-D-17-0211.1

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc (Ser B) 57(1):289–300

Bröcker J, Ben Bouallègue Z (2020) Stratified rank histograms for ensemble forecast verification under serial dependence. Q J R Meteorol Soc. https://doi.org/10.1002/qj.3778

Clark RT, Bett PE, Thornton HE, Scaife AA (2017) Skilful seasonal predictions for the European energy industry. Environ Res Lett 12(2):024002. https://doi.org/10.1088/1748-9326/aa94a7

DelSole T, Yang X, Tippett MK (2013) Is unequal weighting significantly better than equal weighting for multi-model forecasting? Q J R Meteorol Soc 139(670):176–183. https://doi.org/10.1002/qj.1961

Demirel MC, Booij M, Hoekstra A (2015) The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models. Hydrol Earth Syst Sci 19:275–291. https://doi.org/10.5194/hess-19-275-2015

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13(3):253–263. https://doi.org/10.1080/07350015.1995.10524599

Doblas-Reyes FJ, Hagedorn R, Palmer T (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting-II. Calibration and combination. Tellus A Dyn Meteorol Oceanogr 57(3):234–252. https://doi.org/10.3402/tellusa.v57i3.14658

Elmore KL (2005) Alternatives to the Chi-square test for evaluating rank histograms from ensemble forecasts. Weather Forecast 20(5):789–795. https://doi.org/10.1175/WAF884.1

Epstein ES (1969) A scoring system for probability forecasts of ranked categories. J Appl Meteorol 8(6):985–987

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378. https://doi.org/10.1198/016214506000001437

Gneiting T, Raftery AE, Westveld AH III, Goldman T (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon Weather Rev 133(5):1098–1118. https://doi.org/10.1175/MWR2904.1

Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA (2008) Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Test 17(2):211. https://doi.org/10.1007/s11749-008-0114-x

Gong X, Barnston AG, Ward M (2003) The effect of spatial aggregation on the skill of seasonal precipitation forecasts. J Clim 16(18):3059–3071. https://doi.org/10.1175/1520-0442(2003)016%3C3059:TEOSAO%3E2.0.CO;2

Hagedorn R, Doblas-Reyes FJ, Palmer T (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting-I. Basic concept. Tellus A Dyn Meteorol Oceanogr 57(3):219–233. https://doi.org/10.3402/tellusa.v57i3.14657

Harris I, Jones P, Osborn T, Lister D (2014) Updated high-resolution grids of monthly climatic observations-the CRU TS3.10 dataset. Int J Climatol 34(3):623–642. https://doi.org/10.1002/joc.3711

Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 15(5):559–570. https://doi.org/10.1175/1520-0434(2000)015%3C0559:DOTCRP%3E2.0.CO;2

Johnson SJ, Stockdale TN, Ferranti L, Balmaseda MA, Molteni F, Magnusson L, Tietsche S, Decremer D, Weisheimer A, Balsamo G, Keeley SPE, Mogensen K, Zuo H, Monge-Sanz BM (2019) SEAS5: the new ECMWF seasonal forecast system. Geosci Model Dev 12(3):1087–1117. https://doi.org/10.5194/gmd-12-1087-2019

Jollife IT, Primo C (2008) Evaluating rank histograms using decompositions of the chi-square test statistic. Mon Weather Rev 136(6):2133–2139. https://doi.org/10.1175/2007MWR2219.1

Kharin VV, Merryfield WJ, Boer GJ, Lee WS (2017) A postprocessing method for seasonal forecasts using temporally and spatially smoothed statistics. Mon Weather Rev 145(9):3545–3561. https://doi.org/10.1175/MWR-D-16-0337.1

Knutti R, Sedláček J, Sanderson RL BM, Fischer EM, Eyring V (2017) A climate model projection weighting scheme accounting for performance and interdependence. Geophys Res Lett 44(4):1909–1918. https://doi.org/10.1002/2016GL072012

Mächler M (2017) nor1mix: Normal (1-d) Mixture Models (S3 Classes and Methods). R package version 1.2-3. https://CRAN.R-project.org/package=nor1mix. Accessed 9 June 2020

MacLachlan C, Arribas A, Peterson KA, Maidens A, Fereday D, Scaife AA, Gordon M, Vellinga M, Williams A, Comer RE et al (2015) Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. Q J R Meteorol Soc 141(689):1072–1084. https://doi.org/10.1002/qj.2396

Manzanas R, Gutiérrez JM, Bhend J, Hemri S, Doblas-Reyes FJ, Torralba V, Penabad E, Brookshaw A (2019) Bias adjustment and ensemble recalibration methods for seasonal forecasting: a comprehensive intercomparison using the C3S dataset. Clim Dyn 53(3–4):1287–1305. https://doi.org/10.1007/s00382-019-04640-4

Manzanas R, Gutiérrez JM, Bhend J, Hemri S, Doblas-Reyes FJ, Penabad E, Brookshaw A (2020) Statistical adjustment, calibration and downscaling of seasonal forecasts: a case-study for Southeast Asia. Clim Dyn 54:2869–2882. https://doi.org/10.1007/s00382-020-05145-1

Mishra N, Prodhomme C, Guemas V (2018) Multi-model skill assessment of seasonal temperature and precipitation forecasts over Europe. Clim Dyn. https://doi.org/10.1007/s00382-018-4404-z

Murphy AH (1969) On the "ranked probability score". J Appl Meteorol 8(6):988–989

Murphy AH (1971) A note on the ranked probability score. J Appl Meteorol 10:155–156

Ouédraogo M, Zougmoré RB, Barry S, Somé L, Grégoire B (2015) The value and benefits of using seasonal climate forecasts in agriculture: evidence from cowpea and sesame sectors in climate-smart villages of Burkina Faso. Climate Change, Agriculture and Food Security Info Note pp 01–04

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. Mon Weather Rev 133(5):1155–1174. https://doi.org/10.1175/MWR2906.1

Ramírez-Rodrigues MA, Alderman PD, Stefanova L, Cossani CM, Flores D, Asseng S (2016) The value of seasonal forecasts for irrigated, supplementary irrigated, and rainfed wheat cropping systems in northwest Mexico. Agric Syst 147:76–86. https://doi.org/10.1016/j.agsy.2016.05.005

Rodriguez D, Voil PD, Hudson D, Brown JN, Hayman P, Marrou H, Meinke H (2018) Predicting optimum crop designs using crop models and seasonal climate forecasts. Nat Sci Rep 8(1):2231. https://doi.org/10.1038/s41598-018-20628-2

Roudier P, Alhassane A, Baron C, Louvet S, Sultan B (2016) Assessing the benefits of weather and seasonal forecasts to millet growers in Niger. Agric For Meteorol 223:168–180. https://doi.org/10.1016/j.agrformet.2016.04.010

Salles R, Mattos P, Dubois AMDI, Bezerra E, Lima L, Ogasawara E (2016) Evaluating temporal aggregation for predicting the sea surface temperature of the Atlantic Ocean. Ecol Inform 36:94–105

Sanderson BM, Xu Y, Tebaldi C, Wehner M, O'Neill BC, Jahn A, Pendergrass AG, Lehner F, Strand WG, Lin L et al (2017) Community climate simulations to assess avoided impacts in 1.5 and 2 C futures. Earth Syst Dyn 8(3):827–847. https://doi.org/10.3929/ethz-b-000191578

Scaife AA, Smith D (2018) A signal-to-noise paradox in climate science. NPJ Clim Atmos Sci. https://doi.org/10.1038/s41612-018-0038-4

Scheuerer M (2014) Probabilistic quantitative precipitation forecasting using ensemble model output statistics. Q J R Meteorol Soc 140(680):1086–1096. https://doi.org/10.1002/qj.2183

Siegert S (2017a) SpecsVerification: forecast verification routines for ensemble forecasts of weather and climate. R package version 0.5-2. https://CRAN.R-project.org/package=SpecsVerification. Accessed 9 June 2020

Siegert S (2017b) Simplifying and generalising Murphy's Brier score decomposition. Q J R Meteorol Soc 143(703):1178–1183. https://doi.org/10.1002/qj.2985

Siegert S, Stephenson DB (2019) Forecast recalibration and multi-model combination. In: Roberton A, Vitart F (eds) Sub-seasonal to seasonal prediction: the gap between weather and climate forecasting. Elsevier, pp 321–336. https://doi.org/10.1016/b978-0-12-811714-9.00015-2

Stockdale T (2013) The EUROSIP system-a multi-model approach. In: Seminar on seasonal prediction: science and applications, 3–7 September 2012. ECMWF, Reading, UK, pp 257–268. https://www.ecmwf.int/sites/default/files/elibrary/2013/12429-eurosip-system-multi-model-approach.pdf. Accessed 9 June 2020

Torralba V, Doblas-Reyes FJ, MacLeod D, Christel I, Davis M (2017) Seasonal climate prediction: a new source of information for the management of wind energy resources. J Appl Meteorol Climatol 56(5):1231–1247. https://doi.org/10.1175/JAMC-D-16-0204.1

Ventura V, Paciorek CJ, Risbey JS (2004) Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. J Clim 17(22):4343–4356

Vitart F, Huddleston MR, Déqué M, Peake D, Palmer TN, Stockdale TN, Davey MK, Ineson S, Weisheimer A (2007) Dynamically-based seasonal forecasts of Atlantic tropical storm activity issued in June by EUROSIP. Geophys Res Lett. https://doi.org/10.1029/2007GL030740

Weigel AP, Liniger MA, Appenzeller C (2007) The discrete Brier and ranked probability skill scores. Mon Weather Rev 135(1):118–124. https://doi.org/10.1175/MWR3280.1

Weigel AP, Liniger MA, Appenzeller C (2009) Seasonal ensemble forecasts: are recalibrated single models better than multimodels? Mon Weather Rev 137(4):1460–1479. https://doi.org/10.1175/2008MWR2773.1

Weigel AP, Knutti R, Liniger MA, Appenzeller C (2010) Risks of model weighting in multimodel climate projections. J Clim 23(15):4175–4191. https://doi.org/10.1175/2010JCLI3594.1

Wilks DS (2016) "The stippling shows statistically significant grid points": how research results are routinely overstated and over-interpreted, and what to do about it. Bull Am Meteorol Soc 97(12):2263–2273. https://doi.org/10.1175/BAMS-D-15-00267.1

Yuan X, Wood EF, Ma Z (2015a) A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. Wiley Interdiscip Rev Water 2(5):523–536. https://doi.org/10.1002/wat2.1088

Yuan X, Roundy JK, Wood EF, Sheffield J (2015b) Seasonal forecasting of global hydrologic extremes: system development and evaluation over GEWEX basins. Bull Am Meteorol Soc 96(11):1895–1912. https://doi.org/10.1175/BAMS-D-14-00003.1