Bias adjustment and ensemble recalibration methods for seasonal forecasting: A comprehensive intercomparison using the C3S dataset

R. Manzanas · J. M. Gutiérrez · J. Bhend · S. Hemri · F. J. Doblas-Reyes · V. Torralba · E. Penabad · A. Brookshaw

Received: date / Accepted: date

Abstract This work presents a comprehensive intercomparison of different alternatives for the calibration of seasonal forecasts, ranging from simple bias adjustment (BA) —e.g. quantile mapping— to more sophisticated ensemble recalibration (RC) methods —e.g. non-homogeneous Gaussian regression,— which build on the temporal correspondence between the climate model and the corresponding observations to generate reliable predictions. To be as critical as possible, we validate the raw model and the calibrated forecasts in terms of a number of metrics which take into account different aspects of forecast quality (association, accuracy, discrimination and reliability). We focus on one-month lead forecasts of precipitation and temperature from four state-of-the-art seasonal forecasting systems, three of them included in the Copernicus Climate Change Service (C3S) dataset (ECMWF-SEAS5, UK Met Office-GloSea5 and Météo France-System5) for boreal winter and summer over two illustrative regions with different skill characteristics (Europe and Southeast Asia). Our results indicate that both BA and RC methods effectively correct the large raw model biases, which is of paramount importance for users, particularly when directly using the climate model outputs to feed their impact models, or when computing climate indices depending on absolute values/thresholds. However, except for particular regions and/or seasons (typically with high skill), there is only marginal added value beyond this bias removal. For those cases, RC methods can outperform BA ones, mostly due to an improvement in reliability. Finally, we also show that whereas an increase in the number of members only modestly affects the results obtained from calibration, longer hindcast periods lead to improved forecast quality, particularly for RC methods.

R. Manzanas (✉) · J. M. Gutiérrez
Meteorology Group. Institute of Physics of Cantabria (IFCA), CSIC-University of Cantabria
Santander, 39005, Spain. E-mail: rmanzanas@ifca.unican.es

J. Bhend · S. Hemri
Federal Office of Meteorology and Climatology MeteoSwiss, Switzerland

F. J. Doblas-Reyes · V. Torralba
Barcelona Supercomputing Center (BSC), Spain

E. Penabad · A. Brookshaw
European Centre for Medium-Range Weather Forecasts (ECMWF), UK

Keywords Seasonal forecasting, C3S, bias adjustment, ensemble recalibration, forecast quality, reliability, ensemble size, hindcast length

1 Introduction

The current state-of-the-art General Circulation Models (GCMs) used for seasonal forecasting have horizontal resolutions which are typically coarser than those needed for practical applications, and suffer from substantial systematic biases and drifts (see Doblas-Reyes et al, 2013, and references therein). This poses a risk to the use of raw model outputs in many sectors which require seasonal predictions with similar statistical properties to those observed at the regional/local scale (e.g. energy, hydrology, agriculture or health). Nowadays, it is well established that some form of post-processing is needed prior to the direct use of raw model seasonal data, which constitutes a challenging problem for the development of high-quality climate services (see, e.g., Torralba et al, 2017). A number of different approaches aiming at reducing the systematic model errors have been proposed, ranging from bias adjustment (BA) and ensemble recalibration (RC) methods —both acting directly on the variable of interest,— to more complex statistical downscaling techniques building on large-scale predictors (Maraun et al, 2010). Whilst statistical downscaling has been extensively analyzed in the literature in the framework of seasonal forecasting (see, e.g., Gutiérrez et al, 2005; Pavan et al, 2005; Manzanas et al, 2018; Manzanas and Gutiérrez, 2018; Nikulin et al, 2018), little attention has been put to-date into BA and RC methods (the focus of this work).

BA methods adapt the raw model outputs (e.g. predicted precipitation for a target season and a lead time) towards the corresponding observational reference to make them compatible with the local climatology. This is typically done by mapping the distribution of predicted values onto the corresponding observed one, based on a sufficiently long historical/hindcast period. These techniques, which do not use information about temporal correspondence between predictions and observations, range from simple adjustments in the mean and/or variance to more complex quantile mapping alternatives which can adjust higher order moments or even the entire distribution. Whereas the former have a long tradition in seasonal forecasting (see, e.g., Barnston, 1994; Doblas-Reyes et al, 2005), the latter have been introduced in the context of climate change projections (see, e.g., Piani et al, 2010) and their application in seasonal forecasting is quite recent (see, e.g., Zhao et al, 2017; Manzanas et al, 2018; Manzanas and Gutiérrez, 2018). One of the main advantages of BA techniques is that they can be applied to correct daily data, even for variables which do not follow standard (e.g. Gaussian) distributions, which is often required by users (note that here we focus exclusively in the adjustment of seasonal means). For this reason, quantile mapping is rapidly becoming the method of preference by operational agencies and end-users (see, e.g., Bedia et al, 2018). Nevertheless, the fact that their application is straightforward in most of the cases favors the misuse of BA techniques (Maraun et al, 2017).

RC methods transform the raw model outputs building on the temporal correspondence between the ensemble mean predictions and the corresponding observations (see Sansom et al, 2016, for a comprehensive review). They range from relatively simple implementations such as climate conserving recalibration —CCR (see, e.g., Doblas-Reyes et al, 2005; Weigel et al, 2009)— or the ratio of predictable

components —RPC (Eade et al, 2014)— to more general ensemble model output statistics (EMOS) methods such as non-homogeneous Gaussian regression (see, e.g., Gneiting et al, 2005; Tippett and Barnston, 2008; Sansom et al, 2016). The main advantage of RC techniques for seasonal forecasting is that they are designed to produce reliable predictions. However, as opposite to BA methods, they are not suitable for the adjustment of daily data —because of the lack of model skill at this particular time-scale.— Moreover, an important constraint of RC techniques in the context of seasonal forecasting is that the underlying parameters have to be estimated using a limited amount of data and, as a consequence, they are prone to overfitting. Note that, due to the enormous computational requirements and the lack of long observational datasets required to initialize the forecasting system, state-of-the-art hindcasts typically have around 30 years of data (i.e. a sample size of 30 values for calibration).

Recent studies have reported some limitations for BA methods (Manzanas et al, 2018) for seasonal forecasting, and even the preferable choice of RC techniques (Zhao et al, 2017). However, to the authors' knowledge, there is no comprehensive intercomparison of BA and RC methods for this type of predictions. The main goal of this paper is therefore to fill this knowledge gap. To do this, we apply a set of sate-of-the-art BA and RC methods to calibrate one-month lead seasonal predictions of temperature and precipitation from four different forecast systems. Three of these systems are included in the Copernicus Climate Change Service (C3S) seasonal service (`http://climate.copernicus.eu/seasonal-forecasts`), whereas the fourth (the ECMWF System4) is used to test the sensitivity of the results obtained to the hindcast length and the ensemble size. The raw model and calibrated predictions are validated in terms of a number of verification metrics which take into different aspects of forecast quality (association, accuracy, discrimination and reliability).

The paper is organized as follows. In Section 2 we describe all the data used and introduce the different BA and RC methods applied (the implementation details are given in the Annex) and the verification metrics considered. Results obtained are presented through Section 3. The main conclusions and some interesting discussion are outlined in Section 4.

2 Data and Methods

2.1 Data Used

In this work we focus on precipitation and temperature for boreal winter (DJF) and summer (JJA) over two illustrative regions spanning tropical and extra-tropical latitudes: Europe and Southeast Asia (EU and SA, hereafter). Note that whereas low-to-moderate skill is in general acknowledged for the former, overall good skill has been documented in the latter (see, e.g., Manzanas et al, 2014).

We analyze one-month lead seasonal forecasts (i.e. predictions initialized in November and May for DJF and JJA, respectively) from the C3S seasonal multi-system ensemble, which consists of three state-of-the-art models with a common hindcast period of 22 years, 1993-2014 (see Table 1). For DJF, a total of 21 seasons are available (starting with D1993-JF1994, which we refer to as DJF 1994). Therefore, for the sake of comparability we use a common 21-year period for both

Table 1 Seasonal hindcasts used in this study. The last two columns show the ensemble size (members) and the period covered for each dataset.

Source	Institution	Model	Code	Members	Period
ECMWF	ECMWF	System4	System4	51	1982-2016
C3S	ECMWF	SEAS5	SEAS5	25	1993-2016
C3S	UK Met Office	GloSea5	SYSTEM12	12	1993-2015
C3S	Météo France	System5	SYSTEM5	15	1993-2014

DJF and JJA, and only the 12 first members of each model (minimum number of members common across all models) are considered. Additionally, we have also used the 51-member version of the ECMWF-System4 (Molteni et al, 2011) — the longest to-date seasonal hindcast— for testing the sensitivity of the results obtained to the ensemble size and the hindcast length (see Section 3.3).

The ERA-Interim reanalysis (Dee et al, 2011) is used as observational reference dataset for both the calibration of the BA and RC methods and also for the verification of all seasonal forecasts involved in this work. For consistency, all the C3S models and ERA-Interim have been bi-linearly interpolated from their distinct native horizontal resolutions to a common 1° regular grid.

2.2 Bias Adjustment and Ensemble Recalibration Methods

Table 2 shows the BA and RC methods intercompared in this work (see the Annex for details on the particular implementations), which have been already used in the context of seasonal forecasting (see the references in the third column of the table). On the one hand, two BA methods were considered: a simple mean and variance adjustment (MVA) and a standard implementation of empirical quantile-quantile mapping (EQM). Note that we also considered an even simpler method consisting of adjusting only the mean; however, the results were very similar to those obtained for MVA and are thus not shown for brevity. Also, for EQM, we tested the suitability of both monthly and seasonal data for the mapping, obtaining very similar conclusions in both cases. For coherence with the rest of methods, we only show results for the case of seasonal values. On the other hand, four RC methods were considered: climate conserving recalibration (CCR), ratio of predictable components (RPC) and two EMOS choices using linear regression (LR) and non-homogeneous Gaussian regression (NGR). Note that, although more sophisticated RC methods exist (Sansom et al, 2016), we have selected here some standard parsimonious ones (already used in seasonal forecasting studies) which are preferable to avoid overfitting problems.

All the methods considered for this work (with the exception of EQM) have been implemented in an R-package called *calibratoR* (`http://github.com/SantanderMetGroup/calibratoR`), which is publicly available as part of the *climate4R* framework (Iturbide et al, 2018). The method EQM (as well as other BA and downscaling techniques) are available in the *downscaleR* package (`http://github.com/SantanderMetGroup/downscaleR`), which is also part of *climate4R*. A detailed description of each method is given in the Annex. Here, all BA and RC methods have been applied at a gridbox level under a leave-one year-out (LOO) cross-validation scheme (Lachenbruch and Mickey, 1968). Note that proper cross-validation is

Table 2 Bias adjustment (BA) and ensemble recalibration (RC) methods used in this work. See the Annex for implementation details.

Approach	Method	Code	Reference(s)
BA	Mean/variance adjustment	MVA	Doblas-Reyes et al (2005), Torralba et al (2017)
BA	Empirical quantile-quantile mapping	EQM	Zhao et al (2017), Manzanas et al (2018)
RC	Climate conserving recalibration	CCR	Weigel et al (2009)
RC	Ratio of predictable components	RPC	Eade et al (2014)
RC	Linear regression	LR	Marcos et al (2018)
RC	Non-homogeneous Gaussian regression	NGR	Tippett and Barnston (2008)

mandatory in order to avoid artificial skill (Manzanas et al, 2017), especially when working with small sample sizes like in this case (21 years of data).

2.3 Forecast Quality Metrics

The validation of seasonal predictions is a multi-faceted problem, which requires the use of several performance metrics to analyze different aspects of forecast quality such as association, accuracy, discrimination and reliability. Association reflects the strength of the relationship between the forecasts and the corresponding observations, which is measured here by the Pearson correlation between the ensemble mean and the observed interannual time-series.

Accuracy measures the average distance between forecasts and observations. We consider here two standard scores which are typically used to characterize this property: the Continuous Ranked Probability Score (CRPS) and the Ranked Probability Score (RPS). The CRPS (Hersbach, 2000) is a metric that allows to assess the performance of probabilistic forecasts of a continuous variable based on the integrated squared difference between the observed and the predicted cumulative distribution functions (which would correspond to the mean absolute error for deterministic forecasts). To allow for direct comparison across the different BA and RC methods, we use the associated skill score (CRPSS), which is computed as $1-(CRPS_{cal}/CRPS_{ref})$, being $CRPS_{ref}$ the CRPS obtained for the raw model forecasts and $CRPS_{cal}$ the one for the calibrated predictions. Thus, values above (below) 0, indicate that the particular calibration method improves (degrades) the raw model prediction. The RPS (Epstein, 1969) is the discrete version of the CRPS and measures the sum of squared differences in cumulative probability space for a multi-category probabilistic forecast (for two-category forecasts, it would be the Brier Score). As for the case of the CRPS, we also use here the associated skill score (the RPSS), which is computed as $1-(RPS_{cal}/RPS_{ref})$.

Discrimination measures the ability of the forecasts to distinguish between an event and the corresponding non-event, which is assessed here by means of the area under the ROC curve Kharin and Zwiers (2003) (simply referred to as ROC hereafter). Again, we use the associated skill score (ROCSS), which is computed as $(ROC_{cal}-ROC_{ref})/1-ROC_{ref}$. This metric is recommended by the Lead Centre for the Standardized Verification System of Long Range Forecasts and has been used in many previous studies for the verification of seasonal forecasts (see, e.g., Manzanas et al, 2014). Note that RPS and ROC are used here for tercile-based probabilistic predictions. In both cases, terciles are independently computed for the observations and the predictions, which implicitly introduces

a bias adjustment in the forecasts. Therefore, as opposite to CRPS, these two metrics are bias-insensitive, allowing thus to explore the added value of BA and RC method beyond the expected (by construction) model bias reduction.

Finally, reliability measures how closely the forecast probabilities of a certain event correspond to the observed frequency of that event (for instance a particular tercile category). Here reliability is analyzed in two different ways. On the one hand, we separate the RPS into its three components (reliability, resolution and uncertainty) following the Brier decomposition introduced in Murphy (1973). On the other hand, we use the reliability categories introduced in Weisheimer and Palmer (2014), which are based on the relative position of the best-guess reliability line and the uncertainty range around it in the reliability diagram. In particular, we use the extended classification proposed by Manzanas et al (2018), which includes the five original categories —*perfect* (green), *still very useful* (blue), *marginally useful* (yellow) *not useful* (orange) and *dangerously useless* (red)— plus a new one —*marginally useful +* (dark yellow).—

Notice also that, whereas CRPS and RPS are sensitive to changes in reliability, ROC is not. Thus, the latter also allows to assess the potential usefulness of the different calibration methods beyond the (possible) gain in reliability.

3 Results

3.1 Validation of Raw Model Outputs

As a result of their limited spatial resolution and the corresponding misrepresentation of important local features (e.g. complex topography and land-sea contrasts), global models typically exhibit significant mean errors (biases) when compared with observations. Figure 1 shows bias between the one-month lead ensemble mean of the four models considered and ERA-Interim for precipitation (top) and temperature (bottom). As explained, the common period 1994-2014 is used and only the first 12 members are considered for all models. Important variable- and season-dependent biases are found for all models, with values over 4°C (200mm/year) for temperature (precipitation) in many locations. Although there are regional differences among models, there exists a certain common spatial pattern of bias, especially over EU (being SYSTEM5 the most dissimilar model). These systematic errors are due to the important simplifications that need to be done when building the global models as a consequence of the lack of observations and knowledge, which lead to important errors in circulation, energy exchanges, etc.

As expected (by construction), all the BA and RC methods effectively correct the raw model biases, leading to mean errors that are smaller than 15mm/year for precipitation and 0.05°C for temperature in all cases (not shown for brevity). This is of paramount importance for users, particularly when using climate model outputs to feed their impact models, or when computing climate indices depending on absolute values/thresholds, and proves that some form of calibration is necessary prior to the direct use of the raw predictions.

The temporal association between the raw ensemble mean and the corresponding observations is a key parameter used by the RC methods in the calibration process (see the Annex). Figure 2 shows the interannual Pearson correlation between the ensemble mean of the four models considered and ERA-Interim for

precipitation (top) and temperature (bottom). As in Figure 1, the common period 1994-2014 and the first 12 available members were considered in all cases. Only significantly positive correlations at a 90% confidence level (according to a t-test) are shown. Note that, in the following, the results found for all quality metrics are only shown for these "skillful" regions. We do this in order to avoid misinterpretation of the results obtained for the RC methods, which can lead to artificial skill in regions of small (or negative) raw model correlations (see, e.g., Eade et al, 2014). Correlations are higher for temperature than for precipitation, and also higher for tropical latitudes (SA) than for extratropical ones (EU). In general, all models exhibit a similar spatial pattern of correlations, particularly for temperature.

We want to remark that all the results shown in Figures 1 and 2 are almost identical if all the available members are considered for each model (not shown).

3.2 Performance of BA and RC Methods

For brevity, we first focus in this section on the illustrative case of temperature in DJF, for which the highest added value from calibration has been found (Figures 3 to 7). Then, for a comprehensive analysis we summarize the results obtained for all other cases in Figures 8 and 9.

Figure 3 shows the results obtained for the CRPS over EU (top) and SA (bottom). In particular, columns 2-3 (4-7) show the CRPSS obtained for the different BA (RC) methods, computed with respect to the CRPS for the raw outputs, which is considered as reference (column 1). Thus, values above (below) 0, shown in blue (red), indicate that the particular method improves (degrades) the raw model prediction. As a consequence of effectively adjusting the existing model biases (see the previous section), all methods are found to clearly improve the raw forecasts in all cases. Moreover, all methods perform similarly.

Figure 4 is the equivalent to Figure 3, but for the RPS (raw model outputs; column 1) and RPSS (for BA and RC methods; columns 2-7). Again, within each approach (BA or RC), all methods are found to perform very similarly. However, whereas BA methods lead in general to degraded results, RC ones provide a benefit for some particular regions (this is especially visible in SA), being this a robust feature across all models. Nevertheless, as we will show later, this benefit can not be directly generalized to other variables and/or seasons.

To better understand the origin of this benefit found for RC methods (as compared to BA ones), Figure 5 shows the reliability (top) and resolution (bottom) components of the RPSS shown in Figure 4. For simplicity, the results for a single model (System4) are shown; however, the same conclusions hold for the rest of models. The smaller (larger) the reliability (resolution) term is, the lower the RPS is. Therefore, the darker the color, the better in both panels. This figure proves that the improvement of RPS attained by RC methods comes from an increase in reliability (see top panel), a crucial property for the usability of seasonal forecasts.

Figure 6 shows the ROC (and ROCSS) for the cold and warm tercile categories (T1 and T3, top and bottom) of DJF temperature over SA. As in Figures 3 and 4, the ROC found for the raw forecasts (column 1) is considered as reference for all BA and RC methods (columns 2-7). Differently to the case of the RPSS, no added value is attained for this metric, neither for BA nor for RC methods. Moreover, results are generally degraded after calibration, particularly for RC methods —as we shall

see later, this can be in part explained by the short hindcast available for the C3S models.— This points out the complexity and multifaceted character of verification of seasonal forecasts, which needs to be carefully performed so that results are not misinterpreted (Doblas-Reyes et al, 2005). In particular, these results suggest that both RPS and ROC are necessary to fully assess the usefulness of multi-category probabilistic predictions.

Finally, we analyze how association between the predictions and observations varies with calibration. For each model (in rows), the maps in the first column of Figure 7 show the interannual Pearson correlation between ERA-Interim and the raw outputs for DJF temperature over EU (top) and SA (bottom) —this has been already shown in Figure 2.— For each of the BA and RC methods (columns 2-7), results are shown as the difference (in correlation units) with respect to the maps in column 1. As for the ROC, in general all methods are shown to degrade the correlation values attained by the raw forecasts (this is more evident for RC than for BA). It is worth to mention that this degradation in correlation is a consequence of the LOO cross-validation setting used here (all BA and RC roughly maintain the correlations exhibited by the raw outputs if cross-validation is not applied; not shown). Note the importance of this result for the potential use of BA and RC methods in operational seasonal forecasting setups.

In order to provide a comprehensive analysis, Figures 8 (for EU) and 9 (for SA) summarize the results obtained in terms of the different skill scores considered (CRPSS, RPSS, ROCSS and correlation differences; in columns) for all cases analyzed in this work. The two variables (precipitation and temperature) and seasons (DJF and JJA) are shown in different rows. In all cases, results for the four available models are displayed along the x-axis. For each model, the two (four) red (black) boxplots indicate the P25-75 range for each BA (RC) method, with blue corresponding to the P10-P90 range. For EU (with a low-to-moderate skill), BA methods are in general preferable, and especially the simplest MVA. As compared to the EQM, this method is found to provide a similar adjustment of biases, whilst yielding a smaller degradation of accuracy and association measures. Differently, in SA (with high skill in some regions), RC methods yield better reliability than BA ones, although the latter (and particularly MVA) are more robust in terms of accuracy and association. Nevertheless, as we will see in the next section, this can be partially due to the short hindcast period available here.

3.3 Results' Sensitivity to Hindcast Length and Ensemble Size

Taking into account the limited ensemble size (12 members) and hindcast length (21 years) available for this work, we analyze the robustness of the results shown in the previous sections by assessing how the different verification metrics considered may change for larger ensemble sizes and longer hindcast periods. To do this, we use the 51-member version of the System4 (the longest hindcast to-date), and consider the period 1982-2014. For the illustrative case of DJF temperature over SA, Figure 10 shows, in different panels from top to bottom, the CRPSS, RPSS, ROCSS (only for the warm tercile category) and interannual Pearson correlation obtained for three different configurations: the 12-member ensemble for the period 1994-2014 used in the previous sections (top row), a 51-member ensemble for 1994-2014 (middle row) and a 51-member ensemble for 1982-2014 (bottom row).

Whereas a larger ensemble does not play a significant role for any of the metrics analyzed (compare top and middle rows in each panel), there is a large influence coming from the length of the hindcast period available for the RPSS and the ROCSS, and, to a lesser extent, also for correlation (compare middle and bottom rows). Note that the best results for these metrics are obtained for 1982-2014, which points out the importance of having long hindcasts for suitable calibration of seasonal forecasts. On the contrary, note also that neither the ensemble size nor the hindcast length strongly affect the results obtained for the CRPSS, which indicates that small ensembles and short hindcasts (e.g. 12 members and 21 years) are enough to robustly characterize and adjust the main systematic model errors (e.g. mean biases).

Additionally, Figure 11 shows the reliability categories obtained for the different configurations of the System4 considered in Figure 10. For simplicity, results are only shown for the warm tercile category (T3). Reliability is computed for each of the 20 subregions introduced in Figure 1 of Sheau et al (2017), provided there is at least a 25% of points with significantly positive interannual correlations for the ensemble mean (see Figure 2). Within each subregion, we pool together all gridboxes for both observations and predictions.

In agreement with the results found for the decomposition of the RPSS (Figure 5), Figure 11 shows that, whereas in general BA methods do not improve (or even degrade) the reliability of the raw model outputs, RC methods tend to yield better results for particular regions. Moreover, for the case of RC methods, both ensemble size and hindcast length have an impact on reliability, being the latter the dominant factor. In particular, as compared to 1994-2014, reliability is clearly improved for the case of RC methods when considering the period 1982-2010, which suggests, again, the importance of having long hindcasts for suitable calibration.

3.4 Computational Requirements

Although not strictly decisive from a scientific point of view, the analysis of the computational requirements demanded by the different methods is important from a practical perspective, especially regarding their potential usability for climate services and other user-tailored applications. Figure 12 shows, for the illustrative case of DJF temperature from the System4 (12-member, 21-year version), the execution times (in minutes) required by the BA and RC methods used in this work, according to their implementation in *calibratoR*. Dark (light) gray correspond to the LOO cross-validation setting used here for EU (SA) —note that computing times drastically reduce if cross-validation is not applied; not shown.— Among the BA methods, MVA is very rapid, being therefore a suitable option for real-time applications (e.g. interactive webpages). In particular, it is much much faster than EQM, which is widely used nowadays for different sectoral tasks. Among the RC methods, CCR and RPC are computationally inexpensive choices (also potentially exploitable in real-time applications), with LR still providing reasonable times (less than 2 minutes for EU). Differently, the long execution times required by NGR make this method unusable for real-time operations. In the light of these results, MVA and/or CCR could be considered as benchmarking methods which provide a good compromise between performance and computational cost for the calibration of seasonal mean values.

4 Discussion and Conclusions

This work presents a comprehensive intercomparison of different alternatives for the calibration of seasonal forecasts, ranging from simple bias adjustment (BA) to more sophisticated ensemble recalibration (RC) methods, which build on the temporal correspondence between the climate model and the corresponding observations to produce reliable forecasts. A broad set of verification metrics has been applied, accounting for different aspects of forecast quality (association, accuracy, discrimination and reliability). We focus precipitation and temperature from the three available C3S seasonal forecasting models (ECMWF-SEAS5, UK Met Office-GloSea5 and Météo France-System5) and validate the raw and calibrated predictions obtained for boreal winter (DJF) and summer (JJA) over two illustrative regions with different skill characteristics (Europe and Southeast Asia).

Our main conclusions are the following:
1) Both approaches (BA and RC) effectively correct the large biases exhibited by raw model predictions, with the corresponding improvement in bias-sensitive metrics such as the Continuous Ranked Probability Score. This is of paramount importance for users, particularly when using climate model outputs to feed their impact models, or when computing climate indices depending on absolute values/thresholds, and proves that some form of calibration is needed prior to the direct use of the raw predictions in sectoral applications.
2) For particular cases, RC methods can outperform BA ones due to an improvement in reliability (other aspects of forecast quality remain unaltered, or are even deteriorated). However, these situations are confined to regions and seasons with high model skill (as shown here for winter temperature in Southeast Asia).
3) As a result of the leave-one year-out cross-validation setting followed here, bias-insensitive measures are in general degraded by all calibration methods (particularly by RC ones), suggesting some degree of over-fitting due to the short hindcast available. A sensitivity analysis with a longer hindcast exhibited smaller degradation, enhancing the improvement of RC results. This indicates that longer hindcast periods than those available in state-of-the-art seasonal forecasting systems (e.g. C3S dataset) are needed for the robust application of RC methods.
4) Within the RC approach, all methods perform similarly, so the particular implementation does not play a key role. Differently, within the BA approach, the EQM method (applied here to seasonal values) is found to perform worse than the simpler MVA, particularly in terms of discrimination. Moreover, there are significant differences among distinct methods in terms of computational cost, being NGR and EQM the most time-consuming ones. This may be especially relevant for the potential usability of the different methods analyzed in real-time applications for climate services.

In this paper we have focused on the calibration of seasonal mean values using both BA and RC methods. However, as opposite to RC, one of the potential advantages of BA methods —not explored in this work— is their suitability for daily data, which is often demanded in a variety of sectoral applications in order to run impact (crop, hydrology, etc.) models or to compute specific indices (heat waves, length of growing index, thermal comfort index, fire weather index, etc.). As a future work, we plan thus to extend the analysis presented here for BA methods to the daily scale.

Finally, we do not analyze here the sensitivity of the results to the observational reference used to calibrate and validate the different methods (instead, we use a single reference dataset, ERA-Interim). However, the results and conclusions may be sensitive to this particular choice (especially in regions with high observational uncertainty) so we plan to undertake a proper assessment of this factor's impact in a future work. Additionally, note that the choice of reference may also affect the comparison across forecasting systems. Therefore, we do not recommend to use the results presented here for a ranking of the different models.

Annex: Description of BA and RC Methods

All the methods described in this Annex have been applied gridbox by gridbox considering seasonal interannual series. We use the following notation: $y_{m,t}$ and $y'_{m,t}$ denote the original and calibrated values of the ensemble member m at time (season/year) t, $\hat{y}$ is the average of the ensemble mean ($\bar{y}_t$) on all times t, $\hat{o}$ is the average of the observations on all times t, σ_f is the standard deviation of the complete ensemble (pooling all member interannual time-series) and σ_o is the standard deviation of the observed interannual time-series. Finally, ρ is the interannual correlation between the ensemble mean and the observational reference.

Mean (and Variance) Adjustment (MVA)

This is the simplest adjustment method, with a long tradition in the context of seasonal forecasting (see, e.g., Leung et al, 1999). The ensemble mean and variance are adjusted towards the corresponding observational ones in the following form:

$$y'_{m,t} = (y_{m,t} - \hat{y})\frac{\sigma_o}{\sigma_f} + \hat{o} \tag{1}$$

A simpler version consists of correcting just the mean (MA) and has the same formulation, but excluding the term σ_o/σ_f.

Empirical Quantile Mapping (EQM)

Despite quantile mapping is commonly applied to daily data, it has been recently used to correct monthly precipitation from the POAMA seasonal forecasting model (Zhao et al, 2017). Here we have considered an empirical quantile mapping (EQM) method participating in the VALUE downscaling intercomparison initiative (Gutiérrez et al, 2018) which has been also recently applied to correct seasonal precipitation forecasts (Manzanas et al, 2018; Manzanas and Gutiérrez, 2018). This method calibrates the predicted empirical probability density function (PDF) by adjusting a number of quantiles based on the empirical observed PDF (Déqué, 2007). In particular, here we adjust percentiles 1 to 99 and linearly interpolate every two consecutive percentiles inside this range. Outside this range, a constant extrapolation (using the correction obtained for the 1st or 99th percentile) is applied. This method was applied here at a ensemble-wise level; that is, the mapping was done based on all contributing members which were pooled

together (all members are supposed to be statistically indistinguishable). Then, the so-obtained unique correction factor was applied to each individual member. Note that ensemble- and member-wise approaches have been recently reported to provide very similar results (Manzanas et al, 2018).

Climate Conserving Recalibration (CCR)

Also known as variance inflation, this method was first introduced in Doblas-Reyes et al (2005). It modifies the predictions to have the same interannual variance as the observational reference, while preserving their interannual correlation, and can be expressed as:

$$y'_{m,t} = \rho \frac{\sigma_o}{std(\bar{y}_t)} \bar{y}_t + \sqrt{1-\rho^2} \frac{\sigma_o}{\sigma_f} (y_{m,t} - \bar{y}_t) + \bar{o} \tag{2}$$

After Weigel et al (2009), this method has been commonly referred to as climate conserving recalibration.

Ratio of Predictable Components (RPC)

We have also considered for this work the method introduced by Eade et al (2014), which uses the ensemble to reduce noise and adjust the forecast variance so that the ratio of predictable components in the model and in the observations is the same (see the paper for details). In particular, they applied the following correction to adjust seasonal forecasts of the North Atlantic Oscillation (NAO), temperature and pressure in the North Atlantic region:

$$y'_{m,t} = \rho \frac{\sigma_o}{std(\bar{y}_t)} (\bar{y}_t - \hat{y}) + \sqrt{1-\rho^2} \frac{\sigma_o}{\sqrt{var(y_{m,t} - \bar{y}_t)}} (y_{m,t} - \bar{y}_t) + \hat{y} + \hat{o} \tag{3}$$

Linear Regression Recalibration (LR)

This method performs a linear regression between the ensemble mean (i.e. the time-series of y_t) and the corresponding observations:

$$o_t = \alpha + \beta \bar{y}_t + \epsilon \tag{4}$$

To correct the forecast variance, the standardized anomalies are rescaled by the standard deviation of the predictive distribution from the linear fit, so $y'_{m,t} = \alpha + \beta \bar{y}_t + \gamma_t (y_{m,t} - \bar{y}_t)$, where

$$\gamma_t = std(\epsilon_{fit}) \sqrt{1 + 1/n + \frac{(y_t - \bar{y}_t)^2}{(n-1) var(\epsilon_{obs})}}, \tag{5}$$

ϵ_{fit} and ϵ_{obs} are the residuals from the regression and the observations respectively and the number of samples used.

Non-Homogeneous Gaussian Regression (NGR)

This method (Gneiting et al, 2005) uses a constant term and the ensemble mean signal as predictors for the calibrated forecast mean and a constant term and the ensemble spread for the inflation (shrinkage) of the ensemble spread. The correction has the following form:

$$y'_{m,t} = \alpha + \beta(\bar{y}_t - \hat{y}) + (y_{m,t} - \bar{y}_t)\sqrt{\gamma^2 + \delta^2 var(y_t)} \quad (6)$$

The parameters α, β, γ and δ are optimized by minimizing the ensemble CRPS. NGR approaches have been applied in many previous works, but mostly in the context of short-term forecasts (see, e.g., Wilks and Hamill, 2007; Thorarinsdottir and Johnson, 2012; Feldmann et al, 2015; Scheuerer and Möller, 2015; Markus et al, 2017). To our knowledge, only Tippett and Barnston (2008) have used it in the context of seasonal forecasting.

Acknowledgements This work has been funded by the C3S activity on Evaluation and Quality Control for seasonal forecasts. JMG was partially supported by the project MULTI-SDM (CGL2015-66583-R, MINECO/FEDER). FJDR was partially funded by the H2020 EUCP project (GA 776613).

References

Barnston AG (1994) Linear statistical short-term climate predictive skill in the northern hemisphere. Journal of Climate 7(10):1513–1564, DOI 10.1175/1520-0442(1994)007⟨1513:LSSTCP⟩2.0.CO;2, URL http://dx.DOI.org/10.1175/1520-0442(1994)007<1513:LSSTCP>2.0.CO;2

Bedia J, Golding N, Casanueva A, Iturbide M, Buontempo C, Gutiérrez JM (2018) Seasonal predictions of Fire Weather Index: Paving the way for their operational applicability in Mediterranean Europe. Climate Services 9:101 – 110, DOI https://doi.org/10.1016/j.cliser.2017.04.001, URL http://www.sciencedirect.com/science/article/pii/S2405880716300826

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Holm EV, Isaksen L, Kallberg P, Koehler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay P, Tavolato C, Thepaut JN, Vitart F (2011) The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society 137(656):553–597, DOI 10.1002/qj.828

Déqué M (2007) Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. Global and Planetary Change 57(1-2):16–26, DOI 10.1016/j.gloplacha.2006.11.030, URL http://www.sciencedirect.com/science/article/pii/S0921818106002748

Doblas-Reyes FJ, Hagedorn R, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting II. Calibration and combination. Tellus A 57(3):234–252, DOI 10.1111/j.1600-0870.2005.00104.x, URL http://dx.doi.org/10.1111/j.1600-0870.2005.00104.x

Doblas-Reyes FJ, García-Serrano J, Lienert F, Biescas AP, Rodrigues LRL (2013) Seasonal climate predictability and forecasting: Status and prospects. Wiley Interdisciplinary Reviews: Climate Change 4(4):245–268, DOI 10.1002/wcc.217, URL http://dx.doi.org/10.1002/wcc.217

Eade R, Smith D, Scaife A, Wallace E, Dunstone N, Hermanson L, Robinson N (2014) Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? Geophysical Research Letters 41(15):5620–5628, DOI 10.1002/2014GL061146, URL http://DOI.wiley.com/10.1002/2014GL061146

Epstein ES (1969) A scoring system for probability forecasts of ranked categories. Journal of Applied Meteorology 8(6):985–987, DOI 10.1175/1520-0450(1969)008⟨0985:ASSFPF⟩2.0.CO;2, URL https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2

Feldmann K, Scheuerer M, Thorarinsdottir TL (2015) Spatial postprocessing of ensemble forecasts for temperature using non-homogeneous gaussian regression. Monthly Weather Review 143(3):955–971, DOI 10.1175/MWR-D-14-00210.1, URL https://doi.org/10.1175/MWR-D-14-00210.1

Gneiting T, Raftery AE, Westveld AH, Goldman T (2005) Calibrated probabilistic forecasting using Ensemble Model Output Statistics and minimum CRPS estimation. Monthly Weather Review 133(5):1098–1118, DOI 10.1175/MWR2904.1, URL https://doi.org/10.1175/MWR2904.1

Gutiérrez JM, Cano R, Cofiño AS, Sordo C (2005) Analysis and downscaling multi-model seasonal forecasts in Peru using self-organizing maps. Tellus A 57(3):435447, DOI 10.1111/j.1600-0870.2005.00128.x

Gutiérrez JM, Maraun D, Widmann M, Huth R, Hertig E, Benestad R, Roessler O, Wibig J, Wilcke R, Kotlarski S, San Martín D, Herrera S, Bedia J, Casanueva A, Manzanas R, Iturbide M, Vrac M, Dubrovsky M, Ribalaygua J, Pórtoles J, Räty O, Räisänen J, Hingray B, Raynaud D, Casado MJ, Ramos P, Zerenner T, Turco M, Bosshard T, Štěp´nek P, Bartholy J, Pongracz R, Keller DE, Fischer AM, Cardoso RM, Soares PMM, Czernecki B, Pagé C (2018) An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor crossvalidation experiment. International Journal of Climatology pp 1–36, DOI 10.1002/joc.5462, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5462

Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting 15(5):559–570, DOI 10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2, URL https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Iturbide M, Bedia J, Herrera S, Baño J, Fernández J, Frías MD, Manzanas R, San-Martín D, Cimadevilla E, Cofiño AS, Gutiérrez JM (2018) climate4R: An R-based framework for climate data access, post-processing and bias correction. Under review in Environmental Modelling and Software

Kharin VV, Zwiers FW (2003) On the ROC score of probability forecasts. Journal of Climate 16(24):4145–4150, DOI 10.1175/1520-0442(2003)016⟨4145:OTRSOP⟩2.0.CO;2, URL http://dx.doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2

Lachenbruch PA, Mickey MR (1968) Estimation of error rates in discriminant analysis. Technometrics 10(1):1–11, DOI 10.2307/1266219, URL http://www.jstor.org/stable/1266219

Leung LR, Hamlet AF, Lettenmaier DP, Kumar A (1999) Simulations of the ENSO hydroclimate signals in the Pacific Northwest Columbia river basin. Bulletin of the American Meteorological Society 80(11):2313–2330, DOI 10.1175/1520-0477(1999)080⟨2313:SOTEHS⟩2.0.CO;2, URL https://doi.org/10.1175/1520-0477(1999)080<2313:SOTEHS>2.0.CO;2

Manzanas R, Frías MD, Cofiño AS, Gutiérrez JM (2014) Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill. Journal of Geophysical Research: Atmospheres 119(4):1708–1719, DOI 10.1002/2013JD020680, URL http://onlinelibrary.wiley.com/doi/10.1002/2013JD020680/abstract

Manzanas R, Gutiérrez JM, Fernández J, van Meijgaard E, Calmanti S, Magariño ME, Cofiño AS, Herrera S (2017) Dynamical and statistical downscaling of seasonal temperature forecasts in Europe: Added value for user applications. Climate Services DOI 10.1016/j.cliser.2017.06.004, URL http://www.sciencedirect.com/science/article/pii/S2405880717300067

Manzanas R, Lucero A, Weisheimer A, Gutiérrez JM (2018) Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts? Climate Dynamics 50(3):1161–1176, DOI 10.1007/s00382-017-3668-z, URL https://link.springer.com/article/10.1007/s00382-017-3668-z

Manzanas R, Gutiérrez JM (2018) Process-conditioned bias correction for seasonal forecasting: a case-study with ENSO in Peru. Climate Dynamics pp 1–11, DOI 10.1007/s00382-018-4226-z, URL https://doi.org/10.1007/s00382-018-4226-z

Maraun D, Wetterhall F, Ireson AM, Chandler RE, Kendon EJ, Widmann M, Brienen S, Rust HW, Sauter T, Themessl M, Venema VKC, Chun KP, Goodess CM, Jones RG, Onof C, Vrac M, Thiele-Eich I (2010) Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. Reviews of Geophysics 48(3):n/a–n/a, DOI 10.1029/2009RG000314, URL http://dx.doi.org/10.1029/2009RG000314

Maraun D, Shepherd TG, Widmann M, Zappa G, Walton D, M GJ, Hagemann S, Richter I, Soares PMM, Hall A, Mearns LO (2017) Towards process-informed bias correction of climate change simulations. Nature Climate Change 7:764–773, DOI 10.1038/nclimate3418

Marcos R, Llasat MC, Quintana-Seguí P, Turco M (2018) Use of bias correction techniques to improve seasonal forecasts for reservoirs: A case-study in northwestern Mediterranean. Science of The Total Environment 610-611:64 – 74, DOI https://doi.org/10.1016/j.scitotenv.2017.08.010, URL http://www.sciencedirect.com/science/article/pii/S0048969717320089

Markus D, J MG, W MJ, Achim Z (2017) Spatial ensemble postprocessing with standardized anomalies. Quarterly Journal of the Royal Meteorological Society 143(703):909–916, DOI 10.1002/qj.2975, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2975

Molteni F, Stockdale T, Balmaseda M, Balsamo G, Buizza R, Ferranti L, Magnusson L, Mogensen K, Palmer T, Vitart F (2011) The new ECMWF seasonal forecast system (System 4). European Centre for Medium-Range Weather Forecasts, URL http://climate.ncas.ac.uk/people/allan/Fire_Risk_Insurance_Papers/Moltini\%20etal\%202011.pdf

Murphy AH (1973) A new vector partition of the probability score. Journal of Applied Meteorology 12(4):595–600, DOI 10.1175/1520-0450(1973)012⟨0595:

ANVPOT)2.0.CO;2, URL https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

Nikulin G, Asharafb S, Magariño ME, Calmanti S, Cardoso RM, Bhend J, Fernández J, Frías MD, Fröhlichb K, Frühb B, Herrera S, Manzanas R, Gutiérrez JM, Hanssona U, Kolaxa M, Liniger M, Soares PMM, Spirig C, Tome R, Wysera K (2018) Dynamical and statistical downscaling of a global seasonal hindcast in eastern Africa. Climate Services 9:72 – 85, DOI https://doi.org/10.1016/j.cliser.2017.11.003, URL http://www.sciencedirect.com/science/article/pii/S2405880717300055

Pavan V, Marchesi S, Morgillo A, Cacciamani C, Doblas-Reyes FJ (2005) Downscaling of DEMETER winter seasonal hindcasts over Northern Italy. Tellus A 57(3):424–434, DOI 10.1111/j.1600-0870.2005.00111.x

Piani C, Haerter JO, Coppola E (2010) Statistical bias correction for daily precipitation in regional climate models over Europe. Theoretical and Applied Climatology 99(1-2):187–192, DOI 10.1007/s00704-009-0134-9, URL http://dx.doi.org/10.1007/s00704-009-0134-9

Sansom PG, Ferro CAT, Stephenson DB, Goddard L, Mason SJ (2016) Best practices for postprocessing ensemble climate forecasts. Part I: Selecting appropriate recalibration methods. Journal of Climate 29(20):7247–7264, DOI 10.1175/JCLI-D-15-0868.1, URL https://doi.org/10.1175/JCLI-D-15-0868.1

Scheuerer M, Möller D (2015) Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. The Annals of Applied Statistics 9(3):1328–1349, DOI 10.1214/15-AOAS843, URL https://doi.org/10.1214/15-AOAS843

Sheau TN, Tangang F, Juneng L (2017) Bias correction of global and regional simulated daily precipitation and surface mean temperature over Southeast Asia using quantile mapping method. Global and Planetary Change 149:79 – 90, DOI https://doi.org/10.1016/j.gloplacha.2016.12.009, URL http://www.sciencedirect.com/science/article/pii/S0921818116301266

Thorarinsdottir TL, Johnson MS (2012) Probabilistic wind gust forecasting using non-homogeneous gaussian regression. Monthly Weather Review 140(3):889–897, DOI 10.1175/MWR-D-11-00075.1, URL https://doi.org/10.1175/MWR-D-11-00075.1

Tippett MK, Barnston AG (2008) Skill of multimodel ENSO probability forecasts. Monthly Weather Review 136(10):3933–3946, DOI 10.1175/2008MWR2431.1

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. Journal of Applied Meteorology and Climatology 56(5):1231–1247, DOI 10.1175/JAMC-D-16-0204.1, URL https://doi.org/10.1175/JAMC-D-16-0204.1

Weigel AP, Liniger MA, Appenzeller C (2009) Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? Monthly Weather Review 137(4):1460–1479, DOI 10.1175/2008MWR2773.1

Weisheimer A, Palmer TN (2014) On the reliability of seasonal climate forecasts. Journal of the Royal Society Interface 11(96), DOI 10.1098/rsif.2013.1162

Wilks DS, Hamill TM (2007) Comparison of Ensemble-MOS methods using GFS reforecasts. Monthly Weather Review 135(6):2379–2390, DOI 10.1175/MWR3402.1, URL https://doi.org/10.1175/MWR3402.1

Zhao T, Bennett JC, Wang QJ, Schepen A, Wood AW, Robertson DE, Ramos MH (2017) How suitable is quantile mapping for postprocessing GCM precipitation

forecasts? Journal of Climate 30(9):3185–3196, DOI 10.1175/JCLI-D-16-0652.1,
URL https://doi.org/10.1175/JCLI-D-16-0652.1

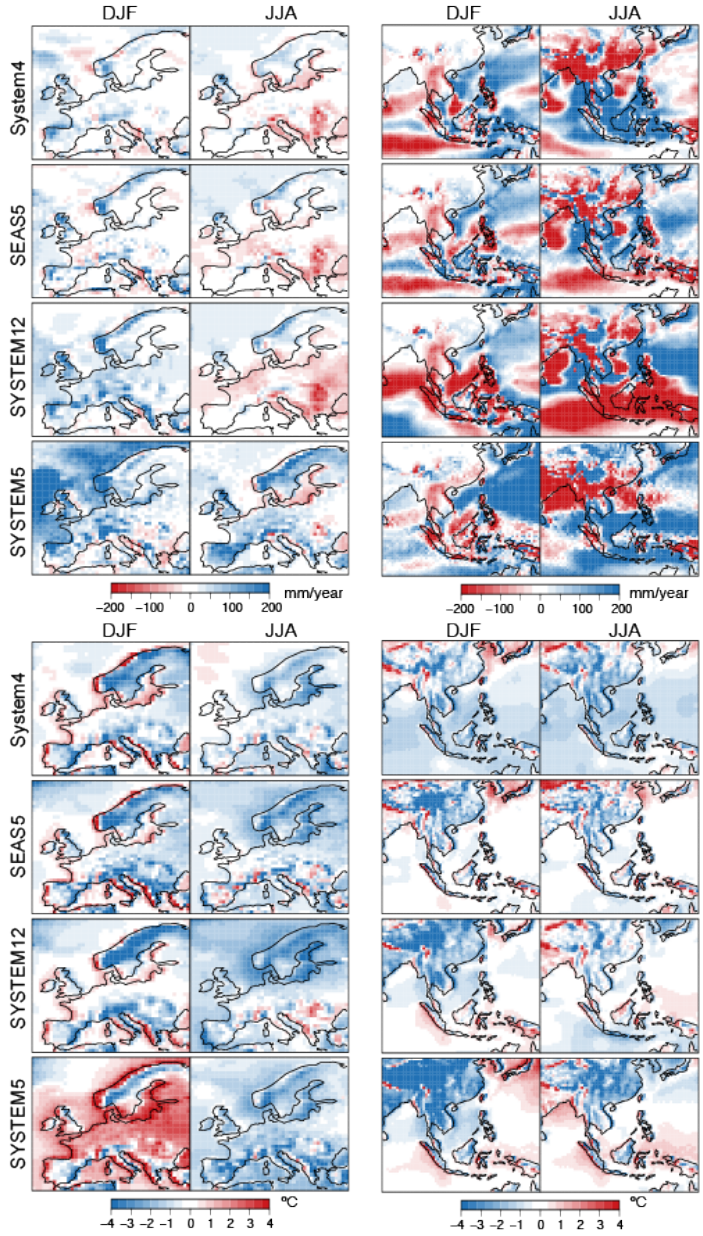

Fig. 1 Bias between the ensemble mean of the four models of Table 1 and ERA-Interim verifying observations for precipitation (top) and temperature (bottom) over EU (left) and SA (right), in DJF and JJA. The errors are expressed as mm/year (°C) for the case of precipitation (temperature).

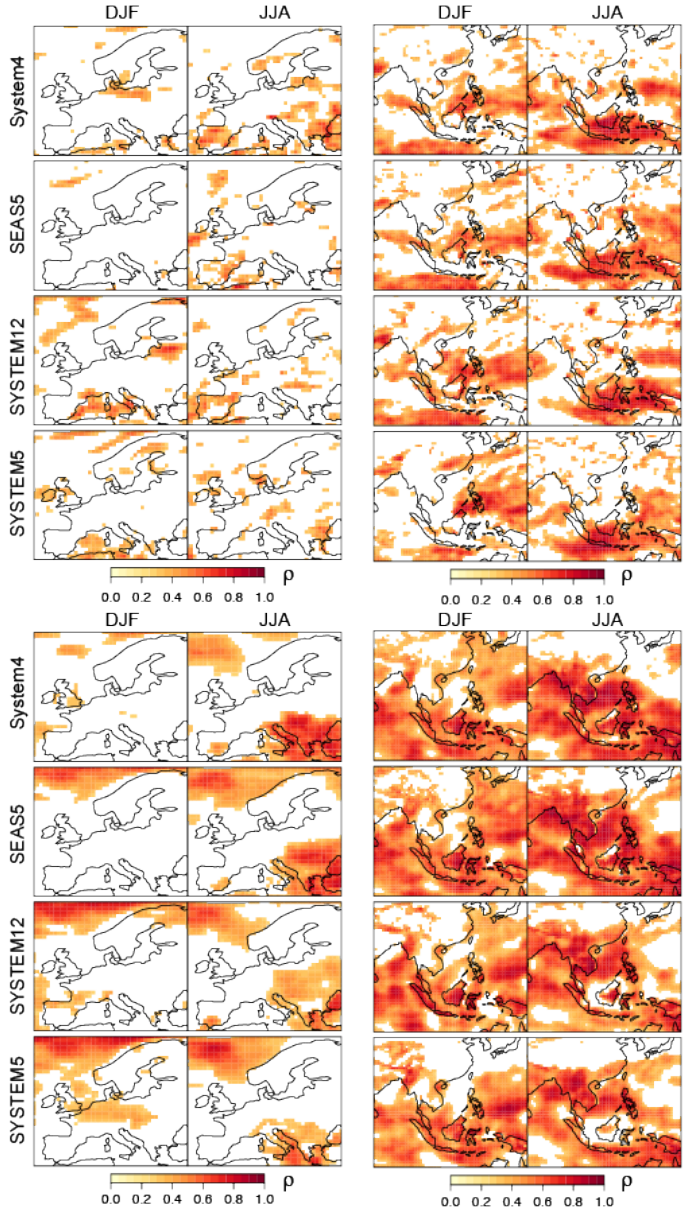

Fig. 2 As Figure 1 but for interannual Pearson correlation. Only significant correlations (90% confidence level, according to a t-test) are shown.

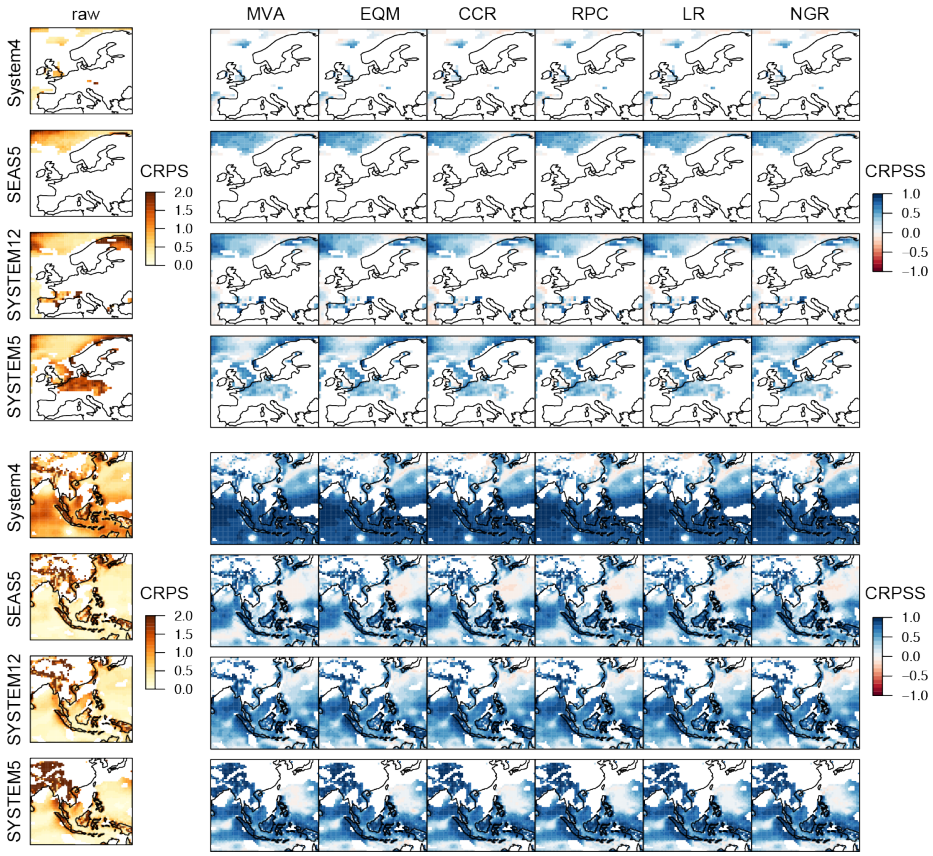

Fig. 3 CRPSS for temperature over EU (top) and SA (bottom) in DJF, as obtained from applying the different BA and RC methods of Table 2 (columns 2-7) to the four models of Table 1 (in rows). In all cases, the CRPS obtained for the raw outputs (column 1) is considered as reference.

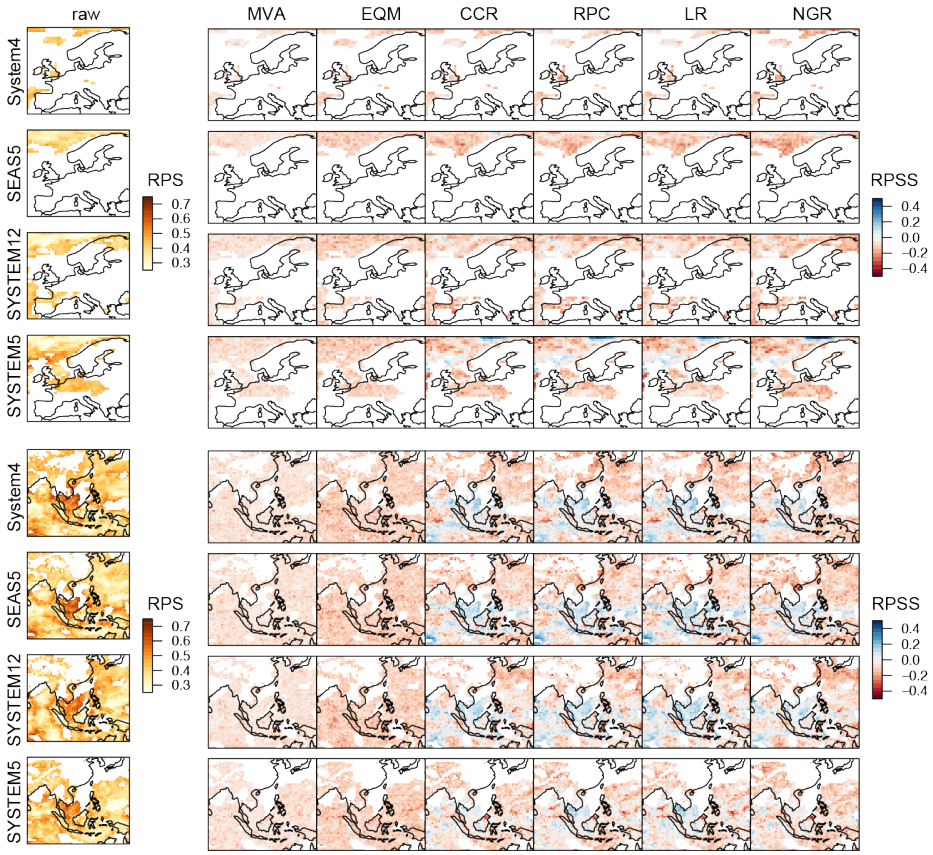

Fig. 4 As Figure 3, but for the RPSS.

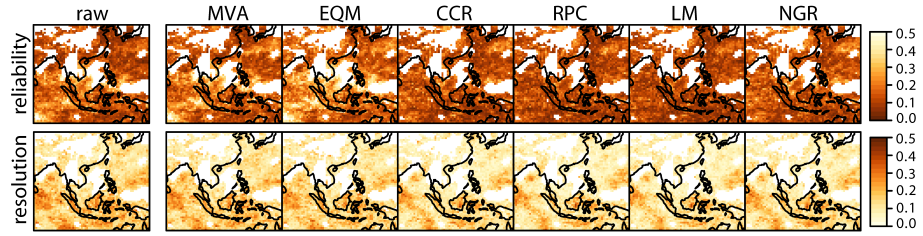

Fig. 5 Reliability and resolution components (top and bottom row, respectively) of the RPS for temperature over SA in DJF, as obtained from applying the BA and RC methods of Table 2 (in columns) to the System4 (12-member, 21-year version).

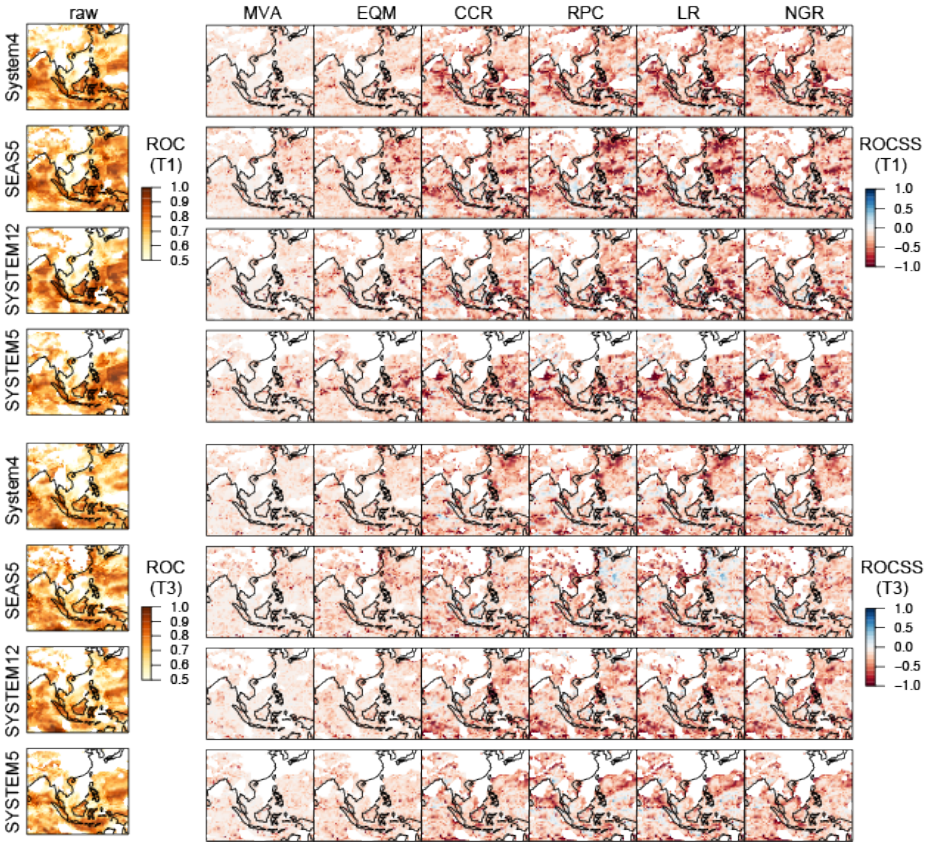

Fig. 6 ROCSS for the cold (top) and warm (bottom) tercile categories of DJF temperature over SA, as obtained from applying the BA and RC methods of Table 1 (columns 2-7) to the four models of Table 1 (in rows). In all cases, the ROC obtained for the raw outputs (column 1) is considered as reference.

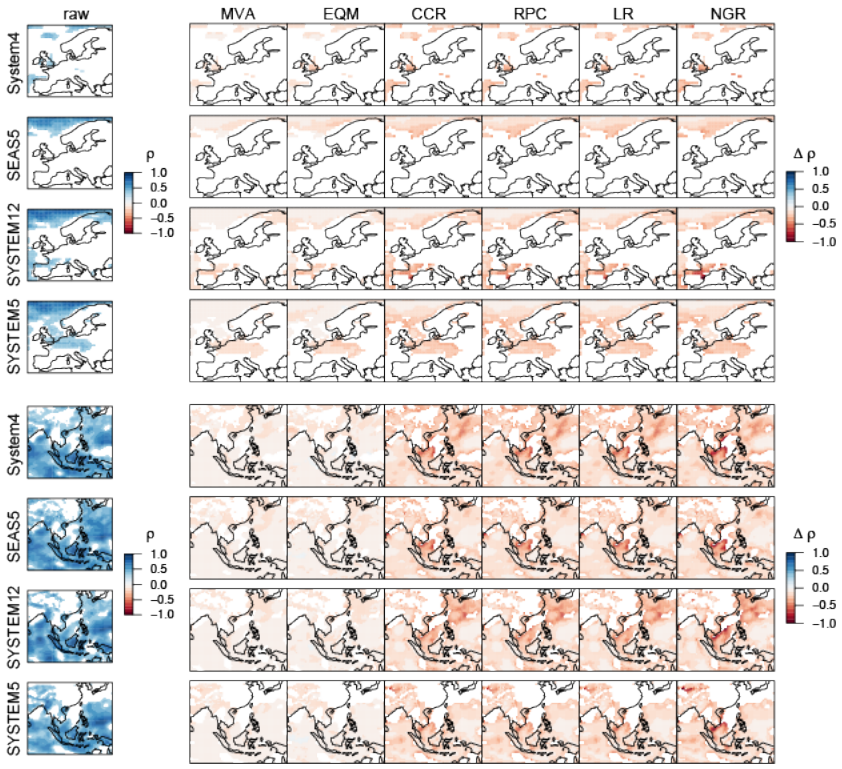

Fig. 7 Column 1: Interannual Pearson correlation between ERA-Interim and the ensemble mean of the four available models (in rows) for DJF temperature, as given by the raw forecasts over EU (top) and SA (bottom). Columns 2-7: Difference (in correlation units) with respect to column 1, as obtained from the application of the BA and RC methods of Table 2.

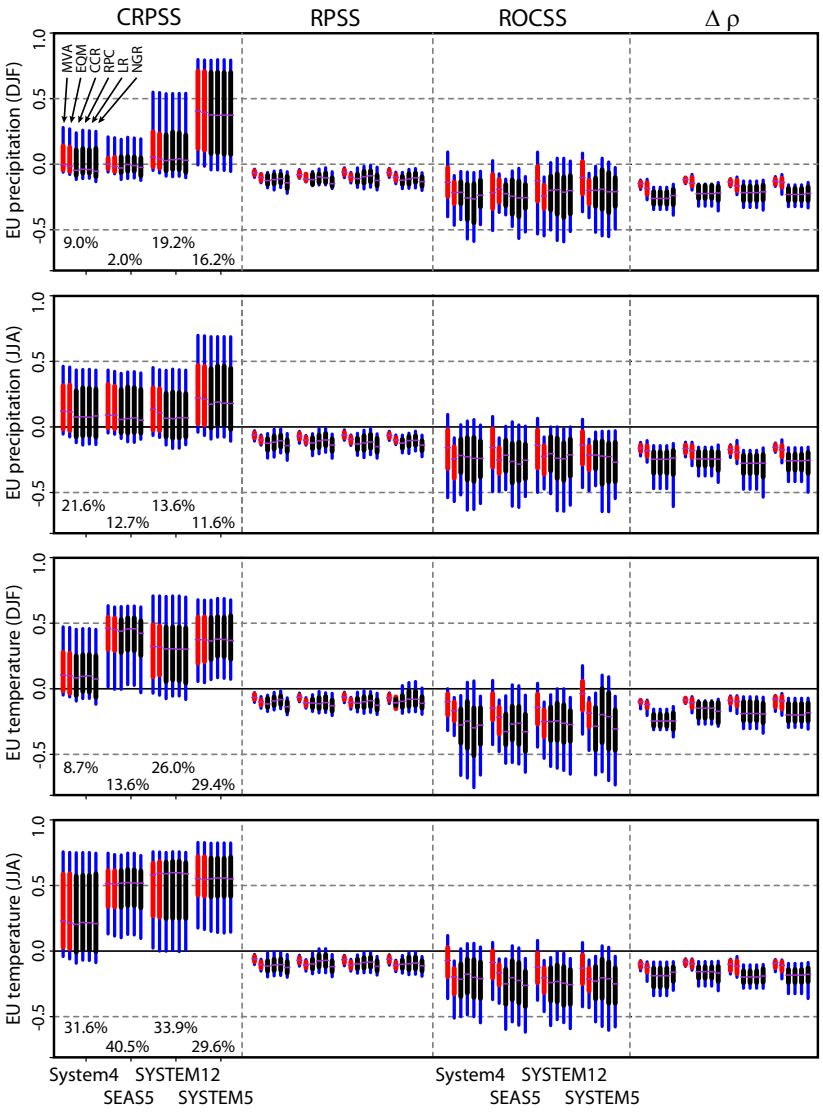

Fig. 8 Summary of the results obtained over EU, in terms of the different skill scores considered (CRPSS, RPSS, ROCSS and correlation differences; in columns). The two variables (precipitation and temperature) and seasons (DJF and JJA) analyzed are shown in different rows. In all cases, results for the four available models (System4, SEAS5, SYTEM12 and SYSTEM5) are displayed along the x-axis. For each model, the two (four) red (black) boxplots indicate the P25-75 range for each BA (RC) method, with blue corresponding to the P10-P90 range. The numbers in the first column correspond to the percentage of skillful gridboxes over which the methods were applied and tested (see Figure 2).

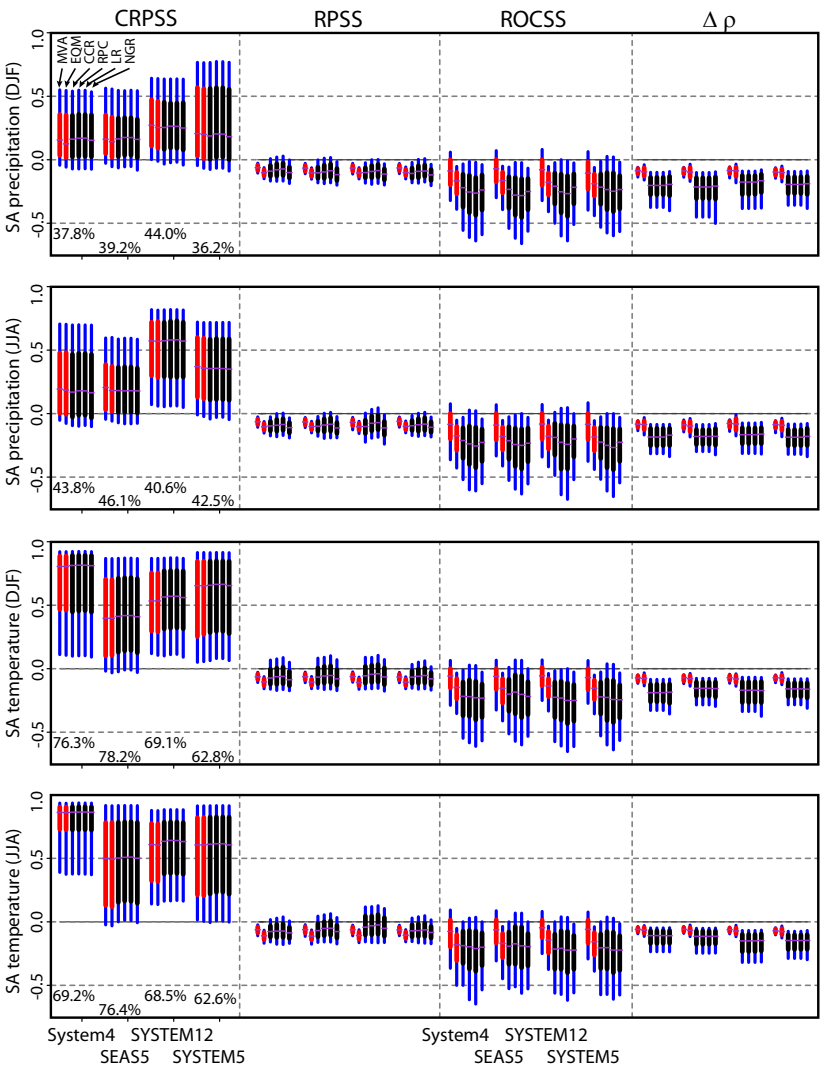

Fig. 9 As Figure 8, but for SA.

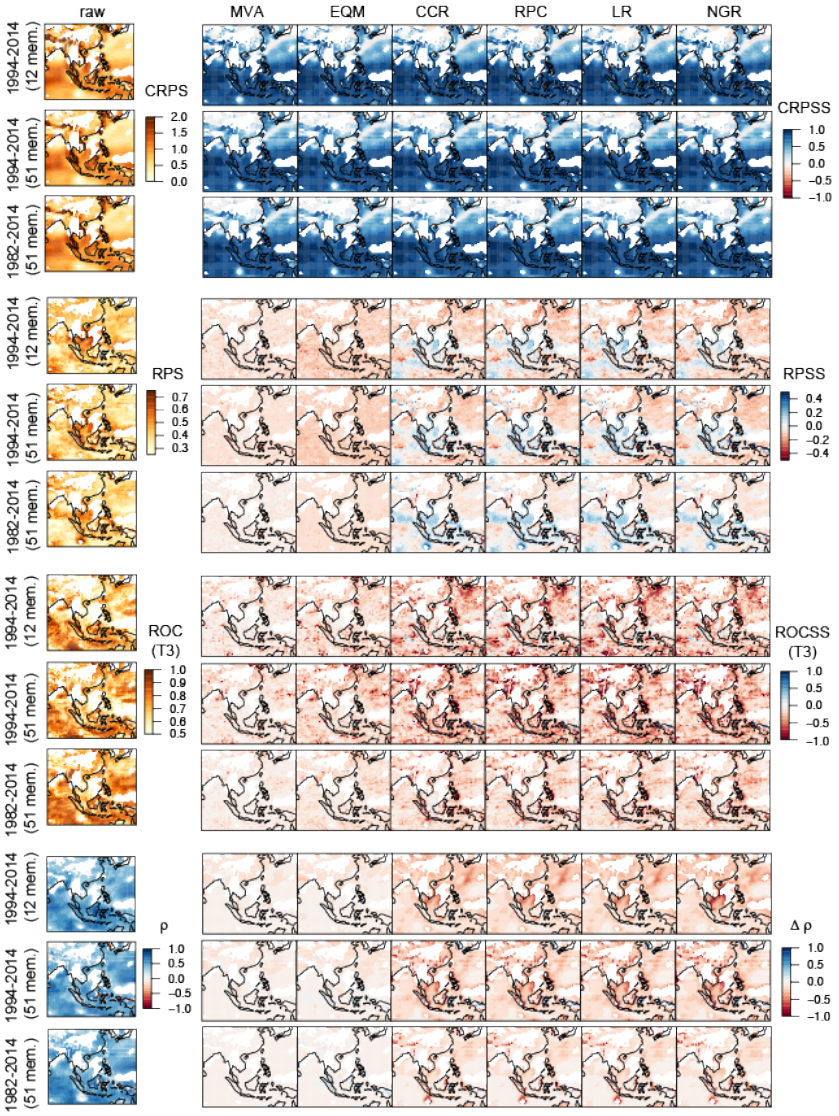

Fig. 10 Results obtained for the CRPSS, the RPSS, the ROCSS (only for the warm tercile category) and the interannual Pearson correlation —in different panels from top to bottom— for temperature over SA in DJF, as obtained from applying the BA and RC methods of Table 2 to the System4. Within each panel, the top row corresponds to a 12-member ensemble for the period 1994-2014 (same as in Figures 3, 4, 6 and 7, displayed here again to facilitate comparison). The middle (bottom) row correspond to a 51-ensemble member for 1994-2014 (1982-2014).

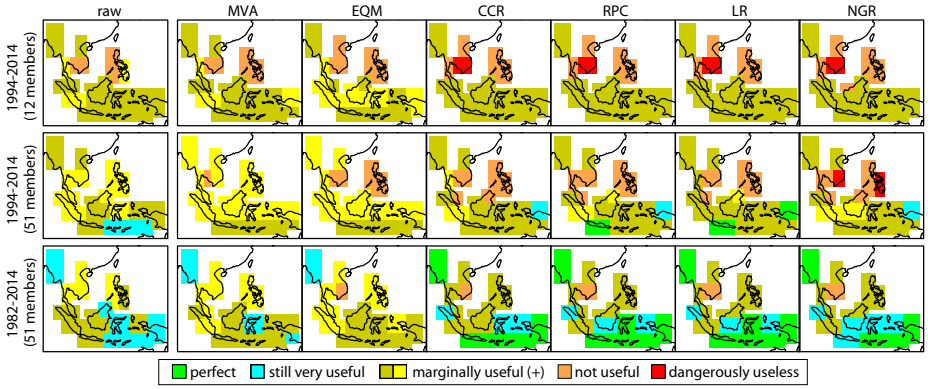

Fig. 11 Reliability categories, as obtained from applying the different BA and RC methods of Table 2 to correct DJF temperature from the System4 over SA. The top row corresponds to a 12-member ensemble for the period 1994-2014. Middle (bottom) row correspond to a 51-ensemble member for 1994-2014 (1982-2014).

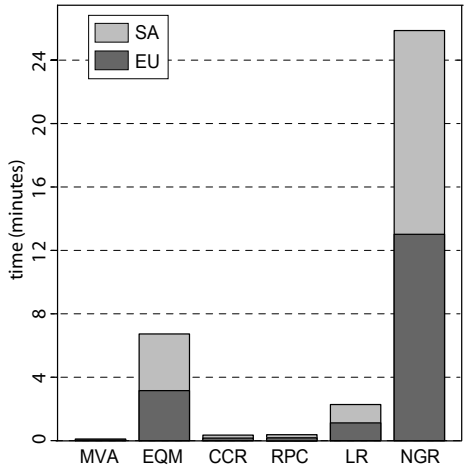

Fig. 12 Execution times —in a personal computer— for the different BA and RC methods of Table 2 for the illustrative case of temperature in DJF over EU (dark gray) and SA (light gray) for System4 (12-member and 21-year version), according to their implementation in the R-package *calibratoR*.