

[Click here to view linked References](#)

Climate Dynamics manuscript No.
(will be inserted by the editor)

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

4 **R. Manzananas · J. M. Gutiérrez · J.**
5 **Bhend · S. Hemri · F. J. Doblas-Reyes ·**
6 **V. Torralba · E. Penabad · A. Brookshaw**

7
8 Received: date / Accepted: date

9 **Abstract** This work presents a comprehensive intercomparison of different alter-
10 natives for the calibration of seasonal forecasts, ranging from simple bias adjust-
11 ment (BA) —e.g. quantile mapping— to more sophisticated ensemble recalibration
12 (RC) methods —e.g. non-homogeneous Gaussian regression,— which build on the
13 temporal correspondence between the climate model and the corresponding obser-
14 vations to generate reliable predictions. To be as critical as possible, we validate
15 the raw model and the calibrated forecasts in terms of a number of metrics which
16 take into account different aspects of forecast quality (association, accuracy, dis-
17 crimination and reliability). We focus on one-month lead forecasts of precipitation
18 and temperature from four state-of-the-art seasonal forecasting systems, three of
19 them included in the Copernicus Climate Change Service (C3S) dataset (ECMWF-
20 SEAS5, UK Met Office-GloSea5 and Météo France-System5) for boreal winter and
21 summer over two illustrative regions with different skill characteristics (Europe and
22 Southeast Asia). Our results indicate that both BA and RC methods effectively
23 correct the large raw model biases, which is of paramount importance for users,
24 particularly when directly using the climate model outputs to run impact mod-
25 els, or when computing climate indices depending on absolute values/thresholds.
26 However, except for particular regions and/or seasons (typically with high skill),
27 there is only marginal added value —with respect to the raw model outputs—

R. Manzananas (✉) · 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, Zurich, Switzerland

F. J. Doblas-Reyes
Barcelona Supercomputing Center (BSC), Barcelona, Spain
ICREA, Pg. Lluís Companys 23 08010, Barcelona, Spain

V. Torralba
Barcelona Supercomputing Center (BSC), Barcelona, Spain

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

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.

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). As a result, the use of raw model outputs poses a risk for 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 to make the raw model seasonal forecasts usable, 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 paid to-date to 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. 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, since their application is rather straightforward, BA methods

73 may be used in an uninformed way in some cases. For instance, as a result of
74 inheriting the model circulation biases —e.g., errors in the position of the inter-
75 tropical convergence zone,— these techniques may lead to meaningless results for
76 some regions (Maraun et al, 2017).

77 RC methods transform the raw model outputs building on the temporal cor-
78 respondence between the ensemble mean predictions and the corresponding ob-
79 servations (see Sansom et al, 2016, for a comprehensive review). They range from
80 relatively simple implementations such as climate conserving recalibration —CCR
81 (see, e.g., Doblas-Reyes et al, 2005; Weigel et al, 2009)— or the ratio of predictable
82 components —RPC (Eade et al, 2014)— to more general ensemble model output
83 statistics (EMOS) methods such as non-homogeneous Gaussian regression (see,
84 e.g., Gneiting et al, 2005; Tippett and Barnston, 2008; Sansom et al, 2016). The
85 main advantage of RC techniques for seasonal forecasting is that they are de-
86 signed to produce reliable predictions. However, as opposite to BA methods, they
87 are not suitable for the adjustment of daily data since, at this particular time-
88 scale, the signal-to-noise ratio of seasonal forecasts starts to sharply decrease a
89 few weeks after the initialization moment. Therefore, as a result of working with
90 monthly/seasonal data, an important constraint of RC techniques in the context
91 of seasonal forecasting is that the underlying parameters have to be estimated
92 using a limited amount of data and, as a consequence, they are prone to over-
93 fitting —due to the enormous computational requirements and the lack of long
94 observational datasets required to initialize the forecasting system, state-of-the-
95 art hindcasts typically have around 30 years of data (i.e. a sample size of 30 values
96 for calibration).— Note that, as we focus here on the adjustment of seasonal means,
97 the BA methods used in this work suffer from the same constraint mentioned for
98 the RC ones. However, this could be avoided in BA methods when working with
99 daily data.

100 Recent studies have reported some limitations for BA methods (Manzanas et al,
101 2018) for seasonal forecasting, and even the preferable choice of RC techniques
102 (Zhao et al, 2017). However, to the authors’ knowledge, there is no comprehensive
103 intercomparison of BA and RC methods for this type of predictions. The main
104 goal of this paper is therefore to fill this knowledge gap. To do this, we apply a
105 set of state-of-the-art BA and RC methods to calibrate one-month lead seasonal
106 predictions of temperature and precipitation from four different forecast systems.
107 Three of these systems are included in the Copernicus Climate Change Service
108 (C3S) seasonal service (<http://climate.copernicus.eu/seasonal-forecasts>),
109 whereas the fourth (the ECMWF System4) is used to test the sensitivity of the
110 results obtained to the hindcast length and the ensemble size. The raw model and
111 calibrated predictions are validated in terms of a number of verification metrics
112 which take into account different aspects of forecast quality (association, accuracy,
113 discrimination and reliability).

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

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

119 2 Data and Methods

120 2.1 Data Used

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

126 We analyze one-month lead seasonal forecasts (i.e. predictions initialized in
 127 November and May for DJF and JJA, respectively) from the C3S seasonal multi-
 128 system ensemble —which consists of three state-of-the-art models with a common
 129 hindcast period of 22 years, 1993-2014 (see Table 1)— together with the ECMWF-
 130 System4 (Molteni et al, 2011) —which provides the longest available hindcast,
 131 starting in 1981.— For the C3S models, a total of 21 seasons are available for DJF
 132 (starting with D1993-JF1994, which we refer to as DJF 1994). Therefore, for the
 133 sake of comparability we use a common 21-year period for both DJF and JJA, and
 134 only the 12 first members of each model (minimum number of members common
 135 across all models) are considered. Additionally, in order to test the sensitivity of
 136 the results obtained to the ensemble size and the hindcast length, we have also
 137 used the full hindcast period and the 51-member version of the ECMWF-System4
 138 (see Section 3.3).

139 The ERA-Interim reanalysis (Dee et al, 2011) is used as observational refer-
 140 ence dataset for both the calibration of the BA and RC methods and also for
 141 the verification of all seasonal forecasts involved in this work. ECMWF-System4
 142 and ERA-Interim have been bi-linearly interpolated from their native horizontal
 143 resolutions to the common 1° regular grid in which the C3S models are pro-
 144 vided through the Climate Data Store (see <https://climate.copernicus.eu/seasonal-forecasts>).

146 2.2 Bias Adjustment and Ensemble Recalibration Methods

147 Table 2 shows the BA and RC methods intercompared in this work (see the Annex
 148 for details on the particular implementations), which have been already used in
 149 the context of seasonal forecasting (see the references in the third column of the
 150 table). On the one hand, two BA methods were considered: a simple mean and
 151 variance adjustment (MVA) and an empirical quantile mapping (EQM) which
 152 adjusts percentiles 1 to 99. Note that we also considered an even simpler method
 153 consisting of adjusting only the mean; however, the results were very similar to
 154 those obtained for MVA and are thus not shown for brevity. Also, for EQM,

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 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)

155 we tested the suitability of both monthly and seasonal data for the mapping,
 156 obtaining very similar conclusions in both cases. For coherence with the rest of
 157 methods, we only show results for the case of seasonal values. On the other hand,
 158 four RC methods were considered: climate conserving recalibration (CCR), ratio of
 159 predictable components (RPC) and two EMOS choices using linear regression (LR)
 160 and non-homogeneous Gaussian regression (NGR). Whilst CCR and RPC only use
 161 statistics of the predicted ensemble and the observations, LR and NGR also involve
 162 some parameters which need to be estimated by regression/optimization taking
 163 into account the correlation between the raw ensemble mean and observations.
 164 Note however that, although more sophisticated RC methods exist (Sansom et al,
 165 2016), we have selected here some standard parsimonious ones (already used in
 166 seasonal forecasting studies) which are preferable to avoid overfitting problems.

167 All the methods considered for this work (with the exception of EQM) have
 168 been implemented in an R-package called *calibratoR* ([http://github.com/SantanderMetGroup/
 169 calibratoR](http://github.com/SantanderMetGroup/calibratoR)), which is publicly available as part of the *climate4R* framework
 170 (Iturbide et al, 2019). The method EQM (as well as other BA and downscal-
 171 ing techniques) are available in the *downscaleR* package ([http://github.com/
 172 SantanderMetGroup/downscaleR](http://github.com/SantanderMetGroup/downscaleR)), which is also part of *climate4R*. Here, all the
 173 BA and RC methods have been applied at a gridbox level considering seasonal
 174 interannual time-series. The statistics/parameters involved in each method are
 175 first obtained based on the complete ensemble and subsequently applied to ad-
 176 just/calibrate each individual member —a detailed description of each method is
 177 given in the Annex.— Moreover, all methods are applied under a leave-one year-
 178 out (LOO) cross-validation scheme (Lachenbruch and Mickey, 1968). Note that
 179 proper cross-validation is mandatory in order to avoid artificial skill (Manzanas
 180 et al, 2017), especially when working with small sample sizes like in this case (21
 181 years of data).

182 2.3 Forecast Quality Metrics

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

189 Accuracy measures the average distance between forecasts and observations.
 190 We consider here two standard scores which are typically used to characterize

191 this property: the Continuous Ranked Probability Score (CRPS) and the Ranked
 192 Probability Score (RPS). The CRPS (Hersbach, 2000) is a metric that allows to
 193 assess the performance of probabilistic forecasts of a continuous variable based on
 194 the integrated squared difference between the observed and the predicted cumu-
 195 lative distribution functions (which would correspond to the mean absolute error
 196 for deterministic forecasts). The perfect value for this score is therefore 0. To allow
 197 for direct comparison across the different BA and RC methods, we also use the
 198 associated skill score (CRPSS), which is computed as $1 - (CRPS_{cal}/CRPS_{ref})$,
 199 being $CRPS_{ref}$ the CRPS obtained for the raw model forecasts and $CRPS_{cal}$ the
 200 one for the calibrated predictions. The RPS (Epstein, 1969) is the discrete version
 201 of the CRPS and measures the sum of squared differences in cumulative proba-
 202 bility space for a multi-category probabilistic forecast (for two-category forecasts,
 203 it would be the Brier Score), being thus its perfect value 0. As for the case of the
 204 CRPS, we also use here the associated skill score (the RPSS), which is computed
 205 as $1 - (RPS_{cal}/RPS_{ref})$.

206 Discrimination measures the ability of the forecasts to distinguish between an
 207 event and the corresponding non-event, which is assessed here by means of the area
 208 under the ROC curve Kharin and Zwiers (2003) (simply referred to as ROC here-
 209 after). Again, in addition to the direct score (whose perfect is 1), we also use the as-
 210 sociated skill score (ROCSS), which is computed as $(ROC_{cal} - ROC_{ref})/1 - ROC_{ref}$.
 211 This metric is recommended by the Lead Centre for the Standardized Verification
 212 System of Long Range Forecasts and has been used in many previous studies for
 213 the verification of seasonal forecasts (see, e.g., Manzanas et al, 2014).

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

226 Note that, as a result of using skill scores (instead of the direct scores), CRPSS,
 227 RPSS and ROCSS values above (below) 0 indicate that the particular calibration
 228 method improves (degrades) the raw model prediction. Moreover, RPS and ROC
 229 are used here for tercile-based probabilistic predictions. In both cases, terciles are
 230 independently computed for the observations and the predictions, which implic-
 231 itly introduces a bias adjustment in the forecasts. Therefore, as opposite to CRPS,
 232 these two metrics are bias-insensitive, allowing thus to explore the added value of
 233 BA and RC method beyond the expected (by construction) model bias reduc-
 234 tion. Also, whereas CRPS and RPS are sensitive to changes in reliability, ROC is
 235 not. Thus, the latter also allows to assess the potential usefulness of the different
 236 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 ($200\text{mm}/\text{season}$) for temperature (precipitation) in many gridboxes. 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 $15\text{mm}/\text{season}$ 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 run impact models, or when computing climate indices depending on absolute values/thresholds, and proves that some form of calibration is needed to make the raw model predictions usable.

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). For this reason, Figure 2 shows the interannual Pearson correlation between the raw 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 (“non-skillful” gridboxes are depicted in white). 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 has been found for the RC methods

(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 skillful gridboxes. 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 cannot 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 RPS 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 partly 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), which is a consequence of the the LOO cross-validation setting used here (see, e.g., Smith et al, 2013). Indeed, 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. The four available models are displayed along the x-axis, with a boxplot for each of the methods considered. In particular, for BA (RC) methods, the red (black) boxplots show the interquartile range (P25-P75) of the values found along all the skillful gridboxes within the region. Blue boxplots represent 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 (as reflected by the CRPSS), whilst yielding a smaller degradation of accuracy and association measures. The same conclusions hold for SA, although in this case (with high skill in some regions) RC methods yield slightly better reliability —as represented by the RPSS— than BA ones for temperature. Nevertheless, as shown in the next section, the poor performance of RC methods (as compared to BA ones) can be partially due to the short hindcast period available here.

3.3 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. This result is in agreement with those found in Smith et al (2013), who also noted the importance of having large enough sample sizes to robustly estimate the post-processing parameters. 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

374 are only shown for the warm tercile category (T3). Reliability is computed for each
375 of the 20 subregions introduced in Figure 1 of Sheau et al (2017), provided there
376 is at least a 25% of points with significantly positive interannual correlations for
377 the ensemble mean (see Figure 2). Within each subregion, we pool together all
378 gridboxes for both observations and predictions.

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

387 3.4 Computational Requirements

388 Although not strictly decisive from a scientific point of view, the analysis of the
389 computational requirements demanded by the different methods is important from
390 a practical perspective, especially regarding their potential usability for climate
391 services and other user-tailored applications. Figure 12 shows, for the illustrative
392 case of DJF temperature from the System4 (12-member, 21-year version), the
393 execution times (in minutes) required by the BA and RC methods used in this
394 work, according to their implementation in *calibratoR* —these times have been
395 computed in a personal computer with two cores and two CPUs (3Ghz) attached
396 to each core, with a RAM memory of 16Gb.— Dark (light) gray correspond to
397 the LOO cross-validation setting used here for EU (SA) —note that computing
398 times drastically reduce if cross-validation is not applied; not shown.— Among the
399 BA methods, MVA is very rapid, being therefore a suitable option for real-time
400 applications (e.g. interactive webpages). In particular, it is much much faster than
401 EQM, which is widely used nowadays for different sectoral tasks. Among the RC
402 methods, CCR and RPC are computationally inexpensive choices (also potentially
403 exploitable in real-time applications), with LR still providing reasonable times (less
404 than 2 minutes for EU). Differently, the long execution times required by NGR
405 make this method unusable for real-time operations. In the light of these results,
406 MVA and/or CCR could be considered as benchmarking methods which provide a
407 good compromise between performance and computational cost for the calibration
408 of seasonal mean values.

409 4 Discussion and Conclusions

410 This work presents a comprehensive intercomparison of different alternatives for
411 the calibration of seasonal forecasts, ranging from simple bias adjustment (BA)
412 to more sophisticated ensemble recalibration (RC) methods, which build on the
413 temporal correspondence between the climate model and the corresponding ob-
414 servations to produce reliable forecasts. A broad set of verification metrics has
415 been applied, accounting for different aspects of forecast quality (association, ac-
416 curacy, discrimination and reliability). We focus on precipitation and temperature

417 from the three available C3S seasonal forecasting models (ECMWF-SEAS5, UK
418 Met Office-GloSea5 and Météo France-System5) plus the ECMWF-System4 and
419 validate the raw and calibrated predictions obtained for boreal winter (DJF) and
420 summer (JJA) over two illustrative regions with different skill characteristics (Eu-
421 rope and Southeast Asia).

422 Our main conclusions are the following:

423 1) Both approaches (BA and RC) effectively correct the large biases exhibited by
424 raw model predictions, with the corresponding improvement in bias-sensitive met-
425 rics such as the Continuous Ranked Probability Score. This is of paramount impor-
426 tance for users, particularly when using climate model outputs to run impact mod-
427 els, or when computing climate indices depending on absolute values/thresholds,
428 and proves that some form of calibration is needed to make the raw model predic-
429 tions usable for sectoral applications.

430 2) For particular cases, RC methods can outperform BA ones due to an improve-
431 ment in reliability (other aspects of forecast quality remain unaltered, or are even
432 deteriorated). However, these situations are confined to regions and seasons with
433 high model skill (as shown here for winter temperature in Southeast Asia).

434 3) As a result of the leave-one year-out cross-validation setting followed here,
435 bias-insensitive measures are in general degraded by all calibration methods (par-
436 ticularly by RC ones), suggesting some degree of over-fitting due to the short
437 hindcast available. A sensitivity analysis with a longer hindcast exhibited smaller
438 degradation, enhancing the improvement of RC results. This indicates that longer
439 hindcast periods than those available in state-of-the-art seasonal forecasting sys-
440 tems (e.g. C3S dataset) are needed for the robust application of RC methods.

441 4) Within the RC approach, all methods perform similarly, so the particular im-
442 plementation does not play a key role. Differently, within the BA approach, the
443 EQM method (applied here to seasonal values) is found to perform worse than the
444 simpler MVA, particularly in terms of discrimination. Note however that, when
445 applied on daily data, the EQM could potentially provide some added value (as
446 compared to the MVA) for the adjustment of extremes or threshold-based indi-
447 cators. Finally, there are significant differences among distinct methods in terms
448 of computational cost, being NGR and EQM the most time-consuming ones. This
449 may be especially relevant for the potential usability of the different methods an-
450 alyzed in real-time applications for climate services.

451

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

460 Finally, we do not analyze here the sensitivity of the results to the observational
461 reference used to calibrate and validate the different methods (see, e.g., Kotlarski
462 et al, 2017; Herrera et al, 2018); instead, we use a single reference dataset, ERA-
463 Interim. However, the results and conclusions may be sensitive to this particular
464 choice (especially in regions with high observational uncertainty) so we plan to un-
465 dertake a proper assessment of this factor’s impact in a future work. Additionally,

466 note that the choice of reference may also affect the comparison across forecasting
 467 systems. Therefore, we do not recommend to use the results presented here for a
 468 ranking of the different models.

469 **Annex: Description of BA and RC Methods**

470 All the methods described in this Annex have been applied gridbox by gridbox
 471 considering seasonal interannual series. We use the following notation: $y_{m,t}$ and
 472 $y'_{m,t}$ denote the original and calibrated values for the ensemble member m at time
 473 (season/year) t , \hat{y} is the average of the ensemble mean (\bar{y}_t) on all times t , \hat{o} is
 474 the average of the observations on all times t , σ_f is the standard deviation of the
 475 complete ensemble (pooling all member interannual time-series) and σ_o is the stan-
 476 dard deviation of the observed interannual time-series. Finally, ρ is the interannual
 477 correlation between the ensemble mean and the observational reference.

478 Mean (and Variance) Adjustment (MVA)

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

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

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

484 Empirical Quantile Mapping (EQM)

485 We have considered an empirical quantile mapping (EQM) method participating in
 486 the VALUE downscaling intercomparison initiative (Gutiérrez et al, 2018) which
 487 has been recently applied to correct seasonal precipitation forecasts (Manzananas
 488 et al, 2018; Manzananas and Gutiérrez, 2018). This method calibrates the predicted
 489 empirical probability density function (PDF) by adjusting a number of quantiles
 490 based on the empirical observed PDF (Déqué, 2007). In particular, here we adjust
 491 percentiles 1 to 99 and linearly interpolate every two consecutive percentiles inside
 492 this range. Outside this range, a constant extrapolation (using the correction ob-
 493 tained for the 1st or 99th percentile) is applied. This method was applied here at
 494 a ensemble-wise level; that is, the mapping was trained based on all contributing
 495 members which were pooled together (all members are supposed to be statistically
 496 indistinguishable). Then, the so-obtained unique correction factor was applied to
 497 each individual member. Note that ensemble- and member-wise approaches have
 498 been recently reported to provide very similar results (Manzananas et al, 2018).

499 Climate Conserving Recalibration (CCR)

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

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

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

506 Ratio of Predictable Components (RPC)

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

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

513 Linear Regression Recalibration (LR)

514 This method performs a linear regression between the ensemble mean (i.e. the
 515 time-series of \bar{y}_t) and the corresponding observations:

$$o_t = \alpha + \beta \bar{y}_t + \epsilon \quad (4)$$

516 To correct the forecast variance, the standardized anomalies are rescaled by
 517 the standard deviation of the predictive distribution from the linear fit, so $y'_{m,t} =$
 518 $\alpha + \beta \bar{y}_t + \gamma_t (y_{m,t} - \bar{y}_t)$, where

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

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

521 Non-Homogeneous Gaussian Regression (NGR)

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

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

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

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

536 **References**

- 537 Barnston AG (1994) Linear statistical short-term climate predictive skill in
 538 the northern hemisphere. *Journal of Climate* 7(10):1513–1564, DOI 10.
 539 1175/1520-0442(1994)007<1513:LSSTCP>2.0.CO;2, URL [http://dx.doi.org/
 540 10.1175/1520-0442\(1994\)007<1513:LSSTCP>2.0.CO;2](http://dx.doi.org/10.1175/1520-0442(1994)007<1513:LSSTCP>2.0.CO;2)
- 541 Bedia J, Golding N, Casanueva A, Iturbide M, Buontempo C, Gutiérrez JM (2018)
 542 Seasonal predictions of Fire Weather Index: Paving the way for their operational
 543 applicability in Mediterranean Europe. *Climate Services* 9:101 – 110, DOI [https:
 544 //doi.org/10.1016/j.cliser.2017.04.001](https://doi.org/10.1016/j.cliser.2017.04.001), URL [http://www.sciencedirect.com/
 545 science/article/pii/S2405880716300826](http://www.sciencedirect.com/science/article/pii/S2405880716300826)
- 546 Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U,
 547 Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L,
 548 Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L,
 549 Healy SB, Hersbach H, Holm EV, Isaksen L, Kallberg P, Koehler M, Matricardi
 550 M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay
 551 P, Tavolato C, Thepaut JN, Vitart F (2011) The ERA-Interim reanalysis: Con-
 552 figuration and performance of the data assimilation system. *Quarterly Journal
 553 of the Royal Meteorological Society* 137(656):553–597, DOI 10.1002/qj.828
- 554 Déqué M (2007) Frequency of precipitation and temperature extremes over
 555 France in an anthropogenic scenario: Model results and statistical correction
 556 according to observed values. *Global and Planetary Change* 57(1-2):16–26,
 557 DOI 10.1016/j.gloplacha.2006.11.030, URL [http://www.sciencedirect.com/
 558 science/article/pii/S0921818106002748](http://www.sciencedirect.com/science/article/pii/S0921818106002748)
- 559 Doblas-Reyes FJ, Hagedorn R, Palmer TN (2005) The rationale behind the success
 560 of multi-model ensembles in seasonal forecasting – II. Calibration and combi-
 561 nation. *Tellus A* 57(3):234–252, DOI 10.1111/j.1600-0870.2005.00104.x, URL
 562 <http://dx.doi.org/10.1111/j.1600-0870.2005.00104.x>

- 563 Doblás-Reyes FJ, García-Serrano J, Lienert F, Biescas AP, Rodrigues LRL (2013)
564 Seasonal climate predictability and forecasting: Status and prospects. *Wiley*
565 *Interdisciplinary Reviews: Climate Change* 4(4):245–268, DOI 10.1002/wcc.217,
566 URL <http://dx.doi.org/10.1002/wcc.217>
- 567 Eade R, Smith D, Scaife A, Wallace E, Dunstone N, Hermanson L, Robinson N
568 (2014) Do seasonal-to-decadal climate predictions underestimate the predictabil-
569 ity of the real world? *Geophysical Research Letters* 41(15):5620–5628, DOI
570 10.1002/2014GL061146, URL <http://DOI.wiley.com/10.1002/2014GL061146>
- 571 Epstein ES (1969) A scoring system for probability forecasts of ranked
572 categories. *Journal of Applied Meteorology* 8(6):985–987, DOI 10.1175/
573 1520-0450(1969)008<0985:ASSFPF>2.0.CO;2, URL [https://doi.org/10.1175/
574 1520-0450\(1969\)008<0985:ASSFPF>2.0.CO;2](https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2)
- 575 Feldmann K, Scheuerer M, Thorarinsdottir TL (2015) Spatial postprocessing of
576 ensemble forecasts for temperature using non-homogeneous gaussian regression.
577 *Monthly Weather Review* 143(3):955–971, DOI 10.1175/MWR-D-14-00210.1,
578 URL <https://doi.org/10.1175/MWR-D-14-00210.1>
- 579 Gneiting T, Raftery AE, Westveld AH, Goldman T (2005) Calibrated probabilistic
580 forecasting using Ensemble Model Output Statistics and minimum CRPS esti-
581 mation. *Monthly Weather Review* 133(5):1098–1118, DOI 10.1175/MWR2904.1,
582 URL <https://doi.org/10.1175/MWR2904.1>
- 583 Gutiérrez JM, Cano R, Cofiño AS, Sordo C (2005) Analysis and downscaling
584 multi-model seasonal forecasts in Peru using self-organizing maps. *Tellus A*
585 57(3):435–447, DOI 10.1111/j.1600-0870.2005.00128.x
- 586 Gutiérrez JM, Maraun D, Widmann M, Huth R, Hertig E, Benestad R, Roessler O,
587 Wibig J, Wilcke R, Kotlarski S, San Martín D, Herrera S, Bedia J, Casanueva A,
588 Manzanar R, Iturbide M, Vrac M, Dubrovsky M, Ribalaygua J, Pórtoles J, Rätty
589 O, Räisänen J, Hingray B, Raynaud D, Casado MJ, Ramos P, Zerenner T, Turco
590 M, Bosshard T, Štěpánek P, Bartholy J, Pongracz R, Keller DE, Fischer AM,
591 Cardoso RM, Soares PMM, Czernecki B, Pagé C (2018) An intercomparison
592 of a large ensemble of statistical downscaling methods over Europe: Results
593 from the VALUE perfect predictor cross-validation experiment. *International*
594 *Journal of Climatology* pp 1–36, DOI 10.1002/joc.5462, URL [https://rmets.
595 onlinelibrary.wiley.com/doi/abs/10.1002/joc.5462](https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5462)
- 596 Herrera S, Kotlarski S, Soares PMM, Cardoso RM, Jaczewski A, Gutiérrez JM,
597 Maraun D (2018) Uncertainty in gridded precipitation products: Influence of
598 station density, interpolation method and grid resolution. *International Journal*
599 *of Climatology* DOI 10.1002/joc.5878, URL [https://rmets.onlinelibrary.
600 wiley.com/doi/abs/10.1002/joc.5878](https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5878)
- 601 Hersbach H (2000) Decomposition of the continuous ranked probability score for
602 ensemble prediction systems. *Weather and Forecasting* 15(5):559–570, DOI 10.
603 1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, URL [https://doi.org/
604 10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2)
- 605 Iturbide M, Bedia J, Herrera S, Baño J, Fernández J, Frías MD, Manzanar
606 R, San-Martín D, Cimadevilla E, Cofiño AS, Gutiérrez JM (2019) The R-
607 based climate4R open framework for reproducible climate data access and post-
608 processing. *Environmental Modelling & Software* 111:42–54, DOI 10.1016/j.
609 envsoft.2018.09.009
- 610 Kharin VV, Zwiers FW (2003) On the ROC score of probability forecasts.
611 *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](http://dx.doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2)
- Kotlarski S, Szabó P, Herrera S, Rätty O, Keuler K, Soares PMM, Cardoso RM, Bosshard T, Pagé C, Boberg F, Gutiérrez JM, Isotta FA, Jaczewski A, Kreienkamp F, Liniger MA, Lussana C, Pianko-Kluczyńska K (2017) Observational uncertainty and regional climate model evaluation: a pan-european perspective. *International Journal of Climatology* DOI 10.1002/joc.5249, URL <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5249>
- 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](https://doi.org/10.1175/1520-0477(1999)080<2313:SOTEHS>2.0.CO;2)
- 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>
- 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>
- 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>

- 661 Markus D, J MG, W MJ, Achim Z (2017) Spatial ensemble post-processing with
662 standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*
663 143(703):909–916, DOI 10.1002/qj.2975, URL <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2975>
- 664
665 Molteni F, Stockdale T, Balmaseda M, Balsamo G, Buizza R, Ferranti L,
666 Magnusson L, Mogensen K, Palmer T, Vitart F (2011) The new ECMWF
667 seasonal forecast system (System 4). European Centre for Medium-Range
668 Weather Forecasts, URL http://climate.ncas.ac.uk/people/allan/Fire_Risk_Insurance_Papers/Moltini%20etal%202011.pdf
- 669
670 Murphy AH (1973) A new vector partition of the probability score. *Journal of*
671 *Applied Meteorology* 12(4):595–600, DOI 10.1175/1520-0450(1973)012<0595:
672 ANVPOT>2.0.CO;2, URL [https://doi.org/10.1175/1520-0450\(1973\)](https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2)
673 [012<0595:ANVPOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2)
- 674 Nikulin G, Asharafb S, Magariño ME, Calmanti S, Cardoso RM, Bhend J,
675 Fernández J, Frías MD, Fröhlichb K, Frühb B, Herrera S, Manzananas R, Gutiérrez
676 JM, Hanssona U, Kolaxa M, Liniger M, Soares PMM, Spirig C, Tome R, Wysera
677 K (2018) Dynamical and statistical downscaling of a global seasonal hindcast
678 in eastern Africa. *Climate Services* 9:72 – 85, DOI [https://doi.org/10.1016/](https://doi.org/10.1016/j.cliser.2017.11.003)
679 [j.cliser.2017.11.003](https://doi.org/10.1016/j.cliser.2017.11.003), URL [http://www.sciencedirect.com/science/article/](http://www.sciencedirect.com/science/article/pii/S2405880717300055)
680 [pii/S2405880717300055](http://www.sciencedirect.com/science/article/pii/S2405880717300055)
- 681 Pavan V, Marchesi S, Morgillo A, Cacciamani C, Doblas-Reyes FJ (2005) Down-
682 scaling of DEMETER winter seasonal hindcasts over Northern Italy. *Tellus A*
683 57(3):424–434, DOI 10.1111/j.1600-0870.2005.00111.x
- 684 Piani C, Haerter JO, Coppola E (2010) Statistical bias correction for daily precip-
685 itation in regional climate models over Europe. *Theoretical and Applied Clima-*
686 *tology* 99(1-2):187–192, DOI 10.1007/s00704-009-0134-9, URL [http://dx.doi.](http://dx.doi.org/10.1007/s00704-009-0134-9)
687 [org/10.1007/s00704-009-0134-9](http://dx.doi.org/10.1007/s00704-009-0134-9)
- 688 Sansom PG, Ferro CAT, Stephenson DB, Goddard L, Mason SJ (2016) Best prac-
689 tices for postprocessing ensemble climate forecasts. Part I: Selecting appropri-
690 ate recalibration methods. *Journal of Climate* 29(20):7247–7264, DOI 10.1175/
691 [JCLI-D-15-0868.1](https://doi.org/10.1175/JCLI-D-15-0868.1), URL <https://doi.org/10.1175/JCLI-D-15-0868.1>
- 692 Scheuerer M, Möller D (2015) Probabilistic wind speed forecasting on a grid based
693 on ensemble model output statistics. *The Annals of Applied Statistics* 9(3):1328–
694 1349, DOI 10.1214/15-AOAS843, URL <https://doi.org/10.1214/15-AOAS843>
- 695 Sheau TN, Tangang F, Juneng L (2017) Bias correction of global and regional
696 simulated daily precipitation and surface mean temperature over Southeast
697 Asia using quantile mapping method. *Global and Planetary Change* 149:79
698 – 90, DOI <https://doi.org/10.1016/j.gloplacha.2016.12.009>, URL [http://www.](http://www.sciencedirect.com/science/article/pii/S0921818116301266)
699 [sciencedirect.com/science/article/pii/S0921818116301266](http://www.sciencedirect.com/science/article/pii/S0921818116301266)
- 700 Smith DM, Eade R, Pohlmann H (2013) A comparison of full-field and anomaly
701 initialization for seasonal to decadal climate prediction. *Climate Dynamics*
702 41(11-12):3325–3338, DOI 10.1007/s00382-013-1683-2, URL [https://link.](https://link.springer.com/article/10.1007/s00382-013-1683-2)
703 [springer.com/article/10.1007/s00382-013-1683-2](https://link.springer.com/article/10.1007/s00382-013-1683-2)
- 704 Thorarinsdottir TL, Johnson MS (2012) Probabilistic wind gust forecast-
705 ing using non-homogeneous gaussian regression. *Monthly Weather Review*
706 140(3):889–897, DOI 10.1175/MWR-D-11-00075.1, URL [https://doi.org/10.](https://doi.org/10.1175/MWR-D-11-00075.1)
707 [1175/MWR-D-11-00075.1](https://doi.org/10.1175/MWR-D-11-00075.1)
- 708 Tippett MK, Barnston AG (2008) Skill of multimodel ENSO probability forecasts.
709 *Monthly Weather Review* 136(10):3933–3946, DOI 10.1175/2008MWR2431.1

- 710 Torralba V, Doblas-Reyes FJ, MacLeod D, Christel I, Davis M (2017) Sea-
711 sonal climate prediction: A new source of information for the management
712 of wind energy resources. *Journal of Applied Meteorology and Climatology*
713 56(5):1231–1247, DOI 10.1175/JAMC-D-16-0204.1, URL [https://doi.org/
714 10.1175/JAMC-D-16-0204.1](https://doi.org/10.1175/JAMC-D-16-0204.1)
- 715 Weigel AP, Liniger MA, Appenzeller C (2009) Seasonal ensemble forecasts: Are
716 recalibrated single models better than multimodels? *Monthly Weather Review*
717 137(4):1460–1479, DOI 10.1175/2008MWR2773.1
- 718 Weisheimer A, Palmer TN (2014) On the reliability of seasonal climate forecasts.
719 *Journal of the Royal Society Interface* 11(96), DOI 10.1098/rsif.2013.1162
- 720 Wilks DS, Hamill TM (2007) Comparison of Ensemble-MOS methods using
721 GFS reforecasts. *Monthly Weather Review* 135(6):2379–2390, DOI 10.1175/
722 MWR3402.1, URL [https://doi.org/10.1175/
723 MWR3402.1](https://doi.org/10.1175/MWR3402.1)
- 724 Zhao T, Bennett JC, Wang QJ, Schepen A, Wood AW, Robertson DE, Ramos MH
725 (2017) How suitable is quantile mapping for postprocessing GCM precipitation
726 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>

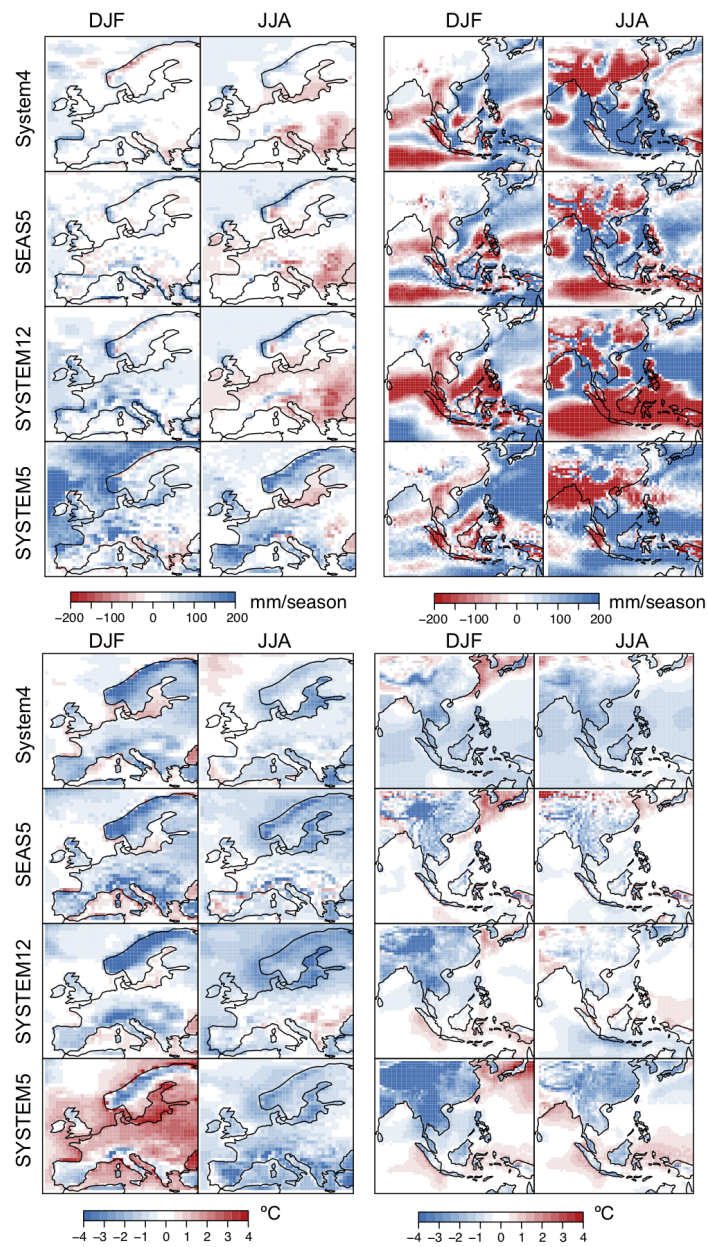


Fig. 1 Bias between the ensemble mean of the four models of Table 1 and ERA-Interim verifying observations for precipitation (top) and temperature (bottom) over EU (left) and SA (right), in DJF and JJA. The errors are expressed as mm/season ($^{\circ}\text{C}$) for the case of precipitation (temperature).

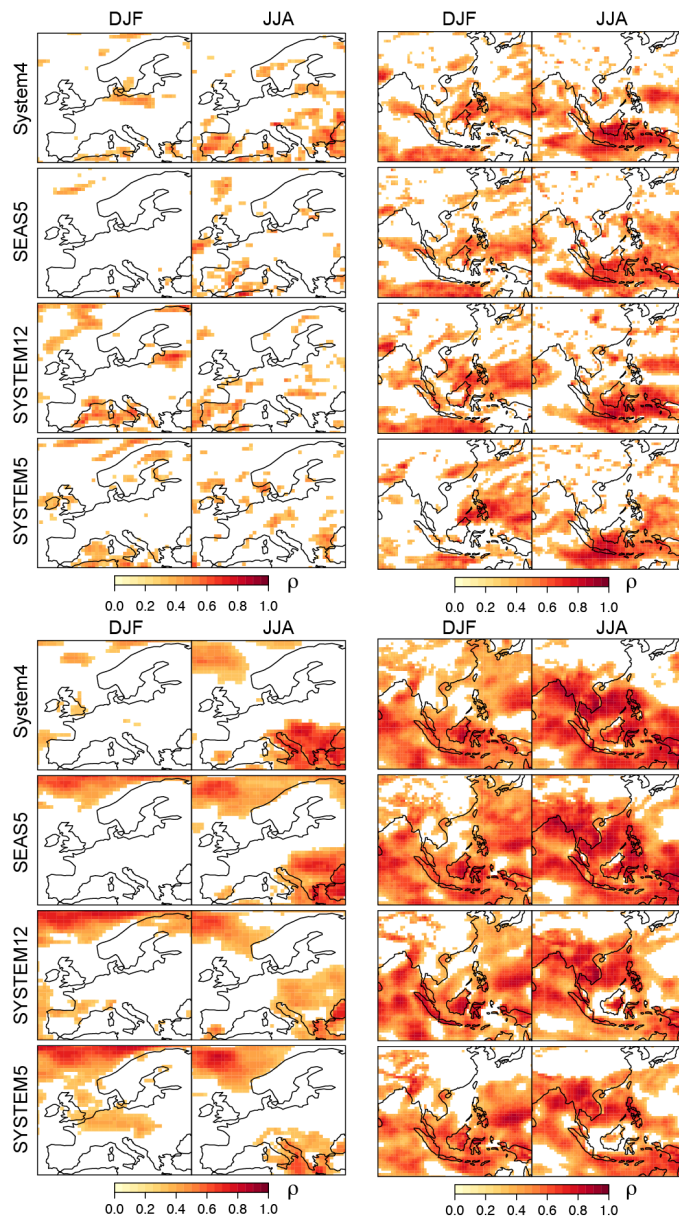


Fig. 2 As Figure 1 but for interannual Pearson correlation. Only significant correlations (90% confidence level, according to a t-test) are shown.

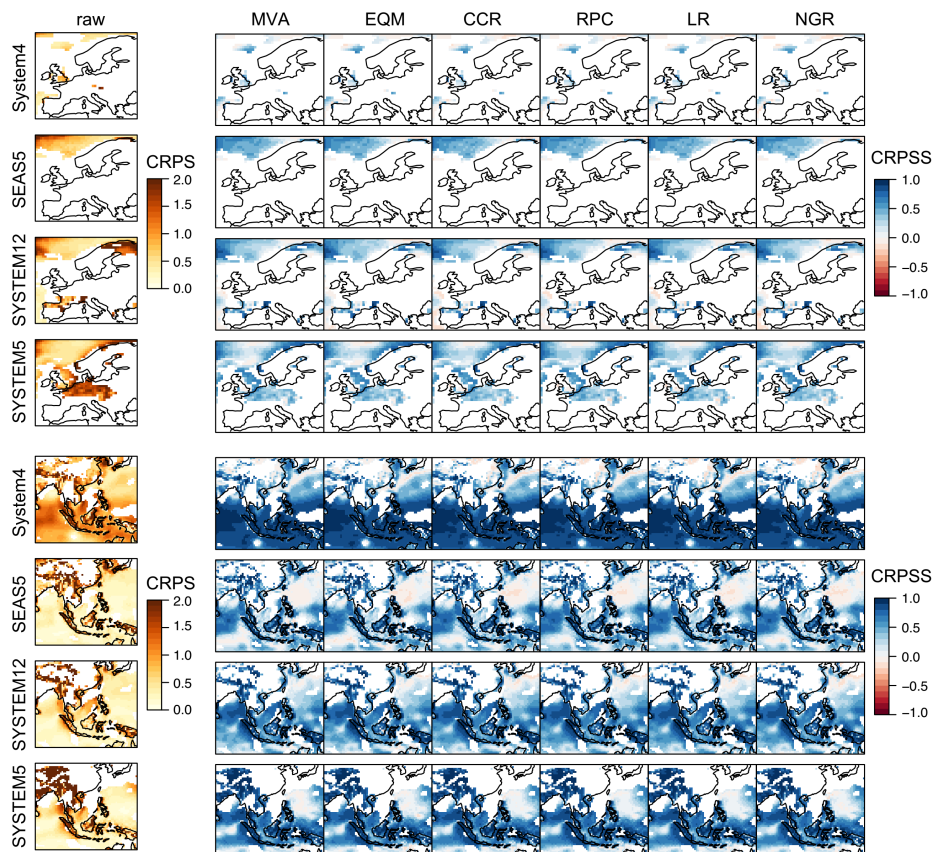


Fig. 3 CRPSS for temperature over EU (top) and SA (bottom) in DJF, as obtained from applying the different BA and RC methods of Table 2 (columns 2-7) to the four models of Table 1 (in rows). In all cases, the CRPS obtained for the raw outputs (column 1) is considered as reference.

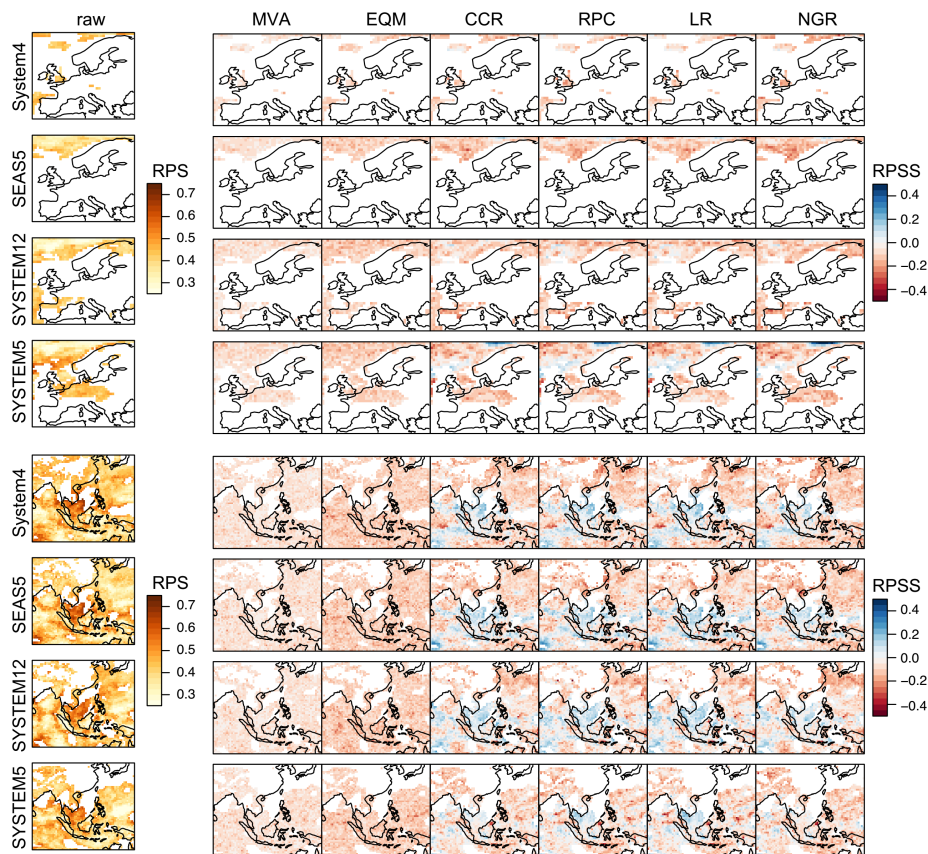


Fig. 4 As Figure 3, but for the RPSS.

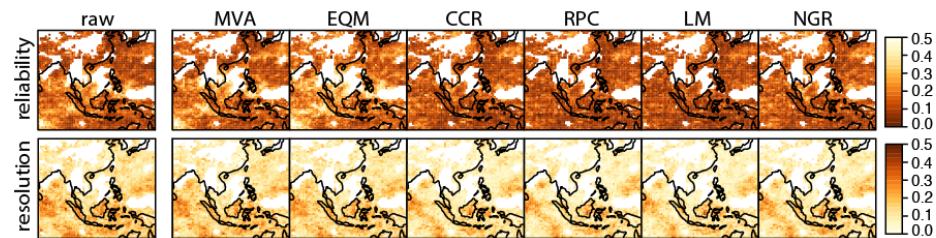


Fig. 5 Reliability and resolution components (top and bottom row, respectively) of the RPS for temperature over SA in DJF, as obtained from applying the BA and RC methods of Table 2 (in columns) to the System4 (12-member, 21-year version).

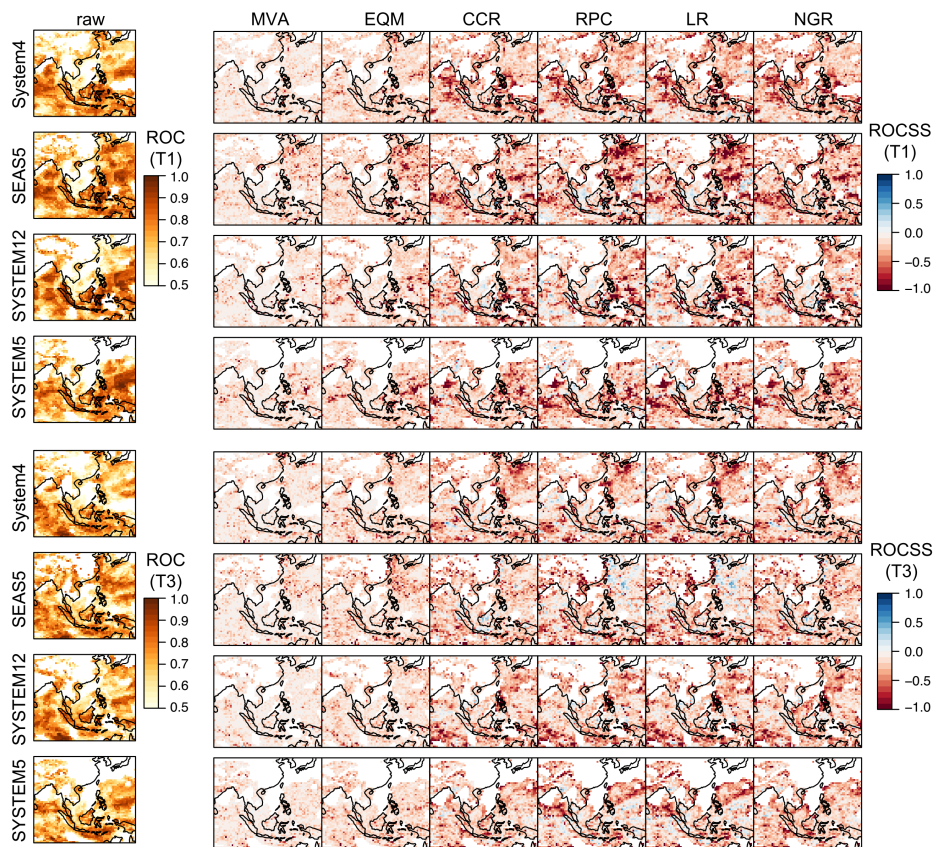


Fig. 6 ROCSS for the cold (top) and warm (bottom) tercile categories of DJF temperature over SA, as obtained from applying the BA and RC methods of Table 1 (columns 2-7) to the four models of Table 1 (in rows). In all cases, the ROC obtained for the raw outputs (column 1) is considered as reference.

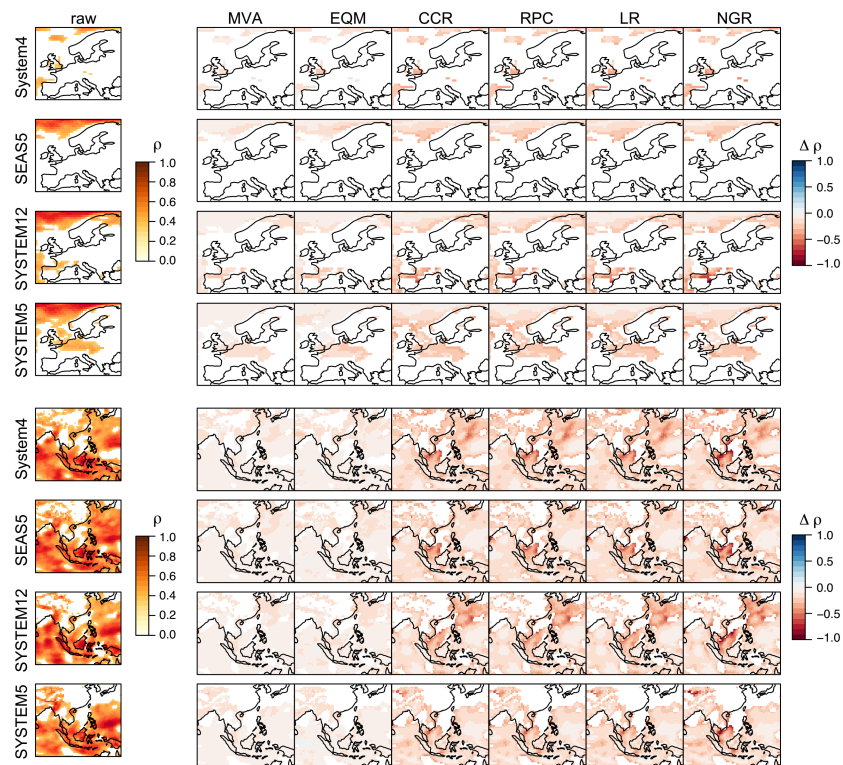


Fig. 7 Column 1: Interannual Pearson correlation between ERA-Interim and the ensemble mean of the four available models (in rows) for DJF temperature, as given by the raw forecasts over EU (top) and SA (bottom). Columns 2-7: Difference (in correlation units) with respect to column 1, as obtained from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

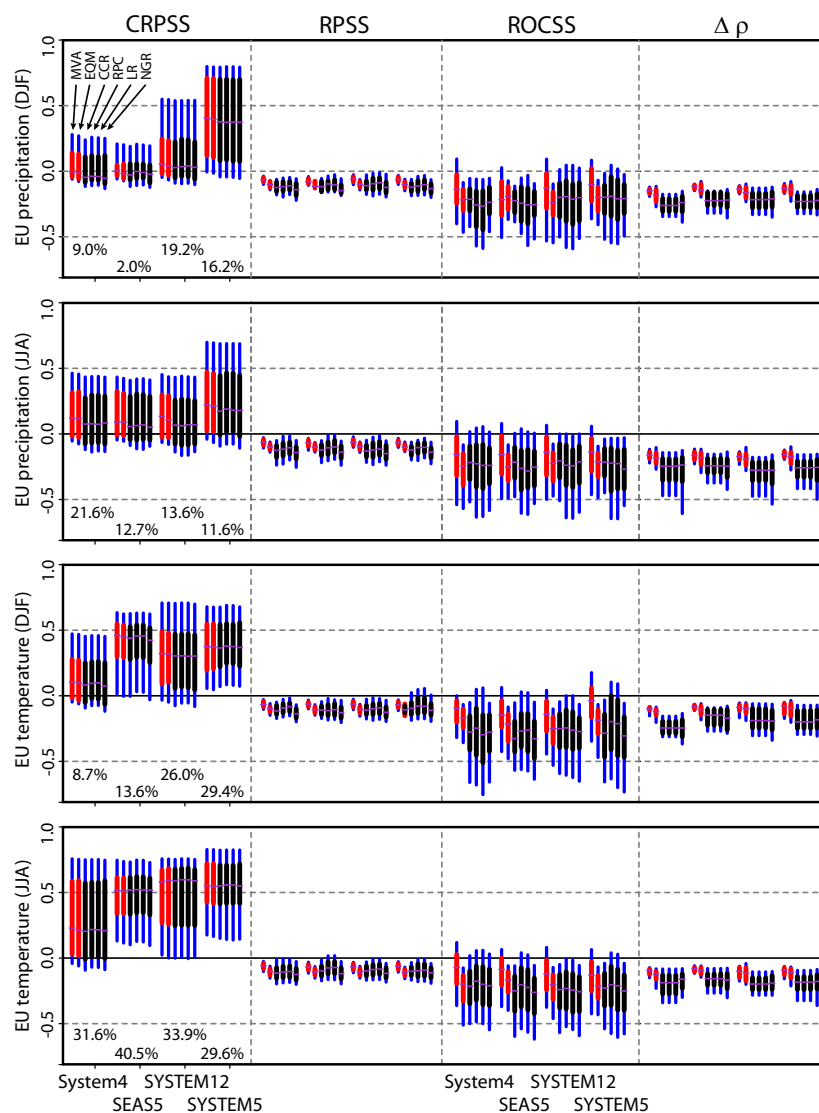


Fig. 8 Summary of the results obtained over EU, in terms of the different skill scores considered (CRPSS, RPSS, ROCSS and correlation differences; in columns). The two variables (precipitation and temperature) and seasons (DJF and JJA) analyzed are shown in different rows. In all cases, results for the 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).

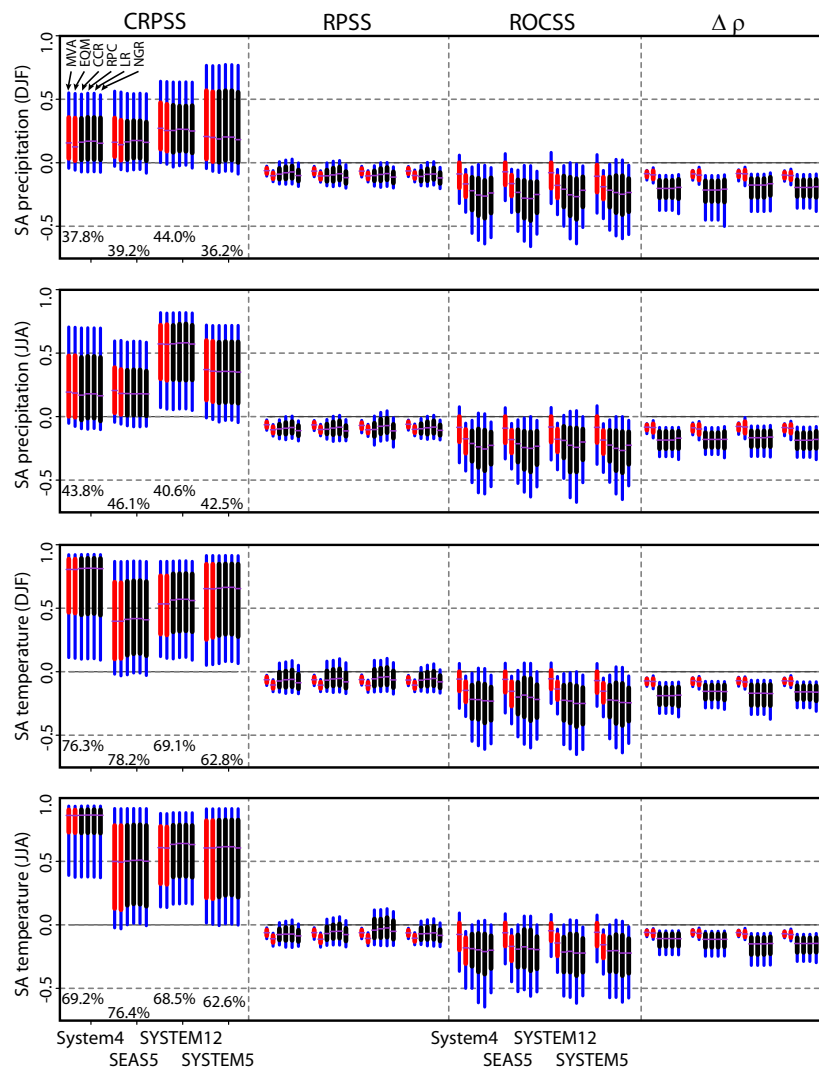


Fig. 9 As Figure 8, but for SA.

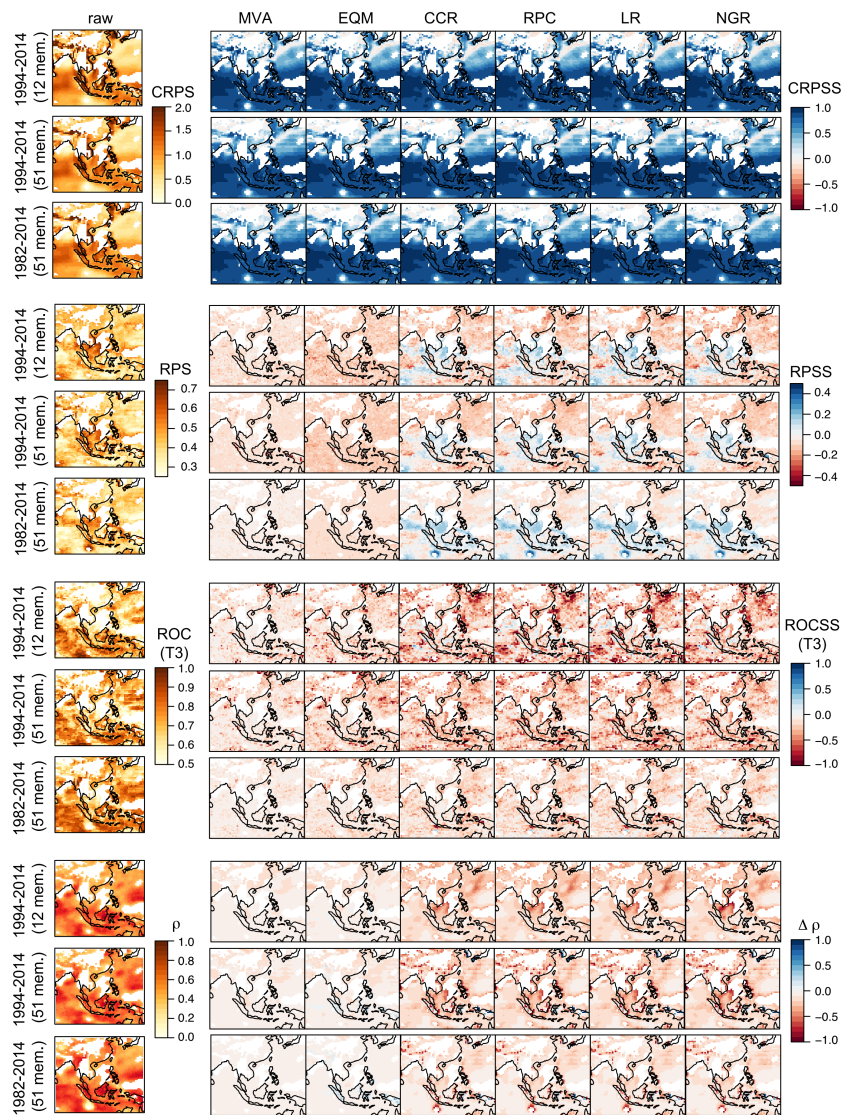


Fig. 10 Results obtained for the CRPSS, the RPSS, the ROCSS (only for the warm tercile category) and the interannual Pearson correlation—in different panels from top to bottom—for temperature over SA in DJF, as obtained from applying the BA and RC methods of Table 2 to the System4. Within each panel, 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-member ensemble for 1994-2014 (1982-2014).

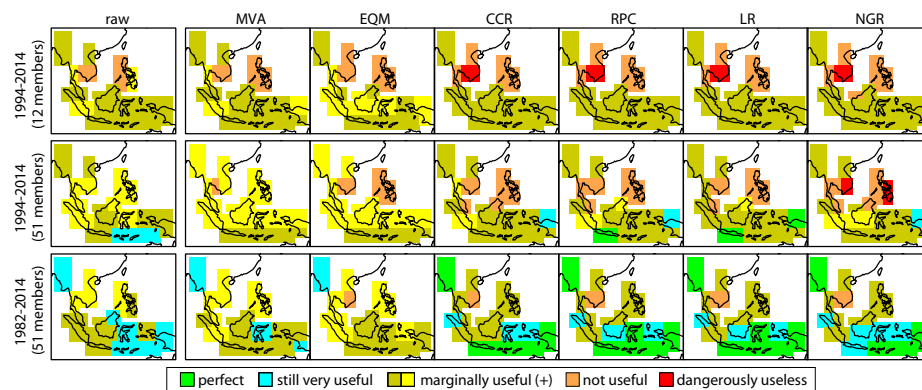


Fig. 11 Reliability categories, as obtained from applying the different BA and RC methods of Table 2 to correct DJF temperature from the System4 over SA. The top row corresponds to a 12-member ensemble for the period 1994-2014. Middle (bottom) row correspond to a 51-ensemble member for 1994-2014 (1982-2014).

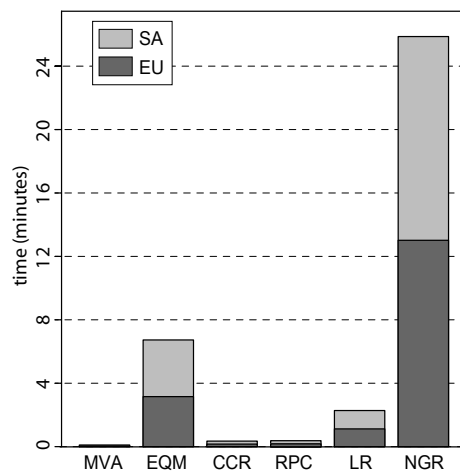


Fig. 12 Execution times—in a personal computer with two cores and two CPUs (3Ghz) attached to each core, with a RAM memory of 16Gb—for the different BA and RC methods of Table 2, for the illustrative case of temperature in DJF over EU (dark gray) and SA (light gray) for System4 (12-member and 21-year version). The MVA and all the RC methods (EQM method) used are implemented in the R-package *calibratoR* (*downscaleR*).