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

Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts?

R. Manzanas · A. Lucero · A. Weisheimer · J. M. Gutiérrez

Received: date / Accepted: date

Abstract Statistical downscaling methods are popular post-processing tools which are widely used in many sectors to adapt the coarse-resolution biased outputs from global climate simulations to the regional-to-local scale typically required by users. They range from simple and pragmatic Bias Correction (BC) methods, which directly adjust the model outputs of interest (e.g. precipitation) according to the available local observations, to more complex Perfect Prognosis (PP) ones, which indirectly derive local predictions (e.g. precipitation) from appropriate upper-air large-scale model variables (predictors). Statistical downscaling methods have been extensively used and critically assessed in climate change applications; however, their advantages and limitations in seasonal forecasting are not well understood yet. In particular, a key problem in this context is whether they serve to improve the forecast quality/skill of raw model outputs beyond the adjustment of their systematic biases.

In this paper we analyze this issue by applying two state-of-the-art BC and two PP methods to downscale precipitation from a multimodel seasonal hindcast in a challenging tropical region, the Philippines. To properly assess the potential added value beyond the reduction of model biases, we consider two validation scores which are not sensitive to changes in the mean/variance (correlation and reliability categories). Our results show that, whereas BC methods maintain or worsen the skill of the raw model forecasts, PP methods can yield significant skill improvement (worsening) in cases for which the large-scale predictor variables considered are better (worse) predicted by the model than precipitation. For instance,

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

A. Lucero
Philippine Atmospheric, Geophysical and Astronomical Services Administration (PAGASA). Quezon City, Philippines.

A. Weisheimer
Department of Physics, National Centre for Atmospheric Science (NCAS), University of Oxford. Oxford OX1 3PU, UK.
European Centre for Medium-Range Weather Forecasts (ECMWF). Reading RG2 9AX, UK.

PP methods are found to increase (decrease) model reliability in nearly 40% of the stations considered in summer (autumn). Therefore, the choice of a convenient downscaling approach (either BC or PP) depends on the region and the season.

Keywords Statistical downscaling, perfect prognosis, bias correction, seasonal forecasting, precipitation, skill, correlation, reliability categories

1 Introduction

Different Statistical Downscaling (SD) methods have been developed since the early 1990s (see, e.g., von Storch et al, 1993) to bridge the gap between the coarse-resolution biased climate information provided by Global Circulation Models (GCMs) and the regional-to-local scale required in different socio-economic sectors such as hydrology, agriculture, energy, etc. These methods rely on empirical/statistical models which link the local observed predictands of interest, here precipitation, with explicative large-scale GCM predictors over the area of interest. These models are first calibrated and tested (i.e., cross-validated) using data from a historical representative period (*training phase*) and subsequently applied to obtain the downscaled local predictions from new GCM predictors (*prediction/downscaling phase*). According to the nature of predictors in the training phase, two different approaches for SD exist (see, e.g. Maraun et al, 2010; Gutiérrez et al, 2013a): Perfect Prognosis (PP) and Model Output Statistics (MOS), the latter including the increasingly popular Bias Correction (BC) methods.

Under the PP approach, quasi-observed predictors from reanalysis are used to train the statistical models (e.g. regression or analog methods), based on their temporal correspondence with the observed precipitation. Afterwards, the resulting models are applied to GCM predictor data in the prediction phase. Therefore, variables well represented by both reanalyses and GCMs (Wilby et al, 2004; Hanssen-Bauer et al, 2005; Brands et al, 2013) accounting for a major part of the variability in the predictands are typically chosen as predictors in this approach (usually large-scale variables at different vertical levels), whereas variables directly influenced by model parameterizations and/or orography, such as precipitation, are usually discarded. As a result, one of the most time-consuming tasks in PP methods is the selection of a suitable combination of predictors, which must be defined over an appropriate geographical domain which encompasses the main synoptic phenomena influencing the climate of the region of interest.

Differently, under the MOS approach, predictors are taken from the same GCM for both the training and the prediction phases. In the context of seasonal forecasting, MOS methods have been traditionally applied establishing an empirical link (e.g. regression or canonical correlation analysis) between large-scale circulation predictors and pairwise observations at a monthly/seasonal time-scale. However, simpler MOS alternatives based on BC methods are becoming increasingly popular (see, e.g., Themeßl et al, 2012a). BC methods directly adjust the distribution of GCM predicted precipitation against local observations (e.g. local scaling or quantile mapping), to ensure that their statistical properties are similar. The main advantage of these methods is their simplicity, since no predictor/domain screening is required (typically, GCM output from the closest model gridbox is considered as unique predictor). For instance, in local scaling methods (the simplest case of BC), a linear transformation is applied to the model output to adjust the first and/or second order moments of the predicted distribution.

A considerable body of research on the application of SD methods to climate change simulations already exists (see, e.g., Gutiérrez et al, 2013b; Vaittinada et al, 2016; Maraun, 2016; San-Martín et al, 2017). Beyond the adjustment of systematic biases (Maraun et al, 2015), however, the advantages and limitations of these methods in seasonal forecasting are not well understood yet, in particular in what refers to their effect on forecast quality/skill. To measure this skill (which is understood as forecast association and reliability here), we focus on correlation and reliability categories. Note that, differently to other scores such as the mean absolute error and the continuous ranked probability score, these two metrics are not sensitive to changes in the mean. Therefore, they allow to properly assess the added value of the SD methods applied beyond the effect of bias reduction.

Some prospects on the potential added value of BC methods can be envisaged for the most simple ones. For instance, local scaling preserve the temporal structure of the original model predictions and do not affect neither correlation nor reliability. However, more sophisticated distributional BC methods such as quantile mapping can introduce arbitrary temporal changes (Maraun, 2013) and thus, their effect on correlation and reliability is difficult to estimate in advance. Differently, PP methods do rely on the temporal correspondence between the predictand and the predictors considered, so there might be windows of opportunity for improving correlation and/or reliability in cases where large-scale variables are better predicted by the model than local precipitation.

In this paper we analyze this problem focusing on a challenging tropical region, the Philippines, which has been identified as an ideal test-bed for SD studies due to the complex topography and land-sea contrasts which determine local rainfall (Moron et al, 2009; Robertson et al, 2012; Manzanas et al, 2015). Moreover, its climate is largely influenced by ENSO (see, e.g., Lyon et al, 2006; Manzanas et al, 2014) and it is located in a region of the world where seasonal forecasts are particularly skillful (Manzanas et al, 2014). As a result, there may be special potential for the application of SD methods to seasonal forecasts in this area. We focus on downscaling methods providing daily data and refer the interested reader to the existing literature (Kang et al, 2007; Robertson et al, 2012) for details on the application of seasonal MOS methods in the Philippines. In particular, we analyze and intercompare the results from two state-of-the-art BC (parametric and empirical quantile mapping) and two PP (analogs and Generalized Linear Models, GLMs) methods when applied to the seasonal hindcast provided by the ENSEMBLES project (Weisheimer et al, 2009) for the period 1981-2005. To our knowledge, this work provides the most comprehensive study on the added value of the BC and PP approaches for downscaling of seasonal forecasts to-date.

The paper is organized as follows. In Section 2 we introduce the data used (both predictand and predictors). Sections 3 and 4 describe the statistical downscaling methods that are applied and the verification metrics which are considered to assess their performance, respectively. The results obtained are presented through Section 5. Finally, the most important conclusions are given in Section 6.

2 Data

2.1 Precipitation in the Philippines: Predictands and Verifying Observations

The Philippines is an archipelago of 7107 islands with complex topography (see Figure 1*a*) located between the monsoonal and inner tropics (4°N and 20°N). Apart from ENSO (Lyon et al, 2006; Manzanas et al, 2014), the climate of this region is affected by important large-scale processes such as the southwest summer and northeast winter monsoons of the western North Pacific Ocean (Wang, 2002), but also by local forcing related to the presence of mountains and the complex land-sea constrast (Robertson et al, 2012). As a result, the country exhibits a rich regional climate composition which has been commonly classified into four different Climatic Types (CTs) in previous studies (Coronas, 1920; Manzanas et al, 2015).

For a good characterization of this variability, daily precipitation from 42 gauges maintained by the Philippine Atmospheric, Geophysical and Astronomical Services Administration (PAGASA: `http://www.pagasa.dost.gov.ph`), which are uniformly distributed across the country (see Figure 1*b*), was considered for this work for the period 1981-2005. The percentage of missing data within this period was less than 5% in all cases (less than 1% in most of the stations) so missing values were ignored in the calibration/training and verification processes. Panels *c-f* in Figure 1 show the interannual variability of spatial average precipitation totals for each CT (see colors in the legend) for the four standard boreal seasons: winter (DJF), spring (MAM), summer (JJA) and autumn (SON). Note that precipitation along the coastlines of the northern part of the archipelago (CT1 and CT2) exhibits a strong seasonal cycle, which is driven by alternating monsoonal winds. In particular, during the southwest monsoon (summer), precipitation peaks at the stations pertaining to CT1 while CT2 is affected by relative dryness. The opposite situation occurs during the northeast monsoon (winter). During the dry months (spring), easterly winds prevail, leading to orographic precipitation along the mountain ranges in the east of the archipelago and to relatively high precipitation amounts for the stations pertaining to CT2. At the stations belonging to CT3 and CT4 (mainly situated in the center and south of the archipelago), precipitation is mainly driven by meso-scale dynamics rather than by large-scale phenomena such as the monsoon circulation, leading to a weak seasonal cycle (rains uniformly distributed along the year). For a more comprehensive description of the climate of the Philippines, the interested reader is referred to Coronas (1920); Flores and Balagot (1969); Kintanar (1984) as well as to the PAGASA website.

2.2 Model Data: Predictors

In this work we consider both reanalysis and seasonal forecast data for the upper-air variables used as predictors (zonal wind component U at 850 and 200 hPa, specific humidity Q and temperature T at 850 hPa; see Section 3) as well as for surface precipitation, the target variable.

On the one hand, and following the recommendation by Manzanas et al (2015) —who carried out an assessment of reanalysis uncertainty over the region of study,— the ERA-Interim reanalysis (Dee et al, 2011) was chosen for the training phase of the PP methods. On the other hand, seasonal forecasts were obtained from four of the GCMs contributing to the ENSEMBLES multimodel seasonal hindcast (Weisheimer et al, 2009), which were produced at the following centres: The European Centre for Medium-Range Weather Forecasts (ECMWF), the Leibniz Institute of Marine Sciences (IFM-GEOMAR), the Euro-Mediterranean Centre for Climate Change (CMCC-INGV) and Météo France (MF). Each of these models —whose main components are summarized in Table 1— ran an ensemble of nine initial conditions (nine equiprobable members), produced by perturbing the realistic estimates of the observed initial state four times a year (the first of February, May, August and November) within the period 1960-2005, providing seven month-long retrospective forecasts. For this work, one-month lead seasonal forecasts were considered. Note that, although the ENSEMBLES models are several years older than state-of-the-art seasonal forecasting systems, they form the most homogeneous and comprehensive multimodel ensemble publicly available to-date.

Centre	**Atmospheric model and resolution**	**Ocean model and resolution**
ECMWF	IFS CY31R1 (T159 $\approx$ 80km/L62)	HOPE (0.3° $-$ 1.4°/L29)
IFM-GEOMAR	ECHAM5 (T63 $\approx$ 180km/L31)	MPI-OM1 (1.5°/L40)
CMCC-INGV	ECHAM5 (T63 $\approx$ 180km/L19)	OPA8.2 (2.0°/L31)
MF	ARPEGE4.6 (T63 $\approx$ 180km)	OPA8.2 (2.0°/L31)

Table 1 Main components of the four global models used in this work, which contributed to the ENSEMBLES multimodel seasonal hindcast.

To keep consistency among reanalysis and the ENSEMBLES models, all predictor data were re-gridded to a common regular 2° grid applying a nearest neighbour interpolation scheme. Moreover, daily instantaneous values at 00 UTC were chosen in all cases. The common period for the available predictands and predictors, 1981-2005, was considered for this work. Note that, according to the WMO Lead Centre for the Long Range Forecast Verification (`http://www.bom.gov.au/wmo/lrfvs`), a 25-years long period is suitable for the proper verification of seasonal forecasts.

Finally, in order to properly harmonize the reanalysis and the ENSEMBLES model data used respectively in the training and prediction phases of the PP methods, a simple local scaling correction was applied to the latter. In particular, for every large-scale model predictor, monthly mean values were adjusted towards the corresponding reanalysis values, gridbox by gridbox, avoiding thus problems that may arise due to the models mean biases.

3 Downscaling Methods

As representative of the PP approach we considered Generalized Linear Models (GLMs) and the analog technique, and relied on the optimum downscaling configuration found for the region of study in Manzanas et al (2015). In particular, they used as predictors a combination of two circulation (U at 850 and 300 hPa) and two thermodynamic (Q and T at 850 hPa) variables over a domain spanning from 114°E to 132°E and from 2°N to 22°N. Here, $U300$ has been replaced by the closest available variable in the ENSEMBLES models, $U200$.

GLMs were formulated by Nelder and Wedderburn (1972) in the 1970's and are an extension of the classical linear regression which allows to model the expected value for non-normally distributed variables. While GLMs have been widely used for statistical downscaling of climate change scenarios (e.g., Brandsma and Buishand, 1997; Chandler and Wheater, 2002; Abaurrea and Asín, 2005; Fealy and Sweeney, 2007; Hertig et al, 2013), they have been rarely applied to seasonal forecasts. Given the dual (occurrence and amount) character of precipitation, we followed in this work the common two-stage implementation (see, e.g., Chandler and Wheater, 2002; Manzanas et al, 2015) in which a GLM with Bernoulli error distribution and *logit* canonical link-function (also known as logistic regression) is used to downscale daily precipitation occurrence (as characterized by a threshold of 0.1mm) and a GLM with gamma error distribution and *log* canonical link-function is applied to downscale daily precipitation amount. A stochastic component could be introduced in both GLMs to increase the predicted variance, which is usually underestimated in deterministic ones (Enke, 1997). However, in order to keep this stochastic effect away from the validation results, the two GLMs considered in this work were deterministic, i.e., predictions were based on the expected values. For this method (denoted as PP1 hereafter), we considered as predictors the 15 leading principal components (PCs, see Preisendorfer, 1988) over the above mentioned domain. PCs were obtained, both for the reanalysis and for the seasonal forecasts, by projecting the corresponding standardized fields onto the Empirical Orthogonal Functions obtained from the reanalysis, which were computed simultaneously on all predictor variables, considering the joined vector of standardized fields. The number of PCs retained, which explain over 80% of the predictor variance, was selected as a trade-off between model parsimony and goodness-of-fit (after a sensitivity study testing models with an increasing number of PCs).

The popular analogue technique (Lorenz, 1963, 1969) estimates the local downscaled values corresponding to a particular atmospheric configuration (as represented by a number of model predictors defined over a certain geographical domain) from the local observations corresponding to a set of similar (or analog) atmospheric configurations within a historical catalog formed by a reanalysis. Here, similarity was measured in terms of the Euclidean distance (Matulla et al, 2008), which was computed over the complete predictor fields. Analog-based methods have been applied in several previous studies to downscale precipitation in the context of seasonal forecasting (see, e.g., Frías et al, 2010; Wu et al, 2012; Shao and Li, 2013). In spite of its simplicity, the analog technique performs as well as other more sophisticated ones (Zorita and von Storch, 1999) and it is one of the most widely used. Here, a deterministic version of the technique (Zorita et al, 1995; Cubasch et al, 1996) which considers the closest analog is used. This will be referred to as PP2 hereafter.

As representative of the BC approach we used two quantile mapping methods, one parametric and one empirical. In the parametric case (referred to as BC1 henceforth) daily predicted and observed rainfall intensities are fitted to gamma distributions and then daily predicted values are corrected according to the differences of the corresponding quantiles from the fitted distributions (Piani et al, 2010; Themeßl et al, 2012a). Note that the parameters of the gamma distribution can be estimated from the first two moments and, therefore, in practice, this method is similar to a local scaling. The empirical method (denoted as BC2 hereafter) consists of calibrating the predicted empirical probability density function (PDF) by adjusting a number of quantiles based on the empirical observed PDF (see, e.g., Déqué, 2007). In particular, we proceed by adjusting percentiles 1 to 99 and linearly interpolating inside this range every two consecutive percentiles. Outside this range a constant extrapolation (using the correction obtained for the 1st or 99th percentile) is applied. Moreover, in cases when the predicted frequency of dry days is larger than the observed one, the frequency adaptation proposed by Themeßl et al (2012b) is applied.

The two BC and the two PP methods described above were separately calibrated/trained and applied for each of the four seasons. We followed a k-fold cross-validation approach (Gutiérrez et al, 2013b) for the period 1981-2005, splitting the whole 25-year period into $k = 5$ random test sets (folds) of 5 years each. Each of these sets was independently used for the prediction phase, using the remaining 20 years for training. For each model, the two BC methods were separately calibrated and applied for each of the nine available ensemble members. However, it is worth to notice here that other configurations were also analyzed for these methods. For instance, we tested cross-validated versus not cross-validated methods and member- versus ensemble-wise calibrated ones (the latter considering the joined nine members series), obtaining very similar results in all cases (not shown). Thus, the conclusions obtained in this work for the BC methods do not depend on the particular experimental configuration followed. Differently, note that the two PP methods were trained just once (based on reanalysis predictor data and local observed precipitation). Afterwards, the (unique) resulting statistical model was separately applied to each of the nine members.

4 Verification Metrics

In order to validate the forecast quality of the raw seasonal precipitation outputs from the ENSEMBLES models and the possible added value of the corresponding downscaled results (beyond the adjustment of systematic biases) we considered two scores recommended by the WMO Lead Centre for the Long Range Forecast Verification (`http://www.bom.gov.au/wmo/lrfvs`): The interannual Anomaly Correlation Coefficient (ACC) and a measure of reliability based on the different categories introduced by Weisheimer and Palmer (2014).

ACC is a simple metric of forecast association which allows to assess the ability of raw/downscaled precipitation to reproduce the observed interannual seasonal anomalies. For each particular model, it is applied here to the deterministic forecast resulting from averaging the nine (either raw or downscaled) available members. In addition, a multimodel (MM) was also constructed by considering the 36 (4 models x 9 members) available predictions (either raw or downscaled), thus giving equal weights to all models and members.

Reliability measures how closely the forecast probabilities of a certain event correspond to the actual chance of observing that event. It is applied here for probabilistic forecasts of each of the three precipitation terciles: dry (T1), normal (T2) and wet (T3). For each model (the MM), probabilities are computed based on the nine (36), either raw or downscaled, available members. Reliability diagrams (see the illustrative examples shown in Figure 2) plot the observed frequencies of the event considered (e.g. T1, T2 or T3) as a function of its forecast probability, as represented by a determined number of bins (see Doblas-Reyes et al, 2008, for details). For a perfectly reliable forecasting system, the curve obtained would match the diagonal (perfect reliability line). Points falling within the so-called skill region (in gray), i.e., the region contained between the no-resolution line (which indicates the expected frequency of the event: 1/3 for terciles) and the no-skill line (halfway between the no-resolution line and the diagonal) positively contribute to the forecast skill (Brier Skill Score > 0). Weisheimer and Palmer (2014) proposed a methodology to translate the information provided by these diagrams to an easy-to-interpret scale with five reliability categories: *perfect* (green), *still very useful* (blue), *marginally useful* (yellow), *not useful* (orange) and *dangerously useless* (red). In particular, they performed a weighted linear regression as a best-guess estimate on all data points in the diagram (using the number of forecasts in each probability bin as weights) and defined the different reliability categories based on the relative position of the so derived reliability line with respect to the perfect reliability (diagonal), no-skill and no-resolution lines, as well as on the uncertainty range around it (as obtained by bootstrapping with 1000 samples). Here, we slightly modified this original classification by Weisheimer and Palmer (2014) for a better adaptation to our particular regional study (see Section 5.3).

Note that the two validation metrics considered for this work are insensitive to data scaling and, therefore, are suitable to assess the added value of the downscaling methods beyond the improvement of systematic biases in the mean and variance. Thus, we assess here the relevant aspects which can provide added value for seasonal forecasting.

5 Results

5.1 Performance of Raw Models

In order to obtain an estimation of the performance of the ENSEMBLES models over the region of study, we carried out a regional validation considering as reference the observed precipitation at the 42 PAGASA stations (model precipitation was bi-linearly interpolated to these gauges). Figure 3 shows the results obtained in terms of local biases, which are in general strong (as compared with the observed climatologies, shown in the first row). Note that in spite of local differences, all models (and as a result the MM) exhibit similar spatial patterns for the different seasons, which reflect their inability to properly represent the local features in this region of complex orography and land-sea contrast. Notice that, by construction, all the statistical downscaling methods here considered reduce the mean biases, yielding absolute biases smaller than 10 mm/year in all cases (not shown). Although this is a clear advantage for end users, here we focus on the added value in terms of skill (as characterized by forecast association and reliability). The reader is referred to (Maraun et al, 2015) for further information on the performance of the different downscaling methods from the point of view of biases and marginal statistics.

Figure 4 shows the local interannual ACC values obtained. In general, significant correlations are found for all models throughout the year (especially in DJF and MAM) except for JJA. This marked seasonality in forecast skill is a consequence of the large influence exerted by the ENSO interannual oscillations in this region (Manzanas et al, 2014). However, important local-to-regional differences can be found for different models in some seasons. For instance, the ECMWF model exhibits a superior performance for the CT1 region in JJA. This could be a consequence of the higher resolution of this model, as compared to the other three (see Table 1).

5.2 Correlation of Downscaled Results

For the different seasons (in rows) and CTs (in columns), panels in Figure 5 show the interannual ACC values obtained for each of the ENSEMBLES models (see the colors in the legend). Boxplots display the results along the different stations for the raw/direct model output (DMO henceforward), which is indicated by a light gray shadow, and for all the downscaling methods considered (right after the DMO). Overall, results vary mainly among seasons, but also among CTs, models and downscaling methods. For the latter, results are in general more sensitive to the approach considered (BC or PP) than to the particular technique used within each approach. As already explained in Section 5.1, the highest scores for the DMO are obtained for DJF and MAM, whereas the worst results are found for JJA, with no significant correlations for any model except for the ECMWF in the CT1 region. In general, the DMO outperforms the BC methods (note that the correlation gain found for the latter in some cases is limited to a few stations and is counteracted by the loss found in others, so no robust signal of added value is obtained for the BC approach). Nonetheless, PP methods can either improve or spoil the correlations attained by the DMO, depending on the case.

More in detail, whereas the BC methods do not improve (or even worsen) the correlations reached by the DMO in general for DJF and MAM, there are a few cases in which PP methods can add important value (indicated by black dotted boxes). In particular, PP methods are shown to improve raw precipitation from the relatively bad performing models (those exhibiting small ACC values, as compared to the rest of models), as occurs for the MF model in DJF (CT4) and the IFM-GEOMAR model in MAM (CT1). Moreover, as marked with red dotted boxes, PP methods can also add important local value for some particular outlier stations (those in which the correlation for the raw model precipitation drops, as compared with the rest of locations). See, for instance, the case of the CMCC-INGV model in MAM (CT2 and CT3). Notice that, as opposite to the DMO and the BC methods —which depend on model precipitation at the nearest gridbox and can be affected by local features such as wrong orographical gradients, land-sea interfaces, etc.,— PP methods rely on large-scale predictors to infer local precipitation, which might allow in turn to properly reproduce the observed interannual variability in these cases.

With respect to JJA and SON, whereas BC methods do not clearly improve (or even worsen) the correlations attained by the DMO, PP methods provide in general better (worse) results than the DMO in the former (latter) season. In particular, notice that PP methods yield large correlation improvements in JJA for the stations pertaining to CT1 for all models (with the exception of the ECMWF), which exhibit nearly-zero ACC values in this season.

In order to summarize the results from Figure 5 and to better quantify the added value of BC and PP methods, Figure 6 shows in bar charts the percentage of stations with significant ACC values for the DMO and for the different downscaling approaches (BC and PP), for the different seasons. Within each approach, the two methods applied are jointly considered. Moreover, all models except the

MM (which is excluded for clarity) and all CTs are also jointly considered. This figure shows that BC methods do not outperform (or slightly reduce) the correlations attained by the DMO for any season. However, PP methods yield higher (lower) correlations than the DMO does for JJA (SON). In particular, whereas the percentage augments from 10% to 30% in JJA, it drops from more than 60% to less than 30% in SON.

5.3 Reliability of Downscaled Results

In Weisheimer and Palmer (2014), the confidence interval around the best-guess reliability line was estimated by randomly resampling members, gridboxes and years, and the 75% of the total range was considered. Here, we analyzed the sensitivity of their classification to different confidence intervals (the same bootstrapping procedure was used) and found that the ensemble size had a large influence, as higher uncertainty around the best-guess reliability line was obtained for smaller ensembles. As a result, *still very useful* (blue) categories may pass to *marginally useful* (yellow) ones due to an enlargement of the confidence region (see Weisheimer and Palmer, 2014, for details on the definition of the different categories). Therefore, in this work we considered a smaller confidence interval given by the central 50% of the total range, which is more suitable for the nine members of the ENSEMBLES models used —note that the original classification was developed for the 51 members version of the ECMWF System 4 model (Molteni et al, 2011).— Moreover, in order to introduce further discrimination power, within the original *marginally useful* (yellow) category, we differentiate those cases in which the best-guess reliability line is above the no skill line, assigning to this new category (denoted as *marginally useful +*) the dark yellow color. See, for instance, panels *g* and *h* in Figure 2 —note that both cases would correspond to the same category in the original definition.—

Figure 7 shows the reliability categories (in colors) obtained after applying the methodology described above for the different models (in columns) and seasons (in rows), by CT (note that the joined series of the different stations falling within each CT are considered). From left to right, each block shows the results for the DMO, the two BC and the two PP methods considered, for the three terciles. Overall, this figure is in good correspondence with the results found for correlation (Figures 5 and 6), with the best reliability obtained in DJF and MAM and the worst in JJA. Moreover, the results for the two BC methods are very similar to those obtained for the DMO, with slight differences due to spurious changes of category (as illustrated in the top row of Figure 2). However, the two PP methods exhibit major reliability differences with respect to the DMO, especially for JJA and SON. In particular, both PP1 and PP2 improve the results of the DMO in the former season, especially for the CT1, where *marginally useful* or *marginally useful +* categories are obtained instead of *not useful* and *dangerously useless* ones. Yet, the opposite situation is found for SON. Additionally, this figure also shows some well-known results (see, e.g., Manzanas et al, 2014), such as the higher performance attained for the extreme terciles (as compared to the normal one) and the superiority of the MM, which in general outperforms any single model.

In order to summarize the results from Figure 7 and to better quantify the added value of the different approaches for statistical downscaling, Figure 8 shows in stacked bar charts the percentage of reliability categories obtained from the DMO and the different downscaling approaches (BC and PP) for the different seasons. Within each approach, the two methods applied are jointly considered. For clarity, the results from the MM and from the normal tercile are excluded from this analysis. This figure shows that BC methods do not provide clear added value (or even worsen the DMO) for any season. However, PP methods yield substantial added value for JJA, leading to *marginally useful* or *marginally useful +* categories in over 50% of the cases, as compared to less than 10% for the DMO (and for the BC methods). In contrast, the opposite situation is found for the PP methods in SON, with *not useful* or *dangerously useless* categories obtained in nearly 50% of the cases (as compared with 10% for the DMO and 20% for the BC methods).

Remarkably, the good alignment between the results found for reliability and those found for correlation points out the suitability and usefulness of the methodology proposed by Weisheimer and Palmer (2014) —which is slightly modified here— for regional studies. Note that the original work was undertaken for the 21 global regions defined in Giorgi and Francisco (2000).

5.4 An Explanation for the Added Value of PP Methods

As already mentioned, PP methods rely on large-scale predictors to infer local precipitation. As such, the above presented cases leading to a gain (loss) of skill for the PP approach could be explained by situations where large-scale variables, defined over a synoptic domain, are better (worse) predicted by the model than the target precipitation, which is more affected by particular local features (as usually represented by parametrizations). In order to check this premise, we focus here on the climate region CT1, where PP methods were shown to improve (deteriorate) the skill of the DMO in JJA (SON). Figure 9 displays the interannual ACC values obtained between observed precipitation at the 13 stations pertaining to this CT and the ERA-Interim and ENSEMBLES models outputs —the nearest gridbox is considered— for precipitation (PR) and the different predictors used (U850, U200, Q850 and T850) for the period 1981-2005. For benchmarking purposes, ERA-Interim is indicated by a light gray shadow.

The gain of skill found in JJA for all models except the ECMWF (Figures 5 and 7) is in agreement with the results shown in the top panel. In particular, whereas significant ACC values for precipitation are only found for the ECMWF model, mostly significant correlations (similar to the benchmark provided by ERA-Interim) are found for all models for U850 and T850, the large-scale predictors most correlated with observed precipitation (as indicated by the reanalysis). This suggests that PP methods might be able to exploit the model ability for reproducing upper-air predictor variables to indirectly obtain improved precipitation forecasts in cases of a poor skill for model precipitation.

The opposite situation is found for SON (bottom panel). In this season, the ACC values found for precipitation are significant (although smaller than the benchmark provided by ERA-Interim) in most cases. However, the results found for the large-scale predictors are in general not significant. Moreover, opposite correlations with observations (as compared to the reanalysis) are found in some cases. The combined effect of these errors could result in wrong downscaled predictions, as occurs for the ECMWF model, which leads to negative ACC values (see the corresponding boxplots in Figure 5) and *dangerously useless* reliability categories (see the corresponding extreme terciles in Figure 7).

6 Conclusions

In order to assess the advantages and limitations of different approaches for statistical downscaling in the context of seasonal forecasting, two state-of-the-art Bias Correction (BC) and two Perfect Prognosis (PP) methods were applied to obtain local precipitation at 42 stations in the Philippines, considering one-month lead forecasts from the ENSEMBLES multimodel seasonal hindcast for the four boreal seasons over the period 1981-2005.

As expected by construction, BC and PP methods were shown to be successful in reducing the systematic model biases over the area of study, which are in general strong (as compared to the local climatologies). In particular, both approaches lead to very small biases after downscaling. However, and even though this is a clear advantage for users, we focus here on the methods' ability to predict interannual anomalies, which is the basis of seasonal forecasting. Therefore, we assess forecast quality/skill in terms of interannual correlation and reliability categories. Note that these two metrics are not sensitive to changes in the mean and allow therefore to properly assess the added value of the downscaling methods beyond the effect of bias reduction.

On the one hand, BC methods were shown to provide no added value in terms of skill, maintaining or worsening both correlation and reliability. These methods directly transform model precipitation (by correcting different quantiles of the distribution) without relying on any additional information about the underlying physical phenomena (e.g. large-scale circulation). As a consequence, BC methods can arbitrarily modify the temporal structure of the raw model output, with the overall result of degrading the skill (Maraun, 2013). Noticeably, the conclusions obtained here for the BC methods are quite general and do not depend on the particular experimental configuration followed. For instance, we tested cross-validated versus not cross-validated methods and member- versus ensemble-wise calibrated ones, obtaining very similar results in all cases.

On the other hand, we found that PP methods can either substantially improve or deteriorate correlation and reliability. As opposite to BC ones, PP methods rely on physically-based large-scale model predictors to infer local precipitation. Thus, this provides an opportunity for improving the original model skill in those cases for which orographic and land-sea contrasts limit the local representativeness of model precipitation, but the model is yet skillful in reproducing the large-scale predictors. In this work, we show that those conditions are met for certain regions and/or seasons. For instance, reliability was increased by PP methods in nearly 40% of the stations considered in summer.

Therefore, we conclude that the choice of an appropriate statistical downscaling method is not trivial and depends on factors such as the region, the season, the strength of the connection between the large- and the local-scale climate and the model skill for predicting surface/upper-air variables. Moreover, this selection should be based on the requirements of the particular user and/or application. In general, it is advisable to test the added value of PP methods as a first choice, particularly in regions with complex orography and/or large local variability. However, BC methods could be a cost-effective and pragmatic choice in applications for which the main concern is just reducing model biases, even at the cost of degrading the skill.

Acknowledgements This study was partially supported by the SPECS and EUPORIAS projects, funded by the European Commission through the Seventh Framework Programme for Research under grant agreements 308378 and 308291, respectively. JMG acknowledges partial support from the project MULTI-SDM (CGL2015-66583-R, MINECO/FEDER). Also, the authors are grateful to PAGASA for the data provided.

References

Abaurrea J, Asín J (2005) Forecasting local daily precipitation patterns in a climate change scenario. Climate Research 28(3):183–197, DOI 10.3354/cr028183

Brands S, Herrera S, Fernández J, Gutiérrez JM (2013) How well do CMIP5 Earth System Models simulate present climate conditions in Europe and Africa? Climate Dynamics 41(3):803–817, DOI 10.1007/s00382-013-1742-8

Brandsma T, Buishand TA (1997) Statistical linkage of daily precipitation in Switzerland to atmospheric circulation and temperature. Journal of Hydrology 198(1-4):98–123, DOI 10.1016/S0022-1694(96)03326-4

Chandler RE, Wheater HS (2002) Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. Water Resources Research 38(10):1–11, DOI 10.1029/2001WR000906

Coronas J (1920) The climate and weather of the Philippines, 1903-1918, Bureau of Printing, Manila, pp 291–467

Cubasch U, von Storch H, Waszkewitz J, Zorita E (1996) Estimates of climate change in Southern Europe derived from dynamical climate model output. Climate Research 7(2):129–149, DOI 10.3354/cr007129

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

Doblas-Reyes FJ, Coelho CAS, Stephenson DB (2008) How much does simplification of probability forecasts reduce forecast quality? Meteorological Applications 15(1):155–162, DOI 10.1002/met.50

Enke SA W (1997) Downscaling climate model outputs into local and regional weather elements by classification and regression. Climate Research 8(3):195–207

Fealy R, Sweeney J (2007) Statistical downscaling of precipitation for a selection of sites in Ireland employing a generalised linear modelling approach. International Journal of Climatology 27(15):2083–2094, DOI 10.1002/joc.1506

Flores JF, Balagot VF (1969) World Survey of Climatology, Climates of Northern and Eastern Asia, vol 8, Arakawa, chap Climate of the Philippines, pp 159–213

Frías MD, Herrera S, Cofiño AS, Gutiérrez JM (2010) Assessing the skill of precipitation and temperature seasonal forecasts in spain: Windows of opportunity related to ENSO events. Journal of Climate 23(2):209–220, DOI 10.1175/2009JCLI2824.1

Giorgi F, Francisco R (2000) Uncertainties in regional climate change prediction: A regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. Climate Dynamics 16(2-3):169–182, DOI 10.1007/PL00013733

Gutiérrez JM, Bedia J, Benestad R, Pagé C (2013a) Review of the different statistical downscaling methods for s2d prediction. Tech. rep., SPECS deliverable 5.2.1, URL http://www.specs-fp7.eu/sites/default/files/u1/SPECS_D52.1.pdf

Gutiérrez JM, San-Martín D, Brands S, Manzanas R, Herrera S (2013b) Reassessing statistical downscaling techniques for their robust application under climate change conditions. Journal of Climate 26(1):171–188, DOI 10.1175/JCLI-D-11-00687.1

Hanssen-Bauer I, Achberger C, Benestad RE, Chen D, Forland EJ (2005) Statistical downscaling of climate scenarios over Scandinavia. Climate Research 29(3):255–268, DOI 10.3354/cr029255

Hertig E, Seubert S, Paxian A, Vogt G, Paeth H, Jacobeit J (2013) Changes of total versus extreme precipitation and dry periods until the end of the twenty-first century: Statistical assessments for the Mediterranean area. Theoretical and Applied Climatology 111(1-2):1–20, DOI 10.1007/s00704-012-0639-5

Kang H, An KH, Park CK, Solís ALS, Stitthichivapak K (2007) Multimodel output statistical downscaling prediction of precipitation in the Philippines and Thailand. Geophysical Research Letters 34(15):n/a–n/a, DOI 10.1029/2007GL030730

Kintanar RL (1984) Climate of the Philippines. Tech. rep., PAGASA

Lorenz EN (1963) Deterministic nonperiodic flow. Journal of the Atmospheric Sciences 20(2):130–141

Lorenz EN (1969) Atmospheric predictability as revealed by naturally occurring analogues. Journal of the Atmospheric Sciences 26(4):636–646, DOI 10.1175/1520-0469(1969)26⟨636:APARBN⟩2.0.CO;2

Lyon B, Cristi H, Verceles ER, Hilario FD, Abastillas R (2006) Seasonal reversal of the ENSO rainfall signal in the Philippines. Geophysical Research Letters 33(24):n/a–n/a, DOI 10.1029/2006GL028182

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

Manzanas R, Brands S, San-Martín D, Lucero A, Limbo C, Gutiérrez JM (2015) Statistical downscaling in the tropics can be sensitive to reanalysis choice: A case study for precipitation in the Philippines. Journal of Climate 28(10):4171–4184, DOI 10.1175/JCLI-D-14-00331.1

Maraun D (2013) Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. Journal of Climate 26(6):2137–2143, DOI 10.1175/

JCLI-D-12-00821.1

Maraun D (2016) Bias correcting climate change simulations: A critical review. Current Climate Change Reports 2(4):211–220, DOI 10.1007/s40641-016-0050-x

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

Maraun D, Widmann M, Gutiérrez JM, Kotlarski S, Chandler RE, Hertig E, Wibig J, Huth R, Wilcke RA (2015) VALUE: A framework to validate downscaling approaches for climate change studies. Earth's Future 3(1):1–14, DOI 10.1002/2014EF000259

Matulla C, Zhang X, Wang X, Wang J, Zorita E, Wagner S, von Storch H (2008) Influence of similarity measures on the performance of the analog method for downscaling daily precipitation. Climate Dynamics 30(2-3):133–144, DOI 10.1007/s00382-007-0277-2

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). Tech. rep., ECMWF

Moron V, Lucero A, Hilario F, Lyon B, Robertson AW, DeWitt D (2009) Spatio-temporal variability and predictability of summer monsoon onset over the Philippines. Climate Dynamics 33(7-8):1159–1177, DOI 10.1007/s00382-008-0520-5

Nelder JA, Wedderburn RWM (1972) Generalized linear models. Journal of the Royal Statistical Society Series A (Statistics in Society) 135(3):370–384, DOI 10.2307/2344614

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

Preisendorfer R (1988) Principal component analysis in meteorology and oceanography, 1st edn. Elsevier

Robertson AW, Qian JH, Tippett MK, Moron V, Lucero A (2012) Downscaling of seasonal rainfall over the Philippines: Dynamical versus statistical approaches. Monthly Weather Review 140(4):1204–1218, DOI 10.1175/MWR-D-11-00177.1

San-Martín D, Manzanas R, Brands S, Herrera S, Gutiérrez JM (2017) Reassessing Model Uncertainty for Regional Projections of Precipitation with an Ensemble of Statistical Downscaling Methods. Journal of Climate 30(1):203–223, DOI 10.1175/JCLI-D-16-0366.1

Shao Q, Li M (2013) An improved statistical analogue downscaling procedure for seasonal precipitation forecast. Stochastic Environmental Research and Risk Assessment 27(4):819–830, DOI 10.1007/s00477-012-0610-0

von Storch H, Zorita E, Cubasch U (1993) Downscaling of global climate change estimates to regional scales: An application to Iberian rainfall in wintertime. Journal of Climate 6(6):1161–1171, DOI 10.1175/1520-0442(1993)006⟨1161:DOGCCE⟩2.0.CO;2

Themeßl MJ, Gobiet A, Heinrich G (2012a) Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal. Climatic Change 112(2):449–468, DOI 10.1007/s10584-011-0224-4

Themeßl MJ, Gobiet A, Heinrich G (2012b) Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal. Climatic Change 112(2):449–468, DOI 10.1007/s10584-011-0224-4

Vaittinada AP, Vrac M, Bastin S, Carreau J, Déqué M, Gallardo C (2016) Intercomparison of statistical and dynamical downscaling models under the EURO- and MED-CORDEX initiative framework: Present climate evaluations. Climate Dynamics 46(3-4):1301–1329, DOI 10.1007/s00382-015-2647-5

Wang B (2002) Rainy season of the Asian-Pacific summer monsoon. Journal of Climate 15(4):386–398, DOI 10.1175/1520-0442(2002)015⟨0386:RSOTAP⟩2.0.CO;2

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

Weisheimer A, Doblas-Reyes FJ, Palmer TN, Alessandri A, Arribas A, Déqué M, Keenlyside N, MacVean M, Navarra A, Rogel P (2009) ENSEMBLES: A new multi-model ensemble for seasonal-to-annual prediction. Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. Geophysical Research Letters 36(21):n/a–n/a, DOI 10.1029/2009GL040896

Wilby RL, Charles S, Zorita E, Timbal B, Whetton P, Mearns L (2004) Guidelines for use of climate scenarios developed from statistical downscaling methods. Tech. rep., IPCC-TGCIA

Wu W, Liu Y, Ge M, Rostkier-Edelstein D, Descombes G, Kunin P, Warner T, Swerdlin S, Givati A, Hopson T, Yates D (2012) Statistical downscaling of climate forecast system seasonal predictions for the Southeastern Mediterranean. Atmospheric Research 118:346–356, DOI 10.1016/j.atmosres.2012.07.019

Zorita E, von Storch H (1999) The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. Journal of Climate 12(8):2474–2489, DOI 10.1175/1520-0442(1999)012⟨2474:TAMAAS⟩2.0.CO;2

Zorita E, Hughes JP, Lettemaier DP, von Storch H (1995) Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. Journal of Climate 8(5):1023–1042, DOI 10.1175/1520-0442(1995)008⟨1023:SCORCP⟩2.0.CO;2

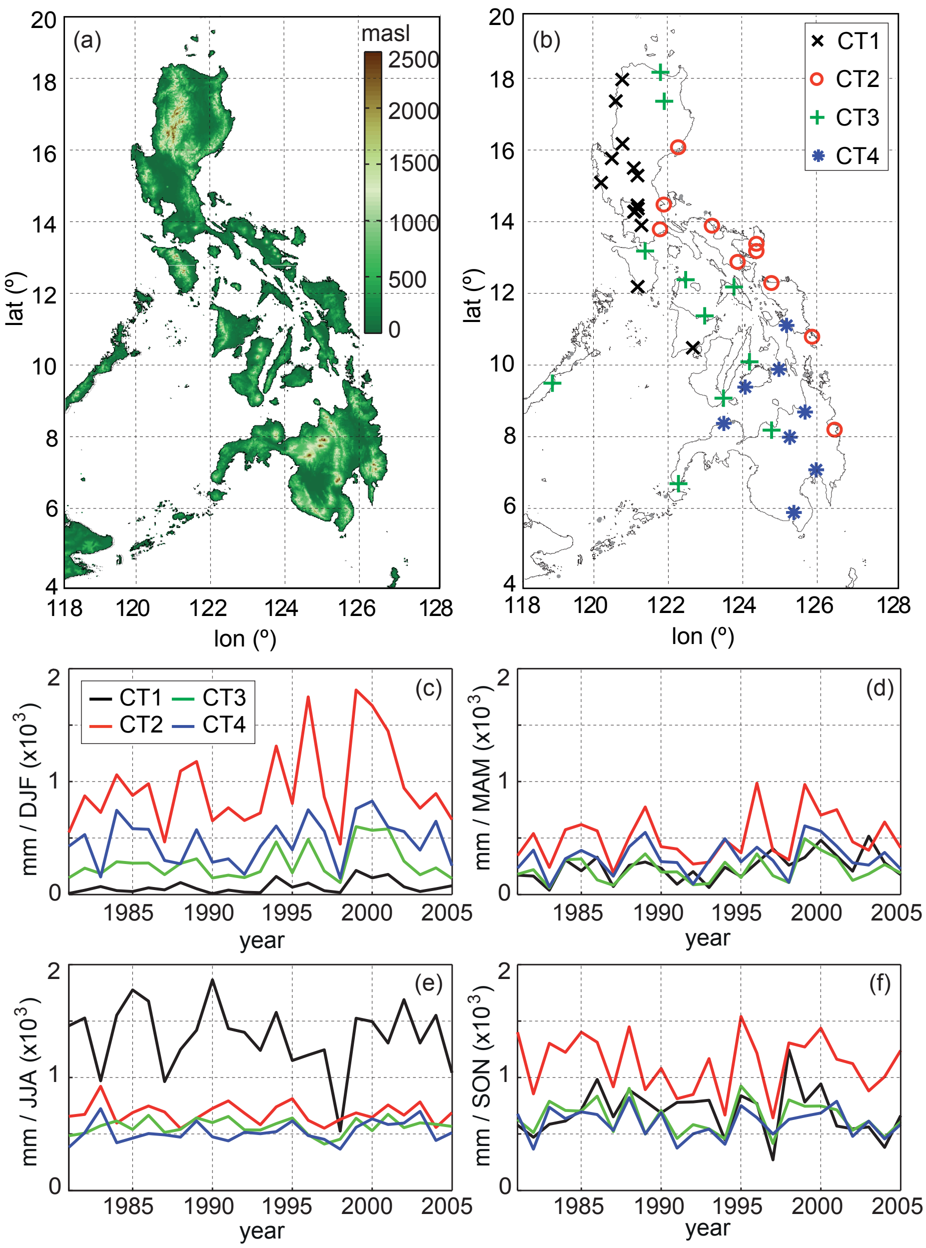

Fig. 1 (a) Topography of the Philippines. (b) Location of the 42 PAGASA gauges considered, classified into the four precipitation climatic types (CTs) defined in Coronas (1920), in colors. (c)-(f) Interannual variability of spatial average precipitation totals for each CT (see colors in the legend) for the period 1981-2005, by seasons.

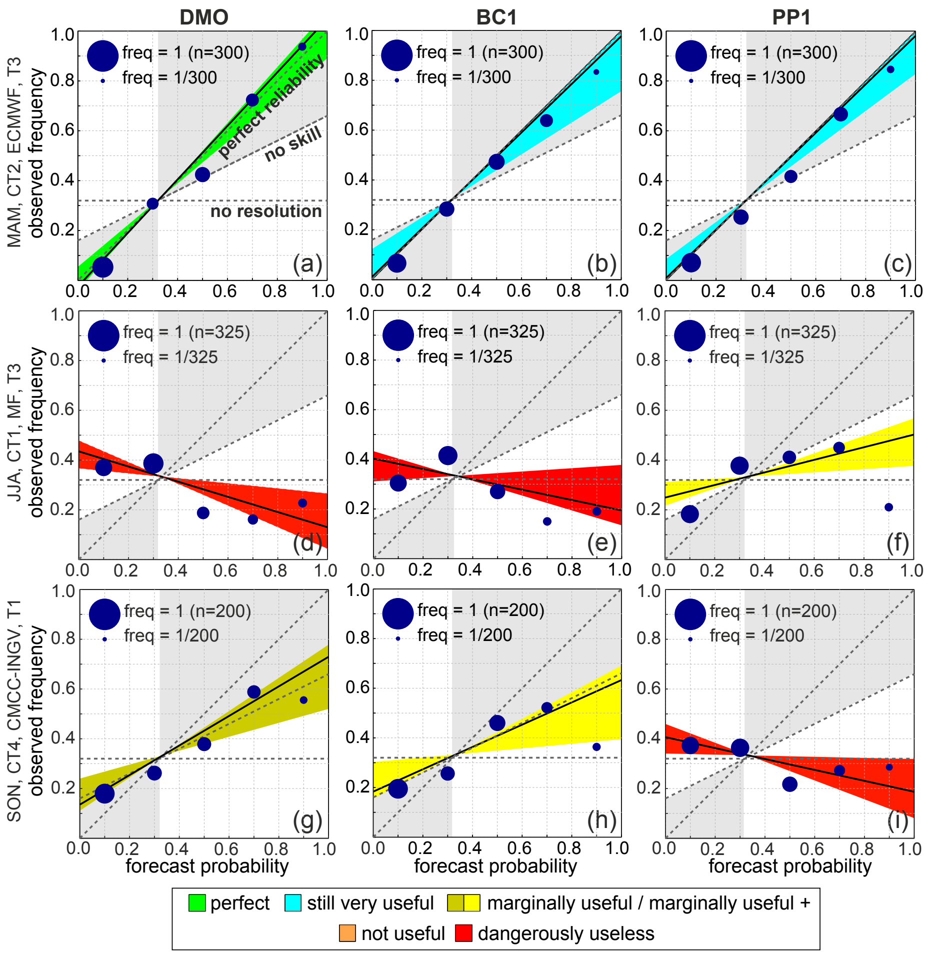

Fig. 2 Reliability diagrams for the raw/direct model output (DMO), the BC1 and the PP1 method (in columns), for three different illustrative examples of seasonal forecasts in MAM, JJA and SON (in rows), for different CTs and models (see the labels on the left-hand side). The gray area defines the region contributing positively to the forecast skill (Brier Skill Score > 0). The *perfect reliability*, *no skill* and *no resolution* lines are indicated in panel *a*. Colors correspond to the different categories used, which are based on the original scale proposed by Weisheimer and Palmer (2014) (see the text for details). Note that the joined series of the different stations falling within each CT are considered. The sample size used in each case is indicated in the upper left corner.

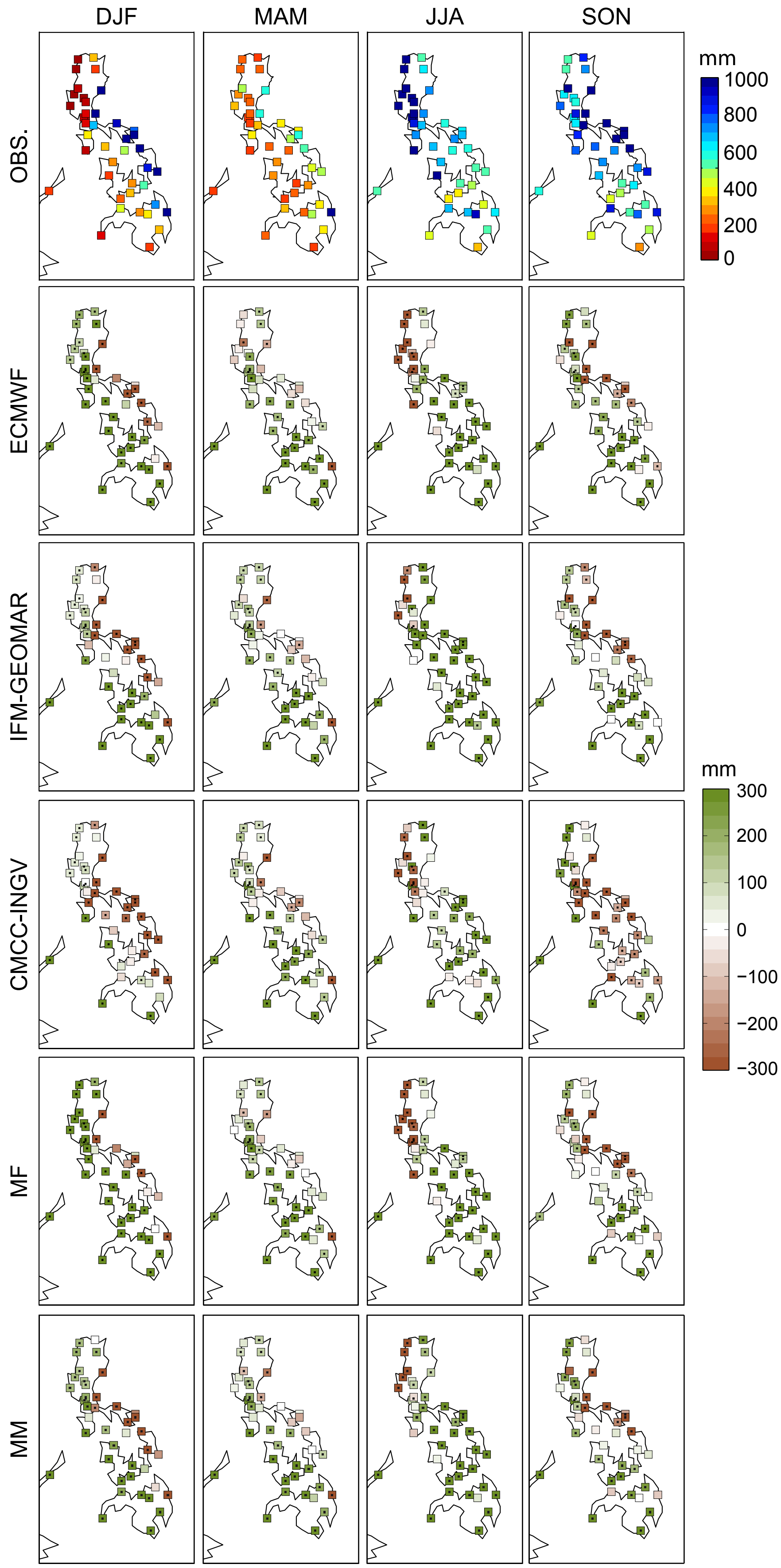

Fig. 3 First row: Observed seasonal climatologies (in mm/season) at the 42 PAGASA stations. Rest of rows: Bias (in mm/season) for the four ENSEMBLES models and the multimodel, by seasons (in columns). Significant ($\alpha = 0.05$, according to a Student's t-test) values are indicated with a black dot.

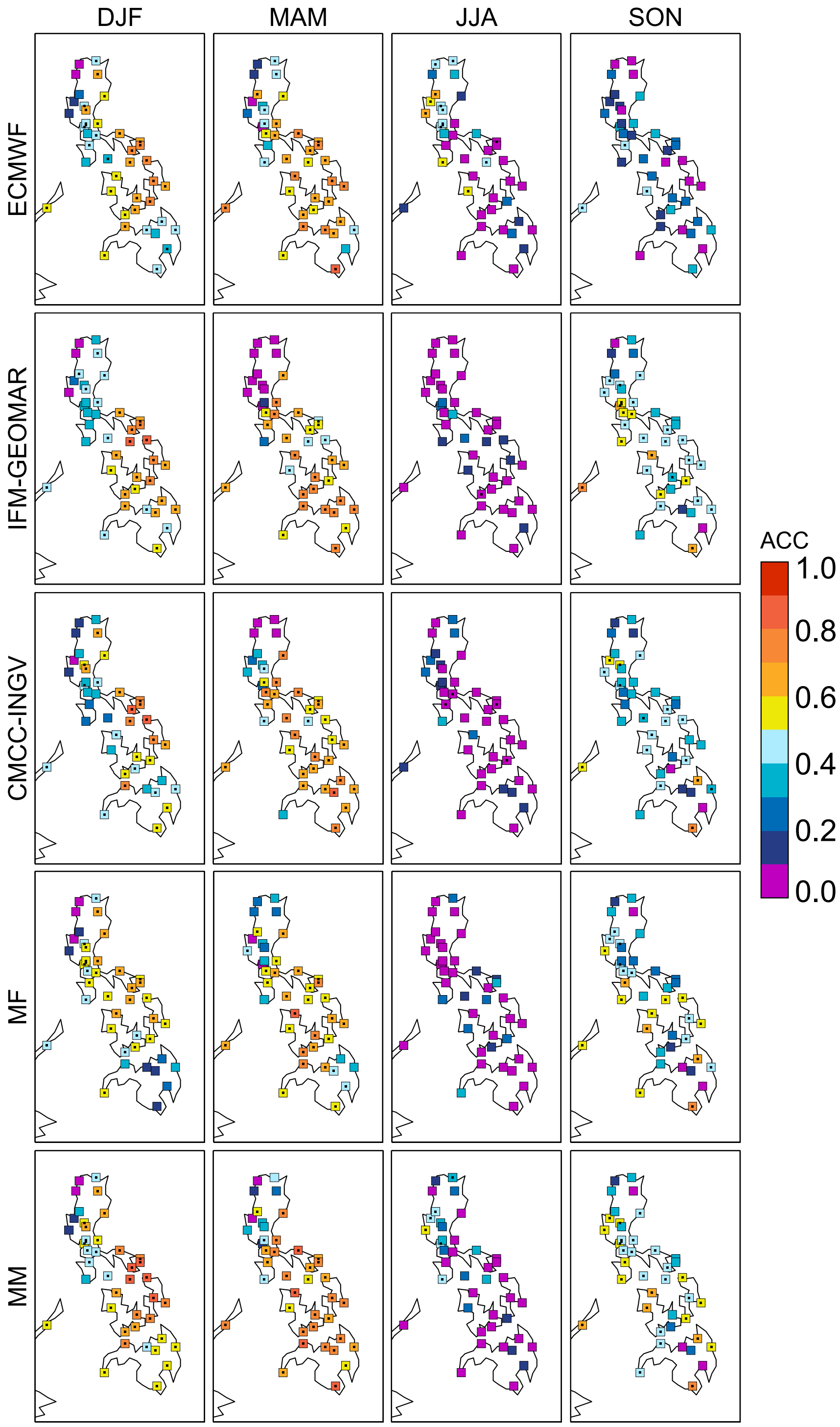

Fig. 4 Interannual ACC values obtained at the 42 PAGASA stations for the four ENSEMBLES models and the multimodel (in rows), by seasons (in columns). Significant ($\alpha = 0.05$, according to a Student's t-test) values are indicated with a black dot.

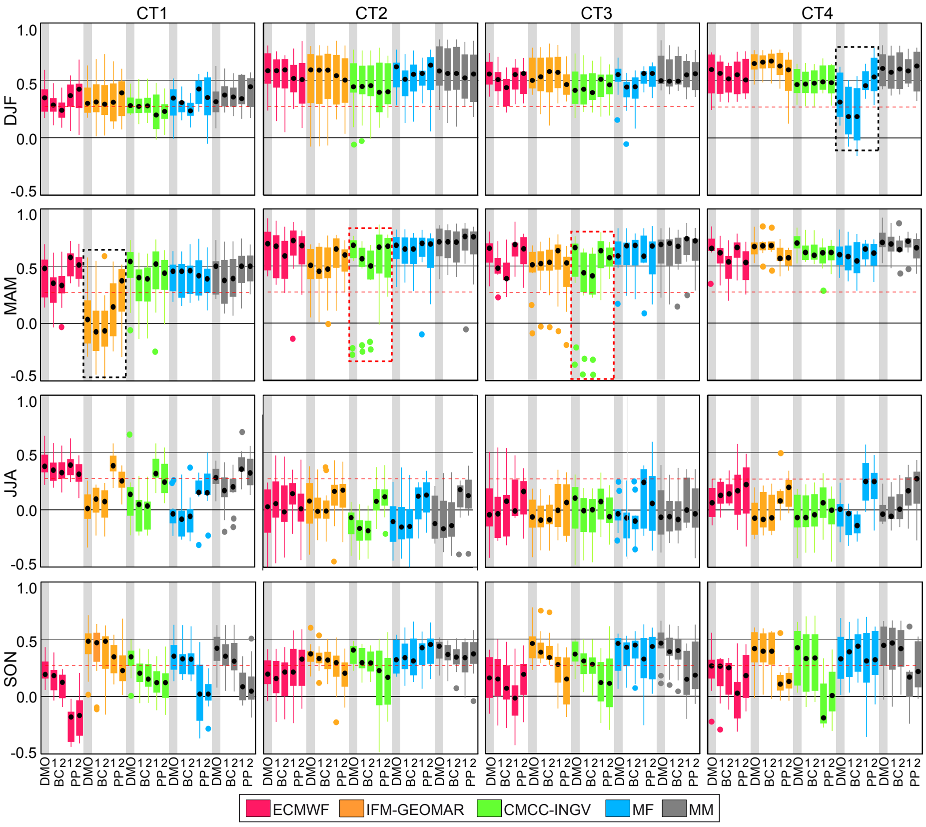

Fig. 5 Interannual ACC obtained for the different seasons (in rows) and CTs (in columns). In each panel, results for each model are shown in different colors (see the legend). From left to right, boxplots display the correlations obtained along the different stations for the DMO (indicated by a light gray shadow) and the BC1, BC2, PP1 and PP2 methods. Significant ($\alpha = 0.1$, according to a Student's t-test) values are those above the red dashed lines. Dashed boxes indicate particular situations which are described in the text.

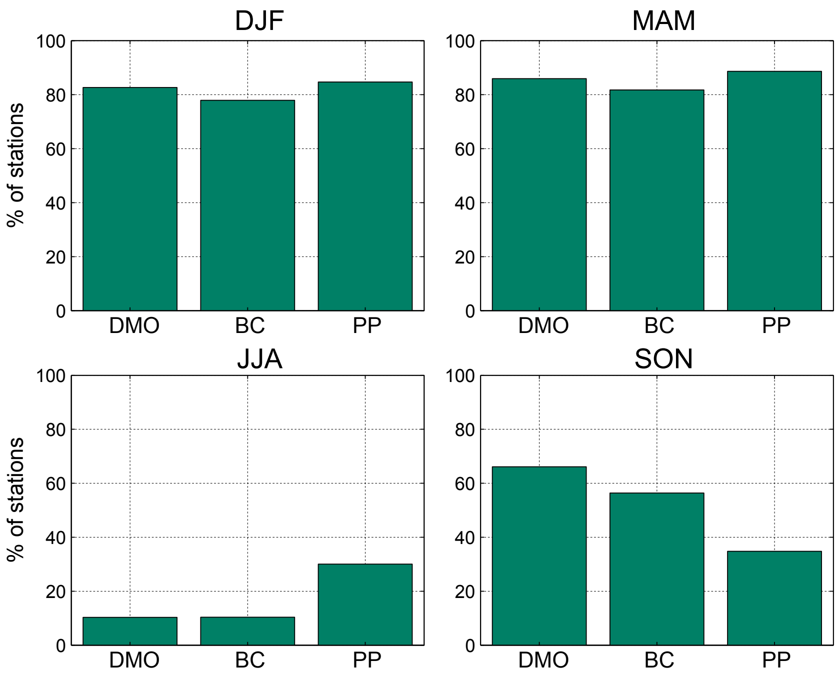

Fig. 6 Summary of Figure 5 showing in bar charts the percentage of stations with significant (α = 0.1, according to a Student's t-test) interannual ACC for the DMO and the BC and PP downscaling approaches, for the different seasons. Within each approach, the two methods considered are jointly analyzed. Moreover, all models except the MM (which is excluded for clarity) and all CTs are also jointly considered.

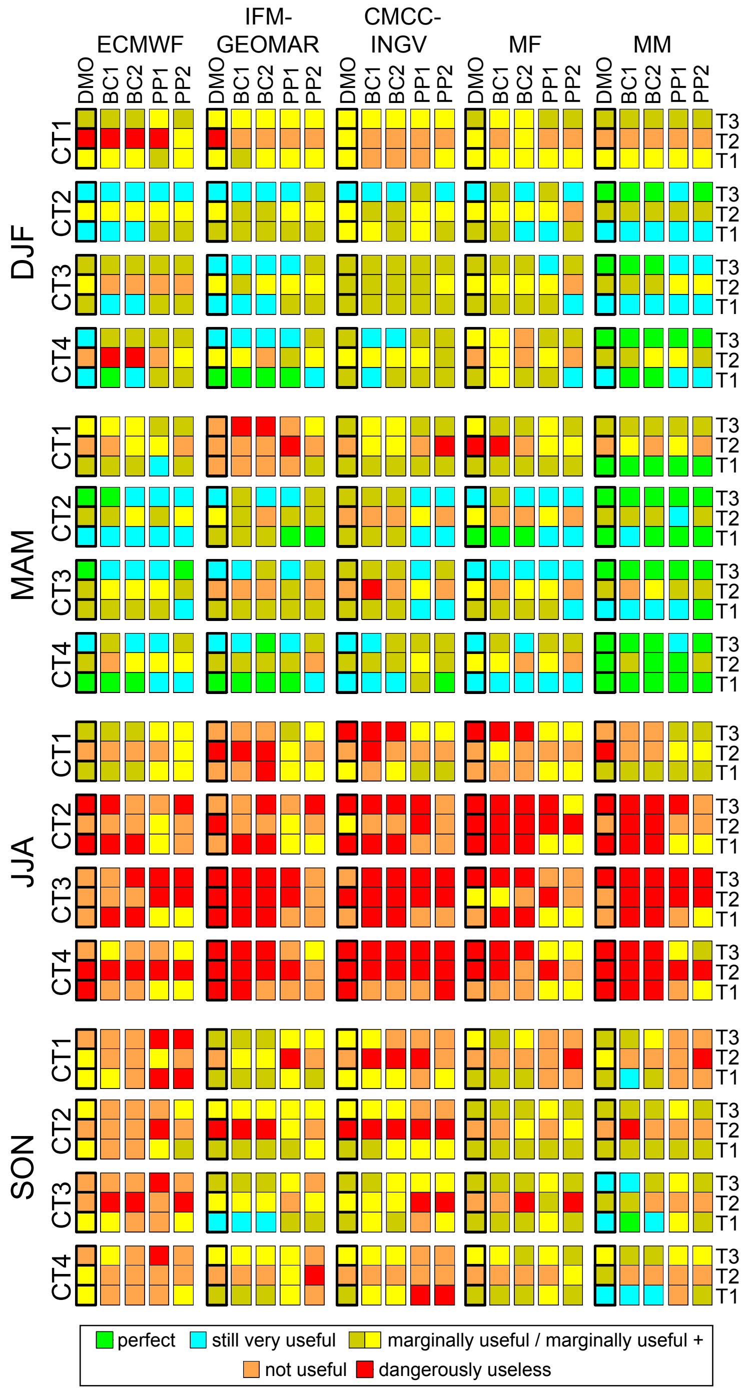

Fig. 7 Reliability categories obtained for the different ENSEMBLES models (in columns) along the different seasons and CTs (in rows). Each block shows the results obtained for the DMO, the two BC and the two PP methods considered, for the three terciles (T1, T2 and T3). Colors correspond to the different categories used, which are based on the original classification proposed by Weisheimer and Palmer (2014) (see the text for details).

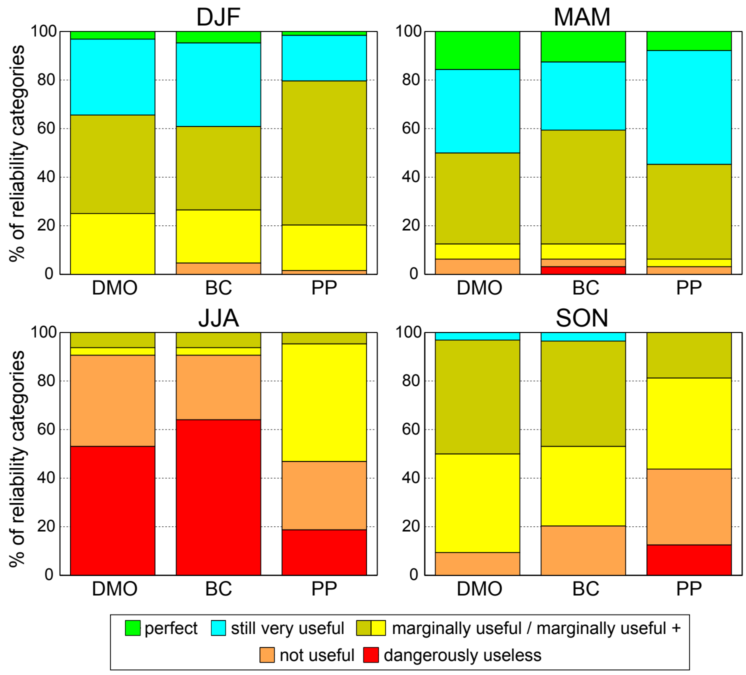

Fig. 8 Stacked bar charts with the percentage of reliability categories (in colors) for the DMO and the BC and PP approaches (within each approach, the two methods considered are jointly analyzed) for the different seasons. For clarity, results from the MM and from the normal tercile (T2) are excluded from this analysis.

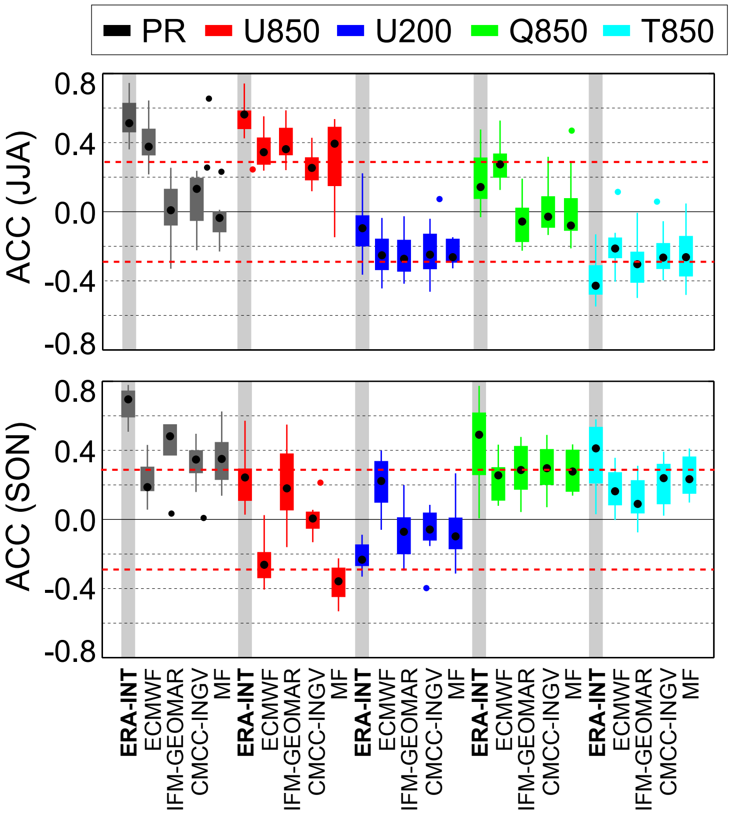

Fig. 9 Interannual ACC values between observed precipitation at the 13 stations pertaining to CT1 and the corresponding ERA-Interim and ENSEMBLES models outputs —the nearest gridbox is considered— for precipitation (PR) and the different predictors used (U850, U200, Q850 and T850) for (top) JJA and (bottom) SON. Significant ($\alpha = 0.1$) positive (negative) values are those above (below) the upper (lower) red dashed line.