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

Statistical adjustment, calibration and downscaling of seasonal forecasts: A case-study for Southeast Asia

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

Received: date / Accepted: date

Abstract The present paper is a follow-on of the work presented in Manzanas et al (2019) which provides a comprehensive intercomparison of alternatives for the post-processing (statistical adjustment, calibration and downscaling) of seasonal forecasts for a particularly interesting region, Southeast Asia. To answer the questions that were raised in the preceding work, apart from Bias Adjustment (BA) and ensemble Re-Calibration (RC) methods —which transform directly the variable of interest,— we include here more complex Perfect Prognosis (PP) and Model Outputs Statistics (MOS) downscaling techniques —which operate on a selection of large-scale model circulation variables linked to the local observed variable of interest.— Moreover, we test the suitability of BA and PP methods for the post-processing of daily —not only seasonal— time-series, which are often needed in a variety of sectoral applications (crop, hydrology, etc.) or to compute specific climate indices (heat waves, fire weather index, etc.). In addition, we also undertake an assessment of the effect that observational uncertainty may have for statistical post-processing.
Our results indicate that PP methods (and to a lesser extent MOS) are highly case-dependent and their application must be carefully analyzed for the region/season/application of interest, since they can either improve or degrade the raw model outputs. Therefore, for those cases for which the use of these methods cannot be carefully tested

R. Manzanas (✉)
Meteorology Group. Dpto. de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Santander, Spain. E-mail: rodrigo.manzanas@unican.es

J. M. Gutiérrez
Meteorology Group. Institute of Physics of Cantabria (IFCA), CSIC-University of Cantabria, Santander, Spain

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

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

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

by experts, our overall recommendation would be the use of BA methods, which seem to be a safe, easy to implement alternative that provide competitive results in most situations. Nevertheless, all methods (including BA ones) seem to be sensitive to observational uncertainty, especially regarding the reproduction of extremes and spells. For MOS and PP methods, this issue can even lead to important regional differences in interannual skill. The lessons learnt from this work can substantially benefit a wide range of end-users in different socio-economic sectors, and can also have important implications for the development of high-quality climate services.

1 Introduction

The state-of-the-art General Circulation Models (GCMs) used for seasonal forecasting suffer from important systematic biases (mean errors) and drifts (leadtime-dependent biases) and have horizontal resolutions which are typically coarser than those needed for practical applications (see, e.g., Doblas-Reyes et al, 2013; Manzanas et al, 2014a). Therefore, some form of post-processing (i.e. adjustment, calibration and/or downscaling) is needed in order to make their raw outputs usable. In a recent study, Manzanas et al (2019) intercompared the performance of Bias Adjustment (BA) —e.g. quantile mapping— and ensemble ReCalibration (RC) —e.g. non-homogeneous Gaussian regression— methods for the adjustment/calibration of seasonal aggregated forecasts. At this particular timescale, they found that the RC methods can result in modest improvement of some quality aspects (in particular reliability), although other aspects can be degraded. Nevertheless, these improvements are restricted to regions/seasons with high model skill. In addition, these methods can be negatively affected by the limited length of state-of-the-art seasonal hindcasts (which typically have less than 30 years). They also found that, beyond removing their systematic biases, BA methods can not improve the skill of the raw model forecasts (even more, some quality aspects can be degraded), since they do not modify their temporal structure. However, the application of these methods is straightforward and may constitute a pragmatic and simple alternative when the resolution of the model is similar to that of the observational reference (BA methods are not suitable for downscaling), or for regions with no expected potential for downscaling (e.g. flat inland regions). Moreover, beyond the adjustment of monthly/seasonal values, Manzanas et al (2019) pointed out the fact that BA techniques can be also applied to adjust daily data, which are often demanded in a variety of sectoral applications in order to run impact models (crop, hydrology, etc.) or to compute specific climate indices (heat waves, length of growing index, thermal comfort index, fire weather index, etc.).

Therefore, we put a special focus in this work on the post-processing of daily (rather than monthly/seasonal) values. For this aim, we consider not only BA methods acting directly on the variable of interest, but also more complex Perfect Prognosis (PP) downscaling techniques (see, e.g., Gutiérrez et al, 2013) which operate on a selection of large-scale model circulation variables (predictors) linked to the local observed variable of interest (predictand). Although there has been some indication that PP methods may add some value in terms of skill (e.g. interannual correlation) for cases where the dynamical model is better at reproducing the relevant large-scale features than the target variable being predicted (Manzanas et al, 2018), they have the extra complexity of building the predictor-predictand relationship at a daily basis using reanalysis data (which provide day-to-day correspondence with observations). Typically, this requires a highly time-consuming screening process to detect robust predictors which are similarly represented in both the reanalysis and hindcast datasets. Moreover, PP methods may suffer from reanalysis uncertainty, which is particularly relevant in tropical regions (Brands et al, 2012; Manzanas et al, 2015). Therefore, in this type of methods, the existing windows of opportunity for improvement can be so narrow that the effort may be disproportionate to the benefit.

Moreover, we also include in this study Model Output Statistics (MOS) downscaling methods (see, e.g., Vannitsem and Nicolis, 2008), which are trained with predictors taken from the same GCM that is being postprocessed. A simple implementation of these methods considers as the only predictor variable the target predictand, e.g., coarse GCM precipitation for local precipitation. Following Manzanas et al (2019), these methods are included as part of the RC approach in this work. Standard downscaling MOS implementations consider large-scale variables from the GCM as predictors (see, e.g., Manzanas et al, 2017). These are referred to as MOS hereafter. Note that, as the relationship between the large-scale seasonal forecasts and observational reference records is established using directly the hindcast (without passing through reanalysis), the complexity and requirements for MOS methods are much lower than for PP ones. However, as for the case of RC methods, the main shortcoming of these techniques is that they can only be applied on monthly/seasonal data, since GCM predictors do not keep temporal correspondence with the local observations at the daily scale.

Given the complexity of this panorama, the relative merits and limitations of the approaches and techniques available for post-processing of seasonal forecasts need to be properly assessed. This is done here by intercomparing the performance of the alternatives described above based on different aspects of forecast quality: association, accuracy and discrimination for seasonally aggregated times-series and reproduction of extremes and spells for daily time-series. Besides, following from the fact that all the adjustment/calibration/downscaling methods rely on observations for the training process, observational uncertainty (see, e.g. Kotlarski et al, 2017; Herrera et al, 2018) may play a role in the statistical post-processing of model forecasts. To shed some light on this potential issue, we also undertake here a comprehensive assessment of the effect of this kind of uncertainty in the context of seasonal forecasting.

Jointly with the work done in Manzanas et al (2019), this study provides practical recommendations for the suitable post-processing of seasonal forecasts, which can substantially benefit a wide range of end-users in different socio-economic sectors, and can also have important implications for the development of high-quality climate services (see, e.g., Torralba et al, 2017).

The paper is organized as follows. In Section 2 we describe the data used and introduce the different methods applied and the verification metrics considered. The results obtained are presented through Section 3. The main conclusions obtained and a set of practical user recommendations are outlined in Section 4.

Table 1 Potential predictor variables considered for the MOS and PP methods.

Code	Variable	Levels
SLP	Mean sea level pressure	Surface
Z	Geopotential height	850, 500, 300 (mb)
T	Temperature	850, 500, 300 (mb)
Q	Specific humidity	850, 500, 300 (mb)
U	Zonal component of wind	850, 500, 300 (mb)
V	Meridional component of wind	850, 500, 300 (mb)

2 Data and Methods

2.1 Data Used

We focus in this work on one illustrative region (Southeast Asia: 95-140° E, 10° S-20° N) and season (boreal winter: DJF), for which overall good skill has been documented (see, e.g., Manzanas et al, 2014b). As explained later, the choice of this region is also supported by the fact that a high-quality observational grid is available —SA-OBS (van den Besselaar et al, 2017),— which allows for an interesting analysis of the effect of observational uncertainty on the results obtained from the different post-processing techniques (see Section 3.2).

We consider one-month lead seasonal forecasts (i.e. predictions initialized in November) of both temperature and precipitation from the ECMWF-System4 (Molteni et al, 2011), which provides the longest seasonal hindcast to-date —note that one of the main conclusions of Manzanas et al (2019) is that as long as possible hindcasts are needed for robust adjustment/calibration.— In particular, we use here all the 51 members that are available for the November initialization (only 15 members are available for other initializations) along the period 1982-2014.

Besides the target variables of interest (temperature and precipitation) used for BA and RC methods, the large-scale variables listed in Table 1 were considered as potential predictors for MOS and PP methods in this work. For the training phase of the PP methods, these predictor variables are taken from ERA-interim reanalysis (Dee et al, 2011). In this case, ERA-Interim and ECMWF-System4 data are harmonized by performing a simple local scaling 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 model mean biases.

We consider ERA-Interim as the common observational reference along the study. However, for the assessment of the effect of observational uncertainty undertaken in Section 3.2, we also consider two other datasets for precipitation: SA-OBS and MSWEP. SA-OBS a high-quality observational dataset which provides daily gridded (0.25° spatial resolution) temperature and precipitation over land for Southeast Asia. It has been built based on more than 8000 meteorological stations and can be freely downloaded from `http://sacad.database.bmkg.go.id`. MSWEP (version 1) (Beck et al, 2017) is a global terrestrial precipitation dataset with a high 3-hourly temporal and 0.25° spatial resolution which combines gauge, satellite and reanalysis information. For the sake of comparability with the results shown in Manzanas et al (2019), all the different datasets used here (ECMWF-System4, ERA-Interim, SA-OBS and MSWEP) have been bi-linearly interpo-

Table 2 Validation metrics considered in this work.

Code	Description	Variable
Cor.	Correlation	Temp., precip.
CRPS	Continuous Ranked Probability Score	Temp., precip.
RPS	Ranked Probability Score	Temp., precip.
ROCA	ROC Skill Area	Temp., precip.
P2, P98	Percentile 2, percentile 98	Temp.
P98-wet	Percentile 98 of wet (precip. >= 1 mm) days	Precip.
R01	Frequency (in %) of wet days	Precip.
ColdSpellP90	Percentile 90 of the length of cold spells	Temp.
WarmSpellP90	Percentile 90 of the length of warm spells	Temp.
WetSpellP90t	Percentile 90 of the length of wet spells	Precip.
DrySpellP90t	Percentile 90 of the length of dry spells	Precip.

lated from their native horizontal resolutions to the common 1° regular grid in which the C3S models are provided through the Climate Data Store (see `http:/climate.copernicus.eu/seasonal-forecasts`). Moreover, daily data have been used in all cases.

2.2 Validation Metrics

We have used for this study the Continuous Ranked Probability Score (CRPS), the Ranked Probability Score (RPS), the ROC Skill Area (ROCA) and the Pearson correlation to validate the interannual series (the daily results from BA and PP are seasonally aggregated in this case). RPS and ROCA are used for tercile-based probabilistic predictions, being the terciles independently computed for the observations and the predictions. Therefore, whereas CRPS is sensitive to changes in the mean and variance (and hence to the effect of bias adjustment), the rest of measures are not so they allow to explore the added value of the post-processing techniques beyond the model bias removal. The reader is referred to Manzanas et al (2019) for further details about the metrics considered. Moreover, for those methods providing daily outputs, we also focus on further aspects of the forecasts such as extremes and spells, which are of special interest for many practical applications. In particular, we have considered the 2nd and 98th percentiles for daily temperature and the 98th percentile for daily precipitation (for the latter, only wet days are considered). Additionally, for the case of precipitation, the frequency of rainy days is also validated. Besides, the 90th percentile of the length of spells is also analyzed. As in Maraun et al (2018), a cold/warm (dry/wet) spell is defined as an episode of two or more consecutive days with values below/above the 10/90th percentile (1 mm). These indicators are computed separately for each ensemble member and the results are validated in a deterministic way based on the ensemble mean. All the validation metrics considered in this work are shown in Table 2.

2.3 Methods

Among BA methods, we have considered two different implementations of quantile mapping; one parametric and one empirical. The latter corresponds to the EQM method presented in Manzanas et al (2019), which is applied here on daily (instead of seasonal) data. The former (referred to as PQM henceforth) is based on the assumption that both observations and raw GCM outputs are well approximated by a given distribution (Gaussian for temperature and Gamma for precipitation), so only the parameters of the theoretical distributions are mapped (see, e.g., Themeßl et al, 2012). For the case of precipitation, the EQM method used here incorporates a frequency adaptation which is thought to alleviate the problem that arises when the frequency of dry days is larger in the model than in the observations (Themeßl et al, 2012). Note that quantile mapping is able to correct automatically the excess of light precipitation frequency or “drizzle effect”.

As representative of the RC family, we have considered the LR method introduced in Manzanas et al (2019), which performs a linear regression between the ensemble mean and the corresponding observations. To correct the forecast variance, the standardized anomalies are rescaled by the standard deviation of the predictive distribution from the linear fit. LR was shown in Manzanas et al (2019) to provide in general good results with a relatively low computational cost. Recall that this method calibrates directly the model temperature (precipitation), based on observed temperature (precipitation). Besides, we have also considered a MOS downscaling configuration in which this same LR method is applied considering T850 (Q300) —see Table 1— as unique predictor to forecast temperature (precipitation). As a compromise between capturing some skill in the model predictors (e.g. correlation with reanalysis data) and retaining a sufficiently large sample size for calibration, the LR method is applied in this work on the monthly means in both cases (referred hereafter to as LR and MOS-LR, respectively).

Among the wide range of alternatives proposed in the literature for PP downscaling, we have selected three of the most representative ones: Multiple Linear Regression (MLR), Generalized Linear Models (GLMs) and the analog technique. MLR (GLMs) are used in this work to downscale temperature (precipitation). The analog technique is common to both predictand variables. MLR is an extension of simple linear regression which attempts to model the relationship between two or more explanatory predictors and the predictand by fitting a linear equation by minimizing the sum of the residuals between the regression line and the observed data. A detailed description on the theory of this technique is provided by Helsel and Hirsch (2002). Regression-based methods have also been used in previous works to downscale seasonal forecasts of temperature (see, e.g., Pavan et al, 2005). GLMs were formulated by Nelder and Wedderburn (1972) in the 1970s and are an extension of the classical linear regression which allows to model the expected value for non-normally distributed variables. GLMs have been already applied to downscale seasonal forecasts (Manzanas et al, 2018). We follow here the two-stage implementation used in the latter reference, in which a GLM with Bernoulli error distribution and logit canonical link-function (also known as logistic regression) is applied to downscale daily precipitation occurrence (as characterized by a threshold of 1mm) and a GLM with gamma error distribution and log canonical link-function is used to downscale daily precipitation amount. In order to increase the predicted variance, which is usually underestimated in deterministic configurations (Enke, 1997), we introduce here a stochastic component in both GLMs (see Manzanas, 2016, for details). For this method, we considered as predictors the standardized anomalies of the predictors considered at the nearest model gridbox (for each predictand location). The popular analog technique (Lorenz, 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, only the closest analog is considered (Zorita et al, 1995; Cubasch et al, 1996). Analogs are defined based on the standardized anomalies of the predictors considered at the 16 nearest model gridboxes (i.e., over a 4x4 square centered around each predictand location which allows to encompass the main synoptic phenomena influencing the local climate) and the Euclidean norm is considered. 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; Manzanas et al, 2018). 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.

To avoid the artificial performance that may derive from model overfitting, all the methods considered in this work are applied under a Leave-One year-Out (LOO) cross-validation (Lachenbruch and Mickey, 1968) scheme, in which each year was separately considered for test, whilst the remaining ones were kept for training. Note that this is the most adequate framework to test the potential usefulness of any method for operational seasonal forecasting.

2.4 Selection of predictors for MOS and PP methods

To cope with the issue of predictor selection in PP methods (see, e.g., Gutiérrez et al, 2013; San-Martín et al, 2016), Figure 1 shows the existing correlation between each of the large-scale variables listed in Table 1 and local temperature (left) and precipitation (right), computed on the daily time-series. The idea behind this analysis is that the higher the correlation (either positive or negative), the stronger the physical link between predictor and predictand is, which allows to make an initial selection of explicative predictors for PP downscaling. However, Manzanas et al (2018) have shown that the results coming out from PP methods in the context of seasonal forecasting also depend on the skill of the model predictors considered. Therefore, both the strength of the predictor-predictand relationship and the skill of the model in reproducing the large-scale should be taken into account when making the final selection of predictors for PP methods.

Figure 2 shows the interannual correlation between ERA-Interim and ECMWF-System4 for each of the variables listed in Table 1. Whereas high skill (understood as the agreement between model and reanalysis) is found for SLP, geopotential height and temperatures, significant discrepancies appear for some humidity fields (in particular Q850) and winds (both U and V). For this reason, we have excluded Q850 and winds from the set of potential predictor variables, since they might negatively affect the results obtained from PP (and MOS) methods. With this limitation in mind, and with the idea of keeping the predictor sets as simple as possible, the final combination considered for temperature (precipitation) was SLP+T850 (SLP+Q300). Note that, for the particular case of precipitation, although Q850 may be more explicative than Q300 (Figure 1), the former variable was discarded in favor of the latter since it is not well reproduced by the ECMWF-System4 (Figure 2).

For consistency with the LR method, T850 (Q300) is considered as unique predictor in the MOS configuration used here to predict temperature (precipitation).

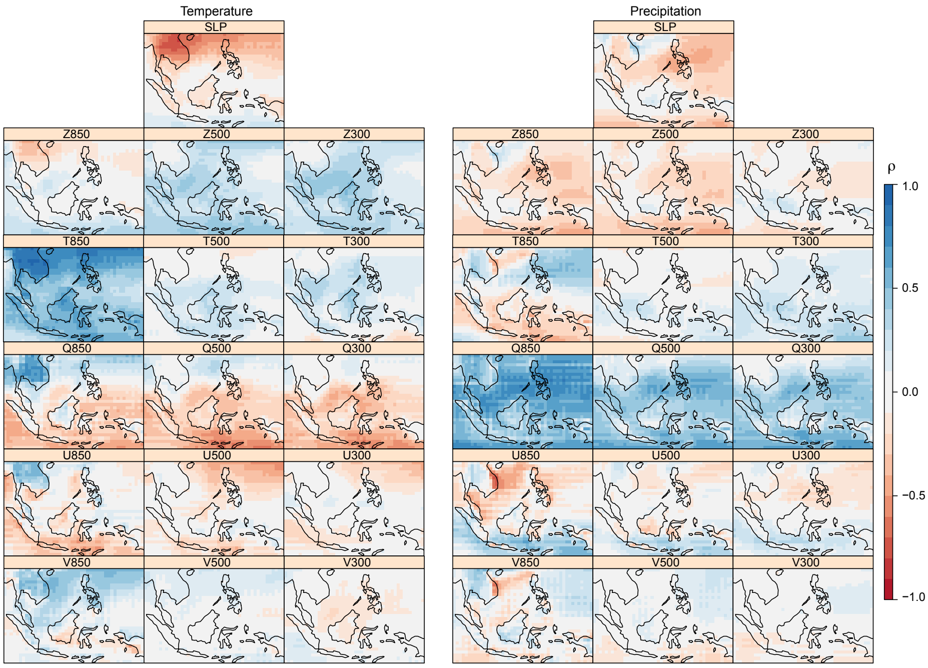

Fig. 1 Correlation between each of the large-scale predictors listed in Table 1 and local temperature (left) and precipitation (right), computed on the daily time-series.

3 Results

3.1 Intercomparison of approaches and methods

The top/bottom panel in Figure 3 shows the validation results obtained for the raw and post-processed interannual predictions of temperature/precipitation, in terms of different metrics (in rows). In all cases, column 1 refers to the raw model outputs. The rest of columns correspond to the different methods considered from the different approaches (BC: columns 2-3, RC: column 4, MOS: column 5 and PP: columns 6-7). For all of them, results are expressed with respect to those shown in column 1, either as skill scores (CRPSS, RPSS and ROCSS) or as direct differences (for correlation). Thus, values above (below) 0, shown in blue (red), indicate that the particular method improves (degrades) the raw model prediction. Note that the RPSS and the ROCSS are computed for probabilistic forecasts of tercile categories, which are separately computed for the observations and the predictions (this entails an implicit bias adjustment in the forecasts).

This figure indicates that all the methods tested here provide a clear benefit in the CRPSS, which is a consequence of effectively removing the important model biases present over the region (see Figure 1 in Manzanas et al (2019)). Note that this result —which was already found for BA and RC methods in Manzanas et al (2019)— is key, since unbiased predictions are needed by many different communities to run their seasonal impact models. However, beyond this improvement in the CRPSS, neither BA nor RC techniques (the latter represented by the LR method) are able to outperform the raw forecasts for any of the remaining metrics, leading in general to slightly worse results over the entire domain for all of them. This deterioration is even more evident for the LR method, and especially for correlation —note that RC methods can lead to artificial anti-skill (i.e. anti-correlations) in regions of small (or negative) raw model correlations (Eade et al, 2014).— It is worth to mention that the EQM tested here (and also the PQM) lead only to slightly better results than those shown for the same method in Manzanas et al (2019), where it was applied on the seasonal (instead of daily) time-series. Moreover, to assess the dependency of the results provided by BA methods on the temporal resolution considered, both EQM and PQM were also applied on the monthly time-series, finding only slightly worse (better) results than in the daily (seasonal) case. Therefore, we do not recommend the application of BA methods on daily data in case only monthly/seasonal data is needed (note that the slight improvement found for higher temporal resolutions does not compensate the increasing computational costs).

Differently from BA and RC, MOS and PP methods provide much more local results, being possible to find areas where the downscaled predictions either outperform or degrade (notably in some cases) the raw model forecasts. These results are in agreement with those found in Manzanas et al (2018), who suggested that the suitable application of PP methods was subjected to particular (and limited) windows of opportunity for which 1) there exists a strong link between the large- and the local-scale and 2) the model is better at reproducing the relevant large-scale predictors considered for downscaling than the local predictand of interest (this can typically happen for variables needing some kind of parametrization, such as precipitation). Again, the results from this work warn on the unexpert use of MOS and PP methods, as they must be carefully analyzed for the particular case-study of interest.

Figure 4 shows the results obtained for the extreme and spell indicators. Whereas column 1 corresponds to the observations, column 2 corresponds to the raw model outputs and columns 3-7 to the different the methods considered. In columns 2-7, the results are expressed as differences (e.g. bias) with respect to the observed values of column 1. Note that neither the RC nor the MOS version of the LR method are considered for this analysis since it cannot be applied at a daily scale. For temperature, the cold bias exhibited by the model in the analyzed percentiles is corrected by all methods except the MLR, which exhibits a warm (cold) bias for the 2nd (98th) percentile. This is due to an underestimation of the predicted variance which is typical of these methods, and could be alleviated by introducing some inflation procedure (see, e.g., Huth, 1999). For spells, the two BA methods maintain the same errors exhibited by the model (the more green/brown, the longer/shorter the predicted spell is, as compared to observations), since they are not able to modify its temporal structure. Differently, since PP methods can alter this temporal structure, they are found to modify the spatial patterns exhibited for the model, being possible to find some areas where the model error is reduced. However, they can also introduce errors in new regions which can be even higher than those present in the raw model.

For precipitation, the two BA methods lead to different results. In particular, similarly as for temperature, the PQM method inherits a great part of the errors exhibited by the raw model, which are only partially corrected (see the results obtained for the frequency of rainy days and the percentile 98th of rainy days). However, as a consequence of the frequency adaptation implemented, these errors are corrected to a higher extent in the EQM method. Despite they lead in general to higher errors than the EQM, the spatial patterns found for the PP methods are, in some cases, more uniform (see, e.g., the results obtained for the 98th percentile of rainy days in the GLM method). Note that, in such situations, simple a-posteriori corrections (e.g. scaling) could be easily applied to further improve the results obtained for PP methods.

In summary, despite correcting marginal aspects such as extreme percentiles, our results indicate that BA methods are not in principle a good candidate to correct spells, since they mostly inherit the errors present in the model. However, for the particular case of precipitation, and provided that some form of frequency adaptation is applied, these methods can be a good alternative (see the results for the EQM). However, as main shortcoming, these methods do not improve (or even slightly degrade) the interannual model skill (see the results obtained for correlation, RPSS and ROCSS in Figure 3). Differently, PP methods are highly case-dependent and their application must be carefully analyzed for the case-study of interest, since they can either improve or degrade the raw model outputs. The strongest advantage of PP methods is that, whilst being competitive (as compared to BA ones) over some regions for predicting extremes and spells, it is possible to find windows of opportunity for which interannual model skill can be also improved (regions/seasons for which the model skill is higher for the large-scale than for the target predictand). Nevertheless, when the predictors selected for downscaling are not well reproduced by the model, PP methods can also lead to unsuitable results. For instance, if Q300 is substituted by Q850 in the predictor set used to downscale precipitation, the results shown in Figures 3 and 4 strongly worsen (not shown). As suggested in Manzanas et al (2018), an explanation for this behaviour comes from the fact that the model skill for reproducing Q850 is more limited (see Figure 2). As a result, the statistical link that is learnt using reanalysis data in PP methods becomes meaningless when applied to model predictors (the use of Q850 instead of Q300 leads to much better cross-validated results when using reanalysis predictors; not shown).

3.2 The effect of observational uncertainty

Observational uncertainty has been identified as one of the factors that may play a role in the statistical post-processing of model forecasts (see, e.g. Kotlarski et al, 2017; Herrera et al, 2018), since all the adjustment/calibration/downscaling methods rely on observations for the training process. To assess the potential impact of this factor, we repeat in this section some of the analysis above presented but replacing ERA-Interim by both SA-OBS and MSWEP.

In particular, we focus on precipitation —for which observational uncertainty is known to be larger— and consider SA-OBS (the only dataset purely based on gauge data) as the ground truth, since it has been found to closely resemble punctual gauge-based measures in terms of dry/wet frequency, timing of rainy days and extremes (van den Besselaar et al, 2017). Figure 5 provides a comparison between ERA-Interim/MSWEP and SA-OBS (left/middle column), in terms of their interannual time-series. In addition, ERA-Interim and MSWEP are also compared (right column). Whereas ERA-Interim and MSWEP show in general good agreement (with correlation values above 0.8 in most of the gridboxes), important differences are found between ERA-Interim and SA-OBS (with rather low, or even negative values over certain parts such as Sumatra). Comparison between ERA-Interim and MSWEP yields intermediate results. These findings point out the limitations of reanalysis data to reproduce the actual climate of the region, which presents thousands of islands, strong land-sea contrasts and a complex topography. In this regard, note that the inclusion of satellite information in MSWEP helps to correct the deviations from reality found in ERA-Interim.

For each of the metrics shown in Figure 6 (7), the middle/bottom row would be the equivalent to those shown in Figure 3 (4) but using SA-OBS/MSWEP instead of ERA-Interim for both training and verification of the different methods. For direct comparison, the top row shows the same results presented in Section 3.1, but only over land. Whereas the results for the interannual time-series (Figure 6) are almost identical for ERA-Interim and MSWEP —note from the comparison against raw model outputs (left column) that both datasets are very similar,— some regional differences (see, e.g., over Borneo and Papua) appear with respect to the results found for SA-OBS, in particular for MOS and PP methods (this effect is less pronounced for BA ones). However, when it comes to the extreme and spell indicators (Figure 7), these differences become more relevant and not only for MOS and PP methods, but also for BA ones. For instance, important performance discrepancies are found for most of the indicators for the case of the PQM method depending on the reference considered (even between ERA-Interim and MSWEP). Although analyzing in detail all the differences found region by region and method by method is not the purpose here, Figures 6 and 7 reveal that the choice of observational dataset can have important effects for the post-processing of seasonal forecasts. This issue seems to be specially relevant for MOS and PP methods, for which notable differences are found even in terms of interannual skill. This poses an important challenge for seasonal forecasting; in particular over the tropics, where large observational uncertainty has been identified, not only for observations but also for reanalysis (see, e.g., Brands et al, 2012; Manzanas et al, 2015). Moreover, seasonal models tend to exhibit the highest interannual skill in tropical latitudes (see, e.g., Manzanas et al, 2014b), being thus difficult to improve their raw forecasts there. As a consequence of these limitations, BA methods may be, in general, a more secure alternative for downscaling in the tropics. Nevertheless, beyond interannual skill, it is very important to warn on the potential conflicts that may arise related to the choice of observational uncertainty, even for BA methods, in terms of other forecast aspects such as extremes and spells.

4 Conclusions and User Recommendations

This section summarizes the main conclusions obtained in Manzanas et al (2019) and in this work and provides a set of recommendations for practitioners on the advantages and limitations of the different approaches available for the appropriate post-processing of dynamical seasonal forecasts. These approaches, which aim to reduce the systematic model biases and increase their skill (as measured by different quality aspects), range from bias adjustment (BA) and ensemble recalibration (RC) methods —both acting directly on the variable of interest; e.g., model precipitation— to more complex statistical downscaling techniques such as Model Output Statistics (MOS) and Perfect Prognosis (PP) methods —which operate on a selection of large-scale circulation predictor variables (e.g. model geopotential and humidity at different vertical levels) linked to the predictand variable of interest (e.g. observed precipitation).—

Besides the nature of the predictor/s used, one of the key differences between these approaches is the suitable temporal scale/s of application: daily for BA and PP and monthly/seasonal for RC and MOS methods (BA can be also directly applied to monthly/seasonal data; being thus the most versatile alternative). Note that MOS and PP are the most complex ones since they involve the selection of suitable large-scale predictors, which is typically a hard, time-consuming task that may require the guidance of an expert.

In terms of performance, all these approaches effectively adjust the large biases exhibited by the raw model predictions, which is of paramount importance for users, particularly when climate information is needed to run impact models for different sectors (e.g. hydrology, agriculture, health, etc.) or for the computation of indices that depend on absolute values/thresholds. However, there is no single approach/technique that systematically provides further benefits in terms of bias-insensitive metrics. In case of BA methods, this is due to their incapability to modify the temporal structure of the raw model forecasts (see, e.g., Maraun et al, 2017). However, the application of these methods is straightforward and constitutes a pragmatic and versatile simple choice in cases where a quick post-processing is needed, no expert knowledge on the regional climate is available, the resolution of the model is similar to that of the observational reference considered (BA does not perform downscaling) and/or for regions with no expected potential for downscaling (e.g. flat inland areas). Moreover, although this approach suffers from some limitations (Maraun et al, 2017), its application to seasonal forecasting does not build on strong extrapolation assumptions as in the case of climate change applications.

As compared to BA methods, RC ones can result in modest improvement of some quality aspects (in particular reliability, although other aspects can be degraded). Nevertheless, these improvements are restricted to regions/seasons with high model skill. In addition, since they operate on a monthly/seasonal basis, RC methods can be negatively affected by the limited length of state-of-the-art seasonal hindcasts (which typically have less than 30 years; e.g. the C3S dataset) and, therefore, appropriate cross-validation (typically leave one-year out) is required in order to avoid overfitting and spurious skill. Note however that this is not a worrying factor neither in PP methods nor in BA ones working with daily data.

Differently from BA and RC methods, MOS and PP methods can improve all quality aspects for particular and limited spatial regions for which the skill of the model is weaker for the target variable (e.g. precipitation) than for the informative predictors used in the downscaling process (e.g. humidity and/or winds). Nevertheless, the reverse situation is also possible (see Manzanas et al, 2018, for a case study for PP methods), which warns on the uniformed use of these methods, as they must be carefully analyzed for the particular case-study of interest. Note that, although both MOS and PP methods rely on large-scale predictors, the complexity and requirements for the former are much lower than for the latter. Whereas MOS methods establish the relationship between the large-scale seasonal forecasts and observational reference records using directly the hindcast (with correspondence with observations at a monthly/seasonal scale), PP methods have the extra complexity of building the relationships at a daily basis using reanalysis data (with day-to-day correspondence with observations). This typically requires a comprehensive screening process in order to detect robust predictors similarly represented in both the reanalysis and the model hindcast. Moreover, PP methods may suffer from reanalysis uncertainty, which is particularly relevant in the tropics (see, e.g., Brands et al, 2012; Manzanas et al, 2015), where seasonal forecasts exhibit the highest skill (see, e.g., Manzanas et al, 2014b). This supposes an extra overhead which needs to be appropriately assessed and planned before applying these techniques since, sometimes, the windows of opportunity for improvement are so narrow that the effort may result useless.

Based on all these findings, our overall recommendation would be the use of versatile, easy to implement BA methods for those cases for which the use of MOS and PP methods cannot be carefully tested by experts. Note that BA are suitable for both daily and monthly timescales and provide competitive results in most situations (especially over the tropics). However, we want to remark the fact that the choice of observational dataset can have important effects for the post-processing of seasonal forecasts. Even though MOS and PP methods seem to be more affected by this issue (which can lead to important regional differences in term of interannual skill), also BA methods may be sensitive to observational uncertainty, especially regarding the reproduction of extreme and spell indicators, which are important for many practical applications.

Finally, from a more practical point of view, it is also important to note that there are significant differences in terms of computational cost among distinct approaches (and even among different methods within the same approach) for adjustment/calibration/downscaling, which may be especially relevant for their potential usability in real-time user-tailored applications (e.g. certain climate services).

Acknowledgements This work has been funded by the C3S activity on Evaluation and Quality Control for seasonal forecasts and the EU project AfriCultuReS (H2020-EU.3.5.5, GA 774652). JMG was partially supported by the project MULTI-SDM (CGL2015-66583-R, MINECO/FEDER). FJDR was partially funded by the H2020 EUCP project (GA 776613). The authors also acknowledge the SA-OBS dataset and the data providers in the SACA&D project (`http://saca-bmkg.knmi.nl`).

References

Beck HE, van Dijk AIJM, Levizzani V, Schellekens J, Miralles DG, Martens B, de Roo A (2017) MSWEP: 3-hourly 0.25 global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data. Hydrology and Earth System Sciences 21(1):589–615, DOI 10.5194/hess-21-589-2017, URL `https://www.hydrol-earth-syst-sci.net/21/589/2017/`

van den Besselaar EJM, van der Schrier G, Cornes RC, Iqbal AS, Klein Tank AMG (2017) SA-OBS: A Daily Gridded Surface Temperature and Precipitation Dataset for Southeast Asia. Journal of Climate 30(14):5151–5165, DOI 10.1175/JCLI-D-16-0575.1, URL `https://doi.org/10.1175/JCLI-D-16-0575.1`

Brands S, Gutiérrez JM, Herrera S, Cofiño AS (2012) On the use of reanalysis data for downscaling. Journal of Climate 25(7):2517–2526, DOI 10.1175/JCLI-D-11-00251.1, URL `http://dx.doi.org/10.1175/JCLI-D-11-00251.1`

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, URL `http://www.int-res.com/abstracts/cr/v07/n2/p129-149/`

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

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

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

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

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

Gutiérrez JM, San-Martín D, Brands S, Manzanas R, Herrera S (2013) 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, URL `http://dx.doi.org/10.1175/JCLI-D-11-00687.1`

Helsel DR, Hirsch RM (2002) Statistical Methods in Water Resources, U.S. Geological Survey

Herrera S, Kotlarski S, Soares PMM, Cardoso RM, Jaczewski A, Gutiérrez JM, Maraun D (2018) Uncertainty in gridded precipitation products: Influence of station density, interpolation method and grid resolution. International Journal of Climatology DOI 10.1002/joc.5878, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5878

Huth R (1999) Statistical downscaling in central europe: evaluation of methods and potential predictors. Climate Research 13(2):91–101, DOI 10.3354/cr013091, URL http://www.int-res.com/abstracts/cr/v13/n2/p91-101/

Kotlarski S, Szabó P, Herrera S, Räty 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

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, URL http://dx.doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2

Manzanas R (2016) Statistical downscaling of precipitation in seasonal forecasting: Advantages and limitations of different approaches. PhD thesis, University of Cantabria, iSBN: 978-84-617-4627-9

Manzanas R, Fernández J, Magariño ME, Gutiérrez JM, Doblas-Reyes FJ, Nikulin G, Buontempo C (2014a) Assessing the drift of seasonal forecasts. Poster at the EGU General Assembly

Manzanas R, Frías MD, Cofiño AS, Gutiérrez JM (2014b) 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, 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, URL http://journals.ametsoc.org/DOI/abs/10.1175/JCLI-D-14-00331.1

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

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

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

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

Maraun D, Widmann M, Gutiérrez JM (2018) Statistical downscaling skill under present climate conditions: A synthesis of the VALUE perfect predictor experiment. International Journal of Climatology DOI 10.1002/joc.5877, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5877

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

Nelder JA, Wedderburn RWM (1972) Generalized linear models. JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES A-GENERAL 135(3):370–384, DOI 10.2307/2344614, URL http://www.jstor.org/stable/2344614

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

San-Martín D, Manzanas R, Brands S, Herrera S, Gutiérrez J (2016) Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods. Journal of Climate , submitted

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, URL http://link.springer.com/article/10.1007/s00477-012-0610-0

Themeßl MJ, Gobiet A, Heinrich G (2012) 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

Torralba V, Doblas-Reyes FJ, MacLeod D, Christel I, Davis M (2017) Seasonal climate prediction: A new source of information for the management of wind energy resources. Journal of Applied Meteorology and Climatology 56(5):1231–1247, DOI 10.1175/JAMC-D-16-0204.1, URL https://doi.org/10.1175/JAMC-D-16-0204.1

Vannitsem S, Nicolis C (2008) Dynamical properties of model output statistics forecasts. Monthly Weather Review 136(2):405–419, DOI 10.1175/2007MWR2104.1, URL http://dx.doi.org/10.1175/2007MWR2104.1

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, URL http://www.sciencedirect.com/science/article/pii/S0169809512002554

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, URL http://dx.doi.org/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, URL http://dx.doi.org/10.1175/1520-0442(1995)008<1023:SCORCP>2.0.CO;2

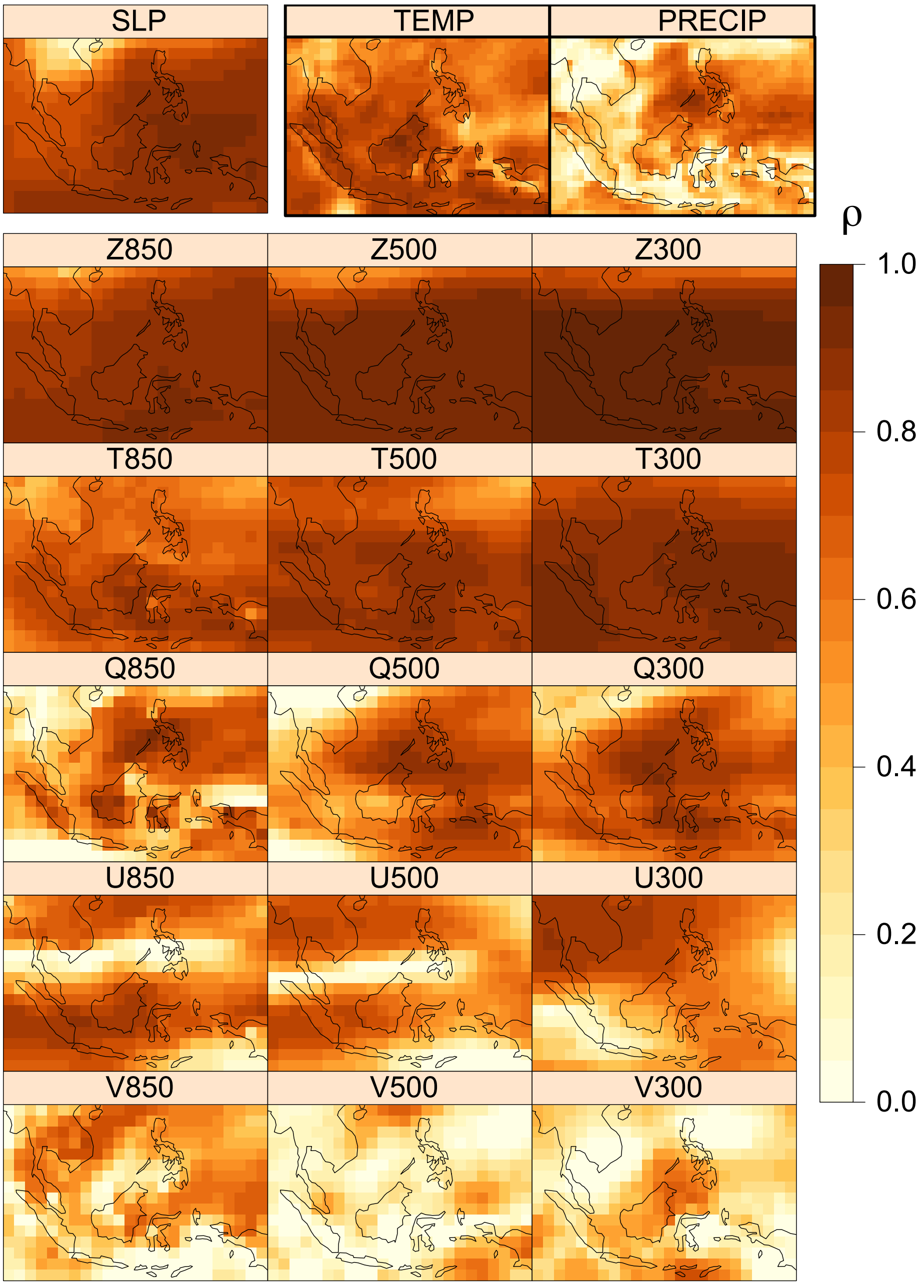

Fig. 2 Interannual correlation between ECMWF-System4 and ERA-Interim for each of the variables (potential predictors) listed in Table 1. For completeness, results are also shown for temperature and precipitation (marked with a black border).

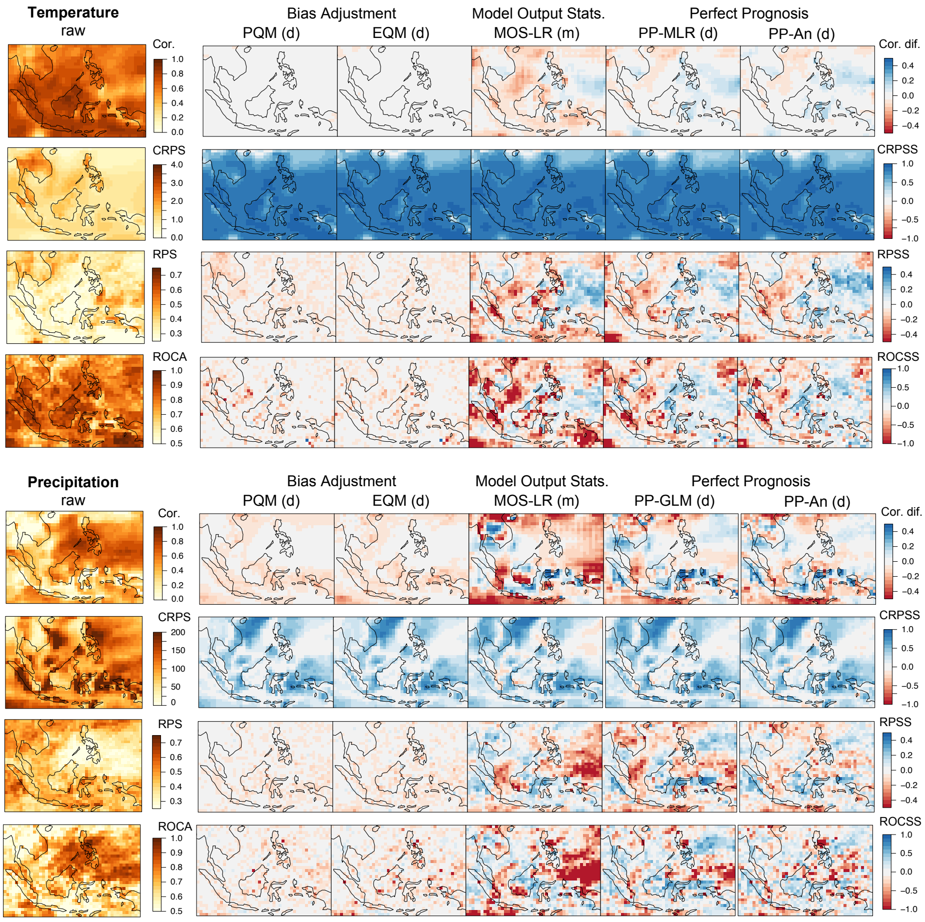

Fig. 3 Validation results obtained for the interannual series of temperature (top) and precipitation (bottom). See the text for details.

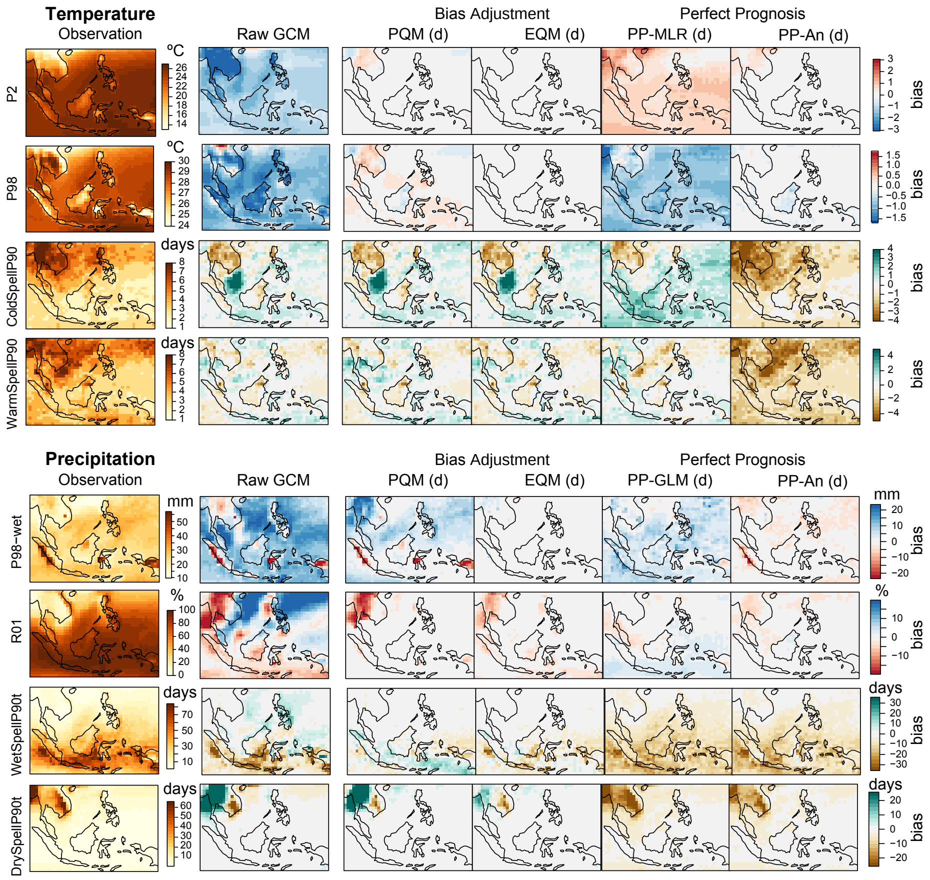

Fig. 4 Validation results for a number of extreme indices obtained for the daily series of temperature (top) and precipitation (bottom). See the text for details.

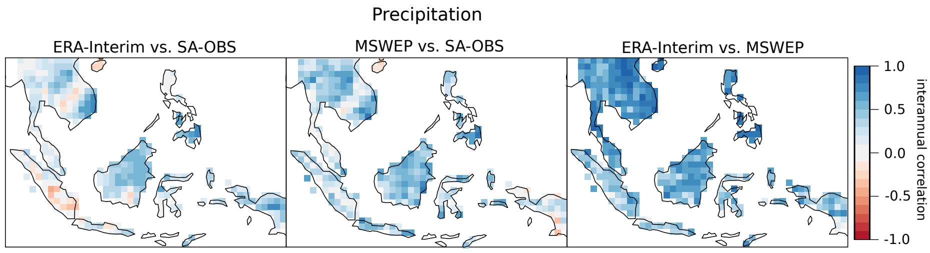

Fig. 5 Comparison of ERA-Interim, SA-OBS and MSWEP precipitation, in terms of correlation for the interannual time-series.

Fig. 6 As bottom panel of Figure 3, but including the results obtained when using SA-OBS/MSWEP for both training and verification of the different methods (middle/bottom row of each metric). For direct comparison, the results shown in Figure 3 for ERA-Interim (top row of each metric) are only displayed over land.

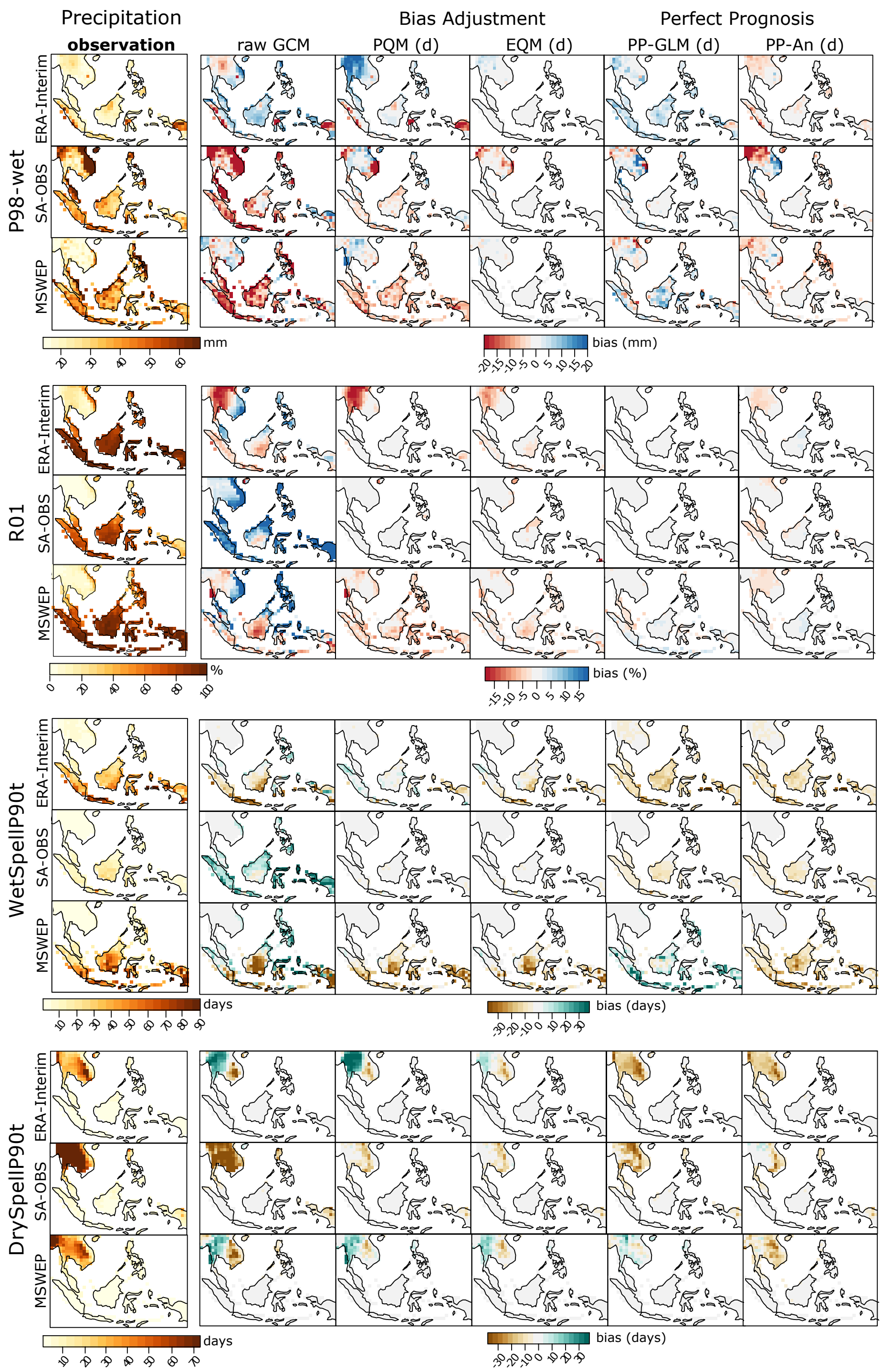

Fig. 7 As bottom panel of Figure 4, but including the results obtained when using SA-OBS/MSWEP for both training and verification of the different methods (middle/bottom row of each metric). For direct comparison, the results shown in Figure 4 for ERA-Interim (top row of each metric) are only displayed over land.