

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

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

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

6
7 Received: date / Accepted: date

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

R. Manzananas (✉)
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

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

35 1 Introduction

36 The state-of-the-art General Circulation Models (GCMs) used for seasonal fore-
37 casting suffer from important systematic biases (mean errors) and drifts (leadtime-
38 dependent biases) and have horizontal resolutions which are typically coarser
39 than those needed for practical applications (see, e.g., Doblas-Reyes et al, 2013;
40 Manzanas et al, 2014a). Therefore, some form of post-processing (i.e. adjust-
41 ment, calibration and/or downscaling) is needed in order to make their raw out-
42 puts usable. In a recent study, Manzanas et al (2019) intercompared the per-
43 formance of Bias Adjustment (BA) —e.g. quantile mapping— and ensemble Re-
44 Calibration (RC) —e.g. non-homogeneous Gaussian regression— methods for the
45 adjustment/calibration of seasonal aggregated forecasts. At this particular time-
46 scale, they found that the RC methods can result in modest improvement of
47 some quality aspects (in particular reliability), although other aspects can be de-
48 graded. Nevertheless, these improvements are restricted to regions/seasons with
49 high model skill. In addition, these methods can be negatively affected by the lim-
50 ited length of state-of-the-art seasonal hindcasts (which typically have less than 30
51 years). They also found that, beyond removing their systematic biases, BA meth-
52 ods can not improve the skill of the raw model forecasts (even more, some quality
53 aspects can be degraded), since they do not modify their temporal structure.
54 However, the application of these methods is straightforward and may constitute
55 a pragmatic and simple alternative when the resolution of the model is similar to
56 that of the observational reference (BA methods are not suitable for downscal-
57 ing), or for regions with no expected potential for downscaling (e.g. flat inland
58 regions). Moreover, beyond the adjustment of monthly/seasonal values, Manzanas
59 et al (2019) pointed out the fact that BA techniques can be also applied to adjust
60 daily data, which are often demanded in a variety of sectoral applications in order
61 to run impact models (crop, hydrology, etc.) or to compute specific climate indices
62 (heat waves, length of growing index, thermal comfort index, fire weather index,
63 etc.).

64 Therefore, we put a special focus in this work on the post-processing of daily
65 (rather than monthly/seasonal) values. For this aim, we consider not only BA
66 methods acting directly on the variable of interest, but also more complex Perfect
67 Prognosis (PP) downscaling techniques (see, e.g., Gutiérrez et al, 2013) which op-
68 erate on a selection of large-scale model circulation variables (predictors) linked to
69 the local observed variable of interest (predictand). Although there has been some
70 indication that PP methods may add some value in terms of skill (e.g. interan-
71 nual correlation) for cases where the dynamical model is better at reproducing the
72 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)

118 2 Data and Methods

119 2.1 Data Used

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

127 We consider one-month lead seasonal forecasts (i.e. predictions initialized in
 128 November) of both temperature and precipitation from the ECMWF-System4
 129 (Molteni et al, 2011), which provides the longest seasonal hindcast to-date —note
 130 that one of the main conclusions of Manzanas et al (2019) is that as long as
 131 possible hindcasts are needed for robust adjustment/calibration.— In particular,
 132 we use here all the 51 members that are available for the November initialization
 133 (only 15 members are available for other initializations) along the period 1982-
 134 2014.

135 Besides the target variables of interest (temperature and precipitation) used
 136 for BA and RC methods, the large-scale variables listed in Table 1 were considered
 137 as potential predictors for MOS and PP methods in this work. For the training
 138 phase of the PP methods, these predictor variables are taken from ERA-interim
 139 reanalysis (Dee et al, 2011). In this case, ERA-Interim and ECMWF-System4 data
 140 are harmonized by performing a simple local scaling to the latter. In particular,
 141 for every large-scale model predictor, monthly mean values were adjusted towards
 142 the corresponding reanalysis values, gridbox by gridbox, avoiding thus problems
 143 that may arise due to the model mean biases.

144 We consider ERA-Interim as the common observational reference along the
 145 study. However, for the assessment of the effect of observational uncertainty un-
 146 dertaken in Section 3.2, we also consider two other datasets for precipitation:
 147 SA-OBS and MSWEP. SA-OBS a high-quality observational dataset which pro-
 148 vides daily gridded (0.25° spatial resolution) temperature and precipitation over
 149 land for Southeast Asia. It has been built based on more than 8000 meteorological
 150 stations and can be freely downloaded from <http://sacad.database.bmkg.go.id>.
 151 MSWEP (version 1) (Beck et al, 2017) is a global terrestrial precipitation dataset
 152 with a high 3-hourly temporal and 0.25° spatial resolution which combines gauge,
 153 satellite and reanalysis information. For the sake of comparability with the results
 154 shown in Manzanas et al (2019), all the different datasets used here (ECMWF-
 155 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.

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

160 2.2 Validation Metrics

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

182 2.3 Methods

183 Among BA methods, we have considered two different implementations of quantile
 184 mapping; one parametric and one empirical. The latter corresponds to the EQM
 185 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 Manzananas 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 Manzananas 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 (Manzananas 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 Manzananas, 2016, for details). For this method, we considered as predictors the standardized anomalies of the predictors considered at the nearest model grid-box (for each predictand location). The popular analog technique (Lorenz, 1969) estimates the local downscaled values corresponding to a particular atmospheric

235 configuration (as represented by a number of model predictors defined over a cer-
236 tain geographical domain) from the local observations corresponding to a set of
237 similar (or analog) atmospheric configurations within a historical catalog formed
238 by a reanalysis. Here, only the closest analog is considered (Zorita et al, 1995;
239 Cubasch et al, 1996). Analogs are defined based on the standardized anomalies
240 of the predictors considered at the 16 nearest model gridboxes (i.e., over a 4x4
241 square centered around each predictand location which allows to encompass the
242 main synoptic phenomena influencing the local climate) and the Euclidean norm
243 is considered. Analog-based methods have been applied in several previous studies
244 to downscale precipitation in the context of seasonal forecasting (see, e.g., Frías
245 et al, 2010; Wu et al, 2012; Shao and Li, 2013; Manzanas et al, 2018). In spite of
246 its simplicity, the analog technique performs as well as other more sophisticated
247 ones (Zorita and von Storch, 1999) and it is one of the most widely used.

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

254 2.4 Selection of predictors for MOS and PP methods

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

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

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

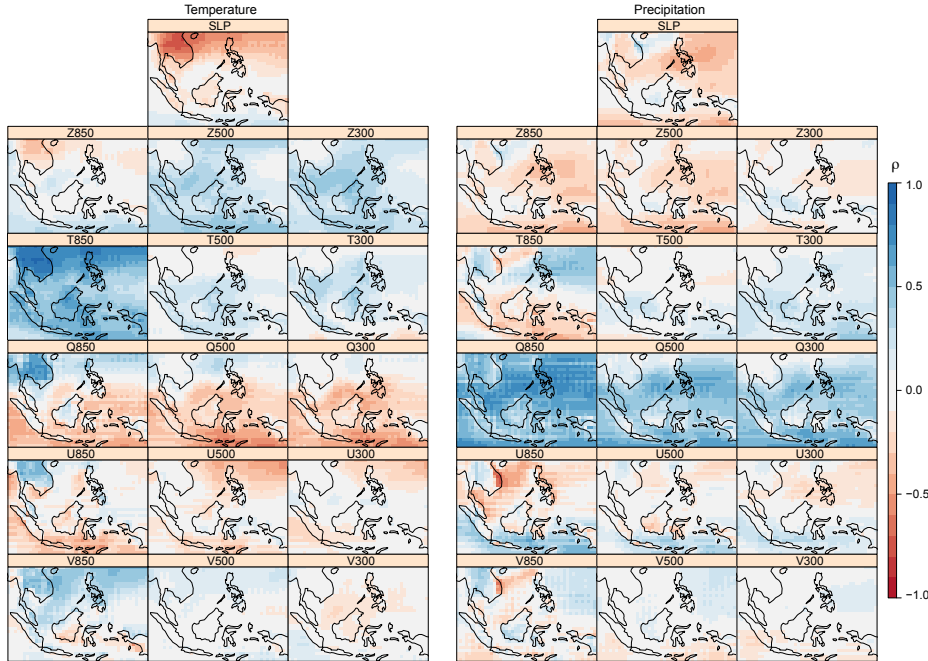


Fig. 1 Correlation between each of the large-scale predictors listed in Table 1 and local temperature (left) and precipitation (right), computed on the daily time-series.

282 3 Results

283 3.1 Intercomparison of approaches and methods

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

296 This figure indicates that all the methods tested here provide a clear benefit in
297 the CRPSS, which is a consequence of effectively removing the important model
298 biases present over the region (see Figure 1 in Manzanas et al (2019)). Note that
299 this result—which was already found for BA and RC methods in Manzanas et al
300 (2019)—is key, since unbiased predictions are needed by many different commu-
301 nities to run their seasonal impact models. However, beyond this improvement
302 in the CRPSS, neither BA nor RC techniques (the latter represented by the LR
303 method) are able to outperform the raw forecasts for any of the remaining met-
304 rics, leading in general to slightly worse results over the entire domain for all of
305 them. This deterioration is even more evident for the LR method, and especially
306 for correlation—note that RC methods can lead to artificial anti-skill (i.e. anti-
307 correlations) in regions of small (or negative) raw model correlations (Eade et al,
308 2014).—It is worth to mention that the EQM tested here (and also the PQM) lead
309 only to slightly better results than those shown for the same method in Manzanas
310 et al (2019), where it was applied on the seasonal (instead of daily) time-series.
311 Moreover, to assess the dependency of the results provided by BA methods on
312 the temporal resolution considered, both EQM and PQM were also applied on the
313 monthly time-series, finding only slightly worse (better) results than in the daily
314 (seasonal) case. Therefore, we do not recommend the application of BA meth-
315 ods on daily data in case only monthly/seasonal data is needed (note that the
316 slight improvement found for higher temporal resolutions does not compensate
317 the increasing computational costs).

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

330 Figure 4 shows the results obtained for the extreme and spell indicators.
331 Whereas column 1 corresponds to the observations, column 2 corresponds to the
332 raw model outputs and columns 3-7 to the different the methods considered. In
333 columns 2-7, the results are expressed as differences (e.g. bias) with respect to the
334 observed values of column 1. Note that neither the RC nor the MOS version of
335 the LR method are considered for this analysis since it cannot be applied at a
336 daily scale. For temperature, the cold bias exhibited by the model in the analyzed
337 percentiles is corrected by all methods except the MLR, which exhibits a warm
338 (cold) bias for the 2nd (98th) percentile. This is due to an underestimation of the
339 predicted variance which is typical of these methods, and could be alleviated by
340 introducing some inflation procedure (see, e.g., Huth, 1999). For spells, the two BA
341 methods maintain the same errors exhibited by the model (the more green/brown,
342 the longer/shorter the predicted spell is, as compared to observations), since they
343 are not able to modify its temporal structure. Differently, since PP methods can
344 alter this temporal structure, they are found to modify the spatial patterns ex-

hibited 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.

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

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

435 4 Conclusions and User Recommendations

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

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

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

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

482 Differently from BA and RC methods, MOS and PP methods can improve all
483 quality aspects for particular and limited spatial regions for which the skill of the
484 model is weaker for the target variable (e.g. precipitation) than for the informative
485 predictors used in the downscaling process (e.g. humidity and/or winds). Never-
486 theless, the reverse situation is also possible (see Manzanas et al, 2018, for a case
487 study for PP methods), which warns on the uniformed use of these methods, as
488 they must be carefully analyzed for the particular case-study of interest. Note that,
489 although both MOS and PP methods rely on large-scale predictors, the complexity
490 and requirements for the former are much lower than for the latter. Whereas MOS
491 methods establish the relationship between the large-scale seasonal forecasts and
492 observational reference records using directly the hindcast (with correspondence
493 with observations at a monthly/seasonal scale), PP methods have the extra com-
494 plexity of building the relationships at a daily basis using reanalysis data (with
495 day-to-day correspondence with observations). This typically requires a compre-
496 hensive screening process in order to detect robust predictors similarly represented
497 in both the reanalysis and the model hindcast. Moreover, PP methods may suf-
498 fer from reanalysis uncertainty, which is particularly relevant in the tropics (see,
499 e.g., Brands et al, 2012; Manzanas et al, 2015), where seasonal forecasts exhibit
500 the highest skill (see, e.g., Manzanas et al, 2014b). This supposes an extra over-
501 head which needs to be appropriately assessed and planned before applying these
502 techniques since, sometimes, the windows of opportunity for improvement are so
503 narrow that the effort may result useless.

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

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

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

527 **References**

- 528 Beck HE, van Dijk AIJM, Levizzani V, Schellekens J, Miralles DG, Martens B,
529 de Roo A (2017) MSWEP: 3-hourly 0.25 global gridded precipitation (1979–
530 2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth*
531 *System Sciences* 21(1):589–615, DOI 10.5194/hess-21-589-2017, URL <https://www.hydrol-earth-syst-sci.net/21/589/2017/>
532
- 533 van den Besselaar EJM, van der Schrier G, Cornes RC, Iqbal AS, Klein Tank
534 AMG (2017) SA-OBS: A Daily Gridded Surface Temperature and Precipitation
535 Dataset for Southeast Asia. *Journal of Climate* 30(14):5151–5165, DOI 10.1175/
536 JCLI-D-16-0575.1, URL <https://doi.org/10.1175/JCLI-D-16-0575.1>
- 537 Brands S, Gutiérrez JM, Herrera S, Cofiño AS (2012) On the use of reanaly-
538 sis data for downscaling. *Journal of Climate* 25(7):2517–2526, DOI 10.1175/
539 JCLI-D-11-00251.1, URL <http://dx.doi.org/10.1175/JCLI-D-11-00251.1>
- 540 Cubasch U, von Storch H, Waszkewitz J, Zorita E (1996) Estimates of cli-
541 mate change in Southern Europe derived from dynamical climate model out-
542 put. *Climate Research* 7(2):129–149, DOI 10.3354/cr007129, URL <http://www.int-res.com/abstracts/cr/v07/n2/p129-149/>
543
- 544 Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U,
545 Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L,
546 Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L,
547 Healy SB, Hersbach H, Holm EV, Isaksen L, Kallberg P, Koehler M, Matricardi
548 M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay
549 P, Tavolato C, Thepaut JN, Vitart F (2011) The ERA-Interim reanalysis: Con-
550 figuration and performance of the data assimilation system. *Quarterly Journal*
551 *of the Royal Meteorological Society* 137(656):553–597, DOI 10.1002/qj.828
- 552 Doblas-Reyes FJ, García-Serrano J, Lienert F, Biescas AP, Rodrigues LRL (2013)
553 Seasonal climate predictability and forecasting: Status and prospects. *Wiley*
554 *Interdisciplinary Reviews: Climate Change* 4(4):245–268, DOI 10.1002/wcc.217,
555 URL <http://dx.doi.org/10.1002/wcc.217>
- 556 Eade R, Smith D, Scaife A, Wallace E, Dunstone N, Hermanson L, Robinson N
557 (2014) Do seasonal-to-decadal climate predictions underestimate the predictabil-
558 ity of the real world? *Geophysical Research Letters* 41(15):5620–5628, DOI
559 10.1002/2014GL061146, URL <http://DOI.wiley.com/10.1002/2014GL061146>
- 560 Enke SA W (1997) Downscaling climate model outputs into local and regional
561 weather elements by classification and regression. *Climate Research* 8(3):195–
562 207
- 563 Frías MD, Herrera S, Cofiño AS, Gutiérrez JM (2010) Assessing the skill of
564 precipitation and temperature seasonal forecasts in Spain: Windows of op-
565 portunity related to ENSO events. *Journal of Climate* 23(2):209–220, DOI
566 10.1175/2009JCLI2824.1
- 567 Gutiérrez JM, San-Martín D, Brands S, Manzanas R, Herrera S (2013) Re-
568 assessing statistical downscaling techniques for their robust application under
569 climate change conditions. *Journal of Climate* 26(1):171–188, DOI 10.1175/
570 JCLI-D-11-00687.1, URL <http://dx.doi.org/10.1175/JCLI-D-11-00687.1>
- 571 Helsel DR, Hirsch RM (2002) *Statistical Methods in Water Resources*, U.S. Geo-
572 logical Survey
- 573 Herrera S, Kotlarski S, Soares PMM, Cardoso RM, Jaczewski A, Gutiérrez JM,
574 Maraun D (2018) Uncertainty in gridded precipitation products: Influence of

- 575 station density, interpolation method and grid resolution. *International Journal of*
576 *of Climatology* DOI 10.1002/joc.5878, URL <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5878>
- 578 Huth R (1999) Statistical downscaling in central europe: evaluation of methods and
579 potential predictors. *Climate Research* 13(2):91–101, DOI 10.3354/cr013091,
580 URL <http://www.int-res.com/abstracts/cr/v13/n2/p91-101/>
- 581 Kotlarski S, Szabó P, Herrera S, Rätty O, Keuler K, Soares PMM, Cardoso
582 RM, Bosshard T, Pagé C, Boberg F, Gutiérrez JM, Isotta FA, Jaczewski A,
583 Kreienkamp F, Liniger MA, Lussana C, Pianko-Kluczyńska K (2017) Obser-
584 vational uncertainty and regional climate model evaluation: a pan-european
585 perspective. *International Journal of Climatology* DOI 10.1002/joc.5249, URL
586 <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5249>
- 587 Lachenbruch PA, Mickey MR (1968) Estimation of error rates in discriminant
588 analysis. *Technometrics* 10(1):1–11, DOI 10.2307/1266219, URL <http://www.jstor.org/stable/1266219>
- 590 Lorenz EN (1969) Atmospheric predictability as revealed by naturally occur-
591 ring analogues. *Journal of the Atmospheric Sciences* 26(4):636–646, DOI 10.
592 1175/1520-0469(1969)26<636:APARBN>2.0.CO;2, URL [http://dx.doi.org/](http://dx.doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2)
593 [10.1175/1520-0469\(1969\)26<636:APARBN>2.0.CO;2](http://dx.doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2)
- 594 Manzanas R (2016) Statistical downscaling of precipitation in seasonal forecasting:
595 Advantages and limitations of different approaches. PhD thesis, University of
596 Cantabria, ISBN: 978-84-617-4627-9
- 597 Manzanas R, Fernández J, Magariño ME, Gutiérrez JM, Doblas-Reyes FJ, Nikulin
598 G, Buontempo C (2014a) Assessing the drift of seasonal forecasts. Poster at the
599 EGU General Assembly
- 600 Manzanas R, Frías MD, Cofiño AS, Gutiérrez JM (2014b) Validation of 40
601 year multimodel seasonal precipitation forecasts: The role of ENSO on the
602 global skill. *Journal of Geophysical Research: Atmospheres* 119(4):1708–1719,
603 DOI 10.1002/2013JD020680, URL [http://onlinelibrary.wiley.com/doi/10.](http://onlinelibrary.wiley.com/doi/10.1002/2013JD020680/abstract)
604 [1002/2013JD020680/abstract](http://onlinelibrary.wiley.com/doi/10.1002/2013JD020680/abstract)
- 605 Manzanas R, Brands S, San-Martín D, Lucero A, Limbo C, Gutiérrez JM (2015)
606 Statistical Downscaling in the Tropics Can Be Sensitive to Reanalysis Choice: A
607 Case Study for Precipitation in the Philippines. *Journal of Climate* 28(10):4171–
608 4184, DOI 10.1175/JCLI-D-14-00331.1, URL [http://journals.ametsoc.org/](http://journals.ametsoc.org/DOI/abs/10.1175/JCLI-D-14-00331.1)
609 [DOI/abs/10.1175/JCLI-D-14-00331.1](http://journals.ametsoc.org/DOI/abs/10.1175/JCLI-D-14-00331.1)
- 610 Manzanas R, Gutiérrez JM, Fernández J, van Meijgaard E, Calmanti S, Mag-
611 ariño ME, Cofiño AS, Herrera S (2017) Dynamical and statistical downscal-
612 ing of seasonal temperature forecasts in Europe: Added value for user appli-
613 cations. *Climate Services* DOI 10.1016/j.cliser.2017.06.004, URL [http://www.](http://www.sciencedirect.com/science/article/pii/S2405880717300067)
614 [sciencedirect.com/science/article/pii/S2405880717300067](http://www.sciencedirect.com/science/article/pii/S2405880717300067)
- 615 Manzanas R, Lucero A, Weisheimer A, Gutiérrez JM (2018) Can bias correction
616 and statistical downscaling methods improve the skill of seasonal precipitation
617 forecasts? *Climate Dynamics* 50(3):1161–1176, DOI 10.1007/s00382-017-3668-z,
618 URL <https://link.springer.com/article/10.1007/s00382-017-3668-z>
- 619 Manzanas R, Gutiérrez JM, Bhend J, Hemri S, Doblas-Reyes FJ, Torralba V,
620 Penabad E, Brookshaw A (2019) Bias adjustment and ensemble recalibra-
621 tion methods for seasonal forecasting: A comprehensive intercomparison us-
622 ing the C3S dataset. *Climate Dynamics* 53(3–4):1287–1305, DOI 10.1007/
623 s00382-019-04640-4

- 624 Maraun D, Shepherd TG, Widmann M, Zappa G, Walton D, M GJ, Hagemann S,
625 Richter I, Soares PMM, Hall A, Mearns LO (2017) Towards process-informed
626 bias correction of climate change simulations. *Nature Climate Change* 7:764–
627 773, DOI 10.1038/nclimate3418
- 628 Maraun D, Widmann M, Gutiérrez JM (2018) Statistical downscaling skill un-
629 der present climate conditions: A synthesis of the VALUE perfect predictor
630 experiment. *International Journal of Climatology* DOI 10.1002/joc.5877, URL
631 <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5877>
- 632 Molteni F, Stockdale T, Balmaseda M, Balsamo G, Buizza R, Ferranti L,
633 Magnusson L, Mogensen K, Palmer T, Vitart F (2011) The new ECMWF
634 seasonal forecast system (System 4). European Centre for Medium-Range
635 Weather Forecasts, URL [http://climate.ncas.ac.uk/people/allan/Fire_](http://climate.ncas.ac.uk/people/allan/Fire_Risk_Insurance_Papers/Moltini%20etal%202011.pdf)
636 [Risk_Insurance_Papers/Moltini%20etal%202011.pdf](http://climate.ncas.ac.uk/people/allan/Fire_Risk_Insurance_Papers/Moltini%20etal%202011.pdf)
- 637 Nelder JA, Wedderburn RWM (1972) Generalized linear models. *JOURNAL OF*
638 *THE ROYAL STATISTICAL SOCIETY SERIES A-GENERAL* 135(3):370–
639 384, DOI 10.2307/2344614, URL <http://www.jstor.org/stable/2344614>
- 640 Pavan V, Marchesi S, Morgillo A, Cacciamani C, Doblas-Reyes FJ (2005) Down-
641 scaling of DEMETER winter seasonal hindcasts over Northern Italy. *Tellus A*
642 57(3):424–434, DOI 10.1111/j.1600-0870.2005.00111.x
- 643 San-Martín D, Manzanas R, Brands S, Herrera S, Gutiérrez J (2016) Reassessing
644 model uncertainty for regional projections of precipitation with an ensemble of
645 statistical downscaling methods. *Journal of Climate* , submitted
- 646 Shao Q, Li M (2013) An improved statistical analogue downscaling procedure
647 for seasonal precipitation forecast. *Stochastic Environmental Research and*
648 *Risk Assessment* 27(4):819–830, DOI 10.1007/s00477-012-0610-0, URL [http:](http://link.springer.com/article/10.1007/s00477-012-0610-0)
649 [//link.springer.com/article/10.1007/s00477-012-0610-0](http://link.springer.com/article/10.1007/s00477-012-0610-0)
- 650 Themeßl MJ, Gobiet A, Heinrich G (2012) Empirical-statistical downscaling and
651 error correction of regional climate models and its impact on the climate change
652 signal. *Climatic Change* 112(2):449–468, DOI 10.1007/s10584-011-0224-4
- 653 Torralba V, Doblas-Reyes FJ, MacLeod D, Christel I, Davis M (2017) Sea-
654 sonal climate prediction: A new source of information for the management
655 of wind energy resources. *Journal of Applied Meteorology and Climatology*
656 56(5):1231–1247, DOI 10.1175/JAMC-D-16-0204.1, URL [https://doi.org/](https://doi.org/10.1175/JAMC-D-16-0204.1)
657 [10.1175/JAMC-D-16-0204.1](https://doi.org/10.1175/JAMC-D-16-0204.1)
- 658 Vannitsem S, Nicolis C (2008) Dynamical properties of model output statis-
659 tics forecasts. *Monthly Weather Review* 136(2):405–419, DOI 10.1175/
660 2007MWR2104.1, URL [http://dx.doi.org/10.1175/](http://dx.doi.org/10.1175/2007MWR2104.1)
661 [2007MWR2104.1](http://dx.doi.org/10.1175/2007MWR2104.1)
- 662 Wu W, Liu Y, Ge M, Rostkier-Edelstein D, Descombes G, Kunin P, Warner T,
663 Swerdlin S, Givati A, Hopson T, Yates D (2012) Statistical downscaling of cli-
664 mate forecast system seasonal predictions for the southeastern mediterranean.
665 *Atmospheric Research* 118:346–356, DOI 10.1016/j.atmosres.2012.07.019, URL
666 <http://www.sciencedirect.com/science/article/pii/S0169809512002554>
- 667 Zorita E, von Storch H (1999) The analog method as a simple statistical down-
668 scaling technique: Comparison with more complicated methods. *Journal of Cli-*
669 *mate* 12(8):2474–2489, DOI 10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.
670 CO;2, URL [http://dx.doi.org/10.1175/1520-0442\(1999\)012<2474:TAMAAS>](http://dx.doi.org/10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2)
671 [2.0.CO;2](http://dx.doi.org/10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2)
- 672 Zorita E, Hughes JP, Lettemaier DP, von Storch H (1995) Stochastic charac-
673 terization of regional circulation patterns for climate model diagnosis and es-

673 timation of local precipitation. *Journal of Climate* 8(5):1023–1042, DOI 10.
674 1175/1520-0442(1995)008<1023:SCORCP>2.0.CO;2, URL <http://dx.doi.org/>
675 [10.1175/1520-0442\(1995\)008<1023:SCORCP>2.0.CO;2](http://dx.doi.org/10.1175/1520-0442(1995)008<1023:SCORCP>2.0.CO;2)

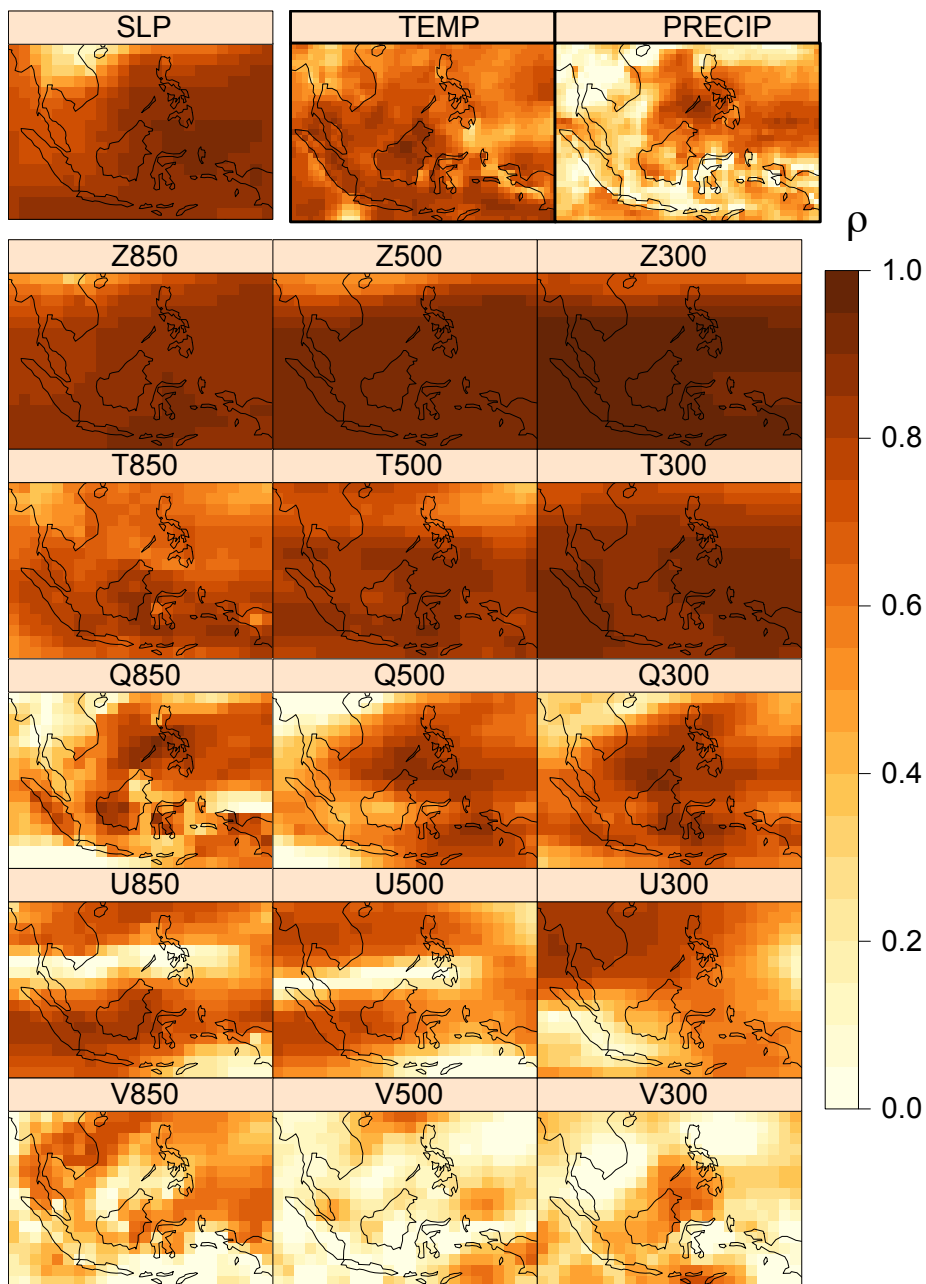


Fig. 2 Interannual correlation between ECMWF-System4 and ERA-Interim for each of the variables (potential predictors) listed in Table 1. For completeness, results are also shown for temperature and precipitation (marked with a black border).

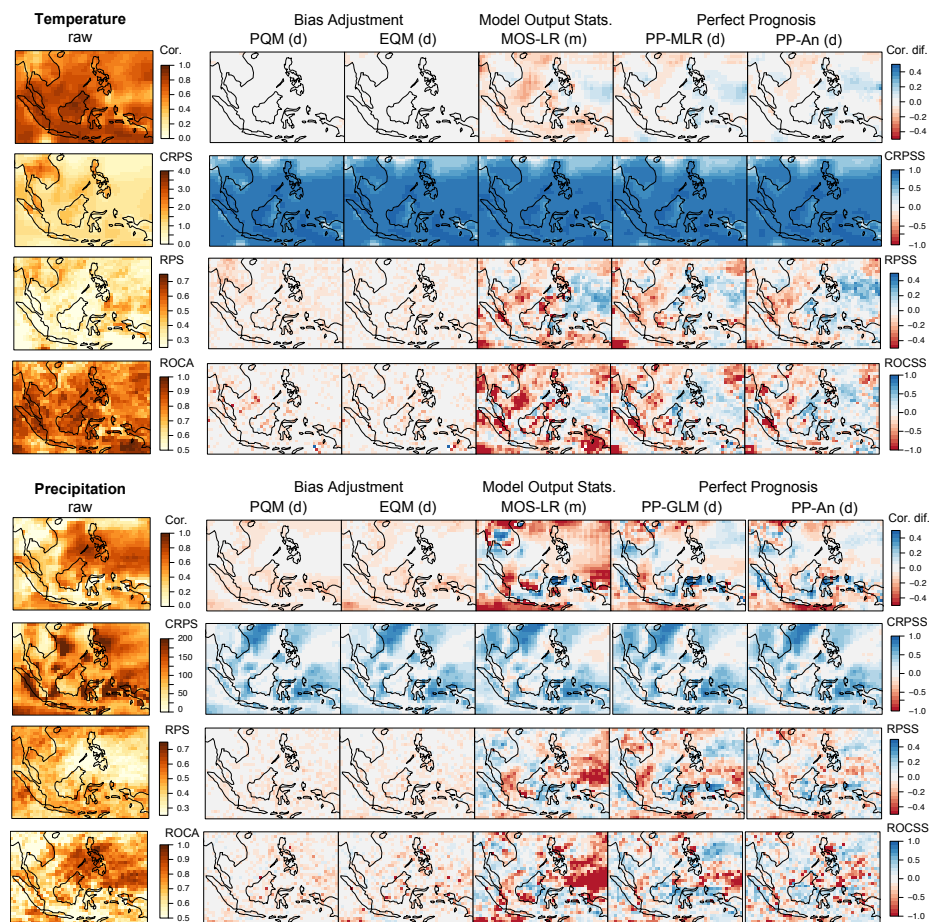


Fig. 3 Validation results obtained for the interannual series of temperature (top) and precipitation (bottom). See the text for details.

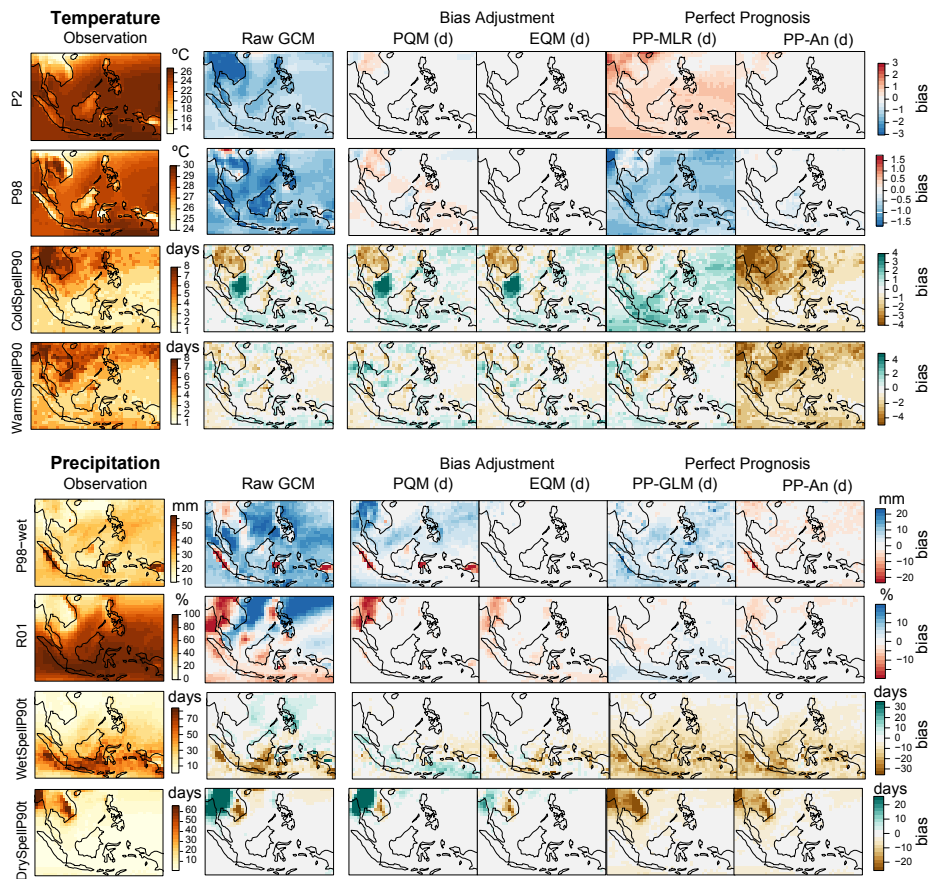


Fig. 4 Validation results for a number of extreme indices obtained for the daily series of temperature (top) and precipitation (bottom). See the text for details.

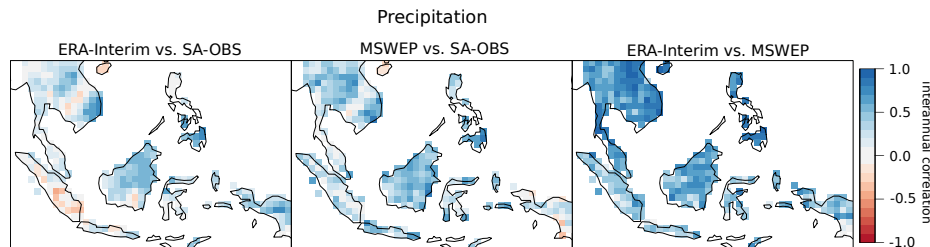


Fig. 5 Comparison of ERA-Interim, SA-OBS and MSWEP precipitation, in terms of correlation for the interannual time-series.

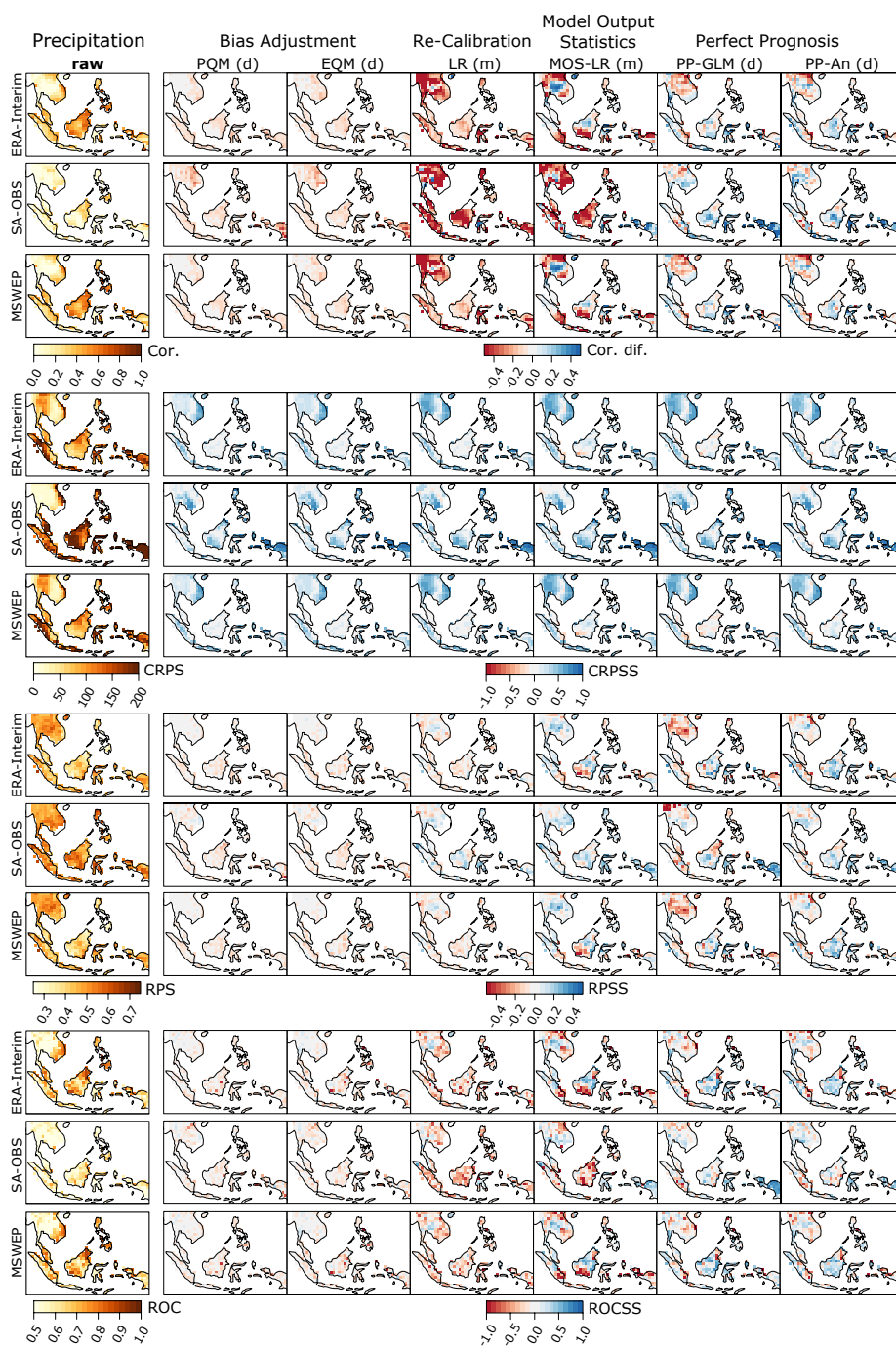


Fig. 6 As bottom panel of Figure 3, but including the results obtained when using SA-OBS/MSWEP for both direct training and verification of the different methods (middle/bottom row of each metric). For direct comparison, the results shown in Figure 3 for ERA-Interim (top row of each metric) are only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

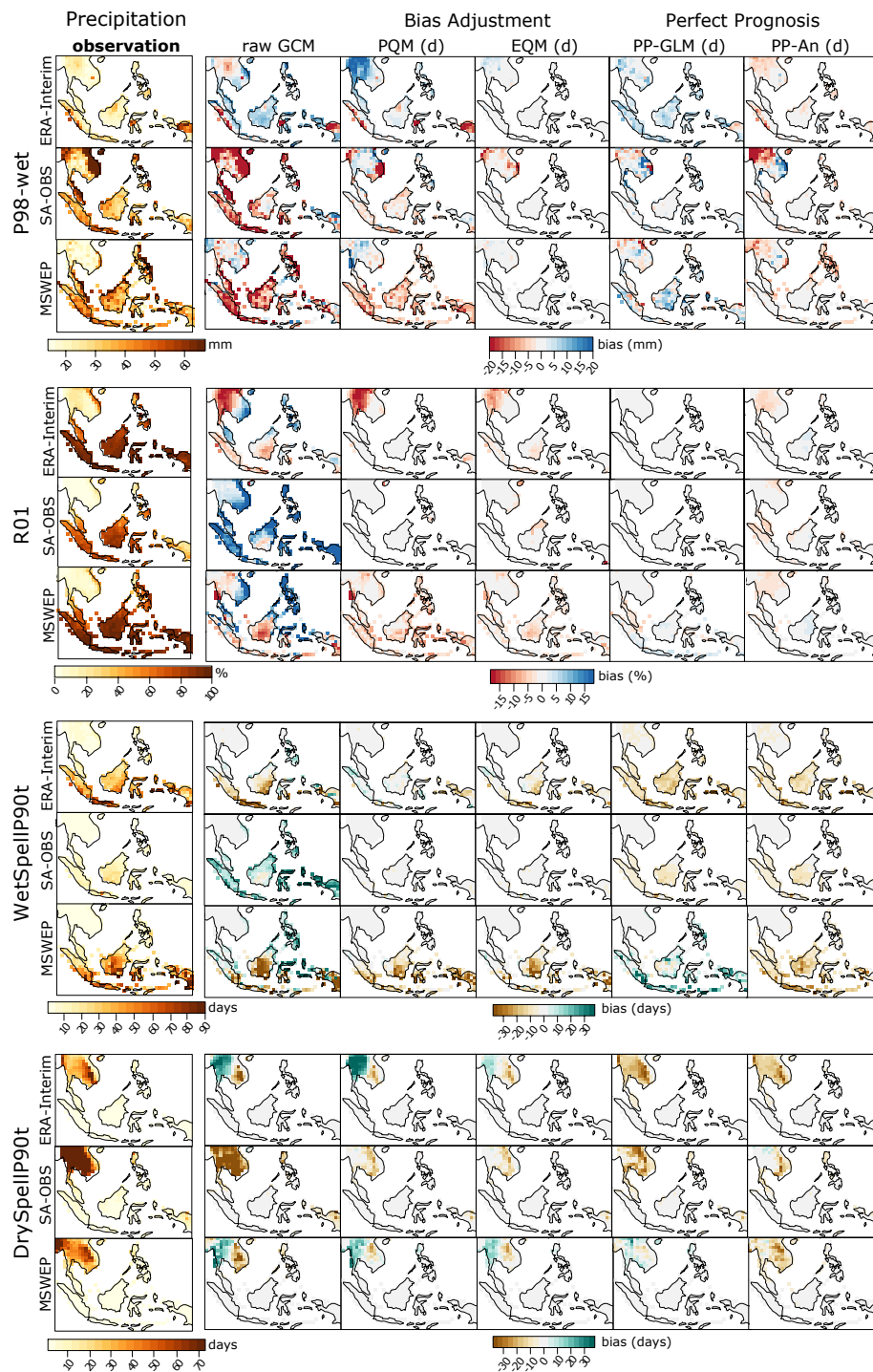


Fig. 7 As bottom panel of Figure 4, but including the results obtained when using SA-OBS/MSWEP for both training and verification of the different methods (middle/bottom row of each metric). For direct comparison, the results shown in Figure 4 for ERA-Interim (top row of each metric) are only displayed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