Climate Dynamics
https://doi.org/10.1007/s00382-020-05589-5

An inter-comparison performance assessment of a Brazilian global sub-seasonal prediction model against four sub-seasonal to seasonal (S2S) prediction project models

Bruno dos Santos Guimarães[1] · Caio Augusto dos Santos Coelho[1] · Steven James Woolnough[2] · Paulo Yoshio Kubota[1] · Carlos Frederico Bastarz[1] · Silvio Nilo Figueroa[1] · José Paulo Bonatti[1] · Dayana Castilho de Souza[1]

Received: 2 July 2020 / Accepted: 15 December 2020
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract
This paper presents an inter-comparison performance assessment of the newly developed Centre for Weather Forecast and Climate Studies (CPTEC) model (the Brazilian Atmospheric Model version 1.2, BAM-1.2) against four sub-seasonal to seasonal (S2S) prediction project models from: Japan Meteorological Agency (JMA), Environmental and Climate Change Canada (ECCC), European Centre for Medium-range Weather Forecasts (ECMWF) and Australian Bureau of Meteorology (BoM). The inter-comparison was performed using hindcasts of weekly precipitation anomalies and the daily evolution of Madden–Julian Oscillation (MJO) for 12 extended austral summers (November–March, 1999/2000–2010/2011), leading to a verification sample of 120 hindcasts. The deterministic assessment of the prediction of precipitation anomalies revealed ECMWF as the model presenting the highest (smallest) correlation (root mean squared error, RMSE) values among all examined models. JMA ranked as the second best performing model, followed by ECCC, CPTEC and BoM. The probabilistic assessment for the event "positive precipitation anomaly" revealed that ECMWF presented better discrimination, reliability and resolution when compared to CPTEC and BoM. However, these three models produced overconfident probabilistic predictions. For MJO predictions, CPTEC crosses the 0.5 bivariate correlation threshold at around 19 days when using the mean of 4 ensemble members, presenting similar performance to BoM, JMA and ECCC. Overall, CPTEC proved to be competitive compared to the S2S models investigated, but with respect to ECMWF there is scope to improve the prediction system, likely by a combination of including coupling to an interactive ocean, improving resolution and model parameterization schemes, and better methods for ensemble generation.

Keywords Sub-seasonal prediction · Forecast verification · Intraseasonal variability · Madden-Julian oscillation

1 Introduction

Sub-seasonal prediction has been recognized in recent years as an activity with potential for producing essential information for various sectors, although performance levels of the currently used general circulation models (GCM) for producing these predictions is still modest. Sub-seasonal predictions fill the gap between numerical weather and seasonal predictions, being of great relevance for disaster prevention and reduction of human and economic losses in various sectors of society (White et al. 2017). Another important aspect making sub-seasonal predictions attractive is the improved ability of GCMs in simulating the Madden–Julian Oscillation (MJO), which is the main source of predictability on the sub-seasonal time scale (Vitart 2014). The MJO is a key intraseasonal variability phenomena responsible for modulating the patterns of circulation, convective activity and precipitation over the tropical region, which is also able to induce tropical-extratropical interaction (Woolnough 2019). Improved ability of GCMs in simulating other sources of sub-seasonal predictability has also been reported

✉ Bruno dos Santos Guimarães
guimara.bruno@gmail.com

[1] Center for Weather Forecast and Climate Studies, National Institute for Space Research, Rodovia presidente Dutra, km 39, Cachoeira Paulista, SP 12630-000, Brazil

[2] Department of Meteorology, National Centre for Atmospheric Science, University of Reading, Reading, Berkshire, UK

Published online: 03 January 2021

(Vitart 2017; de Andrade et al. 2019), including phenomena such as the North Atlantic Oscillation (NAO) (Hurrell et al. 2003), midlatitude tropospheric blocking (Tyrlis and Hoskins 2008), stratospheric warming events (Kidston et al. 2015) and tropical-extratropical teleconnections (Matthews et al. 2004), to name a few. These improvements contribute to increasing the quality of sub-seasonal predictions. For all these reasons, there has been an increasing interest in both the scientific and operational communities in further developing and evaluating sub-seasonal prediction systems (Newman et al. 2003; Vitart 2004, 2008; Hudson et al. 2011a; Mastrangelo et al. 2012; Liu et al. 2017; Weber and Mass 2017; Liang and Lin 2018; Coelho et al. 2018; de Andrade et al. 2019; Guimarães et al. 2020; Klingaman et al. 2020).

This study is primarily an assessment of a model developed for use in South America, which is considered a prominent continent for sub-seasonal predictions, particularly because of being affected by the MJO (Grimm 2019) and tropical-extratropical (Gonzalez and Vera 2014) teleconnections. The ability of GCMs to represent the impacts of these phenomena over South America is likely the reason why these models tend to have better prediction quality over South America when compared to other continental regions, with low to moderate prediction performance at weeks 3–4 (Li and Robertson 2015; de Andrade et al. 2019; Pegion et al. 2019). These findings motivated the Centre for Weather Forecast and Climate Studies [Centro de Previsão de Tempo e Estudos Climáticos (CPTEC)] of the National Institute for Space Research (INPE) to start configuring and evaluating its global model for a future implementation of sub-seasonal predictions (Guimarães et al. 2020). Currently, CPTEC produces operationally numerical weather, extended-range and seasonal predictions.

Guimarães et al. (2020) described the evaluation of several possible configurations of the Brazilian Atmospheric Model version 1.2 (BAM-1.2) to determine the optimal choice for producing sub-seasonal predictions. Characteristics such as vertical resolution, deep convection and boundary layer parameterizations and initialization of soil moisture were evaluated through sensitivity tests and comparison of historical performance based on a series of retrospective predictions (hindcasts) for all investigated model configurations were performed in order to identify the most suitable configuration. With the most suitable CPTEC model configuration determined in Guimarães et al. (2020), the performance of BAM-1.2 sub-seasonal hindcasts over South America have recently been assessed and compared with other models by Klingaman et al. (2020). The latter study showed that CPTEC (BAM-1.2) and the other three investigated models [the US National Centres for Environmental Prediction (NCEP), the United Kingdom Met Office (UKMO) and the European Centre for Medium-range Weather Forecasts (ECMWF)] have moderate precipitation performance over South America for predictions produced three weeks ahead of time. Klingaman et al. (2020) also showed CPTEC (BAM-1.2) was able to represent a large portion of the MJO and El Niño-Southern Oscillation (ENSO) teleconnection patterns over South America up to five weeks ahead of time, with better representation at short lead times. These results are encouraging because the models that were compared with CPTEC have ocean-atmosphere coupling and some with more refined spatial resolution (features not present in BAM-1.2).

The study aims to perform a global assessment of CPTEC model (BAM-1.2) at sub-seasonal time-scale, focusing on an inter-comparison with four Sub-seasonal to seasonal (S2S) prediction project (Vitart et al. 2017) models: the Japan Meteorological Agency (JMA) model (Japan Meteorological Agency 2019), the Environmental and Climate Change Canada (ECCC) model (Liang and Lin 2018), ECMWF model (Vitart 2014) and the Australian Bureau of Meteorology (BoM) model (Hudson et al. 2011a). This study is therefore complementary to both Guimarães et al. (2020) and Klingaman et al. (2020) and compares global precipitation anomaly and MJO hindcast performance of CPTEC model with these four S2S models. JMA, ECCC, ECMWF and BoM models were chosen for comparison with CPTEC because they are (or have been recently) providing sub-seasonal predictions routinely, with two of these models (JMA and ECCC) being also atmosphere-only models as CPTEC. Such a comparative assessment provides the opportunity to explore common virtues and deficiencies of CPTEC with respect to the four S2S models here investigated. Although the main focus of the performed assessment is global, this study also devotes specific attention to South America.

The article is organized as follows. Section 2 presents the datasets, models and verification metrics used in the study. Section 3 describes and compares the performance of sub-seasonal precipitation hindcasts from both the deterministic and probabilistic perspectives, and of MJO hindcasts, produced with CPTEC model and the four S2S models investigated. Conclusions are presented in Sect. 4.

2 Methods

2.1 Datasets and models

Daily data from the Global Precipitation Climatology Project (GPCP) were used to assess precipitation hindcasts quality (Huffman 2001). GPCP is a product derived from observed rainfall data and precipitation estimates by geostationary and polar-orbiting satellites. The spatial resolution of GPCP is $1° \times 1°$ degrees in latitude and longitude. For assessing the model ability to represent the MJO, we used estimates of Outgoing Longwave Radiation (OLR) from the

National Oceanic and Atmospheric Administration (NOAA) satellites, with a spatial resolution of 2.5° × 2.5° degrees in latitude and longitude (Liebmann and Smith 1996), and zonal wind at 850 and 200 hPa from ERA-Interim (Dee et al. 2011).

CPTEC model (BAM-1.2) is a spectral model with triangular quadratic truncation TQ126 (about 100 km) and 42 vertical levels (Guimarães et al. 2020). CPTEC hindcasts were initialized with ERA-Interim reanalysis (Dee et al. 2011). As the CPTEC model version used in this study is not coupled to an ocean model, the total sea-surface temperature (SST) ERA-Interim field (not the anomaly) of each start date was kept constant during the 35 days of integration (persisted SST). A total of 11 hindcast ensemble members were produced for CPTEC through the use of an Empirical Orthogonal Function (EOF) perturbation methodology (Mendonça and Bonatti 2009). The used burst initialization methodology produces optimally perturbed analyses by applying the EOFs to *n* time series formed by the differences between a model run initialized with a control initial condition and *n* model runs initialized with randomly perturbed initial condition. Additional information about BAM-1.2 and the ensemble generation methodology used in this study is provided in Guimarães et al. (2020). In this paper, CPTEC model is compared with JMA, ECCC, ECMWF and BoM models. All data for these four models were obtained from the S2S project database hosted at ECMWF. A brief description of these models and the rationale for their selection is provided below.

The ECMWF model version used in this study is CY43R1, which was in operations during the period from 2016 to 2017, is a spectral model with Tco639 spatial resolution (about 16 km) up to day 15 and Tco319 (about 32 km) after day 15, and 91 vertical levels, initialized with ERA-Interim reanalysis. The atmosphere is coupled with the Nucleus for European Modeling of the Ocean version 3.4.1 (NEMO3.4.1) with a 0.25 degree horizontal resolution and 75 vertical levels, initialized with the Ocean Re-Analysis System version 5 (ORAS5, Zuo et al. 2019). As for CPTEC, ECMWF also has a total of 11 hindcast ensemble members. This model was chosen to be compared with CPTEC because the ECMWF model has been reported in the literature to have the best performance in comparison with other models at the sub-seasonal time-scale (Li and Robertson 2015; de Andrade et al. 2019). It is important to highlight that more recent (updated) versions of ECMWF model are available. However, the ECMWF CY43R1 version was used in this paper because the start (initialization) dates coincide with CPTEC start dates. BoM has its atmosphere coupled to the Australian Community Ocean Model version 2 (ACOM2), which is based on the Geophysical Fluid Dynamics Laboratory (GFDL) Modular Ocean Model version 2 (MOM2), with the atmospheric component being a spectral model with T47 horizontal resolution (about 250 km) and 17 vertical levels. The BoM atmospheric component was initialized with analysis generated by the Atmosphere and Land Initialization (ALI) scheme (Hudson et al. 2011b). The ocean was initialized with the Predictive Ocean Atmosphere Model for Australia (POAMA) Ensemble Ocean Data Assimilation System (PEODAS) reanalysis (Yin et al. 2011). The BoM sub-seasonal prediction system is a multi-model ensemble system with 3 (similar) model versions (24a, 24b and 24c), each with 11 ensemble members making up a total of 33 members. In this study, the BoM version used was the 24a version with its 11 available ensemble members. BoM was chosen because of being a coupled ocean-atmosphere model run operationally until 2020, having a lower spatial resolution than CPTEC. The JMA model version (GEPS1701) used in this study, which was used in operations during the period from 2017 to 2020, is a spectral model with TL479 (about 40 km) horizontal resolution up to 18 days and TL319 (about 55 km) after 18 days and 100 vertical levels. As CPTEC, JMA is also an atmosphere-only model (i.e., not a coupled ocean-atmosphere model) and it was initialized with Japanese 55-year reanalysis (Kobayashi and Iwasaki 2016). JMA produced hindcasts consist of an ensemble of 5 members. JMA hindcasts used the persistence of the SST anomaly of the day prior to the initialization date throughout the integration period, with the SST anomaly derived from the Japanese Merged Satellite and in-situ Data Global Daily Sea Surface Temperature (MGDSST) analysis. JMA was chosen because in addition to being an operational model it stands out in terms of performance among the S2S models that do not have coupling with the ocean (de Andrade et al. 2019). Unlike the other models described above, ECCC is a finite differences model, with a Yin-Yang grid at 0.35° uniform resolution (about 39 km) and 45 vertical levels. ECCC was initialized with ERA-Interim reanalysis and produced hindcasts consisting of an ensemble of four members. ECCC hindcasts used the persistence of the averaged ERA-Interim SST anomaly of the 30 days prior to the initialization date throughout the integration period. The ECCC version used here (GEPS5) is the one that was in operation between 2018 and 2019, and as CPTEC and JMA models, it is also an atmosphere-only model. For this reason, it was also chosen to be compared with CPTEC. Additional information about the S2S models can be found at https://confluence.ecmwf.int/display/S2S/.

Hindcast samples have the same size for all models here investigated, corresponding to a total of 120 retrospective forecasts produced for each model, comprising 120 start (initialization) dates within the extended austral summer (November to March) from 1999/2000 to 2010/2011 (i.e., 12 extended summers), with two retrospective forecasts produced for each month, one initialized at the beginning and the other around the middle of each month. CPTEC and

ECMWF have the same start (initialization) dates, which vary from one month to the next and are presented in Table 1 in Guimarães et al. (2020). The chosen BoM start dates were the 1st and 16th, and for JMA the 1st and 10th days, regardless of the month of the hindcast. As for CPTEC and ECMWF, ECCC start dates vary according to the month of the hindcast. For ECCC, the start dates for the months of November, February and March were selected on the 1st and 15th days of the month. The hindcasts for December were started on the 6th and 20th days and for January on the 4th and 18th days of the month.

2.2 Precipitation verification metrics

Precipitation hindcasts of the five models were evaluated using determinist and probabilistic metrics. Each metric was calculated for four weekly average lead times counting from the hindcast start dates of each model as follows: days 1–7 (week-1), days 8–14 (week-2), days 15–21 (week-3) and days 22–28 (week-4). Lead time dependent hindcast weekly averages for these days (and for the corresponding matching observations) were computed considering the start dates of each model. For the deterministic assessment, the correlation and root mean square error (RMSE) between predicted and observed precipitation anomalies at each grid point were selected to measure two prediction quality attributes: linear association and accuracy, respectively. These two metrics are widely used for sub-seasonal prediction assessment (Liu et al. 2014; Li and Robertson 2015; Coelho et al. 2018; de Andrade et al. 2019; Guimarães et al. 2020; Klingaman et al. 2020). Following Coelho et al. (2018), the probabilistic assessment was performed using the area under the Relative Operating Characteristic curve (AROC), which measure a prediction quality attribute known as discrimination. The AROC was computed for each grid point and for aggregated grid points over the tropical region (30° N–30° S) and over part of South America (0°–30° S and 55° W–35° W). For this South American region, only grid points over the continent were considered. The reliability diagrams were produced for aggregated grid points over the same two above mentioned regions. The reliability diagram is a powerful visual tool for probabilistic predictions assessment because it provides information about three prediction quality attributes (reliability, resolution and sharpness) of the ensemble prediction system in a single diagram. As in Coelho et al. (2018), the probabilistic assessment was performed for the event "positive precipitation anomaly". That is, the probabilistic metrics were computed for dichotomous predictions (occurrences or non-occurrence of positive anomaly) produced with CPTEC, BoM and ECMWF models. The ECCC and JMA models were not included in this analysis because of the limited number of available hindcast ensemble members for these two models (4 and 5 members, respectively) compared to the other three investigated models, making it challenging the estimation of probabilities for the event of interest. More details about the verification metrics can be found in Coelho et al. (2019).

As highlighted earlier, all five models have a sample size of 120 hindcasts, with the same start dates for CPTEC and ECMWF. For the other three models, the start dates are slightly different, but as for CPTEC and ECMWF, the start dates are located at beginning and middle of each extended austral summer month. This start dates selection procedure was designed in order to have a fair comparison of the five models, leading to samples of the same size and close start dates for all models. Ensemble size (i.e. number of ensemble members) is known to affect model performance (de Andrade et al. 2019). Therefore, in order to perform a fair comparison of the five investigated models, the deterministic assessment was performed by computing the mean of the first four ensemble members for each model. This procedure was performed because this is the number of available ensemble members for ECCC, which is the model with the smallest ensemble size among all investigated models. However, in order to also document the performance of five investigated models at full ensemble size configurations, the deterministic assessment was also performed by computing the ensemble mean of all available members for each used model version (11 members for CPTEC, ECMWF and BoM, five members for JMA and 4 members for ECCC). For the probabilistic assessment, the common ensemble size of 11 members among CPTEC, ECMWF and BoM hindcasts was used to allow the comparison of these three models. Probabilities were computed by dividing the number of ensemble members predicting positive anomalies by the total number of 11 ensemble members.

2.3 MJO index calculation and verification metrics

The MJO was assessed using the Real-Time Multivariate MJO (RMMs) indices (Wheeler and Hendon 2004). These indices, referred to as RMM1 and RMM2, are obtained by projecting meridionally averaged 15° N–15° S OLR and zonal winds at 200 and 850 hPa onto the two leading combined EOFs, which can be used to describe the amplitude and phase of the MJO. The model hindcasts and observational RMMs were computed following the six steps described in Gottschalck et al. (2010). The observational RMMs were computed for the same period of the hindcasts using zonal wind at 200 and 850 hPa from ERA-Interim reanalysis and OLR from NOAA. Both model hindcasts and observational zonal winds and OLR were projected into the same two leading combined EOFs of Wheeler and Hendon (2004) when computing the RMMs.

The metrics used to assess MJO prediction quality were the bivariate correlation (BC) and bivariate mean square

error (BMSE), first introduced in the literature by Lin et al. (2008). As discussed in Coelho et al. (2019), the bivariate correlation is an association measure that examines the strength of agreement (or disagreement) between the observed and predicted MJO phases, but is insensitive to MJO amplitude errors (biases in the magnitude of the predicted MJO signal), and the BMSE is a simultaneous accuracy measure of both phase and amplitude of MJO predictions. The BC and BMSE metrics were chosen because they are widely used in MJO hindcast evaluation studies (Lin et al. 2008; Gottschalck et al. 2010; Rashid et al. 2011; Vitart 2017; Lim et al. 2018). The BMSE can be decomposed in its amplitude-error (BMSEa) and phase-error (BMSEp) components, which are both accuracy measures of phase and amplitude of MJO predictions. In this study, we used equations 9 and 10 of Lim et al. (2018) for computing the BMSEa and BMSEp components of the BMSE, respectively. This decomposition is useful because it allows the assessment of the contribution of these two components to the total error (BMSE). All MJO verification metrics were computed for the mean of 4 ensemble members (i.e., the common ensemble size for the five investigated models) and also for mean of the total number of available ensemble members for each model configuration. MJO hindcasts were assessed over the same extended austral summer period and 120 start (initialization) dates as used for the precipitation assessment discussed earlier. However, the MJO assessment was performed for daily predictions with lead times from 1 to 35 days ahead.

3 Hindcasts quality assessment

In this section, we assess and compare CPTEC model hindcast performance with four other S2S models (as described in Sect. 2.1). The analyses comprise deterministic and probabilistic assessments for weekly precipitation anomalies (as described in Sect. 2.2), and a deterministic assessment for the daily MJO evolution (as described in Sect. 2.3).

3.1 Deterministic assessment of weekly precipitation anomalies

Figure 1 shows the spatial distribution of correlation values between the predicted and observed (GPCP) precipitation anomalies for the five investigated model (rows) produced for the first to the fourth weeks (columns). Predicted anomalies used for producing the correlation maps of Fig. 1 are based on the ensemble mean of 4 hindcast members for each model. Applying a two-sided Student's t test, correlation values equal to or larger than 0.2 are statistically significant and different from zero at the 5% level. CPTEC model, as well as the other four models, shows the highest correlation values during the first week in most regions and these values drop rapidly as lead time increases. This is a well documented feature in weekly precipitation predictions at the sub-seasonal time-scale as during the first two weeks the initial conditions used to run the models contribute more importantly to prediction performance than during the last two weeks (Li and Robertson 2015; de Andrade et al. 2019). The five models also poorly simulate small precipitation rates over subtropical high pressure and desert regions, a feature highlighted for predictions produced for the first week as reported in de Andrade et al. (2019). CPTEC shows larger correlation values over the Northern Hemisphere when compared to the Southern Hemisphere due to their better ability to represent winter baroclinic weather systems (Zhu et al. 2014) and tropics-extratropics teleconnections (de Andrade et al. 2019). This feature is also noticed in ECMWF, JMA, ECCC and BoM. All models also show better performance over oceanic than continental regions, likely due to the fact that precipitation anomalies are partially modulated by slowly evolving SST. The spatial correlation patterns are similar among the five models for all four investigated weekly lead times, although the correlation values vary in terms of intensity among models.

Figure 1 shows that for predictions for the first week (first column), the performance of CPTEC, JMA, ECCC are broadly comparable. BoM has much lower correlation values at this lead time when compared to the other models. ECMWF shows larger correlation values than all the other models, illustrating that ECMWF predictions have the highest quality in terms of association between the predicted and observed precipitation anomalies. Figure 1 also shows that the correlation values drop off in extra-tropical regions more rapidly than in tropical regions (especially over the tropical oceans) as lead time increases. All models also show reduced performance in extra-tropical regions when predicting weekly precipitation anomalies beyond the second week. Correlation values above 0.2 are confined to tropical regions, particularly over the Pacific and Atlantic Oceans for predictions produced for the third and fourth weeks. However, it is noticed that all evaluated models show correlation values above 0.2 over specific continental regions for predictions produced for the third and fourth weeks. This feature is noticed over South America. For example, CPTEC, JMA and ECMWF show large correlation values over the eastern South American sector for predictions produced for the third week, and all models show correlation values above 0.2 over northern South America. The correlation values drop off more rapidly in ECCC, CPTEC, JMA than in BoM between week 1 and week 2, and again at weeks 2–3 and 3–4, likely as a result of the coupling in BoM compared to the other three models, such that at weeks 3 and 4 BoM, CPTEC, JMA, and ECCC are comparable. At weeks 3 and

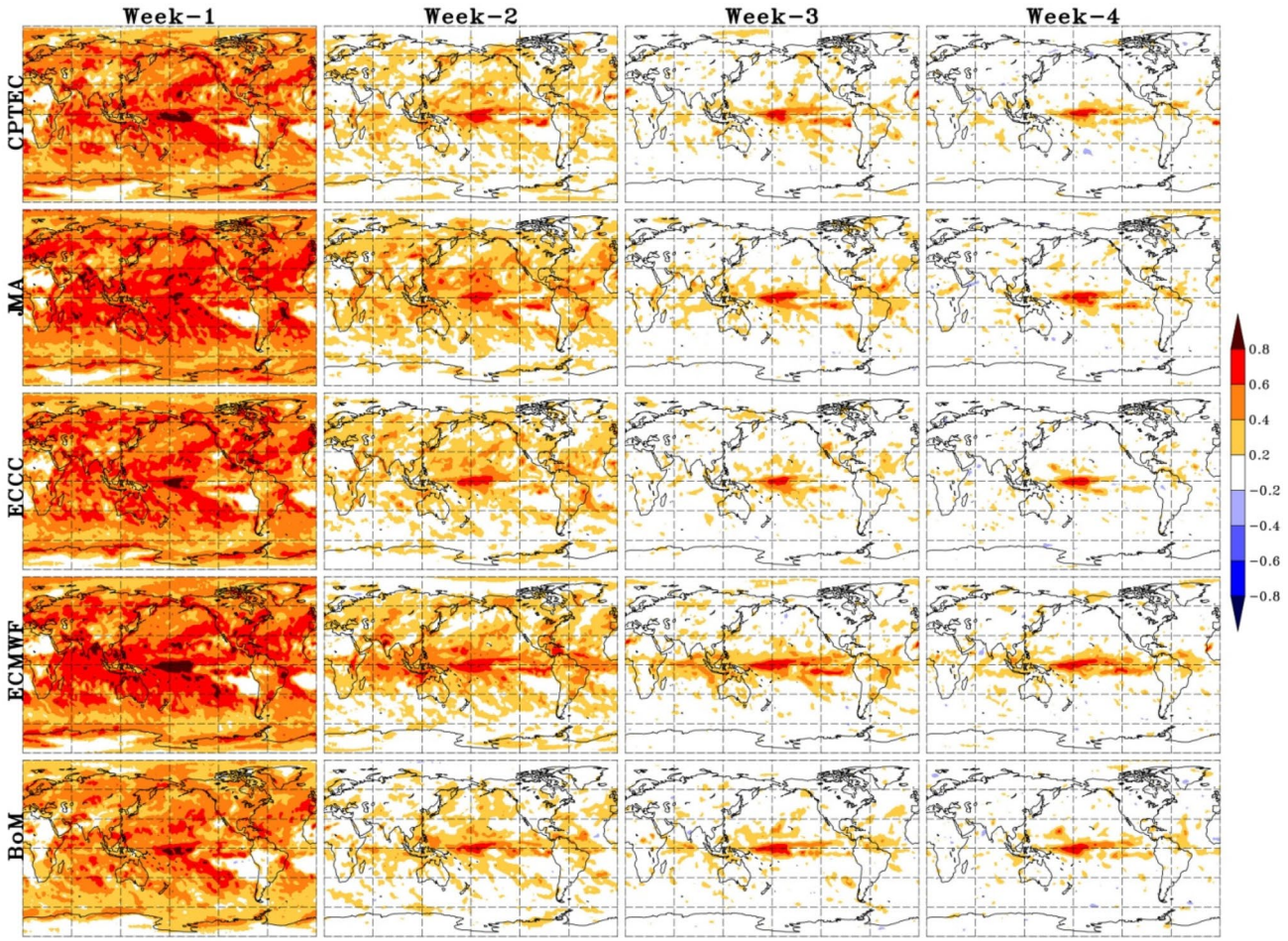

Fig. 1 Correlation between the ensemble mean of 4 members and observed (GPCP) precipitation anomalies for CPTEC, JMA, ECCC, ECMWF and BoM models (rows) for week-1, week-2, week-3 and week-4 (columns). The hindcasts were initialized twice a month within the extended austral summer period (November–March, 1999/2000–2010/2011) leading to a sample size of 120 hindcasts

4, the positive correlation values are largely confined to the tropical Pacific Ocean.

Figure 2 shows similar correlation maps to Fig. 1 but now the ensemble mean precipitation anomalies were computed using the total number of available ensemble members for each of the five investigated model versions. Although the spatial pattern of Figs. 1 and 2 are similar, the correlation values are noticed to be increased for all four lead times when computed with the total (larger) number of available ensemble members, as previously reported in the literature (e.g. de Andrade et al. 2019), except for ECCC that has the same ensemble size (four members) in both figures. This increase is also noticed over South America. Such an increase in correlation values is noticed more pronouncedly for models with larger ensemble size. Increased correlation values are noticed for CPTEC, ECMWF and BoM with 11 ensemble members (first, fourth and fifth rows in Fig. 2) when compared with the assessment performed with four ensemble members for these three models (first, fourth and fifth rows in Fig. 1), particularly over extra-tropical regions for predictions produced for the second week and over tropical regions for predictions produced for the third and fourth weeks. Little change is noticed for JMA because the ensemble size increase is of just one member for this model (from 4 to 5 members) from Figs. 1 and 2.

Figure 3 shows the precipitation anomaly RMSE spatial features for the five models (rows) for the four investigated weekly lead times (columns). RMSE values were computed for the ensemble mean of 4 ensemble members for each model. As for the spatial correlation patterns (Fig. 1), the spatial RMSE patterns are also similar among the investigated models for the four lead times, although RMSE magnitudes vary greatly across the models. As reported in Liu et al. (2014), high RMSE values are noticed over regions where sub-seasonal variability is stronger. These regions are located mostly over the oceans, where convergence zones are usually observed, such as the Intertropical Convergence Zone (ITCZ), the South Pacific Convergence Zone (SPCZ)

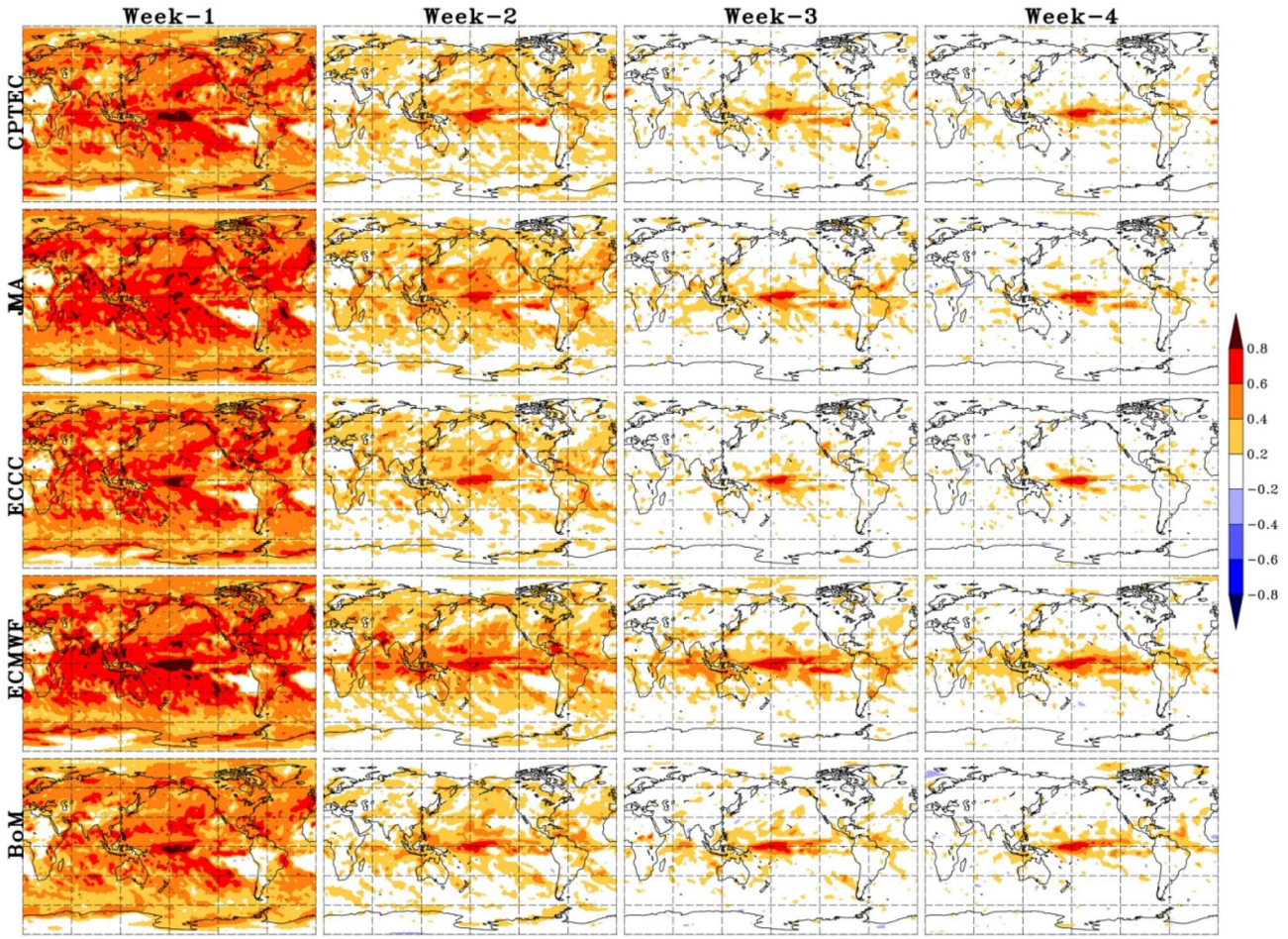

Fig. 2 Same as Fig. 1, but using the total number of available ensemble members for each model when computing the ensemble mean: 11 ensemble members for CPTEC, 5 ensemble members for JMA, 4 ensemble members for ECCC, 11 ensemble members for ECMWF and 11 ensemble members for BoM

and the South Atlantic Convergence Zone (SACZ), and also over the Indian Ocean and Maritime Continent. The RMSE values increase as lead time increases for all models, particularly over the above mentioned regions. Such an increase is the largest when transitioning from predictions produced for the first to the second week (first two columns). ECMWF shows the best performance, with the smallest RMSE values among all investigated models at all four lead times.

CPTEC shows large RMSE values over Madagascar, northern Australia and the SPCZ for predictions produced for the first week. A large increase in RMSE is noticed for predictions produced for the second week when compared to those produced for the first week, particularly over the Indian Ocean, Western Pacific and the SACZ. The RMSE shows little change for CPTEC predictions produced for the third week when compared to predictions produced for the second week. However, an increase in RMSE is noticed over the tropical North Pacific Ocean. The RMSE spatial pattern and magnitudes for predictions produced for the fourth week is similar to the pattern and magnitude for predictions produced for the third week. JMA and ECCC show similar RMSE spatial patterns to CPTEC for all four investigated lead times. However, JMA generally shows smaller RMSE values than CPTEC and ECCC. BoM and ECCC show large RMSE values for predictions produced since the first week lead time, particularly over the Indian Ocean and Western Pacific, and these values are larger than the RMSE values of CPTEC.

To summarize the previous figures and facilitate the ranking of the evaluated models, Fig. 4 shows the global mean correlation between the predicted and observed precipitation anomalies (a) and RMSE (b) averaged between 60°N and 60°S, and similar correlation (c) and RMSE (d) averaged over part of South America (0°−30° S and 55° W–35° W) for the mean of 4 ensemble members for the five investigated models at four weekly lead times. Vertical bars represent 95% confidence intervals, computed using a bootstrap resampling procedure with replacement with 1000 samples.

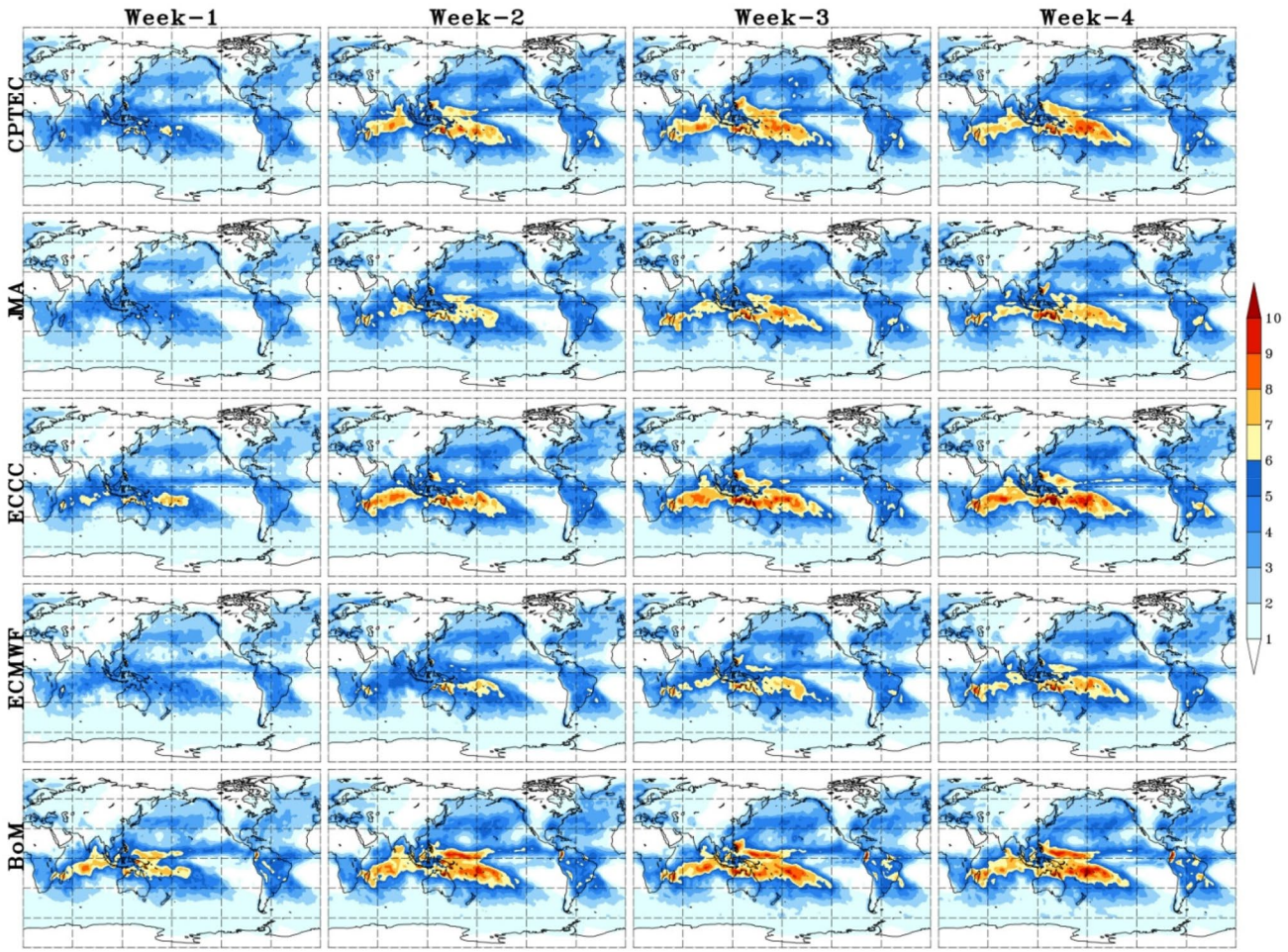

Fig. 3 RMSE of precipitation anomaly (units are mm·day-1) for CPTEC, JMA, ECCC, ECMWF and BoM models (rows), computed for the ensemble means of 4 members for each model, for week-1, week-2, week-3 and week-4 (columns). The hindcasts were initialized twice a month within the extended austral summer period (November–March, 1999/2000–2010/2011) leading to a sample size of 120 hindcasts

Figure 4a shows that the five models exhibit a near-exponential drop in correlation as a function of lead time as reported in Li and Robertson (2015). ECMWF shows the largest correlation values (at all lead times), with JMA appearing as the second model in the rank. Depending on the lead time, the third place in the rank is shared between CPTEC and ECCC. CPTEC correlation values are higher (lower) than ECCC for predictions produced for the last (first) 2 weeks. The 95% confidence intervals for the correlation values for CPTEC and ECCC overlap for predictions produced for the second, third and fourth weeks, illustrating no significant difference in performance between these two models at these longer lead times. BoM shows the lowest correlation values among the five investigated models, with similar performance to ECCC and CPTEC for the last two weekly lead times (with overlapping 95% confidence intervals).

The global mean RMSE between 60° N and 60° S (Fig. 4b) shows a rank similar to the obtained with the correlation assessment (Fig. 4a). ECMWF presents the lowest RMSE values, followed by JMA. CPTEC and ECCC show similar RMSE values with overlapping 95% confidence intervals for predictions produced for the second, third and fourth weekly lead times, while BoM generally shows the highest RMSE values. As previously mentioned, spatial resolution and ocean-atmosphere coupling are important ingredients influencing model performance at the sub-seasonal time scale. This feature is illustrated in Fig. 4. The three models with the highest spatial resolution (ECMWF, JMA and ECCC) show the best performance for the first two weekly lead times, while the models with ocean-atmosphere coupling (ECMWF and BoM) show slower degradation in correlation and RMSE values, particularly BoM, as lead time increases when compared with atmosphere-only models (CPTEC, JMA and ECCC). Figure 4a, b suggest an impact of spatial resolution on prediction performance for shorter lead times. For example, CPTEC shows lower correlation values and larger RMSE values when

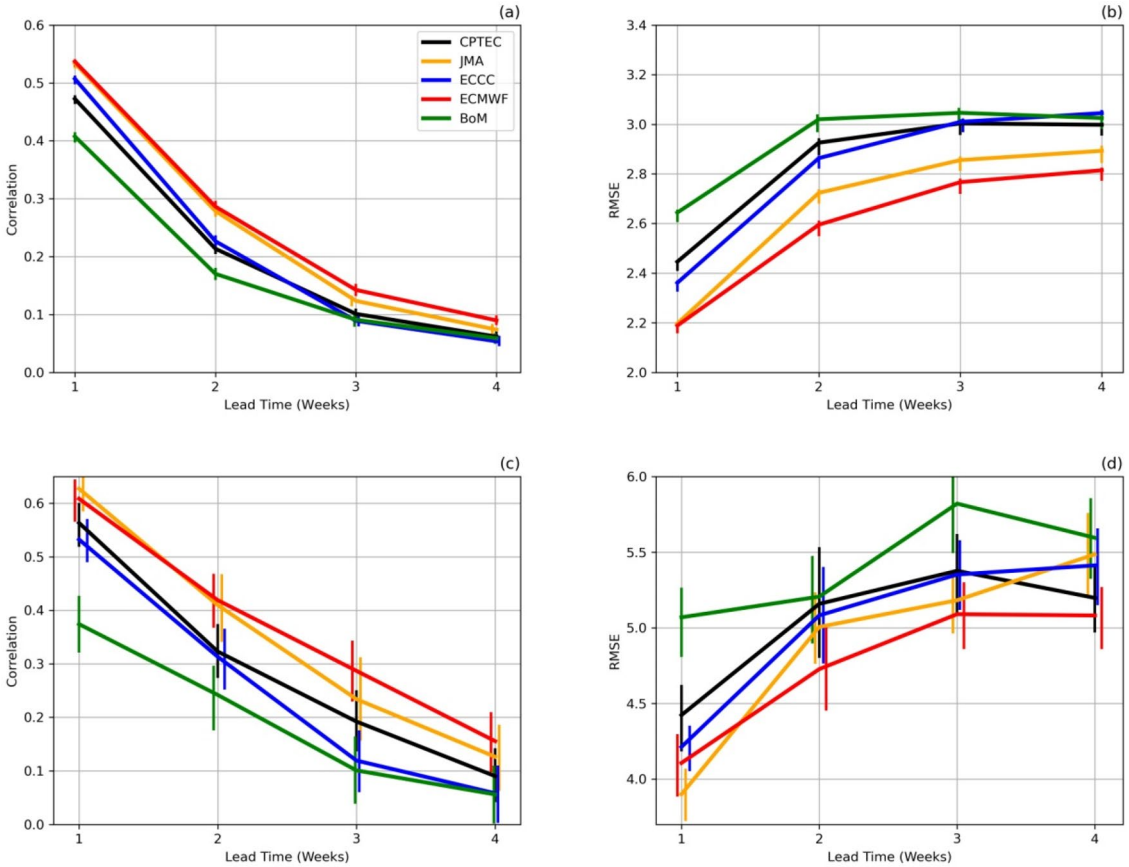

Fig. 4 Global mean correlation (**a**) and RMSE (**b**) between predicted and observed precipitation anomalies averaged over the latitudinal band 60° N–60° S, and similar (**c**) correlation and RMSE (**d**) averaged over part of South America (0°–30° S and 55° W–35° W) for CPTEC, JMA, ECCC, ECMWF and BoM models (computed for the mean of 4 ensemble members of each model) assessed against GPCP, for four lead times (weeks 1 to 4). The hindcasts were initialized twice a month within the extended austral summer period (November–March, 1999/2000–2010/2011) leading to a sample size of 120 hindcasts. The vertical bars plotted around the four lead times represent 95% confidence intervals produced using a bootstrap resampling procedure with replacement with 1000 samples

compared to models with more refined spatial resolution (ECMWF, JMA and ECCC) for predictions produced for the first week. However, when compared to BoM, which presents the lowest spatial resolution among the investigated models, CPTEC shows larger correlation values and lower RMSE values.

Looking specifically at the South American region (0°–35° S and 55° W–35° W), CPTEC and the other investigated models exhibit a near-linear drop in correlation values (Fig. 4c) unlike Fig. 4a. However, the rank is similar to the identified in Fig. 4a, except for the best model for predictions produced for the first week, where JMA shows slightly higher correlation values than ECMWF. The RMSE values averaged over the South American region (Fig. 4d) show a rank similar to Fig. 4c, with values increasing with lead time for most models.

3.2 Probabilistic assessment of weekly precipitation anomalies

The probabilistic assessment is fundamental for a more comprehensive evaluation of sub-seasonal predictions as performed in previous studies (Hudson et al. 2011a; Sun et al. 2018; Coelho et al. 2018). In this section, CPTEC, ECMWF and BoM models are evaluated probabilistically through the spatial distribution of the area under the ROC curve, and ROC curves and reliability diagrams for a collection of predictions aggregated over pre-defined regions. As mentioned in Sect. 2.2, JMA and ECCC models were not evaluated here because of the limited number of available hindcast ensemble members for these two models, making it difficult to estimate prediction probabilities.

Figure 5 shows maps of AROC for the three investigated models for probabilistic predictions for the event positive precipitation anomalies computed with 11 ensemble members for each model, for four weekly lead times. The AROC

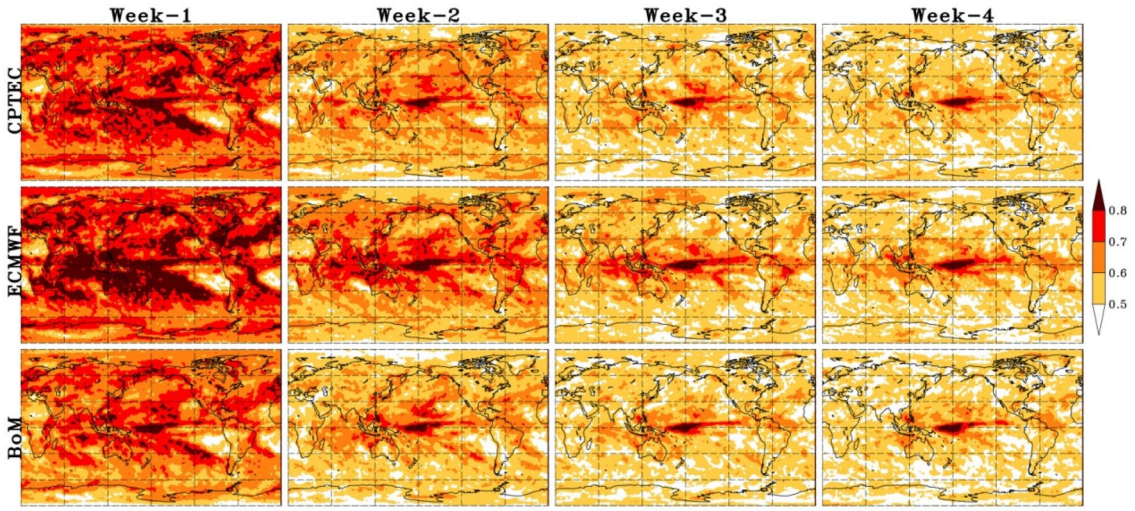

Fig. 5 Maps of area under the ROC curve (AROC) for CPTEC, ECMWF and BoM models (rows), for prediction probabilities for the event positive precipitation anomalies, computed with 11 ensemble members for each model for week-1, week-2, week-3 and week-4 (columns). The hindcasts were initialized twice a month within the extended austral summer period (November–March, 1999/2000–2010/2011) leading to a sample size of 120 hindcasts

measures the ensemble prediction system ability to distinguish the occurrence from the non-occurrence of a certain event (positive anomalies in our case here). Therefore, an area equal to 1 denotes a perfect system able to distinguish the event of interest in all issued probabilistic predictions. An AROC equal to 0.5 indicates that the prediction system has "no discrimination" ability, with probability for successfully discriminating events from non-events equivalent to a random choice (i.e., 50%) of occurrence or no occurrence of the event of interest being predicted. Figure 5 shows that the largest values of the AROC are found over the regions where the largest correlation values were previously identified (Fig. 2). This illustrates that the three investigated model ensemble prediction systems are able to better discriminate events from non-events over the regions where these models also previously demonstrated reasonable performance measured by the association prediction quality attribute. AROC values of the three models drops as lead time increases. CPTEC ranks as the second best model in terms of discrimination ability to distinguish positive from negative precipitation anomalies for all four investigated weekly lead times, with slightly higher AROC values than BoM and slightly lower AROC values than ECMWF.

Figure 6 shows AROC values for probabilistic predictions for the event positive precipitation anomalies, for aggregated grid points over the tropical region (30° N–30° S, panel a) and over South America (0°–35° S and 55° W–35° W, panel b), for the CPTEC (black lines), EMCWF (red lines) and BoM (green lines), for four weekly lead times. For the tropical region, the AROC values are higher for predictions produced for the first week and drop with increase in lead time for the three models, as also noted in Fig. 5. However, all models show values above 0.5 for all four lead times. ECMWF appears as the model with best performance over the tropical region, with AROC values of 0.76 for predictions produced for the first week, 0.7 for predictions produced for the second week, 0.64 for predictions produced for the third week and 0.62 for predictions produced for the fourth week. CPTEC ranks as the second best model, with values of 0.73 for the first week, 0.65 for the second week, 0.61 for the third week and 0.60 for the fourth week, followed by BoM with values of 0.69 for the first week, 0.62 for the second week, 0.59 for the third week and 0.58 for the fourth week. Figure 6b shows that the three models have slightly higher AROC values over the investigated South American region when compared to the tropical region shown in Fig. 6a, particularly for predictions produced for the first two weeks. The model's rank identified through Fig. 6a remains unchanged for this assessment over South America. That is, ECMWF ranks as the best model, followed by CPTEC and BoM.

Figures 7 and 8 show reliability diagrams produced by aggregating all grid points over the tropical region (30° N–30° S) and part of South America (0°–30° S and 55° W–35° W), respectively. These diagrams were produced for CPTEC, ECMWF and BoM by plotting the issued hindcast probabilities against the corresponding observed relative frequencies using ten bins for all four investigated lead times. As in Coelho et al. (2018), the ten used bins were as follows: 0 to 10%, 10 to 20%, 20 to 30%, 30 to 40%, 40 to 50%, 50 to 60%, 60 to 70%, 70 to 80%, 80 to 90% and 90 to 100%. The dashed line plotted in this diagram allows the

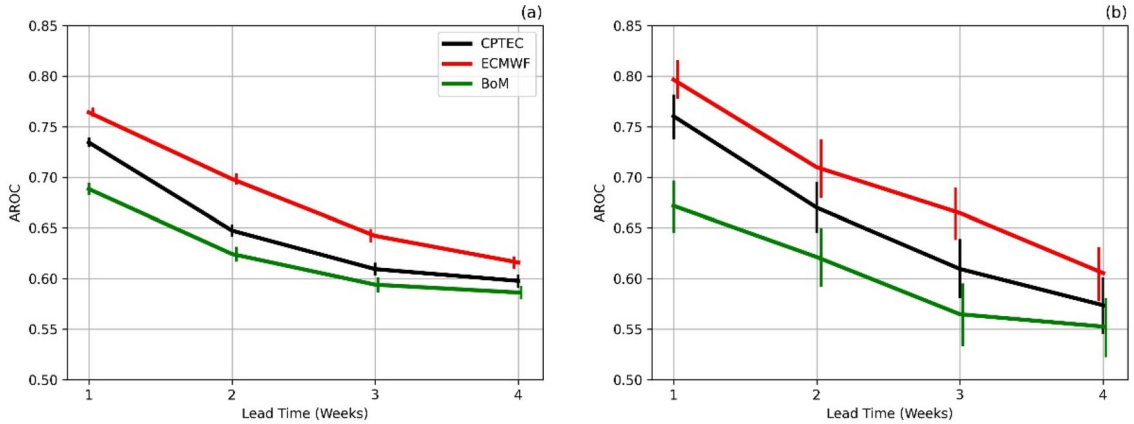

Fig. 6 Area under the ROC curves (AROC) for probabilistic predictions for the event positive precipitation anomalies for CPTEC, ECMWF and BoM models for four lead times (week-1, week-2, week-3 and week-4). The AROC was computed with 11 ensemble members for aggregated grid points over **a** the tropical region (30° N–30° S) and **b** part of South America (0°–30° S and 55° W–35° W). The hindcasts were initialized twice a month within the extended austral summer period (November–March, 1999/2000–2010/2011) leading to a sample size of 120 hindcasts. The vertical bars plotted around the four lead times represent 95% confidence intervals produced using a bootstrap resampling procedure with replacement with 1000 samples

investigation of two prediction quality attributes: reliability and resolution. The histogram plotted at the bottom of the diagram (vertical bars) allows assessment of the prediction sharpness attribute. The diagonal line represents perfectly reliable and well resolved predictions, which have the corresponding observed relative frequencies different from the climatological frequency of occurrence of the event of interest represented in the diagram by the horizontal line.

For both tropical (Fig. 7) and South America (Fig. 8) regions the three models loose quality in terms of reliability, resolution and sharpness as lead time increases. The loss of reliability is noticed as the distance between the dashed lines and the diagonal line increases as prediction lead time increases, moving away from the perfectly reliable diagonal line. The loss of resolution is noticed because the dashed line tends to tilt towards the horizontal line as lead time increases, suggesting that the three models have poor resolution particularly for predictions produced beyond the second week of lead time. Models with good sharpness are expected to present u-shaped prediction probabilities histograms (i.e., blue bars peaking at the two extreme probability bins). This feature is primarily noticed for the two lead times, indicating that probabilistic predictions produced by CPTEC, ECMWF and BoM for the third and fourth weeks lack sharpness. In addition to the characteristics highlighted above, the slopes of dashed lines for all models indicate that they are all overconfident for the four lead times and for both investigated regions, with high (low) probabilities predicted more (less) frequently than observed. Coelho et al. (2018) found similar results for ECMWF precipitation sub-seasonal predictions over South America produced for the austral autumn.

3.3 Deterministic assessment of the daily MJO evolution

The MJO is characterized as a tropical eastward propagating large-scale convective system with associated circulation features, taking around 30–60 days to make a full cycle around the globe (Madden and Julian 1971, 1972), which is able to influence extratropical precipitation through atmospheric teleconnections manifested via Rossby wave trains (Matthews et al. 2004). The MJO is recognized as a key source of predictability on the sub-seasonal timescale (Vitart 2017). Therefore, good quality sub-seasonal precipitation predictions require the realistic representation and prediction of the MJO in the models used for producing these predictions. This section assesses how well the five investigated models predict the daily MJO evolution up to 35 days ahead.

Figure 9 shows the comparative MJO prediction quality assessment for the five investigated models, using the bivariate correlation, BMSE and its amplitude-error (BMSEa) and phase-error (BMSEp) components, as described in Sect. 2.3. Following Lim et al. (2018), the lead time for which MJO predictions are considered of adequate quality can be identified by the time when the bivariate correlation crosses the 0.5 threshold (horizontal black line in Fig. 9a) and when the BMSE crosses the 2.0 threshold (horizontal black line in Fig. 9b). Figure 9a shows that the bivariate correlation values decrease with lead time for all five models. MJO performance varies across models. For the assessment performed, using the mean of four ensemble members for each model (solid lines), ECMWF ranks as the best performing model (as reported in Vitart 2017; Lim et al. 2018), with the

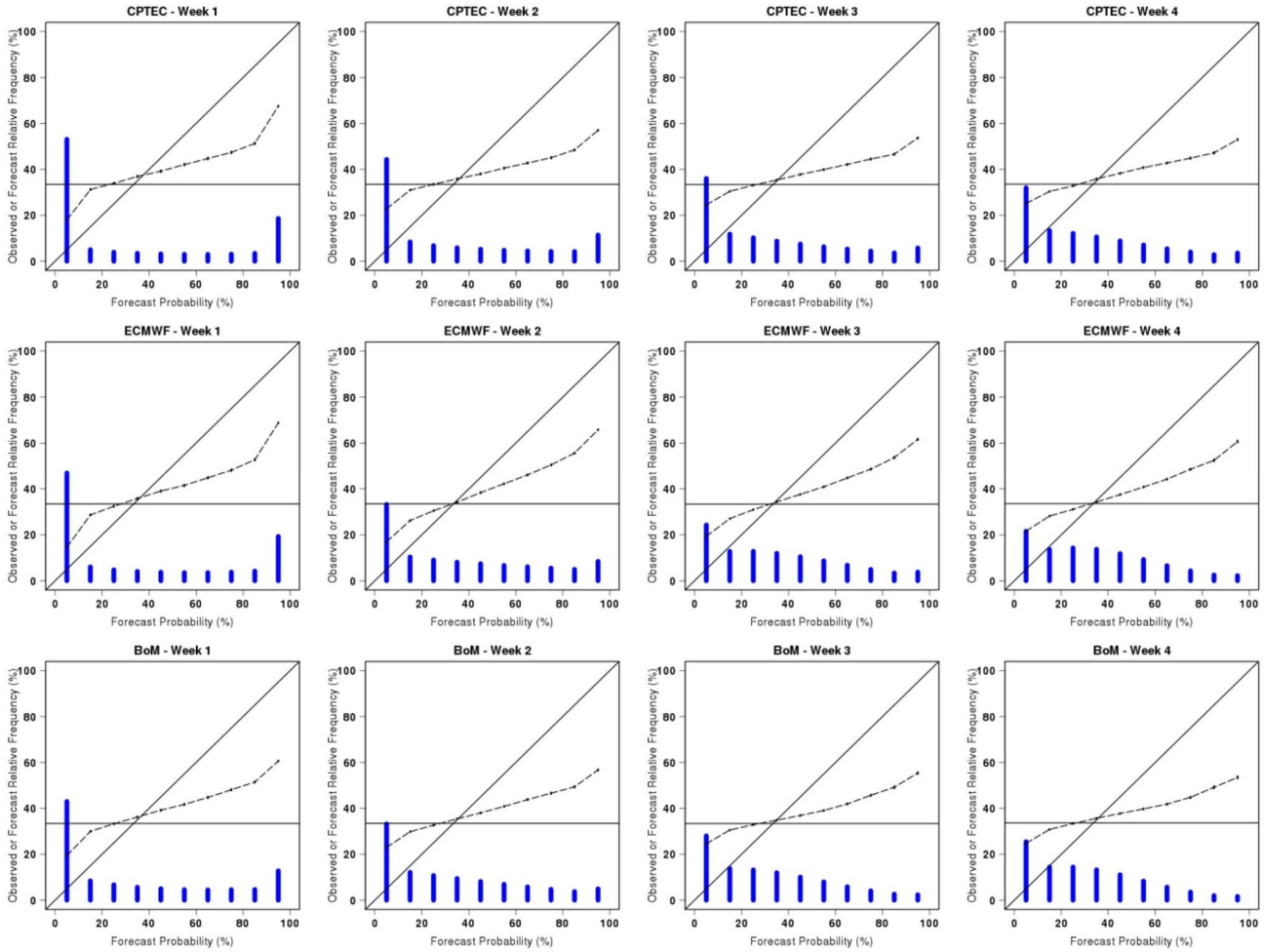

Fig. 7 CPTEC, ECMWF and BoM reliability diagrams (rows) for probabilistic predictions for the event positive precipitation anomalies, for aggregated grid points over the tropical region (30° N–30° S), produced with 11 ensemble members for each model for week-1, week-2, week-3 and week-4 (columns). The vertical bars displayed on top of the reliability curves represent the 95% confidence intervals for the observed relative frequencies produced with a bootstrap resampling procedure (with replacement) with 1000 samples. The vertical blue bars represent the predictions relative frequency histogram. The hindcasts were initialized twice a month within the extended austral summer period (November–March, 1999/2000–2010/2011) leading to a sample size of 120 hindcasts

slowest decrease of bivariate correlation values as lead time increases, reaching the 0.5 threshold around the lead time of 26 days. JMA appears as the second best model, with the 0.5 threshold being reached at the 23 days lead time. CPTEC and BoM share the third position in this performance rank, with the 0.5 threshold being reached around the 19 days lead time. ECCC reaches the 0.5 threshold around 17 days lead time. When using all available members for each model for computing the ensemble mean (dashed lines), one notices an increase in bivariate correlation values, except for JMA, which has only one additional ensemble member (i.e., increasing from 4 to 5 members). CPTEC gains 1 day in prediction performance when using all 11 available hindcast ensemble members, as the 0.5 threshold is now reached around day 20. ECMWF gains 6 days in prediction performance when using all 11 available hindcast ensemble members, as the 0.5 threshold is now reached around day 32. BoM gains 4 days in prediction performance when using all 11 available hindcast ensemble members, as the 0.5 threshold is now reached around 23 days.

Figure 9b shows the BMSE for the five investigated models. The BMSE increases with lead time and the lead time for crossing the 2.0 threshold varies across models. Such variability leads to changes in the model ranking previously identified through Fig. 9a. CPTEC and BoM are the first models to reach the 2.0 threshold at 20 days lead time. ECCC reaches the 2.0 threshold at 22 days lead time. This threshold is achieved at 28 days lead time by JMA, which is the second best model. ECMWF remains the top performing model, with BMSE values remaining below the

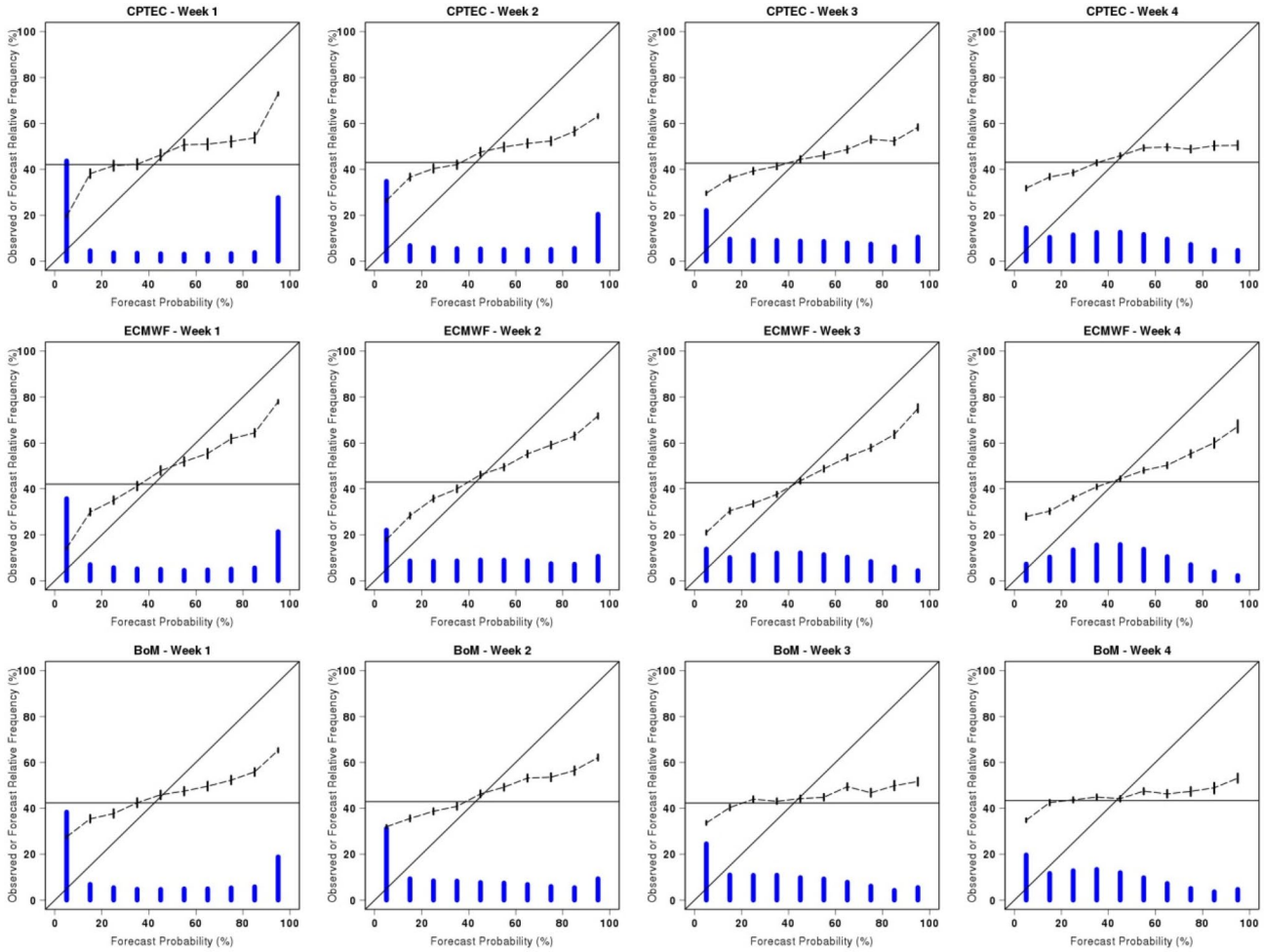

Fig. 8 Same as Fig. 8, but for the region over part of South America (0°–30° S and 55° W–35° W). The grid points over oceanic region were discarded

2.0 threshold at the 35 days lead time here evaluated. As for bivariate correlation, BMSE values decrease when ensemble size is increased from 4 members to the total number of available members for each model (dashed lines), particularly for ECMWF, BoM and CPTEC.

Although the model rank identified through Fig. 9a, b might not necessarily be considered meaningful because of the large confidence intervals on the computed bivariate correlation and BMSE values (e.g., the vertical bars in these two figure panels for CPTEC, ECCC, JMA and BoM overlap for all lead times in the bivariate correlation and BMSE assessment), the results found in this study are similar to findings of Vitart (2017) and Lim et al. (2018). It is interesting to note that four models (ECMWF, JMA, CPTEC and ECCC) have bivariate correlation values very close to 1 and BMSE close to 0 on the first day lead time. These four models show similar performance until approximately the fifth day of lead time, when bivariate correlation and BMSE values of CPTEC and ECCC models start to degrade more strongly than for the other models. Although BoM starts at shorter lead times presenting reduced performance compared to the other models (probably due to low spatial resolution), the degradation of the bivariate correlation and BMSE values is noticed to evolve at a slower rate than for the other atmosphere-only models, reaching ECCC performance level close to 15 days of lead time and CPTEC performance level close to 19 days of lead time. Similar findings were reported by Lim et al. (2018) and Vitart (2017). Increasing ensemble size appears to play an important role in improving MJO prediction performance. BoM, ECMWF and CPTEC performance improved by increasing the number of ensemble members used for computing the ensemble mean from 4 to the total number of available ensemble members of each model.

Figure 9c and d show the BMSEa (amplitude) and BMSEp (phase) components, respectively. These two components show the role of amplitude and phase errors in the growth of BMSE (Fig. 9b). The BMSEa values for CPTEC

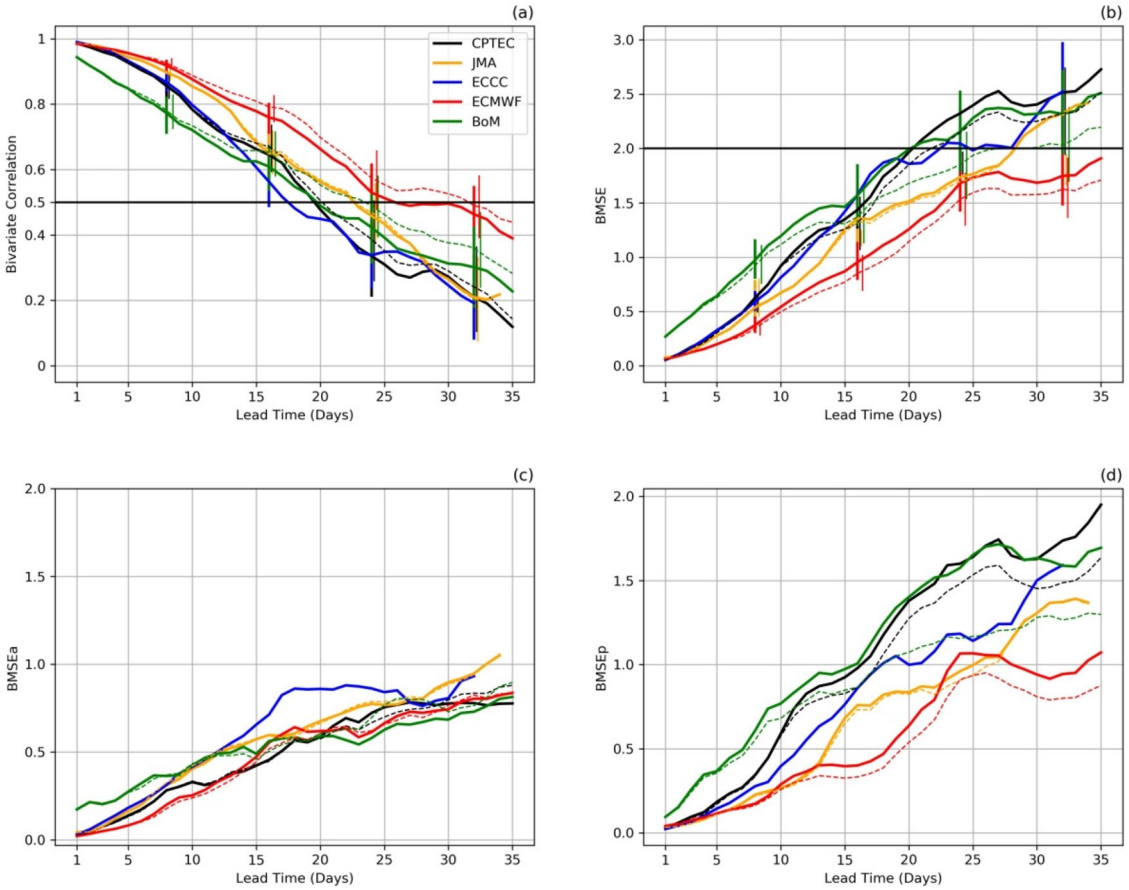

Fig. 9 **a** Bivariate correlation, **b** BMSE, **c** BMSEa and **d** BMSEp for the five investigated models computed with the mean of four ensemble members (solid lines) and with the mean of the total number of available ensemble member (dashed lines) for CPTEC (11 members), JMA (5 members), ECCC (4 members), ECMWF (11 members) and BoM (11 members), as a function of forecast lead time (in days). The hindcasts were initialized twice a month within the extended austral summer period (November–March, 1999/2000–2010/2011) leading to a sample size of 120 hindcasts. The vertical bars in **a** and **b** around lead times 8–32 days plotted every 8 days represent 95% confidence intervals produced using a bootstrap resampling procedure with replacement with 1000 samples for the bivariate correlation and BMSE

are generally close to the values for ECMWF and JMA (the two best ranked models as discussed earlier when describing Fig. 9a, b). ECCC shows the largest BMSEa values for lead times form days 10–26. BoM shows the highest BMSEa at the initial lead times, but after 20 days of lead times becomes the model with the smallest BMSEa. It is curious to note that the increase in the ensemble size shows little impact on BMSEa values, as no major differences are identified when comparing the dashed and continuous lines of each model.

As noted in Lim et al. (2018), in general, BMSEp values (Fig. 9d) are greater than BMSEa values. This indicates that models have more difficulty in characterizing the correct phase (i.e., the location) than the amplitude (i.e., the intensity) of the MJO. ECMWF shows the lowest BMSEp values among all evaluated models. CPTEC and BoM show the highest BMSEp values, and particularly from the day 18 of lead time, these two models show values much higher than of the three other evaluated models. This feature contributes to the identified reduced performance of CPTEC and BoM in terms of BMSE (Fig. 9b) when compared to the other models. Unlike BMSEa, BMSEp is sensitive to ensemble size, as BMSEp values decrease when increasing the number of ensemble members used for computing the ensemble mean. This is clearly noticed in CPTEC, ECMWF and BoM models, which show much lower BMSEp values when using 11 members (dashed black, red and green lines), compared to four members (solid black, red and green lines).

4 Conclusion

This study presented an inter-comparison performance assessment of CPTEC sub-seasonal model against JMA, ECCC, ECMWF and BoM models, which are part of

the S2S prediction project (Vitart et al. 2017). The inter-comparison was performed using retrospective predictions (hindcasts) of weekly precipitation anomalies and daily MJO indices (RMM1 and RMM2) for assessing the models ability in predicting the evolution of the MJO produced during the extended austral summer (November to March). CPTEC model was compared with these four models that are (or have recently been) routinely used for real time predictions in world leading centres, having a range of relevant characteristics to be examined. For example, some models are atmosphere-only as CPTEC model, while other models are coupled ocean-atmosphere. Some have similar spatial horizontal resolution to CPTEC and others have more refined or coarser spatial horizontal resolution compared to CPTEC.

The assessment of weekly precipitation anomaly hindcasts revealed that all five investigated models presented the largest correlation values and the smallest RMSE for predictions produced for the first week after initialization. Correlation and RMSE values degraded with increase of prediction lead time. It is worth noting that the global spatial pattern of correlation values was found to be similar among the investigated models for all four examined weekly lead times, although some variability was noticed across models in terms of the magnitude of these correlation values. The similar spatial pattern of correlation values among the investigated models result from the contribution of phenomena such as MJO, ENSO and tropical-extratropical teleconnections to sub-seasonal prediction performance of all investigated models as reported in previous studies (e.g. de Andrade et al. 2019). For example, CPTEC, ECMWF, JMA and ECCC models show positive correlation values over the southeast South America region characterized by a precipitation seesaw (or dipole) between southeastern Brazil and Uruguay (Nogués-Paegle and Mo 1997) for predictions produced up to the third weekly lead time here investigated. This region is strongly influenced by intraseasonal variability during the austral summer in association with the MJO activity and tropical-extratropical teleconnection (Gonzalez and Vera 2014). The global spatial pattern of RMSE values was also found to be similar among the five assessed models. The largest RMSE values were found over regions where sub-seasonal variability is usually strong (e.g. over the SACZ, SPCZ and ITCZ). For predictions produced for the second, third and fourth weeks the RMSE values over these regions were substantially increased compared to the values for predictions produced for the first week, illustrating models deficiencies in representing this variability for longer leads.

The inter-comparison assessment for weekly precipitation anomalies averaged between 60° N–60° S also revealed CPTEC as a competitive model with performance comparable to the other investigated models for the four examined lead times. The linear association assessment performed by calculating the correlation between predicted and observed weekly precipitation anomalies highlighted ECMWF as the top performing model with correlation values larger than those for CPTEC, JMA, ECCC and BoM models. Similar finding was also reported in previous comparative assessment studies (Li and Robertson 2015; Wheeler et al. 2017; de Andrade et al. 2019). The high ECMWF performance level for sub-seasonal predictions has been attributed primarily to the implementation of an advanced physics model component, with sophisticated parameterization, particularly for deep convective (Wheeler et al. 2017; Vitart 2017). JMA was ranked as the second best performing model despite being an atmosphere-only model. CPTEC and ECCC showed similar performance, with CPTEC (ECCC) ranked as the third best performing model at longer (shorter) lead times. BoM showed the smallest correlation values among the investigated models. This inter-comparison showed the importance of spatial resolution and ocean-atmosphere coupling in the models performance. During the first two weekly lead times, CPTEC and ECCC performed better than BoM, probably due to the higher horizontal spatial resolution of CPTEC and ECCC when compared to BoM. However, the degradation of BoM correlation values was less than the degradation of CPTEC and ECCC as lead time increased, likely due to BoM´s ocean-atmosphere coupling. This feature was also noted when comparing ECMWF and JMA. At first week, these two models showed similar correlation values. However, because the JMA is an atmospheric-only model, the correlation values degraded more quickly in JMA than in ECMWF. The accuracy assessment performed through the calculation of mean RMSE values averaged between 60° N–60° S corroborated the above described findings and ranking of models.

The probabilistic weekly precipitation assessment for the event "positive precipitation anomaly" revealed, through the calculation of the area under the ROC curve, modest discrimination ability for CPTEC, ECMWF and BoM with larger values obtained for the first two weekly lead times when compared to the last two weekly lead times investigated. On the other hand, reliability diagrams produced for the tropical region and for a region over part of South America (0°–30° S and 55° W–35° W) showed that the three investigated models produced overconfident probabilistic predictions, and therefore need to be calibrated. The overconfidence may depend on the ensemble size. For example, if a similar assessment had been performed using all 33 ensemble members from the 3 BOM model versions (each with 11 members), ensemble spread would likely have been increased and the overconfidence feature reduced. It is also worth noting that models tend to have a larger number of ensemble members when run for producing real time predictions compared to when run for producing hindcasts. It is therefore challenging to assess or make inference about

features such as overconfidence of the real time ensemble predictions from the available hindcasts. The performed probabilistic inter-comparison assessment based on the hindcasts revealed that as lead time increases, the predictions loose reliability (the predicted probabilities do not match the observed relative frequencies), resolution (the predicted relative frequencies, regardless of the predicted probabilities, become closer to the observed climatological frequency) and also sharpness (the sharpness histograms become less u-shaped) over these regions. As for the deterministic intercomparison assessment, ECMWF probabilistic predictions tended to perform best among the three investigated models for the four examined lead times, particularly for the assessment over part of South America (0°–30° S and 55° W–35° W). CPTEC and BoM showed similar probabilistic predictions performance.

The MJO predictions assessment performed by computing the bivariate correlation and BMSE, which were computed with the model predicted and observed RMM1 and RMM2 indices, showed that CPTEC crossed the threshold usually considered for predictions having adequate quality at around 19–20 days of lead time when using the mean of four ensemble members, and 20–21 days of lead time when using the mean of 11 ensemble members. These results show that CPTEC MJO performance is similar to BoM, JMA and ECCC. However, such performance level is still considerably distant from ECMWF, which crossed the above mentioned threshold at around 33 days of lead time when using the mean of 11 ensemble members. ECMWF and BoM were found to be the two models presenting larger improvements in MJO predictions performance when increasing the number of ensemble members from 4 to 11 members. The BMSE was decomposed into the BMSEa (amplitude) and BMSEp (phase) components in order to determine of the contribution of these two components to the total MJO error. The phase error (BMSEp) was found to be greater than amplitude error (BMSEa) for all investigated models. This feature was particularly noticed for CPTEC and BoM, which were identified as the models with the largest phase errors.

This study showed that CPTEC sub-seasonal retrospective predictions, expressed as weekly precipitation anomalies and daily MJO indices (RMM1 and RMM2), produced with BAM-1.2 have comparable performance to similar models of world leading centres, particularly ECCC, JMA and BoM. However, such performance level is still distant from ECMWF. This suggests that there is scope to improve the CPTEC prediction system, likely by a combination of including coupling to an interactive ocean, improving resolution and model parameterization schemes, and better methods for ensemble generation.

Acknowledgements We thank Matthew Wheeler and an anonymous reviewer for providing valuable comments and suggestions that contributed for improving the quality of this manuscript. The first author was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), process 88881.189179/2018-01, and University of Reading (ref GS18-179). CASC thanks CNPq, process 305206/2019-2, and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), process 2015/50687-8 (CLIMAX Project) for the support received. SJW was supported by the National Centre for Atmospheric Science ODA national capability programme ACREW (NE/R000034/1), which is supported by NERC and the GCRF. This research was partially supported by the Climate Science for Services Partnership Brazil project (CSSP-Brazil) funded by the Newton Fund. DCS was supported by CNPq (process 167804/2018-9) and CAPES.

References

Coelho CAS, Firpo MA, de Andrade FM (2018) A verification framework for South American sub-seasonal precipitation predictions. Meteorol Z 27(6):503–520

CoelhoCAS, Brown B, Wilson L, Mittermaier M, Casati B (2019) Forecast verificationfor S2S time scales. In: Robertson AW, Vitart F (eds) Sub-seasonal to seasonalprediction: the gap between weather and climate forecasting, Book Chap. 17, 1stedn. Elsevier, Amsterdam, pp 337–361. Accessed 24th Oct 2018 (ISBN: 9780128117149. eBook ISBN: 9780128117156)

de Andrade FM, Coelho CAS, Cavalcanti IF (2019) Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. Clim Dyn 52:5451–5475

Dee DP, Uppala SM, Simmons AJ et al (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorol Soc 137:553–597

Gonzalez PL, Vera CS (2014) Summer precipitation variability over South America on long and short intraseasonal timescales. Clim Dyn 43(7–8):1993–2007

Gottschalck J et al (2010) A framework for assessing operational Madden–Julian oscillation forecasts: a CLIVAR MJO working group project. Bull Am Meteor Soc 91:1247–1258

Guimarães BS et al (2020) Configuration and hindcast quality assessment of a brazilian global sub-seasonal prediction system. Q J R Meteorol Soc 146(728):1067–1084

Grimm AM (2019) Madden–Julian Oscillation impacts on South American summer monsoon season: precipitation anomalies, extreme events, teleconnections, and role in the MJO cycle. Clim Dyn 53(1–2):907–932

Hudson D, Alves O, Hendon HH, Marshall AG (2011a) Bridging the gap between weather and seasonal forecasting: intraseasonal forecasting for Australia. Q J R Meteorol Soc 137:673–689

Hudson D, Alves O, Hendon HH, Wang G (2011b) The impact of atmospheric initialisation on seasonal prediction of tropical Pacific SST. Clim Dyn 36:1155–1171

Huffman GJ, Adler RF, Morrissey MM, Bolvin DT, Curtis S, Joyce R, McGavock B, Susskind J (2001) Global precipitation at one-degree daily resolution from multisatellite observations. J Hydrometeorol 2:36–50

Hurrell JW, Kushnir Y, Ottersen G, Visbeck M (2003) An overview of the North Atlantic oscillation. Geophys Monogr Am Geophys Union 134:1–36

Japan Meteorological Agency (2019) Outline of the operational numerical weather prediction at the Japan Meteorological Agency. Appendix to WMO Technical Progress Report on The Global Data-Processing and Forecasting SYSTEM (GDPFS) and Numerical Weather Prediction (NWP) Research, Japan Meteorological Agency, Tokyo, Japan, 229 pp. https://www.jma.go.jp/jma/

jma-eng/jma-center/nwp/outline2019-nwp/pdf/outline2019_all.pdf. Accessed 10 Jan 2020
Karpechko AY, Charlton-Perez A, Balmaseda M, Tyrrell N, Vitart F (2018) Predicting sudden stratospheric warming 2018 and its climate impacts with a multimodel ensemble. Geophys Res Lett 45:13–538
Kidston J et al (2015) Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. Nat Geosci 8.6:433–440
Klingaman NP, Young M, Chevuturi A, Guimarães BS, Guo L, Woolnough SJ, Coelho CAS, Kubota PY, Holloway CE (2020) Subseasonal prediction performance for austral summer South American rainfall. Weather Forecast. https://doi.org/10.1175/WAF-D-19-0203.1
Kobayashi C, Iwasaki T (2016) Brewer–Dobson circulation diagnosed from JRA-55. J Geophys Res Atmos 121(4):1493–1510
Li S, Robertson AW (2015) Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. Mon Weather Rev 143:2871–2889
Liang P, Lin H (2018) Sub-seasonal prediction over East Asia during boreal summer using the ECCC monthly forecasting system. Clim Dyn 50:1007–1022
Liebmann B, Smith CA (1996) Description of a complete (interpolated) outgoing longwave radiation dataset. Bull Am Meteor Soc 77(6):1275–1277
Lim Y, Son SW, Kim D (2018) MJO prediction skill of the subseasonal-to-seasonal prediction models. J Clim 31(10):4075–4094
Lin H, Brunet G, Derome J (2008) Forecast skill of the Madden–Julian Oscillation in two Canadian atmospheric models. Mon Weather Rev 136:4130–4149
Liu X, Yang S, Li Q, Kumar A, Weaver S, Liu S (2014) Subseasonal forecast skills and biases of global summer monsoons in the NCEP Climate Forecast System version 2. Clim Dyn 42:1487–1508
Liu X, Wu T, Yang S, Li T, Jie W, Zhang L, Wang Z, Liang X, Li Q, Cheng Y, Ren H (2017) MJO prediction using the sub-seasonal to seasonal forecast model of Beijing Climate Center. Clim Dyn 48:3283–3307
Madden RA, Julian PR (1971) Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. J Atmos Sci 28:702–708
Madden RA, Julian PR (1972) Description of global-scale circulation cells in the tropics with a 40–50 day period. J Atmos Sci 29:1109–1123
Mastrangelo D, Malguzzi P, Rendina C, Drofa O, Buzzi A (2012) First outcomes from the CNR-ISAC monthly forecasting system. Adv Sci Res 8:77–82
Matthews AJ, Hoskins BJ, Masutani M (2004) The global response to tropical heating in the Madden–Julian oscillation during the northern winter. Q J R Meteorol Soc 130(601):1991–2011
Mendonça AM, Bonatti J (2009) Experiments with EOF-based perturbation methods and their impact on the CPTEC/INPE ensemble prediction system. Mon Weather Rev 137:1438–1459
Newman M, Sardeshmukh PD, Winkler CR, Whitaker JS (2003) A study of subseasonal predictability. Mon Weather Rev 131(8):1715
Nogués-Paegle J, Mo KC (1997) Alternating wet and dry conditions over South America during summer. Mon Weather Rev 125(2):279–291
Pegion K et al (2019) The subseasonal experiment (SubX): a multimodel subseasonal prediction experiment. Bull Am Meteor Soc 100(10):2043–2060
Rashid HA, Hendon HH, Wheeler MC, Alves O (2011) Prediction of the Madden–Julian Oscillation with the POAMA dynamical prediction system. Clim Dyn 36:649–661
Sun S, Green BW, Bleck R, Benjamin SG (2018) Subseasonal forecasting with an icosahedral, vertically quasi-Lagrangian coupled model. Part II: Probabilistic and deterministic forecast skill. Mon Weather Rev 146(5):1619–1639
Tyrlis E, Hoskins BJ (2008) Aspects of a Northern Hemisphere atmospheric blocking climatology. J Atmos Sci 65:1638–1652
Vitart F (2004) Monthly forecasting at ECMWF. Mon Weather Rev 132:2761–2779
Vitart F (2014) Evolution of ECMWF sub-seasonal forecast skill scores. Q J R Meteorol Soc 140:1889–1899
Vitart F (2017) Madden–Julian Oscillation prediction and teleconnections in the S2S database. Q J R Meteorol Soc 143(706):2210–2220
Vitart F. Ardilouze C, Bonet A, Brookshaw A, Chen M, Codorean C, Déqué M, Ferranti L, Fucile E, Fuentes M, Hendon H (2017) The subseasonal to seasonal (S2S) prediction project database. Bull Am Meteor Soc 98:163–173
Vitart F et al (2008) The new VAREPS-monthly forecasting system: a first step towards seamless prediction. Q J R Meteorol Soc 134(636):1789–1799
Weber NJ, Mass CF (2017) Evaluating CFSv2 subseasonal forecast skill with an emphasis on tropical convection. Mon Weather Rev 145:3795–3815
Wheeler MC, Zhu H, Sobel AH, Hudson D, Vitart F (2017) Seamless precipitation prediction skill comparison between two global models. Q J R Meteorol Soc 143(702):374–383
Wheeler MC, Hendon HH (2004) An all-season real-time multivariate MJO index: development of an index for monitoring and prediction. Mon Weather Rev 132:1917–1932
White CJ et al (2017) Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorol Appl 24.3:315–325
Woolnough SJ (2019) The Madden-Julian Oscillation. In: Robertson AW, Vitart F (eds) Sub-seasonal to seasonal prediction: the gap between weather and climate forecasting. Elsevier, Amsterdam, pp 93–117
Yin Y, Alves O, Oke PR (2011) An ensemble ocean data assimilation system for seasonal prediction. Mon Weather Rev 139:786–808
Zhu H, Wheeler MC, Sobel AH, Hudson D (2014) Seamless precipitation prediction skill in the tropics and extratropics from a global model. Mon Weather Rev 142:1556–1569
Zuo H, Balmaseda MA, Tietsche S, Mogensen K, Mayer M (2019) The ECMWF operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment. Ocean Sci 15(3):779–808

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.