Climate Dynamics
https://doi.org/10.1007/s00382-018-4457-z

Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models

Felipe M. de Andrade[1] · **Caio A. S. Coelho**[1] · **Iracema F. A. Cavalcanti**[1]

Received: 8 December 2017 / Accepted: 19 September 2018

Abstract
This study assessed subseasonal global precipitation hindcast quality from all Subseasonal to Seasonal (S2S) prediction project models. Deterministic forecast quality of weekly accumulated precipitation was verified using different metrics and hindcast data considering lead times up to 4 weeks. The correlation scores were found to be higher during the first week and dropped as lead time increased, confining meaningful signals in the tropics mostly due to El Niño–Southern Oscillation and Madden–Julian Oscillation-related effects. The contribution of these two phenomena to hindcast quality was assessed by removing their regressed precipitation patterns from predicted fields. The model's rank showed ECMWF, UKMO, and KMA as the top scoring models even when using a single control member instead of the mean of all ensemble members. The lowest correlation was shared by CMA, ISAC, and HMCR for most weeks. Models with larger ensemble sizes presented noticeable reduction in correlation when subsampled to fewer perturbed members, showing the value of ensemble prediction. Systematic errors were measured through bias and variance ratio revealing in general large positive (negative) biases and variance overestimation (underestimation) over the tropical oceans (continents and/or extratropics). The atmospheric circulation hindcast quality was also examined suggesting the importance of using a relatively finer spatial resolution and a coupled model for resolving the tropical circulation dynamics, particularly for simulating tropical precipitation variability. The extratropical circulation hindcast quality was found to be low after the second week likely due to the inherent unpredictability of the extratropical variability and errors associated with model deficiencies in representing teleconnections.

Keywords Subseasonal prediction · S2S prediction project models · Hindcast quality · Precipitation · Teleconnections

1 Introduction

Weather predictions up to the first 2 weeks are mainly affected by the atmospheric initial conditions (e.g., Lorenz 1963; Kalnay 2003). On the other hand, monthly and seasonal climate predictions are substantially influenced by slowly varying boundary conditions, for instance sea surface temperature (SST), soil moisture and snow cover (e.g., Palmer and Anderson 1994; Shukla 1998). In between these time scales is the subseasonal time range, that is an intermediate time scale considered until recently as a "predictability desert" (Vitart et al. 2012) where much of the memory of the atmospheric initial conditions is lost and where the boundary forcings do not have substantial influence on the outcomes (Vitart et al. 2017). Therefore, producing skillful subseasonal predictions has been a great challenge for operational forecasters (Johnson et al. 2014).

Recently, motivated by the need to fill the "predictability gap" between weather and climate predictions, some studies have investigated the role played by possible sources of subseasonal predictability. This investigation included several processes in the climate system, such as the Madden–Julian Oscillation (MJO) (e.g., Li and Robertson 2015; Liang and Lin 2018), El Niño–Southern Oscillation (ENSO) (e.g., Li and Robertson 2015; Liang and Lin 2018), teleconnections (e.g., Black et al. 2017; Vitart 2017), land surface initial conditions (e.g., Koster et al. 2010; Kumar et al. 2014), snow initial conditions (e.g., Lin and Wu 2011; Jeong et al. 2013),

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00382-018-4457-z) contains supplementary material, which is available to authorized users.

✉ Felipe M. de Andrade
felipestratus@gmail.com

[1] Centro de Previsão de Tempo e Estudos Climáticos (CPTEC), Instituto Nacional de Pesquisas Espaciais (INPE), Rodovia Presidente Dutra, Km 40, SP-RJ, Cachoeira Paulista, SP 12630-000, Brazil

Published online: 28 September 2018

and stratosphere-troposphere interaction (e.g., Baldwin et al. 2003; Garfinkel and Schwartz 2017). A more comprehensive understanding and a better representation of these potential sources of predictability together with improvements in observation systems, computing resources, model development, and data assimilation techniques should result in potentially skillful subseasonal predictions (Vitart et al. 2012).

In this context, the World Weather Research Programme (WWRP) jointly with the World Climate Research Programme (WCRP) implemented the Subseasonal to Seasonal (S2S) prediction project with the main goal of improving prediction skill and physical understanding on the S2S time scale (Vitart et al. 2012; Robertson et al. 2015). For this purpose, an extensive database (Vitart et al. 2017) has been developed and currently archives near real-time ensemble predictions and hindcasts (reforecasts) up to approximately 60 days from ten operational centres and one research institute, which are: the Australian Bureau of Meteorology (BoM), the China Meteorological Administration (CMA), the European Centre for Medium-Range Weather Forecasts (ECMWF), the Environment and Climate Change Canada (ECCC), the Institute of Atmospheric Sciences and Climate of the National Research Council (ISAC), the Hydrometeorological Centre of Russia (HMCR), the Japan Meteorological Agency (JMA), the Korea Meteorological Administration (KMA), the Météo-France/Centre National de Recherche Météorologiques (CNRM), the National Centers for Environmental Prediction (NCEP), and the United Kingdom's Met Office (UKMO).

The S2S database has been used to investigate different scientific and modelling issues, such as the subseasonal prediction for the July 2015 West-European heat wave with the CNRM prediction system (Ardilouze et al. 2017), the subseasonal predictions of the December 2013 southern South America heat wave produced by BoM and CMA models (Osman and Alvarez 2017), the subseasonal predictions of California precipitation during the unusual winters of 2015–2016 and 2016–2017 using ECMWF and NCEP prediction systems (Wang et al. 2017), the ability of the ECMWF prediction system to predict atmospheric rivers activity at S2S time scales (Baggett et al. 2017), the subseasonal probabilistic predictions of boreal summer monsoon rainfall (Vigaud et al. 2017b) and North American precipitation (Vigaud et al. 2017a) using extended logistic regression applied to ECMWF, NCEP, and CMA models, the MJO forecast in the CMA prediction system (Liu et al. 2017), the effect of the MJO on the Northern Hemisphere stratospheric polar vortex subseasonal predictions using ECMWF and BoM prediction systems (Garfinkel and Schwartz 2017), and simulations of the Asian summer monsoon (Jie et al. 2017) and of the MJO with its teleconnections (Vitart 2017) considering all S2S models, except the KMA model which was not available in the S2S database at the time of writing. Prior to the full implementation of the S2S database, global precipitation predictions have been analyzed through a seamless verification approach in the tropics and extratropics with BoM and ECMWF models (Zhu et al. 2014; Wheeler et al. 2017) and also considering a submonthly prediction skill evaluation with ECMWF, NCEP, and JMA prediction systems (Li and Robertson 2015). Additionally, only a few studies have investigated the relationship between subseasonal precipitation and atmospheric circulation forecast quality by assessing, for instance, the NCEP (Zhang et al. 2016), ECCC (Liang and Lin 2018), and ECMWF (Olaniyan et al. 2018) prediction systems.

Although the aforementioned studies have addressed various subseasonal prediction issues there are several other research issues that still require investigation. For example, a comparative global precipitation hindcast quality assessment, exploring the common virtues and deficiencies in the subseasonal prediction range of all S2S models, is still undocumented. Furthermore, an evaluation of possible connections between the subseasonal global precipitation and atmospheric circulation hindcast quality for providing insight and a better understanding of subseasonal tropical–tropical and tropical-extratropical interaction mechanisms and their potential predictability also deserves a detailed assessment. Therefore, the aim of this study is to perform a quality assessment of subseasonal global precipitation hindcasts of all 11 S2S models, providing an unprecedented comparative framework for investigating the mechanisms and possible key sources of subseasonal predictability. Such an assessment also provides the opportunity for identifying the strengths and weaknesses in each model. The analysis is conducted using a weekly time frame for more properly evaluating the feasibility of subseasonal predictions (Li and Robertson 2015). Particular attention is devoted to prediction leads of 3–4 weeks because there is a rising interest from the scientific, operational and applications communities in improving predictions with a time range longer than the weather predictability limit and shorter than that of climate predictions (White et al. 2017).

Section 2 presents a brief description of the S2S database and the datasets used for hindcast quality assessment, along with the verification methods employed to evaluate the subseasonal hindcast quality from all S2S prediction systems. Section 2 also provides a description of the methods used for investigating the contribution of particular sources of subseasonal predictability on hindcast quality. The results are described in Sect. 3 and Sect. 4 provides a summary and discussions.

2 Data and methods

2.1 S2S database

This study used interpolated hindcasts at the regular 1.5° in latitude and longitude spatial resolution from 11 global prediction systems participating in the S2S prediction project, which are made available through the S2S database maintained by the ECMWF. Each S2S model has a control member (using a single unperturbed initial condition) and a number of perturbed members produced for sampling uncertainty in the initial condition. Table 1 shows the main features of S2S models hindcasts (see Vitart et al. 2017 for additional information). Some models are coupled with an ocean and sea ice component, and others are uncoupled using persisted SST and sea ice anomalies [except ISAC that is the only prediction system with a slab ocean (Malguzzi et al. 2011)]. Additionally, the 11 S2S prediction systems have different prediction time range, spatial resolution, hindcast frequency, hindcast period, and ensemble size, except UKMO and KMA models that have exactly the same configurations. The only difference between UKMO and KMA prediction systems is the atmospheric initial condition (Noh et al. 2016). Furthermore, some hindcasts have a fixed date for model version (e.g., NCEP), whilst others have an on the fly production cycle (e.g., ECMWF) in which hindcasts are produced with the most up to date model version at the time of forecast issuance. For the latter configuration, we used hindcasts corresponding to model versions used in the years of 2016 (ECMWF, UKMO, KMA, ECCC, and HMCR) and 2017 (UKMO and KMA). Besides, JMA and ISAC have two model versions, but only those corresponding to the year of 2017 were used. In spite of the differences among S2S models, it is possible to make inter-comparisons because there are enough common features (Jie et al. 2017; Vitart 2017).

In general, the S2S models do not have the same hindcast frequency and start dates preventing a multi-model evaluation. Notwithstanding, it was still possible to select hindcasts from two initial conditions around the 1st and 15th day of each extended season month [i.e., extended boreal winter (from November to March) and extended boreal summer (from May to September)] over the 1999–2009 period to match all models with the same degrees of freedom and common period making the single model inter-comparison meaningful. This procedure resulted in a total of 110 pairs of hindcasts and corresponding observations for each of the 11 S2S models. This is a considerably larger sample than usually available for performing similar hindcast quality assessment with seasonal prediction models, which typically have around 30 pairs of hindcasts and corresponding observations.

The assessment was performed using 24-h accumulated precipitation and stream function at 200 hPa computed from 24-h instantaneous zonal and meridional wind components. Both zonally symmetric (ZSPSI) and asymmetric (ZAPSI) stream function components were analyzed in order to provide a full assessment of the global rotational circulation. The symmetric component represents the 200 hPa stream function itself prior to removing the zonal mean, which is required for obtaining the asymmetric component. The tropical atmospheric circulation shows a substantial zonally symmetric response due to locally forced equatorial Kelvin and Rossby waves (Clarke 2008). On the other hand, the observed extratropical mean flow varies strongly with longitude, which in turn has a profound effect on the location and strength of teleconnections (e.g., Branstator 1983). Thus,

Table 1 The main features of the S2S models hindcasts (Vitart et al. 2017)

S2S models	Time range (days)	Spatial resolution	Hindcast frequency	Hindcast period	Ensemble size	Ocean coupled	Sea-ice coupled
ECMWF[a]	46	Tco639/Tco319, L91	2/week	Past 20 years	11	Yes	No
UKMO[a]	60	N216, L85	4/month	1996–2009	3	Yes	Yes
NCEP	44	T126, L64	Daily	1999–2010	4	Yes	Yes
BoM	62	T47, L17	6/month	1981–2013	33 (3×11)	Yes	No
CNRM	61	T255, L91	2/month	1993–2014	15	Yes	Yes
CMA	60	T106, L40	Daily	1994–2014	4	Yes	Yes
KMA[a]	60	N216, L85	4/month	1996–2009	3	Yes	Yes
ISAC	31	0.75°×0.56°, L54	Every 5 days	1981–2010	5	No	No
JMA	33	TL479/TL319, L100	3/month	1981–2010	5	No	No
ECCC[a]	32	0.45°×0.45°, L40	Weekly	1995–2012	4	No	No
HMCR[a]	61	1.1°×1.4°, L28	Weekly	1985–2010	10	No	No

[a]Hindcasts are produced on the fly (model version is not fixed)

removing the zonal mean of stream function by computing ZAPSI at 200 hPa can better reveal essential extratropical atmospheric circulation characteristics, such as barotropic Rossby wave propagation (e.g., Jin and Hoskins 1995; Held et al. 2002; Grimm and Reason 2015).

2.2 Reference datasets used for verification

The observational reference used for verifying precipitation hindcasts was the daily analysis data of the Global Precipitation Climatology Project (GPCP) version 1.2 (Huffman et al. 2001) with regular 1° in latitude and longitude spatial resolution over the 1999–2009 period. The GPCP dataset is a blended product derived from both observational station data and satellite measurements provided by the National Center for Atmospheric Research (NCAR). This dataset was regridded to the regular 1.5° in latitude and longitude spatial resolution through linear interpolation to match the S2S models grid here investigated and used to obtain weekly accumulated precipitation. Following Zhu et al. (2014) and Wheeler et al. (2017), the analysis is limited to the range between 80°N and 80°S due to known problems in the GPCP data at high latitudes (Bolvin et al. 2009). Atmospheric circulation hindcasts were verified using 200 hPa ZSPSI and ZAPSI computed from zonal and meridional wind components at 00Z provided by the ERA-Interim reanalysis (Dee et al. 2011) at the same regular 1.5° in latitude and longitude spatial resolution as the investigated S2S models over the 1999–2009 period. The sensitivity level to the choice of the reanalysis was evaluated using two other datasets linearly interpolated from a $2.5° \times 2.5°$ to a $1.5° \times 1.5°$ horizontal grid resolution: the Japanese 55-year reanalysis (JRA55) (Ebita et al. 2011) and the National Centers for Environmental Prediction (NCEP)—Department of Energy (DOE) reanalysis 2 (Kanamitsu et al. 2002).

2.3 Forecast quality assessment

Deterministic hindcast quality was assessed using linear correlation (Pearson's correlation coefficient), mean error (or bias) and variance ratio (i.e., the ratio between the predicted variance and the observed variance). These three deterministic measures were calculated between the ensemble mean of all hindcast members (including the control member and the perturbed members, hereafter referred to as the ensemble mean) and the corresponding observation/reanalysis. Linear correlation evaluates prediction quality by looking at the phase relationship between the predicted and observed data. In other words the correlation coefficient provides an assessment of the strength of the linear association between predictions and observations. The other two measures diagnose systematic errors that are not revealed with the correlation coefficient alone. It is important to note that large differences in ensemble size among models may affect model intercomparison when using the ensemble mean by favoring models with larger ensemble size (Vitart 2017). Hence, in order to have a fair comparison, we also analyzed the same verification metrics using only the control member and a three members ensemble mean of each model (i.e., the minimum common number of available ensemble member among the 11 investigated S2S models as shown in Table 1, which is hereafter referred to as the 3M ensemble mean). This comparison of the hindcast quality dependence on the ensemble size provides relevant information about the value of ensemble prediction and its generation methods (Jie et al. 2017; Vitart 2017). The limited ensemble size for some models prevented the expansion of the comparative hindcast quality assessment using probabilistic measures.

All hindcast quality measures were computed over 110 selected start dates, applied separately for each grid point, lead time, and extended season, considering four weekly precipitation accumulation periods and average atmospheric circulation: days 1–7 (week-1), 8–14 (week-2), 15–21 (week-3), and 22–28 (week-4). Li and Robertson (2015) used the same four weekly periods to evaluate the precipitation hindcasts produced by ECMWF, NCEP, and JMA ensemble prediction systems. According to these authors, as the subseasonal range is beyond the weather prediction limit, the weekly scale is an appropriate choice to remove part of the weather noise. The hindcast anomalies were obtained by subtracting their respective climatology depending on the start date and lead time. The climatology was computed in a cross-validated way leaving 1 year out (Vitart 2017) for each ensemble sampling (i.e., for the ensemble mean of all available members, for the control member, and for the ensemble mean of three members). In this cross-validation procedure the climatological means of hindcasts initialized at the same start date and lead time are computed over the common period (1999–2009) excluding the target year to be verified. For example, for the week-1 hindcast prediction initialized on 1 January 1999, all week-1 hindcast predictions initialized on 1 January 2000–2009 are used to compute the climatological mean, which is then used to determine the anomaly for week-1 in 1999. The observed and reanalysis anomalies were calculated in the same way. Moreover, the corresponding hindcast climatologies of each of the three BoM model versions (Hudson et al. 2013) were computed separately, but the anomalies are the same when using the BoM multi-model climatology as also reported by Zhu et al. (2014). Statistical significance of the computed correlation coefficients was analyzed using a two-sided Student's t test (Wilks 2006) with $n-2$ degrees of freedom for the rejection of the null hypothesis of null correlation considering a p value of 0.05 (5% level). The effective number of

independent samples (n) was estimated taking into account autocorrelation properties of temporal series as proposed by Livezey and Chen (1983).

2.4 Sources of subseasonal predictability

ENSO and the MJO are recognized as the dominant modes of interannual (e.g., Drosdowsky and Chambers 2001) and subseasonal (e.g., Mo and Nogues-Paegle 2005) tropical climate variability, respectively, with well established impact on global precipitation (e.g., Ropelewski and Halpert 1987; Jones et al. 2004). The role of these particular climate drivers on the S2S models hindcast quality was investigated using indices commonly employed for evaluating ENSO and MJO activities (e.g., Li and Robertson 2015; Liang and Lin 2018).

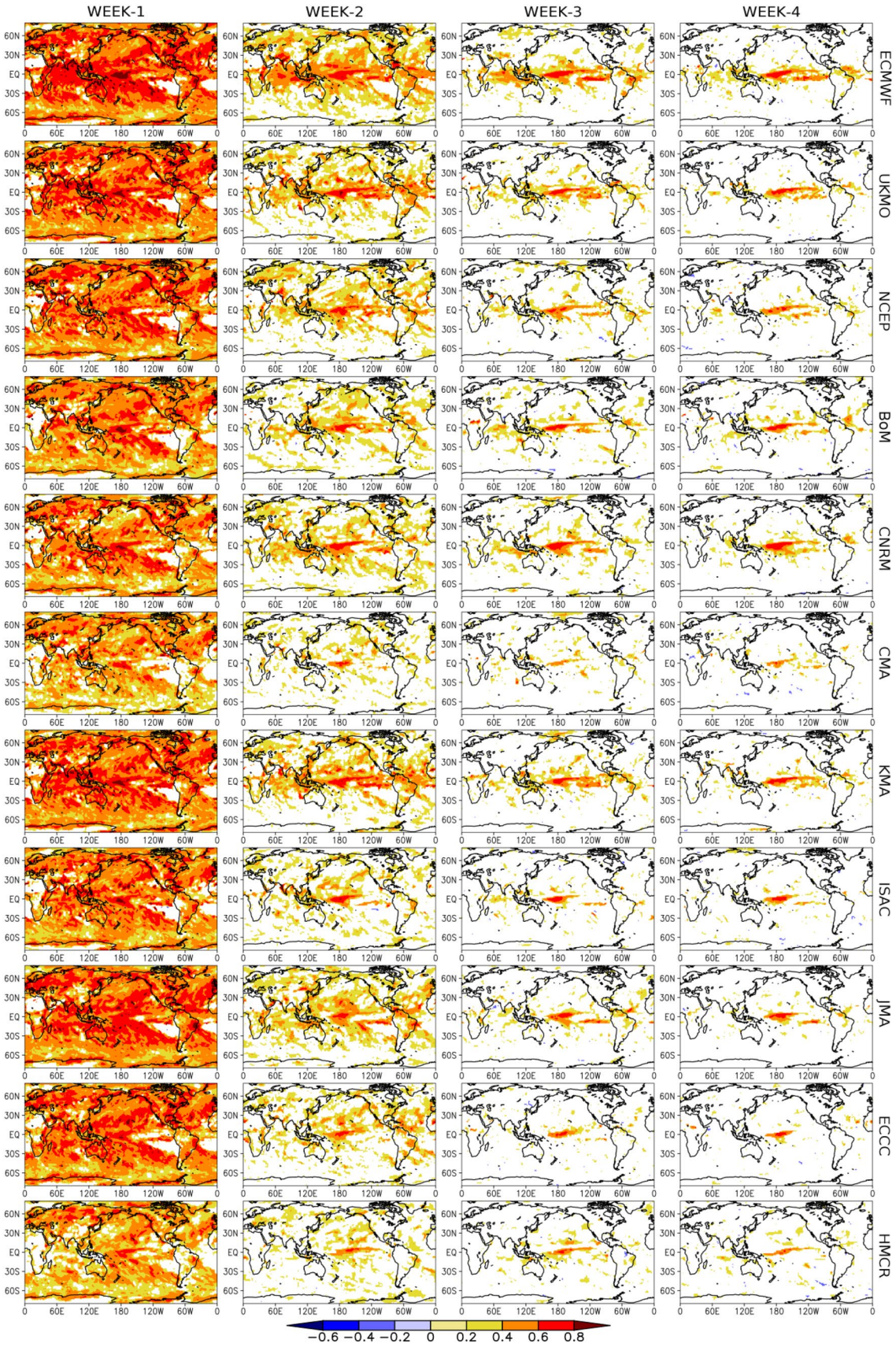
The ENSO activity was measured by the Niño-3.4 index defined as the SST anomaly over the Niño-3.4 region (5°S–5°N, 120°W–170°W) (Trenberth 1997). The ENSO index was computed using weekly mean SST obtained from the daily Optimum Interpolation SST version 2 (OISST.v2) (Reynolds et al. 2007) dataset sourced by the National Oceanic and Atmospheric Administration (NOAA) at the regular 0.25° in latitude and longitude spatial resolution over the 1999–2009 period. The daily OISST.v2 was linearly interpolated to the regular 1.5° in latitude and longitude spatial resolution and used to obtain 7-day mean SST periods for every single day of the year. Additionally, the SST seasonal cycle was removed by computing cross-validated anomalies (i.e., subtracting the time mean leaving 1 year out) and the ENSO index was normalized by the weekly cross-validated standard deviation.

The MJO activity was quantified by the Real-time Multivariate MJO (RMM) daily index components (RMM1 and RMM2) (Wheeler and Hendon 2004, hereafter referred as WH04). The index was obtained by projecting near-equatorially averaged normalized daily anomalies of zonal wind (850 hPa and 200 hPa) and outgoing longwave radiation (OLR) at the top of the atmosphere onto the two pre-computed leading empirical orthogonal functions from WH04. The computation of the RMM index follows the methodology described in WH04 and Gottschalck et al. (2010). The used zonal wind data was from the reanalysis datasets mentioned in Sect. 2.2 and the interpolated OLR was from NOAA (Liebmann and Smith 1996) covering the 1999–2009 period. The data were analyzed using a horizontal grid resolution of 2.5° × 2.5°, corresponding to the same 144 longitudinal points of WH04's pre-computed eigenvectors. Daily anomalies were obtained by subtracting the cross-validated seasonal cycle from the total field and next the 120-day mean of the previous 120 days of each day was removed in order to filter the low-frequency variability such as that associated to ENSO. Before data projection, the zonal wind and OLR anomalies were near-equatorially averaged (15°S–15°N) and divided by its respective observed normalization factor (computed by WH04 having the values of 15.1 W m^{-2} for OLR, 1.81 m s^{-1} for 850 hPa zonal wind, and 4.81 m s^{-1} for 200 hPa zonal wind) as described in Gottschalck et al. (2010). Finally, the projected components were normalized by their respective observed standard deviations estimated by WH04 to generate the RMM1 and RMM2 time series. Weekly values for each RMM component were also computed using a similar procedure for obtaining weekly SST.

In order to diagnose the impact of ENSO and MJO on subseasonal global precipitation variability, a linear regression analysis between GPCP weekly accumulated precipitation anomalies and observed weekly mean ENSO and MJO indices was conducted. This approach allows the identification of ENSO and MJO-related linear patterns in observed weekly precipitation anomalies and then the investigation of their possible contributions to S2S models hindcast quality by removing the identified variability that is linearly dependent on ENSO and MJO from the predicted fields. Regressed values were scaled to a value of one standard deviation of the corresponding index as in Lo and Hendon (2000). The statistical significance for the obtained linear regression coefficients was assessed through a two-sided Student's t test considering a p value of 0.05 (5% level). To account for the autocorrelation in the weekly temporal series, the sample size was corrected to the effective sample size following Livezey and Chen (1983).

3 Results

In this section, subseasonal global precipitation hindcast quality is assessed for all S2S models as a function of weekly lead times using the metrics described in Sect. 2.3. The hindcast quality assessment for the May–September period generally corroborates findings for the November–March period. For this reason and for brevity, this paper mainly discusses the results for the November–March period, with the results for the May–September period mentioned when necessary throught the paper and made available in the Online Resource 1. The assessment starts with the correlation between hindcast and observed accumulated precipitation anomalies because this is the simplest and most commonly used deterministic verification metric (e.g., Zhu et al. 2014; Li and Robertson 2015; Wheeler et al. 2017). Then, systematic errors of the hindcast accumulated precipitation are investigated. In addition, an assessment of the subseasonal atmospheric circulation hindcast quality is presented in order to evaluate the model's ability to represent the observed atmospheric circulation patterns and possible connections with subseasonal precipitation prediction quality.


WEEK-1
WEEK-2
WEEK-3
WEEK-4
ECMWF
UKMO
NCEP
BoM
CNRM
CMA
KMA
ISAC
JMA
ECCC
HMCR
60N
30N
EQ
30S
60S
0
60E
120E
180
120W
60W
−0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8

◂**Fig. 1** Correlation between the ensemble mean and observed (GPCP) accumulated precipitation anomalies for each S2S model (rows) during weeks 1–4 (columns) for hindcasts initialized from November to March over the 1999–2009 period. Correlation coefficients statistically significant at the 5% level are shaded

3.1 Linear association assessment

Figure 1 displays the correlation between the hindcast ensemble mean and the observed accumulated precipitation anomalies for each S2S model in different weekly lead times initialized during November–March over the 1999–2009 period. A similar figure for the hindcasts initialized during May–September is available in the Online Resource 1—Fig. 1. Figure 1 shows the highest correlation scores for week-1. The scores drop as lead time increases, attaining meaningful signals mainly on the tropical oceanic regions as reported by Li and Robertson (2015). This feature is more evident after week-2 when the weather prediction limit is reached. Correlation scores over the continents are generally low after week-2, though some models (especially coupled models, such as ECMWF) exhibit remarkable scores over the northernmost portion of northeast Brazil up to week-4 during both extended seasons. This suggests that the rainy season over the northernmost portion of northeast Brazil (i.e., from March to July; Marengo and Bernasconi 2015) can potentially be predicted on the subseasonal time scale up to 4 weeks in advance. All models show for all weeks negligible correlation scores near subtropical dry zones over Africa, and the eastern Atlantic and eastern Pacific Oceans as also reported by Zhu et al. (2014) and Wheeler et al. (2017). Besides, in weeks 3–4 the correlation scores are consistently high on the equatorial Pacific most likely related to the predictability provided by ENSO (Zhu et al. 2014; Li and Robertson 2015; Wheeler et al. 2017), particularly during its mature phase (Fig. 1). Most coupled models show moderate correlation scores over the tropical Indian Ocean and some regions of the Maritime Continent likely due to a better ability to predict the MJO evolution up to about 3 weeks (Vitart 2017). Likewise, almost all models show correlation scores over the equatorial Atlantic Ocean resembling the Intertropical Convergence Zone-like precipitation pattern (e.g., Schneider et al. 2014), which may be modulated by ENSO (e.g., Marengo and Hastenrath 1993), MJO (e.g., Tomaziello et al. 2016), and the local cross-equatorial SST gradient (e.g., Chiang et al. 2002).

In the extratropical region, the week-2 correlation scores are more pronounced in the extended winter hemisphere when compared to the extended summer hemisphere (compare Fig. 1 with the Online Resource 1—Fig. 1), likely due to the most prominent dynamics of teleconnections during the winter period (e.g., Hoskins and Ambrizzi 1993; Ambrizzi et al. 1995; Grimm and Ambrizzi 2009). It is worth noting that even for the coupled models shown in the top seven rows of Fig. 1 (see also the Online Resource 1—Fig. 1) the extratropical prediction ability in weeks 3–4 is low, especially on the continental areas, probably owing to the inherent unpredictability of the extratropical variability and model deficiencies in simulating land surface processes (e.g., Kumar et al. 2014) and tropical-extratropical interactions, such as the Pacific-South American (PSA) pattern (e.g., Mo and Paegle 2001) and the Pacific-North American (PNA) pattern (e.g., Barnston and Livezey 1987), which can both be related to ENSO (e.g., Horel and Wallace 1981; Mo and Paegle 2001) and the MJO (e.g., Mo and Nogues-Paegle 2005).

The contribution of ENSO and MJO to S2S models precipitation hindcast quality started with an evaluation of the observed relationship between indices representing these two phenomena and precipitation (Fig. 2). This figure shows the linear regression maps between GPCP weekly accumulated precipitation anomalies and weekly mean ENSO (Niño-3.4) and MJO (RMM1 and RMM2) indices for four weekly periods defined for each extended boreal winter month: days 1–7 (first period), 8–14 (second period), 15–21 (third period), and 22–28 (fourth period). The regression analysis using RMM components showed a very similar feature independently of the used reanalysis dataset. Thus, for the sake of brevity, only results using ERA-Interim data are presented. The spatial pattern of Fig. 2a shows consistent ENSO characteristics with previous studies, such as opposite signals between the Maritime Continent and central-eastern equatorial Pacific (e.g., Kumar et al. 2013) and between northern and southeastern South America (e.g., Grimm and Ambrizzi 2009). Furthermore, coherent regression values are also noted over the tropical/subtropical Indian Ocean representing a basinwide pattern (e.g., Taschetto et al. 2011) and over the south/southeastern North America related to the PNA pattern activity (e.g., Horel and Wallace 1981; Coleman and Rogers 2003). The linear regression maps of Fig. 2b, c show MJO features manifested through opposite signals between Maritime Continent and northeastern South America (e.g., Carvalho et al. 2004) for the RMM1 index (Fig. 2b) and a dipole-like pattern between tropical Indian and west Pacific Oceans (e.g., Matthews 2008) for the RMM2 index (Fig. 2c). Significant regression values are also found in the extratropical region probably associated with barotropic Rossby wave propagation on the Pacific-North America and Pacific-South America sectors (e.g., Mo and Nogues-Paegle 2005). Similar linear regression maps were obtained for the extended boreal summer (Online Resource 1—Fig. 2) consistent with ENSO and MJO-related variability patterns previously identified (Li and Robertson 2015). All of these features indicate a considerable global correspondence with the correlation assessment shown in Fig. 1 and in the Online Resource 1—Fig. 1, particularly over the

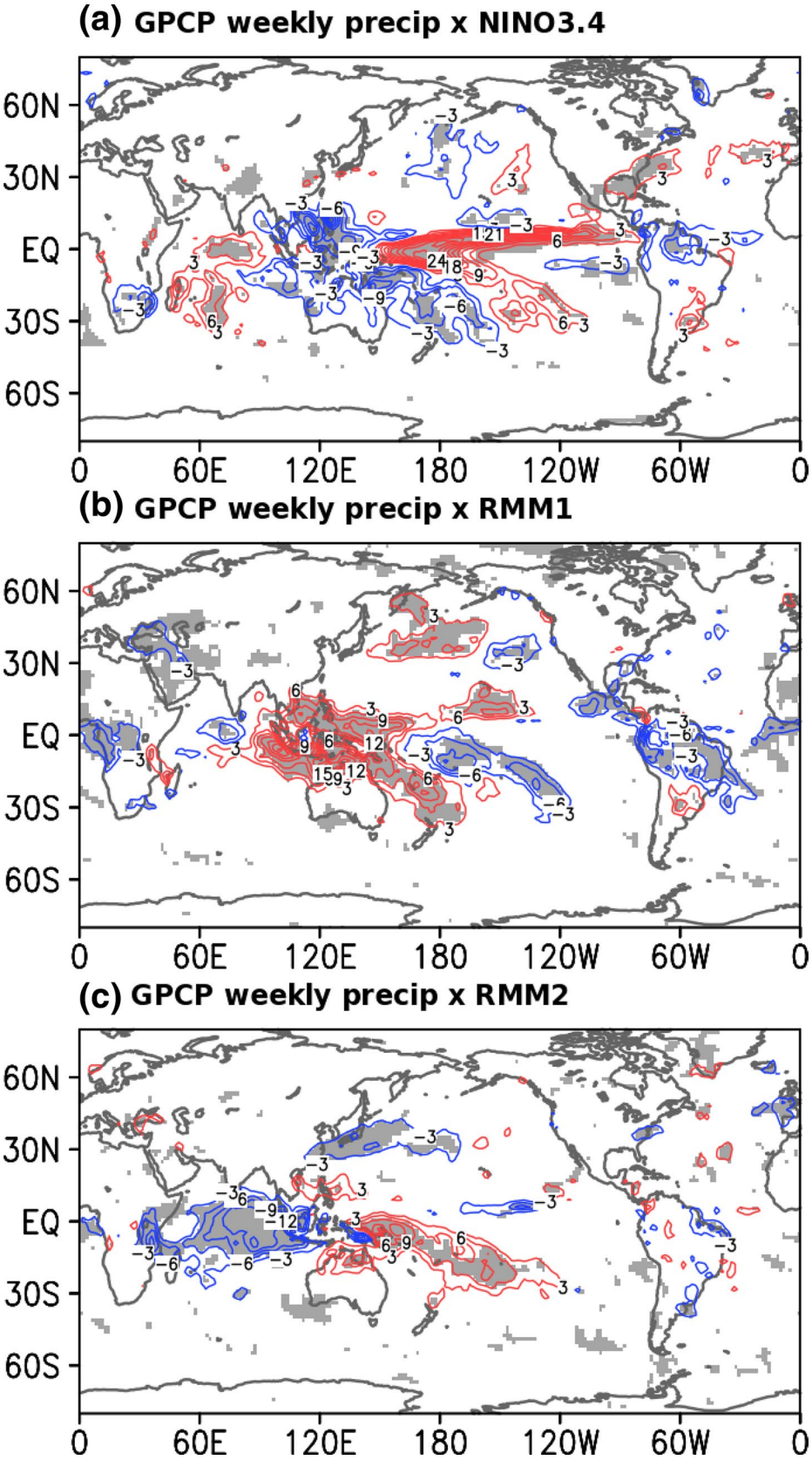


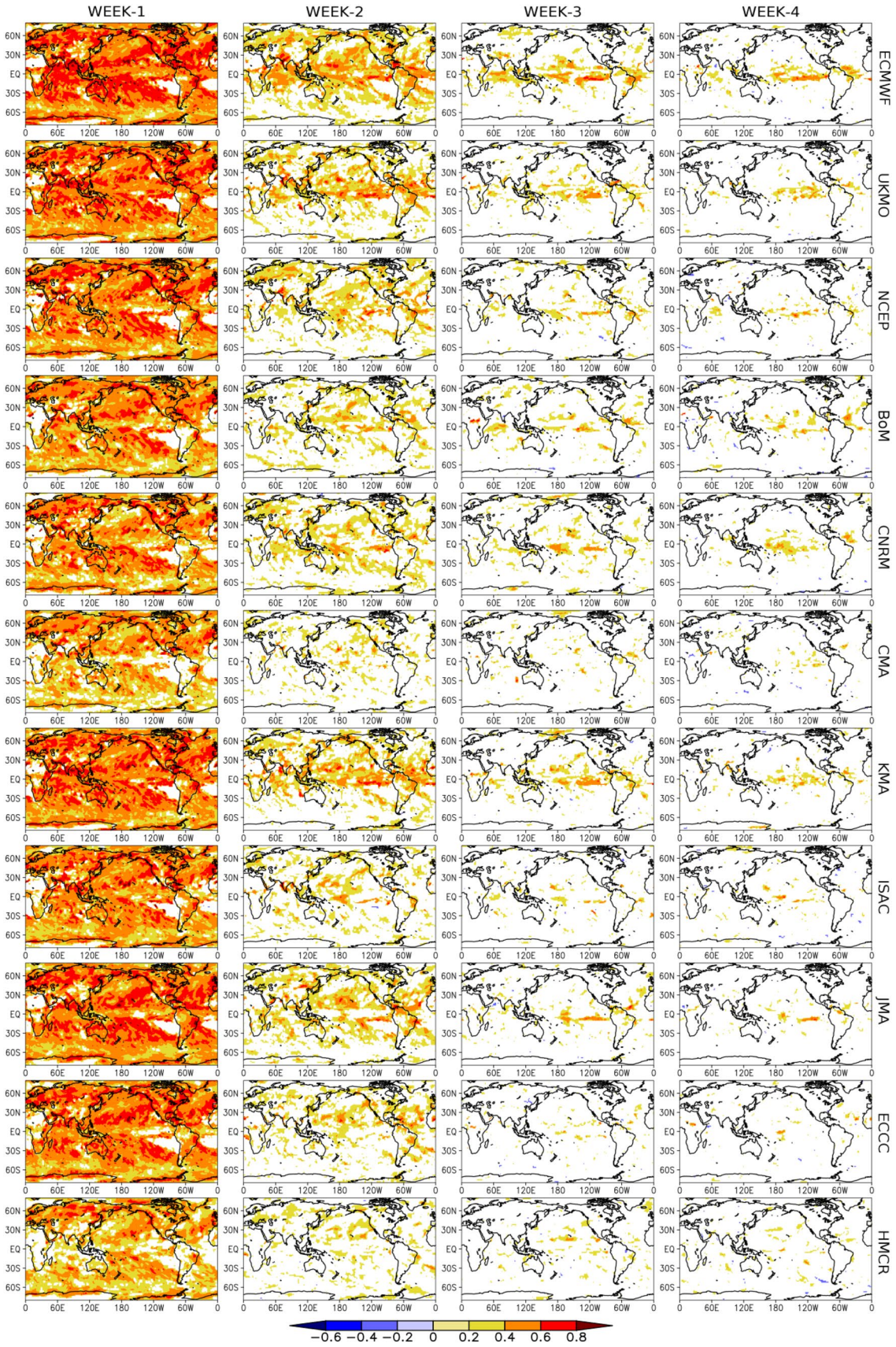
Fig. 2 Linear regression between GPCP weekly accumulated precipitation anomalies and weekly mean of **a** Niño-3.4 index and real-time multivariate MJO index (RMM) components, **b** RMM1 and **c** RMM2 with respect to reanalysis ERA-Interim, considering 4 weekly periods for each extended boreal winter month (November–March) [days 1–7 (first period), 8–14 (second period), 15–21 (third period), and 22–28 (fourth period)] over the 1999–2009 period. The grey shading highlights regression coefficients statistically significant at the 5% level according to a two-sided Student's *t* test. Contour interval is 3 mm/week with positive (negative) values in red (blue)

tropical region. This correspondence suggests that ENSO and the MJO are relevant contributing phenomena for the identified S2S models prediction ability.

The S2S models hindcast quality was also investigated by removing ENSO and MJO projected linear regression patterns from the predicted fields. In this case, the resulting fields to be removed were obtained by regressing out the observed weekly mean indices with hindcasts weekly accumulated precipitation anomalies depending on both lead time and model start date. According to Fig. 3 (Online Resource 1—Fig. 3), which shows the correlation scores when the ENSO and MJO linear regression patterns were sequentially removed from the predicted fields, all S2S models show lower prediction ability over the tropical region when compared to Fig. 1 (Online Resource 1—Fig. 1). This holds true for all weeks and more pronouncedly for the extended boreal winter. During November–March, larger differences are noticed over the Maritime Continent, South America, and the eastern tropical Indian and central equatorial Pacific Oceans, which are regions influenced by ENSO and MJO activities (Fig. 2). Moreover, some regions are slightly modified by removing ENSO and MJO-related effects, notably east equatorial Pacific Ocean, west tropical Indian Ocean, and equatorial Atlantic Ocean. The hindcast quality over these almost unchanged tropical regions are likely modulated by other phenomena, for instance, eastern Pacific ENSO episodes (e.g., Tedeschi et al. 2013), the Indian Ocean Dipole (e.g., Saji et al. 1999), and the Atlantic Meridional Dipole (e.g., Chiang and Vimont 2004). Although tropical correlation scores were clearly modified after removing ENSO and MJO precipitation patterns from the hindcasts, slightly changes are also revealed in the extratropics, particularly for weeks 1–2. This suggests that subseasonal hindcasts prediction ability over the extratropical region is likely modulated by teleconnections triggered by ENSO and MJO. Additional analyses were performed in order to investigate the individual contribution of ENSO and MJO on forecast quality by removing the linear regression patterns of each of these phenomena separately. In the tropics, the correlation scores over the central equatorial Pacific were found to be modulated by ENSO (Niño-3.4 index) whilst over the tropical east Indian Ocean and the Maritime Continent/northeastern Brazil the modulation was evidenced through MJO phases corresponding to RMM2 and RMM1 activity, respectively. In contrast, in the extratropics, for example over south Brazil, both phenomena are responsible to the modulation of the hindcast quality during the first 2 weeks of forecast (not shown).

In a regional scale over South America, several studies have shown that a prominent rainfall dipole pattern between the South Atlantic Convergence Zone (e.g., Cunningham and Cavalcanti 2006) and the subtropical region of southeastern South America generally affects the most densely populated areas in central-east, southeastern, and southern Brazil, Paraguay, Uruguay, and northeastern Argentina during the extended boreal winter (or austral summer) (e.g., Gonzalez and Vera 2014). As displayed in Fig. 1, for most models the South American rainfall dipole pattern is well represented in the first 2 weeks showing high correlation on both of its action centers. This feature is maintained by the ECMWF model up to week-3 likely due to a good MJO extended-range prediction ability (Vitart 2017), as this phenomenon has influence on subseasonal precipitation variability over South America through tropics–tropics and tropics-extratropics teleconnections (Cunningham and Cavalcanti 2006; Grimm and Ambrizzi 2009). Besides, the regression map between GPCP weekly accumulated precipitation anomalies and weekly means of the RMM1 component (Fig. 2b) reveals a spatial pattern over South America resembling the intraseasonal dipole pattern associated with the MJO activity (Gonzalez and Vera 2014). This result also shows that during the extended boreal winter part of the subseasonal precipitation prediction ability over South America is related to MJO phases producing anomalous convective activity over the Maritime Continent (e.g., Carvalho et al. 2004; Cunningham and Cavalcanti 2006), as pointed out in the previous discussion on Fig. 3. For week-4, the dipole-like correlation pattern disappears and statistically significant scores are confined over the far north of South America in almost all models (particularly for the coupled models). This suggests the maintenance of prediction ability due to the effect of local SST anomalies (e.g., Chiang et al. 2002), as the influence of ENSO and MJO does not seem to impact hindcast quality in this region (compare Figs. 1, 3).

As mentioned in Sect. 2.3, the hidcast quality assessment may be sensitive to different ensemble sizes among models and this may affect model inter-comparison. In order to investigate this impact, Figs. 4 and 5 (Online Resource 1—Figs. 4, 5) show correlation scores for hindcast precipitation anomalies using the 3M ensemble mean and the control member for each model. These figures reveal lower correlation scores using reduced ensemble sizes instead of full ensemble sizes (Fig. 1; Online Resource 1—Fig. 1), particularly for models with a larger number of perturbed members (ECMWF, BoM, and CNRM). This sensitivity level is more pronounced over the tropical region for the control member and even for models having the smallest ensemble size (UKMO and KMA). The hindcast quality increase with a larger ensemble size is also noticeable in the extratropics for the first 2 weeks. Curiously, the HMCR model has one of the largest numbers of perturbed members, but it does not show considerable hindcast quality difference when evaluated using reduced ensemble size as reported by Vitart (2017). Additionally, among the uncoupled models, the ECCC model shows the sharpest change in hindcast quality between the control member and the ensemble mean


WEEK-1
WEEK-2
WEEK-3
WEEK-4
ECMWF
UKMO
NCEP
BoM
CNRM
CMA
KMA
ISAC
JMA
ECCC
HMCR
60N
30N
EQ
30S
60S
0
60E
120E
180
120W
60W
−0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8

◂**Fig. 3** Same as Fig. 1, except for removing the regression patterns between hindcast weekly accumulated precipitation anomalies and observed weekly mean Niño-3.4, RMM1, and RMM2 indices from hindcasts

correlation scores during the first 2 weeks, indicating great sensitivity to the ensemble generation method even for a relatively small ensemble size (4 members).

Ranking models according to performance is challenging because the prediction ability varies depending on the geographic location, lead time, and ensemble size, as previously noted. However, a hindcast quality analysis based on the zonal average of anomaly correlation scores for different latitudinal bands can provide some insight in this aspect. Starting from the 80°S–80°N latitudinal band for all longitudes (global domain), Fig. 6 (Online Resource 1—Fig. 6) shows that the ECMWF model presents the highest correlation scores for all 4 weeks when the ensemble mean is considered. The second place is shared by UKMO and KMA models until the CNRM model exceeds these models after week-3. During week-1, JMA and ECCC models show better performance (higher correlation scores) than some coupled models, but for some of these coupled models (CNRM, NCEP, and BoM) this ranking reverses as the lead time increases, particularly for weeks 3–4. On the other hand, CMA, ISAC, and HMCR models have the lowest correlation scores for practically all lead times. This reflects the fact that some models (including the coupled models, such as CMA) seem not to benefit from ensemble forecasting (Vitart 2017). When considering only the control member or the 3M ensemble mean, the correlation score level variation among models is less pronounced and the top scores are now shared by ECMWF, UKMO, and KMA models, showing the importance of coupling the ocean with the atmosphere and refining the spatial resolution. Nevertheless, these ensemble size sampling place the CNRM and BoM coupled models at a worse ranking position when compared to the ensemble mean, particularly during the last 2 weeks, suggesting that its performance is improved when using a larger number of perturbed members. Again, CMA, ISAC, and HMCR models remain at the bottom of the rank in almost all weeks. Actually, the HMCR control member, 3M ensemble mean, and ensemble mean correlation scores seem to be almost identical possibly due to problems of the HMCR perturbed hindcasts as described by Jie et al. (2017). Another important previously noted feature is the great decrease of the ECCC control member correlation scores in weeks 1–2 when compared to the ensemble mean or even with the 3M ensemble mean. This is a striking feature because the other uncoupled models (JMA and HMCR) lack similar behavior. The 20°S-20°N latitudinal band for all longitudes (tropical domain) shows that the model's rank is roughly unchanged in comparison to the global domain. However, correlation scores differences are noticed particularly because correlation scores level variations among models are more pronounced in the tropical domain. Two additional latitudinal bands in the extratropical domain (20°S–80°S and 20°N–80°N) for all longitudes were analyzed and the ECMWF model was identified as the top scoring model for the ensemble mean (not shown). This additional analysis also revealed very low and similar correlation scores for almost all remaining models even when considering the ensemble mean (not shown).

Still regarding to the model's rank, the assessment displays that the best ranking places are shared by coupled models with relatively finer spatial resolution, particularly after week-2. Nonetheless, the JMA model showed good performance for all weeks regardless of its uncoupled configuration. Therefore, the effect of using ocean–atmosphere coupling for improving subseasonal hindcast quality can be explored by investigating the relationship between weekly MJO and precipitation hindcast quality of coupled and uncoupled models with approximately similar spatial resolution (e.g., ECMWF and JMA). The computation of the MJO index for both observations and hindcasts follows the methodology previously described when evaluating the MJO contribution on hindcast quality (Sect. 2.4). The only difference is that the subtraction of the previous 120 day running mean from predicted fields was treated as follows. A combination of observed and hindcast information was used since the hindcasts for the previous 120 days of the initialization time are not available. In order to overcome this lack of information, observations of the 120 days prior to the model's start dates were used to fill those days missing. Following Lin et al. (2008), the removal of the 120-day mean for the forecast at day n was computed by subtracting the average of the $120-n+1$ days of observational data preceding the forecast and the forecast data from day 1 to day $n-1$. Again, the observed MJO index was fully evaluated using ERA-Interim as verifying data since the sensitivity level to the choice of reanalysis was minor providing very similar results for all investigated reanalysis datasets. Figure 7 (Online Resource 1—Fig. 7) shows scatter plots of weekly MJO bivariate correlation scores (Rashid et al. 2011) and correlation scores of the weekly accumulated precipitation anomalies for the ensemble mean of ECMWF and JMA models for different selected regions. The same scatter plots were produced using the control member and the 3M ensemble mean revealing similar characteristics, but with lower prediction ability mainly for ECMWF (not shown). The ECMWF model shows a good correspondence between MJO and precipitation prediction ability, especially for tropical west Pacifc and central-east Indian Oceans where much of subseasonal precipitation variability can be explained by MJO propagation (Mo and Nogues-Paegle 2005). On the extratropics (southeastern United States and south Brazil),

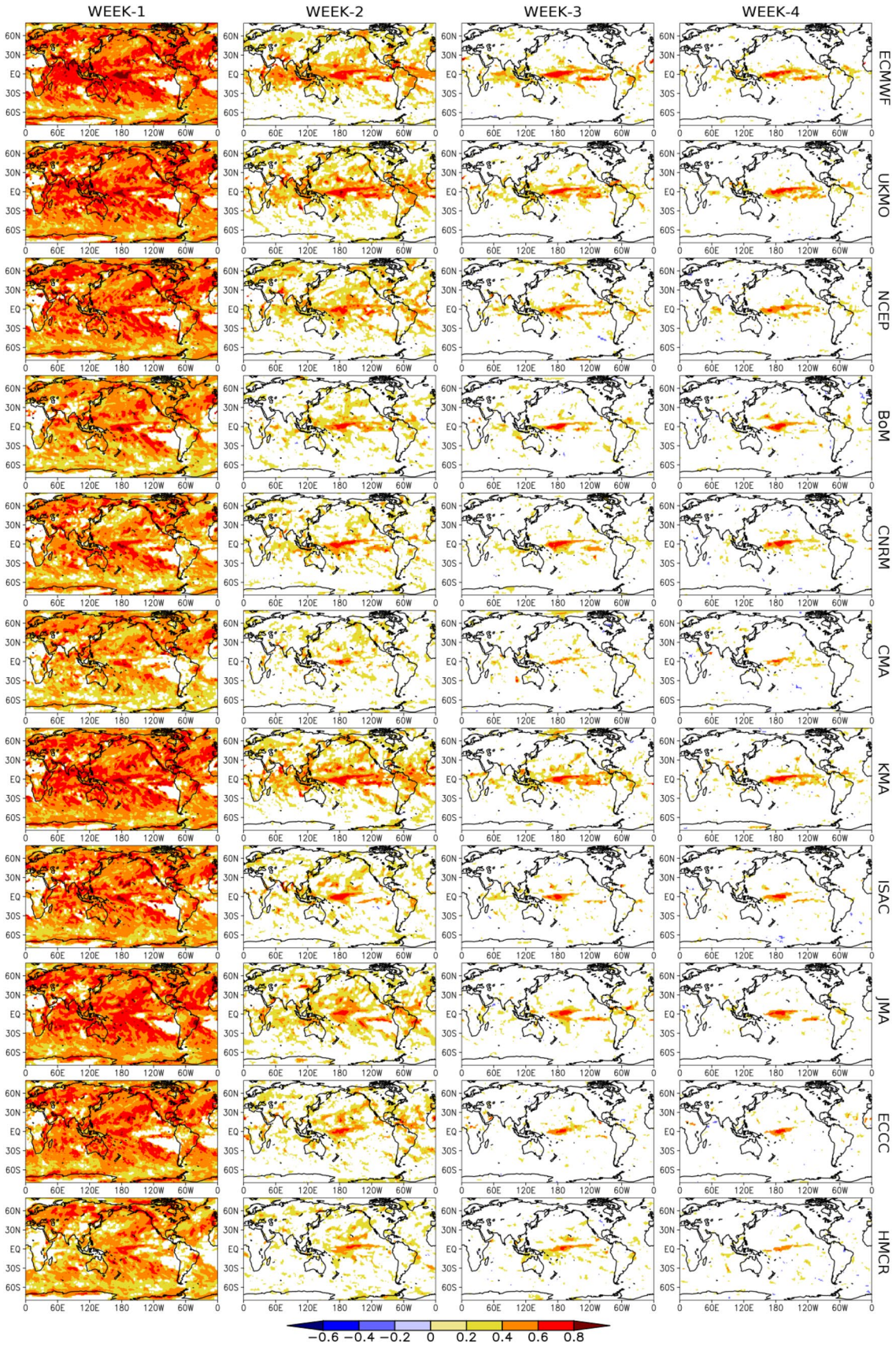


Fig. 4 Same as Fig. 1, except by using the ensemble mean of three ensemble members (referred to as the 3M ensemble mean)

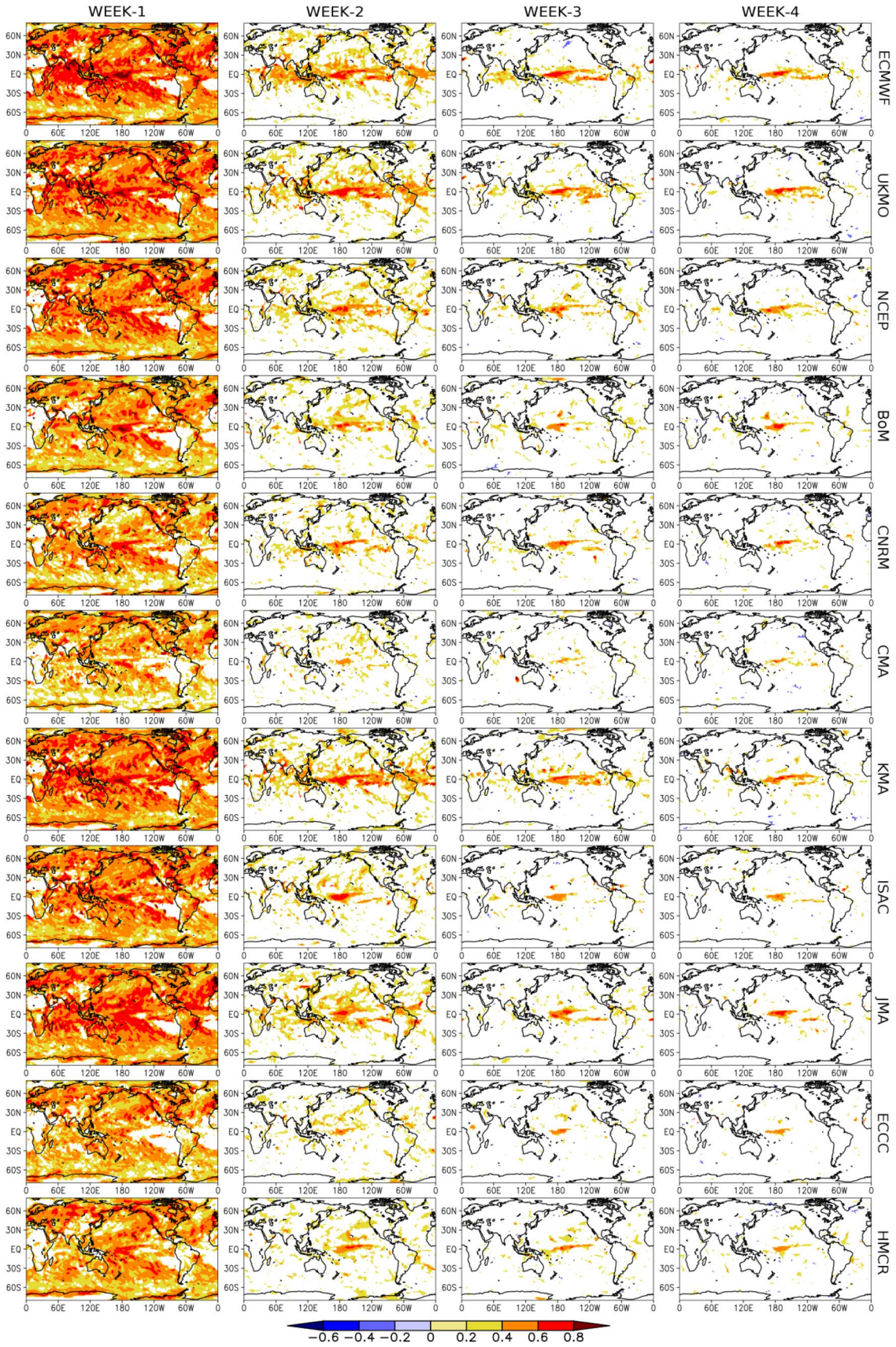


Fig. 5 Same as Fig. 1, except by using the single control member

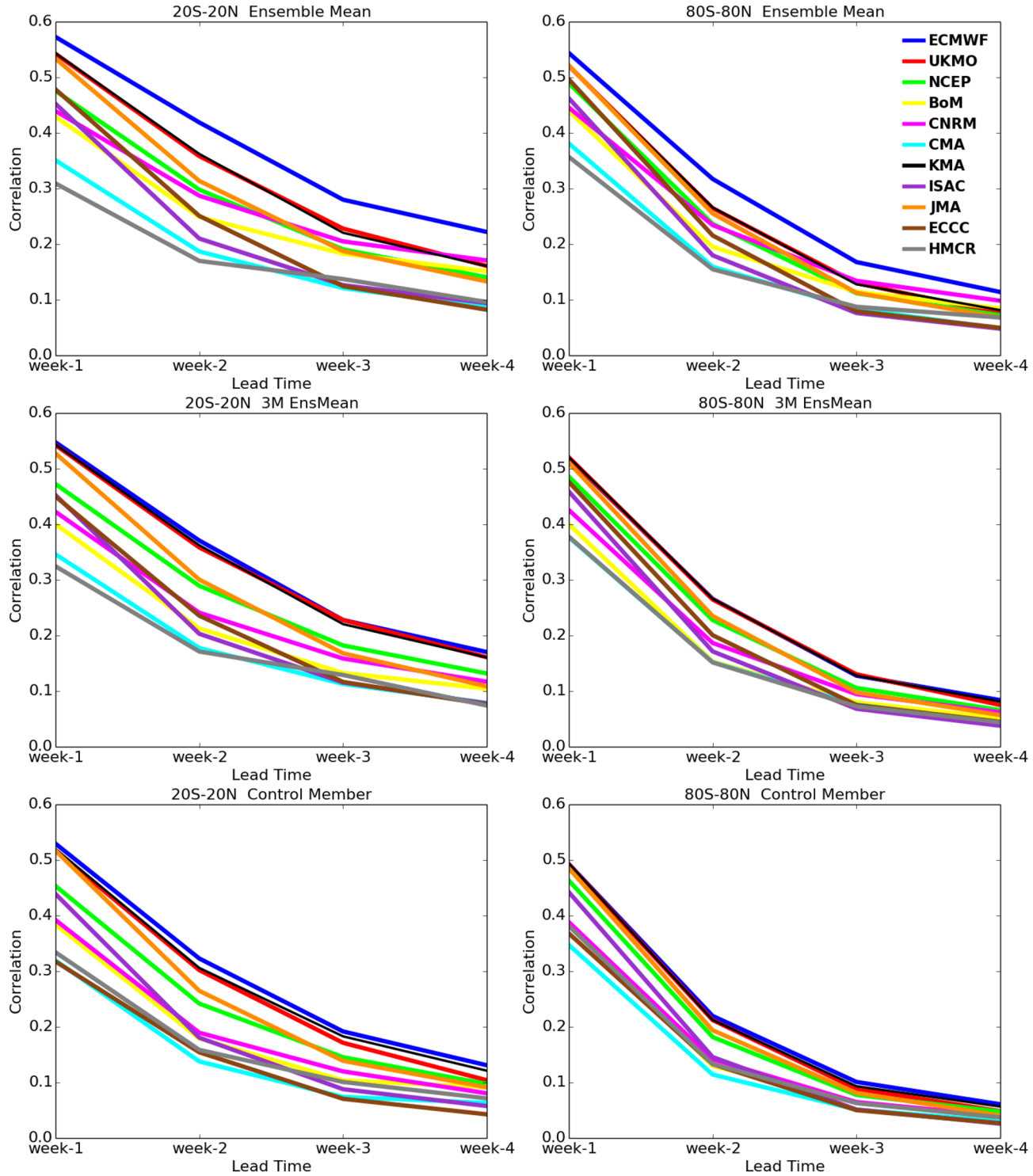


Fig. 6 Zonal average of correlation between hindcast and observed (GPCP) accumulated precipitation anomalies for different latitudinal bands [20°S–20°N (left column) and 80°S–80°N (right column)] during weeks 1–4 (lead time) for hindcasts initialized from November to March over the 1999–2009 period considering the ensemble mean (upper row), three members ensemble mean (middle row), and control member (lower row)

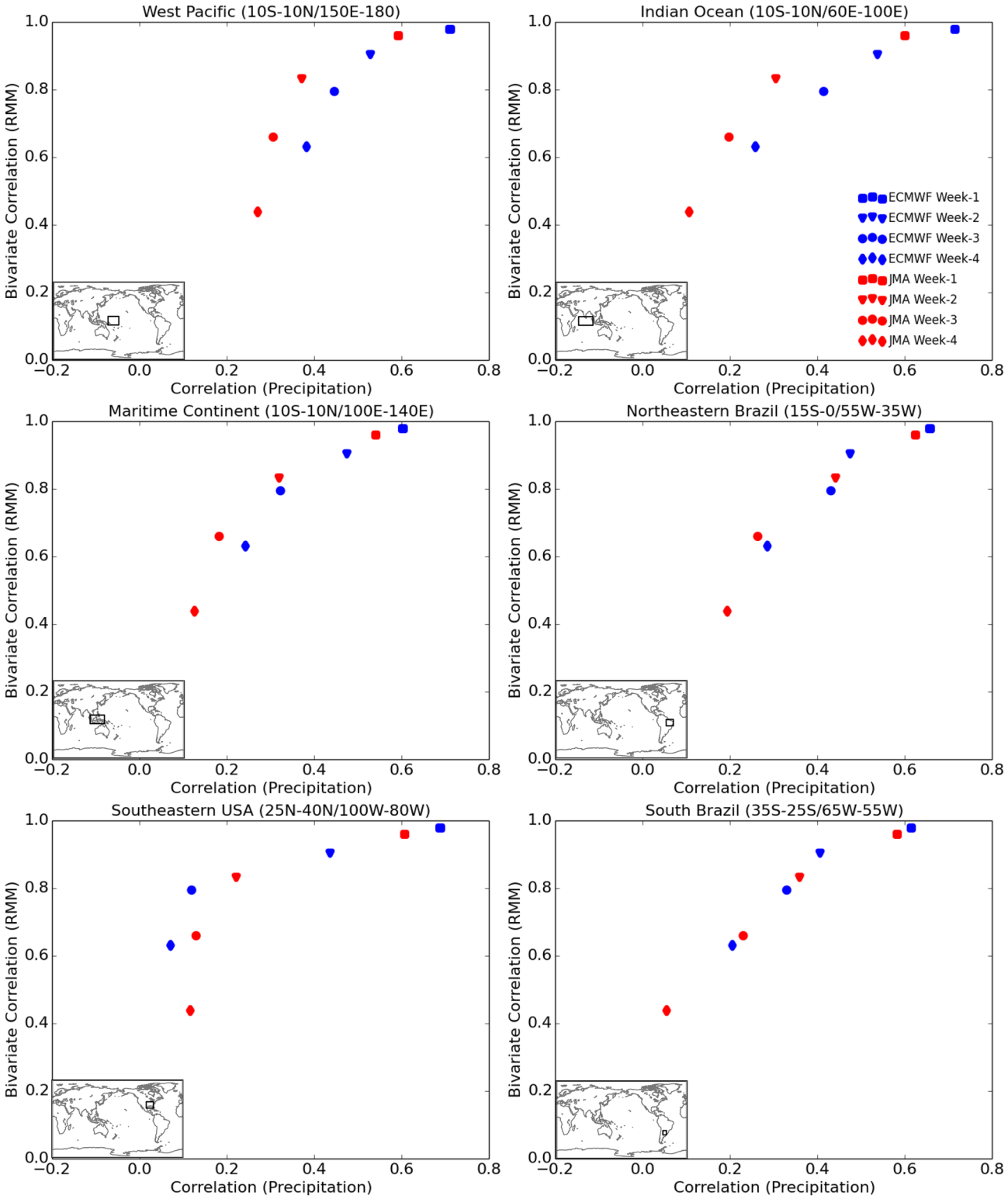
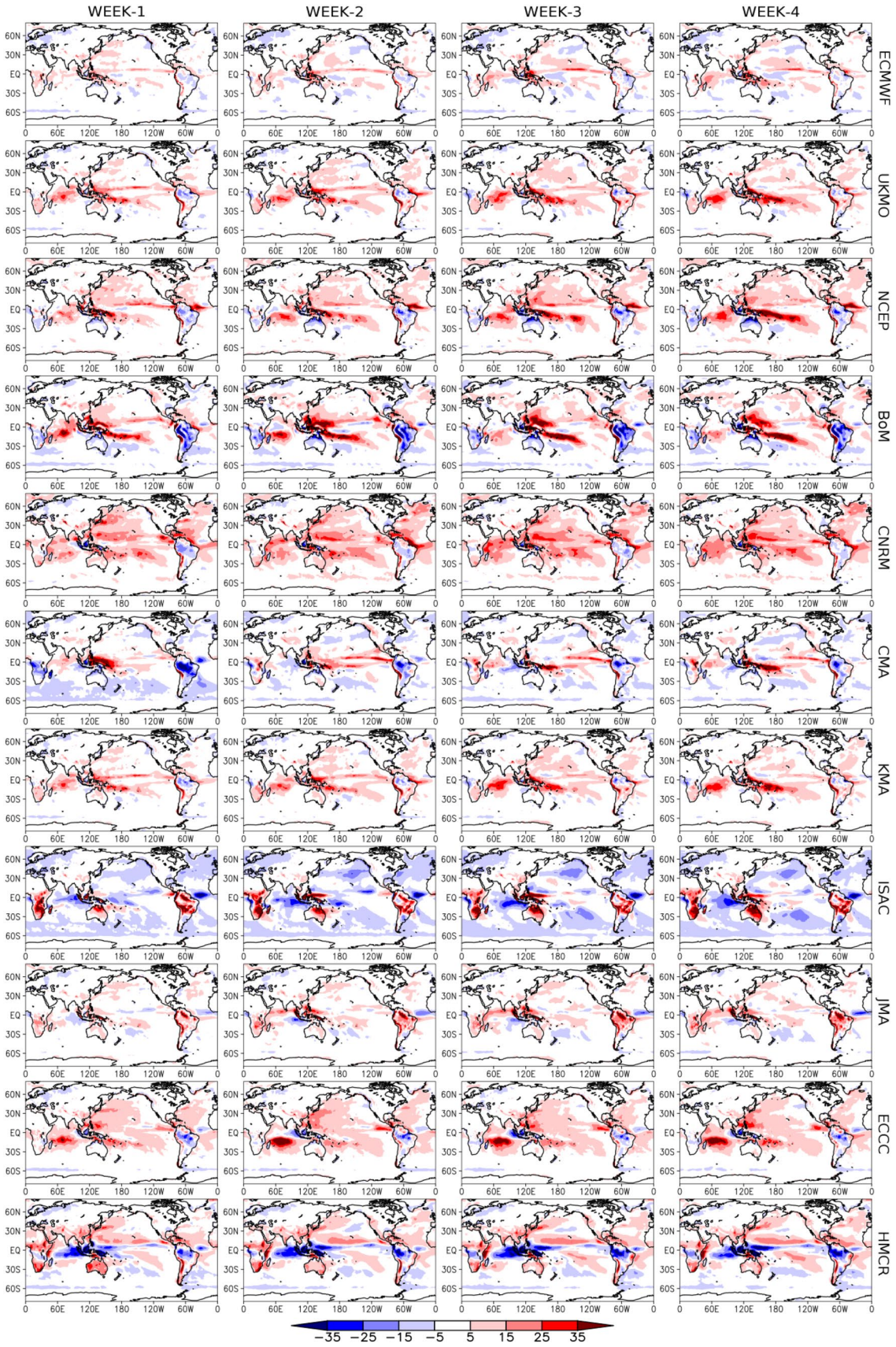


Fig. 7 Scatter plots between average precipitation correlation scores (GPCP×hindcasts) and MJO (RMM index) bivariate correlation for different spatial domains (shown in the top legend and bottom thumbnail image of each plotting) during weeks 1–4 (marker: square for week-1; triangle down for week-2; circle for week-3 and diamond for week-4) for hindcasts initialized from November to March over the 1999–2009 period considering the ensemble mean of ECMWF (blue marker) and JMA (red marker) models


WEEK-1
WEEK-2
WEEK-3
WEEK-4
ECMWF
UKMO
NCEP
BoM
CNRM
CMA
KMA
ISAC
JMA
ECCC
HMCR
60N
30N
EQ
30S
60S
0
60E
120E
180
120W
60W
−35 −25 −15 −5 5 15 25 35

◂**Fig. 8** Mean error (bias) of the ensemble mean accumulated precipitation totals for each S2S model (rows) during weeks 1–4 (columns) for hindcasts initialized from November to March over the 1999–2009 period. The bias was computed using GPCP data as observational reference. Unit: accumulated millimeter per week

this feature is less pronounced after week-2, but still shows better prediction ability association than JMA. Thus, the ocean–atmosphere coupling likely has an important contribution for providing better subseasonal MJO and precipitation prediction ability, particularly on the tropical region.

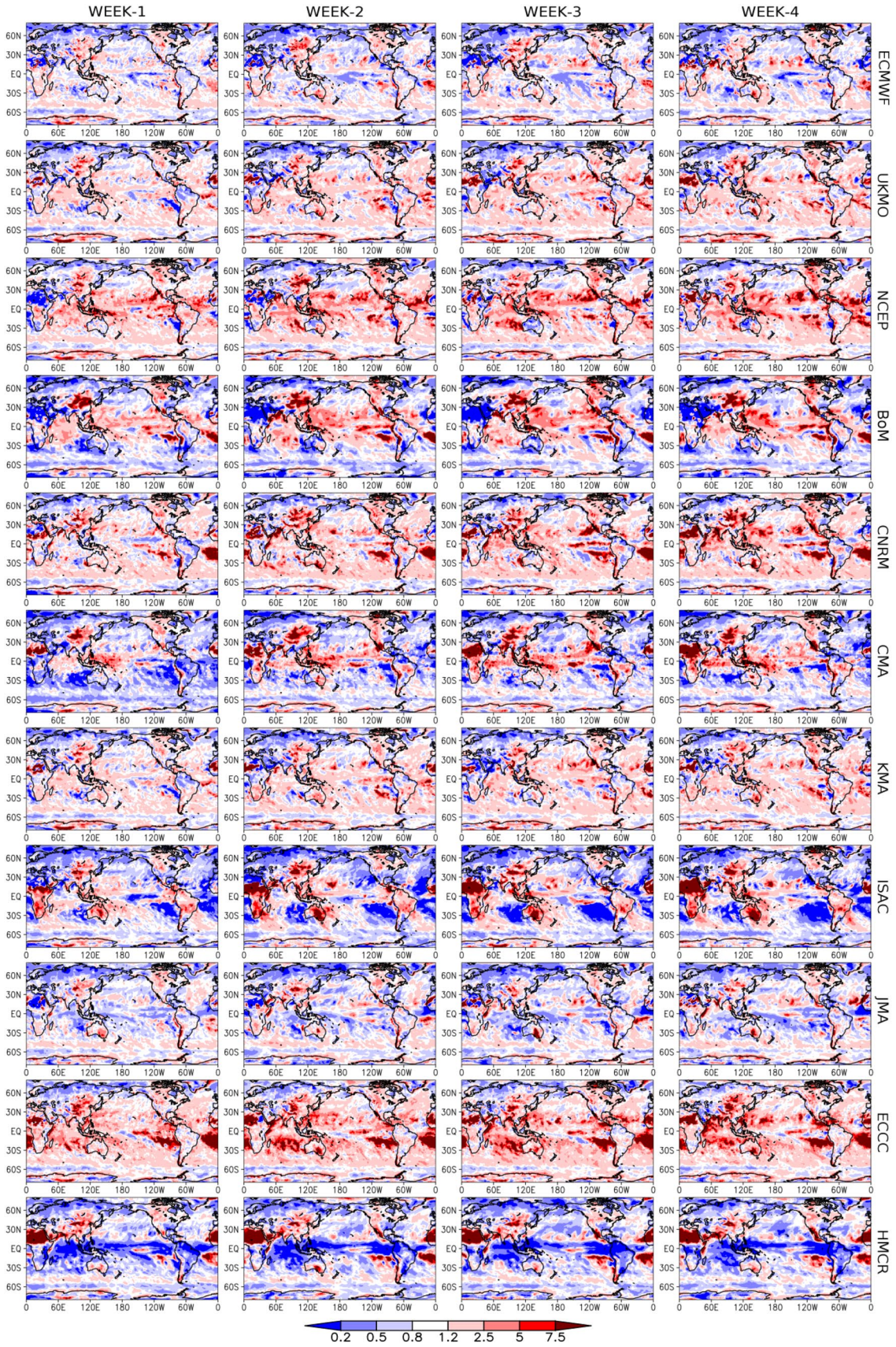
3.2 Systematic error assessment

The bias of the hindcast ensemble mean accumulated precipitation totals was evaluated for each S2S model (Fig. 8; Online Resource 1—Fig. 8). The bias is a measure of forecast accuracy that summarises forecast deficiencies, such as overestimation or underestimation not revealed by the correlation coefficient. Globally, almost all models have a gradual bias increase from week-1 to week-4 in agreement with the findings of Liang and Lin (2018) that also noted similar characteristics over East Asia using the ECCC model. Figure 8 (Online Resource 1—Fig. 8) shows that the tropical region has the largest biases, suggesting that these mean errors are likely related to model deficiencies in representing convective precipitation. Another discernible aspect is that almost all models display larger positive biases (i.e., the mean prediction is on average larger than the mean observation) over the tropical oceans and larger negative biases (i.e., the mean prediction is on average smaller than the mean observation) over the continents and/or extratropics than over other regions. These results should not be related to the ensemble size effect because for the other ensemble size sampling (control member and 3M ensemble mean) the bias spatial pattern is overall similar to the bias spatial pattern of the ensemble mean, although slightly stronger signals are noticed over some tropical regions when considering the same ensemble size for all models in the bias assessment (not shown). Additionally, larger biases are noticed on the extended summer hemisphere, possibly owing to stronger convective activity (compare Fig. 8 with Online Resource 1—Fig. 8). Large positive biases are also noticed in the Eastern Hemisphere, particularly on the warm pool region, including the South Pacific Convergence Zone (Matthews 2012) during November–March (Fig. 8) and the Asian monsoon region during May–September (Online Resource 1—Fig. 8). Jie et al. (2017) also identified large precipitation bias over south/southeast Asia, tropical Indian Ocean, Maritime Continent, and tropical west Pacific Ocean when analyzing 10 S2S models (except KMA). Moreover, the HMCR model shows negative biases around the Maritime Continent, differing from almost all other models. Thus, all the aforementioned features suggest the possibility of deficient representation of tropical convection and the associated diabatic heating, leading to impacts on the subseasonal global precipitation hindcast quality. All these deficiencies are likely related to misrepresentation of atmospheric teleconnection patterns in models predictions.

Systematic errors can be further explored by examining how the predicted variance compares with the observed variance by computing the ratio of these quantities. As the ensemble mean of different perturbed members is known to reduce the variance (not shown), Fig. 9 (Online Resource 1—Fig. 9) displays the variance ratio assessment using the control member of weekly accumulated precipitation anomalies for each S2S model. Similar to bias (Fig. 8; Online Resource 1—Fig. 8), almost all models have a gradual variability error increasing from week-1 to week-4. Some models (such as ECMWF, UKMO, KMA, and JMA) have the predicted variance close to the observed variance (i.e., a variance ratio close to 1) over a large portion of the global domain, especially during week-1. In contrast, CMA, ISAC, and HMCR models show one of the largest variance deficiencies among the S2S models. The latter model also has reversed variance ratio characteristics over some regions, such as the Maritime Continent and the equatorial Pacific Ocean. Moreover, one can also notice large variance underestimation (i.e., the predicted variance is smaller than the observed variance) in the extratropical region and/or continental areas, and variance overestimation (i.e., the predicted variance is larger than the observed variance) over the tropical oceans, including dry zones over eastern Atlantic and eastern Pacific, North Africa, and east/southeast Asia near the Himalaya. These findings may be different for some models showing reversed or even miscellaneous signals over certain regions (e.g., compare ISAC model with BoM model).

3.3 Connections with the subseasonal atmospheric circulation hindcast quality

The previous two sections discussed different deterministic metrics used to evaluate subseasonal precipitation hindcasts produced by S2S models. The correlation assessment indicated that the predictive ability is larger for ECMWF, UKMO, and KMA than for the other models. The highest correlation scores of these three models could possibly be due to the fact that they are coupled models run at higher (more refined) spatial resolution. Furthermore, the highest correlation scores were located in the tropical region whereas the lowest scores were located in extratropical latitudes, particularly for forecasts after week-2. Positive biases were generally identified over the tropical oceans and negative biases over the continents and/or extratropics, whilst prediction variance overestimation


WEEK-1
WEEK-2
WEEK-3
WEEK-4
ECMWF
UKMO
NCEP
BoM
CNRM
CMA
KMA
ISAC
JMA
ECCC
HMCR
60N
30N
EQ
30S
60S
0
60E
120E
180
120W
60W
0.2
0.5
0.8
1.2
2.5
5
7.5

◂**Fig. 9** Variance ratio of the control member accumulated precipitation anomalies for each S2S model (rows) during weeks 1–4 (columns) for hindcasts initialized from November to March over the 1999–2009 period. The variance ratio was computed using GPCP data as observational reference

(underestimation) was identified on the tropical oceans (extratropical region and/or continental areas). Therefore, these results suggest that the subseasonal forecast quality should be further explored through an analysis of the atmospheric circulation, with the aim of investigating S2S models ability in representing equatorial wave dynamics and tropical-extratropical interaction, which may be related to global precipitation hindcast quality.

Figures 10 and 11 (Online Resource 1—Figs. 10, 11) display weekly 200 hPa ZSPSI and ZAPSI prediction quality metrics for the ensemble mean (linear correlation and bias) and control member (variance ratio) of 3 S2S models (ECMWF, BoM, and JMA). The 11 S2S models were analyzed but, for brevity, only the results of the 3 selected models are discussed to illustrate particularly the role of ocean–atmosphere coupling and spatial resolution. The sensitivity level regarding reanalysis choice was analyzed and the main difference was verified in the bias assessment with larger signals when using NCEP-DOE reanalysis 2, particularly for ZSPSI (not shown). Thus, only the results using ERA-Interim dataset are presented. In accordance with the previously examined precipitation correlation patterns, there are also higher correlation scores (Figs. 10, 11; Online Resource 1—Figs. 10, 11—item A) during the first 2 weeks, maintaining higher and meaningful signals mainly over the tropical region in the subsequent weeks. This feature was also identified by Liang and Lin (2018) in both real-time forecasts and hindcasts 500 hPa geopotential height from the ECCC prediction system. The correlation scores in week-2 are generally large in the Northern Hemisphere during November–March (Figs. 10, 11—item A), suggesting a possible association with precipitation through teleconnections, such as the PNA pattern (e.g., Zhu et al. 2014). A similar feature is also identified in the Southern Hemisphere for hindcasts produced during May–September (Online Resource 1—Figs. 10, 11—item A), which highlights for some models large correlation scores on the Pacific-South America sector likely associated with the PSA pattern. For weeks 3–4, there is correspondence between the subseasonal precipitation and atmospheric circulation hindcast quality over the tropical region. This result indicates the ability of S2S models, particularly the coupled models with finer spatial resolution (ECMWF), to represent the connections between tropical rainfall and equatorially-trapped wave dispersion, such as Kelvin and Rossby baroclinic waves (e.g., Kiladis et al. 2009). All of these atmospheric circulation aspects are more pronounced when removing the zonal mean of stream function (ZAPSI), particularly over the extratropical region.

During the first 2 weeks, model resolution seems to have a relevant contribution to hindcast quality. For example, while the JMA model performs better than the low resolution BoM model, it performs worse than the ECMWF model, which has higher spatial resolution than JMA and BoM models. According to Vitart (2017), higher resolution S2S models tend to produce more realistic teleconnections, at least until week-2. In contrast, in weeks 3–4 the best performing models are the coupled models independently of spatial resolution.The analysis for the control member or the 3M ensemble mean is very similar to the ensemble mean analysis for week-1, but is considerably weaker in the subsequent weeks, especially when using the control member over the extratropical region for models with large ensemble size such as the BoM model (not shown). This result suggests the importance of using ensemble predictions for improving subseasonal atmospheric circulation hindcast quality.

As depicted by the correlation scores, the best subseasonal atmospheric circulation hindcast quality after week-2 is located over the tropical region likely due to the medium-range predictability barrier that exists in extratropical weather forecasting. Items B and C in Figs. 10 and 11 (Online Resource 1—Figs. 10, 11) show atmospheric circulation hindcast systematic errors for ECMWF, BoM, and JMA models. As for the mean errors (biases), again it is noteworthy that model resolution has a great impact on hindcast quality, with higher resolution models (ECMWF and JMA) generally showing lower mean errors even for uncoupled models such as the JMA model. This is an intriguing feature which deserves additional investigation.

Item B of Figs. 10 and 11 (Online Resource 1—Figs. 10, 11) shows that the most pronounced mean errors generally increase over the tropical region after week-1, and reach the extratropics after week-2. The latter is more pronounced in the extended winter hemisphere (compare Figs. 10, 11 with Online Resource 1—Figs. 10, 11—item B) for the lowest spatial resolution model (i.e., BoM). Furthermore, the largest extratropical errors resemble the PNA pattern during the extended boreal winter (Figs. 10, 11—item B) and the PSA pattern during the extended boreal summer (Online Resource 1—Figs. 10, 11—item B), when the teleconnections are more pronounced due to more favorable basic state configurations (e.g., Hoskins and Ambrizzi 1993; Ambrizzi et al. 1995; Grimm and Ambrizzi 2009). These results are more evident for ZAPSI and are consistent with the global atmospheric circulation response to tropical diabatic heating, which establishes within a week for convectively coupled equatorial waves dispersion and in an additional week for extratropical Rossby waves propagation (e.g., Jin and Hoskins 1995; Seo and Son 2012). These dynamic patterns are most likely related to tropical precipitation biases

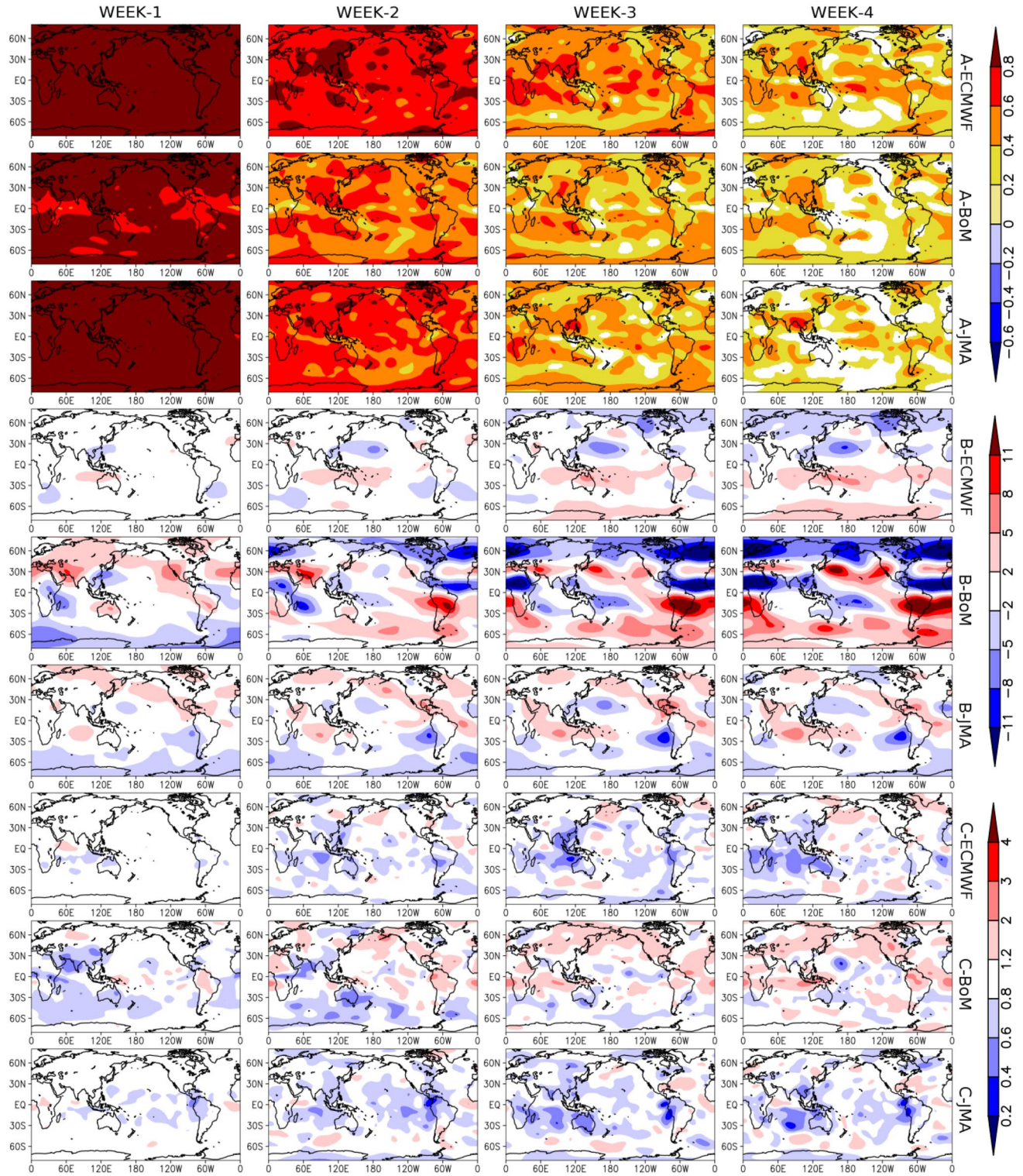


Fig. 10 200 hPa ZSPSI prediction quality metrics for the ensemble mean [rows: item A—correlation of anomalies; item B—bias of totals (unit: 10^6 m^2/s)] and control member (item C—variance ratio of anomalies) for three S2S models (ECMWF, BoM, and JMA) with respect to ERA-Interim reanalysis during weeks 1–4 (columns) for hindcasts initialized from November to March over the 1999–2009 period. Correlation coefficients statistically significant at the 5% level are shaded

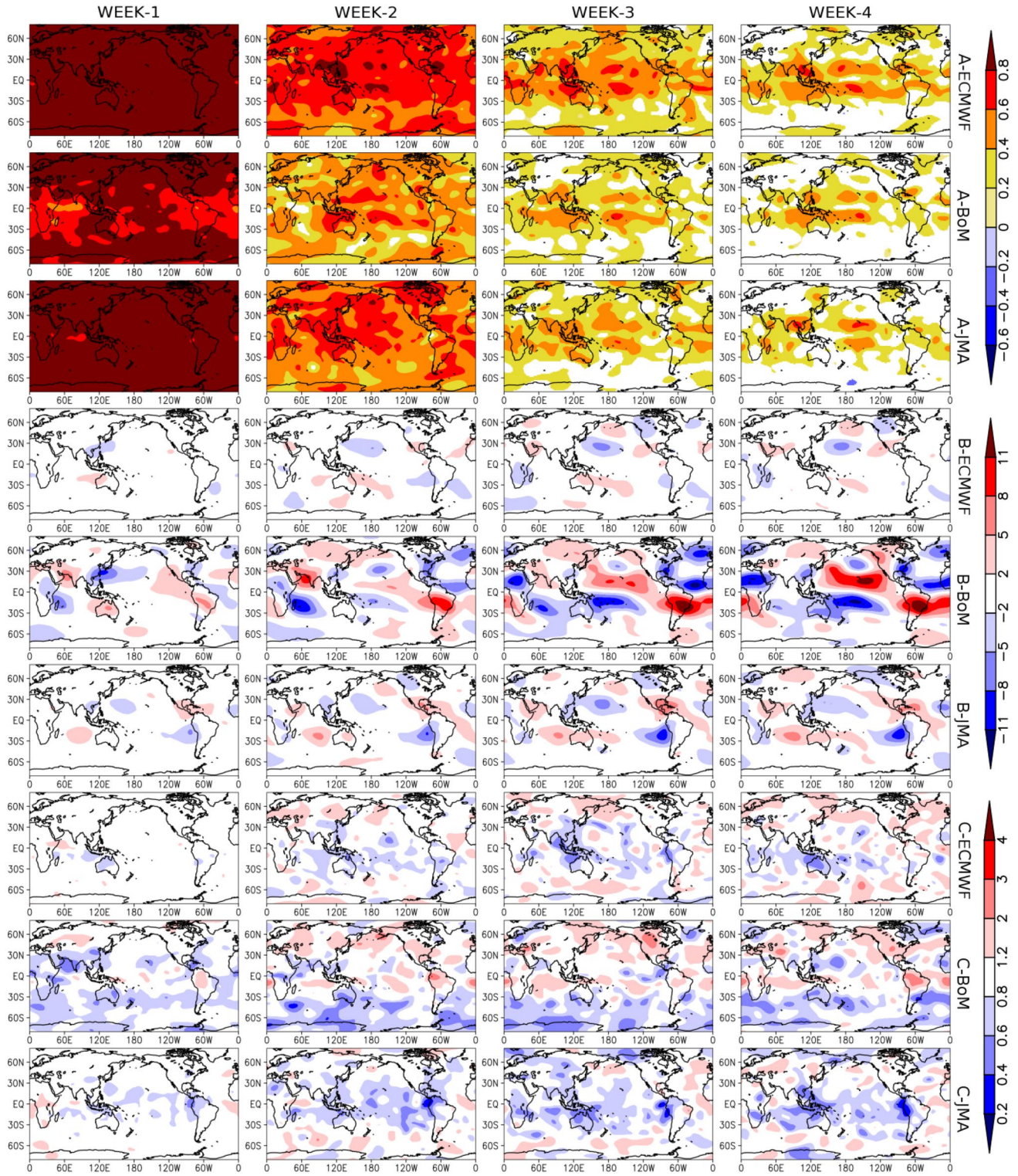


Fig. 11 Same as Fig. 10, except for 200 hPa ZAPSI

(Fig. 8; Online Resource 1—Fig. 8), which disturb latent heat release and consequently the Hadley and Walker circulations that may be responsible for triggering the teleconnections (e.g., Grimm and Ambrizzi 2009). Corresponding to the previously discussed precipitation biases, the 200 hPa ZSPSI and ZAPSI mean biases of the control member and the 3M ensemble mean are overall similar to those of the all ensemble members, but the two smallest ensemble size

samplings generally have slightly larger errors (not shown). The 200 hPa ZSPSI and ZAPSI anomalies variance ratio are approximately similar showing larger variability deficiencies after week-1 especially for BoM model (Figs. 10, 11; Online Resource 1—Figs. 10, 11—item C), which can be associated with precipitation variance ratio (Fig. 9; Online Resource 1—Fig. 9).

4 Summary and discussion

The subseasonal global precipitation hindcast quality was evaluated using all models participating in the WWRP/WCRP S2S prediction project. This unprecedented comparative analysis provided a unique opportunity for improving the current knowledge about the ability of these models in representing precipitation variability at the weekly time scale and the mechanisms behind the sources of subseasonal predictability, elucidating possible shortcomings in S2S ensemble prediction systems. Different deterministic forecast quality metrics (linear correlation, bias, and variance ratio) were employed for verifying precipitation hindcasts anomalies in 4 consecutive weeks (weeks 1–4) during two contrasting extended seasons (November–March and May–September) over the 1999–2009 common period. The quality of the ensemble mean (computed by considering all ensemble size of each model) was contrasted to the quality of the ensemble size sampling using the same number of members for each model [i.e., the single control member and three members ensemble mean (referred to as the 3M ensemble mean)] in order to assess the importance of ensemble prediction.

All models showed higher correlation in week-1, with rapid scores decrease in the following weeks, maintaining meaningful scores mainly in the tropical region. This hindcast quality in the tropics was identified to be related to ENSO and the MJO by removing their linearly regressed precipitation patterns from predicted fields. Both phenomena have been found to be providing potential information for predicting subseasonal precipitation with S2S models, particularly during November–March over the Maritime Continent, the northeastern/south Brazil, and the eastern tropical Indian and central equatorial Pacific Oceans, which are regions influenced by ENSO and MJO activities. For weeks 3–4, the extratropical correlation scores were found to be low, especially on the continental areas, likely due to the inherent unpredictability of the extratropical variability, and to model deficiencies in simulating tropical-extratropical interactions (e.g., PNA and PSA teleconnection patterns) and land surface processes. It was noted that when the control member or the 3M ensemble mean were used the correlation scores were considerably reduced for models with a larger number of ensemble members (ECMWF, BoM and CNRM), demonstrating the relevance of using perturbed ensemble members for improving hindcast quality. The model's rank was outlined through a hindcast quality analysis in different latitudinal bands by computing the zonal average of anomaly correlation scores. In general, the best scores were found over the tropical region and the ranking was roughly unchanged when compared to the global domain, indication that the ECMWF model has the highest scores in all weeks followed by UKMO and KMA models, especially for the ensemble mean. This result indicates that some models were benefited from being coupled and having refined spatial resolution, although some uncoupled models (JMA and ECCC) also presented good ranking performance (i.e., higher scores than lower resolution coupled models, such as BoM) at least until week-2. On the other hand, the lowest correlation scores were shared by CMA, ISAC, and HMCR models during almost all weeks, suggesting that these models have larger deficiencies than the other S2S models here investigated. The contribution of ocean–atmosphere coupling was assessed by examining the relationship between MJO and precipitation hindcast quality for coupled and uncoupled models with approximately similar spatial resolution (ECMWF and JMA). This relationship was found to be stronger for ECMWF model, indicating the importance of coupling for producing improved quality subseasonal MJO and precipitation hindcasts, particularly on the tropical region.

Systematic errors were also analyzed, elucidating deficiencies in the S2S prediction systems. The bias assessment revealed a gradual increase of this measure from week-1 to week-4, with the largest values concentrated over the tropics, possibly due to model deficiencies in representing convective precipitation. This hypothesis is plausible because the ECMWF model showed one of the lowest tropical biases likely due to model physics improvements implemented over the past years, especially in convective parameterization as highlighted in Vitart (2017). Nonetheless, additional analysis is essential to precisely identify which aspect of the ECMWF prediction system (e.g., initialization, model physics, model resolution, ensemble generation method, and ensemble size) contributed for its improved performance. Another remarkable feature presented by most models was the manifestation of large positive biases over the tropical oceans and large negative biases over the continents and/or extratropics. Furthermore, the performed variance ratio assessment provided a new insight about possible failures in a given S2S model in representing subseasonal (weekly) precipitation variability. In general, this assessment indicated overestimation (underestimation) of the variance of the control member in the tropical (extratropical and/or continental) regions.

As subseasonal precipitation hindcast quality is expected to be related to the models ability in representing

atmospheric circulation, we also investigated how S2S models represented the equatorial wave dynamics and tropical-extratropical interactions. Hindcasts of zonally symmetric (ZSPSI) and asymmetric (ZAPSI) streamfunction at 200 hPa were analyzed for three models (ECMWF, BoM, and JMA) for assessing the role played by ocean–atmosphere coupling and model resolution in simulating global atmospheric circulation. As for precipitation, for all models the circulation correlation scores were found to be higher during the first week, dropping in most regions in the subsequent weeks, maintaining meaningful signals mainly over the tropical region. During the first 2 weeks, the model resolution showed to have a more relevant contribution to hindcast quality, with higher resolution models (ECMWF and JMA) producing more realistic teleconnections. On the other hand, for weeks 3–4 the coupled models (ECMWF and BoM) showed the highest correlation scores independently of spatial resolution. For almost all models, the circulation bias increased over the tropical region after week-1 and gradually reached the extratropics after week-2, in particular for the lowest resolution model (BoM) during the extended winter hemisphere when the teleconnections are generally more pronounced. It is hyphothesized that these dynamic patterns are related to the identified systematic errors (bias and variance ratio) of tropical precipitation that may affect both tropical and extratropical atmospheric circulation.

This study was restricted to a deterministic hindcast quality assessment due to the limited number of available perturbed ensemble members of some S2S models. In the future, when more S2S models start producing hindcasts with a large number of perturbed ensemble members it will be possible to perform a similar comparative assessment as performed here using a probabilistic approach. Besides, multi-model ensemble prediction has been recognized as a more robust procedure than the single model ensemble prediction (e.g., Hagedorn et al. 2005). Again, when S2S models hindcasts start to be produced in a more harmonized way, particularly in terms of consistency between initialization dates, it will be possible to start developing procedures for combining and calibrating subseasonal forecast information from a large number of models. An initial step in this direction has recently been given by Vigaud et al. (2017a, b) using a selection of 3 S2S models. Moreover, a more detailed assessment including the combined ENSO-MJO effect on subseasonal precipitation hindcast quality should help improve the current understanding of subseasonal predictability mechanisms as ‘‘windows of opportunity’’ (i.e., periods of potential increase in hindcast quality) can exist as a function of ENSO and the MJO activity (Li and Robertson 2015; Liang and Lin 2018). Finally, an evaluation regarding the contribution of other climate drivers (such as the Indian Ocean Dipole and the Atlantic Meridional Dipole) and additional sources of subseasonal predictability (such as land–atmosphere coupling and stratosphere-troposphere interaction) to S2S models prediction skill is left for future work.

Acknowledgements We thank the two anonymous reviewers for their constructive comments that were helpful in improving the overall quality of the manuscript. The first author was supported by São Paulo Research Foundation (FAPESP), Grant #2016/18156-5. We also acknowledge the support provided by FAPESP CLIMAX project (2015/50687-8). CASC was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (Process: 304586/2016-1). IFAC thanks CNPq (Process: 308451/2014-7) for research support. We thank NCAR (GPCP precipitation: https://rda.ucar.edu/datasets/ds728.3/; JRA55 reanalysis: https://rda.ucar.edu/datasets/ds628.0/), ECMWF (Hindcasts from S2S database: http://apps.ecmwf.int/datasets/data/s2s; ERA-Interim reanalysis: http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=pl/), and NOAA (OISST.v2: ftp://ftp.cdc.noaa.gov/Datasets/noaa.oisst.v2.highres/; NCEP-DOE reanalysis 2: ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis2/pressure/; OLR: https://www.esrl.noaa.gov/psd/data/gridded/data.interp_OLR.html) for making available the dataset used in this study. This work is based on S2S data. S2S is a joint initiative of the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP). The original S2S database is hosted at ECMWF as an extension of the TIGGE database.

Funding The first author was supported by São Paulo Research Foundation (FAPESP), Grant #2016/18156-5. CASC (Process: 304586/2016-1) and IFAC (Process: 308451/2014-7) were supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

Ambrizzi T, Hoskins BJ, Hsu HH (1995) Rossby wave propagation and teleconnection patterns in the austral winter. J Atmos Sci 52:3661–3672

Ardilouze C, Batté L, Déqué M (2017) Subseasonal-to-seasonal (S2S) forecasts with CNRM-CM: a case study on the July 2015 West-European heat wave. Adv Sci Res 14:115–121

Baggett CF, Barnes EA, Maloney ED, Mundhenk BD (2017) Advancing atmospheric river forecasts into subseasonal-to-seasonal time scales. Geophys Res Lett 44:7528–7536

Baldwin MP, Stephenson DB, Thompson DWJ, Dunkerton TJ, Charlton AJ, O'Neill A (2003) Stratospheric memory and extended-range weather forecasts. Science 301:636–640

Barnston AG, Livezey RE (1987) Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. Mon Weather Rev 115:1083–1126

Black J, Johnson NC, Baxter S, Feldstein SB, Harnos DS, L'Heureux ML (2017) The predictors and forecast skill of northern hemisphere teleconnection patterns for lead times of 3–4 weeks. Mon Weather Rev 145:2855–2877

Bolvin DT, Adler RF, Huffman GJ, Nelkin EJ, Poutiainen JP (2009) Comparison of GPCP monthly and daily precipitation estimates with high-latitude gauge observations. J Appl Meteorol Climatol 48:1843–1857

Branstator G (1983) Horizontal energy propagation in a barotropic atmosphere with meridional and zonal structure. J Atmos Sci 40:1689–1708

Carvalho LMV, Jones C, Liebmann B (2004) The South Atlantic convergence zone: intensity, form, persistence, and relationships with

intraseasonal to interannual activity and extreme rainfall. J Clim 17:88–108
Chiang JCH, Vimont DJ (2004) Analogous Pacific and Atlantic meridional modes of tropical atmosphere–ocean variability. J Clim 17:4143–4158
Chiang JCH, Kushnir Y, Giannini A (2002) Deconstructing Atlantic Intertropical Convergence Zone variability: influence of the local cross-equatorial sea surface temperature gradient, and remote forcing from the eastern equatorial Pacific. J Geophys Res 107:4004
Clarke AJ (2008) An introduction to the dynamics of El Niño and the Southern Oscillation. Elsevier, Academic Press, Cambridge
Coleman JSM, Rogers JC (2003) Ohio River Valley winter moisture conditions associated with the Pacific–North American teleconnection pattern. J Clim 16:969–981
Cunningham CAC, Cavalcanti IFA (2006) Intraseasonal modes of variability affecting the South Atlantic Convergence Zone. Int J Climatol 26:1165–1180
Dee DP et al (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorol Soc 137:553–597
Drosdowsky W, Chambers LE (2001) Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall. J Clim 14:1677–1687
Ebita A et al (2011) The Japanese 55-year reanalysis "JRA-55": an interim report. SOLA 7:149–152
Garfinkel CI, Schwartz C (2017) MJO-related tropical convection anomalies lead to more accurate stratospheric vortex variability in subseasonal forecast models. Geophys Res Lett 44:10054–10062
Gonzalez PLM, Vera CS (2014) Summer precipitation variability over South America on long and short intraseasonal timescales. Clim Dyn 43:1993–2007
Gottschalck J et al (2010) A framework for assessing operational Madden–Julian Oscillation forecasts: a CLIVAR MJO Working Group project. Bull Am Meteorol Soc 91:1247–1258
Grimm AM, Ambrizzi T (2009) Teleconnections into South America from the tropics and extratropics on interannual and intraseasonal timescales. In: Vimeux F, Sylvestre F, Khodri M (eds) Past climate variability in South America and surrounding regions. Springer, Apeldoorn, pp 159–191
Grimm AM, Reason CJC (2015) Intraseasonal teleconnections between South America and South Africa. J Clim 28:9489–9497
Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept. Tellus A 57(3):219–233
Held IM, Ting M, Wang H (2002) Northern winter stationary waves: Theory and modeling. Journal of climate 15:2125–2144
Horel JD, Wallace JM (1981) Planetary-scale atmospheric phenomena associated with the Southern Oscillation. Mon Weather Rev 109:813–829
Hoskins BJ, Ambrizzi T (1993) Rossby wave propagation on a realistic longitudinally varying flow. J Atmos Sci 50:1661–1671
Hudson D, Marshall AG, Yin Y, Alves O, Hendon HH (2013) Improving intraseasonal prediction with a new ensemble generation strategy. Mon Weather Rev 141:4429–4449
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
Jeong J-H, Linderholm HW, Woo S-H, Folland C, Kim B-M, Kim S-J, Chen D (2013) Impacts of snow initialization on subseasonal forecasts of surface air temperature for the cold season. J Clim 26:1956–1972
Jie W, Vitart F, Wu T, Liu X (2017) Simulations of Asian summer monsoon in the sub-seasonal to seasonal prediction project (S2S) database. Q J R Meteorol Soc 143:2282–2295
Jin F, Hoskins BJ (1995) The direct response to tropical heating in a baroclinic atmosphere. J Atmos Sci 52:307–319
Johnson NC, Collins DC, Feldstein SB, L'Heureux ML, Riddle EE (2014) Skillful wintertime North American temperature forecasts out to 4 weeks based on the state of ENSO and the MJO. Weather Forecast 29:23–38
Jones C, Waliser DE, Lau KM, Stern W (2004) Global occurrences of extreme precipitation and the Madden–Julian oscillation: observations and predictability. J Clim 17:4575–4589
Kalnay E (2003) Atmospheric modeling, data assimilation and predictability. Cambridge University Press, New York
Kanamitsu M et al (2002) NCEP–DOE AMIP-II reanalysis (R2). Bull Am Meteorol Soc 83:1631–1643
Kiladis GN, Wheeler MC, Haertel PT, Straub KH, Roundy PE (2009) Convectively coupled equatorial waves. Rev Geophys. https://doi.org/10.1029/2008RG000266
Koster RD et al (2010) Contribution of land surface initialization to subseasonal forecast skill: first results from a multi-model experiment. Geophys Res Lett. https://doi.org/10.1029/2009GL041677
Kumar A, Chen M, Wang W (2013) Understanding prediction skill of seasonal mean precipitation over the tropics. J Clim 26:5674–5681
Kumar S, Dirmeyer PA, Lawrence DM, DelSole T, Altshuler EL, Cash BA, Fennessy MJ, Guo Z, Kinter JL III, Straus DM (2014) Effects of realistic land surface initializations on subseasonal to seasonal soil moisture and temperature predictability in North America and in changing climate simulated by CCSM4. J Geophys Res Atmos. https://doi.org/10.1002/2014JD022110
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
Liebmann B, Smith CA (1996) Description of a complete (interpolated) outgoing longwave radiation dataset. Bull Am Meteorol Soc 77:1275–1277
Lin H, Wu Z (2011) Contribution of the autumn Tibetan Plateau snow cover to seasonal prediction of North American winter temperature. J Clim 24:2801–2813
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 et al (2017) MJO prediction using the sub-seasonal to seasonal forecast model of Beijing Climate Center. Clim Dyn 48:3283–3307
Livezey RE, Chen WY (1983) Statistical field significance and its determination by Monte Carlo techniques. Mon Weather Rev 111:46–59
Lo F, Hendon HH (2000) Empirical extended-range prediction of the Madden–Julian oscillation. Mon Weather Rev 128:2528–2543
Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141
Malguzzi P, Buzzi A, Drofa O (2011) The meteorological global model GLOBO at the ISAC-CNR of italy assessment of 1.5 year of experimental use for medium-range weather forecasts. Weather Forecast 26:1045–1055
Marengo JA, Bernasconi M (2015) Regional differences in aridity/drought conditions over Northeast Brazil: present state and future projections. Clim Change 129:103–115
Marengo JA, Hastenrath S (1993) Case studies of extreme climatic events in the Amazon basin. J Clim 6:617–627
Matthews AJ (2008) Primary and successive events in the Madden–Julian oscillation. Q J R Meteorol Soc 134:439–453
Matthews AJ (2012) A multiscale framework for the origin and variability of the South Pacific Convergence Zone. Q J R Meteorol Soc 138:1165–1178

Mo KC, Nogues-Paegle J (2005) Pan-America. In: Lau WKM, Waliser DE (eds) WaliserIntraseasonal variability in the atmosphere–ocean climate system. Springer praxis books (environmental sciences). Springer, Berlin. https://doi.org/10.1007/3-540-27250-X_4

Mo KC, Paegle JN (2001) The Pacific–South American modes and their downstream effects. Int J Climatol 21:1211–1229

Noh Y-C, Sohn B-J, Kim Y, Joo S, Bell W (2016) Evaluation of temperature and humidity profiles of unified model and ECMWF analyses using GRUAN radiosonde observations. Atmosphere. https://doi.org/10.3390/atmos7070094

Olaniyan E, Adefisan EA, Oni F, Afiesimama E, Balogun AA, Lawal KA (2018) Evaluation of the ECMWF sub-seasonal to seasonal precipitation forecasts during the peak of West Africa monsoon in Nigeria. Front Environ Sci 6:4

Osman M, Alvarez MS (2017) Subseasonal prediction of the heat wave of December 2013 in Southern South America by the POAMA and BCC-CPS models. Clim Dyn. https://doi.org/10.1007/s00382-017-3582-4 **(ISSN:1432-0894)**

Palmer TN, Anderson DLT (1994) The prospects for seasonal forecasting. A review paper. Q J R Meteorol Soc 120:755–793. https://doi.org/10.1002/qj.49712051802

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

Reynolds RW, Smith TM, Liu C, Chelton DB, Casey KS, Schlax MG (2007) Daily high-resolution-blended analyses for sea surface temperature. J Clim 20:5473–5496

Robertson AW, Kumar A, Peña M, Vitart F (2015) Improving and promoting subseasonal to seasonal prediction. Bull Am Meteorol Soc 96:ES49–ES53

Ropelewski CF, Halpert MS (1987) Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. Mon Weather Rev 115:1606–1626

Saji NH, Goswami BN, Vinayachandran PN, Yamagata T (1999) A dipole mode in the tropical Indian Ocean. Nature 401:360

Schneider T, Bischoff T, Haug GH (2014) Migrations and dynamics of the intertropical convergence zone. Nature 513:45–53

Seo K-H, Son S-W (2012) The global atmospheric circulation response to tropical diabatic heating associated with the Madden–Julian oscillation during northern winter. J Atmos Sci 69:79–96

Shukla J (1998) Predictability in the midst of chaos: a scientific basis for climate forecasting. Science 282:728–731

Taschetto AS, Gupta AS, Hendon HH, Ummenhofer CC, England MH (2011) The contribution of Indian Ocean sea surface temperature anomalies on Australian summer rainfall during El Niño events. J Clim 24:3734–3747

Tedeschi RG, Cavalcanti IF, Grimm AM (2013) Influences of two types of ENSO on South American precipitation. Int J Climatol 33:1382–1400

Tomaziello ACN, Carvalho LMV, Gandu AW (2016) Intraseasonal variability of the Atlantic Intertropical Convergence Zone during austral summer and winter. Clim Dyn 47:1717–1733

Trenberth KE (1997) The definition of El Niño. Bull Am Meteorol Soc 78:2771–2777

Vigaud N, Robertson AW, Tippett MK (2017a) Multimodel ensembling of subseasonal precipitation forecasts over North America. Mon Weather Rev 145:3913–3928

Vigaud N, Robertson AW, Tippett MK, Acharya N (2017b) Subseasonal predictability of boreal summer monsoon rainfall from ensemble forecasts. Front Environ Sci. https://doi.org/10.3389/fenvs.2017.00067

Vitart F (2017) Madden–Julian oscillation prediction and teleconnections in the S2S database. Q J R Meteorol Soc 143:2210–2220

Vitart F, Robertson AW, Anderson DLT (2012) Subseasonal to seasonal prediction project: bridging the gap between weather and climate. Bull World Meteorol Organ 61:23–28

Vitart F et al (2017) The subseasonal to seasonal (S2S) prediction project database. Bull Am Meteorol Soc 98:163–173

Wang S, Anichowski A, Tippett MK, Sobel AH (2017) Seasonal noise versus subseasonal signal: forecasts of California precipitation during the unusual winters of 2015–2016 and 2016–2017. Geophys Res Lett 44:9513–9520

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

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 143:374–383

White CJ et al (2017) Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorol Appl 24:315–325

Wilks DS (2006) Statistical methods in the atmospheric sciences. Academic Press, London

Zhang T, Yang S, Jiang X, Dong S (2016) Sub-seasonal prediction of the maritime continent rainfall of wet-dry transitional seasons in the nCEP climate forecast version 2. Atmosphere 7(2):28. https://doi.org/10.3390/atmos7020028

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