Climate Dynamics (2021) 57:2889–2904
https://doi.org/10.1007/s00382-021-05845-2

Calibration and combination of seasonal precipitation forecasts over South America using Ensemble Regression

Marisol Osman[1,2,3] · **Caio A. S. Coelho**[4] · **Carolina S. Vera**[1,2,3]

Received: 9 December 2020 / Accepted: 11 June 2021 / Published online: 23 June 2021

Abstract
Models participating in the North American Multi Model Ensemble project were calibrated and combined to produce reliable precipitation probabilistic forecast over South America. Ensemble Regression method (EREG) was chosen as it is computationally affordable and uses all the information from the ensemble. Two different approaches based on EREG were applied to combine forecasts while different ways to weight the relative contribution of each model to the ensemble were used. All the consolidated forecast obtained were confronted against the simple multi-model ensemble. This work assessed the performance of the predictions initialized in November to forecast the austral summer (December–January–February) for the period 1982–2010 using different probabilistic measures. Results show that the consolidated forecasts produce more skillful forecast than the simple multi-model ensemble, although no major differences were found between the combination and weighting approaches considered. The regions that presented better results are well-known to be impacted by El Niño Southern Oscillation.

Keywords Climate prediction · NMME · Multi-model ensemble

1 Introduction

Seasonal climate predictions are produced by operational centers and internationally coordinated activities worldwide, such as the National Oceanic and Atmospheric Administration (NOAA), the European Center for Medium Range Weather Forecast (ECMWF), the North American Multi-Model Ensemble (NMME), the World Meteorological Organization Lead Centre for Long-Range Forecast Multi-Model Ensemble (WMO LC-LRFMME) and the Asian-Pacific Economic Cooperation Climate Center (APEC). These forecasts are mostly presented in terms of probabilities, as a result of uncertainties arising due to the chaotic nature of the atmosphere and the errors associated to the initial conditions as well as numerical formulation of the dynamical models used.

To address the forecast uncertainty problem, the development of ensemble prediction has been the strategy adopted by most forecast centers (e.g. Buizza 2006). Furthermore, the multi-model ensemble (MME) technique has been widely adopted to account for the uncertainty due to model errors and several studies have documented its advantage over the single model approach (Doblas-Reyes et al. 2005; Hagedorn et al. 2005). In this context, different ways to calibrate and combine forecasts have been applied to aggregate forecasts from different sources. For example, Min et al. (2009) calibrate and combine operational models from APCC using Gaussian fitting and weighting models inversely proportional to the errors associated with the model retrospective errors, improving the performance against the equal weighted multi-model ensemble forecast. On subseasonal timescales, Vigaud et al. (2017) apply Extended Logistic Regression to three different models and combine the resulting calibrated individual

✉ Marisol Osman
osman@cima.fcen.uba.ar

1 Departamento de Ciencias de la Atmósfera y los Océanos, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina

2 CONICET – Universidad de Buenos Aires, Centro de Investigaciones del Mar y la Atmósfera (CIMA), BuenosAires, Argentina

3 CNRS – IRD – CONICET – UBA, Instituto Franco-Argentino para el Estudio del Clima y sus Impactos (IRL3351 IFAECI), Buenos Aires, Argentina

4 Centro de Previsão de Tempo e Estudos Climáticos, Instituto National de Pesquisas Espaciais, São Paulo, Brazil

model probabilistic forecasts with equal weights obtaining more reliable week3–week4 precipitation and temperature forecasts over North America. Extended Logistic Regression is also currently used by the International Research Institute for Climate and Society (IRI) to produce its real time precipitation and temperature seasonal forecasts. In South America, Coelho et al. (2006, 2007) apply a Bayesian approach to combine summer and winter precipitation predictions over South America derived from several dynamical forecast systems. Coelho et al. (2006) used a relatively long set of hindcasts and found that the resulting combined and calibrated forecasts showed improved skill over the equal weighted (multi-model ensemble mean) forecast in terms of the reliability and resolution.

Among the multiple calibration techniques developed for correcting forecast errors, there is Ensemble Regression (EREG, Unger et al. 2009). EREG is easy to implement and has been shown to provide competitive performance in comparison with other calibration techniques, retaining information from the individual ensemble members while obtaining the calibration parameters from the ensemble mean (Unger et al. 2009). Some examples of the application of EREG to a single model are available in the literature (e.g. Unger et al. 2009; Ou et al. 2016) but, to the authors knowledge, its application in the context of MME predictions has only been addressed by Collins (2017) for North America seasonal precipitation forecasts, improving the forecast compared to pooling all models ensemble members together. In this paper we expand the work of Collins (2017) by applying EREG to combine South America seasonal precipitation forecasts produced by eight models.

The main objective of this work is then to assess the performance of calibration and combination approaches based on EREG applied to retrospective multi model ensemble precipitation forecasts over South America. We work with the forecasts produced by the models participating in the North America Multi Model Ensemble project (NMME, Kirtman et al. 2014) and apply EREG. To do this, two different approaches based on EREG are explored (see Sect. 2.3) while the impact of different weighting techniques is also assessed. The paper is organized as follows: Sect. 2 describes the reference and forecast data, introduces the Ensemble Regression technique to calibrate forecasts and describes the combination framework. Section 3 presents the main outcomes while in Sect. 4 the main conclusions are discussed.

2 Data and methodology

2.1 Data

2.1.1 Reference data

Precipitation data from the Climate Prediction Center (CPC) global daily Unified Raingauge Database (URD, Xie et al. 2010) and the CPC Merged Analysis of Precipitation (CMAP, Xie and Arkin 1997) were used as observational references. The CPC-URD database was used for land gridpoints while CMAP was used for ocean gridpoints. The combined dataset was obtained for a $1.0° \times 1.0°$ grid through the IRI Data Library (IRIDL). This data was used to produce the calibrated forecast as well as to verify the retrospective forecasts.

2.1.2 Forecast data

Monthly precipitation forecasts from the North America Multi-Model Ensemble (NMME) project were used (Kirtman et al. 2014). The NMME project consists of a set of coupled models from USA and Canadian modeling centers available to the community through the IRIDL. The NMME project is being used operationally by, for example, the CPC and the IRI. In addition, several studies have assessed the performance of NMME in the context of ENSO predictions (e.g. Tippett et al. 2019; Landman et al. 2019), monthly European temperature and precipitation predictions (Rodrigues et al. 2019),

Table 1 Models from NMME participating in the study

Model	Institution	Atmospheric component	Oceanic component	Ensemble size	Hindcast period
Canadian-CanCM3	Environment Canada	CanAMa T63L31	CanOM4 L40 .94°Eq	10	(1982–2010)
Canadian-CanCM4	Environment Canada	CanAM4 T63L35	CanOM4 L40 .94°Eq	10	(1982–2010)
NCAR-CCSM4	NCAR	CAM4 0.9 x 1.25° L26	POPL60 .25°	10	(1982–2010)
GFDL-CM2p1a	NOAA/GFDL	CM2.1 1x2.5° L24	MOM4 L50 .3°Eq	10	(1982–2010)
GFDL-FLOR-A05	NOAA/GFDL	CM2.5 C18L32	MOM5 L50 .3°Eq	12	(1982–2010)
GDFL-FLOR-B01	NOAA/GFDL	CM2.5 C18L32	MOM5 L50 .3°Eq	12	(1982–2010)
NASA-GEOS5	NASA Goddard Space Flight Center	GEOS5 AGCM 0.5° L72	MOM5 L40 .5°Eq	4	(1982–2017)
NCEP-CFSV2	NOAA/NCEP	GFS T126L64	MOM4 L40 .25°Eq	24	(1982–2010)

and climate extreme predictions (Slater et al. 2019). Table 1 lists the eight models from NMME used in this study along with the available number of ensemble members, lead times and the retrospective (hindcast) period. All model outputs have a resolution of 1° × 1° and forecast lead times of at least up to 9 months. All the models considered were run retrospectively and the hindcast period of 1982–2010 was commonly available.

2.2 Ensemble Regression

We used Ensemble Regression (EREG) to calibrate and combine the forecasts. In this section, we briefly introduce EREG in the context of a single model prediction. The methodology is described in detail in Unger et al. (2009). We choose EREG because it is easy to apply, it is computationally cheap and it uses all the ensemble information. The Ensemble Regression equation is equivalent to linear regression between the ensemble mean and the observation, but is applied to each member of the ensemble to obtain a Probability Density Function (PDF) that represents the prediction of the entire ensemble. EREG retains the ensemble spread to represent conditional uncertainty of forecasts, to the extent that spread is found to be a reliable indicator of the average mean square error of a model's forecast (Collins 2017). Mathematically, the implementation of EREG begins with the analysis of the linear relation between the forecast ensemble mean F_m and the observation, defined as Y:

$$Y = \alpha_0 + \alpha_1 F_m + \epsilon$$

where α_1 and α_2 represent the regression coefficients and ϵ the residuals. The application of the linear regression consists in minimizing the quantity $\langle (F_m - Y)^2 \rangle$, where brackets mean temporal average, to estimate α_s and to obtain the equation $F'_m = \alpha_0 + F_m \alpha_1$, where F'_m is the regression estimation. The coefficients are obtained through:

$$\alpha_1 = R_m \frac{S_Y}{S_m}, \quad \alpha_0 = \langle Y \rangle - \alpha_1 \langle F_m \rangle \tag{1}$$

where S_Y and S_m are the sample standard deviation of Y and F_m, respectively; and R_m is the correlation between Y and F_m.

If the following inequality is satisfied:

$$\frac{T-1}{T-2} S_Y^2 (1 - R_m^2) \geq \alpha_1^2 \langle E^2 \rangle \tag{2}$$

where T is the number of cases (years in our work), S_Y^2 is the observed variance and $\langle E^2 \rangle$ is the mean spread of the model, defined as $\langle E^2 \rangle = \langle \frac{1}{N} \sum_{i=1}^{N} (F_i - F_m)^2 \rangle$ with F_i the forecast of member i and N the number of ensemble members, then the ensemble members satisfy the assumptions needed to apply the linear regression developed for the ensemble mean into each of the ensemble members. The standard deviation of the regression of each individual ensemble member can be obtained through

$$S_\epsilon = S_Y \left[\frac{T-1}{T-2} (1 - R_m^2) \right]^{1/2}. \tag{3}$$

The PDF that represents the N ensemble members takes the form of N kernels with Gaussian distributions, each of them centered in the individual estimation associated to each individual ensemble member and its width determined through Eq. 3. The final PDF is then the simple normalized sum of all the error distributions, each of them representing $1/N$ of the total distribution, as is shown in Fig. 1. If Eq. 2 is not satisfied, EREG is only applied to the ensemble mean of the model. Therefore, EREG adjusts the model PDF, becoming a standard lineal regression on the ensemble mean and collapsing the spread if the information added by individual ensemble members worsens the forecast based on the ensemble mean alone (Unger et al. 2009).

2.3 Calibration and combination techniques

The calibration of each model is a necessary but not a sufficient step if we wish to obtain a unified forecast from the contribution of each multiple member forecast system. To achieve this goal, it is necessary to combine all the forecasts involved into a consolidated forecast. In this work, we applied two different approaches for calibration and combination, both based on EREG. Below, we give a brief explanation of both.

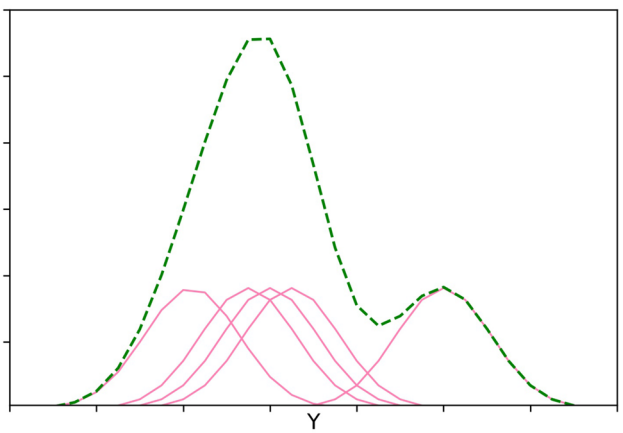

Fig. 1 Example of the consolidated PDF obtained from the application of EREG to an ensemble of five members. The PDF of each member is centered at the regression estimation of that member while the PDF width is common to all members (pink lines). The consolidated PDF is the sum of the Gaussian kernels (green dashed line)

2.3.1 Averaged PDFs (APDFs)

The Averaged PDFs methodology simply consists of obtaining the final consolidated PDF through the combination of the PDF from each model, after calibrating each of them through EREG. In this sense, the regression parameters used in EREG are obtained for each model separately. Therefore, if we consider each ensemble member from each model as a kernel with an associated error in the regression (which determines the width of each kernel), APDFs assumes that the kernel widths are variable between models. Several studies have suggested weighting each model according to their performance prior to the combination (Weigel et al. 2008; Rajagopalan et al. 2002; Robertson et al. 2004). In this work, when the combination is performed, three approaches are used to assess this subject: (a) we equally weight the individual model calibrated kernels to produce a simple averaged consolidated PDF, (b) we weight the individual model calibrated kernels according to the correlation between each model ensemble mean and observations (mean_cor) to produce a weighted consolidated PDF, and (c) we weight the individual model calibrated kernels proportionally to the number of cases when the kernel calibrated PDF of each model showed the highest probability at the observation point with respect to the total number of forecast cases (pdf_int) to produce a weighted averaged consolidated PDF. Mathematically, in the first case the weight that each model receives, w_i, takes the following form:

$$w_i = \frac{1}{M}$$

where M represents the number of Models used. In the second example the weight that each model receives, w_{imean_cor}, takes the following form:

$$w_{imean\_cor} = \frac{R_i}{\sum_{i=1}^{M} R_i}$$

where R_i is the correlation between model i and observations and M is the number of models. If R_i is negative then $w_i = 0$, unless all correlations are negative, in which case all models receive the same weight. In the third case, the weight that each model receives, w_{ipdf_int}, is defined as:

$$w_{ipdf\_int} = \frac{h}{T}$$

where h represents the number of cases the model i presented the highest PDF value at the observation point, and T is the total number of cases (years). Figure 2 presents an example of how this weight is implemented. For each year and gridpoint, we determine which model presents the maximum intensity of its PDF at the observation point. In this case, the magnitude of the PDF obtained through EREG at the observation value is higher for model blue than for model red. We repeat this process for each year to get the weight for each model that represents the percentage of years that each model presented the highest PDF value at the observation point and reflects the probability of each model of being the best according to its historical performance.

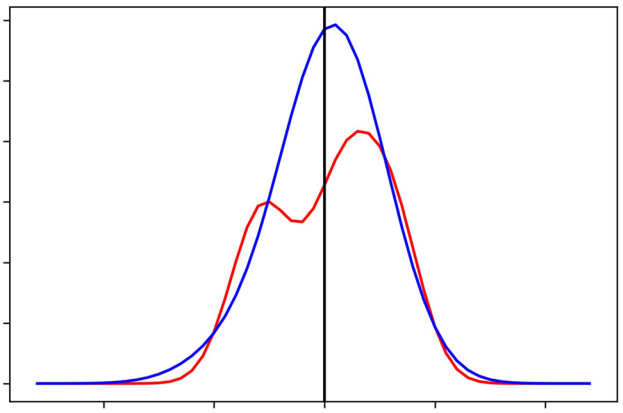

Fig. 2 Illustration of the computation of the weight for PDF_INT for two models (blue and red model). For each year and gridpoint we evaluate the intensity of the calibrated PDF for each model (blue and red line) at the observation value (black vertical line). In the example, the blue model is more intense than the red model at the observation value. The weight each model receives at each gridpoint results from the ratio between the number of years that model presented the most intense PDF at the observation value and the number of years analyzed

Figure 3 shows an example of the implementation of APDFs with the three approaches adopted (panel a, b and c, respectively). In the figure, two models (red model and blue model) with different ensemble sizes are considered. First, each model is calibrated through EREG and the PDF of each model (red and blue dotted lines) is obtained after summing the individual PDF from each ensemble member (thin red and blue lines). The final consolidated PDF (thick green line) results from the combination of the PDFs of each model. From the analysis of the different panels it can be seen that in all the approaches the calibrated PDF of blue and red model are identical. However, in Fig. 3a the final consolidated PDF lies between the PDF of both models whereas in Fig. 3b the blue model receives a greater weight than the red model, because the ensemble mean of blue model presented a greater correlation against observations, and therefore the final consolidated PDF is closer to the calibrated PDF of blue model. Finally, in Fig. 3c the red model receives a greater weight than the blue model, because its calibrated PDF was highest at the observation point in more cases, and then the final consolidated PDF is similar to the calibrated PDF from red model.

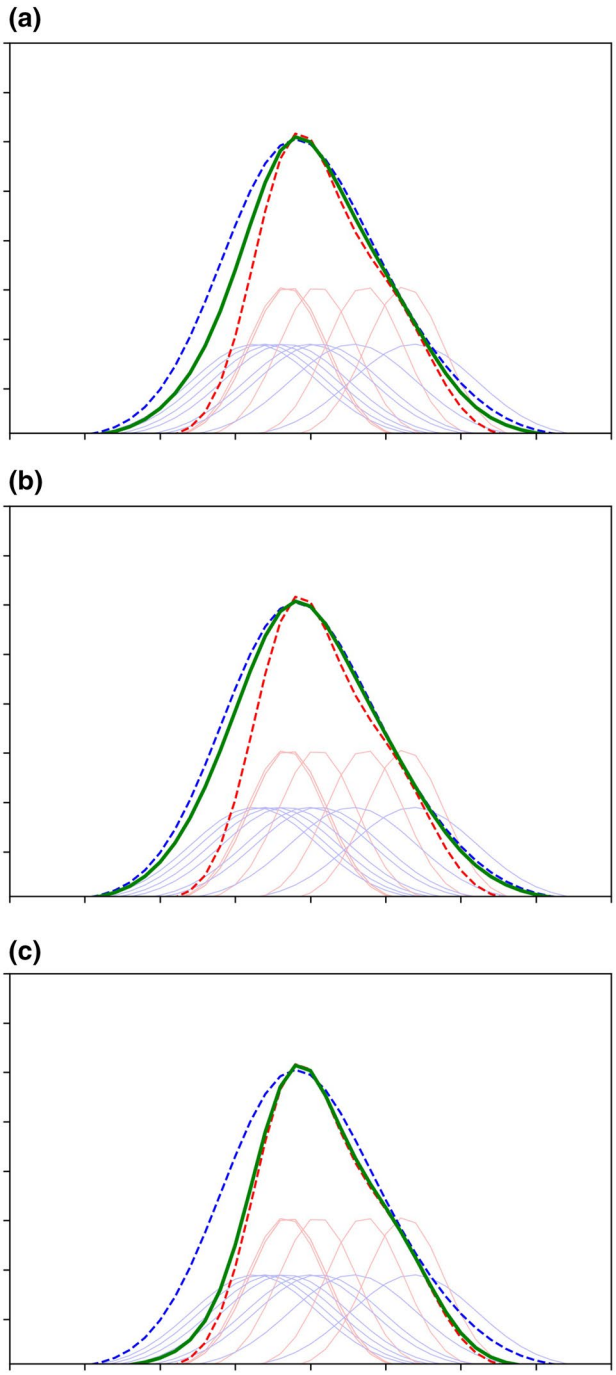

Fig. 3 Illustration of APDFs combination with two models (blue and red model). The thin red and blue lines represent the Gaussian kernels associated to each ensemble member from each model while the dashed blue and red line are the calibrated PDFs of each model obtained through the application of EREG to each model. The final consolidated PDF (thick green line) results from the average of the PDFs from both models. When the averaged is performed, models are weighted differently: **a** both models receive the same weight, **b** the blue model is weighted more heavily than the red model because it has the highest correlation, and in **c** the red model is weighted more heavily than the blue model because it has the largest proportion of cases with forecast PDFs obtained with EREG peaking at the observation point

2.3.2 Multi-model Ensemble Regression (MMEREG)

The multi-model Ensemble Regression technique is a generalization of EREG but for multi-model ensembles. It consists in applying EREG to the entire multi-model ensemble. In this case, the regression parameters are obtained by regressing the observation against the multi-model ensemble mean, which is obtained by averaging all members from all participating models. Therefore, the regression parameters used are common to all models and ensemble members. The consolidated PDF results from the normalized sum of each member or kernel and the associated error is the same for each member (resulting from the errors in the implementation of EREG to the multi-model ensemble as it is done for one model). As was done for APDFs, we use three different approaches for determining the multi-model ensemble mean that is regressed against observations: (a) we equally weight all models for determining the multi-model ensemble mean, (b) we weight models according to the correlation between each model ensemble mean and observations (mean_cor) for determining the multi-model ensemble mean, and (c) we weight models proportionally to the number of cases when the kernel calibrated PDF of each model showed the highest probability at the observation point with respect to the total number of forecast cases (pdf_int) for determining the multi-model ensemble mean.

Figure 4a shows an example of the application of MMEREG to two models (red model and blue model) with different ensemble size and weighting models equally. It can be seen that all the ensemble member PDFs present the same width (thin red and blue lines). The final consolidated PDF (green line) results from the normalized sum of the PDFs of both models. Figure 4b and c are an adaptation of Figure 4a but with models weighted differently. In Fig. 4b, the blue model receives a greater weight than red model because the correlation between the ensemble mean of that model against observations is greater. As a result, the regression parameters are different from those obtained for Fig. 4a, and the position of the kernels, as well as their width and the final consolidated PDF, change. In Fig. 4c red model receives a greater weight than blue model, because its calibrated PDF was highest at the observation point in more cases, and this also modifies the kernels, their width and the consolidated PDF. This method is slightly different from that applied by Collins (2017) in which models are first calibrated through EREG to obtain the calibrated ensemble members for each model and then those calibrated ensemble members are combined by applying EREG a second time to the entire calibrated ensemble to produce a final consolidated PDF.

From the description provided for both methodologies (APDFs and MMEREG) we try to answer whether it is better to apply EREG and obtain the regression parameters to

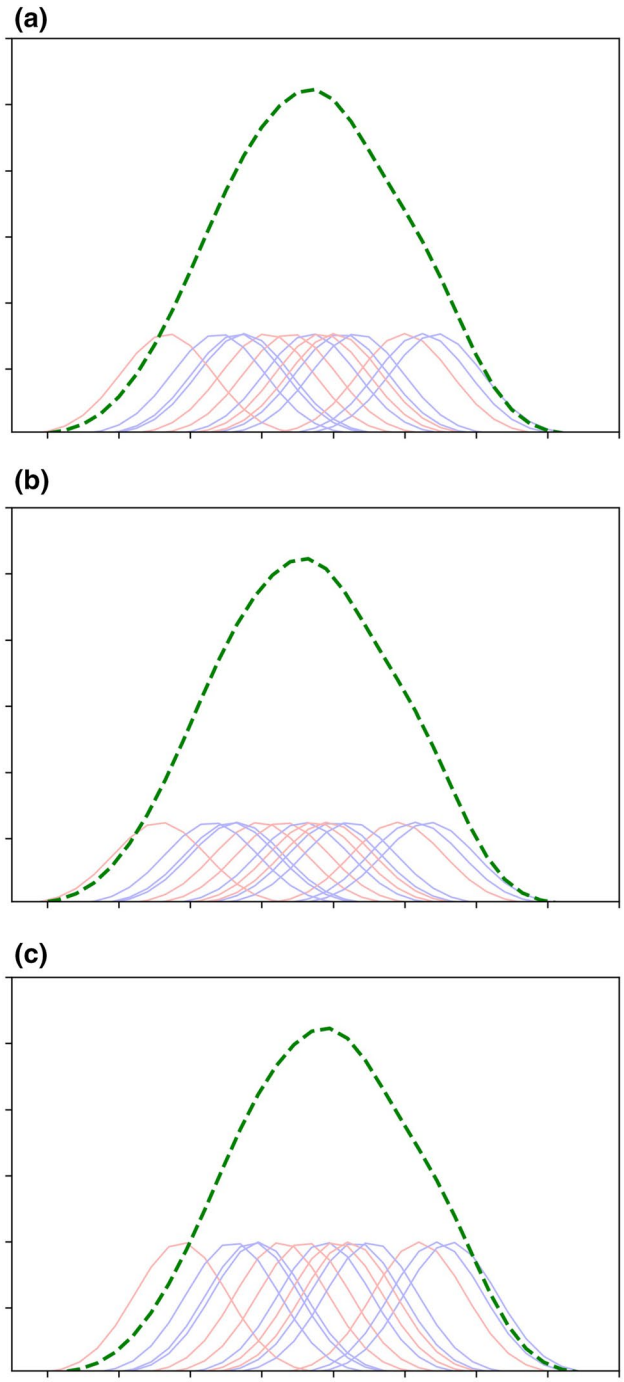

Fig. 4 Illustration of MMEREG combination with two models (blue and red model). Thin blue and red lines represent the PDF of each member. Since MMEREG is applied to the entire MME all the PDFs present the same width and are centered at the regression estimation of each member. The final consolidated PDF (green dashed line) results from the normalized sum of all Gaussian kernels. In each panel, models are weighted differently when the multi-model mean is computed and therefore the widths and the regression estimations are differently in each panel (see text): **a** equal weight to both models, **b** weights are proportional to the correlation, **c** weights are proportional to the cases with PDF peaking at the observational point

each model separately and then combine all models or if it is better to pool all the models together and apply EREG to the entire MME. Most of the currently used techniques first calibrate each model individually and then combine all models through a simple or weighted average (e.g. Min et al. 2009; Vigaud et al. 2017). On the other hand, the Forecast Assimilation technique, applied for instance in Coelho et al. (2006, 2007) and Rodrigues et al. (2019), attempts to calibrate and combine all models at the same time.

2.4 Calibration and combination of seasonal forecasts

The outlined combination techniques were applied to the mean December–January–February (DJF) precipitation forecast made with models initialized in November (lead 1 month) from the NMME project over South America (275° E–330° E; 15° N–60° S) for the period 1982–2010. In all the mentioned steps, the regression coefficients as well as the weights were determined using 1-year-out cross-validation.

We first computed the seasonal mean by simply averaging the forecasts and observations. Then, forecasts and observations were linearly detrended and standardized by dividing the anomalies by the standard deviation. We then determined the weight of each model according to their historical performance, either through mean_cor or pdf_int to use then in the combination procedure. Finally, we combined models either with APDFs or MMEREG to obtained the consolidated PDF associated to each methodology. In total, we obtained 6 different consolidated forecasts. Three of them correspond to the consolidation through APDFs weighting the individually calibrated PDFs obtained through EREG using equal weights, mean_cor or pdf_int. The other three correspond to the consolidation through MMEREG applied to the ensemble weighted either using equal weights, mean_cor or pdf_int. Table 2 summarizes the different approaches studied in this work.

We evaluated the performance of the combined forecasts in forecasting the terciles of the observed precipitation distribution, that is, the equiprobable categories below normal (BN), near normal (NN) and above normal (AN). This evaluation was done through the computation of the Heidke Skill Score (HSS, Wilks 2011), the Ranked Probabilistic Skill Score (RPSS, Epstein 1969), the Brier Skill Score (BSS, Stephenson et al. 2008), the reliability (Hartmann et al. 2002) and the ROC diagrams (Mason and Graham 2002). The HSS assesses discrimination, reliability and resolution of the forecast while the RPSS is a measure of the error of the probabilistic forecasts equivalent to the Mean Square Error but for probabilities. The BSS is a particular case of the RPSS but for each category separately. In this work, the climatological forecast, which assigns to each category

Table 2 Combination and weighting approaches applied

Acronym	Combination technique
APDFs-SAME	Average of PDF of each model obtained through EREG. Final PDF computed given equal weights to each model
APDFs-MEAN_COR	Weighted average of PDF of each model obtained through EREG. PDFs are weighted proportionally to the correlation between the ensemble mean of each model and the corresponding observations
APDFs-PDF_INT	Weighted average of PDF of each model obtained through EREG. PDFs are weighted proportionally to the intensity of the PDF of each model at the observation value
MMEREG-SAME	EREG applied to the MME mean, with the MME mean computed using equal weights for all models
MMEREG-MEAN_COR	EREG applied to the MME mean, with the MME mean computed using weights proportional to the correlation between the ensemble mean of each model and the corresponding observations
MMEREG-PDF_INT	EREG applied to the MME mean, with the MME mean computed using weights proportional to the intensity of the PDF of each model at the observation value

the same chance of occurrence, was used as a reference to obtain the RPSS and BSS. The reliability diagram evaluates the correspondence between the observed frequency of an event and the mean forecast probability of that event whereas the ROC diagram assesses the ability of the forecast in discriminating between events and non-events with different chances of occurrence. We confronted the results obtained with the six consolidated forecasts against that obtained simply by computing the probabilities of each tercile for each model separately and then averaging the probabilities of each tercile across all models (multi-model averaged counting estimate, MME). The tercile probabilities for each model are obtained by counting the number of ensemble members falling in each category.

3 Results

We first describe the characteristics of the NMME system before the calibration and combination were applied. Figure 5 shows the maps of HSS and RPSS and the reliability and ROC diagrams for the DJF forecasts initialized in November and obtained through MME. Values of HSS and RPSS higher than 0 indicate that the forecast is better than the climatology. It is worth pointing out that although we consider this forecast as uncalibrated, we corrected the forecasts for biases in the mean and variance since the probabilities for each model were computed with respect to each model tercile. HSS is highest at tropics, especially over the Atlantic Ocean and Northeastern South America. At extratopics, a local maximum is noticed in South Eastern South America (SESA). Both regions are highlighted with black rectangles in the figure. These regions are well known for being impacted by ENSO teleconnections. On the other hand, there are vast regions, like central Brazil, where the South American Monsoon develops, where the South Atlantic Convergence Zone (SACZ) usually manifests, and over western Argentina, where HSS is less than 40%. The RPSS presents a similar behaviour to HSS with the highest values noticed over the tropics peaking in the Atlantic Ocean. Over the continental extratropics, SESA is the only region with positive RPSS values. Coelho et al. (2007) showed that the mentioned regions presented the highest probabilistic skill for this season according to forecasts made with European models. It should be noticed that the region of the eastern Tropical Pacific Ocean associated to ENSO presents low values of HSS and negative RPSS values. The reliability diagram reveals that NMME system is slightly under-dispersive or overconfident, especially for the above normal category where the event conditioned on a predicted probability of 80–100% was observed only about 50% of the time. The histograms of forecast probabilities (dashed lines) present the characteristic bell shape centered around the climatological (33%) probabilities. Finally, the forecast probabilities for both categories present similar performance in terms of discrimination because the ROC curves and ROC areas for both categories are virtually identical.

We now begin with the evaluation of the performance of the six consolidated forecasts analyzing the HSS (Fig. 6). The maps show that both combination techniques present an improvement with respect to the MME with almost no gridpoints with HSS below than 20%. In addition, in the extratropics there is an increment of the number of gridpoints with HSS values above 40% with respect to MME. As was observed for the MME, HSS is highest at tropics, especially over the Atlantic Ocean and Northern South America. At extratopics, a local maximum is noticed in South Eastern South America (SESA). The comparison of the different combination techniques and weighting approaches shows that overall APDFs and MMEREG present a similar performance, although equally weighting the models produces a slightly lower performance in comparison to the other weighting methods used.

The RPSS for the six consolidated forecasts (Fig. 7) shows that both employed techniques present an improvement with respect to MME with a much reduced number of gridpoints with negative RPSS values. However, positive

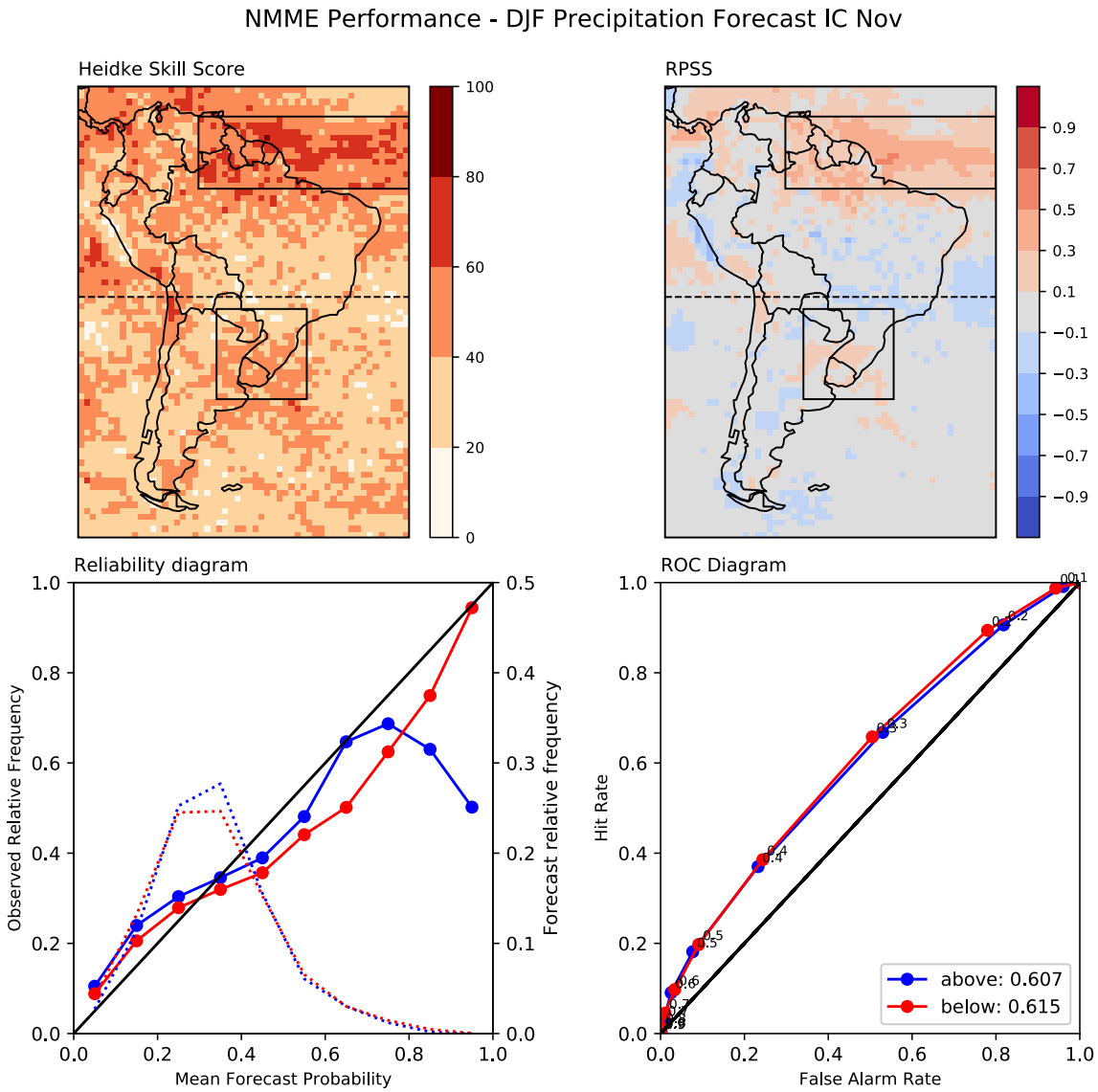

Fig. 5 Skill scores for the uncalibrated DJF precipitation forecast initialized in November: **a** Heidke Skill Score, **b** Ranked Probabilistic Skill Score, **c** reliability diagram (solid lines) and histograms of forecast relative frequency (dotted lines) for the AN (blue) and BN (red) category, **d** ROC diagrams and ROC area for the AN (blue) and BN (red) category. Black rectangles in **a** and **b** denote regions with highest skill while dashed line denotes the 20° S latitude

RPSS values are still low and over the tropical Pacific region associated to ENSO RPSS values remain negative when models are combined through MMEREG. The largest improvements are observed in the tropics over northern South America and the Atlantic Ocean. It is interesting to note that for most of central Brazil and over the region where the SACZ usually manifests the RPSS does not outperform the climatological forecast for any combination technique. As was reported for the HSS, the performance of MMEREG and APDFs are comparable.

Figure 8 shows the temporal evolution of the HSS and the RPSS aggregated over the land gridpoints for the six consolidated forecasts and MME. In addition, we included the index of the Niño 3.4 region since El Niño Southern Oscillation (ENSO) is the main mode of variability that influences austral summer precipitation over South America and previous studies have shown that prediction skill associated to ENSO is high over the regions where ENSO impacts are observed, like northeastern Brazil (Hastenrath et al. 2009; Folland et al. 2001) and SESA (Bombardi et al. 2018; Osman and Vera 2017), which are also the regions with high HSS and RPSS as illustrated in Figs. 6 and 7. For most years the consolidated forecasts perform better than MME, especially in terms of HSS. Even though the performance of the different consolidated forecasts is comparable, those consolidated forecast obtained by equally weighting

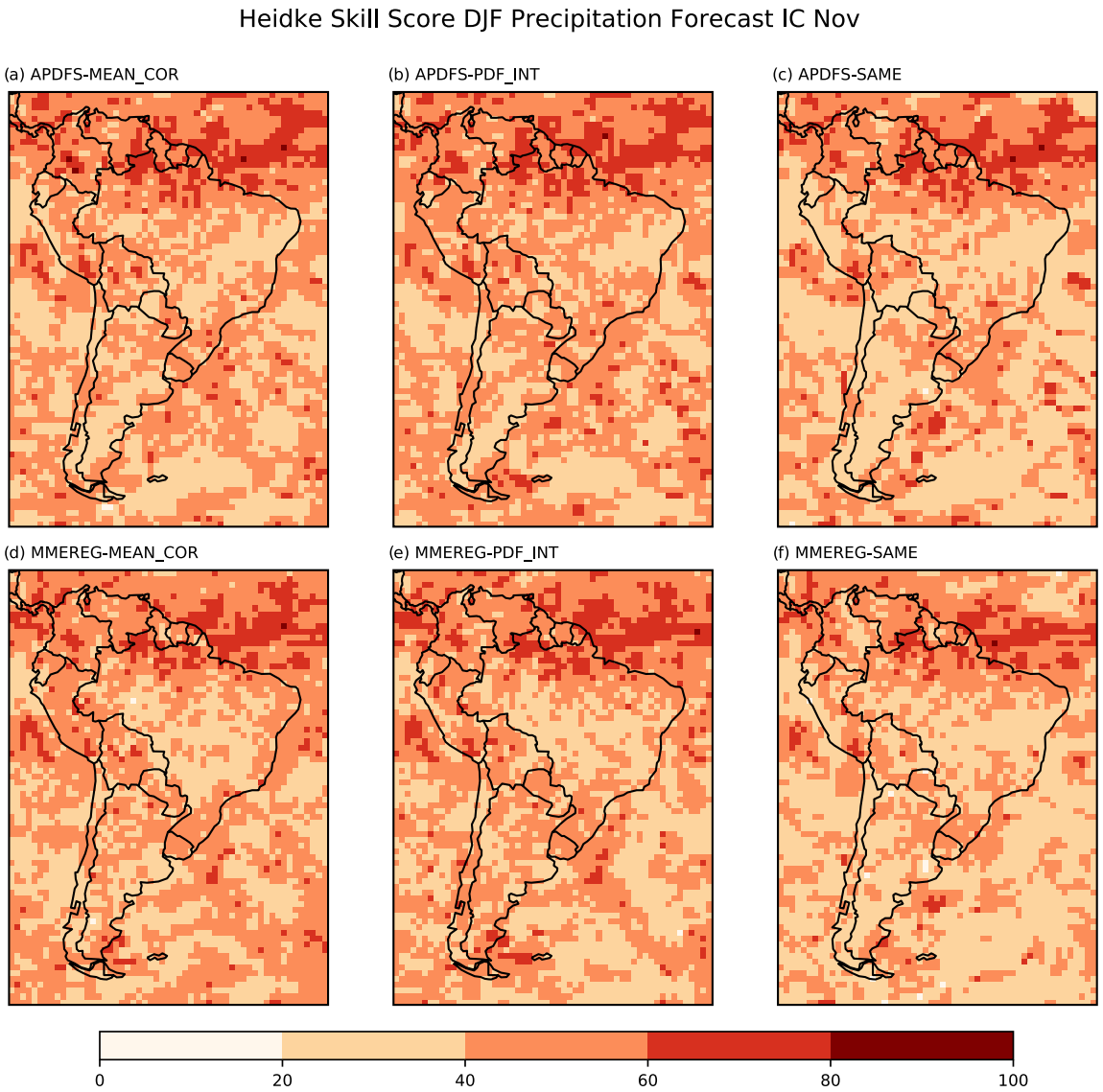

Fig. 6 HSS of the DJF precipitation forecast initialized in Nov for all the combined forecasts

models usually presents lower skill than the rest for most of the years. It is interesting to notice that the performance of the consolidated forecasts is better when the SST anomalies in the Niño 3.4 region are stronger while this is not always noticed for MME. This leads to cases in which the consolidated forecasts show an important improvement in comparison to MME, like the 1982/1983 and 2006/2007 El Niño events. However, there are also non-ENSO years in which calibrated forecasts present similar performance in comparison to MME, like (1985/1986 and 2000/2001). We also notice that MME forecast for La Niña 1988/1989 obtained one of the highest HSS in the entire period and therefore the consolidation degraded the skill in that case. We will analyze this case in more detail later.

Reliability diagrams of the AN and BN forecast event for the six consolidated forecasts are shown in Fig. 9. Overall, the six consolidated forecasts show a similar performance. Combining models with APDFs changes MME forecast from being overconfident to underconfident, regardless of the applied weighting approach. This is more evident for the BN category. In addition, the analysis of the forecast histograms reveals that the APDF methodology pulls inward the forecast between 20–40% bins and therefore the bell shape observed for the MME technique is now sharper, especially when models are equally weighted. This also reduces the frequency of occurrence of the higher probabilities (> 80%) to almost zero. The MMEREG technique applied with any of the weighting methods improves the reliability against MME for the lower probabilities forecast for both events, although

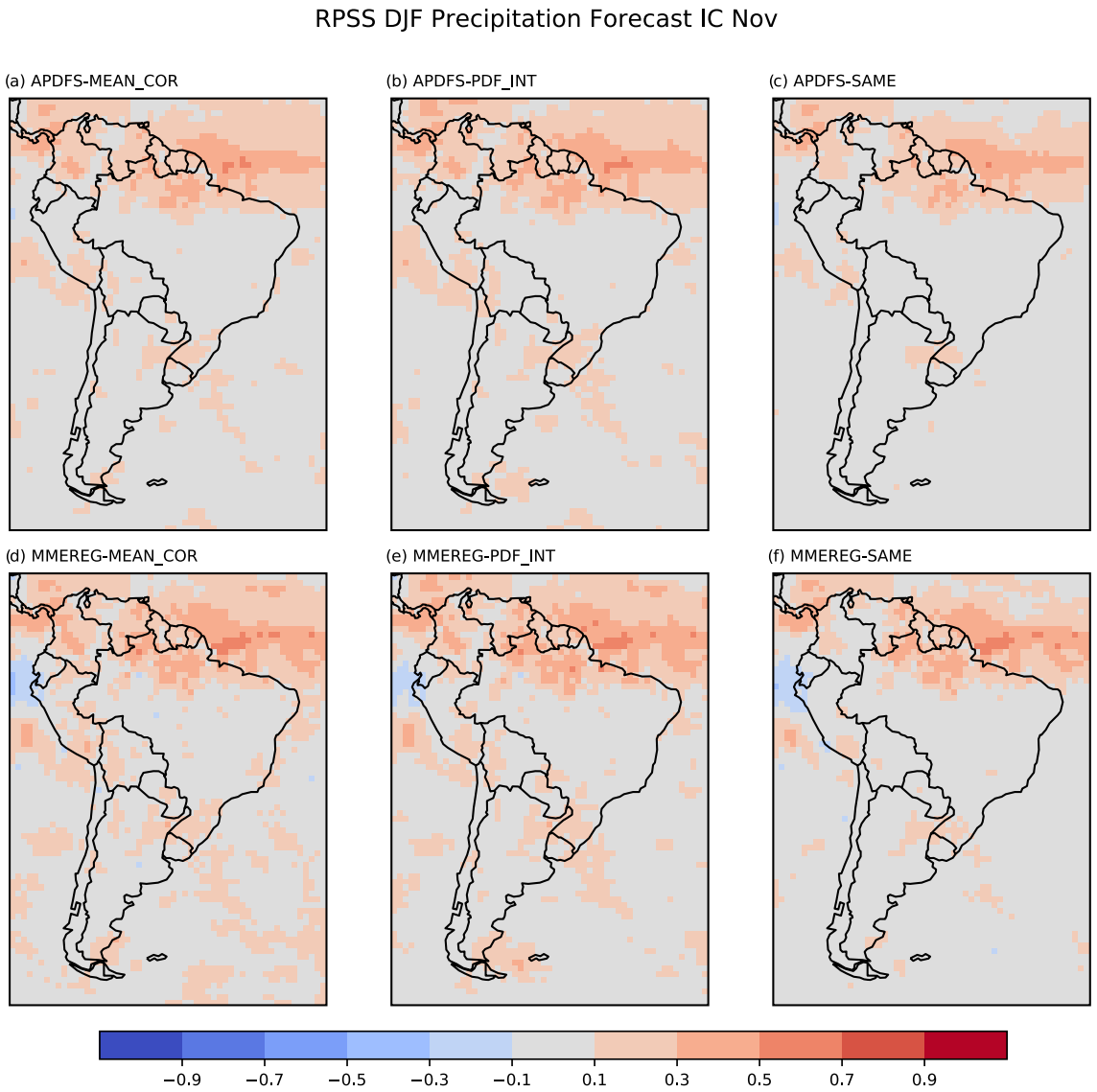

Fig. 7 Same as Fig. 6 but for the RPSS

in the AN category the event conditioned on a high predicted probability is still observed fewer times. In this sense, better results for higher probabilities are obtained when models are weighted according to their performance. When models are combined with MMEREG and weighted according to the correlation the frequency of occurrence of extreme bins increases with respect to MME, while when models are combined with MMEREG but weighted according to the intensity of the PDF or equally weighted the forecast histograms concentrate between the 20–40% bins and show a sharper peak at ∼ 33% than for MME.

The ROC diagrams for the six consolidated forecasts show improvements with respect to MME for both events since ROC areas increase by about 10% in most of the cases (Fig. 10). Both combination techniques present virtually identical performance for both categories, although equally weighting models produces a marginal improvement with respect to MME.

Table 3 shows the Brier Skill Score (BSS) of the categories AN and BN precipitation of all the consolidated forecasts and MME, aggregated over: all the land gridpoints, land gridpoints north of 20° S (Tropics) and land gridpoints south of 20° S (Extratropics). The 20° S latitude circle which splits tropical from extratropical gridpoints is presented in Fig. 5. Overall, all the methods and weighting techniques outperform MME in term of BSS for both categories. However, equally weighting models leads to the lowest BSS in all the domains considered. For all the regions, the BSS for the BN category is higher than for the AN category, regardless the combination and weighting technique used.

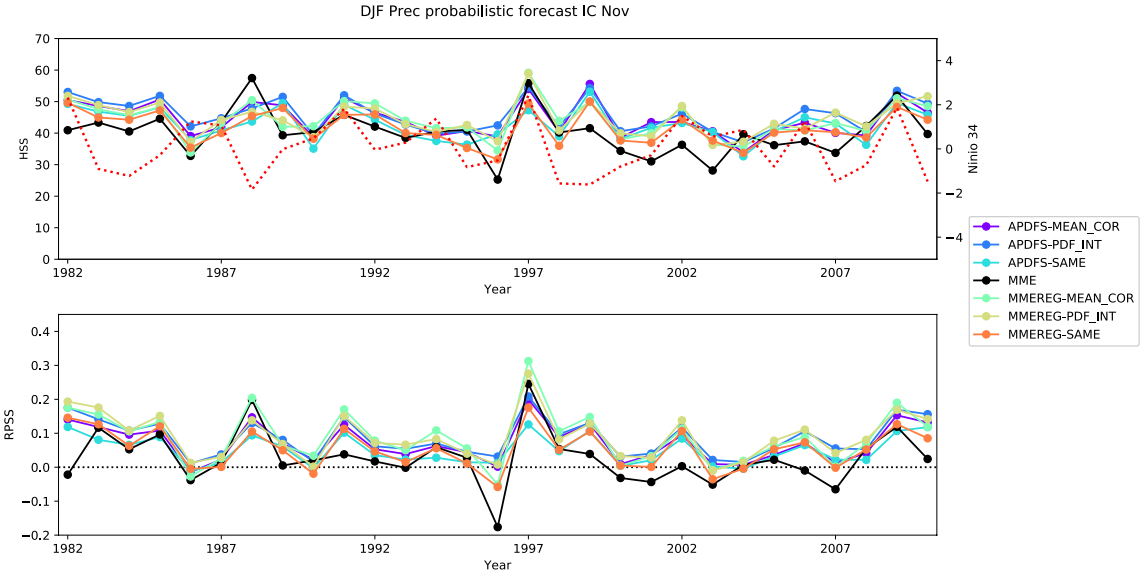

Fig. 8 Temporal evolution of scores. (Top) Mean HSS over land gridpoints in the entire domain for the period of study for all the combined forecasts and MME (filled lines) and DJF Niño 3.4 index (red dotted line). (Bottom) Mean RPSS over land gridpoints in the entire domain for the period of study for all the combined forecasts and MME. The year in the x axis corresponds to that associated to December

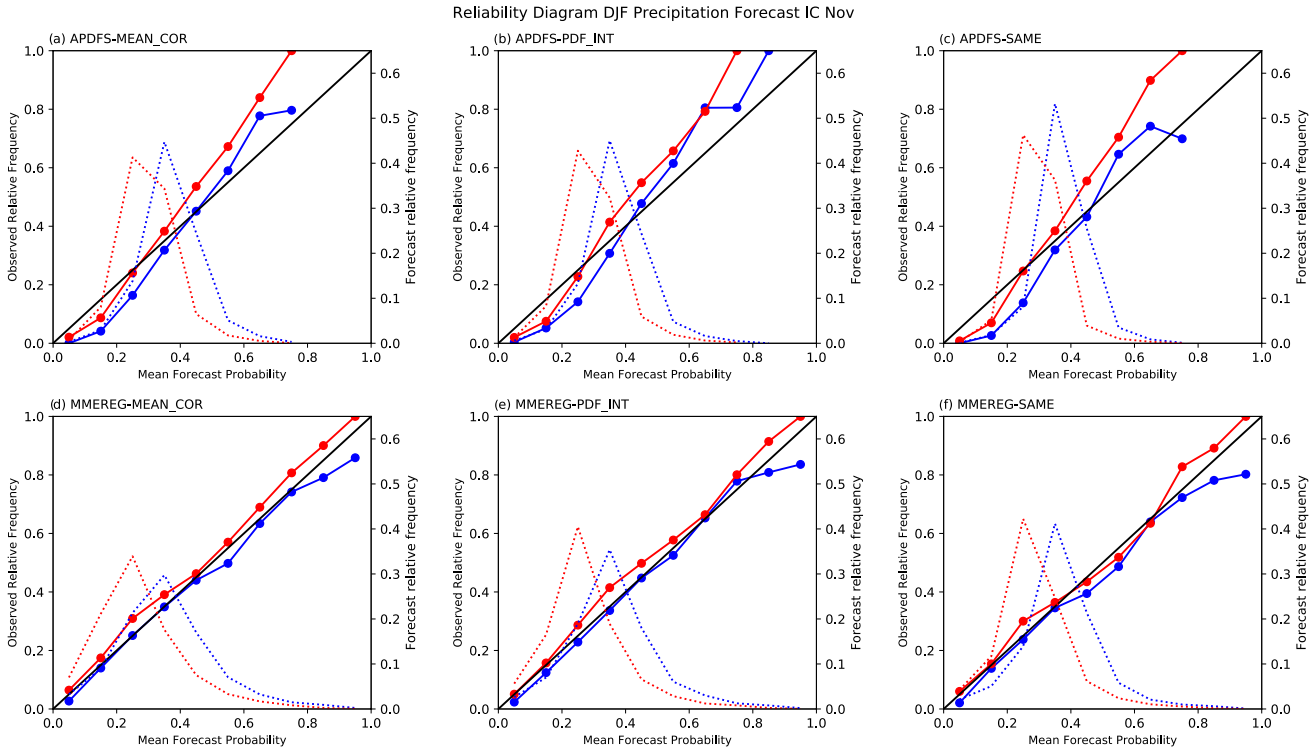

Fig. 9 Reliability diagram (filled lines) and forecast histograms (dotted lines) for the AN (blue) and BN (red) category of the DJF precipitation forecast initialized in Nov for all the combined forecasts

Finally we show two examples of the forecast for all the consolidated forecasts and MME and we confront them with the observed category (Figs. 11 and 12). We present the forecast in the form of probabilities for the most likely category, where categories are in tercile-based format, that is, the probability of the BN, NN, and AN categories, with

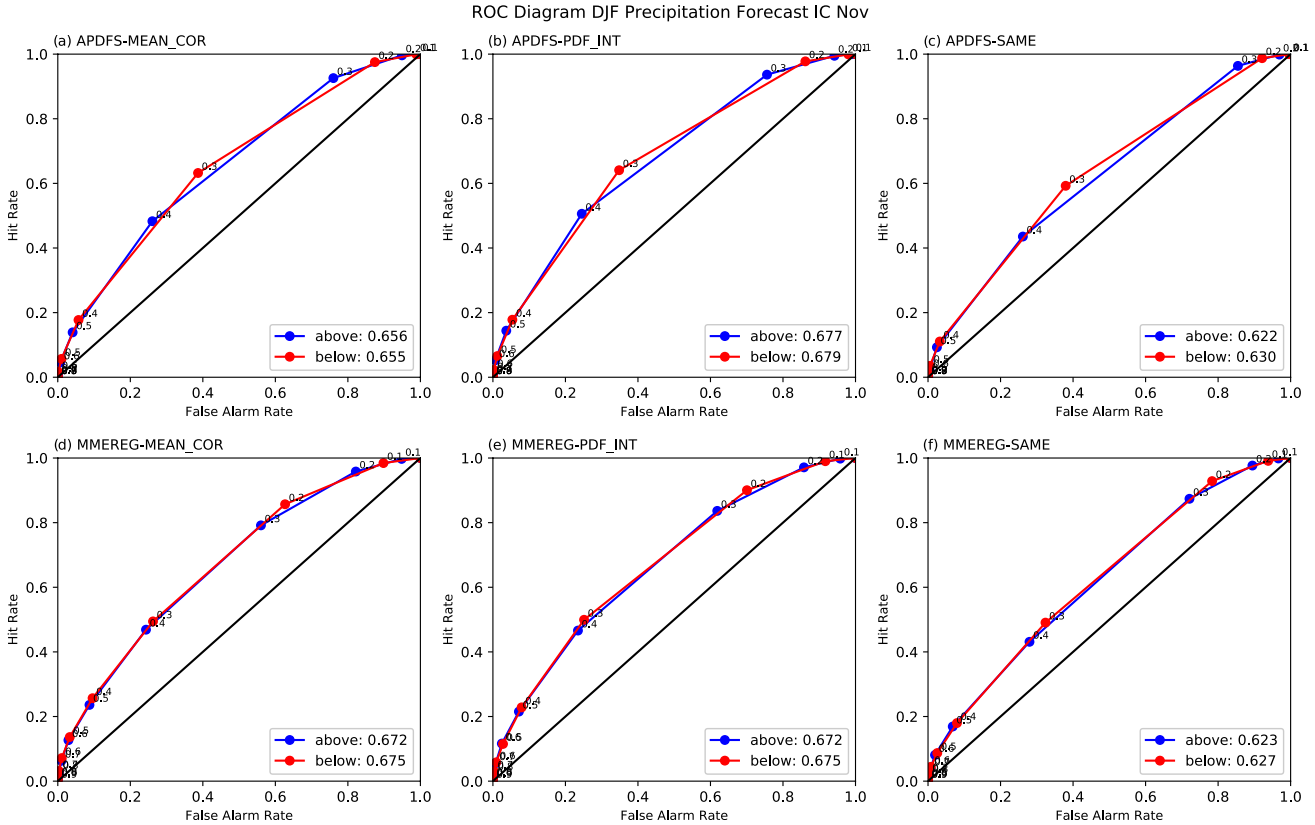

Fig. 10 ROC diagram and ROC area for the AN (blue) and BN (red) category of the DJF precipitation forecast initialized in Nov for all the combined forecasts

Table 3 Mean BSS for the AN and BN category averaged over land gridpoints in the domain of study (All), land gridpoints north of 20° S (Tropics), land gridpoints south of 20° S (Extratropics)

Method	All		Tropics		Extratropics	
	AN	BN	AN	BN	AN	BN
MME	0.7	5.7	3.0	7.5	− 3.8	1.7
APDFs-SAME	**2.9**	**8.2**	**4.4**	**9.7**	**− 0.2**	**5.3**
APDFs-MEAN_COR	**5.1**	**10.1**	**7.0**	**11.8**	**1.2**	**6.4**
APDFs-PDF_INT	**6.8**	**11.3**	**8.4**	**12.9**	**3.7**	***8.4***
MMEREG-SAME	**3.4**	**8.4**	**4.9**	**10.2**	**0.2**	**5.0**
MMEREG-MEAN_COR	**7.2**	***12.3***	***9.2***	***13.5***	**3.2**	**7.0**
MMEREG-PDF_INT	***7.4***	**11.6**	**9.1**	**13.3**	***4.1***	**8.3**

Each row corresponds to a different combination and weighting technique. Bolditalic cells denote the highest BSS for each category and region among all combination techniques while bold cells denote the calibration and combination techniques that outperforms MME for each category and region

respect to climatology. The probability maps are smoothed spatially with a 9×9 Gaussian filter. In addition, we plot the most likely category only when that probability is above 40%. We selected the DJF precipitation during 1982/1983 and 1988/1989 because they corresponds to ENSO events (1982/1983 El Niño and 1988/1989 La Niña) but in one case all the combined forecasts outperforms MME in terms of HSS and RPSS (1982/1983) and in the other case the MME was better than those resulting from combination (1988/1989) (see Fig. 8). The typical influence of El Niño events in austral summer (DJF) precipitation is of negative anomalies over northern South America (Colombia, Venezuela and northern Brazil) and positive anomalies in Ecuador and northern Perú, central Chile and SESA (Cai et al. 2020). Opposite conditions are observed during La Niã events. In 1982/1983 (Fig. 11) BN precipitation is observed in tropical South America, over northern Brazil and Colombia; and Perú. In the extratropics AN precipitation is present

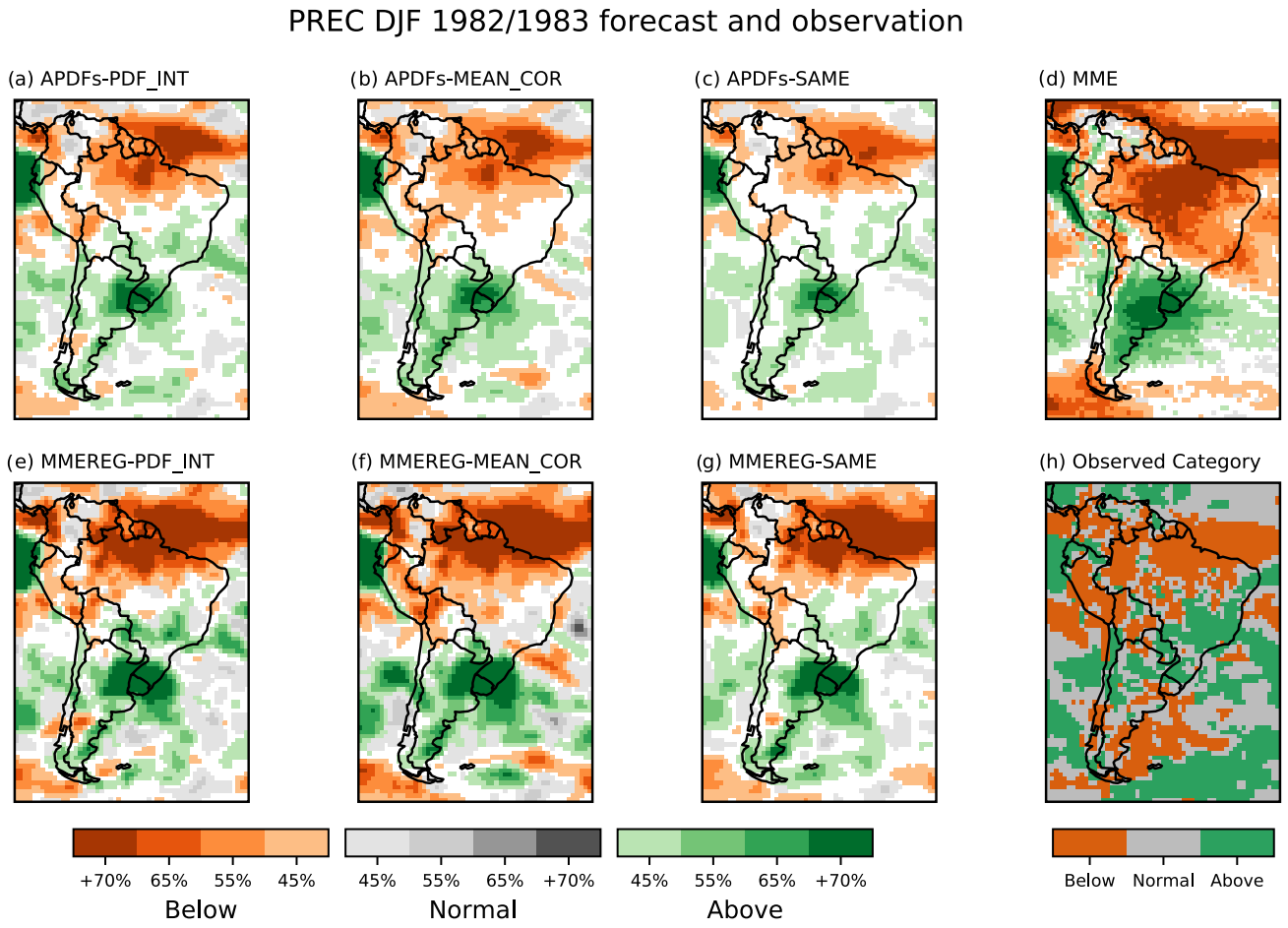

Fig. 11 Probabilistic forecast maps for 1982/1983 DJF precipitation initialized in November made with all the combined forecasts (**a**–**c** and **e**–**g**) and that made with MME (**d**). 1982/1983 DJF observed category (**h**). The forecast maps show the regions where the dominant forecast category presents a probability higher than 40%

over mostly SESA, northern and Central Chile. MME forecasts higher chances for the BN category over most of northern Brazil, in agreement with observations. However, the region with higher probabilities for the BN tercile also extends toward the American Monsoon region and where the SACZ usually manifests, where observed precipitation is AN. On the other hand, in the extratropics the region of highest forecast probabilities for the AN category is south of SESA and Patagonia, although this category is observed only in limited gridpoints there. Over the western coast of South America the most likely category forecast does not match with observations in a large part of Perú and southern Chile. All the combined forecasts improve the performance with respect to MME in the American Monsoon region and SACZ while in the extratropics the region with highest forecast probabilities for the AN category shifted from central Argentina to SESA. Over western South America the combined forecast also presents a higher agreement between the most likely forecast category and the observed one. Overall, the MME forecast for the most likely category resembles the ENSO impacts for El Niño in austral summer over South America although the influence spans larger areas. The calibration and combination of forecasts reduce these areas improving the performance in comparison to MME.

In 1988/1989 (Fig. 12) observed precipitation was mostly associated either to the BN or the AN category while the NN category is less present than in the previous example. The pattern resembles that for La Niña impacts although spans a larger area. MME forecast for the most likely category is in high agreement with the observed category, which explains the high HSS (near 60%) reported. The largest discrepancies are over the western north and central coast. The combined forecasts reduce the probabilities of the most likely category over most of the domain and therefore there are more gridpoints in which all the categories are equiprobable in comparison to MME. This can explain why HSS was lower for the six combined forecasts than for MME.

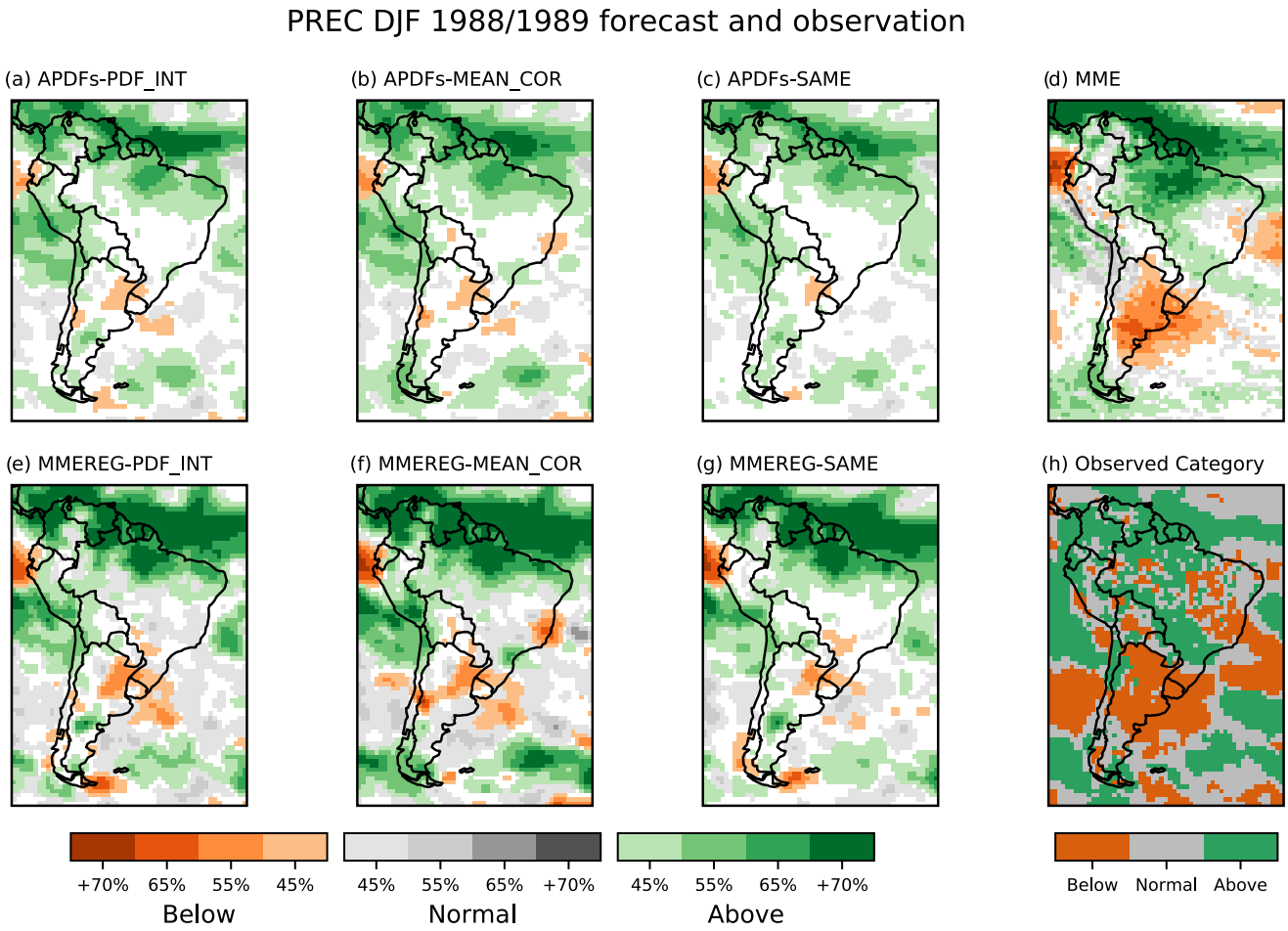

Fig. 12 Same as Fig. 11 but for 1988/1989 DJF precipitation

4 Conclusions

Seasonal forecasts of models participating in the NMME project have been calibrated and combined to provide probabilistic forecasts initialized in November for DJF precipitation in South America and its performance for the three equiprobable categories AN, NN and BN has been assessed using several skill scores. This system is based on EREG that uses the entire ensemble information to obtain reliable forecasts.

In this context, two different ways to combine models were tested. One consists in averaging the PDFs of each model previously calibrated with EREG (APDFs) while the other one implies applying EREG to the entire multi-model ensemble together (MMEREG). In addition, we evaluated the impacts of weighting models according to their performance: on one hand weighting models according to the mean correlation against observations (mean_cor) and on the other weighting models proportionally to the intensity of the calibrated PDF of each model at the observation value (pdf_int). When models are combined through APDFs, the weight is applied when averaging the calibrated PDFs of each model, whereas through MMEREG each model is weighted when the multi-model ensemble mean is computed prior to estimating the regression parameters. We also combined models equally weighting them in both procedures, in APDFs when computing the final consolidated forecast, and in MMEREG when computing the multi-model ensemble mean prior to estimating the regression parameters. We confronted the six options of combination and weights against the average of the probabilities of each tercile across all models obtained by counting the number of ensemble members falling in each category (MME).

Overall the six consolidated options performed as well or better than MME. The performance of both combination techniques was comparable, although MMEREG showed slightly better results than APDFs in several cases. On the other hand, weighting models according to their performance prior to combination led to somewhat more reliable forecasts than equally weighting them. The combination of models through MMEREG when models are weighted according to their correlation produced slightly better results overall.

This could be attributed to the fact that weighting models adjusts the ensemble spread to meet the EREG assumptions and therefore EREG can be applied to individual ensemble members instead of the multi-model ensemble mean, better capturing the uncertainty in the forecasts. Further work is needed in order to evaluate this hypothesis.

Verification of hindcast forecasts showed that the regions with highest skill (in terms of the HSS and the RPSS) are the Tropics, especially northeastern Brazil and the adjacent Atlantic Ocean. At the extratropics, SESA is the main region where forecasts outperform the climatological reference. Both regions present the highest ENSO impacts in the continent. The analysis of the yearly evolution of the skill scores revealed that the combined forecast performed better when the amplitude of the El Niño 3.4 index was higher. This was also reported by a previous probabilistic forecast system developed for South America (Coelho et al. 2006) and worldwide (Min et al. 2017). On the other hand, central Brazil, particularly the region encompassing part of the South American Monsoon system and the regions where the SACZ usually manifests, where precipitation peaks in this season, showed much reduced performance. This result observed with a probabilistic forecast was also reported when deterministic skill scores were analyzed (e.g. Osman and Vera 2017). The analysis of the reliability and ROC diagrams showed that the forecasts of the BN category were more reliable than the AN when forecasts were combined through MMEREG while the opposite was observed for the combination through APDF. ROC diagrams were almost identical for both categories, AN and BN.

All in all, our results showed that both employed methodologies resulted in improved probabilistic predictions when compared to the simple multi-model ensemble (MME) in many aspects while being computationally affordable. The forecast combination approaches assessed in this study have been implemented for the twelve overlapping seasons and for both precipitation and temperature. Forecast are available at http://climar.cima.fcen.uba.ar/Estacional.php. These forecasts are being used in the context of the monthly climate briefings jointly organized by the Argentina Weather Service (SMN) and Centro de Investigaciones del Mar y la Atmósfera. In addition, the forecasts serve as a guidance for the official seasonal forecast released by the SMN. In a separate study we will present the performance of the hindcast forecast in other seasons as well as the real-time predictions.

Acknowledgements The research was supported by UBACyT20020170100428BA, PDE_46_2019 and the CLIMAX Project funded by Belmont Forum/ANR-15-JCL/-0002-01. We acknowledge the agencies that support the NMME-Phase II system, and we thank the climate modeling groups (Environment Canada, NASA, NCAR, NOAA/GFDL, NOAA/NCEP, and University of Miami) for producing and making available their model output. NOAA/NCEP, NOAA/CTB, and NOAA/CPO jointly provided coordinating support and led development of the NMME-Phase II system. MO thanks Dan Collins from Climate Prediction Center for helpful discussions and suggestions throughout the investigation. CASC thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), process 305206/2019-2, and Fundação de Amparo à Pesquisa do Estado de S ao Paulo (FAPESP), process 2015/50687-8 (CLIMAX Project) for the support received.

References

Bombardi RJ, Trenary L, Pegion K, Cash B, DelSole T, Kinter JL (2018) Seasonal predictability of summer rainfall over South America. J Clim 31(20):8181–8195

Buizza R (2006) The ECMWF ensemble prediction system. Cambridge University Press, pp 459–488

Cai W, McPhaden MJ, Grimm AM, Rodrigues RR, Taschetto AS, Garreaud RD, Dewitte B, Poveda G, Ham Y-G, Santoso A, Ng B, Anderson W, Wang G, Geng T, Jo H-S, Marengo JA, Alves LM, Osman M, Li S, Wu L, Karamperidou C, Takahashi K, Vera C (2020) Climate impacts of the El Niño–Southern Oscillation on South America. Nat Rev Earth Environ 1(4):215–231

Coelho CAS, Stephenson DB, Balmaseda M, Doblas-Reyes FJ, van Oldenborgh GJ (2006) Toward an integrated seasonal forecasting system for South America. J Clim 19(15):3704–3721

Coelho CA, Stephenson DB, Doblas-Reyes FJ, Balmaseda M, Graham R (2007) Integrated seasonal climate forecasts for South America. CLIVAR Exch 12:13–19

Collins DC (2017) Assessment of ensemble regression to combine and weight seasonal forecasts from multiple models of the NMME. In: Climate prediction S&T digest: NWS science & technology infusion climate bulletin supplement, chapter. National Oceanic Atmospheric Administration. Professional Paper

Doblas-Reyes FJ, Hagedorn R, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. Tellus A 57(3):234–252

Epstein ES (1969) A scoring system for probability forecasts of ranked categories. J Appl Meteorol 8(6):985–987

Folland CK, Colman AW, Rowell DP, Davey MK (2001) Predictability of northeast brazil rainfall and real-time forecast skill, 1987–98. J Clim 14(9):1937–1958

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

Hartmann HC, Pagano TC, Sorooshian S, Bales R (2002) Confidence builders: evaluating seasonal climate forecasts from user perspectives. Bull Am Meteorol Soc 83(5):683–698

Hastenrath S, Sun L, Moura AD (2009) Climate prediction for Brazil's Nordeste by empirical and numerical modeling methods. Int J Climatol 29(6):921–926

Kirtman BP, Min D, Infanti JM, Kinter JL III, Paolino DA, Zhang Q, van den Dool H, Saha S, Mendez MP, Becker E, Peng P, Tripp P, Huang J, DeWitt DG, Tippett MK, Barnston AG, Li S, Rosati A, Schubert SD, Rienecker M, Suarez M, Li ZE, Marshak J, Lim Y-K, Tribbia J, Pegion K, Merryfield WJ, Denis B, Wood EF (2014) The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bull Am Meteorol Soc 95(4):585–601

Landman WA, Barnston AG, Vogel C, Savy J (2019) Use of El Niño–Southern Oscillation related seasonal precipitation predictability in developing regions for potential societal benefit. Int J Climatol 39(14):5327–5337

Mason SJ, Graham NE (2002) Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves:

statistical significance and interpretation. Q J R Meteorol Soc 128(584):2145–2166
Min Y-M, Kryjov VN, Park C-K (2009) A probabilistic multimodel ensemble approach to seasonal prediction. Weather Forecast 24(3):812–828
Min Y-M, Kryjov VN, Oh SM, Lee H-J (2017) Skill of real-time operational forecasts with the APCC multi-model ensemble prediction system during the period 2008–2015. Clim Dyn 49(11):4141–4156
Osman M, Vera CS (2017) Climate predictability and prediction skill on seasonal time scales over South America from CHFP models. Clim Dyn 49(7):2365–2383
Ou MH, Charles M, Collins DC (2016) Sensitivity of calibrated week-2 probabilistic forecast skill to reforecast sampling of the NCEP global ensemble forecast system. Weather Forecast 31(4):1093–1107
Rajagopalan B, Lall U, Zebiak SE (2002) Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. Mon Weather Rev 130(7):1792–1811
Robertson AW, Lall U, Zebiak SE, Goddard L (2004) Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. Mon Weather Rev 132(12):2732–2744
Rodrigues LRL, Doblas-Reyes FJ, Coelho CAS (2019) Calibration and combination of monthly near-surface temperature and precipitation predictions over Europe. Clim Dyn 53(12):7305–7320
Slater LJ, Villarini G, Bradley AA (2019) Evaluation of the skill of North-American multi-model ensemble (NMME) global climate models in predicting average and extreme precipitation and temperature over the continental USA. Clim Dyn 53(12):7381–7396
Stephenson DB, Coelho CAS, Jolliffe IT (2008) Two extra components in the brier score decomposition. Weather Forecast 23(4):752–757
Tippett MK, Ranganathan M, L'Heureux M, Barnston AG, DelSole T (2019) Assessing probabilistic predictions of ENSO phase and intensity from the North American Multimodel Ensemble. Clim Dyn 53(12):7497–7518
Unger DA, van den Dool H, O'Lenic E, Collins D (2009) Ensemble regression. Mon Weather Rev 137(7):2365–2379
Vigaud N, Robertson AW, Tippett MK (2017) Multimodel ensembling of subseasonal precipitation forecasts over North America. Mon Weather Rev 145(10):3913–3928
Weigel AP, Liniger MA, Appenzeller C (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q J R Meteorol Soc 134(630):241–260
Wilks DS (2011) Statistical methods in the atmospheric sciences. Elsevier Academic Press, Amsterdam
Xie P, Arkin PA (1997) Global precipitation: a 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. Bull Am Meteorol Soc 78(11):2539–2558
Xie P, Yatagai A, Chen M, Hayasaka T, Fukushima Y, Liu C, Yang S (2007) A gauge-based analysis of daily precipitation over East Asia. J Hydrometeorol 8:607–626