Meteorologische Zeitschrift, PrePub DOI 10.1127/metz/2018/0898
© 2018 The authors

A verification framework for South American sub-seasonal precipitation predictions

Caio A.S. Coelho[*], Mári A.F. Firpo and Felipe M. de Andrade

Centro de Previsão de Tempo e Estudos Climáticos (CPTEC), Instituto Nacional de Pesquisas Espaciais (INPE), Brazil

(Manuscript received November 30, 2017; in revised form May 2, 2018; accepted June 7, 2018)

Abstract

This paper proposes a verification framework for South American sub-seasonal (weekly accumulated) precipitation predictions produced one to four weeks in advance. The framework assesses both hindcast and near real time forecast quality focusing on a selection of attributes (association, accuracy, discrimination, reliability and resolution). These attributes are measured using deterministic and probabilistic scores. Such an attribute-based framework allows the production of verification information in three levels according to the availability of sub-seasonal hindcasts and near real time forecasts samples. The framework is useful for supporting future routine sub-seasonal prediction practice by helping forecasters to identify model forecast merits and deficiencies and regions where to trust the model guidance information. The three information levels are defined according to the verification sampling strategy and are referred to as target week hindcast verification, all season hindcast verification, all season near real time forecast verification. The framework is illustrated using ECMWF sub-seasonal precipitation predictions. For the investigated period (austral autumn), reasonable accordance was identified between hindcasts and near real time forecast quality across the three levels. Sub-seasonal precipitation predictions produced one to two weeks in advance presented better performance than those produced three to four weeks in advance. The northeast region of Brazil presented favorable sub-seasonal precipitation prediction performance, particularly in terms of association, accuracy and discrimination attributes. This region was identified as a region where sub-seasonal precipitation predictions produced one to four weeks in advance are most likely to be successful in South America. When aggregating all predictions over the South American continent the probabilistic assessment showed modest discrimination ability, with predictions clearly requiring calibration for improving reliability and possibly combination with predictions produced by other models for improving resolution. The proposed framework is also useful for providing feedback to model developers in identifying strengths and weaknesses for future sub-seasonal predictions systems improvements.

Keywords: sub-seasonal prediction, verification precipitation, South America

1 Introduction

Sub-seasonal precipitation predictions – here referred to as precipitation predictions for weekly periods produced one to four weeks in advance – are relevant for strategic decisions and planning in various South American economic sectors (e.g. agriculture, water management and hydropower generation). However, the sub-seasonal time scale, which covers the range between the traditional day to day weather and average seasonal climate conditions, poses various scientific and practical challenges for model developers and forecasters. Among these challenges is the design and implementation of a procedure for performing a quality assessment of the emerging sub-seasonal predictions, which are starting to be routinely produced by a number of world leading modeling centers. The recent availability of sub-seasonal predictions produced within the context of the joint World Weather Research Program (WWRP)/World Climate Research Program (WCRP) Sub-seasonal to Seasonal prediction project (S2S, Vitart et al., 2012; Robertson et al., 2015) allows the investigation of retrospective predictions (hindcast) and real time forecast quality levels of the participating S2S modeling centers. However, a verification strategy is required in order to document the quality of both deterministic and probabilistic predictions in support of future routine sub-seasonal predictions. This strategy is required because verification information detailing past model performance is a key prediction practice component to enhance forecasters' confidence on the available models predictions and also in support of future model developments. This study proposes a verification framework for these purposes.

An important aspect to be considered in the proposed framework is the large degree of differences in some characteristics of sub-seasonal hindcasts and real time forecasts, which directly impact the verification sample size. For example, the number of available sub-seasonal hindcast years (typically of the order of 20 years or

[*]Corresponding author: Caio Coelho, 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, caio.coelho@inpe.br

DOI 10.1127/metz/2018/0898
Gebrüder Borntraeger Science Publishers, Stuttgart, www.borntraeger-cramer.com

less) is usually reduced when compared to the number of available seasonal hindcast years (typically of the order of 30 years). Within the context of the S2S project very few real time subseasonal forecast years are currently available for verification (about 3 years) with a typically much larger ensemble size than usually available for hindcasts. These differences in sub-seasonal hindcasts and real time forecasts highlight the need for a strategy for sub-seasonal prediction verification practice. The verification literature contains various studies assessing sub-seasonal prediction quality (Weigel et al., 2008; Hudson et al., 2011, 2013; Zhu et al., 2014; Li and Robertson, 2015; Wheeler et al., 2017). However, the literature lacks studies proposing a procedure for comparing the quality of sub-seasonal hindcasts and real time forecasts considering the differences described here, which are of fundamental importance for the under development sub-seasonal prediction verification practice. The framework here proposed deals with these differences, by incorporating a strategy for dealing with the currently available sample sizes in order to advance sub-seasonal prediction verification practice.

The manuscript is organized as follows. Section 2 describes the used datasets and the proposed verification framework for sub-seasonal South America precipitation predictions. Section 3 presents a quality assessment of both deterministic and probabilistic sub-seasonal South America precipitation predictions. Section 4 summarizes the main findings and discusses various aspects of the proposed framework.

2 The proposed verification framework

2.1 Elucidation of the sub-seasonal verification problem and associated questions

Let's say one is interested in issuing a forecast for the expected (mean) accumulated precipitation anomalies for the week 18–24 April 2016 (Monday to Sunday) over South America, where anomalies are computed with respect to the climatological mean accumulated precipitation for this week of interest based on a predefined historical period. From the probabilistic point of view the correspondent forecast is the probability for the occurrence of positive accumulated precipitation anomalies for the week 18–24 April 2016. The event of interest here is occurrence of positive precipitation anomalies, which will have as forecast a probability value between 0 and 100 %, and as outcome 1 if the event positive precipitation anomaly is later observed and 0 if the event turns out not to be observed.

In order to illustrate the proposed verification framework for such sub-seasonal prediction let's consider the European Centre for Medium-Range Weather Forecasts (ECMWF) sub-seasonal forecasts available through the S2S prediction project database (Vitart et al., 2017). As part of this project, the ECMWF model provides ensemble forecasts composed by a control (unperturbed) member and 50 perturbed members, comprising a total of 51 ensemble members, for the following 46 days after the initialization date, at a regular 1.5° × 1.5° grid in latitude and longitude. The closest available ECMWF model forecast initialization date is the previous Thursday 14 April 2016, providing a forecast for the week 18–24 April 2016 four days in advance. This forecast is therefore valid for days 5 (Monday 18 April) to 11 (Sunday 24 April 2016) after the initialization date (14 April 2016) and is referred to as a forecast valid for week 1. If one takes the ECMWF model forecast initialized one week earlier on Thursday 7 April 2016, then this earlier initialization provides a forecast for the week 18–24 April 2016 eleven days in advance. This forecast is now valid for days 12 (Monday 18 April) to 18 (Sunday 24 April 2016) after the initialization date (7 April 2016) and is referred to as a forecast valid for week 2. Taking the ECMWF model forecast initialized two weeks earlier on Thursday 31 March 2016, then this even earlier initialization provides a forecast for the week 18–24 April 2016 eighteen days in advance. This forecast is now valid for days 19 (Monday 18 April) to 25 (Sunday 24 April 2016) after the initialization date (31 March 2016) and is referred to as a forecast valid for week 3. Finally, by taking the ECMWF model forecast initialized three weeks earlier on Thursday 24 March 2016, this much earlier initialization provides a forecast for the week 18–24 April 2016 twenty five days in advance. This forecast is now valid for days 26 (Monday 18 April) to 32 (Sunday 24 April 2016) after the initialization date (24 March 2016) and is referred to as a forecast valid for week 4. These forecasts for the week 18–24 April 2016 produced with these four different initialization dates separated one week apart from each other will be referred hereafter as forecasts produced one to four weeks in advance.

Note that other studies, such as Li and Robertson (2015), defined sub-seasonal forecasts for the first four weeks considering days 1 to 7 after the initialization date to define the first week, days 8 to 14 to define the second week, days 15 to 21 to define the third week, and days 22 to 28 to define the fourth week. In this study the first four days after the initialization date were disregarded and the first week was defined considering days 5 to 11 because there exists another ECMWF model version more adequate and better adjusted for issuing medium range forecasts for the first four days than the ECMWF sub-seasonal model version used here.

Fig. 1 (panels a to d) shows ECMWF ensemble mean forecast accumulated precipitation anomalies for the target week of interest (18–24 April 2016) initialized on Thursdays 14 April 2016, 7 April 2016, 31 March 2016 and 24 March 2016, representing forecasts valid for weeks 1 to 4 as described above. Ensemble means were computed using all 51 ensemble members, and anomalies were determined with respect to the 1996–2015 hindcast period (20 years) for which an ensemble of 11 members [one control (unperturbed) and 10 perturbed members] were available for each of these

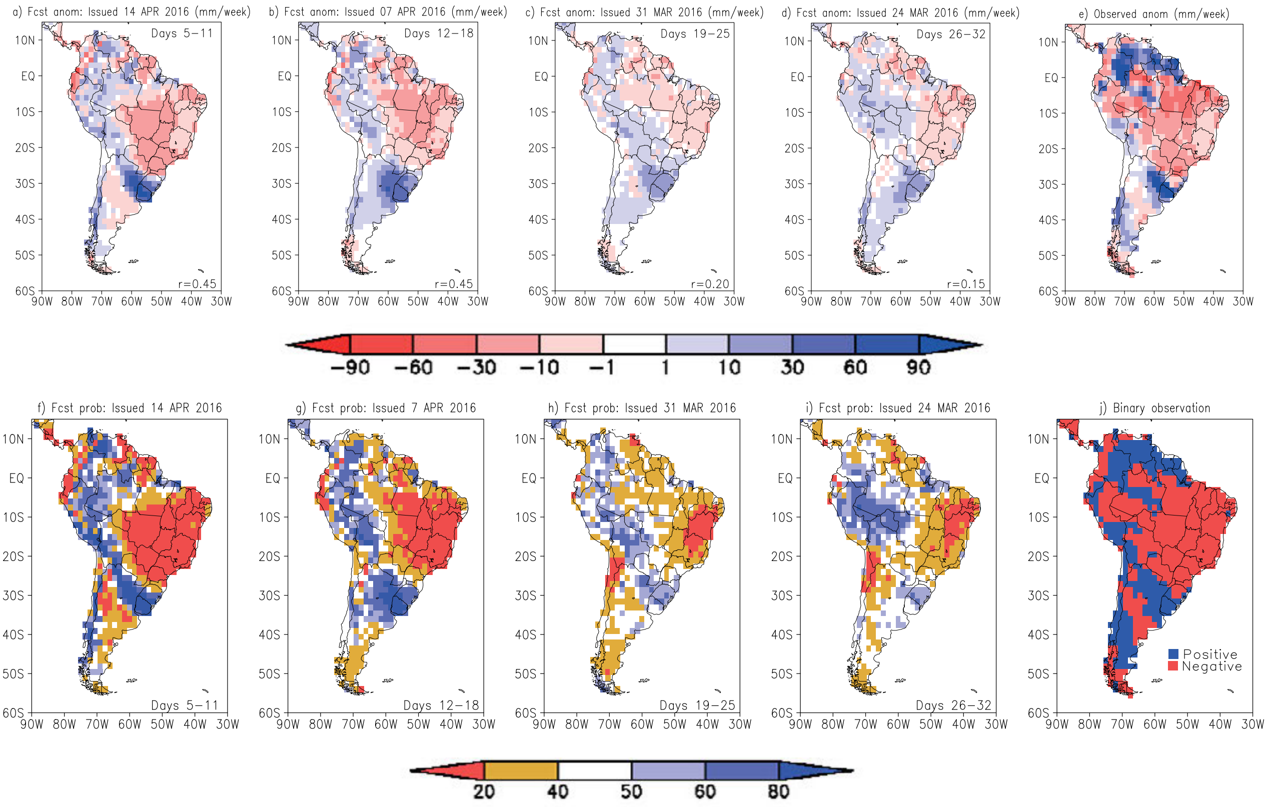
a) Fcst anom: Issued 14 APR 2016 (mm/week)
Days 5−11
r=0.45
b) Fcst anom: Issued 07 APR 2016 (mm/week)
Days 12−18
r=0.45
c) Fcst anom: Issued 31 MAR 2016 (mm/week)
Days 19−25
r=0.20
d) Fcst anom: Issued 24 MAR 2016 (mm/week)
Days 26−32
r=0.15
e) Observed anom (mm/week)
10N
EQ
10S
20S
30S
40S
50S
60S
90W 80W 70W 60W 50W 40W 30W
−90 −60 −30 −10 −1 1 10 30 60 90
f) Fcst prob: Issued 14 APR 2016
Days 5−11
g) Fcst prob: Issued 7 APR 2016
Days 12−18
h) Fcst prob: Issued 31 MAR 2016
Days 19−25
i) Fcst prob: Issued 24 MAR 2016
Days 26−32
j) Binary observation
Positive
Negative
20 40 50 60 80

Figure 1: ECMWF ensemble mean forecast accumulated precipitation anomalies for the target week of interest (18–24 April 2016) initialized on the a) 14 April 2016, b) 7 April 2016, c) 31 March 2016 and d) 24 March 2016, representing forecasts valid for weeks 1 to 4 as described in Section 2.1. A total of 51 ensemble members were used for computing the ensemble mean forecast, and anomalies were calculated with respect to the 1996–2015 hindcast period (20 years) for which an ensemble of 11 members was available for each of these 20 years. e) Observed accumulated precipitation anomalies for the week 18–24 April 2016 with respect to the 1996–2015 period based on the Climate Prediction Center (CPC) daily precipitation dataset (Chen et al., 2008), linearly interpolated to the same 1.5° × 1.5° grid in latitude and longitude as the ECMWF model forecasts. ECMWF forecast probabilities for the occurrence of positive accumulated precipitation anomalies during the target week of interest (18–24 April 2016) initialized on the f) 14 April 2016, g) 7 April 2016, h) 31 March 2016 and i) 24 March 2016, derived from 51 ensemble members. Forecast probabilities were computed as the fraction of ensemble members indicating a positive precipitation anomaly. j) Binary observation indicating where a positive (blue) or a negative (red) precipitation anomaly was recorded during the week 18–24 April 2016 based on the CPC daily precipitation dataset.

20 years. The ECMWF model climatologies used for computing the ensemble mean forecast anomalies were therefore composed by a sample of 220 hindcast model runs for each of the four initialization dates here investigated. Note that for both the computation of ensemble mean anomalies and of verification metrics that are later presented in Section 3 one needs to select the hindcast initialization dates to be consistent with the four real time initialization dates for having adequate lead times (4, 11, 18 and 25 days in advance) for the target week (18–24 April) here investigated. Fig. 1e shows the observed accumulated precipitation anomalies for the week 18–24 April 2016 based on the Climate Prediction Center (CPC) daily precipitation dataset (Chen et al., 2008) linearly interpolated to the same 1.5° × 1.5° grid in latitude and longitude as the ECMWF model, where anomalies were computed with respect to the 1996–2015 period. The visual comparison of the ensemble mean forecast with the observed anomalies reveals an amplitude bias in the forecasts produced three and four weeks in advance, which tend to underestimate the magnitude of the precipitation anomalies over South America. The fact that the ensemble mean anomalies are weaker than the observed anomalies is a sign of large ensemble spread for longer lead forecasts (i.e. for the forecasts produced three and four weeks in advance). Although here the focus of the forecast bias assessment is on the ensemble mean, it is worth noting that not all individual ensemble members may underestimate the magnitude of the precipitation anomalies.

Fig. 1 (panels f to i) shows ECMWF forecast probabilities for the occurrence of positive accumulated precipitation anomalies during the target week of interest (18–24 April 2016) initialized on the 14 April 2016, 7 April 2016, 31 March 2016 and 24 March 2016. Forecast probabilities were computed by counting the number of forecast ensemble members indicating a positive anomaly for the target week and dividing this count by 51 (i.e. the total number of available ensemble members). For determining forecast anomalies for each of the 51 ensemble members, ECMWF model climatologies computed as the mean of the 220 hindcast model runs for each of the four initialization dates here investigated were used. Fig. 1j shows the binary observation indicating where a positive (blue) or a negative (red) precipitation anomaly was recorded during the week 18–24 April 2016 based on the CPC daily precipitation dataset (Chen et al., 2008), where anomalies were computed with respect to the 1996–2015 period.

A few natural questions for the forecasters having access to the forecasts shown in Fig. 1 prior to observing the precipitation accumulation for the week 18–24 April 2016 are: How good are these forecasts for the week 18–24 April 2016 produced one to four weeks in advance in terms of correspondence with the observations? Where spatially over South America can these forecasts be best trusted? How strong is the relationship between the forecast and observed precipitation anomalies? How accurate are the forecast precipitation anomalies compared to the accuracy of a reference naïve forecasting strategy of always issuing a constant forecast value (e.g. null anomaly for the climatological forecast)? How reliable are the issued forecast probabilities? Can the issued forecast probabilities detect the event of interest (i.e. distinguish events from non-events)?

After having observed the precipitation for the week 18–24 April 2016, by comparing the forecast anomalies of Fig. 1 (panels a to d) with the observed anomalies (panel e) and the probabilistic forecasts for the occurrence of positive anomalies (panels f to i) with the binary observation (panel j) one can have a qualitative assessment and visually identify the regions where these forecasts produced one to four weeks in advance were successful. For example, the model successfully indicated the potential for the occurrence of wet conditions (through the indication of high probabilities for the occurrence of positive anomalies) in southern Brazil, northeast Argentina, part of Uruguay and southern Peru up to four weeks in advance. The model also successfully indicated the potential for the occurrence of dry conditions (through the indication of low probabilities for the occurrence of positive anomalies) over northeast Brazil one to four weeks in advance. However, the model failed to indicate the potential for the occurrence of wet conditions over Venezuela, Guyana, Suriname and French Guiana, particularly three to four weeks in advance. Although this is a useful a posteriori forecast assessment, when issuing the forecast for the week 18–24 April 2016, it would also be useful for the forecasters to have available, in addition to the forecast maps shown in Fig. 1 (panels a to d and f to i), some historical performance assessment of the hindcasts and forecasts previously produced for the target week of interest. Such historical information can help the forecasters identify regions where the model consistently shows acceptable performance and regions where the model shows deficiencies, and therefore contribute to building forecasters' confidence on the forecast model guidance information. However, a quantitative approach is required in order to appropriately document past forecast quality and provide support to those using the forecasts. The following Sections 2.2 and 2.3 elaborate and propose a framework for such quantitative sub-seasonal precipitation forecast quality assessment.

2.2 Sampling strategies and information levels for sub-seasonal verification

When issuing seasonal forecasts (forecasts for the expected climate conditions usually for the following three to six months) a common practice is to examine, together with the model forecast guidance information similar to the maps shown in Fig. 1, some verification information in the form of maps or graphics usually constructed with hindcasts produced for a past period of around 30 years and the corresponding observations (i.e. to produce supporting verification information using a sample of around 30 pairs of hindcasts and corresponding observations). A similar procedure can be performed for sub-seasonal forecasting through the use of the available hindcasts. For the sub-seasonal forecast example discussed in this paper 20 years of ECMWF model hindcasts are available, allowing the production of such verification maps and graphics for the target week of interest (18–24 April) for the four Thursday initialization dates (14 April, 7 April, 31 March and 24 March), providing a hindcast quality assessment for the predictions produced for the target week one to four weeks in advance based on a sample of 20 pairs of hindcasts and observations. Although this is a smaller sample than usually available for seasonal predictions, this is enough to provide an initial quality assessment for the sub-seasonal predictions. Such assessment is hereafter referred to as target week hindcast verification, and is the first information level of the proposed verification framework for South America sub-seasonal precipitation predictions.

As ECMWF sub-seasonal hindcasts are generated for initialization dates of every week of the year, the sample of hindcasts and observations pairs can be substantially increased by aggregating surrounding hindcast initialization dates. For example, in addition to the hindcasts produced for the four Thursday initialization dates previously selected (14 April, 7 April, 31 March and 24 March), one can aggregate to the verification sample hindcasts produced for nine additional initialization dates during the weeks of the previous and following month in order to incorporate in the sample all hindcasts initialized on Thursdays of March, April and May of the 2016 calendar. The motivation for aggregating hindcasts initialized in these months is that they represent the austral autumn season, a period marked by similar atmospheric features in various South American regions. By performing this aggregation the hindcasts of the following Thursday initialization dates are incorporated to the verification sample: 3 March, 10 March, 17 March, 21 April, 28 April, 5 May, 12 May, 19 May and 26 May. As a result of this aggregation the verification sample increases from 20 pairs of hindcasts and observations to 260 pairs, obtained by multiplying the 13 initialization dates (the initial four initialization dates plus the nine new initialization dates) by the 20 years of hindcasts (for the period 1996–2015) produced for each initialization date. This new sub-seasonal hindcast sample composed by 260 pairs of hindcasts and observations is much larger than usually available when performing seasonal hindcast verification and provides a solid basis for a robust hindcast quality assessment for the predictions produced one to four weeks in advance during the austral autumn season. Such assessment is hereafter referred to as all season hindcast verification, and is the second information level of the proposed verification framework for South America sub-seasonal precipitation predictions. Note that the sample could have been doubled to 520 pairs of hindcasts and observations as ECMWF also has sub-seasonal hindcasts initialized on Mondays. However, for maintaining consistency with the choice of Thursday initialization dates used in the first level of the proposed verification framework (target week hindcast verification) the second level also solely uses hindcasts produced with Thursday initialization dates, which already provide a reasonably large sample size for verification.

In addition to hindcasts, near real time ECMWF sub-seasonal forecasts are also available through the S2S project database since 2015. The previously discussed Fig. 1 shows illustrations of these forecasts. The availability of near real time forecasts for a selected number of past years provides an opportunity for verifying these forecasts and comparing their performance with the quality of hindcasts. Following the same reasoning for aggregating hindcasts initialized during the weeks of the austral autumn season as earlier elaborated, for near real time forecast verification one can aggregate the forecasts produced on Thursdays during the 13 weeks of March, April and May of each of the past three years (2015, 2016 and 2017). This aggregation leads to a verification sample of 39 pairs of near real time forecasts and observations, obtained by multiplying the 13 initialization dates by the 3 years of forecasts produced for each initialization date. This sub-seasonal near real time forecast sample composed by 39 pairs of near real time forecasts and observations is still larger than usually available when performing seasonal hindcast verification and provides a good basis for a quality assessment of the near real time forecasts produced one to four weeks in advance during the austral autumn seasons of 2015, 2016 and 2017. Such assessment is hereafter referred to as all season near real time forecast verification, and is the third information level of the proposed verification framework for South America sub-seasonal precipitation predictions. Note that the sample could have been doubled to 78 pairs of near real time forecasts and observations as ECMWF also has sub-seasonal forecasts initialized on Mondays. However, for maintaining consistency with the choice of Thursday initialization dates used in the first two levels of the proposed verification framework the third level also solely uses forecasts produced with Thursday initialization dates.

The three level strategy of the proposed verification framework has strengths and weaknesses, making it challenging to decide and choose a single approach out of the three options. For this reason it is important to recognize the differences, merits and limitations of the three approaches and, most importantly, consider then as complementary verification strategies. For example, the hindcast datasets of level two (all season hindcast verification) provide extensively large samples compared to the reduced samples of level one (target week hindcast verification). The level two hindcast quality is, however, likely to be lower than level three (all seasonal near real time forecast verification) quality, particularly because the initial conditions from the reanalysis dataset used for producing hindcasts are of poorer quality than the operational analysis used as initial conditions for producing real time forecasts. This is due to the fact that the observing system has improved over the past 20 years and the reanalysis used for initializing hindcasts is based on a model and data assimilation system that are outdated compared to the operational model version used for producing real time forecasts. Besides, the ensemble size for the produced hindcasts is smaller than for real time forecasts (11 hindcast ensemble members against 51 real time forecast ensemble members for the ECMWF sub-seasonal predictions here investigated). Additionally, the level three (all season near real time forecast verification) quality assessment cannot be considered comprehensive either due to the limited number of available forecast years (3 years for the ECMWF sub-seasonal predictions here investigated). In this third level forecast quality is heavily affected by the specific interannual variability due to the El Niño Southern-Oscilattion (ENSO) and the Madden-Julian Oscilation (MJO) activity, and the numerous model version changes that occurred during this three year period.

2.3 Attribute-based forecast quality assessment

Murphy (1993) defined a number of aspects, so-called attributes, for assessing forecast quality. This section describes the proposed procedures for assessing sub-seasonal South American precipitation predictions based on a selection of some of the most fundamental of these attributes, namely, association, accuracy, discrimination, reliability and resolution.

In order to quantify the degree of correspondence between the deterministic forecasts shown in Fig. 1 (panels a to d) and the observations shown in Fig. 1e, it is proposed, as an initial assessment, to perform the comparison of the forecast (ensemble mean) precipitation anomaly pattern with the observed precipitation anomaly pattern over South America for the selected target week (18–24 April 2016), with the strength of correspondence between the two patterns quantified with the spatial pattern correlation. This comparison measures the so-called forecast quality association attribute commonly used in forecast verification studies, here proposed to be assessed with the linear Pearson correlation coefficient.

Additionally, the following procedure is proposed for each of the three information levels of the verification framework for South America sub-seasonal precipitation predictions, namely, target week hindcast verification, all season hindcast verification, all season near real time forecast verification, described in the previous Section 2.2, with hindcast and observed anomalies produced in cross-validation mode by removing the target re-forecast or observed year being verified when computing the model or observed climatological means needed for computing anomalies:

a) Construction of maps showing the correlation between the predicted ensemble mean and observed precipitation anomalies at each grid point over the available hindcast/forecast period, with the aim of assessing the strength of the linear association between the predicted and observed anomalies. Note that here the association attribute is assessed locally (at each grid point) rather than spatially as described above. However, the final verification diagnostics is a spatial map with the Pearson correlation coefficient shown for each grid point over South America.

b) Construction of maps showing the mean squared error skill score ($\mathrm{MSSS} = 1 - \mathrm{MSE}/\mathrm{MSE}_c$), where MSE is the mean squared error of the predicted precipitation anomalies computed at each grid point over the available hindcast/forecast period, and MSE_c is the mean squared error for a reference prediction. In this paper the constant climatological (null) precipitation anomaly prediction is used as the reference prediction. The MSE is a deterministic accuracy measure computed as the average square difference between predictions and observations. The MSSS therefore measures deterministic prediction accuracy relative to the accuracy of the reference (climatological) prediction. Positive values of MSSS indicate greater (less) accurary than the reference (climatological) prediction.

c) Construction of maps showing the ratio of the predicted precipitation ensemble mean anomaly standard deviation and the observed precipitation anomaly standard deviation at each grid point over the available hindcast/forecast period in order to complement the information provided in the first two verification maps above (correlation and MSSS). This is due to the fact that, as shown in equation 12 of Murphy (1988), the MSSS incorporates information about the phase error (through the correlation between the predicted ensemble mean and the observed precipitation anomalies), the mean error (through the overall bias) and the amplitude error (through the ratio of the predicted ensemble mean to the observed standard deviation). It is therefore important to consider these different MSSS components separately for a complete assessment. As we are dealing with anomalies the mean error (the overall bias) is null and therefore does not require an assessment. The correlation component for evaluating association (or phase error) is already assessed in a) above. The final missing component (the above described standard deviation ratio) is here proposed to be computed for assessing amplitude error.

d) Use of ensemble predictions for constructing maps showing the area under the Relative Operating Characteristic (ROC) curve (Mason, 1982; Mason and Graham, 2002) for probabilistic predictions for the event positive precipitation anomaly issued for each grid point over the available hindcast/forecast period, with the aim of assessing discrimination ability (i.e. ability to successfully distinguish events from non-events). Discrimination is considered one the most fundamental attributes of forecast quality, and therefore important to be measured.

e) Construction of ROC curves for ensemble derived probabilistic predictions issued for the event positive precipitation anomaly collected over all South American grid points, with the aim of assessing overall discrimination after aggregating all available hindcasts/forecasts in space and time.

f) Construction of reliability diagrams for ensemble derived probabilistic predictions issued for the event positive precipitation anomaly collected over all South American grid points, with the aim of assessing reliability (i.e. how well calibrated the issued probabilities are) and resolution (i.e. how the frequency of occurrence of the event differs as the issued probability changes) after aggregating all available hindcasts/forecasts in space and time. Reliability and resolution are also considered fundamental attributes of forecast quality, required to be measured.

By providing all these maps and graphics produced in the above proposed three level verification framework, together with the forecasts shown in Fig. 1, to the forecaster, he/she will be able to have a comprehensive quality assessment based on hindcasts and previously issued near real time forecasts, and therefore answer the questions posed in Section 2.1, which will help building confidence in the model forecast guidance information when issuing sub-seasonal forecasts.

3 Sub-seasonal prediction quality assessment

3.1 Deterministic (ensemble mean) prediction quality

As proposed in Section 2.3 the first assessment for quantifying the quality of sub-seasonal forecasts for a selected target week can be obtained by computing the correlation between the forecast and observed precipitation anomaly patterns over South America. A perfect spatial pattern match would result in a pattern correlation value equals to unity. The numbers on the bottom right hand corner of Fig. 1 (panels a to d) show the correlation between the forecast patterns for the week 18–24 April 2016 produced one to four weeks in advance and the observed precipitation anomaly pattern for the same week shown in Fig. 1 (panel e). These pattern correlation values are positive and larger for shorter forecast leads than for longer forecast leads, but far from unity. This indicates that the spatial patterns of forecasts produced one to two weeks in advance have a closer spatial match to the observed spatial pattern than forecasts produced three to four weeks in advance. The larger association between the forecast and observed patterns for forecast produced one to two weeks in advance compared to forecasts produced three to four weeks in advance is also confirmed when visually comparing the forecasts (panels a to d) with the observed (panel e) anomaly pattern.

In order to further assess association within the context of the proposed three information level verification framework for South America sub-seasonal precipitation predictions, Fig. 2 shows maps of correlation between the predicted ensemble mean and observed precipitation anomalies at each grid point for the three hindcasts/forecasts sampling strategies discussed in Section 2.3 produced one to four weeks in advance. These maps assess the temporal strength (over the available hindcast/forecast sample) of the linear association between the predicted and observed anomalies at each grid point. Perfect association (i.e., a correlation coefficient equal unity) is obtained when the forecasts and observations are in phase with each other and oscillate exactly in the same direction. This measure, however, only provides an indication of potential prediction ability because correlation is insensitive to forecast amplitude error such as differences in hindcast/forecast versus observed standard deviation. Fig. 2 reveals that all three verification sampling strategies applied to hindcasts/forecasts produced one to two weeks in advance show better association than hindcasts/forecasts produced three to four weeks in advance. This feature is noticed by the larger portion of the South American continent presenting statistically significant (at the 5 % level) and different from zero correlation coefficients for shorter lead when compared to longer lead predictions. Another feature worth noting is the maintenance of such statistically significant correlation levels for the forecasts produced three to four weeks in advance in tropical South America, particularly over northeast Brazil.

Comparing the correlation maps of Fig. 2 for the two hindcast sampling verification strategies of the first two levels of the proposed verification framework, namely target week hindcast verification (panels a to d) and all season hindcast verification (panels e to h), one can notice a reasonable consistency between the spatial patterns for the four investigated weeks. This suggests that although the hindcast sample size is considerably reduced in the first level compared to the second, such an assessment does provide valuable and meaningful information about regions where the predictions have good linear association. The correlation maps for the third level, all season near real time forecast verification strategy (panels i to l), are also largely spatially consistent with the maps of the first two levels (panels a to h), even though the number of ensemble members available for the realtime forecasts used in third level is considerably larger than for the other two levels. When producing the correlation maps for the third level with 11 ensemble members (the same number of ensemble members used for the first and second levels) the spatial patterns remain similar to the patterns shown in panels i to l (not shown), with the magnitudes of the correlation coefficients slightly reduced particularly for forecasts produced three and four weeks in advance, suggesting that the use of a larger ensemble size contributes for improving forecast quality. However, increasing the sample size for the third level when more real time forecasts become available in the future is likely to make more apparent the impact of increased ensemble size on forecast quality. For predictions produced three to four weeks in advance some differences are noticed between the correlation maps for the third level (panels k and l) and the correlation maps for the other two levels (panels c, d, g and h), for example in northern Brazil and Argentina. As earlier discussed in Section 2.2 these differences might be due to the modulation of specific interannual variability phenomenon (e.g. ENSO and MJO) during the three year period here investigated. But overall the correlation maps for the three levels are consistent and complementary to each other providing a reasonable indication for the regions where the model predictions have good linear association.

Fig. 3 shows maps of the mean squared error skill score (MSSS) for hindcast/forecast precipitation anomalies for the three verification sampling strategies computed with respect to the reference (climatological) prediction as described in Section 2.3. There is overall a good match of the regions showing positive MSSS in Fig. 3 (i.e. improved accuracy compared to the reference prediction) and the regions that showed the largest correlation coefficients (i.e. the smallest phase error) in Fig. 2. This suggests that the correlation component of the MSSS decomposition [equation 12 of Murphy (1988)] contributes considerably for the identified positive skill in Fig. 3. The MSSS maps for the first (panels a to d) and third (panels i to l) levels of the proposed verification framework show larger positive values than for the second level (panels e to h). This is partially due to the much increased sample size of the second level compared to the other two levels, which makes it harder to produce predictions with improved accuracy compared to the reference (climatological) prediction. When producing the MSSS maps for the third level with 11 ensemble members (the same number of ensemble members used for the first and second levels) the spatial patterns remain similar to the patterns shown in panels i to l

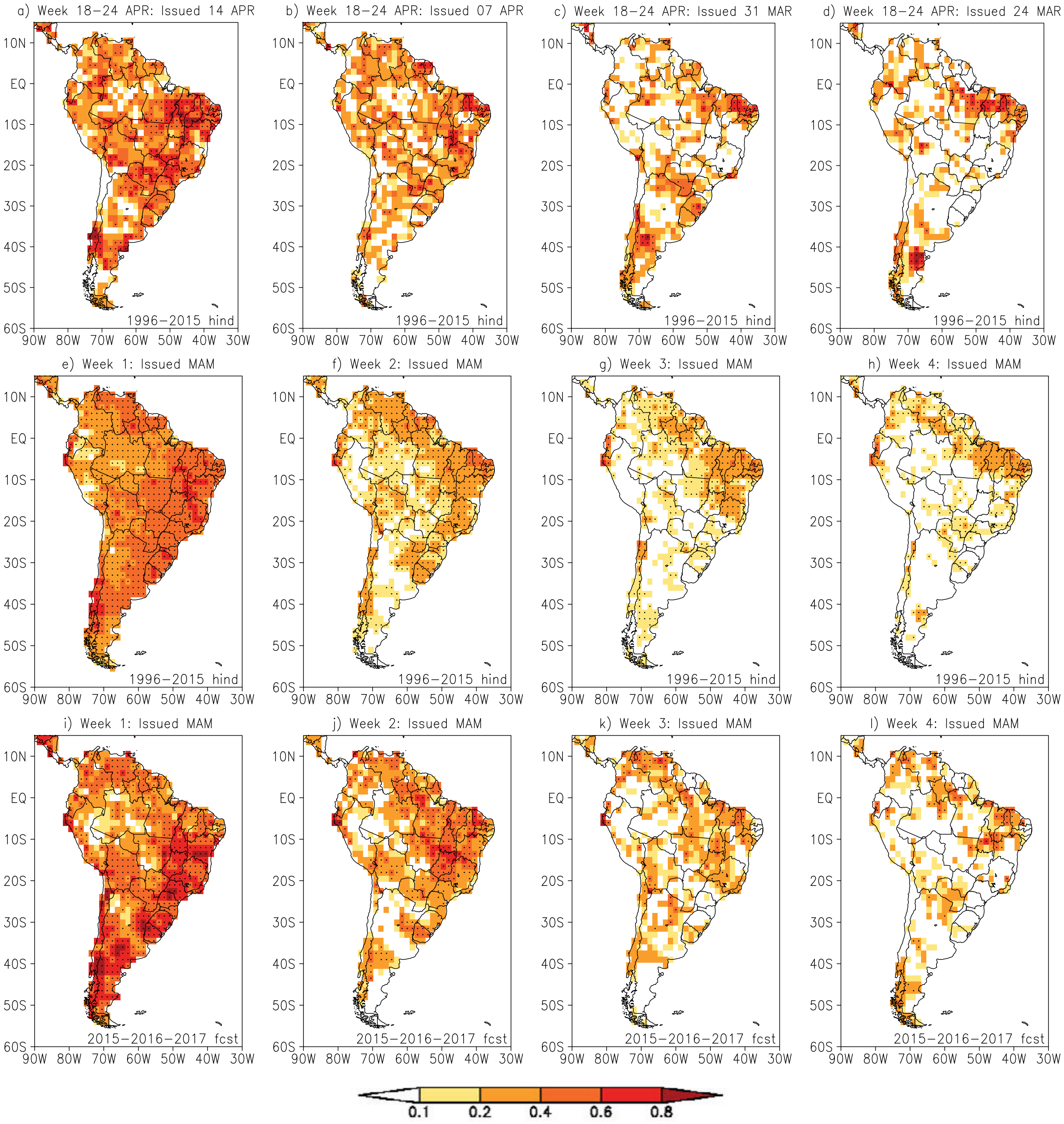

Figure 2: Maps of correlation between the ECMWF ensemble mean precipitation anomaly prediction produced one to four weeks in advance (1st to 4th columns) and the corresponding observed (CPC) precipitation anomalies at each grid point for (panels a to d) the target week hindcast verification sampling strategy (20 samples), (panels e to h) the all season hindcast verification sampling strategy (260 samples) and (panels i to l) the all season near real time forecast verification sampling strategy (39 samples) described in Section 2.3. ECMWF ensemble mean anomalies for the two hindcast verification sampling strategies were computed with respect to the 1996–2015 hindcast period produced with the 2016 model version in cross-validation (leaving one year out). ECMWF ensemble mean anomalies for the all season near real time forecast verification sampling strategy were computed using three different sets of hindcasts as follows: 1995–2014 hindcasts for the near real time forecast for 2015, 1996–2015 hindcasts for the near real time forecast for 2016, and 1997–2016 hindcasts for the near real time forecast for 2017, all representing the 20 years prior to the forecast year for which hindcasts were produced with the model versions available in 2015, 2016 and 2017, respectively. The dots mark grid points where the computed correlation coefficient was found to be statistically significantly different from zero at the 5 % level using a two-sided Student's *t* test.

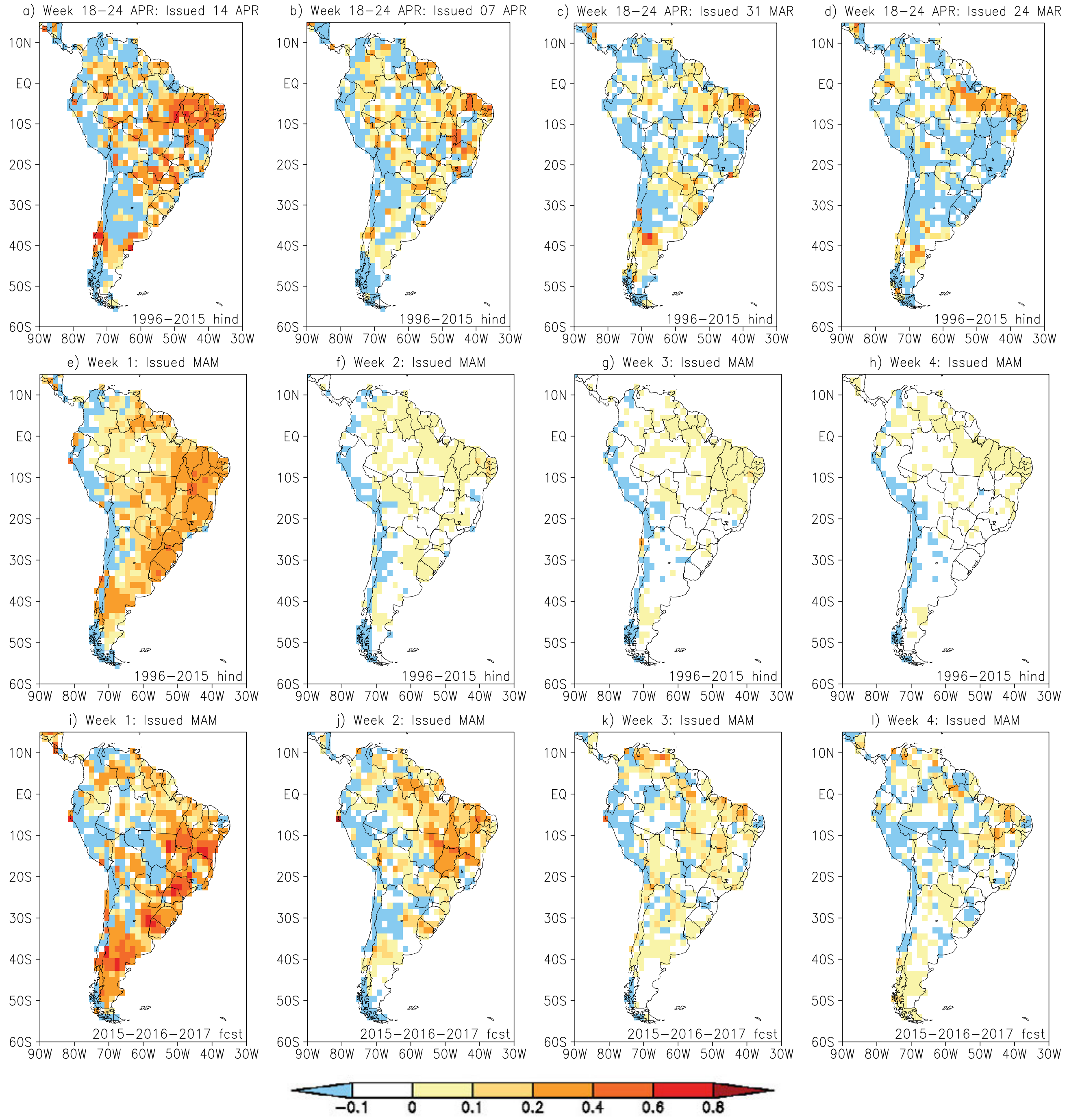

Figure 3: Maps of MSSS with respect to climatology the ECMWF ensemble mean precipitation anomaly predictions produced one to four weeks in advance (1st to 4th columns) for (panels a to d) the target week hindcast verification sampling strategy (20 samples), (panels e to h) the all season hindcast verification sampling strategy (260 samples) and (panels i to l) the all season near real time forecast verification sampling strategy (39 samples) described in Section 2.3.

(not shown), with the magnitudes of the MSS much reduced particularly for forecasts produced three and four weeks in advance, suggesting that the use of a larger ensemble size contributes for improving forecast accuracy for longer lead predictions.

Fig. 4 shows maps of the ratio of the predicted precipitation ensemble mean anomaly standard deviation and the observed precipitation anomaly standard deviation. Over most South America, for hindcasts/forecasts produced one to four weeks in advance, and for all three verification sampling strategies, this ratio is less than unity. This indicates that the predicted ensemble mean precipitation anomaly standard deviation is predominantly less than the observed precipitation anomaly standard deviation for all three levels of the proposed verification framework. In other words, there exists an ampli-

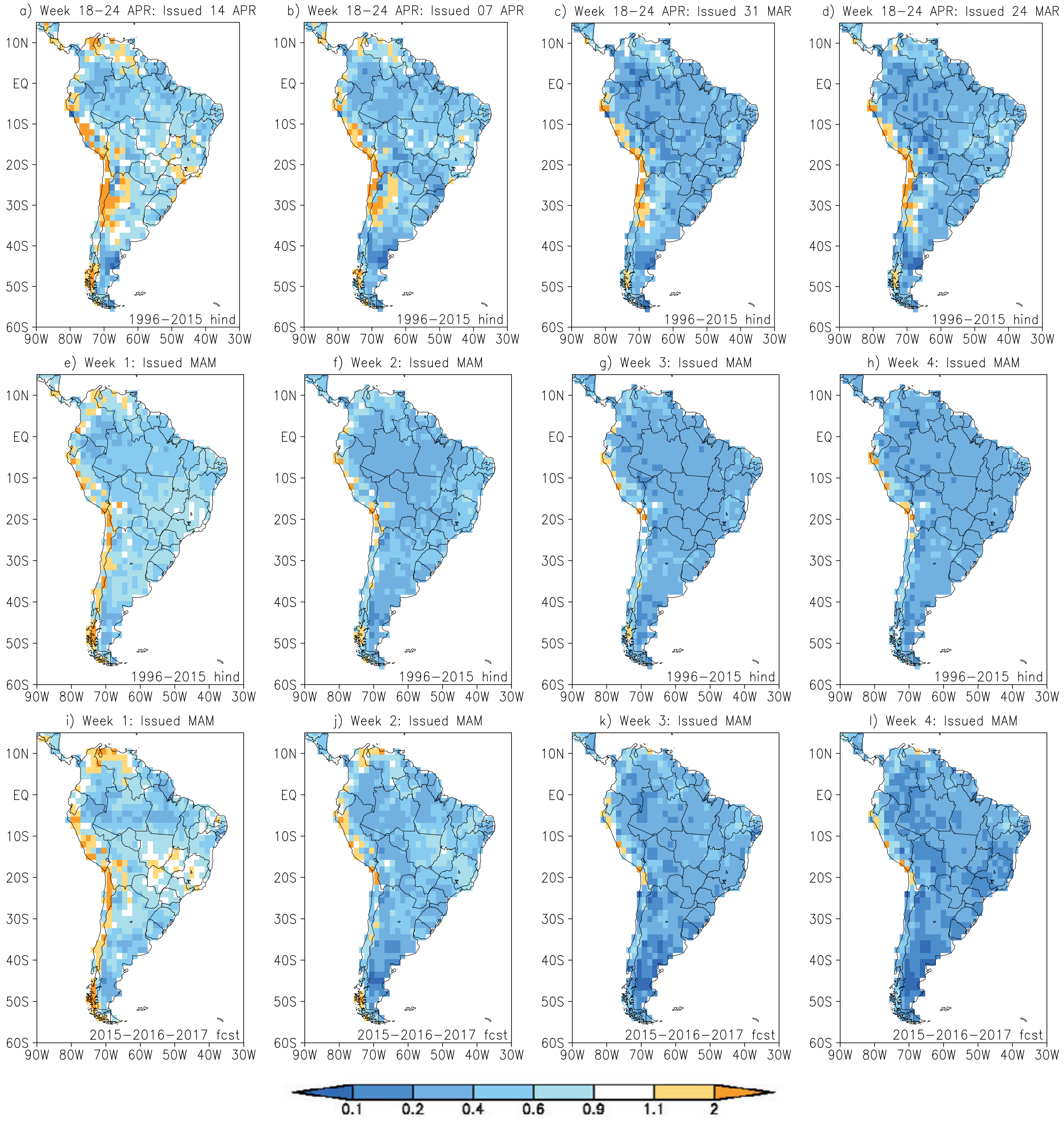

Figure 4: Maps of the ratio of the predicted ECMWF ensemble mean precipitation anomaly standard deviation and the observed precipitation anomaly standard deviation for predictions produced one to four weeks in advance (1st to 4th columns) for (panels a to d) the target week hindcast verification sampling strategy (20 samples), (panels e to h) the all season hindcast verification sampling strategy (260 samples) and (panels i to l) the all season near real time forecast verification sampling strategy (39 samples) described in Section 2.3.

tude error with the predicted ensemble mean anomalies presenting a reduced variability compared to the variability of the observed anomalies. A few exceptions to this predominant pattern are noticed mostly along the western South American boundary, where the opposite pattern is observed. The availability of a larger number of ensemble members for the third level compared to the other two levels makes the ensemble mean time series for the third level smoother (less variable) than for the other two levels. The effect of such a smoothing is noticed in the maps of Fig. 4 that show larger areas with smaller ratio for level three (panels i to l) when compared to the other two levels (panels a to h). This smoothing effect was further diagnosed by recomputing the standard deviation ratio maps for the third level using 11 ensemble members, which resulted in increased ratio values (not shown). The comparison of the correlation maps of Fig. 2 and the MSSS maps of Fig. 3 with the maps of Fig. 4 reveals that most regions presenting negative skill are those with reduced correlation coefficients and/or large amplitude errors.

3.2 Probabilistic (derived from ensemble members) prediction quality

In order to assess discrimination [the ability of the model in issuing forecast probabilities that successfully distinguish wet (positive precipitation anomaly) from dry (negative precipitation anomaly) events], Fig. 5 shows maps of the area under the ROC curve computed for hindcast/forecast probabilities for the occurrence of the event positive precipitation anomaly at each grid point for the three hindcasts/forecasts sampling strategies discussed in Section 2.3 produced one to four weeks in advance. Hindcast/forecast probabilities were derived from the available ensemble members and determined by computing the fraction of ensemble members indicating a positive precipitation anomaly. The ROC curve was constructed by plotting a graph of the hit rate against the false alarm rate, with these two rates computed from contingency tables obtained using as thresholds all issued hindcast/forecast probability values. The area under the ROC curve was computed by triangulation (applying the trapezium rule), and provide the probability of successfully discriminating (distinguishing) wet (positive precipitation anomaly) from dry (negative precipitation anomaly) events. As such events are binary (i.e. a wet/dry anomaly can occur or not occur, therefore having two possible outcomes) the reference probability of correctly discriminating (distinguishing) events from non-events is 50 %, and is represented by the area under the diagonal (45° line) of the ROC curve, known as the "no discrimination" line. Probabilistic hindcasts/forecasts that lead to a ROC curve falling near to or over this diagonal "no discrimination" line are those for which the distribution of issued probabilities for the occasions when the event of interest (positive precipitation anomaly) was observed is similar and overlaps with the distribution of issued probabilities for the occasions when the event of interest was not observed. Therefore in Fig. 5 the forecast probabilities issued for grid points with an area under the ROC curve larger than 0.5, which are colored, are better able to discriminate (distinguish) wet from dry events than unskillful forecasts with equal (50 %) probability of distinguishing (discriminating) events from non-events. Grid points marked with dots are those for which the computed area under the ROC curve was found to be significantly different from 0.5 at the 5 % confidence level. For these grid points the distribution of issued probabilities for the occasions when the event of interest (positive precipitation anomaly) was observed is shifted with respect to the distribution of issued probabilities for the occasions when the event of interest was not observed. Fig. 5 shows that the discrimination ability is larger for short lead predictions produced one to two weeks in advance (with most South America presenting the area under the ROC curve larger than 0.6) than for longer lead predictions produced three to four weeks in advance (with a reduced portion of South America presenting the area under the ROC curve larger than 0.6), although over some regions (e.g. parts of north and northeast Brazil) a similar level of discrimination ability in maintained for the forecasts produced three to four weeks in advance.

The spatial pattern of the maps shown in Fig. 5 for the first (panels a to d) and second (panels e to h) levels of the proposed verification framework are largely consistent, reconfirming that even with the much reduced sample size of the first level compared to the second level a meaningful (this time probabilistic) hindcast quality assessment is achieved. The maps for the third level (panels i to l) are also largely consistent with the maps of the other two levels (panels a to h), except in parts of northern Brazil where a reduction in discrimination ability is noticed. These differences are likely to be due to the modulation of specific interannual variability during the three years investigated in the third level. Recomputing the maps for the third level but now using 11 ensemble members resulted in similar spatial patterns with comparable magnitudes to those shown in panels i to l (not shown), suggesting that the verification sample size needs to be expanded to include additional years in order to effectively demonstrate the benefit of the larger number of ensemble members when assessing discrimination of near real time forecasts.

Fig. 6 shows ROC curves for ensemble derived probabilistic predictions issued for the event positive precipitation anomaly collected over all South American grid points, aggregating all available hindcasts/forecasts in space and time. This procedure allows assessment of overall discrimination for the hindcasts/forecasts produced one to four weeks in advance for the target week hindcast verification (panels a to d), the all season hindcast verification (panels e to h), and the all season near real time forecast verification (panels i to l) sampling strategies. The larger area under the ROC curve (around 0.72 and 0.63) for hindcasts/forecasts produced one to two weeks in advance compared to (around 0.58 and 0.56) for hindcasts/forecasts produced three to four weeks in advance corroborates the previous assessment discussed and described above that indicated higher discrimination ability for shorter lead predictions when compared to longer lead predictions. These results are consistent among the three levels of the proposed verification framework as the ROC curves and areas are similar for the three hindcast/forecast sampling strategies. Addressing sampling uncertainty in the computed scores is an important aspect in hindcast/forecast verification practice. In order to address sampling uncertainty in the computed area under the ROC curve scores Table 1 shows the 95 % confidence intervals for the three verification sampling strategies here investigated for the hindcasts/forecasts produced one to four weeks in advance.

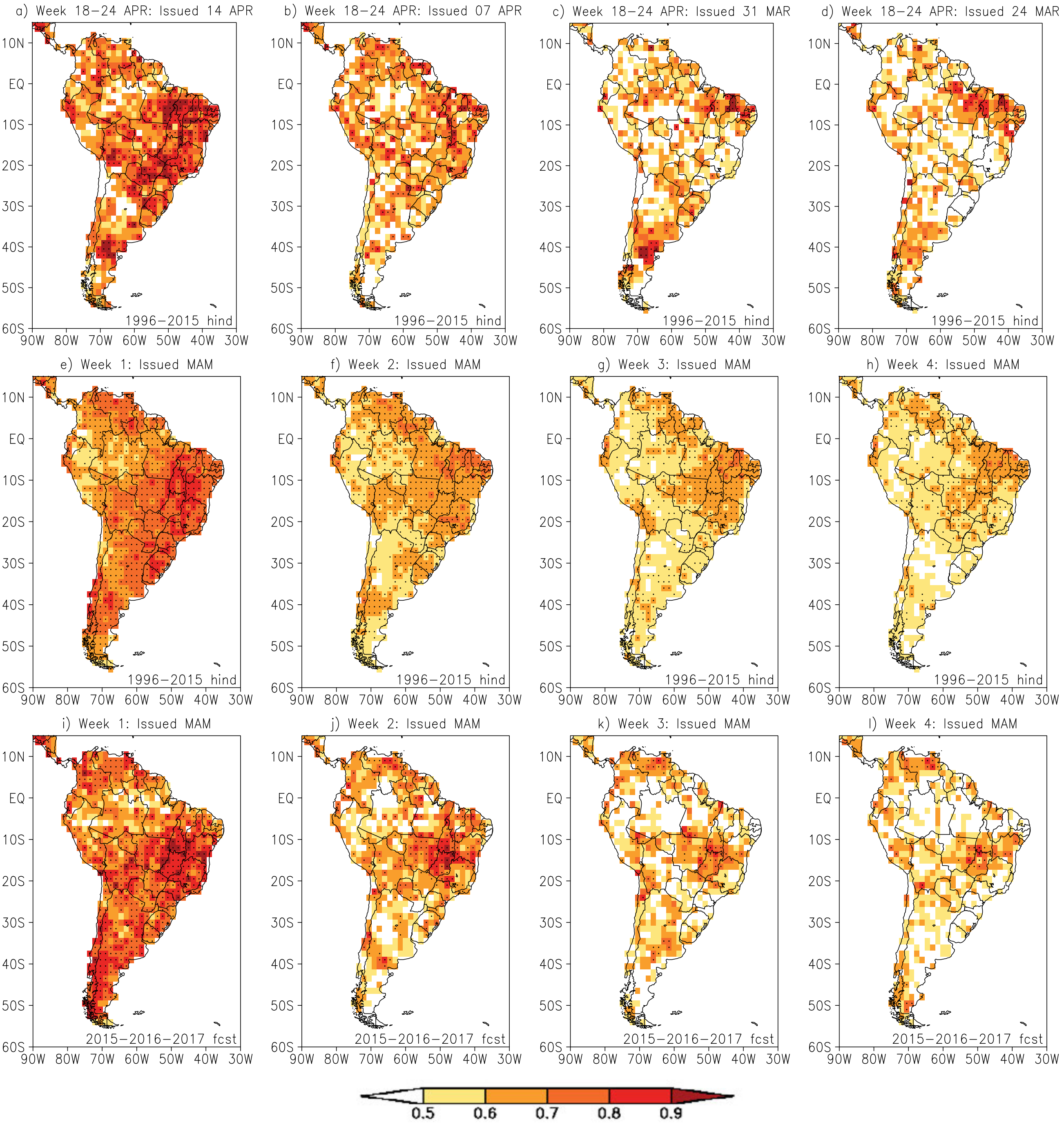

Figure 5: Maps of area under the ROC curve computed for ECMWF forecast/hindcast probabilities for the occurrence of the event positive precipitation anomaly produced one to four weeks in advance (1st to 4th columns) at each grid point for (panels a to d) the target week hindcast verification sampling strategy (20 samples), (panels e to h) the all season hindcast verification sampling strategy (260 samples) and (panels i to l) the all season near real time forecast verification sampling strategy (39 samples) described in Section 2.3. Forecast/hindcast probabilities were derived using the available ensemble members for each sampling strategy and determined by computing the fraction of ensemble members indicating a positive precipitation anomaly. ECMWF probabilities for the two hindcast verification sampling strategies were computed with respect to the 1996–2015 hindcast period produced with the 2016 model version in cross-validation (leaving one year out). ECMWF probabilities for the all season near real time forecast verification sampling strategy were computes using three different sets of hindcasts as follows: 1995–2014 hindcasts for the near real time forecast for 2015, 1996–2015 hindcasts for the near real time forecast for 2016, and 1997–2016 hindcasts for the near real time forecast for 2017, all representing the 20 years prior to the forecast year for which hindcasts were produced with the model versions available in 2015, 2016 and 2017, respectively. The dots mark grid points where the computed area under the ROC curve was found to be significantly different from 0.5 at the 5 % confidence level using a statistical hypothesis test based on the Mann–Whitney U-distribution.

Table 1: ECMWF ROC area for ensemble derived probabilistic predictions issued for the event positive precipitation anomaly collected over all South American grid points, aggregating all available forecasts/hindcasts in space and time, produced one to four weeks in advance (2nd to 5th columns), with the corresponding 95 % confidence intervals (in brackets) estimated from a 1000 bootstrap resampling procedure, for the target week hindcast verification (level 1), the all season hindcast verification (level 2) and the all season near real time forecast verification (level 3) sampling strategies described in Section 2.3.

Verification strategy	**Week 1**	**Week 2**	**Week 3**	**Week 4**
Level 1	0.714 (0.706, 0.722)	0.632 (0.623, 0.641)	0.583 (0.574, 0.592)	0.562 (0.553, 0.573)
Level 2	0.719 (0.716, 0.721)	0.626 (0.624, 0.629)	0.597 (0.594, 0.600)	0.582 (0.579, 0.584)
Level 3	0.732 (0.726, 0.738)	0.631 (0.624, 0.638)	0.576 (0.569, 0.583)	0.563 (0.555, 0.570)

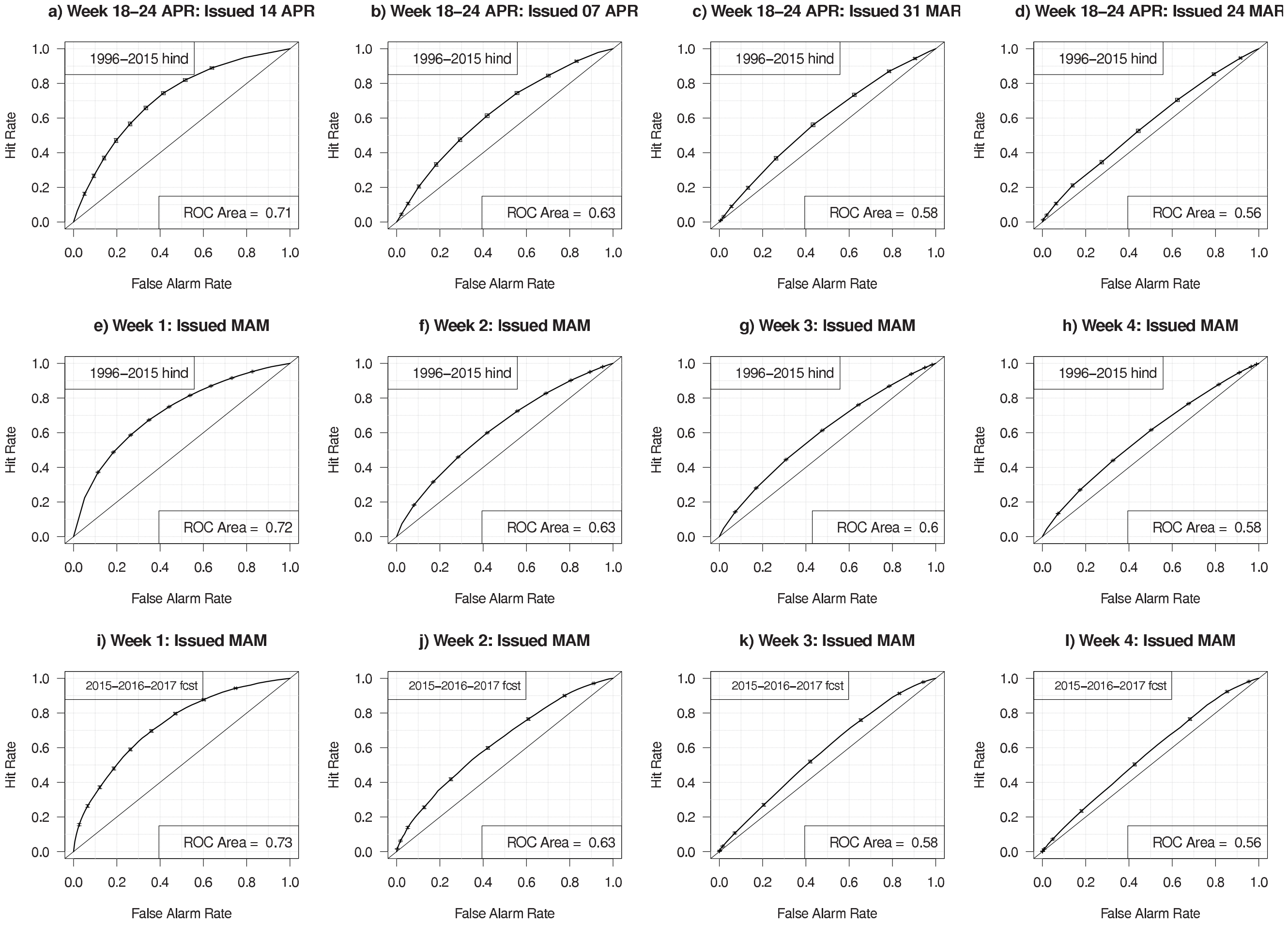

Figure 6: ECMWF ROC curves for ensemble derived probabilistic predictions issued for the event positive precipitation anomaly collected over all South American grid points, aggregating all available forecasts/hindcasts in space and time, produced one to four weeks in advance (1st to 4th columns), for (panels a to d) the target week hindcast verification sampling strategy (20 samples), (panels e to h) the all season hindcast verification sampling strategy (260 samples) and (panels i to l) the all season near real time forecast verification sampling strategy (39 samples) described in Section 2.3. Forecast/hindcast probabilities were derived using the available ensemble members for each sampling strategy and determined by computing the fraction of ensemble members indicating a positive precipitation anomaly. ECMWF probabilities for the two hindcast verification sampling strategies were computed with respect to the 1996–2015 hindcast period produced with the 2016 model version in cross-validation (leaving one year out). ECMWF probabilities for the all season near real time forecast verification sampling strategy were computes using three different sets of hindcasts as follows: 1995–2014 hindcasts for the near real time forecast for 2015, 1996–2015 hindcasts for the near real time forecast for 2016, and 1997–2016 hindcasts for the near real time forecast for 2017, all representing the 20 years prior to the forecast year for which hindcasts were produced with the model versions available in 2015, 2016 and 2017, respectively. The bars displayed on top of the ROC curves represent the 95 % confidence intervals for the hit rates and false alarm rates constructed from 1000 bootstrap samples extracted (with replacement) from the available sample.

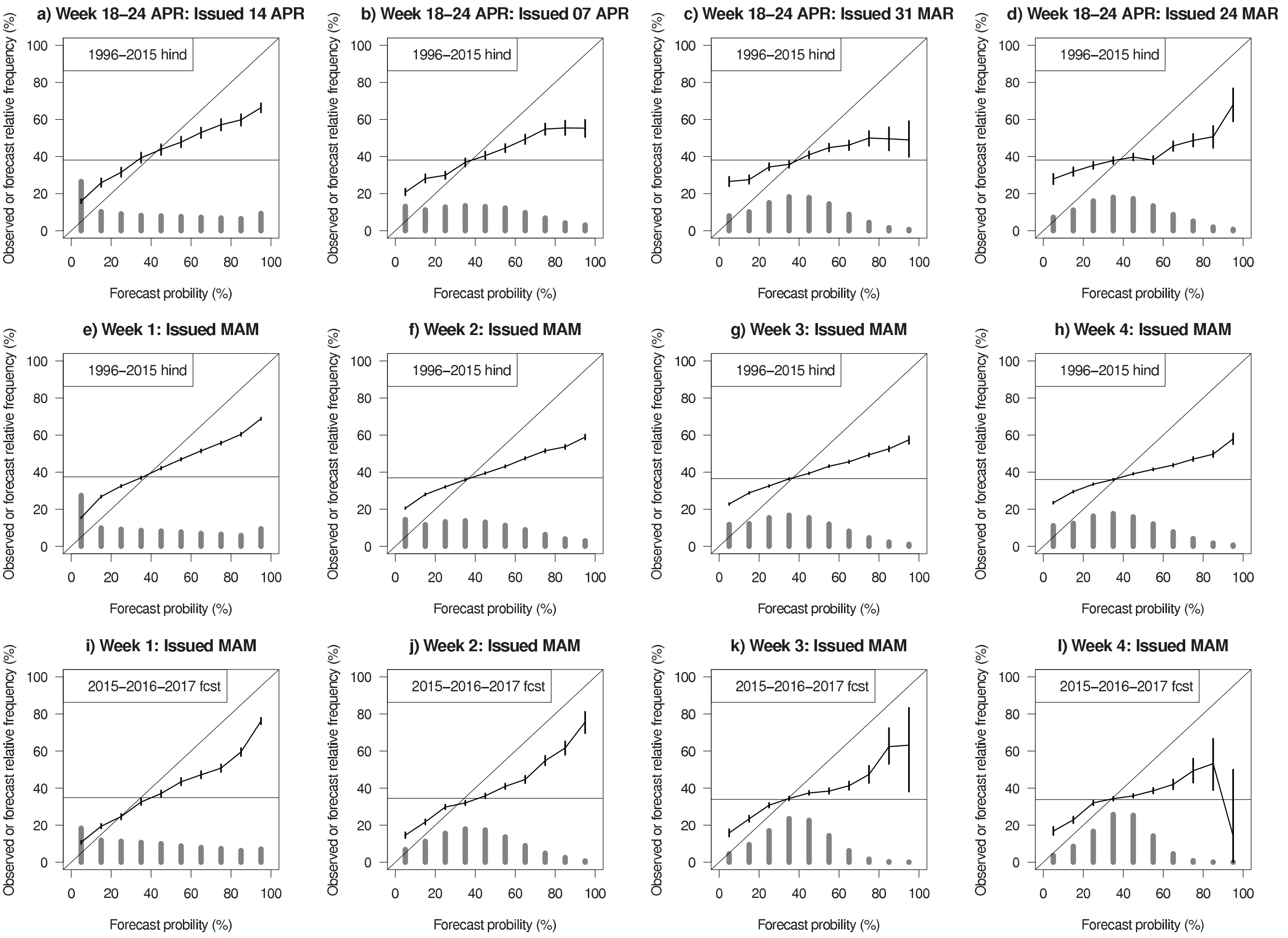

Figure 7: ECMWF reliability diagrams for ensemble derived probabilistic predictions issued for the event positive precipitation anomaly collected over all South American grid points, aggregating all available forecasts/hindcasts in space and time, produced one to four weeks in advance (1st to 4th columns), for (panels a to d) the target week hindcast verification sampling strategy (20 samples), (panels e to h) the all season hindcast verification sampling strategy (260 samples) and (panels i to l) the all season near real time forecast verification sampling strategy (39 samples) described in Section 2.3. Forecast/hindcast probabilities were derived using the available ensemble members for each sampling strategy and determined by computing the fraction of ensemble members indicating a positive precipitation anomaly. ECMWF probabilities for the two hindcast verification sampling strategies were computed with respect to the 1996–2015 hindcast period produced with the 2016 model version in cross-validation (leaving one year out). ECMWF probabilities for the all season near real time forecast verification sampling strategy were computes using three different sets of hindcasts as follows: 1995–2014 hindcasts for the near real time forecast for 2015, 1996–2015 hindcasts for the near real time forecast for 2016, and 1997–2016 hindcasts for the near real time forecast for 2017, all representing the 20 years prior to the forecast year for which hindcasts were produced with the model versions available in 2015, 2016 and 2017, respectively. The vertical bars displayed on top of the reliability curves represent the 95 % confidence intervals for the observed relative frequencies constructed from 1000 bootstrap samples extracted (with replacement) from the available sample.

These intervals were estimated from 1000 bootstrap samples extracted from the available samples. This procedure allows the computation of 1000 values of the area under the ROC curve by re-sampling (with replacement) the available hindcasts/forecasts. The 95 % confidence intervals were then estimated using the 2.5th and 97.5th percentiles of the resulting empirical distribution constructed with the 1000 computed values of the area under the ROC curve. The aggregation of hindcasts/forecasts over space and time resulted in considerably large samples for computing the area under the ROC curves leading to relatively small confidence intervals as reported in Table 1.

As the final assessment of the proposed verification framework for South America sub-seasonal precipitation predictions, Fig. 7 shows reliability diagrams for ensemble derived probabilistic predictions issued for the event positive precipitation anomaly collected over all South American grid points, aggregating all available hindcasts/forecasts in space and time for the three verification sampling strategies. For the construction of the diagrams shown in Fig. 7 the verification sample was first binned according to the issued hindcast/forecast probabilities, and next it was computed the observed event frequency for all of the hindcast/forecast probabilities in each bin. The reliability diagram is a graph of the issued hindcast/forecast probabilities plotted against the corresponding observed frequencies for each bin. Ten bins were used (0 to 10 %, 10 to 20 %, 20 to 30 %, 30 to 40 %, 40 to 50 %, 50 to 60 %, 60 to 70 %, 70 to 80 %, 80 to 90 % and 90 to 100 %). The points were plotted at the mid points of the ten bins. This diagram provides a graphical interpretation of probabilistic forecast quality in terms of reliability (how well forecast probabilities match the observed frequency of the event of interest) and resolution (how the observed frequency varies when the data sample is stratified by the hindcast/forecast probabilities). Perfectly reliable hindcasts/forecasts should result in a diagram represented by a 45° diagonal line where the issued probabilities exactly match the observed frequencies (for example, the event must occur on 30 % of the occasions that the 30 % forecast probability was issued). Well resolved hindcasts/forecasts should also result in a diagram represented by a 45° diagonal line because at this diagonal line the observed frequencies are well distinct from the climatological frequency of occurrence of the event of interest, represented by the horizontal line in the reliability diagram. The horizontal line illustrates fully unresolved hindcasts/forecasts with the same observed frequency regardless of the hindcasts/forecasts probabilities.

Fig. 7 shows that the black lines in the reliability diagrams for all three hindcast/forecast sampling strategies here considered are away from the perfect reliability diagonal (45°) line, indicating lack of reliability in the issued hindcast/forecast probabilities. For example, when let's say high 70–80 % hindcast/forecast probabilities were issued for the occurrence of positive precipitation anomalies, in fact positive precipitation anomalies were observed in much less than 70–80 % of the occasions. Similarly, when low 10–20 % hindcast/forecast probabilities were issued for the occurrence of positive precipitation anomalies, in fact positive precipitation anomalies were observed in more than 10–20 % of the occasions. This is the so-called overconfidence feature also commonly identified in seasonal forecasts usually due to the fact that the ensemble spread is not large enough to encompass the observations. The larger the distance between the black line and the diagonal line the worse is the reliability of the issued hindcast/forecast probabilities. Therefore, the area between these two lines can be used to assess reliability. The smaller this area/distance the better is the reliability of the issued probabilities. Fig. 7 shows that for all lead times and sampling strategies the issued hindcast/forecast probabilities do not match the observed frequencies, all presenting the over confidence feature described above. This demonstrates the need for applying a procedure for calibrating the hindcast/forecast probabilities to make them more reliable. The vertical bars displayed on top of the reliability curves represent the 95 % confidence intervals for the observed relative frequencies constructed from 1000 bootstrap samples extracted (with replacement) from the available samples. Note that the confidence intervals for the first (panels a to d) and third (panels i to l) levels of the proposed verification framework are larger than for the second level (panels e to h) due to the fact that the second level has a much larger sample of aggregated hindcasts/forecasts over space and time than the other two levels.

Fig. 7 also illustrates that the black lines tend to be tilted towards the horizontal "no resolution" line, suggesting that the hindcasts/forecasts have poor resolution. The smaller the distance between the black line and the horizontal line the worse is the resolution of the issued hindcast/forecast probabilities. Therefore, the area between these two lines can be used to assess resolution. The larger this area/distance the better is the resolution of the issued probabilities. Fig. 7 shows that, for all three verification sampling strategies, resolution is generally poorer for hindcasts/forecasts produced three to four weeks in advance than for hindcasts/forecasts produced one to two weeks in advance.

The vertical bars in the form of a histogram at the bottom of each panel in Fig. 7 represent the percentage of the issued forecast probabilities falling into each of the ten probability bins used for constructing the reliability diagrams. This histogram is known as the "sharpness diagram". Sharp forecasts have *u*-shaped histograms with high frequencies for near 0 and 100 % forecast probabilities. Fig. 7 shows that sub-seasonal precipitation hindcasts/forecasts produced two to four weeks in advance are not particularly sharp, with slight better sharpness noticed for hindcasts/forecasts produced one week in advance. These features are largely consistent for the assessment here performed with the three verification sampling strategies.

4 Summary and discussions

This paper proposed a verification framework for South American sub-seasonal precipitation predictions produced one to four weeks in advance. Due to the complexity and large degree of differences in some characteristics of currently available datasets for sub-seasonal hindcast and near real time forecast verification, which directly impact the verification sample sizes, such a framework is both attractive and necessary for advancing the developments in this new research area. The proposed framework was designed to assess hindcast/forecast quality focusing on a selection of the most fundamental attributes (association, accuracy, discrimination, reliability and resolution). These attributes were measured using various deterministic and probabilistic verification scores in order to provide a complete hindcast/forecast quality assessment. This attribute-based framework allows the production of verification information for the assessment of hindcast/forecast quality in three levels according to the availability of sub-seasonal hindcasts and near real time forecasts samples. The three information levels were referred to as target week hindcast verification, all season hindcast verification, and all season near real time forecast verification.

The first level (target week hindcast verification) used a similar sampling strategy used in seasonal hindcast verification, where verification is targetted to a particular seasonal of interest, and in the framework here proposed it was target to a particular week of interest. In this first level, the sample size for performing verification was limited to the number of years for which hindcasts were produced (20 years for the South American sub-seasonal predictions discussed in this paper). However, it is worth noting that for models with short hindcast periods (e.g. less than 15 years) it may not be appropriate to generate this first level of verification information due to the excessively reduced sample for a meaningful assessment. Although the hindcast sample size was considerably reduced in this first level compared to the other two levels of the proposed verification framework, the results reported in this study indicated that such an assessment did provide valuable, comparable and meaningful information about regions where the predictions presented good quality.

The second level (all season hindcast verification) considerably expanded the sample size by considering hindcasts produced in different weeks within a season presenting common atmospheric features, allowing increasing the robustness of the performed hindcast quality assessment. For the South American sub-seasonal predictions discussed in this paper, focused on the austral autumn season, the sample size was increased to 260 samples. One may question if the aggregation performed in this second level is seasonally specific enough for the verification assessment to be considered physically meaningful. As indicated in Section 2.2 such aggregation was motivated by fact that this sample contains hindcasts initialized over the austral autumn season, which is a period marked by similar atmospheric features in various South American regions. The second level of the proposed verification framework can therefore be considered seasonally specific enough for the verification quality assessment performed in this study to be considered physically meaningful.

The third level (all season near real time forecast verification) was based on the collection of near real time forecasts for a selected number of years, providing an opportunity for comparing forecasts with hindcasts quality. For the South American precipitation example discussed in this paper the sample size for this third level of verification information was intermediate (39 samples, representing real time forecasts produced for three years, 2015, 2016 and 2017) between the relatively small sample used for the first level (20 samples) and the considerably large sample of size 260 for the second level. Although the spatial patterns of the computed scores and the verification plots for the third level were found to be largely coherent with the other two levels, some differences were noticed in northern Brazil and Argentina. These differences were suggested to be associated with the modulation of specific interannual variability during the three years investigated in the third level. The number of available ensemble members for the assessment performed in the third level was considerably larger (51 members) compared to the number of ensemble member available and used in the first and second levels (11 members) of the proposed verification framework. In order to investigate the effect of this larger ensemble size, the scores were re-computed for the third level using 11 ensemble members for consistency with the number of ensemble members used in the first two levels. The resulting spatial patterns of the re-computed scores for the third level remained similar to the patterns of the other two levels, although the magnitudes of the computed deterministic and probabilistic scores were comparable/slightly reduced particularly for forecasts produced three and four weeks in advance. This suggests that the use of a larger ensemble size has the potential to contribute for improving forecast quality. However, we hypothesize that increasing the sample size by including additional forecast years for the third level when more real time forecasts become available in the future is likely to make more apparent the impact of increased ensemble size on forecast quality.

The proposed framework allowed answering the questions initially posed in Section 2.1. In order to answer these questions maps of correlation between ensemble mean hindcasts/forecasts and observed anomalies, maps of the area under the ROC curve, ROC curves and reliability diagrams for the event positive precipitation anomaly were constructed using ECMWF hindcasts/forecasts available through the S2S project database for all three verification information levels of the proposed framework. An outstanding and unique feature of the S2S project database worth emphasizing is the availability of both hindcasts and realtime forecasts for a number of models in a single platform, which provides an extraordinary rich database facilitating and enabling the implementation of the proposed verification framework using various sampling strategies.

As indicated in Section 2.2, the three level strategy of the proposed verification framework has strengths and weaknesses, making the decision and choice a single approach out of the three options challenging. It is therefore important to recognize the differences, merits and limitations of the three approaches and, most importantly, consider then as complementary verification strategies. Overall, for the investigation carried out in this study, reasonable hindcast/forecast quality accordance was identified across the three levels of verification information produced, illustrating the complementarity of the performed assessment. The ECMWF subseasonal precipitation hindcasts/forecasts produced one to two weeks in advance presented better performance than those produced three to four weeks in advance. ENSO is likely playing a substantial role in the performance of predictions produced three to four weeks in advance investigated in this study. The MJO is another important source of sub-seasonal precipitation predictability (Li and Robertson, 2015). The large scale circulation and atmospheric teleconnections associated with these two phenomenon either favor periods with enhanced vertical motion and increased precipitation or periods with inhibited vertical motion (subsidence) and reduced precipitation, particularly over tropical South America (Grimm and Ambrizzi, 2009; Gonzalez and Vera, 2014). The northeast region of Brazil, which is affected by both ENSO and MJO, consistently presented favorable sub-seasonal precipitation prediction performance through the computed verification scores, specifically in terms of association, accuracy and discrimination attributes. This region was therefore identified as a region where sub-seasonal predictions produced one to four weeks in advance with the ECMWF model are most likely to be successful in South America. When aggregating all hindcasts/forecasts over the South American continent the assessed probabilistic predictions showed modest discrimination ability, with hindcasts/forecasts clearly requiring calibration for improving reliability and possibly combination with forecasts produced by other models for improving resolution.

The assessment of ECMWF forecasts produced for the week 18–24 April 2016 indicated that the model was able to signalize up to four weeks in advance the possibility for occurrence of dry conditions over the northeast region of Brazil and occurrence of wet conditions over parts of northeast Argentina, Uruguay and southern Brazil and Peru. However, the model failed to indicate the possibility for occurrence of wet conditions over Venezuela, Guyana, Suriname and French Guyana, particularly three to four weeks in advance.

The same verification scores used in this paper are commonly used in weather and seasonal hindcast/forecast verification for assessing the association, accuracy, discrimination, reliability and resolution attributes. This suggests that forecast verification practice is naturally moving towards the seamless concept in terms of common metrics and attributes being assessed across different time scales.

The availability of both hindcasts and near real time forecasts through the S2S project database allows the implementation of the proposed three level verification framework in support of future routine sub-seasonal prediction practice, by helping forecasters to identify model forecast merits and deficiencies and regions where to best trust the model guidance information. However, the proposed framework is useful not only in support of sub-seasonal prediction practice, but also to provide feedback to model developers in identifying strengths and weaknesses for future sub-seasonal predictions systems improvements. Although the framework proposed here was illustrated with South American sub-seasonal precipitation predictions, it is also applicable for other regions and variables. Note also that the proposed verification framework is not limited to the probabilistic predictions used as illustration in this paper, but it is also applicable to the very popular tercile category probabilistic forecasts often issued by the seasonal and sub-seasonal forecast communities. The proposed framework has potential for stimulating the implementation of coordinated sub-seasonal prediction verification practice following pre-defined protocols for the use of the available hindcasts and near real time forecasts, and also for promoting and advancing research in this new verification area.

Acknowledgments

The WWRP/WCRP S2S project and ECMWF are acknowledged for developing the S2S project database and making available the model hindcast and near real time forecast data used in this study. CASC thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), process 304586/2016-1, and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), process 2015/50687-8 (CLIMAX Project) for the support received for the development of this research. Two anonymous reviewers are acknowledged for providing valuable comments, suggestions and constructive criticism that substantially contributed for improving the quality of this manuscript.

References

Chen, M., W. Shi, P. Xie, V.B. S. Silva, V.E. Kousky, R. Wayne Higgins, J.E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation, – J. Geophys. Res. **113**, D04110, DOI:10.1029/2007JD009132.

Gonzalez, P.L.M., C. Vera, 2014: Summer precipitation variability over South America on long and short intraseasonal timescales. – Climate Dyn. **43**, 7–8.

Grimm, A.M., T. Ambrizzi, 2009: Teleconnections into South America from the tropics and extratropics on interannual and intraseasonal timescales. – In: Past Climate Variability in South America and Surrounding Regions. – Springer Netherlands, 159–191.

Hudson, D., O. Alves, H.H. Hendon, A.G. Marshall, 2011: Bridging the gap between weather and seasonal forecasting: Intraseasonal forecasting for Australia. – Quart. J. Roy. Meteor. Soc. **137**,673–689, DOI:10.1002/qj.769, http://onlinelibrary.wiley.com/doi/10.1002/qj.769/abstract.

Hudson, D., A.G. Marshall, Y. Yin, O. Alves, H.H. Hendon, 2013: Improving intraseasonal prediction with a new ensemble generation strategy. – Mon. Wea. Rev. **141**, 4429–4449. DOI:10.1175/MWR-D-13-00059.1.

Li, S., A.W. Robertson, 2015: Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. – Mon Wea. Rev **143**, 2871–2889. https://journals.ametsoc.org/doi/abs/10.1175/MWR-D-14-00277.1

Mason, I., 1982: A model for assessment of weather forecasts. – Aust. Meteor. Mag. **30**, 291–303.

Mason, S.J., N.E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves. – Statistical significance and interpretation. – Quart. J. Roy. Meteor. Soc. **128**, 2145–2166.

Murphy, A.H., 1988: Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. – Mon. Wea. Rev. **116**, 2417–2424.

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. – Wea. Forecast. **8**, 281–293.

Robertson, A.W., A. Kumar, M. Peña, F. Vitard, 2015: Improving and promoting subseasonal to seasonal prediction. – Bull. Amer. Meteor. Soc. **96**, ES49–ES53.

Vitart, F., A.W. Robertson, D.L.T. Anderson, 2012: Subseasonal to Seasonal Prediction Project: Bridging the gap between weather and climate. – Bull. World Meteor. Org. **61**, 23.

Vitart, F., C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti, E. Fucile, M. Fuentes, H. Hendon, J. Hodgson, H. -S. Kang, A. Kumar, H. Lin, G. Liu, X. Liu, P. Malguzzi, I. Mallas, M. Manoussakis, D. Mastrangelo, C. MacLachlan, P. McLean, A. Minami, R. Mladek, T. Nakazawa, S. Najm, Y. Nie, M. Rixen, A.W. Robertson, P. Ruti, C. Sun, Y. Takaya, M. Tolstykh, F. Venuti, D. Waliser, S. Woolnough, T. Wu, D. -J. Won, H. Xiao, R. Zaripov, L. Zhang, 2017: The Subseasonal to Seasonal (S2S) Prediction Project Database. – Bull. Amer. Meteor. Soc. **98**, 163–173, DOI: 10.1175/BAMS-D-16-0017.1.

Weigel, A., D. Baggenstos, M.A. Liniger, F. Vitart, C. Appenzeller, 2008: Probabilistic verification of monthly temperature forecasts. – Mon. Wea. Rev. **136**, 5162–5182. DOI: 10.1175/2008MWR2551.1.

Wheeler, M.C., H. Zhu, A.H. Sobel, D. Hudson, F. Vitart, 2017: Seamless precipitation prediction skill comparison between two global models. – Quart. J. Roy. Meteor. Soc. **143**, 374–383, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2928.

Zhu, H., M.C. Wheeler, A.H. Sobel, D. Hudson, 2014: Seamless Precipitation Prediction Skill in the Tropics and Extratropics from a Global Model. – Mon. Wea. Rev. **142**, 1556–1569.