METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **20**: 236–245 (2013)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/met.1407

RMetS
Royal Meteorological Society

Comparative skill assessment of consensus and physically based tercile probability seasonal precipitation forecasts for Brazil

Caio A. S. Coelho*

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

ABSTRACT: This study aimed to perform a comparative skill assessment of consensus and physically based tercile probability seasonal precipitation forecasts for Brazil produced during the last decade. Two fundamental forecast attributes have been examined: discrimination and reliability. The discrimination assessment revealed that forecast quality is seasonally dependent and that consensus and physically based forecasts are complementary. During spring and summer consensus forecasts were generally found to have better discrimination ability than physically based forecasts. During autumn and winter physically based forecasts were found to have better discrimination ability than consensus forecasts. However, discrimination is a necessary but not sufficient forecast skill attribute, and therefore only provides indication of potential forecast quality provided forecasts are reliable (i.e. well calibrated). The analysis of tendency diagrams has revealed that both consensus and physically based forecasts suffer from systematic errors (biases) for the three forecast categories. Both forecasts under-forecasted the below-normal category and over-forecasted the above normal category. This over-forecasting feature was stronger for physically based forecasts when compared to consensus forecasts. The normal category was more severely over-forecast for consensus forecasts when compared to physically based forecasts. The assessment through the computation of the reliability component of the Brier Score has revealed that consensus forecasts are better calibrated than CPTEC/AGCM physically based forecasts.

KEY WORDS seasonal forecasting; consensus; verification; tercile probabilities

Received 14 September 2012; Revised 14 February 2013; Accepted 27 March 2013

1. Introduction

Tercile probability consensus seasonal (3 month mean) precipitation forecasts for Brazil for the categories below normal, normal and above normal precipitation have been produced since early 2000 and displayed every month as spatial maps by the Centre for Weather Forecasts and Climate Studies (CPTEC) of the Brazilian National Institute for Space Research (INPE). These forecasts are currently produced as a joint collaborative effort between CPTEC, the National Meteorological Service of Brazil (INMET) and meteorological offices of various regions in Brazil. The production process of these forecasts consists of three steps:

- diagnostics of global and regional weather and climate conditions recently observed. These diagnostics have particular emphasis on (1) surface and subsurface ocean temperature conditions; (2) the corresponding tropical convective activity; and (3) the associated high and low level atmospheric circulation response;
- examination of forecasts produced by physically based global and regional dynamical models and by empirical (statistical) models. The dynamical models include atmospheric-only models forced with forecast and persisted sea surface temperatures, coupled ocean–atmosphere models and regional models, the latter performing downscaling of the forecast produced by global atmospheric models. The empirical models use the most recently observed sea surface temperature conditions in the Pacific and Atlantic oceans as predictors for precipitation over Brazil for the next 3 month season; and
- use of climate expertise of all partners involved in this process to define the forecast. The diagnostics and model forecast information previously examined, including retrospective skill assessment, is used to determine subjectively the consensus tercile probability forecast for Brazil. In this final step, forecast areas (i.e. regions where forecasts are issued) and the probabilities assigned to each tercile category are determined by consensus agreement among partners.

Despite the drawbacks of consensus seasonal forecasts (e.g. subjectivity when defining both forecast areas and tercile forecast probabilities), these forecasts can potentially be more skillful than climate model physically based tercile probability seasonal forecasts. This is because the consensus process aggregates climate expertise with climate model forecast information when preparing the final forecast. This forecast is sometimes also referred to as outlook. However, Berri *et al*. (2005) assessed the skill of climate outlook fora's consensus seasonal precipitation tercile probability forecasts for southeast South America during 1998–2002 against physically based seasonal forecasts and concluded that consensus forecasts do not add skill to physically based forecasts. Nevertheless the reduced number of forecasts (only 16) used in this assessment made it difficult to draw definitive conclusions about the skill of consensus seasonal forecasts in this region. Moreover, forecasts for different seasons have been aggregated to compose the sample

* Correspondence: C. A. S. Coelho, Centro de Previsão de Tempo e Estudos Climáticos (CPTEC), Instituto Nacional de Pesquisas Espaciais (INPE), Rodovia Presidente Dutra, Km 40, SP-RJ, 12630-000 Cachoeira Paulista, SP, Brazil. E-mail: caio.coelho@cptec.inpe.br

of 16 forecasts for verification, preventing the investigation of seasonal dependence in forecast skill.

Given the availability of a considerably larger sample of 120 never previously verified forecasts (i.e. a decade of consensus-based seasonal precipitation forecasts for Brazil issued every month), and the debate in the literature about the advantage these forecasts can provide when compared to physically based objective tercile probability forecasts (O´Lenic *et al*., 2008), this study aims to perform a comparative skill assessment of tercile probability seasonal precipitation forecasts for Brazil produced by these two approaches. Consensus forecasts were assessed and compared to CPTEC atmospheric general circulation model (AGCM) physically based forecasts (Cavalcanti *et al*., 2002; Marengo *et al*., 2003). This comparative assessment will help provide guidance on future practices for seasonal forecasting in Brazil.

The paper is organized as follows. Section 2 presents the strategy adopted for verifying the forecasts. Section 3 describes how well consensus and physically based forecasts can discriminate different forecast situations. Section 4 describes how reliable (i.e. how well calibrated) consensus and physically based forecasts are. Section 5 summarizes the main findings and presents the final remarks.

2. Verification strategy

This study assessed the skill of both consensus and CPTEC/AGCM physically based half-month lead tercile probability precipitation forecasts for Brazil (i.e. forecasts issued around the 15th of each month and valid for the following 3 month season). Verification was performed against the observed precipitation from the Brazilian meteorological network. This network is composed by stations maintained by CPTEC/INPE, INMET and several regional institutions interpolated to a regular $0.25^{o} \times 0.25^{o}$ grid. The upper and lower precipitation limits defining the below normal, normal and above normal categories were determined for each grid point by computing the 33.33 and 66.66 percentiles of the 1960–2000 climatological empirical distribution.

The forecasts were aggregated in four groups to allow the investigation of skill seasonal dependence as follows: austral spring (August–October/September–November/October–December 2002–2011), austral summer (November–January/December–February/January–March 2001–2010), austral autumn (February–April/March–May/April–June 2002–2011) and austral winter (May–July/June–August/July–September 2002–2011), each group containing a total of 30 forecasts. Figure 1(a) shows an example of consensus forecast map where forecast probabilities were issued for the below normal, normal and above normal precipitation categories of specific areas in Brazil. Forecast probabilities for the categories below normal, normal and above normal precipitation were estimated subjectively by expert assessment. Prior to verification each of the 120 consensus forecast maps was digitalized to the same regular grid of the observations. In this processes it was assumed that the tercile forecast probabilities for all grid points falling inside an area where a forecast was issued was identical. For each of these grid points the forecast was given by the forecast probabilities for the categories below normal, normal and above normal issued for the entire area.

The physically based forecasts for comparison with consensus forecasts were produced with the official operational CPTEC/AGCM version used to generate forecast products as 1 of the 12 designated World Meteorological Organization (WMO) Global Producing Centre for Long-Range Forecasts (GPC, http://www.clima1.cptec.inpe.br/gpc/). This model uses a deep cloud convection parameterization scheme developed

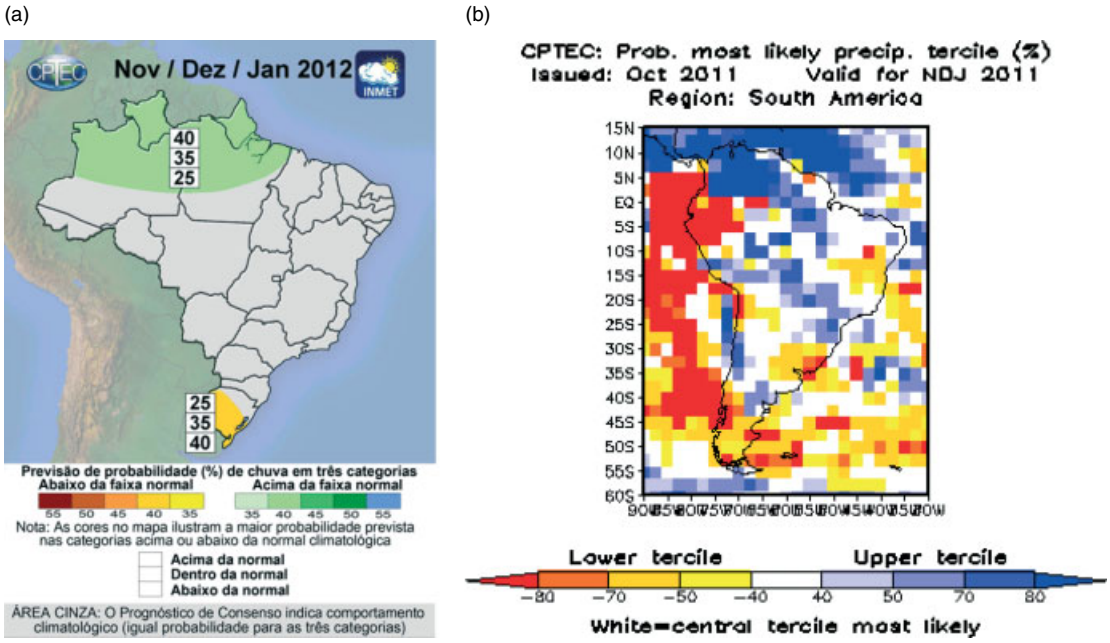

Figure 1. (a) Example of tercile probability consensus forecast for Brazil issued in mid-October 2011 and valid for November–January 2011/2012. Forecast probabilities for the categories below normal (bottom number in the rectangle), normal (central number in the rectangle) and above normal (top number in the rectangle) precipitation were estimated subjectively by expert assessment. Grey areas mean climatological forecast (i.e. forecast indicates equal probability for the three categories). (b) Example of physically based CPTEC/AGCM forecast probability map for the most likely precipitation tercile issued in mid-October 2011 and valid for November–January 2011/2012. Forecast probabilities were estimated objectively from an ensemble of 15 ensemble member forecasts. See text for additional information.

by Kuo (1974) and the short wave radiation scheme as described in Barbosa *et al*. (2008). Ensemble hindcasts for the period 1979–2008 were produced using a lagged initialization approach based on 10 atmospheric initial conditions from NCEP/NCAR reanalysis (Kalnay *et al*., 1996; Kanamitsu *et al*., 2002). For real-time forecasts produced after 2009 15 atmospheric initial conditions were used for ensemble generation as described in Coelho *et al*. (2012). Persisted observed sea surface temperature anomalies (Reynolds *et al*., 2002) of the month immediately prior to the first forecast month were used as ocean boundary conditions for the AGCM when producing both hindcasts and real-time forecasts for the following season.

Figure 1(b) shows an example of physically based CPTEC/AGCM forecast probability map for the most likely precipitation tercile. In order to issue tercile probability forecasts from an ensemble of forecasts produced by a climate model one needs to define a procedure for converting the available and finite number of ensemble member forecasts into probabilities. This procedure allows the production of forecast probabilities for the below normal, normal and above normal categories. The procedure adopted by CPTEC was to determine the upper and lower precipitation limits that define the three categories for each forecast grid point. These limits were obtained by computing the 33.33 and 66.66 percentiles of the 1979–2000 climatological empirical hindcast distribution. This distribution contains a total of 220 values for each grid point, resulting from 10 ensemble members for each of the 22 years of hindcasts. Once these limits were determined forecast probabilities for each of the three categories of each forecast produced during the period 2001–2011 were obtained by first counting the number of ensemble member falling in each category and next dividing these counts by the total number of produced ensemble member forecasts.

It is well recognized that no single metric is adequate to summarize forecast quality. Two essential forecast quality attributes are discrimination and reliability. Discrimination measures forecast ability to distinguish between different observed situations. In other words, discrimination measures whether forecasts differ when the corresponding observations differ. Reliability measures how well calibrated forecasts are. Even a perfectly calibrated forecast system is effectively useless if it lacks discrimination ability (Weigel and Mason, 2011). The following two sections describe and compare both discrimination and reliability of consensus and CPTEC/AGCM physically based tercile probability precipitation forecasts for Brazil.

3. Discrimination assessment

Forecast discrimination ability was assessed in this study using the generalized discrimination score (Mason and Weigel, 2009). This score is applicable for a number of forecast types including tercile probability precipitation forecasts as described in the previous section. The corresponding observation for each tercile probability forecast is either 1 (if precipitation was observed in the below normal category), 2 (if precipitation was observed in the normal category) or 3 (if precipitation was observed in the above normal category). Following Mason and Weigel (2009), for computing the score a set of two forecast-observation pairs was considered, and the question of whether the forecasts can be used successfully to distinguish between the observations was asked. The aim was to compare all possible sets of two forecast-observation pairs asking the same question each time and calculating the proportion of times the question was answered correctly. Each time the question was asked there was a 50% chance of identifying the correct observation in the absence of any useful information, but if the forecasts were skillful the proportion of correctly identifying the observations (i.e. the generalized discrimination score) would exceed 50%. The generalized discrimination score ranges from 0% for unskillful forecasts consistently unable to distinguish between two different observations to 100% for skillful (perfect) forecasts able to always distinguish two different observations. This score can be simplistically interpreted as how often the forecasts were correct. The scores presented in Figures 2 and 3 were computed using equation 16 of Mason and Weigel (2009) for discrete probabilistic forecasts. These were the type of seasonal forecasts issued for Brazil in the last decade during consensus forecast discussions and objective derivation from physically based CPTEC/AGCM ensemble forecasts. Weigel and Mason (2011) provide guidance on how to compute the generalized discrimination score for ensemble forecasts.

Figures 2 and 3 show generalized discrimination score maps for CPTEC/AGCM physically based (first column) and consensus (second column) tercile probability precipitation forecasts for Brazil for austral spring (Figure 2(a), (b), (c)), summer (second row of Figure 2(d), (e), (f)), autumn (first row of Figure 3(a), (b), (c)) and winter (second row of Figure 3(d), (e), (f)). The difference between CPTEC/AGCM physically based and consensus generalized discrimination scores for the four seasons is shown in the third column. Locations where the differences are statistically significant at the 5% level are marked with a black dot. Statistical significance on score difference has been assessed by first computing 95% confidence intervals for the difference between physically based CPTEC/AGCM and consensus forecast scores using a bootstrap resampling procedure with replacement. Next it has been checked whether or not these intervals include zero as recommended by Jolliffe (2007). For locations where the confidence interval did not include zero there was sufficient evidence that the two scores were different.

During spring (Figure 2(a)–(c)) and summer (Figure 2(d)–(f)) consensus forecasts are generally more skillful than physically based CPTEC/AGCM forecasts as illustrated by the predominance of blue areas in the maps of the third column indicating negative score differences. Exceptions to this pattern are noticed for the northeast (NE) and part of southeast (SE) regions of Brazil where CPTEC/AGCM forecasts are more skillful as illustrated by the yellow and orange areas indicating positive score difference. In parts of south (S) and north (N) Brazil regions consensus forecast achieve in certain areas generalized discrimination scores of the order of 70% that are about 20–30% larger than CPTEC/AGCM forecasts. A similar feature is also found during the summer (Figure 2(f)) for the northern part of northeast (NE) Brazil. During spring (Figure 2(c)) an opposite feature is noticed for northeast Brazil where CPTEC/AGCM forecasts achieve generalized discrimination scores of the order of 70% in certain regions that are about 10–20% larger than consensus forecasts.

During autumn (Figure 3(a)–(c)) and winter (Figure 3(d)–(f) physically based CPTEC/AGCM forecasts are generally slightly more skillful than consensus forecasts as illustrated by the predominance of yellow, orange and red areas in the maps of the third column indicating positive score differences. Exceptions to this pattern are noticed for parts of northern north (N) Brazil, and small areas of southern southeast (SE) and central-east (CE) regions of Brazil where consensus forecasts are more skillful. This feature is illustrated by the blue areas indicating negative score difference. In parts of south (S),

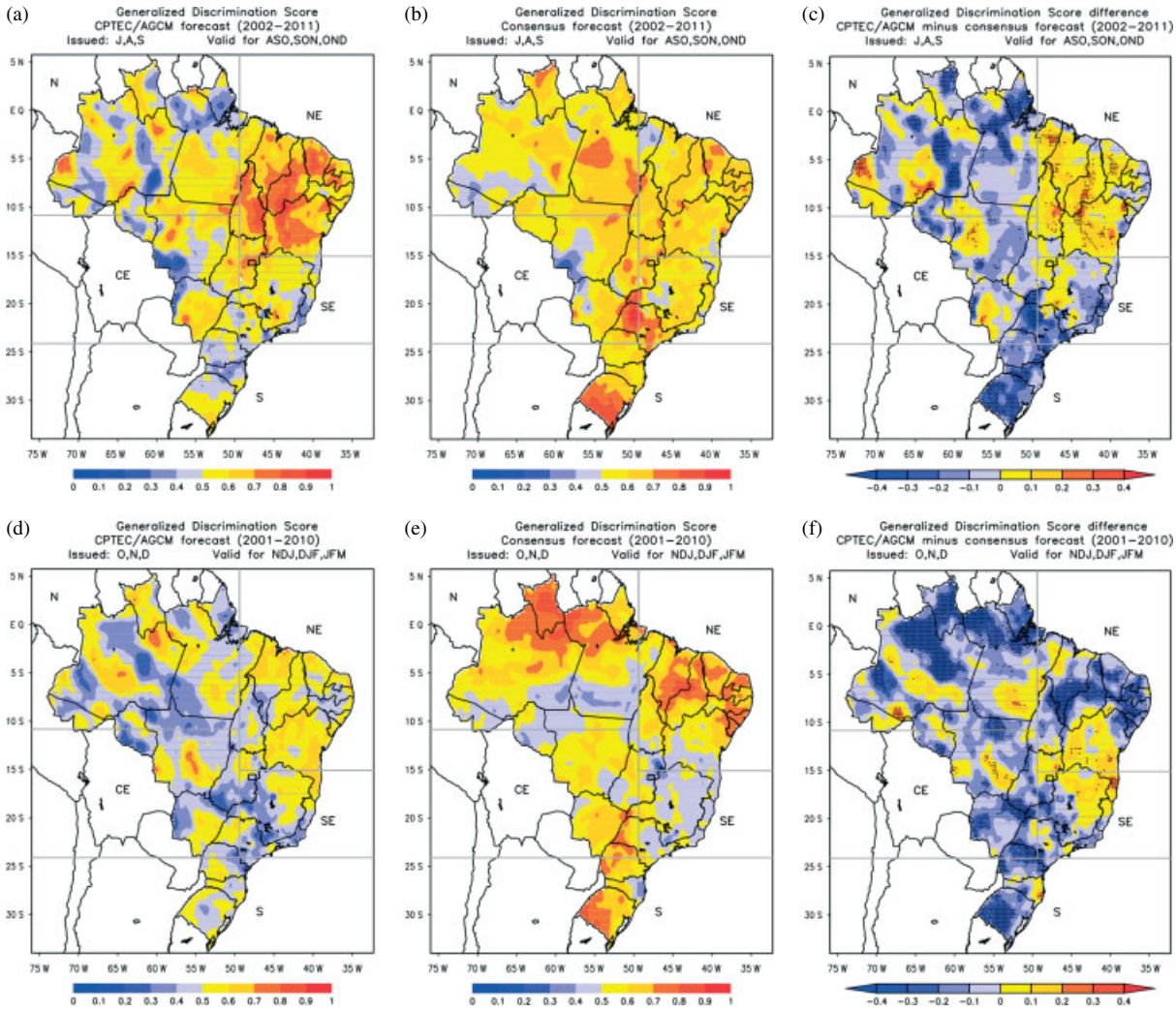

Figure 2. Generalized discrimination score maps for CPTEC/AGCM physically based (panels (a) and (d)) and consensus (panels (b) and (e)) tercile probability precipitation forecasts for Brazil for austral spring (panels (a) to (c)) and summer (panels (d) to (f)). The difference between CPTEC/AGCM physically based and consensus generalized discrimination scores for the four seasons is shown in the forth column. Locations where the differences are statistically significant at the 5% level are marked with a black dot. See text for additional information.

north and particularly northeast Brazil regions CPTEC/AGCM forecasts achieve generalized discrimination scores of the order of 70% that are about 20–30% larger than consensus forecasts. In northern north Brazil an opposite feature is noticed with consensus forecasts achieving generalized discrimination scores of the order of 60–70% that are about 10–20% larger than CPTEC/AGCM forecasts.

With a sample of 30 forecasts for each investigated season one can expect considerable sampling uncertainty in the computed generalized discrimination scores. As highlighted by Mason (2008) knowing the sampling uncertainty in the verification statistics not only provides an indication as to whether the computed scores may be misleading but also help address the question of whether the forecasts are good. Forecasts can confidently be considered good if they score well and if the uncertainty in the score is small. In order to estimate the score sampling uncertainty one should attempt to find the possible range of scores given different sets of forecasts from the same forecast system. This has been done using bootstrap resampling with replacement. In this procedure the original forecast-observation pairs were randomly sampled to generate a pre-defined number of forecast-observation samples (e.g. 500) of the same size (30) of the original sample. Next the verification score was computed for each of these new samples and confidence intervals were estimated from the empirical distribution of the obtained new scores.

Table 1 shows the area averaged generalized discrimination scores and corresponding 95% confidence intervals (in brackets) estimated using the bootstrap resampling procedure described above. This procedure was applied point by point (i.e. for each grid point) and the results presented in Table 1 are area averages for five regions in Brazil for the four investigated seasons, for physically based CPTEC/AGCM (top of Table 1) and consensus (bottom of Table 1) forecasts. The five regions are defined and marked with grey lines in the panels of Figure 2 as follows: North (N, west of 49.1 °W and north of 10.9 °S); Northeast (NE, east of 49.1 °W and north of 15.1 °S); South (S, south of 24.4 °S); Southeast (SE, east of 49.1 °W, south of 15.1 °S, and north of 24.4 °S); and Central-East (CE, west of 49.1 °W, south of 10.9 °S, and north of 24.4 °S). The scores presented in this table support the previous finding of Figure 2, indicating that for spring and summer consensus forecasts are generally more skillful than physically based CPTEC/AGCM forecasts, and that for autumn and winter physically based CPTEC/AGCM

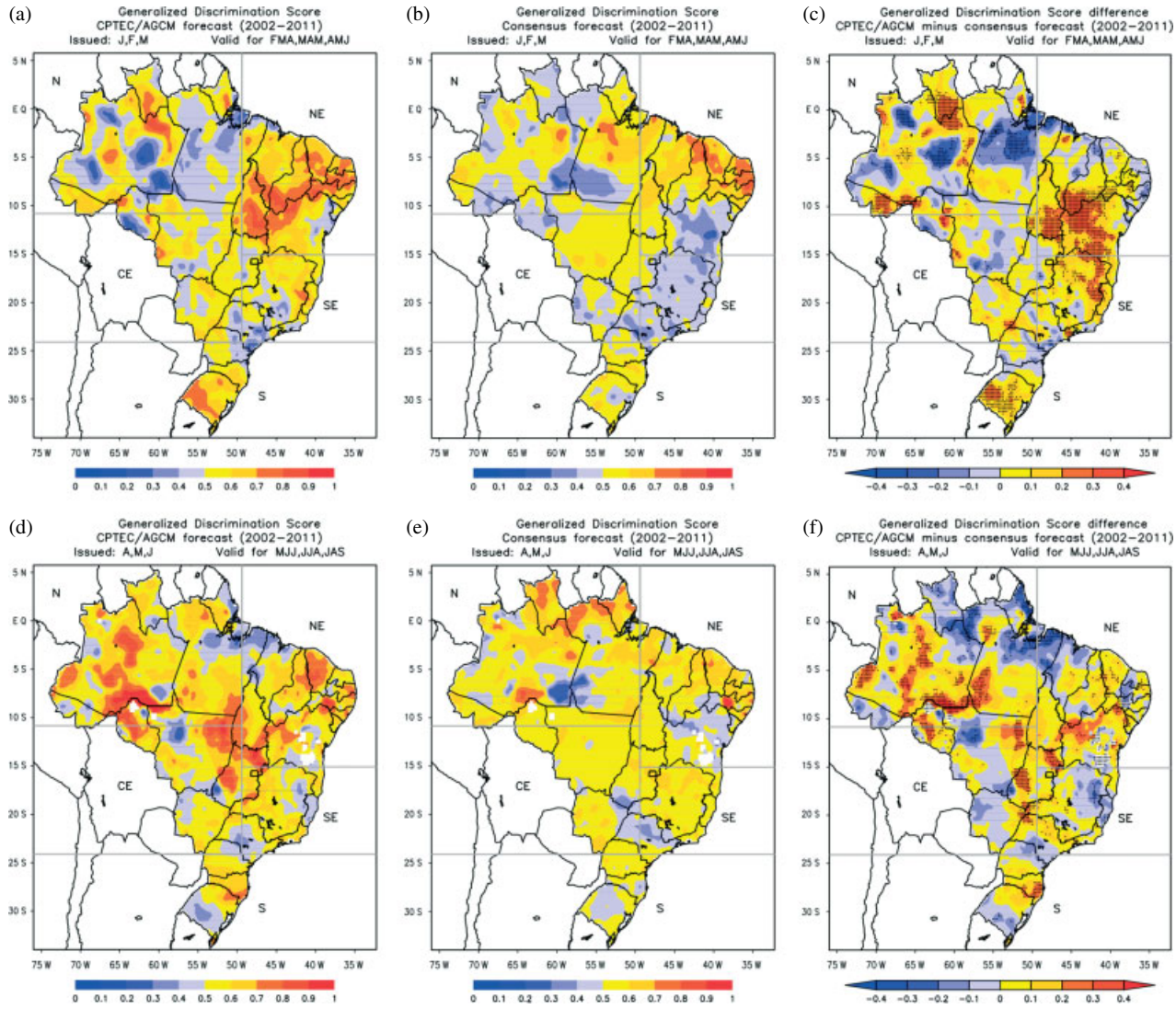

Figure 3. Generalized discrimination score maps for CPTEC/AGCM physically based (panels (a) and (d)) and consensus (panels (b) and (e)) tercile probability precipitation forecasts for Brazil for austral autumn (panels (a) to (c)) and winter (panels (d) to (f)). The difference between CPTEC/AGCM physically based and consensus generalized discrimination scores for the four seasons is shown in the forth column. Locations where the differences are statistically significant at the 5% level are marked with a black dot. See text for additional information.

Table 1. Area averaged generalized discrimination scores for CPTEC/AGCM physically based (top of table) and consensus (bottom of table) forecasts for the four seasons as defined in Section 2, for five regions in Brazil as defined and marked with grey lines in the panels of Figure 2.

CPTEC/AGCM	Spring	Summer	Autumn	Winter
N	0.52 (0.33, 0.70)	0.47 (0.29, 0.66)	0.50 (0.31, 0.70)	0.60 (0.37, 0.81)
NE	0.68 (0.47, 0.87)	0.54 (0.35, 0.72)	0.64 (0.46, 0.81)	0.59 (0.36, 0.80)
S	0.49 (0.31, 0.66)	0.49 (0.31, 0.67)	0.59 (0.41, 0.76)	0.55 (0.37, 0.73)
SE	0.53 (0.32, 0.73)	0.46 (0.28, 0.65)	0.53 (0.34, 0.72)	0.55 (0.32, 0.77)
CE	0.52 (0.30, 0.73)	0.46 (0.28, 0.65)	0.53 (0.31, 0.74)	0.58 (0.29, 0.83)
Consensus	Spring	Summer	Autumn	Winter
N	0.58 (0.42, 0.74)	0.59 (0.43, 0.75)	0.51 (0.34, 0.67)	0.57 (0.37, 0.76)
NE	0.61 (0.45, 0.77)	0.61 (0.44, 0.77)	0.56 (0.40, 0.72)	0.55 (0.38, 0.72)
S	0.68 (0.52, 0.83)	0.64 (0.47, 0.80)	0.54 (0.35, 0.72)	0.52 (0.35, 0.69)
SE	0.60 (0.41, 0.77)	0.48 (0.32, 0.65)	0.45 (0.31, 0.59)	0.55 (0.43, 0.67)
CE	0.59 (0.41, 0.75)	0.56 (0.41, 0.71)	0.51 (0.38, 0.64)	0.55 (0.44, 0.65)

North (N, west of 49. 1° W and north of 10.9° S); Northeast (NE, east of 49. 1° W and north of 15. 1° S); South (S, south of 24.4° S); Southeast (SE, east of 49. 1° W, south of 15.1° S, and north of 24.4° S); and Central-East (CE, west of 49.1° W, south of 10.9° S, and north of 24.4° S).

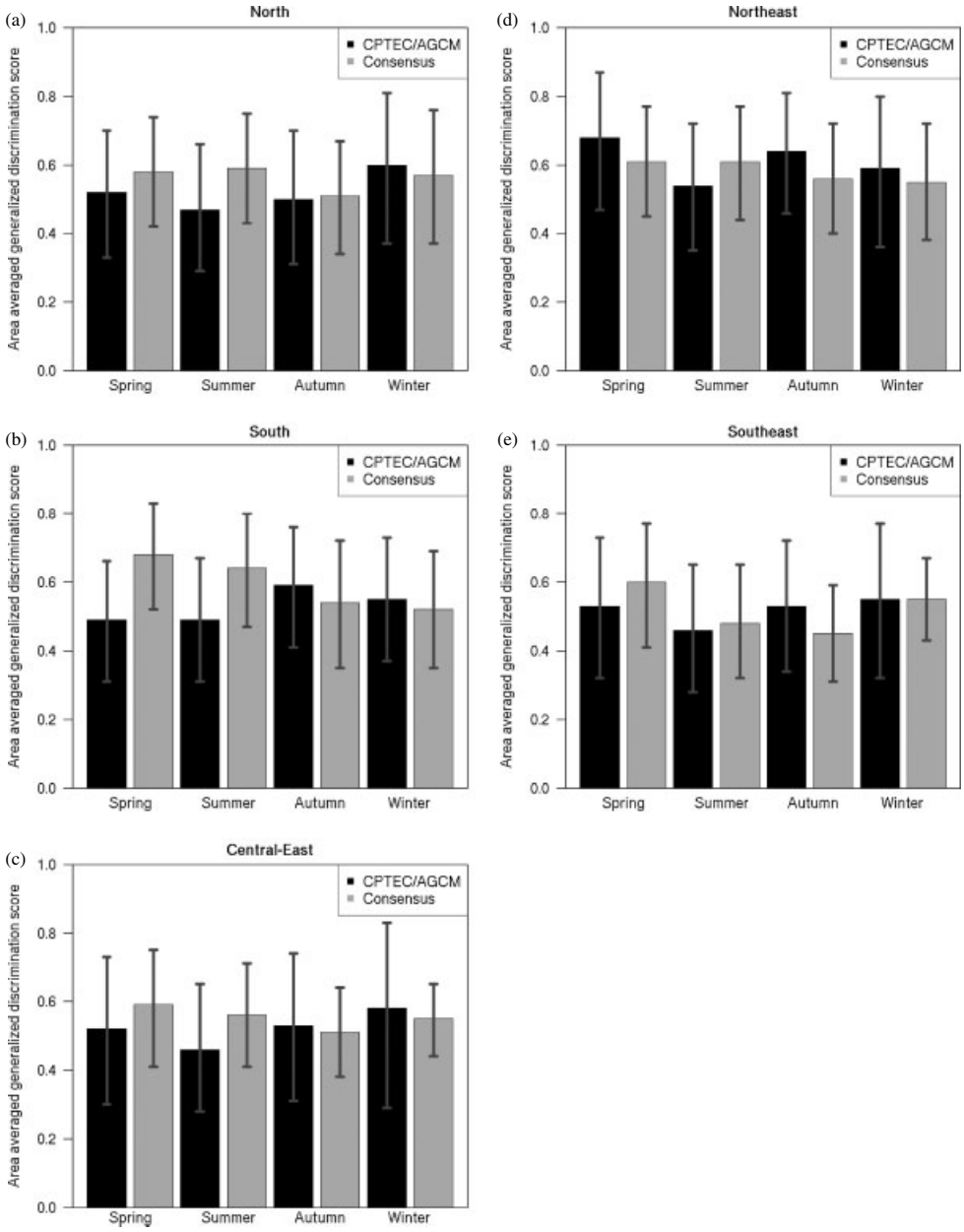

Figure 4. Area averaged generalized discrimination scores for CPTEC/AGCM (black bars) and consensus (grey bars) forecasts for the five investigated regions in Brazil (North in panel (a), South in panel (b), Central-East in panel (c), Northeast in panel (d) and Southeast in panel (e)) for austral summer, spring, autumn and winter. The vertical whiskers (dark grey) show the 95% confidence interval for the area averaged scores (see text for additional information).

forecasts are slightly more skillful than consensus forecasts. Most scores are larger than 50% indicating potential forecast usefulness. However, the uncertainty in the computed scores given by the 95% confidence intervals shown in brackets is quite large, ranging on average from around 0.3 to 0.8. This indicates that in fact a large part of the identified potential skill might not necessarily be translated into good quality forecasts for the five investigated regions. In other words, due to the large sampling uncertainty not all forecasts are confidently useful to discriminate between two different observed situations at the investigated confidence level.

Figure 4 shows a graphical representation of Table 1 scores. This figure allows a straightforward comparison between CPTEC/AGCM (black bars) and consensus (grey bars) area averaged discrimination scores for the five investigated regions in Brazil. It highlights the previous findings: (1) during spring and summer consensus forecasts are more skillful than CPTEC/AGCM forecasts (grey bars generally longer than black bars); and (2) during autumn and winter CPTEC/AGCM forecasts are slightly more skillful than consensus forecasts (black bars generally slightly longer than grey bars). It should be noticed, however, that the differences between the averaged forecast scores for CPTEC/AGCM and consensus forecasts are generally modest (i.e. small) and that scores confidence intervals (dark grey whiskers) are considerably large. The bootstrap resampling computation for the difference in the area averaged scores between CPTEC/AGCM and consensus forecasts revealed that these differences are not statistically significant at the 5% level. This procedure was performed similarly to the

procedure used in Figure 2 by checking whether or not the computed 95% confidence interval for the score difference included zero. For all five regions the confidence intervals did include zero, suggesting that there was not sufficient evidence that area averaged CPTEC/AGCM and consensus forecast scores were different for the five regions.

The generalized discrimination score presented in this section measured the degree of correct probabilistic forecast discrimination. In other words, it measured probabilistic forecast ability to distinguish one observed tercile category outcome from another even if the forecast probabilities were biased or poorly calibrated. The next section assesses whether or not there exist biases (systematic differences between forecasts and the correspondent observations) in these forecasts and how well calibrated these forecasts are.

4. Reliability assessment

Probabilistic forecast biases can be assessed using the so-called tendency diagrams (Mason, 2012). The diagram shows for each tercile category the average forecast probability and the corresponding observed relative frequency over the verification period as vertical bars. Both quantities were computed for the entire forecast domain (all Brazil). The diagram then contains a total of six bars, two for each tercile category as illustrated in Figure 5. The two bars for each category display the same information presented in reliability diagrams, i.e. forecast probabilities in the horizontal axis and observed relative frequencies in the vertical axis. Therefore, the tendency diagram is useful for assessing forecast reliability. It is particularly useful to identify if the forecasting system under-or over-forecast any particular category. A perfectly well calibrated and unbiased system should have exactly the same heights for the two bars for each of the three categories in the tendency diagram (equivalent to a 45^{o} diagonal line in the reliability diagram). The tendency diagram also indicates whether or not there exists any shift in the observed climate with respect to the previously defined climatology. For example, an observed relative frequency of below normal precipitation larger than 33.3% indicates that recent climate conditions are becoming drier when compared to previous years. The ability of the forecasting system in identifying this change is assessed by comparing the heights of the mean forecast probability and observed relative frequency bars for the below normal category and checking whether or not the mean forecast probability is also larger than 33.3%.

Figure 5(a)–(d) shows tendency diagrams for consensus tercile probability precipitation forecasts over Brazil for spring, summer, autumn and winter, respectively. Average forecast probabilities (black bars) for the below normal, normal and above normal categories for all fours seasons are around 30, 40 and 30%, respectively. These probability values illustrate the traditionally conservative nature of consensus forecasts. During consensus forecast discussions forecasters tend to issue forecast probabilities for the three categories close to the climatological distribution, often favouring the normal category as the most likely. This latter fact can be identified by comparing the back and grey bars for the normal category for the four seasons in Figure 5(a)–(d). This comparison reveals that consensus forecasts always over-forecast the normal category (black bars are longer than grey bars). The same over-forecasting feature was also found for above normal category consensus forecasts, although during summer (Figure 5(b)) and autumn (Figure 5(c)) consensus forecasts only slightly over-forecast this category. These figures also revealed that consensus forecasts under-forecast the below-normal category (black bars are shorter than grey bars).

Figure 5(e)–(h) shows tendency diagrams for physically based CPTEC/AGCM tercile probability precipitation forecasts over Brazil for spring, summer, autumn and winter, respectively. CPTEC/AGCM forecasts tend on average to favour the above-normal category as most likely and always over-forecast this category (black bars longer than grey bars). Such an over-forecasting feature is even larger than for consensus forecasts described above. Contrasting the previous finding for consensus forecasts for the normal category, CPTEC/AGCM forecasts show much less severe biases for this category (black and grey bar heights nearly equal, expect for winter when the over-forecasting feature is noticed). As for consensus forecasts, CPTEC/AGCM also under-forecast the below-normal category (black bars are shorter than grey bars).

Forecast reliability can also be objectively measured by computing the reliability component of the Brier Score (Sanders, 1963; Murphy, 1971, 1973, 1986). Perfectly reliable (i.e. well calibrated) forecasts have null reliability component of the Brier Score. Table 2 shows the reliability component of the Brier Score for probability forecasts for the events precipitation in the below-normal, normal and above-normal categories produced by CPTEC/AGCM physically based and consensus forecasts for Brazil produced during the last decade. Values in brackets are the 95% confidence intervals for the computed scores applying the bootstrap resampling procedure described in Section 3. The comparison of the scores shown in Table 2 revealed that consensus forecasts have smaller and closer to null scores than CPTEC/AGCM forecasts. This result is valid for all seasons and categories investigated and indicate that consensus forecasts are better calibrated than CPTEC/AGCM physically based forecasts.

5. Summary and conclusions

This study has assessed and compared the skill of consensus and CPTEC/AGCM physically based tercile probability precipitation forecasts for Brazil produced during the last decade. The comparative assessment focused on two fundamental forecast attributes: discrimination (the ability of the forecasting systems in discriminating between different forecast situations) and reliability (a measure of forecast calibration). Discrimination has been assessed with the so-called generalized discrimination score, which is a skill measure relevant for both administrative purposes and general public. This is because it can be easily interpreted as how often the forecasts are correct. It should be noted, however, that the generalized discrimination score is a measure of potential skill, and therefore cannot be used as a unique measure to summarize forecast quality. This is because forecasts with high generalized discrimination score can still have systematic errors (e.g. biases leading to under or over confident forecasts), and therefore require calibration to remove these errors. In order to complement the assessment reliability has been investigated through the use of tendency diagrams (similar to reliability diagrams) and by computing the reliability component of the Brier Score (a measure of how well calibrated the forecasts are). The combined presentation of discrimination ability and reliability for both consensus and CPTEC/AGCM physically based forecasts provided a good forecast quality summary. It is also worth emphasizing that even perfectly calibrated forecasts may still be of effective little or no use if forecasts lack discrimination ability.

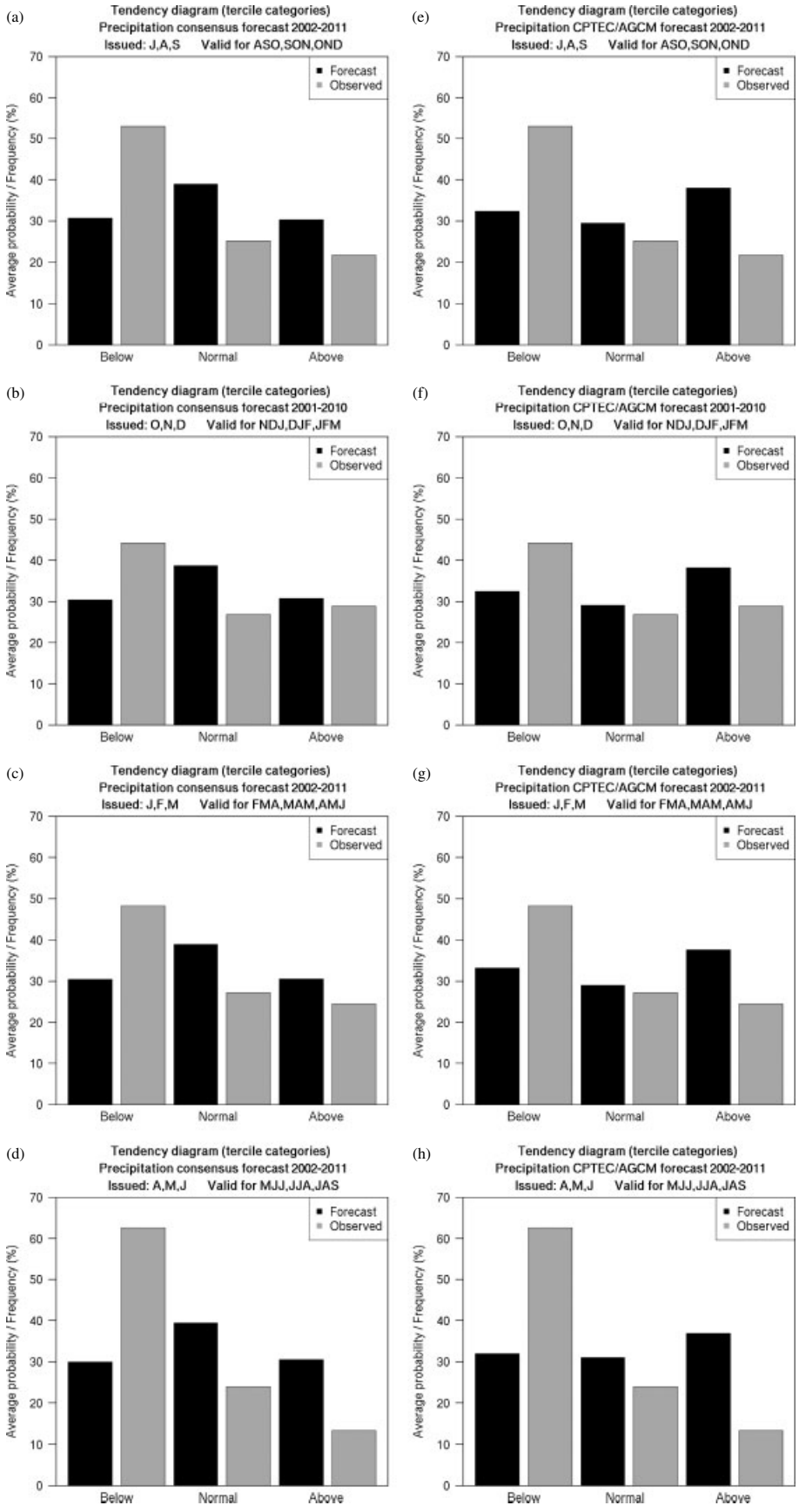

Figure 5. Tendency diagrams for consensus ((a) to (d)) and physically based CPTEC/AGCM ((e) to (h)) tercile probability precipitation forecasts over Brazil for spring (first row), summer (second row), autumn (third row) and winter (forth row). Black bars show average forecast probabilities. Grey bars show the observed relative frequency.

Table 2. Reliability component of the Brier Score for probability forecasts for the events precipitation in the below normal, normal and above normal categories produced by CPTEC/AGCM physically based (top of table) and consensus (bottom of table) forecasts for Brazil produced during the last decade.

CPTEC/AGCM	Below	Normal	Above
Spring	0.1276 (0.1265, 0.1289)	0.0456 (0.0449, 0.0462)	0.0899 (0.0890, 0.0908)
Summer	0.0961 (0.0951, 0.0972)	0.0400 (0.0394, 0.0406)	0.0871 (0.0862, 0.0881)
Autumn	0.1100 (0.1089, 0.1113)	0.0497 (0.0490, 0.0504)	0.0775 (0.0765, 0.0784)
Winter	0.1875 (0.1861, 0.1888)	0.0672 (0.0665, 0.0680)	0.1118 (0.1107, 0.1127)
Consensus	Below	Normal	Above
Spring	0.0567 (0.0559, 0.0575)	0.0214 (0.0210, 0.0219)	0.0093 (0.0090, 0.0095)
Summer	0.0201 (0.0197, 0.0206)	0.0151 (0.0147, 0.0154)	0.0012 (0.0011, 0.0013)
Autumn	0.0371 (0.0365, 0.0377)	0.0170 (0.0166, 0.0174)	0.0058 (0.0056, 0.0060)
Winter	0.1151 (0.1140, 0.1163)	0.0267 (0.0262, 0.0272)	0.0316 (0.0313, 0.0320)

Values in brackets are the 95% confidence intervals for the computed scores.

The discrimination skill assessment revealed that during spring and summer consensus forecasts were generally more skillful than CPTEC/AGCM physically based forecasts. For these two seasons consensus forecasts showed generalized discrimination scores exceeding 50% in a large portion of Brazil, indicating potential forecast usefulness. On the other hand, during autumn and winter CPTEC/AGCM physically based forecasts were generally more skillful than consensus forecasts. For these two seasons physically based forecasts presented larger generalized discrimination scores than consensus forecasts in a large portion of Brazil. These results suggest that consensus and physically based forecasts are seasonally complementary. Particularly for spring and summer climate knowledge expertise does add value to the final forecasts produced during consensus forecast discussions. During these two seasons expert knowledge about tropical convection in response to tropical sea surface temperatures appeared to be important to produce improved quality consensus forecasts when compared to physically based forecasts, particularly for northern and south Brazil. During autumn and winter climate expert assessment in consensus forecast discussions appeared to be more challenging, making it more difficult to produce improved quality forecasts for a large area in Brazil when compared to physically based forecasts. During these two seasons synoptic systems intra-seasonal variability most likely made physically based forecasts generally better able to reproduce the observed seasonal rainfall variability in parts of Brazil than consensus forecasts based on expert assessment.

Although consensus forecasts for Brazil were found to be potentially useful when compared to CPTEC/AGCM physically based forecasts, particularly during spring and summer, as highlighted in the previous paragraph, the tendency diagrams revealed that these forecasts suffer from systematic errors (biases). Consensus forecasts were found to be on average too conservative, generally not differing too much from the climatological distribution. The normal category of these forecasts was always over-forecast, as well as the above-normal category, while the below-normal category was always under-forecast. Tendency diagrams for CPTEC/AGCM physically based forecasts also revealed important systematic biases. As for consensus forecasts, CPTEC/AGCM physically based forecasts also over-forecasted the above-normal category and under-forecasted the below-normal category. However, the over-forecasting feature of CPTEC/AGCM forecasts was found to be even greater than for consensus forecasts. On the other hand, CPTEC/AGCM physically based forecasts showed better and much less biased forecast probabilities for the normal category than consensus forecasts. The assessment through the computation of the reliability component of the Brier-Score revealed that consensus forecasts are better calibrated than CPTEC/AGCM physically based forecasts.

This comparative assessment will help provide guidance on future practices for seasonal forecasting in Brazil. The systematic errors (biases) in consensus forecasts were found to be most likely due to the subjectivity involved in the process of attributing forecast probabilities to tercile categories during consensus forecast discussions. In order to address this problem and produce better calibrated and reliable forecasts, at the time of writing this manuscript, CPTEC, INMET and Ceará State Meteorology and Hydrology Foundation (FUNCEME) are testing an objective procedure for attributing forecast probabilities to tercile categories during consensus forecast discussions. This procedure is based on linear regression of past forecasts and past observations. It takes into account the magnitude of the forecast anomalous signal and the past forecast skill both from a multi-model ensemble forecast system. It is expected that this new procedure will allow the production of improved quality tercile probability precipitation consensus seasonal forecasts for Brazil in the future.

Acknowledgement

CASC was supported by Conselho Nacional de Desenvolvimento Cientrfico e Tencológico (CNPq) process 306664/2010-0. Two anonymous reviewers are acknowledged for their valuable suggestions and constructive criticism that helped to substantially improve the quality of the paper.

References

Barbosa HMJ, Tarasova TA, Cavalcanti IFA. 2008. Impacts of a new solar radiation parameterization on the CPTEC AGCM climatological features. *J. Appl. Meteorol. Climatol.* **47**: 1377–1392.

Berri GJ, Antico PL, Goddard L. 2005. Evaluation of the climate outlook forums´ seasonal precipitation forecasts of southeast South America during 1998–2002. *Int. J. Climatol.* **25**: 365–377.

Cavalcanti IFA, Marengo JA, Satyamurty P, Nobre CA, Trosnikov I, Bonatti JP, Manzi AO, Tarasova T, Pezzi LP, D'Almeida C, Sampaio G, Castro CC, Sanches MB, Camargo H. 2002. Global climatological features in a simulation using the CPTEC-COLA AGCM. *J. Clim.* **15**: 2965–2988.

Coelho CAS, Cavalcanti IAF, Costa SMS, Freitas SR, Ito ER, Luz G, Santos AF, Nobre CA, Marengo JA, Pezza AB. 2012. Climate diagnostics of three major drought events in the Amazon and

illustrations of their seasonal precipitation predictions. *Meteorol. Appl.* **19**: 237–255.
Jolliffe IT. 2007. Uncertainty and inference for verification measures. *Weather Forecast.* **22**: 637–650.
Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Wollen J, Zhu Y, Letman A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Moo KC, Ropelewski C, Wang J, Jenne R, Josepf D. 1996. The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**(3): 437–471.
Kanamitsu M, Ebisuzaki W, Woollen J, Yang S-K, Hnilo JJ, Fiorino MG, Potter GL. 2002. NCEP-DOE AMIP-II reanalysis (R-2). *Bull. Am. Meteorol. Soc.* **83**: 1631–1643.
Kuo HL. 1974. Further studies of the parameterization of the influence of cumulus convection on a large-scale flow. *J. Atmos. Sci.* **31**: 1232–1240.
Marengo JA, Cavalcanti IFA, Satyamurty P, Trosnikov I, Nobre CA, Bonatti JP, Camargo H, Sampaio G, Sanches MB, Manzi AO, Castro CAC, D'almeida C, Pezzi LP, Candido L. 2003. Assessment of regional seasonal rainfall predictability using the CPTEC/COLA atmospheric GCM. *Clim. Dyn.* **21**: 459–475.
Mason SJ. 2008. Understanding forecast verification statistics. *Meteorol. Appl.* **15**: 31–40.
Mason SJ. 2012. Seasonal and longer range forecasts. In *Book Chapter of Forecast Verification: A Practioner's Guide in Atmospheric Science*, 2nd edn, Jolliffe IT, Stephenson DB (eds). Wiley-Blackwell: Chichester, UK, 203–220.
Mason SJ, Weigel AP. 2009. A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.* **137**: 331–349.
Murphy AH. 1971. A note on the ranked probability score. *J. Appl. Meteorol.* **10**: 155–156.
Murphy AH. 1973. A new vector partition of the probability score. *J. Appl. Meteorol.* **12**: 595–600.
Murphy AH. 1986. A new decomposition of the Brier score: formulation and interpretation. *Mon. Weather Rev.* **114**: 2671–2673.
O´Lenic EA, Unger DA, Halpert MS, Pelman KS. 2008. Developments in operational long-range climate prediction at CPC. *Weather Forecast.* **23**: 496–515.
Reynolds RW, Rayner NA, Smith TM, Stokes DC, Wang W. 2002. An improved in situ and satellite SST analysis for climate. *J. Clim.* **15**: 1609–1625.
Sanders F. 1963. On subjective probability forecasting. *J. Appl. Meteorol.* **2**: 191–201.
Weigel AP, Mason SJ. 2011. The generalized discrimination score for ensemble forecasts. *Mon. Weather Rev.* **139**: 3069–3074.

