Assessment of ECMWF Subseasonal Temperature Predictions for an Anomalously Cold Week Followed by an Anomalously Warm Week in Central and Southeastern South America during July 2017

M. S. ALVAREZ

Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Ciencias de la Atmósfera y los Océanos and CONICET–Universidad de Buenos Aires, Centro de Investigaciones del Mar y la Atmósfera (CIMA) and CNRS–IRD–CONICET–UBA, Instituto Franco-Argentino para el Estudio del Clima y sus Impactos (UMI 3351 IFAECI), Buenos Aires, Argentina

C. A. S. COELHO

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

M. OSMAN

Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Ciencias de la Atmósfera y los Océanos and CONICET–Universidad de Buenos Aires, Centro de Investigaciones del Mar y la Atmósfera (CIMA) and CNRS–IRD–CONICET–UBA, Instituto Franco-Argentino para el Estudio del Clima y sus Impactos (UMI 3351 IFAECI), Buenos Aires, Argentina

M. Â. F. FIRPO

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

C. S. VERA

Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Ciencias de la Atmósfera y los Océanos and CONICET–Universidad de Buenos Aires, Centro de Investigaciones del Mar y la Atmósfera (CIMA) and CNRS–IRD–CONICET–UBA, Instituto Franco-Argentino para el Estudio del Clima y sus Impactos (UMI 3351 IFAECI), Buenos Aires, Argentina

(Manuscript received 3 October 2019, in final form 24 June 2020)

ABSTRACT

The demand of subseasonal predictions (from one to about four weeks in advance) has been considerably increasing as these predictions can potentially help prepare for the occurrence of high-impact events such as heat or cold waves that affect both social and economic activities. This study aims to assess the subseasonal temperature prediction quality of the European Centre for Medium-Range Weather Forecasts (ECMWF) against the Japan Meteorological Agency reanalyses. Two consecutive weeks of July 2017 were analyzed, which presented anomalously cold and warm conditions over central South America. The quality of 20 years of hindcasts for the two investigated weeks was compared to that for similar weeks during the JJA season and of 3 years of real-time forecasts for the same season. Anomalously cold temperatures observed during the week of 17–23 July 2017 were well predicted one week in advance. Moreover, the warm anomalies observed during the following week of 24–30 July 2017 were well predicted two weeks in advance. Higher linear association and discrimination (ability to distinguish events from nonevents), but reduced reliability, was found for the 20 years of hindcasts for the target week than for the hindcasts produced for all of the JJA season. In addition, the real-time forecasts showed generally better performance over some regions of South America than the hindcasts.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-19-0200.s1.

Corresponding author: Alvarez M. S., alvarez@cima.fcen.uba.ar

DOI: 10.1175/WAF-D-19-0200.1

The assessment provides robust evidence about temperature prediction quality to build confidence in regional subseasonal forecasts as well as to identify regions in which the predictions have better performance.

1. Introduction

The interest in predicting well in advance cold and warm weekly conditions, and even cold and heat waves, has been increasing in recent years. In this way, one of the main goals of the Subseasonal to Seasonal Prediction project (S2S; Vitart et al. 2017) is to diagnose and improve forecast skill and understand the sources of predictability on the S2S time scale with special emphasis on the prediction of high-impact events. Heat waves are one of these events that have societal impacts on, for example, public health, including loss of life, and on the agriculture sector, as heat waves condition irrigation schedules and application of pesticide or fertilizers (White et al. 2017).

The forecast quality assessment of heat waves on subseasonal time scales has been recently addressed in several studies motivated by the S2S Project. Hudson et al. (2016) considered the Australian Bureau of Meteorology (BoM) model (POAMA-2) to evaluate its performance in forecasting the three most extreme heat events over Australia in 2013. Ardilouze et al. (2017) assessed the performance of real-time Météo-France model (CNRM-CM) forecasts for an intense heat wave that struck west Europe in early July 2015 and found limited utility of the forecast system beyond 12 days. Using the CNRM-CM model retrospective forecasts, that is, forecasts produced after the events were observed and that are also known as hindcasts, Batté et al. (2018) studied the prediction quality of boreal spring heat wave events over West Africa and Sahel. Over South America, Osman and Alvarez (2018) evaluated two models from the S2S database, the POAMA and BCC-CPS models, from the BoM and the Chinese Meteorological Agency, respectively, in predicting an intense heat wave during December 2013, finding promising performance of probabilistic forecasts for leads longer than a week. Other studies for the South American region focused on the subseasonal precipitation forecast quality assessment or on extreme rainfall event analysis (Hirata and Grimm 2018; Doss-Gollin et al. 2018; Coelho et al. 2018). However, a systematic subseasonal temperature forecast quality assessment focused on South America has not been done yet and it still remains largely undocumented.

During July 2017, two consecutive weeks showed remarkably contrasting temperature anomalies over central and southeastern South America (Figs. 2e and 9e). As an example of the daily evolution of temperature anomalies, Fig. 1a shows the daily mean temperature anomaly time series during July 2017 for three meteorological stations: Resistencia, in central-northern Argentina, Sao Luiz Gonzaga in south Brazil, and Sao Paulo in southeast Brazil (Fig. 1b). A sharp temperature decrease started between 16 and 18 July 2017, and the anomaly remained negative during 6 days in Resistencia and Sao Luiz Gonzaga and during 3 days in Sao Paulo. Moreover, 3 days in Resistencia and Sao Luiz Gonzaga recorded temperatures below one standard deviation as well as two days in the Sao Paulo. The spatial distribution of temperature anomalies for 17–23 July (Fig. 2e) shows that negative anomalies covered a large region over most central and southeastern South America and contrasting with positive anomalies over the Patagonia region in southern South America. On 23 or 24 July 2017, above-normal temperatures were restored in Resistencia and Sao Luiz Gonzaga, respectively. Such anomalously warm conditions lasted for the whole week with every day surpassing the one standard deviation threshold (Fig. 1a). Figure 9e shows that the warm anomalies extended across central and southern South America, being most intense over central and northern Argentina, Uruguay, Paraguay, and southern Brazil.

From Fig. 1 the alert reader might wonder whether the colder-than-normal week of 17–23 July 2017 was actually a disruption of a long-lasting warm event, as the previous and following weeks showed intense warm anomalies over northern Argentina and South Brazil. We then computed the temperature anomalies for each of the weeks analyzed along the June–July–August (JJA) 2017 season (see Fig. S1 in the online supplemental material) and found that along the 3 previous weeks to 17–23 July, positive temperature anomalies prevailed over the region. Moreover, most of the JJA weeks resulted in warmer-than-normal conditions over the region comprising northern Argentina, Paraguay, and southern Brazil, and that positive anomaly stands out when computing the JJA seasonal average, together with a less intense warm anomaly over the Brazilian Amazon region (Fig. S1).

One of the sources of predictability on the subseasonal time scale is the Madden–Julian oscillation (MJO; Madden and Julian 1994). Along the second fortnight of July 2017 here analyzed, the RMM index (Wheeler and Hendon 2004) reflected a marginally active MJO, with amplitudes larger than 0.75 during most

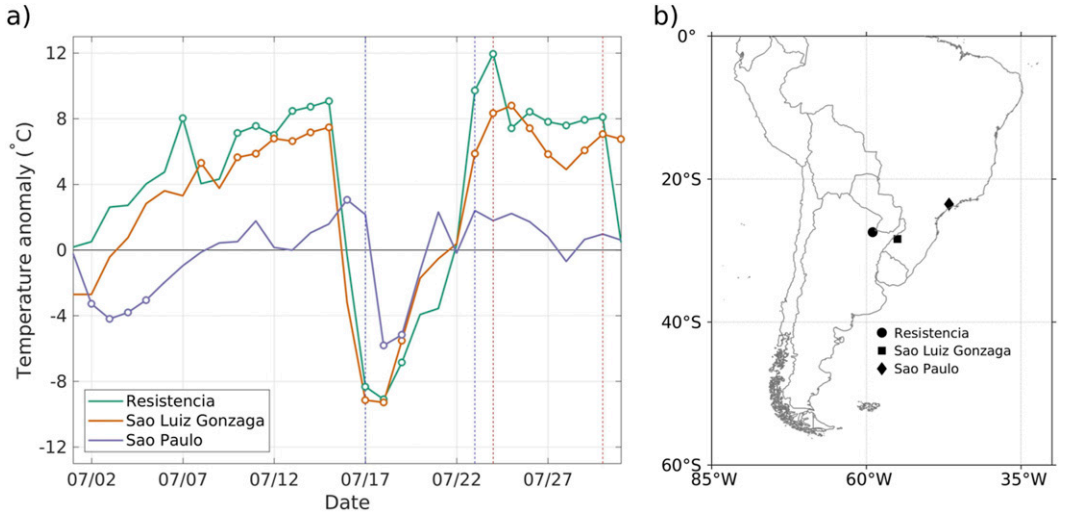

FIG. 1. (a) Daily temperature anomaly time series for July 2017 for Resistencia (87155; 27.45°S, 59°W), Sao Luiz Gonzaga (83907; 28.4°S, 55.01°W), and Sao Paulo (83781; 23.5°S, 46.61°W) meteorological stations. Anomalies were computed with respect to the 1980–2016, 31-point smoothed daily climatological mean. White circles represent days with temperature anomalies in magnitude larger than one standard deviation for each station, with the standard deviation computed using the historical daily temperature values for the 1980–2016 period. The vertical blue dashed lines mark the cold week period from 17 to 23 Jul 2017. The vertical red dashed lines mark the warm week period from 24 to 30 Jul 2017. (b) Locations of the weather stations used in (a).

of the period, and spanning MJO phases 3 and 4 during the first week and mostly 5 and 6 during the second week (not shown). The upper-level circulation response (not shown) resulted in a similar pattern over the Pacific Ocean and South America to the composites presented in Alvarez et al. (2016), and therefore the MJO might have had an influence in the circulation anomalies and ultimately over the temperature anomalies.

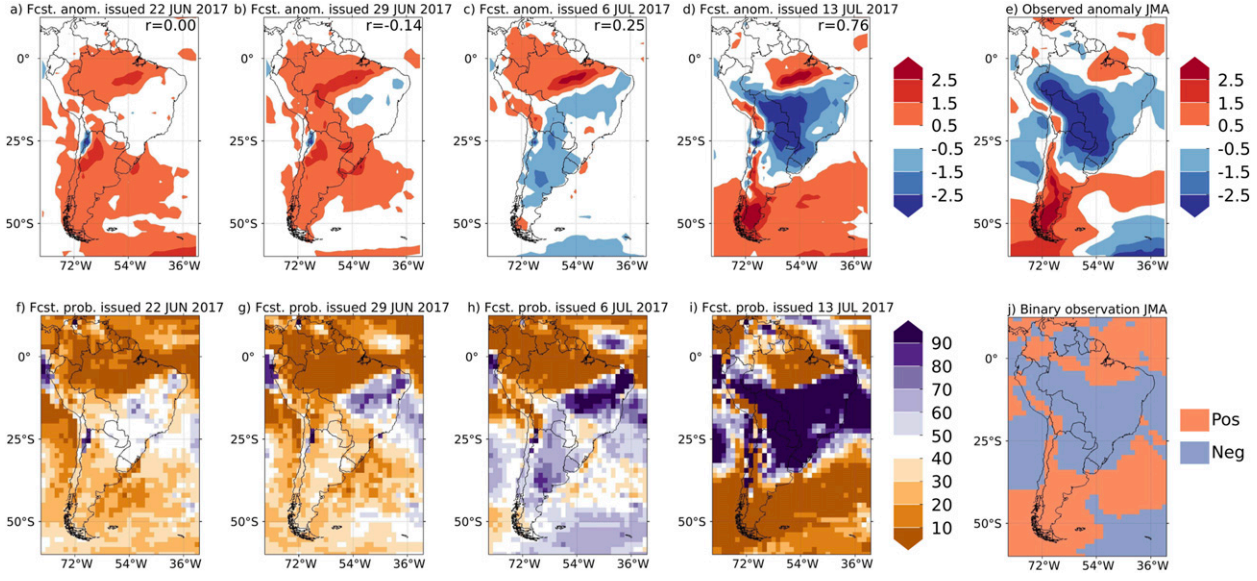

FIG. 2. ECMWF ensemble mean real-time forecast temperature anomalies for the target week of interest (17–23 Jul 2017) initialized on the (a) 22 Jun 2017, (b) 29 Jun 2017, (c) 6 Jul 2017, and (d) 13 Jul 2017, representing forecasts produced from 4 to 1 week in advance as described in the text. Anomalies were calculated with respect to the 1997–2016 hindcast period (20 years). (e) Observed temperature anomalies for the week 17–23 Jul 2017 with respect to the 1995–2017. ECMWF forecasts probabilities for the occurrence of negative temperature anomaly during the target week of interest (17–23 Jul 2017) initialized on the (f) 22 Jun 2017, (g) 29 Jun 2017, (h) 6 Jul 2017, and (i) 13 Jul 2017, derived from 51 ensemble members. (j) Binary observation indicating where a negative (blue) or a positive (red) temperature anomaly was recorded during the week.

The main objective of this paper is to assess the quality of the subseasonal temperature forecasts for this sequence of anomalously cold and warm weeks of July 2017. To achieve this objective, the verification framework proposed by Coelho et al. (2018) is used. This framework provides verification information together with forecast information at the time of issuing the forecast for a particular week of interest. Particularly, we seek to answer the following questions: How many weeks in advance were the observed temperature anomalies predicted? Do retrospective forecasts produced for the target weeks have better performance than aggregated retrospective forecasts produced for similar weeks during the JJA season? And how does retrospective forecast performance compare to real-time forecast performance? Retrospective and real-time subseasonal forecasts of the European Centre for Medium-Range Weather Forecasts (ECMWF) model are used in this study because this state-of-the-art model provides a reasonably large ensemble size in the hindcast period and an even larger for the real-time forecasts.

This document is organized as follows: section 2 describes the data used and summarizes the verification framework used in this study. Section 3 presents the verification results for both the cold and the warm week here investigated. Finally, a discussion and summary is presented in section 4.

2. Data and methodology

The ECMWF subseasonal forecasts available through the S2S prediction database (Vitart et al. 2017) were used in this study. The model provides an ensemble of real-time forecasts composed by 51 members for the following 46 days after the initialization date since 2015, which are sometimes referred as ''near-real-time forecasts'' as they are made available after 3 weeks of issued. For each version of the model, which is updated every year, a set of retrospective forecasts (hindcasts) for the previous 20 years is also provided, but with a reduced number of ensemble members (11 ensemble members). The ECMWF model versions used in this study are as follows: CY41R1 for 2015 real-time forecasts and respective hindcasts to compute the climatology, CY41R2 for 2016 real-time forecasts and respective hindcasts to compute the climatology, and CY43R1 for June and CY43R3 since 11 July for 2017 real-time forecasts and respective hindcasts to compute the climatology. The hindcast evaluation was done using the 2017 versions of the model, which were operational until June 2018.

The ECMWF forecasts and hindcasts are initialized on Mondays and Thursdays of every week, but in this study only the Thursday initializations are used in as in Coelho et al. (2018). Following Weigel et al. (2008), Vitart and Molteni (2010) and Coelho et al. (2018), in this study we define the first week considering forecast/hindcast days 5–11, the second week as forecast/hindcast days 12–18, forecast/hindcast days 19–25 define the third week, and forecast/hindcast days 26–34 define the fourth week. These 4 weeks are referred to through the text of the manuscript as forecasts/hindcasts produced from one to four weeks in advance. As discussed in Coelho et al. (2018) the first four forecast/hindcast days after the initialization date are disregarded because these are considered daily medium-range forecast on the weather time scale and here the interest is on the extended-range (subseasonal) time scale.

All daily forecasts and hindcasts were weekly averaged (nonoverlapping). Weekly anomalies for each real-time forecast start date were computed with respect to the climatology of the same start date of the 20-yr weekly hindcasts. Also, the weekly hindcasts anomalies for each start date were computed with respect to the climatology of the same start date and the same set of 20-yr weekly hindcasts in a cross-validation (i.e., leaving one year out) framework.

Two target weeks were particularly analyzed, following what was described in the previous section and choosing the initialization dates according to the available real-time and associated retrospective forecasts. First, the week from Monday 17 July to Sunday 23 July 2017 was chosen, when intense cold anomalies were observed in central South America (Fig. 2e). Then, the week from Monday 24 July to Sunday 30 July 2017 was analyzed, during which warm anomalies extended through central and northern Argentina, Paraguay, Bolivia, Uruguay, and southern Brazil (Fig. 9e). Tables 1 and 2 show for each of these two target weeks, the initialization date, week number, valid days of forecasts, and days in advance in which the forecasts were issued.

To perform the verification study, the Japanese Meteorological Agency (JMA) 55-year Reanalysis (JRA-55) project (Kobayashi et al. 2015), postprocessed with a resolution of 1.25° (Japan Meteorological Agency 2013) were considered as observational reference. They were linearly interpolated to the same 1.5° × 1.5° grid in latitude and longitude as the ECMWF model. The verification framework proposed by Coelho et al. (2018), designed to assess hindcast and real-time forecast quality, was followed in this study. In this framework, three information levels are defined:

1) *Target week hindcast verification*: It provides a quality assessment of the predictions (11 ensemble members) produced for the target weeks (17–23 July and

TABLE 1. For the target week of 17–23 Jul 2017, their associated week number, initialization date, valid days of the forecast, and days in advance in which the first day of that week was forecast.

Week No.	Initialization date	Valid week	Valid days	Days in advance
1	Thursday 13 Jul 2017	Monday 17 Jul–Sunday 23 Jul	5–11	4
2	Thursday 6 Jul 2017	Monday 17 Jul–Sunday 23 Jul	12–18	11
3	Thursday 29 Jun 2017	Monday 17 Jul–Sunday 23 Jul	19–25	18
4	Thursday 22 Jun 2017	Monday 17 Jul–Sunday 23 Jul	26–32	25

24–30 July), from one to four weeks in advance, and based on the weeks available in the hindcasts (one for each hindcast year). Therefore, in this first verification level, the sample size is limited to the number of years for which hindcasts were produced (20 years).

2) *All-season hindcast verification*: The second verification level considers hindcasts (11 ensemble members) produced in all Thursdays start dates within the JJA season (14 Thursday start dates—of the ECMWF model version of 2017—over 20 years of hindcasts leading to a sample size equal to 280). This sample is considerably large and therefore allows a robust hindcast quality assessment.
3) *All-season real-time forecast verification*: The third verification level consists in aggregating the forecasts (51 ensemble members) produced on Thursdays during the JJA season and has a sample size equal to 40 (13 Thursdays for the 2015 and 2016 versions, and 14 Thursdays for the 2017 version). ECMWF ensemble mean anomalies for this level are computed using three different sets of hindcasts as follows: 1995–2014 hindcasts for the real-time forecast for 2015, 1996–2015 hindcasts for the real-time forecast for 2016, and 1997–2016 hindcasts for the 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.

In the all-season hindcast and real-time forecast verification levels, all Thursdays start dates within the JJA season are used. Even though this means that some of the verifying weeks (those start dates near the end of August) do not fall within JJA, we consider this does not affect the results considerably, as September is a month usually considered within the extended winter season when separating the year in two halves and the summer circulation is yet not established. We disregarded the option of evaluating only weeks of forecast within JJA to avoid the reduction of the sample size for the longer leads.

In each of the three levels of verification, the ensemble mean is used to perform the deterministic assessment and all the ensemble members are used to compute probabilities of the event of interest when performing the probabilistic assessment. The purpose of such a three levels verification framework is to provide supporting verification information in the form of maps and graphics to be examined together with the forecast maps at the time of issuing the forecast for the particular week of interest.

Following Coelho et al. (2018) a selection of metrics was used to evaluate some of the most fundamental forecast quality attributes. These include the correlation, given by the linear Pearson correlation coefficient to assess the strength of linear association between the ensemble mean forecast and the observed anomalies. The correlation coefficient was tested using a two-tailed Student's t test, reducing the sample size based on autocorrelation of the observations using the effective sample size proposed by Wilks (2011). Also, the mean squared error skill score, $\mathrm{MSSS} = (1 - \mathrm{MSE})/\mathrm{MSE}_c$, is used to assess deterministic skill with respect to climatology, where MSE is the mean squared error of the predicted ensemble mean temperature anomalies computed at each grid point over the available hindcast/forecast period and MSE_c is the mean squared error for a reference prediction. The constant climatological (null) temperature anomaly prediction was used

TABLE 2. For the target week of 24–30 Jul 2017, their associated week number, initialization date, valid days of the forecast, and days in advance in which the first day of that week was forecast.

Week No.	Initialization date	Valid week	Valid days	Days in advance
1	Thursday 20 Jul 2017	Monday 24 Jul–Sunday 30 Jul	5–11	4
2	Thursday 13 Jul 2017	Monday 24 Jul–Sunday 30 Jul	12–18	11
3	Thursday 6 Jul 2017	Monday 24 Jul–Sunday 30 Jul	19–25	18
4	Thursday 29 Jun 2017	Monday 24 Jul–Sunday 30 Jul	26–32	25

as reference in this paper. Moreover, a ratio of standard deviations was computed to assess the amplitude error of the ensemble mean predictions. The ensemble mean anomaly standard deviation is defined as the standard deviation among the ensemble mean of the temperature anomaly hindcast/forecast for each grid point and lead time. The observed standard deviation is computed among the observed temperature anomaly for every grid point for the 20 years of hindcasts or 3 years of real-time forecasts, according to the level of verification. The ratio between this two is used. This metric is used to complement the linear association assessment provided by the correlation between the ensemble mean forecast anomalies and the observed anomalies, as this ratio is part of one of the components of the MSSS decomposition.

The relative operating characteristic (ROC) curves of the probabilistic predictions for the event negative temperature anomaly collected over all South American grid points were also used. The aim is to assess overall discrimination (i.e., ability to successfully distinguish events from nonevents) after aggregating all available hindcasts/forecasts in space and time. Maps for the area under the ROC curve are also presented, which together with its p value were computed using NCAR's R verification package, which follows Mason and Graham (2002) using the Mann–Whitney U statistic to assess statistical significance. Last, reliability diagrams for ensemble derived probabilistic predictions issued for the event negative temperature anomaly collected over all South American grid points were computed 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.

3. Results

a. *Cold week: 17–23 July*

During the week of 17–23 July cold anomalies were observed in eastern-central and northern Argentina, as well as in Uruguay, Paraguay, Bolivia, and a large portion of Brazil, while warm anomalies were observed in the Patagonia region of Argentina in southern South America (Fig. 2e). Figures 2a–d show the deterministic forecasts represented by the ensemble mean temperature anomalies produced four to one week in advance and valid for that particular week. The pattern of observed anomalies could only be deterministically predicted by the ECMWF model one week in advance. The forecast issued one week in advance (Fig. 2d) displays temperature anomalies of similar magnitude to the observed ones (Fig. 2e). However, the forecast produced two weeks in advance (Fig. 2c) exhibits a cold anomaly considerably weaker than that observed and extended over most Argentina, hindering the temperature contrast observed between the north and the south parts of the country. On the other hand, the forecasts issued three to four weeks in advance (Figs. 2b and 2a, respectively) show predominantly positive temperature anomalies, with a pattern completely different from the observed anomalies (Fig. 2e).

To quantify the degree of correspondence between the deterministic forecasts in Figs. 2a–d and the observations in Fig. 2e, the area-weighted spatial linear Pearson correlation was computed, and is presented in the top right of each of the panels of Figs. 2a–d. As expected, the pattern correlation values are larger for shorter lead forecasts; however a modest correlation value of 0.65 was obtained for forecasts produced one week in advance, diminishing considerably for forecasts produced two weeks in advance (0.25). Longer-lead forecasts resulted in no spatial correlation.

Figures 2f–i show the forecast probability for the occurrence of negative anomalies during the target week of 17–23 July 2017. Forecast probabilities were computed as the fraction of the ensemble members (51) indicating a negative temperature anomaly. Blue (brown) colors represent regions where the model showed high (low) probabilities of cold anomalies. The model successfully indicated high probabilities for the occurrence of negative anomalies in tropical South America one week in advance (Fig. 2i). However, large probabilities of cold conditions were only forecast two weeks in advance over some regions in eastern Brazil and with reduced values in northeastern Argentina (Fig. 2h), while over Patagonia the model was unable to indicate the potential for the occurrence of anomalously warm conditions with that anticipation. The model also failed to forecast the potential for the occurrence of cold conditions from three to four weeks in advance (Figs. 2f,g).

The performance of the ECMWF model in forecasting the week of 17–23 July 2017 can be compared to the historical performance of the ECMWF deterministic (ensemble mean) predictions produced for past years using the three levels verification framework: (i) using the retrospective forecasts for the same target week (target week hindcast verification), (ii) using the retrospective forecasts for all weeks within the JJA season (all-season hindcast verification), and (iii) using the forecast produced on real time for all weeks within the JJA season (all-season real-time forecast verification). Such an assessment when performed in conjunction

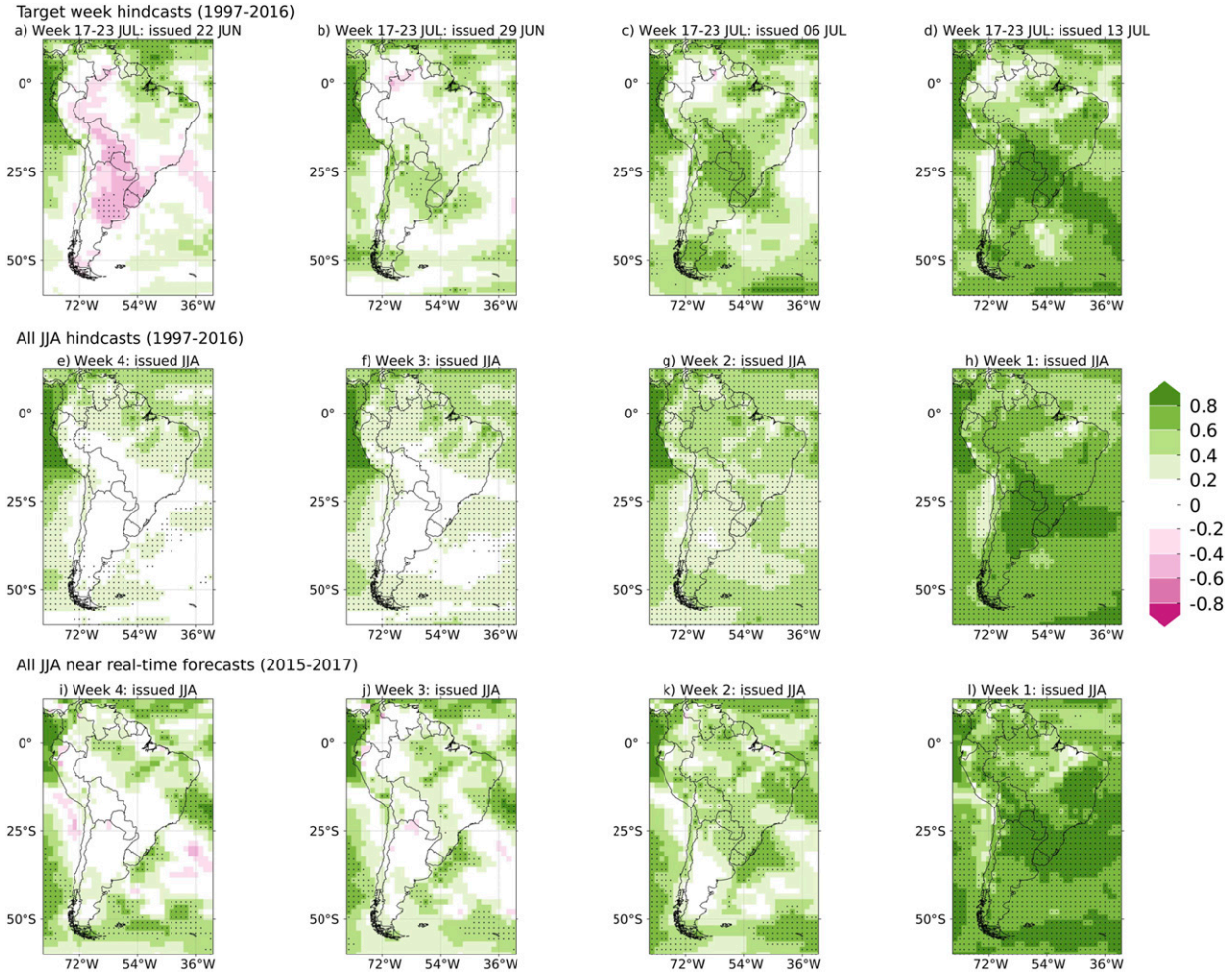

FIG. 3. Maps of correlation between the ECMWF ensemble mean temperature anomaly prediction produced from four to one week in advance (shown from left to right) and the corresponding observed (JRA-55) temperature anomalies at each grid point for (a)–(d) the target week hindcast verification sampling strategy (20 samples), (e)–(h) the all-season hindcast verification sampling strategy (280 samples), and (i)–(l) the all-season real-time forecast verification sampling strategy (40 samples) described in the text. 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.

with the forecast to be issued for the particular week of interest is useful to reinforce and expand the confidence levels underpinning the weekly forecasts (Coelho et al. 2018).

To assess the strength of the linear association, maps of correlation between the observed temperature anomalies and the predicted ensemble mean produced four to one week in advance are presented in Fig. 3 for the three levels of verification: target week hindcast verification (Figs. 3a–d), all-season hindcast verification (Figs. 3e–h) and all-season real-time forecasts verification (Figs. 3i–l). As expected, all levels of verification applied to hindcasts/forecasts produced 1–2 weeks in advance show larger and statistically significant linear association than those for hindcasts/forecasts produced 3–4 weeks in advance, mainly over northern Argentina, southeastern Brazil, and Paraguay. The real-time forecasts for the JJA 2015–17 period produced one week in advance (Fig. 3l) show larger correlation in eastern Brazil than the JJA 1997–2016 hindcasts (Fig. 3h). Figure 3c shows that the hindcasts produced two weeks in advance for the target week in the 1997–2016 period presented better association in northeastern Argentina, Paraguay and southern Brazil than the hindcasts considering all JJA Thursday initializations (Fig. 3g). This feature was also noticed for forecasts produced three weeks in advance, but with only a restricted number of grid points showing statistically significant correlation values (Fig. 3b). Figure 3a shows negative correlation coefficients over southeastern South America for

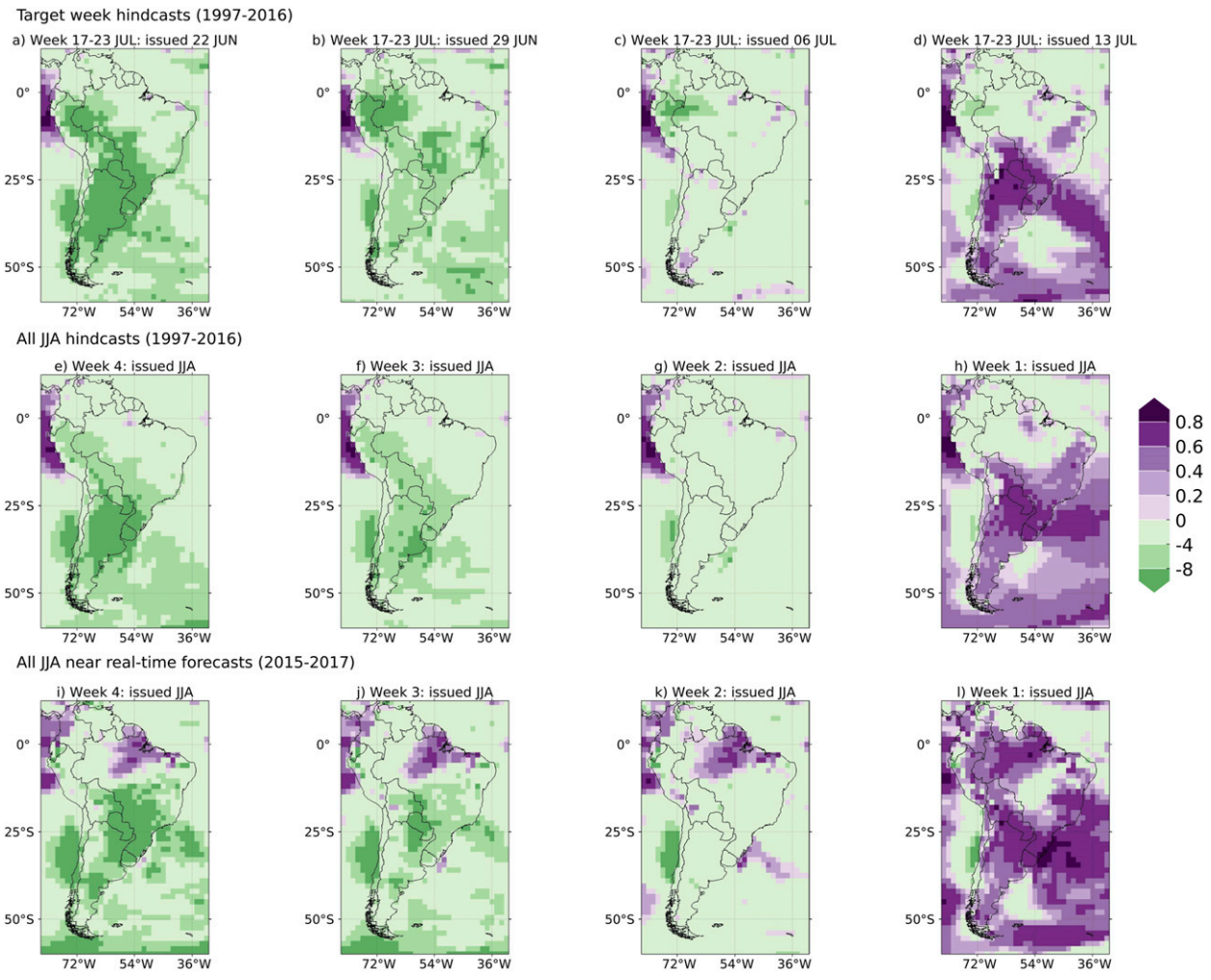

FIG. 4. Maps of MSSS with respect to climatology for the ECMWF ensemble mean temperature anomaly predictions produced from four to one week in advance (shown from left to right) for (a)–(d) the target week hindcast verification sampling strategy (20 samples), (e)–(h) the all-season hindcast verification sampling strategy (280 samples), and (i)–(l) the all-season real-time forecast verification sampling strategy (40 samples) described in the text.

forecasts produced four weeks in advance for the week of 17–23 of July, indicating an opposite phase between the forecast temperature anomalies and the observations. Neither in hindcasts nor in real-time forecasts produced three and four weeks in advance during the JJA season a positive statistically significant correlation was obtained over the central and southeastern South America region affected by the marked negative temperature anomalies (Figs. 3e,f,i,j).

The MSSS was computed for the three verification sampling strategies to study the skill of the hindcasts/forecasts compared to the climatological prediction. Most of subtropical and extratropical South America show positive values of MSSS for hindcasts/forecasts produced one week in advance, revealing improved accuracy respect to climatology (Fig. 4). For the three levels of verification, the regions of positive MSSS for forecasts produced one week in advance (Figs. 4d,h,l) match those of largest correlation coefficients (Figs. 3d,h,l). This accordance between the regions with improved accuracy and smaller phase error (i.e., the forecast/hindcast anomalies oscillate in phase with the observed anomalies) suggest that the correlation component of the MSSS decomposition contributes considerably for the identified positive skill shown with the MSSS (Fig. 4). On the other hand, the target week and all-season hindcast verification levels present no skill over most of the region when producing the forecasts 2–4 weeks in advance (Figs. 4a–c,e–g). There is, however, a region in northern Brazil in which the real-time forecasts verification present forecast skill up to four weeks in advance (Figs. 4i–k).

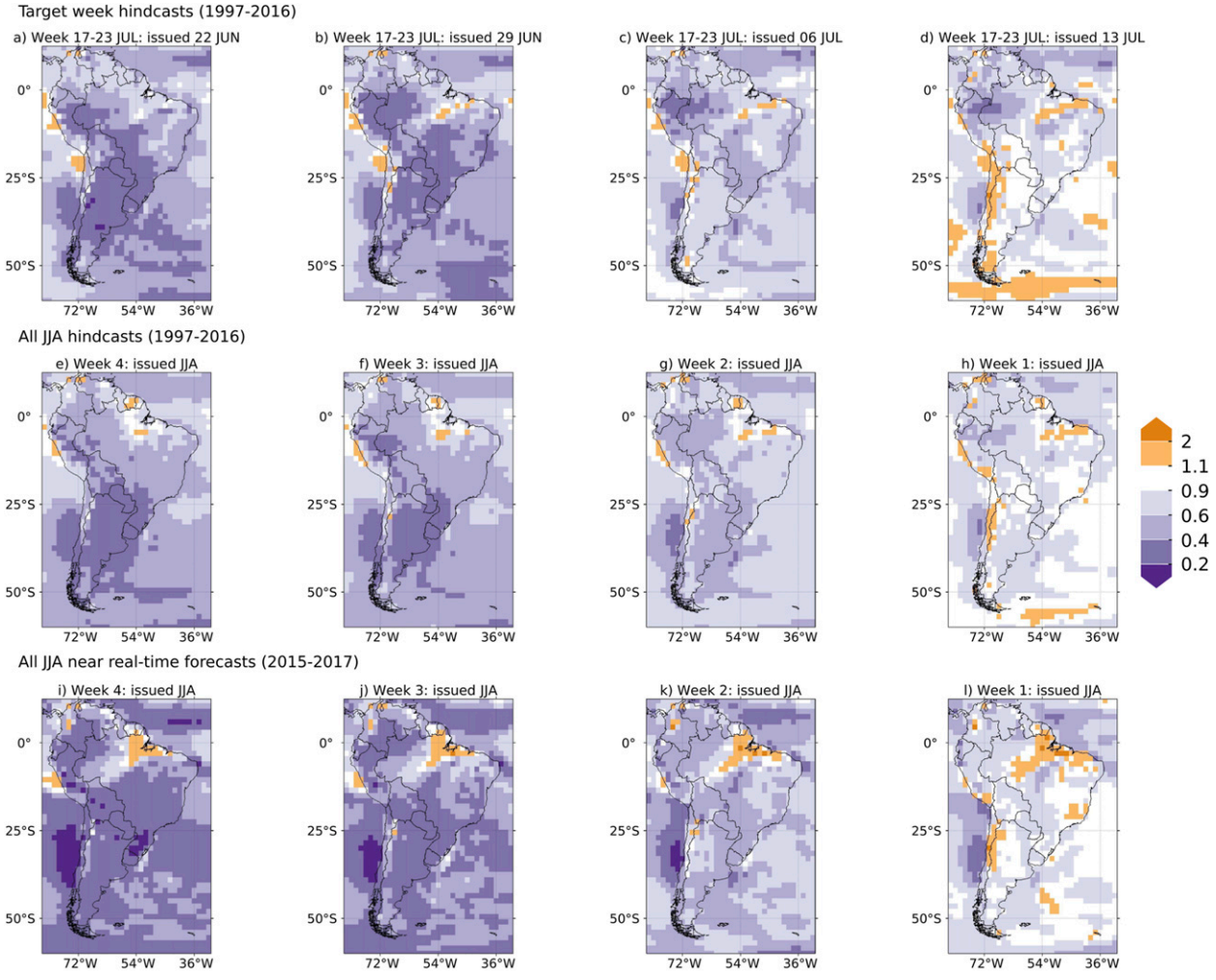

FIG. 5. Maps of the ratio of the predicted ECMWF ensemble mean temperature anomaly standard deviation and the observed temperature anomaly standard deviation for predictions produced from four to one week in advance (shown from left to right) for (a)–(d) the target week hindcast verification sampling strategy (20 samples), (e)–(h) the all-season hindcast verification sampling strategy (280 samples), and (i)–(l) the all-season real-time forecast verification sampling strategy (40 samples) described in the text.

The amplitude error of the ensemble mean forecast was studied through computing the ratio between the predicted temperature ensemble mean anomaly standard deviation and the observed temperature anomaly standard deviation (Fig. 5). This amplitude error of the ensemble mean is different (and larger) than the amplitude error of the individual ensemble members. In regions where this ratio is smaller than unity a large amplitude error exists in the forecasts, and the predicted temperature anomalies present lower variability than the variability of the observed anomalies. That is the case over most of the continent for forecasts issued 2–4 weeks in advance (Fig. 5), with some exceptions mainly in northern Brazil and particularly for the real-time forecasts (Figs. 5i–k). The amplitude error is smaller for forecasts issued one week in advance for all verification sampling strategies (Figs. 5d,h,l).

The ability of the model to successfully discriminate cold (negative anomaly) from warm (positive anomaly) events over South America was assessed by analyzing maps of the area under the ROC curve at each grid point for the three verification levels, which are presented in Fig. 6. The forecast is better at distinguishing cold from warm events in the regions where the area under the ROC curve is larger than 0.5. This was generally the case over most of the continent for forecasts produced one and two weeks in advance in all verification levels. For the target week hindcast verification, forecasts issued one week in advance show values overall larger than 0.7 and even greater than 0.9 in central, northeastern, and

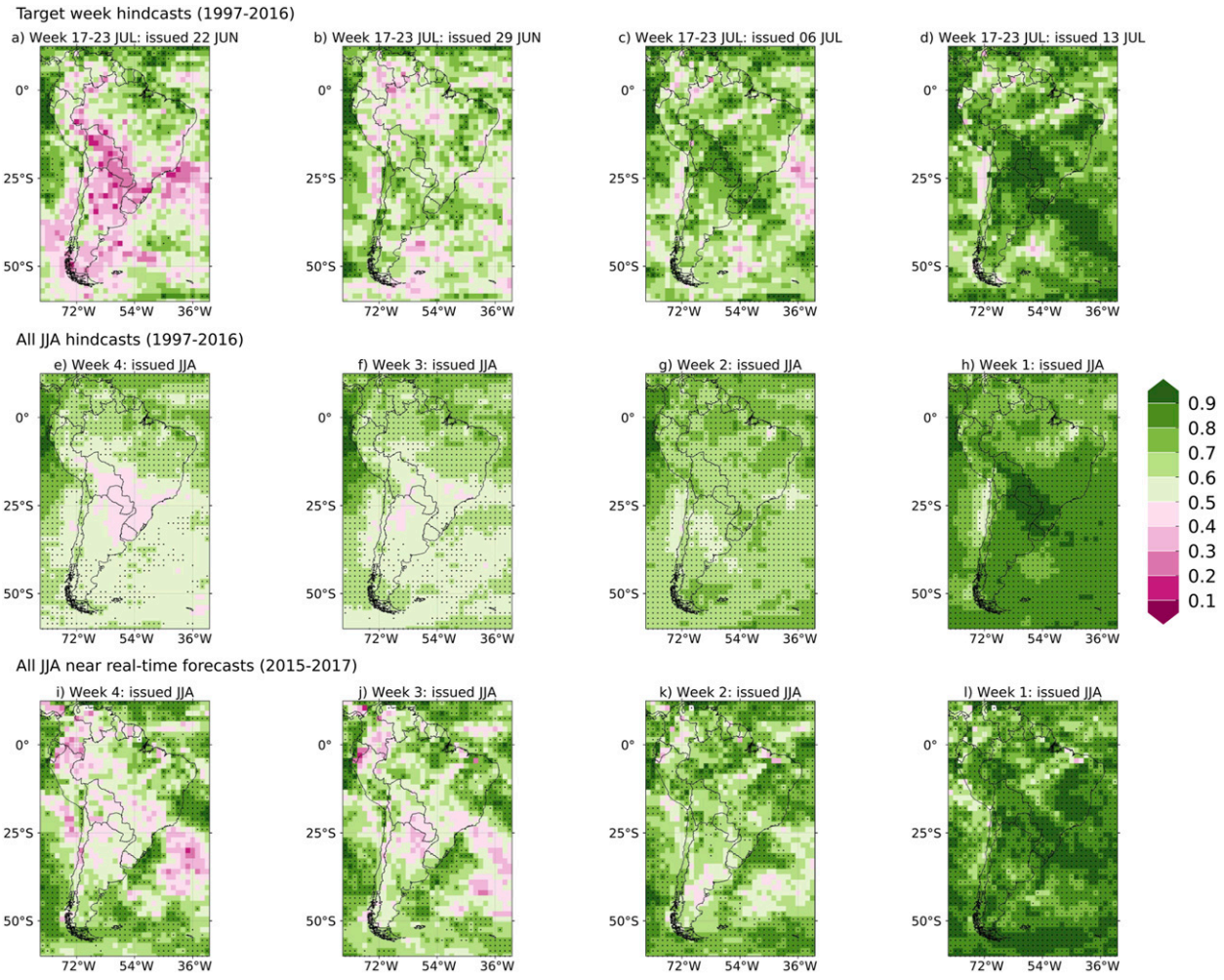

FIG. 6. Maps of area under the ROC curve computed for ECMWF forecast/hindcast probabilities for the occurrence of the event negative temperature anomaly produced from four to one week in advance (shown from left to right) at each grid point for (a)–(d) the target week hindcast verification sampling strategy (20 samples), (e)–(h) the all-season hindcast verification sampling strategy (280 samples), and (i)–(l) the all-season real-time forecast verification sampling strategy (40 samples) described in the text. 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.

southeastern South America (Fig. 6d), and values mostly greater than 0.6 for forecasts produced two weeks in advance (Fig. 6c). Discrimination of cold events for forecasts for the target week issued 3 weeks in advance show more deficient discrimination ability in central and southeastern South America. There is much reduced discrimination ability for those issued 4 weeks in advance except for northern Brazil (Figs. 6b,a). For the all-season hindcast strategy, the larger sample accounts for smoother patterns, and forecasts issued one week in advance were able to discriminate a cold event from a warm event over most of the continent, particularly over northeastern Argentina, Paraguay, and southeastern Brazil (Fig. 6h), and similarly for forecasts issued two weeks in advance (Fig. 6g). For longer leads the best discrimination ability is achieved north of 15°S. Last, the third level of verification (Figs. 6i–l) shows good discrimination for forecasts issued one week in advance, particularly over eastern and southeastern Brazil, and for forecasts issued 2–4 weeks in advance show a more scattered behavior, with less-coherent spatial patterns.

A similar assessment was performed for the events ''temperature anomaly in the lower tercile'' and ''temperature anomaly in the upper tercile'' to analyze differences in the discrimination ability of the model for those events. The all-season hindcast verification level for these events is shown in Fig. S2. The area under the ROC curve for both events is overall similar and to Figs. 6e–h, with a slight difference for hindcasts

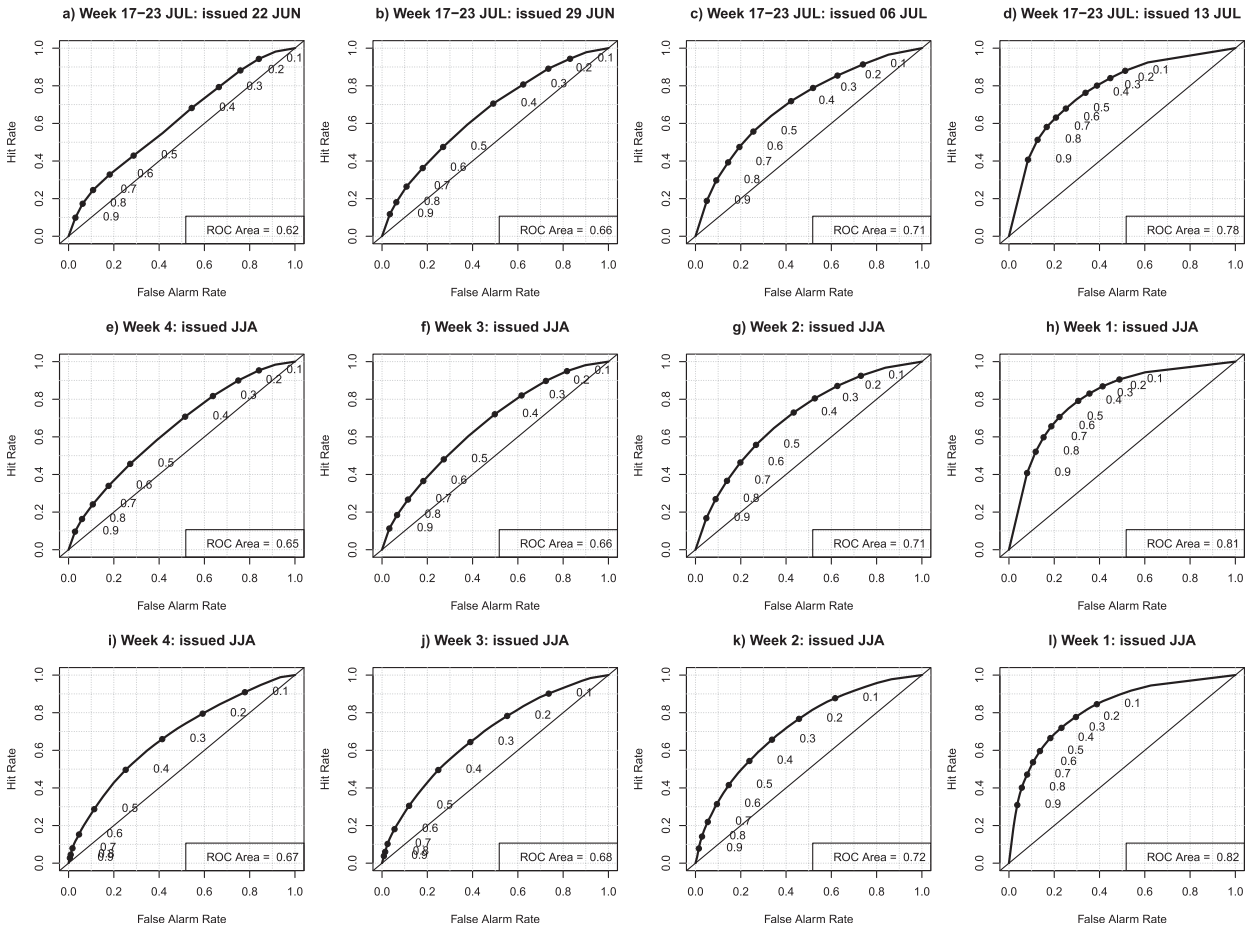

FIG. 7. ECMWF ROC curves for ensemble derived probabilistic predictions issued for the event negative temperature anomaly collected over all South American grid points, aggregating all available forecasts/hindcasts in space and time, produced from four to one week in advance (shown from left to right), for (a)–(d) the target week hindcast verification sampling strategy (20 samples), (e)–(h) the all-season hindcast verification sampling strategy (280 samples), and (i)–(l) the all-season real-time forecast verification sampling strategy (40 samples) described in the text. 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 negative temperature anomaly.

produced one week in advance, which shows a higher score over northeastern Argentina and Uruguay for warm events respect to the cold events (Fig. S2).

In addition, overall discrimination was assessed by computing the ROC curves for ensemble derived probabilistic prediction issued for the event negative temperature anomaly collected over all grid points within the domain shown in the figures, aggregating all available hindcasts/forecasts in space and time, as presented in Fig. 7. Therefore, the ROC curves represent the hit rate versus false alarm rate when evaluating probability forecast of negative temperature anomalies, considering the cases when the anomaly was forecast with at least 10% of probability, 20%, up to at least 90%. Hindcast/forecast probabilities were determined by computing the fraction of ensemble members indicating a negative temperature anomaly. In general, there are larger areas under the ROC curve for shorter leads, irrespective to the three verification levels here investigated. When analyzing only the ROC curves for the target week (17–23 July, Figs. 7a–d) presented for each lead (one to four weeks in advance), there is slightly less or equivalent discrimination ability than the all-season hindcasts (Figs. 7e–h) and real-time forecasts (Figs. 7i–l) verification levels.

Considering the probabilistic predictions issued for the event negative temperature anomaly, the forecast is considered to be reliable if the forecast probability of the event corresponds to the expected frequency of observing negative temperature anomalies. Figure 8 shows the reliability diagrams for the ensemble derived probabilistic predictions issued for the event negative temperature anomaly collected over all grid points, aggregating all available hindcasts/forecasts in space and

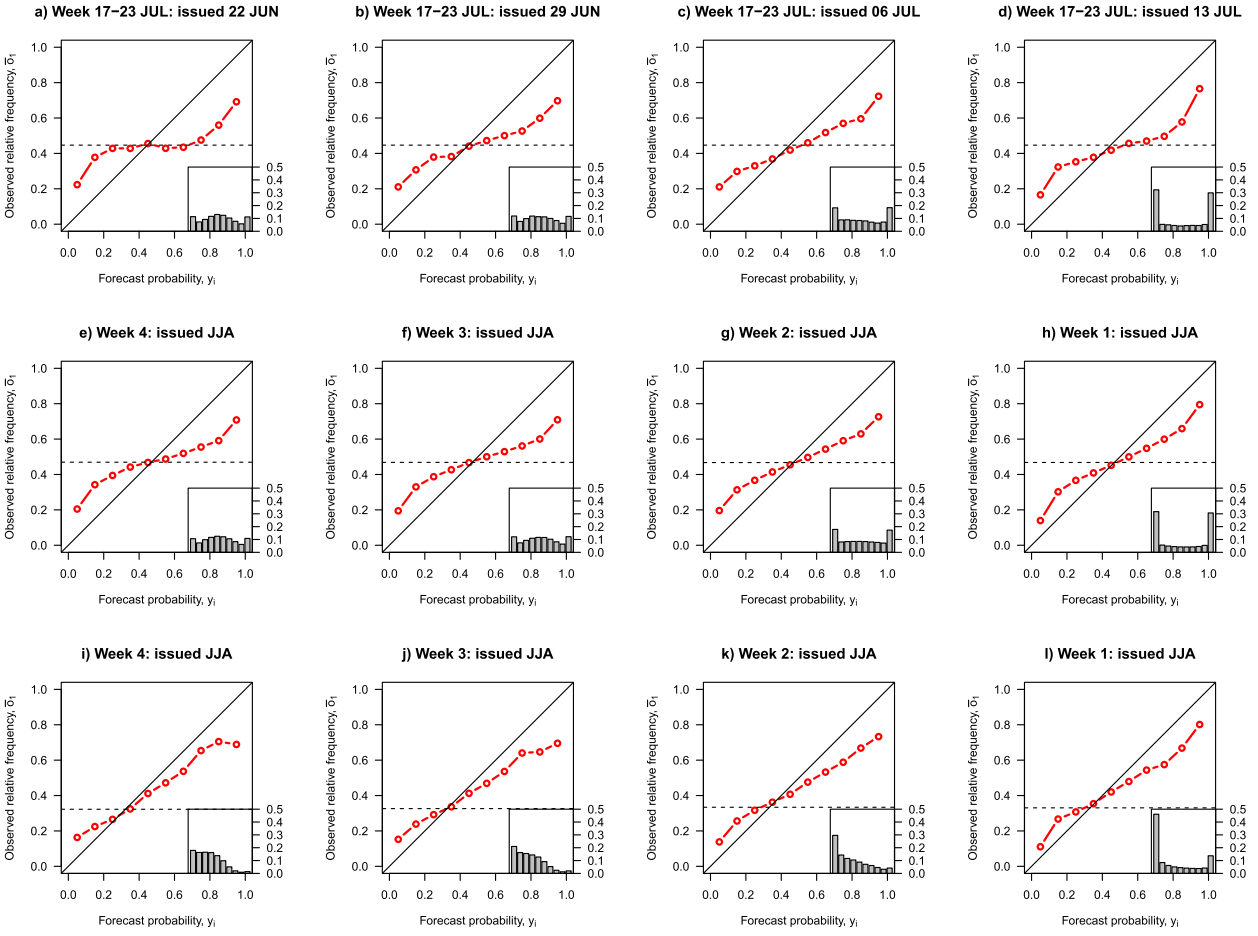

FIG. 8. ECMWF reliability diagrams for ensemble derived probabilistic predictions issued for the event negative temperature anomaly collected over all South American grid points, aggregating all available forecasts/hindcasts in space and time, produced from four to one week in advance (shown from left to right), for (a)–(d) the target week hindcast verification sampling strategy (20 samples), (e)–(h) the all-season hindcast verification sampling strategy (280 samples), and (i)–(l) the all-season real-time forecast verification sampling strategy (40 samples) described in the text. Sharpness diagrams for each case are plotted in the bottom right of the reliability diagrams, and relative frequencies of the predicted event are shown in each of 10 bins: 0%–10%, 10%–20%, 20%–30%, 30%–40%, 40%–50%, 50%–60%, 60%–70%, 70%–80%, 80%–90%, and 90%–100%.

time. The closer the curves to the diagonal, the more reliable the forecast is; if the curve is below (above) the diagonal, forecast probabilities are higher (lower) than observed frequencies, and therefore indicates overforcasting (underforcasting). If the curve is flat, the forecast presents no resolution. Overall, predictions are overconfident and show poor resolution particularly for hindcasts produced 4 weeks in advance. Better reliability is observed for the real-time forecasts issued for all JJA season (Figs. 8i–l).

The histograms in the bottom right of each panel in Fig. 8 represent the relative frequency of the issued forecast probabilities of the event negative temperature anomaly falling into each of the 10 bins of forecast probability (0%–10%, 10%–20%, and so on up to 90%–100%): the sharpness diagram. The forecasts are sharper for shorter leads, with the histograms peaking at the two extreme bins, particularly when produced one week in advance. For this lead time the event is forecast around 30% of the times with a 0%–10% and also 30% of the time with 90%–100% probability in the target week and all-season hindcasts levels (Figs. 8d,h). When produced two weeks in advance, the event is forecast around 20% of the times with those probabilities, losing sharpness but still forecasting the extreme bins more often than the rest (Figs. 8c,g). Differently, the sharpness diagram for the real-time forecasts show that the event negative temperature anomaly is forecast around 50% of the times with a probability of 0%–10%, and around 13% of the times with a probability of 90%–100% (Fig. 8l). This sampling verification level also shows that higher relative

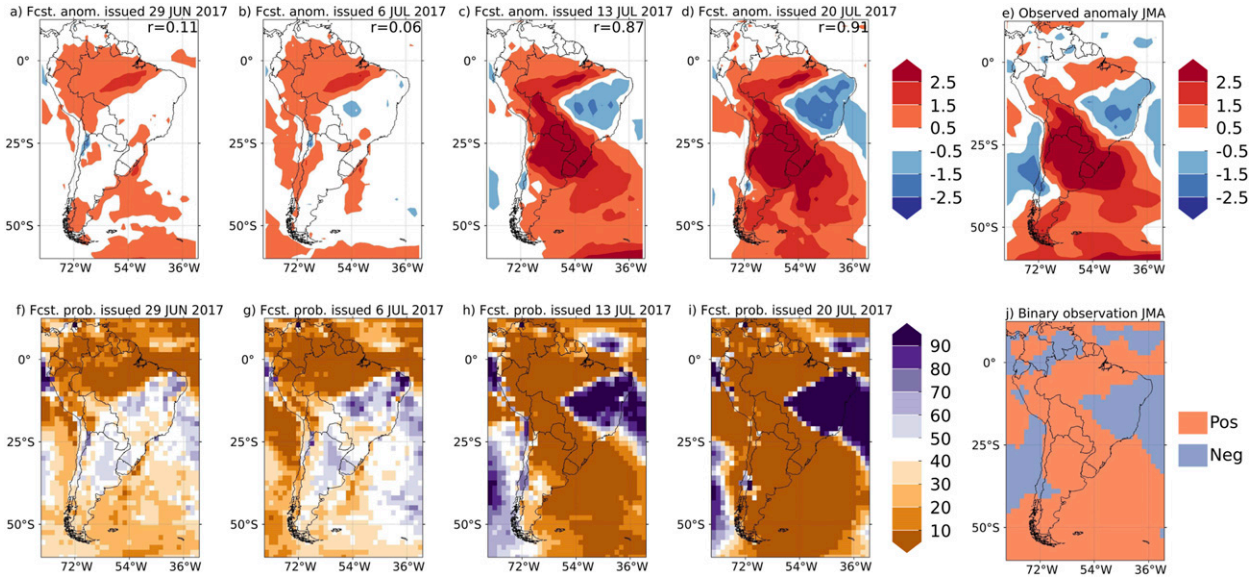

FIG. 9. ECMWF ensemble mean forecast temperature anomalies for the target week of interest (24–30 Jul 2017) initialized on (a) 29 Jun 2017, (b) 6 Jul 2017, (c) 13 Jul 2017, and (d) 20 Jul 2017, representing forecasts produced from 4 to 1 week in advance as described in the text. (e) Observed temperature anomalies for the week 24–30 Jul 2017 with respect to the 1995–2017 period. ECMWF forecasts probabilities for the occurrence of negative temperature anomaly during the target week of interest (24–30 Jul 2017) initialized on the (f) 29 Jun 2017, (g) 6 Jul 2017, (h) 13 Jul 2017, and (i) 20 Jul 2017. (j) Binary observation indicating where a negative (blue) or a positive (red) temperature anomaly was recorded during the week of 24–30 Jul 2017.

forecast frequencies are noticed for the lower probability bins (Figs. 8i–k).

b. Warm week: 24–30 July

A warm week followed the cold conditions of the week of 17–23 July in central South America. Between 24 and 30 July, warm anomalies were observed in central and north Argentina, Uruguay, Paraguay, Bolivia, and southeastern Brazil, while cold anomalies were observed in eastern Brazil (Fig. 9e). In this section we assess the performance of ECMWF model in forecasting the temperature anomalies for this anomalously warm week and compare it with the performance of the previous anomalously cold week presented in the preceding section. For this assessment the three level verification framework used for evaluating forecast quality for the previous cold week and providing supporting verification information together with forecast information at the time of issuing the forecast now for the warm target week is also used.

The deterministic forecasts given by the ensemble mean temperature anomalies (Figs. 9a–d) show a relatively good similarity for forecasts issued one and two weeks in advance, with forecast anomalies (Figs. 9c,d) of similar magnitude to the observed anomalies (Fig. 9e). Nevertheless, the forecasts produced three to four weeks in advance (Figs. 9b,a) failed to simulate the observed large positive temperature anomalies (Fig. 9e). For this week, the linear Pearson pattern correlation, shown in the top right of each panel of Figs. 9a–d, was found to be large for forecasts produced one and two weeks in advance (0.90 and 0.86, respectively); however, for longer-lead forecasts the spatial correlation was much reduced. Compared to the spatial correlations for forecasts produced one to two weeks in advance for the previous cold week (Figs. 2d,c; 0.65 and 0.26, respectively), the warm week resulted in a much better forecast pattern match than the cold week. As warm temperature anomalies were observed previous to the cold week of 17–23 July and also after that, this long-lasting warmer-than-normal event might have been responsible of the better forecast for longer leads for this target week.

The probabilistic forecasts were also found to be good when issued up to two weeks in advance, but not for longer leads, in central and southern South America. Figures 9f–i show the forecast probability of occurrence of negative anomalies during the target week. Blue (brown) colors indicate regions where the model showed high (low) probabilities for the occurrence of cold anomalies and low (high) probabilities for the occurrence of warm anomalies. The model successfully indicated low probabilities for the occurrence of negative anomalies in most South America, and therefore, high probabilities for the occurrence of positive anomalies, and high probabilities for the occurrence of

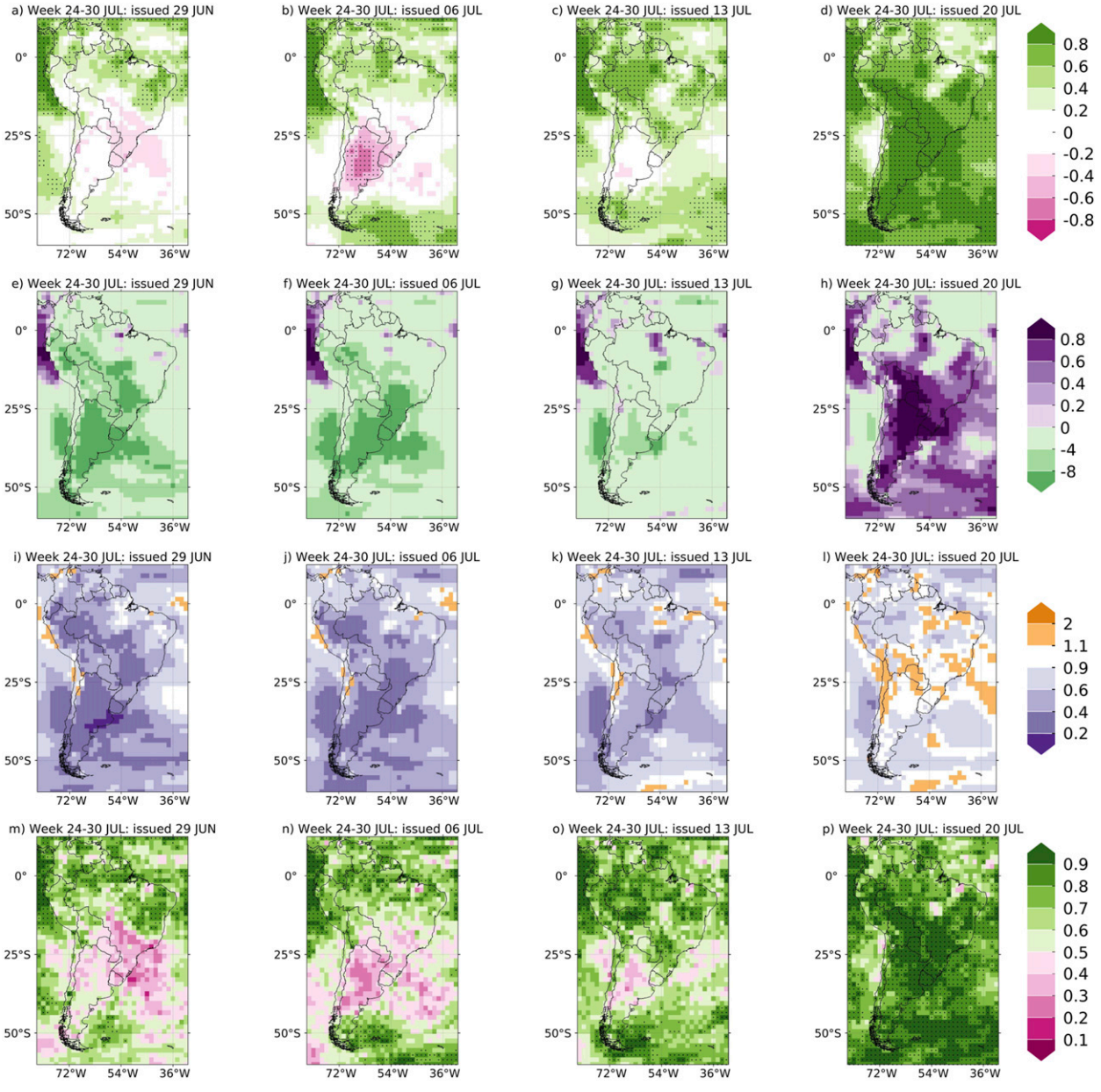

FIG. 10. Verification metrics for the week 24–30 Jul 2017 for the target week hindcast verification sampling strategy (20 samples): (a)–(d) maps of correlation between the ECMWF ensemble mean temperature anomaly prediction produced from four to one week in advance and the corresponding observed temperature anomalies (JRA-55) at each grid point, (e)–(h) maps of MSSS with respect to climatology the ECMWF ensemble mean temperature anomaly predictions produced four to one week in advance, (i)–(l) maps of the ratio of the predicted ECMWF ensemble mean temperature anomaly standard deviation and the observed temperature anomaly standard deviation for predictions, and (m)–(p) maps of area under the ROC curve computed for ECMWF hindcast probabilities for the occurrence of the event negative temperature anomaly at each grid point.

cold anomalies in eastern Brazil 1–2 weeks in advance (Figs. 9i,h). The forecasts (Figs. 9g,f) failed to indicate the potential for the occurrence of warm conditions in central and southeastern South America three to four weeks in advance. As when analyzing the deterministic forecast of the ensemble mean, the probabilistic forecast produced one week in advance and particularly the forecast produced two weeks in advance were better for this target week (Figs. 9i–h) compared to the previous cold one (Figs. 2i,h).

The verification metrics for hindcasts produced from four to one week in advance of the week of 24–30 July

and for the target week hindcast verification level are shown in Fig. 10. These include maps of correlation between the observed temperature anomalies and the predicted ensemble mean (Figs. 10a–d), maps of the MSSS for hindcast temperature anomalies computed with respect to the climatological prediction (Figs. 10e–h), maps of the ratio between the predicted temperature ensemble mean anomaly standard deviation and the observed temperature anomaly standard deviation (Figs. 10i–l) and maps of the area under the ROC curve at each grid point (Figs. 10m–p). In this case, only the first sampling strategy is computed, as the all-season hindcast and all-season real-time forecast verification levels are the same as shown in Figs. 3–6 for the previously investigated week, as both belong to the JJA season. However, in the proposed framework used here verification information of the three levels are compared and examined together with the forecast information for the target week of interest.

There is a vast region of large positive correlation spanning most of Argentina, southern Brazil, Uruguay, Paraguay and Bolivia for the target week forecast verification level produced one week in advance (Fig. 10d), which reveals a small phase error (i.e., the forecast anomalies generally oscillate in the direction of the observed anomalies), but the correlation values decrease considerably over this region for hindcasts produced two to four weeks in advance, with regions presenting statistically significant positive correlation values scattered over specific regions (Figs. 10a–c). Comparing to the week of 17–23 July of the hindcasts analyzed in the previous section (Figs. 3a–d) produced one week in advance, larger regions of correlation greater than 0.8 are noticed for 24–30 July week (Fig. 10d). Hindcasts produced two weeks in advance show statistically significant correlation values greater than 0.4 mostly over western Brazil (Fig. 10c), as opposed to eastern Bolivia, Paraguay, northern Argentina, and southern Brazil for the earlier week (Fig. 3c). When comparing to the other verification levels (all-season hindcasts and real-time forecasts, Figs. 3e–l), the target week verification level for hindcasts produced one week in advance show the largest region with positive correlation values (Fig. 10d). Moreover, for hindcasts produced two weeks in advance western Brazil presented the highest correlation values (Fig. 10c), and this region also appear showing moderate correlation in the all-season hindcast sampling strategy (Fig. 3g) but not in the real-time forecasts (Fig. 3k).

The maps of MSSS (Figs. 10e–h) show that regions with improved accuracy compared to the climatological prediction are only noticed for hindcasts produced one week in advance and mostly concentrated over the same region as in the other two verification levels (Figs. 4h,l). Maps of the ratio between the predicted temperature ensemble mean anomaly standard deviation and the observed temperature anomaly standard deviation (Figs. 10i–l) for hindcasts produced one week in advance show, as for the previous target week, that for the week of 24–30 July the ratio is smaller than unity (and then, there is a large amplitude error in the predicted anomalies) only in some regions, particularly over northern South America (Fig. 10l). Forecasts produced two to four weeks in advance show large amplitude errors in most South America (Figs. 10i–k). Finally, the maps of the area under the ROC curve (Figs. 10m–p) show that for hindcasts produced one week in advance, the model can discriminate a cold from a warm event over most of South America, and only to the north of approximately 15°S for longer leads. The target week verification level for 24–30 July indicates overall lower discrimination ability at grid point level over central South America (Figs. 10m–o) than the previous week (Figs. 6a–d) and the all-season hindcast (Figs. 6e–h) and real-time forecasts (Figs. 6i–l), particularly for forecasts issued 2–4 weeks in advance. When comparing forecasts produced 3 and 4 weeks in advance for the first level of verification for the cold and warm target weeks, some similarities arise in the verification scores maps. This may be associated to the forecast quality of the investigated model.

ROC curves, reliability and sharpness diagrams for ensemble derived probabilistic prediction issued for the event negative temperature anomaly collected over all South American grid points for the target week hindcast sample strategy are presented in Fig. 11. Shorter lead forecasts show the larger areas under the ROC curve and the area for hindcasts produced one week in advance is greater for the week of 24–30 July (Fig. 11d) than for the week of 17–23 (Fig. 7d). The reliability diagrams for the week of 24–30 July reveal better reliability and resolution for shorter leads, which are also better than for the previously investigated week of 17–23 July (Figs. 8a–d), but the hindcasts seem to be equally sharp.

4. Discussion and conclusions

This paper presented an assessment of ECMWF subseasonal temperature predictions for two anomalous cold and warm weeks of July 2017 in South America, following the three level verification framework designed by Coelho et al. (2018). Ensemble mean forecasts were able to predict the cold anomaly of the week of 17–23 July 2017 one week in advance and the warm anomaly of the week of 24–30 July 2017 up to two weeks in advance.

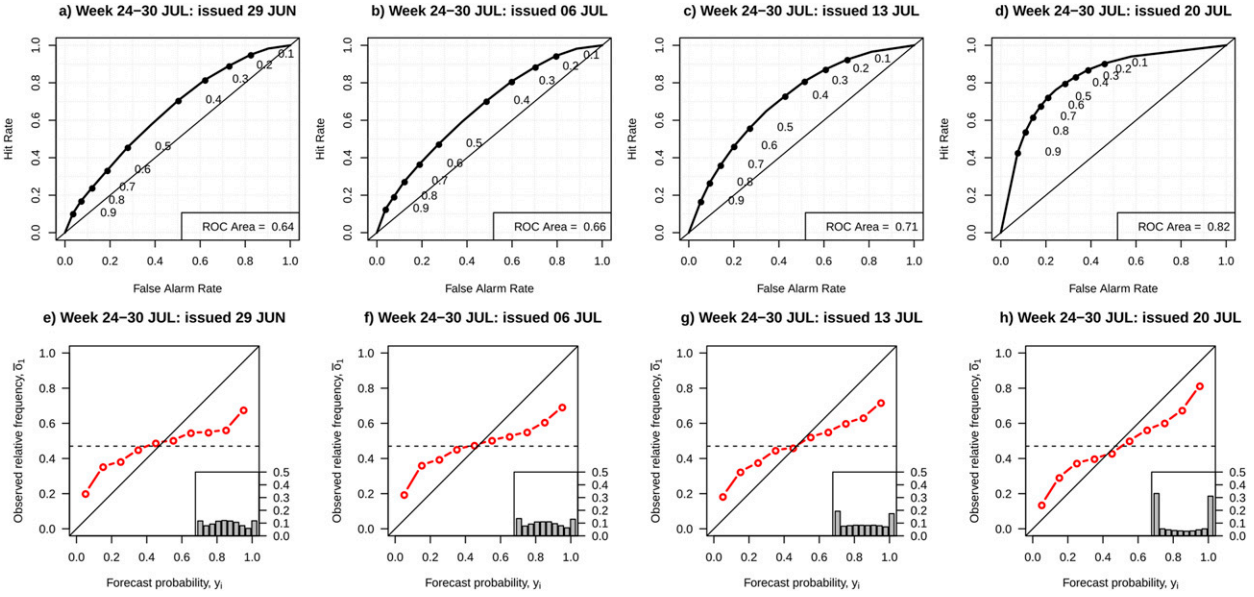

FIG. 11. Verification curves and diagrams for the week 24–30 Jul 2017 for the target week hindcast verification sampling strategy (20 samples): (a)–(d) ECMWF ROC curves for ensemble derived probabilistic predictions issued for the event negative temperature anomaly collected over all South American grid points, aggregating all available forecasts/hindcasts in space and time, produced four to one week in advance, and (e)–(h) ECMWF reliability diagrams for ensemble derived probabilistic predictions issued for the event negative temperature anomaly collected over all South American grid points, aggregating all available forecasts/hindcasts in space and time, produced from four to one week in advance.

These findings were consistent in both deterministic and probabilistic forecasts here assessed.

A set of verification metrics were computed for three verification levels: considering all target weeks in the 20 years of hindcast period, aggregating all hindcasts produced on Thursdays within the JJA season along 20 years and collecting all real-time forecasts produced during the JJA season for 2015, 2016, and 2017. The metrics included correlation, mean squared skill score, the ratio of the standard deviations of the ensemble mean anomaly and the observed temperature, area under the ROC curve, ROC curves and reliability diagrams. These metrics were computed to be used together with the forecasts produced for the target week of interest in order to help identify regions where the forecasts have best performance. As the cold and warm anomalies were most pronounced over central and southeastern South America during the two weeks here investigated, and forecasts produced three and four weeks in advance generally showed poor performance, the following discussion is focused in those regions and in forecasts produced one and two weeks in advance (valid days 5–11 and 12–18, respectively), which showed better performance, in order to answer the questions posed in the introduction.

A smaller phase error was noticed in central and southeastern South America for the target week 17–23 July hindcast verification strategy compared to the all-season JJA hindcast verification strategy, particularly when produced two weeks in advance. The MSSS resulted mostly similar in both verification strategies and lead times, which revealed that in both cases the skill with respect to the (null) climatology is similar. Also, the ratios of standard deviations was similar and mostly lower than 1, revealing large amplitude errors in the hindcasts of the target week and in the all-season verification strategies. The discrimination between cold and warm event for the target week 17–23 July hindcast verification was higher than for the all-season hindcast verification strategy, as shown by the area under the ROC curve, and particularly over central and southeastern South America for hindcasts produced two weeks in advance. When collecting all grid points in the study region, the all-season hindcast verification strategy resulted in more reliable predictions than the target week hindcast verification strategy produced one week in advance.

On the other hand, when comparing the target week of 24–30 July to the all JJA hindcasts verification levels, the linear association was higher for the target week over southern South America only and for forecasts produced one week in advance. However, for forecasts produced two weeks in advance, the phase error resulted smaller when considering all JJA initializations.

Forecasts produced one week in advance of 24–30 July showed locally some regions with ratios of standard deviation greater than 1.1, revealing that the predicted temperature anomalies present higher variability than that of the anomalies, mostly over Paraguay and eastern Bolivia. This is not seen in the all-season hindcasts sampling strategy. Forecast discrimination of positive versus negative anomaly in forecasts produced two weeks in advance of the target week was not good in central and northern Argentina, only in southeastern Brazil two weeks in advance, and reliability for forecasts produced one and two weeks in advance were as reliable for the target week as for the all JJA season verification level.

The all-season hindcast verification strategy and the all-season real-time forecast verification strategy for 2015–17 were compared using the same verification metrics used earlier. The correlation between the forecast and observed temperature anomalies was larger for the real-time forecast verification strategy over central South America and eastern Brazil than for the all-season hindcast verification strategy for predictions produced two weeks in advance. There is also a region off the coast of southern Brazil and Uruguay with remarkable correlation levels even for predictions produced 3 weeks in advance. The MSSS is quite similar in both cases, except for northern Brazil for all leads, where the score is positive (better skill than climatology) in the real-time forecasts. Recently, Gubler et al. (2020) found that the ECMWF SEAS5 model seasonal forecasts achieve a good performance for seasonal temperature predictions over this region, which the authors attribute to the relatively high influence of ENSO there. ENSO influence may also be providing a source of predictability for the Amazon region. Real-time forecasts produced two weeks in advance show larger area under the ROC curve over northern and eastern Brazil and off the coast, compared to the all-season hindcast verification strategy. The performed assessment using ROC curves and reliability diagrams revealed slightly better discrimination, reliability, and resolution for the real-time forecast verification strategy compared to the all-season hindcast verification strategy. However, the sharpness of the predictions of these two strategies is quite different, with the real-time forecasts presenting higher frequency for low probabilities, dropping to near zero frequency for forecasts produced two or more than two weeks in advance.

As discussed in Coelho et al. (2018), the verification framework used in this study should be used being aware of its advantages but also its limitations. When comparing the target week and the all-season hindcast verification, sampling might be responsible for some differences in the skill scores, and therefore one good or bad forecast may have larger influence in the first level of verification. Nonetheless, the 20 samples are above the suggested number of hindcast seasons when verifying seasonal forecasts, according to the World Meteorological Organization. The third level of verification, which uses all-season real-time forecasts, should be analyzed considering that sampling due to interannual variability, quality and ensemble size influence the results when comparing to the 20-yr hindcast analysis (Coelho et al. 2018). Sampling of this third level of verification spans only three years and therefore its quality is affected by interannual variability (e.g., El Niño–Southern Oscillation, year-to-year activity of the MJO); the better quality of the initial conditions used to initialize the real-time forecasts might also be one of the reasons of the higher skill respect to the all-season hindcast; and the 51 member of ensemble size is considerably larger than the 11 member size of the hindcast.

The robustness of the results was analyzed in the following aspects. Ensemble size of the real-time forecast was reduced using a subsample of 11 members to assess the third level of verification (Fig. S3) and we found that the spatial patterns observed in the skill maps resulted mostly unchanged, and was therefore discarded as the main reason of the differences against the all-season hindcast verification level. Also, the ability of the ECMWF model to discriminate a cold event—which was assessed using the area under the ROC curve and defining a cold event as a negative temperature anomaly—was also determined defining a cold event as those weeks in which the temperature anomaly is in the lower tercile and in the lower quintile. The differences were small and mostly only observed in the target week hindcast verification, for example a higher ROC area was obtained for central and northern Argentina in week 1 (not shown). Finally, a different dataset was used to verify the ECMWF model hindcast/forecast, the ERA5 reanalysis (Copernicus Climate Change Service 2017). We found that our results still stand, as only some specific regions show slight differences in the magnitude of the scores but they present overall the same pattern (e.g., Figs. S4 and S5).

As was previously noted, ECMWF ensemble mean forecasts could predict the cold anomaly of the week of 17–23 July 2017 one week in advance and the warm anomaly of the week of 24–30 July 2017 two weeks in advance. To our knowledge it has not been studied whether the ECMWF model forecasts are more skillful for warm anomalies than for cold anomalies within the same season (in this particular case, austral winter) and over South America. A recent study by Lavaysse et al. (2019) has analyzed the predictability of heat waves

during summer and cold waves during winter over Europe for the ECMWF model and found overall higher predictive skill for cold waves. However, the dynamics of the atmospheric circulation are also important for case studies. For example, Vitart and Robertson (2018) have shown that the cold anomalies in the Northern Hemisphere during March 2013 were more predictable for those CFSv2 model (Saha et al. 2014) members which predicted accurately the phase and amplitude of the Madden–Julian oscillation (MJO). In this way, the ability of a model to forecast the alternation of cold and warm weeks might also be linked to its ability to predict transitions between circulation patterns, and future studies should address the link between the skill in the prediction of circulation and temperature anomalies in South America and the skill in predicting the evolution of circulation patterns such as the MJO.

Finally, advancing in the verification of forecasting systems in the subseasonal time scale is key for supporting adequate use of these predictions for identified regions and lead times for which these forecasts show best quality. Calibration techniques and the construction of multimodel ensembles are pathways to improve subseasonal prediction performance in the future.

Acknowledgments. The research was supported by UBACyT20020170100428BA and the CLIMAX Project funded by Belmont Forum/ANR-15-JCL/-0002-01. The WWRP/WCRP S2S project and ECMWF are acknowledged for developing the S2S project database and making available the model hindcast and real-time forecast data used in this study. CASC thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico(CNPq), Process 305206/2019-2, and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Process 2015/50687-8 (CLIMAX Project) for the support received for the development of this research. We also thank the three anonymous reviewers for their comments and suggestions that have improved this manuscript.

REFERENCES

Alvarez, M. S., C. S. Vera, G. N. Kiladis, and B. Liebmann, 2016: Influence of the Madden Julian Oscillation on precipitation and surface air temperature in South America. *Climate Dyn.*, **46**, 245–262, https://doi.org/10.1007/s00382-015-2581-6.

Ardilouze, C., L. Batté, and M. Déqué, 2017: Subseasonal-to-seasonal (S2S) forecasts with CNRM-CM: A case study on the July 2015 West-European heat wave. *Adv. Sci. Res.*, **14**, 115–121, https://doi.org/10.5194/asr-14-115-2017.

Batté, L., C. Ardilouze, and M. Déqué, 2018: Forecasting West African heat waves at subseasonal and seasonal time scales. *Mon. Wea. Rev.*, **146**, 889–907, https://doi.org/10.1175/MWR-D-17-0211.1.

Coelho, C. A., M. A. Firpo, and F. M. de Andrade, 2018: A verification framework for South American sub-seasonal precipitation predictions. *Meteor. Z.*, **27**, 503–520, https://doi.org/10.1127/metz/2018/0898.

Copernicus Climate Change Service, 2017: ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS), accessed 16 January 2020, https://cds.climate.copernicus.eu/cdsapp#!/home.

Doss-Gollin, J., A. G. Muñoz, S. J. Mason, and M. Pastén, 2018: Heavy rainfall in Paraguay during the 2015/16 austral summer: Causes and subseasonal-to-seasonal predictive skill. *J. Climate*, **31**, 6669–6685, https://doi.org/10.1175/JCLI-D-17-0805.1.

Gubler, S., and Coauthors, 2020: Assessment of ECMWF SEAS5 seasonal forecast performance over South America. *Wea. Forecasting*, **35**, 561–584, https://doi.org/10.1175/WAF-D-19-0106.1.

Hirata, F. E., and A. M. Grimm, 2018: Extended-range prediction of South Atlantic convergence zone rainfall with calibrated CFSV2 reforecast. *Climate Dyn.*, **50**, 3699–3710, https://doi.org/10.1007/s00382-017-3836-1.

Hudson, D., A. G. Marshall, O. Alves, G. Young, D. Jones, and A. Watkins, 2016: Forewarned is forearmed: Extended-range forecast guidance of recent extreme heat events in Australia. *Wea. Forecasting*, **31**, 697–711, https://doi.org/10.1175/WAF-D-15-0079.1.

Japan Meteorological Agency, 2013: JRA-55: Japanese 55-year reanalysis, daily 3-hourly and 6-hourly data. National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed 12 October 2018, https://doi.org/10.5065/d6hh6h41.

Kobayashi, S., and Coauthors, 2015: The JRA-55 reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan*, **93**, 5–48, https://doi.org/10.2151/jmsj.2015-001.

Lavaysse, C., G. Naumann, L. Alfieri, P. Salamon, and J. Vogt, 2019: Predictability of the European heat and cold waves. *Climate Dyn.*, **52**, 2481–2495, https://doi.org/10.1007/s00382-018-4273-5.

Madden, R. A., and P. R. Julian, 1994: Observations of the 40–50-day tropical oscillation: A review. *Mon. Wea. Rev.*, **122**, 814–837, https://doi.org/10.1175/1520-0493(1994)122<0814:OOTDTO>2.0.CO;2.

Mason, S. J., and 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, https://doi.org/10.1256/003590002320603584.

Osman, M., and M. S. Alvarez, 2018: Subseasonal prediction of the heat wave of December 2013 in southern South America by the POAMA and BCC-CPS models. *Climate Dyn.*, **50**, 67–81, https://doi.org/10.1007/s00382-017-3582-4.

Saha, S., and Coauthors, 2014: The NCEP climate forecast system version 2. *J. Climate*, **27**, 2185–2208, http://doi.org/10.1175/JCLI-D-12-00823.1.

Vitart, F., and F. Molteni, 2010: Simulation of the Madden–Julian Oscillation and its teleconnections in the ECMWF forecast system. *Quart. J. Roy. Meteor. Soc.*, **136**, 842–855, https://doi.org/10.1002/qj.623.

——, and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate Atmos. Sci.*, **1**, 3, https://doi.org/10.1038/s41612-018-0013-0.

——, and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

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

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, https://doi.org/10.1175/1520-0493(2004)132<1917:aarmmi>2.0.co;2.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

Wilks, D., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. **100**, Academic Press, 704 pp.