Verification of Two Years of CNR-ISAC Subseasonal Forecasts

DANIELE MASTRANGELO AND PIERO MALGUZZI

CNR-ISAC, Bologna, Italy

(Manuscript received 1 June 2018, in final form 21 January 2019)

ABSTRACT

The monthly forecasting system of the Institute of Atmospheric Sciences and Climate of the National Research Council (CNR-ISAC) of Italy is operationally run on a weekly basis in the framework of the Subseasonal-to-Seasonal (S2S) project to produce 41-member ensemble forecasts. The first two years of forecasts, covering 106 weeks from April 2015, are verified against ERA-Interim as weekly averages starting from the first forecast day. Nonprobabilistic scores of 500-hPa geopotential height and 850-hPa temperature anomalies are computed for the extratropical hemispheres. The anomaly correlation coefficient shows enhanced predictive skill during the cold months, when favorable values are occasionally obtained beyond week 2. The root-mean-square error saturates toward the climatological value between weeks 2 and 3. Reliability diagrams are used to evaluate the probabilistic forecast skill of 2-m temperature over Northern Hemisphere extratropical land points, in terms of above- and below-normal events. The forecasting system loses reliability and resolution beyond week 2, but well reproduces the observed 2-yr mean frequency up to week 4, proving to be unbiased. The reliability of the forecasting system systematically outperforms that obtained by persisting the previous week forecast. Beyond week 2, the forecast distribution of below-normal events shows low confidence. However, a reliability diagram based on equally populated bins of forecast probabilities highlights residual resolution up to week 4 at low probabilities. ROC diagrams confirm that the modeling system has greater discrimination capability for below-normal events. The reliability analysis of accumulated precipitation shows minor differences between below- and above-normal events, with lower skill than 2-m temperature.

1. Introduction

Subseasonal prediction usually refers to forecasts with a lead time beyond two weeks and less than one season that share theoretical and practical aspects with both medium-range weather and seasonal predictions (e.g., Vitart 2004; Hudson et al. 2011). The predictability issue underlying subseasonal forecasts stands between the initial and boundary condition predictability problems, or predictability of the first and the second kind as originally defined by Lorenz (1975). Improved model formulation and resolution, ensemble strategies, data observation and assimilation techniques, and the increase of observations and computing capacity and availability, have allowed tackling the intricacies associated with the two predictability kinds with growing success in the last decades (Bauer et al. 2015). These improvements suggest that there is predictive signal potentially available at all atmospheric space–time scales (Hoskins 2013), fostering investigations on the feasibility of subseasonal forecasts (Brunet et al. 2010).

The initial state of the atmosphere and underlying boundaries contributes to the predictability of an anomalous future state, with the atmospheric initial conditions affecting predictability beyond the medium range of the integration (Reichler and Roads 2003). Atmospheric phenomena whose life cycle length is comparable to the subseasonal scale are a privileged source of predictability: in this context, the Madden–Julian oscillation (MJO) is considered of primary importance (Waliser 2011). The MJO affects the atmospheric intraseasonal variability in the tropics, where it originates, and propagates its influence on the medium latitudes through teleconnections (Cassou 2008; Lin et al. 2009), affecting some major modes of internal variability and their predictability, such as the North Atlantic Oscillation (Vitart and Molteni 2010). Predictive skill also benefits from the state of climatic circulation modes such as El Niño–Southern Oscillation (ENSO; White et al. 2014). Moreover, the interaction of the anomalies of the initial stratospheric condition with the troposphere (e.g., Thompson et al. 2002)

Corresponding author: Daniele Mastrangelo, d.mastrangelo@isac.cnr.it

DOI: 10.1175/WAF-D-18-0091.1

has been shown to enhance the predictability in the subseasonal scale (Tripathi et al. 2015). Although more relevant on longer time scales, forcing from the initial state of land and ocean influences subseasonal predictability (e.g., Chen et al. 2010). For example, numerical studies have recently highlighted the role of initialization of frozen and liquid soil moisture and snow cover over particular regions and periods (e.g., Thomas et al. 2016).

The possibility to obtain predictive skill and the increasing interests in applications of predictions on the subseasonal scale (White et al. 2017) have promoted dynamical subseasonal forecasting that, going beyond the first experiments in the 1980s, has become one of the numerical products progressively developed in recent years (e.g., Vitart 2014). The forecasting activities undertaken by a growing number of centers have been recently gathered in, and favored by, the ongoing World Weather Research Programme/World Climate Research Programme initiative on Subseasonal-to-Seasonal (S2S) prediction (Robertson et al. 2015). Since 2015, as one of the main goals of the project (Vitart et al. 2017), 11 operational and research centers contribute reforecast and quasi–real time forecast simulations to the S2S database. A dataset of reforecast simulations is necessary for calibration purposes (Hagedorn et al. 2008) and to evaluate model bias. Moreover, reforecasts allow for a robust evaluation of the model predictive skill, feasible even in cases of limited ensemble members because of the long reference period usually covered (Weigel et al. 2008).

Recent studies addressed the deterministic and probabilistic performance of subseasonal forecasting systems. Hudson et al. (2011) used a 27-yr reforecast dataset to assess the predictive skill of the second fortnight of precipitation and minimum/maximum temperature with the Predictive Ocean Atmosphere Model for Australia. Vitart (2014) compared reforecast datasets from different versions of the European Centre for Medium-Range Weather Forecasts (ECMWF) subseasonal forecasting system, showing the increasing capacity to predict MJO and associated extratropical teleconnections. Li and Robertson (2015) compared reforecasts of three different modeling systems highlighting the difficulties in forecasting weekly averaged precipitation anomalies beyond the first week during boreal summers.

The forecasting system operated at the Institute of Atmospheric Sciences and Climate of the National Research Council (CNR-ISAC) of Italy participates in the S2S project producing a 41-member ensemble, 31-day forecast on a weekly basis. Temperature and precipitation forecasts are also routinely provided to the Italian Civil Protection Agency. The reforecast dataset used to calibrate the operational forecasts is not produced in ensemble mode and does not allow for an assessment of the probabilistic predictive skill of the system. On the contrary, the ensemble forecasts are suitable for such verification. A preliminary evaluation, limited to the 2-m temperature, has recently been presented (Mastrangelo and Malguzzi 2017). In this work, a collection of about 2 years of operational ensemble forecasts is used with the aim of performing an overall evaluation of the CNR-ISAC subseasonal forecasting system through deterministic and probabilistic diagnostics. As argued by Vitart et al. (2012), the number of relevant purposes tackled by forecast verification makes such activity a major goal of the S2S project.

The modeling system and the strategy adopted to produce the analyzed ensemble forecasts are described in section 2. The main results obtained for each of the verification methods are analyzed in section 3. A summary and a conclusive discussion are given in section 4.

2. Modeling system and data

The ensemble forecasts analyzed in this work are obtained through the atmospheric general circulation model GLOBO (Malguzzi et al. 2011). This model, developed at CNR-ISAC, has been used since the first, experimental version of the subseasonal forecasting system operated at the same institute (Mastrangelo et al. 2012). In this section, some features of the model set up and forecasting system design are summarized.

The GLOBO model solves the atmospheric equations on a regular grid with a horizontal spacing of 0.56° × 0.80° latitude/longitude and 54 vertical sigma-hybrid levels. The soil moisture and temperature processes are modeled on seven levels. A slab ocean model parameterizes the sea surface temperature (SST) evolution taking into account the sum of turbulent and radiative surface fluxes and relaxing the initial analyzed anomaly to a prescribed climatological field [see Eq. (1) of Mastrangelo et al. (2012)]. The sea ice cover is kept constant if the initial analyzed anomaly is positive (negative) during the growing (decaying) phase of the sea ice cover climatological seasonal cycle. Otherwise, the sea ice cover is relaxed to the climatological seasonal cycle with a rate of 3% day^{-1}. The climatological fields used in both forecast and reforecast simulations (SST, sea ice cover, deep-soil temperature, and water content) are computed from the ECMWF ERA-Interim reanalyses (Dee et al. 2011) as 5-day means over the 30-yr reference period ending in 2010.

The forecasts verified in this work have been operationally produced on a weekly basis starting on 29 March 2015. Each forecast consists of an ensemble of 41 members obtained with a mixed lagged–perturbed

initialization technique. The initialization data are obtained by interpolating on the GLOBO grid the lead time 0 of the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS; e.g., Zhou et al. 2017) forecasts, freely available in real time. Specifically, up to the forecast issued on 15 January 2017, 10 perturbed members have been initialized every 6 h (0000, 0600, 1200, 1800 UTC) on Sunday by ingesting the GEFS fields for the same initialization hour; the control run has been initialized at 0000 UTC on Monday, which is therefore considered as the lead time 0 of the ensemble forecast. The initialization day was shifted to Wednesday–Thursday starting on 19 January 2017. This change was introduced to line up with the submission day of most of the centers contributing to the S2S database.

ERA-Interim provides the initial conditions for reforecast simulations, obtained with the same GLOBO version and set up used for forecast simulations. Reforecasts cover the 1981–2010 period that constitutes the reference climate used to compute calibrated forecast anomalies and tercile thresholds. A single 31-day reforecast run is initialized on the 73 calendar days ranging from 1 January to 27 December every 5 days. The whole reforecast dataset is therefore made up of 2190 simulations. To compensate for the lack of ensemble members, several runs are combined through a weighted average to estimate, for each meteorological variable, a model climatology function of forecast initialization date, lead time, and position. Specifically, the Gaussian weights $W_i = e^{-(|c-C_i|\Delta)^2}$ assign greater importance to reforecasts initialized on the calendar days C_i closest to the forecast initialization date c. The half-width Δ, which basically controls the number of reforecasts involved in the average, has been set to $\Delta = 22$ days for all meteorological fields but total precipitation for which $\Delta = 10$. This weighting average technique efficiently accomplishes two goals: it defines a model climatology for any forecast initialization day c and acts to filter out the shorter-scale, unpredictable signal (Mastrangelo et al. 2012). The resulting mean climatology is used to perform a bias-reduction calibration of the ensemble forecast anomalies.

The weighted mean climatology described above can be interpreted as the mean value of a mixture distribution (Wilks 2011), a composite probability distribution function (PDF) where the ith component is made up of the 30 reforecasts initialized on calendar day C_i. The weights W_i are therefore the probabilities that the values forming the mixture distribution come from each of the i component distributions. The resulting PDF represents the reference climatological distribution that provides the tercile thresholds used to compute the forecast probability of the associated dichotomous events.

3. Forecast verification

A continuous series of 106 forecasts of 850-hPa temperature (T850), 500-hPa geopotential height (Z500), 2-m temperature (T2m), and total accumulated precipitation are verified against the ERA-Interim on the 1.5° × 1.5° latitude/longitude grid used to store the simulations in the S2S database. These parameters are among the standard products of an extended-range global numerical prediction system recommended for verification by the World Meteorological Organization (2017). T2m and precipitation forecasts are evaluated only over land where reanalyses tend to be better thanks to the higher number of available observations (Dee et al. 2011).

To keep the verifying climate dataset as homogeneous as possible with the modeled one, ERA-Interim data have been collected for the same reforecast calendar days and treated in the same way. The forecasts are mainly evaluated as weekly averages computed from the first day of the forecast range. The averaging on a weekly basis, widely used in the S2S community, provides a further filter to reduce the shorter-scale, noisy signal in the atmospheric fields.

a. *Nonprobabilistic scores*

A first evaluation is performed through two nonprobabilistic scores such as the anomaly correlation coefficient (ACC) and the root-mean-square error (RMSE). These scores have been traditionally adopted for medium-range forecast verification and used since the first extended-range forecast experiments (e.g., Déqué and Royer 1992). Being based on the ensemble mean, ACC and RMSE do not account for the probabilistic information of the forecast distribution.

The ACC is here defined according to its "uncentered" version (Wilks 2011),

$$\mathrm{ACC} = \frac{\sum f' o'}{\sqrt{\sum f'^2 o'^2}}, \tag{1}$$

with the forecast f' and verifying o' anomalies computed from the reforecast and reanalysis climatology, respectively. Parameter f' is therefore the calibrated forecast anomaly. If the sums in Eq. (1) are performed over the grid points of a given domain, then ACC measures the correlation between the anomaly patterns on that domain.

The time series of ACC and RMSE of Z500, evaluated on the two extratropical hemispheres, are shown in Fig. 1. In terms of ACC most of the predictive skill is lost between the second and third week, as clearly shown by the mean values of the time series. In the two last forecast weeks, the mean ACC is low, but some skillful cases

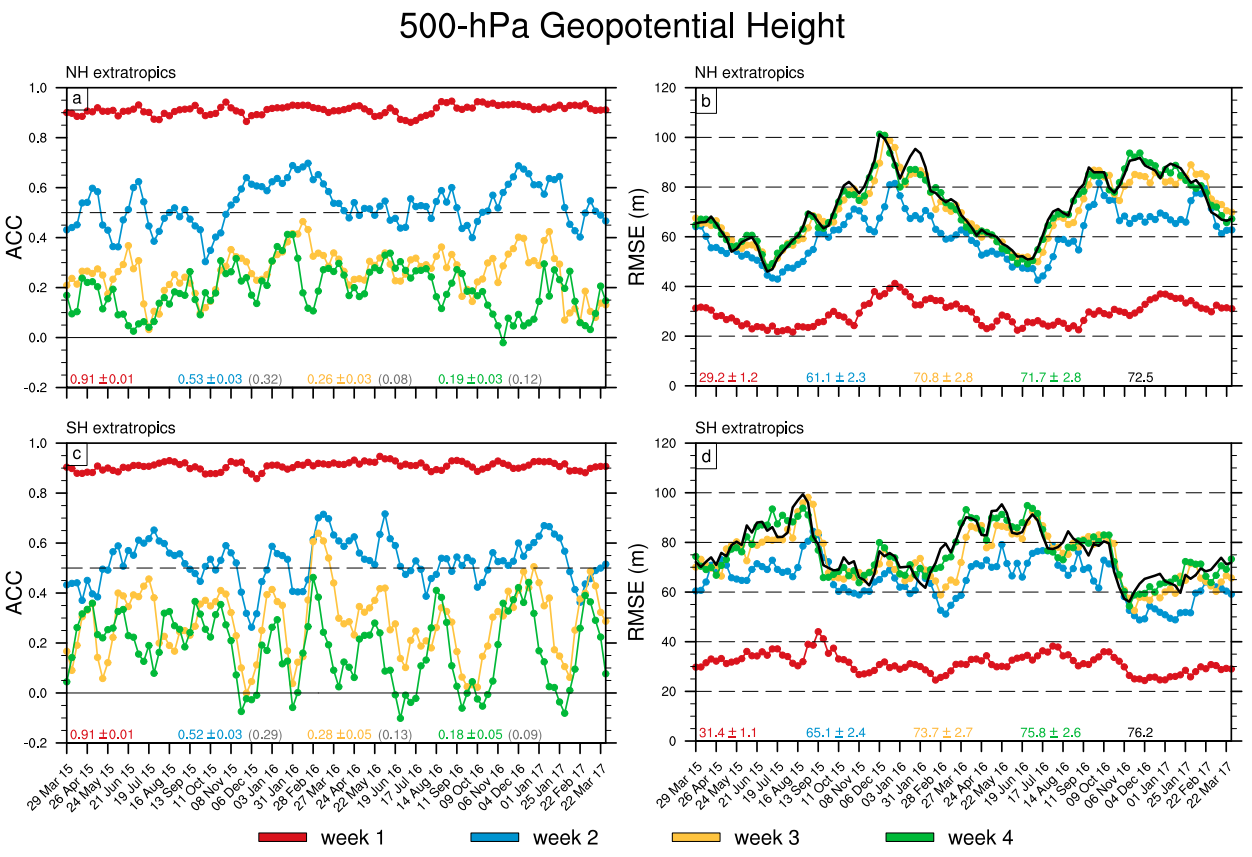

FIG. 1. (a),(c) Anomaly correlation coefficient and (b),(d) RMSE of weekly means of Z500 averaged over the (top) Northern (>20°N) and (bottom) Southern (<20°S) Hemisphere extratropics. The black time series in (b) and (d) represents the climatological RMSE (a four-point moving average is applied to all time series). For each curve, time mean and 95% confidence intervals, computed with a 10 000 bootstrap resampling procedure, are reported at the bottom. The climatological forecast ACC values, obtained by persisting the reanalysis anomaly of the week preceding the verifying week, are reported in parentheses.

are still observed. As already found with other subseasonal modeling systems (e.g., Lin et al. 2016), the highest ACC is obtained during the cold season of both the hemispheres. This seasonal cycle seems more detectable on the Northern Hemisphere (NH) and up to the third forecast week. The time-mean ACC values show a statistically significant skill reduction between the third and fourth week, larger in the Southern Hemisphere (SH). Figure 1a also suggests that three main spells account for most of the fourth week loss of skill in the NH.

The RMSE also shows a seasonal cycle more evident in the NH: the error is higher during the cold season following the climatological trend represented by the black time series. Although not always statistically significant, the mean RMSE of the NH is systematically slightly lower than SH, consistent with the respective climatological values. The RMSE for week 3 is comparable to week 4 and, in turn, to climate, suggesting that errors approach saturation from the third week. The asymptotic behavior of the error growth in the extended range indicates that the forecasting system is correctly calibrated.

Figure 2 shows the same scores but for T850. This field has been masked in areas with elevation greater than 1500 m. The masking reduces possible inconsistencies of T850 caused by the different orography of the GLOBO and ECMWF model, used to create ERA-Interim. The predictive skill, evaluated as ACC, is systematically lower than Z500 except for weeks 3 and 4 in the NH. The less skillful forecasts are obtained in the SH, where some negative values of ACC occur in week 4. On the same hemisphere, the RMSE matches the climatological value from the second week. The seasonal cycle is less evident over the SH both in terms of ACC and RMSE.

The time-mean ACC values are, in all cases, higher than the ACC of the climatological forecast, obtained by persisting the reanalysis anomaly averaged over the week ending the day before the forecast initialization day. These results indicate that the model is overall closer to the verifying anomalies than climatological

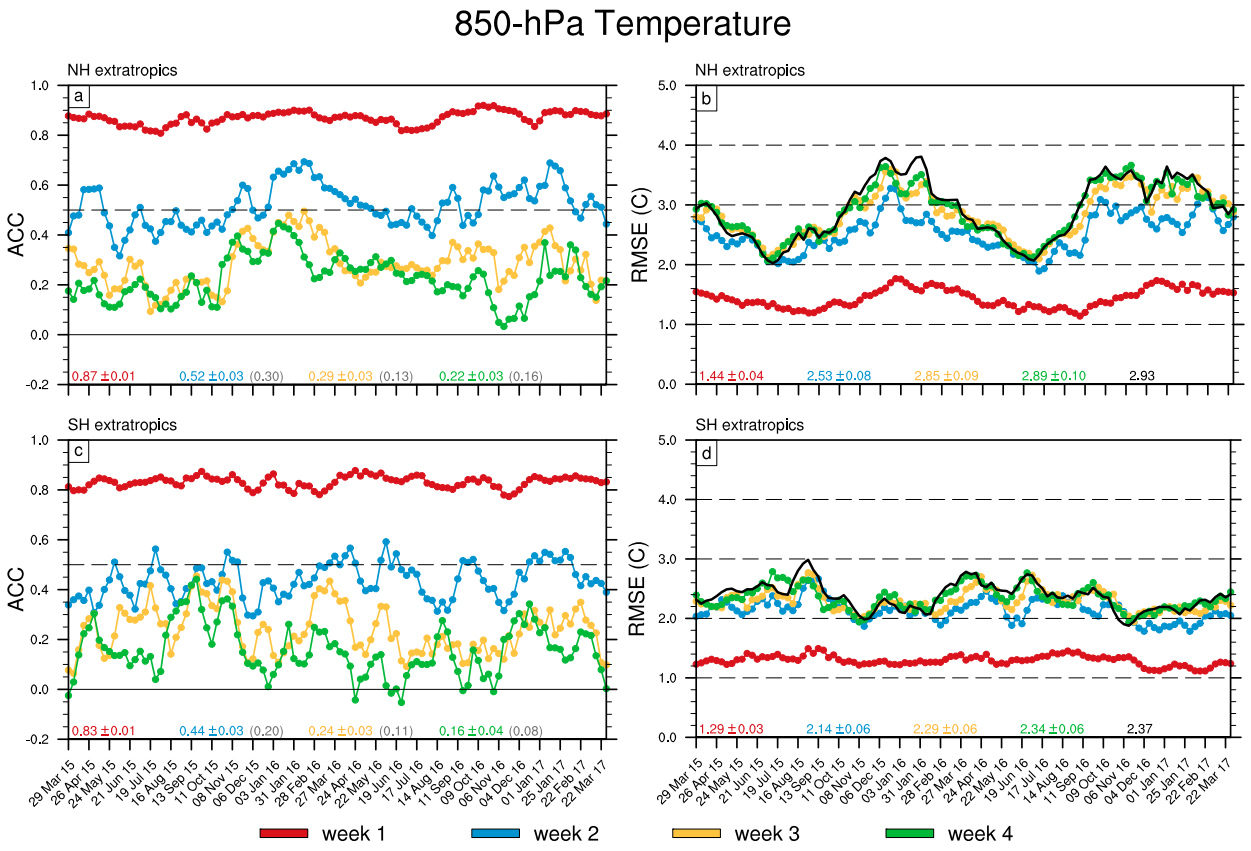

FIG. 2. As in Fig. 1, but for T850.

forecasts. However, in practical terms, some useful information on the forecast pattern beyond week 2 is obtained only in some cases.

To get the spatial distribution of the correlation between the anomaly patterns, a different score is computed by performing the sums in Eq. (1) on the 106 forecast cases. Being summed over forecast initialization dates, this score is usually named the temporal anomaly correlation. Figure 3 shows the temporal anomaly correlation of weekly averaged Z500. Several areas are affected by a continuous reduction of the predictive skill. In particular, the low-skill regions over the extratropical zones of the Pacific and SH oceans reach the lowest values in week 4. A greater variability of the skill pattern for weeks 1–3 is instead evident over North Atlantic and Eurasian regions. For instance, the high skill (0.7) obtained in the second week over Siberia is mostly lost in the following forecast week. Over the northern portion of the North Atlantic the skill is low already in week 2 and decreases subsequently, suggesting a relevant loss of predictive skill on the European side of the Atlantic storm track region. Areas with relatively high values in weeks 3 and 4 are evident over the equatorial belt, as expected, northeastern Asia, Australia, part of the southern Indian Ocean, and North America mainly east of the Rocky Mountains.

The temporal anomaly correlation of weekly averaged T850 is shown in Fig. 4. The areas with low values appearing from week 2 partially match the Z500 maps but, overall, the skill is lower. Some marginal skill in weeks 3 and 4 is obtained again over North America and northeastern Asia. Also, the western North Atlantic shows a region of higher values emanating from the tropical latitudes. The surface processes partially influence the predictive skill of this low-tropospheric field. This is mainly evident in the tropical areas, in particular over most of the tropical oceans with, for example, the prominent signature of the El Niño forcing over the eastern Pacific. In this context, the relatively low skill observed over different oceanic areas and Western Africa even in the first week indicate the occurrence of a mismatch with the reanalyses not compensated for by calibration.

b. Probabilistic scores

In this section, the probabilistic predictive skill of T2m and precipitation forecasts is assessed. These meteorological parameters, closely related to everyday human life, are at the basis of atmospheric events

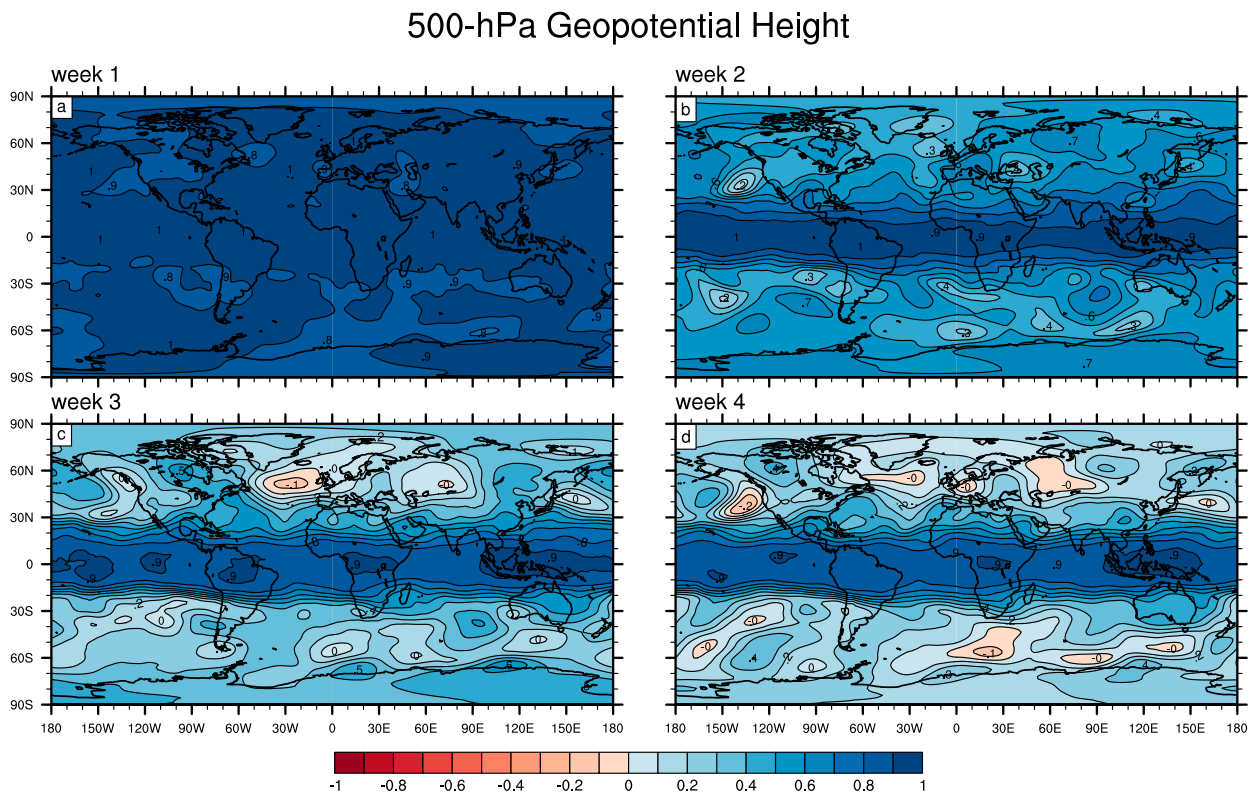

FIG. 3. Temporal anomaly correlation averaged on the 106 forecasts of Z500 weekly means.

relevant in subseasonal forecasting such as heat waves (e.g., Hudson et al. 2016) or dry or wet spells (e.g., Wang et al. 2017).

The evaluation is performed on the forecast probabilities of the dichotomous events defined through the identification of tercile thresholds as, for example, above-normal (warm) events, with T2m in the upper tercile, and below-normal (cold) events, with T2m in the lower tercile. The threshold values are computed from the reforecast distribution to take into account the model systematic error, whereas the observed events are referred to the tercile thresholds computed from the reanalysis distribution—treated consistently with reforecasts as described in section 2. As a result, for each of the four forecast weeks and grid points, a joint distribution of 106 forecast–observation pairs can be drawn for both the warm and cold events.

1) TEMPERATURE AT 2 M

A graphical representation giving the whole information contained in the calibration–refinement factorization of the joint distribution can be obtained through the reliability diagram (Wilks 2011). Figure 5 shows the reliability diagrams of the probabilities of weekly averaged T2m computed over land points of the extratropical NH for the above- and below-normal events. The observed frequencies (y axis) define the calibration distribution of the observations conditioned to the forecast probabilities (x axis) that, in this work, have been subdivided into five equivalent bins of 0.2. For a perfectly reliable forecasting system, each observed frequency would be equal to the conditional forecast probability, and the resulting curve would lie along the diagonal (dashed black line). A reliability curve intercepting the diagonal with a shallower slope indicates, in particular, a lack of resolution, which is the capability to sort the range of observed frequencies for different classes of forecast probabilities. The refinement distribution given by the relative forecast frequency of each probability bin is reported in the histograms of Fig. 5. When the distribution populates the probabilities classes and appreciably differs from the average values, the forecasting system is said to be confident.

The four panels indicate a decrease of reliability with the forecast lead time for both kinds of event, especially from the third week. However, the most reliable forecast probabilities keep close to the mean observed frequency (marks on the left edge of the diagrams and associated dotted lines) up to the fourth week. Indeed, the mean forecast probability (marks on the bottom edge of the diagrams) nearly matches the mean observed frequency indicating that, overall, the forecasting system is not

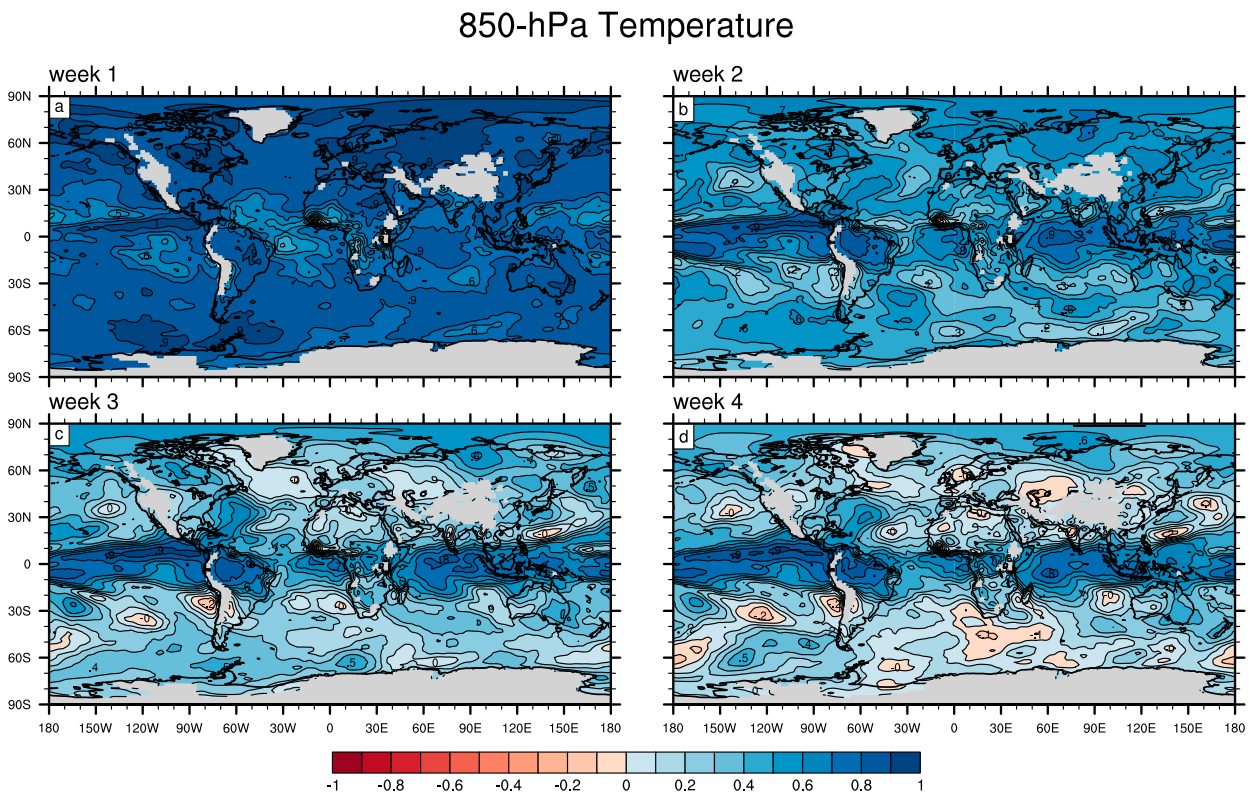

FIG. 4. As in Fig. 3, but for T850. Areas with GLOBO orography > 1500 m are shaded in gray.

biased and reliably predicts the observed mean frequency, especially for the cold-event category. Note that, for this category, the observed frequency differs from the expected climatological value of 1/3, reflecting a climate anomaly of the analyzed period with respect to the reference period (1981–2010). The histograms show a loss of sharpness in the last two weeks, when the relative forecast frequencies peak around the average observed frequencies. This feature, which indicates a loss of confidence of the forecasting system in the extended range, is also associated with a general reduction of resolution, as clearly indicated by the slope of the reliability curves that get closer to the respective no-resolution lines. Nonetheless, both in terms of reliability and resolution, the predictive skill is marginally greater than what obtained by persisting the probabilities of the previous forecast week (dashed curves). Although, as visually evident in Figs. 5c and 5d, the calibration curves are below the zero-skill line (by definition, the line bisecting the sector between the perfect reliability and no-resolution lines; Hsu and Murphy 1986), this persistence test indicates that the modeling system is capable of producing a slightly improved probabilistic forecast up to the last week.

As already stated, in Figs. 5b–d the refinement distributions of cold events indicate low confidence: the lowest bin (0.0–0.2 forecast probability) includes the mean forecast probability and accounts for most of the forecasts. Also, very few forecasts fall in the higher probability bins, making the corresponding portion of the reliability curves rather unstable. Therefore, most of the information crucial for the verification of the cold-event category is compressed in the lowest probability bin.

To cope with this drawback, an adaptive binning based on forecast frequency rather than forecast probability is adopted (as suggested by Bröcker and Smith 2007). For each week and grid point, the 106 forecasts are sorted and distributed in five equally populated bins (the last bin includes 22 forecasts instead of 21), and the conditional observed frequencies are recomputed accordingly. The mean forecast probabilities of each bin constitute the abscissa values in the resulting reliability diagram. Figure 6 shows the results for weeks 3 and 4 and must be compared with Figs. 5c and 5d. The clustering underlying the new reliability evaluation highlights enhanced resolution for very low forecast probabilities, indicating that the forecast of cold events is slightly more confident than what is suggested by the histograms in Fig. 5. The loss of confidence, typically observed in the extended range, could be partially corrected through a logistic regression technique, as shown by Ferrone et al. (2017) in a recent work on a multimodel

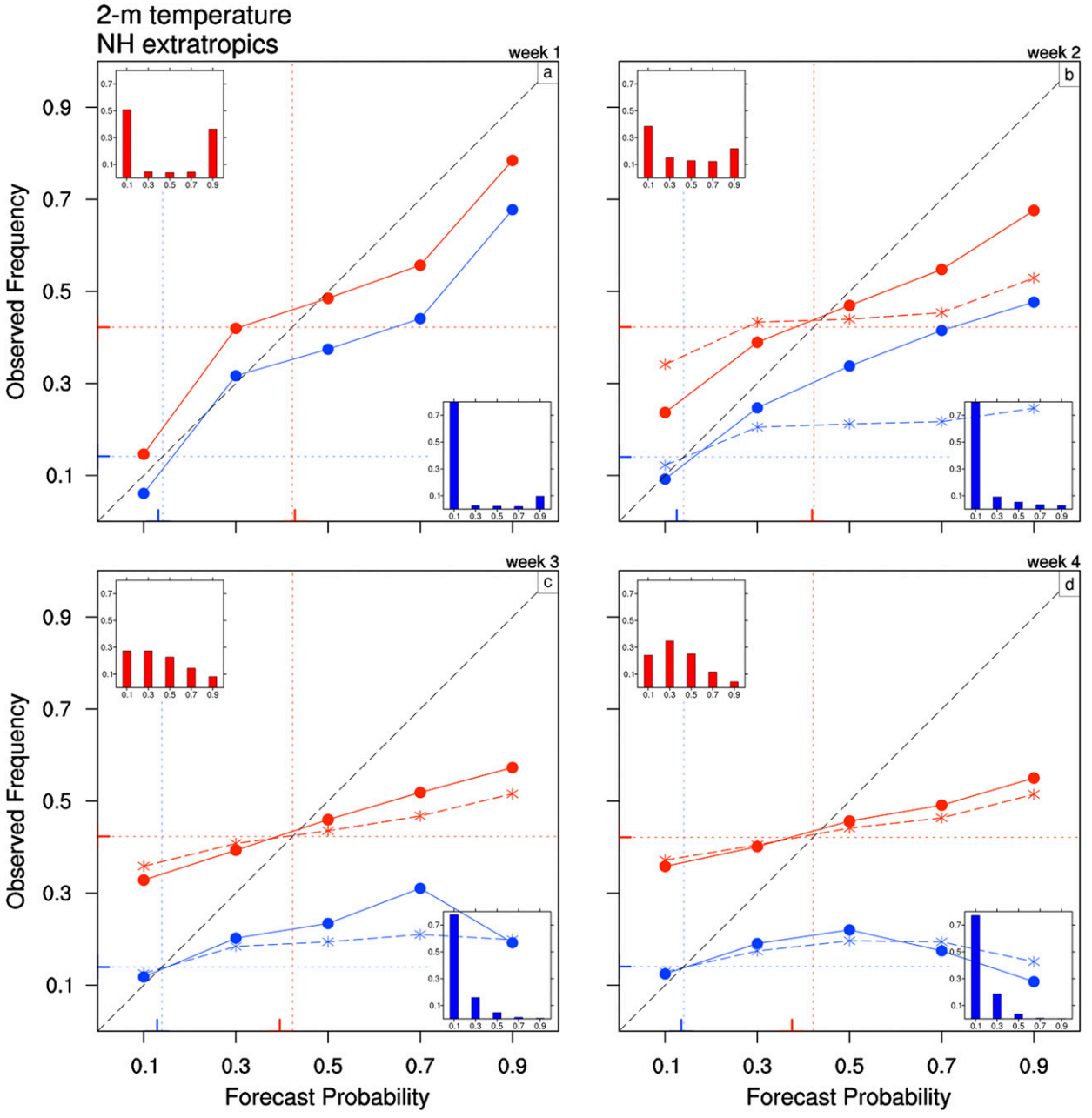

FIG. 5. Reliability diagrams computed from 106 weekly averaged 2-m temperature forecasts, over land points of the extratropical Northern Hemisphere, for (a) week 1, (b) week 2, (c) week 3, and (d) week 4. The red color refers to above-normal (warm) events and the blue color to below-normal (cold) events. The histograms of the relative forecast probability frequencies are shown as insets in the top-left (warm events) and bottom-right (cold events) corner of each panel. The dashed curves in (b)–(d) are obtained by persisting the probabilities for the previous week of the same forecast. Marks on the left and bottom edges indicate the mean observed frequency and forecast frequency, respectively. The dotted lines starting from the mean observed frequency marks (y axis) are drawn from both axes to indicate the no-resolution boundaries.

ensemble involving the CNR-ISAC and ECMWF forecasting systems.

The new clustering has a small impact on the reliability curve of warm-event prediction due to the greater confidence of its refinement distribution (Fig. 5), which implies significant occupation of all probability bins. As a final consideration, showing the actual range of forecast probabilities, this technique avoids instability problems in the statistical evaluation of low populated bins.

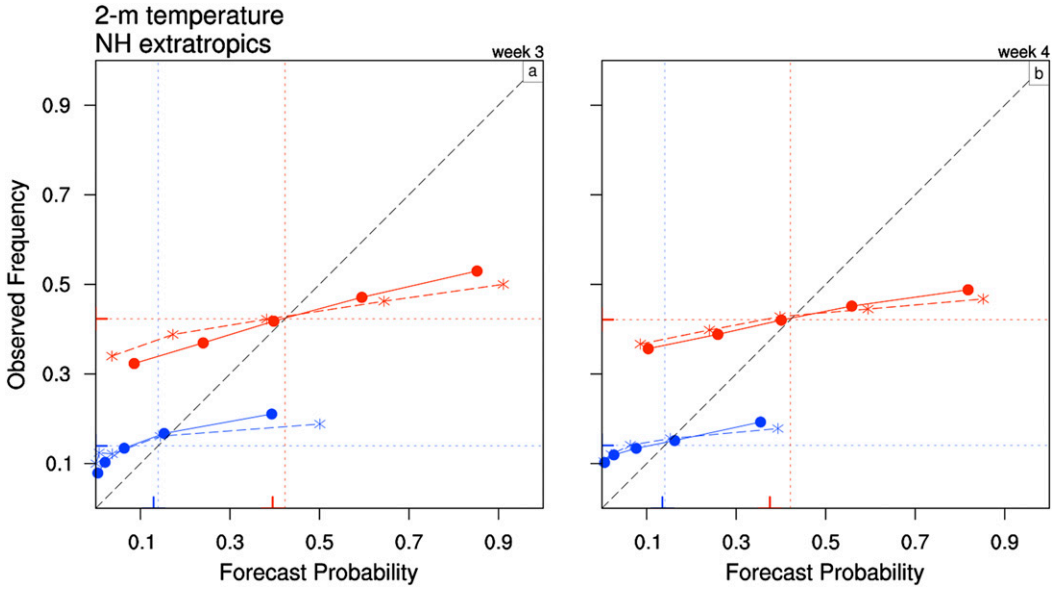

FIG. 6. As in Figs. 5c and 5d, but the curves are computed using equally populated bins.

The resolution of forecast distributions is also related to the discrimination attribute of a forecasting system, which is the ability to correctly forecast different probabilities for the different occurrences of a selected event. Some aspects of this forecast property are graphically evaluated through the relative operating characteristic (ROC) diagram (Wilks 2011), shown here for T2m forecasts. In this diagram, the forecast outcomes conditioned to the occurrence and nonoccurrence of the event are expressed as a hit rate (y axis) and false alarm rate (x axis), respectively. These quantities are computed by making the forecast distribution dichotomous through probability thresholds and then the obtained values are merged to draw the ROC curves. A forecasting system producing random probabilities would yield an equal rate of hits and false alarms, with the ROC curve lying along the diagonal. Therefore, a forecasting system is more skillful when the ROC curve lies closer to the upper-left corner of the diagram, where the hit rate is maximized and false alarm rate is minimized. The ROC score is the area under the curve and is typically used as a quantitative measure of the forecasting skill (Mason and Graham 2002). In this work, the ROC score is computed by summing the area of the trapezoids forming the curve, and the value is normalized so that it ranges from 0, the forecasting skill equivalent to random forecasts, to 1, perfectly discriminating forecasts.

Figure 7 shows the ROC diagrams for the same forecast events of Fig. 5. In both panels, the loss of the forecast skill mainly occurs by the third week, for which the area under the curve is only slightly greater than the fourth week. Similar to the results of Fig. 5, the ROC scores for weeks 2–4 are systematically greater than those obtained by persisting the probabilities of the previous forecast week (values are reported at the bottom of each panel of Fig. 7). The smallest ROC score difference between persistence and actual forecasts, obtained on week 4, has a level of significance of about 93%, according to a Wilcoxon–Mann–Whitney test (Wilks 2011). To apply this test, it has been taken into account that tercile forecast probabilities are not spatially decorrelated. For each grid point, the number of correlated points N_c has been estimated by averaging the one-point correlation matrix. Then, $1/N_c$ are used as weights to sum over all extratropical NH land grid points to obtain the number of independent data.

Although ROC scores are similar for the four weeks, cold-event curves are bounded to the left portion of the diagram and, mainly in the first two weeks, are more aligned toward the y axis: the latter feature indicates that the modeling system better discriminates among the occurrence and nonoccurrence of below-normal events. However, in particular for the last two weeks, most of the discrimination capability is quickly lost as the probability threshold increases. For above-normal events, ROC curves are more symmetric around the negative diagonal, a feature that, as argued by Marzban (2004), is associated with a similar dispersion of the (likelihood) distributions from which false alarm and hit rates are derived. These two quantities have a similar, gradual decrease, and a slightly improved discrimination is obtained for intermediate probability thresholds. In conclusion, the forecasting system shows some

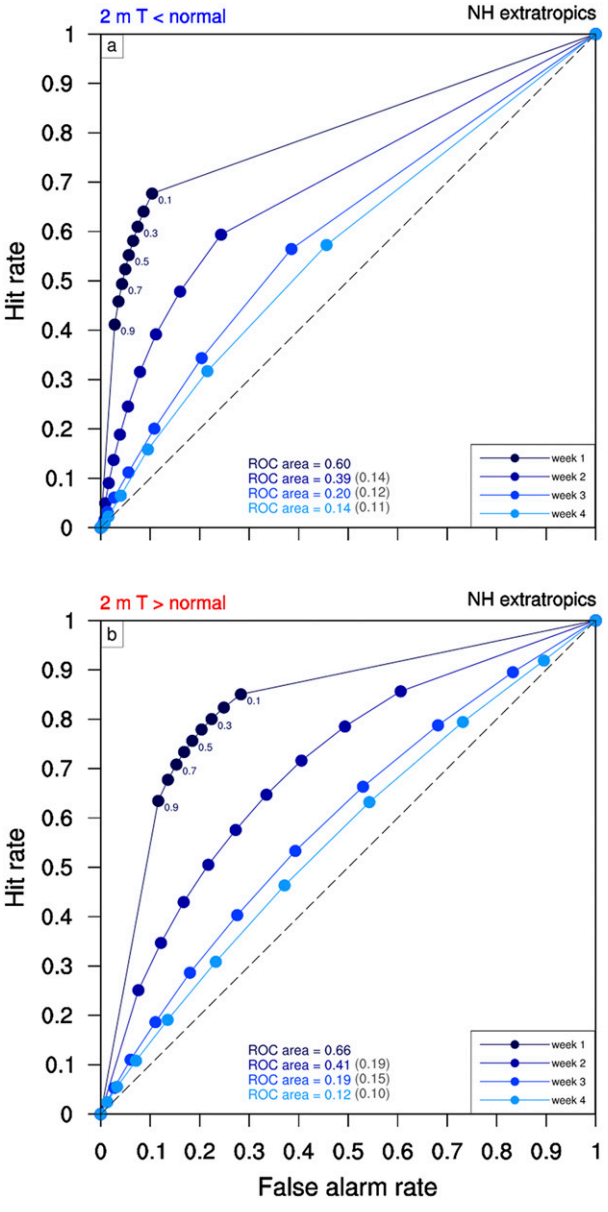

FIG. 7. ROC diagrams computed from 106 weekly averaged forecasts of 2-m temperature over land points of the extratropical Northern Hemisphere, for (a) below-normal and (b) above-normal events. The area under the curve for each forecast week is reported at the bottom of each panel; the same value obtained by persisting the probabilities for the previous week of the same forecast is reported in parentheses for weeks 2–4.

difference in discriminating between the occurrence and nonoccurrence of below- and above-normal events. In particular, the ROC curves indicate that there is a slightly better discrimination in forecasting cold-event cases, mainly for low predicted probabilities. This aspect is associated with the increased resolution visible in Fig. 6 for very low forecast probabilities.

2) PRECIPITATION

The categorical skill of weekly averaged forecasts of above- and below-normal total accumulated precipitation, computed over the land points of the extratropical NH, is analyzed through the reliability diagrams shown in Fig. 8. Similarly to T2m, for both below- and above-normal cases the mean forecast probability is reliably predicted, being very close to the respective mean observed frequency. However, both in terms of reliability and resolution, the predictive skill for the last two weeks is not systematically better than what is obtained by persisting the probabilities of the previous week. Different from T2m, for the whole forecasting range, the modeling system shows similar reliability and resolution properties between the two kinds of events. The excess of false alarms occurring for mid to high probabilities affects the first two weeks, whereas the loss of resolution is the main feature in the last two weeks for the same portion of the curves. Some residual resolution is instead suggested in the lower probability range for below-normal events up to week 4.

The similarity between the predictions of below- and above-normal cases is also evident in the histograms. In particular, in the last two weeks, the refinement distributions are poorly populated in the high-probability bins and narrow around the respective mean observed frequencies. A reliability diagram based on the adaptive frequency binning, as adopted for T2m, is shown in Fig. 9 and compared with Fig. 8d. This diagram confirms that the two distributions are very similar. The adaptive frequency binning shows a general lack of resolution also for the lowest probabilities of both categories, different from what is suggested by Fig. 8d. In conclusion, the loss of resolution affecting precipitation prediction is more severe than for T2m and involves the whole forecast probability range.

4. Summary and conclusions

The CNR-ISAC subseasonal forecasting system routinely provides outputs to the database of the WWRP/WCRP Subseasonal-to-Seasonal (S2S) project. Real-time ensemble forecasts, made up of 41 members, are calibrated through a combination of single-member reforecasts covering the 1981–2010 reference period. In this work, the first two years of forecasts issued on a weekly basis since April 2015 have been verified against ERA-Interim. The evaluation is based on week averages starting from the first forecast day and is performed both in a nonprobabilistic and probabilistic framework.

The 500-hPa geopotential height (Z500; Fig. 1) and 850-hPa temperature (T850; Fig. 2) anomalies are evaluated through the anomaly correlation coefficient (ACC)

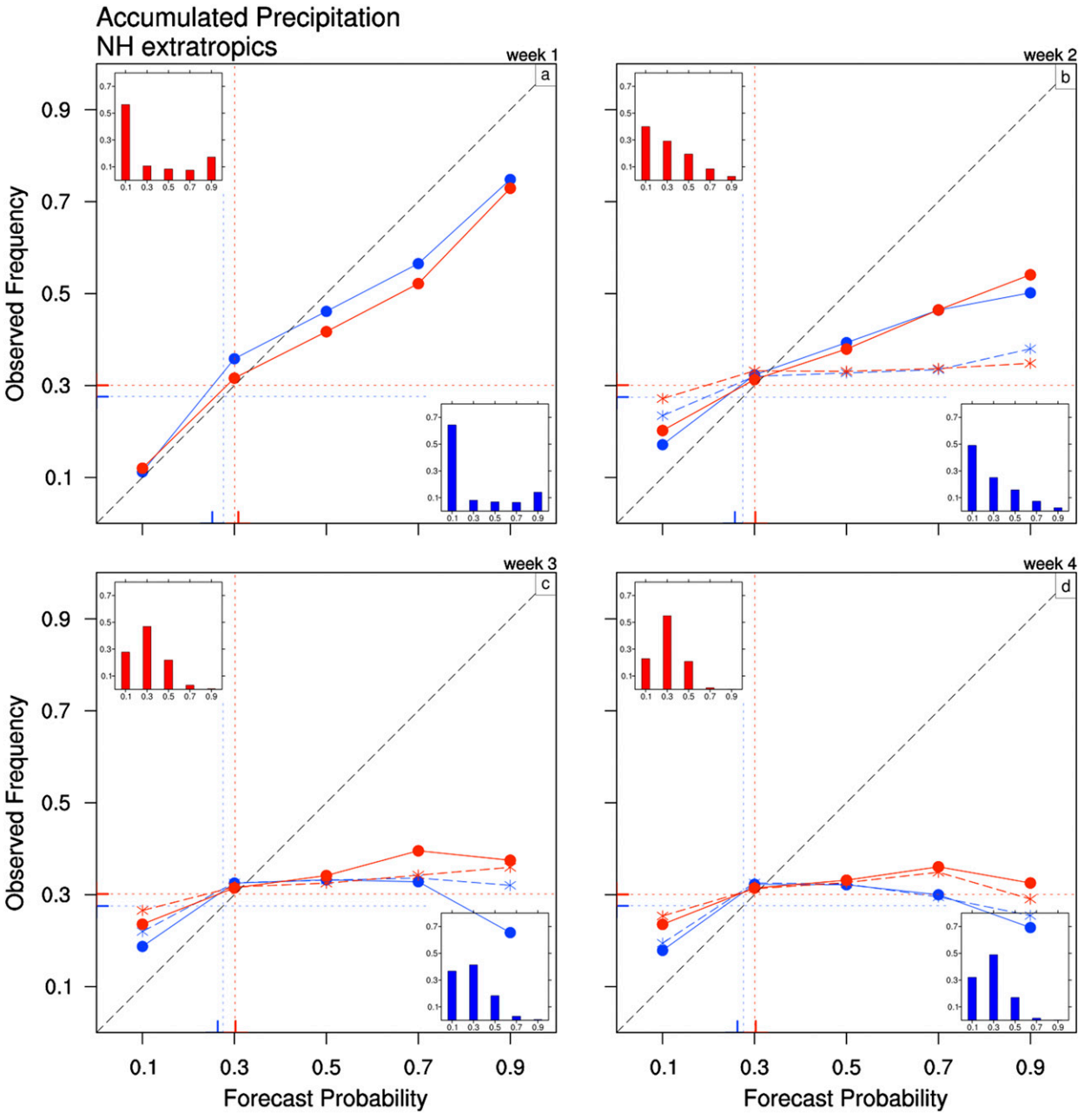

FIG. 8. As in Fig. 5, but for total accumulated precipitation.

and the root-mean-square error (RMSE) to get an overview of the model predictive performance on the global domain. Although the 2-yr areal means of ACC roughly halve between weeks 2 and 3, enhanced predictive skill is observed over both extratropical hemispheres during the cold months, when favorable ACC values are occasionally recorded beyond week 2. The seasonal modulation is more evident for Z500 in the extratropical NH. Similar results are obtained in terms of RMSE, which saturates toward the climatological value between weeks 2 and 3.

The temporal anomaly correlation maps (Figs. 3, 4) show the loss of predictive skill beyond week 2 over most of the extratropical regions and, in particular, over northwestern Europe. However, some areas with enhanced predictability are still detected. The forecast skill over the equatorial belt is higher, as expected, and T850 maps show the signature of the prominent ENSO event that started in late 2015 (e.g., Xue and Kumar 2017; L'Heureux et al. 2017). The possible, positive feedback of ENSO on the extended-range predictive skill of the

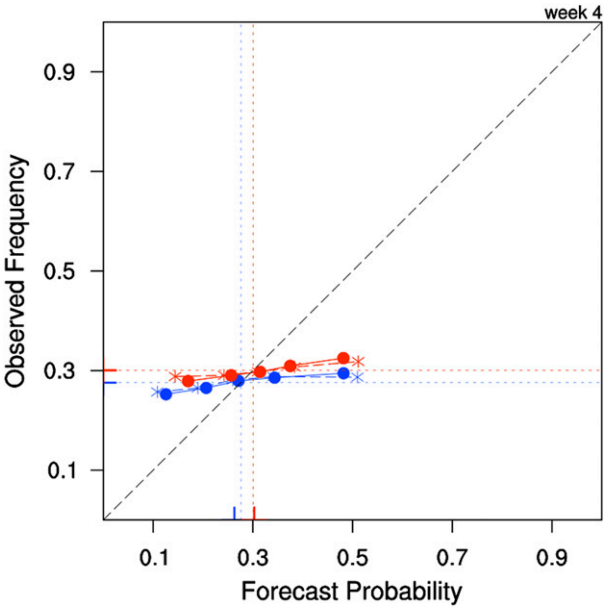

FIG. 9. As in Fig. 8d, but the curves are computed using equally populated bins.

northern extratropics' T850 is also detectable in the ACC time series shown in Fig. 2a.

Reliability diagrams are used to evaluate the probabilistic forecast skill of tercile-based categorical events of 2-m temperature (T2m) and total accumulated precipitation computed over the land grid points of the extratropical NH. A decrease of reliability and resolution is more evident from the third week and affects the forecast of both above- and below-normal cases. In the extended range, the forecast distributions lose sharpness, but the mean forecast probabilities are still close to the mean observed frequencies, in particular for the below-normal category. This appears to be a minimal but still positive result in terms of reliability of the modeling system.

The mean observed frequencies reveal that the 2-yr period here analyzed was characterized by greater (lower) frequency of warm (cold) events than the climatological frequency. The combination of the low value of the observed frequency of the below-normal category, together with the adopted binning interval (0.2), hinders most of the information contained in the reliability curves of cold events. To overcome this issue, a reliability diagram created with equally populated bins of forecast probabilities has been adopted for weeks 3 and 4 (Fig. 6). This diagram reveals some residual resolution for very low forecast probabilities and turns out to be a useful tool to improve the reliability analysis in the case of low confident forecasts.

To evaluate these results in terms of discrimination capability, ROC diagrams have been computed for the same quantities (Fig. 7). ROC scores suggest a very similar performance of the modeling system for both the kind of events and along the whole forecasting range. However, the slope of the ROC curves indicates that the modeling system discriminates between occurrence and nonoccurrence of below-normal cases better than above-normal cases, with the discrimination capability being greater for very low forecast probabilities. Even if less markedly so, this property still occurs in forecast weeks 3 and 4, consistent with the residual resolution identified in the reliability diagram of Fig. 6. The marginal probabilistic skill up to the last forecast week has also been suggested by the comparison of reliability and ROC curves obtained by persisting the probabilities of the previous forecast week.

Reliability diagrams for total accumulated precipitation (Fig. 8) show that, different from T2m, the performance of the forecasting system is similar for the two categorical events. In the extended range, a drop in the forecast skill occurs, with a loss of resolution and confidence. As for T2m, the mean forecast probabilities of the two kinds of events are very close to the respective mean observed frequencies. However, in the last two forecast weeks the predictive skill is not systematically better than persistence.

For the probabilistic verification of precipitation and 2-m temperature, ERA-Interim fields have been used as a proxy of observations. Especially for 2-m temperature, this can raise some concerns since this field may depend more on the soil scheme of the underlying model rather than on observations. Furthermore, observations are subject to uncertainties that affect the analyses themselves. For instance, long-lasting errors in the observed low-level air temperature can result in inconsistencies with the analysis of sea surface temperature, thus producing poor ACC even for week 1, as evident in Fig. 4 for T850. This issue is more relevant for temperature at 2 m, for which negative ACC has been detected in week 1 over some areas of the tropical Pacific Ocean (not shown). Similar considerations apply to the accumulated precipitation analysis.

A longer dataset must be used to broaden the results presented in this work. For instance, as argued by Li and Robertson (2015), the analysis of more years featuring ENSO events could highlight how ENSO's interaction with MJO phases enhances the precipitation predictability in some regions. This is one of the purposes that can be accomplished through the new 30-yr reforecast dataset made up of five ensemble members, recently made available on the S2S database.

Acknowledgments. The authors are grateful to NCEP for providing forecast initialization data (http://nomads.ncep.noaa.gov/pub/data/nccf/com/gens/prod/)

and ECMWF for providing reanalysis data (http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=ml/). Three anonymous reviewers are acknowledged for their comments and suggestions that helped to improve the manuscript. All figures were generated with the NCAR Command Language (version 6.4.0) (UCAR/NCAR/CISL/TDD 2019).

REFERENCES

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, https://doi.org/10.1175/WAF993.1.

Brunet, G., and Coauthors, 2010: Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bull. Amer. Meteor. Soc.*, **91**, 1397–1406, https://doi.org/10.1175/2010BAMS3013.1.

Cassou, C., 2008: Intraseasonal interaction between the Madden–Julian oscillation and the North Atlantic Oscillation. *Nature*, **455**, 523–527, https://doi.org/10.1038/nature07286.

Chen, M., W. Wang, and A. Kumar, 2010: Prediction of monthly-mean temperature: the roles of atmospheric and land initial conditions and sea surface temperature. *J. Climate*, **23**, 717–725, https://doi.org/10.1175/2009JCLI3090.1.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Déqué, M., and J. F. Royer, 1992: The skill of extended-range extratropical winter dynamical forecasts. *J. Climate*, **5**, 1346–1356, https://doi.org/10.1175/1520-0442(1992)005<1346:TSOERE>2.0.CO;2.

Ferrone, A., D. Mastrangelo, and P. Malguzzi, 2017: Multimodel probabilistic prediction of 2 m-temperature anomalies on the monthly timescale. *Adv. Sci. Res.*, **14**, 123–129, https://doi.org/10.5194/asr-14-123-2017.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619, https://doi.org/10.1175/2007MWR2410.1.

Hoskins, B., 2013: The potential for skill across the range of the seamless weather-climate prediction problem: A stimulus for our science. *Quart. J. Roy. Meteor. Soc.*, **139**, 573–584, https://doi.org/10.1002/qj.1991.

Hsu W., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecast.*, **2**, 285–293, https://doi.org/10.1016/0169-2070(86)90048-8.

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

——, 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.

L'Heureux, M. L., and Coauthors, 2017: Observing and predicting the 2015/16 El Niño. *Bull. Amer. Meteor. Soc.*, **98**, 1363–1382, https://doi.org/10.1175/BAMS-D-16-0009.1.

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

Lin, H., G. Brunet, and J. Derome, 2009: An observed connection between the North Atlantic Oscillation and the Madden–Julian oscillation. *J. Climate*, **22**, 364–380, https://doi.org/10.1175/2008JCLI2515.1.

——, N. Gagnon, S. Beauregard, R. Muncaster, M. Markovic, B. Denis, and M. Charron, 2016: GEPS-based monthly prediction at the Canadian Meteorological Centre. *Mon. Wea. Rev.*, **144**, 4867–4883, https://doi.org/10.1175/MWR-D-16-0138.1.

Lorenz, E. N.: 1975: Climate predictability: The physical basis of climate modeling. WMO GARP Publ. Ser. 16, 132–136.

Malguzzi, P., A. Buzzi, and O. Drofa, 2011: The meteorological global model GLOBO at the ISAC-CNR of Italy: Assessment of 1.5 years of experimental use for medium range weather forecasts. *Wea. Forecasting*, **26**, 1045–1055, https://doi.org/10.1175/WAF-D-11-00027.1.

Marzban, C., 2004: The ROC curve and the area under it as performance measures. *Wea. Forecasting*, **19**, 1106–1114, https://doi.org/10.1175/825.1.

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.

Mastrangelo, D., and P. Malguzzi, 2017: CNR-ISAC 2 m temperature monthly forecasts: a first probabilistic evaluation. *Adv. Sci. Res.*, **14**, 85–88, https://doi.org/10.5194/asr-14-85-2017.

——, ——, C. Rendina, O. Drofa, and A. Buzzi, 2012: First outcomes from the CNR-ISAC monthly forecasting system. *Adv. Sci. Res.*, **8**, 77–82, https://doi.org/10.5194/asr-8-77-2012.

Reichler, T. J., and J. O. Roads, 2003: The role of boundary and initial conditions for dynamical seasonal predictability. *Nonlinear Processes Geophys.*, **10**, 211–232, https://doi.org/10.5194/npg-10-211-2003.

Robertson, A. W., A. Kumar, M. Peña, and F. Vitart, 2015: Improving and promoting subseasonal to seasonal prediction. *Bull. Amer. Meteor. Soc.*, **96**, ES49–ES53, https://doi.org/10.1175/BAMS-D-14-00139.1.

Thomas, J. A., A. A. Berg, and W. J. Merryfield, 2016: Influence of snow and soil moisture initialization on sub-seasonal predictability and forecast skill in boreal spring. *Climate Dyn.*, **47**, 49–65, https://doi.org/10.1007/s00382-015-2821-9.

Thompson, D. W. J., M. P. Baldwin, and J. M. Wallace, 2002: Stratospheric connection to Northern Hemisphere wintertime weather: Implications for prediction. *J. Climate*, **15**, 1421–1428, https://doi.org/10.1175/1520-0442(2002)015<1421:SCTNHW>2.0.CO;2.

Tripathi, O. P., A. Charlton-Perez, M. Sigmond, and F. Vitart, 2015: Enhanced long-range forecast skill in boreal winter following stratospheric strong vortex conditions. *Environ. Res. Lett.*, **10**, 104007, https://doi.org/10.1088/1748-9326/10/10/104007.

UCAR/NCAR/CISL/TDD, 2019: NCAR Command Language (version 6.6.2). UCAR/NCAR/CISL/TDD, Boulder, CO, https://doi.org/10.5065/D6WD3XH5.

Vitart, F., 2004: Monthly forecasting at ECMWF. *Mon. Wea. Rev.*, **132**, 2761–2779, https://doi.org/10.1175/MWR2826.1.

——, 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quart. J. Roy. Meteor. Soc.*, **140**, 1889–1899, https://doi.org/10.1002/qj.2256.

——, 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.

——, A. W. Robertson, and D. L. T. Anderson, 2012: Subseasonal to Seasonal Prediction Project: Bridging the gap between weather and climate. *WMO Bull.*, **61**, 23–28.

——, and Coauthors, 2017: The Sub-seasonal to Seasonal Prediction (S2S) Project Database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Waliser, D. E., 2011. Predictability and forecasting. *Intraseasonal Variability of the Atmosphere-Ocean Climate System*. 2nd ed. W. K.-M. Lau and D. E. Waliser, Eds., Springer, 433–468.

Wang, S., A. Anichowski, M. K. Tippett, and A. H. Sobel, 2017: Seasonal noise versus subseasonal signal: Forecasts of California precipitation during the unusual winters of 2015–2016 and 2016–2017. *Geophys. Res. Lett.*, **44**, 9513–9520, https://doi.org/10.1002/2017GL075052.

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.

White, C. J., D. Hudson, and O. Alves, 2014: ENSO, the IOD and intraseasonal prediction of heat extremes across Australia using POAMA-2. *Climate Dyn.*, **43**, 1791–1810, https://doi.org/10.1007/s00382-013-2007-2.

——, and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Met. Apps*, **24**, 315–325, https://doi.org/10.1002/met.1654.

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

World Meteorological Organization, 2017: Manual on the Global Data-processing and Forecasting System: Annex IV to the WMO Technical Regulations. WMO-485, 119 pp., https://library.wmo.int/doc_num.php?explnum_id=4246.

Xue, Y., and A. Kumar, 2017: Evolution of the 2015/16 El Niño and historical perspective since 1979. *Sci. China Earth Sci.*, **60**, 1572–1588, https://doi.org/10.1007/s11430-016-0106-9.

Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the New NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.