Clim Dyn
DOI 10.1007/s00382-015-2757-0

Long-lead ENSO predictability from CMIP5 decadal hindcasts

Paula L. M. Gonzalez[1] · **Lisa Goddard**[1]

Received: 13 February 2015 / Accepted: 4 July 2015

Abstract Using decadal prediction experiments from the WCRP/CMIP5 suite that were initialized every year from 1960-onward, we explore long-lead predictability of ENSO events. Both deterministic and probabilistic skill metrics are used to assess the ability of these decadal prediction systems to reproduce ENSO variability as represented by the NINO3.4 index (EN3.4). Several individual systems as well as the multi-model mean can predict ENSO events 3–4 years in advance, though not for every event during the hindcast period. This long-lead skill is beyond the previously documented predictability limits of initialized prediction systems. As part of the analysis, skill in reproducing the annual cycle of EN3.4, and the annual cycle of its interannual variability is examined. Most of the prediction systems reproduce the seasonal cycle of EN3.4, but are less able to capture the timing and magnitude of the variability. However, for the prediction systems used here, the fidelity of annual cycle characteristics does not appear to be related to the system's ability to predict ENSO events. In addition, the performance of the multi-model ensemble mean is explored and compared to the multi-model mean based solely on the most skillful systems; the latter is found to yield better results for the deterministic metrics. Finally, an analysis of the near-surface temperature and precipitation teleconnections reveals that the ability of the systems to detect ENSO events far in advance could translate into predictive skill over land for several lead years, though with reduced amplitudes compared to observations.

Keywords ENSO predictability · Decadal hindcasts · Deterministic and probabilistic skill · Teleconnections

1 Introduction

The limits for the predictability of the El Niño–Southern Oscillation (ENSO) have been long discussed. Even when ENSO prediction skill is expected to be limited, questions remain as to which are the controlling factors. The role of atmospheric noise for ENSO initiation, the growth of initial errors and inadequate models have been identified as key elements (e.g., Chen et al. 2004; Chen and Cane 2008; Jin et al. 2008; Guilyardi et al. 2009).

Past studies have used retrospective forecasts from seasonal prediction systems to evaluate the predictability of ENSO for up to 24 months (e.g., Chen et al. 2004; Chen and Cane 2008; Ludescher et al. 2014). More recently, Wittenberg et al. (2014) have shown, using a 4000-years control run and a set of reforecasts from the model GFDL CM2.1, that in this setting, free from external forcings, the potential predictability is lost after the 3–4 year range.

For several decades it has been recognised that ensemble prediction is fundamental for an adequate representation of the probabilistic nature of forecast information (e.g., Tippett and Barnston 2008 and references therein). More recently, it has been acknowledged that multi-model ensembles provide a more accurate representation of the forecast uncertainty than single model ensembles, resulting in more skillful prediction systems (e.g., Tippett and Barnston 2008; Kirtman et al. 2014).

✉ Paula L. M. Gonzalez
gonzalez@iri.columbia.edu

Lisa Goddard
goddard@iri.columbia.edu

[1] International Research Institute for Climate and Society, Columbia University, 61 Route 9W, Monell 103, Palisades, NY 10964, USA

Published online: 23 July 2015

Table 1 Description of the subset of CMIP5 decadal hindcasts used in the study

Name	Modeling Center	Ensemble size	Initialization	Starts	Atmospheric component	Oceanic component	References
BCC-CSM1.1	BCC, China	3	Full field	1960:1:2006	BCC AGCM2.1 26 Vertical layers, T42	MOM4 L40 40 Vertical layers, tripolar grid l° × (1–1/3)	http://www.lasg.ac.cn/C20C/UserFiles/File/C20C-xin.pdf
CanCM4	CCCMA, Canada	10	Full field	1960:1:2011	CAM4 26 Vertical layers, 1.25° × 0.9°	POP2 60 Vertical layers, l.11° × (0.27–0.54)°	http://www.cccma.ec.gc.ca/models
EC-Earth i1	Consortium, Europe	5	Full field	1960:1:2005	IFS 62 Vertical layers, T159	NEMO v2 42 Vertical layers, ORCA 1°	Hazeleger et al. (2011)
EC-Earth i3	Consortium, Europe	8	Anomaly	1960:1:2005	IFS 62 Vertical layers, T159	NEMO v2 42 Vertical layers, ORCA 1°	Hazeleger et al. (2011)
GFDL CM2.1	GFDL, USA	10	Full field	1960:1:2011	AM2 2° lat × 2.5° lon	MOM l° × (0.34–1.25)°	Delworth et al. (2006)
HadCM3 i2	Hadley Center, UK	10	Anomaly	1960:1:2009	HadAM3 19 Vertical layers, 3.75° × 2.5°	HadOM3 20 Vertical layers, 1.25° × l.25°	Collins et al. (2011)
HadCM3 i3	Hadley Center, UK	10	Full field	1960:1:2009	HadAM3 19 Vertical layers, 3.75° × 2.5°	HadOM3 20 Vertical layers, 1.25° × l.25°	Collins et al. (2011)
MIROC5	MIROC, Japan	6	Anomaly	1960:1:2010	CCSR/NIES/FRCGC AGCM 40 Vertical layers, T85	COCO v4.5 49 Vertical layers, l.4° × (0.5–1.4)°	Watanabe et al. (2010)
MPI-ESM-LR	MPI-M, Germany	3	Anomaly	1960:1:2010	ECHAM6 47 Vertical layers, T63	MPI-OM 40 Vertical layers, ~1.5° × 1.5°	Raddatz et al. (2007)

More information can be obtained here: http://www.wcrp-climate.org/decadal/cmip5.shtml

The present study makes use of the innovative WCRP/CMIP5 multi-model decadal hindcasts (Taylor et al. 2012) to explore ENSO predictability beyond the 24-months threshold. These hindcasts were initialized every year starting in 1960 and run for 120 months, with varying ensemble sizes, providing a unique opportunity for the assessment of long-lead ENSO predictability in state-of-the-art coupled general circulation models (CGCMs). Additionally, this study objectively addresses the relationship between the ability of the hindcast to reproduce the annual cycle of tropical Pacific sea surface temperature and their skill to represent ENSO activity.

The paper is organized as follows: Sect. 2 describes the data and the methodologies employed in this study and Sect. 3 presents a detailed description of the results. Section 4 summarizes the most relevant discoveries and discusses their implications.

2 Data and methodology

This analysis focuses on the characteristics of the El Niño 3.4 index (hereafter EN3.4), which has been shown to be a good descriptor of ENSO variability (Barnston et al. 1999). It is defined as the average of sea surface temperature anomalies (SSTA) in the box 5°S–5°N, 190°E–240°E. Observations from the Extended Reconstructed Sea Surface Temperature Version 3b (ERSSTv3b, Smith et al. 2008) dataset are considered for the period starting in 1961 until the present (October 2014). In addition, near-surface temperature anomalies from the NOAA CPC CAMS dataset (CAMS, Ropelewski et al. 1985) and precipitation from the WMO/DWD GPCC Version 4 at 2.5 resolution (GPCCv4, Schneider et al. 2010) are used to evaluate the decadal systems' skill to represent the ENSO teleconnections. The three datasets were accessed through the IRI Data Library (Blumenthal et al. 2014).

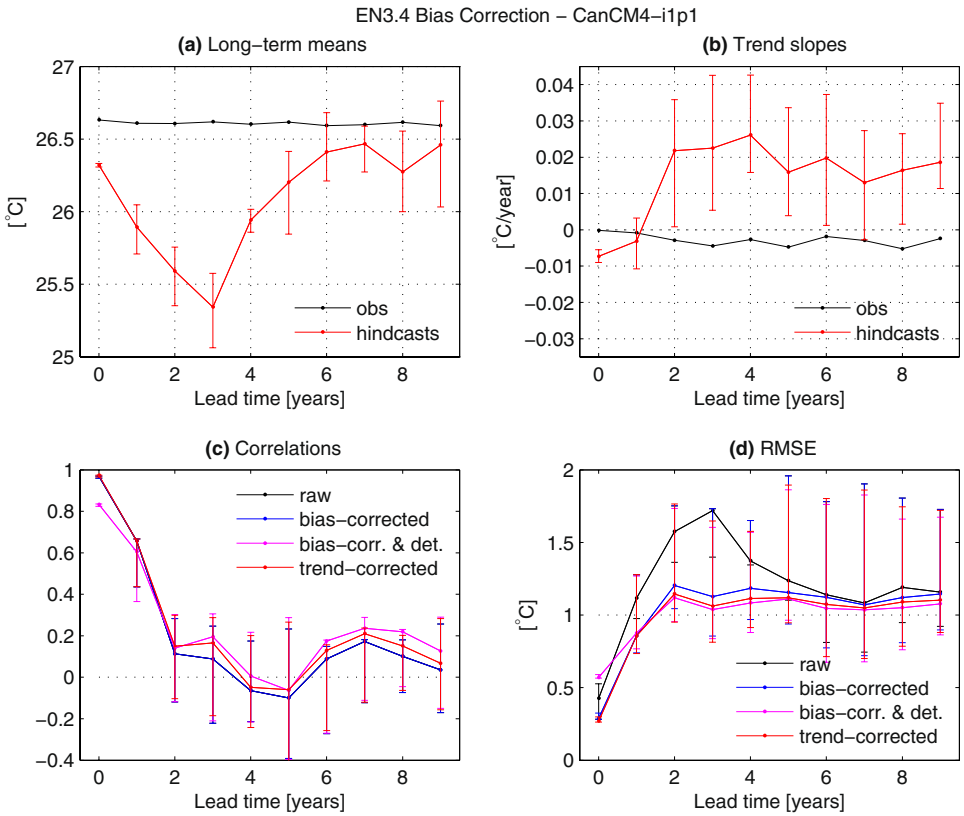

Fig. 1 Comparison between bias correction methodologies applied to the EN3.4 monthly index from the CanCM4 system. Long-term means (*upper left panel*) and trend slope coefficients (*upper right panel*) as a function of lead time, for the observations (*black*) and the hindcasts (*red*). The deterministic skill resulting from three bias correction methodologies is shown for the correlations (*lower left panel*) and the root mean square error (*lower right panel*). In every case, the *vertical bars* indicate the full ensemble spread

2.1 Decadal prediction systems and experimental design

A subset of nine decadal hindcasts from the WCRP/CMIP5 ensemble (Taylor et al. 2009, 2012) from seven different modeling centers (Table 1) are compared with the observations in order to assess the long-lead predictability of ENSO. The decadal predictions were generated with state-of-the-art coupled CGCMs that were initialized every year starting from 1960 to near present, and run for 10 years. Only CMIP5 decadal hindcasts with start times every year are considered in order to avoid sampling errors in the estimation of the variability and trend, as well as in the bias correction, that would be introduced by the standard decadal hindcast design in which predictions were initialized only every 5 years (Meehl et al. 2014). For consistency, only the systems that were initialized in December of each start year and for which the first simulated month was January are considered. The number of start years and ensemble members depend on the specific modeling system and are detailed in Table 1, alongside other specifications. Additionally, two different initialization methodologies—full field and anomaly initialization—are considered in the experimental design (Table 1).

The multi-model ensemble mean (MMM) is constructed as the standard average of the ensemble means of the 9 systems. A multi-model ensemble that included the 'best' models (BMM, see Sect. 3.4) is also analyzed.

2.2 Index definition and bias correction

The monthly raw sea surface temperature (model variable 'tos') from each system is averaged over the EN3.4 box. Unless identified as 'raw', the EN3.4 indices from each of the models are bias-corrected. When compared with monthly EN3.4 SSTA observed time series, some of the hindcasts exhibit unrealistic long-term trends (Fig. 1). This drift depends on the forecast lead time, and this is due to the fact that the standard bias correction methodology (WCRP 2011) does not consider trends in the model drifts (Meehl et al. 2014), resulting in states that remain biased during the forecast period. This suggests that in order to optimize the predictive skill, a time-varying bias correction should be considered (Goddard et al. 2013; van Oldenborgh et al. 2012; Meehl et al. 2014). Therefore, the trend correction methodology proposed by Kharin et al. (2012) is applied. This methodology provides a time-dependent trend adjustment by fitting linear trend coefficients for each forecast lead time. The resulting EN3.4 indices are referred to as trend-corrected.

The trend correction for EN3.4 employed in this study is found to be more skillful than other approaches to

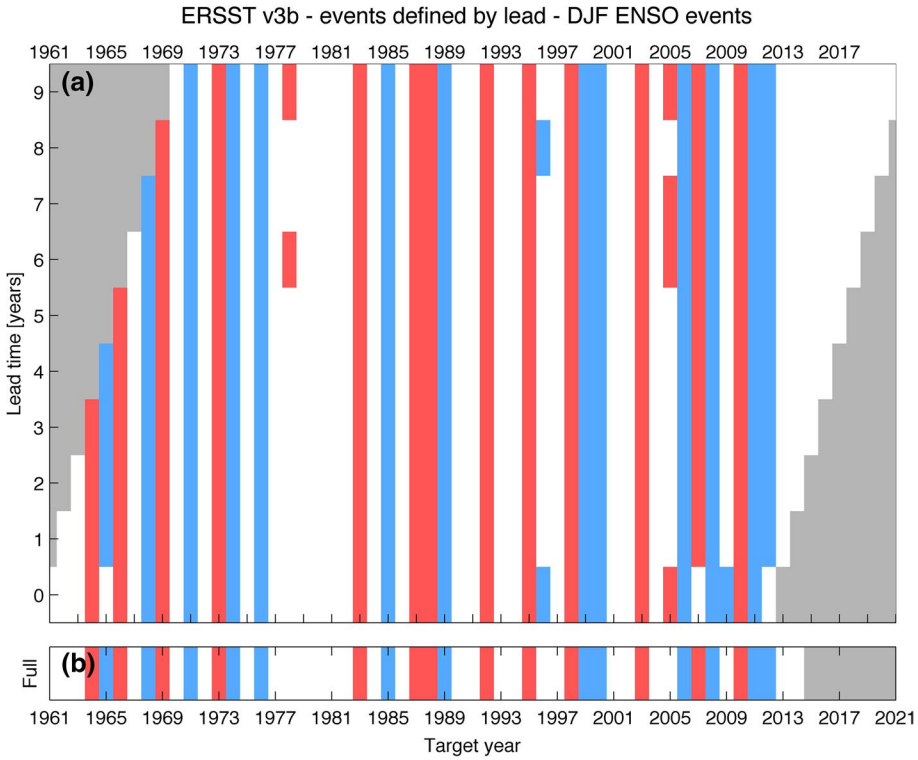

Fig. 2 Effect of the moving climatological period on ENSO event detection. Detection of El Niño events (*red bars*) and La Niña events (*blue bars*) in the hindcast lead year versus target year space, using the ERSSTv3b observed EN3.4 index (*top plot*). The *lower panel* shows the events that are detected using the complete 1861-present time series from the same dataset

bias correction. In addition to the standard methodology (WCRP 2011), it is compared with a detrending procedure applied to each bias-corrected ensemble member and each lead year independently (Fig. 1), as proposed for example in García-Serrano and Doblas-Reyes (2012). The bias corrected SSTA are detrended at each grid point by removing their lineal regression with the global mean near-surface temperature (e.g., Lienert and Doblas-Reyes 2013). A potential caveat of applying this methodology for the study of ENSO predictability lies in the fact that ENSO leaves a significant imprint in the global mean temperature (GMT) time series (Trenberth and Fasullo 2013). Therefore, by removing the regression between the EN3.4 index and GMT, the system is likely to lose some of the ENSO signal.

Due to the fact that the specific climatological reference period for these hindcasts is a function of lead time, some variations can be detected in the long-term mean and trend slopes. Nonetheless, the variations in these parameters observed for the CanCM4 hindcast (Fig. 1) are significantly larger and different to those in the observations. This example illustrates the need for bias and trend corrections of these systems in order to maximize their potential predictive power. Although relatively small, it was found that the trend correction presents improvements against the two other methodologies and the raw hindcasts for lead years 0–3 for every decadal prediction system (not shown).

2.3 ENSO events definition

ENSO events are defined using the EN3.4 time series. Following Coelho and Goddard (2009), El Niño (EN) events are identified when EN3.4 falls in the upper quartile, and La Niña (LN) events when it falls in the lower quartile of variability for a particular prediction system or for the observations. The quartile definition is calculated separately for each lead year's climatology in a cross-validated way. The climatological reference period is determined by the hindcast design (Table 1), and is a function of the lead year. The cross-validation, as in the generation of the bias-corrected anomalies, was calculated based on these moving climatological periods and omitting the year under analysis. For comparison and skill calculations, consistent climatological reference periods are used in the case of the observations. As a result of the changes in the reference climatological period, occasionally an event that would be detected using the full climatology (1961–2014) does not qualify as such for all lead years (Fig. 2).

An example of DJF EN3.4 evolution and event detection is shown for the CanCM4 system (Fig. 3). The events detected from the complete observational period 1961–2014 are shown in the center panel (Fig. 3b) to ease the comparison. However, the skill of each system to represent the events is assessed by comparing the events detected

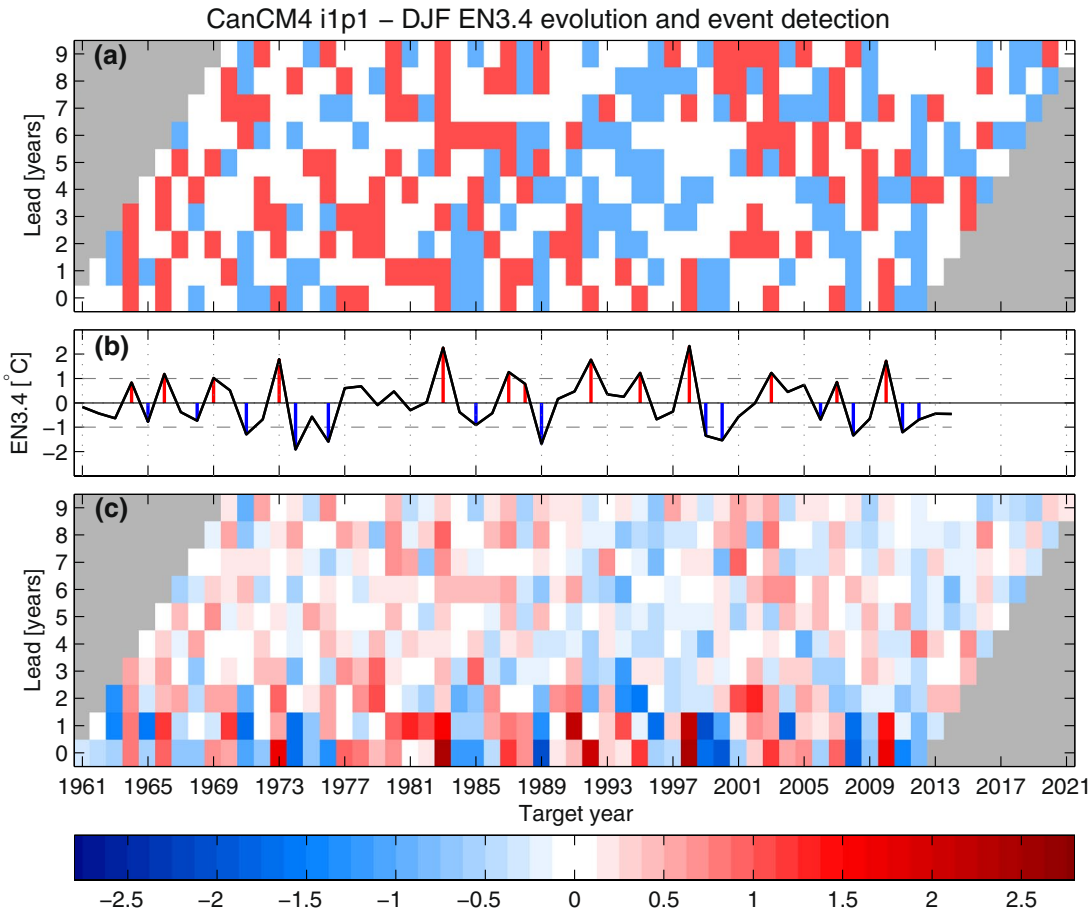

Fig. 3 Evolution and event detection of the DJF EN3.4 index for the CanCM4 decadal system. **a** Binary detection of the EN and LN events in the hindcast system, for each target year as a function of the lead year; **b** evolution of the observed DJF EN3.4 anomalies. *Vertical red* and *blue bars* indicate the years that qualify as El Niño (EN) and La Niña (LN) events, respectively; **c** trend-corrected hindcast anomalies obtained for each target year as a function of the lead year. The *middle panel* presents

with a changing reference period (as in Fig. 3a) against the observed events similarly defined (i.e. lead-year dependent, as in Fig. 2a).

3 Results

3.1 Seasonal cycle

The fidelity of CGCMs to represent the tropical Pacific mean state and its seasonal cycle has been previously linked to their ability to reproduce ENSO variability (e.g., Jin et al. 2008). Other studies have shown that the ENSO skill of CGCMs appears to be independent of their skill to reproduce the basic state or the annual cycle (e.g., Latif et al. 2001; Mechoso et al. 1995). This study objectively addresses the link between fidelity of the annual cycle and the skill of ENSO prediction for this subset of CMIP5 decadal hindcasts. Additionally, since the seasonal cycle of EN3.4 is tied to complex interactions between the components of the climate system (i.e., low-level stratus clouds, dynamics of ocean and atmosphere, etc.), it constitutes an advantageous parameter to test coupled GCMs performance (Jin et al. 2008).

To investigate the skill of the selected hindcasts to represent the activity of EN3.4, the seasonal cycle of the raw index is compared in each case with the observations (Fig. 4). For lead year 0, most hindcasts have a reasonable representation of the seasonal cycle (Fig. 4a), but they are not equivalently good at reproducing the maximum of interannual variability during the boreal winter months (Fig. 4c). Some systems do not represent the exact timing of the maxima and minima in the climatological series, but the overall representation of the seasonal cycle is consistent with that in state-of-the-art coupled GCMs (e.g., Guilyardi et al. 2009). The structure of the seasonal cycle is equally well represented for lead year 2 (Fig. 4b, d), with the exception of the BCC CSM1.1 system. In some

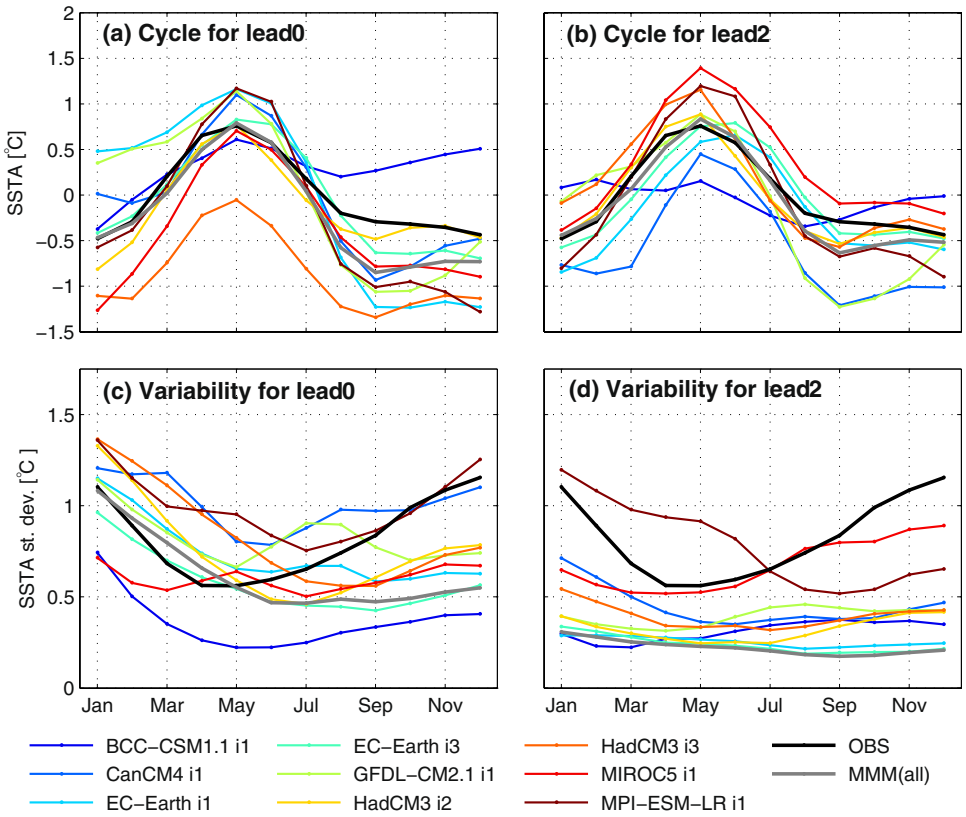

Fig. 4 Skill of the decadal systems to represent the seasonal cycle of the EN3.4 index. **a** Ensemble mean seasonal cycle for lead year 0; **b** same as **a** but for lead year 2; **c** annual cycle of the magnitude of monthly interannual variability for lead year 0; **d** same as **c** but for lead year 2. The multi-model ensemble mean (MMM) is included as a *gray curve*

cases, like the HadCM3 i3 system, the seasonal cycle at the longer lead-time is even closer than that of lead year 0 to the observed one, which can be attributed to the fact that this system used a full field initialization (Table 1). In the case of the year-to-year monthly EN3.4 activity (Fig. 4d), however, most systems show damped variability by lead year 2, which tends to continue decreasing with higher leads (not shown). The diminished quality in the representation of tropical Pacific SST with increased lead-time has been documented for seasonal forecasting systems (e.g., Jin et al. 2008). The problems in the representation of the maxima and the damping of the activity at longer lead times are inherited by the MMM (grey curves in Fig. 4c, d); this is one performance measure that is not improved through multi-model ensembling.

Figure 5 presents different skill metrics for the seasonal cycles of the EN3.4 index (Fig. 4a, c, e) and the magnitude of the EN3.4 interannual variability (Fig. 4b, d, f) as a function of the lead year (shown as colored dots). Most systems show some spread in the amplitude of the seasonal cycle (Fig. 5a) across different lead times; however, this may just be sampling. With the exception of BCC-CSM2.1, the amplitude of the seasonal cycle is better represented for higher lead years (Fig. 5a). The systems that are closer to the observed amplitude are HadCM3, EC-Earth, CanCM4 and MIROC5 (Fig. 5a). In the case of the first two, it is seen that the hindcasts with an anomaly initialization technique (EC-Earth i3, HadCM3 i2) show smaller spreads than their full field initialization counterparts (EC-Earth i1, HadCM3 i3), which is likely related to the initialization shock of the full field initialization technique. The average error in the EN3.4 seasonal cycle remains smallest in EC-Earth i3 and HadCM3 i2 (Fig. 5c) and comparable to the error of the MMM. The correlations between the simulated and observed EN3.4 annual cycles (Fig. 5e), which reflects the ability of the systems to capture the timing of the annual evolution, reveals the HadCM3 i2 system to best represent the observations, with a spread that is even smaller than that of the MMM. The monthly standard deviation of the EN3.4 interannual variability (Fig. 5b) shows that the variability becomes more constant through the year for longer lead times. This is true for every hindcast system as well as for the MMM. The corresponding root mean square error (Fig. 5d) shows that the systems with the smallest errors are MPI-ESM-LR i1 and MIROC5 i1, followed by the GDFL-CM2.1 i1 system. In most systems, the error tends to grow very rapidly from lead years 0–3, in agreement with the observed variability damping (Fig. 4). Finally, the correlations (Fig. 5f) show the largest spreads of all the metrics, with the MIROC5 i1 system displaying the highest overall skill. It can be noted that these correlations do not always decrease with increased lead-time, and in several cases the

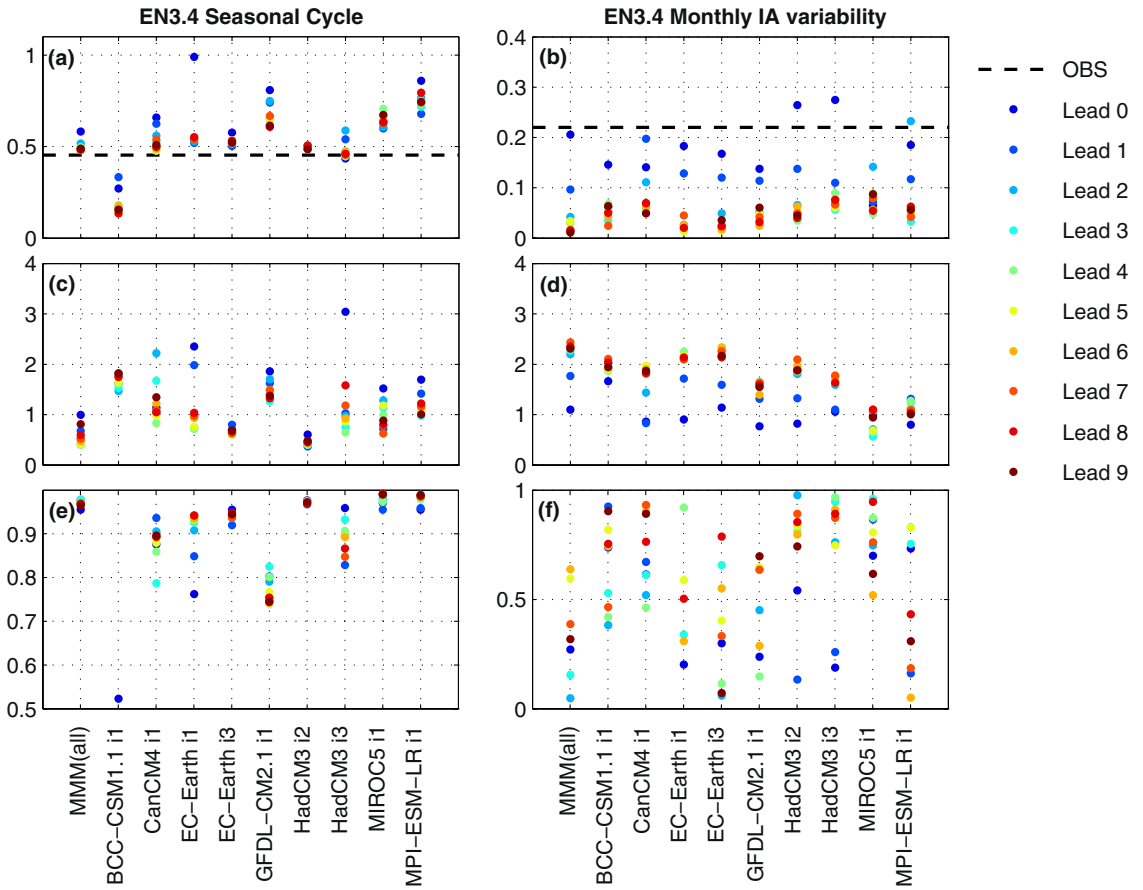

Fig. 5 Skill of the decadal systems to represent the seasonal cycle of the EN3.4 index and its monthly interannual variability as a function of lead year. The panels on the *left column* present: **a** standard deviation; **c** root mean square error and **e** correlation of the hindcasts' seasonal cycle compared to the observed one. The panels on the *right column* present: **b** the standard deviation; **d** the root mean square error; and **e** correlation of the month-by-month interannual variability of EN3.4. In *each plot*, the skill of the decadal systems is compared with that of the multi-model ensemble mean (MMM), which corresponds to the first set of *dots*

opposite might occur. This means that although the amplitude of the cycle decreases (the variability is damped), its timing actually improves. Overall, the problems presented here for the decadal prediction systems to represent EN3.4, its annual cycle and variability are consistent with those previously identified for state-of-the-art models' uninitialized climate change projections (e.g., Guilyardi et al. 2009) and seasonal-to-interannual forecasting systems (e.g., Jin et al. 2008).

3.2 Boreal winter deterministic skill

The December–January–February trimester (DJF) is considered to assess the ENSO prediction skill of the decadal systems, since this is the season when ENSO events typically reach their peak magnitude (Fig. 4). This trimester also exhibits strong global teleconnections (e.g., Ropelewski and Halpert 1987; Trenberth et al. 1998). For these seasonal averages, the lead year 0 value is estimated using the January–February average from the first year of simulation. Therefore, lead year 0 represents forecast leads of *1–2 months*; lead year 1 represents forecast leads of 12–14 months; lead year 2 accounts for forecast leads of 24–26 months; etc. The notation of these events will be such that, the season December 1960–February 1961 is hereafter referred to as DJF 1961; the season December 1961–February 1962 as DJF 1962; etc. (e.g. Figs. 2, 3).

The average deterministic skill of each system as a function of lead year is evaluated using two statistics: anomaly correlation (ACC) and root mean square error (RMSE, Fig. 6). All systems show anomaly correlation values between approximately 0.85 and 0.95 for lead year 0 (Fig. 6a). For lead year 1 (months 12–14) the spread is much larger but all the correlations are statistically significant, in agreement with those shown in Chen and Cane (2008) for seasonal prediction systems for forecasts extending up to 12 months. Subsequently, some systems still show

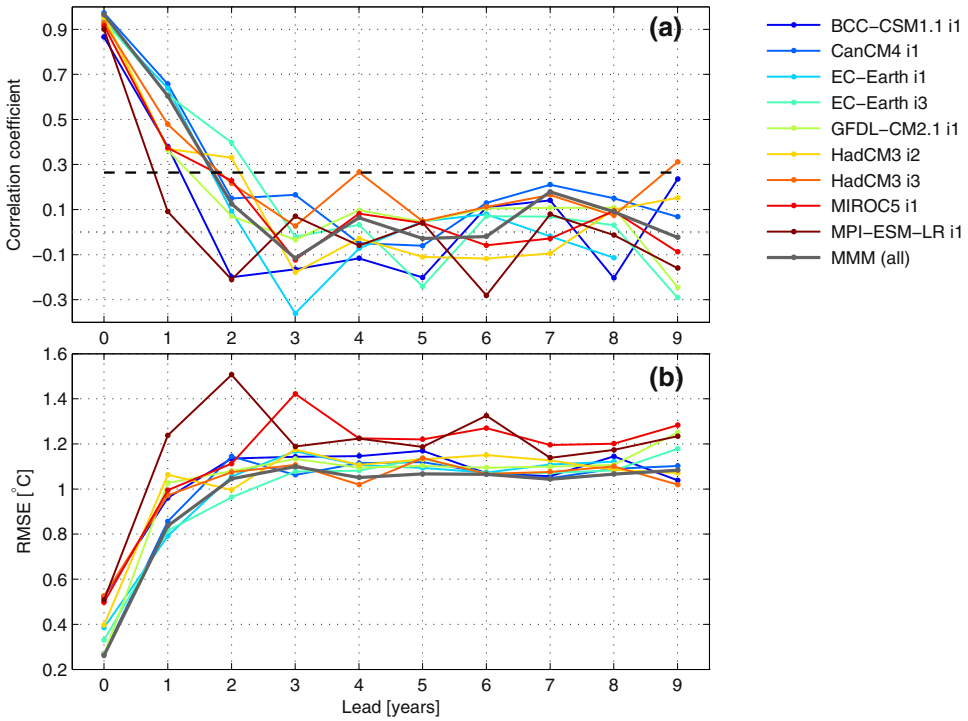

Fig. 6 Deterministic skill of the decadal prediction systems for the DJF EN3.4 index. **a** Anomaly correlations; and **b** the root mean square error, as a function of lead year for each hindcast and the multi-model ensemble mean (MMM, *grey*). The *dashed line* on **a** indicates the 95 % significance threshold through a Student *t* test

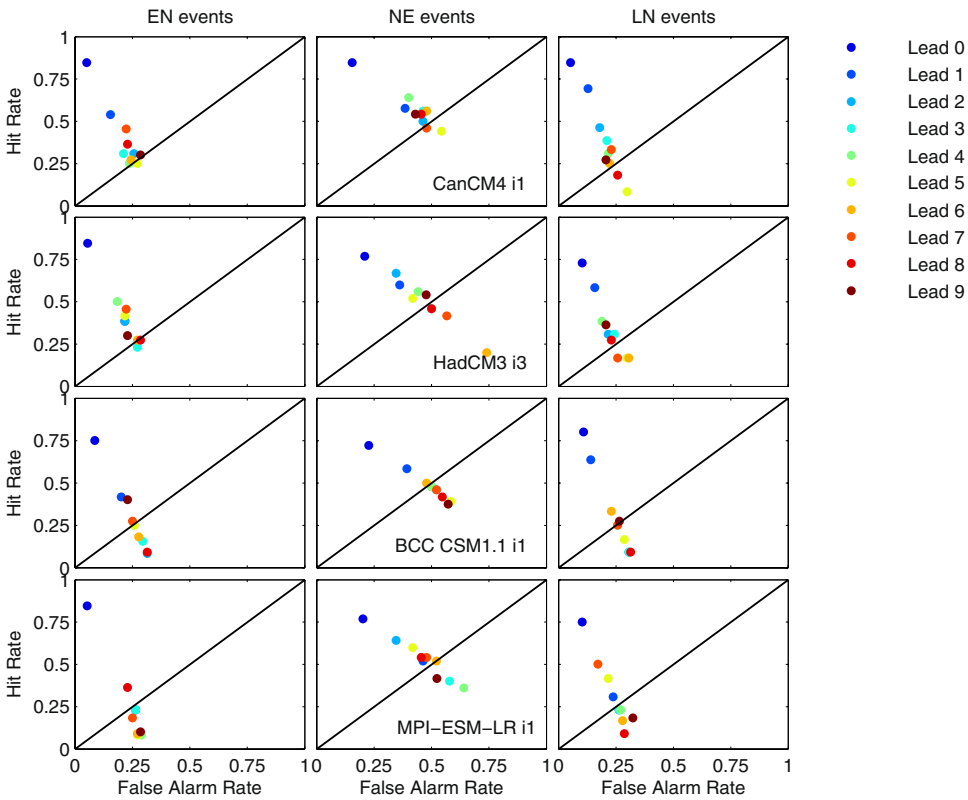

Fig. 7 Skill of the decadal prediction systems' ensemble mean to detect El Niño (EN), neutral (NE) and La Niña (LN) events. *Each row* presents the deterministic relative operating characteristic (ROC) diagrams for a different hindcast system (from *top* to *bottom*: CanCM4, HadCM3 i3, BCC-CSM1.1 i1 and MPI-ESM-LR i1) and for each, the *left panel* corresponds to the EN, the *central panel* to the NE and the *right panel* to the LN events. Lead years are indicated in *different colors*

significant values for lead year 2, which extends beyond 24 months of simulation (EC-Earth i3, HadCM3 i2, followed closely by HadCM3 i3 and MIROC5 i1). No consistent skill is found beyond that lead year. The MMM (grey line) shows improved skill with respect to most individual systems only for lead years 0 and 1.

The RMSE for year lead 0 (Fig. 6b) exhibits values between 0.3 and 0.6 °C, which is in agreement with results

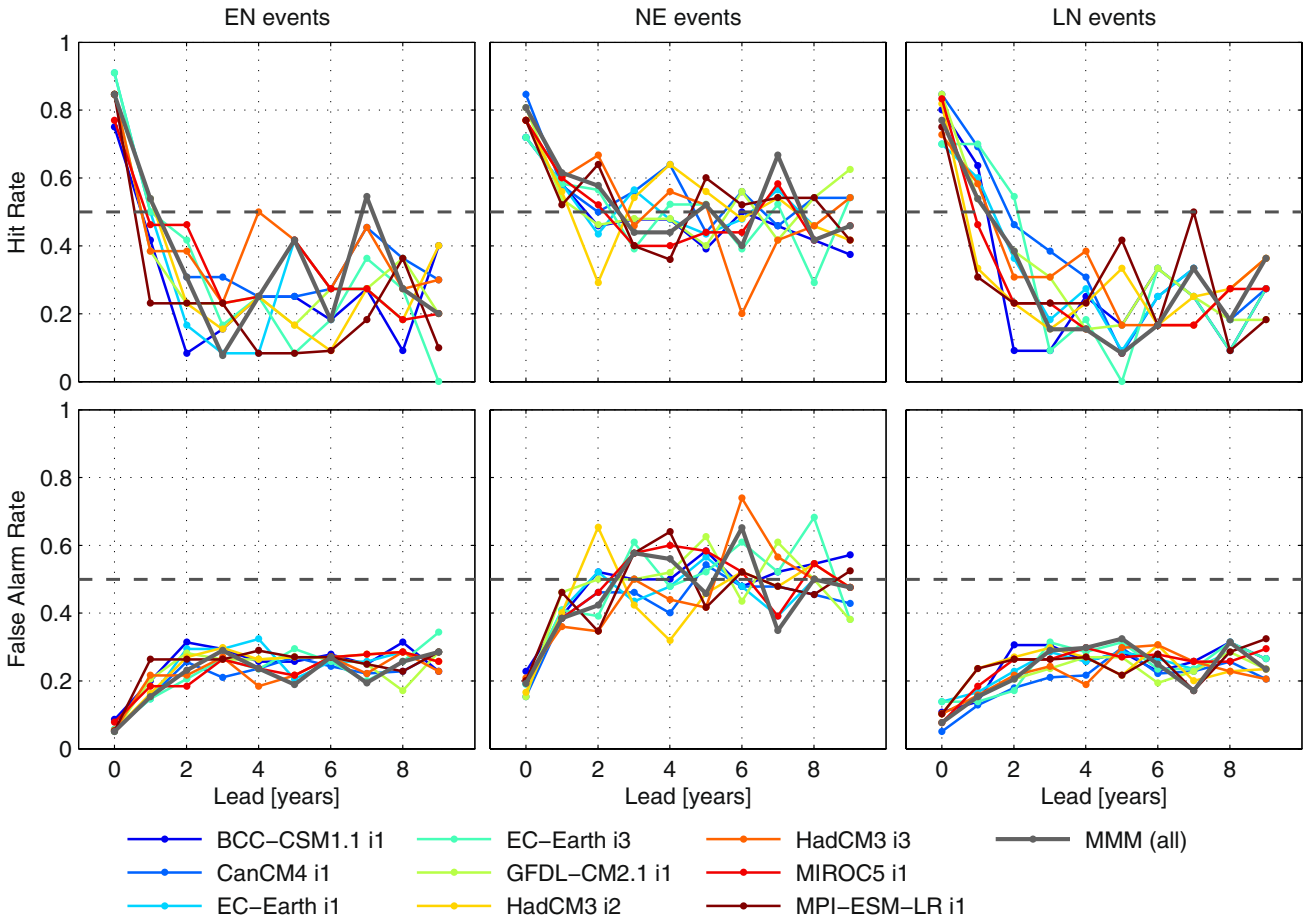

Fig. 8 Deterministic skill of the decadal prediction systems for the detection of EN (*left panel*), NE (*center panel*) and LN (*right panel*) as a function of lead year. In each case, the *top graph* presents the event hit rate and the bottom one the event false alarm rate for each system (*colors*) and for the multi-model ensemble mean (MMM, *grey*)

observed for seasonal forecasting systems for leads of 1–3 months in Jin et al. 2008. With the exception of the MPI-ESM-LR hindcast, which shows very high errors for lead years 1 and 2, and some fluctuations in MIROC5, the error grows consistently for higher leads. After lead year 3, the errors stabilize at a value of around 1 °C, showing less spread than the ACC for higher leads. Unlike the case of the ACC, the MMM is consistently better than most individual systems for all lead years, which may be linked to the loss of variance of the MMM for longer lead times (Fig. 4).

The use of correlation and RMSE to assess the quality of an ENSO prediction system is a necessary first step to put the analysis on common ground with the numerous existing analyses of seasonal prediction systems. However, ENSO events are episodic, and the ability to detect these events, may be even more critical than assessment of a system's general performance over all conditions. To assess and compare objectively the deterministic skill of these systems to detect ENSO events as a function of lead year, the hit rates and false alarm rates were studied along with the relative operating characteristic (ROC) diagrams. For illustrative purposes, and graphical clarity, the ROC diagrams are presented for just four of the decadal systems; two which seem to present high deterministic skill: CanCM4 and HadCM3 i3; and two that show a poorer performance: BCC-CSM1.1 i1 and MPI-ESM-LR i1. In these diagrams, skillful event detection is observed if the dot is located above the diagonal, where the hit rate is higher than the false alarm rate. The cases in which these parameters are equal (values on the diagonal) imply that the system has no skill. Furthermore, those systems for which the false alarm rate is higher than the hit rate (values under the diagonal) are said to have negative skill. The closer the dot is to the upper left corner, the higher its skill to discriminate events. A quick examination of Fig. 7 reveals that CanCM4 and HadCM3 i3 are more skillful than BCC-CSM1.1 i1 and MPI-ESM-LR i1, given that they are skillful for more lead years for the detection of EN, LN and Neutral-ENSO (NE) events. Additionally, the skill to detect EN events (left column) generally tends to be the largest, though closely

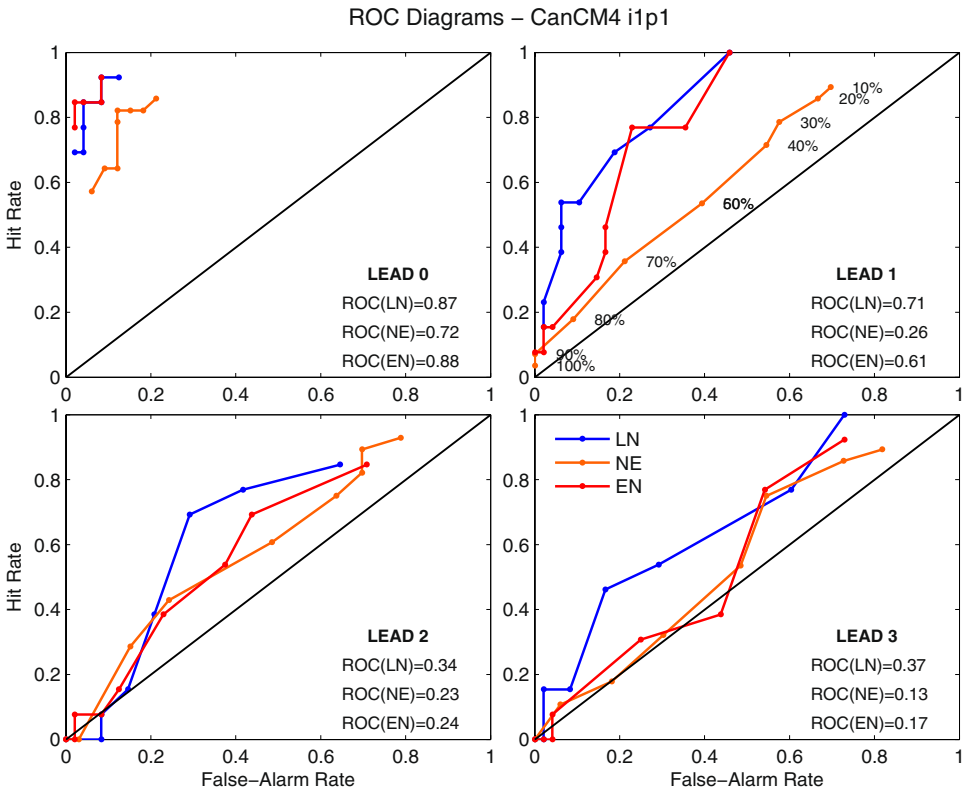

Fig. 9 Probabilistic ROC diagrams for the CanCM4 decadal system. Probabilistic relative operating characteristic (ROC) diagrams are presented for lead years 0 (*top left*), 1 (*top right*), 2 (*bottom left*) and 3 (*bottom right*). The ROC curves are plotted in each case for the EN (*red*), NE (*orange*) and LN (*blue*) events, and the percentages for each point correspond to the forecast probabilities. The ROC scores, defined as the normalized area beneath *each curve* are indicated on the *lower right hand side* of the diagrams

followed by that of LN events (right column). The ability of all systems to simulate NE events (center column) is smaller.

The time evolution of the hit rates (HRs) and false alarm rates (FARs) for each ENSO category (Fig. 8) further elaborates these metrics for the decadal prediction systems in our analysis and for the MMM. The largest HRs are observed for EN events at lead year 0, ranging from 0.75 to 0.9. The HRs for NE and LN events at lead 0 are very similar. For the three ENSO categories, the HRs decrease for higher lead years. Only in the case of NE events (middle diagram) all the HRs for lead year 1 are above 0.5 for every decadal system, whereas the EN and LN cases show a larger dispersion and generally smaller values. This is consistent with the reduced variability at longer leads, such that neutral conditions are more consistently forecast. Nonetheless, the MMM shows HR values above 0.5 for LN, EN and EN events for lead years 0 and 1, and only in the case of NE events for lead year 2. The individual systems with the highest skill according to these parameters are CanCM4, EC-Earth and HadCM3 i3.

The FARs for EN and LN events show much smaller dispersions, and even though they initially increase with higher lead times, the individual systems stabilize at a value in the vicinity of 0.25. In the case of the NE events, the spread is larger for higher lead years and the values don't stabilize oscillating in a range between 0.3 and 0.7. These large variability is also observed for the MMM in lead years higher than 3.

These results are in agreement with what is observed in the case of the CanCM4 system (Fig. 3), one of the models with the higher deterministic skill. For this system, most EN and LN events are detected for lead year 0, and a large number of them also for lead year 1 (lead times higher than 12 months, Fig. 3a). More LN events than EN events (e.g., DJF 1971, DJF 1985, DJF 1989) are correctly detected by this hindcast set for lead year 2 (lead times of more than 24 months). There are, however, a number of false alarms, even for lead year 0 (e.g., DJF 1977, DJF 1984). It is worth noting that some significant events such as the EN in DJF 1983 and the LN in DJF 2012 were captured by CanCM4 for a large number of lead years. This super long-lead skill does not seem to be related to the event strength for this system, however, since some weak events are predicted several years ahead (e.g., DJF 1964) and other strong cases such as the EN in DJF 1998 were only detected for lead years 0 and 1 (Fig. 3c). Even the amplitude of EN3.4, after trend correction, appears to be reasonably predicted for the first 3–4 lead years (Fig. 3c). For the neutral ENSO years (NE), even though the HRs seems to have values that are

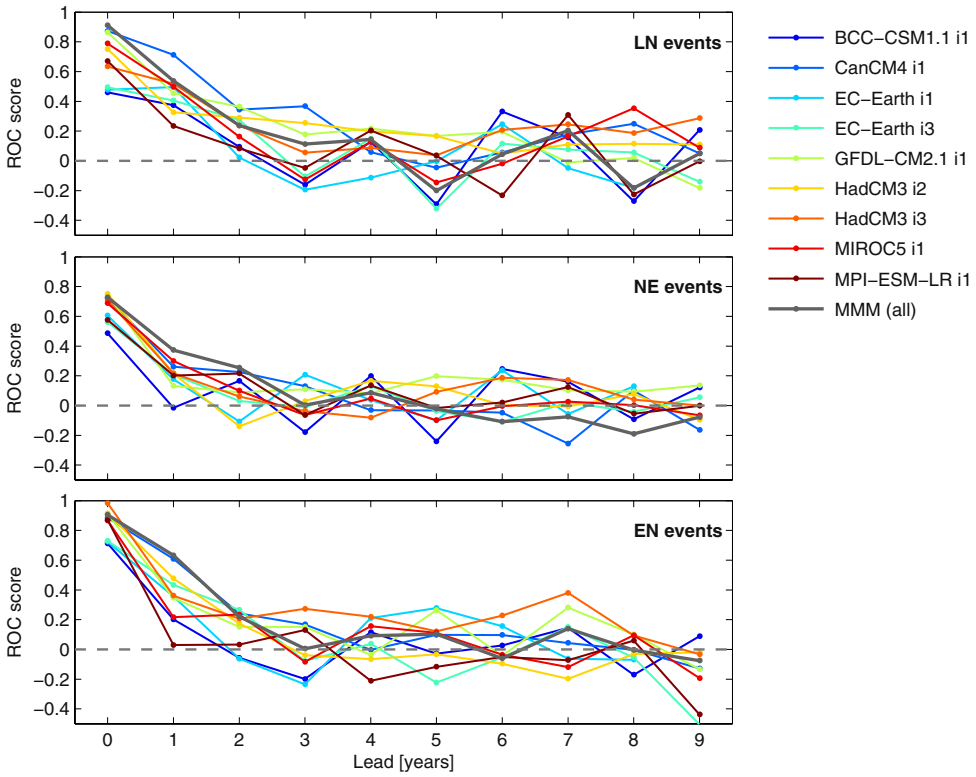

Fig. 10 Probabilistic ROC scores as a function of lead year. Probabilistic ROC scores for LN (*top panel*), NE (*center panel*) and EN (*top panel*) events. Each decadal prediction system is indicated by a *color* and the multi-model ensemble mean (MMM) is included in *grey*

overall higher than those for EN and LN events at long lead times, the FARs also exhibit higher values, meaning that the ability to effectively detect NE is limited.

The analysis of the diagrams corresponding to all systems (not shown) shows that though there is no perfect system –i.e., one that detects each event for all lead years and has no false alarms– some of the systems appear to be able to skillfully detect some events for long leads, beyond the previously documented predictability limits.

3.3 Boreal winter probabilistic skill

Given that these hindcasts are actually ensembles of predictions of varying sizes (Table 1), the probabilistic skill of each individual decadal prediction system and that of the MMM was also assessed. A probabilistic ROC diagram (Fig. 9) is similar to the deterministic ROC diagrams shown before (Fig. 7) except that the HR and FAR values are determined for different levels of forecast probability rather than for the ensemble mean. These specific probabilities are determined by the agreement between ensemble members and are therefore dependent on the system's ensemble size. In the case of CanCM4, which has 10 ensemble members, the probabilities go from 0 to 100 % every 10 % (Fig. 9, upper right quadrant). As in the case of the deterministic ROC diagrams (Fig. 7), skillful systems are those for which the ROC curve stays above the diagonal. In the case of CanCM4 (Fig. 9) this occurs for most forecast probabilities from lead years 0–3. To summarize the information contained in the ROC diagram and allow for an easier comparison between systems, a score can be defined (Mason 1982) using the area under the curve. To create the ROC score, the area is normalized so that a perfect forecast system has a score of 1, a curve lying along the diagonal (no skill) has a score of 0 and the systems that generate bad forecasts result in negative ROC scores.

The ROC scores for CanCM4 (lower right corners of Fig. 9, panels) indicate that the system is skillful for the detection of EN, LN and NE years for lead years 0–2. For lead year 0, the score for EN and LN events are similar, but for the longer leads, the system is better able to capture the LN events. The NE events are harder to detect in every case.

Comparison of the probabilistic ROC scores for all the decadal systems and for the MMM (Fig. 10) reveals that all the systems show skill to predict the occurrence of LN for lead years 0–2, whereas the MMM shows positive scores for lead years up to 4. Additionally, some systems such as CanCM4, HadCM3 i2 and i3, and GFDL-CM2.1 show positive scores for lead years as high as 5. In the case of the NE events, some systems lose skill quickly. Nonetheless, most of them remain skillful up to lead year 2 and the MMM as well as the CanCM4 and GFDL-CM2.1 systems do so until lead year 3. For EN events, even though the spread

Table 2 EN3.4 skill ranking table

Model	BCC-CSM1.1	CanCM4	EC-Earth i1	EC-Earth i3	GFDL-CM2.1	HadCM3 i2	HadCM3 i3	MIROC5	MPI-ESM-LR
Properties of seasonal cycle									
Lowest RMSE of cycle				2nd		1st		3rd	
Lowest RMSE for amplitude				3rd		1st	2nd		
Lowest RMSE of Ivar		3rd						1st	2nd
Seasonal cycle points	0	1	0	3	0	6	2	4	2
Deterministic EN3.4 index skill									
Deterministic ENSO skil									
Highest ACC		1st		2nd			3rd		
Lowest RMSE		2nd	3rd	1st					
EN									
Highest HR		1st		2nd				3rd	
Lowest FAR		1st		2nd				3rd	
NE									
Highest HR		2nd	3rd				1st		
Lowest FAR		2nd	3rd				1st		
LN									
Highest HR		1st		3rd	2nd				
Lowest FAR		1st			2nd		3rd		
Probabilistic ENSO skill									
EN									
Highest ROC score		1st			3rd		2nd		
NE									
Highest ROC score		1st			2nd			3rd	
LN									
Highest ROC score		1st			2nd	3rd			
Enso skill points	0	30	3	10	9	1	10	3	0
Total points	0	31	3	13	9	7	12	7	2

The 1st, 2nd and 3rd labels indicate which decadal system showed the highest, 2nd highest and 3rd highest skills for lead years 0–3

of the ROC scores across systems is relatively larger, these hindcasts remain skillful for the 2-year horizon, adding the HadCM3 i3 system. A comparison between the three panels confirms that on average, the CanCM4 (Fig. 9) results hold true for the other prediction systems, with higher scores for LN events than for EN events, and the smallest scores for NE events for the first lead years. Additionally, it is seen that even though some skillful systems might stand out, the MMM has a higher ability to detect ENSO events than most systems. Overall, the decadal hindcast ensemble shows skill to detect the events for lead years 0–2 for individual systems, and the MMM results skillful up to lead year 4.

3.4 Identification of the most skilful systems ('best' models ensemble)

A subset of decadal systems is selected based on their overall EN3.4 skill to assess the impact of vetting the models included in the MMM on the long-lead ENSO prediction. The hypothesis being that more skillful individual models should lead to a more skillful multi-model ensemble, which would improve the seasonal-to-decadal ENSO prediction skill as proposed, for instance, by Guilyardi et al. 2009. Tippett and Barnston (2008) showed that a better performance for ENSO prediction was observed when the multi-model ensembles were selected based on the skill of the

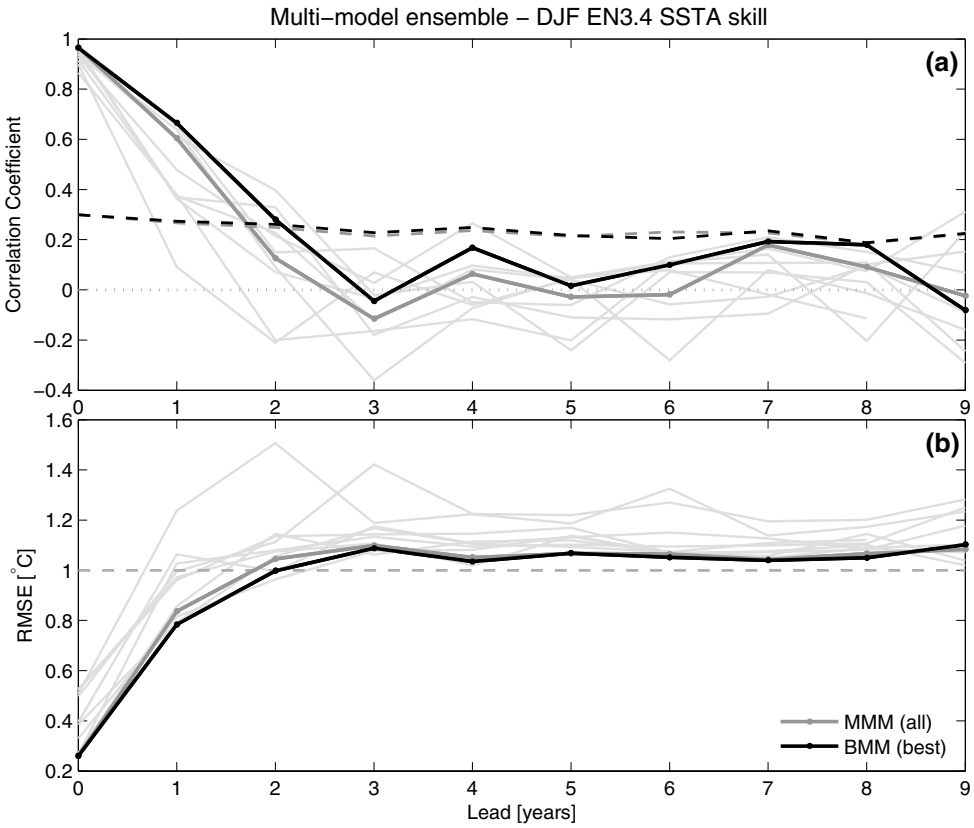

Fig. 11 EN3.4 multi-model ensemble mean deterministic skill. Comparison of the EN3.4 index deterministic skill, as measured by **a** anomaly correlation, and **b** root mean squared error (RMSE), as a function of lead year between the multi-model ensemble mean considering all systems (MMM, *thick grey line*) and the one considering the 'best' systems as defined on the text (BMM, *thick black curve*). The *thin grey lines* represent the ensemble mean skill of each individual system. The *dashed lines* on panel (**a**) indicate the 95 % significance threshold through a Student *t* test corrected for serial autocorrelation

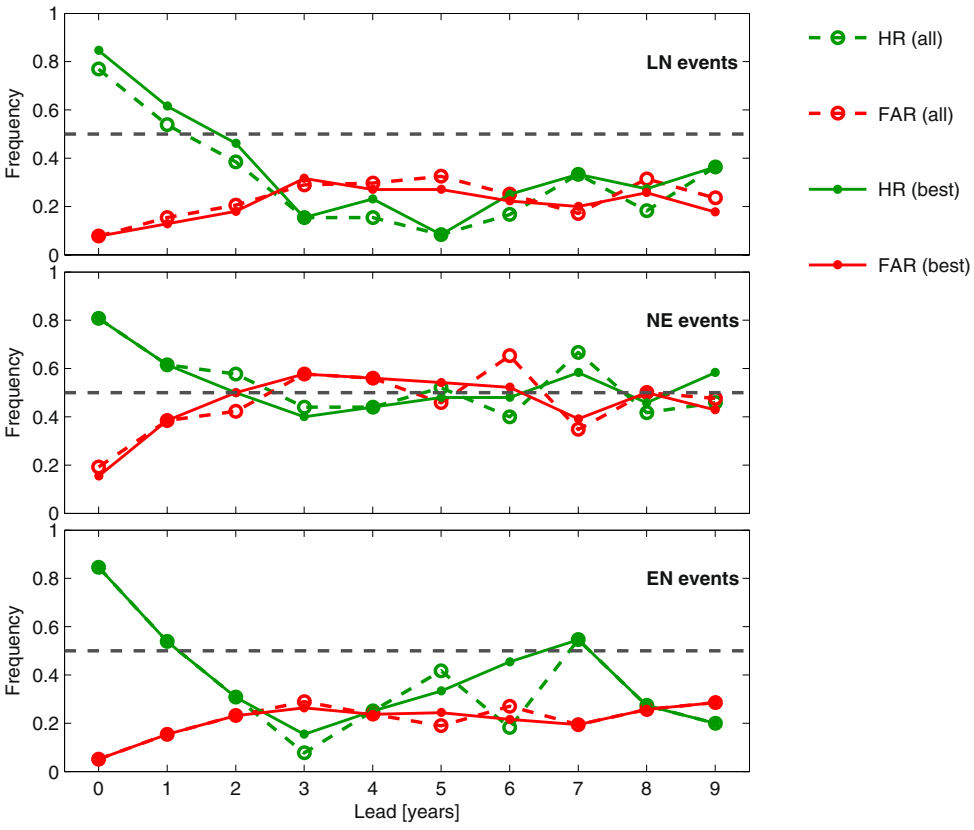

Fig. 12 EN3.4 multi-model ensemble mean deterministic skill for event detection. Comparison of the ENSO event detection deterministic skill as a function of lead year between the multi-model ensemble mean considering all systems (MMM, *dotted lines*, *open circles*) and the one considering the 'best' systems as defined on the text (BMM, *continuous lines*, *full circles*). The *green curves* represent the hit rate (HR) and the *red curves* the false alarm rate (FAR) for the LN (*top panel*), NE (*center panel*) and EN (*bottom panel*) events

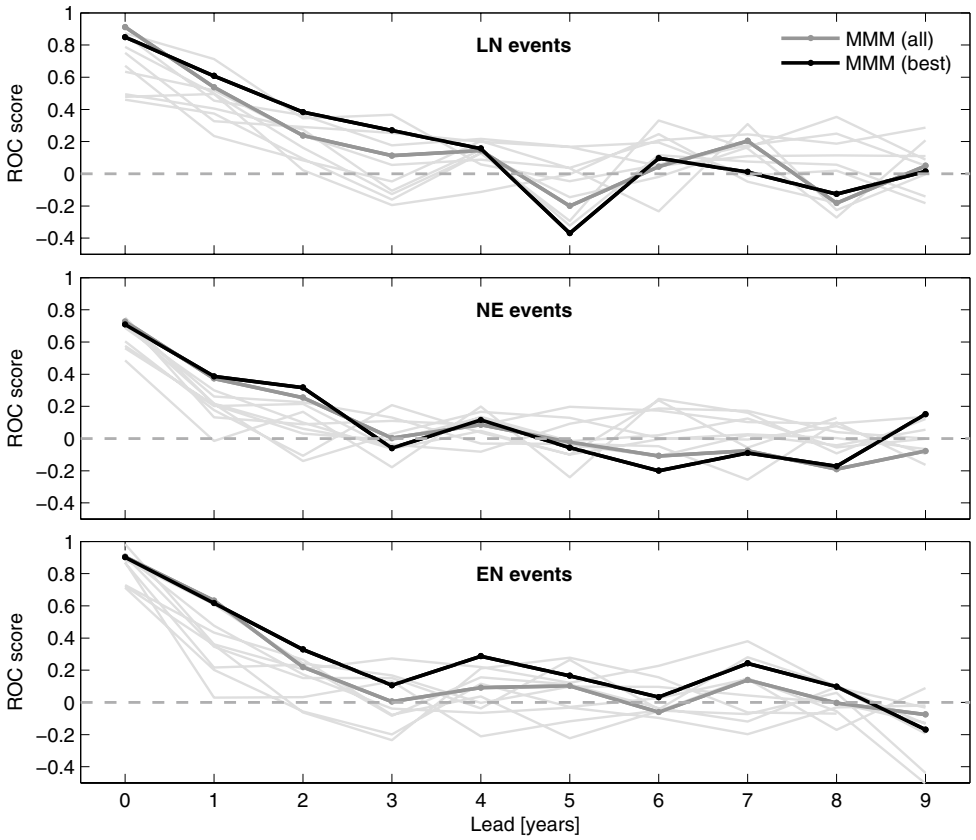

Fig. 13 EN3.4 multi-model ensemble mean probabilistic skill for event detection. Comparison of the ENSO events probabilistic ROC scores as a function of lead year between the multi-model ensemble mean considering all systems (MMM, *thick grey line*) and the one considering the 'best' systems as defined on the text (BMM, *thick black line*). The *thin grey lines* represent the ensemble mean skill of each individual system. The ROC scores are displayed for the LN (*top panel*), NE (*center panel*) and EN (*bottom panel*) events

individual models. In this case, the 'best' models (Table 2) are identified based on a set of skill metrics for the seasonal cycle of EN3.4, and its deterministic and probabilistic skills (first column of Table 2). A ranking system is created based on the highest skill for lead years 0–3, for each metric. In each case, the three systems with the highest skills (labeled '1st', '2nd' and '3rd') are identified and assigned points: 3, 2 and 1, respectively. Although redundancy exists in the metrics, a ranking based on a reduced set of metrics would yield similar results. The five decadal systems with the highest scores are selected for the 'best' models ensemble: CanCM4, EC-Earth i3, HadCM3 i3, GFDL-CM2.1 and MIROC5. Even though there is a tie in the points between the MIROC5 and the HadCM3 i2, the first was chosen based on two criteria: to avoid repeating the HadCM3 model components in the ensemble; and to favor the ENSO skill over representation of the EN3.4 seasonal cycle. The fact that these top five models are the same whether the selection is based on all metrics or if it is based solely on ENSO skill implies robustness. One can notice that the ability of the decadal systems to represent the EN3.4 seasonal cycle doesn't seem to be tied to the skill to represent ENSO variability. For example, even though CanCM4 is the system that leads the ranking by a large margin, it is not the one with the highest ability to represent the seasonal cycle of EN3.4.

A comparison between the deterministic skill metrics of DJF EN3.4 variability shows that the 'best' models ensemble does present an improvement both in higher correlations (Fig. 11a) for lead years up to 7, and in the lower RMSE (Fig. 11b) for lead years up to 3. The most substantial improvement is found in the ACC for lead year 2, for which the correlation for the 'best' models ensemble mean (BMM) becomes statistically significant (Fig. 11a).

The improvements in the deterministic skill for ENSO event detection (Fig. 12) that arise from the selection of the 'best' models ensemble are largest for the LN events at lead times from 0 to 2 years (Fig. 12, top panel). The performance in the case of NE events (Fig. 12, center panel) is very similar between ensembles, and for lead year 2 it actually worsens. For EN events (Fig. 12, bottom panel), the skill is almost identical for lead years 0–2 and shows some improvement for lead year 3.

Probabilistic ROC scores show similar results (Fig. 13) to the deterministic results. For the LN events (Fig. 13, top panel), the ROC score is very similar but slightly less for the 'best' models ensemble compared to the full MMM for lead year 0. After that, the 'best' model ensemble yields relative improvement for lead years 1–4. In the case of the NE events (Fig. 13, center panel) the scores are very similar, with slight improvement for lead year 2. Finally, the EN events (Fig. 13, bottom panel) show higher scores in

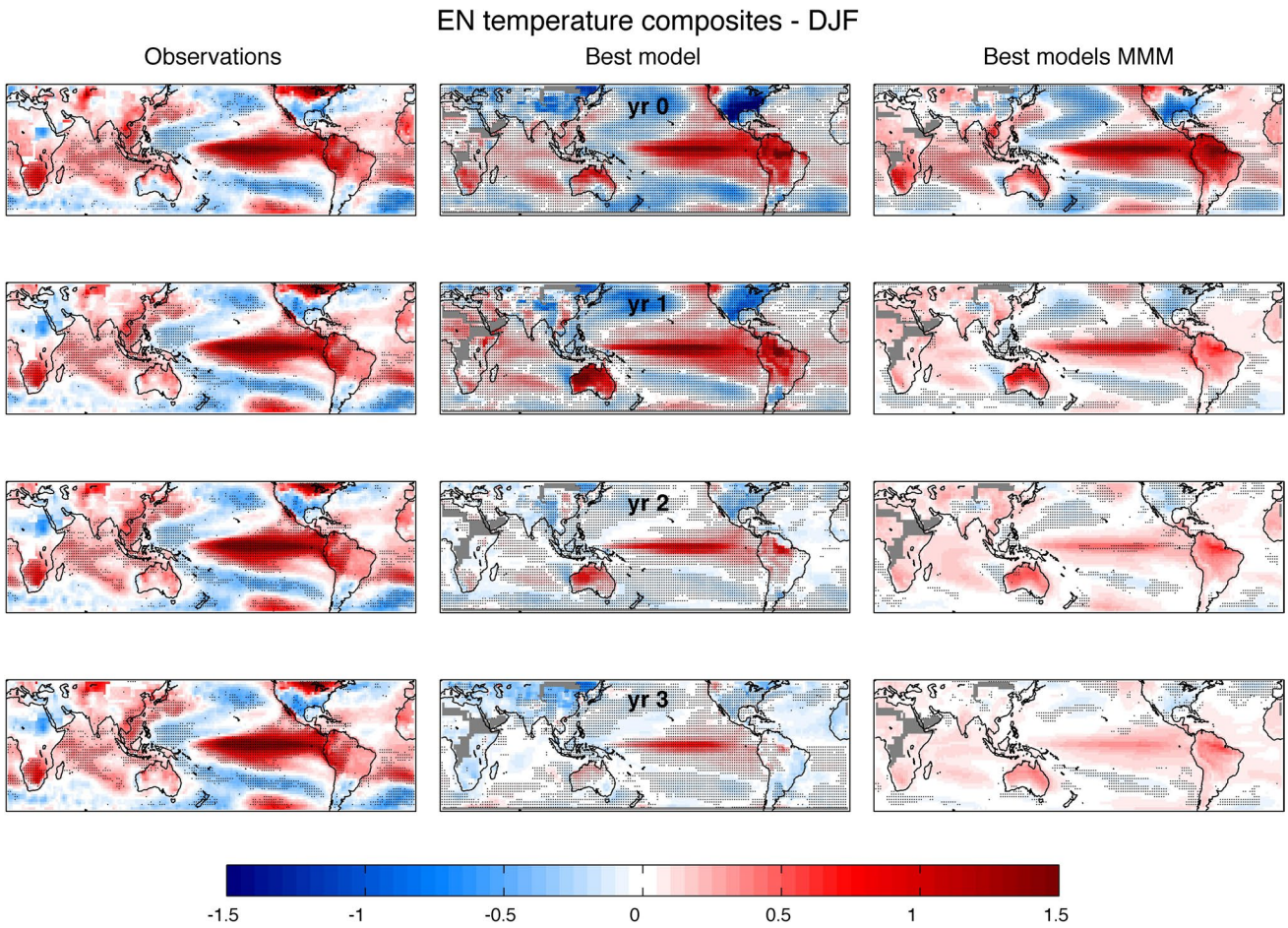

Fig. 14 Comparison of the composites of near-surface temperature anomalies for EN events. The *left column* presents the near surface temperature composites for the observational CAMS dataset. The *central column* corresponds to the best system (CanCM4) and the *right column* to the multi-model ensemble mean of the subset of systems identified as the 'best' models (BMM, see text). From *top* to *bottom*, *rows* correspond to lead years 0, 1, 2 and 3. The *grey hatching* indicates statistical significance. In the case of the observations, a hypergeometric test with 95 % confidence level was applied. In the case of the models, the hatching indicates sign agreement among ensemble members of at least 60 %

the case of the 'best' models ensemble for lead times higher than 1 year and up to 8 years out.

Different tests were applied to determine if the skill improvement observed here is statistically significant. Firstly, following DelSole and Tippett (2014), the results from four tests were compared: the sign test, the Wilkinson test, the Morgan–Granger–Newbold test, and a permutation test. They all revealed that the improvements in the deterministic skill observed for lead years 1 and 2 were statistically significant at the 95 % confidence level (with the exception of lead year 2 in the sign test, which was significant at the 90 % level). Secondly, these results were compared with those obtained from a bootstrapping approach applied to the difference in correlation coefficients, as applied in Goddard et al. (2013). Consistently, lead year 1 resulted significant at the 90 % level and lead year 2 at the 95 % level. Finally, a similar bootstrapping approach was applied to the difference in probabilistic ROC scores, but no statistical significance was achieved in this case. Nonetheless, this does not contradict the fact that the process of generating multi-model ensembles yields higher probabilistic skills than most individual hindcasts, as seen in Sect. 3.3.

3.5 ENSO teleconnections

To assess if the ability of the decadal prediction systems to reproduce the EN3.4 activity is reflected in temperature and precipitation signals over land, the teleconnections are explored using composites. For both variables, and for both EN and LN events, composites of anomalies were created using observational datasets, the most skillful system (CanCM4) and the 'best' models ensemble (BMM), and for lead years from 0 to 4. The statistical significance of the

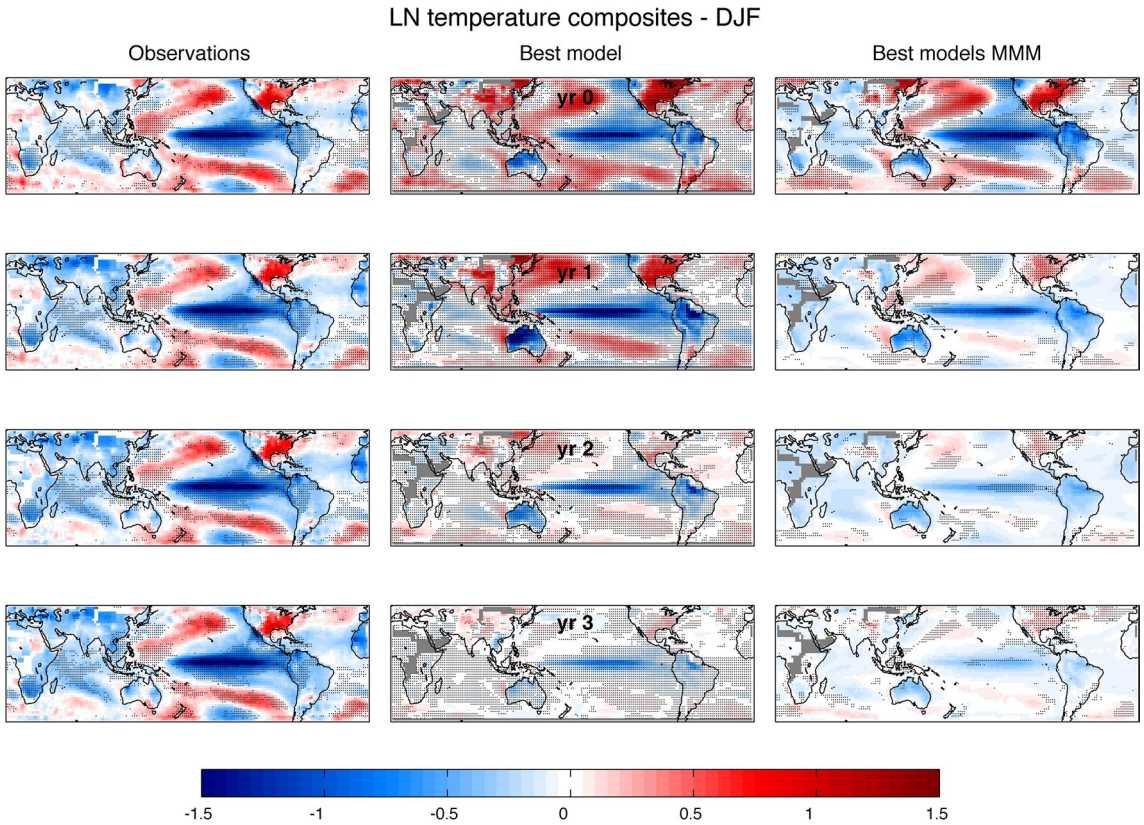

Fig. 15 Same as Fig. 14 but for LN events

composites was assessed through a hypergeometric tests in the case of the observations, following Mason and Goddard (2001), and through ensemble member sign agreement in the case of the models. Confidence levels are indicated in the figure captions.

The three composites of near-surface temperature during EN events (Fig. 14) show very similar patterns for lead year 0 (Fig. 14, top row), especially in the vicinity of the tropical Pacific basin. The results are better for the BMM mean because it corrects some of the biases observed in the CanCM4 system, such as those over Asia and the overestimation of the negative anomalies over North America. Nonetheless, some discrepancies remain, such as the overestimation of the anomalies over the Amazon core and the extension of the negative anomalies over North America. In the case of lead year 1 (Fig. 14, second row from top), one can notice some differences in the observed composites (left panel) with respect to lead year 0, such as over Africa and the Middle East, and over Australia. This suggests that there is some sensitivity to the climatological period, probably due to the diverse responses of EN events of varied magnitudes and structures. The magnitudes of the anomalies decrease more dramatically in the BMM than for CanCM4, likely due to the multi-model averaging effect. A notable bias that is seen for lead year 1 and longer is the displacement of the tropical Pacific temperature anomalies to the West. The shift is somewhat stronger in the case of CanCM4 and less pronounces in the BMM. Over land, the largest differences between the predictions and the observed composites are seen over North America. For lead year 2, the same anomalies in the composites are found, but the amplitudes of the simulated anomalies decrease further. Finally, for lead year 3, the discrepancies with the observed composites increase, especially in the case of CanCM4, where unrealistic cold anomalies dominate most regions over land. The BMM shows realistic patterns even for lead year 3, though with attenuated magnitudes.

For the temperature composites for LN events (Fig. 15), the lead year 0 maps show that the main structures are well captured both by CanCM4 and by the BMM. The former has the strongest disagreements over Asia and in the overestimation of the anomalies over North America, Australia and southeastern South America, and in the pattern over the Atlantic Ocean and Africa. These problems are milder in the BMM, for which the largest differences are seen over Europe, northern Africa and northeastern Asia.

As lead years evolve, smaller changes are seen in the observed composites than in the case of EN events

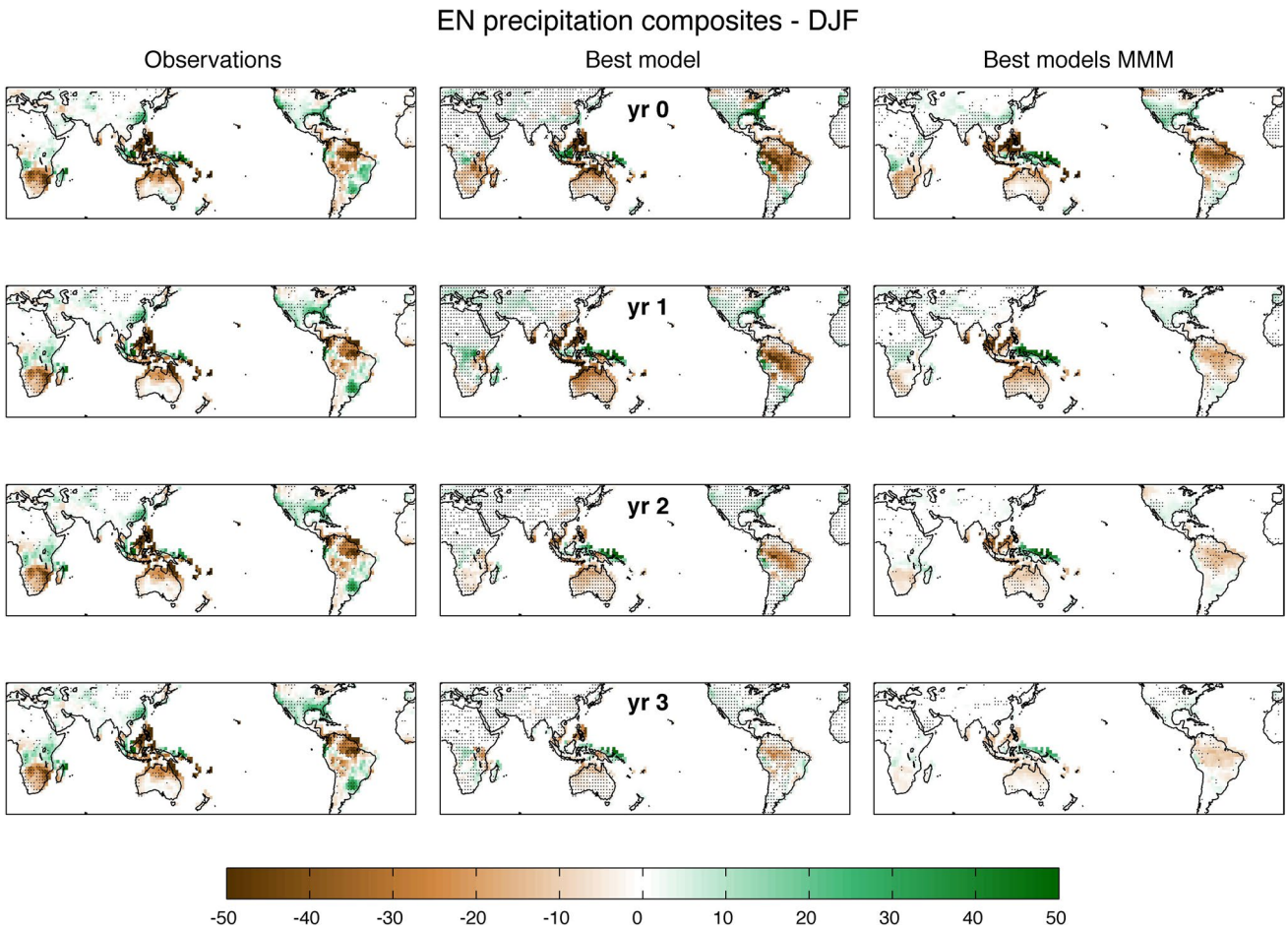

Fig. 16 Comparison of the composites of precipitation anomalies for EN events. The *left column* presents the near surface temperature composites for the observational GPCCv4 dataset. The *central column* corresponds to the best system (CanCM4) and the *right column* to the multi-model ensemble mean of the subset of systems identified as the 'best' models (BMM, see text). From *top* to *bottom*, *rows* correspond to lead years 0, 1, 2 and 3. The *grey hatching* indicates statistical significance. In the case of the observations, a hypergeometric test with 95 % confidence level was applied. In the case of the models, the *hatching* indicates sign agreement among ensemble members of at least 60 %

(Fig. 14), meaning that there is a smallest sensitivity to the climatological period. In the case of the hindcast composites, the largest systematic change is the shift of the temperature anomalies over the Pacific basin to the west, which is slightly weaker for the BMM. The latter, however, shows a slightly stronger decrease in the amplitude of the anomalies, especially going from lead years 0–1. Up to lead year 2, the pattern of the BMM composite are realistic (except over central Asia) but of much smaller amplitude than the observed ones. CanCM4 shows larger discrepancies.

Precipitation composites for the EN events (Fig. 16) show the strongest anomalies in the tropical band and in the Southern Hemisphere, with the exception of the anomalies over North America. Overall, the patterns observed for lead year 0 are well captured by both CanCM4 and the BMM. The latter is slightly better, though it underestimates the anomalies over eastern South America, Australia, and southern Africa, and it shows other minor differences over the Maritime Continent. As lead time grows, the patterns are preserved up to lead year 3, though with smaller amplitudes. The loss of signal is particularly noticeable over Asia and western North America. In the case of the LN events precipitation composites (Fig. 17), some discrepancies are present from lead year 0, such as over West Africa, northern Australia and eastern South America. Even though some features of the observed composite are preserved by lead years 2 and 3 in the BMM composites, with damped magnitudes, some significant differences are observed such as the predominance of dry anomalies over the eastern Maritime Continent.

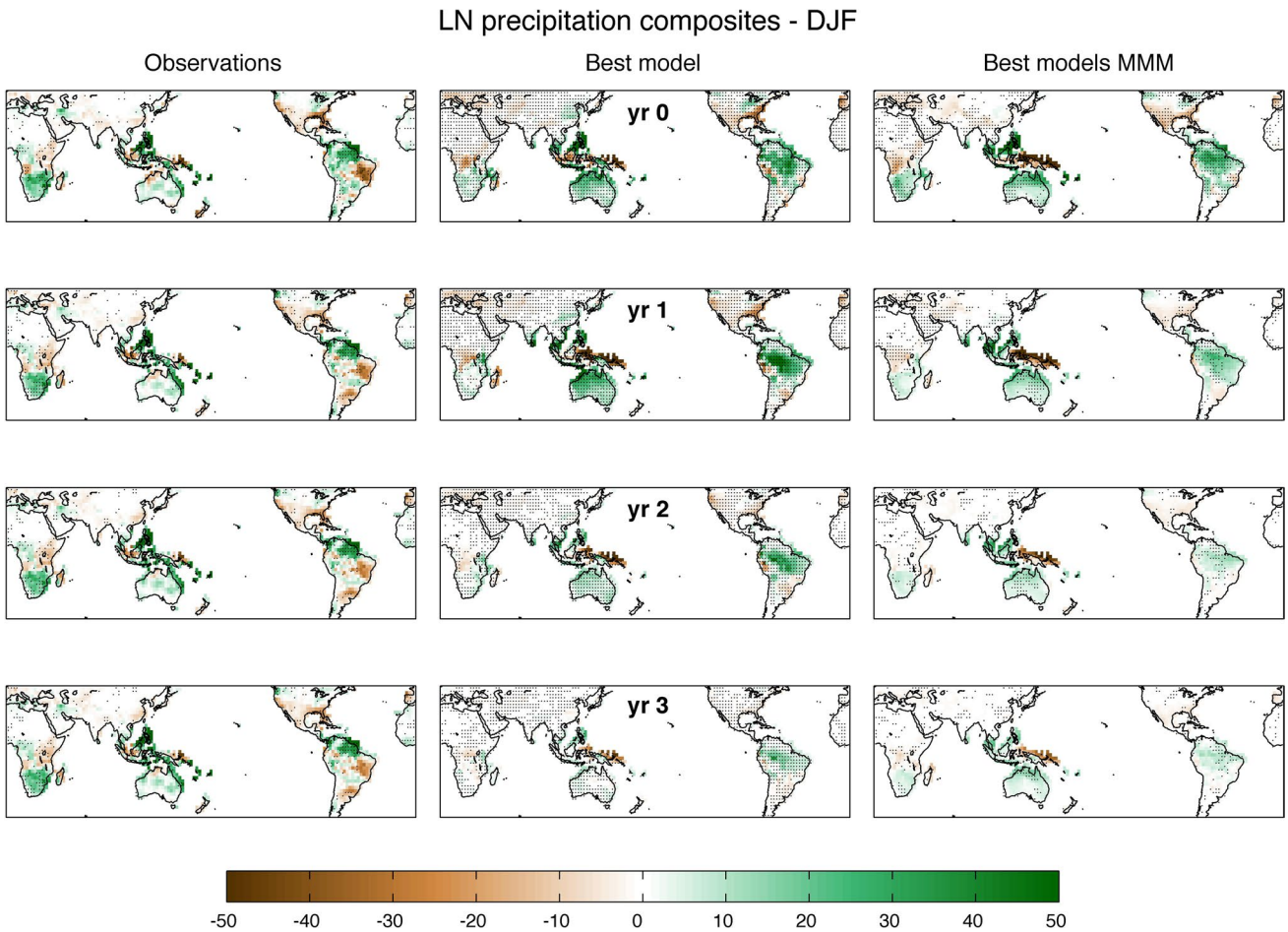

Fig. 17 Same as Fig. 16 but for LN events

4 Summary and discussion

This work presents an analysis of the potential long-lead ENSO predictability based on the CMIP5 decadal hindcasts. The ability of nine state-of-the-art decadal prediction systems to represent some features of the EN3.4 index, such as its seasonal cycle and the annual cycle of its interannual variability was assessed. Additionally, through the definition of EN and LN events for the DJF season, the deterministic and probabilistic skills of the decadal systems were analyzed for lead years 0 (months 0–2) to 9.

In agreement with previous results for seasonal prediction systems and climate change projections (e.g., Guilyardi et al. 2004; Guilyardi 2006; Jin et al. 2008; Guilyardi et al. 2009), most of the decadal prediction systems are able to reproduce the seasonal cycle of the EN3.4 index to some extent, with adequate amplitude and timing, for several lead years, though with some spread among systems and lead times. Also in congruence, the annual cycle of the EN3.4 variability is not well captured by the systems. Most models fail to represent both the timing and the amplitude of this cycle, and the discrepancies typically increase with the lead year. It has been suggested that this problem is linked to the inability of the systems to correctly model tropical wind variability, such as the westerly wind bursts (e.g., Guilyardi et al. 2009; Hu et al. 2014). However, the results of this study suggest that the skill for ENSO in decadal systems is not tied to their ability to reproduce the seasonal cycle of EN3.4 and its properties, as it was stated in previous works (e.g., Jin et al. 2008).

An analysis of the deterministic skill to represent the EN3.4 index—that of each system's ensemble mean—revealed that most models have significant ability (i.e., assessed through the ACC and the RMSE) for lead years 0 and 1. Some of them, such as the HadCM3 and EC-Earth i3 systems, are also skillful at lead year 2 (24–26 months). These results are comparable to those found for seasonal prediction systems for up to 12 months (e.g., Chen and Cane 2008; Barnston et al. 2012). The deterministic skill of the MMM is similar to that of the individual systems, with stronger improvements observed for the RMSE. An analysis of the deterministic ENSO ROC diagrams, hit rates

(HRs) and false alarm rates (FARs) showed that most systems are skillful for up to lead year 1 (12–14 months), and some models maintain their ability to detect events for lead year 2, such as CanCM4, EC-earth and HadCM3 i3. The deterministic event detection skill was found to be largest for EN events, closely followed by that of LN events, with the NE events being the hardest to discriminate.

Making use of all the information contained in the full decadal prediction ensembles, the probabilistic skill for ENSO event detection was also assessed. Consistently, the decadal prediction systems are skillful for lead years 0–2 and the MMM for lead years up to 3 and 4. Unlike the deterministic case, the probabilistic skill for event detection is larger for LN events than for EN events. NE years are the harder to predict in every case.

A subset of five hindcast systems was chosen based on their overall EN3.4 skill. The ability of this new 'best' models ensemble (BMM) to represent the EN3.4 variability is greater than that of the standard MMM for lead years up to 3 for the deterministic skill metrics. In the case of the probabilistic skill, some improvements were seen for lead years higher than 1, although not statistically significant.

The EN and LN teleconnections were computed for near-surface temperature and precipitation, to assess if the ability of the systems to capture the ENSO variability could be translated into potential predictive skill over land. In the case of the temperature anomalies, a comparison between the observed composites and those corresponding to the best system (CanCM4), and to the BMM show good agreement for lead year 0, especially for the latter. Both hindcasts simulate the composites, though they lose amplitude with the increasing lead time. The largest bias over the tropical oceans consists of a westward displacement of the warm (cold) anomalous center observed over the central Pacific for EN (LN) events for increasing lead time, which has previously been detected in seasonal prediction systems (Chen et al. 2004; Jin et al. 2008). The largest biases over land are found over Asia for all lead years, though milder in the BMM. In the case of precipitation, the observed composite patterns are reasonably well represented though with decreasing amplitudes for lead years up to 2, and with some discrepancies over the Southern Hemisphere landmasses. For precipitation, BMM is also slightly better than CanCM4. Nonetheless, particular regions with strong ENSO teleconnections such as South Eastern South America show reasonable teleconnections for all analyzed lead years, in agreement with previous results (Gonzalez and Goddard 2013).

Discrepancies between the observed and modeled ENSO teleconnections are likely due, in part, to biases in the location of maximum SST anomalies, leading to biases in the location of coupling between the tropical SST anomalies and the atmospheric responses, such as the anomalies of the trade winds (e.g., Guilyardi 2006). Small discrepancies in the magnitude and pattern of ENSO SST anomalies can significantly degrade the associated teleconnections (e.g., Langenbrunner and Neelin 2013; Coelho and Goddard 2009). Other important sources of errors are the misrepresentation of topography, which induces displacements in wavetrains, and of land-sea contrasts. Additionally, small discrepancies in the magnitude and pattern of ENSO SST anomalies can significantly degrade the associated teleconnections (e.g., Langenbrunner and Neelin 2013; Coelho and Goddard 2009).

Although not a focus of this study, the presence of two pairs of decadal systems, for which the only difference is the initialization strategy (i.e., EC-Earth i1 vs. i3 and HadCM3 i2 vs. i3), provides an initial estimate of its impact on ENSO representation. The systems that applied an anomaly initialization scheme (i.e., EC-Earth i3 and HadCM3 i2) exhibited smaller errors in the representation of the EN3.4 seasonal cycle and its variability (Fig. 5) and an overall greater seasonal cycle skill (Table 2), which were computed using raw modeled data. However, the distinction between paired prediction systems did not hold for the ENSO skill, which was computed after trend correction.

Contrary to previous results (e.g., Chen et al. 2004; Jin et al. 2008), the analysis presented here suggests that the ability of these systems to detect ENSO events is not necessarily tied to the event's strength. However, the number of events captured by the study period does not allow a robust quantitative assessment, and so this is still an open question. A longer study period would also allow for an assessment of the spectral properties of ENSO variability in these decadal hindcast systems. It is important to note, that even the present study would not have been possible with the standard experimental design required for the decadal hindcasts from the CMIP5 ensemble, with only one start time every 5 years (e.g., Taylor et al. 2012; Meehl et al. 2014).

A similar study of the sub-surface mechanisms associated with ENSO might reveal a more extended predictability limit (e.g., Jin et al. 2008; Hu et al. 2014), but if that skill does not manifest itself at the surface, it cannot be translated into prediction ability over land.

The results presented in this study support the idea that improvements in CGCMs data assimilation, initialization schemes and the representation of surface fluxes can further improve ENSO prediction skill (e.g., Chen and Cane 2008; Jin et al. 2008).

Questions remain as to whether the use of decadal hindcasts, which may have the potential to better represent low frequency climate variability (e.g. Goddard et al. 2013; Meehl et al. 2014), leads to increased ability to predict the decadal modulation of ENSO (Wittenberg et al. 2014). Even though limitations exist due to the relatively small number of events captured by the hindcasts design, this will be the focus of future research. In addition, the design of this set

of decadal hindcasts is suitable to capture predictability increases due to the improvements in CGCMs themselves (e.g., parameterizations), in their assimilation and initialization schemes, their increase resolution, and even in the observing systems (e.g., Guilyardi et al. 2009; Barnston et al. 2012).

There is no perfect prediction system, but some of the decadal prediction systems analyzed here appear skillful and able to detect some of the events for long leads—up to 4 years. This is beyond the previously documented predictability limit of seasonal prediction systems. Given the fact that ENSO is not only the main driver of tropical climate variability but also drives major global teleconnections that impact agriculture, water resources and livelihoods around the world, these results are of utmost relevance for decision-making and disaster preparedness.

Acknowledgments The authors would like to thank the Climate Forcasting Unit (CFU) at the Institut Català de Ciències del Clima (IC3) for providing access to their CMIP5 decadal hindcasts archive. We acknowledge the guidance of M. Tippett and S. Mason to assess the statistical significance of skill improvements. We thank two anonymous reviewers for their valuable help in the improvement of the original manuscript. This research was funded by NSF EaSM award 1049066 (Multi-scale Climate Information for Agricultural Planning in Southeastern South America for Coming Decades).

References

Barnston AG, Glantz M, He Y (1999) Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997–98 El Niño and the 1998 La Niña onset. Bull Am Meteorol Soc 80:217–243

Barnston AG, Tippett MK, L'Heureux ML, Li S, DeWitt DG (2012) Skill of real-time seasonal ENSO model predictions during 2002–11: is our capability increasing? Bull Amer Meteorol Soc 93:631–651. doi:10.1175/BAMS-D-11-00111.1

Blumenthal MB, Bell M, del Corral J, Cousin R, Khomyakov I (2014) IRI data library: enhancing accessibility of climate knowledge. Earth Perspect 1(1):1–12

Chen D, Cane MA (2008) El Niño prediction and predictability. J Comput Phys 227(7):3625–3640

Chen D, Cane MA, Kaplan A, Zebiak SE, Huang D (2004) Predictability of El Niño over the past 148 years. Nature 428(6984):733–736

Coelho CAS, Goddard L (2009) El Niño-induced tropical droughts in climate change projections. J Clim 22:6456–6476. doi:10.1175/2009JCLI3185.1

Collins WJ, Bellouin N, Doutriaux-Boucher N et al (2011) Development and evaluation of an Earth-System model – HadGEM2. Geosci Model Dev 4:1051–1075. doi:10.5194/gmd-4-1051-2011

DelSole T, Tippett MK (2014) Comparing forecast skill. Mon Weather Rev 142:4658–4678

Delworth T, Broccoli A, Rosati A et al (2006) GFDL's CM2 global coupled climate models. Part I: formulation and simulation characteristics. J Clim (19):643–674. doi:10.1175/JCLI3629.1

García-Serrano J, Doblas-Reyes FJ (2012) On the assessment of near-surface global temperature and North Atlantic multi-decadal variability in the ENSEMBLES decadal hindcast. Clim Dyn 39:2025–2040. doi:10.1007/s00382-012-1413-1

Goddard L et al (2013) A verification framework for interannual-to-decadal predictions experiments. Clim Dyn 40:245–272. doi:10.1007/s00382-012-1481-2

Gonzalez PLM, Goddard L (2013) Seasonal-to-interannual variability of precipitation over southeastern South America in CMIP5 decadal hindcasts. Presented at international workshop on seasonal to decadal prediction. Toulouse, France, 13–16 May 2013

Guilyardi E (2006) El Niño-mean state-seasonal cycle interactions in a multi-model ensemble. Clim Dyn 26:329–348

Guilyardi E et al (2004) Representing El Niño in coupled ocean–atmosphere GCMs: the dominant role of the atmospheric component. J Clim 17:4623–4629. doi:10.1175/JCLI-3260.1

Guilyardi E, Wittenberg A, Fedorov A, Collins M, Wang C, Capotondi A, Oldenborgh GJ, Stockdale T (2009) Understanding El Niño in ocean–atmosphere general circulation models. Bull Am Meteorol Soc 90:324–340

Hazeleger W, Wang X, Severijns A et al (2011) EC-Earth V2.2: description and validation of a new seamless earth system prediction model. Clim Dyn 39(11):2611–2629. doi:10.1007/s00382-011-1228-5

Hu S, Fedorov AV, Lengaigne M, Guilyardi E (2014) The impact of westerly wind bursts on the diversity and predictability of El Niño events: an ocean energetics perspective. Geophys Res Lett 41:4654–4663. doi:10.1002/2014GL059573

Jin EK, Kinter JL III, Wang B et al (2008) Current status of ENSO prediction skill in coupled ocean–atmosphere models. Clim Dyn 31:647–664

Kharin VV, Boer GJ, Merryfield WJ, Scinocca JF, Lee W-S (2012) Statistical adjustment of decadal predictions in a changing climate. Geophys Res Lett 39:L19705. doi:10.1029/2012GL052647

Kirtman BP, Min D, Infanti JM et al (2014) The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bull Am Meteorol Soc 95:585–601. doi:10.1175/BAMS-D-12-00050.1

Langenbrunner B, Neelin JD (2013) Analyzing ENSO teleconnections in CMIP models as a measure of model fidelity in simulating precipitation. J Clim 26:4431–4446. doi:10.1175/JCLI-D-12-00542.1

Latif M, Sperber K, Arblaster J et al (2001) ENSIP: the El Niño simulation intercomparison project. Clim Dyn 18:255–276

Lienert F, Doblas-Reyes FJ (2013) Decadal prediction of interannual tropical and North Pacific sea surface temperature. J Geophys Res Atmos 118:5913–5922. doi:10.1002/jgrd.50469

Ludescher J, Gozolchiani A, Bogachev MI, Bunde A, Havlin S, Schellnhuber HJ (2014) Very early warning of next El Niño. PNAS 111(6):2064–2066. doi:10.1073/pnas.1323058111

Mason I (1982) A model for assessment of weather forecast. Aust Meteorol Mag 30:291–303

Mason SJ, Goddard L (2001) Probabilistic precipitation anomalies associated with ENSO. Bull Amer Meteor Soc 82:619–638

Mechoso CR, Robertson AW, Barth N et al (1995) The seasonal cycle over the tropical Pacific in coupled atmosphere–ocean generalcirculation models. Mon Weather Rev 123:2825–2838

Meehl GA et al (2014) Decadal climate prediction: an update from the trenches. Bull Am Meteorol Soc 95:243–267. doi:10.1175/BAMS-D-12-00241.1

Raddatz TJ, Reick CH, Knorr W et al (2007) Will the tropical land biosphere dominate the climate-carbon cycle feedback during the twenty-first century? Clim Dyn 29(6):565–574

Ropelewski CF, Halpert MS (1987) Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. Mon Weather Rev 115:1606–1626

Ropelewski CF, Janowiak JE, Halpert MS (1985) The analysis and display of real time surface climate data. Mon Weather Rev 113:1101–1106

Schneider U, Becker A, Meyer-Christoffer A, Rudolf B (2010) Global precipitation analysis products of the GPCC. DWD, Germany, Internet Publication, 1–12. http://www.dwd.de. Accessed 14 Dec 2014

Smith TM, Reynolds RW, Peterson TC, Lawrimore J (2008) Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). J Clim 21:2283–2296. doi:10.1175/2007JCLI2100.1

Taylor KE, Stouffer RJ, Meehl GA (2009) A summary of the CMIP5 experiment design. http://www-pcmdi.llnl.gov/. Accessed 14 Dec 2014

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498

Tippett MK, Barnston AG (2008) Skill of multimodel ENSO probability forecasts. Mon Weather Rev 136:3933–3946. doi:10.1175/2008MWR2431.1

Trenberth KE, Fasullo JT (2013) An apparent hiatus in global warming? Earth's Future. doi:10.1002/2013EF000165

Trenberth KE, Branstator GW, Karoly D, Kumar A, Lau N-C, Ropelewski C (1998) Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. J Geophys Res 103:14291–14324

van Oldenborgh GJ, Doblas-Reyes FJ, Wouters B, Hazeleger W (2012) Decadal prediction skill in a multi-model ensemble. Clim Dyn 38:1263–1280. doi:10.1007/s00382-012-1313-4

Watanabe M, Suzuki T, O'ishi R et al (2010) Improved climate simulation by MIROC5: mean states, variability, and climate sensitivity. J Clim 23:6312–6335. doi:10.1175/2010JCLI3679.1

WCRP (2011) Data and bias correction for decadal climate predictions. International CLIVAR Project Office Publication Series 150. www.wcrp-climate.org/decadal/references/DCPP_Bias_Correction.pdf. Accessed 14 Dec 2014

Wittenberg AT, Rosati A, Delworth TL, Vecchi GA, Zeng F (2014) ENSO modulation: is it decadally predictable? J Clim 27:2667–2681. doi:10.1175/JCLI-D-13-00577.1