ENSO Precipitation and Temperature Forecasts in the North American Multimodel Ensemble: Composite Analysis and Validation

LI-CHUAN CHEN

Earth System Science Interdisciplinary Center/Cooperative Institute for Climate and Satellites, University of Maryland, College Park, and Climate Prediction Center, NOAA/NWS/NCEP, College Park, Maryland

HUUG VAN DEN DOOL

Climate Prediction Center, NOAA/NWS/NCEP, College Park, Maryland

EMILY BECKER

Climate Prediction Center, NOAA/NWS/NCEP, College Park, and Innovim, LLC, Greenbelt, Maryland

QIN ZHANG

Climate Prediction Center, NOAA/NWS/NCEP, College Park, Maryland

(Manuscript received 18 December 2015, in final form 16 September 2016)

ABSTRACT

In this study, precipitation and temperature forecasts during El Niño–Southern Oscillation (ENSO) events are examined in six models in the North American Multimodel Ensemble (NMME), including the CFSv2, CanCM3, CanCM4, the Forecast-Oriented Low Ocean Resolution (FLOR) version of GFDL CM2.5, GEOS-5, and CCSM4 models, by comparing the model-based ENSO composites to the observed. The composite analysis is conducted using the 1982–2010 hindcasts for each of the six models with selected ENSO episodes based on the seasonal oceanic Niño index just prior to the date the forecasts were initiated. Two types of composites are constructed over the North American continent: one based on mean precipitation and temperature anomalies and the other based on their probability of occurrence in a tercile-based system. The composites apply to monthly mean conditions in November, December, January, February, and March as well as to the 5-month aggregates representing the winter conditions. For anomaly composites, the anomaly correlation coefficient and root-mean-square error against the observed composites are used for the evaluation. For probability composites, a new probability anomaly correlation measure and a root-mean probability score are developed for the assessment. All NMME models predict ENSO precipitation patterns well during wintertime; however, some models have large discrepancies between the model temperature composites and the observed. The fidelity is greater for the multimodel ensemble as well as for the 5-month aggregates. February tends to have higher scores than other winter months. For anomaly composites, most models perform slightly better in predicting El Niño patterns than La Niña patterns. For probability composites, all models have superior performance in predicting ENSO precipitation patterns than temperature patterns.

1. Introduction

El Niño–Southern Oscillation (ENSO) has a large influence on the seasonal precipitation P and temperature T patterns over the United States and across the globe (Ropelewski and Halpert 1986, 1987; Kiladis and Diaz 1989; Trenberth et al. 1998; Dai and Wigley 2000; Yang and DelSole 2012). The 1997/98 El Niño had record-breaking sea surface temperature anomalies in the tropical Pacific and a profound impact on the global climate, resulting in many extreme events around the world (Bell et al. 1999; Barnston et al. 1999). For example, flooding in the central and northeastern United States and the U.S. West Coast (Bell et al. 1999; Persson et al. 2005), the Mexican drought (Bell et al. 1999), the Yangtze River flood in China (Lau and Weng 2001), Indonesian forest fires (Gutman et al. 2000;

Corresponding author e-mail: Li-Chuan Chen, lichuan.chen@noaa.gov

DOI: 10.1175/JCLI-D-15-0903.1

Parameswaran et al. 2004), and excessive rainfall in southern Africa (Lyon and Mason 2007), all attributed to the 1997/98 event.

At the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC), a large effort is devoted to monitoring and forecasting of Niño-3.4 sea surface temperature and the tropical Pacific Ocean conditions in order to provide the most up-to-date information on the phase of the ENSO cycle. Statistical tools have been developed for objective seasonal prediction using Niño-3.4 region sea surface temperature forecasts in conjunction with observed temperature and precipitation composites keyed to phases of the ENSO cycle (Higgins et al. 2004). On the other hand, many studies (e.g., Kumar et al. 1996; Rowell 1998; Shukla et al. 2000; Mathieu et al. 2004; Saha et al. 2014; Yang and Jiang 2014) have shown that improved skill of P and T prediction in climate models can be attributed to the known impacts of ENSO signals, especially during the Northern Hemisphere (NH) cold season. Recent developments in multimodel ensembles provide a promising way to increase P and T predictive skill using dynamical model forecasts (Graham et al. 2000; Hagedorn et al. 2005; Weisheimer et al. 2009; Kirtman et al. 2014).

With a warm or cold event approaching, one not only wants to know whether a climate model can predict the onset of an ENSO event but also whether the model can adequately predict its impacts on remote P and T patterns if an ENSO event is in progress. In other words, to what extent does the real-time forecast by model M resemble (its own version of) ENSO composites. We here provide a tool to answer that question. We intend to take advantage of the large ensemble of the North American Multimodel Ensemble (NMME) and examine how well NMME (or its constituent models) forecasts ENSO events by comparing the model-based ENSO composites to the observed. The study of composites based on model forecast data has been attempted before. Smith and Ropelewski (1997) studied rainfall composites based on the NCEP Medium-Range Forecast Model spectral T40 version (Ji et al. 1994; Kumar et al. 1996) and found substantial discrepancies. Since then, much has advanced in atmospheric general circulation models and thus a reassessment of ENSO–precipitation (or temperature) relationships from climate models is needed.

In this study, we construct two types of composites over the North American continent: one based on mean precipitation and temperature anomalies in physical units, the other based on the probability of occurrence in a three-class forecast system. They are referred as anomaly and probability composites, respectively, hereafter. The composite analyses are conducted using the 1982–2010 hindcasts from six models in NMME with selected ENSO episodes based on the seasonal oceanic Niño index (ONI; Kousky and Higgins 2007) just prior to the date the forecasts were initiated. The composites apply to monthly mean conditions in November, December, January, February, and March (NDJFM) as well as to the 5-month aggregates representing the winter conditions.

To analyze how well the model composites resemble the observed, we compute performance scores for each model and month as well as the NMME ensemble and 5-month aggregates. For anomaly composites, we use the anomaly correlation coefficient (ACC) and root-mean-square error (RMSE) against the observed composites for evaluation. For probability composites, unlike conventional probabilistic forecast verification assuming binary outcomes in the observations, both model and observed composites are expressed in probability terms. Performance metrics for such validation are limited. Therefore, we develop a probability anomaly correlation (PAC) measure and a root-mean probability score (RMPS) for assessment. Our study is focused on land where ENSO impacts are the greatest (in terms of the population affected) and forecasts are most needed.

In the following, section 2 introduces the NMME forecast data and the precipitation and temperature observations used in the study. Section 3 describes the methodology for constructing the composites. Section 4 explains the performance metrics, including the development of the new scores. Section 5 presents the anomaly composite analysis and evaluation. The examination of probability composites is shown in section 6. Section 7 carries out a sensitivity analysis of the performance scores to the sample used for constructing the composites. Section 8 discusses the results and challenges of ENSO forecast validation. Finally, section 9 summarizes the major findings from the investigation.

2. Data

a. NMME seasonal forecast data

NMME is an experimental multimodel forecasting system consisting of coupled climate models from U.S. modeling centers (including NCEP, GFDL, NASA, and NCAR) and the Canadian Meteorological Centre (CMC), aimed at improving intraseasonal to interannual prediction capability as recommended by the National Research Council (NRC 2010). The multimodel ensemble approach has proven effective at quantifying prediction uncertainty due to uncertainty in model formulation and has proven to produce better forecast quality (on average) than the constituent single model

ensembles (Weisheimer et al. 2009; Kirtman et al. 2014; Becker et al. 2014). The NMME seasonal system currently contains eight climate models that provide various periods of hindcasts from 1981 to 2012 and real-time forecasts starting from August 2011. In this study, we selected six models, Climate Forecast System, version 2 (CFSv2; Saha et al. 2006, 2014); Canadian Centre for Climate Modelling and Analysis (CCCma) Third and Fourth Generation Canadian Coupled Global Climate Model (CanCM3 and CanCM4, respectively; Merryfield et al. 2013); the Forecast-Oriented Low Ocean Resolution (FLOR) version of GFDL CM2.5 (Vecchi et al. 2014; Jia et al. 2015); Goddard Earth Observing System model, version 5 (GEOS-5; Vernieres et al. 2012); and Community Climate System Model, version 4 (CCSM4; Danabasoglu et al. 2012), that have a common period of hindcasts from 1982 to 2010 for the evaluation. The number of ensemble members ranges from 10 (for CanCM3, CanCM4, and CCSM4) to 24 (for CFSv2 and FLOR), and NMME has a total of 89 ensemble members. Despite the original spatial resolution of the participating models, all NMME forecasts are remapped to a common grid system of $1° \times 1°$ resolution covering the globe. More detailed information about the NMME project and data can be found on NOAA Climate Test Bed website (http://www.nws.noaa.gov/ost/CTB/nmme.htm).

b. *Observed precipitation data*

The CPC precipitation reconstruction (PREC) global land analysis is used to construct the observed ENSO composites for comparison. PREC is a gridded monthly precipitation product that interpolated gauge observations from over 17 000 stations collected in the Global Historical Climatology Network (GHCN) and the Climate Anomaly Monitoring System (CAMS). Details of the PREC dataset and the optimal interpolation technique used to create this dataset are described in Chen et al. (2002). The PREC product is reprocessed to the $1° \times 1°$ NMME grid system from its original $0.5° \times 0.5°$ resolution using bilinear interpolation. Monthly data from January 1950 to December 2010 are used in this study.

c. *Observed temperature data*

The observed temperature composites are computed using a global monthly land surface temperature analysis—the GHCN–CAMS gridded 2-m temperature data. This dataset combines station observations from the GHCN and CAMS and employed the anomaly interpolation approach with spatially and/or temporally varying temperature lapse rates derived from the reanalysis for topographic adjustment (Fan and Van den Dool 2008). Similar to the PREC data, the GHCN–CAMS data are also reprocessed to the $1° \times 1°$ NMME grid system from its original $0.5° \times 0.5°$ resolution to be consistent in the analysis. Monthly data from January 1950 to December 2010 are employed as well.

3. ENSO composites

Two types of model composites are constructed in this study: one based on forecast anomalies and the other based on forecast probabilities. Their procedures are described below.

a. *Anomaly composites*

For each model, monthly ensemble mean P and T forecasts are first obtained by averaging all members. The P and T anomalies for a given start and lead time are then computed as the difference between the ensemble mean P and T forecasts and the lead-specific model climatology derived from the hindcast mean of all members and all years excluding the forecast year. The P and T anomaly composites for the warm ENSO (El Niño) events and cold ENSO (La Niña) events are simply the average of the ensemble P and T anomaly maps of selected years. The years are chosen based on the historical ONI (starting from 1950) published on the CPC website (at http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml). If the seasonal ONI just prior to the date the forecasts were initiated indicates a warm or cold ENSO episode, the forecasts are selected for the composite analysis. For example, July–September (JAS) 1982 ONI indicates a warm ENSO episode is in progress, and thus the forecasts with initial condition (IC) of 1 October 1982 are chosen for the El Niño composite analysis of November. In doing so, we avoid the question whether the model itself is actually predicting the |ONI| to be larger than 0.5.

Table 1 specifies all the years used in the composite analysis for each initial condition. Only the ENSO events that occurred between 1982 and 2010 are used for computing the model composites. The number of ENSO episodes varies with initial condition from 7 to 10 cases depending on the month. In this paper, we only present lead-1 month forecasts in the figures. For example, the November composites are the average of 8 yr of forecasts with IC of 1 October for El Niño (warm) events. Since the analysis applies to lead-1 forecasts, we do stay fairly close to the real world's classification of events. (Lead-7 composites based on the model's predicted ONI might look rather different.) We acknowledge that the quality of model composites may depend on lead (or seasonal mean), but from a practical standpoint it is easier to deal with

TABLE 1. Selected years used in the ENSO composite analysis. The years are chosen based on |ONI| ≥ 0.5 on average for the three consecutive months prior to the initial time of model integration. The 1982–2010 set is used for model and observed composites. The 1950–2010 set is used for observed composites only.

IC Month ENSO	1 Oct November		1 Nov December		1 Dec January		1 Jan February		1 Feb March	
	Warm	Cold	Warm	Cold	Warm	Cold	Warm	Cold	Warm	Cold
1950–81	1951	1950	1951	1954	1951	1950	1952	1951	1952	1951
	1953	1954	1953	1955	1953	1954	1954	1955	1954	1955
	1957	1955	1957	1956	1957	1955	1958	1956	1958	1956
	1963	1956	1963	1964	1963	1956	1959	1957	1959	1957
	1965	1964	1965	1970	1965	1964	1964	1965	1964	1965
	1968	1970	1968	1971	1968	1970	1966	1971	1966	1971
	1969	1971	1969	1973	1969	1971	1969	1972	1969	1972
	1972	1973	1972	1975	1972	1973	1970	1974	1970	1974
		1975	1976		1976	1974	1973	1975	1973	1975
			1977		1977	1975	1977	1976	1977	1976
							1978		1978	
1982–2010	1982	1985	1982	1983	1982	1983	1983	1984	1983	1984
	1986	1988	1986	1985	1986	1984	1987	1985	1987	1985
	1987	1998	1987	1988	1987	1988	1988	1989	1988	1989
	1991	1999	1991	1995	1991	1995	1992	1996	1992	1996
	1997	2000	1994	1998	1994	1998	1995	1999	1995	1999
	2002	2007	1997	1999	1997	1999	1998	2000	1998	2000
	2004	2010	2002	2000	2002	2000	2003	2001	2003	2001
	2009		2004	2007	2004	2007	2005	2006	2005	2006
			2006	2010	2006	2010	2007	2008	2007	2008
			2009		2009		2010	2009	2010	2009
Total No. of events from 1982 to 2010	8	7	10	9	10	9	10	10	10	10
Total No. of events from 1950 to 2010	16	16	20	17	20	19	21	20	21	20

short leads because the ENSO classification based on ONI in the real world applies better to forecasts for short leads.

The composites apply to monthly mean conditions in NDJFM as well as the 5-month aggregates to represent the winter conditions. We focus on NH winter only, when most ENSO cases in nature have happened and the extratropical impact should be the largest, according to theory (Opsteegh and Van den Dool 1980; Hoskins and Karoly 1981). The NMME composites are the equally weighted mean of the six models' composites.

b. *Probability composites*

For each model, P and T forecasts for a given start and lead time are classified into three categories (above, near, and below normal) based on the terciles derived from the hindcasts of all members excluding the forecast year. For precipitation forecasts, the tercile thresholds are the 33rd and 67th percentiles determined by fitting a gamma distribution to the hindcasts. For temperature forecasts, the tercile thresholds are set as mean ± 0.431 multiplied by the standard deviation by assuming a Gaussian distribution. The classification applies to each individual member forecast, and the number of ensemble members that fell into the three categories under the warm (El Niño) and cold (La Niña) events are counted for the selected ENSO years. For model composites, the years (between 1982 and 2010) are chosen based on the ONI criterion discussed in section 3a. At each grid point, the probability of occurrence for each category under the El Niño (or La Niña) condition is then calculated by dividing the total number of counts by the product of the number of the selected ENSO years and the number of ensemble members for each model.

The ENSO probability composites for NDJFM are the combination of all five winter months, that is, the probability of occurrence for each category is calculated by summing all counts in each of the five months (all at lead 1) divided by the total number of events from all five months. Similarly, the NMME probability composites are the combination of all six models by adding all counts in each category from the six models together, but note that the classification of each model is determined separately in respect to the model's own hindcast distribution for a particular month.

c. *Observed composites*

The model composites provide a general picture of how NMME models predict ENSO impacts on P and T

patterns. To examine if NMME models can adequately reproduce ENSO signals in their forecasts, we also create observed ENSO composites using historical P and T observations for comparison. The observed composites are computed using the same procedures as those used to derive model composites. For instance, November 1982 is part of the observed El Niño composite because the ONI satisfies the threshold just prior to 1 October. Different from model composites, observed composites are constructed based on a single realization, and thus the sample size is much smaller than that of model composites. For observed probability composites, calculations based solely on 1982–2010 events (7–10 cases) are not sufficient to yield statistically meaningful results and show sudden category changes in adjacent areas and discontinuities in spatial patterns for individual month composites. To increase the sample size to reach a more stable result for observed probability calculations, we selected ENSO events from the period of 1950–2010. Depending on month, the criterion gives 16–21 ENSO events in that month in this 61-yr period (also listed in Table 1). These events provide a better estimate of the probability of occurrence from limited observations [given that CPC places higher confidence in years after 1950 and no longer uses ENSO cases before 1950 as in Ropelewski and Halpert (1986, 1987)]. For observed anomaly composites, we also explore two scenarios: one based on the 1982–2010 events to coincide with the hindcast period and the other based on the 1950–2010 events to have a larger sample.

Our observed composites do not follow exactly the method used at CPC (Higgins et al. 2004) for making ENSO composites. In particular, we did not attempt to separate the signal into high and low frequency [a debatable activity in Higgins et al. (2004), attempting to deal with global change]. Our concern is mainly whether models resemble observations, and both include unspecified trends. The CPC case selection is furthermore ''simultaneous'' with no lag in time—that may also lead to minor differences. Our goal, similar to Smith and Ropelewski (1997), is to diagnose the models' ability in reproducing P and T patterns under ENSO conditions.

4. Performance metrics

The ENSO composite of a given variable (P or T) for a given model and a given month is validated against the P or T composite derived from the observations for the same target month. For example, the El Niño T anomaly composite for NMME February forecasts (with IC of 1 January) is validated with the El Niño T anomaly composite derived from the observations for the selected Februarys (given the ONI classification just prior to 1 January; see Table 1 for participating years). Under this framework, the evaluation is straightforward for the anomaly composites; note that there is no dimension time in calculating performance metrics after compositing. We employ the ACC and RMSE, commonly used in forecast verification, as the performance metrics but summing in space only.

The ACC measures the linear association between the model anomaly and the observed anomaly across a given domain with area weighting. It is calculated using the formula

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} (w_i \times X_{m_i} \times X_{o_i})}{\sqrt{\sum_{i=1}^{n} (w_i \times X_{m_i}^2) \times \sum_{i=1}^{n} (w_i \times X_{o_i}^2)}}, \tag{1}$$

where X_{m_i} is the model ensemble mean anomaly (either P or T) at grid i, X_{o_i} is the observed anomaly at grid i, n is the total number of land grid points within the North American domain, and w_i is the weighting coefficient based on the latitude (y) of grid i, that is,

$$w_i = \cos(y_i). \tag{2}$$

The RMSE calculates the average of the squared differences between the model ensemble mean anomaly and the observed anomaly over the North American domain with area weighting, and thus it has the same unit as the measurements. For the P anomaly, the unit is in millimeters per day, and for the T anomaly, the unit is degrees Celsius. The equation is written as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} w_i (X_{m_i} - X_{o_i})^2}{\sum_{i=1}^{n} w_i}}. \tag{3}$$

For the probability composites, both model and observed composites are expressed in probability terms—a unique case in verification study. Classical probabilistic forecasts in a tercile-based system are usually validated under the assumption that the verifying quantities are exact and by assigning binary outcomes to the observations. Most standard metrics formulated under this assumption, such as Brier score and ranked probability score, cannot be directly applied to our case without modifications

(Candille and Talagrand 2008). To have similar measures to the anomaly composites for comparison, we develop a PAC and RMPS for the assessment.

At each grid, the model and observed probabilities are given in three categories: above, near, and below normal. We define the probability anomaly for a category as the difference between the model (or observed) probability and the climatology value (i.e., 0.333) for the given category. The PAC, mimicking the ACC, quantifies the strength of the linear association between the model probability anomaly and the observed probability anomaly across all three forecast categories with area weighting. It is computed using the formula

$$\mathrm{PAC} = \frac{\sum_{i=1}^{n} w_i (A_{m_i} \times A_{o_i} + N_{m_i} \times N_{o_i} + B_{m_i} \times B_{o_i})}{\sqrt{\sum_{i=1}^{n} w_i (A_{m_i}^2 + N_{m_i}^2 + B_{m_i}^2) \times \sum_{i=1}^{n} w_i (A_{o_i}^2 + N_{o_i}^2 + B_{o_i}^2)}}, \tag{4}$$

where A_m, N_m, and B_m are the probability anomalies of the above-, near-, and below-normal categories from the model composite, respectively, and A_o, N_o, and B_o are the probability anomalies of the above-, near-, and below-normal categories from the observed composite, respectively.

The PAC and ACC calculated in this study are spatial correlations (aggregated/averaged across space). One way to assess the validity of spatial correlations is through statistical significance tests. Because of the dependency and inhomogeneity of climate fields, significance testing for spatial correlations is still an open research topic. Here, we adopt the approach described in Van den Dool (2007) and use the degrees of freedom (dof; or effective sample size) to determine a significance threshold based on Student's t test or Fisher z test (Wilks 2011). By doing so, we implicitly assume that probability anomalies and anomaly correlations are Gaussian distributions. The dof is obtained from Wang and Shen (1999), who compared four different methods for estimating spatial dof and suggested that dof is around 60–80 in the NH winter months. Using this approach, the significance threshold is about ± 0.2 for both the Student's t test and Fisher z test.

Analogous to the RMSE, the RMPS measures the difference between the model composite and the observed composite with area weighting. Specifically, it is the root-mean-square error between the model and observed probability anomalies from all three forecast categories, that is,

$$\mathrm{RMPS} = \sqrt{\frac{\sum_{i=1}^{n} w_i [(A_{m_i} - A_{o_i})^2 + (N_{m_i} - N_{o_i})^2 + (B_{m_i} - B_{o_i})^2]}{\sum_{i=1}^{n} 3w_i}}. \tag{5}$$

5. Anomaly composite analysis

Figure 1 presents the NMME El Niño P anomaly composites for November, December, January, February, and March forecasts, and these are shown individually, so any evolution of the ENSO response pattern throughout the winter can be judged. In the display, Fig. 1f for NDJFM has been added that combines all five winter months at lead 1. For each model (figures not shown but are available on CPC NMME website at http://www.cpc.ncep.noaa.gov/products/NMME/enso/), this is the typical winter ENSO composite about 1–2 months after initiation of the forecast. In Fig. 1f (NDJFM), the sample size is attractively large. Five winter months times (about) nine cases times the number of ensemble members would be a sample size of around 4000 for NMME, 1100 for CFSv2 and FLOR, 500 for GEOS-5, and 450 for CCSM4 and both CanCM models. For individual months the sample size is 5 times smaller. For the 1982–2010 observed composites, the sample size for NDJFM is only about 45, and it is very possible that the NDJFM composite is a better prediction for an independent new case in January than a January composite alone (based on about nine cases).

Figure 2 shows the El Niño P anomaly composites of December, February, and NDJFM based on the 1950–2010 and 1982–2010 observations. There are slight differences in magnitude between the two sets of composites because of the differences in sample size and period. Despite that, both sets of composites closely resemble the El Niño precipitation pattern characterized by Ropelewski and Halpert (1986, 1987) using station data from 1875 to 1980, with enhanced rainfall over the southern United States and northern Mexico and drier conditions over the Pacific Northwest and Ohio River valley.

Comparing NMME P anomaly composites (Fig. 1) to the observed, we can see that NMME is able to capture the evolution of ENSO response and reproduce El Niño precipitation patterns well, and

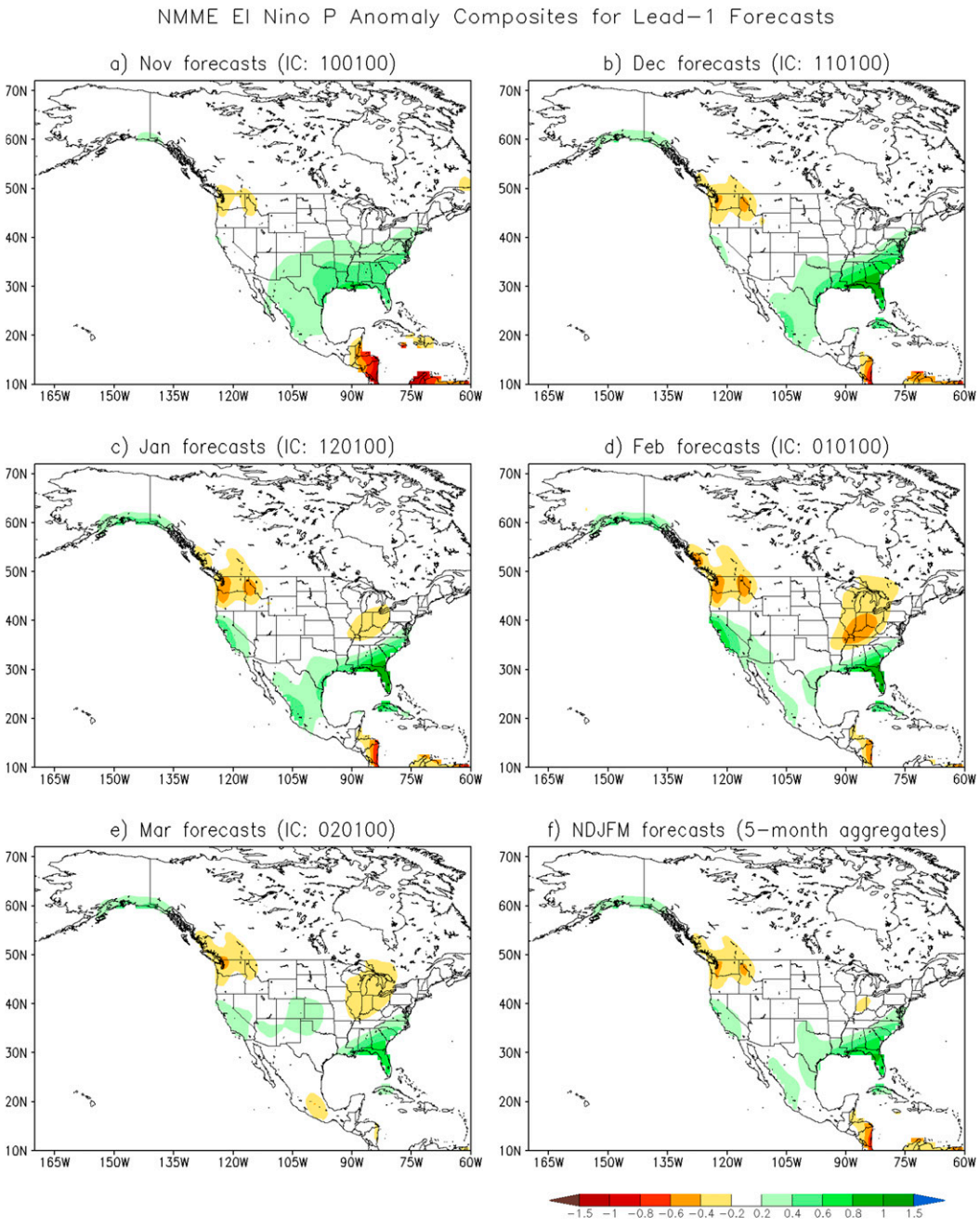

FIG. 1. NMME El Niño precipitation anomaly (mm day^{-1}) composites for lead-1 forecasts with initial conditions of (a) 1 Oct, (b) 1 Nov, (c) 1 Dec, (d) 1 Jan, and (e) 1 Feb, and for (f) 5-month (NDJFM) aggregates.

this is true for all models. There are subtle differences between the NMME and observed composites throughout the winter months. The most apparent difference is over the Pacific Northwest. In the NMME composites, negative anomalies exist in this region from November to March. In the observed composites (both the 1950–2010 and 1982–2010 sets), a strong dry signal appears in November over the Pacific Northwest, then it switches to wet conditions in January (not shown) and back to dry conditions in February and March.

In a more compact display, Fig. 3 shows the La Niña *P* anomaly composites for NDJFM based on 1982–2010 and 1950–2010 observations, NMME, and the six

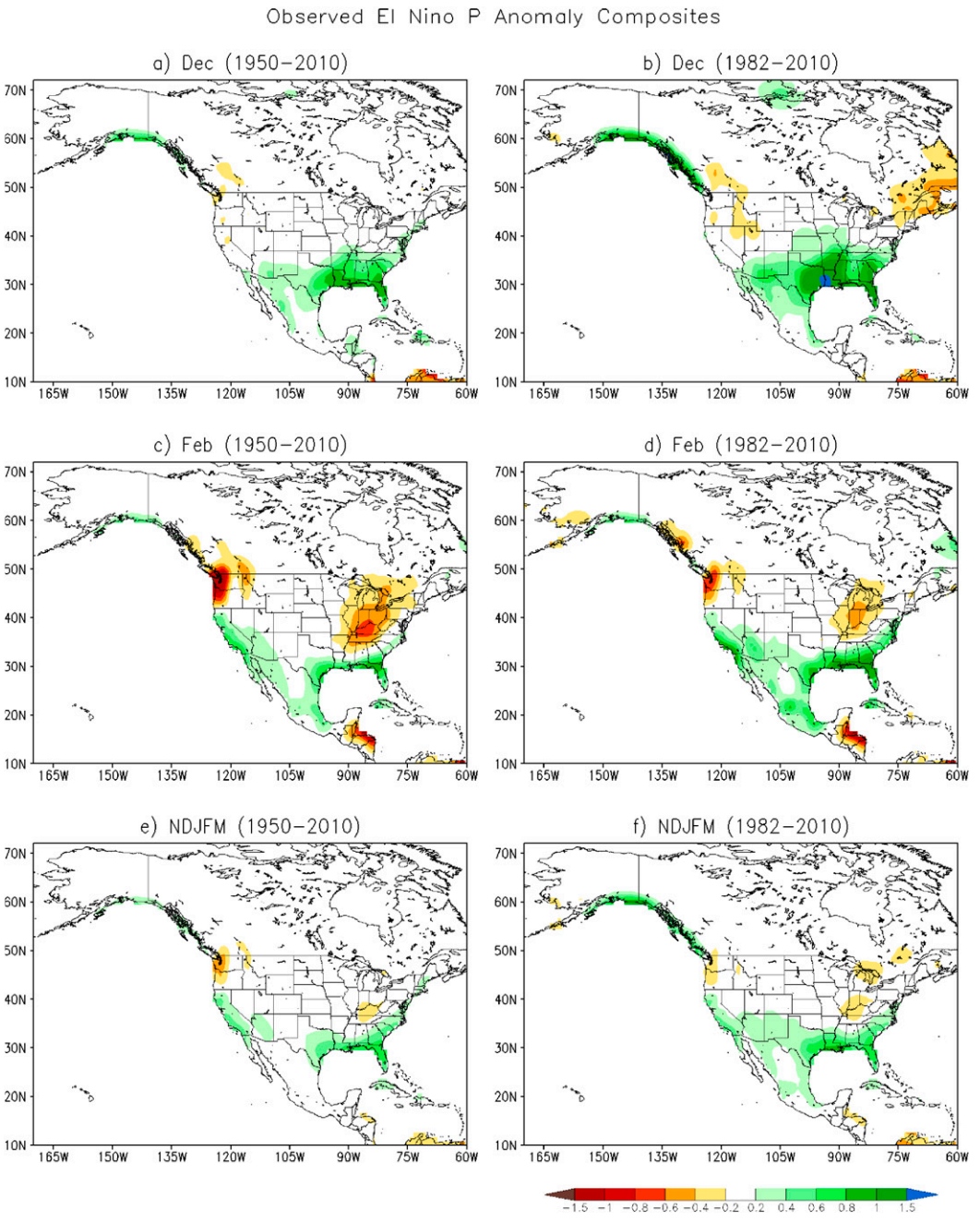

FIG. 2. El Niño precipitation anomaly (mm day^{-1}) composites based on (a) December 1950–2010, (b) December 1982–2010, (c) February 1950–2010, (d) February 1982–2010, (e) NDJFM 1950–2010, and (f) NDJFM 1982–2010 observations.

models. All models and the 1950–2010 observed composites present drier than normal conditions over the southern United States and enhanced rainfall over the Pacific Northwest, consistent with the pattern suggested by Ropelewski and Halpert (1986, 1987). The 1982–2010 observed NDJFM *P* anomaly composite also displays a similar La Niña pattern to the 1950–2010 observed. In contrast to the NMME and 1950–2010 observed composites, the 1982–2010 observed has below-normal rainfall over the Pacific Northwest. There are some variations among the six models but all models are reasonably good. CFSv2 has the biggest north–south contrast in the anomalies and its dry area is spread farther into central Mexico, while

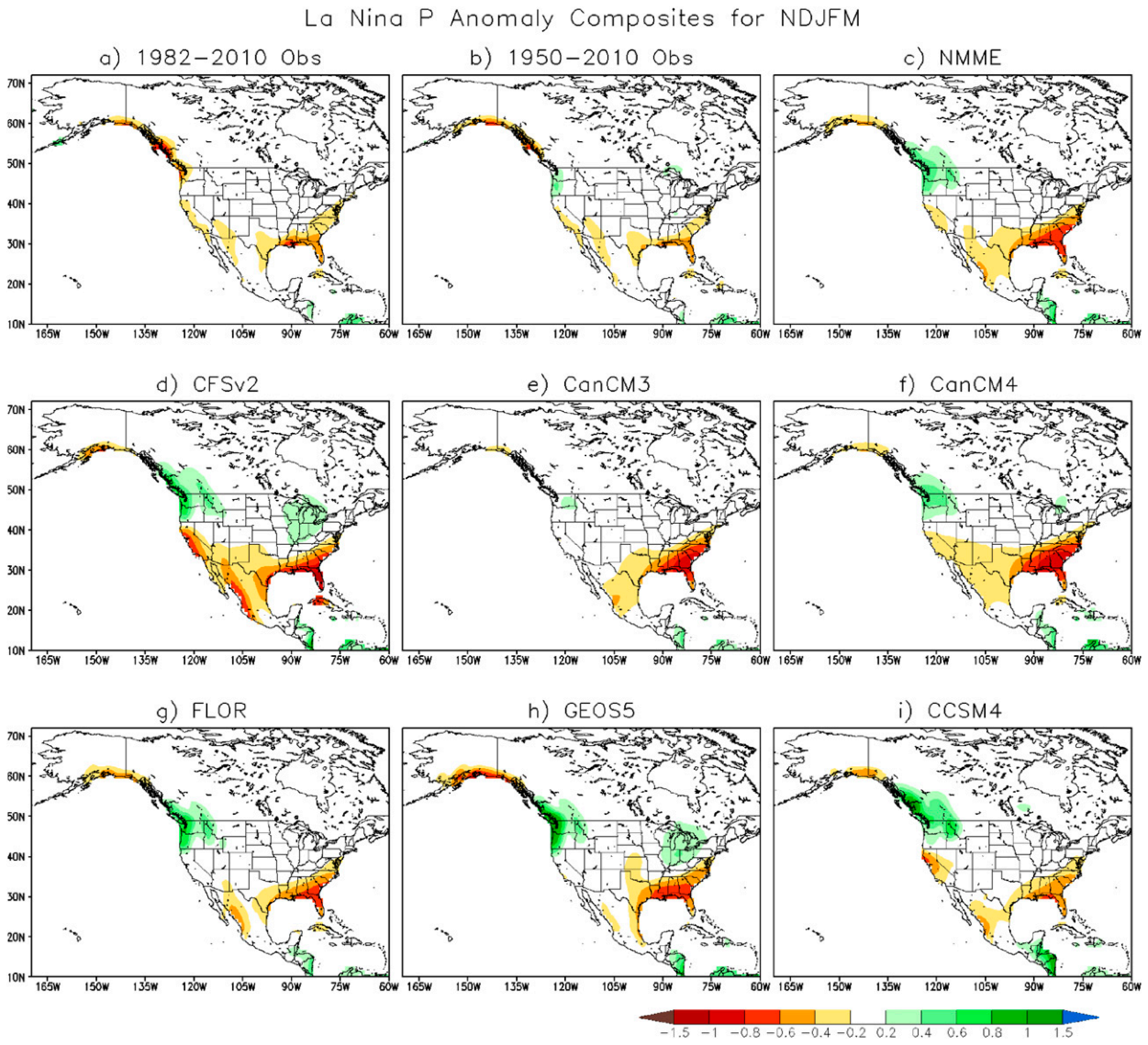

FIG. 3. La Niña precipitation anomaly (mm day^{-1}) composites for NDJFM based on (a) 1982–2010 observations, (b) 1950–2010 observations, (c) NMME, (d) CFSv2, (e) CanCM3, (f) CanCM4, (g) FLOR, (h) GEOS-5, and (i) CCSM4 forecasts over the North American continent.

both CanCM models produce a large negative deviation over the southeastern United States. Despite these subtle differences, the remarkable similarity between the NMME and observed *P* anomaly composites under both El Niño and La Niña conditions demonstrates the significant progress in ENSO–precipitation relationships from seasonal dynamical models since Smith and Ropelewski (1997).

Figure 4 presents the La Niña *T* anomaly composites for NDJFM based on 1982–2010 and 1950–2010 observations, NMME, and the six models. Unlike the *P* anomaly composites, there are major differences between the model and 1950–2010 observed composites. The differences are even greater when compared to the 1982–2010 observed composites. All six models feature large cold anomalies (in some cases exceeding 2°C) over Alaska and western Canada (oriented west–east rather than northwest–southeast as in the observed), allowing warm air to extend from the southeastern United States into central United States. In some models (e.g., the GEOS-5, CanCM4, and FLOR models), positive anomalies are seen over more than half of the United States, resulting in a large area of warming in the NMME composite, in contrast to the small warming area over the Gulf states and northeastern Mexico in the observed. It should be noted

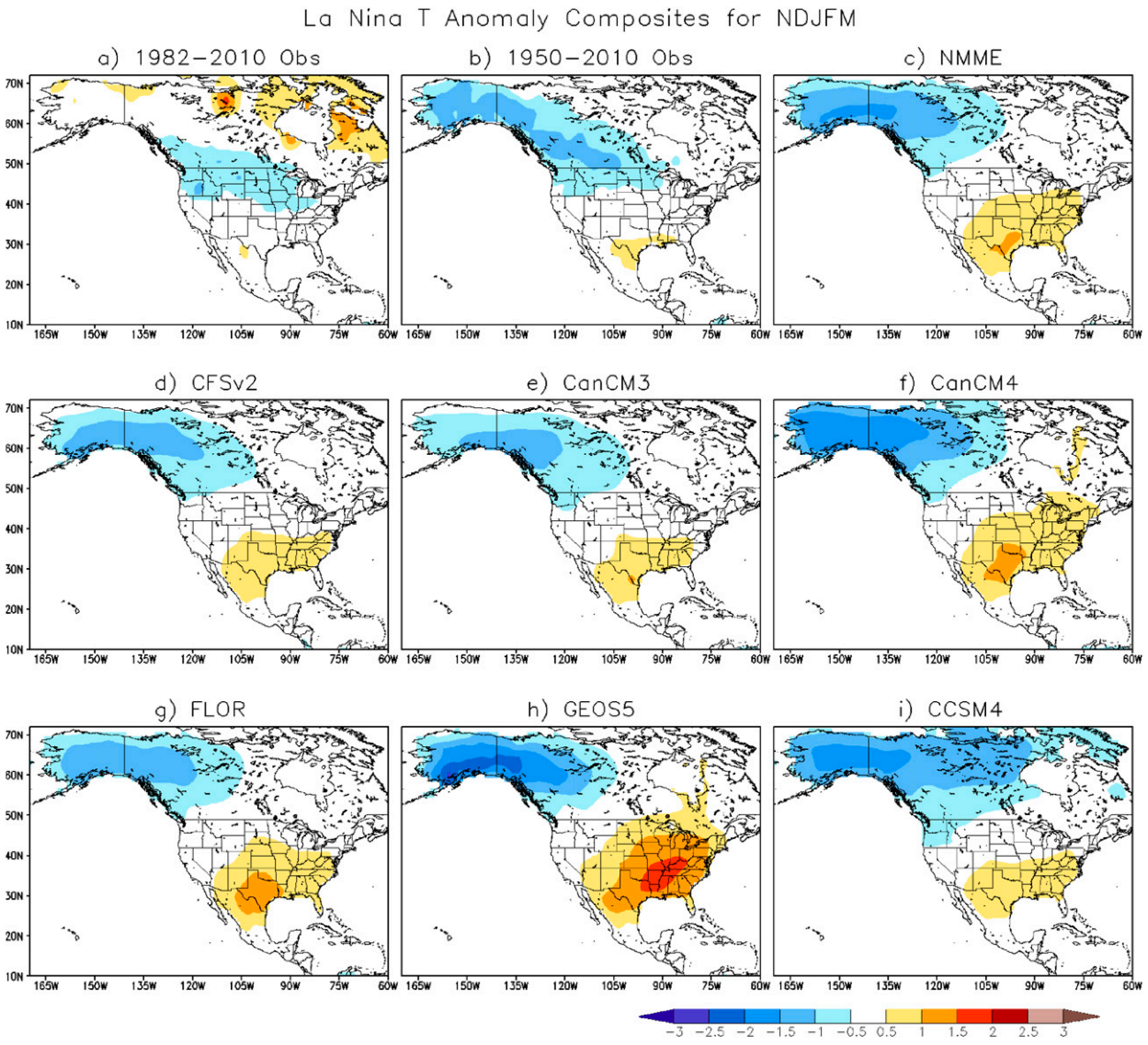

FIG. 4. As in Fig. 3, but for temperature anomaly (°C).

that because of different samples, the warming area in the 1982–2010 observed T anomaly composite is much less than what Ropelewski and Halpert (1986, 1987) found, which covered most of the southeastern United States.

To present a quantitative evaluation of how well NMME models predict P and T patterns under ENSO conditions, we compute the ACC and RMSE for P and T anomaly composites. Figure 5 shows the matrix charts of ACC for all models and months, including NMME and NDJFM, using the 1950–2010 observations for validation. ACC values greater than 0.2 are significantly different from zero at the 90% confidence level based on the Student's t test (Wilks 2011). In Fig. 5, matrix grids are shaded with green colors indicating the level of skill. For example, the ACC for the El Niño P anomaly composite of CFSv2 NDJFM (row 1, column 6 in Fig. 5a) is 0.81, shaded with the darkest green. Matrix charts are frequently used in climate ensemble evaluation (e.g., Gleckler et al. 2008) and biological sciences and statistical communities to identify the dominant factors among (or describe the relationships between) two or more groups of variables.

Several features are worth highlighting in Fig. 5. First, the fidelity is generally higher for NMME composites as well as for NDJFM composites. Second, predictive skill varies with month. All models, as well as NMME, have greater ACC for February prediction,

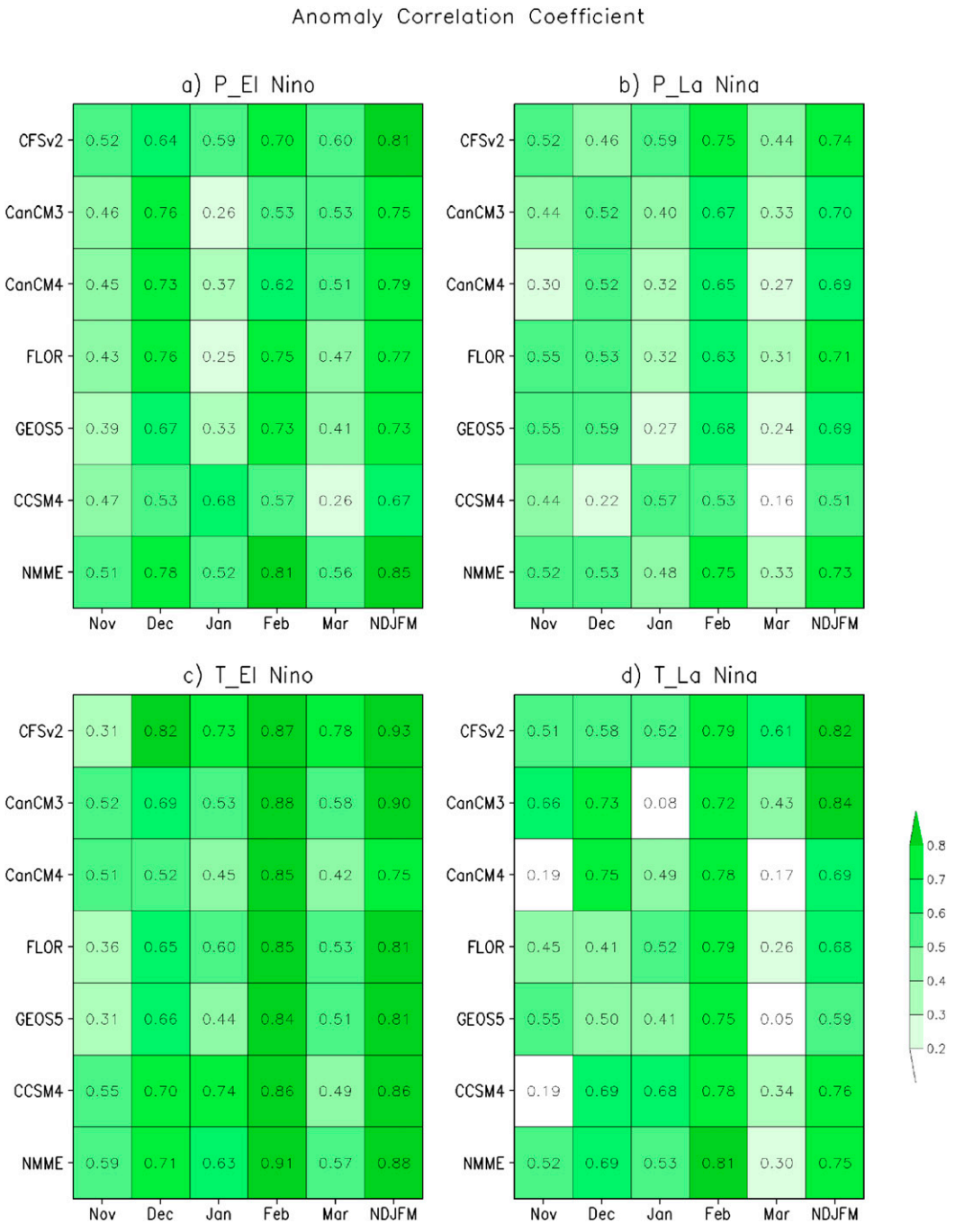

FIG. 5. ACC of all models and months for anomaly composites of (a) El Niño precipitation, (b) La Niña precipitation, (c) El Niño temperature, and (d) La Niña temperature, validated with 1950–2010 observations. Values >0.2 are significant at the 90% confidence level based on Student's t test. The level of green shading corresponds to the range of ACC values indicated by the color bar.

and this is seen for both P and T anomaly composites under either El Niño or La Niña condition. Third, for NDJFM composites, all models perform better in predicting El Niño P and T anomaly patterns than La Niña patterns. This result is consistent with the literature. The El Niño response is known to be stronger than the La Niña response (Frauen et al. 2014), and the higher the signal-to-noise ratio, the better the prediction skill in the way we measure skill. Fourth and last, based on the sample at hand, the

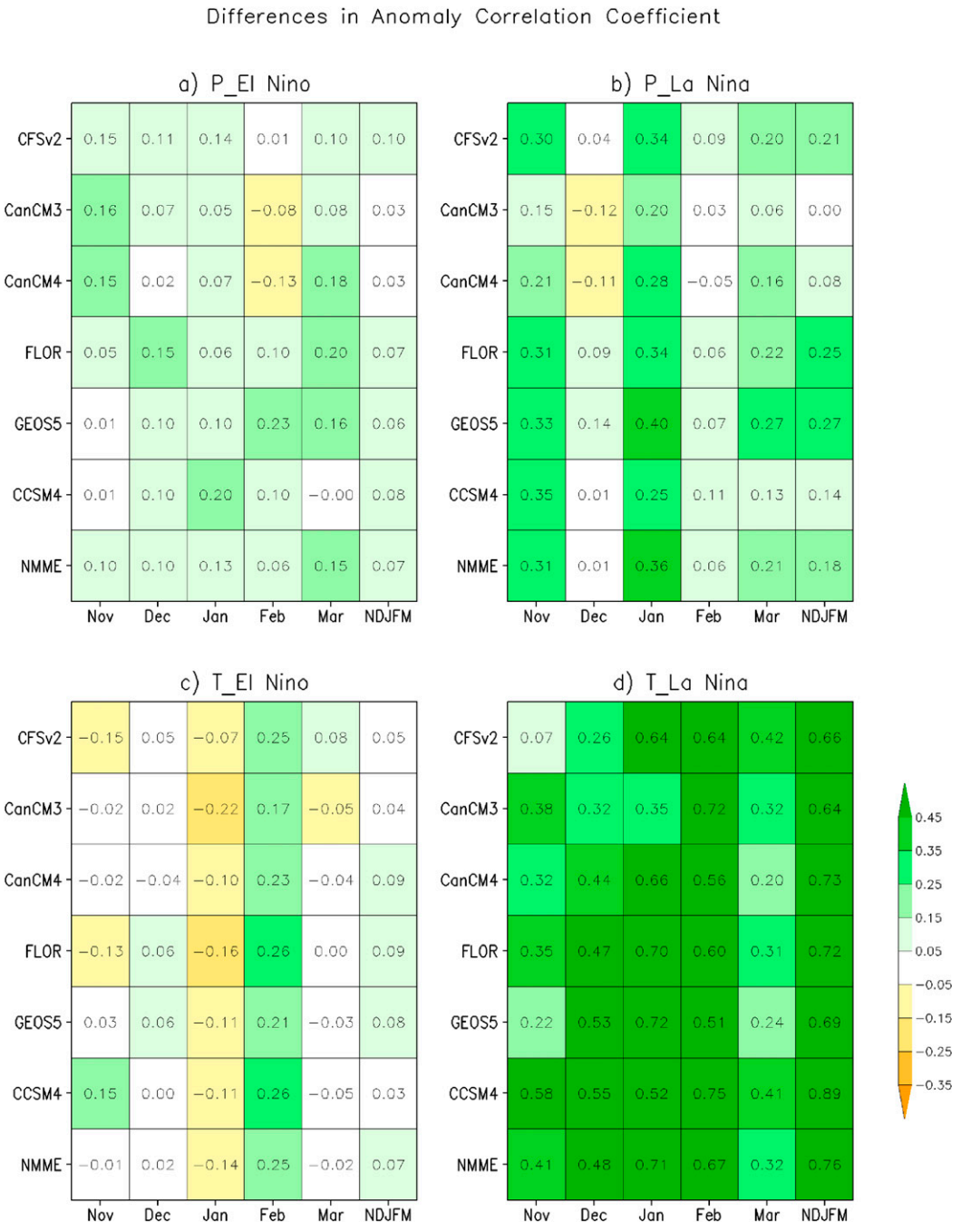

FIG. 6. As in Fig. 5, but for differences in ACC between validated with 1950–2010 observations and 1982–2010 observations. Values >0.2 or <−0.2 are significant at the 90% confidence level based on Fisher z test.

CFSv2 model did quite well, with CanCM3 as a close competitor for T anomaly composites.

The above findings also hold true for the validation with the 1982–2010 observed composites; however, their ACC scores are generally lower. The differences in ACC between validations with the 1950–2010 and 1982–2010 observations are shown in Fig. 6. Positive values (green colors) indicate greater ACC if validated with the 1950–2010 observations, and differences within the range from −0.2 to 0.2 are statistically insignificant based on Fisher z test (Wilks 2011) at the 90% confidence level. Since most numbers in Fig. 6 are positive, it is evident that model P and T anomaly composites correspond better with the

TABLE 2. RMSE for NMME anomaly composites for P (mm day^{-1}) and T (°C) of selected target months. Area of validation is the North American continent within the domain of 10°–72°N, 60°–170°W. The normalized RMSE for a given month (shown in parentheses) is the ratio of the RMSE to the observed standard deviation of a given variable (P or T) averaged over the North American continent. The corresponding ACC scores are shown in Fig. 5.

	P El Niño	P La Nina	T El Niño	T La Nina
RMSE (normalized RMSE) validated with 1982–2010 observations				
November	0.41 (0.37)	0.39 (0.35)	0.68 (0.22)	0.89 (0.29)
December	0.28 (0.27)	0.29 (0.28)	1.13 (0.32)	1.04 (0.29)
January	0.26 (0.27)	0.37 (0.39)	0.62 (0.18)	0.99 (0.29)
February	0.21 (0.22)	0.23 (0.24)	0.86 (0.25)	1.21 (0.35)
March	0.25 (0.31)	0.29 (0.36)	0.79 (0.28)	1.19 (0.42)
NDJFM	0.15 (0.15)	0.20 (0.20)	0.46 (0.14)	0.85 (0.26)
RMSE (normalized RMSE) validated with 1950–2010 observations				
November	0.29 (0.26)	0.30 (0.27)	0.43 (0.15)	0.54 (0.18)
December	0.17 (0.17)	0.24 (0.23)	0.53 (0.15)	0.61 (0.17)
January	0.24 (0.24)	0.25 (0.25)	0.69 (0.18)	0.68 (0.18)
February	0.17 (0.18)	0.18 (0.19)	0.50 (0.14)	0.57 (0.16)
March	0.18 (0.22)	0.24 (0.29)	0.80 (0.26)	0.95 (0.31)
NDJFM	0.11 (0.11)	0.15 (0.15)	0.38 (0.11)	0.45 (0.13)

1950–2010 observed composites, which have a larger sample size. The improvement in La Niña T anomaly composites is substantial. No model has skill in predicting La Niña T anomaly patterns if validated with the 1982–2010 observed composites, pointing to a major challenge in temperature forecast and verification under ENSO conditions that will be discussed more in section 8.

While ACC provides a measure of the linear association between the model and observation, RMSE is the overall accuracy metric. The assessment based on RMSE is consistent with the results from ACC. Table 2 lists the RMSE values of all target months for NMME composites validated with the 1982–2010 (top of Table 2) and 1950–2010 (bottom of Table 2) observed anomaly composites. Because RMSE has the same unit as the variable (mm day^{-1} for P anomaly and °C for T anomaly), RMSE from P and T composites cannot be directly compared. Therefore, we also calculate the normalized RMSE (shown in parentheses) by dividing RMSE by the observed standard deviation of a given variable (P or T) for a given month averaged over the North American continent. Similar to the ACC analysis, RMSE values vary with month and NMME has the lowest normalized RMSE in February for predicting ENSO P patterns regardless of the validation period. Under ENSO conditions, NDJFM composites have the best performance compared to any single month for both P and T anomaly composites. Performance is very poor for predicting La Niña T patterns when validated with the 1982–2010 observed composites (top of Table 2, last column), and its RMSE for February composite is the largest among all five winter months, contrary to those validated with the 1950–2010 observed and for P anomaly composites.

6. Probability composite analysis

Conventional atmospheric anomaly composites are derived as an arithmetic mean based on a selected sample under a specific condition and thus provide a mean state for that condition in physical units. The simplicity of this approach has made it widely used in many climatological studies to provide a typical pattern under a certain condition, such as El Niño or La Niña. However, arithmetic mean is strongly affected by outliers (large deviations) in the sample, especially when the sample size is small, and this situation is frequently encountered in ENSO composite analysis. To reduce the influence from outliers and take advantage of NMME's large ensemble size, we develop a new type of composites based on the probability of occurrence in a three-class forecast system commonly used in operational seasonal prediction (Higgins et al. 2004). The idea is to provide explicit information on the likelihood of a specific category (i.e., above, near, or below normal) to occur under ENSO conditions.

Figure 7 shows the El Niño P probability composites for NDJFM based on 1982–2010 and 1950–2010 observations, NMME, and the six models. In the maps, the above-normal shading (green) at a grid point is shown only when its probability is greater than 38% and the probability of below normal at the same location is lower than 33%. In contrast, below-normal shading (brown) is shown when its probability is greater than 38%, and the probability of above normal at the same

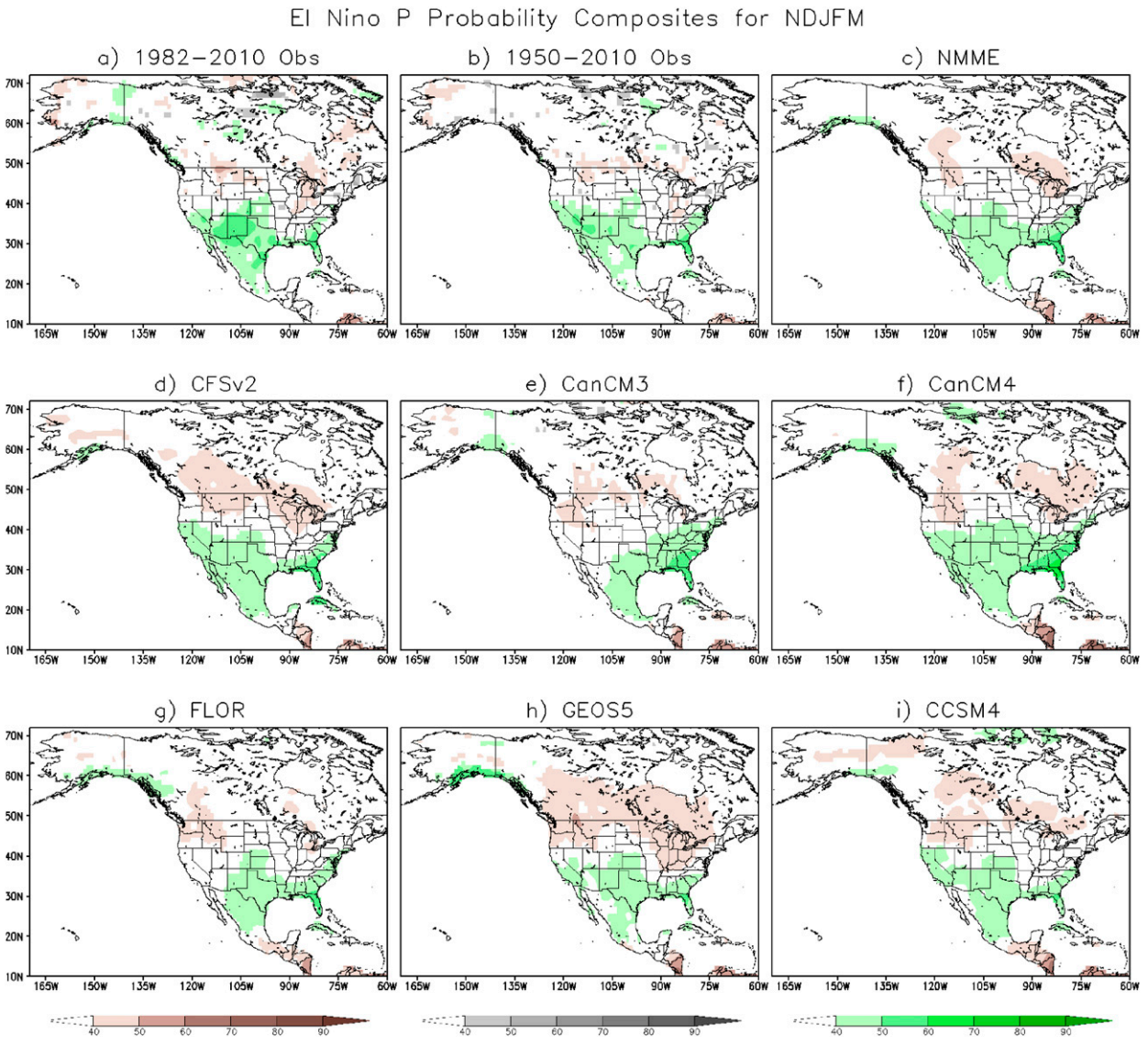

FIG. 7. El Niño precipitation probability composites for NDJFM based on (a) 1982–2010 observations, (b) 1950–2010 observations, (c) NMME, (d) CFSv2, (e) CanCM3, (f) CanCM4, (g) FLOR, (h) GEOS-5, and (i) CCSM4 forecasts over the North American continent. Brown, gray, and green colors indicate the probability of below-normal, near-normal, and above-normal categories, respectively. Forecast category displayed in model composites, where colors are shown, is significant at the 90% confidence level.

location is lower than 33%. Near-normal condition is shown when more than 38% of the counts fell into the neutral tercile and the probabilities of above normal and below normal are both less than 33%. When no class is dominant (either all categories are under 38% or both above and below normal are over 33%), no shading is shown. This set of rules for displaying probability composites is the same as that used for CPC's operational probabilistic forecasts. The 38% threshold, with an estimated margin of error of 5%, is placed to present a forecast category that is significant at the 90% confidence level in model composites.

Generally, *P* probability composites resemble similar spatial patterns to *P* anomaly composites (Figs. 1f, 2e,f), but the dry signal over the Pacific Northwest shifts more inland. This is because a probability composite implies a normalization so that a relatively large anomaly signal over the Pacific Northwest coast, where rainfall climatology and variability is high, is not as large in terms of probability. The spatial patterns of the 1982–2010 and 1950–2010 observed *P* probability composites are very much alike with slight differences in magnitude. As noted in the anomaly composite analysis, there are only small variations among the models. CFSv2 again

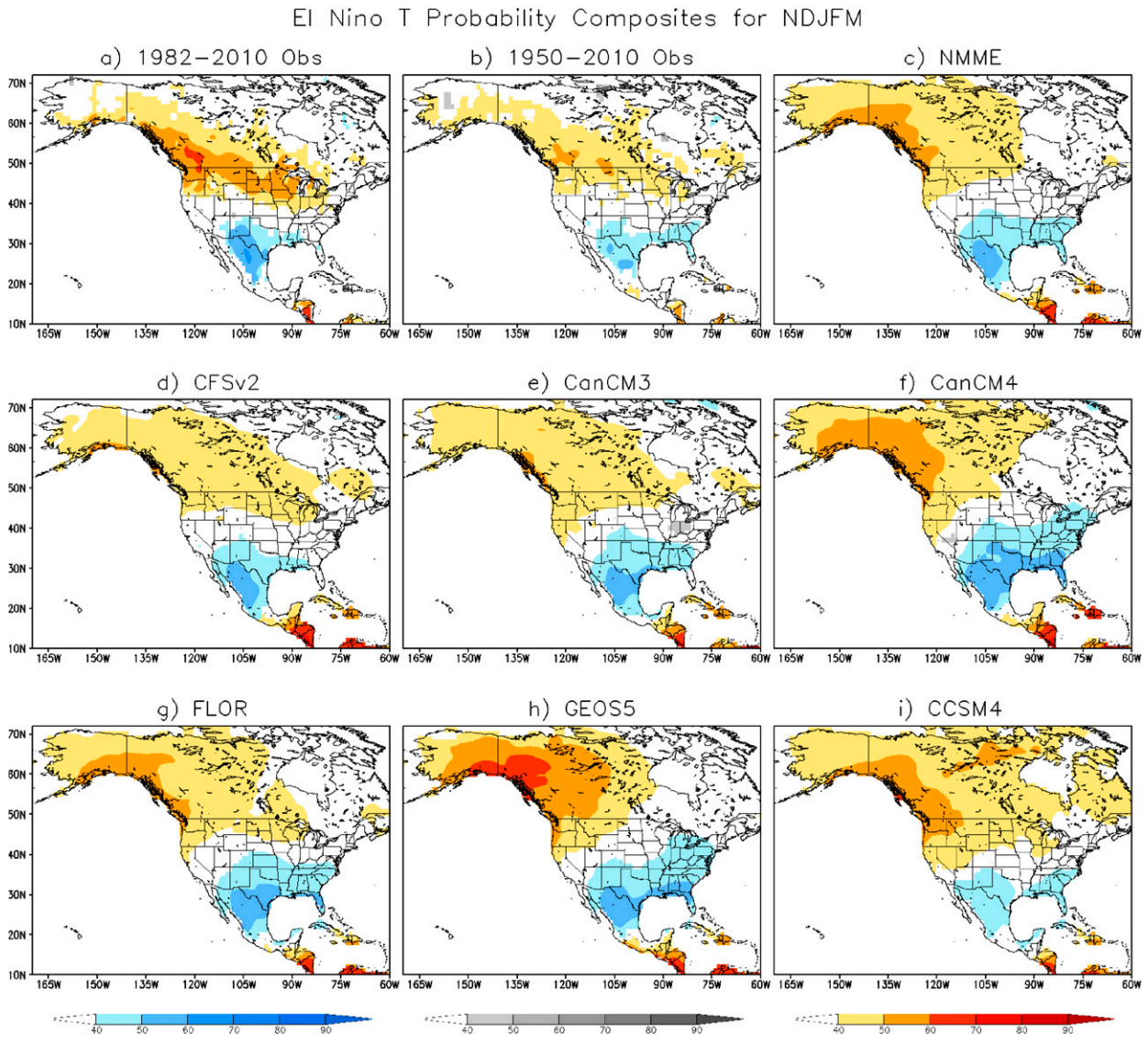

FIG. 8. As in Fig. 7, but for temperature. Here, blue, gray, and yellow-to-red colors indicate the probability of below-normal, near-normal, and above-normal categories, respectively.

generates spatial patterns most similar to the observed, while CanCM3 and FLOR models produce less rainfall over the southwestern United States, and CanCM4 overproduces wetness south of 40°N. Overall, NMME probability composite reproduces the wet–dry pattern and magnitude as seen in the 1950–2010 observed *P* probability composite.

Figure 8 presents the El Niño *T* probability composites for NDJFM based on 1982–2010 and 1950–2010 observations, NMME, and the six models. Unlike the observed *P* probability composites, there are larger differences between the 1982–2010 and 1950–2010 observed *T* probability composites. The 1982–2010 observed composite has a bigger warm–cold (north–south) contrast, and its below-normal area is centered over Texas and northern Mexico and does not cover the southeastern United States. Similar to the findings from the La Niña *T* anomaly composites (Fig. 4), *T* probability composites vary with model. Again, GEOS-5, CanCM4, and FLOR models have the largest deviations and are the main contributors to the difference between the NMME and observed probability composites.

For a formal validation, we compute the PAC and RMPS for all models and months. Figure 9 shows the matrix charts of PAC for *P* and *T* probability composites

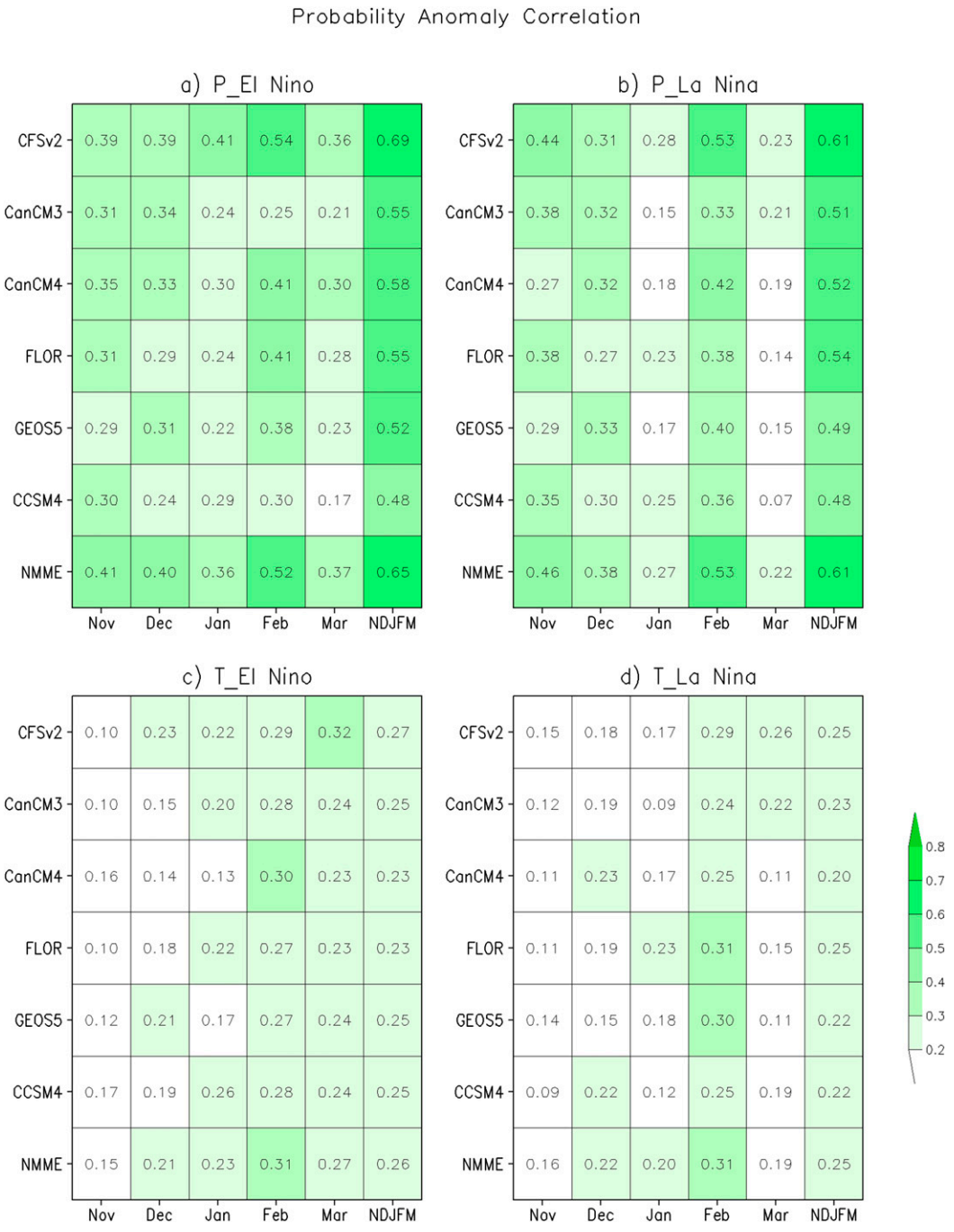

FIG. 9. PAC of all models and months for probability composites of (a) El Niño precipitation, (b) La Niña precipitation, (c) El Niño temperature, and (d) La Niña temperature, validated with 1950–2010 observations. Values >0.2 are significant at the 90% confidence level based on Student's t test. The level of green shading corresponds to the range of PAC values indicated by the color bar.

under El Niño and La Niña conditions. Consistent with the findings from the ACC analysis (Fig. 5), the fidelity for NMME and NDJFM composites is generally greater than that for individual models and months, although a particular model in a specific month may still outperform NMME prediction. Among all months, February tends to have higher scores than other months for both P and T probability composites under either El Niño or La Niña condition. Different from the ACC analysis, PAC is able to discriminate the

TABLE 3. RMPS for NMME probability composites of selected target months. Area of validation is the North American continent within the domain of 10°–72°N, 60°–170°W. Note that the corresponding PAC scores are shown in Fig. 9.

	P El Niño	*P* La Nina	*T* El Niño	*T* La Nina
RMPS validated with 1982–2010 observations				
November	0.163	0.160	0.211	0.212
December	0.149	0.147	0.222	0.205
January	0.131	0.161	0.197	0.214
February	0.135	0.142	0.215	0.209
March	0.144	0.150	0.207	0.216
NDJFM	0.074	0.078	0.170	0.175
RMPS validated with 1950–2010 observations				
November	0.118	0.110	0.184	0.182
December	0.104	0.111	0.183	0.180
January	0.101	0.115	0.183	0.183
February	0.098	0.106	0.185	0.186
March	0.099	0.111	0.185	0.196
NDJFM	0.053	0.058	0.166	0.164

performance between the *P* and *T* prediction more and shows larger scores for *P* composites than *T* composites under both El Niño and La Niña conditions.

The high predictive skill in February is also seen in the ACC analysis and can be explained by the steady-state linear response of the atmosphere to thermal forcing in the tropics (Hoskins and Karoly 1981). Opsteegh and Van den Dool (1980) found that the impact of tropical heating anomalies on the midlatitudes is achieved via Rossby wave propagation. Rossby waves excited by heating anomalies are trapped in the deep tropics if the upper-level winds are from the east. In the climatological annual cycle the upper-level westerly wind in the NH, conducive for Rossby wave propagation and typical for midlatitude and subtropics, push farther equatorward in late winter (January–February) than in any other season (Van den Dool 1983, his Fig. 2), although the precise reason for a favorable waveguide in February may be more complicated to describe in a realistic zonally varying basic state (Newman and Sardeshmukh 1998). In contrast, broad upper-level easterlies in NH summer and early fall reduce the potential for any direct impact of ENSO on the midlatitudes.

Table 3 provides the RMPS values of selected target months for NMME probability composites validated with the 1982–2010 (top of Table 3) and 1950–2010 (bottom of Table 3) observed. Since both *P* and *T* composites are expressed in probability terms, their RMPS values can be directly compared. Here, we can clearly see that NMME has higher performance in predicting *P* patterns than *T* patterns under both El Niño and La Niña conditions, and the NDJFM composite is more accurate than any single month composite, regardless of the validation period. However, because of the smaller sample size, each count is weighted more in the probability calculations and hence RMPS is constantly larger for the validation with the 1982–2010 observed probability composites. Different from the anomaly composite analysis, NMME has indistinguishable skill (in terms of probability accuracy) in predicting El Niño and La Niña patterns, for both *P* and *T* probability composites.

7. Sensitivity analysis

In the anomaly composite analysis, we have noticed some discrepancies between the 1982–2010 and 1950–2010 observed composites. The differences are mainly caused by the sample used to construct the composites. To examine how sensitive the validation is to the selected sample, we carry out a numerical experiment to illustrate the effects by removing one major El Niño episode and one major La Niña episode from the event list in Table 1. During the 1982–2010 period, the strongest El Niño event occurred in 1997/98, and the largest La Niña event happened in 1988/89. Therefore, we choose to delete these two biggest events from the list and then recompute the composites from both observations and model hindcasts following the same procedures described in section 3. After the composites are reconstructed, we recalculate the performance scores: ACC and RMSE for anomaly composites and PAC and RMPS for probability composites. Because the 1997/98 El Niño and 1988/89 La Niña events are not included in either model or observed composites, the new scores measure the performance from a new sample slightly different from the original one.

Figure 10 shows the differences in ACC (validated with 1982–2010 observations) after the two events were removed for both *P* and *T* anomaly composites under El Niño or La Niña condition. A positive number indicates an increase in score after the event was deleted and vice versa. There are clearly large differences after the 1997/98 El Niño and 1988/89 La Niña events were removed. For *P* anomaly composites, most models and months have lower ACC if the 1997/98 and 1988/89 events were not included in the composite analysis. CFSv2 and CCSM4 models have the greatest decrease (−0.21 for CFSv2 and −0.16 for CCSM4 February prediction) under El Niño condition. The influence is stronger for El Niño composites than La Niña composites. For *T* anomaly composites, the differences are even more pronounced. The changes can be as large as −0.30 for El Niño composites (CanCM4 March prediction) and −0.47 for La Niña composites (NMME and CanCM4 January prediction). Yet, some increases in

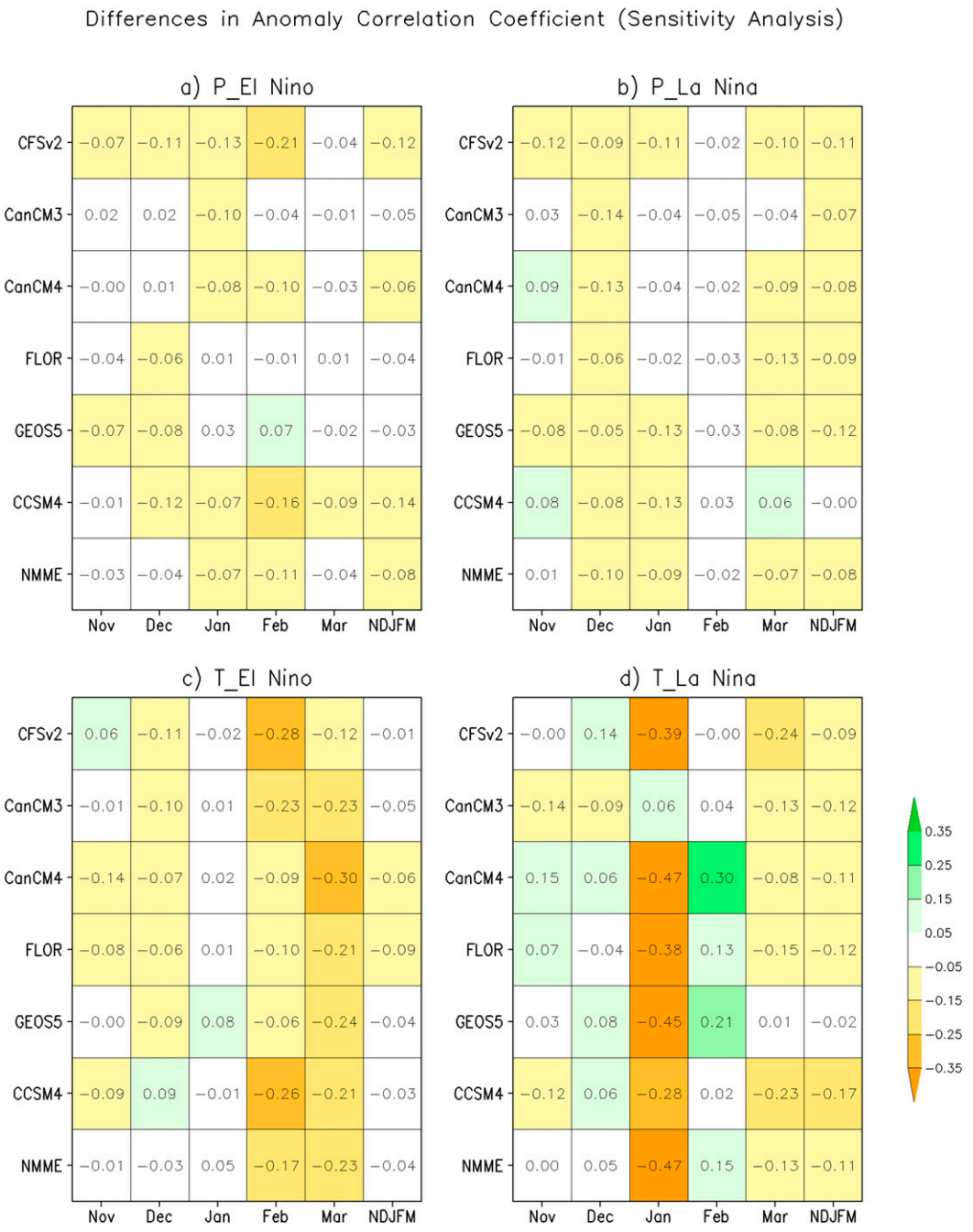

FIG. 10. Differences in ACC (validated with 1982–2010 observations) after the 1997/98 El Niño and 1988/89 La Niña events were removed from the composite analysis for anomaly composites of (a) El Niño precipitation, (b) La Niña precipitation, (c) El Niño temperature, and (d) La Niña temperature. Values >0.2 or <-0.2 are significant at the 90% confidence level based on Fisher z test.

ACC can be seen for December and February prediction under La Niña condition.

The same experiment is repeated for the validation with 1950–2010 observations. Similar to the above findings, most models and months show decreases in ACC for P and T anomaly composites after the two events were deleted, and the impact is greater for T composites than P composites. However, because of the larger sample size, the differences in ACC are not as big as those validated with the 1982–2010 observed anomaly composites. For P anomaly composites, the changes span from -0.17 (CCSM4 December prediction under

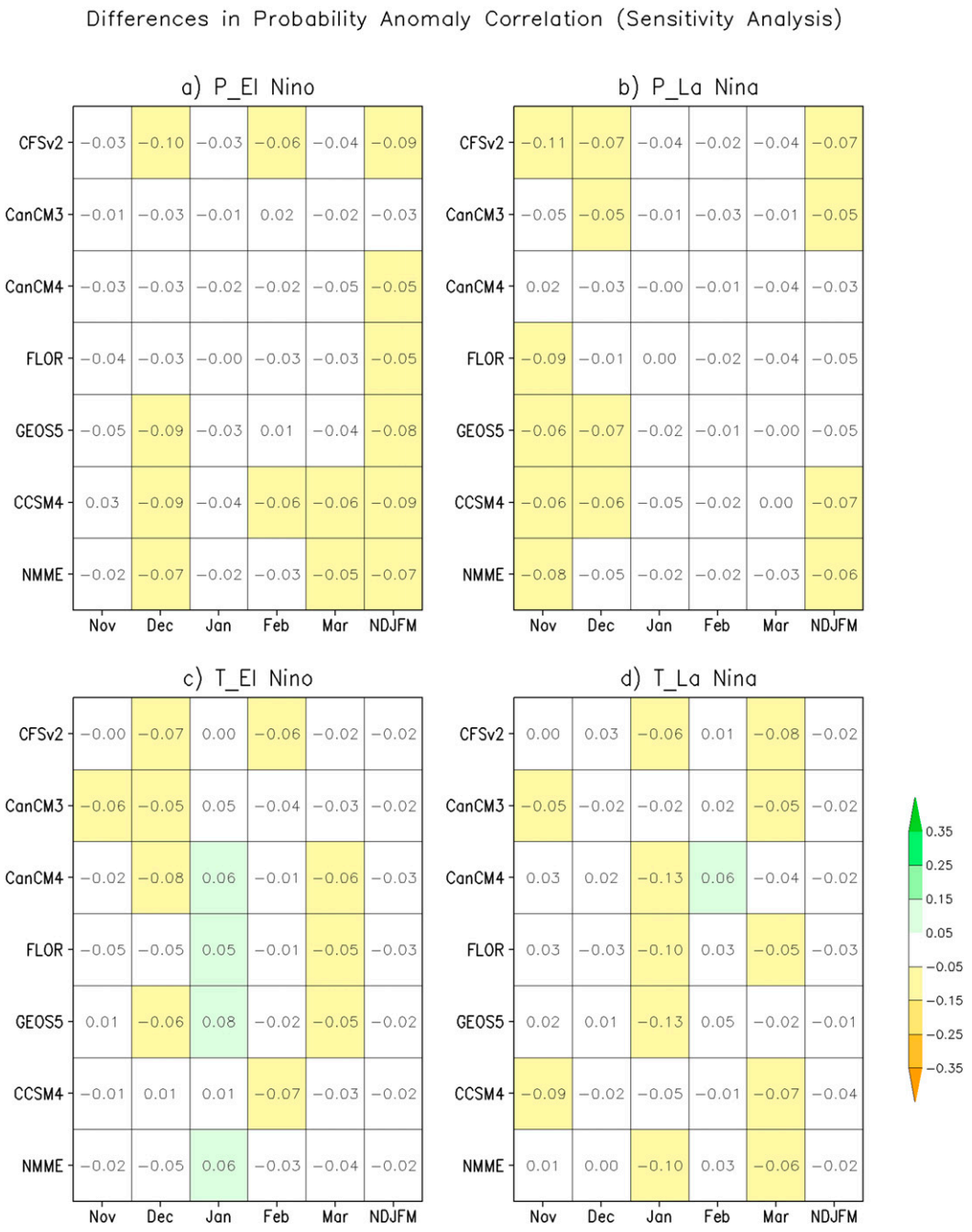

FIG. 11. As in Fig.10, but differences in PAC.

El Niño condition) to 0.10 (FLOR January prediction under La Niña condition). For T anomaly composites, ACC differences vary from −0.38 (CanCM3 January prediction) to 0.06 (GEOS-5 February prediction) under La Niña condition. This result demonstrates the importance of sample size for the ENSO validation study. When sample size is small, performance assessment based on anomaly composites is largely influenced by strong ENSO events.

The sensitivity analysis is also carried out for probability composite validation. Figure 11 presents the differences in PAC (validated with 1982–2010 observations) after the two events were removed for both P and T probability composites under El Niño or La Niña condition. In contrast to the results from the anomaly composites (Fig. 10), the differences in PAC for the probability composites are small: within −0.13 and 0.08 for all cases. The differences in PAC when validated with 1950–2010 observed probability

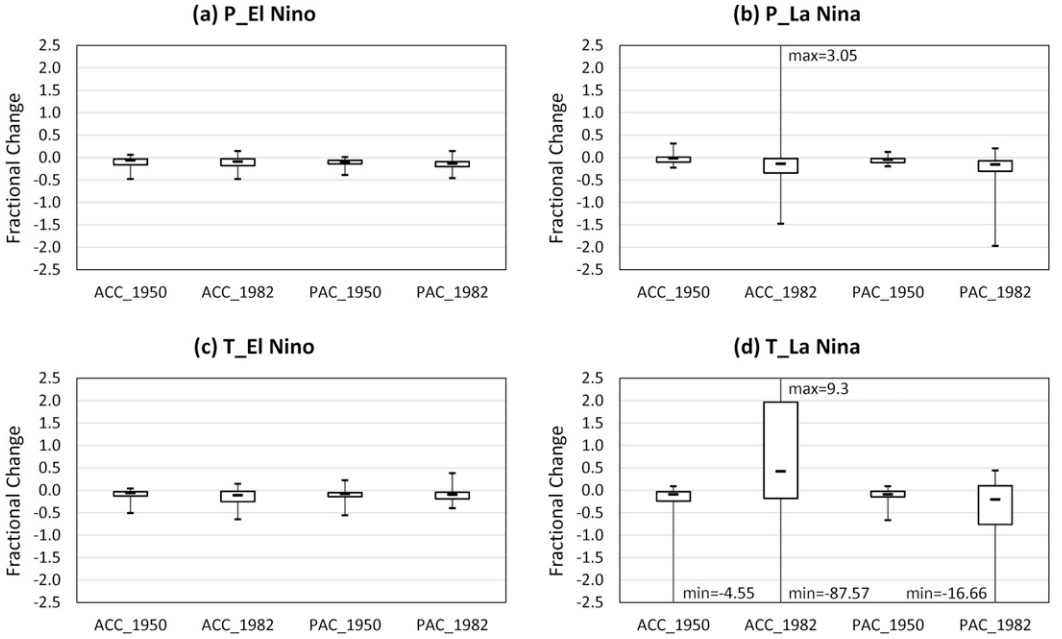

FIG. 12. Box-and-whisker plots of fractional change after the 1997/98 El Niño and 1988/89 La Niña events were removed from the composite analysis for composites of (a) El Niño precipitation, (b) La Niña precipitation, (c) El Niño temperature, and (d) La Niña temperature. The number after the underscore with ACC or PAC on the *x* axes indicates the starting year of the validation period. Indicated for each box are the median (horizontal line through the box middle), the 25th and 75th percentiles (top and bottom edges of the box), and the minimum and maximum (lower and upper ends of the vertical whisker line).

composites are even smaller. In fact, the differences in PAC (validated with 1950–2010 observations) are within the range from −0.05 to 0.05 for most models and months, except for December and NDJFM *P* probability composites under El Niño condition and a few others.

To have a level comparison between ACC and PAC, we calculate the fractional changes, defined as $(AC_{removed} - AC_{original})/AC_{original}$, where AC is ACC or PAC, for the four sets of experiments. Their box-and-whisker plots are displayed in Fig. 12. Each box and whisker represents the distribution of the 42 combinations from six choices of month (including NDJFM) and seven choices of model (including NMME) for each panel in the matrix charts (Figs. 5, 6, and 9–11). Indicated for each box are the median (horizontal line through the box middle), the 25th and 75th percentiles (top and bottom edges of the box), and the minimum and maximum (lower and upper ends of the vertical whisker line). It is clearly seen that the range of fractional changes for PAC usually is smaller than that for ACC when validated with the same period of observations. For PAC, the fractional change after removing a major ENSO episode is always smaller for validation with 1950–2010 observations than that validated with 1982–2010 observations. This result suggests that the probability composite is less sensitive to the particular sample used to construct the composite and thus gives a more robust estimate of the true ENSO impact. Because of that, sample size is a more critical factor for probability composite validation than the sample period.

In addition to the above advantage, there are several benefits of probability composites. First, they naturally unify *P* and *T* composites through the use of probability (0–1) as a unit. Second, they directly provide probability distribution information for three category outcomes (as used in CPC's operational seasonal prediction). Third, by using the tercile thresholds, each count is treated and contributed equally and thus the effect of outliers is reduced. Fourth, because both model and observed composites are derived with respect to their own distributions, we bypass the question of whether the model and observation have the same distribution. In cases when a model cannot reproduce the distribution as the observed, the probability composite provides a better depiction of the possible deviations closer to the observed. These advantages indicate that the probability composite is a far more robust and effective tool than the anomaly composite for describing and predicting ENSO impacts over the North American continent.

8. Discussion

In previous sections, we have illustrated one major challenge in ENSO validation study—limited observations!

This situation becomes problematic when there is significant contrast in the sample size of the large ensemble prediction (such as NMME) and the single verifying quantity. The validation of model-based ENSO composites, although based on 1982–2010 hindcast data, fares better, by all measures, against observed ENSO composites if the latter are based on as many years as possible. While there may be some inherent merit in using matching years, that merit is outweighed by the lack of sample in the observations.

By using the composite approach, we implicitly assume that the sample is drawn from the same population invariant in time, and thus the larger the sample size, the more stable the mean is. Under this assumption, it is justifiable to use ENSO events from a longer period of time to derive a more stable observed composite (climatology) for validation. This strategy works well for variables that meet the requirement, such as precipitation that has marginal climatic changes over the 1950–2010 period, as seen in Fig. 2. However, for nonstationary variables, such as temperature, this strategy may be questionable. Smith and Ropelewski (1997) did not provide an assessment on ENSO–temperature relationships in the climate model, and to our knowledge we are the first to attempt such evaluation in multimodel ensemble forecasts.

We have noticed greater differences between the 1982–2010 and 1950–2010 observed T composites, especially for La Niña patterns. One factor contributing to the differences are the strong outliers that occurred within the 1982–2010 period. Another factor is the global warming effect. Several studies (e.g., Collins et al. 2010; Bayr et al. 2014) have proposed a theory on possible influences on ENSO due to global warming. Limited by observations, its actual effects remain unknown. Among the six NMME models, only a few models (e.g., CFSv2) have displayed some warming trends in their temperature forecasts but not as large as the observed. How climate models simulate this trend and its effects on ENSO is beyond the scope of this paper and requires further investigations. On top of that, how to combine and adjust model forecasts with diverse trends (and no trend) is a challenging topic. We conduct the validation without any adjustments to the temperature forecasts and observations as the first step to understand the models' ability in predicting ENSO impacts. We hope our study will inspire more research on nonstationarity in multimodel ensemble forecasts.

In spite of the focus on ENSO model composites here, we do NOT suggest that model forecasts should be replaced by ENSO composites in years when a warm or cold event is in progress. Neither do we suggest that observed ENSO composites are the best a model can do. There may be legitimate case-to-case variations in ENSO (flavors of ENSO), and models may attempt to include other conditions that apply only to the year in question. One thing one can learn from a large (model) ensemble is that there are considerable variations from a composite based on ensemble member j versus ensemble member k. This adds a note of caution in the use of observed composites for seasonal prediction, which are based on a single realization.

The similarity of model and observed composites (broadly speaking and precipitation in particular) does suggest that models are quite good at simulating teleconnections (the response to ENSO over the United States is an obvious teleconnection). To the extent that teleconnections can be explained from the dispersion of Rossby waves, this should have been expected. However, errors in the mean state and misplacement of the jet stream can cause Rossby wave trains to take different routes (Hoskins and Karoly 1981). The results are thus encouraging.

9. Summary and conclusions

We have compared and validated precipitation and temperature forecasts under ENSO conditions in six NMME models with long-term climate observations. Our aim is to understand whether coupled climate models can adequately predict ENSO's impacts on North American precipitation and temperature patterns while an El Niño or La Niña event is in progress. We focus on the overall model performance and provide a comprehensive analysis and validation of both the anomaly and probability composites constructed from selected warm or cold ENSO episodes based on the tropical Pacific Ocean conditions during the Northern Hemisphere winter season. The key findings from the study are summarized below. These findings are robust regardless of the validation period or the type of composite used in the analysis:

- NMME predicts ENSO precipitation patterns well during wintertime. All models are reasonably good. CFSv2 performs particularly well. This result gives us confidence in NMME precipitation forecasts during an ENSO episode and the models' ability in simulating teleconnections.
- There are some discrepancies between the NMME and observed composites for temperature forecasts in terms of both magnitude and spatial distribution. The differences are mainly contributed by the GEOS-5, CanCM4, and FLOR models, and thus the NMME aggregates have difficulties in reproducing the ENSO–temperature relationships.
- For all ENSO precipitation and temperature composites, the fidelity is greater for the multimodel ensemble

as well as for the 5-month aggregates. February tends to have higher performance scores than other winter months.
- For anomaly composites, most models perform slightly better in predicting El Niño patterns than La Niña patterns.
- For probability composites, all models have superior performance in predicting ENSO precipitation patterns than temperature patterns.
- Compared to the anomaly composite, the probability composite is less sensitive to the particular sample used to construct the composite and has several advantages, suggesting that probability composite is a more robust and effective tool for describing and predicting ENSO's impacts over the North American continent.

Our findings are encouraging. We have demonstrated the progress of ENSO precipitation forecasts made in atmospheric models since Smith and Ropelewski (1997) and yet identified some deficiencies in temperature forecasts in the current NMME models. We hope this study will inspire more research to improve our understanding on how ENSO is simulated in climate models and lead to model enhancement, advanced ensemble techniques, and better forecasts. In addition to the above findings, we have developed two new performance metrics, PAC and RMPS, for verifying probabilistic forecasts when both prediction and observation are expressed in probability terms. These metrics can also be applied to validate ensemble prediction systems when observational errors (or uncertainty) are taken into consideration. We have also produced global anomaly and probability composites using the described methodology. The complete set of ENSO composites for all models and months (including all the figures not shown in this manuscript), along with the global composites, are available on the CPC NMME website (at http://www.cpc.ncep.noaa.gov/products/NMME/enso/).

Acknowledgments. This research was supported by the North American Multimodel Ensemble (NMME) project, a multiagency and multi-institutional research effort led by NOAA National Weather Service (NWS) Climate Test Bed (CTB) and Climate Program Office (CPO) Modeling, Analysis, Predictions, and Projections (MAPP) program in partnership with DOE, NSF, and NASA, under NOAA Grant NA14OAR4310188 to Huug van den Dool and Grant NA14NES4320003 [Cooperative Institute for Climate and Satellites (CICS)] at the University of Maryland/ESSIC. We greatly appreciate the editor and three anonymous reviewers for their thorough comments and suggestions to help improve the manuscript.

REFERENCES

Barnston, A. G., A. Leetmaa, V. E. Kousky, R. E. Livezey, E. A. O'Lenic, H. van den Dool, A. J. Wagner, and D. A. Unger, 1999: NCEP forecasts of the El Nino of 1997–98 and its impacts. *Bull. Amer. Meteor. Soc.*, **80**, 1829–1852, doi:10.1175/1520-0477(1999)080<1829:NFOTEN>2.0.CO;2.

Bayr, T., D. Dommenget, T. Martin, and S. Power, 2014: The eastward shift of the Walker circulation in response to global warming and its relationship to ENSO variability. *Climate Dyn.*, **43**, 2747–2763, doi:10.1007/s00382-014-2091-y.

Becker, E., H. van den Dool, and Q. Zhang, 2014: Predictability and forecast skill in NMME. *J. Climate*, **27**, 5891–5906, doi:10.1175/JCLI-D-13-00597.1.

Bell, G. D., M. S. Halpert, V. E. Kousky, M. E. Gelman, C. F. Ropelewski, A. V. Douglas, and R. C. Schnell, 1999: Climate assessment for 1998. *Bull. Amer. Meteor. Soc.*, **80**, 1040, doi:10.1175/1520-0477(1999)080<1040:CAF>2.0.CO;2.

Candille, G., and O. Talagrand, 2008: Impact of observational error on the validation of ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **134**, 959–971, doi:10.1002/qj.268.

Chen, M., P. Xie, J. E. Janowiak, and P. A. Arkin, 2002: Global land precipitation: A 50-yr monthly analysis based on gauge observations. *J. Hydrometeor.*, **3**, 249–266, doi:10.1175/1525-7541(2002)003<0249:GLPAYM>2.0.CO;2.

Collins, M., and Coauthors, 2010: The impact of global warming on the tropical Pacific Ocean and El Niño. *Nat. Geosci.*, **3**, 391–397, doi:10.1038/ngeo868.

Dai, A., and T. M. L. Wigley, 2000: Global patterns of ENSO-induced precipitation. *Geophys. Res. Lett.*, **27**, 1283–1286, doi:10.1029/1999GL011140.

Danabasoglu, G., S. C. Bates, B. P. Briegleb, S. R. Jayne, M. Jochum, W. G. Large, S. Peacock, and S. G. Yeager, 2012: The CCSM4 ocean component. *J. Climate*, **25**, 1361–1389, doi:10.1175/JCLI-D-11-00091.1.

Fan, Y., and H. van den Dool, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.*, **113**, D01103, doi:10.1029/2007JD008470.

Frauen, C., D. Dommenget, N. Tyrrell, M. Rezny, and S. Wales, 2014: Analysis of the nonlinearity of El Niño–Southern Oscillation teleconnections. *J. Climate*, **27**, 6225–6244, doi:10.1175/JCLI-D-13-00757.1.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, doi:10.1029/2007JD008972.

Graham, R. J., A. Evans, K. Mylne, M. Harrison, and K. Robertson, 2000: An assessment of seasonal predictability using atmospheric general circulation models. *Quart. J. Roy. Meteor. Soc.*, **126**, 2211–2240, doi:10.1256/smsqj.56711.

Gutman, G., I. Csiszar, and P. Romanov, 2000: Using NOAA/AVHRR products to monitor El Niño impacts: Focus on Indonesia in 1997–98. *Bull. Amer. Meteor. Soc.*, **81**, 1189–1205, doi:10.1175/1520-0477(2000)081<1189:UNPTME>2.3.CO;2.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus*, **57A**, 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Higgins, R. W., H.-K. Kim, and D. Unger, 2004: Long-lead seasonal temperature and precipitation prediction using tropical Pacific SST consolidation forecasts. *J. Climate*, **17**, 3398–3414, doi:10.1175/1520-0442(2004)017<3398:LSTAPP>2.0.CO;2.

Hoskins, B. J., and D. J. Karoly, 1981: The steady linear response of a spherical atmosphere to thermal and orographic forcing.

J. Atmos. Sci., **38**, 1179–1196, doi:10.1175/1520-0469(1981)038<1179:TSLROA>2.0.CO;2.

Ji, M., A. Kumar, and A. Leetmaa, 1994: An experimental coupled forecast system at the National Meteorological Center. *Tellus*, **46A**, 398–418, doi:10.1034/j.1600-0870.1994.t01-3-00006.x.

Jia, L., and Coauthors, 2015: Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model. *J. Climate*, **28**, 2044–2064, doi:10.1175/JCLI-D-14-00112.1.

Kiladis, G., and H. Diaz, 1989: Global climatic anomalies associated with extremes in the Southern Oscillation. *J. Climate*, **2**, 1069–1090, doi:10.1175/1520-0442(1989)002<1069:GCAAWE>2.0.CO;2.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble (NMME): Phase-1 seasonal to interannual prediction, phase-2 toward developing intra-seasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, doi:10.1175/BAMS-D-12-00050.1.

Kousky, V. E., and R. W. Higgins, 2007: An alert classification system for monitoring and assessing the ENSO cycle. *Wea. Forecasting*, **22**, 353–371, doi:10.1175/WAF987.1.

Kumar, A., M. Hoerling, M. Leetmaa, and P. Sardeshmukh, 1996: Assessing a GCM's suitability for making seasonal predictions. *J. Climate*, **9**, 115–129, doi:10.1175/1520-0442(1996)009<0115:AAGSFM>2.0.CO;2.

Lau, K.-M., and H. Weng, 2001: Coherent modes of global SST and summer rainfall over China: An assessment of the regional impacts of the 1997–98 El Niño. *J. Climate*, **14**, 1294–1308, doi:10.1175/1520-0442(2001)014<1294:CMOGSA>2.0.CO;2.

Lyon, B., and S. J. Mason, 2007: The 1997–98 summer rainfall season in southern Africa. Part I: Observations. *J. Climate*, **20**, 5134–5148, doi:10.1175/JCLI4225.1.

Mathieu, P.-P., R. T. Sutton, B. Dong, and M. Collins, 2004: Predictability of winter climate over the North Atlantic European region during ENSO events. *J. Climate*, **17**, 1953–1974, doi:10.1175/1520-0442(2004)017<1953:POWCOT>2.0.CO;2.

Merryfield, W. J., and Coauthors, 2013: The Canadian Seasonal to Interannual Prediction System. Part I: Models and initialization. *Mon. Wea. Rev.*, **141**, 2910–2945, doi:10.1175/MWR-D-12-00216.1.

Newman, M., and P. D. Sardeshmukh, 1998: The impact of the annual cycle on the North Pacific/North American response to remote low-frequency forcing. *J. Atmos. Sci.*, **55**, 1336–1353, doi:10.1175/1520-0469(1998)055<1336:TIOTAC>2.0.CO;2.

NRC, 2010: *Assessment of Intraseasonal to Interannual Climate Prediction and Predictability*. National Academies Press, 181 pp.

Opsteegh, J. D., and H. M. van den Dool, 1980: Seasonal differences in the stationary response of a linearized primitive equation model: Prospects for long-range weather forecasting? *J. Atmos. Sci.*, **37**, 2169–2185, doi:10.1175/1520-0469(1980)037<2169:SDITSR>2.0.CO;2.

Parameswaran, K., S. K. Nair, and K. Rajeev, 2004: Impact of Indonesian forest fires during the 1997 El Nino on the aerosol distribution over the Indian Ocean. *Adv. Space Res.*, **33**, 1098–1103, doi:10.1016/S0273-1177(03)00736-1.

Persson, P. O., P. J. Neiman, B. Walter, J.-W. Bao, and F. M. Ralph, 2005: Contributions from California coastal-zone surface fluxes to heavy coastal precipitation: A CALJET case study during the strong El Niño of 1998. *Mon. Wea. Rev.*, **133**, 1175–1198, doi:10.1175/MWR2910.1.

Ropelewski, C. F., and M. S. Halpert, 1986: North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation. *Mon. Wea. Rev.*, **114**, 2352–2362, doi:10.1175/1520-0493(1986)114<2352:NAPATP>2.0.CO;2.

——, and ——, 1987: Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Wea. Rev.*, **115**, 1606–1625, doi:10.1175/1520-0493(1987)115<1606:GARSPP>2.0.CO;2.

Rowell, D. P., 1998: Assessing potential seasonal predictability with an ensemble of multidecadal GCM simulations. *J. Climate*, **11**, 109–120, doi:10.1175/1520-0442(1998)011<0109:APSPWA>2.0.CO;2.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate*, **19**, 3483–3517, doi:10.1175/JCLI3812.1.

——, and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, doi:10.1175/JCLI-D-12-00823.1.

Shukla, J., D. Paoline, D. Straus, D. DeWitt, M. Fennessy, J. Kinter, L. Marx, and R. Mo, 2000: Dynamical seasonal predictions with the COLA atmospheric model. *Quart. J. Roy. Meteor. Soc.*, **126**, 2265–2299, doi:10.1256/smsqj.56713.

Smith, T. M., and C. F. Ropelewski, 1997: Quantifying Southern Oscillation–precipitation relationships from an atmospheric GCM. *J. Climate*, **10**, 2277–2284, doi:10.1175/1520-0442(1997)010<2277:QSOPRF>2.0.CO;2.

Trenberth, K. E., G. Branstator, D. Karoly, A. Kumar, N. Lau, and C. Ropelewski, 1998: Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.*, **103**, 14 291–14 324, doi:10.1029/97JC01444.

Van den Dool, H. M., 1983: A possible explanation of the observed persistence of monthly mean circulation anomalies. *Mon. Wea. Rev.*, **111**, 539–544, doi:10.1175/1520-0493(1983)111<0539:APEOTO>2.0.CO;2.

——, 2007: *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, 215 pp.

Vecchi, G. A., and Coauthors, 2014: On the seasonal forecasting of regional tropical cyclone activity. *J. Climate*, **27**, 7994–8016, doi:10.1175/JCLI-D-14-00158.1.

Vernieres, G., M. M. Rienecker, R. Kovach, and C. L. Keppenne, 2012: The GEOS-iODAS: Description and evaluation. NASA Tech. Rep. Series on Global Modeling and Data Assimilation NASA/TM-2012-104606/Vol. 30, 73 pp. [Available online at https://gmao.gsfc.nasa.gov/pubs/docs/Vernieres589.pdf.]

Wang, X., and S. S. P. Shen, 1999: Estimation of spatial degrees of freedom of a climate field. *J. Climate*, **12**, 1280–1291, doi:10.1175/1520-0442(1999)012<1280:EOSDOF>2.0.CO;2.

Weisheimer, A., and Coauthors, 2009: ENSEMBLES: A new multimodel ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.*, **36**, L21711, doi:10.1029/2009GL040896.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

Yang, S., and X. Jiang, 2014: Prediction of eastern and central Pacific ENSO events and their impacts on East Asian climate by the NCEP Climate Forecast System. *J. Climate*, **27**, 4451–4472, doi:10.1175/JCLI-D-13-00471.1.

Yang, X., and T. DelSole, 2012: Systematic comparison of ENSO teleconnection patterns between models and observations. *J. Climate*, **25**, 425–446, doi:10.1175/JCLI-D-11-00175.1.