# The Northern Hemisphere Extratropical Atmospheric Circulation Response to ENSO: How Well Do We Know It and How Do We Evaluate Models Accordingly?

CLARA DESER, ISLA R. SIMPSON, KAREN A. MCKINNON, AND ADAM S. PHILLIPS

*National Center for Atmospheric Research,[a] Boulder, Colorado*

## ABSTRACT

Application of random sampling techniques to composite differences between 18 El Niño and 14 La Niña events observed since 1920 reveals considerable uncertainty in both the pattern and amplitude of the Northern Hemisphere extratropical winter sea level pressure (SLP) response to ENSO. While the SLP responses over the North Pacific and North America are robust to sampling variability, their magnitudes can vary by a factor of 2; other regions, such as the Arctic, North Atlantic, and Europe are less robust in their SLP patterns, amplitudes, and statistical significance. The uncertainties on the observed ENSO composite are shown to arise mainly from atmospheric internal variability as opposed to ENSO diversity. These observational findings pose considerable challenges for the evaluation of ENSO teleconnections in models. An approach is proposed that incorporates both pattern and amplitude uncertainty in the observational target, allowing for discrimination between true model biases in the forced ENSO response and apparent model biases that arise from limited sampling of non-ENSO-related internal variability. Large initial-condition coupled model ensembles with realistic tropical Pacific sea surface temperature anomaly evolution during 1920–2013 show similar levels of uncertainty in their ENSO teleconnections as found in observations. Because the set of ENSO events in each of the model composites is the same (and identical to that in observations), these uncertainties are entirely attributable to sampling fluctuations arising from internal variability, which is shown to originate from atmospheric processes. The initial-condition model ensembles thus inform the interpretation of the single observed ENSO composite and vice versa.

## 1. Introduction

The El Niño–Southern Oscillation (ENSO) phenomenon is the leading mode of coupled variability of the tropical ocean–atmosphere system on interannual time scales (e.g., Philander 1990; Neelin 2012; and references therein). During its warm phase (El Niño), sea surface temperatures (SSTs) in the central and eastern tropical Pacific increase, the trade winds weaken, and the equatorial Pacific thermocline flattens. Opposite-signed changes occur during the cool phase (La Niña), with a tendency for weaker but longer-lived anomalies compared with those during the warm phase. ENSO events typically last 1–2 years and repeat irregularly at intervals of approximately 3–8 years. Driven primarily by coupled air–sea processes within the tropical Indo-Pacific, the effects of ENSO are transmitted to the rest of the world via changes in the large-scale atmospheric circulation, with ensuing impacts on climate and ecosystems.

Recent progress in understanding, observing, and modeling ENSO has resulted in skillful dynamical predictions of this phenomenon up to two seasons in advance (Barnston et al. 2012; Kirtman et al. 2014; McPhaden et al. 2015; Stock et al. 2015; Gonzalez and Goddard 2016; Infanti and Kirtman 2016; Kumar et al. 2016). Similarly, improved understanding of the dynamical processes governing the extratropical atmospheric circulation response to ENSO has provided a physical basis for seasonal climate prediction over North America and Eurasia (Ropelewski and Halpert 1987; Shukla et al. 2000; Tippett et al. 2012; L'Heureux et al. 2015; Scaife et al. 2014; Dunstone et al. 2016). Such forecasts, although less skillful than those for ENSO itself because of the presence of inherently unpredictable atmospheric variability (i.e., atmospheric noise), are of great economic value and benefit to society.

The extratropical Northern Hemisphere atmospheric circulation response to ENSO includes regional teleconnection patterns over the Pacific–American and

Atlantic–European sectors, as well as a zonal-mean component manifest in the position and strength of the subtropical and midlatitude jet streams (e.g., Horel and Wallace 1981; Alexander et al. 2002; Seager et al. 2003; Spencer and Slingo 2003; L'Heureux and Thompson 2006; Lu et al. 2008; and many others). Dynamical understanding of these responses is well advanced, albeit still incomplete. In particular, zonally asymmetric ENSO teleconnections can be understood within the general theoretical framework of poleward-propagating Rossby waves forced from the tropics by anomalous upper-tropospheric divergence associated with latent heat release in deep convection and subsequently modified by wave–mean flow interactions (Hoskins and Karoly 1981; Simmons et al. 1983; Held et al. 1989; Hoskins and Ambrizzi 1993; Sardeshmukh and Hoskins 1988; Trenberth et al. 1998; Garfinkel and Hartmann 2010; and many others). Stratosphere–troposphere coupling may also play a role in the response over the North Atlantic–European sector (e.g., Manzini et al. 2006; Garfinkel and Hartmann 2008; Ineson and Scaife 2009; Cagnazzo and Manzini 2009; Nishii et al. 2010; Richter et al. 2015).

How well do we know the observed atmospheric circulation response to ENSO? It may seem surprising to ask this question, given the long history of empirical ENSO research discussed above. However, growing appreciation of the challenge of isolating the forced response to a given perturbation in the presence of unrelated internal variability prompts a revisiting of this question (e.g., Deser et al. 2012a; Wallace et al. 2013; McGraw et al. 2016). Typically, empirical studies apply compositing or regression analysis to a particular period of record, almost always <50 yr, to estimate the observed extratropical atmospheric circulation response to ENSO (e.g., Horel and Wallace 1981; Trenberth and Caron 2000; Hoerling et al. 2001; DeWeaver and Nigam 2002; Garfinkel et al. 2013; Frauen et al. 2014). If the sample size of ENSO events in the period analyzed is sufficiently large, the noise due to variability that exists in the absence of ENSO will be minimized, revealing the forced response. The question then becomes how many events (or what length of record) are needed to identify the forced response without significant aliasing of unrelated variability? Further, the fact that no two El Niño events (and no two La Niña events) are identical poses additional challenges for estimating the forced response to ENSO based on observations (see also Garfinkel et al. 2013).

The longest observational records that exist for determining the atmospheric circulation response to ENSO are those based on station pressure. Indeed, the earliest studies of global atmospheric teleconnection patterns utilized surface pressure measurements at land stations (i.e., Exner 2015; Walker 1923). Merchant ships also collected data on sea level pressure (SLP), providing complementary coverage over the oceans starting in the mid-nineteenth century (e.g., Woodruff et al. 2008). However, reliable marine data coverage in the Northern Hemisphere (NH) did not begin until the early 1920s, with some areas of the North Pacific not well sampled until after World War II (Woodruff et al. 2008; Trenberth and Paolino 1980; Deser et al. 2010; Raible et al. 2014) and the Southern Ocean not until at least the 1960s (Fan et al. 2014). Thus, empirical determination of the observed atmospheric response to ENSO relies on a record that is, at best, approximately 100 years long. To what degree this record provides a sufficient sampling of ENSO events is one of the questions this study will address.

Atmospheric modeling studies have long known the importance of large sample sizes to obtain robust statistics on the response to a given forcing. For example, Sardeshmukh et al. (2000) examined the global atmospheric circulation responses to the 1986/87 El Niño and 1988/89 La Niña events using 180-member ensembles of simulations with an atmospheric general circulation model (AGCM). With such large ensembles, it was possible to obtain detailed information on the remote response to ENSO, including changes in the probability of occurrence of extreme events and asymmetries between El Niño and La Niña. In this case, the ensemble spread was entirely attributable to internal atmospheric variability. Hoerling and Kumar (1997) used a similar approach to study the atmospheric response to the seven strongest El Niño events during 1950–94 based on a 13-member AGCM ensemble. By comparing the ensemble spread for each individual event and for the composite of all seven events, they concluded that internal atmospheric processes contributed to the majority of the uncertainty in the forced response (given by the seven-event composite), more so than differences among individual El Niño events.

While modeling studies have the luxury of large ensemble sizes to identify the forced response in the presence of noise, observational studies are limited to a single realization for any given ENSO event or any particular composite of ENSO events. However, conclusions regarding the uncertainty in the model's forced response to ENSO are only valid to the extent that the model realistically simulates internal variability unrelated to ENSO. Indeed, Scaife et al. (2014) and Dunstone et al. (2016) suggest that, for the case of the North Atlantic Oscillation (NAO), models may have too much atmospheric noise and therefore spuriously small signal-to-noise ratios and thus may require larger ensemble sizes for skillful predictions than would be

needed in the real world. The question of whether models' responses to ENSO are subject to realistic levels of uncertainty arising from atmospheric noise shall be addressed in this study.

Renewed interest in ENSO diversity (e.g., Capotondi et al. 2015) has led to a proliferation of studies examining whether different types of events lead to distinct circulation responses. For example, some studies claim that El Niño events that maximize in the central Pacific drive different extratropical responses from those that peak in the eastern Pacific (e.g., Graf and Zanchettin 2012; Hegyi and Deng 2011; Xie et al. 2012; Feng et al. 2016; Yu et al. 2012, 2015). Others question whether these findings are statistically robust because of the limited sample of events in each category (Garfinkel et al. 2013). Similarly, the issue of whether El Niño and La Niña produce nonlinear responses in terms of amplitude and pattern remains controversial, with some studies reporting significant nonlinearities (Hoerling et al. 1997, 2001) and others finding no significant differences once internal decadal-scale climate variability is accounted for (DeWeaver and Nigam 2002). We note that the nonlinear ENSO responses identified in Frauen et al. (2014) are mainly confined to the tropics and subtropics, not the middle and high latitudes that are the subject of this investigation.

Given the limited span of the observational record, sampling requirements suggested from modeling studies, and ENSO diversity, we revisit the issue of how well the observed atmospheric circulation response to ENSO is known and how to evaluate models accordingly. While this subject is not new (i.e., all of the studies cited above include estimates of statistical significance on their ENSO teleconnections), our approach provides a more integrated perspective by incorporating information on both pattern and amplitude uncertainty on the observed ENSO response. We use as long of a record as feasible within the observational data constraints (1920–2013, a period that contains 18 El Niño events and 14 La Niña events by our criteria) and standard random sampling techniques to construct synthetic ENSO composites, each of which could have plausibly occurred had a different temporal sequence of natural variability unrelated to ENSO occurred. These synthetic observationally based ENSO composites provide important context for, and uncertainty bounds on, the one composite that actually did occur. Issues related to ENSO diversity and nonlinearity within these synthetic composites are also addressed.

In addition to our observational analysis, we analyze a suite of initial-condition coupled and atmosphere-only model ensembles with realistic tropical Pacific sea surface temperature anomaly evolution during 1920–2013.

We then construct ENSO composites for each ensemble member using the same set of events as in our observational analysis. The resulting range of composite ENSO teleconnections across the individual members of a given model ensemble provides a direct assessment of the uncertainties associated with any single composite sample (i.e., the model's ensemble spread is the counterpart of the spread across the observationally based synthetic composites for which there is only one actual composite sample). That is, each composite within a given model ensemble represents the model's true forced response to ENSO combined with a different sampling of its (unrelated) internal variability. We then propose an approach for evaluating the models' ENSO teleconnections that incorporates both pattern and amplitude uncertainty in the observational target, allowing for discrimination between true model biases in the forced response to ENSO and apparent model biases that arise from limited sampling of non-ENSO-related natural variability alone. Our focus is on ENSO teleconnections over the extratropical NH in boreal winter, defined here as December–February (DJF). A forthcoming companion study will deal with the impacts of these ENSO teleconnections on surface air temperature and precipitation over North America.

The rest of this study is organized as follows. The observational datasets, model simulations, and methodology are described in section 2. Results are presented in section 3 and summarized and discussed in section 4.

## 2. Data and methods

### a. Observational data

Our primary SLP dataset is the Twentieth Century Reanalysis (20CR), version 2, (Compo et al. 2011) on a 2° latitude × 2° longitude grid. The 20CR is an atmospheric reanalysis product that assimilates only surface pressure reports using an ensemble Kalman filter procedure and observed monthly mean SSTs and sea ice concentrations as surface boundary conditions, taken from the Met Office HadISST1.1 (Rayner et al. 2003). We verify that our results are robust to choice of dataset by comparing the 20CR results with the ECMWF reanalysis product, a combination of the ECMWF twentieth-century reanalysis (ERA-20C; Poli et al. 2016) prior to 1979 and ERA-Interim (Dee et al. 2011) after 1979. This dataset is provided on a 1° latitude × 1° longitude grid, and like 20CR assimilates only surface atmospheric fields (pressure and winds). We also verify that the results based on 20CR and the ECMWF reanalyses are consistent with those obtained with the Trenberth and Paolino (1980) SLP dataset. The latter

is a purely observational dataset with some missing data; it is provided on a 5° latitude × 5° longitude grid.

### b. Model simulations

We use a hierarchical set of model configurations and simulations to examine the sensitivity of our results to the degree of ocean–atmosphere coupling and model structural uncertainty. These include: (i) tropical Pacific pacemaker simulations, which make use of fully coupled climate models but with the historical evolution of SST anomalies in the eastern tropical Pacific relaxed toward observations; these yield the global coupled ocean–atmosphere response to observed ENSO events; (ii) tropical ocean–global atmosphere (TOGA) simulations, which make use of atmosphere-only models with the observed historical evolution of SSTs prescribed throughout the tropics and the observed climatological seasonal cycle of SSTs prescribed elsewhere; these yield the atmospheric response in the absence of air–sea coupling to the tropical-wide ENSO signal; and (iii) an atmosphere-only control simulation configured by prescribing the climatological seasonal cycle of SSTs globally; this provides an assessment of internal atmospheric variability, which can confound the forced response to ENSO. With the exception of (iii), all simulations include the historical (and RCP8.5 after 2005) radiative forcing protocols of CMIP5 (Taylor et al. 2012). More detailed information on each type of simulation is provided below.

#### 1) TROPICAL PACIFIC PACEMAKER COUPLED MODEL SIMULATIONS

A coordinated set of experiments was performed with CESM1, GFDL CM2.1, and MIROC5 in which SST anomalies in the eastern tropical Pacific (10°S–10°N, 160°–90°W, with a linearly tapering buffer zone that extends to 20°S and 20°N, 180°W to the American coast) are nudged to those from the NOAA Extended Reconstruction Sea Surface Temperature, version 3, (ERSSTv3b) dataset. This protocol follows that of Kosaka and Xie (2013) except for the choice of SST dataset (they used HadISST1). In this way, the observed evolution of ENSO is maintained in each simulation (i.e., ENSO is the pacemaker), with the rest of the model's coupled climate system free to evolve. Note that only the SST anomalies, not the total SST, are nudged to observations, maintaining the model's mean state, including any model biases. Both CESM1 and GFDL CM2.1 performed 10-member ensembles, while MIROC5 conducted a five-member ensemble. Each models' ensemble members are identical except for an initial atmospheric temperature perturbation introduced on the first day of the simulation, of a magnitude that is at the

level of round-off error [$10^{-14}$ °C; see also Kay et al. (2015) for further explanation]. This small initial perturbation serves to create spread among the individual members after a few months because of unpredictable chaotic dynamics of the system (e.g., Lorenz 1963). The GFDL and MIROC pacemaker runs begin in 1880, while the CESM ones begin in 1920, and they all end in 2013. We shall use the CESM1 pacemaker simulations as our primary model dataset and therefore base all of our analyses on the period 1920–2013. For the purpose of our study, we consider the pacemaker simulations to be the most realistic setting for evaluating the models' response to ENSO, since observed SST anomalies are prescribed only in the eastern tropical Pacific, leaving the rest of the global coupled ocean–atmosphere–land system to respond.

#### 2) TOGA ATMOSPHERIC MODEL SIMULATIONS

Before the advent of the pacemaker protocol, the typical approach for studying a model's response to ENSO was to prescribe the observed evolution of tropical SSTs (either confined to the eastern Pacific or encompassing all ocean basins), and the observed climatological seasonal cycle of SSTs elsewhere, to an atmospheric model coupled to a land surface model [the so-called TOGA configuration (Lau and Nath 1994)]. A 10-member TOGA ensemble was conducted with the atmosphere–land model components of CESM1, using SSTs from ERSSTv4 and the same initial-condition procedure for generating the individual members as the CESM1 pacemaker runs. While these simulations have the advantage of a realistic tropical-wide distribution of SST anomalies during ENSO (and a realistic SST mean state), they are more idealized in the sense that they lack two-way ocean–atmosphere coupling. We include them here to examine whether the results based on the coupled pacemaker simulations hold for these more traditional AMIP-style experiments. We note that a slightly different version of the ERSST dataset was used for the TOGA and pacemaker experiments. Version 4 of this dataset, which was used for the TOGA simulations, has slightly smaller (~10%) variability in the eastern equatorial Pacific than version 3b; the two datasets are highly correlated in time ($r > 0.99$).

#### 3) ATMOSPHERIC CONTROL SIMULATION

Finally, we make use of a 2600-yr preindustrial (constant 1850 radiative conditions) control simulation conducted with CESM1's atmosphere–land model components using a prescribed repeating seasonal cycle of SSTs and sea ice conditions taken from the long-term climatology of a companion 2200-yr preindustrial control run of the fully coupled CESM1 (see Kay et al. 2015).

TABLE 1. The 18 El Niño events used for the ENSO composites, identified by the years spanning the season November–January (e.g., "25/26" indicates the El Niño event of 1925/26). Shown are the number of times each event was sampled in the synthetic ENSO composites based on the 10th- and 90th-percentile values of the Aleutian low (AL10 and AL90, respectively) and Arctic (AR10 and AR90, respectively) SLP indices shown in Fig. 4 (see text for details). The last column (bold font, labeled UQ) shows the number of unique events in each composite.

| El Niño | 25/26 | 30/31 | 40/41 | 41/42 | 57/58 | 63/64 | 65/66 | 68/69 | 72/73 | 82/83 | 86/87 | 87/88 | 91/92 | 94/95 | 97/98 | 02/03 | 06/07 | 09/10 | UQ 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AL10 | — | 1 | 2 | 1 | 1 | 2 | — | — | — | 3 | 1 | 1 | 1 | 1 | 1 | — | 2 | 1 | **13** |
| AL90 | 1 | — | — | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | — | 1 | — | 1 | **14** |
| AR10 | 2 | — | 4 | — | 1 | 2 | 1 | — | 2 | — | — | 4 | — | 1 | — | — | — | 1 | **9** |
| AR90 | 1 | — | 2 | 1 | 1 | 1 | 2 | — | 2 | 2 | 3 | 2 | 1 | — | — | — | — | — | **11** |

This lengthy atmosphere-only control simulation provides robust statistics on the variability created solely by the atmosphere and land. It does not include any nonseasonal variation in marine surface boundary conditions, including ENSO-related SST anomalies. As such, it provides a benchmark for the simulated level of unpredictable atmospheric circulation variability that exists in the absence of ENSO.

*c. Methods*

We compute monthly anomalies by subtracting the long-term monthly means based on the period 1920–2013 from the corresponding month of each year. We then form winter (DJF) averages from the monthly anomalies and linearly detrend the DJF time series. We evaluate statistical significance using a two-sided Student's *t* test as well as a random sampling approach, depending on the null hypothesis being tested (see section 3). Next, we describe our procedures for constructing the ENSO composites, including the application of random sampling techniques.

1) ENSO COMPOSITING

We identify El Niño (EN) and La Niña (LN) events using a standard approach based on the Niño-3.4 SST index [SST anomalies averaged over 120°–170°W, 5°S–5°N (Barnston et al. 1997)]. Here we use ERSSTv3b to construct the Niño-3.4 index. We first smooth the monthly Niño-3.4 SST index with a three-point binomial filter, average it over November–January (NDJ), linearly detrend the time series, and identify EN (LN) events when the detrended index exceeds (falls below minus) one standard deviation (following, e.g., Alexander et al. 2002; Okumura and Deser 2010; Deser et al. 2012b). This procedure yields 18 EN and 14 LN events during 1920–2013; these events are listed in Tables 1 and 2. Note that the Niño-3.4 index is based on NDJ, one month ahead of the DJF season used for the extratropical SLP composites to accommodate the time scale of the Rossby wave response to tropical heating [e.g., Alexander et al. (2002) and references therein]. We form ENSO composites by subtracting the average of the 14 LN events from the average of the 18 EN events. Similar ENSO composites are obtained using the Niño-4 SST index (SST anomalies averaged over 160°E–150°W, 5°S–5°N) in place of Niño-3.4 (not shown).

2) RANDOM SAMPLING

Our aim is to evaluate to what degree sampling variability influences our observed ENSO composites of SLP over the extratropical NH. That is, to what extent can we interpret the ENSO-composite circulation anomalies as an accurate representation of nature's response to

TABLE 2. As in Table 1, but for 14 La Niña events.

| La Niña | 24/25 | 33/34 | 38/39 | 42/43 | 49/50 | 55/56 | 73/74 | 75/76 | 84/85 | 88/89 | 98/99 | 99/00 | 07/08 | 10/11 | UQ 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AL10 | — | — | 1 | 2 | 3 | 1 | 1 | 2 | — | 1 | 1 | — | 2 | — | **9** |
| AL90 | — | 2 | 1 | — | 1 | — | 1 | 1 | 3 | — | 3 | 2 | — | — | **8** |
| AR10 | — | — | — | 2 | — | 3 | — | 1 | — | — | 2 | — | 2 | 4 | **6** |
| AR90 | 1 | 2 | — | — | 1 | — | 1 | 1 | — | 3 | 1 | 2 | 1 | 1 | **10** |

ENSO, given that we only have 18 EN and 14 LN years to examine? In the model world, we can examine this question directly by running an ensemble of simulations forced by the observed evolution of ENSO to eliminate (or at least greatly reduce) the noise from non-ENSO variability. In the real world, we have no such recourse. If we assume for the moment that the 18 EN events are exchangeable and that the 14 LN events are exchangeable (in the sense that event-to-event differences are unimportant), we can form synthetic ENSO composites by randomly sampling with replacement from among the 18 EN events and the

14 LN events, always retaining 18 samples for the former and 14 samples for the latter (these samples will necessarily omit some events and repeat others). The resulting "bootstrapped" ENSO composites, which always consist of 18 EN and 14 LN events, albeit nonunique ones, yield an estimate of the degree to which non-ENSO sampling variability influences the one ENSO composite we have actually observed. Of course, this procedure breaks down to the extent that event-to-event differences are important, or that the number of events is not sufficient to fully characterize the variability. Both of these issues will be
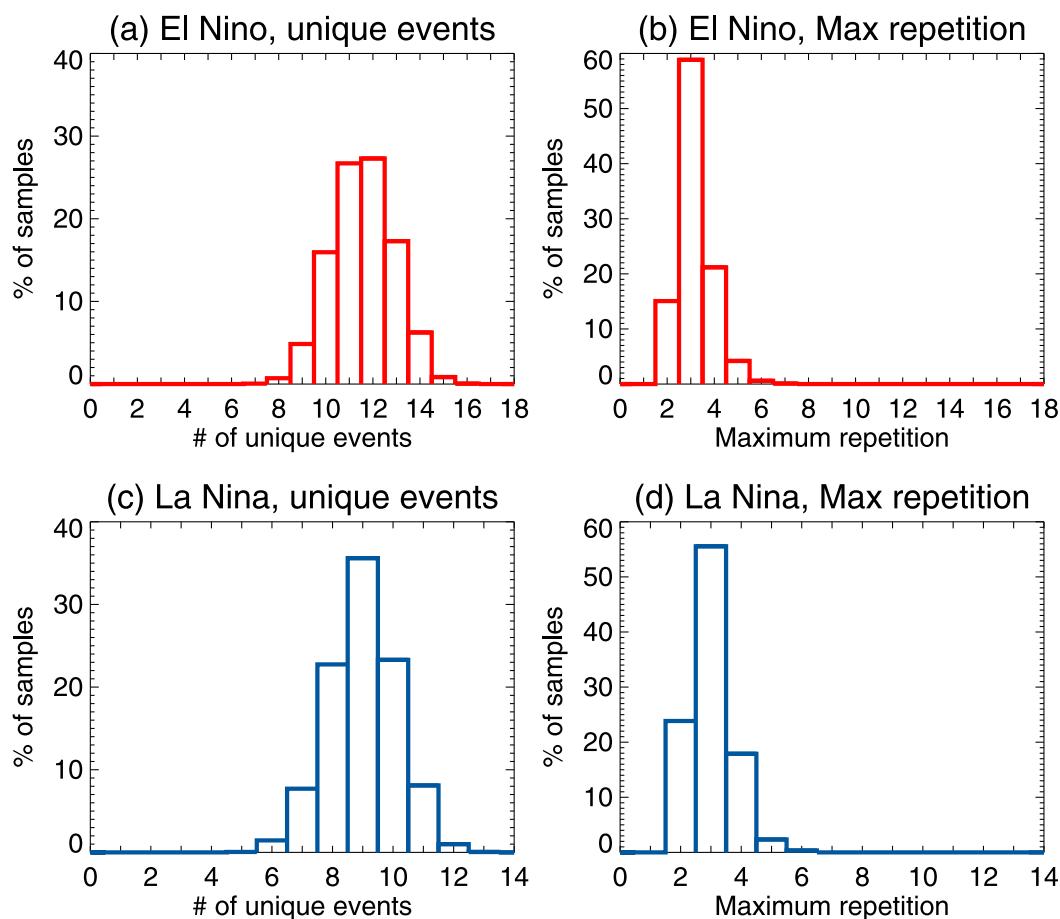


FIG. 1. Distribution of (a) the number of unique El Niño events, (b) the maximum number of times a single El Niño event is repeated, (c) the number of unique La Niña events, and (d) the maximum number of times a single La Niña event is repeated, in the 2000 synthetic ENSO composites based on observations. See text for details.
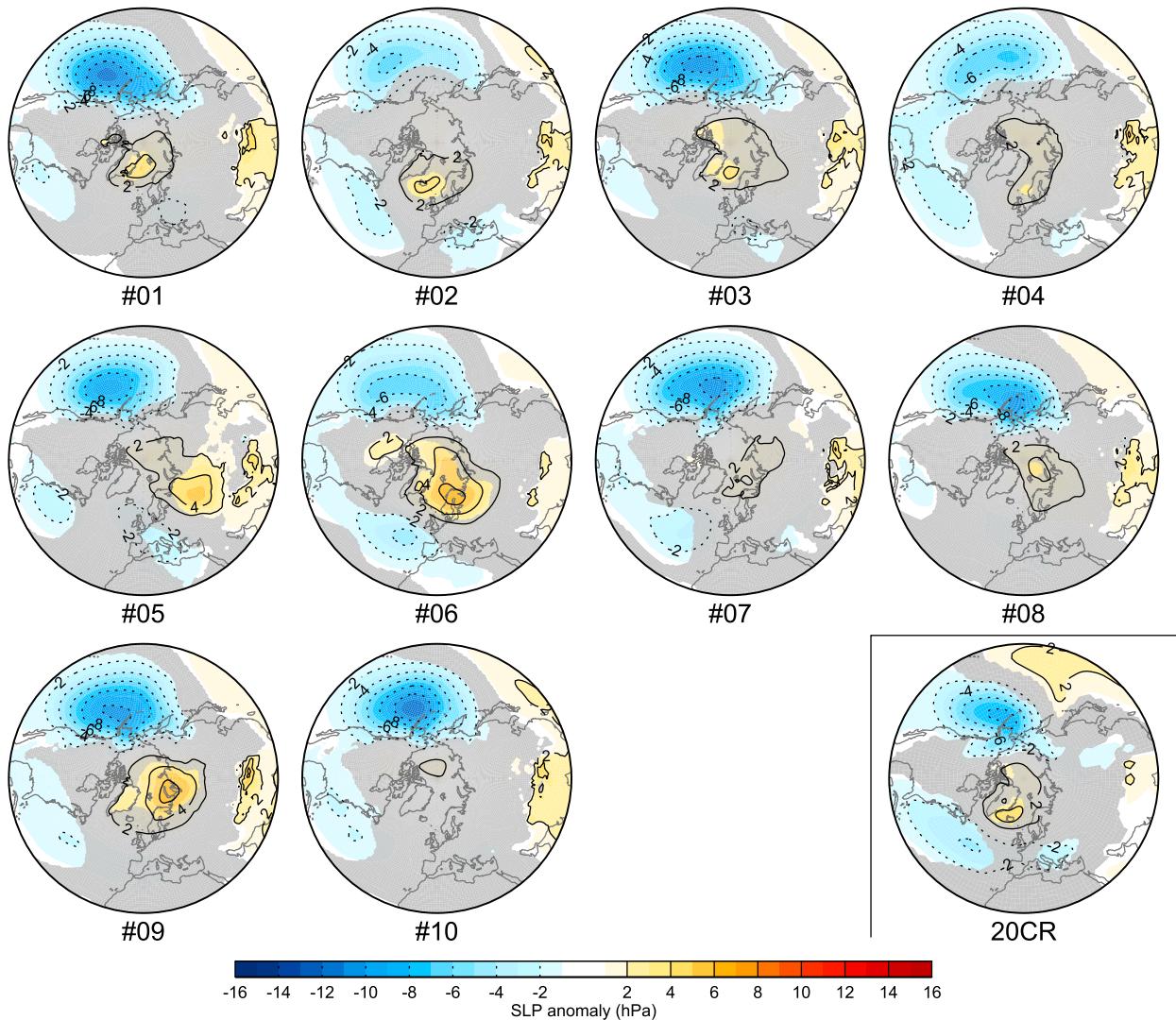
FIG. 2. ENSO composites of DJF SLP from each of the 10 CESM1 pacemaker simulations (labeled #01, . . . , #10) and from (bottom right) 20CR. Each composite is based on the same set of 18 El Niño events minus 14 La Niña events during 1920–2013. Values not significant at the 5% confidence level based on a two-sided *t* test are shaded in gray. Contour interval is 2 hPa.

addressed below. We have generated 2000 bootstrapped samples following this procedure for both the observational record and for each model's ensemble of simulations individually and combined. Note that the mean of the 2000 bootstrapped composites based on observations (or any individual model simulation) is centered on the actual observed (simulated) ENSO composite. Finally, to assess whether the 18 EN and 14 LN events sampled during our period of study are sufficient to accurately characterize the true distribution of internal variability unrelated to ENSO, we also estimate non-ENSO-related internal variability by considering the full 93-yr observational record, in addition to using the atmospheric model control simulation (which omits ENSO by design).

To give an idea of the sampling characteristics of ENSO events in our observed bootstrapped composites, we show the distribution of the number of unique EN and unique LN events across all 2000 samples, as well as the distribution of the maximum number of times a single EN event or a single LN event is repeated (Fig. 1). The majority of observed bootstrapped composites consist of 11–12 unique EN events and 9 unique LN events, or approximately 64% of the total number of unique events available in the record (18 EN and 14 LN). Similarly, the maximum number of times a single event is repeated in either the EN or the LN observed composite is most commonly 3, with <5% of samples having more than a maximum of 4 repetitions. Results

are similar for the model simulations when applying the random sampling procedure to each one individually (not shown). When the procedure draws from all members of a given model ensemble simultaneously, the sampling of unique events and the maximum number of repetitions of a given event are much higher and lower, respectively (not shown).

## 3. Results

### a. Observed and simulated ENSO SLP composites over the NH

Figure 2 illustrates the diversity of patterns and magnitudes in ENSO composites of extratropical NH SLP across the 10 CESM1 pacemaker simulations. Similar diversity is apparent in the GFDL and MIROC pacemaker ensembles (see Figs. A1 and A2) and in the CESM1 TOGA ensemble (see Fig. A3). Recall that each composite is based on the same set of observed ENSO events (18 EN and 14 LN) so that any differences among ensemble members are due entirely to sampling fluctuations (i.e., contamination of the ENSO response by internal variability). Here we evaluate statistical significance (at a level of 5%) against a null hypothesis of zero response using a two-sided Student's *t* test. All ensemble members show a significant negative SLP anomaly over the North Pacific, but the amplitude of this response varies by approximately a factor of 2 (e.g., member 1 shows maximum amplitude of 12 hPa compared with 6 hPa in member 2). There is also considerable diversity in other aspects of the composites. For example, about half of the ensemble members show a significant positive SLP anomaly in the polar region, while the others show no significant polar response. In addition, the significant negative SLP anomaly over the North Atlantic varies considerably in amplitude and location across the runs: in some members, it is confined to the southwestern portion of the basin, while in others (notably runs 2 and 6) it extends to the central and eastern Atlantic, forming a dipole structure with the polar anomaly of opposite sign reminiscent of the NAO (Hurrell et al. 2003). Thus, while there is some commonality to the ENSO composites across the ensemble members, such as the significant negative SLP anomalies over the central North Pacific and southwestern North Atlantic, their amplitudes vary by approximately a factor of 2; in addition, significant SLP anomalies over the polar region and central North Atlantic are evident only in some members. Similar results are found based on the period since 1950, which consists of 14 EN and 9 LN events (see Fig. A4).

It is interesting to view the observed (20CR) composite in the context of the diversity of the CESM1 pacemaker simulations (Fig. 2, bottom-right panel). We

note that nearly identical results to 20CR are obtained with the ECMWF reanalyses (see Fig. A2) and the Trenberth and Paolino (1980) dataset (see Fig. A3). The main features of the observed composite can be identified in the various members of the pacemaker ensemble (although not necessarily in any individual member, although numbers 2 and 6 come close), such as the significant negative SLP anomalies over the central North Pacific and across the Atlantic and the significant positive SLP anomalies near Iceland and the subtropical western Pacific. However, the amplitude of the latter feature appears to be larger than that found in any of the 10 CESM1 pacemaker simulations.

The diversity of ENSO composites across the pacemaker ensemble raises the following related questions:

- How well do we know the observed response to ENSO given a sample of 18 EN and 14 LN events during 1920–2013? That is, to what extent is the observed composite subject to aliasing from unrelated variability?
- What might the observed composite have looked like had a different temporal sequence of variability unrelated to ENSO occurred?
- How should one go about evaluating the ENSO response in models given uncertainty in the observational target?
- How realistic is the representation of internal variability unrelated to ENSO in models?

We address these questions below, beginning with model evaluation and then turning to uncertainty in the observed ENSO composite.

### b. Evaluating CESM1's response to ENSO

The ensemble mean of the 10 CESM1 pacemaker composites provides an accurate determination of the model's response to ENSO, as it is based on a total of 180 EN events and 140 LN events. Over most of the NH, this SLP response is significantly different from zero at the 5% level according to a two-sided Student's *t* test: only those areas with near-zero values remain insignificant (Fig. 3a). This forced SLP response to ENSO consists of negative values over most of the North Pacific centered southeast of the Aleutian Islands (maximum amplitude of approximately 9 hPa), a center of positive values over the Arctic extending into northern Eurasia (maximum amplitude ~2–3 hPa), negative values of ~1 hPa extending from the Gulf of Mexico and the southeastern United States across the central North Atlantic into the Mediterranean, and positive values of ~1 hPa in the western subtropical Pacific, and a positive center over the Tibetan Plateau region. Compared with the observed composite (Fig. 3b), the model's ensemble-mean response shows a larger expanse of negative SLP
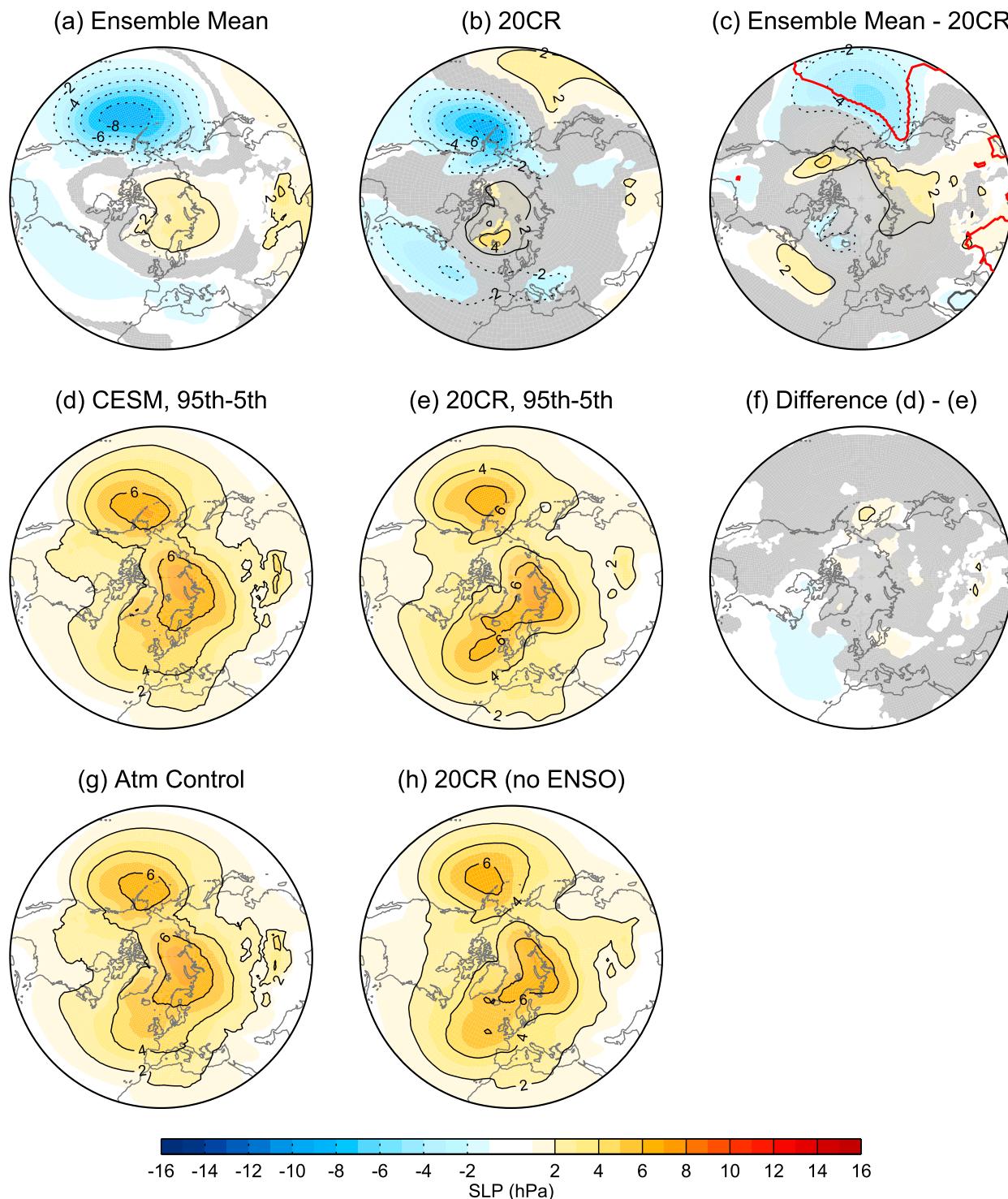
FIG. 3. ENSO composites of DJF SLP (hPa) for (a) ensemble-mean of the 10 CESM1 pacemaker simulations, (b) 20CR, and (c) their difference. Values not significant at the 5% confidence level based on a two-sided *t* test are shaded in gray. Red contours in (c) indicate regions where the 20CR lies outside any of the individual simulations. The 5%–95% CIs (hPa) on the SLP ENSO composites for (d) CESM1 pacemaker simulations, (e) 20CR, and (f) their difference. CIs are based on 2000 bootstrapped samples for both the model and 20CR. Gray shading in (f) indicates regions where the observed value falls within the spread of values from the individual simulations. (g) As in (d), but for the CAM5 atmospheric control simulation. (h) As in (e), but for 20CR after removing ENSO. See text for details.
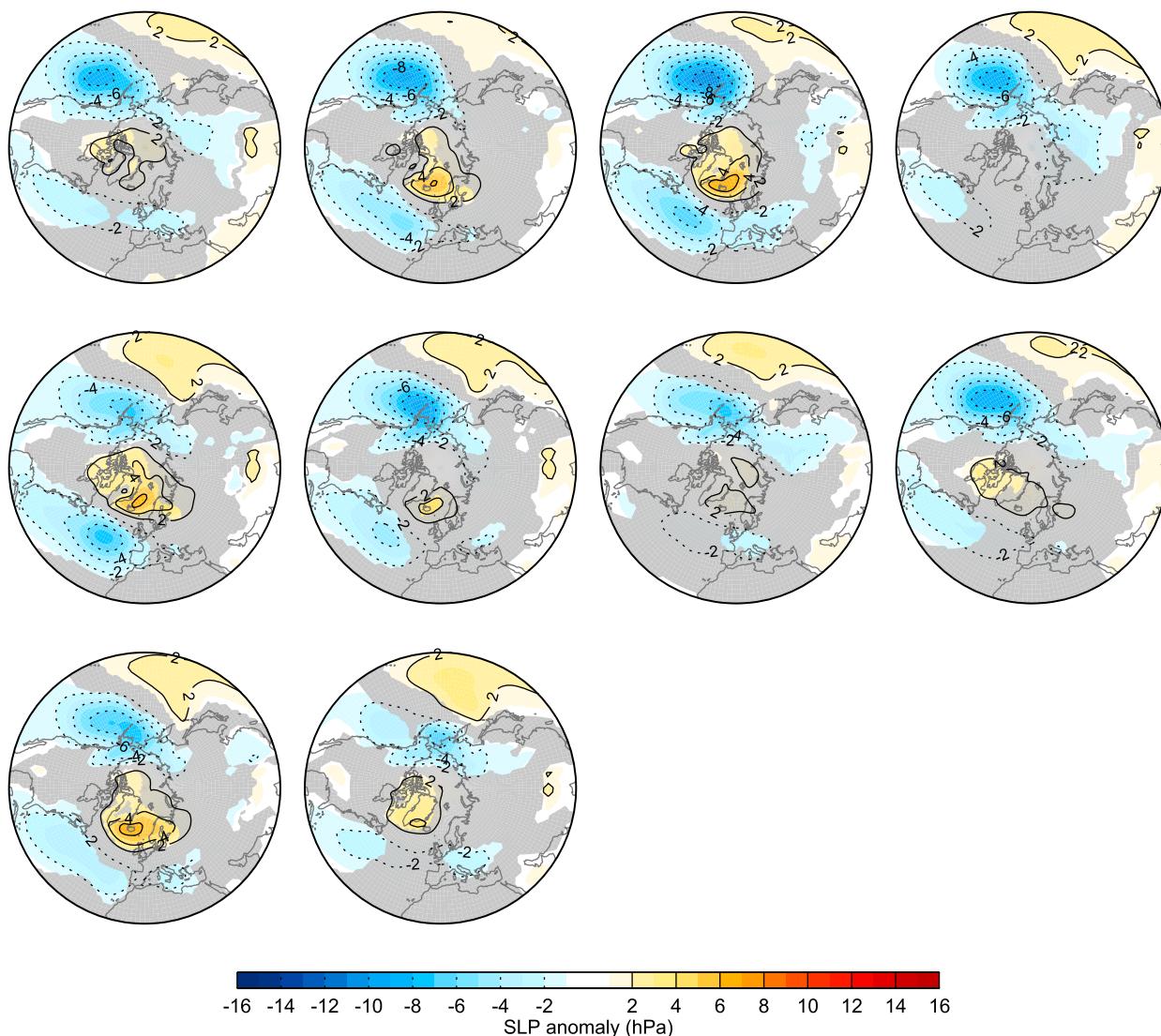
FIG. 4. As in Fig. 2, but for 10 randomly selected bootstrapped ENSO composites based on observations. See text for details.

anomalies over the North Pacific, a weaker and more spatially confined positive anomaly over the western subtropical Pacific, a weaker negative anomaly across the Atlantic, and a larger positive anomaly over the central Arctic and northern Eurasia.

The differences between the model's ensemble-mean ENSO response and the observed ENSO composite are shown in Fig. 3c. Areas without gray shading indicate regions where the observed composite has values lower than the 5th percentile or greater than the 95th percentile of all 2000 bootstrapped ENSO composites obtained by randomly sampling from among all 10 pacemaker ensemble members: that is, in these regions there is a less than 5% chance that a value as low, or as high, as seen in the observations would be obtained by sampling from the

model distribution. The red contours enclose areas where the single observational composite lies entirely outside of any of the 2000 bootstrapped model samples; in other words, there is a <1/2000 chance that the observed and model composites were drawn from the same distribution. By these measures, the southward and westward expansion of the model's North Pacific SLP response is clearly a model bias, as is the strength of the response over parts of the Middle East and far Southeast Asia; the weaker response over the Atlantic and stronger response over parts of northeast Asia and the southeastern United States are likely model biases.

However, to conclude that the differences noted above are indeed indicative of model biases, one has to assess whether the characteristics of the internal variability are
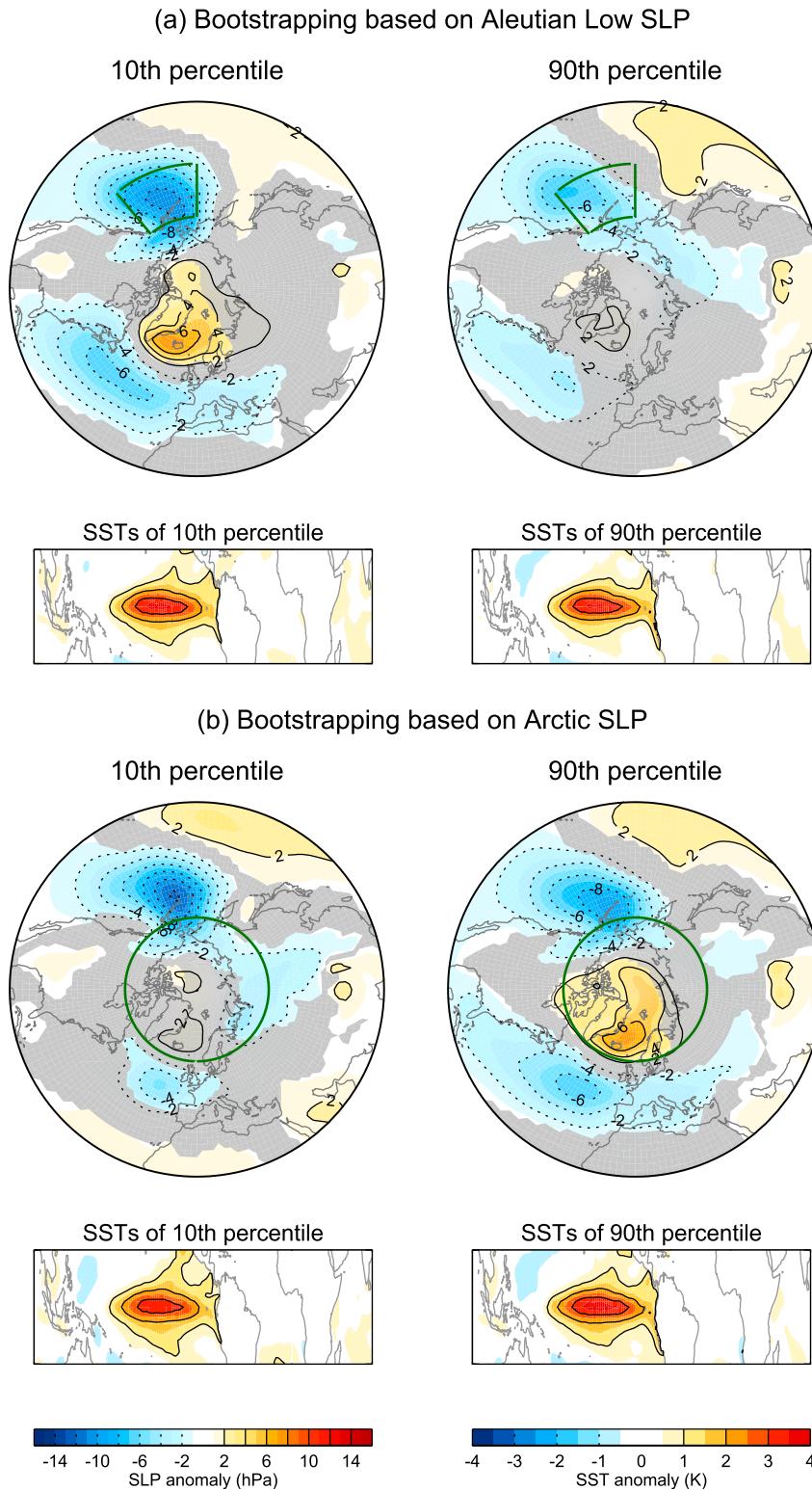
FIG. 5. Sample bootstrapped ENSO composites of (top) SLP (hPa) and (bottom) SST (°C) based on observations. The (left) 10th- and (right) 90th-percentile composites based on the (a) AL SLP index and (b) AR SLP index (index regions outlined in green). SLP values not significant at the 5% confidence level are shaded in gray.

## (a) AL vs Nino3.4      (b) AR vs Nino3.4
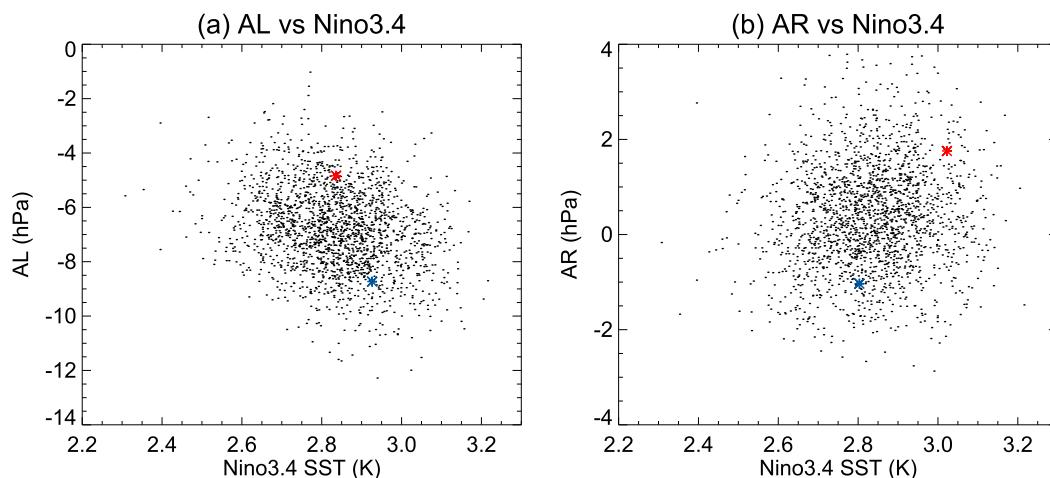
FIG. 6. Scatterplots of ENSO composite values of the (a) AL and (b) AR SLP indices (hPa) against the Niño-3.4 SST index (°C) based on 2000 observed bootstrapped samples. Blue and red symbols show the 10th- and 90th-percentile values, respectively, of the AL and AR indices.

the same in the model and the real world. For example, one might incorrectly determine a bias in the model's forced response if the model's variability is unrealistically small, or, conversely, one might not detect a bias in the model's forced response if the model's variability is unrealistically large. To address whether the model's variability is realistic, (and whether the observed ENSO composite is subject to the same uncertainty one would infer from the pacemaker ensemble), we examine maps of the confidence intervals (CIs) on the simulated and observed ENSO composites (Figs. 3d and 3e, respectively). These CIs are obtained by taking the difference between the 95th and 5th percentiles of the 2000 bootstrapped ENSO composites at each grid point independently (i.e., the CIs encompass 90% of the bootstrapped distributions at each grid point). Recall that the spread of the 2000 bootstrapped ENSO composites represents internal variability not driven by ENSO [Section 2c(2)]. Here, the model's CI is obtained by randomly sampling 18 EN and 14 LN events from all available ensemble members, calculating the composite mean difference between EN and LN, ordering the 2000 composite mean differences from smallest to largest at each grid point, and identifying the 5th and 95th percentiles. The spatial distribution and amplitude of the model's CIs are very similar to observations, with the largest values (approximately 6–7 hPa) occurring over the Gulf of Alaska, the Arctic, and the subpolar Atlantic, suggesting that the observed ENSO composite might be subject to a similar level of uncertainty as seen in the pacemaker ensemble. The differences between the simulated and observed CIs are shown in Fig. 3f. To identify regions where the model's variability may differ from that of the real world, we assess whether the observed CI at each grid point is

greater than the maximum CI or less than the minimum CI of any of the 10 pacemaker runs individually (the CIs for the individual pacemaker simulations are shown in Fig. A5). Regions shaded in gray in Fig. 3f indicate where the model's CIs encompass the observed value. The only substantial region where the model's CIs lie outside the observed value is the central North Atlantic; here, the model exhibits less internal variability (of non-ENSO origin) than nature (by about 1–2 hPa).

In summary, with the exception of the central North Atlantic where CESM1 underestimates the uncertainty on the ENSO SLP composite, the model's internal variability is in good agreement with observations. Because of this agreement in internal variability, we are able to attribute cases where the observed composite lies outside of the model ensemble to biases in the model's forced response. It also suggests that the observed ENSO composite is subject to the same level of uncertainty as one would infer from the spread among the 10 pacemaker ENSO composites shown in Fig. 2.

Next we compute the analogous CI from the 2600-yr atmospheric control simulation to assess the contribution of internal atmospheric variability alone to the CI on the simulated ENSO composites (Fig. 3g). To do so, we have randomly selected two groups of years from the control run: one consisting of 18 winters and the other of 14 winters. We then average the SLP anomaly fields within each group and take their difference, in a manner analogous to how we formed the ENSO composites. We repeat this procedure with replacement 2000 times and use these 2000 random samples to compute the CIs. By comparing the CI maps from the coupled model's pacemaker simulations (Fig. 3d) with those from the atmosphere-only control run (Fig. 3g), one may conclude
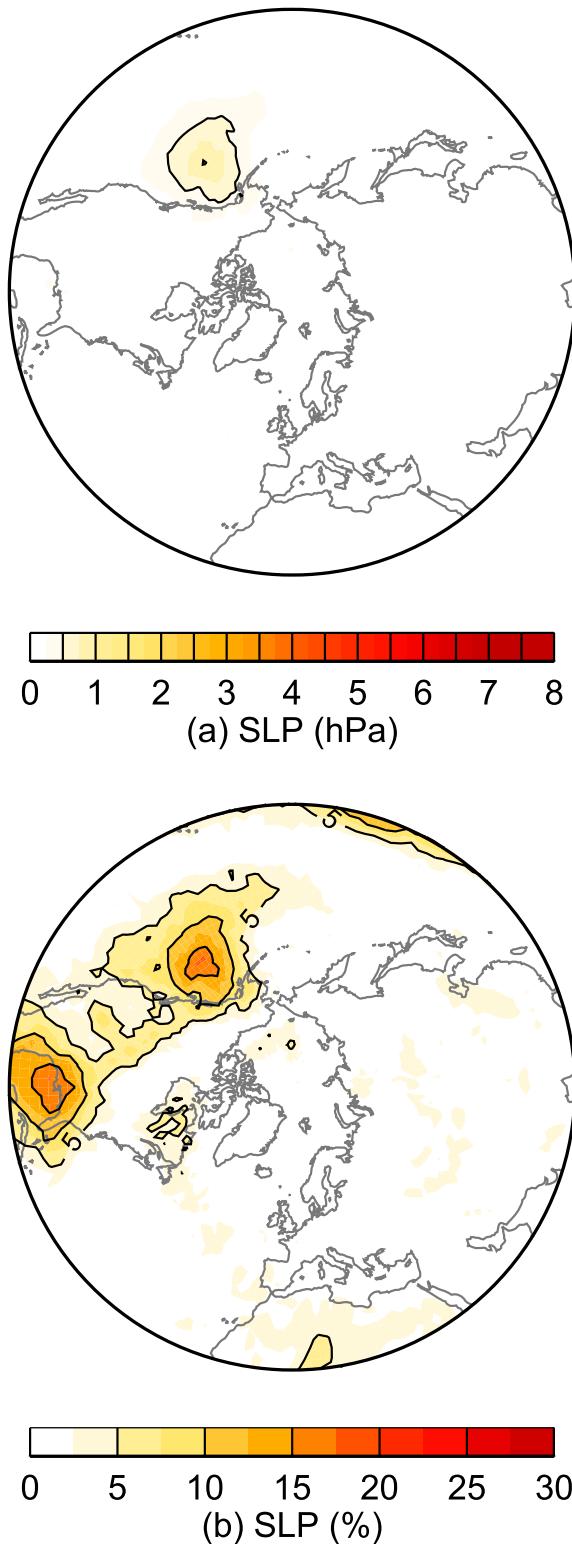
FIG. 7. Linear contribution of the Niño-3.4 SST index to the 5%–95% confidence intervals on ENSO composites of DJF SLP based on 2000 observed bootstrapped samples, expressed (a) in units of hectopascals (contour interval = 1 hPa) and (b) as a percentage of the total confidence interval shown in Fig. 3e (contour interval = 5%). See text for details.

that internal atmospheric variability accounts for virtually all of the uncertainty in the model's bootstrapped ENSO composites. In other words, the reason why the individual pacemaker runs differ in their ENSO composites of NH SLP is because of atmospheric variability that is inherently unpredictable on these time scales (e.g., from one winter to the next). Put another way, the random superposition of internal atmospheric circulation anomalies on the forced ENSO response causes the diversity of ENSO SLP composites within the pacemaker ensemble. The only way to reduce this atmospheric noise is to sample more ENSO events, which is possible in the model world by running more ensemble members but of limited practical application in the real world.

While we cannot isolate the contribution of atmospheric dynamics to the CIs calculated from the observations, we can evaluate the contribution of non-ENSO-related SLP variability. To address this, we have computed analogous CIs by randomly sampling from all 93 DJF seasons in the 1920–2012 record after linearly regressing out the Niño-3.4 SST index (Fig. 3h). Similar results are obtained by computing CIs from the 61 ENSO-neutral years (not shown). It is clear that the CI results based on the ENSO sample (18 EN events and 14 LN events; Fig. 3e) are very similar to that based on the ENSO residual sample (Fig. 3h). While this analysis does not allow us to definitively conclude that internal atmospheric variability underlies the uncertainty in the observed ENSO composite, it is certainly highly suggestive given the model-based results. The differences between the CIs from the atmospheric control simulation (Fig. 3g) and the ENSO-residual observations (Fig. 3h) are even smaller than those shown in Fig. 3f (i.e., the difference between the CIs from the pacemaker simulations and the total observations), with a similar pattern (not shown).

## c. Range of ENSO composites using bootstrapped observations

In this section, we address two related questions: 1) how well do we know the observed ENSO response and 2) what might the observed composite have looked like had a different sequence of natural variability occurred? Here we make use of the 2000 bootstrapped ENSO composites based on observations. Figure 4 shows 10 of these composites selected at random. In analogy with the 10 pacemaker simulations shown in Fig. 2, these synthetic observational composites display a range of SLP anomaly amplitudes and patterns. For example, some include an NAO-like response, while others exhibit no significant anomalies over the eastern North Atlantic, and the amplitude of the negative SLP values over the northeastern Pacific can vary by more than a factor of 2.
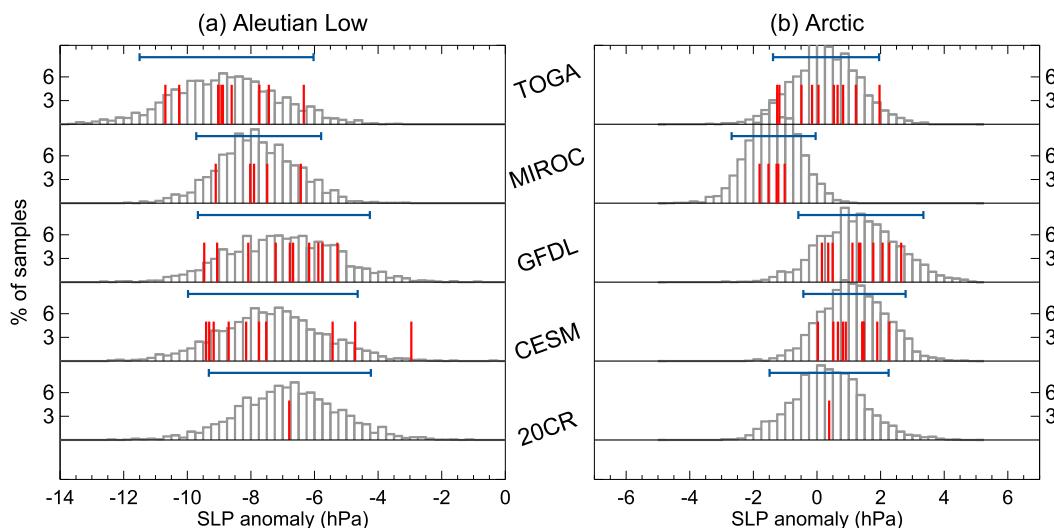
FIG. 8. Histograms of ENSO composite values of the (a) AL and (b) AR SLP indices (hPa). Gray bars denote values from the 2000 bootstrapped samples, and red bars indicate actual values. (bottom)–(top) The 20CR, pacemaker simulations from CESM1 (10 members), GFDL CM2.1 (10 members) and MIROC5 (5 members), and CAM5 TOGA simulations (10 members). The horizontal blue bar at the top of each histogram indicates the 5%–95% confidence interval based on 2000 bootstrapped samples.

Next, we sort the 2000 observational bootstrapped ENSO composites according to their area-weighted amplitude in two regions: the Aleutian low (AL; 140°W–180°, 40°–60°N) and the Arctic (AR; poleward of 60°N). For illustrative purposes, we display the 10th- and 90th-percentile composite samples based on the AL sorting in Fig. 5a and those based on the AR sorting in Fig. 5b. The particular EN and LN events and the number of times they are sampled for each of the bootstrapped composites shown in Fig. 5 are listed in Tables 1 and 2. For now, we note that a minimum of nine distinct EN events and six distinct LN events make up each of the composites shown in Fig. 5 and that no single event is sampled more than four times. Additional discussion of ENSO event sampling in the observed bootstrapped composites is provided below.

Figure 5a (top panels) shows the observed bootstrapped ENSO composites that lie at the 10th and 90th percentiles based on the AL index. As expected, the magnitude of the anomalous deepening of the Aleutian low differs by approximately a factor of 2 (−10 hPa for the 10th and −6 hPa for the 90th) between the two composites, in keeping with the choice of index region and the amplitude of the CIs shown previously. In addition, however, there are striking differences in their spatial patterns outside of the Pacific. In particular, the 10th-percentile composite shows a hemispheric pattern with significant positive anomalies over the polar region (maximum values ~6 hPa near Iceland) and significant negative anomalies stretching across the North Atlantic

(maximum amplitudes ~6 hPa) and extending into Europe; the dipole pattern between the Atlantic and Arctic is reminiscent of the NAO. In contrast, the 90th-percentile composite shows very limited areas of significant response beyond the Pacific and western Atlantic. Interestingly, the positive SLP response over the southwestern North Pacific is larger in the 90th-percentile composite compared with the 10th-percentile composite, while the magnitude of the AL response is weaker. The 10th- and 90th-percentile composites based on the AR response show similarities and differences with their AL counterparts. In particular, the 90th-percentile AR composite shows a similar hemispheric pattern of significant responses as the 10th-percentile AL composite, but the amplitudes are weaker over the North Pacific and stronger over the southwestern Pacific. Significant responses in the 10th-percentile AR composite are mainly confined to the Pacific.

The diversity of patterns and amplitudes in the observed bootstrapped ENSO composites shown in Figs. 5a and 5b is reminiscent of that seen in the 10-member CESM pacemaker ensemble (Fig. 2). Recall that the pacemaker composites are based on the identical set of 18 EN and 14 LN events, ruling out the possibility that differences in ENSO sampling give rise to their diversity. For observations, to the extent that our random sampling methodology does not introduce additional diversity as a result of differences between ENSO events, the range of ENSO composites in Figs. 4 and 5 illustrates what nature might have produced

### (a) Aleutian Low

### (b) Arctic
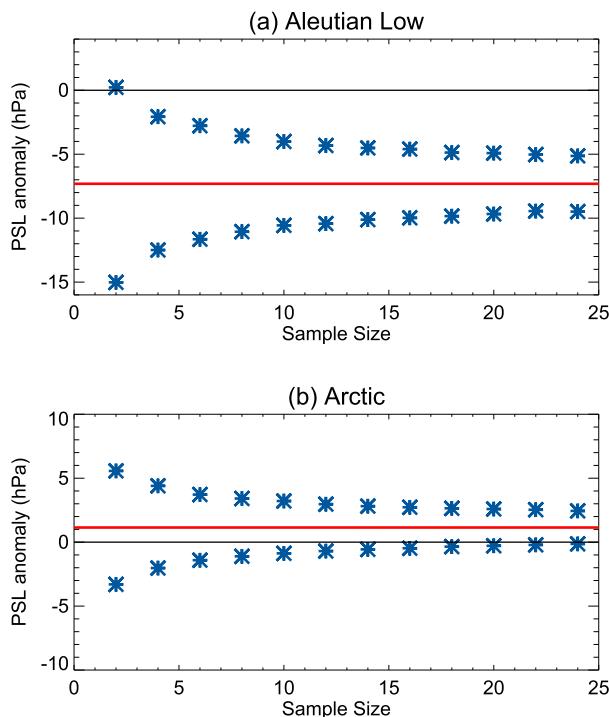
FIG. 9. The 5%–95% confidence intervals (blue asterisks; hPa) on ENSO composites of the (a) AL and (b) AR SLP indices as a function of sample size based on bootstrapping the CESM pacemaker ensemble. A sample size of two indicates two El Niño and two La Niña events, and so on. Red line denotes the ensemble-mean value of the SLP index (hPa).

given a different sequence of internal variability independent of ENSO. That is, even with a sample of 18 EN and 14 LN events, the amplitude, structure, and relative spatial emphasis of the winter extratropical NH SLP response to ENSO are subject to considerable uncertainty.

One way to address to what extent the different sampling of EN and LN events in the observed bootstrapped composites causes their SLP patterns to diverge is to compare their tropical SST anomalies (bottom panels of Figs. 5a,b). The overall patterns and amplitudes of the equatorial Pacific SST anomalies that comprise the 10th- and 90th-percentile composites based on the AL and AR sorting are very similar, with only small differences in peak magnitude and structure. Although the Niño-3.4 SST index is slightly larger for the 10th-percentile AL composite compared with the 90th-percentile one (2.77°C vs 2.65°C, respectively), there is a large amount of scatter between the AL SLP and Niño-3.4 SST indices when considering all 2000 bootstrapped ENSO composite values (Fig. 6a), indicating that the precise values of any particular pair, and by extension the pair of spatial patterns shown in Fig. 5, are likely due to random chance. For example, for a given value of the AL index such as −9 hPa, which is close to the

value of the 90th-percentile sample (−8.7 hPa), there is a wide range of possible Niño-3.4 SST values (from 2.5° to 3.05°C) across the 2000 bootstrapped ENSO composites. Thus, the difference in Niño-3.4 SST values (0.1°C) between the 10th- and 90th-percentile AL composites is unlikely to be the cause of the nearly twofold difference in magnitude of their AL values (−4.8 vs −8.7 hPa). Conversely, for a given value of the Niño-3.4 SST index, say 2.7°C, the AL index can range from −12 hPa to −3 hPa, which exceeds the difference between the 10th- and 90th-percentile AL samples. Similar remarks apply to the 10th- and 90th-percentile composites based on the AR SLP index (Fig. 6b). Finally, although there is a weak linear dependence of the AL index on the Niño-3.4 SST index across the 2000 bootstrapped ENSO composites (correlation coefficient = 0.24), and to a lesser extent of the AR index on Niño-3.4 (correlation coefficient = 0.09), removing this dependency via linear regression analysis has virtually no effect on the results (not shown), underscoring that differences between the 2000 individual bootstrapped composites are unlikely to be the result of sampling slightly different sets of ENSO events.

To extend this analysis to the full NH domain, we show the contribution to the observed CI that arises from the linear dependence of the SLP composite values at each grid box upon the Niño-3.4 composite values across the 2000 bootstrapped samples, expressed in units of hectopascals (Fig. 7a) and as a percentage of the total CI (Fig. 7b). To obtain this "ENSO contribution," we first compute the CIs using the 2000 SLP values of the bootstrapped composites from which the Niño-3.4 SST index has been linearly removed via regression analysis and then subtract it from the original CIs. The ENSO contribution to the CIs is <0.5 hPa everywhere except the Gulf of Alaska, where it reaches 1–1.5 hPa (Fig. 7a) or 10%–15% of the total CI (Fig. 7b). The southeastern United States also shows maximum values of 10%–15%, and the low-latitude western Pacific reaches 5%–10% (Fig. 7b). These results demonstrate that, while there is some effect associated with sampling different sets of EN and LN events in the observed bootstrapped composites, it does not make a large contribution to the uncertainty in the SLP ENSO composites. That is, the diversity of pattern, amplitude, and relative spatial emphasis among the SLP composite maps shown in Figs. 4 and 5 is primarily due to (atmospheric) internal variability rather than slightly different samples of ENSO events. In this context, it is worth recalling that a similar diversity is apparent across the individual members of the pacemaker ensembles for which the set of ENSO events is identical.

Finally, we address the issue of whether differences in the character of El Niño events as portrayed by the east Pacific (EP) and central Pacific (CP) dichotomy
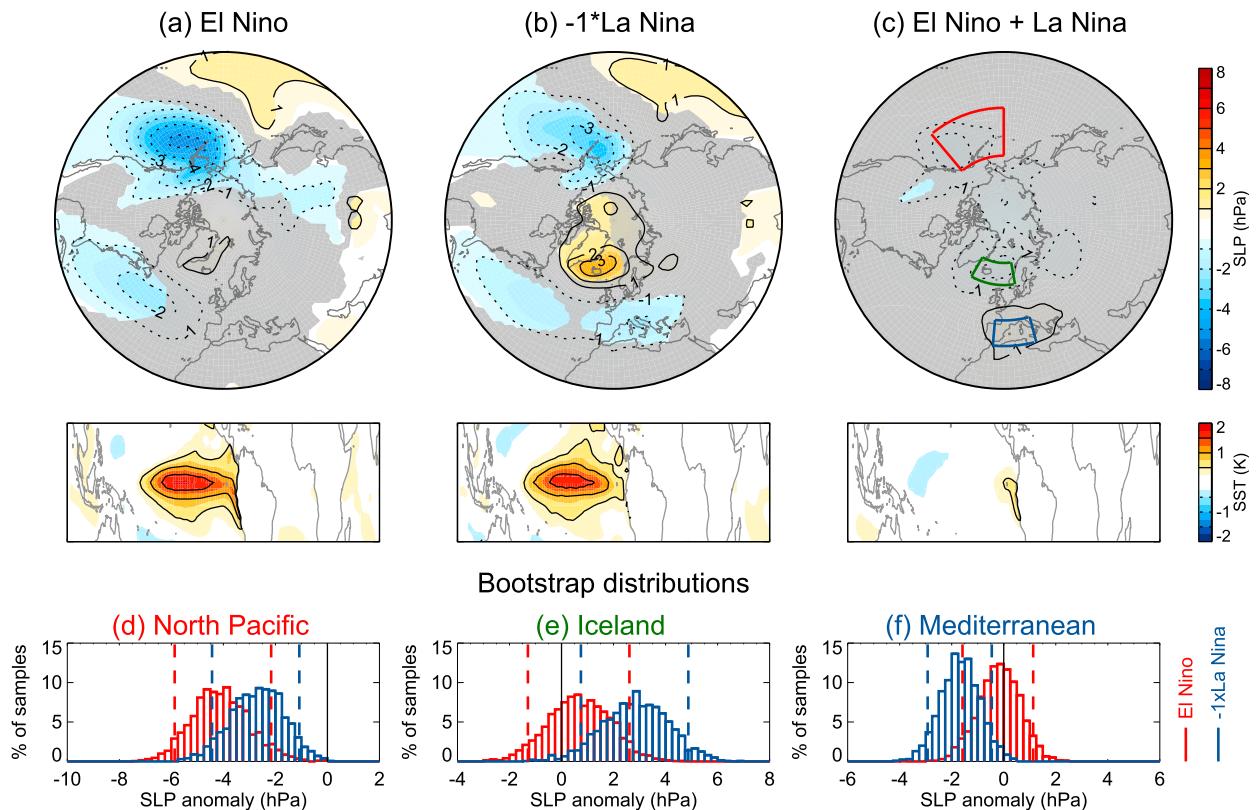
FIG. 10. Observed (a) El Niño and (b) −1 × La Niña composites of DJF (top) SLP (contour interval of 2 hPa) and (middle) SST (contour interval of 1°C) based on 18 and 14 events, respectively, during 1920–2013. (c) Sum of El Niño and La Niña composites. Gray shading in the top row indicates values not significant at the 5% confidence level based on a two-sided $t$ test. Histograms of the 2000 bootstrapped SLP composite values (hPa) for the (d) Aleutian low, (e) Icelandic Low, and (f) Mediterranean region [areas outlined in (c)]. Red and blue bars are for El Niño and −1 × La Niña, respectively. Dashed vertical lines indicate the 5% and 95% values.

(e.g., Graf and Zanchettin 2012; Yu et al. 2012, 2015) may affect the uncertainty on our observed ENSO composite by constructing CIs based on two additional 2000-member sets of bootstrapped composites. These differ from the original set by restricting the random sampling of all 18 El Niño events to those that fall in the EP category (7) and to those that fall in the CP category (11); nothing is changed for the sampling of La Niña events. Note that we maintain a total of 18 El Niño events in these CP and EP bootstrapped composites for consistency with the original "all El Niño" bootstrapped composites. The CI maps based on the 2000 CP and 2000 EP bootstrapped composites are very similar in both pattern and amplitude (Fig. A6). The EP set (Fig. A6b) displays maximum values that are ~15% larger than those in the CP set (Fig. A6a), but these differences are within the range of what could be expected by chance, based upon an examination of the CIs from individual pacemaker simulations that all have the same ENSO events (Fig. A5). Similar results are obtained using the east Pacific nonconvecting

(EPN) and east Pacific convecting (EPC) dichotomy of El Niño events defined by Johnson and Kosaka (2016; not shown). Taken together, the results shown in Figs. 6 and 7 (see also Fig. A6), reinforce the notion that ENSO diversity, either in the form of different magnitudes of ENSO events or in the form of different "flavors" of El Niño, does not have an appreciable effect on our quantification of uncertainty for the observed ENSO SLP composite.

### d. Comparison across models

We summarize the amplitudes of AL and AR SLP indices across all 2000 bootstrapped ENSO composites from observations and models in Fig. 8. Each panel shows the histograms of the composite SLP anomaly values from the 2000 bootstrapped samples (open bars) as well as the actual values as determined from observations or the individual model ensemble members (red bars). Note that there is only one actual value for observations, 10 actual values from each of the CESM1 and GFDL CM2.1 pacemaker and CESM1 TOGA ensembles, and 5 actual
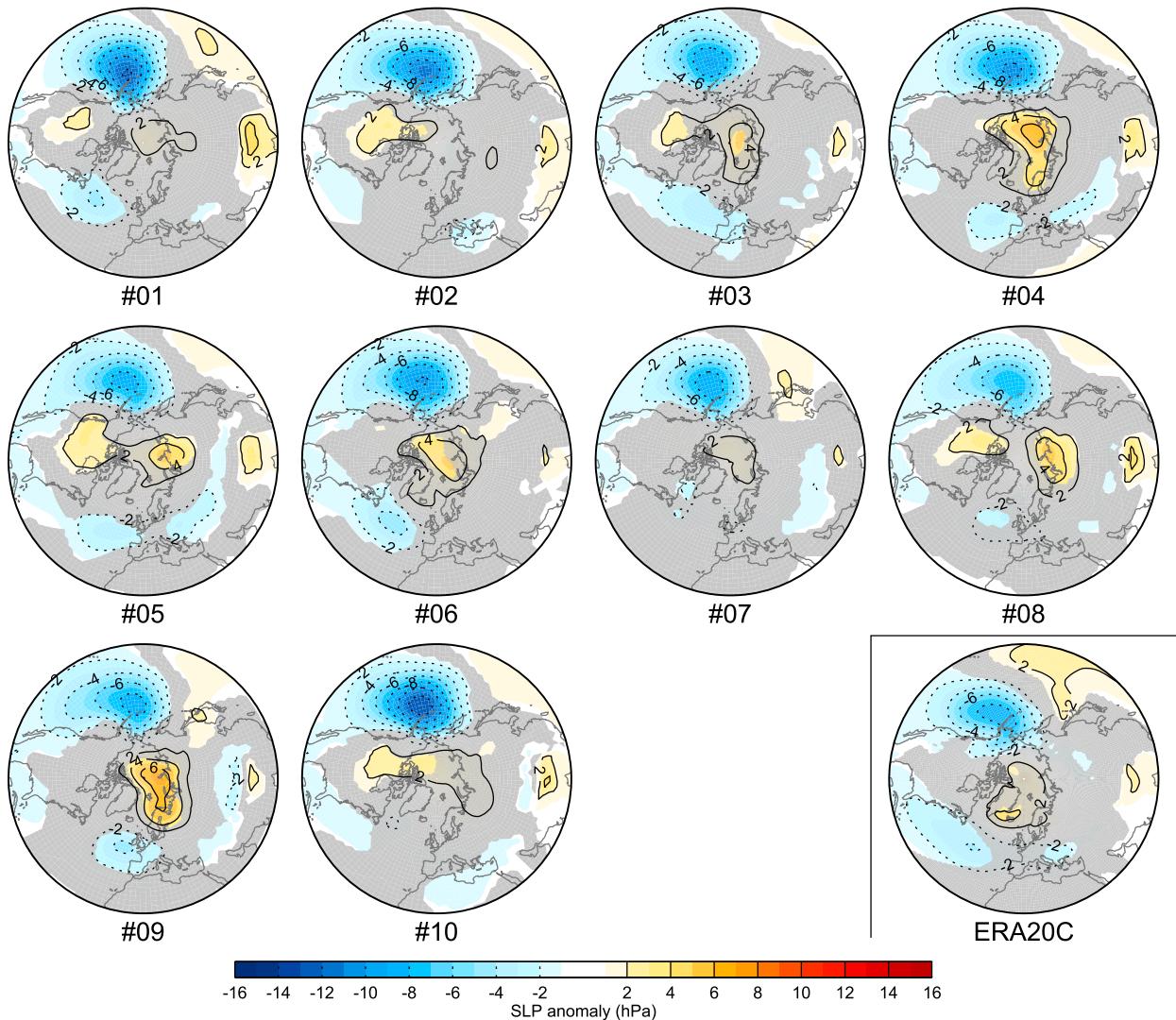
FIG. A1. ENSO composites of DJF SLP from each of the 10 GFDL CM2.1 pacemaker simulations (labeled #01, . . . , #10) and from (bottom right) ERA-20C. Each composite is based on the same set of 18 El Niño events minus 14 La Niña events during 1920–2013. Values not significant at the 5% confidence level based on a two-sided *t* test are shaded in gray. Contour interval is 2 hPa.

values from the MIROC5 pacemaker ensemble. The associated CI range based on the 2000 bootstrapped samples is indicated by the horizontal blue bar above each dataset. We note that, while the observed value must lie in the middle of its bootstrapped samples by construction, this need not be the case for the models since their bootstrapped samples were constructed by drawing from among all ensemble members (although the average of the individual ensemble members will lie at the peak of the distribution of the bootstrapped samples for a given model). As expected based on the results already presented, the observed and CESM1 pacemaker distributions are very similar in terms of width, mean value, and CI for both SLP indices, as indicated by the nearly complete

overlap in their CIs. Similar results are found for the GFDL CM2.1 and MIROC5 pacemaker ensembles and the CESM1 TOGA ensemble, with the possible exception of the AR index in MIROC5. This portrayal of the bootstrapped ENSO composites highlights the need for large model ensembles, since a single simulation from a particular model can alter the mean value (and to a lesser extent the CI) of the distribution just by chance.

To further explore the influence of a limited sample size on uncertainty in ENSO composites, we resample the full pacemaker ensemble using prespecified numbers of EN and LN events. The 5%–95% range of the AL and AR indices as a function of the number of events used for the ENSO composite is shown in Fig. 9. With only two
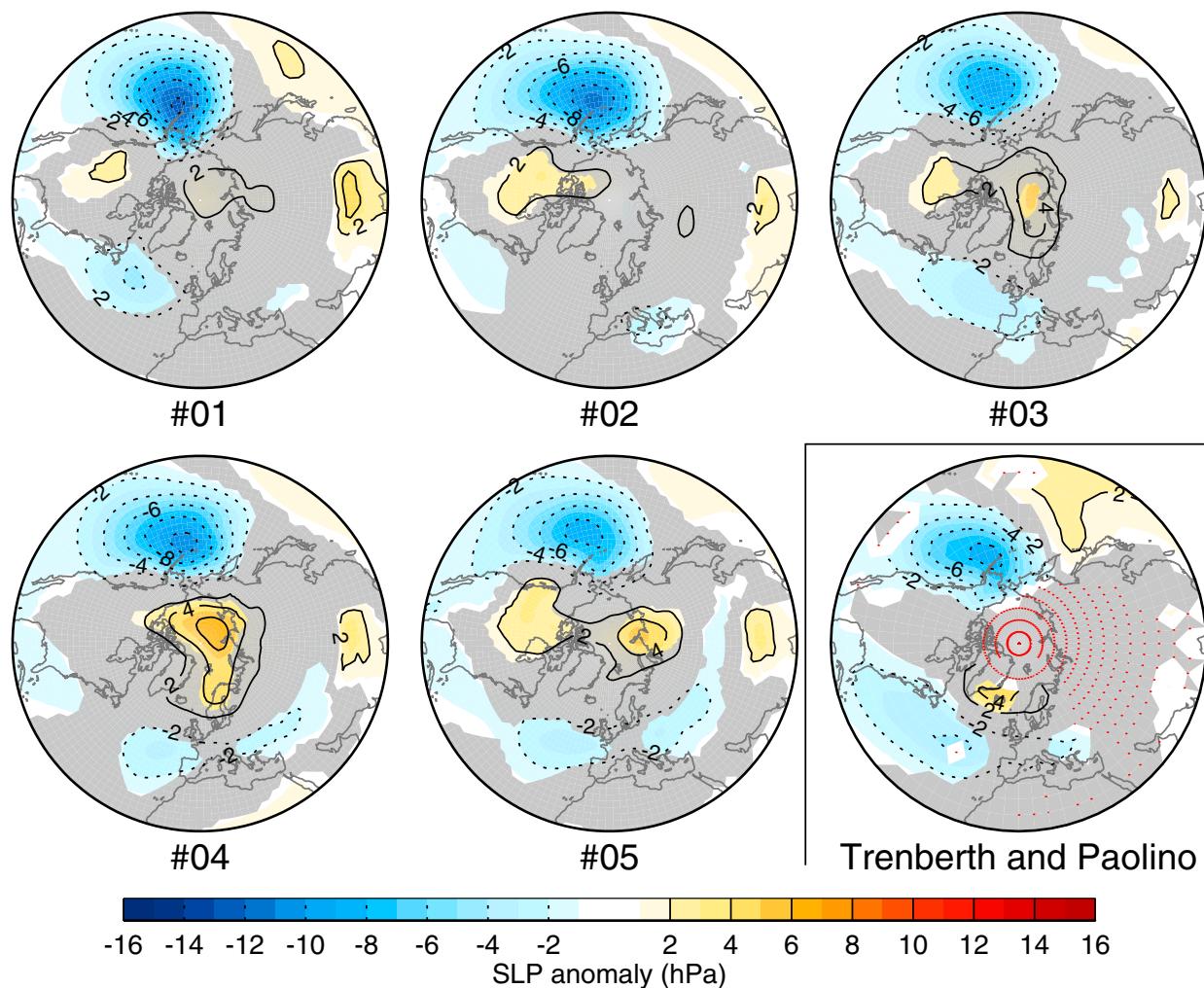
FIG. A2. ENSO composites of DJF SLP from each of the 5 MIROC5 pacemaker simulations (labeled #01, . . . , #05) and (bottom right) from observations based on the Trenberth and Paolino (1980) dataset (missing data indicated by red stippling). Each composite is based on the same set of 18 El Niño events minus 14 La Niña events during 1920–2013. Values not significant at the 5% confidence level based on a two-sided $t$ test are shaded in gray. Contour interval is 2 hPa.

events of each sign (i.e., two EN and two LN), the CI on the AL index ranges from −15 to 0 hPa, and thus the AL response to ENSO (−7 hPa) is not distinguishable from zero (Fig. 9a). With four events of each sign, the 5%–95% range of the bootstrapped samples remains wide (from −12.5 to −3 hPa) but no longer includes zero (Fig. 9a). The 5%–95% CI range diminishes slowly with increasing sample size beyond approximately 10 events of each sign, with values from −9.5 to −5 hPa for a sample size of 24; this sample size is well beyond that of the 16 in our study period 1920–2013 (Fig. 9a). In contrast, the 5%–95% CI range on the AR index includes zero even at a sample size of 24, such that it is more difficult to determine with high confidence that the forced trend is positive (Fig. 9b). The CI ranges on the AL and AR indices for a sample size of 16

are consistent with the results shown in Fig. 3d, noting that the region used to define the AR index extends beyond the area with a significant ensemble-mean response (Fig. 3a).

### e. El Niño versus La Niña composites

Up to now, we have focused on the linear component of ENSO. Here we examine whether there are any appreciable nonlinearities by examining composites of the 18 EN events and the 14 LN events separately. The observed composite SLP patterns for EN and LN, shown in Figs. 10a and 10b, respectively (with the sign inverted for LN for ease of comparison), are largely similar, with negative anomalies over the northeast Pacific and across the Atlantic, and positive anomalies over the southwest Pacific and the Arctic. All aforementioned features differ
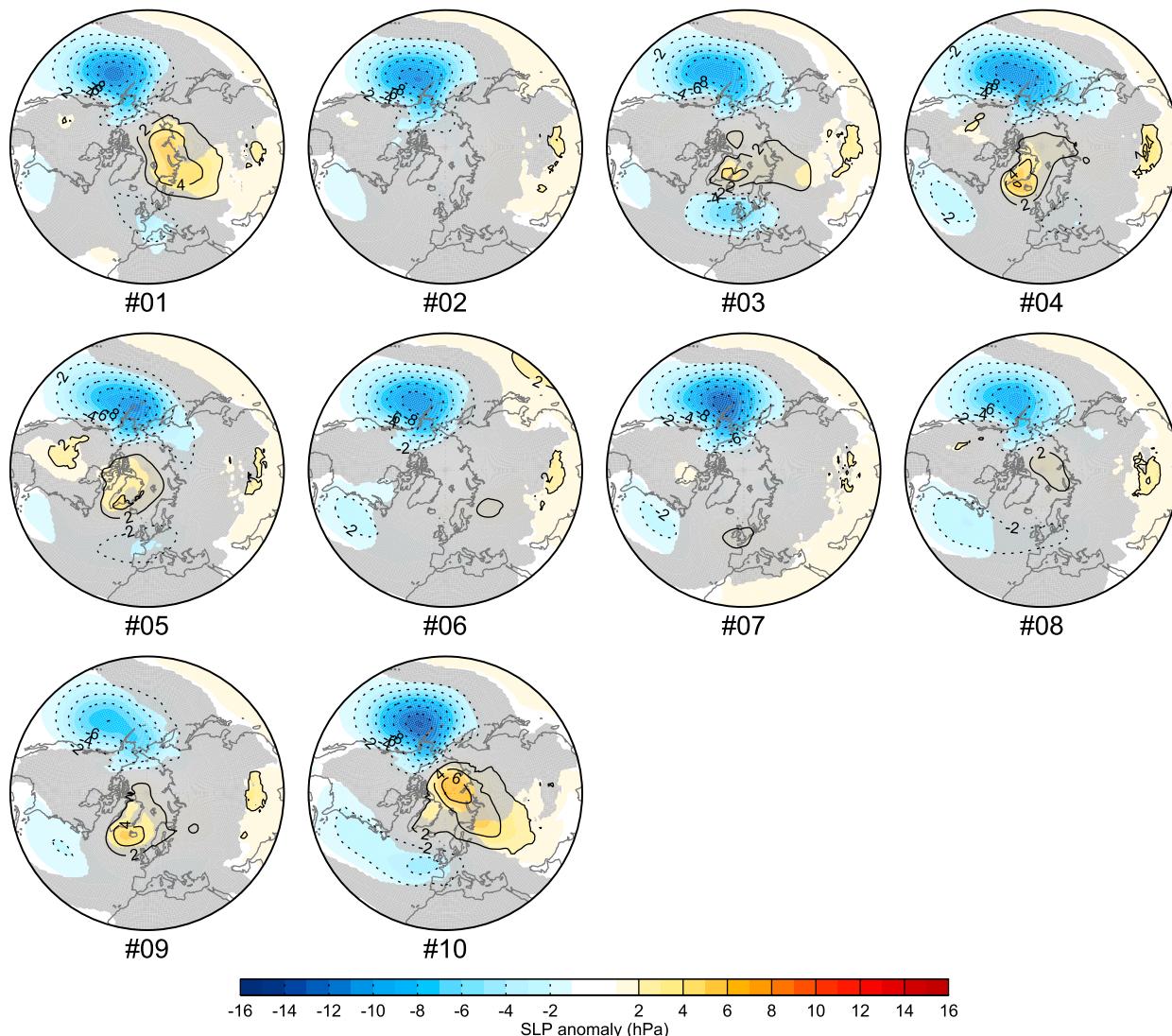
FIG. A3. As in Fig. A1, but for the 10 CAM5 TOGA simulations.

significantly from zero at the 5% confidence level, except for the Arctic anomaly in the EN composite. While the EN and LN composites show regional differences in amplitude, they exhibit no significant nonlinearities, as shown by their sum in Fig. 10c. Notably, the presence of a significant NAO-like SLP dipole in the (inverted) LN composite, consisting of positive anomalies in the region around Iceland and negative anomalies over the eastern Atlantic extending into Europe, a feature that is largely absent in the EN composite, cannot be interpreted as a significant nonlinearity based on the set of events analyzed here (Fig. 10c). The lack of significant nonlinearity in the observed EN and LN composites is further confirmed by the substantial overlap in their regional SLP distributions based on 2000 bootstrapped samples for both EN and LN (Fig. 10, bottom row). In particular, the bootstrapped SLP composites

averaged over the Aleutian low (Fig. 10d), Icelandic (Fig. 10e), and Mediterranean (Fig. 10f) regions (areas outlined in Fig. 10c) exhibit relatively broad distributions whose 5%–95% CIs overlap. For completeness, we include maps of the observed tropical SST anomalies in each composite (Fig. 10, middle row). The (inverted) LN SST composite shows a slightly weaker maximum value in the central Pacific, a westward extension toward New Guinea, and weaker anomalies off South America compared with EN, as found in previous studies (e.g., Okumura and Deser 2010).

## 4. Summary and discussion

Even with nearly 100 years of data comprising 18 El Niño and 14 La Niña events, we have shown that
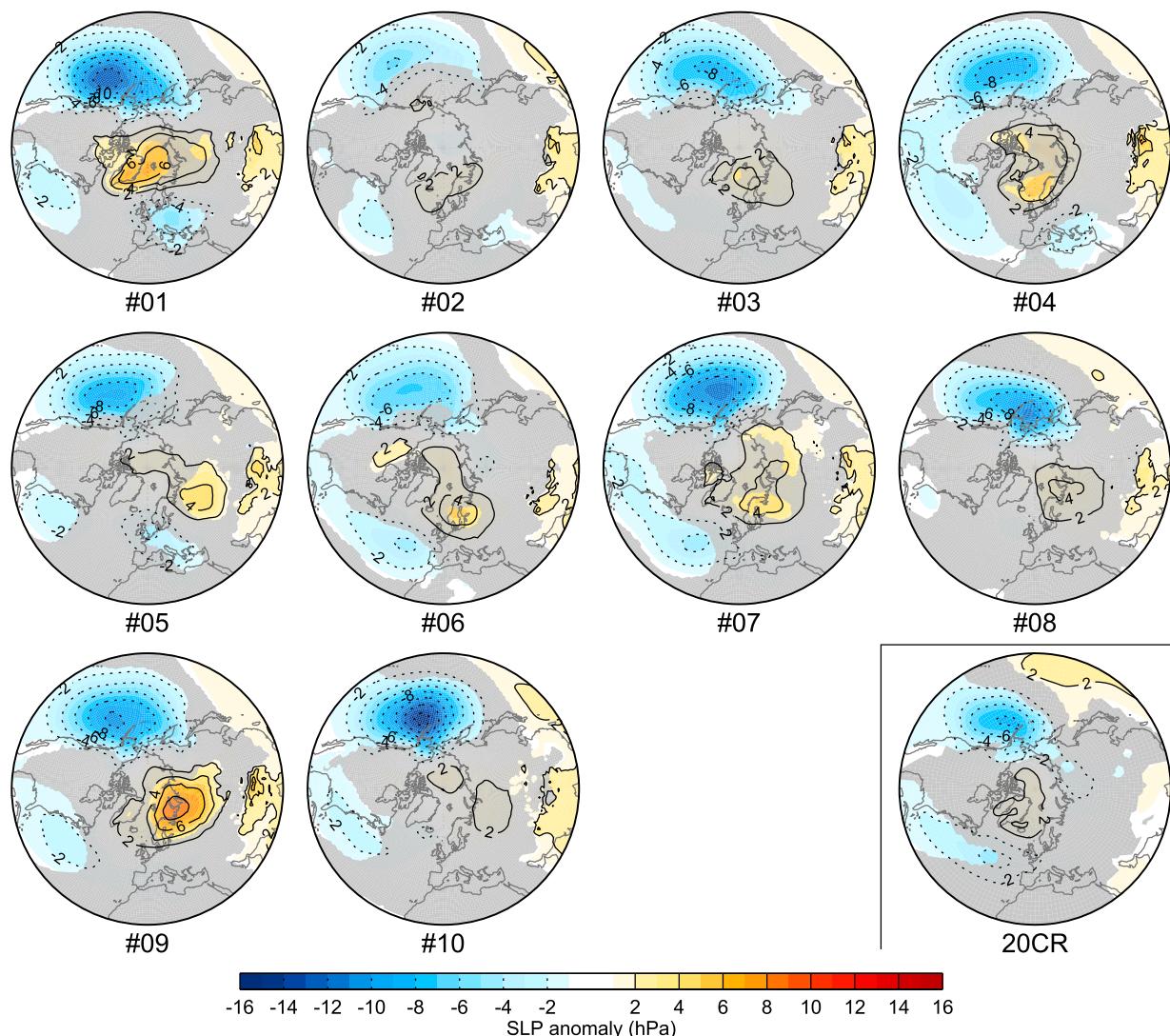
FIG. A4. As in Fig. 2, but for the period 1950–2013.

the observed extratropical NH SLP response to ENSO in boreal winter (DJF) is subject to considerable uncertainty in pattern and amplitude as a result of sampling fluctuations associated with variability that is independent of ENSO and likely atmospheric in origin. Our results are based on composite differences between the 18 El Niño and 14 La Niña events observed during the period 1920–2013 and 2000 synthetic composites obtained by randomly sampling these events in various combinations and frequencies. The observed SLP composite shows a robust ENSO response over the North Pacific and North America that is statistically significant (different from zero) in all 2000 synthetic composites, but its amplitude is uncertain by approximately a factor of 2. Other regions, such as the Arctic, North Atlantic, and Europe, show a larger range of patterns, amplitudes,

and statistical significance across the bootstrapped samples. Although the synthetic ENSO composites are based on different combinations of El Niño and La Niña events, we find that ENSO diversity makes only a minor contribution to their spread and thus to the uncertainty in the actual observed ENSO SLP composite. We also find no significant nonlinearities between the SLP composites for El Niño and La Niña separately. Our results pose considerable challenges for the evaluation of ENSO teleconnections in models. In particular, uncertainty in the pattern and amplitude of the observational target necessitates an approach to model assessment that considers not only the forced ENSO response but also the internal variability.

In addition to our observational analysis, we have investigated ENSO teleconnections in several large
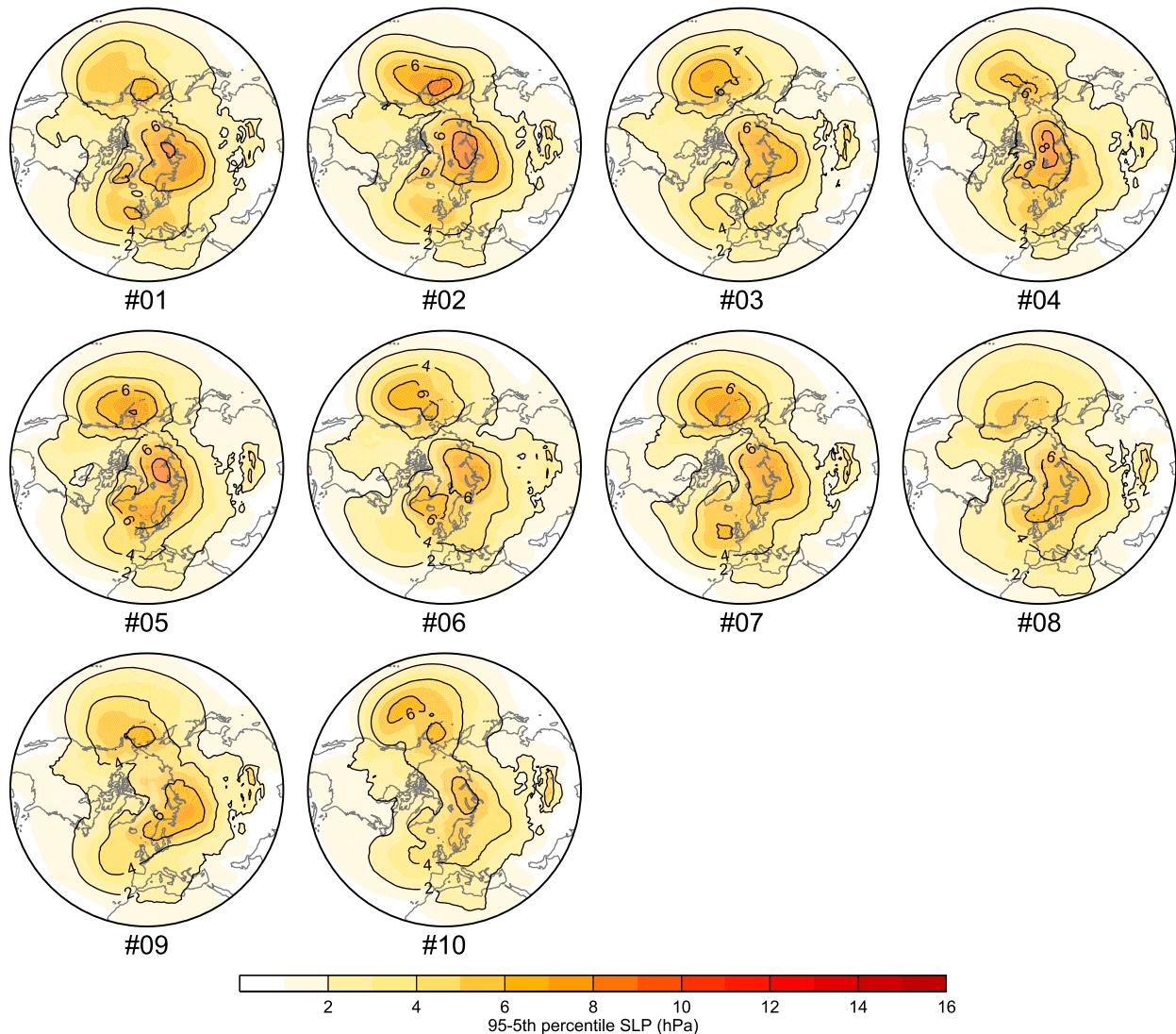
FIG. A5. The 5%–95% CIs (hPa) on the DJF SLP ENSO composites in each of the 10 CESM1 pacemaker simulations (contour interval = 2 hPa), based on 2000 bootstrapped samples for each.

initial-condition ensembles with different models and configurations of ocean–atmosphere coupling. Each of these simulations is subject to the observed evolution of tropical Pacific SST anomalies during 1920–2013. As for the observations, we formed composite differences between the 18 El Niño events and 14 La Niña events for each model simulation. The ENSO SLP composites from the model ensembles show a similar diversity of pattern and amplitude as that found in the observationally derived synthetic composites. Because each composite within a given model ensemble is based on the same observed set of ENSO events (unlike the synthetic composites for observations), their diversity is entirely a result of sampling fluctuations of unrelated variability, which we have shown arises primarily from atmospheric

processes. These results are robust across the three coupled model ensembles we have analyzed (CESM1, GFDL CM2.1, and MIROC5) and between the atmosphere-only and coupled configurations of CESM1.

Finally, we have demonstrated an approach for evaluating ENSO teleconnections in models that incorporates both pattern and amplitude uncertainty in the observational target and allows for discrimination between true model biases in the forced response to ENSO and apparent model biases that arise from limited sampling of internal variability unrelated to ENSO. This approach has mutual benefits for both observational analyses and model evaluation of ENSO teleconnections. In particular, the spread across the model ensembles informs the interpretation of the single observed ENSO

## (a) Sampling from CP



## (b) Sampling from EP



SLP (hPa)

FIG. A6. As in Fig. 3e, but using only the (a) 11 CP and (b) 7 EP El Niño events to form the observational bootstrapped ENSO composites. See text for details.

of a southwestward-expanded SLP response over the North Pacific and a weaker meridional dipole response over the North Atlantic.

Our statistical approach to the analysis of ENSO teleconnections in models and observations, while illuminating, does not alleviate the need for a more dynamically oriented investigation. In particular, a dynamical assessment of model biases in Rossby wave excitation by tropical convection and subsequent propagation characteristics mediated by eddy–mean flow interactions would serve to advance physical understanding and predictability of ENSO-induced circulation anomalies around the globe and their attendant impacts on surface climate.

### APPENDIX

### Additional Model Results

Figures A1–A6 show additional results from the model simulations as discussed in the main text.

#### REFERENCES

Alexander, M. A., I. Blade, M. Newman, J. Lanzante, N.-C. Lau, and J. D. Scott, 2002: The atmospheric bridge: The influence of ENSO teleconnections on air–sea interaction over the global oceans. *J. Climate*, **15**, 2205–2231, doi:10.1175/1520-0442(2002)015<2205:TABTIO>2.0.CO;2.

Barnston, A. G., M. Chelliah, and S. B. Goldenberg, 1997: Documentation of a highly ENSO-related SST region in the equatorial Pacific. *Atmos.–Ocean*, **35**, 367–383, doi:10.1080/07055900.1997.9649597.

——, M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt, 2012: Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing? *Bull. Amer. Meteor. Soc.*, **93**, 631–651, doi:10.1175/BAMS-D-11-00111.1.

Cagnazzo, C., and E. Manzini, 2009: Impact of the stratosphere on the winter tropospheric teleconnections between ENSO and the North Atlantic and European region. *J. Climate*, **22**, 1223–1238, doi:10.1175/2008JCLI2549.1.

Capotondi, A., and Coauthors, 2015: Understanding ENSO diversity. *Bull. Amer. Meteor. Soc.*, **96**, 921–938, doi:10.1175/BAMS-D-13-00117.1.

composite, and the single observed ENSO composite informs the evaluation of model biases in both the forced response and internal variability. In the case of CESM1, we have shown that the model's internal SLP variability is generally realistic over the extratropical NH in DJF, which in turn allows us to determine biases in its forced ENSO response. The latter consists mainly

Compo, G. P., and Coauthors, 2011: The Twentieth Century Re-analysis Project. *Quart. J. Roy. Meteor. Soc.*, **137**, 1–28, doi:10.1002/qj.776.

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

Deser, C., M. A. Alexander, S. -P. Xie, and A. S. Phillips, 2010: Sea surface temperature variability: Patterns and mecha-nisms. *Annu. Rev. Mar. Sci.*, **2**, 115–143, doi:10.1146/annurev-marine-120408-151453.

——, A. S. Phillips, V. Bourdette, and H. Teng, 2012a: Uncertainty in climate change projections: The role of internal variability. *Climate Dyn.*, **38**, 527–546, doi:10.1007/s00382-010-0977-x.

——, and Coauthors, 2012b: ENSO and Pacific decadal variability in Community Climate System Model version 4. *J. Climate*, **25**, 2622–2651, doi:10.1175/JCLI-D-11-00301.1.

DeWeaver, E., and S. Nigam, 2002: Linearity in ENSO's atmo-spheric response. *J. Climate*, **15**, 2446–2461, doi:10.1175/1520-0442(2002)015<2446:LIESAR>2.0.CO;2.

Dunstone, N., D. Smith, A. Scaife, L. Hermanson, R. Eade, N. Robinson, M. Andrews, and J. Knight, 2016: Skilful pre-dictions of the winter North Atlantic Oscillation one year ahead. *Nat. Geosci.*, **9**, 809–814, doi:10.1038/ngeo2824.

Exner, F. M., 2015: On monthly weather anomalies in the Northern Hemisphere in winter. *Meteor. Z.*, **24**, 107–111, doi:10.1127/metz/2015/0654.

Fan, T., C. Deser, and D. P. Schneider, 2014: Recent Antarctic sea ice trends in the context of Southern Ocean surface climate variations since 1950. *Geophys. Res. Lett.*, **41**, 2419–2426, doi:10.1002/2014GL059239.

Feng, J., W. Chen, and Y. Li, 2016: Asymmetry of the winter extra-tropical teleconnections in the Northern Hemisphere associ-ated with two types of ENSO. *Climate Dyn.*, **48**, 2135–2151, doi:10.1007/s00382-016-3196-2.

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

Garfinkel, C. I., and D. L. Hartmann, 2008: Different ENSO tele-connections and their effects on the stratospheric polar vortex. *J. Geophys. Res.*, **113**, D18114, doi:10.1029/2008JD009920.

——, and ——, 2010: The influence of the quasi-biennial oscillation on the North Pacific and El Niño teleconnections. *J. Geophys. Res.*, **115**, D20116, doi:10.1029/2010JD014181.

——, M. M. Hurwitz, D. W. Waugh, and A. H. Butler, 2013: Are the teleconnections of central Pacific and eastern Pacific El Niño distinct in boreal wintertime? *Climate Dyn.*, **41**, 1835–1852, doi:10.1007/s00382-012-1570-2.

Gonzalez, P. M., and L. Goddard, 2016: Long-lead ENSO pre-dictability from CMIP5 decadal hindcasts. *Climate Dyn.*, **46**, 3127–3147, doi:10.1007/s00382-015-2757-0.

Graf, H.-F., and D. Zanchettin, 2012: Central Pacific El Niño, the "subtropical bridge," and Eurasian climate. *J. Geophys. Res.*, **117**, D01102, doi:10.1029/2011JD016493.

Hegyi, B. M., and Y. Deng, 2011: A dynamical fingerprint of tropical sea surface temperatures on the decadal-scale vari-ability of cool-season Arctic precipitation. *J. Geophys. Res.*, **116**, D20121, doi:10.1029/2011JD016001.

Held, I. M., S. W. Lyons, and S. Nigam, 1989: Transients and the extratropical response to El Niño. *J. Atmos. Sci.*, **46**, 163–174, doi:10.1175/1520-0469(1989)046<0163:TATERT>2.0.CO;2.

Hoerling, M. P., and A. Kumar, 1997: Why do North American climate anomalies differ from one El Niño event to another? *Geophys. Res. Lett.*, **24**, 1059–1062, doi:10.1029/97GL00918.

——, ——, and M. Zhong, 1997: El Niño, La Niña, and the nonlinearity of their teleconnections. *J. Climate*, **10**, 1769–1786, doi:10.1175/1520-0442(1997)010<1769:ENOLNA>2.0.CO;2.

——, ——, and T. Xu, 2001: Robustness of the nonlinear climate response to ENSO's extreme phases. *J. Climate*, **14**, 1277–1293, doi:10.1175/1520-0442(2001)014<1277:ROTNCR>2.0.CO;2.

Horel, J. D., and J. M. Wallace, 1981: Planetary-scale atmospheric phenomena associated with the Southern Oscillation. *Mon. Wea. Rev.*, **109**, 813–829, doi:10.1175/1520-0493(1981)109<0813:PSAPAW>2.0.CO;2.

Hoskins, B. J., and D. 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.

——, and T. Ambrizzi, 1993: Rossby wave propagation on a realistic longitudinally varying flow. *J. Atmos. Sci.*, **50**, 1661–1671, doi:10.1175/1520-0469(1993)050<1661:RWPOAR>2.0.CO;2.

Hurrell, J. W., Y. Kushnir, M. Visbeck, and G. Ottersen, 2003: An overview of the North Atlantic Oscillation. *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*, Geophys. Monogr., Vol. 134, Amer. Geophys. Union, 1–35, doi:10.1029/134GM01.

Ineson, S., and A. Scaife, 2009: The role of the stratosphere in the European climate response to El Niño. *Nat. Geosci.*, **2**, 32–36, doi:10.1038/ngeo381.

Infanti, J. M., and B. P. Kirtman, 2016: North American rainfall and temperature prediction response to the diversity of ENSO. *Climate Dyn.*, **46**, 3007–3023, doi:10.1007/s00382-015-2749-0.

Johnson, N. C., and Y. Kosaka, 2016: The impact of eastern equatorial Pacific convection on the diversity of boreal winter El Niño teleconnection patterns. *Climate Dyn.*, **47**, 3737–3765, doi:10.1007/s00382-016-3039-1.

Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) Large Ensemble Project: A community re-source for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.*, **96**, 1333–1349, doi:10.1175/BAMS-D-13-00255.1.

Kirtman, and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal to interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bull. Amer. Me-teor. Soc.*, **95**, 585–601, doi:10.1175/BAMS-D-12-00050.1.

Kosaka, Y., and S.-P. Xie, 2013: Recent global-warming hiatus tied to equatorial Pacific surface cooling. *Nature*, **501**, 403–407, doi:10.1038/nature12534.

Kumar, A., Z.-Z. Hu, B. Jha, and P. Peng, 2016: Estimating ENSO predictability based on multi-model hindcasts. *Climate Dyn.*, **48**, 39–51, doi:10.1007/s00382-016-3060-4.

Lau, N.-C., and M. J. Nath, 1994: A modeling study of the relative roles of tropical and extratropical SST anomalies in the variability of the global atmosphere–ocean system. *J. Climate*, **7**, 1184–1207, doi:10.1175/1520-0442(1994)007<1184:AMSOTR>2.0.CO;2.

L'Heureux, M. L., and D. W. J. Thompson, 2006: Observed re-lationships between the El Niño–Southern Oscillation and the extratropical zonal-mean circulation. *J. Climate*, **19**, 276–287, doi:10.1175/JCLI3617.1.

——, M. K. Tippett, and A. G. Barnston, 2015: Characterizing ENSO coupled variability and its impact on North American seasonal precipitation and temperature. *J. Climate*, **28**, 4231–4245, doi:10.1175/JCLI-D-14-00508.1.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Lu, J., G. Chen, and D. Frierson, 2008: Response of the zonal mean atmospheric circulation to El Niño versus global warming. *J. Climate*, **21**, 5835–5851, doi:10.1175/2008JCLI2200.1.

Manzini, E., M. A. Giorgetta, L. Kornbluth, and E. Roeckner, 2006: The influence of sea surface temperatures on the northern winter stratosphere: Ensemble simulations with the MAECHAM5 model. *J. Climate*, **19**, 3863–3881, doi:10.1175/JCLI3826.1.

McGraw, M. C., E. A. Barnes, and C. Deser, 2016: Reconciling the observed and modeled Southern Hemisphere circulation response to volcanic eruptions. *Geophys. Res. Lett.*, **43**, 7259–7266, doi:10.1002/2016GL069835.

McPhaden, M. J., A. Timmermann, M. J. Widlansky, M. A. Balmaseda, and T. N. Stockdale, 2015: The curious case of the El Niño that never happened: A perspective from 40 years of progress in climate research and forecasting. *Bull. Amer. Meteor. Soc.*, **96**, 1647–1665, doi:10.1175/BAMS-D-14-00089.1.

Neelin, J. D., 2012: *Climate Change and Climate Modeling.* Cambridge University Press, 304 pp.

Nishii, K., H. Nakamura, and Y. J. Orsolini, 2010: Cooling of the wintertime Arctic stratosphere induced by the western Pacific teleconnection pattern. *Geophys. Res. Lett.*, **37**, L13805, doi:10.1029/2010GL043551.

Okumura, Y. M., and C. Deser, 2010: Asymmetry in the duration of El Niño and La Niña. *J. Climate*, **23**, 5826–5843, doi:10.1175/2010JCLI3592.1.

Philander, S. G., 1990: *El Niño, La Niña and the Southern Oscillation.* Academic Press, 293 pp.

Poli, P., and Coauthors, 2016: ERA-20C: An atmospheric reanalysis of the twentieth century. *J. Climate*, **29**, 4083–4097, doi:10.1175/JCLI-D-15-0556.1.

Raible, C. C., F. Lehner, J. F. González-Rouco, and L. Fernández-Donado, 2014: Changing correlation structures of the Northern Hemisphere atmospheric circulation from 1000 to 2100 AD. *Climate Past*, **10**, 537–550, doi:10.5194/cp-10-537-2014.

Rayner, N. A., and Coauthors, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, doi:10.1029/2002JD002670.

Richter, J. H., C. Deser, and L. Sun, 2015: Effects of stratospheric variability on El Niño teleconnections. *Environ. Res. Lett.*, **10**, 124021, doi:10.1088/1748-9326/10/12/124021.

Ropelewski, C. F., and M. S. Halpert, 1987: Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Wea. Rev.*, **115**, 1606–1626, doi:10.1175/1520-0493(1987)115<1606:GARSPP>2.0.CO;2.

Sardeshmukh, P. D., and B. J. Hoskins, 1988: The generation of global rotational flow by steady idealized tropical divergence. *J. Atmos. Sci.*, **45**, 1228–1251, doi:10.1175/1520-0469(1988)045<1228:TGOGRF>2.0.CO;2.

——, G. Compo, and C. Penland, 2000: Changes of probability associated with El Niño. *J. Climate*, **13**, 4268–4286, doi:10.1175/1520-0442(2000)013<4268:COPAWE>2.0.CO;2.

Scaife, A. A., and Coauthors, 2014: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514–2519, doi:10.1002/2014GL059637.

Seager, R., N. Harnik, Y. Kushnir, W. Robinson, and J. Miller, 2003: Mechanisms of hemispherically symmetric climate variability.

*J. Climate*, **16**, 2960–2978, doi:10.1175/1520-0442(2003)016<2960:MOHSCV>2.0.CO;2.

Shukla, J., and Coauthors, 2000: Dynamical seasonal prediction. *Bull. Amer. Meteor. Soc.*, **81**, 2593–2606, doi:10.1175/1520-0477(2000)081<2593:DSP>2.3.CO;2.

Simmons, A. J., J. M. Wallace, and G. Branstator, 1983: Barotropic wave propagation and instability, and atmospheric teleconnection patterns. *J. Atmos. Sci.*, **40**, 1363–1392, doi:10.1175/1520-0469(1983)040<1363:BWPAIA>2.0.CO;2.

Spencer, H., and J. M. Slingo, 2003: The simulation of peak and delayed ENSO teleconnections. *J. Climate*, **16**, 1757–1774, doi:10.1175/1520-0442(2003)016<1757:TSOPAD>2.0.CO;2.

Stock, C., and Coauthors, 2015: Seasonal sea surface temperature anomaly prediction for coastal ecosystems. *Prog. Oceanogr.*, **137**, 219–236, doi:10.1016/j.pocean.2015.06.007.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Tippett, M. K., A. G. Barnston, and S. Li, 2012: Performance of recent multimodel ENSO forecasts. *J. Appl. Meteor. Climatol.*, **51**, 637–654, doi:10.1175/JAMC-D-11-093.1.

Trenberth, K. E., and D. A. Paolino, 1980: The Northern Hemisphere sea-level pressure data set: Trends, errors and discontinuities. *Mon. Wea. Rev.*, **108**, 855–872, doi:10.1175/1520-0493(1980)108<0855:TNHSLP>2.0.CO;2.

——, and J. M. Caron, 2000: The Southern Oscillation revisited: Sea level pressures, surface temperatures, and precipitation. *J. Climate*, **13**, 4358–4365, doi:10.1175/1520-0442(2000)013<4358:TSORSL>2.0.CO;2.

——, G. W. Branstator, D. Karoly, A. Kumar, N.-C. 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.

Walker, G. T., 1923: Correlation in seasonal variation of weather, VIII: A preliminary study of world weather. *Mem. India Meteor. Dep.*, **24**, 75–131.

Wallace, J. M., C. Deser, B. V. Smoliak, and A. S. Phillips, 2013: Attribution of climate change in the presence of internal variability. *Climate Change: Multidecadal and Beyond*, C. P. Chang et al., Eds., World Scientific Series on Asia-Pacific Weather and Climate, Vol. 6, World Scientific, 1–29.

Woodruff, S. D., H. F. Diaz, E. C. Kent, R. W. Reynolds, and S. J. Worley, 2008: The evolving SST record from ICOADS. *Climate Variability and Extremes During the Past 100 Years*, S. Brönnimann et al., Eds., Advances in Global Change Research, Vol. 33, Springer, 65–83, doi:10.1007/978-1-4020-6766-2_4.

Xie, F., J. P. Li, W. S. Tian, and J. Feng, 2012: Signals of El Niño Modoki in the tropical tropopause layer and stratosphere. *Atmos. Chem. Phys.*, **12**, 5259–5273, doi:10.5194/acp-12-5259.

Yu, B., X. Zhang, H. Lin, and J.-Y. Yu, 2015: Comparison of wintertime North American climate impacts associated with multiple ENSO indices. *Atmos.–Ocean*, **53**, 426–445, doi:10.1080/07055900.2015.1079697.

Yu, J.-Y., Y. Zou, S. T. Kim, and T. Lee, 2012: The changing impact of El Niño on US winter temperatures. *Geophys. Res. Lett.*, **39**, L15702, doi:10.1029/2012GL052483.