An "Observational Large Ensemble" to Compare Observed and Modeled Temperature Trend Uncertainty due to Internal Variability

KAREN A. MCKINNON

Advanced Study Program and Climate and Global Dynamics Division, National Center for Atmospheric Research, Boulder, Colorado

ANDREW POPPICK

Department of Mathematics and Statistics, Carleton College, Northfield, Minnesota

ETIENNE DUNN-SIGOUIN

Department of Earth and Environmental Science, Lamont Doherty Earth Observatory, Columbia University, New York, New York

CLARA DESER

Climate and Global Dynamics Division, National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 22 December 2016, in final form 6 June 2017)

ABSTRACT

Estimates of the climate response to anthropogenic forcing contain irreducible uncertainty due to the presence of internal variability. Accurate quantification of this uncertainty is critical for both contextualizing historical trends and determining the spread of climate projections. The contribution of internal variability to uncertainty in trends can be estimated in models as the spread across an initial condition ensemble. However, internal variability simulated by a model may be inconsistent with observations due to model biases. Here, statistical resampling methods are applied to observations in order to quantify uncertainty in historical 50-yr (1966–2015) winter near-surface air temperature trends over North America related to incomplete sampling of internal variability. This estimate is compared with the simulated trend uncertainty in the NCAR CESM1 Large Ensemble (LENS). The comparison suggests that uncertainty in trends due to internal variability is largely overestimated in LENS, which has an average amplification of variability of 32% across North America. The amplification of variability is greatest in the western United States and Alaska. The observationally derived estimate of trend uncertainty is combined with the forced signal from LENS to produce an "Observational Large Ensemble" (OLENS). The members of OLENS indicate the range of observationally constrained, spatially consistent temperature trends that could have been observed over the past 50 years if a different sequence of internal variability had unfolded. The smaller trend uncertainty in OLENS suggests that is easier to detect the historical climate change signal in observations than in any given member of LENS.

1. Introduction

Anthropogenic radiative forcing is associated with a range of observed climatic changes including increased near-surface air temperature. Observed temperatures, however, are a combination of a forced climate change signal and random sampling of internal variability (Hawkins and Sutton 2009; Solomon et al. 2011; Deser et al. 2012b, 2014; Screen et al. 2014). Evaluating the relative contribution of each is challenging using observations alone.

The forced and internal components of temperature trends can be more easily separated within "initial condition" ensembles of climate model simulations (Rowell 1998; Collins and Allen 2002; Deser et al. 2012b; Fischer et al. 2013; Kay et al. 2015; Sanderson et al. 2017; Hawkins et al. 2015). Such an ensemble is constructed

Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/JCLI-D-16-0905.s1.

Corresponding author: Karen A. McKinnon, karen.mckinnon@post.harvard.edu

DOI: 10.1175/JCLI-D-16-0905.1

using a single climate model and estimate of external forcing; however, the initial conditions of each ensemble member are perturbed randomly at the start of their integration. The resulting differences in behavior of the ensemble members can be interpreted as due to simulated internal variability alone. Similarly, the average across ensemble members provides an estimate of the forced response of the model.

In such model ensembles, the range of multidecadal *global* average temperature trends across the individual members is small compared to the trend induced by external forcing over the recent past (Dai et al. 2015) and in future projections (Deser et al. 2012a). In contrast, *regional*-scale temperature trends can be highly variable, primarily due to the influence of atmospheric circulation on temperature (e.g., Hurrell 1996; Deser et al. 2012b; Holmes et al. 2016). For example, Deser et al. (2016) showed that recent 50-yr linear trends in winter [December–February (DJF)] North American temperature across members of the NCAR CESM1 Large Ensemble (LENS) had a large spread around the ensemble mean. While the ensemble mean suggested a forced signal of warming primarily between 1° and 2°C $(50\,\mathrm{yr})^{-1}$ across North America, individual members could exhibit large regions of cooling of the same magnitude, or warming greater than 4°C $(50\,\mathrm{yr})^{-1}$.

How should observed trends be interpreted given the presence of internal variability? If an observed trend is found to be within the model spread simulated by an initial-condition ensemble such as LENS, one might conclude that the two are consistent with each other, and that any other member of the ensemble could have also been observed given a different sampling of internal variability. However, the spread in modeled trends due to the internal variability must be assessed in the context of model biases and uncertainties in external forcing (Hawkins and Sutton 2009; Collins et al. 2012; Forster et al. 2013). In particular, internal variability in the model and its influence on uncertainty in trends may be biased (Thompson et al. 2015), complicating the interpretation of spread across ensemble members.

A complementary approach to using initial condition ensembles is to rely on historical observations to simulate fields consistent with the covariance structure of the observations. This idea has been applied previously in stochastic weather generation (e.g., Wilks and Wilby 1999) and statistical downscaling of climate model output (e.g., Teutschbein and Seibert 2012). With respect to multidecadal trends, Thompson et al. (2015) suggested that the observations provide a strong constraint on uncertainty due to internal variability for temperature and precipitation over land. Here, we extend upon their work by using statistical resampling methods that largely preserve the spatial and temporal correlation structure of the observations in order to create a synthetic ensemble of winter temperatures in North America. Unlike an initial condition ensemble from a climate model, the synthetic ensemble cannot be used to estimate the forced response. Instead, the synthetic ensemble is used to provide an observationally derived estimate of the magnitude of uncertainty in surface temperature trends due to internal variability. The synthetic ensemble is also compared to LENS to identify model biases. Finally, the synthetic ensemble is combined with the forced response from LENS to create an ''Observational Large Ensemble'' (OLENS) with spatially consistent estimates of the range of temperature trends that could have been observed in the past 50 years due to internal variability alone.

2. Datasets and model output

Model output is from the NCAR Large Ensemble (LENS; Kay et al. 2015), which, at the time of writing, comprises 40 simulations of CESM1 spanning at least 1920–2100. The initial condition ensemble was constructed by adding random perturbations of order 10^{-14} K to the air temperature fields of a single parent simulation. The simulations are driven by historical forcing from 1920 to 2005 (Lamarque et al. 2010) and by the RCP8.5 scenario for the subsequent years (Meinshausen et al. 2011). The atmosphere in the simulations has a horizontal resolution of approximately 1°. We use both near-surface air temperature and sea level pressure (SLP) from the model.

Three different observational temperature datasets are used for comparison to LENS: Berkeley Earth Surface Temperature (BEST), available at 1° resolution (Rohde et al. 2013), NASA GISTEMP, available at 2° resolution (Hansen et al. 2010), and HadCRUT4, available at 5° resolution (Morice et al. 2012). Each dataset is produced from in situ temperature data through different methods of averaging and interpolation. The primary results in the manuscript rely on the BEST dataset, but results using all other datasets are similar and can be found in the online supplemental material (Figs. S1–S4).

We link regional-scale variability in temperature with circulation through analysis of SLP. Two reanalysis datasets are used as estimates of observational SLP: the NCEP–NCAR reanalysis, available at 2.5° resolution (Kalnay et al. 1996) and the Twentieth Century Reanalysis, version 2c (20CRv2c), available at 2° resolution (Compo et al. 2011).

For each model–observation comparison, either the model output or the observations are regridded to the coarser of the two grids using bilinear interpolation. Analysis focuses on average wintertime (DJF) temperatures over the 50-yr period spanning 1966–2015.

3. Trend model

The primary goal of our analysis is to quantify the uncertainty in observed trends due to internal variability, and compare the result to what is suggested by LENS. Such a task requires choosing an appropriate trend model. Uncertainty in the character of, and dynamical response to, regional radiative forcing precludes modeling the regional trends as a function of past radiative forcing (Shindell and Faluvegi 2009; Wang et al. 2016). Instead, consistent with previous work (Thompson et al. 2015; Deser et al. 2016), we use a linear-in-time model reflecting the fact that global radiative forcing increased approximately linearly between 1966 and 2015 (Prather et al. 2013). However, results are very similar when the forced trend is instead assumed to scale with the ensemble-mean global-mean temperature from LENS [methodology from Dai et al. (2015); see Figs. S5 and S6]. The use of a longer, or future, time period would likely require a more sophisticated trend model. The methods for uncertainty quantification described in the remainder of the paper remain applicable to other choices for modeling the forced trend.

The linear model is written as

$$T_y = \alpha + \beta y + \varepsilon_y, \tag{1}$$

where T_y is the seasonal mean temperature in year y at a given grid box. The term ε_y represents the internal variability around the linear trend, β. In particular, ε_y is assumed to be "unforced"—that is, independent of anthropogenic radiative forcing, with a mean of zero and constant variance, σ^2. Because the true forced trend, β, and the character of the variability, ε_y, are not known, the ordinary least squares (OLS) empirical estimate of the trend, denoted by $\hat{\beta}$, will typically be influenced by both. This has minimal effect on our ability to properly characterize the effect of ε_y, as demonstrated by the sensitivity tests discussed in section 5b. Our primary focus is on the uncertainty in the true value of β, the forced 50-yr time trend.

Uncertainty in β emerges due to the presence of internal variability, represented by ε_y, combined with the limited data record. For a given trend length (50 years in our case), uncertainty grows with larger internal variability and greater autocorrelation. Note that nonnegligible uncertainty in a 50-yr trend can emerge from even minimally autocorrelated data (e.g., Fig. 2 in Thompson et al. 2015). Information about both internal variability and autocorrelation are contained within the covariance matrix of ε_y. Given perfect knowledge of that covariance matrix, $\mathbf{\Sigma}$, the variance in the trend estimator can be calculated exactly as [also see line 2 of Eq. (A.23) in Weisberg (2005)]

$$\mathrm{var}(\hat{\beta}) = [(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}]_{2,2}, \tag{2}$$

where the indices (2,2) indicate the lower right-hand corner of the resulting matrix. The matrix $\mathbf{X}$ is a 50 × 2 matrix containing a column of ones and a column of the index of the years. In reality, we do not have perfect knowledge of $\mathbf{\Sigma}$, so we resort to other approaches to estimate $\mathrm{var}(\hat{\beta})$.

4. Internal variability and autocorrelation in observations and the CESM1 Large Ensemble

a. *Internal variability*

We first examine the internal variability of DJF temperatures across North America during 1966–2015 in the observations and LENS. The internal variability in the observations is estimated as the standard deviation of the detrended time series ($\hat{\sigma}_{\mathrm{obs}}$; Fig. 1a). The internal variability in LENS is estimated by the pooled standard deviation across ensemble members ($\hat{\sigma}_{\mathrm{LENS}}$; Fig. 1b). Both the observations and LENS exhibit larger DJF temperature variability at higher latitudes, with a band of high variability stretching from the Great Lakes to coastal Alaska. The most variable 10% of grid boxes in the observations exhibit time series with standard deviations ranging from 2.7° to 3.5°C, whereas those in LENS are larger at 3.3° to 4.3°C. The differences are smaller in less variable regions, with the least variable 10% of grid boxes in the observations having standard deviations of 0.34°–1.0°C versus 0.31°–1.4°C in LENS. Nevertheless, CESM1 still simulates greater *spatial* contrasts of the temporal standard deviations in DJF temperature. The spatial pattern of internal variability in LENS is also distinct from the observations, with local maxima in the interior west of the United States and in southwestern Alaska that are not present in the observations.

The larger internal variability in LENS is most easily seen by examining the ratio of the standard deviations of internal variability, $\hat{\sigma}_{\mathrm{LENS}}$ to $\hat{\sigma}_{\mathrm{obs}}$ (Fig. 1c). Internal variability in LENS is larger than that in the observations across the majority of North America, with 90% of grid boxes exhibiting a ratio greater than one. On average, LENS has a standard deviation 26% larger than the observations, with the amplification reaching

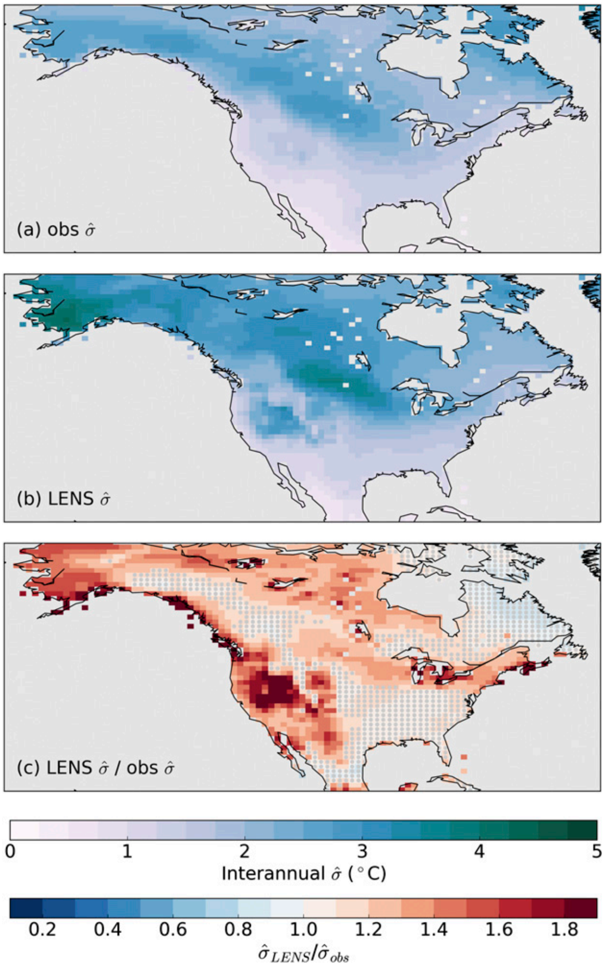
(a) obs $\hat{\sigma}$
(b) LENS $\hat{\sigma}$
(c) LENS $\hat{\sigma}$ / obs $\hat{\sigma}$
0
1
2
3
4
5
Interannual $\hat{\sigma}$ (°C)
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
$\hat{\sigma}_{LENS}/\hat{\sigma}_{obs}$

FIG. 1. Internal variability of detrended DJF temperatures in the NCAR CESM1 Large Ensemble (LENS) and observations. (a) The standard deviation of detrended DJF temperatures in the BEST dataset. (b) The pooled standard deviation of detrended DJF temperatures across the members of LENS. (c) The ratio of variability in LENS to that in the observations [i.e., (b)/(a)]. Stippling indicates grid boxes that are not significant based on use of a false discovery rate of 10% (see main text).

greater than 80% in the interior West, along the western coast of Canada and Alaska, and adjacent to the Great Lakes.

The internal variability in LENS is estimated by pooling across ensemble members, in contrast to the use of a single 50-yr time series for the observational estimate. As such, it is necessary to determine whether the differences between $\hat{\sigma}_{\mathrm{obs}}$ and $\hat{\sigma}_{\mathrm{LENS}}$ are distinguishable from those that could occur solely due to the comparison between an ensemble and a single time series. To do so, we compare the internal variability from each member of the ensemble to that from the observations, where $\hat{\sigma}$ is estimated in both cases as the standard deviation of a detrended 50-yr time series. In particular, we compare the ratio $\hat{\sigma}_{\mathrm{LENS}}/\hat{\sigma}_{\mathrm{obs}}$ to the distribution of the ratios $\hat{\sigma}_{\mathrm{LENS}}/\hat{\sigma}_{\mathrm{member}_i}$, where $\hat{\sigma}_{\mathrm{member}_i}$ is the standard deviation estimated using a single member of LENS. If each member of LENS had *identical* behavior to the ensemble as a whole, the ratio of $\hat{\sigma}_{\mathrm{LENS}}$ to $\hat{\sigma}_{\mathrm{member}_i}$ would be unity at each grid box. As expected, the ratio has deviations from unity due to imperfect estimation of the variance of a short time series. The distribution of these deviations indicates the range of ratios that could result from estimation uncertainties alone rather than systematic differences in the representation of variability. The values of $\hat{\sigma}_{\mathrm{LENS}}/\hat{\sigma}_{\mathrm{obs}}$ can then be compared to the 40 sets of values of $\hat{\sigma}_{\mathrm{LENS}}/\hat{\sigma}_{\mathrm{member}_i}$ to determine the extent to which the observations are truly inconsistent with the model. A rough estimate of a p value at each grid box is calculated as the proportion of the $\hat{\sigma}_{\mathrm{LENS}}/\hat{\sigma}_{\mathrm{member}_i}$ values that are at least as large as $\hat{\sigma}_{\mathrm{LENS}}/\hat{\sigma}_{\mathrm{obs}}$. Since each grid box then has a separate p value, significance is assessed through controlling the false discovery rate across grid boxes (Wilks 2006, 2016). We use an FDR of 10%. Based on this metric, the larger internal variability in LENS than in the observations is found to be significant across most of North America (unstippled regions in Fig. 1c), with the exception of the southeastern United States, eastern Canada, and a swath of western Canada roughly aligning with the Canadian Rockies.

A complete assessment of the source of biases in internal temperature variability is beyond the scope of this work. However, cognizant of the influence of atmospheric circulation on temperature (e.g., Hurrell 1996; Deser et al. 2012b; Holmes et al. 2016; Deser et al. 2016), we briefly assess the simulation of internal SLP variability in LENS by comparing to NCEP–NCAR (Fig. 2) and 20CRv2c (Fig. S7) reanalyses. Both reanalyses suggest that the simulated SLP variability is too large in the North Pacific, extending into western Canada, Alaska, and the western United States, which could lead to augmented temperature variability in these areas as a result of enhanced zonal and meridional temperature advection. In contrast, model biases in temperature variability in central and eastern Canada are not obviously related to biases in SLP variability, and there is enhanced SLP variability in the southeastern United States where temperature variability appears unbiased. The comparison suggests that an overly variable circulation likely plays an important but incomplete role in the modeled internal temperature variability.

b. Autocorrelation

In addition to the magnitude of internal variability, the uncertainty in the forced trend is controlled by the

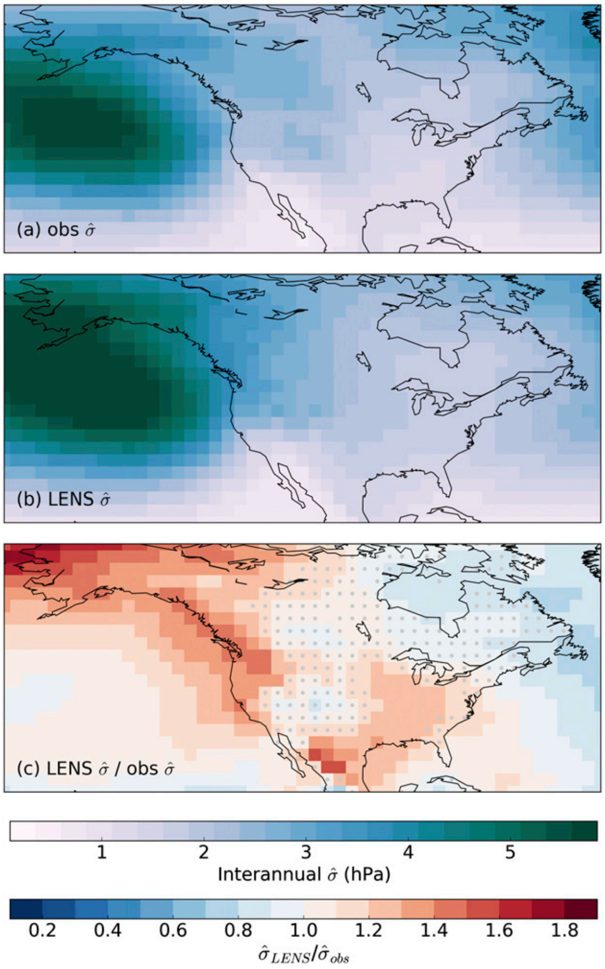
(a) obs $\hat{\sigma}$
(b) LENS $\hat{\sigma}$
(c) LENS $\hat{\sigma}$ / obs $\hat{\sigma}$
1
2
3
4
5
Interannual $\hat{\sigma}$ (hPa)
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
$\hat{\sigma}_{LENS}/\hat{\sigma}_{obs}$

FIG. 2. As in Fig. 1, but for SLP. Observations are from the NCEP–NCAR reanalysis.

autocorrelation structure of the variability (or ''noise''). Noise that is positively correlated in time will contribute to more variability in trends and therefore greater uncertainty, and noise that is negatively correlated in time will lead to smaller uncertainty in the trends. Empirical estimates of the autocorrelation function using short time series tend to be highly variable (Fig. S8; Property 3.10 in Shumway and Stoffer 2015; Deser et al. 2017a), so comparing LENS and the observations is less straightforward than for the magnitude of the internal variability. As an approximation, we summarize the autocorrelation function with the empirical lag–one year autocorrelation coefficient for DJF temperatures. When pooling across space, the observations tend to show larger autocorrelation coefficients than LENS (Fig. 3a); however, the differences are not generally found to be significant at the gridbox scale due to the high variance of the empirical estimator (Fig. S9). The smaller empirical autocorrelations in LENS will reduce the variability in the modeled trend estimates, potentially compensating for the enhanced internal variability.

5. Bootstrapping

How do model biases in internal variability and autocorrelation map onto biases in trend uncertainty? At the gridbox level, the link can be calculated analytically if a parametric time series model for the noise is assumed (e.g., Thompson et al. 2015). This approach, however, neglects the spatial covariance structure present in the data. Here, we instead rely on bootstrapping methods to produce a synthetic ensemble that retains both the spatial and temporal structures of the observations. Trend uncertainty can be assessed in the synthetic ensemble by calculating the standard deviation of the trends across members, analogous to the approach used for LENS. Unlike a true initial condition ensemble, however, the bootstrap trends will be centered around the empirical trend obtained from the actual observations, $\hat{\beta}$, rather than the ''true'' forced trend, β.

a. *Overview of approach*

A typical bootstrapping method for atmospheric time series is the so-called block bootstrap (Kunsch 1989; Politis and Romano 1992; Wilks 1997). In this method, synthetic observations are created by detrending a time series, resampling the residuals in time blocks, and then adding these resampled residuals back to the estimated trend. A new trend is estimated for each synthetic time series, and the uncertainty in the original trend estimate is assessed using a metric such as the standard deviation across the bootstrap estimates. Block bootstrapping is meant to create new synthetic time series that retain most of the correlation structure of the original data. In our setting, residuals are resampled in time only, thereby entirely preserving their *spatial* structure in any given year.

The assumptions underlying block bootstrapping are 1) the residuals are stationary in time; 2) the time blocks are suitably large compared to the scale of temporal autocorrelation; and 3) the number of separate time blocks is also large enough to generate sufficient variability between bootstrap samples. If these conditions are met, the block bootstrap produces variability across bootstrap samples that is comparable to the variability that would be seen in, for example, an initial condition ensemble.

The assumptions of the block bootstrap appear to be reasonable for gridbox-level DJF temperatures over the past 50 years for the following reasons. Although there is evidence of changes in *subseasonal* winter temperature variability in some regions such as northern North

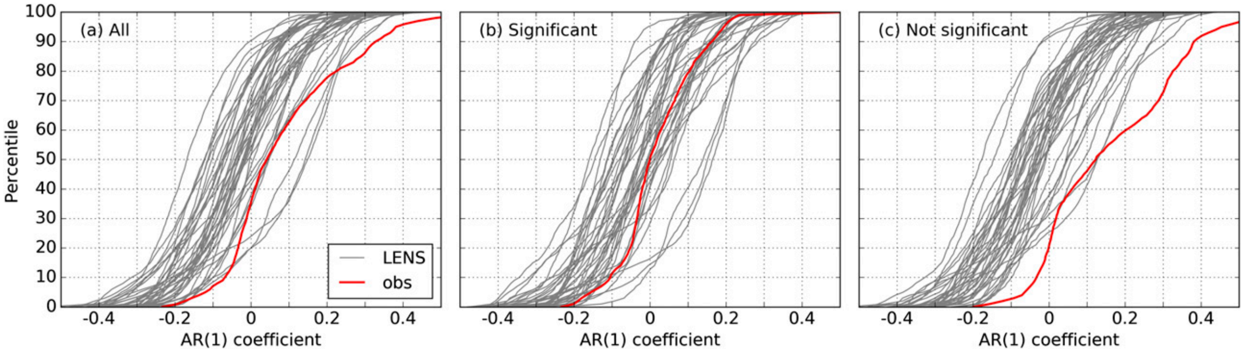

FIG. 3. Lag–1 year autocorrelation in the NCAR CESM1 LENS and observations. (a) Cumulative distribution functions for the empirical lag–1 year autocorrelations in each member of LENS (gray lines) and the observations (red) for all grid boxes over land in the domain. (b) As in (a), but for the grid boxes that are identified as significant in Fig. 5c. (c) As in (a), but for the grid boxes that are identified as not significant in Fig. 5c.

America, perhaps due to Arctic amplification (Screen 2014; Rhines et al. 2016), evidence of significant, forced changes in *interannual* variability over the same period has not been demonstrated to the best of our knowledge, supporting the assumption of stationarity in the residuals. Gridbox-scale DJF temperatures over North America appear only weakly autocorrelated (Fig. 3a) so that a reasonable block length can be chosen that achieves the competing goals of assumptions 2 and 3 above. We use a 2-yr block that best achieved this balance. The validity of this choice of block size is further discussed in the next subsection.

After generating many bootstrap samples of trend estimates, trend uncertainty in the observations can be compared to trend uncertainty in LENS by comparing the variability in trends across the model ensemble and the observationally based synthetic ensemble.

b. *Sensitivity tests*

Before proceeding with our results, we consider the effects of two known issues with the bootstrap methodology. First, estimates of temporal autocorrelation are spatially variable across North America (Figs. S8 and S9), but we use a single block size for the region in order to easily preserve the spatial relationships in the data. Second, block bootstrapping tends to give estimates of trend variability that are biased low when the data are positively correlated and the data record is short. This can lead to an underestimation of the spread across trends [see example 8.4 in Davison and Hinkley (1997), for a similar illustration about the standard error of the mean] because 1) the bootstrapped time series will be more weakly correlated than the original time series because correlation between blocks is destroyed and 2) the finite nature of the data leads to greater similarity among the resampled time series than would occur if they were newly generated from the original underlying system.

To quantify the estimation biases that result from both of these issues, we first perform bootstrapping on a set of synthetic data whose properties are known. We create random Gaussian time series of length 50 with prespecified noise variances and autocorrelation coefficients using an order-1 autoregressive model [AR(1) process]. The true trend, β, of each time series is zero, although they will tend to have a nonzero empirical trend, $\hat{\beta}$. Regardless, the choice of zero trend is arbitrary because the uncertainty in trends due to internal variability is independent of the true value of the trend [see Eq. (2)]. A single random time series can be viewed as analogous to a 50-yr temperature time series at a single grid box.

The uncertainty in the trend for each time series due to its variability and autocorrelation is estimated through bootstrapping in the manner described in section 5a. To acquire a more stable estimate of the bootstrap-based estimates of trend uncertainty, the full bootstrap process is repeated with 1000 random time series for each variance–autocorrelation coefficient pair. The final estimate of the bootstrap-based trend variance is calculated by pooling across the 1000 times series. These estimates can be compared to the true trend uncertainty because, for the AR(1) time series considered in the synthetic analysis, the entries of $\mathbf{\Sigma}$ in Eq. (2) can be calculated exactly as [also see Eqs. (9.19) and (9.21) in Wilks (2011)]

$$\mathbf{\Sigma}_{i,j} = \sigma^2 \phi^{|i-j|}, \qquad (3)$$

where ϕ is the first-order autocorrelation and σ^2 is the noise variance.

To assess the bias introduced by the bootstrapping methodology, we calculate the ratio of the true trend uncertainty to the bootstrap-based estimates of trend uncertainty, where both are measured in standard deviations. We consider the ratio rather than the difference so the reader can compare the magnitude of the effect to the differences between LENS and the observations (shown in Fig. 5c). In the case of no bias, the ratio would be exactly one. As expected, the bootstrap is generally a conservative estimator, leading to an underestimation of the spread of trends due to internal variability (and a ratio value greater than one) when the AR(1) coefficient is greater than −0.05 (Fig. 4a). For a reasonable set of $\hat{\phi}$ and $\hat{\sigma}$ values estimated from the observations and model simulations, the range of ratios that could emerge due to the use of a 2-yr block bootstrap would primarily be between 0.95 and 1.1.

Since temperature may not behave as a Gaussian AR(1) process, we also quantify biases produced by our methodology through the use of the 1800-yr preindustrial control simulation conducted with the same model as was used to create LENS, but with constant external forcing. Like our prior example with synthetic data, the "true" forced trend, β, is zero; any nonzero empirical estimates of the trend, $\hat{\beta}$, are due sampling of internal variability.

We divide the control simulation into 1000 50-yr segments, and calculate the empirical linear trend at each grid box in each segment. The spread of empirical trends across the 1000 segments, quantified using the standard deviation, indicates the uncertainty in 50-yr trends based on the model-simulated internal variability. We next apply the bootstrapping procedure 1000 times for each segment; the average bootstrap-based estimate of trend uncertainty is calculated by pooling across the 1000 50-yr segments. The two estimates of uncertainty in the linear trend are similar, with the ratio of their values having a one standard deviation range around the mean of 0.94–1.1 (Fig. 4b), comparable to what was inferred using the Gaussian AR(1) time series above. This test demonstrates that most of the internal variability relevant for 50-yr DJF temperature trends in this region is preserved in the 2-yr blocks used for bootstrapping. Furthermore, the test shows that this variability can be accounted for through removing the empirical, rather than true, forced trend from a given time series.

6. Spread in trends due to sampling of internal variability

Encouraged by the results of our sensitivity tests, we now return to our analysis of uncertainty in DJF temperature trends. Observed DJF temperatures over North America are bootstrapped 1000 times using a block size of two years to produce the synthetic ensemble as described in section 5a. The spread across the synthetic ensemble indicates the uncertainty in the forced trend due to internal variability as estimated from observations, which can also be compared to that in LENS to identify model biases.

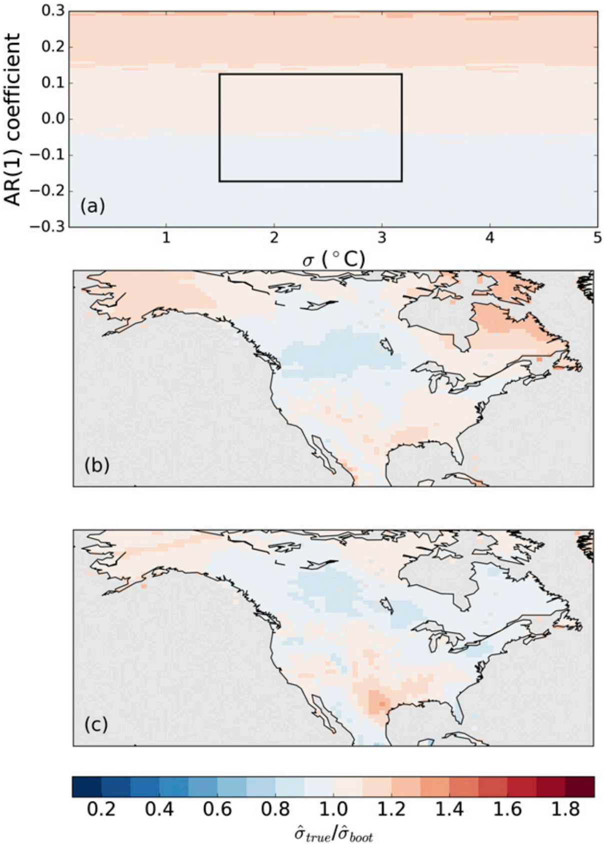

FIG. 4. Validation metrics for the bootstrapping methodology. In all subpanels, color indicates the ratio of the "true" standard deviation of trends due to internal variability to that inferred using block bootstrapping with a 2-yr block. The magnitude of the values can be compared to those in Fig. 5c. Values close to unity indicate that the bootstrap is nearly unbiased. (a) The ratio of trend variability calculated from synthetic AR(1) time series with specified variability and lag–1 year autocorrelations to that inferred from applying the 2-yr block bootstrapping procedure. The black box outlines the typical range (±1 standard deviation around the mean) of autocorrelations and noise variances that we estimate for DJF temperatures across the observations and all members of the NCAR CESM1 Large Ensemble. (b) The ratio of trend variability calculated from 1000 50-yr segments of the NCAR CESM1 1800-yr control simulation to that inferred from applying the 2-yr block bootstrapping procedure to each segment. (c) The ratio of trend variability calculated from the 40 members of the NCAR CESM1 LENS to that inferred from applying the 2-yr block bootstrapping procedure to each member of LENS.

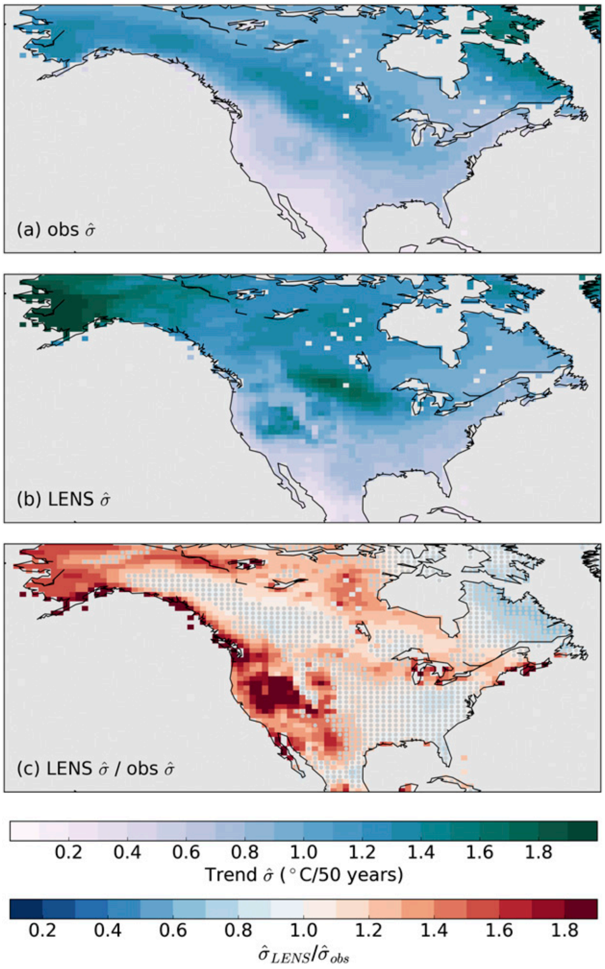
(a) obs $\hat{\sigma}$
(b) LENS $\hat{\sigma}$
(c) LENS $\hat{\sigma}$ / obs $\hat{\sigma}$
0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8
Trend $\hat{\sigma}$ (°C/50 years)
0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8
$\hat{\sigma}_{LENS}/\hat{\sigma}_{obs}$

FIG. 5. Spread across 50-yr trends (1966–2015) in the NCAR CESM1 LENS and bootstrapped observations. (a) The standard deviation of 50-yr trends in DJF temperature based on 1000 bootstrap samples of the observations. (b) The standard deviation of 50-yr trends in DJF temperature across all 40 members of LENS. (c) The ratio of the spread of trends in LENS to that in the observations [i.e., (b)/(a)]. Stippling indicates grid boxes that are not significant based on use of a false discovery rate of 10% (see main text).

As expected based upon the analysis of internal variability, the spread in trends across the synthetic observational ensemble (Fig. 5a) tends to be less than that across LENS (Fig. 5b). In both cases, the spread is calculated as the standard deviation of the trends at each grid box across the members of the ensemble. The ratio of the standard deviation in LENS to that from the synthetic ensemble is greater than one at 78% of grid boxes (Fig. 5c), with an average amplification across grid boxes of 32%. The regions where trends from LENS are more variable than those from the synthetic ensemble align with the regions where LENS exhibits more internal variability than in the observations (cf. Figs. 1c and 5c).

To assess the significance of differences in trend variability between the synthetic ensemble and the model simulations, each member of LENS is bootstrapped in the same way as was done for the observations to account for uncertainty and biases emerging from the bootstrap process itself. As in section 4a, significance is assessed by calculating the proportion of individual LENS members that have a greater trend variance than the observations, and limiting the false discovery rate across grid boxes to 10%.

The trend variability inferred from bootstrapping individual members of LENS can also be compared to the spread across the actual members of LENS for an additional validation of the methodology. If the two estimates are similar, we can conclude that the bootstrapping methodology reproduces the behavior of a true initial conditional ensemble. The ratios of the two estimates across grid boxes have a one standard deviation range around the mean of 0.92–1.1 (Fig. 4c), consistent with our two other prior estimates of potential bias (see section 5b).

Returning to Fig. 5, fewer grid boxes are identified as exhibiting a significant difference in trend variability between LENS and the observations compared to the analysis of internal variability, suggesting the presence of a compensating factor. Recalling that the autocorrelation structure of the data also influences the variability in trends, we again compare the distributions of empirical lag–one year autocorrelation coefficients in LENS to the observations after subdividing the domain into grid boxes identified as having significant versus not significant differences in trend variability (Figs. 3b,c). We find that for the grid boxes where no significant difference is identified between LENS and the observations with respect to variability in trends, the empirical lag-one autocorrelation in LENS tends to be less than that of the observations. Conversely, in the regions where there are significant differences between LENS and the observations, the empirical autocorrelations are relatively consistent between the model and observations. This result suggests that the regions of consistency in trend variance between LENS and the observations may be due, in part, to compensating errors: higher internal variability but reduced autocorrelation in the model.

Finally, we repeat our trend uncertainty analysis using the analytical model of Thompson et al. (2015), which assumed an AR(1) model for temperature at each grid box independently (Fig. S10). The enhancement of trend variability across the western United States and Canada is also identified using the analytical model, with

69% of grid boxes showing a ratio greater than one of the spread in trends in LENS to the spread inferred using the AR(1) model. The primary difference between the two approaches is with respect to the fraction of grid boxes found to be significant. Significance is assessed in the same manner as was used for both internal and trend variability, namely by comparing the result with a null hypothesis for which each members of LENS is compared, in turn, to the ensemble as a whole to quantify differences that could emerge due comparing a single time series and an ensemble. When using the Thompson et al. (2015) model, however, only the western United States is found to exhibit significantly different behavior in the observations compared to LENS. The reason for the difference is that the spread of trends inferred from the analytical model depends on an estimate of the autocorrelation coefficient that will itself be variable given only a 50-yr record (Figs. S8 and S11). By contrast, the block bootstrap approach does not involve estimating a parameter like the autocorrelation coefficient and so can produce less variable estimates of the spread. Because the temporal correlations in these data are weak, the block bootstrap also does not appear to introduce any greater bias in this setting compared to an AR(1) model (Fig. S11). The block bootstrapping approach additionally permits for the easy preservation of the spatial covariance in the data, thereby allowing us to produce spatially consistent temperature fields.

7. An Observational Large Ensemble

The synthetic ensemble can be combined with an estimate of the true forced trend in DJF surface temperatures to create an Observational Large Ensemble that is consistent with the statistics of the observed variability. Because it is challenging to identify the forced trend from the observations alone due to the confounding influence of internal variability [although see, e.g., Smoliak et al. (2015) and Deser et al. (2016) for empirical approaches], we use the ordinary least squares trend of the ensemble mean (EM) of LENS as our best estimate. We then center the synthetic observational ensemble on the EM trend to create OLENS such that its trend variability is based solely on statistics derived from the observations but the forced trend is from the model simulations. We also apply the entire bootstrapping and centering procedure to SLP from the NCEP–NCAR reanalysis.

We extract 36 random members from the 1000-member OLENS for display (Fig. 6). These maps may be compared to the results from LENS in Fig. S12, as well as Fig. 1 in Deser et al. (2016). Each member of OLENS can be interpreted as a temperature history that might have occurred if a different sequence of internal variability had unfolded. By construction, the average trend across members is identical to that of the EM from LENS, which indicates a poleward-amplified forced warming trend across all of North America. Forced temperature trends range from around 1°C $(50\,\mathrm{yr})^{-1}$ in the continental United States to over twice that rate at the higher latitudes.

There is considerable diversity in the spatial patterns and magnitudes of DJF temperature trends within OLENS. For example, member 22 shows warming throughout North America, but exhibits greater warming than the EM over western Canada and Alaska. The pattern of warming in this member is similar to what was actually observed, although the magnitude is slightly weaker. In contrast to member 22, member 3 shows cooling in western Canada, and amplified warming in eastern Canada (also see members 11 and 30). Member 10 has a more spatially uniform warming trend than either member 3 or member 22, with a close resemblance to the EM.

The spread across all 1000 members of OLENS is summarized through identifying the ensemble members that bookend the 95% range of temperature trends averaged across North America. The ensemble member associated with the 2.5th percentile of North American temperature trends shows a similar spatial pattern of trends as member 3, with cooling up to 2°C $(50\,\mathrm{yr})^{-1}$ in western Canada, and muted warming elsewhere. The ensemble member that is at the 97.5th percentile of North American warming shows the greatest temperature increases in north-central Canada; regions of warming are generally shifted eastward compared to member 22 and the observations. The contrast between these ‘‘bookend’’ maps of temperature trends over the past 50 years underscores the importance of accounting for internal variability when interpreting the observational record.

Many of the temperature trend patterns in OLENS can be related to trends in circulation, as summarized by SLP (shown as contours in Fig. 6). Note that LENS shows little evidence for forced trends in DJF SLP, so the circulation trends in each member of OLENS (and LENS) result primarily from sampling of internal variability. A similar lack of forced SLP trends is found in the CMIP5 ensemble mean (Deser et al. 2017a). In member 22, the SLP shows widespread decreases over the North Pacific, with maximum amplitude of more than 8 hPa $(50\,\mathrm{yr})^{-1}$ near the Aleutian Islands. This anomalous cyclonic circulation trend results in anomalous southerly flow over western Canada, causing amplified warming in that region. Conversely, member 3 shows a north–south dipole in SLP trends, with an

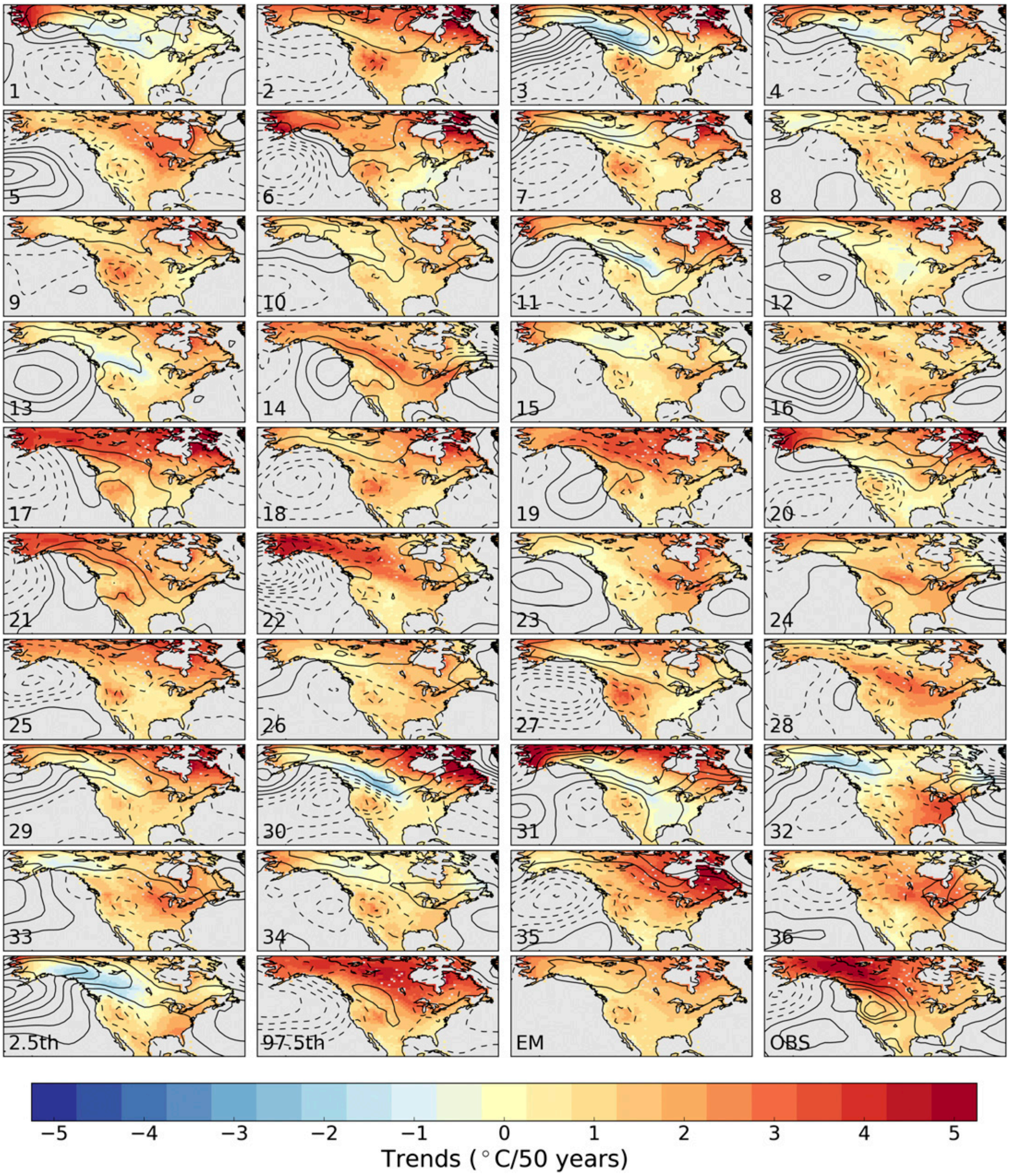

FIG. 6. Sample realizations of 50-yr temperature trends from 36 members of the OLENS. The forced signal is taken as the ensemble mean (EM, bottom row, third column) of the NCAR CESM1 LENS. The variability around the forced signal in temperature (colors) and sea level pressure (contours) is based upon bootstrapping the observations. The contour interval is 1 hPa $(50\,\mathrm{yr})^{-1}$, starting at ±0.5 hPa $(50\,\mathrm{yr})^{-1}$. The spread across the ensemble is summarized by the members associated with the 2.5th and 97.5th percentile of average North American temperature change (bottom row, first and second columns). The observed trend (OBS) is shown in the lower right-hand corner.

increase of over 6 hPa $(50\,\mathrm{yr})^{-1}$ spanning the North Pacific and Canada, and a smaller decrease of over 2 hPa $(50\,\mathrm{yr})^{-1}$ to the south. The positive SLP trends to the north result in anomalous northerly flow, whereas the negative SLP trends to the south result in southerly flow, leading to additional cooling in western Canada and warming in the western United States, respectively.

The SLP trends associated with the ensemble members at the 2.5th and 97.5th percentile of average North American temperature change have notable differences (Fig. 6, lower left panels). The cooler member shows positive trends in SLP greater than 6 hPa $(50\,\mathrm{yr})^{-1}$ over the North Pacific that extend into Alaska and Canada. Conversely, the warmer member shows weaker North

Pacific trends that have minimal extension onto North America beyond Alaska. The difference in SLP trend patterns, especially their different magnitudes over land, suggests a role for other processes such as those related to sea ice and snow cover in causing the high temperature trends. A quantitative assessment of the role of the atmospheric circulation in the spread of temperature trends across OLENS is beyond the scope of this study, but see Deser et al. (2016) for a more complete analysis of LENS.

When visually comparing temperature trend maps from LENS (Fig. S12) and OLENS (Fig. 6), a striking difference is the presence of large regions of cooling in many of the members of LENS. While some members of OLENS also suggest that a 50-yr cooling trend would have been possible given a different sampling of internal variability, the cooling in OLENS tends to be smaller and confined to western Canada and the northern Great Plains. One way to quantify this difference in behavior is through calculating the probability that each grid box had a positive trend in temperature over the past 50 years in each ensemble (Figs. 7a,b). OLENS suggests that over half (54% of grid boxes) of North America had at least a 95% probability of exhibiting warming, whereas the same is true of only 29% of grid boxes in LENS. Note, however, that OLENS contains more members than LENS; thus, the value of this metric will be noisier for LENS than OLENS. To address this issue, and recalling that the members of LENS have themselves been bootstrapped as part of the analysis, we also calculate the fraction of grid boxes with at least a 95% chance of warming across many random sets of 1000 bootstrap samples from LENS, yielding a range (95% confidence interval) of 31%–40% for LENS (Fig. 7c). In both LENS and OLENS, the region with the lowest probability (<80%) of having a positive temperature trend over the past 50 years spans southern Alaska to the north-central United States, which is related to the high internal variability of temperature in the same region (Fig. 1).

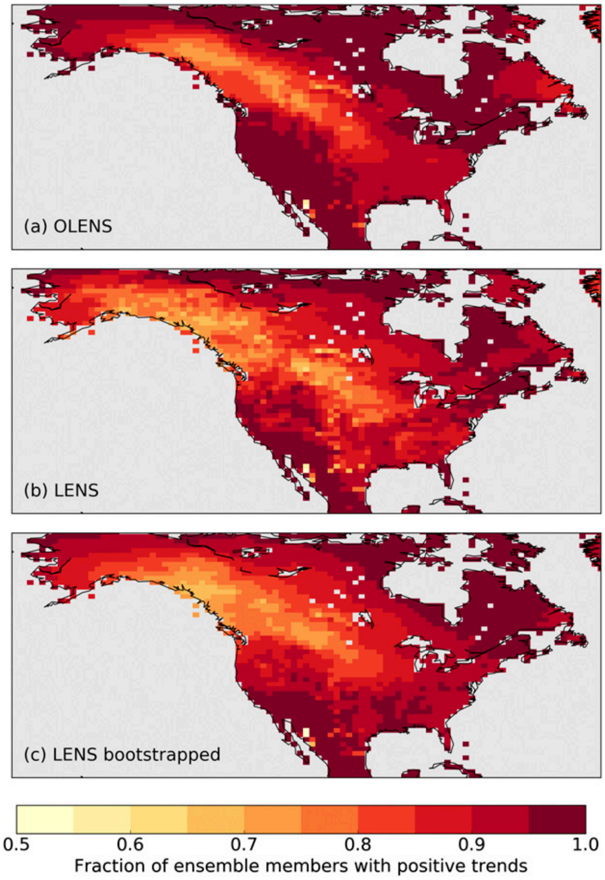

FIG. 7. The fraction of ensemble members that show a positive temperature trend between 1966–2015 in (a) OLENS and (b) the NCAR CESM1 LENS. The spatial pattern for LENS is noisier than OLENS because the latter has a much larger number of ensemble members than the former (1000 vs 40). (c) A smoother estimate through sampling a random set of 1000 members from the bootstrapped versions of LENS.

8. Discussion and conclusions

Determining which climate trends are "forced" versus internal is critical for contextualizing observed climate change and making predictions for the future. Underlying any detection and attribution study, for example, is an estimation of internal variability (Allen and Stott 2003); however, as has been demonstrated here and in prior studies (e.g., Laepple and Huybers 2014; Thompson et al. 2015), there may be large biases in model-generated variability. We have shown that the NCAR CESM1 LENS tends to overestimate internal variability in winter temperatures over North America, which leads to corresponding overestimates of the possible spread in 50-yr temperature trends due to internal variability alone. The overestimation of internal variability in LENS means that it is more difficult to identify a climate change signal in any given model simulation than in the observations using a 50-yr record. The presence of biases in variability during the historical period suggests that the spread of future projections are also likely to be biased. Conceptually similar conclusions regarding model variability were reached by Eade et al. (2014) in the context of North Atlantic seasonal forecasts using an ensemble composed of CMIP5 models.

We do not formally assess which frequency band is leading to the overamplification of variance in LENS. Prior work focusing on sea surface temperatures demonstrated that many models *underestimate* variance on multidecadal and longer time scales, with the underestimation becoming larger at longer periods (Laepple and Huybers 2014). If the same behavior holds over land, then the positive biases in internal variability we identify may emerge from a compensation between increased variability at short time scales and damped variability at longer time scales. This conclusion is consistent with our finding that temperature in LENS is more weakly correlated in time than the observations. Further work, however, is required to compare the full spectra of modeled and observed land temperatures.

The bootstrapping approach taken here is more generally applicable to other regions, seasons, time periods, and variables for which the autocorrelation time scale is small compared to the length of the data record. In these cases, resampling methods can provide an intuitive, observationally based method to estimate the uncertainty in the forced signal due to internal variability. While we have focused on climate change as the signal, it is also possible to apply similar methods to quantify the uncertainty in the climate response to internal modes of variability such as El Niño–Southern Oscillation (Deser et al. 2017b). The bootstrapping method, however, is inappropriate for studies of variables with more substantial autocorrelation, such as sea surface temperatures or global mean surface temperature, for which it would be more useful to use other types of stochastic models in order to create synthetic observations (e.g., Navarra et al. 1998; Brown et al. 2015). In some cases, the short observational record will not contain sufficient long-time-scale information to create synthetic observations at all, although it may be possible to draw upon paleoclimate information instead (Ault et al. 2013; Laepple and Huybers 2014).

Neither models nor observations can provide a complete picture of internal variability due to the influence of model biases and the lack of complete and long instrumental records, respectively. Nevertheless, insights from both can be combined in order to better understand past climate changes and make more robust projections for the future. In this study, we have developed and demonstrated a method for estimating the contribution of internal variability to 50-yr DJF temperature trends over North America. Combined with the estimate of the forced climate change signal from LENS, the method was used to create an Observational Large Ensemble that illustrates counterfactual temperature trends that could have occurred given a different sampling of internal variability, assuming the accuracy of the forced trend from LENS and minimal influence of multidecadal variability, consistent with model behavior (Fig. 4). Similar methods could be applied to future projections in order to provide improved information about the expected variability in regional temperature trends.

Acknowledgments. The authors acknowledge Sloan Coats, Flavio Lehner, and Reto Knutti for their helpful comments on the manuscript, as well as the Rossbypalooza workshop at University of Chicago in 2016 for providing the initial impetus for the project. K.A.M. is supported by the Advanced Study Program at the National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National Science Foundation. E.D.S. acknowledges support from the Natural Sciences and Engineering Research Council of Canada. Code to reproduce the analysis can be found at https://github.com/karenamckinnon. Samples from the synthetic ensemble are available online (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/7CPJPQ).

REFERENCES

Allen, M., and P. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting, Part I: Theory. *Climate Dyn.*, **21**, 477–491, doi:10.1007/s00382-003-0313-9.

Ault, T. R., J. E. Cole, J. T. Overpeck, G. T. Pederson, S. St. George, B. Otto-Bliesner, C. A. Woodhouse, and C. Deser, 2013: The continuum of hydroclimate variability in western North America during the last millennium. *J. Climate*, **26**, 5863–5878, doi:10.1175/JCLI-D-11-00732.1.

Brown, P. T., W. Li, E. C. Cordero, and S. A. Mauget, 2015: Comparing the model-simulated global warming signal to observations using empirical estimates of unforced noise. *Sci. Rep.*, **5**, 9957, doi:10.1038/srep09957.

Collins, M., and M. R. Allen, 2002: Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. *J. Climate*, **15**, 3104–3109, doi:10.1175/1520-0442(2002)015<3104:ATRROI>2.0.CO;2.

——, R. E. Chandler, P. M. Cox, J. M. Huthnance, J. Rougier, and D. B. Stephenson, 2012: Quantifying future climate change. *Nat. Climate Change*, **2**, 403–409, doi:10.1038/nclimate1414.

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

Dai, A., J. C. Fyfe, S.-P. Xie, and X. Dai, 2015: Decadal modulation of global surface temperature by internal climate variability. *Nat. Climate Change*, **5**, 555–559, doi:10.1038/nclimate2605.

Davison, A. C., and D. V. Hinkley, 1997: *Bootstrap Methods and Their Application*. Vol. 1. Cambridge University Press, 594 pp.

Deser, C., R. Knutti, S. Solomon, and A. S. Phillips, 2012a: Communication of the role of natural variability in future North American climate. *Nat. Climate Change*, **2**, 775–779, doi:10.1038/nclimate1562.

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

——, A. S. Phillips, M. A. Alexander, and B. V. Smoliak, 2014: Projecting North American climate over the next 50 years: Uncertainty due to internal variability. *J. Climate*, **27**, 2271–2296, doi:10.1175/JCLI-D-13-00451.1.

——, L. Terray, and A. S. Phillips, 2016: Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications. *J. Climate*, **29**, 2237–2258, doi:10.1175/JCLI-D-15-0304.1.

——, J. W. Hurrell, and A. S. Phillips, 2017a: The role of the North Atlantic Oscillation in European climate projections. *Climate Dyn.*, doi:10.1007/s00382-016-3502-z, in press.

——, I. R. Simpson, K. A. McKinnon, and A. S. Phillips, 2017b: The Northern Hemisphere extratropical atmospheric circulation response to ENSO: How well do we know it and how do we evaluate models accordingly? *J. Climate*, **30**, 5059–5082, doi:10.1175/JCLI-D-16-0844.1.

Eade, R., D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, **41**, 5620–5628, doi:10.1002/2014GL061146.

Fischer, E., U. Beyerle, and R. Knutti, 2013: Robust spatially aggregated projections of climate extremes. *Nat. Climate Change*, **3**, 1033–1038, doi:10.1038/nclimate2051.

Forster, P. M., T. Andrews, P. Good, J. M. Gregory, L. S. Jackson, and M. Zelinka, 2013: Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models. *J. Geophys. Res. Atmos.*, **118**, 1139–1150, doi:10.1002/jgrd.50174.

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, doi:10.1029/2010RG000345.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.*, **90**, 1095–1107, doi:10.1175/2009BAMS2607.1.

——, R. S. Smith, J. M. Gregory, and D. A. Stainforth, 2015: Irreducible uncertainty in near-term climate projections. *Climate Dyn.*, **46**, 3807–3819, doi:10.1007/s00382-015-2806-8.

Holmes, C. R., T. Woollings, E. Hawkins, and H. De Vries, 2016: Robust future changes in temperature variability under greenhouse gas forcing and the relationship with thermal advection. *J. Climate*, **29**, 2221–2236, doi:10.1175/JCLI-D-14-00735.1.

Hurrell, J. W., 1996: Influence of variations in extratropical wintertime teleconnections on Northern Hemisphere temperature. *Geophys. Res. Lett.*, **23**, 665–668, doi:10.1029/96GL00459.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kay, J., and Coauthors, 2015: The Community Earth System Model (CESM) large ensemble project: A community resource 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.

Kunsch, H. R., 1989: The jackknife and the bootstrap for general stationary observations. *Ann. Stat.*, **17**, 1217–1241, doi:10.1214/aos/1176347265.

Laepple, T., and P. Huybers, 2014: Ocean surface temperature variability: Large model–data differences at decadal and longer periods. *Proc. Natl. Acad. Sci. USA*, **111**, 16 682–16 687, doi:10.1073/pnas.1412077111.

Lamarque, J.-F., and Coauthors, 2010: Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: Methodology and application. *Atmos. Chem. Phys.*, **10**, 7017–7039, doi:10.5194/acp-10-7017-2010.

Meinshausen, M., and Coauthors, 2011: The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, **109**, 213–241, doi:10.1007/s10584-011-0156-z.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, doi:10.1029/2011JD017187.

Navarra, A., M. Ward, and N. Rayner, 1998: A stochastic model of SST for climate simulation experiments. *Climate Dyn.*, **14**, 473–487, doi:10.1007/s003820050235.

Politis, D. N., and J. P. Romano, 1992: A circular block-resampling procedure for stationary data. *Exploring the Limits of Bootstrap*, R. LePage and L. Billard, Eds., Wiley, 263–270.

Prather, M., and Coauthors, 2013: Annex II: Climate system scenario tables. *Climate Change 2013: The Physical Science Basis*, T. Stocker et al., Eds., Cambridge University Press, 1397–1445. [Available online at https://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_AnnexII_FINAL.pdf.]

Rhines, A., K. A. McKinnon, M. P. Tingley, and P. Huybers, 2016: Seasonally resolved distributional trends of North American temperatures show contraction of winter variability. *J. Climate*, **30**, 1139–1157, doi:10.1175/JCLI-D-16-0363.1.

Rohde, R., and Coauthors, 2013: Berkeley Earth temperature averaging process. *Geoinf. Geostat. Overview*, **1** (2), 1–13, doi:10.4172/2327-4581.1000103.

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.

Sanderson, B. M., K. W. Oleson, W. G. Strand, F. Lehner, and B. C. O'Neill, 2017: A new ensemble of GCM simulations to assess avoided impacts in a climate mitigation scenario. *Climatic Change*, doi:10.1007/s10584-015-1567-z, in press.

Screen, J. A., 2014: Arctic amplification decreases temperature variance in northern mid- to high-latitudes. *Nat. Climate Change*, **4**, 577–582, doi:10.1038/nclimate2268.

——, C. Deser, I. Simmonds, and R. Tomas, 2014: Atmospheric impacts of Arctic sea-ice loss, 1979–2009: Separating forced change from atmospheric internal variability. *Climate Dyn.*, **43**, 333–344, doi:10.1007/s00382-013-1830-9.

Shindell, D., and G. Faluvegi, 2009: Climate response to regional radiative forcing during the twentieth century. *Nat. Geosci.*, **2**, 294–300, doi:10.1038/ngeo473.

Shumway, R. H., and D. S. Stoffer, 2015: *Time Series Analysis and Its Applications: With R Examples.* 3rd ed. Springer, 202 pp.

Smoliak, B. V., J. M. Wallace, P. Lin, and Q. Fu, 2015: Dynamical adjustment of the Northern Hemisphere surface air temperature field: Methodology and application to observations. *J. Climate*, **28**, 1613–1629, doi:10.1175/JCLI-D-14-00111.1.

Solomon, A., and Coauthors, 2011: Distinguishing the roles of natural and anthropogenically forced decadal climate variability: Implications for prediction. *Bull. Amer. Meteor. Soc.*, **92**, 141–156, doi:10.1175/2010BAMS2962.1.

Teutschbein, C., and J. Seibert, 2012: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *J. Hydrol.*, **456–457**, 12–29, doi:10.1016/j.jhydrol.2012.05.052.

Thompson, D. W., E. A. Barnes, C. Deser, W. E. Foust, and A. S. Phillips, 2015: Quantifying the role of internal climate variability in future climate trends. *J. Climate*, **28**, 6443–6456, doi:10.1175/JCLI-D-14-00830.1.

Wang, H., S.-P. Xie, and Q. Liu, 2016: Comparison of climate response to anthropogenic aerosol versus greenhouse gas forcing: Distinct patterns. *J. Climate*, **29**, 5175–5188, doi:10.1175/JCLI-D-16-0106.1.

Weisberg, S., 2005: *Applied Linear Regression*. 3rd ed. Wiley Series in Probability and Statistics, Vol. 528, John Wiley & Sons, 336 pp.

Wilks, D. S., 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate*, **10**, 65–82, doi:10.1175/1520-0442(1997)010<0065:RHTFAF>2.0.CO;2.

——, 2006: On ''field significance'' and the false discovery rate. *J. Appl. Meteor. Climatol.*, **45**, 1181–1189, doi:10.1175/JAM2404.1.

——, 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Academic Press, 676 pp.

——, 2016: ''The stippling shows statistically significant gridpoints'': How research results are routinely overstated and over-interpreted, and what to do about it. *Bull. Amer. Meteor. Soc.*, **97**, 2263–2273, doi:10.1175/BAMS-D-15-00267.1.

——, and R. L. Wilby, 1999: The weather generation game: A review of stochastic weather models. *Prog. Phys. Geogr.*, **23**, 329–357, doi:10.1191/030913399666525256.