JOURNAL OF CLIMATE

Pattern Recognition Methods to Separate Forced Responses from Internal Variability in Climate Model Ensembles and Observations

ROBERT C. J. WILLS*, DAVID S. BATTISTI, KYLE C. ARMOUR

University of Washington, Seattle, Washington

TAPIO SCHNEIDER

California Institute of Technology, Pasadena, California

CLARA DESER

National Center for Atmospheric Research, Boulder, Colorado

ABSTRACT

Ensembles of climate model simulations are commonly used to separate externally forced climate change from internal climate variability. However, much of the information gained from running large ensembles is lost in traditional methods of data reduction such as linear trend analysis or large-scale spatial averaging. This paper demonstrates a pattern recognition method (forced pattern filtering) that extracts patterns of externally forced climate change from large ensembles and identifies the forced climate response with up to 10 times fewer ensemble members than simple ensemble averaging. It is particularly effective at filtering out spatially coherent modes of internal variability (e.g., El Niño, North Atlantic Oscillation), which would otherwise alias into estimates of regional responses to forcing. This method is used to identify forced climate responses within the 40-member Community Earth System Model (CESM) large ensemble, including an El-Niño-like response to volcanic eruptions and forced trends in the North Atlantic Oscillation. The ensemble-based estimate of the forced response is used to test statistical methods for isolating the forced response from a single realization (i.e., individual ensemble members). Low-frequency pattern filtering is found to effectively identify the forced response within individual ensemble members and is applied to the HadCRUT4 reconstruction of observed temperatures, whereby it identifies slow components of observed temperature changes that are consistent with the expected effects of anthropogenic greenhouse gas and aerosol forcing.

1. Introduction

The observed increase in global temperatures over the past century has not been uniform in space or time. Variability in the rate and pattern of global warming arises from a combination of anthropogenic influences, natural external forcing (e.g., from volcanic sulfur emissions), and internal climate variability arising from processes within (and interactions between) the atmosphere, oceans, cryosphere, and land surface. A primary goal of climate science is to separate the influences of external forcing and internal variability on the global temperature record, as is needed to attribute observed climate changes, to estimate the climate response to future changes in radiative forcing, and to characterize and understand internal climate variability.

Internal climate variability gives rise to uncertainty in climate projections (Hawkins and Sutton 2009; Deser et al. 2012a,b, 2014; Thompson et al. 2015), especially at the regional scale. The separation of externally forced climate change and internal variability has typically been addressed by computing the climate response that is robust across an ensemble of simulations (Harzallah and Sadourny 1995; Hawkins and Sutton 2009; Ting et al. 2009; Solomon et al. 2011; Deser et al. 2014; Frankcombe et al. 2015). Averaging over multiple ensemble members removes internal variability that varies in phase between realizations. Externally forced climate change can be estimated by the ensemble mean, and internal variability can be estimated by deviations from the ensemble mean. However, multi-model ensembles such as the Coupled Model Intercomparison Project (CMIP) conflate model biases with internal variability. This has motivated the use of single-model large ensembles (e.g., Kay et al. 2015), where the same model is run multiple times with

* *Corresponding author address:* Robert C. Jnglin Wills, Department of Atmospheric Sciences, University of Washington, Box 351640, Seattle, WA 98195.
E-mail: rcwills@uw.edu

the same forcing but small differences in the initial condition.

Estimating the climate response to forcing from large ensembles is subject to any model biases in the forced response. This has led to a wide range of conclusions on, for example, the extent to which multi-decadal variability in Atlantic sea-surface temperatures (SSTs) represents true internal variability or is modified by anthropogenic forcing (Ting et al. 2009; Booth et al. 2012; Zhang et al. 2013; Tandon and Kushner 2015; Bellucci et al. 2017; Bellomo et al. 2018; Watanabe and Tatebe 2019) and the extent to which the observed strengthening of the Pacific trade winds and east-west SST gradient since the late 1970s is forced or unforced (McPhaden et al. 2011; England et al. 2014; Takahashi and Watanabe 2016; Coats and Karnauskas 2017; Kohyama et al. 2017; Seager et al. 2019). Comparing across multiple climate models can give insights into which aspects of the forced response are robust and which are not, but this approach becomes computationally intensive as large ensembles are needed for multiple climate models. It is therefore important to identify how many ensemble members are needed to identify forced climate responses and what if anything can be gleaned from individual simulations or from observations.

Seminal work by Deser et al. (2012b, 2014) emphasized that as many as 10-40 ensemble members or more may be needed to identify regional climate responses on timescales up to a few decades, particularly for fields with large internal variability such as precipitation and sea-level pressure (SLP). This has motivated modeling centers to run large ensembles with between 20 and 100 ensemble members (Jeffrey et al. 2013; Kay et al. 2015; Rodgers et al. 2015; Kirchmeier-Young et al. 2017; Sun et al. 2018; Maher et al. 2019; Deser et al. 2020). Now that these large ensembles are available as a testbed, it is possible to revisit the question of how many ensemble members are needed, in order to inform future modeling efforts.

Many studies diagnose the forced response based on the ensemble average of a linear trend or large-scale spatial average. However, this ignores spatiotemporal covariance information that can be valuable in separating forced climate responses from internal variability. A number of studies have demonstrated spatiotemporal analysis methods for isolating the forced climate response from a single realization (Schneider and Held 2001; Wallace et al. 2012; Smoliak et al. 2015; Deser et al. 2016; Frankignoul et al. 2017; Wills et al. 2018; Sippel et al. 2019), with the ultimate goal of isolating the forced component of observed climate changes. However, there has been less focus on the best way to extract forced climate responses from small ensembles (2-10 ensemble members). In this study, we use large ensembles to test statistical methods for isolating forced climate responses, with the goal of identifying the forced response from small ensembles and/or from a single realization. We demonstrate a pattern recognition method that identifies patterns that provide the optimal representation of the forced response (with maximum signal-to-noise ratio) when multiple ensemble members are available and filters out patterns that are temporally incoherent across different ensemble members (i.e., patterns with low signal-to-noise ratio).

Spatiotemporal analysis methods to ascertain the forced response within individual realizations fall into two categories: (i) time-scale separation and (ii) dynamical adjustment. Taking advantage of the fact that forced climate change operates on a longer timescale than most internal variability, time-scale separation methods seek to identify the slowest evolving anomaly patterns and use them to estimate the forced response (Schneider and Held 2001; Frankignoul et al. 2017; Wills et al. 2018). For example, low-frequency component analysis (LFCA, Wills et al. 2018) filters out patterns of anomalies that exhibit primarily high-frequency variability (i.e., that have a small ratio of lowpass filtered variance to total variance). Dynamical adjustment instead estimates the influence of atmospheric internal variability on a target variable by regression against a variable that is representative of the atmospheric circulation (e.g., SLP). This approach has been successful, especially for removing the influence of internal variability on temperature and precipitation changes at midlatitudes (Wallace et al. 2012; Smoliak et al. 2015; Deser et al. 2016; Saffioti et al. 2016; Merrifield et al. 2017; Lehner et al. 2017; Sippel et al. 2019; Guo et al. 2019) and on snowpack or glacier mass balance changes (Christian et al. 2016; Siler et al. 2019; Bonan et al. 2019). However, in cases where atmospheric circulation changes are important to the forced response (see, e.g., Palmer 1999), dynamical adjustment requires a separate method to estimate forced circulation changes (e.g., the mean over a large ensemble). We are interested in a more general method that could, for example, be applied directly to estimate forced changes in atmospheric circulations. Therefore, we do not focus on dynamical adjustment in this paper. We refer the reader to Sippel et al. (2019) for a thorough discussion of how to approach this problem using dynamical adjustment.

This paper is organized as follows. In Section 2, we introduce the pattern recognition methods considered in this study and describe the climate model simulations and observational data analyzed. In Section 3, we demonstrate how identifying forced patterns (FPs) improves estimates of the forced climate response within climate model ensembles compared to a simple ensemble average. We show that it isolates forced responses in quantities with low signal-to-noise ratios such as the east-west SST gradient across the equatorial Pacific, SLP over the North Pacific, and precipitation over the Southwest United States. In Section 4, we show that this method can identify many aspects of the forced response with less than 10 ensemble members. In Section 5, we demonstrate how identifying

low-frequency patterns (LFPs) can be used to estimate the forced climate response from a single ensemble member and apply this method to characterize long-term changes in observed temperatures that are consistent with the expected responses to external forcing. In Section 6, we summarize our conclusions and discuss the generalizability and applications of the statistical methods presented herein.

2. Methods and data

In this paper, we describe statistical methods that identify patterns of externally forced or low-frequency changes. These methods rely on a pattern recognition method called linear discriminant analysis (a type of supervised machine learning) to find spatial patterns, or linear combinations of empirical orthogonal functions (EOFs), that maximize a particular type of variance representing a "signal" compared to "noise" that exists within internal variability or amongst realizations (Déqué 1988; Allen and Smith 1997; Schneider and Griffies 1999; Venzke et al. 1999; Schneider and Held 2001; Ting et al. 2009; DelSole et al. 2011; Wills et al. 2018). This broad category of analyses has variously been referred to as optimal filtering, predictable component analysis, or signal-to-noise-maximizing EOF analysis.

We introduce two types of optimal filtering, which differ in their definition of what type of variance constitutes a signal and what type of variance constitutes noise. In forced pattern (FP) filtering, signal is defined by the mean over an ensemble of simulations; therefore, at least two ensemble members are required. Noise is defined as differences between ensemble members and includes all internal variability, regardless of timescale. It is based on earlier work by Schneider and Griffies (1999; hereafter SG99) and Ting et al. (2009; hereafter T09). Similar to mulitvariate analysis of variance (MANOVA) methods (e.g,. Harzallah and Sadourny 1995; Stern and Miyakoda 1995; Zwiers 1996), it tests whether anomaly patterns within an ensemble are distinct in periods with different external forcing (i.e., predictability of the second kind; Lorenz 1975). In low-frequency pattern (LFP) filtering, signal is defined as variance that makes it through a lowpass filter. Noise is defined as all variability at timescales shorter than the lowpass cutoff. It has also been called low-frequency component analysis (LFCA) and is based on earlier work by Wills et al. (2018; hereafter W18); see also Schneider and Held (2001; hereafter SH01).

In both cases, 'filtering' refers to the retention of only the leading order patterns (i.e., FPs/LFPs), such that patterns of (high-frequency) internal variability are removed from the data set. These methods thus use the spatial structure of covariance in climate noise to optimally filter it out.

a. Forced pattern filtering

The goal of FP filtering is to find anomaly patterns (FPs), for which different ensemble members agree on the temporal evolution [i.e., patterns with a high signal-to-noise ratio (SNR); SG99; T09]. The variability not described by these patterns can then be truncated, such that patterns of ensemble member disagreement (i.e., noise from internal variability) do not alias into the ensemble average.

We seek anomaly patterns associated with timeseries $\mathbf{t}_k$ that maximize the ratio of (ensemble mean) signal to total variance:

$$s_k = \frac{\langle \mathbf{t}_k \rangle^T \langle \mathbf{t}_k \rangle}{\mathbf{t}_k^T \mathbf{t}_k}. \tag{1}$$

Here, angle brackets denote an ensemble average. These timeseries are determined by the projection of a fingerprint pattern $\mathbf{u}_k$ onto the ensemble data matrix $\mathbf{X}$:

$$\mathbf{t}_k = \mathbf{X}\mathbf{u}_k. \tag{2}$$

The $n \cdot n_e \times p$ ensemble data matrix $\mathbf{X}$ is constructed by concatenating the $n \times p$ data matrices $\mathbf{X}_i$ from each ensemble member in the time dimension, where n is the length of timeseries, n_e is the number of ensemble members, and p is the spatial dimension. Each ensemble member data matrix $\mathbf{X}_i$ is weighted by the square root of grid cell area, such that the covariance matrix is area weighted.

To ensure that the identified patterns correspond to variability that actually occurs within the ensemble, the fingerprint patterns $\mathbf{u}_k$ are required to be linear combinations of the N leading ensemble EOFs $\mathbf{a}_k$, with normalized weight vectors $\mathbf{e}_k$:

$$\mathbf{u}_k = \left[\frac{\mathbf{a}_1}{\sigma_1} \;\; \frac{\mathbf{a}_2}{\sigma_2} \;\; \ldots \;\; \frac{\mathbf{a}_N}{\sigma_N} \right] \mathbf{e}_k. \tag{3}$$

The ensemble EOFs $\mathbf{a}_k$ are eigenvectors of the ensemble-mean covariance matrix $\langle \mathbf{C} \rangle$,

$$\langle \mathbf{C} \rangle \mathbf{a}_k = \sigma_k^2 \mathbf{a}_k, \tag{4}$$

where σ_k^2 is the variance associated with the kth EOF. The ensemble-mean covariance matrix $\langle \mathbf{C} \rangle$ (i.e., the pooled covariance matrix) can be computed as:

$$\langle \mathbf{C} \rangle = n_E^{-1} \textstyle\sum_{i=1}^{n_E} \mathbf{C}_i, \tag{5}$$

where $\mathbf{C}_i = (n-1)^{-1} \mathbf{X}_i^T \mathbf{X}_i$ are the individual ensemble member climatological covariance matrices. The ensemble EOFs are normalized such that $||\mathbf{a}_k|| = 1$ and the principal components $\mathbf{c}_k = \sigma_k^{-1} \mathbf{X}\mathbf{a}_k$ have unit variance over the entire ensemble.

We can solve for the linear-combination coefficients $\mathbf{e}_k$ that give $\mathbf{u}_k$ and $\mathbf{t}_k$ that maximize s_k by plugging (2) and

(3) into (1) and using the definition of a principal component $\mathbf{c}_k = \sigma_k^{-1}\mathbf{X}\mathbf{a}_k$ to turn this into an eigenvalue problem, $\mathbf{S}\mathbf{e}_k = s_k\mathbf{e}_k$, where

$$\mathrm{S}_{mn} = \langle \mathbf{c}_m \rangle^T \langle \mathbf{c}_n \rangle \qquad m,n \in [0\ N]. \tag{6}$$

The matrix $\mathbf{S}$ has N eigenvectors $\mathbf{e}_k$, with eigenvalues that give the ratio s_k of signal to total variance. Finally, the FPs $\mathbf{v}_k$ are determined by the regression of the ensemble data matrix $\mathbf{X}$ onto each $\mathbf{t}_k$:

$$\mathbf{v}_k = \mathbf{X}^T\mathbf{t}_k = \mathbf{X}^T\mathbf{X}\mathbf{u}_k = [\sigma_1\mathbf{a}_1\ \ \sigma_2\mathbf{a}_2\ \ \ldots\ \ \sigma_N\mathbf{a}_N]\mathbf{e}_k. \tag{7}$$

In this analysis, the timeseries $\mathbf{t}_k$ retain their orthogonality (like principal components), but the FPs $\mathbf{v}_k$ do not.

The FPs are sorted by s_k such that the leading FPs are patterns of forced response within the ensemble. This is equivalent to sorting by SNR, which is uniquely determined by the eigenvalue s_k:

$$\mathrm{SNR} = s_k(1-s_k)^{-1}. \tag{8}$$

The 1st FP is the linear combination of the leading N EOFs with the maximum possible SNR.

Note the difference between the fingerprint patterns $\mathbf{u}_k$ (Eq. 3) and the forced patterns $\mathbf{v}_k$ (Eq. 7); the fingerprint patterns are a unitless weight vector that is used to detect the signal but has no physical meaning, whereas the forced patterns characterize the signal itself. Fingerprint patterns are used in detection and attribution to detect a model-based signal within observational data (see, e.g., Hasselmann 1993; Santer et al. 1995). Here, in contrast, the signal (as characterized by the FPs) is determined empirically within a single model-based dataset.

Once the FPs have been calculated, the forced response is isolated by constructing a truncated dataset from the M leading FPs:

$$\mathbf{X}_{FP} = \sum_{k=1}^{M} \mathbf{t}_k\mathbf{v}_k^T. \tag{9}$$

We will show that the ensemble average of the truncated dataset $\langle \mathbf{X}_{FP} \rangle$ (i.e., FP filtering) gives a better estimate of the forced response than a simple ensemble average $\langle \mathbf{X} \rangle$. The inclusion of M FPs to construct an estimate of the forced response $\langle \mathbf{X}_{FP} \rangle$ is what distinguishes FP filtering from the method of T09, which focuses on the leading pattern in order to estimate the contribution of forcing to Atlantic multi-decadal variability.

FP filtering has two hyperparameters: N, the number of EOFs retained, and M, the number of FPs used in constructing the truncated dataset. The number of EOFs N should generally not exceed the degrees of freedom in the signal of interest, which in the case of the ensemble mean used here is approximately $n-1$. We pick N to retain 75-95% of the total variance. We choose M empirically to maximize agreement between subsets of the large ensemble (i.e., by comparison to a validation set; see Section 3), but we also compare with methods to choose M based on the eigenvalue spectrum s_k (cf. North et al. 1982). Our results are generally insensitive to these hyperparameter choices for $50 < N < 400$ and $2 < M < 20$ (see Section 3).

A similar method was presented by DelSole et al. (2011) that looks for patterns that maximize the variance in a simulation of forced climate change relative to a pre-industrial control run. This has the advantage of requiring only one forced simulation and one pre-industrial control run (rather than at least two forced simulations). However, it could miss forced responses where forcing only modifies the timing (i.e., phase) of a mode of internal variability. In most other respects these methods would identify similar patterns of forced response.

b. Low-frequency pattern filtering

FP filtering relies on the computation of an ensemble mean to diagnose the variance that is forced within a dataset. In the case that only a single realization is available, it is necessary to come up with a new variance criterion to distinguish forced from unforced variance. Responses to anthropogenic forcing generally differ from most internal variability in terms of their long timescale. We can therefore look for the slowest evolving patterns within a dataset, which will predominantly include the forced response. One method to find the slowest evolving patterns is low-frequency component analysis (LFCA; W18; see also SH01), which solves for patterns with the maximum ratio of low-frequency to total variance (i.e., LFPs).

LFCA uses the same linear algebra machinery as FP filtering, but instead seeks anomaly patterns associated with timeseries $\mathbf{t}_k$ that maximize the ratio of low-frequency signal to total variance:

$$r_k = \frac{\widetilde{\mathbf{t}}_k^T\widetilde{\mathbf{t}}_k}{\mathbf{t}_k^T\mathbf{t}_k}. \tag{10}$$

Low-frequency signal is defined as any variations that makes it through a lowpass filter (denoted by a tilde). Here, we use a linear Lanczos filter with a 10-year lowpass cutoff to focus on variability at decadal and longer timescales (i.e., multi-decadal variability). In lowpass filtering, we do not filter over discontinuities between ensemble members; the data from each ensemble member is filtered separately then concatenated into a single $\widetilde{\mathbf{t}}_k$.

The LFPs $\mathbf{v}_k$ and their timeseries $\mathbf{t}_k$ are determined by Eqs. (7) and (2), respectively, but with weight vectors $\mathbf{e}_k$ that are normalized eigenvectors of the covariance matrix $\mathbf{R}$ of the first N lowpass filtered principal components $\widetilde{\mathbf{c}_k}$:

$$\mathrm{R}_{mn} = \widetilde{\mathbf{c}_m}^T\widetilde{\mathbf{c}_n} \qquad m,n \in [0\ N]. \tag{11}$$

The matrix $\mathbf{R}$ has N eigenvectors, $\mathbf{R}\mathbf{e}_k = r_k\mathbf{e}_k$, with eigenvalues that give the ratio r_k of low-frequency to total variance. The LFPs are sorted by r_k such that the leading LFPs

are the anomaly patterns that maximize the ratio of low-frequency to total variance over the entire ensemble.

Just as in FP filtering, a truncated dataset is created that contains just the variability captured by the leading M LFPs:

$$\mathbf{X}_{LFP} = \sum_{k=1}^{M} \mathbf{t}_k \mathbf{v}_k^T. \quad (12)$$

In addition to the hyperparameters of FP filtering (N and M), LFP filtering depends in general on the properties of the filter used, though we will not explore this particular sensitivity here. A detailed discussion of the robustness of LFPs to the choice of parameters and filter can be found in W18. Unlike principal component analysis of lowpass filtered data, LFCA uses information about spatiotemporal covariance at all timescales (e.g., in computing the EOFs $\mathbf{a}_k$). LFCA thus provides a method to isolate the regions and physical mechanisms important at long timescales while avoiding the issues with attributing lead-lag relationships based on filtered data (Cane et al. 2017; Wills et al. 2019a,b).

c. Model output and observational datasets

We focus primarily on surface temperature anomalies in the 40-member CESM1 large ensemble (CESM-LE, Kay et al. 2015), analyzing years 1920-2005 from the historical simulations and years 2006-2019 from the RCP8.5 simulations. Each ensemble member experiences the same historical and RCP8.5 forcing from greenhouse gases, anthropogenic aerosols, volcanic sulfur emissions, solar variability, and ozone. They differ by machine-precision atmospheric perturbations on 1 January 1920 (so-called micro initialization). Seasonal (3-monthly) anomalies are computed with respect to the each ensemble member's climatological seasonal cycle over 1920-2019. Results are unchanged if the anomalies are computed instead with respect to the ensemble-mean climatology. In Section 3b, we also include analysis of seasonal precipitation and SLP anomalies.

For comparison, we also analyze a 30-member ensemble of the CSIRO-Mk3.6 climate model (CSIRO-LE, Jeffrey et al. 2013), a 20-member ensemble of the GFDL-CM3 climate model (GFDL-LE, Sun et al. 2018), and a 100-member ensemble of the MPI-ESM climate model (MPI-LE, Maher et al. 2019), including years 1920-2005 from the historical simulations and years 2006-2019 from the RCP8.5 simulations. As in the CESM-LE, the GFDL-CM3-LE uses micro initialization in 1920. The ensemble members of the CSIRO-LE and MPI-LE, however, are all started from different ocean states in 1850 (so-called macro initialization). For computational efficiency, all analysis is done on grids that are half the atmospheric models' resolution ($\sim 1^\circ$ in CESM-LE; $\sim 1.8^\circ$ in CSIRO-LE and MPI-LE, $\sim 2^\circ$ in GFDL-LE) such that 4 model grid points are averaged into one analysis grid point. For the observational analysis in Section 5c, we use the infilled surface temperature reconstruction of Cowtan and Way (2014), based on HadCRUT4 data, for the period 1920-2019.

3. Improved identification of forced climate responses

a. Forced surface temperature responses

We begin by identifying the FPs of seasonal (3-monthly) surface temperature anomalies in the 40-member CESM-LE over the time period 1920-2019. FP-1 shows the predominant pattern of long-term global warming (Fig. 1a) and can be detected based on changes in temperature throughout the subtropical oceans (Fig. 2a). All ensemble members show approximately the same timing of its evolution (grey lines in Fig. 1a) and are tightly clustered about the ensemble-mean timeseries (black line in Fig. 1a). FP-1 captures centennial global warming punctuated by volcanically induced global cooling due to the eruptions of Agung in 1963, El Chichón in 1982, and Pinatubo in 1991. However, it is not the only pattern of forced response: FP-2, which shows hemispherically asymmetric temperature anomalies, also has a common temporal evolution in all ensemble members (Fig. 1b). The signal fraction (i.e., the eigenvalue s_k) is only slightly lower for FP-2 than for FP-1 (0.75 vs. 0.95, Fig. 3a) and both have a SNR well above 1. The timing of FP-2 corresponds to Northern Hemisphere cooling between 1940 and 1970, and warming since, consistent with anthropogenic aerosol forcing (Shindell et al. 2013). FP-2 also shows large negative anomalies (cold Northern Hemisphere) following volcanic eruptions. FP-2 can be detected based on the asymmetry in subtropical ocean warming between the Northern and Southern Hemisphere (Fig. 2b).

The next four FPs have $0.2 < s_k < 0.4$ (Fig. 3a), corresponding to a SNR between 0.25 and 0.67. They capture centennial changes in the seasonal cycle of temperature, which manifest themselves in annual cycles in the corresponding ensemble-mean timeseries (black lines in Fig. 1c-f), with opposite phasing in the early and later parts of the simulations (insets in Fig. 1 show ensemble-mean trends separately for each season). These FPs have the largest anomalies in regions of sea-ice cover (Fig. 1c-f), indicating that they are capturing changes in the seasonal extent of sea ice (as discussed in Zhang and Walsh 2006; Eisenman et al. 2011). FP-3 and FP-4 also show non-monotonic long-term changes, with decreasing trends between 1920 and 1960, increasing trends from 1960 to 1980, and weakly decreasing and more seasonal dependent trends since 1980. FP-5 and FP-6 both show some evidence of an El-Niño-like response to volcanic eruptions as well as evolution from a common La-Niña-like initial ocean state in January 1920 (a result of micro initialization). While FP-1 and FP-2 are robust to the choice of N (the number of EOFs retained), the next four FPs show

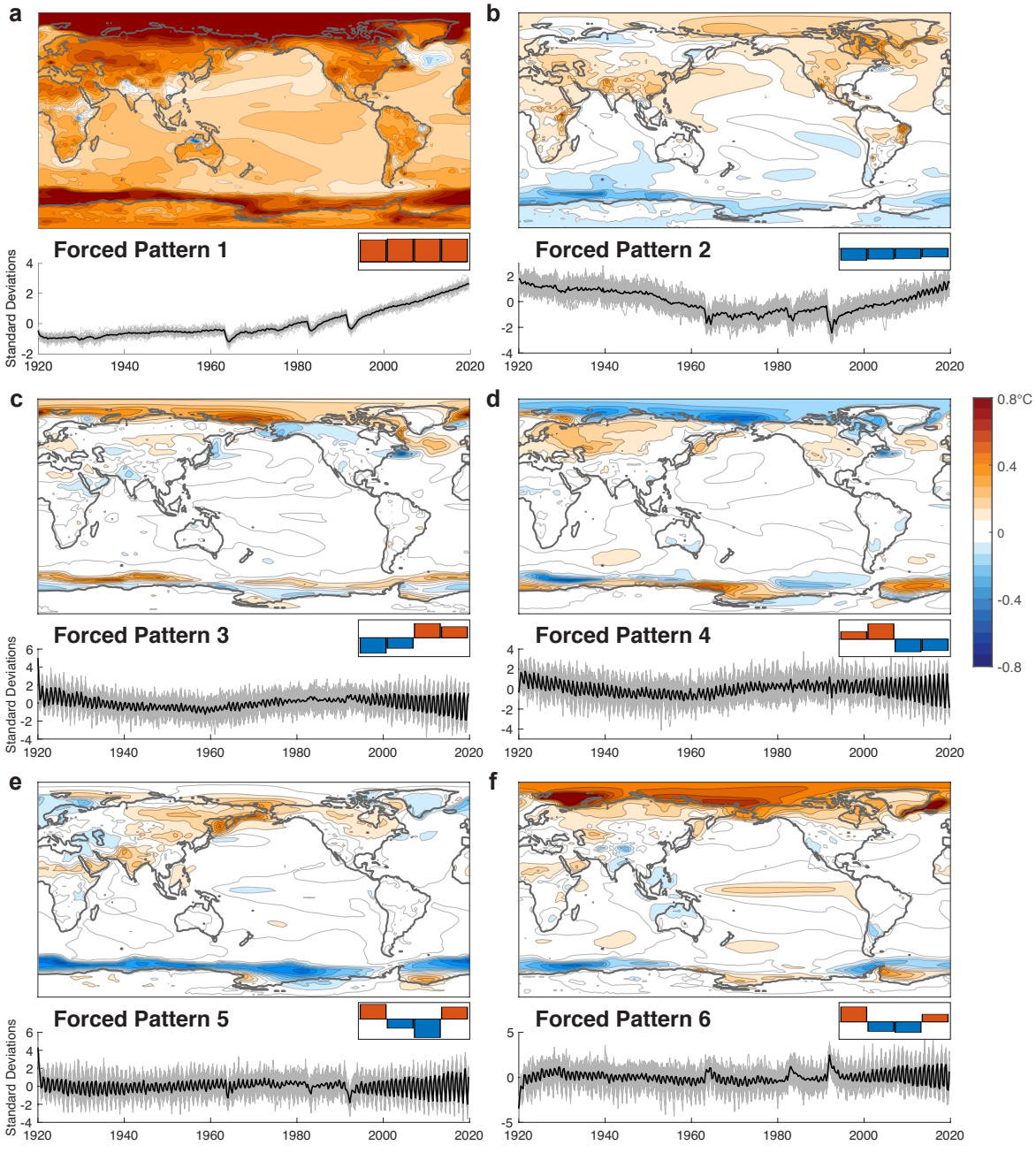

FIG. 1. Forced patterns (FPs) of seasonal-mean surface temperature anomalies in the CESM-LE historical and RCP8.5 simulations over the time period 1920-2019, with $N = 150$ EOFs retained. The time evolution of the FPs in all ensemble members are shown as standard deviation anomalies with grey lines. The black line shows the ensemble-mean time evolution of each pattern (i.e., $\langle \mathbf{t}_k \rangle$). Note that seasonal cycle in the ensemble-mean time evolution indicates forced changes in the seasonality of surface temperature. 100-year ensemble-mean trends in each pattern are shown separately for JFM, AMJ, JAS, and OND (left to right) in bar-chart insets. The y-scale for the bar-chart insets is half that for the timeseries in panel (a).

some rearrangement as N is varied (i.e., they capture the same responses, but partition them in different ways).

In order to construct an estimate of the forced response, we must choose the number of patterns M to retain. Three possible methods for choosing M are: (1) choosing eigen-

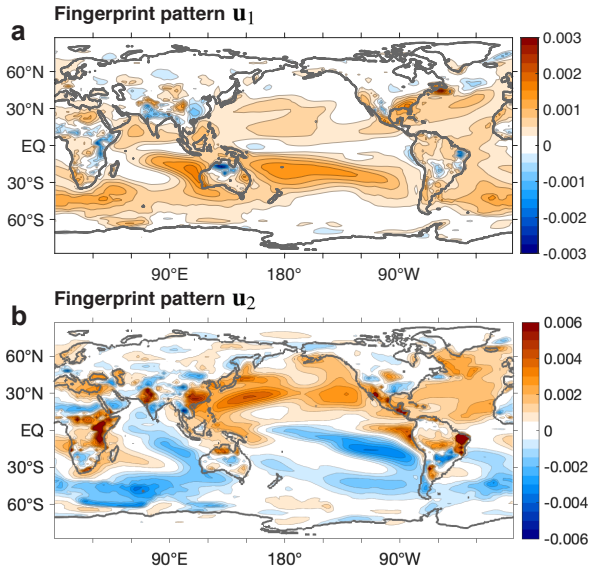

FIG. 2. Unitless fingerprint patterns $\mathbf{u}_k$ of seasonal-mean surface temperature anomalies in the CESM-LE historical and RCP8.5 simulations over the time period 1920-2019, with $N = 150$ EOFs retained (cf. forced patterns $\mathbf{v}_k$ in Figs. 1a and 1b).

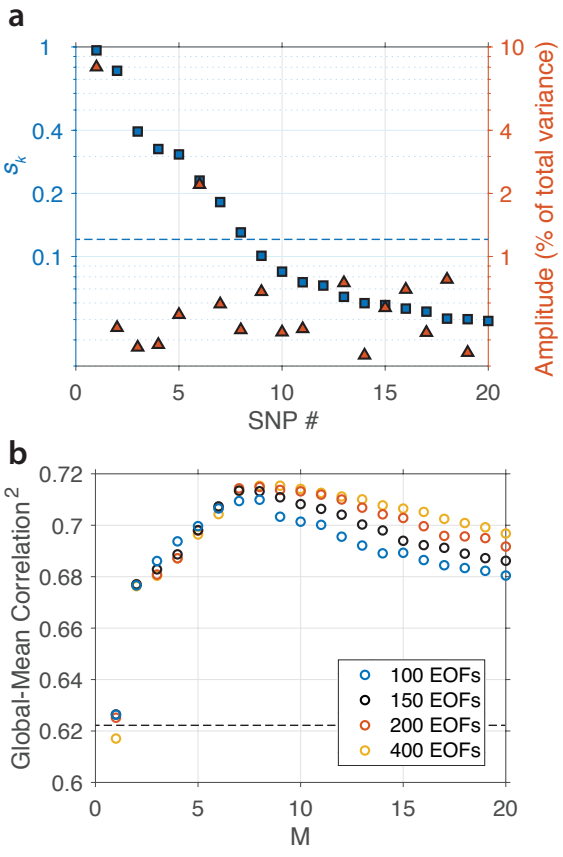

FIG. 3. (a) Signal fraction (s_k) and amplitude of the leading FPs. The dashed line gives the minimum value of s_k that is significant at the 5% significance level computed by block bootstrapping. Note that while the amplitude is expressed as a percentage of the total variance, these percentages do not add to exactly 100% because of the non-orthogonality of the FPs. (b) Global mean of the grid-point squared correlation between the pattern filtered estimate of the forced response $\langle \mathbf{X}_{FP} \rangle$ from one 20-member half-ensemble and the simple ensemble mean $\langle \mathbf{X} \rangle$ of the opposite 20-member half-ensemble, as a function of the number of FPs included M and the number of EOFs retained N. The dashed line gives the global-mean grid-point squared correlation between 20-member half-ensembles when no pattern filtering is applied.

values s_k that are well separated from the continuum of eigenvalues, (2) finding a significance level for s_k with block bootstrapping, or (3) using the large ensemble to empirically determine the number of patterns that works best. The first method is based on the North et al. (1982) test, which is employed in EOF analysis. It looks for a scale break in the eigenvalue spectrum (Fig. 3a). Using this test, we could choose to either retain the first 2 well separated FPs ($s_k > 0.75$), or to include all FPs up to the point where the separation between neighboring eigenvalues becomes small compared to the uncertainty in those eigenvalues, which is estimated in terms of the degrees of freedom (DOF) as $s_k \cdot (2/\mathrm{DOF})^{1/2}$. The DOF of the 40-member ensemble mean (where autocorrelation primarily comes from the forcing itself) is approximately the number of seasonal time steps minus one (i.e. 399). This gives a 7% fractional uncertainty in the eigenvalues, which leads us to conclude that the first 10 eigenvalues are well separated.

We have also tested a block bootstrapping approach, taking random 10-year samples from the 40-member ensemble (with replacement) to construct randomized ensembles where the members should not agree on the timing of climate responses. We then rerun the FP filtering on these randomized ensembles and compute the statistics of the resulting s_1. We find that $s_k > 0.12$ are significant at the 5% significance level. FPs with s_k below this level could occur due to random chance and are not significant. According to this bootstrapping test, 8 FPs are statistically significant (Fig. 3a), roughly in agreement with the simpler but less rigorous North et al. (1982) test. This bootstrapping approach may generalize better to smaller ensembles.

Within a large ensemble, we can also empirically test which value of M best estimates the forced response (which should be the same in all subsets of the large ensemble). To do so, we split the ensemble in half, apply FP filtering to one 20-member half-ensemble (the training set), and test how well the resulting $\langle \mathbf{X}_{FP} \rangle$ agrees with the ensemble mean $\langle \mathbf{X} \rangle$ of the opposite 20-member half-ensemble (the validation set). We test agreement based on the global average of the squared correlation between

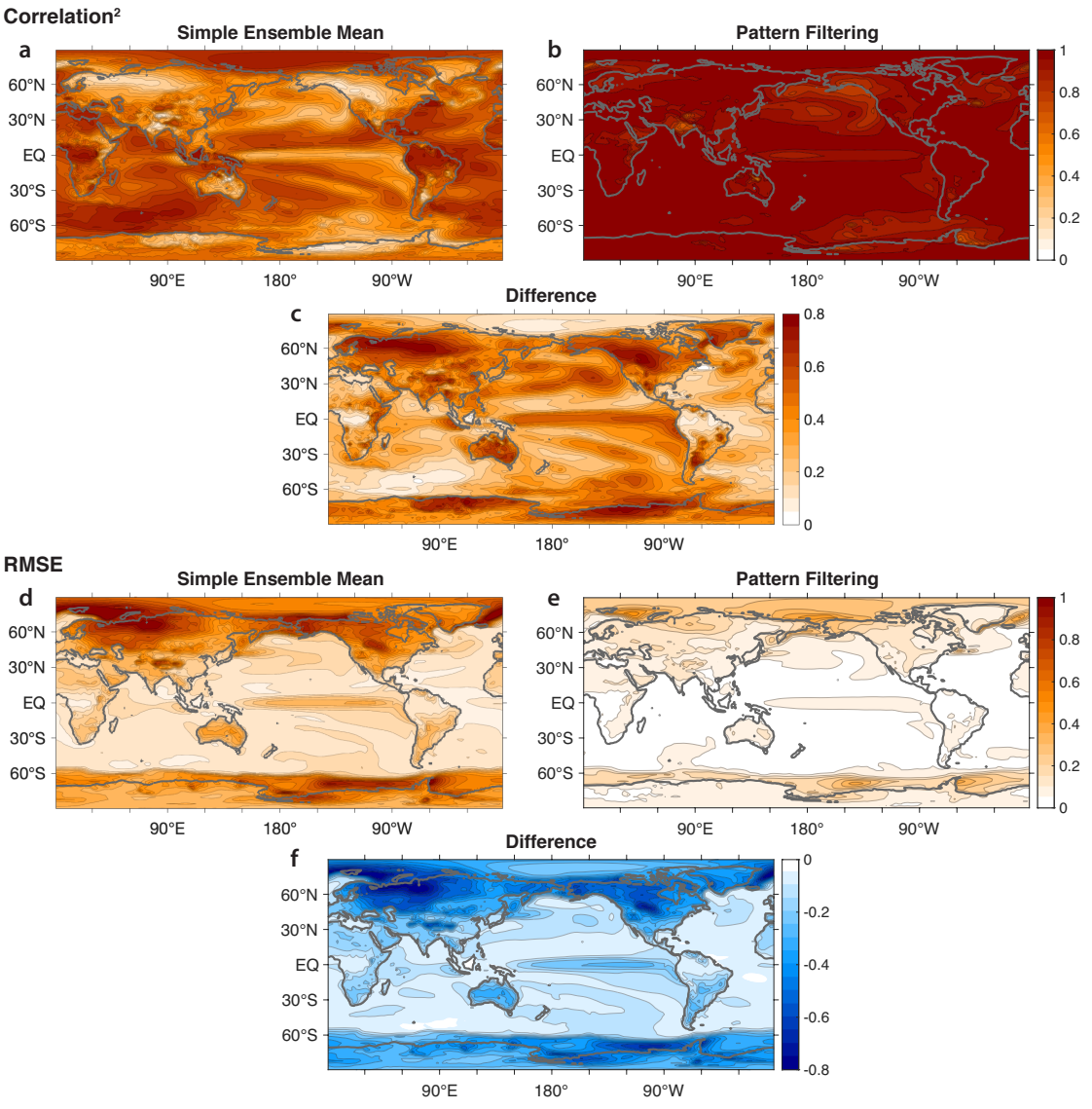

FIG. 4. (a),(b) Spatial maps of the squared correlation between estimates of the forced response in 20-member half-ensembles: (a) when the forced response is estimated by a simple ensemble mean and (b) when the forced response is estimated by FP filtering with $M = 7$ and $N = 150$. (c) Difference between (a) and (b). (d),(e) Spatial maps of the root-mean square error (RMSE) between estimates of the forced response in 20-member half-ensembles: (d) when the forced response is estimated by a simple ensemble mean and (e) when the forced response is estimated by FP filtering with $M = 7$ and $N = 150$. (f) Difference between (d) and (e).

the two estimates of the forced response at a grid point (Fig. 3b). As long as two or more FPs are included, FP filtering improves the agreement with the ensemble mean of the validation set (the agreement between the ensemble means of the opposite half-ensembles is shown with a dashed line in Fig. 3b). The large jump in agreement between $M = 1$ and $M = 2$ means that it is critical to include at least two degrees of freedom (2 patterns) in an estimate of the forced response.

For the case where 150 EOFs (88.9% of the total variance) are included in the analysis, including $M = 7$ FPs maximizes the agreement with the ensemble mean of the validation set.[1] Including further EOFs increases the number M of FPs required to maximize this agreement without substantially improving the maximum value of the global-

[1] In determining the value of M to use, one must compare to the simple ensemble mean of the validation set rather than the pattern filtered validation set, because truncating to a single pattern (i.e. $M = 1$) maximizes the agreement between two pattern filtered sub-ensembles (by construction).

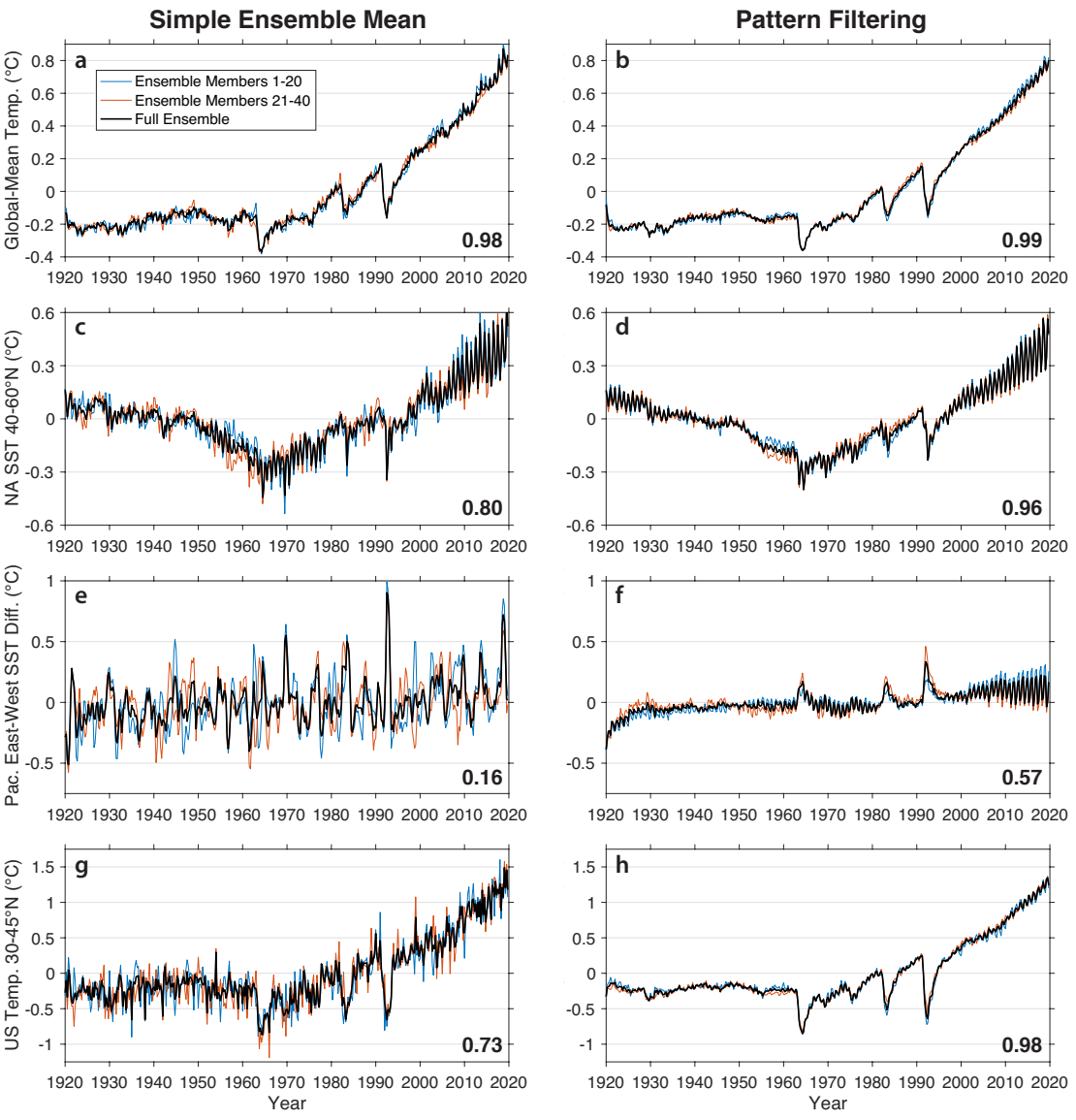

FIG. 5. Forced responses of large-scale temperature indices in the CESM-LE, computed from (left) a simple ensemble mean $\langle \mathbf{X} \rangle$ and (right) the pattern filtered ensemble mean $\langle \mathbf{X}_{FP} \rangle$, from FP filtering of seasonal surface temperature anomalies with $M = 7$ and $N = 150$. Blue and orange lines show the first and second 20-member half-ensembles of the CESM-LE, respectively. The black line shows the full 40-member CESM-LE. The squared correlation between the 20-member half-ensembles is shown in the bottom right of each panel. North Atlantic SST is averaged over 40-60°N and 0-80°W. The Pacific east-west SST difference is the difference between the eastern equatorial Pacific (90-150°W, 6°S-6°N) and the western equatorial Pacific (120°E-180°, 6°S-6°N). US land surface temperatures are averaged over 30-45°N, including most of the contiguous United States and parts of Mexico and Canada.

mean squared correlation. The reduction in agreement beyond $M = 7$-9 is a sign of overfitting to the evolution of anomalies in the particular ensemble members used. We choose representative hyperparameter values of $N = 150$ EOFs and $M = 7$ FPs for most of the analysis that follows.

Spatial maps of the squared grid-point correlation between 20-member half-ensembles, before (Fig. 4a) and after (Fig. 4b) applying FP filtering, show that FP filtering substantially increases the agreement between subensembles. The largest improvements are over the Northern Hemisphere continents, North Pacific, Tropical Pacific, Australia, and Antarctica. We find qualitatively similar results if we instead use the root mean square error (RMSE) between two half-ensembles to measure their agreement (Fig. 4d-f). Note that the FP filtering of each

sub-ensemble is independent and no information (e.g., EOFs) is shared between analyses.

Detecting climate signals in grid-point temperature is significantly harder than detecting climate signals in a large-scale spatial average because of the larger amplitude of internal variability at small scales (e.g., Deser et al. 2012a). We would therefore like to test whether the improved identification of the forced response by FP filtering extends to large-scale averages. Again, we will compare agreement between the two 20-member half-ensembles before and after applying pattern filtering.

The time evolution of global-mean surface temperature is in good agreement between the two half-ensembles, even before applying pattern filtering (squared correlation of 0.98, Fig. 5a). FP filtering improves this agreement (squared correlation of 0.99, Fig. 5b), but only marginally so. The global average already averages out most internal variability, so pattern filtering does not substantially improve the estimate of the forced response in global-mean surface temperature. Note, however, that it does improve the global-mean surface temperature response estimate when fewer ensemble members are available (see Section 4).

The improved identification of climate responses by FP filtering is again apparent if we examine regional temperature anomalies such as the North Atlantic (NA) SST (40-60°N, including the NA warming hole), the SST difference between the eastern and western equatorial Pacific, or the United States land surface temperature averaged over 30-45°N (Fig. 5c-h). Particularly noteworthy is that the 20-member and even 40-member ensemble means of the equatorial Pacific east-west SST difference show substantial noise from the El Niño-Southern Oscillation (ENSO) (Fig. 5e), which is removed in the FP-filtered estimate of the forced response (Fig. 5f). The squared correlation between the two half-ensembles is only 0.16 before FP filtering, but increases to 0.57 after. This reveals an El-Niño-like response to volcanic forcing that was not apparent in the 20- or 40-member ensemble means. This response has been studied elsewhere (Maher et al. 2015; Khodri et al. 2017; Pausata et al. 2020), but has only been identifiable by compositing over hundreds of modeled eruption responses. Pattern filtering also reveals ensemble agreement on evolution from a common La-Niña-like initial state in January 1920 (a result of micro initialization) and a weak El-Niño-like trend since ∼1970 (particularly in the winter half year). In the US-average land temperature, a simple ensemble average shows a long-term warming trend punctuated by cooling in response to volcanic eruptions, but it also has considerable seasonal-to-interannual noise superimposed (Fig. 5g). FP filtering identifies the same forced climate signal, but with almost all of this noise removed (Fig. 5h).

b. Forced precipitation and SLP responses

Identifying climate signals in surface temperature is generally easier than in other variables, because the pattern of global warming differs from dominant modes of temperature variability (see, e.g., Santer et al. 1994). To test whether the improved identification of climate responses by FP filtering extends to other variables, we consider seasonal precipitation and sea-level pressure (SLP) anomalies in the 40-member CESM-LE. For both variables, FP filtering considerably improves the agreement between halves of the CESM-LE on their estimates of the forced response, compared to a simple ensemble mean. Using the metric in Fig. 3b, FP filtering (with 9 patterns retained) improves the skill in identifying the spatiotemporal evolution of the forced response from 0.08 to 0.15 for precipitation and from 0.14 to 0.19 for SLP (cf. from 0.62 to 0.71 for surface temperature, Fig. 3b). While more noise remains in these variables after FP filtering, the fractional improvement is actually greater than for temperature.

Further improvement can be made by performing a combined analysis on all three variables. We will show the results from this three-variable analysis before returning to discuss how it differs from the single-variable analyses at the end of this section. For the multi-variable analysis, seasonal precipitation and SLP anomaly matrices are concatenated with the surface temperature anomaly matrix $\mathbf{X}$ in the spatial dimension (i.e. creating a new data matrix $\mathbf{X}$ with 3 times the spatial dimension). This is analogous to the generalization of EOF analysis to multiple field variables (Bretherton et al. 1992; Deser and Blackmon 1993). Each variable is normalized by the trace of its covariance matrix such that all variables are unitless and weighted equally. The rest of the multi-variable analysis proceeds exactly as in the single-variable case. By using a combined analysis of all three variables, we hope to take advantage of the relatively high SNR in surface temperature anomalies to identify contemporaneous forced responses in precipitation and SLP.

The first two multi-variable FPs show similar temperature anomaly patterns to those found in the single-variable analysis (Fig. 6, cf. Fig. 1). However, the multi-variable analysis additionally identifies contemporaneous precipitation and SLP anomaly patterns. Multi-variable FP-1 shows increasing SLP in the subtropics and midlatitudes and decreasing SLP in the Arctic and Antarctic (Fig. 6a), trends associated with the poleward shift of the storm tracks and jet streams in both hemispheres (Kushner et al. 2001; Yin 2005). The associated precipitation anomaly pattern shows on average that the dry subtropical regions get drier and the wet extratropical regions get wetter (Held and Soden 2006; Seager et al. 2010), but there is also considerable variability with longitude. Multi-variable FP-2 shows positive SLP anomalies in the Pacific and Indian oceans and negative SLP anomalies over SE Asia, North

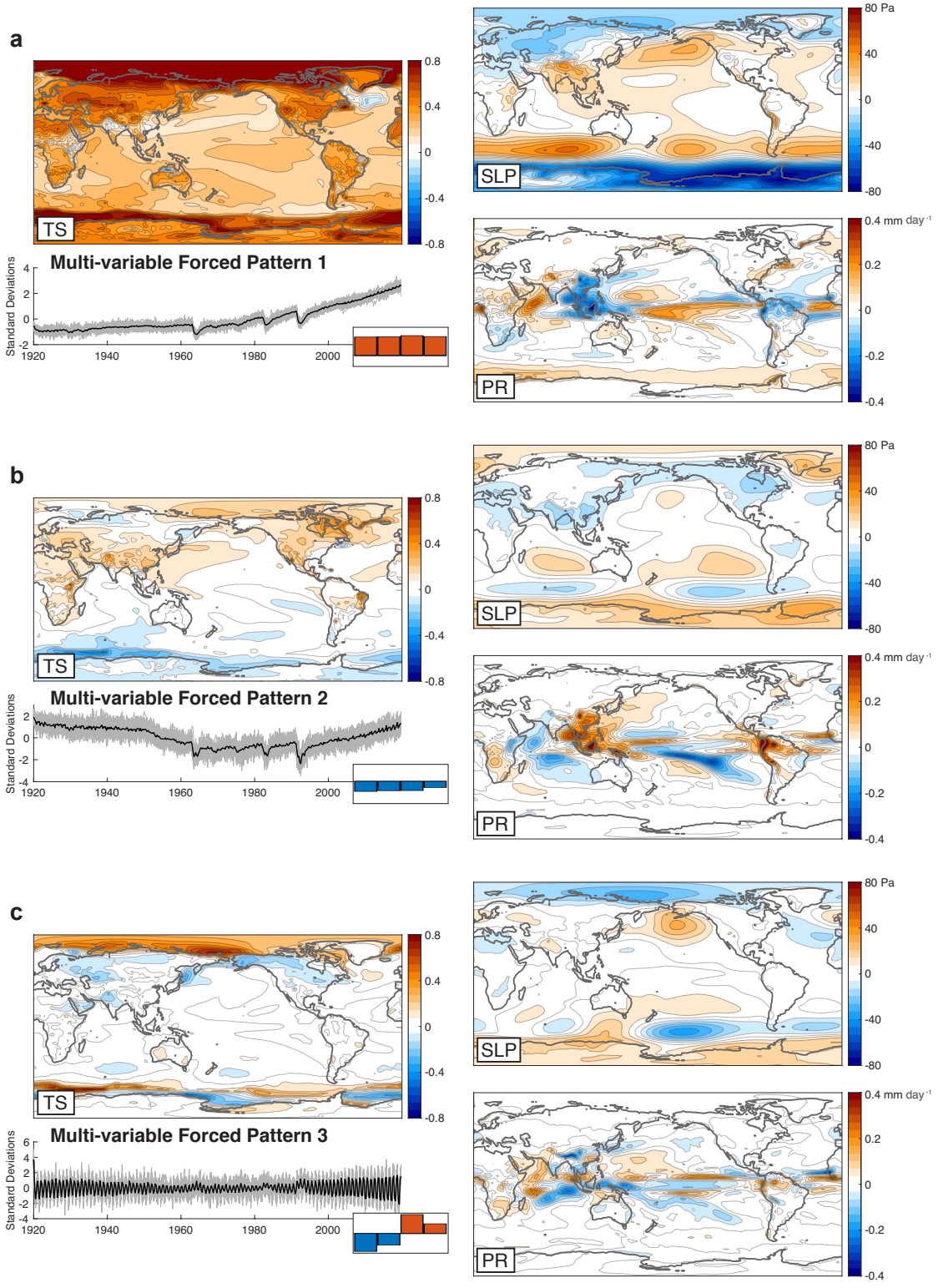

FIG. 6. Multi-variable forced patterns (FPs) of seasonal-mean surface temperature (TS), sea-level pressure (SLP), and precipitation (PR) anomalies in the CESM-LE historical and RCP8.5 simulations over the time period 1920-2019, with $N = 200$ EOFs retained. The time evolution of the FPs in all ensemble members are shown as standard deviation anomalies with grey lines. The black line shows the average ensemble-mean time evolution of each pattern. 100-year ensemble-mean trends in each pattern are shown separately for JFM, AMJ, JAS, and OND (left to right) in bar-chart insets. The y-scale for the bar-chart insets is half that for the timeseries in panel (a).

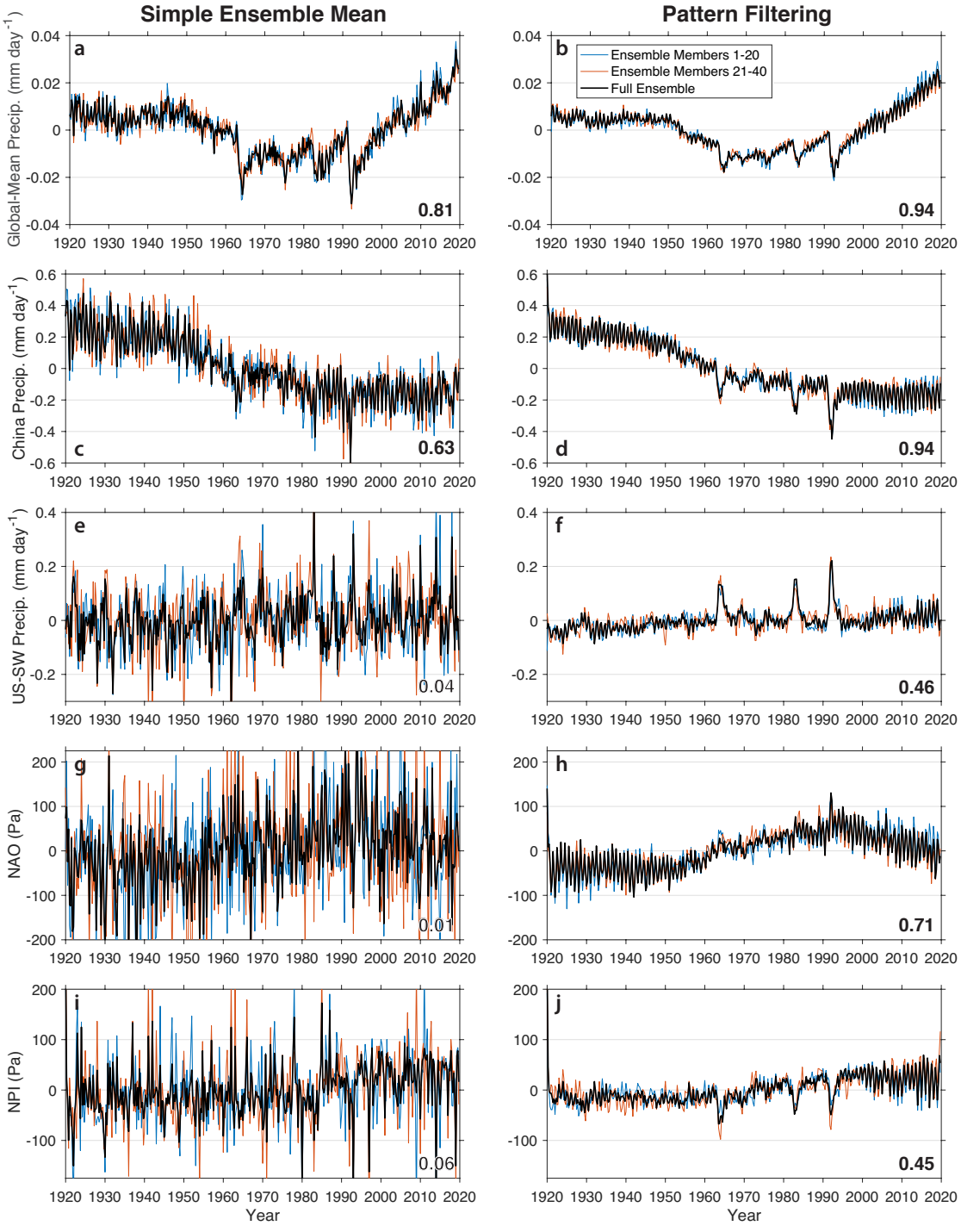

FIG. 7. Forced responses of large-scale precipitation and SLP indices in the CESM-LE, computed from (left) a simple ensemble mean $\langle \mathbf{X} \rangle$ and (right) the pattern filtered ensemble mean $\langle \mathbf{X}_{FP} \rangle$, from multi-variable FP filtering of seasonal surface temperature, precipitation, and SLP anomalies with $M = 8$ and $N = 200$. Blue and orange lines show the first and second 20-member half-ensembles of the CESM-LE, respectively. The black line shows the full 40-member CESM-LE. The squared correlation between the 20-member half-ensembles is shown in the bottom right of each panel. China precipitation is averaged over land in 100-120°E and 20-40°N, which includes small parts of Southeast Asia. US Southwest (US-SW) precipitation is averaged over land in 105-125°W and 30-40°N, which includes small parts of northwest Mexico. An approximate North Atlantic Oscillation (NAO) index is computed from the unnormalized SLP anomaly difference between Lisbon and Reykjavik, such that it is in units of Pa. The North Pacific Index is the average SLP anomaly over 160°E-140°W and 30-65°N, as in Trenberth and Hurrell (1994).

America, and the midlatitude Southern Ocean (Fig. 6b). It also shows a shift of the South Pacific Convergence Zone (SPCZ) towards the southwest and positive precipitation anomalies over China, Southeast Asia, and tropical South America. On average it shows a northward shift in precipitation, consistent with hemispherically asymmetric heating due to anthropogenic Northern Hemisphere aerosol loading (Broccoli et al. 2006; Kang et al. 2008).

The third multi-variable FP also has a somewhat similar temperature anomaly pattern to FP-3 in the single-variable analysis (Fig. 6c, cf. Fig. 1c), but is sufficiently different that its timeseries does not show the non-monotonic long-term trends that were present in the single-variable analysis. It instead shows only changes in seasonality (as evident by the annual cycle in the ensemble mean timeseries), a positive trend in the winter half-year in particular. It is associated with positive SLP anomalies over the Aleutian Low region and negative SLP anomalies over the Pacific sector of the Southern Ocean. Precipitation anomalies are weaker and of smaller spatial scale than those in multi-variable FP-1 and -2.

As with surface temperature, FP filtering improves the identification of forced responses in large-scale precipitation anomalies and SLP indices including global-mean precipitation, precipitation averaged over China (land within 100-120°E and 20-40°N), precipitation averaged over the United States Southwest (US-SW; land within 105-125°W and 30-40°N), the SLP difference between Lisbon and Reykjavik [an unnormalized variant of the North Atlantic Oscillation (NAO) index of Hurrell (1995)], and the North Pacific Index [NPI; SLP averaged over 160°E-140°W and 30-65°N, as in Trenberth and Hurrell (1994)]. Most of these forced responses have a low SNR and are therefore difficult to detect with simple ensemble averaging of 20-member of even 40-member ensembles (left side of Fig. 7). However, by FP filtering with the leading 8 multi-variable FP patterns (which maximizes the agreement with the ensemble mean of a 20-ensemble-member validation set for N = 200 EOFs / 79.9% of the total variance retained), both 20-member half-ensembles find the same forced responses in these precipitation and SLP indices (right side of Fig. 7).

With the exception of changes in global-mean precipitation (Fig. 7a,b), the forced responses uncovered by multi-variable FP filtering would be difficult to detect using more traditional methods. For example, while the long-term decreasing trend in China precipitation would be easy enough to detect in 20-member or even smaller ensembles using standard ensemble averaging or linear trend analysis (Fig. 7c), the reduction in precipitation following volcanic eruptions and the long-term trend in seasonality (towards wetter winters and drier summers) are not apparent until after the FP filtering is applied (Fig. 7d). In US-SW precipitation, the signal is small compared to internal variability such that it is completely swamped by noise, even when averaging over a 40-member ensemble (Fig. 7e). However, a weak but robust signal is found in both 20-member half-ensembles using FP filtering (Fig. 7f): increased precipitation following volcanic eruptions and a very small long-term positive trend (~ 0.1 mm day^{-1} century^{-1}). Recent work by Coats et al. (2015) has investigated whether external forcing, such as from volcanoes, has influenced long-term droughts in this region and concluded that they are dominated by internal variability. While we also find that internal variability is a bigger influence than external forcing on precipitation in this region, we find that volcanic eruptions lead to a detectable shift towards wetter conditions over the subsequent several years (in CESM), likely linked to the El-Niño-like response to eruptions.

SLP anomalies have very high amplitude internal variability, which is aliased into even the 40-member ensemble average (Figs. 7g and 7i). Long-term forced shifts in the NAO or NPI are therefore hard to detect, though there is much interest in knowing the relative contribution of forcing to observed trends (Hurrell 1995; Ulbrich and Christoph 1999; Semenov et al. 2008; Greatbatch et al. 2012; Scaife et al. 2014; Deser et al. 2017). FP filtering provides a means to characterize forced responses in these indices within large ensembles. The CESM-LE shows a forced positive trend in the NAO between 1950 and 1990 (Fig. 7h), corresponding roughly to the timing and magnitude of the observed trend over that period (Hurrell 1995; Ulbrich and Christoph 1999; Semenov et al. 2008), and a forced negative trend in the NAO between 1990 and 2019. In the Pacific, the CESM-LE shows a weak forced positive trend in the NPI over the entire century that is punctuated by negative anomalies following volcanic eruptions (Fig. 7j). Another interesting feature isolated by the FP filtering is a 200 Pa anomaly in the first three months of 1920, a symptom of the micro-initialization.

Multi-variable FP filtering uncovers a rich spatiotemporal complexity within the forced responses of precipitation and SLP in CESM-LE that would be lost on other methods. This does benefit from the use of surface temperature in the analysis, as single-variable analyses of precipitation or SLP alone do not give as good an agreement between the 20-member half-ensembles (reducing the squared correlations given on the right hand side of Fig. 7 to 0.61, 0.89, 0.21, 0.29, and 0.22, from top to bottom, compared to the multi-variable analysis values given in the figure). Single-variable FP filtering is still considerably better than a simple ensemble mean (cf. values on the left side of Fig. 7), with the notable exception of global-mean precipitation, forced changes in which are underestimated by single-variable FP filtering. The forced response of global-mean precipitation is retained in the multi-variable analysis (Fig. 7b), presumably because of its correlation with aspects of the surface temperature response. Overall, multi-variable FP filtering isolates the forced responses of

precipitation and SLP better than single-variable FP filtering, especially for global-mean precipitation.

4. How many ensemble members are needed?

Now that we have shown how FP filtering improves estimates of the forced response (for the specific case of 20-member half-ensembles), we will investigate how many ensemble members are needed to identify this forced response within the CESM-LE. To do so, we reserve one FP-filtered 20-member half-ensemble for comparison (CESM-LE members 21-40), which we will refer to as the reference estimate, and test how well this forced response can be identified within subsets of the remaining ensemble members. Sampling of the remaining 20 ensemble members is kept simple, with each n_E member ensemble constructed from members 1 to n_E of the CESM-LE. We find similar results with randomized ensemble member sampling (without replacement). For simplicity, we use $M = 7$ FPs for all ensemble sizes, though it would be an easy generalization to identify the optimal value of M for each ensemble size.

For the case of identifying the forced evolution of temperature at a grid point, FP filtering gives a dramatic improvement in squared correlation with the reference estimate compared to simple ensemble averaging (Figs. 8a and 8b). This is true for all ensemble sizes between 2 and 20 members. The FP-filtered estimate of the forced response based on 3 ensemble members is better than the simple ensemble average of 20 members, both in terms of squared correlation and root mean square error with the reference estimate. The FP-filtered estimate based on 2 ensemble members is only slightly worse. This means that FP filtering reduces the number of ensemble members needed to estimate the forced response by a factor of $\sim$7-10 compared to simple ensemble averaging.

We can characterize the number of ensemble members needed to estimate the forced response based on where the variance shared with the reference estimate exceeds a threshold (e.g., 80%). Based on the 80% threshold, 5 ensemble members are needed with FP filtering, while significantly greater than 20 ensemble members would be needed with simple ensemble averaging (Fig. 8a). We can map how many ensemble members are needed to detect the forced response in different local temperature anomalies by computing the number of ensemble members at which the sub-ensemble forced response estimate first exceeds an 80% squared correlation with the reference estimate (Fig. 9). Using a simple ensemble mean, more than 20 ensemble members are needed for about two thirds of grid points globally (Fig. 9a), whereas 2-3 ensemble members are generally enough to detect local forced responses with FP filtering (Fig. 9b). Only a few locations, such as the North Pacific, the Pacific sector of the Southern Ocean, the equatorial Pacific between 150°E-180°, India, and some regions of the North Atlantic (regions of small-scale and/or low-frequency variability) require greater than 10 ensemble members when using FP filtering.

Similar results hold for detecting forced responses in large-scale average temperature anomalies. For NA SST anomalies, 3 ensemble members are needed with FP filtering versus 10 with a simple ensemble mean (Fig. 8d); for US average land-surface temperature, 2 members are needed versus 14 (Fig. 8f). Fewer ensemble members are needed to capture the forced response in global-mean surface temperature: 2 ensemble members with FP filtering vs. 3 with simple ensemble averaging (here based on a stricter 95% variance criterion, Fig. 8c). The forced response in the Pacific SST gradient does not satisfy the 80% squared correlation criterion for any choice of ensemble size, but the squared correlation is not increasing further after about 7 ensemble members, suggesting that including more than 7-10 ensemble members in an estimate of the forced response (based on FP filtering) has marginal returns.

Similar results are found for the three other large ensembles (CSIRO-LE, GFDL-LE, and MPI-LE): using FP filtering, these ensembles require 10, 4, and 6 ensemble members, respectively, to meet the 80% threshold in global-mean squared correlation (cf. 8a). They need 2-4 ensemble members to meet the 95% squared correlation threshold for global-mean surface temperature, 2-4 ensemble members to meet the 95% squared correlation threshold for US temperature, and 2-12 ensemble members to meet the 95% squared correlation threshold for North Atlantic SST. None of the other ensembles reaches the 50% squared correlation for the east-west Pacific SST difference found with CESM-LE, not even the two 50-member sub-ensembles of MPI-LE. However, this could simply be a result of these models not having a strong response of the Pacific SST gradient to forcing over the past 100 years.

For all temperature indices except the Pacific SST gradient, FP filtering with 2-3 ensemble members already gives a reasonable estimate of the forced response, which raises the question of what can be done with a single ensemble member. We will answer this question in the next section.

5. Estimating the forced response from a single realization

a. Testing LFP filtering within the CESM-LE

For the case of a single ensemble member, or equivalently, observations, agreement on the timing of evolution of large-scale temperature anomaly patterns can no longer be used as a metric for whether they are forced or unforced. Another major difference between forced changes

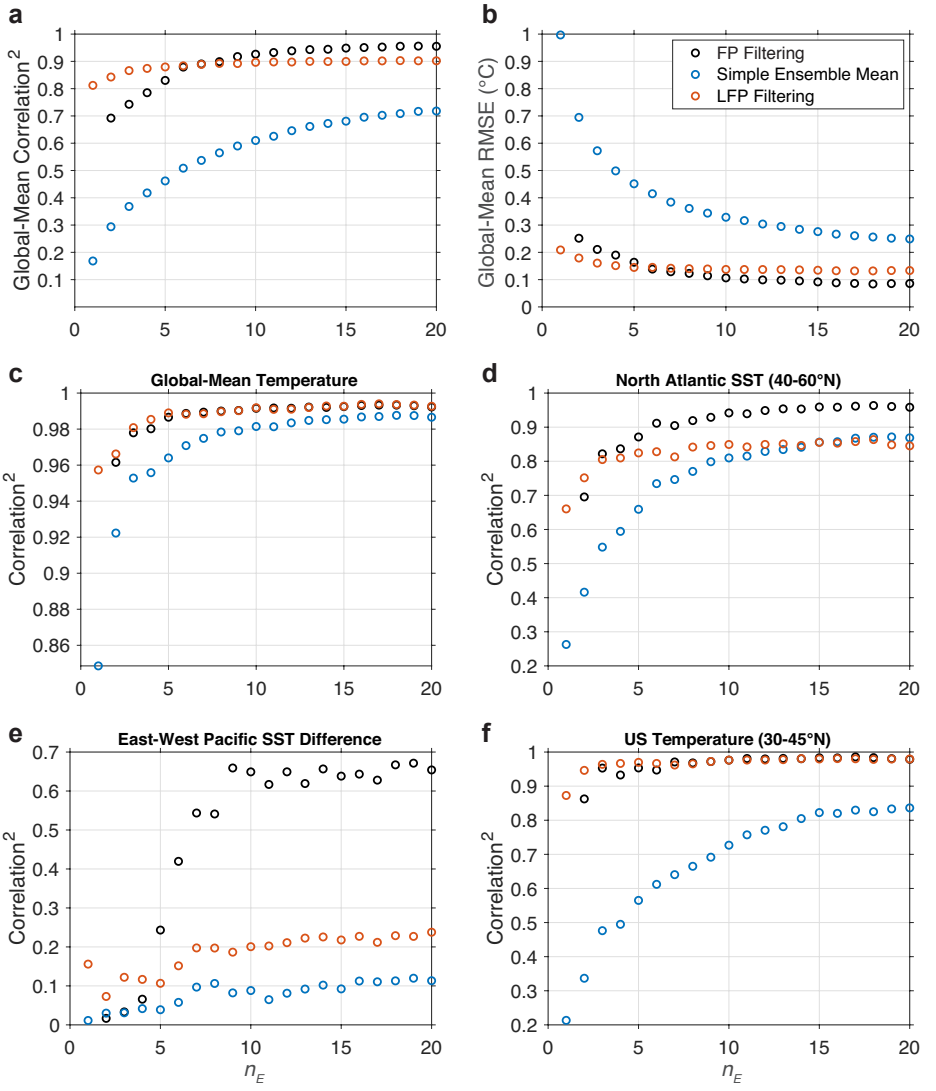

FIG. 8. (a) The global-mean grid-point squared correlation and (b) the root-mean square error (RMSE) between estimates of the forced surface temperature response in n_E-member subensembles and a reference estimate of the forced response, computed from FP filtering of 20 CESM-LE ensemble members that are withheld from the subensembles. Within the subensembles, the forced response is estimated by a simple ensemble mean (blue), FP filtering (black), and LFP filtering (orange). (c)-(f) Same as (a), except with spatial averaging computed before computing the squared correlation between forced response estimates. Spatial averages are computed as in Fig. 5, such that the values shown here for $n_E = 20$ match with those in Fig. 5.

and (most) internal variability is their longer timescale. We can take advantage of this longer timescale to identify patterns that are representative of the forced response. This was first proposed by SH01, who solved for patterns of global surface temperature anomalies that maximize the variance between decadal means relative to the total variance. This was further explored by W18, who solved for patterns of Pacific SST anomalies that maximized the ratio of low-frequency (lowpass filtered) to total variance and found that this can cleanly separate long-term warming from variability associated with the Pacific Decadal Oscillation (PDO) and ENSO. Here, we use the CESM-LE to test how well the method used in W18 (and described in Section 2b) can isolate the forced climate response within a single realization.

First, we show the low-frequency patterns (LFPs) of the full 40-member CESM-LE (Fig. 10). We retain only 50 EOFs in the analysis (vs. 150 in FP filtering), amounting to 76.7% of the total variance, because there are fewer degrees of freedom in a lowpass filtered 100-year timeseries than there are in the full 100-year timeseries. The leading LFP shows a global warming pattern, with am-

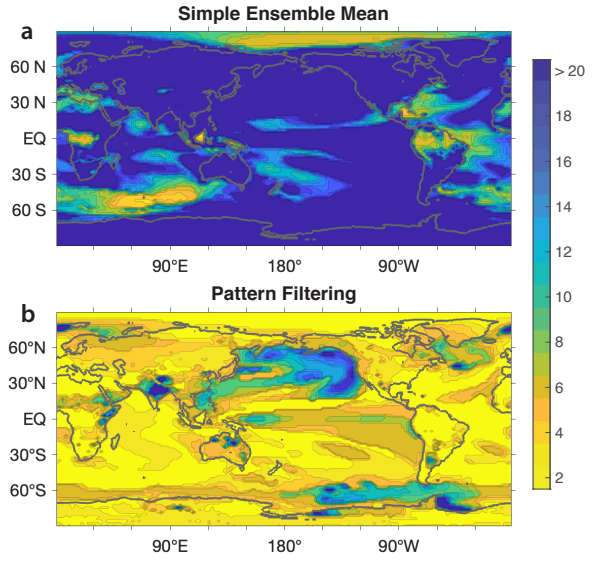

FIG. 9. The number of ensemble members needed to constrain the forced response in local temperature using (a) a simple ensemble mean and (b) FP filtering. The criterion used is that the forced response must share 80% of its variance with the reference estimate (FP filtering of the opposite 20-member half-ensemble, i.e., no ensemble members are shared between the estimate and the reference).

plified warming over land and at high latitudes, similar to the leading FP (Fig. 10a, cf. Fig. 1a; pattern correlation = 0.9995). The second LFP shows cooling of the North Atlantic, Arctic, and Northern Hemisphere land through the 1950s and 60s and a subsequent recovery, as well as opposite signed changes in the Southern Ocean (Fig. 10b), similar to FP-2 (Fig. 1b; pattern correlation = 0.95). The variance amongst ensemble members is somewhat greater for LFP-2 than for FP-2, likely because of the greater projection onto the region of Atlantic multidecadal variability (Enfield et al. 2001; Wills et al. 2019a; Zhang et al. 2019). The third LFP shows low-frequency internal variability associated with the PDO (Fig. 10c) (Mantua et al. 1997; Newman et al. 2016; Wills et al. 2019b). There is only a small excursion in the ensemble mean timeseries, before 1930, resulting from memory of common ocean initial conditions in January 1920. The fourth LFP also shows somewhat PDO-like low-frequency internal variability, but with opposite signed anomalies in the Greenland, Norwegian, Barents, and Kara Seas (Fig. 10d). It shows little agreement on the timing of its evolution amongst ensemble members, except for a small response to the 20th-century volcanic eruptions. The remaining LFPs show internal variability with increasingly shorter timescales.

As with FP filtering, we need to choose how many patterns to include in estimating the forced response. Using the CESM-LE, we can determine the ratio of forced signal to total variance s_k for each LFP. The only LFPs that exceed the $s_k \approx 0.15$ cutoff used in the FP filtering analysis are LFP-1 ($s_k = 0.95$), LFP-2 ($s_k = 0.62$), and LFP-48 ($s_k = 0.27$). LFP-48 is not low-frequency (i.e., it has low r_k); it shows primarily changes in the seasonal cycle and will be excluded here. However, this suggests that a more careful treatment of seasonality (e.g., filtering each season separately) could further improve the isolation of forced responses in the leading LFPs. LFP-3, for comparison, has $s_k = 0.09$. We therefore include the leading 2 LFPs in an estimate of the forced response. Applying LFP filtering to individual ensemble members, we also find that $M = 2$ patterns maximizes the agreement with a reference estimate (the ensemble mean of 20 ensemble members not included in the LFP filtering).

We find that LFP filtering of a single-ensemble member provides a better estimate of the forced response than a 20-member ensemble mean (Fig. 8), capturing more than 80% of the spatiotemporal variations in the forced response as diagnosed by the reference estimate. It remains the best method to estimate the forced response for up to about 3-5 ensemble members (depending on the metric used), beyond which FP filtering is the best method. For global-mean surface temperature (Fig. 8c) and US land surface temperature (Fig. 8f), LFP filtering remains as good an estimate of the forced response as FP filtering for up to 20 ensemble members. The benefits of LFP filtering are not as clear for ocean regions with substantial low-frequency internal variability, such as for the Pacific SST gradient and NA SST anomaly (in terms of squared correlation), but the RMSE is substantially reduced. The reduction in RMSE can be seen in Fig. 11, which shows the distribution of individual ensemble member timeseries before and after applying LFP filtering. LFP filtering reduces the spread in the responses by a factor of 2 for global-mean surface temperature and by as much as a factor of 10 in other metrics (note the different y-axes). LFP filtering does remove some signals, such as the El-Niño-like response to volcanic eruptions and some of the changes in seasonality. The latter would likely be improved by low-pass filtering each season separately within the LFCA.

b. Filtering with Linear Inverse Models

With similar goals in mind, Frankignoul et al. (2017) described an optimal perturbation filter (LIMopt) based on linear inverse models (LIMs), and showed that it is among the best available methods for determining the forced climate response from a single realization. Specifically, they considered methods that do not require multiple ensemble members and compared the LIMopt method to a linear trend, quadratic trend, regression against global-mean SST, and multi-variate ensemble empirical mode decomposition. We have also tested the LIMopt method for the isolation of the forced response from subsets of the

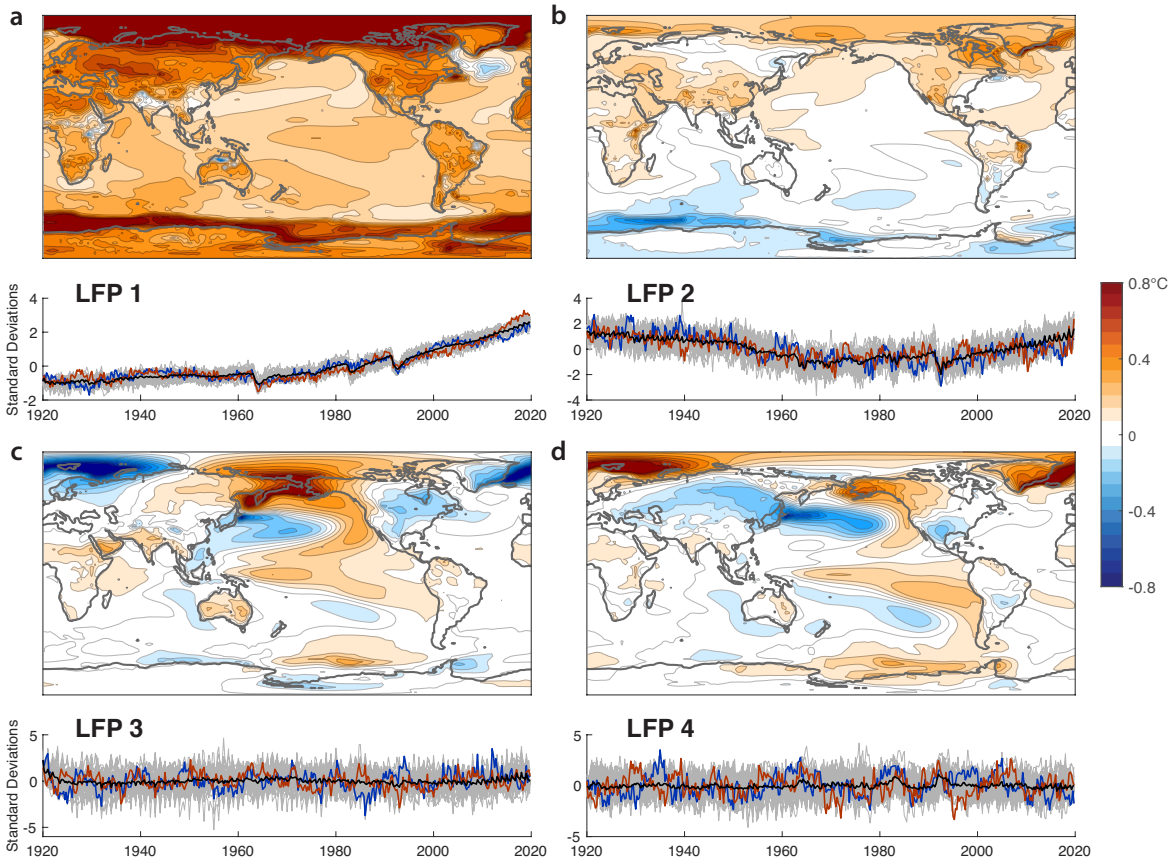

FIG. 10. Low-frequency patterns (LFPs) of seasonal-mean surface temperature anomalies in the CESM-LE historical and RCP8.5 simulations over the time period 1920-2019, with $N = 50$ EOFs retained. The time evolution of the LFPs in all ensemble members are shown as standard deviation anomalies with grey lines. The orange (blue) lines show show the ensemble member with the most (least) change in LFP-1 over 2000-2019. The black line shows the ensemble-mean time evolution of each pattern. Modified from Wills et al. (2017).

CESM-LE (see Supplementary Material). We find that LFP filtering performs better for global-mean surface temperature and for grid-point temperatures, and that it has skill equal to or greater than LIMopt for most large-scale temperature metrics. Furthermore, LFP filtering scales better with the addition of further ensemble members. Comparing with the work of Frankignoul et al. (2017), this also means that LFP filtering isolates the forced response within individual ensemble members better than a linear trend, quadratic trend, regression against global-mean SST, or multi-variate ensemble empirical mode decomposition.

c. Application to HadCRUT4 Observations

Given the success of LFP filtering in estimating the forced response from individual ensemble members (Figs. 8 and 11), we would like to see what this method can tell us about the forced response in observations. We examine the HadCRUT4 infilled observational surface temperature product (Cowtan and Way 2014). We compute the LFPs of seasonal (3-monthly) surface temperature anomalies over the period 1920-2019, retaining 50 EOFs (78.1% of the total variance). While the infilling of missing data can in general lead to biases in the estimated covariance matrix and thus in the LFPs, we find similar results when using HadCRUT3 data imputed with a regularized expectation maximization algorithm (Schneider 2001) (not shown).

LFP-1 and LFP-2 of observed temperature anomalies are similar to LFP-1 and LFP-2 of the CESM-LE (pattern correlations of 0.92 and 0.59, respectively). This suggests that LFP filtering with $M = 2$ LFPs would help to remove variability not associated with the forced response, as in the large ensemble. LFP-3 and LFP-4 are both somewhat PDO-like (cf. W18), giving additional motivation to exclude them from the LFP filtering.

Most long-term trends in observations can be attributed to the first two LFPs (Fig. 13). Over the full century, the influence of the residual is small, and most temperature

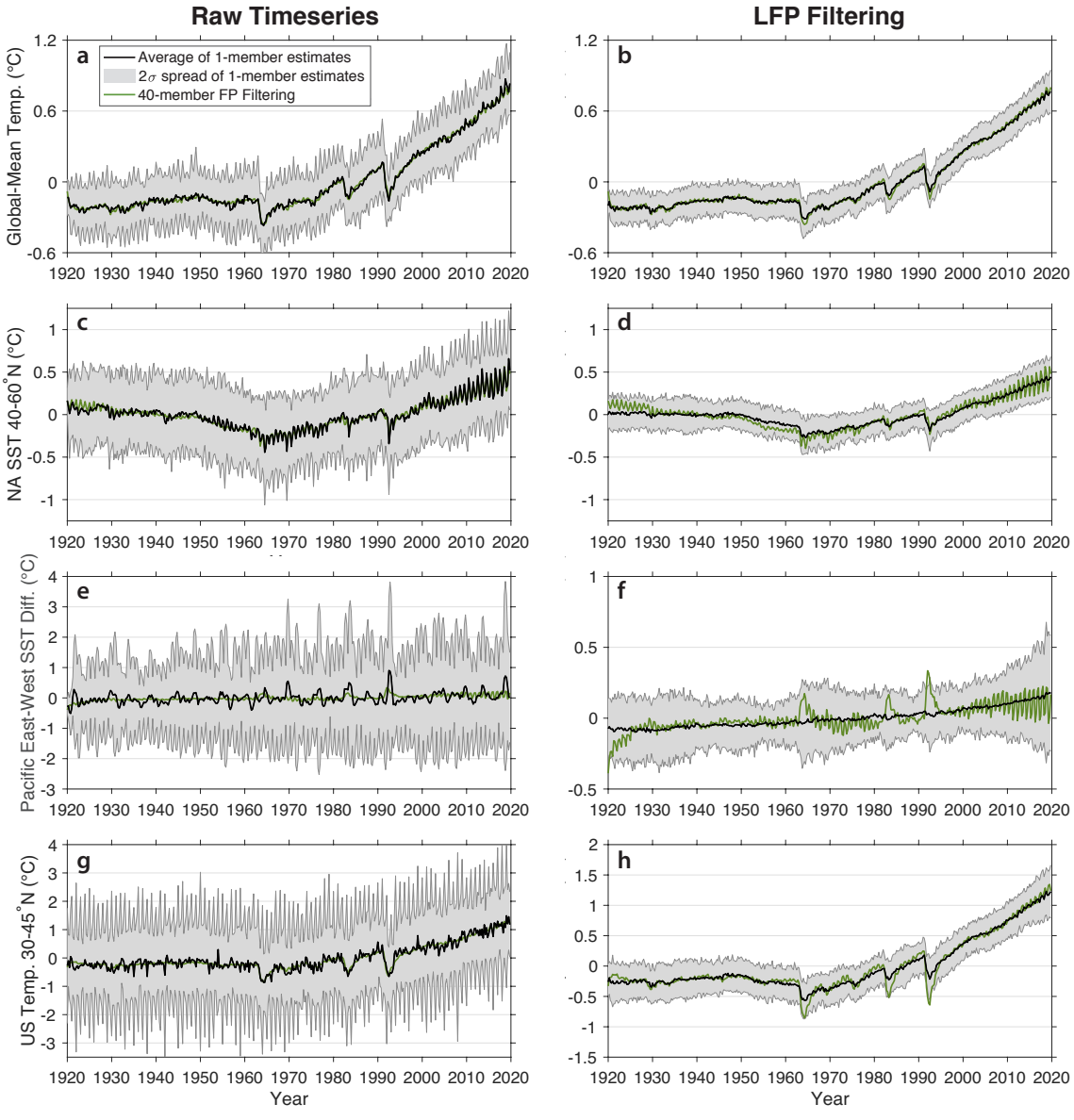

FIG. 11. (left) Spread in time evolutions of large-scale temperature indices in individual members of the CESM-LE and (right) spread in time evolutions of the same large-scale temperature indices after the application of LFP filtering in individual members of the CESM-LE. Averaging regions for the large-scale temperature indices are defined in the caption of Fig. 5. Note the different y-axis scales for the Pacific east-west SST difference and the US land surface temperature. For reference, the forced response estimate from FP filtering of the full 40-member CESM-LE (as in the right-hand side of Fig. 5) is shown in green (same on left and right).

changes are captured by the LFP-filtered data. Over 1939-1978, Northern Hemisphere cooling, which is thought to result in part from aerosol forcing, is retained in the LFP-filtered data. Over this period, there is additionally a negative PDO-like trend in the Pacific and a weak cooling trend in the Atlantic (captured by the residual). The recent trend over 1979-2019 is largely captured by the LFP-filtered data, except for a negative PDO-like trend in the Pacific and a weak cooling trend in the Atlantic.

We also use LFP filtering to examine the slow component of observed changes in key large-scale temperature indices (Fig. 14). Almost all of the observed global-mean surface temperature changes and much of the observed Atlantic multi-decadal variability remain in the LFP-filtered data. The Pacific east-west SST gradient is dominated by high-frequency internal variability (i.e., ENSO), but it also exhibits a slow La-Niña-like trend since 1980. Note, how-

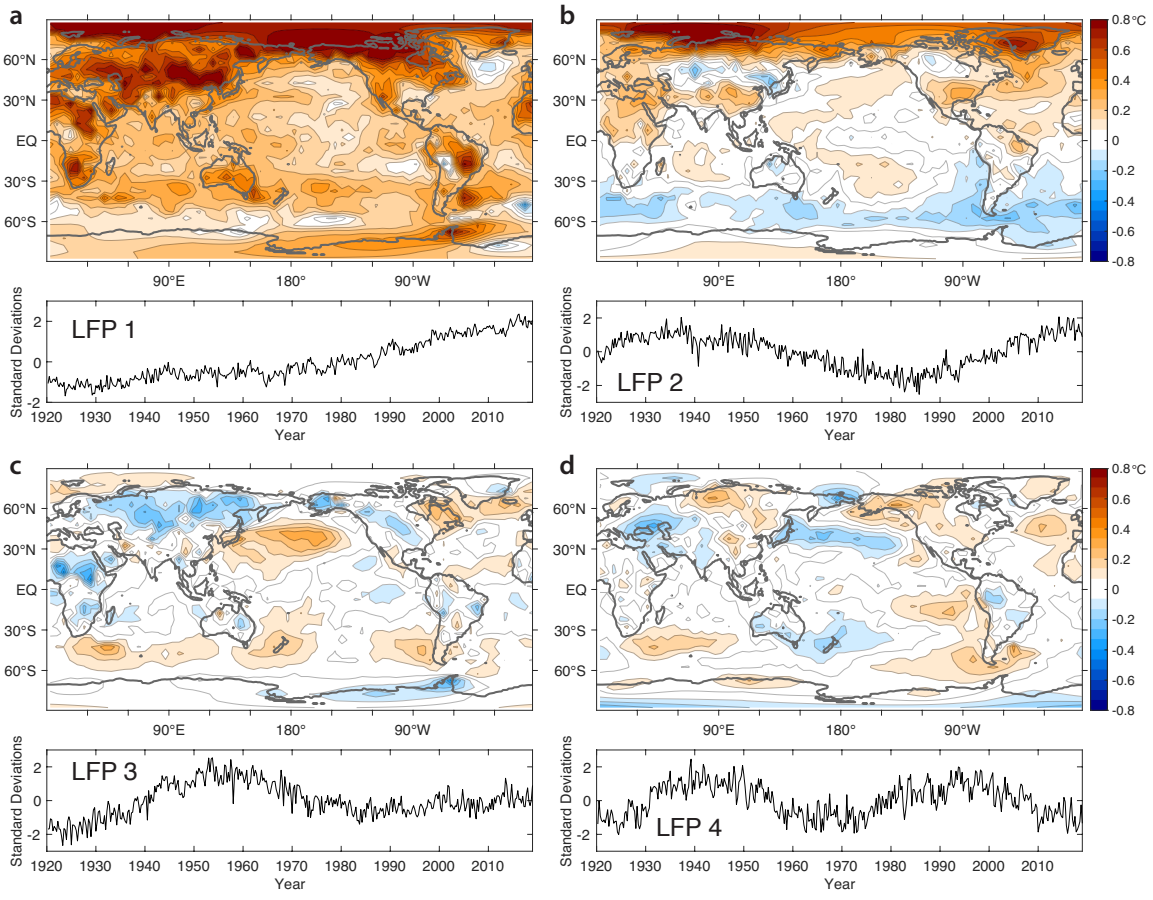

FIG. 12. Low-frequency patterns (LFPs) of seasonal surface temperature anomalies, and their time evolution in standard deviation anomalies, from the infilled HadCRUT4 (Cowtan and Way 2014) observational product over the time period 1920-2019, with $N = 50$ EOFs retained.

ever, that the LFP-filtered trend in the east-west SST gradient is smaller than the trend in the raw data (Fig. 13).

Interpreting this observational analysis in the context of the results from our LFP-filtering analysis of the CESM-LE (Figs. 8, 10, and 11) may give insight into the forced and unforced components of observed temperature changes. In particular, Fig. 8 suggests that the LFP filtering gives a good estimate of the forced component of changes in large-scale temperature indices from a single realization, roughly equivalent to an estimation of the forced response from a 5-member ensemble mean. This means that the LFP-filtered timeseries in Fig. 14 approximate the forced responses in these indices. However, it is important to keep in mind that the analysis is only guaranteed to isolate the slow component, which happens to be a better approximation of the forced response than the full unfiltered dataset in most cases. The LFP-filtered timeseries can still contain some amount of low-frequency internal variability, and should be interpreted with the spread in Figs. 11b, 11d, and 11f in mind.

The LFP-filtered observations are broadly consistent with the forced component (based on FP filtering) of temperature changes in four different large ensembles (Fig. 15): CESM-LE (Kay et al. 2015), CSIRO-LE (Jeffrey et al. 2013), GFDL-LE (Sun et al. 2018), and MPI-LE (Maher et al. 2019). One model (GFDL-CM3) has too much mid-century cooling of both global-mean temperatures and subpolar North Atlantic SSTs, suggesting that its aerosol forcing may be too strong. It also seems to overestimate warming in the past two decades, suggesting that its climate sensitivity may be too high. Another model (MPI-ESM) has too little mid-century cooling of subpolar North Atlantic SSTs, suggesting that its aerosol forcing may be too weak. This is consistent with a diagnosis of aerosol radiative forcing based on simulations with fixed SST (Booth et al. 2018), where these two models span the range of diagnosed aerosol forcing strength in CMIP5 models. In general, the models show mid-century cooling of the subpolar North Atlantic that occurs earlier than in observations (Fig. 15b), though the timing in observa-

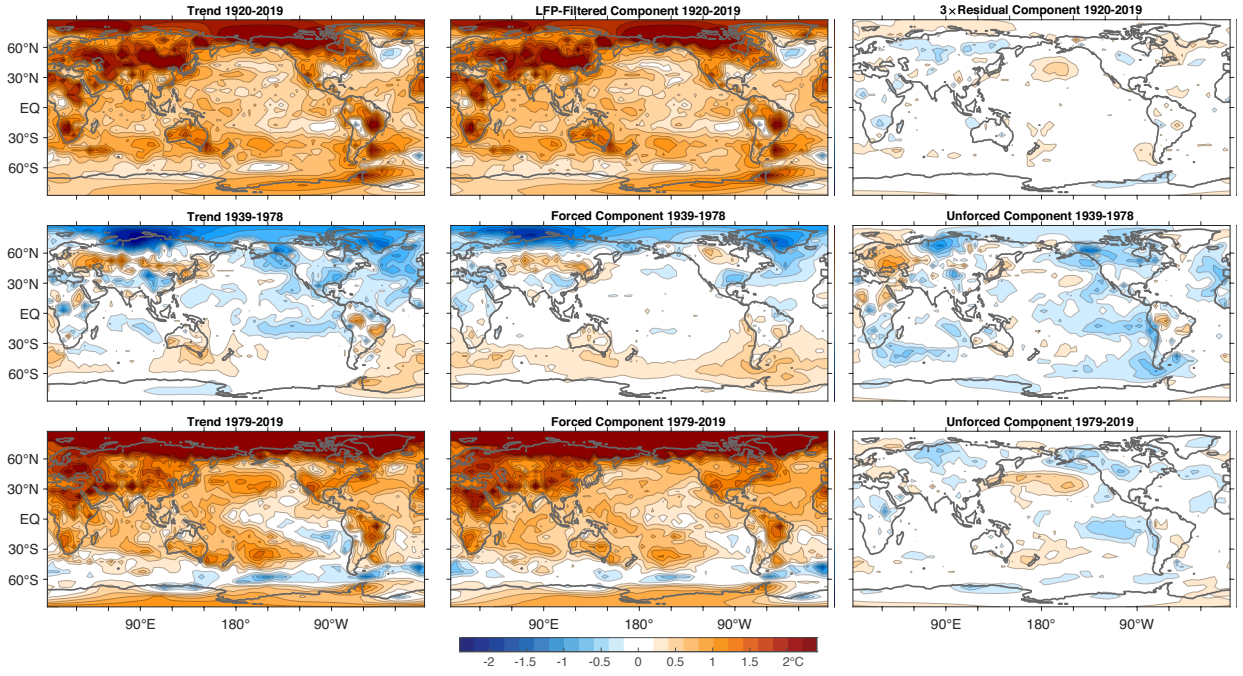

FIG. 13. Partitioning of observed trends into an LFP-filtered component, based on LFP filtering with $M = 2$ LFPs included and $N = 50$ EOFs retained, and a residual. Trends are shown in units of °C per trend length (e.g., °C per 40 yr). Note that the residual component of 1920-2019 temperature trends is multiplied by a factor of 3 for ease of comparison.

tions could also be influenced by Atlantic multi-decadal variability. The response of the Pacific east-west SST difference varies across models from positive (El-Niño-like) to weakly negative (La-Niña-like) (Fig. 15c). None of the other models show as strong of an El-Niño-like response to volcanic eruptions as CESM. Observations show a La-Niña-like trend between the 1970s and present that is outside of the range of model forced responses (Fig. 15c), as has been found in several studies looking at the full 20th century (Cane et al. 1997; Solomon and Newman 2012; Coats and Karnauskas 2017). A response to volcanic eruptions is not apparent in the LFP-filtered observations, but this could be a result of the LFP-filtering itself (cf. Fig. 11). The best agreement between LFP-filtered observations and ensemble-based estimates of the forced response is found with $M = 1$ observational LFP, but the estimate with $M = 2$ observational LFPs remains in good agreement with the ensemble-based forced response estimates (Fig. 16). The reason including LFP-2 reduces agreement with the models might be because the observational LFP-2 reaches its minimum somewhat later than the CESM-LE LFP-2, in the mid 1980s instead of around 1970 (Figs. 10 and 12). Overall, the forced responses in the CESM-LE and the MPI-LE have the highest correlation with the observational record (Fig. 16).

The observed trend in temperature asymmetry between the Northern and Southern Hemispheres during the period 1939-1978 shows up in the LFP-filtered component in our analysis (Fig. 13), but only if 2 LFPs are included. This trend in hemispheric asymmetry could have been caused by anthropogenic aerosols (Booth et al. 2012; Tandon and Kushner 2015; Bellucci et al. 2017; Bellomo et al. 2018; Watanabe and Tatebe 2019), stratospheric ozone changes (Thompson et al. 2011), unforced AMOC variability (Semenov et al. 2010; DelSole et al. 2011; Chen et al. 2017), or a transient response of ocean circulations to climate change (Armour et al. 2016; Stolpe et al. 2018). The key to disentangling the forced and unforced components of observed global temperature changes lies in distinguishing between these hypotheses. LFP filtering provides a potential path forward by identifying the main slowly changing temperature pattern (LFP-2) in need of attribution. Climate model ensembles with individual forcing from greenhouse gasses, aerosols, and ozone may provide utility in attributing these hemispherically asymmetric temperature changes.

Overall, estimates of the forced and unforced components of observed temperature trends based on LFP filtering largely agree with other estimates in the literature (Frankcombe et al. 2015; Frankignoul et al. 2017; Bellucci et al. 2017; Stolpe et al. 2017, 2018; Haustein et al. 2019), with the exception of T09, DelSole et al. (2011), and Chen et al. (2017), who use related statistical analyses but suggest that only the first pattern is forced and there-

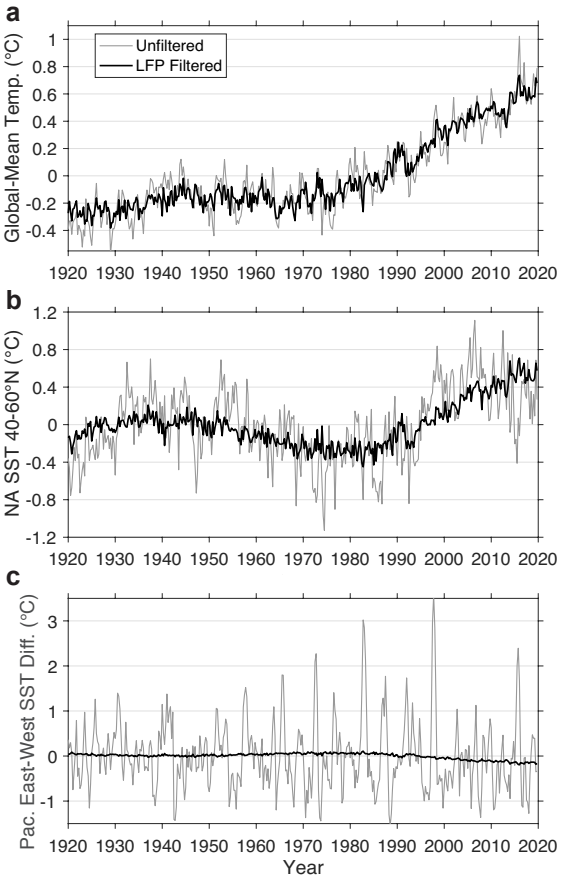

FIG. 14. Time evolution of (a) global-mean surface temperature, (b) North Atlantic SST averaged over 40-60°N (i.e., the North Atlantic warming hole), and (c) the SST difference between the eastern and western equatorial Pacific (averaging regions as in Fig. 5) in HadCRUT4 (Cowtan and Way 2014), before and after applying LFP filtering.

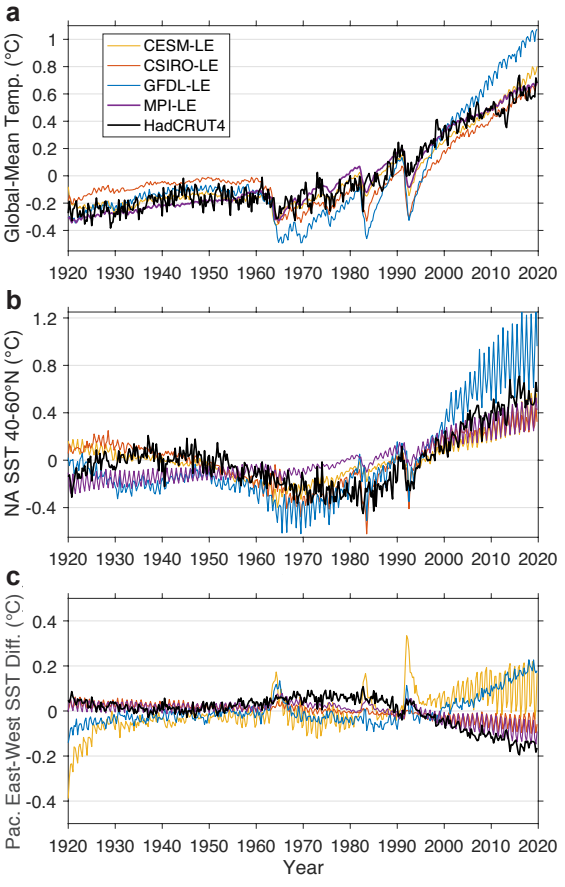

FIG. 15. Comparison across four single-model large ensembles and HadCRUT4 observations of the time evolution of (a) global-mean surface temperature, (b) North Atlantic SST averaged over 40-60°N (i.e., the North Atlantic warming hole), and (c) the SST difference between the eastern and western equatorial Pacific (averaging regions as in Fig. 5). In models, the timeseries shown are averaged over the full ensemble after application of FP filtering. In the analysis of CESM-LE, CSIRO-LE, GFDL-LE, and MPI-LE, we choose a number of EOFs to retain between 89% and 90% of the total variance (150, 200, 135, and 200, respectively); we choose the number of FPs based on a criterion that SNR > 0.15 (7, 6, 7, and 5, respectively). The observations are LFP filtered, as shown in Fig. 14.

fore conclude that a large portion of recent warming can be attributed to internal climate variability. In the case of T09 and DelSole et al. (2011), this comes from requiring that forced responses show up in a multi-model average, which could average out aerosol-forced climate responses that differ in pattern, strength, or timing between models.

6. Discussion and conclusions

a. Summary and conclusions

Here, we have demonstrated how FP filtering reduces the ensemble size needed to identify forced responses. Within the CESM-LE, this uncovers forced responses that were not otherwise apparent, such as an El-Niño-like response to volcanic eruptions, increased (decreased) precipitation in the US-Southwest (China) following volcanic eruptions, forced trends in the NAO, and regional changes in the seasonality of temperature, precipitation, and SLP. While all of these signals have a small SNR in a particular year or season, this method uncovers the time progression of local climate change signals that, when averaged over 30 or so years (or sufficient volcanic eruptions), would be statistically significant. The details of the diagnosed forced responses differ across models, but in all four large ensembles tested, FP filtering identifies the forced response with fewer ensemble members than a simple ensemble average. The inclusion of at least two degrees of freedom (patterns of change) in the forced response is critical in all cases, suggesting that methods that include only one pattern of forced response will generally underestimate the contribution of external forcing to observed tem-

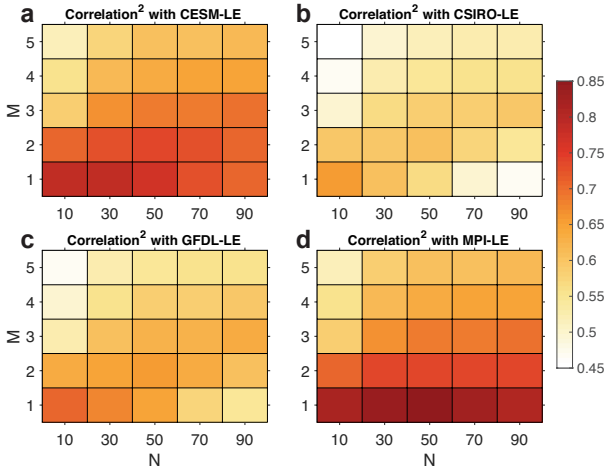

FIG. 16. Global mean of grid-point squared correlation between $\langle \mathbf{X}_{LFP} \rangle$, computed entirely from the HadCRUT4 observational product, and $\langle \mathbf{X}_{FP} \rangle$, computed from 4 difference large ensembles, over the time period 1930-2019 (excluding micro-initialization spin-up period), as a function of the number M of LFPs included and the number N of EOFs retained in the observational LFP filtering. The values of M and N used in the FP filtering are given in the caption of Fig. 15.

perature changes. Dynamical adjustment (Wallace et al. 2012; Smoliak et al. 2015; Deser et al. 2016; Sippel et al. 2019) may perform similarly for some applications, but does not allow for the detection of forced atmospheric circulation responses, as were identified in the CESM-LE.

Using this pattern-recognition-based method for estimating the forced response within climate model ensembles, we revisited the question of how many ensemble members are needed to isolate the forced climate response from internal variability. We tested the number of ensemble members needed (from one half of the CESM-LE) to converge on the same forced response estimate as was obtained from the other half of the CESM-LE. The answer depends on the particular climate response of interest and on the error tolerance level. For global-mean surface temperature, even a simple ensemble mean is able to isolate the forced response with about 3 ensemble members. However, FP filtering is able to isolate the forced global-mean surface temperature response with 2 ensemble members and LFP filtering with a single ensemble member. In order to capture 80% of the full spatiotemporally variable climate response globally, more ensemble members are required (5 when using FP filtering). This is a large improvement over simple ensemble averaging, which would need well over 20 ensemble members to reach this threshold. Even in noisy climate metrics such as the tropical Pacific SST gradient, US-Southwest precipitation, or the NAO, the addition of ensemble members beyond an ensemble size of about 10 has marginal returns for the identification of the forced response. For future modeling efforts, increasing the number and quality (e.g., resolution) of, e.g., 5-member or 10-member ensembles would provide greater benefit than increasing the ensemble size.

Using the CESM-LE as a testbed, we showed that LFP filtering can give an estimate of the forced response from a single realization (ensemble member), although it can miss rapid forced signals such as the response to volcanic eruptions. LFP filtering differs from simple lowpass filtering because it includes information about the spatiotemporal structure of the high-frequency noise in order to optimally filter it out. LFP filtering of a single ensemble member captures more than 80% of the spatiotemporal variance in the ensemble's forced climate response. With these results as motivation, we used LFP filtering to approximate the forced and unforced components of observed temperature trends, without using any model-based information. Our results support the conjecture that most of the multidecadal changes in global-mean surface temperature and North Atlantic SST are forced and that there has been an externally forced strengthening of the tropical Pacific SST gradient over the past four decades. This approach to estimating the forced response from observations provides an alternative to approaches that use both observational and model-based information (e.g., detection and attribution), which are subject to model-biases in the forced response.

b. Generalizability

The number of ensemble members needed to isolate forced climate signals will depend in general on the amplitude of the signal of interest and the characteristics of the noise in the model used. We have focused on simulations of global climate change over 1920-2019, where the forced response is comparable in amplitude to modes of internal variability. Fewer ensemble members would be needed to isolate the forced climate response in simulations with stronger forcing, such as simulations of 21st century climate change or of a quadrupling of CO_2. Properties of the internal variability within climate model ensembles and observations also influence the ability to isolate the forced response. Higher amplitude noise from internal variability does not necessarily make climate responses harder to detect, because this high-amplitude noise could all be contained in a few spatial patterns (e.g., ENSO). The climate variability that is most difficult to remove from estimates of the forced response is that which is on small spatial scales (such that it doesn't show up in the leading EOFs) and/or on long timescales (such that it has fewer temporal DOF).

A number of studies have pointed out that observations are more predictable than expected from comparison to individual members of climate model ensembles despite similar amplitudes of climate variability in models and observations, especially on seasonal-to-decadal timescales in the North Atlantic (Scaife et al. 2014; Eade et al. 2014;

Scaife and Smith 2018). One potential explanation for this so-called 'signal-to-noise paradox' is that the fraction of atmospheric variability driven by variations in SST is larger in observations than in models, such that a model that is able to correctly predict the evolution of SSTs may correctly predict the timing (but not the amplitude) of observed atmospheric variability once the unpredictable atmospheric noise is averaged out (see, e.g., Simpson et al. 2018). The implications of the 'signal-to-noise paradox' for the skill of pattern filtering in isolating the forced climate response are not clear cut; more unpredictable atmospheric noise in models would make it harder to isolate the forced response in models (and therefore overestimate the difficulty in observations), but more multi-decadal coupled atmosphere-ocean variability in observations would pose a challenge for isolating the forced response in observations. Based on this literature, we have no reason to believe that our analysis in Section 5 systematically overestimates or underestimates what can be learned about the forced climate response from a single realization.

One limitation of the pattern filtering methods presented here is that they only consider linear combinations of state variables. This may lead to underestimates of nonlinear climate responses (e.g., in cases where positive and negative anomalies have different patterns or amplitudes). This may be apparent in the estimated El-Niño-like response to volcanic eruptions (Fig. 5f, cf. Fig. 5e). Looking forward, future work should investigate whether nonlinear machine learning methods can be constructed that take advantage of patterns with high signal-to-noise ratio, in a similar spirit to the analyses shown here (e.g., Barnes et al. 2019).

c. *Further applications*

Estimates of forced responses from pattern filtering are complimentary to estimates of the uncertainty in long-term trends, as can be computed from unforced variability in control runs or observations (Thompson et al. 2015; McKinnon et al. 2017). In order to characterize the unforced variability in observations, these studies rely on removing the forced response, either through detrending or the subtraction of a model-based forced response estimate. However, some of the variability about the long-term trend likely comes from aerosol forcing and other non-monotonic forcing, as encompassed in LFC-2 of observed temperatures (Fig. 12). If these non-monotonic forced responses are not fully removed (e.g., if there are biases in the modeled forced response), then this may bias the estimates of unforced variability in observations. By first removing non-monotonic forced responses using LFP filtering, the uncertainty in long-term trends that results from internal variability could be better estimated from observations.

Separating the forced response from the internal variability also helps to understand internal decadal variability, which may lead to better decadal climate predictions (Meehl et al. 2009). Current methods of removing the forced component from indices of internal variability, such as removing the linear trend (Enfield et al. 2001) or global-mean SST (Trenberth and Shea 2006), will become less effective as the forced climate change pattern changes over time (Andrews et al. 2015). LFP filtering provides a way to identify and remove the forced response from indices of climate variability.

Pattern filtering methods can also provide utility for the analysis of multi-model ensembles (e.g., CMIP), as shown in Ting et al. (2009) and DelSole et al. (2011). However, if the timing of a particular forced response pattern differs across models, application of FP filtering to a multi-model ensemble would filter this response out. Therefore, it is generally preferable to apply pattern filtering analyses to each climate model separately in order to analyze inter-model differences in the forced climate response.

Overall, the common framework of FP and LFP filtering provide a powerful tool for separating forced and unforced components of climate change, thereby identifying the full spatiotemporal complexity of the climate system's response to radiative forcing.

Acknowledgments. R.C.J.W. and D.S.B. acknowledge support from the National Science Foundation (Grant AGS-1929775) and the Tamaki Foundation. R.C.J.W. and K.C.A. acknowledge support from the National Science Foundation (Grant AGS-1752796). R.C.J.W. is also supported by the University of Washington eScience Institute. T.S. is supported by Eric and Wendy Schmidt by recommendation of the Schmidt Futures program and by the Earthrise Alliance. The CESM project is supported primarily by the National Science Foundation (NSF). This material is based on work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the NSF under Cooperative Agreement no. 1852977. We thank Dennis Hartmann, Cristian Proistosescu, Flavio Lehner, Elizabeth Maroon, Mingfang Ting, and David Bonan for valuable input on this work. The code for FP filtering is available at github.com/rcjwills/forced-patterns. The code for LFCA is available at github.com/rcjwills/lfca.

References

Allen, M. R., and L. A. Smith, 1997: Optimal filtering in singular spectrum analysis. *Phys. Lett. A*, **234 (6)**, 419–428.

Andrews, T., J. M. Gregory, and M. J. Webb, 2015: The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models. *J. Climate*, **28 (4)**, 1630–1648.

Armour, K. C., J. Marshall, J. R. Scott, A. Donohoe, and E. R. Newsom, 2016: Southern Ocean warming delayed by circumpolar upwelling and equatorward transport. *Nature Geoscience*, **9 (7)**, 549.

Barnes, E. A., J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2019: Viewing forced climate patterns through an AI lens. *Geophys. Res. Lett.*, **46 (22)**, 13 389–13 398.

Bellomo, K., L. N. Murphy, M. A. Cane, A. C. Clement, and L. M. Polvani, 2018: Historical forcings as main drivers of the Atlantic multidecadal variability in the CESM large ensemble. *Clim. Dyn.*, **50 (9-10)**, 3687–3698.

Bellucci, A., A. Mariotti, and S. Gualdi, 2017: The role of forcings in the twentieth-century North Atlantic multidecadal variability: The 1940–75 North Atlantic cooling case study. *J. Climate*, **30 (18)**, 7317–7337.

Bonan, D. B., J. E. Christian, and K. Christianson, 2019: Influence of North Atlantic climate variability on glacier mass balance in Norway, Sweden and Svalbard. *Journal of Glaciology*, 1–15.

Booth, B. B., N. J. Dunstone, P. R. Halloran, T. Andrews, and N. Bellouin, 2012: Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. *Nature*, **484 (7393)**, 228.

Booth, B. B., G. R. Harris, A. Jones, L. Wilcox, M. Hawcroft, and K. S. Carslaw, 2018: Comments on "Rethinking the lower bound on aerosol radiative forcing". *J. Climate*, **31 (22)**, 9407–9412.

Bretherton, C. S., C. Smith, and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Climate*, **5 (6)**, 541–560.

Broccoli, A. J., K. A. Dahl, and R. J. Stouffer, 2006: Response of the ITCZ to Northern Hemisphere cooling. *Geophys. Res. Lett.*, **33 (1)**.

Cane, M. A., A. C. Clement, A. Kaplan, Y. Kushnir, D. Pozdnyakov, R. Seager, S. E. Zebiak, and R. Murtugudde, 1997: Twentieth-century sea surface temperature trends. *Science*, **275 (5302)**, 957–960.

Cane, M. A., A. C. Clement, L. N. Murphy, and K. Bellomo, 2017: Low-pass filtering, heat flux, and Atlantic multidecadal variability. *J. Climate*, **30 (18)**, 7529–7553.

Chen, X., J. M. Wallace, and K.-K. Tung, 2017: Pairwise-rotated EOFs of global SST. *J. Climate*, **30 (14)**, 5473–5489.

Christian, J. E., N. Siler, M. Koutnik, and G. Roe, 2016: Identifying dynamically induced variability in glacier mass-balance records. *J. Climate*, **29 (24)**, 8915–8929.

Coats, S., and K. Karnauskas, 2017: Are simulated and observed twentieth century tropical Pacific sea surface temperature trends significant relative to internal variability? *Geophys. Res. Lett.*, **44 (19)**, 9928–9937.

Coats, S., J. E. Smerdon, B. I. Cook, and R. Seager, 2015: Are simulated megadroughts in the North American Southwest forced? *J. Climate*, **28 (1)**, 124–142.

Cowtan, K., and R. G. Way, 2014: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Q. J. Meteorol. Soc.*, **140 (683)**, 1935–1944.

DelSole, T., M. K. Tippett, and J. Shukla, 2011: A significant component of unforced multidecadal variability in the recent acceleration of global warming. *J. Climate*, **24 (3)**, 909–926.

Déqué, M., 1988: 10-day predictability of the Northern Hemisphere winter 500-mb height by the ECMWF operational model. *Tellus A*, **40 (1)**, 26–36.

Deser, C., and M. L. Blackmon, 1993: Surface climate variations over the North Atlantic Ocean during winter: 1900–1989. *J. Climate*, **6 (9)**, 1743–1753.

Deser, C., J. W. Hurrell, and A. S. Phillips, 2017: The role of the North Atlantic Oscillation in European climate projections. *Clim. Dyn.*, **49 (9-10)**, 3141–3157.

Deser, C., R. Knutti, S. Solomon, and A. S. Phillips, 2012a: Communication of the role of natural variability in future North American climate. *Nat. Clim. Change*, **2 (11)**, 775–779.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012b: Uncertainty in climate change projections: the role of internal variability. *Clim. Dyn.*, **38 (3-4)**, 527–546.

Deser, C., 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 (6)**, 2271–2296.

Deser, C., 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 (6)**, 2237–2258.

Deser, C., and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, 1–10.

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 (15)**, 5620–5628.

Eisenman, I., T. Schneider, D. S. Battisti, and C. M. Bitz, 2011: Consistent changes in the sea ice seasonal cycle in response to global warming. *J. Climate*, **24 (20)**, 5325–5335.

Enfield, D. B., A. M. Mestas-Nuñez, and P. J. Trimble, 2001: The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental US. *Geophys. Res. Lett.*, **28 (10)**, 2077–2080.

England, M. H., and Coauthors, 2014: Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus. *Nat. Clim. Change*, **4 (3)**, 222.

Frankcombe, L. M., M. H. England, M. E. Mann, and B. A. Steinman, 2015: Separating internal variability from the externally forced climate response. *J. Climate*, **28 (20)**, 8184–8202.

Frankignoul, C., G. Gastineau, and Y.-O. Kwon, 2017: Estimation of the SST response to anthropogenic and external forcing and its impact on the Atlantic multidecadal oscillation and the pacific decadal oscillation. *J. Climate*, **30 (24)**, 9871–9895.

Greatbatch, R. J., G. Gollan, T. Jung, and T. Kunz, 2012: Factors influencing Northern Hemisphere winter mean atmospheric circulation anomalies during the period 1960/61 to 2001/02. *Q. J. Meteorol. Soc.*, **138 (669)**, 1970–1982.

Guo, R., C. Deser, L. Terray, and F. Lehner, 2019: Human influence on winter precipitation trends (1921–2015) over North America and Eurasia revealed by dynamical adjustment. *Geophys. Res. Lett.*, **46 (6)**, 3426–3434.

Harzallah, A., and R. Sadourny, 1995: Internal versus SST-forced atmospheric variability as simulated by an atmospheric general circulation model. *J. Climate*, **8 (3)**, 474–495.

Hasselmann, K., 1993: Optimal fingerprints for the detection of time-dependent climate change. *J. Climate*, **6 (10)**, 1957–1971.

Haustein, K., and Coauthors, 2019: A limited role for unforced internal variability in twentieth-century warming. *J. Climate*, **32 (16)**, 4893–4917.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, **90 (8)**, 1095–1108.

Held, I. M., and B. J. Soden, 2006: Robust responses of the hydrological cycle to global warming. *J. Climate*, **19 (21)**, 5686–5699.

Hurrell, J. W., 1995: Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science*, **269 (5224)**, 676–679.

Jeffrey, S., L. Rotstayn, M. Collier, S. Dravitzki, C. Hamalainen, C. Moeseneder, K. Wong, and J. Syktus, 2013: Australia's CMIP5 submission using the CSIRO Mk3. 6 model. *Aust. Meteor. Oceanogr. J*, **63**, 1–13.

Kang, S. M., I. M. Held, D. M. Frierson, and M. Zhao, 2008: The response of the ITCZ to extratropical thermal forcing: Idealized slab-ocean experiments with a GCM. *J. Climate*, **21 (14)**, 3521–3532.

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. Am. Meteorol. Soc.*, **96 (8)**, 1333–1349.

Khodri, M., and Coauthors, 2017: Tropical explosive volcanic eruptions can trigger El Niño by cooling tropical Africa. *Nature Communications*, **8 (1)**, 1–13.

Kirchmeier-Young, M. C., F. W. Zwiers, and N. P. Gillett, 2017: Attribution of extreme events in arctic sea ice extent. *J. Climate*, **30 (2)**, 553–571.

Kohyama, T., D. L. Hartmann, and D. S. Battisti, 2017: La Niña–like mean-state response to global warming and potential oceanic roles. *J. Climate*, **30 (11)**, 4207–4225.

Kushner, P. J., I. M. Held, and T. L. Delworth, 2001: Southern Hemisphere atmospheric circulation response to global warming. *J. Climate*, **14 (10)**, 2238–2249.

Lehner, F., C. Deser, and L. Terray, 2017: Toward a new estimate of "time of emergence" of anthropogenic warming: Insights from dynamical adjustment and a large initial-condition model ensemble. *J. Climate*, **30 (19)**, 7739–7756.

Lorenz, E., 1975: Climatic predictability. *The physical basis of climate and climate modelling*, B. Bolin et al., Eds., GARP Publication Series, Vol. 16, World Meteorological Organization, 132–136.

Maher, N., S. McGregor, M. H. England, and A. S. Gupta, 2015: Effects of volcanism on tropical variability. *Geophys. Res. Lett.*, **42 (14)**, 6024–6033.

Maher, N., and Coauthors, 2019: The Max Planck Institute grand ensemble: Enabling the exploration of climate system variability. *J. Adv. Model. Earth Sy.*

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, 1997: A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Am. Meteorol. Soc.*, **78 (6)**, 1069–1079.

McKinnon, K. A., A. Poppick, E. Dunn-Sigouin, and C. Deser, 2017: An "observational large ensemble" to compare observed and modeled temperature trend uncertainty due to internal variability. *J. Climate*, **30 (19)**, 7585–7598.

McPhaden, M., T. Lee, and D. McClurg, 2011: El Niño and its relationship to changing background conditions in the tropical Pacific Ocean. *Geophys. Res. Lett.*, **38 (15)**.

Meehl, G. A., and Coauthors, 2009: Decadal prediction: can it be skillful? *Bull. Am. Meteorol. Soc.*

Merrifield, A., F. Lehner, S.-P. Xie, and C. Deser, 2017: Removing circulation effects to assess central US land-atmosphere interactions in the CESM large ensemble. *Geophys. Res. Lett.*, **44 (19)**, 9938–9946.

Newman, M., and Coauthors, 2016: The Pacific Decadal Oscillation, revisited. *J. Climate*, **29 (12)**, 4399–4427.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Monthly weather review*, **110 (7)**, 699–706.

Palmer, T. N., 1999: A nonlinear dynamical perspective on climate prediction. *J. Climate*, **12 (2)**, 575–591.

Pausata, F., D. Zanchetti, C. Karamperidou, R. Caballero, and D. Battisti, 2020: ITCZ shift and extratropical teleconnections drive ENSO response to volcanic eruptions. *Science Express, in press*.

Rodgers, K. B., J. Lin, and T. L. Frölicher, 2015: Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an earth system model. *Biogeosciences*, **12 (11)**, 3301–3320.

Saffioti, C., E. M. Fischer, S. C. Scherrer, and R. Knutti, 2016: Reconciling observed and modeled temperature and precipitation trends over Europe by adjusting for circulation variability. *Geophys. Res. Lett.*, **43 (15)**, 8189–8198.

Santer, B. D., W. Brüggemann, U. Cubasch, K. Hasselmann, H. Höck, E. Maier-Reimer, and U. Mikolajewica, 1994: Signal-to-noise analysis of time-dependent greenhouse warming experiments. *Clim. Dyn.*, **9 (6)**, 267–285.

Santer, B. D., K. E. Taylor, T. M. Wigley, J. E. Penner, P. D. Jones, and U. Cubasch, 1995: Towards the detection and attribution of an anthropogenic effect on climate. *Climate Dynamics*, **12 (2)**, 77–100.

Scaife, A., and Coauthors, 2014: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41 (7)**, 2514–2519.

Scaife, A. A., and D. Smith, 2018: A signal-to-noise paradox in climate science. *npj Climate and Atmospheric Science*, **1 (1)**, 1–8.

Schneider, T., 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, **14 (5)**, 853–871.

Schneider, T., and S. M. Griffies, 1999: A conceptual framework for predictability studies. *J. Climate*, **12 (10)**, 3133–3155.

Schneider, T., and I. M. Held, 2001: Discriminants of twentieth-century changes in earth surface temperatures. *J. Climate*, **14 (3)**, 249–254.

Seager, R., M. Cane, N. Henderson, D.-E. Lee, R. Abernathey, and H. Zhang, 2019: Strengthening tropical Pacific zonal sea surface temperature gradient consistent with rising greenhouse gases. *Nat. Clim. Change*, **9 (7)**, 517.

Seager, R., N. Naik, and G. A. Vecchi, 2010: Thermodynamic and dynamic mechanisms for large-scale changes in the hydrological cycle in response to global warming. *J. Climate*, **23 (17)**, 4651–4668.

Semenov, V. A., M. Latif, D. Dommenget, N. S. Keenlyside, A. Strehz, T. Martin, and W. Park, 2010: The impact of North Atlantic–Arctic multidecadal variability on Northern Hemisphere surface air temperature. *J. Climate*, **23 (21)**, 5668–5677.

Semenov, V. A., M. Latif, J. H. Jungclaus, and W. Park, 2008: Is the observed NAO variability during the instrumental record unusual? *Geophys. Res. Lett.*, **35 (11)**.

Shindell, D. T., and Coauthors, 2013: Radiative forcing in the ACCMIP historical and future climate simulations. *Atmos. Chem. Phys*, **13 (6)**, 2939–2974.

Siler, N., C. Proistosescu, and S. Po-Chedley, 2019: Natural variability has slowed the decline in western US snowpack since the 1980s. *Geophys. Res. Lett.*, **46 (1)**, 346–355.

Simpson, I. R., C. Deser, K. A. McKinnon, and E. A. Barnes, 2018: Modeled and observed multidecadal variability in the North Atlantic jet stream and its connection to sea surface temperatures. *J. Climate*, **31 (20)**, 8313–8338.

Sippel, S., N. Meinshausen, A. Merrifield, F. Lehner, A. G. Pendergrass, E. Fischer, and R. Knutti, 2019: Uncovering the forced climate response from a single ensemble member using statistical learning. *J. Climate*, **32 (17)**, 5677–5699.

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 (4)**, 1613–1629.

Solomon, A., and M. Newman, 2012: Reconciling disparate twentieth-century Indo-Pacific ocean temperature trends in the instrumental record. *Nature Climate Change*, **2 (9)**, 691.

Solomon, A., and Coauthors, 2011: Distinguishing the roles of natural and anthropogenically forced decadal climate variability: implications for prediction. *Bull. Am. Meteorol. Soc.*, **92 (2)**, 141–156.

Stern, W., and K. Miyakoda, 1995: Feasibility of seasonal forecasts inferred from multiple GCM simulations. *J. Climate*, **8 (5)**, 1071–1085.

Stolpe, M. B., I. Medhaug, and R. Knutti, 2017: Contribution of Atlantic and Pacific multidecadal variability to twentieth-century temperature changes. *J. Climate*, **30 (16)**, 6279–6295.

Stolpe, M. B., I. Medhaug, J. Sedláček, and R. Knutti, 2018: Multidecadal variability in global surface temperatures related to the Atlantic meridional overturning circulation. *J. Climate*, **31 (7)**, 2889–2906.

Sun, L., M. Alexander, and C. Deser, 2018: Evolution of the global coupled climate response to arctic sea ice loss during 1990–2090 and its contribution to climate change. *J. Climate*, **31 (19)**, 7823–7843.

Takahashi, C., and M. Watanabe, 2016: Pacific trade winds accelerated by aerosol forcing over the past two decades. *Nat. Clim. Change*, **6 (8)**, 768.

Tandon, N. F., and P. J. Kushner, 2015: Does external forcing interfere with the AMOC's influence on North Atlantic sea surface temperature? *J. Climate*, **28 (16)**, 6309–6323.

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 (16)**, 6443–6456.

Thompson, D. W., S. Solomon, P. J. Kushner, M. H. England, K. M. Grise, and D. J. Karoly, 2011: Signatures of the Antarctic ozone hole in Southern Hemisphere surface climate change. *Nature Geoscience*, **4 (11)**, 741.

Ting, M., Y. Kushnir, R. Seager, and C. Li, 2009: Forced and internal twentieth-century SST trends in the North Atlantic. *J. Climate*, **22 (6)**, 1469–1481.

Trenberth, K. E., and J. W. Hurrell, 1994: Decadal atmosphere-ocean variations in the Pacific. *Clim. Dyn.*, **9 (6)**, 303–319.

Trenberth, K. E., and D. J. Shea, 2006: Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.*, **33 (12)**.

Ulbrich, U., and M. Christoph, 1999: A shift of the NAO and increasing storm track activity over Europe due to anthropogenic greenhouse gas forcing. *Clim. Dyn.*, **15 (7)**, 551–559.

Venzke, S., M. R. Allen, R. T. Sutton, and D. P. Rowell, 1999: The atmospheric response over the North Atlantic to decadal changes in sea surface temperature. *J. Climate*, **12 (8)**, 2562–2584.

Wallace, J. M., Q. Fu, B. V. Smoliak, P. Lin, and C. M. Johanson, 2012: Simulated versus observed patterns of warming over the extratropical northern hemisphere continents during the cold season. *Proceedings of the National Academy of Sciences*, **109 (36)**, 14 337–14 342.

Watanabe, M., and H. Tatebe, 2019: Reconciling roles of sulphate aerosol forcing and internal variability in Atlantic multidecadal climate changes. *Clim. Dyn.*, 1–15.

Wills, R. C., D. S. Battisti, D. L. Hartmann, and T. Schneider, 2017: Extracting modes of variability and change from climate model ensembles. *Proceedings of the 7th International Workshop on Climate Informatics: CI 2017*, V. Lyubchich, N. C. Oza, A. Rhines, and S. E, Eds., NCAR Technical Note NCAR/TN-536+PROC, 25–28.

Wills, R. C., T. Schneider, J. M. Wallace, D. S. Battisti, and D. L. Hartmann, 2018: Disentangling global warming, multidecadal variability, and El Niño in Pacific temperatures. *Geophys. Res. Lett.*, **45**, 2487–2496.

Wills, R. C. J., K. C. Armour, D. S. Battisti, and D. L. Hartmann, 2019a: Ocean-atmosphere dynamic coupling fundamental to the Atlantic Multidecadal Oscillation. *J. Climate*, **32 (1)**, 251–272.

Wills, R. C. J., D. S. Battisti, C. Proistosescu, L. Thompson, D. L. Hartmann, and K. C. Armour, 2019b: Ocean circulation signatures of North Pacific decadal variability. *Geophys. Res. Lett.*, **46 (3)**, 1690–1701.

Yin, J. H., 2005: A consistent poleward shift of the storm tracks in simulations of 21st century climate. *Geophys. Res. Lett.*, **32 (18)**.

Zhang, R., R. Sutton, G. Danabasoglu, Y.-O. Kwon, R. Marsh, S. G. Yeager, D. E. Amrhein, and C. M. Little, 2019: A review of the role of the Atlantic Meridional Overturning Circulation in Atlantic multidecadal variability and associated climate impacts. *Reviews of Geophysics*.

Zhang, R., and Coauthors, 2013: Have aerosols caused the observed Atlantic multidecadal variability? *J. Atmos. Sci.*, **70 (4)**, 1135–1144.

Zhang, X., and J. E. Walsh, 2006: Toward a seasonally ice-covered arctic ocean: Scenarios from the IPCC AR4 model simulations. *J. Climate*, **19 (9)**, 1730–1747.

Zwiers, F., 1996: Interannual variability and predictability in an ensemble of AMIP climate simulations conducted with the CCC GCM2. *Clim. Dyn.*, **12 (12)**, 825–847.