UCLA
UCLA Previously Published Works

Title
The Inherent Uncertainty of Precipitation Variability, Trends, and Extremes due to Internal Variability, with Implications for Western US Water Resources

Authors
McKinnon, Karen A
Deser, Clara

Peer reviewed

The Inherent Uncertainty of Precipitation Variability, Trends, and Extremes due to Internal Variability, with Implications for Western U.S. Water Resources✐

Karen A. McKinnon[a] and Clara Deser[b]

[a] *Department of Statistics and Institute of the Environment and Sustainability, University of California, Los Angeles, Los Angeles, California*
[b] *National Center for Atmospheric Research, Boulder, Colorado*

ABSTRACT: The approximately century-long instrumental record of precipitation over land reflects a single sampling of internal variability. Thus, the spatiotemporal evolution of the observations is only one realization of "what could have occurred" given the same climate system and boundary conditions but different initial conditions. While climate models can be used to produce initial-condition large ensembles that explicitly sample different sequences of internal variability, an analogous approach is not possible for the real world. Here, we explore the use of a statistical model for monthly precipitation to generate synthetic ensembles based on a single record. When tested within the context of the NCAR Community Earth System Model version 1 Large Ensemble (CESM1-LE), we find that the synthetic ensemble can closely reproduce the spatiotemporal statistics of variability and trends in winter precipitation over the extended contiguous United States and that it is difficult to infer the climate change signal in a single record given the magnitude of the variability. We additionally create a synthetic ensemble based on the Global Precipitation Climatology Centre (GPCC) dataset, termed the GPCC-synth-LE; comparison of the GPCC-synth-LE with the CESM1-based ensembles reveals differences in the spatial structures and magnitudes of variability, highlighting the advantages of an observationally based ensemble. We finally use the GPCC-synth-LE to analyze three water resource metrics in the upper Colorado River basin: frequency of dry, wet, and whiplash years. Thirty-one-year "climatologies" in the GPCC-synth-LE can differ by over 20% in these key water resource metrics due to sampling of internal variability, and individual ensemble members in the GPCC-synth-LE can exhibit large near-monotonic trends over the course of the last century due to sampling of internal variability alone.

KEYWORDS: North America; Atmospheric circulation; Precipitation; Climate variability; Regression analysis; Uncertainty; Internal variability

1. Introduction

Precipitation is highly variable both spatially and temporally. Understanding and preparing for this variability has always been critical to human societies, which often rely on consistent water supplies throughout and between years. Given the importance of precipitation and water availability, there has been substantial focus on understanding and predicting changes in precipitation in response to anthropogenic radiative forcing (e.g., Held and Soden 2006; Trenberth 2011; Durack et al. 2012; Sarojini et al. 2016; Kooperman et al. 2018). However, the trajectory of the past and future climate system is a function of not only anthropogenic radiative forcing, but also a random sampling of internal variability. The internal variability emerges from processes intrinsic to the coupled climate system and is not generally predictable after any memory of initial conditions is lost. Because of the highly variable nature of precipitation, the ratio of the externally

forced trend in regional precipitation to internal variability is small over the observational record (e.g., McKinnon and Deser 2018), and the "time of emergence"—when the forced signal exceeds the noise—is not likely to occur for multiple decades in many regions (Giorgi and Bi 2009; Mahlstein et al. 2012). Thus, in addition to understanding the forced trend, it is equally critical to properly quantify and model the internal variability of precipitation, which can itself lead to unforced multidecadal trends.

A dominant source of internal variability is the random fluctuations of the atmospheric circulation. Recent work (Deser et al. 2018) demonstrated how different sampling of this unpredictable component of circulation variability, using either climate model ensembles or statistical resampling of the observations, led to different inferences about the influence of El Niño–Southern Oscillation (ENSO) on North American precipitation. Indeed, most of the variability in winter U.S. West Coast precipitation has been attributed to internal atmospheric variability rather than tropical or extratropical sea surface temperature (SST) forcing (Dong et al. 2018; Zhang et al. 2021).

On top of this atmospherically generated noise, which has a large amplitude but minimal year-to-year memory, precipitation can exhibit lower-frequency variations due to ocean influences. These oceanic sources of internal variability, and the atmospheric teleconnections that they induce, are often associated with well-known modes of interannual-to-multidecadal climate fluctuations, particularly ENSO (Ropelewski and Halpert 1987),

the Pacific decadal oscillation (PDO; Deser et al. 2004; Wang et al. 2014; Zhang and Delworth 2015; Newman et al. 2016), and the Atlantic multidecadal variability (AMV; Martin et al. 2014; L'Heureux et al. 2015; Sun et al. 2015; Ruprich-Robert et al. 2018; Simpson et al. 2019).

Cognizant of the important role of internal variability in the climate system—including both atmospheric noise and lower-frequency SST-driven modulations—climate modelers have advanced the concept of "initial-condition large ensembles" for assessing climate variability and change. Such large ensembles consist of multiple simulations (typically 30–100) with a single fully coupled climate model under a particular radiative forcing scenario, but with perturbations to the initial conditions. The resulting ensemble spread can be used to characterize the uncertainty in any given climate parameter at any point in time due to unpredictable sampling of internal variability (e.g., Deser et al. 2012; Kay et al. 2015; Maher et al. 2019; Tél et al. 2020). These large ensembles have been used to demonstrate the irreducible uncertainty in climate trends (see Deser et al. 2020 for a recent review) and can be powerful tools for decision-making (Mankin et al. 2020) but do suffer from biases in their simulation of variability and the forced response (McKinnon and Deser 2018; Suarez-Gutierrez et al. 2020; von Trentini et al. 2020).

A complementary approach to the use of numerical climate simulations is the formulation of statistical models based on the observational record. A common tool used for the purposes of water engineering and planning is the stochastic weather generator (e.g., Wilks and Wilby 1999), which can be used to simulate time series of a quantity such as precipitation that statistically resembles the observed values at a given location. While these generators have historically focused on high-frequency (daily or subdaily) variability at a single location, they are increasingly designed to better capture decadal variability (Chen et al. 2010) and to incorporate the spatial correlation structure of precipitation variability at multiple sites within a single watershed or other entities of comparable size (Kleiber et al. 2012; Steinschneider and Brown 2013; Verdin et al. 2015; Chen et al. 2018). Nevertheless, these approaches have typically focused on the spatial and temporal scales of weather, rather than climate.

Here, we combine the philosophies of climate model large ensembles and stochastic weather generators into a methodology that allows for the production of synthetic ensembles from the information contained in a single climate record, such as the observations. Our synthetic ensembles preserve both the spatial and temporal correlation structure of precipitation variability on seasonal to multidecadal time scales, in contrast to methods that focus only on the temporal variability at each location separately (e.g., Thompson et al. 2015; Castruccio et al. 2019), and so cannot be used to explore variability in large-scale precipitation patterns.

The current work advances the statistical methodology of McKinnon et al. (2017) and McKinnon and Deser (2018) and contains novel results based on the synthetic ensembles. The updated statistical methodology focuses on monthly average precipitation, improves the modeling of the coupled ocean–atmosphere modes including retaining the seasonal cycle of

ENSO amplitudes, and contains an automated, rule-based method to choose the block size used in the statistical resampling process. Further, while our prior work focused solely on 50-yr trends as a metric for variability, in this work we validate our synthetic ensembles using multiple metrics for internal variability and use the validation process to explore the more general topic of the challenges of inferring climate statistics from limited data records. We additionally present an analysis of precipitation and water resource metrics for the upper Colorado River basin, a major source of water for the western United States.

The remainder of the paper proceeds as follows. Section 2 provides an overview of the datasets and statistical model used to create the synthetic climate ensembles. Section 3 examines our ability to estimate certain statistical properties of precipitation from a limited record, provides validation for our synthetic ensemble methodology, and compares the variability in the observations to that in a climate model. Section 4 applies the synthetic ensemble methodology to precipitation in the upper Colorado River basin, with a focus on three key water resource metrics. We discuss the implications of our findings and conclude in section 5.

2. Data sources and statistical model

We first provide details of the climate model output and observational data used in our analysis, and then describe the statistical model used to create the synthetic ensembles.

a. Climate model output and observational data

We use monthly output of precipitation and SST at a nominal spatial resolution of 1° in latitude and longitude from the NCAR Community Earth System Model version 1 Large Ensemble (CESM1-LE). The CESM1-LE is composed of 40 simulations of CESM1 that that branch off a parent simulation on 1 January 1920; the spread across the ensemble was introduced by adding round-off level perturbations (of order 10^{-14} K) to the initial atmospheric temperatures (Kay et al. 2015). The CESM1-LE was forced by the historical forcing scenario from 1920 to 2005 (Lamarque et al. 2010) and the RCP8.5 scenario for 2006–2100 (Meinshausen et al. 2011). We limit our analysis of CESM1-LE to the 1921–2005 period, where we exclude the first year of the simulations (1920) to reduce any influence of land and atmosphere initial condition memory, and do not extend beyond the historical period because of different variability in the forcing related to a lack of episodic volcanic eruptions in the RCP scenarios.

Precipitation observations at 1° resolution are from the Global Precipitation Climatology Centre (GPCC; Schneider et al. 2008), which provides a gridded land-only record of precipitation from 1891 to the near present based on in situ measurements. SST observations used to calculate the time series of ENSO, PDO, and AMV for the observations are from HadISST (Rayner et al. 2003).

Precipitation for both CESM1-LE and GPCC includes solid and liquid forms and is normalized to the daily accumulation rate in millimeters (mm) in all of our analyses. For both the CESM1-LE output and the observations, we use the Climate

Variability Diagnostics Package (Phillips et al. 2014) to calculate the time series of the modes.

We transform precipitation before estimating the parameters of our statistical model using a Box–Cox power transform (Box and Cox 1964), which both guarantees that the modeled precipitation amounts remain nonnegative and increases the symmetry and normality of the precipitation distribution. The increased symmetry improves the estimation of the model parameters by reducing the influence of outliers. The parameter λ that controls the specific form of the transform is selected via maximization of the log likelihood function for the Gaussian distribution for each grid box and month independently and has a typical value of 0.34 in CESM1-LE and 0.37 in GPCC. To allow for intercomparison of model parameters between members of the CESM1-LE, λ is estimated using the first member and then used for all members. The few monthly precipitation values that are exactly zero are increased to a very small positive number so that the transform is valid everywhere.

b. Statistical model

We first propose to describe monthly average precipitation over land as a function of four different categories of terms: 1) the mean state, including the annual cycle; 2) the response to external forcing (forced component); 3) the response to large-scale, coupled ocean–atmosphere modes as summarized by the ENSO, PDO, and AMV indices; and 4) the residual "climate noise":

$$\mathcal{P}^{i,t} = \beta_0^{i,m(t)} + \beta_F^{i,m(t)} F^t + \beta_{\mathrm{ENSO}}^{i,m(t)} \mathrm{ENSO}^t + \beta_{\mathrm{PDO}_\perp}^{i,m(t)} \mathrm{PDO}_\perp^t$$
$$+ \beta_{\mathrm{AMV}}^{i,m(t)} \mathrm{AMV}^t + \epsilon^{i,t} \qquad (1)$$

In our notation, t is time, $m(t)$ is the month, and i is an index of location (grid box). The β coefficients on the right-hand side (except β_0, the mean) describe the monthly varying spatial pattern of sensitivity of $\mathcal{P}$, the transformed monthly average precipitation, to anthropogenic forcing, ENSO, PDO, and AMV, respectively.

Following Dai et al. (2015), we initially define the time series used for the estimation of the forced component, F^t, as the global mean, ensemble mean time series of near-surface air temperature from the CESM1-LE. By projecting $\mathcal{P}$ onto a time series indicative of the evolution of the forced response, the amount of unforced variability aliased onto the forced signal can be reduced compared to assuming a linear trend, although not eliminated. Nevertheless, as will be discussed in section 3a, the forced signal in precipitation at a regional level is so small compared to the noise that we find we cannot estimate $\beta_F^{i,m(t)}$ sufficiently well using a single record, so we remove the forced term from our model, and instead use the simpler model,

$$\mathcal{P}^{i,t} = \beta_0^{i,m(t)} + \beta_{\mathrm{ENSO}}^{i,m(t)} \mathrm{ENSO}^t + \beta_{\mathrm{PDO}_\perp}^{i,m(t)} \mathrm{PDO}_\perp^t + \beta_{\mathrm{AMV}}^{i,m(t)} \mathrm{AMV}^t + \epsilon^{i,t}$$
$$(2)$$

The time series ENSO^t is the standard Niño-3.4 SST index. The PDO time series is calculated as the principal component associated with the leading empirical orthogonal function of SST

anomalies in the North Pacific, poleward of 20°N, where the anomalies are calculated through removing both the climatological annual cycle and global-average SST from each grid box. The time series of ENSO and the PDO are highly correlated, and components of the PDO have been suggested to be midlatitude responses to ENSO (Newman et al. 2016). To more clearly parse the two modes in a statistical sense, we create a version of the PDO time series, $\mathrm{PDO}_\perp^t$ that is orthogonal to the standard Niño-3.4 time series by removing the projection of the PDO time series onto the Niño-3.4 time series. Like the orthogonalized versions of the PDO proposed by Chen and Wallace (2016) and Wills et al. (2018), the SST anomaly pattern associated with $\mathrm{PDO}_\perp^t$ is dominated by the classical PDO pattern in the midlatitudes with a negative region of SST anomalies extending from Japan into the central Pacific and a positive region off the west coast of North America, with a much weaker tropical component (supplementary Fig. 1 in the online supplemental material). The difference between the SST anomaly pattern of the traditional PDO and that associated with $\mathrm{PDO}_\perp^t$ strongly resembles an El Niño pattern and has a pattern correlation with the SST anomaly pattern associated with ENSO^t of 0.95 (not shown).

The AMV time series, AMV^t, is calculated as the average of SST anomalies in the North Atlantic (0°–60°N, 80°W–0°) minus the near-global mean SST (60°S–60°N) (Trenberth and Shea 2006). The time series is smoothed with a low-pass Butterworth filter using a forward and backward digital filter and a cutoff frequency of $1/20\,\mathrm{yr}^{-1}$ in order to isolate the component of the AMV that is primarily ocean driven (Delworth et al. 2017) and has been shown to have downstream impacts on precipitation (Simpson et al. 2018). The forward–backward Butterworth filter is maximally flat in the passband, does not introduce phase shifts, and performs comparably to other low-pass filters (e.g., Roberts and Roberts 1978). The AMV^t time series is not significantly correlated at the 0.05 level with ENSO^t or $\mathrm{PDO}_\perp^t$ at zero lag. At nonzero lags, there are greater correlations between AMV^t and $\mathrm{PDO}_\perp^t$, peaking as high as 0.57 when leads AMV^t by 21 years. However, given the very few degrees of freedom in the AMV^t time series, it is difficult to assess whether the relationships are statistically significant and meaningful using the data alone. All three time series (ENSO^t, $\mathrm{PDO}_\perp^t$, and AMV^t) are normalized to have unit standard deviation.

The model is fit using ordinary least squares regression separately for each month due to the known seasonal dependence of the atmospheric teleconnections associated with the modes. The initial estimates of each model parameter are then projected onto the first two harmonics of the annual cycle to enforce the expected smoothness in seasonality. The residual $\epsilon^{i,t}$ is the "climate noise" that primarily arises from internal atmospheric dynamics but could also reflect influences from other modes of variability that are uncorrelated with those explicitly considered here. In general, the interannual variance in the residual is greater than that linked to the modes: the fraction of variance in the residual exceeds 0.9 for December–February (DJF) in many places across the domain, and still generally exceeds 0.7 even in the regions with strong teleconnections (Fig. 1). As a result, there is considerable scatter
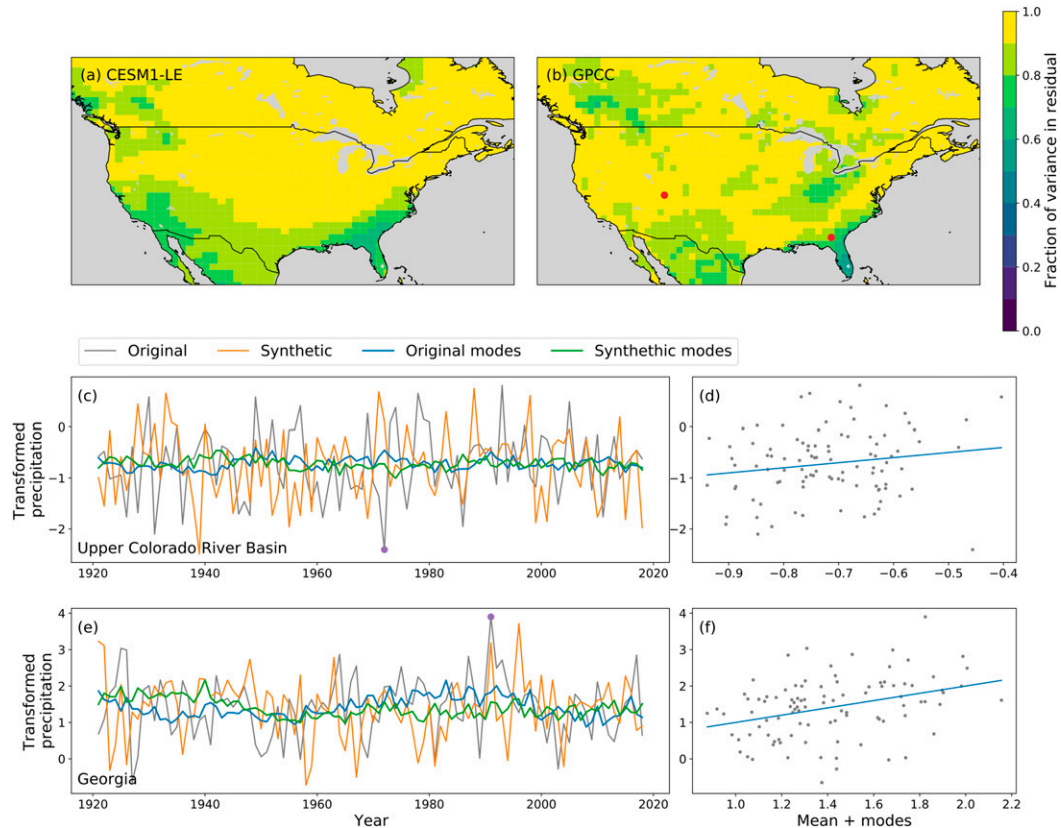
FIG. 1. The relative importance of the residual vs the modes in precipitation variability. (a) The ratio of the interannual variance in DJF precipitation in the residual term to that in the residual and modes for CESM1-LE. Variances are calculated for each member and averaged. (b) As in (a), but for GPCC. (c) The time series of transformed January precipitation (gray) from GPCC at a grid box in the upper Colorado River basin (38.5°N, 109.5°W), the estimated contribution of the modes (blue), a single synthetic time series (orange), and the contribution to that time series from the modes (green). (d) The predicted (transformed) precipitation from the mean value and modes (horizontal axis, blue line) vs the actual transformed precipitation (vertical axis). The scatter around the line is the residual ϵ. (e),(f) As in (c) and (d), but for a grid box in Georgia (31.5°N, 83.5°W) where the relative contribution of the modes is greater. The spatial pattern of the precipitation anomalies in years indicated by purple dots (1972 and 1991) in (c) and (e) are shown in Figs. 4a and 4c, respectively.

around a regression on the mode time series; however, there is not clear evidence of nonlinearity (e.g., Figs. 1d,f).

c. Generating the synthetic ensemble

To move from Eq. (2), which simply describes the dependence of precipitation on our chosen covariates, to the creation of an ensemble, we must appropriately introduce stochasticity. In the context of Eq. (2), we view all β terms as fixed and reflective of the physics of the climate system. In contrast, we view the time series of each mode (ENSO^t, PDO$^t_\perp$, and AMV^t) as well as the residual ($\epsilon^{i,t}$) as stochastic.

To produce alternative versions of the time series of ENSO^t, PDO$^t_\perp$, and AMV^t, we employ the Iterative Amplitude Adjusted Fourier Transform (IAAFT) method (Schreiber and Schmitz 1996), which produces synthetic time series that have the same amplitude distributions and power spectra as the originals. We additionally modify the algorithm to retain the seasonal cycle of ENSO amplitudes. An unavoidable constraint of the IAAFT

approach is the reliance on the empirical amplitude distribution and power spectra, which themselves are limited by what we have observed. For example, a synthetic ENSO time series would never contain an El Niño event that is larger than the largest observed event, even if such an event could theoretically be produced by the climate system. The resulting synthetic time series of the three modes, by design, do not exhibit coherence with each other, which is consistent with the lack of significant synchronous coherence of the modes in the observational record (supplementary Fig. 2).

To produce alternative realizations of $\epsilon^{i,t}$, we perform a moving block bootstrap in space and time, where the spatial block size is the full domain and the temporal block size is constrained to be an integer number of years to retain any seasonality in the variability. To select the (temporal) block size, we first use the methods of Wilks (1997) [see their Eq. (19)], which were developed such that a moving block bootstrap-based significance test on autocorrelated data would

reject the null hypothesis appropriately, to identify a block size for each grid box i and month m. The selected block size for each grid box is summarized in supplementary Fig. 3, which shows the largest block size selected across months averaged across the full CESM1-LE, for a single member of the CESM1-LE, for the GPCC observations, and the average difference between GPCC and individual members of the CESM1-LE. In most regions outside of North America and Europe, the GPCC-based block size is greater than that estimated using individual members of the CESM1-LE. The block size for the full spatial field is then chosen as the 97th percentile across grid boxes and months of the selected block sizes, excluding the dry subtropics (10°–30° in both hemispheres) where it is common to have no precipitation for certain months of the year, leading to estimates of very high autocorrelation. We select a high percentile rather than the maximum to avoid the undue influence of a small number of outliers. The block size identified for CESM1-LE is 2 years, and that for the observed precipitation from GPCC is 4 years. The use of a moving block bootstrap in time allows us to easily retain the complex spatial correlation structure in precipitation without having to rely on a spatial model with strong parametric assumptions, such as isotropy and spatial stationarity (e.g., Beusch et al. 2020 for temperature).

Using the subscript "synth" to indicate synthetic data, we can place our alternative versions of the mode time series and climate noise back into Eq. (2) to produce new versions of the precipitation data as follows:

$$\mathcal{P}_{synth}^{i,t} = \beta_0^{i,m} + \beta_{ENSO}^{i,m} ENSO_{synth}^t + \beta_{PDO_\perp}^{i,m} PDO_{\perp synth}^t$$
$$+ \beta_{AMV}^{i,m} AMV_{synth}^t + \epsilon_{synth}^{i,t} \qquad (3)$$

The approach has been applied across the global landmasses, although we focus our subsequent discussion on a subset of North America centered on the contiguous United States (24°–60°N, 230°–300°E) during DJF.

3. How well do we know the statistics of precipitation from a limited data record?

Analysis of climate model large ensembles has highlighted the challenge of estimating the precipitation response to human influence (Deser et al. 2014) and ENSO teleconnections (Deser et al. 2018) from the observational record or a small number of model simulations due to the influence of internal variability independent of the desired signal. In the context of the synthetic ensemble methodology, these topics surface in two ways: first, in the variability of the model parameters of Eqs. (1) and (2), and second, in any dependence of the statistics of precipitation on the specific record on which a synthetic ensemble is based.

We explore both of these questions using 85 years (1921–2005) of model output from the CESM1-LE as our testbed (see Fig. 4 in Deser et al. 2020); this record length is comparable to that of many in situ–based climate datasets. We fit Eqs. (1) and (2) to each of the 40 members of the CESM1-LE, and then use Eq. (3) to produce 40 synthetic ensemble members from each original member of the CESM1-LE, producing a total of 1600

ensemble members that make up the CESM1-synth-LE. We then compare the parameters and statistical characteristics of the synthetic ensemble those of the actual CESM1-LE, which is viewed as the truth for the purposes of validation.

Since our ultimate goal is to apply the synthetic methodology to the observations, we also fit Eq. (2) to the GPCC dataset, using the mode time series based on observed SSTs. The mode time series and residual are randomized 1000 times to produce the GPCC-synth-LE.

We do not account for observational uncertainty (e.g., sensitivity to the choice of dataset) in our mode time series, or in the precipitation dataset, in order to focus on uncertainty due to sampling internal variability alone. Similarly, we estimate a single set of regression coefficients based on the mode time series and a fixed period of record, and therefore do not propagate uncertainty in the regression coefficients through the model, although it would be straightforward to do so by performing a bootstrap in time, and repeatedly reestimating the parameter values.

a. Uncertainty in parameter estimation

Through the process of creating the CESM1-synth-LE, we have 40 estimates—one from each original member of the CESM1-LE—of each β parameter in the statistical model. The "true" values of the β parameters—uncontaminated by sampling of internal variability—should be the same in each ensemble member; we can therefore use the spread across the 40 estimates as a metric for the uncertainty in each parameter due to sampling of internal variability. Figure 2 shows the average estimate of each parameter across the CESM1-LE, the standard deviation of the parameter estimates, and the signal-to-noise ratio (SNR), defined as the absolute value of average divided by standard deviation.

We first demonstrate the challenge of inferring the forced component using Eq. (1). The SNR of β_F is less than one everywhere and is often close to zero (Fig. 2c). Even regions with seemingly large amplitudes for the forced response such as northern Mexico and the U.S. Southwest have a SNR less than one due to the large spread across the ensemble. To more completely visualize the variability in the β_F estimates, we plot the map of the parameter estimate for each member of the CESM1-LE (supplementary Fig. 4). The large-scale spatial structures in each panel differ dramatically from each other, with no clear and consistent behavior across the members. Given that the forcing is the same in each model, these differences indicate that it is not possible to accurately estimate the forced component using regression methods for precipitation; rather, the estimates of β_F are strongly contaminated by sampling of internal variability. We thus choose to omit the forced component term from our model. This omission does not mean that there is no climate change signal in precipitation, but rather that it is too small compared to the noise to be estimated in the regression model given a single record. All subsequent results do not include estimation of the forced component, that is, are based on Eq. (2). Using an alternative method, low-frequency component analysis (Wills et al. 2018), to estimate the forced response does not increase the SNR (supplementary Fig. 5).
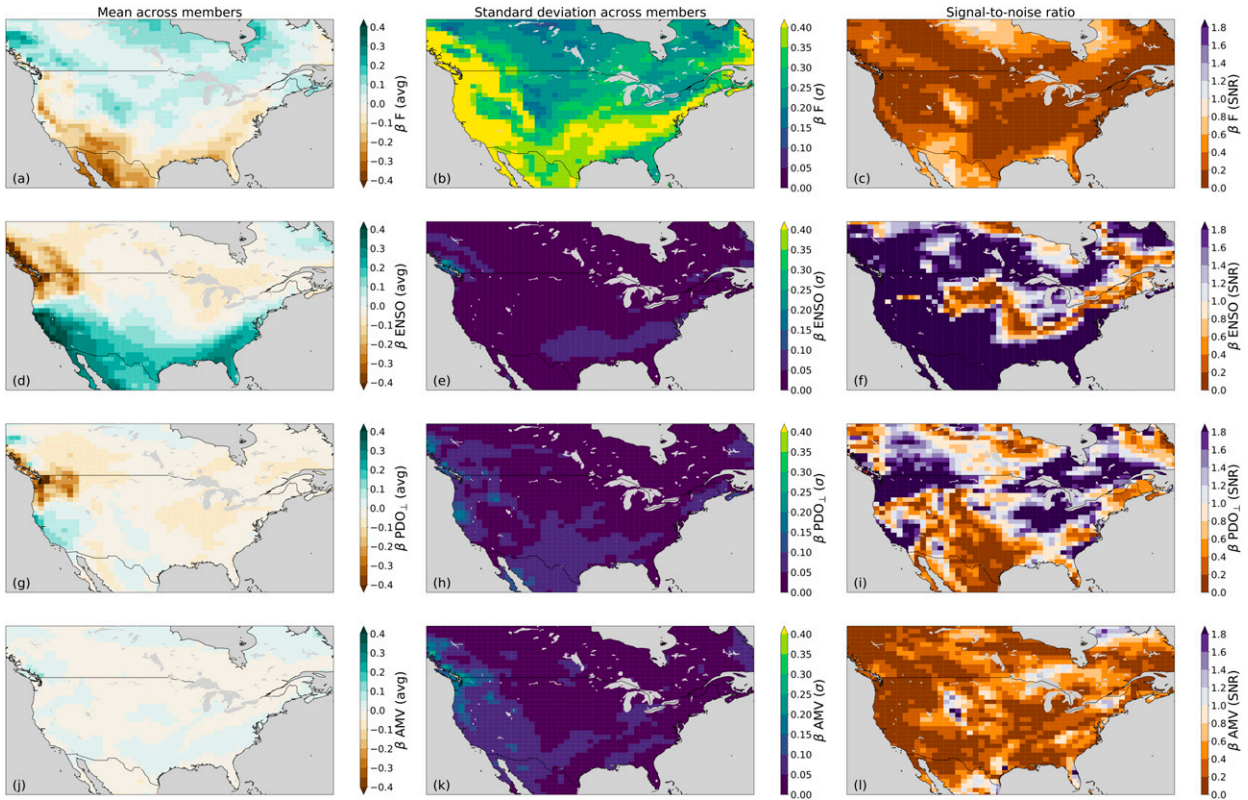
FIG. 2. The (left) mean value, (center) standard deviation, and (right) SNR (absolute value of the mean divided by standard deviation) of the DJF model parameters from Eq. (1) estimated using each member of the CESM1-LE. Parameters are (from top to bottom) $\beta_F^{i,DJF}$, $\beta_{ENSO}^{i,DJF}$, $\beta_{PDO_\perp}^{i,DJF}$, and $\beta_{AMV}^{i,DJF}$. Recall that precipitation is transformed via a Box–Cox power transformation before model fitting, so the parameters cannot be interpreted in terms of standard precipitation units. Model parameters are estimated for each month separately, and then averaged to produce the seasonal-average maps.

In contrast to the forced component, the across-ensemble spread in estimating the β coefficients for each coupled ocean–atmosphere mode is small. For ENSO, the SNR exceeds two across the western and southern parts of the domain where ENSO influences precipitation and is less than one only in regions with a weak ENSO teleconnection (Figs. 2d–f). Similarly, the SNR for the precipitation response to $PDO_\perp$ is greater than one in most regions where there is a signal, with the exception of interior California (Figs. 2g–i). The SNR for the AMV is below one in most places, which reflects the small magnitude of the signal everywhere (Figs. 2j–l). The small signal identified for the AMV, and to some extent $PDO_\perp$, at least within CESM1, is encouraging for our ability to estimate the statistics of precipitation from the observational record,

since we have limited samples of these lower-frequency modes of variability.

The high SNR for the β coefficients (apart from the forced response) suggests that we can estimate these coefficients within the single observational record. The GPCC-based β maps (Fig. 3) share many of the same features as CESM1, with a dipole of wetting/drying along the West Coast and wetting in the Southeast in response to ENSO; a similar but weaker dipole as well as drying in the Midwest in response to $PDO_\perp$, and a weak response to the AMV.

b. Spatiotemporal structures in the synthetic ensembles

Having established that the parameters of the statistical model in Eq. (2) can be well estimated using a single record,
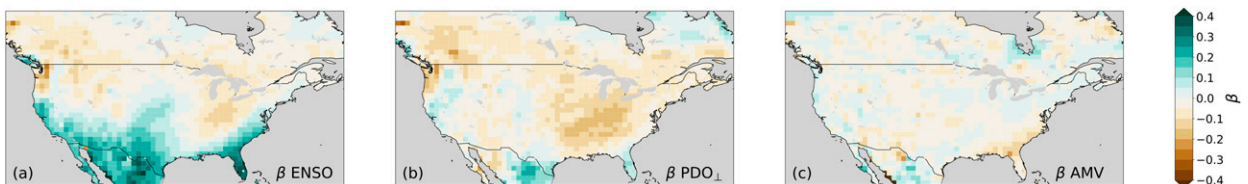

FIG. 3. The DJF-average β coefficient maps from fitting Eq. (2) to the GPCC data.
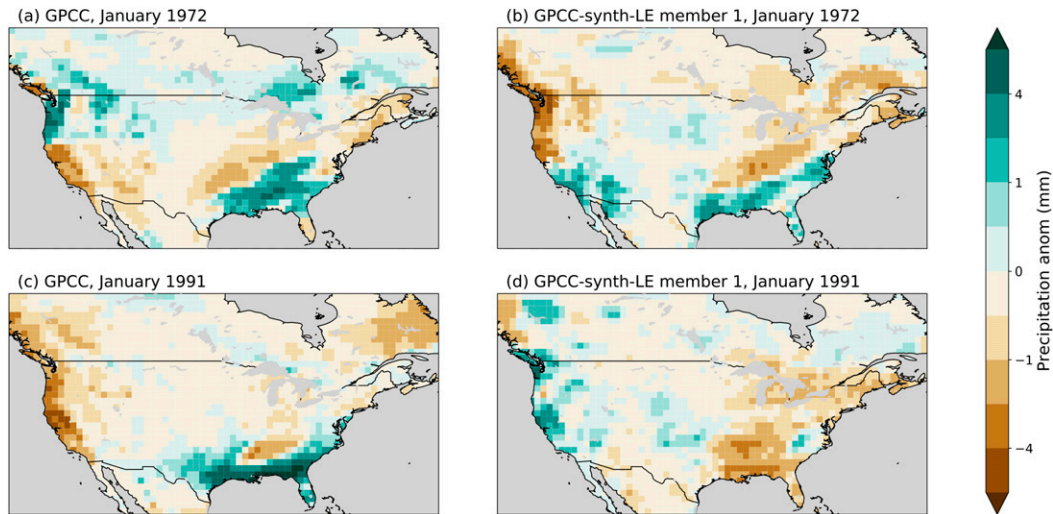
FIG. 4. Example precipitation fields from the GPCC dataset and the GPCC-synth-LE. (a) The observed precipitation anomaly in January 1972, which was relatively dry in the upper Colorado River basin. (b) The precipitation anomaly from the first member of the GPCC-synth-LE in January 1972. (c) The observed precipitation anomaly in January 1991, which was relatively wet in Georgia. (d) The precipitation anomaly from the first member of the GPCC-synth-LE in January 1991. The color bar is nonlinear.

we turn to analysis of the synthetic ensembles. To build intuition for the synthetic ensemble approach, we first show two example fields of January precipitation anomalies from the GPCC dataset and GPCC-synth-LE (Fig. 4). The Januarys were arbitrarily chosen to align with low and high January precipitation in the upper Colorado River basin and Georgia, respectively, from Figs. 1c and 1e. As can be seen in these examples, the synthetic ensemble approach produces spatially coherent precipitation fields that qualitatively reproduce the spatial correlation structures of the original data. The spatial correlation structures are generally retained in the synthetic ensemble due to the spatial structure of the teleconnection maps (Fig. 3) and the use of the block bootstrap approach, which retains the spatial correlation in the residual $\epsilon^{i,t}$.

To generalize this evaluation to the full spatiotemporal variability in DJF precipitation, we perform an empirical orthogonal function (EOF) analysis on DJF-average precipitation in the CESM1-LE, the CESM1-synth-LE, and the GPCC-synth-LE, where the EOFs are calculated in each ensemble by appending the members in time (Fig. 5). For the EOF analysis of the synthetic ensembles, we use the 40-member CESM1-synth-LE constructed from the first member of the CESM1-LE, and similarly only select the first 40 members of the GPCC-synth-LE, to match the sample size of the original CESM1-LE. We also present maps of the total variance in each ensemble for comparison.

The total variance in DJF-precipitation in the CESM1-synth-LE is nearly identical to that in the CESM1-LE (Figs. 5a,b). Further, the first four EOFs in the CESM1-LE and the CESM1-synth-LE are very similar in both their spatial patterns and variance explained, indicating both that the synthetic ensemble methodology retains the spatiotemporal structure of the original

data, and that fitting the synthetic ensemble with only a single member is sufficient to capture these structures.

Comparing the GPCC-synth-LE to the CESM1-based ensembles, the total variance in the GPCC-synth-LE is smaller almost everywhere, although the general structure of high variance along the West Coast and a secondary maximum in the Southeast is present in both GPCC and CESM1. The EOFs from the GPCC-synth-LE share many of the large-scale spatial patterns of those from the CESM1-based ensembles, although some differences are evident, including a different sign of precipitation anomalies in Texas for EOF1, a different location of the eastern negative anomalies in EOF 2, and an opposite phasing of the West Coast precipitation anomalies in EOF 3. Further, the amplitudes of the GPCC-synth-LE EOFs over the western half of the continent tend to be weaker than those of CESM1, consistent with the smaller total variance. The variance explained by the first four EOFs is greater in the CESM1-based ensembles compared to the GPCC-synth-LE (60% vs 53%), and the CESM1-based ensembles have substantially more variance in the first two EOFs than their GPCC-synth-LE counterparts (46%–47% vs 32%).

c. Characteristics of temporal variance in the synthetic ensembles

The large-scale spatiotemporal variability in precipitation appears to be well captured by the synthetic ensembles, even when the synthetic ensemble is based on a single record. We now turn to other common metrics of precipitation variability to continue to evaluate the synthetic ensemble approach and assess the limits of what we can learn from a single record. While there are innumerable metrics to summarize the statistics of precipitation, here we focus on three: high-frequency (<10 years) and low-frequency (>10 years) variability, and the
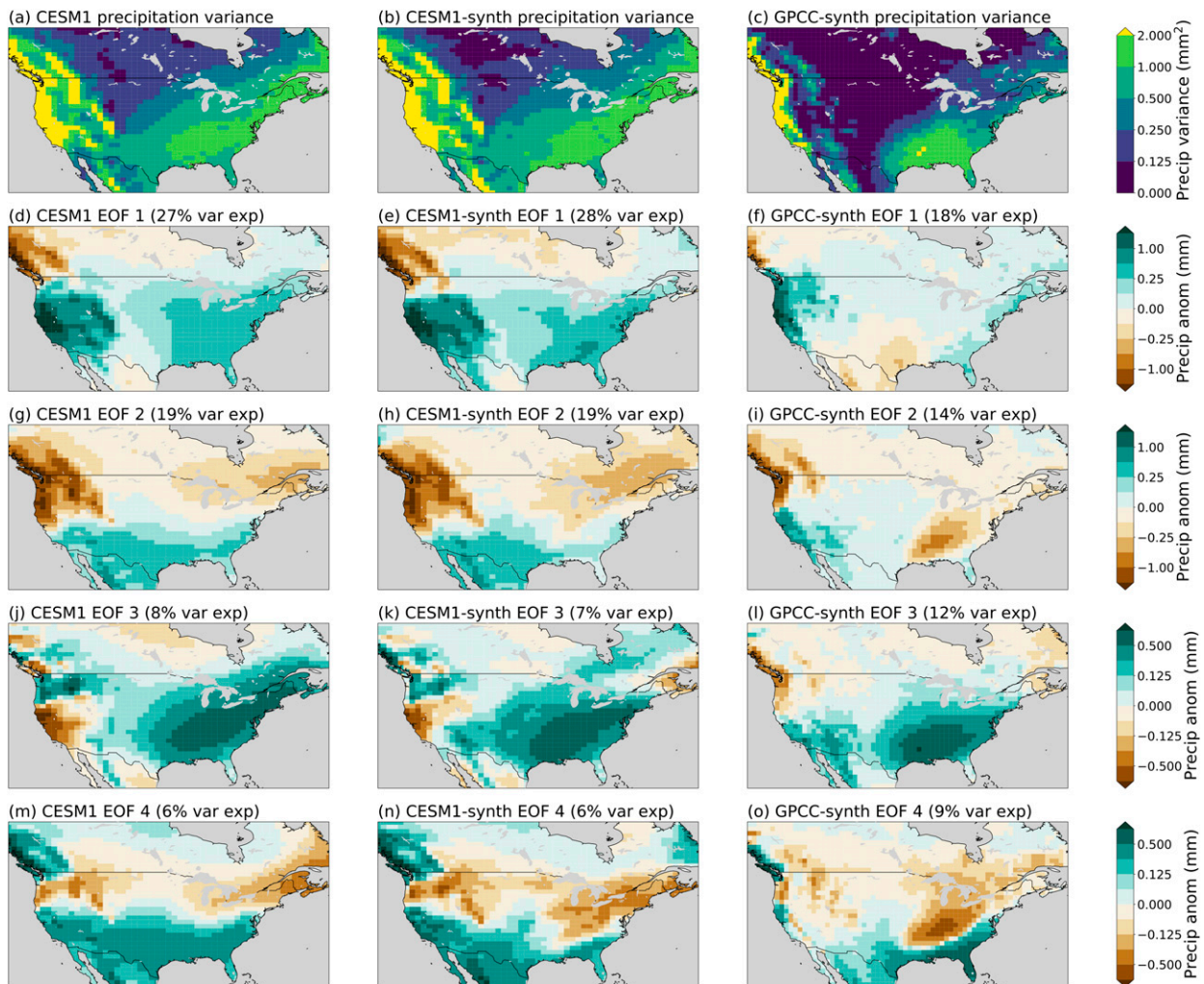
FIG. 5. The total variance and first four EOFs of DJF precipitation in each ensemble: CESM1-LE, CESM1-synth-LE, and GPCC-synth-LE. (a)–(c) The total variance across years and members for each ensemble. (d)–(o) The spatial pattern associated with the first four EOFs from each ensemble, where EOFs are taken across years and members. The ensemble name, EOF number, and variance explained by the EOF is in the associated subtitle. The CESM1-synth-LE EOFs are based on the 40 synthetic ensemble members produced from the first member of the CESM1-LE alone, and the GPCC-synth-LE EOFs are similarly based on the first 40 members of that ensemble. All color scales are nonlinear.

magnitude of extreme (low probability) events. We discuss our findings about low-frequency variability in detail below and refer the reader to the supplement for the other metrics.

Low-frequency variability is calculated as the interannual variance in DJF-average precipitation after using a low-pass forward and backward digital Butterworth filter with a frequency cutoff of $1/10 \, \mathrm{yr}^{-1}$; the interannual variance calculation excludes the first and last 5 years to avoid edge effects. Within the context of the CESM1-LE, the best estimate of the low-frequency variability is the average across the low-pass interannual variance calculated for each ensemble member; the spread across the ensemble indicates the uncertainty in this quantity given an 85-yr record. Low-frequency variability is maximized along the West Coast and has a secondary maximum spanning from the Southeast up the Eastern Seaboard (Fig. 6a). By definition, the average difference between the

estimate of high-frequency variability in any given ensemble member and the ensemble mean estimate is zero (Fig. 6b); the spread across ensemble members as measured by the standard deviation has a similar pattern but about a quarter of the magnitude of the ensemble mean estimate (Fig. 6c; note the different color bar scale compared to Fig. 6a). To assess how well the estimate of variability from a single ensemble member—analogous to having a single observational record—matches the ensemble mean estimate of variability, we show the difference between the low-frequency variability estimated using a single representative ensemble member and the ensemble mean in Fig. 6d. This single ensemble member shows regions of both over- and underestimation of the low-frequency variability as expected, with the largest differences along the West Coast, a region of high variability.

The above analysis was focused on the CESM1-LE alone, but a similar approach can be used to validate our synthetic
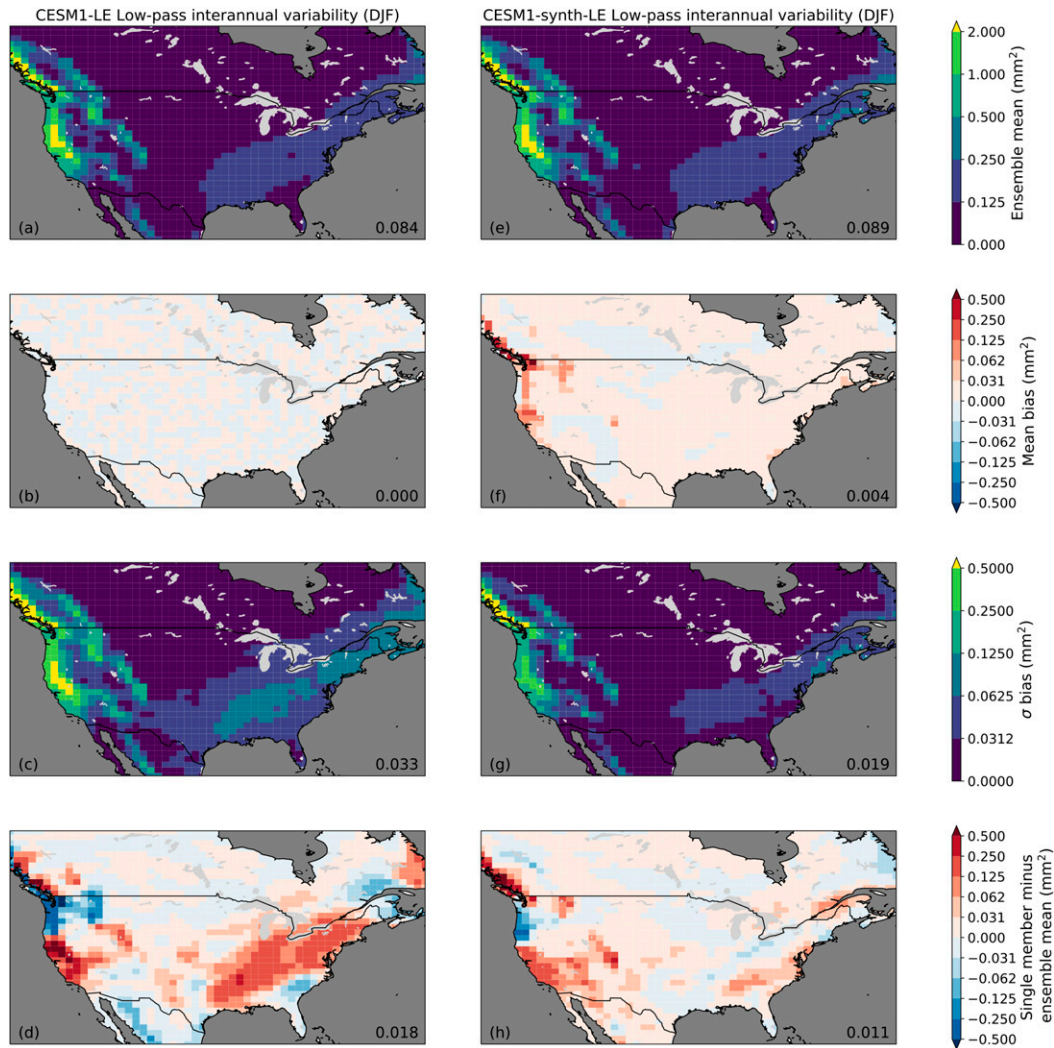
FIG. 6. Low-frequency variability (frequency $< 1/10$ yr^{-1}) of DJF precipitation in the CESM1-LE and the CESM1-synth-LE. (a) The ensemble mean of the CESM1-LE; (b) the mean bias across ensemble members, calculated as the mean difference between each ensemble member and the CESM1-LE ensemble mean; (c) the standard deviation across ensemble members of the difference between each ensemble member and the CESM1-LE ensemble mean; and (d) the difference between the ensemble member with the average bias and the ensemble mean. (e)–(h) As in (a)–(d), but for the CESM1-synth-LE; bias is still assessed against the CESM1-LE ensemble mean shown in (a). In all panels, the number in the lower-right-hand corner is the median value across the grid boxes shown. Note that the color scale in all panels is nonlinear, and that the range in the bottom three rows is one quarter of that in the top row.

ensemble methodology by asking whether the statistics in the CESM1-synth-LE are comparable to those from the original CESM1-LE. For each of the 40 members of the CESM1-synth-LE based on a single member of the CESM1-LE, we estimate the low-frequency variability as the average across the interannual low-pass variance of the 40 members, giving us 40 estimates—each one based on single member of the CESM1-LE—of the variability. We benchmark the variability in the CESM1-synth-LE against the ensemble mean of the CESM1-LE, which is viewed as the ''truth.'' Encouragingly, the CESM1-synth-LE has similar low-frequency variability to the original CESM1-LE (Fig. 6e). While it does exhibit a small

positive bias along the West Coast, the spread across the ensemble is smaller than for the CESM1-LE. Further, the difference between the low-frequency variability estimated using a single ensemble member from CESM1-LE and the true low-frequency variability is reduced by creating the synthetic ensemble (Figs. 6f–h). In other words, the synthetic ensemble produces a slightly biased simulation of low-frequency variability but reduces the variance in the estimates when given a single record.

The same conclusions—the synthetic ensemble successfully captures the spatial structures of variability with a small bias but decreased variance—hold for other metrics, including the
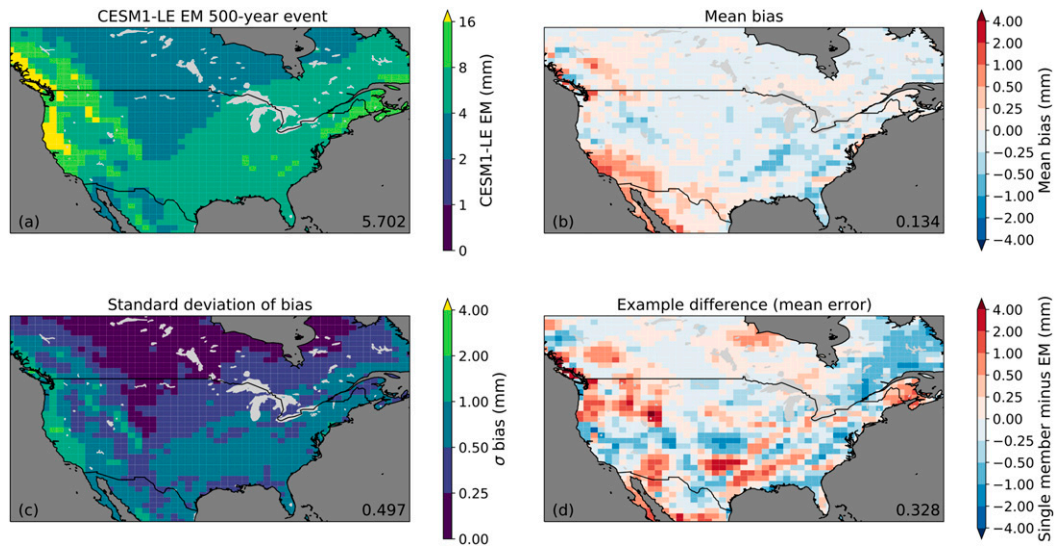
FIG. 7. The magnitude of the 500-yr event in DJF precipitation. (a) The "true" estimate from the full CESM1-LE; (b) the mean bias across ensemble members in the CESM1-synth-LE; (c) the standard deviation of the bias across ensemble members in the CESM1-synth-LE; and (d) the difference between the CESM1-synth-LE ensemble member with the average bias and the "truth" from the CESM1-LE. In all panels, the number in the lower-right-hand corner is the median value across the grid boxes shown. Note that the color scale in all panels is nonlinear, and that the range in (b)–(d) is one quarter of that in (a).

high-frequency variability and the magnitude of the 150-yr event (supplementary Figs. 6 and 7).

Since the synthetic ensemble approach allows us to simulate many "years" of data, we also estimate the magnitude of the 500-yr event across the full CESM1-LE, which contains 3400 member-years of data, and compare it to estimates from the CESM1-synth-LE (Fig. 7). While the bias and variance of the estimates from the CESM1-synth-LE are, as expected, higher than those for the other metrics, the synthetic ensemble estimate is nearly unbiased in the majority of the domain, indicating that the approach can be used to estimate the probability of very extreme events.

Having validated our estimates of the high- and low-frequency variability, and the 150- and 500-yr events, for DJF precipitation in the CESM1-synth-LE, we present each metric based on the GPCC-synth-LE (Fig. 8). In general, the GPCC-synth-LE exhibits substantially less variance and smaller extremes in most of the domain, with the exception of the Southeast, where it shows larger variance and extremes than the CESM1-based ensembles (cf. the panels in Fig. 8 to supplementary Fig. 6a, Fig. 6a, supplementary Fig. 7a, and Fig. 7a, respectively). Using the 500-yr event and the grid box containing Sacramento, California, as a specific example, the CESM1-LE simulates a value of 14.9 mm day^{-1}, and the CESM1-synth-LE simulates a 95% range of 12.5–17.9 mm day^{-1}, depending on which member of the CESM1-LE it was constructed from. In contrast, the GPCC-synth-LE suggests that the true magnitude of the event is much smaller, at 7.2 mm day^{-1}.

d. The relative contribution of the climate noise

All of the results thus far have been based on Eq. (3), in which precipitation variability is a function of the dependence

on the three coupled ocean–atmosphere modes, ENSO, PDO, and AMV, as well as the climate noise. What is the relative role of the climate noise versus the large-scale coupled ocean–atmosphere modes?

To answer this question, we produce "noise only" versions of the CESM1-synth-LE and GPCC-synth-LE, in which the residual climate noise $\epsilon^{i,t}$ is bootstrapped as before, but the mode time series are not randomized. Thus, the spread across the ensemble reflects different samples of the climate noise alone. We then calculate the ratio of high- and low-frequency variance in the noise only ensemble to that in the full ensemble, where the ensemble mean is first removed from both ensembles. This has the effect of removing the influence of the modes from the noise only ensemble, since their time series are identical in all of the synthetic ensemble members. In both ensembles, over 90% of the high-frequency variability and over 80% of the low-frequency variability is due to residual climate noise at most grid boxes (Fig. 9). While this result may seem counterintuitive given the minimal year-to-year memory in the climate noise, it results from the fact that white noise has power across all frequencies, and the contribution of the unpredictable atmospheric circulation to precipitation variance is large (see also Dong et al. 2018).

e. Validation of 50-yr trends

The prior analyses have demonstrated that the variability in the CESM1-synth-LE closely matches the original CESM1-LE. Our final validation focuses on the across-ensemble ranges of 50-yr (1956–2005) trends in DJF precipitation in each ensemble. Given that the synthetic ensembles have, by design, a forced trend of zero, how do their trends compare to the CESM1-LE, which does contain a forced component?
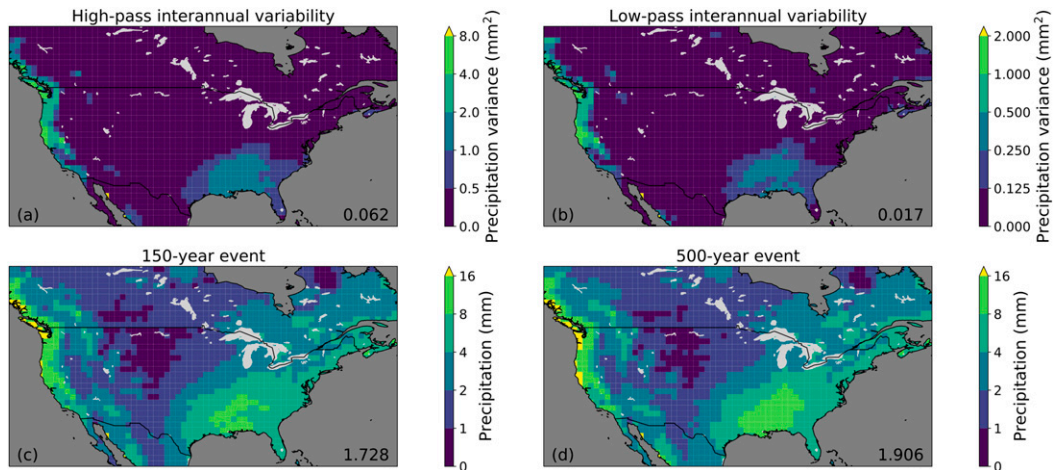
FIG. 8. Maps of (a) high-frequency variability (frequency $> 1/10\text{ yr}^{-1}$), (b) low-frequency variability (frequency $< 1/10\text{ yr}^{-1}$), (c) the magnitude of the 150-yr event, and (d) the magnitude of the 500-yr event in DJF precipitation from the GPCC-synth-LE. In all panels, the number in the lower-right-hand corner is the median value across the grid boxes shown, and the color scales are nonlinear.

That the influence of random sampling of internal variability dominates the forced climate signal for recent 50-yr trends is confirmed by comparing the 5th and 95th percentiles of 50-yr trends in DJF precipitation (calculated independently at each grid box) in the CESM1-LE (Figs. 10a,b) to those in the CESM1-synth-LE (Figs. 10c–f). At all locations, the 5th percentile of trends is negative, while the 95th percentile is positive and of roughly equal magnitude, demonstrating the small contribution of the forced component compared to the random sampling of internal variability.

Thus, not surprisingly, the 5th and 95th percentile trends in the CESM1-synth-LE are very similar in polarity, spatial structure, and magnitude, as those in the CESM1-LE. The slightly larger amplitude in the synthetic ensemble compared to the original ensemble (Figs. 10g,h) is due to the small, but nonzero, forced trend, which is deterministic, being mapped onto the random internal variability in the synthetic ensemble. The 5th and 95th percentiles of the trends in the GPCC-synth-LE tend to have smaller magnitudes (Figs. 10i,j) than those in the CESM1-synth-LE, consistent with CESM1's
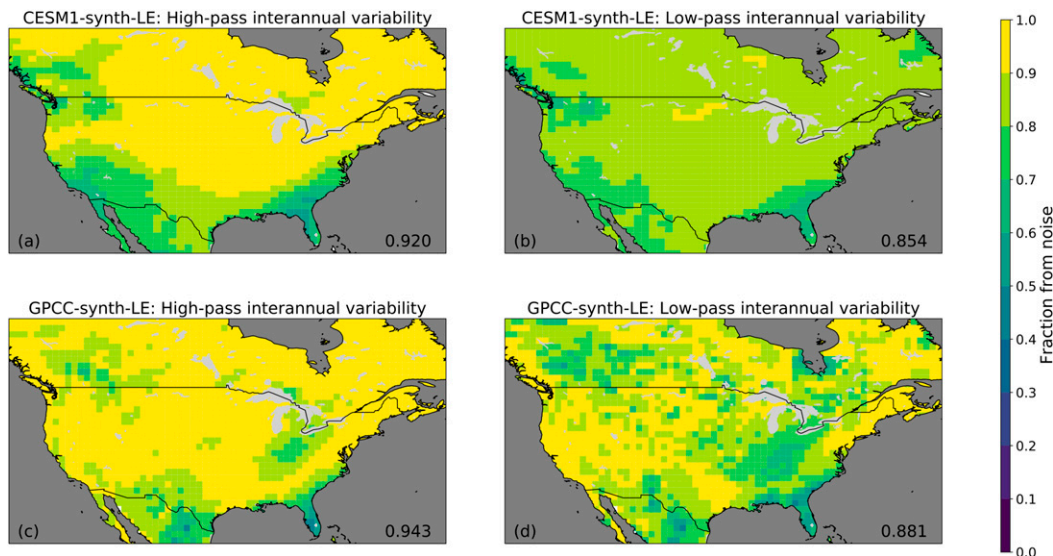

FIG. 9. The fraction of variance from the climate noise term. (a) The ratio of the high-frequency ($>1/10\text{ yr}^{-1}$) variance in the "noise only" CESM1-synth-LE to the high-frequency variance in the full CESM1-synth-LE. (b) As in (a), but for low-frequency ($<1/10\text{ yr}^{-1}$) variance. (c),(d) As in (a),(b), but for the GPCC-synth-LE. The variance is calculated after the ensemble mean of each ensemble is removed, since the "noise only" ensembles share the same temporal evolution of the modes.
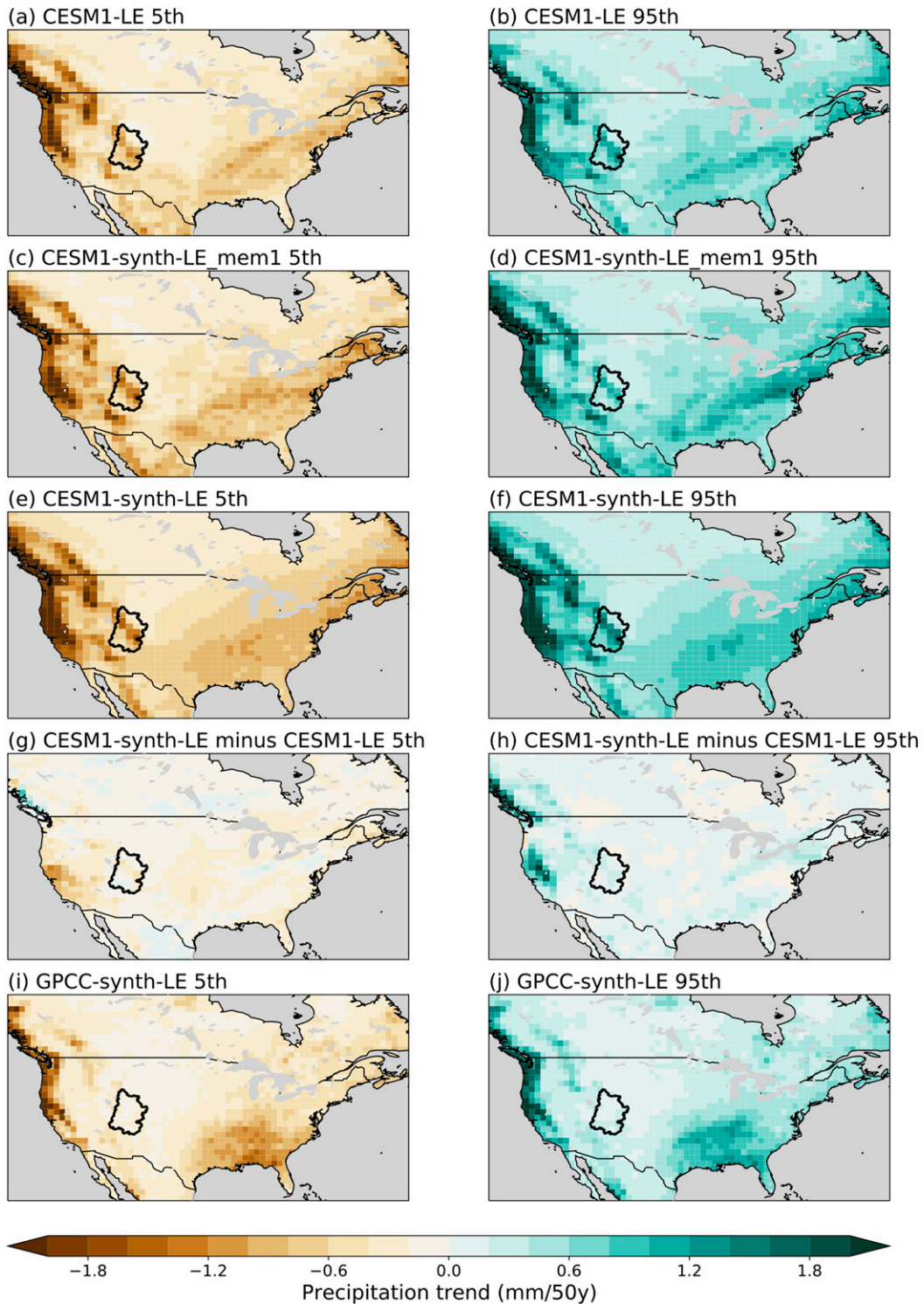
FIG. 10. Maps of the 5th and 95th percentiles (at each grid box) in 50-yr trends in DJF precipitation during 1956–2005 for (a),(b) the CESM1-LE; (c),(d) the CESM1-synth-LE based on the first member of the CESM1-LE; (e),(f) the full CESM1-synth-LE; (g),(h) the difference between the CESM1-synth-LE and CESM1-LE; and (i),(j) the GPCC-synth-LE. The black outline shows the upper Colorado River basin, whose areal-average precipitation statistics are shown in Fig. 11.

overestimation of variance in most regions overall (recall section 3b).

f. Challenges and opportunities of a synthetic ensemble approach

In sum, the evaluation of the CESM1-synth-LE and creation of the GPCC-synth-LE highlight both challenges and opportunities in using a synthetic ensemble approach. The signal-to-noise maps in Fig. 2 demonstrate that, with an 85-yr record, the impact of the coupled ocean–atmosphere modes on precipitation can be estimated well in regions where their influences are relatively large, but weaker teleconnections may be difficult to ascertain due to the sampling effects of independent internal variability. In contrast, it remains challenging to estimate the forced component of precipitation through standard methods, because it is small compared to the variability. Despite these estimation challenges, the CESM1-synth-LE produces variability that is very similar to that of the original CESM1-LE for two reasons. First, signals that are harder to estimate—the response to the AMV or climate change—are also smaller, so contribute less to estimations of variability. Second, a large fraction of both high- and low-frequency precipitation variability is contained in the climate noise term, whose structure remains largely intact through our block bootstrap resampling methodology. The GPCC-synth-LE exhibits spatiotemporal variability that is distinct from the CESM1-based ensembles, with generally less variance overall, particularly in the western United States. These differences underscore the utility and importance of producing observationally based ensembles. Like CESM1-synth-LE, DJF precipitation variance in GPCC-synth-LE is dominated by the climate noise term across a broad range of time scales, with implications for predictability.

4. Application to western U.S. water resources

Precipitation in the upper Colorado River basin (see outline in Fig. 10) supplies water to Arizona, California, Colorado, Nevada, New Mexico, Utah, and Wyoming, and can vary substantially from year-to-year based on the particular sampling of internal variability (e.g., Harding et al. 2012). Here, we use the GPCC-synth-LE to examine the effects of sampling uncertainty on the characteristics of precipitation variability in the upper Colorado River basin; a validation of the synthetic ensemble approach for this region using the CESM1-LE as a testbed is given in supplementary Fig. 8.

The distribution of DJF precipitation in the upper Colorado River basin in the GPCC-synth-LE is similar to that of the actual GPCC record (Fig. 11a), although it is clear that the 96-yr time series provides incomplete sampling of the distribution, especially at the upper tail. This can be seen clearly in Fig. 11b, which shows the 1st to 99th percentiles in actual observed GPCC DJF precipitation against the same percentiles in different members of the GPCC-synth-LE. For the highest quantiles, the GPCC-synth-LE suggests that the range approaches of 0.5 mm day^{-1}, meaning that the wettest winter seasons could range in their total precipitation by around 5 cm due to sampling of internal variability alone.

Motivated by prior work on western water resources (Swain et al. 2018; Persad et al. 2020), we focus on three metrics for the interannual variability in DJF precipitation: the frequency of dry years, the frequency of wet years, and the frequency of pairs of "whiplash" years that alternate from wet-to-dry or dry-to-wet. Wet years are defined as those with precipitation greater than the 80th percentile across the full GPCC record, and dry years are defined as those with precipitation less than the 20th percentile. Frequency is calculated empirically for moving 31-yr periods, the standard period used to calculate climate normals. The cutoff for dry (20th percentile) and wet (80th percentile) years are 0.51 and 0.85 mm day^{-1}. By design, the ensemble mean and ensemble spread of each metric over time is constant, reflective of the assumption of no forced trend and stationary variability. The observed dry, wet, and whiplash frequencies all fall well within the GPCC-synth-LE ensemble during all 31-yr periods (Figs. 11c,e,g). However, it is clear from comparison to individual members of the GPCC-synth-LE that the variability and secular change in the water resource metrics over the last century could have been substantially greater than what actually occurred (e.g., member 1 shown in Figs. 11c–h). In general, the 95% range across the ensemble shows that a given 31-yr period could have a frequency of dry or wet years ranging from less than 10% to more than 35%, and whiplash years from 0% to nearly 20%. As a result, the change in dry, wet, and whiplash year frequency from the beginning to the end of the record ranges from decreases in excess of 20% to increases in excess of 20%. Even without any climate change signal, upper Colorado River basin precipitation has the potential to vary dramatically from one 31-yr period to the next, and proper quantification of this variability is necessary for water resource planning.

5. Discussion and conclusions

Understanding regional precipitation variability and trends is critical for the purposes of planning for stable water supplies and floodwater infrastructure. However, precipitation exhibits substantial internal variability that impacts our ability to estimate its statistics and infer the significance and attribution of trends even given an approximately century-long record. In this work, we presented a statistical model that can be used to create a synthetic ensemble of precipitation fields whose spatial and temporal characteristics are consistent with those in the actual observational record in a statistical sense, but whose chronologies differ due to random sampling of internal variability. Such a synthetic "observational large ensemble" can be analyzed in a manner analogous to that of climate model large ensembles to quantify uncertainties related to sampling of internal variability. An effectively unlimited number of ensemble members can be produced with minimal computation time, although there are finite meaningfully different combinations of internal variability.

Testing of the synthetic ensemble methodology using a climate model large ensemble, the CESM1-LE, highlighted the challenge of inferring a forced component in DJF regional precipitation over the last century using a single climate model simulation or observational record, because it remains small
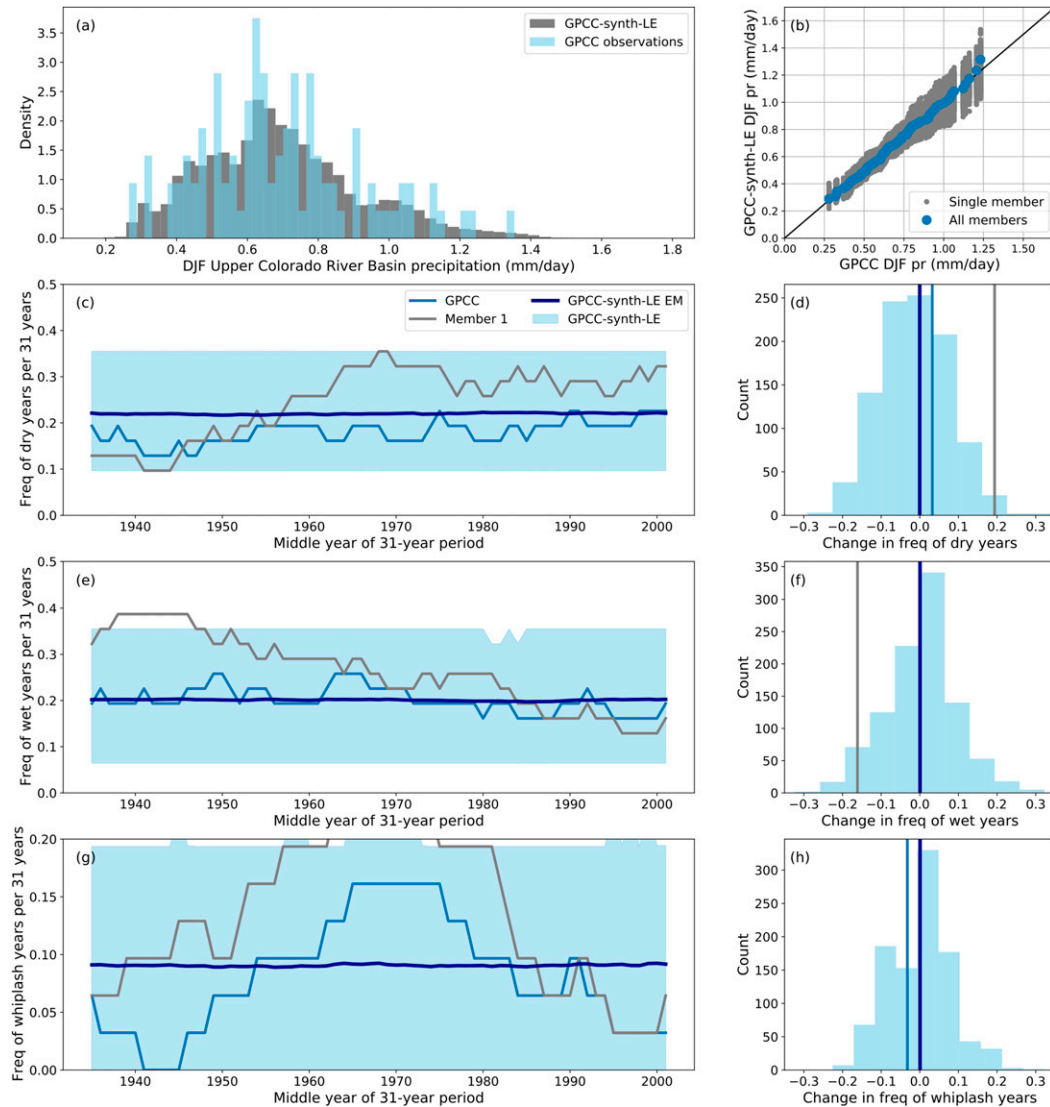
FIG. 11. Changes in water resource metrics in the GPCC-synth-LE due to sampling of internal variability. (a) The distribution of upper Colorado River basin (see outline in Fig. 11) DJF precipitation in the GPCC observational record (blue) and the GPCC-synth-LE (gray). (b) A quantile–quantile plot showing the GPCC observations vs the GPCC-synth-LE. The gray dots show the quantiles in each member of the GPCC-synth-LE, and the blue dots are the average across members. The diagonal line is the one-to-one line. (c) The time series of the frequency of dry years in the GPCC observations (light blue line), the first member of the GPCC-synth-LE (gray line), and the ensemble mean of the GPCC-synth-LE (dark blue line). The light blue shading shows the 5%–95% range across the GPCC-synth-LE. (d) The change in the frequency of dry years from the first (1921–51) to the last (1988–2018) 31-yr period. The vertical lines show the change in the ensemble mean (dark blue), the observations (light blue), and the first member of the GPCC-synth-LE (gray). (e),(f) As in (c) and (d), but for wet years. (g),(h) As in (c) and (d), but for whiplash years.

compared to the magnitude of internal variability. In the context of the synthetic ensembles, we thus chose to remove the forced component from the statistical model and produce synthetic ensembles whose variability is a function only of the precipitation response to three large-scale coupled ocean–atmosphere modes (ENSO, PDO, and AMV), and the residual atmospherically driven climate noise. Nevertheless, the dominant spatio-temporal patterns of variability and the temporal statistics of

precipitation in the CESM1-based synthetic ensemble, the CESM1-synth-LE, remained nearly identical to those in the original CESM1-LE. Further, the CESM1-synth-LE closely reproduces the 5th and 95th percentiles of 50-yr trends in DJF precipitation simulated by the CESM1-LE, despite the synthetic ensemble having no forced component, due to the much greater importance of internal variability over the forced component at the gridbox scale for historical multidecadal trends.

The relative contribution of the ocean–atmosphere modes versus the climate noise was also explored using "noise only" versions of the synthetic ensemble, within which all across-ensemble variability was due to resampling of the climate noise only. The large majority of both high- and low-frequency variability was shown to be contained in the climate noise term, rather than linked to one of the three modes.

While our creation of the CESM1-synth-LE was primarily for the purposes of validating the synthetic ensemble methodology, synthetic ensembles could also be produced in order to expand the number of members in a climate model large ensemble. This hybrid approach of expansion of climate model large ensembles with statistical methods may be particularly useful for analyses of extreme events, which may be undersampled even in an ensemble with tens or hundreds of members. Importantly, the synthetic ensembles are primarily focused on emulation of internal variability; the original climate model large ensembles remain necessary for estimating the forced trend.

When applied to the observations, the synthetic ensemble offers the benefit of more closely capturing the observed structures of variability than may be possible in a global climate model. For example, the CESM1-based ensembles tend to be overly variable in the western half of the domain as compared to the GPCC-synth-LE, resulting in simulation of extremes that are substantially larger than is realistic (e.g., cf. Figs. 7a and 8d).

Motivated by this finding, we analyzed additional metrics of precipitation variability for the upper Colorado River basin, a region of considerable importance to western U.S. water resources. Despite the smaller variability in GPCC compared to CESM1, there is still a wide range of uncertainty in 31-yr "climatologies" of dry, wet and whiplash years due to random sampling of internal variability alone. Any given 31-yr period could have a dry or wet year frequency ranging from 10% to 35% (95% range), and a frequency of whiplash years from 0% to 20%, with the shifts between 31-yr averages being largely unpredictable. Further, individual ensemble members can show nearly monotonic trends in any of the water resource metrics (e.g., Figs. 11c,e) due to sampling of internal variability alone.

We next turn to discussion of key assumptions and limitations that should be considered.

First, we validated our methodology using CESM1 alone. If CESM1 were to exhibit a vastly different structure of variability than the observations, or a substantially different division between the climate change signal and the internal variability (cf. the signal-to-noise paradox in climate models; Scaife and Smith 2018), the validation reported here might not apply to the observational record. In particular, the synthetic ensemble approach will tend to perform less well when there is more low-frequency variability, since it is challenging to estimate this variability from a short record. While there is some evidence that precipitation in the observations has a longer memory than in CESM1—for example, we identify a longer block length for our GPCC-synth-LE than for the CESM1-synth-LE (4 years vs 2 years)—precipitation variability is dominated by high-frequency variations in both the observations and CESM1, so we expect that the CESM1 validation should be sufficiently accurate. To perform a more complete

validation, one could also produce synthetic ensembles for all existing large ensembles, as well as validate the observationally based ensemble against the single observational record using ensemble forecast verification metrics such as rank histograms (Suarez-Gutierrez et al. 2020). In addition, it is important to recognize that there are inherent and unavoidable limitations in validating the methodology as applied to the observations, because there is no true ensemble of observations to compare with. Specifically, the estimation of our model parameters themselves will be influenced by the particular sampling of internal variability that we have observed, and the variability produced by the synthetic ensemble will not perfectly reproduce the "true" (but unobserved) variability in the climate system because it is based on a limited record (see Fig. 6 and supplementary Figs. 6 and 7). Nevertheless, our validation of the approach within the CESM1-LE suggests that, even when the synthetic ensemble is based on a single record, it still performs well in simulating the full ensemble spread.

Second, we have assumed ergodicity, that is, that information about the temporal evolution of the climate system can be used to create an ensemble, as well as stationarity of the variability and teleconnections over the historical record. Due to the magnitude of interannual precipitation variability, changes in its variance (Pendergrass et al. 2017) or in the structure of teleconnections (Van Oldenborgh and Burgers 2005) are typically not detectable within the observational record. However, inasmuch as there is reason to believe that precipitation variability and/or teleconnections will change in a future climate, it would be necessary to modify the approach before applying it to future projections. For example, multiple climate model large ensembles project an increase in upper Colorado River basin precipitation variability over the twenty-first century, despite a lack of agreement about any change in the mean (see Fig. 2 in Deser et al. 2020).

Third, we have assumed that the low-frequency variability in precipitation can be summarized by a linear relationship with the three dominant modes of the climate system, ENSO, PDO, and AMV, and that the modes behave independently of each other (after orthogonalizing the PDO time series with respect to the ENSO time series). Further, we have assumed that the statistics of the modes are stationary in the observational record. Similar to our discussion of precipitation variability and teleconnections above, changes in the modes are not yet detectable in the observational record, and models differ on how they may change in the future (Maher et al. 2018; Beobide-Arsuaga et al. 2021). Nevertheless, inasmuch as it is supported by the observations and/or theory, it would be advantageous to create a more sophisticated model for the coupled ocean–atmosphere modes, including the potential for a change in their statistics in the future. While future modeling work, especially focused on specific regions, may want to incorporate additional predictors, increases in complexity should be justified by significant increases in skill.

In sum, we have expanded upon prior work building "observational large ensembles" to focus on winter precipitation variability in North America. We have found that, given an approximately century-long record, the synthetic ensemble can

reproduce the variability in important precipitation statistics, including those related to water resources. While our analysis in this and prior work has focused only on temperature, precipitation, and sea level pressure, most physical quantities in the climate system also exhibit substantial spatiotemporal variability that challenges our ability to completely characterize the internal variability in a short observational record. This characterization is key for not only the water resource applications we focused on here, but also for detection and attribution of trends (e.g., Hegerl et al. 1996) and validation of climate models (Deser et al. 2020).

Data availability statement. Simulations from the NCAR CESM1-LE are publicly available via https://www.cesm.ucar.edu/projects/community-projects/LENS/. GPCC precipitation is publicly available at https://psl.noaa.gov/data/gridded/data.gpcc.html. Code to create the synthetic ensembles is publicly available at https://github.com/karenamckinnon/observational_large_ensemble/. Members of the GPCC-synth-LE are available by contacting K.A.M.

REFERENCES

Beobide-Arsuaga, G., T. Bayr, A. Reintges, and M. Latif, 2021: Uncertainty of ENSO-amplitude projections in CMIP5 and CMIP6 models. *Climate Dyn.*, **56**, 3875–3888, https://doi.org/10.1007/s00382-021-05673-4.

Beusch, L., L. Gudmundsson, and S. I. Seneviratne, 2020: Emulating Earth system model temperatures with MESMER: From global mean temperature trajectories to grid-point-level realizations on land. *Earth Syst. Dyn.*, **11**, 139–159, https://doi.org/10.5194/esd-11-139-2020.

Box, G. E., and D. R. Cox, 1964: An analysis of transformations. *J. Roy. Stat. Soc.*, **B26**, 211–243, https://doi.org/10.1111/J.2517-6161.1964.TB00553.X.

Castruccio, S., Z. Hu, B. Sanderson, A. Karspeck, and D. Hammerling, 2019: Reproducing internal variability with few ensemble runs. *J. Climate*, **32**, 8511–8522, https://doi.org/10.1175/JCLI-D-19-0280.1.

Chen, J., F. P. Brissette, and R. Leconte, 2010: A daily stochastic weather generator for preserving low-frequency of climate variability. *J. Hydrol.*, **388**, 480–490, https://doi.org/10.1016/j.jhydrol.2010.05.032.

——, H. Chen, and S. Guo, 2018: Multi-site precipitation downscaling using a stochastic weather generator. *Climate Dyn.*, **50**, 1975–1992, https://doi.org/10.1007/s00382-017-3731-9.

Chen, X., and J. M. Wallace, 2016: Orthogonal PDO and ENSO indices. *J. Climate*, **29**, 3883–3892, https://doi.org/10.1175/JCLI-D-15-0684.1.

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

Delworth, T. L., F. Zeng, L. Zhang, R. Zhang, G. A. Vecchi, and X. Yang, 2017: The central role of ocean dynamics in connecting the North Atlantic Oscillation to the extratropical component of the Atlantic multidecadal oscillation. *J. Climate*, **30**, 3789–3805, https://doi.org/10.1175/JCLI-D-16-0358.1.

Deser, C., A. S. Phillips, and J. W. Hurrell, 2004: Pacific interdecadal climate variability: Linkages between the tropics and the North Pacific during boreal winter since 1900. *J. Climate*, **17**, 3109–3124, https://doi.org/10.1175/1520-0442(2004)017<3109:PICVLB>2.0.CO;2.

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

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

——, I. R. Simpson, A. S. Phillips, and K. A. McKinnon, 2018: How well do we know ENSO's climate impacts over North America, and how do we evaluate models accordingly? *J. Climate*, **31**, 4991–5014, https://doi.org/10.1175/JCLI-D-17-0783.1.

——, and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Climate Change*, **10**, 277–286, https://doi.org/10.1038/s41558-020-0731-2.

Dong, L., L. R. Leung, F. Song, and J. Lu, 2018: Roles of SST versus internal atmospheric variability in winter extreme precipitation variability along the U.S. West Coast. *J. Climate*, **31**, 8039–8058, https://doi.org/10.1175/JCLI-D-18-0062.1.

Durack, P. J., S. E. Wijffels, and R. J. Matear, 2012: Ocean salinities reveal strong global water cycle intensification during 1950 to 2000. *Science*, **336**, 455–458, https://doi.org/10.1126/science.1212222.

Giorgi, F., and X. Bi, 2009: Time of emergence (TOE) of GHG-forced precipitation change hot-spots. *Geophys. Res. Lett.*, **36**, L06709, https://doi.org/10.1029/2009GL037593.

Harding, B., A. Wood, and J. Prairie, 2012: The implications of climate change scenario selection for future streamflow projection in the upper Colorado River basin. *Hydrol. Earth Syst. Sci.*, **16**, 3989–4007, https://doi.org/10.5194/hess-16-3989-2012.

Hegerl, G. C., H. von Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones, 1996: Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. *J. Climate*, **9**, 2281–2306, https://doi.org/10.1175/1520-0442(1996)009<2281:DGGICC>2.0.CO;2.

Held, I. M., and B. J. Soden, 2006: Robust responses of the hydrological cycle to global warming. *J. Climate*, **19**, 5686–5699, https://doi.org/10.1175/JCLI3990.1.

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

Kleiber, W., R. W. Katz, and B. Rajagopalan, 2012: Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resour. Res.*, **48**, W01523, https://doi.org/10.1029/2011WR011105.

Kooperman, G. J., Y. Chen, F. M. Hoffman, C. D. Koven, K. Lindsay, M. S. Pritchard, A. L. Swann, and J. T. Randerson, 2018: Forest response to rising CO_2 drives zonally asymmetric rainfall change over tropical land. *Nat. Climate Change*, **8**, 434–440, https://doi.org/10.1038/s41558-018-0144-7.

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

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

Maher, N., D. Matei, S. Milinski, and J. Marotzke, 2018: ENSO change in climate projections: Forced response or internal variability? *Geophys. Res. Lett.*, **45**, 11 390–11 398, https://doi.org/10.1029/2018GL079764.

——, and Coauthors, 2019: The Max Planck Institute Grand Ensemble: Enabling the exploration of climate system variability. *J. Adv. Model. Earth Syst.*, **11**, 2050–2069, https://doi.org/10.1029/2019MS001639.

Mahlstein, I., R. W. Portmann, J. S. Daniel, S. Solomon, and R. Knutti, 2012: Perceptible changes in regional precipitation in a future climate. *Geophys. Res. Lett.*, **39**, L05701, https://doi.org/10.1029/2011GL050738.

Mankin, J. S., F. Lehner, S. Coats, and K. A. McKinnon, 2020: The value of initial condition large ensembles to robust adaptation decision-making. *Earth's Future*, **8**, e2012EF001610, https://doi.org/10.1029/2020EF001610.

Martin, E. R., C. Thorncroft, and B. B. Booth, 2014: The multidecadal Atlantic SST–Sahel rainfall teleconnection in CMIP5 simulations. *J. Climate*, **27**, 784–806, https://doi.org/10.1175/JCLI-D-13-00242.1.

McKinnon, K. A., and C. Deser, 2018: Internal variability and regional climate trends in an observational large ensemble. *J. Climate*, **31**, 6783–6802, https://doi.org/10.1175/JCLI-D-17-0901.1.

——, 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**, 7585–7598, https://doi.org/10.1175/JCLI-D-16-0905.1.

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

Newman, M., and Coauthors, 2016: The Pacific decadal oscillation, revisited. *J. Climate*, **29**, 4399–4427, https://doi.org/10.1175/JCLI-D-15-0508.1.

Pendergrass, A. G., R. Knutti, F. Lehner, C. Deser, and B. M. Sanderson, 2017: Precipitation variability increases in a warmer climate. *Sci. Rep.*, **7**, 17966, https://doi.org/10.1038/s41598-017-17966-y.

Persad, G. G., D. L. Swain, C. Kouba, and J. P. Ortiz-Partida, 2020: Inter-model agreement on projected shifts in California hydroclimate characteristics critical to water management. *Climatic Change*, **162**, 1493–1513, https://doi.org/10.1007/s10584-020-02882-4.

Phillips, A. S., C. Deser, and J. Fasullo, 2014: Evaluating modes of variability in climate models. *Eos, Trans. Amer. Geophys. Union*, **95**, 453–455, https://doi.org/10.1002/2014EO490002.

Rayner, N., D. E. Parker, E. Horton, C. K. Folland, L. V. Alexander, D. Rowell, E. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, https://doi.org/10.1029/2002JD002670.

Roberts, J., and T. D. Roberts, 1978: Use of the Butterworth low-pass filter for oceanographic data. *J. Geophys. Res.*, **83**, 5510–5514, https://doi.org/10.1029/JC083iC11p05510.

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

Ruprich-Robert, Y., T. Delworth, R. Msadek, F. Castruccio, S. Yeager, and G. Danabasoglu, 2018: Impacts of the Atlantic multidecadal variability on North American summer climate and heat waves. *J. Climate*, **31**, 3679–3700, https://doi.org/10.1175/JCLI-D-17-0270.1.

Sarojini, B. B., P. A. Stott, and E. Black, 2016: Detection and attribution of human influence on regional precipitation. *Nat. Climate Change*, **6**, 669–675, https://doi.org/10.1038/nclimate2976.

Scaife, A. A., and D. Smith, 2018: A signal-to-noise paradox in climate science. *npj Climate Atmos. Sci.*, **1**, 28, https://doi.org/10.1038/s41612-018-0038-4.

Schneider, U., T. Fuchs, A. Meyer-Christoffer, and B. Rudolf, 2008: Global precipitation analysis products of the GPCC. DWD Global Precipitation Climatology Centre, 112 pp.

Schreiber, T., and A. Schmitz, 1996: Improved surrogate data for nonlinearity tests. *Phys. Rev. Lett.*, **77**, 635–638, https://doi.org/10.1103/PhysRevLett.77.635.

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**, 8313–8338, https://doi.org/10.1175/JCLI-D-18-0168.1.

——, S. G. Yeager, K. A. McKinnon, and C. Deser, 2019: Decadal predictability of late winter precipitation in western Europe through an ocean–jet stream connection. *Nat. Geosci.*, **12**, 613–619, https://doi.org/10.1038/s41561-019-0391-x.

Steinschneider, S., and C. Brown, 2013: A semiparametric multivariate, multisite weather generator with low-frequency variability for use in climate risk assessments. *Water Resour. Res.*, **49**, 7205–7220, https://doi.org/10.1002/wrcr.20528.

Suarez-Gutierrez, L., N. Maher, and S. Milinski, 2020: Evaluating the internal variability and forced response in large ensembles. *US CLIVAR Variations*, **18**, 27–35, https://doi.org/10.5065/0DSY-WH17.

Sun, C., J. Li, and S. Zhao, 2015: Remote influence of Atlantic multidecadal variability on Siberian warm season precipitation. *Sci. Rep.*, **5**, 16853, https://doi.org/10.1038/srep16853.

Swain, D. L., B. Langenbrunner, J. D. Neelin, and A. Hall, 2018: Increasing precipitation volatility in twenty-first-century California. *Nat. Climate Change*, **8**, 427–433, https://doi.org/10.1038/s41558-018-0140-y.

Tél, T., T. Bódai, G. Drótos, T. Haszpra, M. Herein, B. Kaszás, and M. Vincze, 2020: The theory of parallel climate realizations. *J. Stat. Phys.*, **179**, 1496–1530, https://doi.org/10.1007/s10955-019-02445-7.

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

Trenberth, K. E., 2011: Changes in precipitation with climate change. *Climate Res.*, **47**, 123–138, https://doi.org/10.3354/cr00953.

——, and D. J. Shea, 2006: Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.*, **33**, L12704, https://doi.org/10.1029/2006GL026894.

Van Oldenborgh, G. J., and G. Burgers, 2005: Searching for decadal variations in ENSO precipitation teleconnections. *Geophys. Res. Lett.*, **32**, L15701, https://doi.org/10.1029/2005GL023110.

Verdin, A., B. Rajagopalan, W. Kleiber, and R. W. Katz, 2015: Coupled stochastic weather generation using spatial and generalized linear models. *Stochastic Environ. Res. Risk Assess.*, **29**, 347–356, https://doi.org/10.1007/s00477-014-0911-6.

von Trentini, F., E. E. Aalbers, E. M. Fischer, and R. Ludwig, 2020: Comparing interannual variability in three regional single-model initial-condition large ensembles (SMILEs) over Europe. *Earth Syst. Dyn.*, **11**, 1013–1031, https://doi.org/10.5194/esd-11-1013-2020.

Wang, S., J. Huang, Y. He, and Y. Guan, 2014: Combined effects of the Pacific Decadal Oscillation and El Niño–Southern Oscillation on global land dry–wet changes. *Sci. Rep.*, **4**, 6651, https://doi.org/10.1038/srep06651.

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

——, and R. L. Wilby, 1999: The weather generation game: A review of stochastic weather models. *Prog. Phys. Geogr.*, **23**, 329–357, https://doi.org/10.1177/030913339902300302.

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, https://doi.org/10.1002/2017GL076327.

Zhang, H., R. Seager, J. He, H. Diao, and S. Pascale, 2021: Quantifying atmosphere and ocean origins of North American precipitation variability. *Climate Dyn.*, **56**, 4051–4074, https://doi.org/10.1007/s00382-021-05685-0.

Zhang, L., and T. L. Delworth, 2015: Analysis of the characteristics and mechanisms of the Pacific decadal oscillation in a suite of coupled models from the Geophysical Fluid Dynamics Laboratory. *J. Climate*, **28**, 7678–7701, https://doi.org/10.1175/JCLI-D-14-00647.1.