The inherent uncertainty of precipitation variability, trends, and extremes due to internal variability, with implications for Western US water resources

Karen A. McKinnon\textsuperscript{1*} and Clara Deser\textsuperscript{2}
\textsuperscript{1}Department of Statistics and Institute of the Environment and Sustainability; University of California, Los Angeles; Los Angeles, CA 90095; kmckinnon@ucla.edu
\textsuperscript{2}National Center for Atmospheric Research; Boulder, CO 80305

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. Here, we expand upon a statistical model that partitions the variance in monthly-average precipitation into a forced component, a dependence on large-scale coupled ocean-atmosphere modes, and residual ‘climate noise’. By appropriately introducing stochasticity into the statistical model, we produce synthetic ensembles with coherent spatiotemporal variability. The synthetic ensembles can closely reproduce the high- and low-frequency variability, and magnitude of extreme events, in winter precipitation in the extended contiguous United States when trained and tested with the NCAR Community Earth System Model version 1 Large Ensemble (CESM1-LE). In addition, the validation reveals the inherent challenges in estimating these types of statistics from even a near-century long record. We then create a synthetic ensemble based on the Global Precipitation Climatology Centre (GPCC) dataset, termed the GPCC-synth-LE, and 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 despite a small forced component.
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. Over ocean, energetic constraints suggest a ‘wet get wetter, dry get drier’ (Held and Soden, 2006) pattern, whose zonal-mean structure is dominated by changes in precipitation over those in evaporation. This pattern has been identified in observations (Durack et al., 2012; Polson et al., 2016) and model projections (Trenberth, 2011; Liu and Allan, 2013), although there can be important regional deviations from the theory (Sarojini et al., 2016; Deser et al., 2020a, and references therein). Over land, the picture is more complicated due to changes in temperature gradients and relative humidity (Byrne and O’Gorman, 2015), and the influence of vegetation (Lemordant et al., 2018; 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 multi-decadal trends, for the purposes of decision making and risk management (Mankin et al., 2020).
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 the El Niño-Southern Oscillation (ENSO) on North American precipitation. For example, California precipitation could either show no significant response to El Niño events, or a strong wetting, depending on the phases and amplitudes of the intrinsic atmospheric variability independent of ENSO that were present in the averaging period. Similarly, most of the variability in winter US 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. For example, ENSO, a dominant mode of interannual variability in the climate system, is well-known to affect precipitation in many regions around the world (e.g. Ropelewski and Halpert, 1987). North American precipitation, particularly in the West and South, is sensitive to the phase of the Pacific Decadal Oscillation (PDO) (Deser et al., 2004; Zhang and Delworth, 2015; Newman et al., 2016). Further, the phase of the PDO has been suggested to modulate the impacts of ENSO on precipitation (e.g. Wang et al., 2014), although this interpretation has been questioned by Newman et al. (2016) who note the challenges of inferring this type of modulation from the observational record. The Atlantic multidecadal variability (AMV) has been shown to influence precipitation via changes in atmospheric circulation in regions such as Western Europe (Simpson et al., 2019), the Sahel (Martin et al., 2014), Siberia (Sun et al., 2015), and Southwest North America (L’Heureux et al., 2015; Ruprich-Robert et al., 2018).
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 a large number of simulations (typically 30-100) with a single fully-coupled climate model under a particular radiative forcing scenario, but with small 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., 2019). These large ensembles have been used to demonstrate the irreducible uncertainty in climate trends (see Deser et al., 2020b, 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 for assessing sampling fluctuations due to internal variability is the formulation of statistical models based on the observed 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 (Steinschneider and Brown, 2013; 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 produces synthetic climate ensembles constrained by the observational record. Our synthetic ensembles preserve both the spatial
and temporal correlation structure of precipitation variability on seasonal to multidecadal
timescales, in contrast to methods that focus only on the temporal variability at each loca-
tion separately (e.g. Thompson et al., 2015; Castruccio et al., 2019) so cannot be used to
explore variability in large-scale precipitation patterns.

The current work advances the statistical methodology of McKinnon et al. (2017); McK-
innon and Deser (2018) and contains novel results based on the synthetic ensembles. The up-
dated statistical methodology focuses on monthly-average precipitation; improves the mod-
eling 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-year
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 explores the relative
contributions of different terms in the statistical model to precipitation variability. Section
4 validates and 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 our synthetic ensembles.

2.1 Climate model output and observational data

We use monthly output of precipitation and SST at a nominal spatial resolution of 1 degree 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 January 1, 1920; the spread across the ensemble was introduced by adding round-off level perturbations (order $10^{-14}$) to the initial atmospheric temperatures (Kay et al., 2015). The CESM1-LE was forced by the historical forcing scenario from 1920-2005 (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 are from the Global Precipitation Climatology Centre (GPCC, Schneider et al., 2008), which provides a gridded land-only record of precipitation from 1891-2016 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 non-negative, 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 $\lambda$ that controls the specific form of the transform is selected via maximization of the log likelihood function for the Gaussian distribution for each gridbox and month independently. The few monthly precipitation values that are exactly zero are increased to a very small positive number so that the transform is valid everywhere.

### 2.2 Statistical model

Monthly-average precipitation over land is described 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’. Mathematically, we estimate the contribution of each of these terms through fitting a multiple linear regression model to time series (in monthly time steps $t$) of transformed precipitation, $P^{i,t}_t$, for each month ($m$) and gridbox ($i$) separately,

$$P^{i,t}_t = \beta_0^{i,m} + \beta_F^{i,m} F^t + \beta_{ENSO}^{i,m} ENSO^t + \beta_{PDO}^{i,m} PDO_\perp^t + \beta_{AMV}^{i,m} AMV^t + \epsilon^{i,t}_t$$

The $\beta$ coefficients on the right hand side (except $\beta_0$, the mean) describe the monthly-varying spatial pattern of sensitivity of $P$ to anthropogenic forcing, ENSO, PDO, and AMV, respectively. The model is fit separately for each month due to the known seasonal dependence of the forced response and atmospheric teleconnections associated with the oceanic modes.

The time series associated with the estimation of the forced component, $F^t$, is the global mean, ensemble mean time series of near-surface air temperature from the CESM1-LE following Dai et al. (2015). By projecting $P$ onto a time series indicative of the evolution of the
forced response, the amount of unforced variability aliased onto the forced signal is reduced
compared to assuming a linear trend, although not eliminated. The skill of this method, as
well as a comparison to other approaches, is discussed in Section 3.

The time series ENSO$^f$ is the standard Niño3.4 SST index. The PDO time series is calcu-
lated as the principal component associated with the leading empirical orthogonal function
of SST anomalies in the North Pacific, polewards of 20N, where the anomalies are calculated
through removing both the climatological annual cycle and global-average SST from each
gridpoint. The time series of ENSO and the PDO are highly correlated, and components
of the PDO have been suggested to be mid-latitude 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, PDO$^{-}_{t}$ that is orthogonal to the standard Niño3.4 time series by removing
the projection of the PDO time series onto the Niño3.4 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 PDO$^{-}_{t}$ is dominated by the classical PDO pattern in the
midlatitudes with a negative region of SST anomalies extending from Japan into the cen-
tral Pacific and a positive region off the west coast of North America, with a much weaker
tropical component (Figure S1). The difference between the SST anomaly pattern of the
traditional PDO and that associated with PDO$^{-}_{t}$ strongly resembles an El Niño pattern,
and has a pattern correlation with the SST anomaly pattern associated with ENSO$^f$ of 0.95
(not shown).

The AMV time series, AMV$^t$, is calculated as the average of SST anomalies in the North
Atlantic (0-60N, 80W-0) minus the near-global mean SST (60S-60N) (Trenberth and Shea,
2006). The time series is smoothed with a lowpass Butterworth filter using a forward and
backward digital filter and a cutoff frequency of 1/20 year$^{-1}$ 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 AMV$^t$ time series is
not significantly correlated at the 0.05 level with ENSO$^f$ or PDO$^{-}_{t}$ at zero lag. At nonzero
lags, there are greater correlations between AMV\textsuperscript{t} and PDO\perp\textsuperscript{t}, peaking as high as 0.57 when PDO\perp\textsuperscript{t} leads AMV\textsuperscript{t} by 21 years. However, given the very small number of degrees of freedom in the AMV\textsuperscript{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\textsuperscript{t}, PDO\perp\textsuperscript{t}, and AMV\textsuperscript{t}) are normalized to have unit standard deviation.

The model parameters (\(\beta_0, \beta_F, \beta_{ENSO}, \beta_{PDO\perp}, \beta_{AMV}\)) are estimated using multiple linear regression of \(P\) onto the covariates. 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.

### 2.3 Generating the synthetic ensemble

In order to move from Eqn. (1), 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 Eqn. (1), we view all \(\beta\) terms as fixed and reflective of the physics of the climate system. We additionally view the time series \(F^t\) as deterministic, given that it summarizes the past human influence on the climate. In contrast, we view the time series of each mode (ENSO\textsuperscript{t}, PDO\perp\textsuperscript{t}, and AMV\textsuperscript{t}) as well as the residual (\(\epsilon^{i,t}\)) as stochastic.

To produce alternative versions of the time series of ENSO\textsuperscript{t}, PDO\perp\textsuperscript{t}, and AMV\textsuperscript{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. The resulting time series, 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 (Figure S2).

To produce alternative realizations of \(\epsilon^{i,t}\), we perform a moving block bootstrap in time, where the block size is constrained to be an integer number of years to retain any seasonality in the variability. The block size for each gridbox \(i\) and month \(m\) is selected using the methods...
of Wilks (1997), and is a function of the autocorrelation and length of the residual time series. The selected block size for each gridbox is summarized in Figure S3, 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 chosen as the 97th percentile across gridboxes and months of the selected block sizes; 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 two years, and that for the observed precipitation from GPCC is four 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 Eqn. (1) to produce new versions of our precipitation data as follows:

\[ P_{synth}^{i,t} = \beta_0^{i,m} + \beta_F^{i,m} F^t + \beta_{ENSO}^{i,m} ENSO_{synth}^t + \beta_{PDO}^{i,m} PDO_{synth}^t + \beta_{AMV}^{i,m} AMV_{synth}^t + \epsilon_{synth}^{i,t} \]  

The approach can be applied across the global land masses, although we focus our subsequent discussion on a subset of North America centered on the contiguous United States (24-60N, 230-300E) during boreal winter (December-January-February, or 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 our synthetic ensemble methodology, these challenges surface in two ways: first, the estimation of the model parameters of Eqn. (1) and second, the variability in the statistics of precipitation across members of the synthetic ensemble that are based on a single observational record or model simulation.

We explore both of these questions using 85 years (1921-2005) of model output from the CESM1-LE as our testbed. (See Figure 4 in Deser et al., 2020b, for a schematic summarizing the use of initial condition large ensembles as methodological testbeds.) This record length is comparable to that of many in situ-based climate datasets; in the Discussion, we demonstrate the differences in our results when only 40 years of data are available, analogous to the length of the satellite record. We fit Eqn. (1) to each of the 40 members of the CESM1-LE, and then use Eqn. (2) to produce 40 synthetic ensemble members from each original member of the CESM1-LE, producing a total of 1600 ensemble members that comprise 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.

3.1 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 $\beta$ parameter in Eqn. (1). Because the
same climate model was used to create all members of the CESM1-LE, we would expect that the ‘true’ values of the $\beta$ 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. This spread, measured by the standard deviation, as well as the average estimate of the parameters from the 40 ensemble members and the signal-to-noise ratio (SNR, average divided by spread), is shown in Figure 1 for DJF (Figure S4 for JJA). The SNR exceeds two for the precipitation response to ENSO in the western and southern parts of the domain, and is greater than one for the response to the PDO⊥ in Washington State, along the California coast, and in the non-coastal eastern US and Canada (Figure 1f, i). For both the forced component and the AMV, the SNR is always less than one, and is typically close to zero (Figure 1c, l). The low SNR for the forced component is primarily due to the large noise term (Figure 1b), whereas for the AMV it is due to the small signal (Figure 1j.) The small signal identified for the AMV, and to some extent PDO⊥, 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 results for JJA are similar, although the regions where the SNR is large for ENSO and PDO⊥ differ. Using an alternative method, low-frequency component analysis (Wills et al., 2018), to estimate the forced response does not increase the SNR (Figure S5), perhaps because LFCA does not incorporate information about the evolution of the forced response as we do through our choice of $F^t$.

The SNR maps for the $\beta$ values associated with each coupled ocean-atmosphere mode reflect the spatial structure of the signal: regions with large SNR have large signals, and the converse (Figure 1d-l). In contrast, the SNR map for the forced component is dominated by the noise: even regions with large signals such as northern Mexico and the US Southwest have a SNR less than one due to the large spread across the ensemble (Figure 1a-c). This result is consistent with prior work noting the challenge of detecting trends in precipitation
at the gridbox scale (Fischer and Knutti, 2014). In the context of the synthetic ensemble methodology, an alternative approach, given the large uncertainty in $\beta_F$, is to remove the forced trend from the model. This type of simplification typically induces a bias, but reduces the variance of the output, which can be advantageous for certain applications. Although our primary results will retain the $\beta_F$ term, we find very similar behavior when we create a ‘no forced component’ version of the CESM1-synth-LE by removing the $\beta_F$ term from Eqns. (1) and (2) (Figure S6); differences of interest will be highlighted below.

3.2 Variance in precipitation statistics

We now turn to our second question: how well can we estimate the statistical characteristics of precipitation given a single record? While there are innumerable metrics to summarize the spatiotemporal statistics of precipitation, here we focus on three: high frequency ($< 10$ year), low frequency ($> 10$ year) variability, and the 150-year return period precipitation event. Given the small contribution of the forced component to precipitation variability (Figure S6), we analyze the full 1921-2005 period without consideration of nonstationarity.

High-frequency variability is calculated as the interannual variance in DJF-average precipitation after using a highpass forward and backward digital Butterworth filter with a frequency cutoff of $1/10$ year$^{-1}$. The interannual variance calculation excludes the first and last five years to avoid edge effects. Within the context of the CESM1-LE, the best estimate of the high-frequency variability is the average across the highpass interannual variance calculated for each ensemble member; the spread across the ensemble indicates the uncertainty in this quantity given an 85-year record. High-frequency variability is maximized along the West Coast, and has a secondary maximum spanning from the Southeast up the Eastern Seaboard (Figure 2a). By definition, the average difference between the estimate of high-frequency variability in any given ensemble member and the ensemble mean estimate is zero (Figure 2b); 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.
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 give each ensemble member a score, calculated as the areal median of the absolute difference in high-frequency variability for each ensemble member minus the ensemble mean estimate. We identify the ensemble member with the average score, and show the difference between the high-frequency variability estimated using this ensemble member and the ensemble mean in Figure 2d. This single ensemble member shows regions of both over- and under-estimation of the high-frequency variability as expected, with the largest differences along the West Coast, a region of high variability. Although we have picked out this single ensemble member to demonstrate the limitations of having a single record, the value of the scores across the ensemble members are all very similar (the standard deviation of the scores is 0.0048 mm$^2$, compared to the mean score of 0.045 mm$^2$ of the ensemble member shown), so a single member is reasonably representative of the full ensemble in terms of its error in estimating the high-frequency variability.

The above analysis was focused on the CESM1-LE alone, but a similar approach can be used to validate our synthetic 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 high-frequency variability as the average across the interannual highpass 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 high-frequency variability to the original CESM1-LE (Figure 2e). While it does exhibit a small bias over California, the spread across the ensemble is smaller than for the CESM1-LE and its average score, indicating the median error across the domain, is smaller than in CESM1-LE (Figure 2f, g, h). The lower error is likely because we are able
to average over a greater number of realizations, combined with the fact that our original 85-year record is sufficiently long that we can accurately and precisely estimate the model parameters relevant for high-frequency variability.

We now turn to our next metric, the low-frequency (greater than decadal) variability, calculated in a manner analogous to the high-frequency variability but using a lowpass, rather than highpass, filter. For DJF precipitation in the continental United States, the magnitude of the simulated low-frequency variability is approximately one-quarter of that for high-frequency variability, consistent with the small contribution of low-frequency modes shown in Figure 1. As expected, the fractional errors in estimating the low-frequency variability using a single member of the CESM1-LE are larger than for high-frequency variability (Figure 3a-d). The challenges of estimating greater-than-decadal timescale variability from a 85-year record also map onto the CESM1-synth-LE, which, although it captures the large-scale patterns in low-frequency variance, exhibits a small positive bias across the full domain (Figure 3e-f). Nevertheless, similar to our findings for high-frequency variability, the ability to simulate multiple realizations of the field leads to a reduced spread across the estimates (each one based on a single member of the CESM1-LE; Figure 3g) and generally smaller errors in the estimate of low-frequency variability given a single member (Figure 3h). Notably, removing the forced component from the CESM1-synth-LE slightly reduces both the bias (Figure S6b) and the across-ensemble variance (not shown), indicating a benefit of using a simpler model when parameter estimates are highly variable.

Finally, we examine our third metric, the magnitude of the 150-year return period DJF precipitation event at each gridbox. Return periods are calculated using the Gringorten plotting position as \( \frac{n + 0.12}{m - 0.44} \) (Gringorten, 1963), where \( n \) is the number of years in the record, and \( m \) is the rank of each event. The Gringorten formula is optimized for estimating the probability of extreme values in heavy-tailed distributions, so is appropriate for precipitation extremes. The 150-year return period event in a single member of the CESM1-LE is the maximum value in that simulation. The best estimate of the return period using the full
CESM1-LE is calculated by effectively appending all 40 simulations into a single record; as such, the 150-year return period is the 23rd most extreme event across the full ensemble. Similarly, the 40 members of the CESM1-synth-LE calculated using a single member of the CESM1-LE are grouped together to calculate the 150-year return period event. We again find that the CESM1-synth-LE has a small bias, with overestimation of the magnitude of the 150-year even in the Southwest and underestimation in the Florida Panhandle, but that there is a reduced spread across the ensemble such that we can generally get a better estimate of the CESM1-LE ensemble mean through using the synthetic ensemble approach (Figure 4).

Given an 85-year record, we cannot directly estimate via the Gringorten formula the magnitude of an event that has a return period greater than 150 years. To overcome this obstacle, it is common to fit a parametric distribution to the recurrence interval curve in order to smooth and extrapolate the empirical estimate. Here, we take a complementary approach: through our synthetic ensemble methodology, we can simply simulate more years of data to better sample the tails of the precipitation distributions. In Figure 5, we compare our empirical estimate of the magnitude of the 500-year precipitation event at each gridbox using the CESM1-synth-LE to that simulated using the full CESM1-LE. The results are similar to our findings for the 150-year event, although the bias and variance are somewhat higher, as expected. Importantly, the estimates remain essentially unbiased excepting the Southwest, the Baja Peninsula, and Florida, as before.

### 3.3 The relative contribution of the climate noise

All of the CESM1-synth-LE results thus far have been based on Eqn. (2), in which precipitation variability is a function of the forced component, the dependence on ENSO, PDO, and AMV, and the climate noise. We have demonstrated that the forced component plays a very small role in contributing to the statistics of precipitation variability that we have explored (Figure S6), but what is the relative role of the climate noise versus the large-scale
coupled ocean-atmosphere modes?

To answer this question, we produce a ‘noise only’ version of the CESM1-synth-LE via setting all $\beta$ terms to zero in Eqn. (2). Note that, in contrast to the ‘no forced component’ version in which we remove the $\beta_F$ term from Eqns. (1) and (2), we only remove the forced component and mode dependence in the simulation step. As such, variance is still partitioned to the forced component and modes in the model fitting step using Eqn. (1).

To compare the ‘noise only’ and standard CESM1-synth-LE, we calculate the ratio of high- and low-frequency variability, and the magnitudes of the 150- and 500-year events, in the ‘noise only’ ensemble to those in the standard ensemble (Figure 6). For high-frequency variability and the magnitude of the 150-year and 500-year events, a large majority of the metrics can be explained by the noise alone. Specifically, the median ratio across gridboxes between the two ensembles is 0.94, 0.97, and 0.94 for the high-frequency variability and the magnitude of the 150-year and 500-year events, respectively. As expected based on the prior results, the regions where the modes play the largest role are the Southwest/Baja Peninsula and the Florida Panhandle. The coupled ocean-atmosphere modes are a more significant but not dominant contributor to the low-frequency variability: the ratio of the low-frequency variability in the ‘noise only’ ensemble to the standard ensemble has a median value across gridboxes of 0.79. While this result may seem counterintuitive given the minimal year-to-year memory in the climate noise – recall that we have used a block size of two years – it results from the fact that even 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).

4 Application to Western US water resources

In the prior section, we demonstrated that the synthetic ensemble methodology can be used to replicate the statistical characteristics of a climate model initial-condition large ensemble.

16
We now turn to applying our methodology to the observational record, with a focus on the area-averaged DJF precipitation in the Upper Colorado River basin (see outline in Figure 5). Precipitation in the Basin supplies water to Arizona, California, Colorado, Nevada, New Mexico, Utah, and Wyoming, and has been shown in climate model output to be sensitive to sampling of internal variability (Harding et al., 2012). We first again compare the CESM1-LE and CESM1-synth-LE to validate our approach for this application, and then present results from our observationally-based synthetic ensemble.

The distribution of DJF precipitation across the Upper Colorado River Basin in the CESM1-synth-LE is consistent with that simulated by the CESM1-LE (Figure 7a), as expected based on our prior validation metrics. A quantile-quantile plot (Figure 7b) shows that the CESM1-synth-LE is, on average, an unbiased estimator of the CESM1-LE (the blue dots follow the one-to-one line), but that versions of the CESM1-synth-LE based on individual members of the CESM1-LE can slightly under- or over-estimate the true quantiles, especially at the upper tail. 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 CESM1-LE for 1921-2005, and dry years are defined as those with precipitation less than the 20th percentile. Frequency is calculated empirically for moving 31-year periods, the standard period used to calculate climate normals. The CESM1-LE ensemble mean suggests a forced response of a small decrease in dry year frequency, increase in wet year frequency, and no change in whiplash frequency (Figure 7c-h). However, the spread across the ensemble is large in all cases. For example, the member of the CESM1-LE that shows the largest increase in dry year frequency simulates a 16% frequency over the earliest 31-year period (1921-1951) rising nearly monotonically to a 32% frequency in latest 31-year period (1985-2005) despite the forced trend towards fewer dry years, whereas the member that shows the largest decrease
begins with a 35% frequency and decreases nearly monotonically to a 16% frequency (Figure 7c).

With an eye towards applying the synthetic ensemble methodology to the single observational record, we arbitrarily select the first member of the CESM1-LE, and produce a CESM1-synth-LE based on this member alone, hereafter referred to as CESM1-synth-LE_mem1. We then compare the behavior of CESM1-synth-LE_mem1 with that of the entire (40-member) CESM1-LE to assess how well our synthetic ensemble methodology applied to a single member approximates the statistical behavior of the full CESM1-LE. The first member of the CESM1-LE simulates a decrease in the frequency of dry years and an increase in the frequency of wet years, in line with the ensemble mean, but its changes are over four times greater than the ensemble mean for both metrics (Figure 7c, e). The CESM1-synth-LE_mem1 exhibits a spread in all three metrics similar to that simulated by the full CESM1-LE in both its 31-year frequency estimates (Figure 7c, e, g) and the change in frequency of wet, dry, and whiplash events from the first 31-year period to the last 31-year period (Figure 7d, f, h). Thus, given a single record, our methodology accurately reproduces the variability and spread of trends in important water resource metrics.

While the design of the synthetic ensemble is focused on quantification of variability, the increased sample size can lead to an added benefit of improved estimation of the forced trend. We quantify this by calculating, for each of our three metrics, the root-mean-square difference over time between the ensemble mean of the CESM1-synth-LE_memX, where X indicates the member of the CESM1-LE that the synthetic ensemble is based on, and the true ensemble mean based on all 40 members of the CESM1-LE. We compare these values to the root-mean-square difference of each metric over time between the member upon which the synthetic ensemble was based upon and the true ensemble mean. The comparison thus highlights any added benefit in our estimation of the forced response from creating the synthetic ensemble. The median reduction across ensemble members in the root-mean-square difference using the synthetic ensemble is 45%, 41%, and 60% for the dry
year frequency, wet year frequency, and whiplash years frequency, respectively. As such, for these water resources metrics, the synthetic ensemble not only provides an estimate of the uncertainty due to internal variability but also provides a better estimate of the forced response than is possible using a single record.

Having validated our synthetic ensemble for Upper Colorado River Basin precipitation using the CESM1-LE, we are now in a position to apply it to the observational record. Using the same methodology as for the CESM1-synth-LE, we construct 1000 synthetic ‘ensemble members’ of gridded monthly precipitation fields over global land areas based on the GPCC dataset during 1921-2016. Here, we focus on analyzing the Upper Colorado River Basin region from this GPCC-based synthetic ensemble, termed the GPCC-synth-LE. The average value of DJF precipitation in the Basin from the GPCC observational dataset is 23% of the amount simulated in the CESM1-LE (0.70 mm/day versus 2.98 mm/day), highlighting the challenges of using climate model output directly on regional scales (compare Figures 7a and 8a). The distribution of DJF precipitation in the GPCC-synth-LE is similar to that of the actual GPCC record (Figure 8a), although it is clear that the 96-year record provides incomplete sampling of the distribution, especially at the upper tail. This can be seen clearly in Figure 8b, which plots 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 is in excess of 0.5 mm/day, meaning that the wettest winter seasons could range in their total precipitation by around 5 cm due to sampling of internal variability alone.

We now turn to our three water resource metrics – frequency of dry years, wet years, and pairs of whiplash years. The cutoff for dry (20th percentile) and wet (80th percentile) years within GPCC are 0.51 mm/day and 0.85 mm/day. The ensemble mean suggests a small trend towards an increasing frequency of dry years and decreasing frequency of wet years, consistent with the inferred small forced trend towards decreasing precipitation (not shown). The spread of the GPCC-synth-LE, shown as the 2.5th to 97.5th percentile range,
is effectively constant over time, consistent with our assumption of stationary variability in
the observational record. The observed dry, wet, and whiplash frequencies all fall well within
the GPCC-synth-LE ensemble during all 31-year periods (Figure 8c, 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 Figure 8c-h).
In general, the 95% range across the ensemble shows that a given 31-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-year 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 (single) 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.
The results of the synthetic ensemble methodology highlight the challenge of inferring the forced component, and to some extent the response to coupled ocean-atmosphere modes of variability, from the single observational record, even one as long as 85 years. Despite these challenges, our validation of the synthetic ensemble methodology using the CESM1-LE as a testbed demonstrates that the synthetic ensemble, CESM1-synth-LE, generally performs well in reproducing the statistics of variability in the climate model ensemble on which it was based. Further, we have shown that the synthetic ensemble can, on average, provide a better estimate of the true high- and low-frequency variability and magnitude of the 150-year event than the original record upon which it was based (compare panels d and h in Figures 2, 3, and 4).

The synthetic ensemble offers the benefit of being directly constrained by the observations, which is particularly important for cases in which climate models exhibit large biases, such as the Upper Colorado River Basin example discussed here. Proper quantification of the statistics of precipitation in the Upper Colorado is critical for water resource management, but the fact that CESM1-LE overestimates precipitation in the region by over 300% raises concerns regarding whether the model output should be used directly for planning purposes. Our synthetic ensemble closely matches the distribution of the observed precipitation, and highlights the wide range of uncertainty in 31-year “climatologies” of dry, wet and whiplash years due to random sampling of internal variability alone. Member 1 of the GPCC-synth-LE provides a cautionary tale: although the forced component only suggests an increase from 21% to 23% in dry year frequency per 31-years over the century, the random sample of natural variability in member one leads to an increase from 16% to 42%, which would surely stress our water resources (Figure 8c).

Nevertheless, our approach has a few key assumptions and limitations that should be considered.

First, we validated our methodology using CESM1 alone. Inasmuch as CESM1 exhibits a vastly different structure of variability than the observations, the validation could be mis-
leading. In particular, the synthetic ensemble approach will tend to perform more poorly 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 (four years versus two 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 each of the models in the Multi-Model Large Ensemble Archive (Deser et al., 2020b), 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).

Second, we have assumed ergodicity, i.e. 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 21st century, despite a lack of agreement about any change in the mean (see Figure 2 in Deser et al., 2020b).

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). Inasmuch as it is supported by the observations, it would be advantageous to create a more sophisticated model for the
coupled ocean-atmosphere modes. While future modeling work, especially focused on spe-
cific regions, may want to incorporate additional predictors, increases in complexity should
be justified by significant increases in skill via, e.g. the use of information criteria.

Our analysis focused entirely on records of at least 85 years in length; however, it may
be of interest to use similar methods for variables that have shorter records, such as ocean
chlorophyll (Elsworth et al., 2020). To demonstrate the impact of having a more limited
record, we produce a new version of the CESM1-synth-LE, but using only 40 years of model
output to fit the model parameters, analogous to the length of the satellite record (i.e. since
1979). With this shorter record length, the bias and variance in the synthetic ensemble
increases. For example, using low-frequency variability as a metric, the median bias across
the domain increases from 0.008 mm$^2$ (recall Figure 3f) to 0.021 mm$^2$ (Figure 9b), and the
median across-ensemble spread of the bias increases from 0.025 mm$^2$ (Figure 3h) to 0.071
mm$^2$ (Figure 9c). The decline in performance is linked to the challenges of estimating the
$\beta$ parameters for the forced component and the AMV. Over the shorter 40 year period of
1966-2005, the two are somewhat collinear, since only a partial cycle of the AMV is sampled.
As an alternative, we can make a simpler synthetic ensemble that does not use $F^t$ or AMV$^t$
as covariates in Eqn. (1). This simpler ensemble tends to underestimate the low-frequency
variability, as expected (Figure 9f), but its bias and variance are comparable to the ensemble
created with the full model and 85 years of data. Thus, it is important – and also relatively
straightforward – to modify the synthetic ensemble approach for different types and lengths
of data, although the user must adjust their interpretation of the results based on which
covariates are included.

In sum, we have expanded upon prior work building “observational large ensembles”
to focus on precipitation variability in North America. We have found that, given an ap-
proximately 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., 2020b).

Acknowledgements

C.D. was supported by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the National Science Foundation under the Cooperative Agreement 1852977. Analysis was performed on the Casper cluster supported by by NCAR’s Computational and Information Systems Laboratory.

Data and code availability

Figure 1: The mean value (left column), standard deviation (middle column), and signal-to-noise (mean divided by standard deviation, right column) of the DJF model parameters from Eqn. (1) estimated using each member of the CESM1-LE. From top to bottom, parameters are $\beta_{F,DJF}$, $\beta_{ENSO,DJF}$, $\beta_{PDO,DJF}$, and $\beta_{AMV,DJF}$. 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.
Figure 2: High-frequency variability (less than decadal) 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. Panels (e)-(h) show the same metrics for the CESM1-synth-LE; bias is still assessed against the CESM1-LE ensemble mean shown in panel (a). In all panels, the number in the lower righthand corner is the median value across the gridboxes 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.
Figure 3: As in Figure 2, but for low-frequency variability (more than decadal) of DJF precipitation in the CESM1-LE and the CESM1-synth-LE.
Figure 4: As in Figure 2, but for the magnitude of the 150-year event of DJF precipitation in the CESM1-LE and the CESM1-synth-LE.
Figure 5: The magnitude of the 500-year 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 righthand corner is the median value across the gridboxes 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). The outline of the Upper Colorado River Basin used for the analysis in Figures 7 and 8 is shown in black.
Figure 6: The contribution of the climate noise term to each of the four validation metrics: (a) high-pass interannual variability, (b) low-pass interannual variability, (c) the magnitude of the 150-year event, and (d) the magnitude of the 500-year event. All plots show the CESM1-synth-LE ensemble mean in a version where the contribution of the forced component and modes is set to zero divided by the standard CESM1-synth-LE ensemble mean.
Figure 7: Validation of the Obs-LE methodology for water resource metrics in the Upper Colorado River Basin (see outline in Figure 5). (a, b) The distribution of DJF Upper Colorado River Basin precipitation across the full CESM1-LE and CESM1-synth-LE. (c) The time series of the frequency of dry years in overlapping 31-year periods across the full CESM1-LE (black), the first member of the CESM1-LE (light blue), and the CESM1-synth-LE_mem1, which is based on the first member of the CESM1-LE alone (dark blue). The 5%-95% range across the two ensembles is shown in gray for the CESM1-LE and light blue for the CESM1-synth-LE_mem1. (d) The change in the frequency of dry years from the first (1921-1951) to the last (1985-2005) 31-year period. All colors are the same as in (c). (e, f) As in (c, d) but for wet years. (g, h) As in (c, d) but for whiplash years.
Figure 8: As in Figure 7 but comparing the single observational record from GPCC to the GPCC-synth-LE. The first member of the GPCC-synth-LE is shown for illustration.
Figure 9: Validation of low-frequency variability in two versions of a CESM1-synth-LE based on 40 years of data (1966-2005). (a-d) As in Figure 3e-h, but for a CESM1-synth-LE fit using 40 years of data and the full model in Eqn. (1). (e-h) As in panels (a)-(d), but for a CESM1-synth-LE fit using 40 years of data but excluding $F^t$ or AMV$^t$ as covariates in Eqns. (1) and (2).
References


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 US West Coast, *Journal of Climate, 31*(19), 8039–8058.


Kooperman, G. J., Y. Chen, F. M. Hoffman, C. D. Koven, K. Lindsay, M. S. Pritchard,


Steinschneider, S., and C. Brown (2013), A semiparametric multivariate, multisite weather
generator with low-frequency variability for use in climate risk assessments, *Water Re-
sources Research*, 49(11), 7205–7220.

Suarez-Gutierrez, L., N. Maher, and S. Milinski (2020), Evaluating the internal variability

Sun, C., J. Li, and S. Zhao (2015), Remote influence of Atlantic multidecadal variability on

Swain, D. L., B. Langenbrunner, J. D. Neelin, and A. Hall (2018), Increasing precipitation

Tél, T., T. Bódai, G. Drótos, T. Haszpra, M. Herein, B. Kaszás, and M. Vincze (2019), The

Thompson, D. W., E. A. Barnes, C. Deser, W. E. Foust, and A. S. Phillips (2015), Quantifi-
fying the role of internal climate variability in future climate trends, *Journal of Climate*,
28(16), 6443–6456.


Trenberth, K. E., and D. J. Shea (2006), Atlantic hurricanes and natural variability in 2005,

Van Oldenborgh, G. J., and G. Burgers (2005), Searching for decadal variations in ENSO

variability in three regional single-model initial-condition large ensembles (SMILEs) over


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, Journal of Climate, 28(19), 7678–7701.
Supplementary Figures

Figure S1: The sea surface temperature (SST) anomaly pattern associated with PDO⊥, the component of the traditional Pacific Decadal Oscillation time series that is orthogonal to the El Niño-Southern Oscillation time series. The SST anomaly pattern is calculated by regressing the annual mean PDO⊥ time series onto the annual mean, linearly detrended SST anomalies from HadISST.
Figure S2: The empirical coherence (black line) and 95% range of coherence values from the synthetic mode times series between (a) ENSO and PDO⊥, (b) ENSO and AMV, and (c) PDO⊥ and AMV.
Figure S3: The selected block size at each gridbox for CESM1-LE and the GPCC observations. Block size is chosen for each gridbox and month using the methods of Wilks (1997); the maps show the maximum value across months. The CESM1-LE ensemble mean shows the average selected block size across members. Member 1 is shown as an example of the block size given a single record, which is more readily compared to the GPCC-based observational estimate. The average across ensemble members of the difference between the GPCC- and CESM1-based block sizes shows that, in most regions of the world outside of North America and Europe, the GPCC dataset suggests a larger block size.
Figure S4: As in Fig. 1 but for JJA.

Figure S5: The signal to noise ratio for the forced component for winter (left) and summer (right) estimated using low frequency component analysis (Wills et al., 2018). The forced component is assumed to be contained in the first low frequency pattern (LFP). The LFPs are estimated in each member of CESM1-LE as a linear combination of the first 30 EOFs for monthly precipitation. The signal is defined as the mean of the first LFP across the ensemble, and the noise is the standard deviation across the ensemble.
Figure S6: The fractional difference between a version of the CESM1-synth-LE with no forced component and the standard version based on Eqn. (1) in (a) highpass interannual variability, (b) lowpass interannual variability, (c) the magnitude of the 150-year event, and (d) the magnitude of the 500-year event. Note that the standard version of the CESM1-synth-LE overestimates the lowpass interannual variability compared to the CESM1-LE, so the reduction in lowpass interannual variability in the version with no forced component is an improvement.