A Comprehensive Intermediate-Term Drought Evaluation System and Evaluation of Climate Data Products over the Conterminous United States

ZEYU XUE^a AND PAUL ULLRICH^a

^a *Atmospheric Science Graduate Group, University of California, Davis, Davis, California*

(Manuscript received 23 December 2020, in final form 24 June 2021)

ABSTRACT: Climate models are frequently used tools for adaptation planning in light of future uncertainty. However, not all climate models are equally trustworthy, and so model biases must be assessed to select models suitable for producing credible projections. Drought is a well-known and high-impact form of extreme weather, and knowledge of its frequency, intensity, and duration are key for regional water management plans. Droughts are also difficult to assess in climate datasets, due to the long duration per event, relative to the length of a typical simulation. Therefore, there is a growing need for a standardized suite of metrics addressing how well models capture this phenomenon. In this study, we present a widely applicable set of metrics for evaluating agreement between climate datasets and observations in the context of drought. Two notable advances are made in our evaluation system: first, statistical hypothesis testing is employed for normalization of individual scores against the threshold for statistical significance. And second, within each evaluation region and dataset, principal feature analysis is used to select the most descriptive metrics among 11 metrics that capture essential features of drought. Our metrics package is applied to three characteristically distinct regions in the conterminous United States and across several commonly employed climate datasets (CMIP5/6, LOCA, and CORDEX). As a result, insights emerge into the underlying drivers of model bias in global climate models, regional climate models, and statistically downscaled models.

KEYWORDS: Drought; Climate models; Model evaluation/performance; Statistics; Model output statistics

1. Introduction

Droughts are among the most expensive and devastating extreme weather events, responsible for severe impacts for rural communities and regional agriculture (Andreadis and Lettenmaier 2006; Strzepek et al. 2010). While some regions are anticipated to experience increased precipitation because of warmer temperatures, near-constant relative humidity, and more precipitable water (Held and Soden 2006), drought risk is still anticipated to increase because of warming temperatures, intensified evapotranspiration, and greater climate variability (Dai 2013; Strzepek et al. 2010). Consequently, water managers have turned to climate data and model projections to develop “model droughts” which may be used to ascertain the challenges posed by future droughts (Frumhoff et al. 2007; Strzepek et al. 2010; Seager et al. 2012). At present, climate models are the primary methods to project climate change and its consequences (Frumhoff et al. 2007; Kharin et al. 2007; Wagener et al. 2010). However, all models are known to possess certain biases which can contaminate simulation results (Moon et al. 2018; Nasrollahi et al. 2015). Biases present in a single

modeling system can be quantified and mitigated through the use of an ensemble of global climate models (GCMs), such as the multimodel ensemble featured in the Coupled Model Intercomparison Project phases 5 and 6 (CMIP5/6), and corresponding downscaled climate data products, such as the dynamical downscaled simulations from the Coordinated Regional Climate Downscaling Experiment (CORDEX) or the statistically downscaled Localized Constructed Analogs (LOCA) product. Nonetheless, there is significant value in understanding these biases in singular products, or across an ensemble, particularly before projections are used for climate adaptation planning (Collier et al. 2018; Eyiring et al. 2019; Gleckler et al. 2008, 2016; Lee et al. 2018; Nguyen et al. 2017; Ukkola et al. 2018; Wagener et al. 2010). However, repetitive and redundant evaluations across modeling centers can incur unnecessary efforts and increased costs. Thus, comprehensive and robust evaluation systems (e.g., Eyiring et al. 2019; Gleckler et al. 2008, 2016; Lee et al. 2018) can reduce research costs and provide greater certainty in the employ of climate models for understanding drought character and impacts, especially when it comes to the large ensemble context (Deser et al. 2020).

With a patchwork of evaluation studies focusing on different products and different metrics, there remains significant uncertainty as to which model data can be trusted for correctly representing drought intensity, frequency, and duration for their region of interest (Jagannathan et al. 2021; Moon et al. 2018; Nasrollahi et al. 2015). In general, our capacity to evaluate drought depends on the duration of the drought and the length of the available dataset. While large ensembles (Kay et al. 2015) permit the assessment of multiyear droughts, individual CMIP5/6 models often only

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-20-0314.s1>.

Corresponding author: Zeyu Xue, zyxue@ucdavis.edu

DOI: 10.1175/JHM-D-20-0314.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

provide single realizations. In this paper, our evaluation suite is applied to datasets of 60 years duration, but the model performance scores are shown to be relatively stable with as little as 30 years of data. With this in mind, this paper proposes a system with four key novelties.

First, most past research has only focused on evaluating models using meteorological mean fields such as mean precipitation and temperature (Koutroulis et al. 2016; Nguyen et al. 2017); however, mean performance often does not necessarily correlate with model performance in the extreme (Wehner et al. 2020). In the case of drought, impacts from drought are related to temporal continuity features like consecutive duration which cannot be captured by mean fields. Thus holistic metrics that are more comprehensive in capturing the features of each event are needed. For this, drought indices like the standardized precipitation index (SPI) are proposed to calculate and straightforwardly employed for quantifying the character of droughts (Hayes et al. 2002; Svoboda and Fuchs 2016; Ukkola et al. 2018; Yihdego et al. 2019). Second, most drought evaluation systems (Darand and Sohrabi 2018; Orlowsky and Seneviratne 2013; Teshome and Zhang 2019) rely on a handful of Expert Team on Climate Change Detection and Indices (ETCCDI) metrics established by the World Meteorological Organization (WMO) and the World Climate Research Program (WCRP). Given that ETCCDI metrics use daily mean precipitation, their metrics (e.g., consecutive dry days) capture features on a much shorter time scale than traditionally associated with drought (Mishra and Singh 2010; Wilhite and Glantz 1985). Third, to determine a score for a model's overall performance, most evaluation systems have normalized these metrics using a standard score or min-max feature scaling (Collier et al. 2018; Yang et al. 2014; Koutroulis et al. 2016). Others have evaluated models' performance by assessing relative performance for each kind of metric (Chen and Sun 2015; Meher et al. 2017; Ukkola et al. 2018). However, a key problem that arises in this case is that the scoring does not have an absolute optimum—the score cannot convey whether a group of models are all terrible or all indistinguishable from observations. So, herein we propose to use a normalization that emerges naturally from the theory of statistical hypothesis testing. Namely, we normalize the value of each metric relative to the statistical significance of its null hypothesis (i.e., the hypothesis that the model and observations come from the same distribution or have the same mean value) at the 95% confidence level. This implies that a score of 1 has the same meaning across metrics. Finally, to build a comprehensive evaluation system which scores climate models' consistency with historical simulations, we evaluate multiple feature-specific metrics related to drought, including monthly precipitation, SPI, seasonality, drought spatial coverage, drought duration, intensity, frequency, and probability of initiation/termination. The overall performance of each model is illustrated by the assembling of principal metrics detected by selected by principal feature analysis (PFA).

This study focuses on meteorological droughts with temporal scale between seasonal and annual (herein referred to as

intermediate-term droughts). Although the term “intermediate-term drought” is widely used in numerous studies of droughts (Bhuyan et al. 2017; Kolb et al. 2016; Thomas et al. 2015; Vicente-Serrano et al. 2014; Wilhite 2005), it has no strict definition. But generally, it refers to the droughts that are quantified and detected by drought indices that operate at time scales shorter than a year but longer than a month, especially the 6-month standardized precipitation index (SPI6) (Bhuyan et al. 2017; Thomas et al. 2015; Wilhite 2005; Svoboda et al. 2012; Vicente-Serrano et al. 2014). Intermediate-term drought tends to be a focus of water managers for several reasons. First, following its original definition (McKee et al. 1993; Wilhite 2005), intermediate-term SPI (e.g., SPI6) retains advantages of both short-term and long-term indices. Namely, it not only responds quickly to emerging drought conditions but also provides greater stability and persistence for longer-term droughts; therefore, SPI6 is commonly employed in the drought monitor and associated applications (The National Drought Mitigation Center 2021; Svoboda et al. 2012; Wilhite 2005). Second, although SPI is designed to measure only meteorological drought conditions (Guttman 1999; McKee et al. 1993), some studies (e.g., The National Drought Mitigation Center 2021; McKee et al. 1993; Svoboda and Fuchs 2016; Svoboda et al. 2012) have shown that SPI can capture drought impacts on water resource availability because it is closely related to accumulated precipitation. Namely, soil moisture conditions (agriculture drought) respond to precipitation anomalies on a relatively short scale, while surface runoff and reservoir storage (hydrologic drought) are more so determined by longer-term precipitation anomalies. Therefore, at intermediate time scales, SPI6 can potentially capture the conditions related to agriculture and hydrologic droughts. This motivates its employ in characterizing model droughts for water management agencies (The National Drought Mitigation Center 2021; Svoboda et al. 2012; Zargar et al. 2011). Third, some previous studies have illustrated that there exists a general wetting trend at longer time scales and increased frequency of short-term drought conditions over many of the CONUS regions (Andreadis and Lettenmaier 2006; Orlowsky and Seneviratne 2013; Otkin et al. 2018; Seager et al. 2012; Taylor et al. 2013). However, large uncertainties still persist when it comes to intermediate-term drought conditions (Strzepek et al. 2010; Frumhoff et al. 2007). Therefore, in our present evaluation we focus on drought-feature metrics based on intermediate-term drought (SPI6), although we note that shorter droughts (as could be captured with SPI1 or SPI3) as well as multiyear droughts (SPI24 or SPI36) could also be employed in our package.

In this paper we propose an evaluation system for drought and apply this system to evaluate simulated droughts in three characteristically distinct regions of the United States in the CMIP5/6, CORDEX, and LOCA datasets. The evaluation system aims to select excellent models with less model uncertainties to project the features of drought and is subsequently employed to answer our motivating questions. First, what are the regional and model-specific characteristics that are most relevant for determining the quality of model-simulated drought? Second, when it comes to drought, does

the CMIP6 ensemble outperform its predecessor CMIP5? Third, does dynamical or statistical downscaling provided robust added value in the representation of drought? And if so, how?

The paper is structured as follows. Assessed datasets are described in [section 2](#). The statistical methods employed across the metrics suite are described briefly in [section 3](#). Our proposed suite of monthly precipitation-based drought metrics are defined in [section 4](#). Results and conclusions are then provided in [sections 5 and 6](#), respectively.

2. Datasets

Following the requirements of SPI, monthly mean precipitation from four datasets below will be used as input data and evaluated by our system.

a. CPC Unified Gauge-Based Analysis precipitation data

The Climate Prediction Center (CPC) Unified Gauge-Based Analysis precipitation data are a component of the CPC Unified Precipitation Project underway at NOAA Climate Prediction Center (CPC). The dataset covers the period 1948–2014 with a spatial resolution of 0.25° (NOAA/Physical Sciences Laboratory 2020; Chen et al. 2008; Xie et al. 2007, 2010). For this study, CPC is used as our observational dataset and serves as the “real” values—hereafter, whenever the text refers to “observational data,” we are referring to the CPC data.

b. CMIP5/6 model data

The Coupled Model Intercomparison Project (CMIP) provides a framework for GCMs to produce simulations for understanding climate change (Balaji et al. 2018; Eyring et al. 2016). As the most advanced presently available GCM ensemble, CMIP6 is expected to be frequently employed for understanding future climate impacts. As such, it is particularly important to understand the biases and performance of its component models in capturing features of drought. The resolution of each model simulation is determined by individual modeling centers, varying between 0.5° and 2.8° . Here 33 CMIP6 r1 ensembles are evaluated, which all have more than 60 years overlap with CPC observed data. To understand CMIP6 performance relative to its predecessor (Knutti and Sedláček 2013; Taylor et al. 2012), 33 CMIP5 r1 ensembles are also evaluated.

c. NA-CORDEX dynamical downscaled data

The North American Coordinated Regional Climate Downscaling Experiment (NA-CORDEX) dataset is produced using boundary conditions from global climate model simulations (CMIP5) to drive several regional climate models (RCMs) over North America. The simulations are from 1950 to 2100 with a spatial resolution of $0.22^\circ/25$ km or $0.44^\circ/50$ km (Mearns et al. 2017; Giorgi et al. 2012). Here we use the raw CORDEX data, which does not include bias correction, to ascertain the added value from dynamical downscaling relative to the original CMIP5/6 data.

d. LOCA statistical downscaling data

LOCA is a statistical downscaling technique that uses historical analogs to add fine-scale details to global climate model

simulations. The LOCA dataset includes 28 downscaled CMIP5 models from 1950 to 2005 at a resolution of 0.0625° (Pierce et al. 2014). Bias correction is applied in LOCA based on the Livneh observationally based gridded product (Livneh et al. 2015).

e. Watershed Boundary Dataset

Watershed Boundary Dataset (WBD) is a highly organized and seamless U.S. national hydrologic units dataset describing watershed boundaries across the continental United States (USGS and NRCS 2013). Generally, decisions on water management and policy are made at the watershed level. With this in mind, our evaluation regions are based on the hydrological unit two-digit level (HUC2) of the WBD. These boundaries are used to subset models' data and evaluate their performance in some typical hydrological units with different meteorological and hydrologic conditions.

f. Data preprocessing

In our present study, the four climate model datasets described above are evaluated and compared over three select hydrologic regions. In the context of simulated drought, we aim to determine if CMIP6 outperforms its precursor, CMIP5, and ascertain whether or not dynamical downscaled products perform better than their driver GCMs. Because one of our metrics (viz., fractional drought coverage) requires all compared datasets to have the same number of grid points and hydrologic boundaries, in the evaluation performed in this paper we interpolate all datasets onto a $1^\circ \times 1^\circ$ latitude–longitude grid using conservative interpolation (Zhuang et al. 2020). While a common grid is required for the employ of our drought metrics package, the actual grid employed is an input parameter. This particular grid is chosen as it represents an intermediate resolution among available products. As most metrics employed here are based on regional fields, we expect that the interpolation does not actually have a significant impact on the evaluation results for any metric except fractional drought coverage. To confirm this is the case we have conducted a sensitivity analysis to common grid resolution in section S8 in the online supplemental material. Indeed we find that if we interpolate all datasets onto a 0.22° or 0.44° grid (the resolution of the CORDEX ensemble), the evaluation produces total scores with correlation between 0.90 and 0.94 (Figs. S10, S11, S13 and S14 in the online supplemental material). Fractional drought coverage does indeed produce worse scores at higher resolution, but scores that are in essentially the same rank order as the coarser evaluation.

g. Hydrologic regions

Although our drought evaluation could be easily conducted over all hydrologic regions from WBD, in the present study we focus on three characteristically distinct regions—the California Region (CA), the Lower Mississippi Region (LM), and the New England Region (NE) (Fig. 1). These regions are chosen because of their relatively distinct climatological and topographical characteristics, as well as their potential vulnerability to drought. Specifically, the California Region is well known for its pronounced intra- and interannual climatological variability (a consequence of its Mediterranean climate), rough

FIG. 1. Hydrologic regions used in this paper.

topography, and substantial spatial heterogeneity, with hot deserts, humid coastline, rugged highlands, and a flat central valley (Fig. 1). California has often suffered from droughts and is projected to experience even worse drought conditions in the future (Ullrich et al. 2018; Williams et al. 2015). As a part of the southeastern United States, the Lower Mississippi Region includes parts of Arkansas, Kentucky, Louisiana, Mississippi, Missouri, and Tennessee (Fig. 1) and is one of the most productive and diverse agricultural regions in the United States. It accounts for a quarter of the total U.S. cotton and two-thirds of the total U.S. rice production (USDA 2012; Lower Mississippi River Conservation Committee 2014) and has a humid subtropical climate with ample rainfall and a spatially and temporally homogeneous climatology. However, it is also one of the most heavily irrigated land areas (7.1 million irrigated acres) over the CONUS (USDA 2020) and often experiences droughts that result in substantial crop loss, typically driven by lower rainfall and higher temperatures occurring simultaneously over the growing season (USDA 2020, 2012; Melillo et al. 2014; Mo 2011; U.S. Army Corps of Engineers 2013). The most well-known example of droughts in this region is the 1988/89 drought, which even led to a stoppage of barge traffic over the lower Mississippi River (Changnon 1989; Trenberth et al. 1988; Trenberth and Guillemot 1996). Finally, the New England Region is one of the most populated and developed HUC2 regions in the United States, accounting for about 20% of GDP and population with only 5% of the land area (Hobbs 2008; U.S. Bureau of Economic Analysis 2016; Fig. 1). It has a humid continental climate with abundant precipitation evenly spread throughout the year. However, there is significant spatial variation in precipitation between the inland and coastal areas, driven by its rugged topography. Although the northeastern United States (NEUS) is presently experiencing a wet period that has continued since the unprecedented 1960s drought (Seager et al. 2012), it is largely acknowledged that droughts are not things of the past for NEUS (Hayhoe et al. 2007; Krakauer et al. 2019; Seager et al. 2012). In fact, several studies conclude the risks of severe droughts for the NEUS

remain, with increased likelihood of short-term droughts under global warming (Frumhoff et al. 2007; Hayhoe et al. 2007; Krakauer et al. 2019). Overall, we anticipate the performance and biases of climate models to depend on each region's temporal and spatial variability of precipitation, as well as their local geography. Therefore, these three regions are selected to sample a spread of regional climatologies—large temporal and spatial variability over mountainous topography (CA), small temporal and spatial variability over flat and even topography (LM), and large spatial but small temporal variability over complex topography (NE).

3. Statistical methods

a. Scoring individual metrics

Standard statistical tests, including Kolmogorov–Smirnov (K–S) test and Z test (described in supplement section S1) are used to derive quantitative measures of performance for various characteristics of drought. These statistical tests also provide a means for absolute normalization of our performance metrics: our “score” for each metric is defined as the absolute ratio of each test's statistic to the critical value at 95% confidence level. As such, a score less than or equal to 1 indicates that we cannot reject the null hypothesis at this level. On the other hand, if a score is larger than 1 there is evidence that this particular metric is significantly different from observations. Given that scores are of approximately the same magnitude, this also means that scores convey comparable information across metrics. In general, for continuous metrics we will use K–S test to examine if the model data have the same distribution as observed data. For discrete variables, we will use the Z test to test if the model has the same mean value as observed data. And for those variables involving proportion or probability, we will conduct one proportion Z test to see if model and observed data have the same proportion/probability, with the assumption that the proportion or probability derived from observational data is the real value.

b. Principal feature analysis

Because of the multifaceted character of drought, multiple metrics are needed to capture all aspects of model performance. However, metrics within the same category are highly correlated, and so it is necessary to select a subset of principal metrics so as to reduce redundancy in the total score calculation. In the context of dimension reduction, principal components analysis (PCA) is commonly used (Bryant and Yarnold 1995); however, PCA yields a set of orthogonal vectors consisting of the combination of original features instead of several selected original features. Therefore, here we PFA to select a subset of metrics used to calculate the total score.

PFA is a novel feature selection method based on PCA and the k -means unsupervised clustering algorithm used to select the p original features that are most independent and can best represent the first q principal components produced by PCA (Lu et al. 2007; Song et al. 2010). To apply this technique, we first standardize all the metrics' scores and apply a PCA to see how many principal components (q) are needed to explain 95% variance. By applying the k -means algorithm with $k = p$, we cluster the first q principal components that explain 95% variance into p clusters. Then we choose the original feature that is closest to the mean of each selected cluster as the principal features. As suggested (Lu et al. 2007; Song et al. 2010), we set $p = q + 1$ because the number of features should be slightly higher than the number of principal components to explain the same variability.

c. Total score

To better compare each model's performance within the evaluation region, we also introduce a total score, which is defined as the sum of all principal metrics. The reasonably performant model will have metrics with individual scores between 0 to 2 (with 1 again corresponding to the 95% confidence level). However, sometimes particularly strong disagreement arises, and some metrics produce scores that are larger than 2. Left unattended, such large metrics would dominate the total score, and so we set an upper limit on the score of each metric of 2 in calculating the total score and define the total score as

$$\langle \text{total score} \rangle = \sum_{\text{all metrics } i} \min(\langle \text{score} \rangle_i, 2). \quad (1)$$

Under this modification, models with a significant deficiency in one metric can still achieve decent performance in the total score if they have smaller scores in the other metrics; nonetheless, care should be taken not to ignore such deficiencies as they may still be indicative of more severe issues in the model's treatment of drought. Also note that although a total score consisting of principal metrics is recommended, users can still customize the total score by self-selecting features of drought that they care about the most.

4. Drought metrics

The metrics discussed here are selected to capture a handful of meaningful quantities from data that span two dimensions in space and one in time. We expect the chosen metrics should only be expected to produce reasonable scores with at least 30

TABLE 1. SPI classification following Guttman (1999).

SPI value	Conditions
$\text{SPI} \geq 2$	Extremely wet
$2 > \text{SPI} \geq 1.5$	Very wet
$1.5 > \text{SPI} \geq 1$	Moderately wet
$1 > \text{SPI} > -1$	Nearly normal
$-1.5 < \text{SPI} \leq -1$	Moderately dry
$-2 < \text{SPI} \leq -1.5$	Very dry
$\text{SPI} \leq -2$	Extremely dry

years monthly precipitation as input to meet the requirements of SPI and central limit theorem (supplement section S7; Guttman 1999; Svoboda and Fuchs 2016; Svoboda et al. 2012). These metrics do rely on the specification of a region over which the evaluation takes place, which can (in essence) be any arbitrary shapefile of sufficient size to ensure our statistics are robust. In the text that follows we refer to this as the *evaluation region*.

Generally, drought refers to a prolonged time period with abnormally low precipitation (meteorological drought) and a shortage of water (hydrological drought), usually varies from months to years (Mishra and Singh 2010; Wilhite and Glantz 1985). In this paper we focus specifically on meteorological drought, since hydrological drought indicators such as soil moisture and runoff are less frequently available and often inconsistently defined across models. For meteorological droughts, the *standardized precipitation index* has been widely used as a drought indicator to capture the drought's beginning, end, frequency, intensity and probability at various time scales (Hayes et al. 2002; McKee et al. 1993; Svoboda and Fuchs 2016; Ukkola et al. 2018). Using the SPI classification table (Table 1), drought magnitude can be easily assessed. *Regional SPI* and *gridpoint SPI* are, respectively, calculated from regional mean precipitation and each grid point's precipitation via the Climate Indices Package (Adams 2017). Because of the normalization, dry months ($\text{SPI} \leq -1$) refer to approximately 16% coverage over the whole time period (Hayes et al. 2002; McKee et al. 1993; Yihdego et al. 2019).

Our 11 metrics have been placed into six categories associated with different characteristics of drought. These categories and metrics are discussed in the following sections.

a. Monthly means

1) MONTHLY MEAN REGIONAL PRECIPITATION (A1. MEAN PRECIP.)

As precipitation is the most relevant upstream driver of drought, it is important to determine if models capture the distribution of monthly mean precipitation correctly. To evaluate this, we apply the K-S test [Eq. (S13)] to monthly regional precipitation (viz., the monthly mean precipitation of all grids' within the evaluation region) to evaluate if the comparative distributions of model data and observational data.

2) REGIONAL SPI6 (A2. MEAN SPI6)

Regional SPI6 is the SPI6 calculated by the regional monthly mean 6 months' accumulative precipitation in the evaluation region, and is representative of the nature of intermediate term

droughts. The model performance for this metric is again scored using the K–S test [Eq. (S13)].

3) REGIONAL SPI36 (A3. MEAN SPI36)

Regional SPI36 is also evaluated analogous to SPI6, using the regional monthly mean 3-year accumulated precipitation in the evaluation region. This metric is selected to evaluate the model's performance at capturing longer-term droughts. Evaluation is also conducted by K–S test [Eq. (S13)].

b. Seasonality

1) REGIONAL PRECIPITATION AT EACH MONTH (B1. SEASON PRECIP.)

An issue with the metrics in category A is that they group all months into the same statistical sample and so cannot capture the seasonality of precipitation. However, drought seasonality determines the intra-annual distribution of drought, which is particularly important in regions of high climatic variability such as California. Here drought seasonality is assessed by applying the K–S test to each month separately and using the mean of the normalized score from each month to measure the model's performance.

2) REGIONAL LONG-TERM MONTHLY MEAN (ACCUMULATED SEASONALITY; B2. LTMM)

The long-term monthly mean, a 12-month time series consisting of average precipitation in each month of the year, is commonly used to describe the seasonality of precipitation and droughts. By normalizing the LTMM by the total precipitation over the entire year, we obtain an accumulated fractional contribution through the given month, analogous to a cumulative distribution function. This then provides another means to test seasonality through calculation of the K–S test statistic for the difference between the long term monthly mean of model and observational data. Unlike B1, this metric focuses on the accumulated seasonality, which tends to emphasize the precipitation in the middle of year (the growing season).

c. Spatial character

1) FRACTIONAL DROUGHT COVERAGE (C1. FRAC. COVER.)

Each evaluation region typically contains more than one grid point. Because of the variation in spatial characteristics of the domain, including topography and land surface type, different grid points will also tend to have differences in precipitation and susceptibility to dry conditions. Therefore, the fractional drought coverage metric is intended to evaluate if the model can simulate the spatial distribution of drought within a given region. Here, each grid point is deemed as dry when the SPI6 for that month is no larger than -1 (as Table 1). We define the ratio of dry grids to total number of grids at each month as the *fractional drought coverage*:

$$\langle \text{fractional drought coverage} \rangle = \frac{\langle \text{number of grid points with SPI6} \leq -1 \rangle}{\langle \text{total number of grid points within the evaluation region} \rangle} \quad (2)$$

The monthly dry grid ratio is then evaluated by the K–S test statistic [Eq. (S13)] to assess if model's distribution is the same as observational data.

d. Drought frequency

1) PROPORTION OF DRY MONTHS (D1. DRY FRAC.)

Using the definition of SPI (Table 1), we define a dryness indicator via

$$\langle \text{dry month indicator} \rangle = \begin{cases} 1, & \text{if regional SPI6} \leq -1 \\ 0, & \text{if regional SPI6} > -1 \end{cases} \quad (3)$$

The *proportion of dry months* is then the number of dry months divided by the total number of months:

$$\langle \text{proportion of dry months over all years} \rangle = \frac{\langle \text{number of months with regional SPI6} \leq -1 \rangle}{\langle \text{total number of months} \rangle} \quad (4)$$

To determine the score for this metric, we use the one proportion Z test [Eq. (S14)].

2) ANNUAL NUMBER OF DRY MONTHS (D2. DRY COUNT)

To assess the interannual variability of drought frequency, we define the annual number of dry months as the number of dry months [Eq. (3)] from each year. Because all models have more than 60 years of data, and so meet the criteria of the central limit theorem, we calculate the Z-test statistic [Eq. (S15)] to evaluate and normalize this metric.

e. Drought intensity

1) INTENSITY FROM SPI6 (E1. INTENSITY)

The drought intensity is defined as the regional SPI6 values over all months when SPI6 is no larger than -1 . To test if model can produce the same drought intensity distribution, we employ the K–S test statistic [Eq. (S13)] to obtain the normalized score for this metric.

f. Probability of drought

1) PROBABILITY OF DROUGHT INITIATION (AND AVERAGE NONDRY PERIOD DURATION; F1. PROB. INIT.)

The probability of drought initiation is defined as the probability that the following month is dry given the current month is not dry. This is connected to the average nondry period duration via

$$\begin{aligned} \langle \text{probability of drought initiation} \rangle &= P(\langle \text{dry month} \rangle | \langle \text{previous month is nondry} \rangle) \\ &= \frac{\langle \text{number of dry months when the last month is nondry} \rangle}{\langle \text{total number of months when previous month is nondry} \rangle} \\ &= \langle \text{average nondry period duration in months} \rangle^{-1}. \end{aligned} \quad (5)$$

This connection between drought initiation and the duration of the nondry period emerges naturally if one treats each month

as a sample from a geometric distribution. As is typically done with proportions, to normalize the metric we apply the one proportion Z test to the nondrought duration data.

2) PROBABILITY OF DROUGHT TERMINATION (AND AVERAGE DRY PERIOD DURATION; F2. PROB. TERM.)

We define the probability of drought termination as the probability that the following month is not dry given the current month is dry:

$$\begin{aligned} &\langle \text{probability of drought termination} \rangle \\ &= P(\langle \text{nondry month} \rangle | \langle \text{previous month is dry} \rangle) \\ &= \frac{\langle \text{number of nondry months when the last month is dry} \rangle}{\langle \text{total number of months when previous month is dry} \rangle} \\ &= \langle \text{average drought duration in months} \rangle^{-1}. \end{aligned} \quad (6)$$

This equation is analogous to the probability of drought initiation, again emerging naturally from the geometric distribution. The metrics is also used to assess models' ability in capturing average drought duration. As with probability of drought initiation, we use the one proportion Z test to normalize the model's score.

Notably, the two metrics F1 and F2 are in essence derived from the transition probabilities of a two-state Markov chain, with states corresponding to nondry and dry conditions.

g. Taylor diagram and Taylor score

In addition to the statistical metrics described above, we also compare our evaluation system with a more traditional evaluation method. The Taylor diagram is a widely used mathematical plot in Earth model assessments that depicts correspondence between the model and observational data using a mathematical property that relates the standard deviation, the Pearson correlation coefficient, and the root-mean-square error (RMSE) (Taylor 2001). One way to quantify models' performance on the Taylor diagram is by the Taylor score,

$$\langle \text{Taylor score} \rangle = \exp \left[-\alpha(1 - R) - \beta \left(\frac{\sigma + 1}{\sigma - 2} \right) \right], \quad (7)$$

where α and β are scaling factors that are set to 1 (the default value) in this paper, R is the Pearson correlation between each grid point's mean precipitation, and σ is the model's normalized standard deviation of each grid point's mean precipitation (normalization is conducted by dividing this by the standard deviation from the observational data). Notably, the Taylor score uses the mean precipitation to measure performance but ignores temporal continuity (as we will see later). Unlike our statistical scores above, a higher Taylor score indicates better performance. In this paper, we rank all models from the best to the worst in the Taylor diagram and compare the results with our evaluation system.

5. Results

a. Principal metrics employed

To maintain consistency among study regions, we apply PFA to the CMIP6 dataset over the entirety of the continental United States to select a set of principal metrics to apply

analysis for all case evaluations. These metrics and their total scores will then be our primary means to evaluate each dataset's performance.

PCA shows six principal components are needed to explain 95% variance (Fig. S1). To match this level of variance, the PFA procedure gives a set of seven principal metrics; as apparent from Fig. 2, most principal metrics, have less than 0.6 correlations with each other which confirms the reliability of PFA to select distinguish and significant metrics. Note that in Fig. 2, metrics within one category are highly correlated partially because it is over CONUS. And within each evaluation region (like the California Region), metrics in one category are less correlated as shown by Fig. S2.

PFA's identification of at least one metric from each category also confirms our natural intuition on the independence of these metrics. In the following sections, we will evaluate the performance of CMIP6, CORDEX, and LOCA datasets over the California, the Lower Mississippi, and the New England Regions based on these seven principal metrics and their total score. We conclude a model "performs well" when its total score is less than the number of principal metrics, with a lower total score generally indicative of better performance. In the resulting heat maps, individual metric scores less than 1 and total scores less than 7 are indicated in black font (and white font otherwise).

b. The validation of our evaluation system

To examine if our evaluation system can really distinguish the high-performance climate models, here we evaluate the ECMWF Reanalysis v5 (ERA5) monthly precipitation (ECMWF 2016), which is the most advanced fifth generation ECMWF atmospheric reanalysis dataset, and processes numerous improvements and advances (Hoffmann et al. 2019; Hersbach 2016; Hersbach et al. 2020). From Fig. 3, it is obvious that the highly performant reanalysis data are also identified to be excellent to capture droughts' features over most CONUS hydrologic regions (14 out 18). Even if for those regions where ERA5 has total scores larger than 7, the total scores are still relatively small (ranging from 7.07 to 8.32). This confirms that our system works well in capturing the drought features. Also, for these four hydrologic regions, most CMIP6 models (32/33 for Great Basin Region and Upper Colorado Region, 30/33 for Souris–Red–Rainy Region, and 24/33 for Great Lakes Region) cannot get good total scores (less than 7) too, which might be caused by the biases in CPC observed data, the deficiency to simulate certain atmospheric circulation in climate models over these regions. More studies are needed to explore the reasons why such a huge systematic discrepancy exists between CPC observed data and most climate models and reanalysis data.

c. CMIP6 performance

Our first application of the metrics package focuses on CMIP6 model performance.

1) CMIP6 PERFORMANCE OVER THE CALIFORNIA REGION

Performance for the CMIP6 suite of models in the California Region is tabulated in Fig. 4. From the principal metrics alone,

FIG. 2. Cross-correlation heat map of all metrics calculated over the continental United States from CMIP6 models (bold metrics are principal metrics selected by principal feature analysis).

24 out of 33 CMIP6 models produce a total score less than 7, implying that more than a half CMIP6 models perform well within this region. HadGEM3-GC31-LL is the best model, with the lowest total score and with only one principal metric slightly larger than 1. CESM2-WACCM is also an impressive model as it ranks 4 out of 33 and has all principal metrics less than 1.

Although overall CMIP6 performance is good, most models have a fractional drought coverage score larger than 1. There are two main reasons why CMIP6 models tend to perform so poorly on this metric over the California. First, the California Region is a region of significant spatial variability due to its complex mountainous topography with steep altitude gradient (particularly in the east of the state) which makes it hard to capture the spatial coverage of drought. As we can see from Fig. 5g, CMIP6 tends to obtain much better performance at C1. Frac. Cover. over low-lying regions without complex topography significantly (with p value less than 0.03). Second, compared with the resolution of the CPC observations, CMIP6 models have much coarser spatial resolution that makes it particularly difficult to simulate the finer-scale geographic features present in the CPC data. Examining CMIP6 models' original resolution [defined by $\sqrt{\text{dlon} \times \text{dlat}}$ where dlon is the model's longitudinal grid spacing and dlat is the latitudinal grid spacing] versus fractional drought coverage scores, a weak but positive correlation emerges ranging from 0.14 to 0.28 (with p value 0.12–0.44) in all three evaluation regions (Figs. 5a,c,e). This indicates the models with finer resolution tend to better capture the fractional drought coverage and

agrees with our hypothesis that coarse resolution is a limiting factor behind why CMIP6 cannot simulate drought coverage well. Moreover, nearly all the CMIP6 model have scores less than 1 for their proportion of dry months, probability of drought initiation and termination. One potential reason why CMIP6 models tend to simulate these temporally related metrics well is because the significant intra-annual and interannual climatological variability in California leads to a large normalization factor in the Z test. Also the region's overall dryness and strong seasonal contrast makes it easy for models to correctly capture the probability of drought's duration, initiation and termination. Therefore, CMIP6 models are generally trustworthy when it comes to simulating these drought features in the California Region, as opposed to spatial metrics like fractional drought coverage.

If we compare our results with the Taylor diagram which is based on each grid point's mean precipitation (Fig. 6a), we can see there exists obvious differences between conclusions obtained from the Taylor diagram and our system's assessment. For example, BCC-ESM1 has a total score of 5.16 and is ranked as 5 out of 33 in our system; however, it is assessed as a low-performing model with rank 27 out of 33 in the Taylor diagram. This is mainly due to the low spatial standard deviation of each grid point's mean precipitation, which is captured by the precipitation mean magnitude metrics in our framework (A1. Mean Precip., B1. Season Precip., and D1. Dry Frac.). In Fig. 4, although BCC-ESM1 exhibits poor performance for two mean magnitude metrics (A1. Mean Precip., B1. Season Precip.) and further shows poor performance at C1. Frac. Cover., it shows

FIG. 3. Principal metrics and total score of ERA5 reanalysis data over all 18 hydrologic regions over CONUS [the performance is identified as good when individual metric scores less than 1, and total scores less than 7 are indicated in black font (and white font otherwise)].

better performance at the temporally related metrics (especially F1. Prob. Init. and F2. Prob. Term.). This difference in scoring also confirms the value of our system, which can assess models' performance on drought's temporal features. On the other hand, some models like GISS-E2-1-H perform the worst in terms of both the Taylor diagram (Fig. 6a) and under our metrics (Fig. 4). The Taylor diagram metrics, namely, spatial standard deviation (precipitation mean magnitude) and spatial correlation of mean precipitation at each grid point, are not represented in our system and so it makes sense that there is a not strong correlation (-0.392) between our total score and the Taylor score (Fig. 5b).

2) CMIP6 PERFORMANCE OVER THE LOWER MISSISSIPPI REGION

Performance of the CMIP6 suite of models in the Lower Mississippi Region is tabulated in Fig. 7. As we mentioned, this

region is the flattest one over CONUS with small spatial variability of precipitation, thus it is perhaps not surprising that the CMIP6 models capture fractional drought coverage much better over this relatively flat region than over California; 16 out of 33 models have a score less than 1 at fractional drought coverage in this region, which is the best among all the hydrologic regions of the continental United States. On the other hand, the Lower Mississippi Region has the most intense rainfall within the continental United States, with much less climatological variability. Such a smooth precipitation seasonality means it is difficult for CMIP6 to simulate drought frequency, duration, and probability well. Thus most CMIP6 models exhibit poor performance at D1. Dry Frac., F1. Prob. Init., and F2. Prob. Term. The poor performance among these metrics thus results in slightly worse performance across the CMIP6 ensemble as

FIG. 4. Principal metrics and total score of CMIP6 datasets over the California Region [the performance is identified as good when individual metric scores less than 1, and total scores less than 7 are indicated in black font (and white font otherwise)].

FIG. 5. Scatterplots where correlations appear between model's and region's characteristics and evaluation results. (a) Each CMIP6 model's C1. Frac. Cover. score vs its original resolution in the California Region. (b) Each CMIP6 model's total score vs Taylor score in the California Region. (c) Each CMIP6 model's C1. Frac. Cover. vs its original resolution in the Lower Mississippi Region. (d) Each CMIP6 model's total score vs Taylor score in the Lower Mississippi Region. (e) Each CMIP6 model's C1. Frac. Cover. score vs its original resolution in the New England Region. (f) Each CMIP6 model's total score vs Taylor score in the New England Region. (g) The average scores of all CMIP6 models' C1. Frac. Cover. scores vs evaluation region's elevation mean over all 18 CONUS hydrologic regions.

FIG. 6. Taylor diagram of CMIP6 datasets over three hydrologic regions, which is based on each grid point's monthly mean precipitation. Models are labeled starting from 1 based on their Taylor scores from the highest to the lowest; so the index 1 corresponds to the model with highest Taylor score and best performance identified by Taylor diagram. (a) Over the California Region, (b) over the Lower Mississippi Region, and (c) over the New England Region.

FIG. 7. Principal metrics and total score of CMIP6 datasets over the Lower Mississippi Region (the performance is identified as good when individual metric scores less than 1, and total scores less than 7 are indicated in black font [and white font otherwise]).

compared with the California Region, with 15 out of 33 models have total scores less than 7. Comparing our total score to the Taylor diagram for this region, there still only exists a weak correlation (-0.418) between our total score and the Taylor score (Fig. 5d).

3) CMIP6 PERFORMANCE OVER THE NEW ENGLAND REGION

Performance for the CMIP6 suite of models in the New England Region is tabulated in Fig. 8. This weak temporal variability and precipitation seasonality leads to poor performance for temporal features in the CMIP6 models compared with the California Region (i.e., E1. Intensity and F2. Prob. Term.). The large spatial variability of precipitation explains why CMIP6 models tend to perform poorly with respect to spatial coverage (C1. Frac. Cover.) as well. As a result, only 6 out of 33 models have total scores less than 7. Again, there is only a weak correlation (-0.225) between our evaluation results and the Taylor score (Fig. 5f). We note that the poor performance for drought intensity and fractional drought coverage may be affected by poor representation of tropical cyclones in GCMs due their coarse resolution (Henderson-Sellers et al. 1998). Metrics to evaluate the representation of tropical cyclones and their associated precipitation in model data remain under development (Stansfield et al. 2020; Zarzycki et al. 2021), and can be employed to understand specific model deficiencies in this regard.

d. CMIP5 performance

To quantify any improvements in CMIP6 in capturing the drought, we evaluate the same number (33) of CMIP5 models over all regions. As depicted in Fig. S3, 19 out of the 33 CMIP5 models perform well over the California Region, which is less than CMIP6 (with 24 out of 33). In general, the improvements in CMIP6 are present in almost all principal metrics and total scores, with the multimodel average total score decreasing from 7.20 to 6.66. These improvements are perhaps not surprising given CMIP6 represents a later generation of models (Ahn et al. 2020; Eyring et al. 2016; Li et al. 2020; Priestley et al. 2020). However, others have shown only relatively modest improvement between CMIP5 and CMIP6 (Srivastava et al. 2020). The same process was also applied to the Lower Mississippi Region (Fig. S4) and the New England Region (Fig. S5), with the ratio of performant models increasing from 12/33 to 15/33 in LM but staying steady at 6/33 in NE. However, the means of total scores decreased from 8.08 to 7.59 over the LM Region and from 9.25 to 8.76 over the NE Region. In fact, such improvements are present in nearly all hydrologic regions, suggesting that CMIP6 is the preferred ensemble for study of drought features.

e. CORDEX performance

We now turn our attention to the CORDEX dataset, consisting of downscaled simulations from several regional climate models. In particular, we focus on using our evaluation system to understand whether or not dynamical downscaling improves the quality of simulated drought compared its driving model

(CMIP5) and how these simulations compare to the CMIP6 ensemble.

1) CORDEX PERFORMANCE OVER THE CALIFORNIA REGION

CORDEX performance is evaluated with the same principal metrics mentioned above (Fig. 9). The most obvious finding is that CORDEX achieves much better performance on fractional drought coverage within the California Region compared with CMIP6. There are 26/31 CORDEX models that perform well at C1. Frac. Cover. (compared with CMIP6's 7 out of 33 and CMIP5's 5 out of 33). This is perhaps unsurprising as the low CMIP5/6 model resolution is one main causes of poor performance when it comes to spatial coverage of drought. Namely, the higher model resolution improves how the model captures the spatial variability of precipitation over complex topography. Evidence for this result also emerges from CORDEX products with different resolutions. For example, MPI-M-MPI-ESM-MR.UQAM-CRCM5.22 is the downscaling product based on MPI-M-MPI-ESM-MR, downscaled by UQAM-CRCM5 with a spatial resolution of 0.22° . Its performance when it comes to fractional drought coverage (0.25) is better than MPI-M-MPI-ESM-MR.UQAM-CRCM5.44 (0.47), which uses the same two models but with a coarser resolution. Nonetheless, 3 products do not follow this trend (CCCma-CanESM2.UQAM-CRCM5, UQAM-GEMatm-Can-ESMsea.UQAM-CRCM5 and HadGEM2-ES.NCAR-WRF), because we interpolate all CORDEX models onto the same 1° grid which masks the benefits of finer resolution products in capturing spatial features. In the sensitivity analysis with different resolutions, if we interpolate all models into a 0.22° or 0.44° grid (the original grid used in CORDEX), finer products always get better performance in fractional drought coverage than their coarser analogs (see detailed information in supplement section S8).

At first glance, there appears to be clear added benefit in using higher resolution products for assessing the spatial character of drought/precipitation. However, we also find that for the other principal metrics, CORDEX does not provide improvement compared with CMIP5/6—even yielding worse performance for average drought intensity. For example, the mean average drought intensity score from the CORDEX dataset is 2.20, which is larger than CMIP5's (1.80) and CMIP6's (1.48). Further, while the difference in overall performance is much smaller, CMIP6 still tends to produce the best overall performance with an average total score of 6.66, beating CMIP5 (7.20) and CORDEX (6.83). Therefore, it seems that dynamical downscaling does not provide significant benefit in representation of drought features (and may even degrade some features' representation). The exception is in the fractional drought coverage, which does show some benefit from finer model resolution (supplement section S8). Nonetheless the degradation of performance is perhaps unsurprising, as droughts are primarily a product of synoptic-scale meteorological drivers which are largely prescribed by the driving model. That is, this appears to be a combination of “garbage in, garbage out” and systematic biases from regional downscaling at play (Giorgi and Mearns 1999; Hall 2014; Pontoppidan et al. 2018; Rummukainen 2010; Switanek et al. 2017). Notably, RCMs are known to

FIG. 8. Principal metrics and total score of CMIP6 datasets over the New England Region [the performance is identified as good when individual metric scores less than 1, and total scores less than 7 are indicated in black font (and white font otherwise)].

FIG. 9. Principal metrics and total score of CORDEX datasets and CMIP5 drivers over the California Region [the performance is identified as good when individual metric scores less than 1, and total scores less than 7 are indicated in black font (and white font otherwise)].

possess significant biases when it comes to precipitation over mountainous topography (Caldwell et al. 2009; Maraun and Widmann 2015); consequently we will see later that CORDEX performance is much better over homogeneous and flat regions (e.g., the Lower Mississippi Region). Our sensitivity analysis using a common resolution grid at 0.44° and 0.22° resolution further supports these findings (supplement section S8).

We do acknowledge that the mediocre overall performance of CORDEX compared with CMIP5 over the California Region may also be due to selection bias. That is, only a small subset of the 33 CMIP5 models were chosen as CORDEX drivers, and these may not be representative of the larger ensemble. However, according to Mearns et al. (2017), the CMIP5 drivers used in CORDEX were chosen because they were highly performant. To better understand if this is the case, we selected the 6 CMIP5 models used as drivers in CORDEX and tabulated their performance in Fig. 9. Compared to their CMIP5 drivers we find that CORDEX products all have better performance at fractional drought coverage. However, in most cases, CORDEX is actually worse for the other principal metrics and total scores. For example, the average total score of the 6 CMIP5 drivers is 6.05, far smaller than the CORDEX's (6.83) and the whole set of 33 CMIP5 models' (7.20). This verifies that CORDEX CMIP5 drivers are performant models and supports our claim that dynamical downscaling does not always improve the quality of simulated drought in these models (Giorgi and Mearns 1999; Hall 2014; Pontoppidan et al. 2018; Rummukainen 2010; Switanek et al. 2017). Additionally, these results also appear to confirm that systematic biases from dynamical downscaling do degrade the fidelity of simulated drought in regions of complex terrain (Caldwell et al. 2009; Maraun and Widmann 2015).

Using the Taylor diagram and score to evaluate CORDEX performance (Fig. 10a) versus its CMIP5 drivers' performance (Fig. 10a), we find analogous results to our evaluation. For example, CORDEX models achieve better performance for spatial metrics (i.e., spatial correlation) than their CMIP5 drivers as a result of their finer resolution. However, the mean magnitude metrics (e.g., mean precipitation standard deviation) is not obviously improved and even degraded due to systematic biases in the RCMs.

2) CORDEX PERFORMANCE OVER THE LOWER MISSISSIPPI REGION

As in the California Region, CORDEX simulations in the Lower Mississippi also provide significant improvement in the fractional drought coverage; however, unlike California, here CORDEX also shows improvement for other metrics (Fig. 11). In this flat region the systematic biases from the RCMs are smaller than the improvements that come from higher resolution, which also follows previous studies that systematic biases in regional precipitation simulations are much larger over complex topography (Caldwell et al. 2009; Maraun and Widmann 2015). In total, 21 out of 31 CORDEX models perform well here with a total score less than 7, compared with 15 out of 33 CMIP6 models and 12 out of 33 CMIP5 models. The improvement from the CORDEX models also holds against their CMIP5 drivers (Fig. 11); except HadGEM2-ES, all other CORDEX models perform better than their associated CMIP5 drivers. As

CORDEX improves on the CMIP models for essentially all of our metrics, it is perhaps no surprise that these models also show improvements using the Taylor diagram and score (Fig. 10b) compared with CMIP6's (Fig. 6b) and their CMIP5 drivers (Fig. 10b).

3) CORDEX PERFORMANCE OVER THE NEW ENGLAND REGION

The New England Region has high spatial precipitation variability and complex terrain (mountains and coasts). In line with the other two regions analyzed, we expect CORDEX to have better performance on fractional drought coverage but worse performance for other metrics (Fig. 12). In general, CORDEX produces worse performance for nearly all other metrics, with only 4 out of 31 CORDEX models having a total score less than 7 (compared with 6 out of 33 CMIP5 and CMIP6 models). But perhaps surprisingly, even for the fractional drought coverage, CORDEX does not exhibit any obvious improvement. This is likely because the New England Region has a more complex topography and precipitation distribution, as well as coastal effects contributing to precipitation variability (Agel et al. 2015). In the Taylor diagram (Fig. 10c) we also see although CORDEX is evaluated to be better than its CMIP5 drivers (Fig. 10c), it still has clearly worse performance compared with CMIP6 (Fig. 6c); CORDEX models generally have larger standard deviation and spatial correlation.

f. LOCA performance

Although LOCA is also a downscaled product of CMIP5, LOCA has much finer resolutions (1/16°) compared with CORDEX, and uses a statistical downscaling method based on local analogs combined with bias correction (Pierce et al. 2014). As such it is not surprising that LOCA's climatology tends to match the historical climate better than direct model simulations—nonetheless, this does not necessarily mean that LOCA does a better job of capturing future change as it relies on a functional relationship between coarse GCM and regional climate data that may not hold in the future (i.e., the stationarity problem). Further, unlike dynamical downscaling, this method does not attempt to preserve relationships among different variables as variables are corrected independently (Prasanna 2018).

Over the California Region (Fig. 13), LOCA achieves much better performance than both CMIP6 and CORDEX in not only the fractional drought coverage but across all principal metrics. All LOCA models actually produce total scores below 7, and the best LOCA model (IPSL-CM5A-LR with 2.70 total score) also outperforms the best models from CORDEX (CCCma-CanESM2.UQAM-CRCM5.22 with 4.67 total score), CMIP5 (MPI-ESM-MR with 4.35), and CMIP6 (HadGEM3-GC31-LL with 4.07 total score). There is little doubt that LOCA captures the historical drought climatology better than these other products. Of course, its performance is strongly related to the use of bias correction Pierce et al. (2014), finer resolution and avoidance of biases from direct simulation of physical processes.

From Fig. 14a, it is apparent that when evaluated with the Taylor diagram all models are compressed to the same point and it is impossible to distinguish performance. This problem

FIG. 10. Taylor diagram of CORDEX datasets and their drivers over the three hydrologic regions (CMIP5 drivers are bolder). (a) Over the California Region, (b) over the Lower Mississippi Region, and (c) over the New England Region.

persists in other evaluation regions as well, and arises because LOCA uses statistical downscaling and correction based on the observational data's mean fields of precipitation; however, the Taylor diagram only evaluates precipitation mean fields. Therefore, a Taylor diagram cannot be used for distinguishing

LOCA products—the results are effectively identical to comparing CPC with Livneh.

Results from LOCA are similar in the Lower Mississippi Region (Figs. S7 and 14b) and the New England Region (Figs. S8 and 14c).

FIG. 11. Principal metrics and total score of CORDEX datasets and CMIP5 drivers over the Lower Mississippi Region [the performance is identified as good when individual metric scores less than 1, and total scores less than 7 are indicated in black font (and white font otherwise)].

FIG. 12. Principal metrics and total score of datasets and CMIP5 drivers over the New England Region [the performance is identified as good when individual metric scores less than 1, and total scores less than 7 are indicated in black font (and white font otherwise)].

FIG. 13. Principal metrics and total score of LOCA datasets over the California Region [the performance is identified as good when individual metric scores less than 1, and total scores less than 7 are indicated in black font (and white font otherwise)].

FIG. 14. Taylor diagram of LOCA datasets over three hydrologic regions. (a) Over the California Region, (b) over the Lower Mississippi Region, and (c) over the New England Region.

6. Conclusions

To help with selecting datasets that are more likely to capture the features and process drivers of intermediate-term drought, we have put forward a comprehensive evaluation framework that can be easily used within an arbitrary evaluation region, given more than 30 years monthly precipitation data. In total, 11 metrics have been defined encapsulating drought's monthly mean, seasonality, spatial coverage, frequency, intensity, duration, and probability. Principal feature analysis (PFA) is suggested to select principal metrics which are significant and relatively independent; however, users can customize the metrics used to meet their needs. A framework based on statistical hypothesis testing is used to normalize each metric and determine model performance. By employing this framework to four climate datasets—CMIP5/6, CORDEX, and LOCA over three characteristically distinct hydrologic regions in CONUS—we have developed a better understanding of the sensitivity of model performance to region and dataset. This enables us to provide guidance to stakeholders and data users on the selection of suitable models, and furthers our understanding of how various factors impact model biases in capturing the drought features.

Our evaluation system complements evaluation using ETCCDI drought indices and spatial analysis of precipitation using the Taylor diagram and score. We evaluate a variety of drought characteristics that depend only on monthly precipitation fields, including temporal characteristics such as duration and occurrence/termination probability. We argue that these characteristics are highly relevant for climate adaptation planning, and so should be included in any evaluation system. Notably, the total scores from our proposed system are only weakly correlated with evaluation using Taylor scores, with correlations between -0.12 and -0.42 across the three evaluation regions.

The system proposed here is a unique combination of statistical hypothesis testing, drought indices, and a comprehensive multimetric framework addressing various essential characters of drought. It has several significant advantages:

- 1) The system can evaluate models' performance on temporal continuity and probabilistic features of drought which are often ignored in traditional evaluation. It can also be used on arbitrarily large or small regions (data permitting), accounting for the fact that drought is often widespread.
- 2) By using statistical hypothesis testing, the system builds a standard criterion to assess absolute performance of models for each metric. Consequently, we can easily determine if one model produces drought that is similar to observational data and compare models' performance across multiple categories of models.
- 3) The system provides a method to analyze the performance of statistical downscaling products (like LOCA) while a traditional method like a Taylor score shows essentially identical performance. This is because traditional methods usually only evaluate precipitation means which are modified by the bias correction process.
- 4) The system is flexible depending on user need. Users can either use the principal metrics selected by PFA, which can explain most variance of all metrics, or customize the

metrics used according to which drought features they need. And it is easy to conduct evaluations based on whatever regions (shapefile) or datasets (NetCDF) the users are interested in.

The major conclusions of this study are as follows:

- We have confirmed our system indeed assesses different drought features, and that the metrics selected via PFA represent distinct measures of model performance. Using ERA5 data we have validated that it can successfully identify that a high quality reanalysis product (ERA5) captures drought features well over most hydrologic regions (14 out of 18). Among the four remaining regions, the total scores are still reasonable (from 7.07 to 8.32).
- Evaluations of CMIP5/6 and CORDEX confirm our hypothesis that the performance of climate models in capturing regional drought features depends highly on the characteristics of the evaluation regions. For example, 32 out of 33 CMIP6 models, and even the most advanced reanalysis dataset (ERA5) cannot capture drought features well over the Great Basin Region and Upper Colorado Region (i.e., they get a total score no larger than 7).
- Climate models do an excellent job in simulating the temporal continuity features like drought duration and probability over regions with a clear precipitation seasonality and strong precipitation variability like the California Region, but performance on temporal features is poor over regions with weak temporal variability. Certainly, it is intuitive that initiation and termination of drought is easier to simulate if precipitation follows a strong seasonality. Also, the significant intra-annual and interannual climatological variability can lead to a large normalization factor in the statistical tests employed. On the other hand, CMIP5/6 models generally better simulate the spatial features of droughts over flat regions with less spatial variability since their typically coarse resolution (between 0.5° and 2.8° for CMIP6) prevents them from capturing the finer spatial features represented in the CPC observational data (0.25°) over the regions with complex topography and significant altitude gradients.
- Following recent advancements in climate modeling systems (Balaji et al. 2018; Eyring et al. 2016), we also find CMIP6 outperforms its precursor (CMIP5) in all regions and in nearly all metrics employed. EC-Earth3 demonstrates particularly high performance and is identified as the best CMIP6 model, with strong performance in all three evaluation regions. Overall, even compared with CMIP5-based dynamical downscaling products (CORDEX), CMIP6 still produced comparable or better performance in temporal continuity of droughts. With that said, the coarser resolution of CMIP6 models produces worse performance in the fractional drought coverage.
- For the uncorrected CORDEX runs we find that, compared with CMIP5/6, both our evaluation system and the Taylor diagram indicates that dynamical downscaling significantly improves models' performance on spatial metrics (i.e., spatial correlation in Taylor diagram and fractional drought coverage). However, CORDEX models do not always exhibit better performance on magnitude mean and temporal

continuity metrics. And the degradation of performance mainly exists over regions with complex topography and large latitude gradient. Given that droughts are largely driven by synoptic meteorology, this problem appears to be, at least in part, because of the “garbage in, garbage out” problem (Hall 2014). CORDEX models often produce worse performance on these metrics in regions with complex topography and precipitation inhomogeneity due to systematic errors common in dynamical downscaling systems over mountainous and coastal regions (Caldwell et al. 2009; Maraun and Widmann 2015; Velasquez et al. 2020). But in the flattest hydrologic region examined—namely, the Lower Mississippi Region—the RCMs consistently exhibited overall improvements in their representation of droughts.

- The LOCA dataset produces far lower scores (better performance) across essentially all metrics. This is unsurprising as bias correction has been employed to match the historical climatology. However, care still needs to be taken in the employ of LOCA for future projections, as the statistical relationships employed in LOCA may not hold in the future and LOCA does not attempt to preserve physical relationship among different variables, as variables are corrected independently (Prasanna 2018).
- By examining the sensitivity of our results to different evaluation periods and common grid spacing (supplement sections S7 and S8), we find our conclusions are largely independent of these choices. As long as the study period is larger than 30 years (the minimum requirement of the central limit theorem and SPI), modifying the study period from 1948–2014 to 1970–1999 and 1970–2009 yields total scores that are highly correlated (0.88 and 0.93, supplement section S7). Further, when interpolating CMIP6 and CORDEX to finer resolutions (0.22° or 0.44°) the total score correlations are similarly strong (0.90–0.94, supplement section S8).

Notably, the evaluation system proposed here only takes monthly precipitation as input data, which may not account for the impacts of temperature and evapotranspiration on drought; therefore, future work will focus on expanding the suite of available metrics, and examining how this system can be employed for long-term (multiyear) droughts.

The evaluation system proposed in this paper is available as an open source software package implemented in Python (https://github.com/Zeyu88/Drought_Metrics). Users can easily conduct evaluations by inputting monthly precipitation and observational data fields (NetCDF files) along with the precipitation variable name, evaluation regions (shapefiles), principal metrics to be calculated and the evaluation period. We recommended PFA be performed first over all evaluation regions to select the principal metrics used. This package also provides a function to draw Taylor diagrams for comparison. Prior to evaluation, both model and observational data must be interpolated to the same resolution (provided by our package).

Acknowledgments. This research was supported by the RGMA program area(s) in the U.S. Department of Energy’s Office of Biological and Environmental Research as part of the

multi-program, collaborative Integrated Coastal Modeling (ICoM) project. PAU was supported by Department of Energy Office of Science award number DE-SC0016605, “A Framework for Improving Analysis and Modeling of Earth System and Intersectoral Dynamics at Regional Scales (HyperFACETS).” This project is also supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, hatch project under California Agricultural Experiment Station project accession 1016611. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF. The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript.

REFERENCES

- Adams, J., 2017: Climate-indices: An open source python library providing reference implementations of commonly used climate indices. Github, <https://github.com/monocongo/climate-indices>.
- Agel, L., M. Barlow, J.-H. Qian, F. Colby, E. Douglas, and T. Eichler, 2015: Climatology of daily precipitation and extreme precipitation events in the northeast United States. *J. Hydrometeor.*, **16**, 2537–2557, <https://doi.org/10.1175/JHM-D-14-0147.1>.
- Ahn, M.-S., and Coauthors, 2020: MJO propagation across the Maritime Continent: Are CMIP6 models better than CMIP5 models? *Geophys. Res. Lett.*, **47**, e2020GL087250, <https://doi.org/10.1029/2020GL087250>.
- Andreadis, K. M., and D. P. Lettenmaier, 2006: Trends in 20th century drought over the continental United States. *Geophys. Res. Lett.*, **33**, L10403, <https://doi.org/10.1029/2006GL025711>.
- Balaji, V., and Coauthors, 2018: Requirements for a global data infrastructure in support of CMIP6. *Geosci. Model Dev.*, **11**, 3659–3680, <https://doi.org/10.5194/gmd-11-3659-2018>.
- Bhuyan, U., C. Zang, and A. Menzel, 2017: Different responses of multispecies tree ring growth to various drought indices across Europe. *Dendrochronologia*, **44**, 1–8, <https://doi.org/10.1016/j.dendro.2017.02.002>.
- Bryant, F. B., and P. R. Yarnold, 1995: Principal-components analysis and exploratory and confirmatory factor analysis. *Reading and Understanding Multivariate Statistics*, L. G. Grimm and P. R. Yarnold, Eds., American Psychological Association, 99–136.
- Caldwell, P., H.-N. S. Chin, D. C. Bader, and G. Bala, 2009: Evaluation of a WRF dynamical downscaling simulation over California. *Climatic Change*, **95**, 499–521, <https://doi.org/10.1007/s10584-009-9583-5>.
- Changnon, S. A., 1989: The 1988 drought, barges, and diversion. *Bull. Amer. Meteor. Soc.*, **70**, 1092–1104, [https://doi.org/10.1175/1520-0477\(1989\)070<1092:TDBAD>2.0.CO;2](https://doi.org/10.1175/1520-0477(1989)070<1092:TDBAD>2.0.CO;2).
- Chen, H., and J. Sun, 2015: Assessing model performance of climate extremes in China: An intercomparison between CMIP5 and CMIP3. *Climatic Change*, **129**, 197–211, <https://doi.org/10.1007/s10584-014-1319-5>.

- Chen, M., W. Shi, P. Xie, V. B. Silva, V. E. Kousky, R. Wayne Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.*, **113**, D04110, <https://doi.org/10.1029/2007JD009132>.
- Collier, N., F. M. Hoffman, D. M. Lawrence, G. Keppel-Aleks, C. D. Koven, W. J. Riley, M. Mu, and J. T. Randerson, 2018: The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. *J. Adv. Model. Earth Syst.*, **10**, 2731–2754, <https://doi.org/10.1029/2018MS001354>.
- Dai, A., 2013: Increasing drought under global warming in observations and models. *Nat. Climate Change*, **3**, 52–58, <https://doi.org/10.1038/nclimate1633>.
- Darand, M., and M. M. Sohrabi, 2018: Identifying drought-and flood-prone areas based on significant changes in daily precipitation over Iran. *Nat. Hazards*, **90**, 1427–1446, <https://doi.org/10.1007/s11069-017-3107-9>.
- Deser, C., 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>.
- ECMWF, 2016: The fifth generation ECMWF atmospheric reanalysis dataset. ECMWF, Reading, United Kingdom, <https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5>.
- Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, <https://doi.org/10.5194/gmd-9-1937-2016>.
- , and Coauthors, 2019: Taking climate model evaluation to the next level. *Nat. Climate Change*, **9**, 102–110, <https://doi.org/10.1038/s41558-018-0355-y>.
- Frumhoff, P. C., J. J. McCarthy, J. M. Melillo, S. C. Moser, and D. J. Wuebbles, 2007: Confronting climate change in the U.S. Northeast: Science, impacts, and solutions. Union of Concerned Scientists, 146 pp., <https://www.ucsusa.org/sites/default/files/2019-09/confronting-climate-change-in-the-u-s-northeast.pdf>.
- Giorgi, F., and L. O. Mearns, 1999: Introduction to special section: Regional climate modeling revisited. *J. Geophys. Res.*, **104**, 6335–6352, <https://doi.org/10.1029/98JD02072>.
- , and Coauthors, 2012: RegCM4: Model description and preliminary tests over multiple CORDEX domains. *Climate Res.*, **52**, 7–29, <https://doi.org/10.3354/cr01018>.
- Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, <https://doi.org/10.1029/2007JD008972>.
- , P. J. Durack, R. J. Stouffer, G. C. Johnson, and C. E. Forest, 2016: Industrial-era global ocean heat uptake doubles in recent decades. *Nat. Climate Change*, **6**, 394–398, <https://doi.org/10.1038/nclimate2915>.
- Guttman, N. B., 1999: Accepting the standardized precipitation index: A calculation algorithm 1. *J. Amer. Water Resour. Assoc.*, **35**, 311–322, <https://doi.org/10.1111/j.1752-1688.1999.tb03592.x>.
- Hall, A., 2014: Projecting regional change. *Science*, **346**, 1461–1462, <https://doi.org/10.1126/science.aaa0629>.
- Hayes, M. J., C. Alvord, and J. Lowrey, 2002: Drought indices. National Drought Mitigation Center, University of Nebraska, 18 pp.
- Hayhoe, K., and Coauthors, 2007: Past and future changes in climate and hydrological indicators in the US Northeast. *Climate Dyn.*, **28**, 381–407, <https://doi.org/10.1007/s00382-006-0187-8>.
- 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>.
- Henderson-Sellers, A., and Coauthors, 1998: Tropical cyclones and global climate change: A post-IPCC assessment. *Bull. Amer. Meteor. Soc.*, **79**, 19–38, [https://doi.org/10.1175/1520-0477\(1998\)079<0019:TCAGCC>2.0.CO;2](https://doi.org/10.1175/1520-0477(1998)079<0019:TCAGCC>2.0.CO;2).
- Hersbach, H., 2016: The ERA5 atmospheric reanalysis. 2016 Fall Meeting, San Francisco, CA, Amer. Geophys. Union, Abstract NG33D-01.
- , and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, <https://doi.org/10.1002/qj.3803>.
- Hobbs, J. J., 2008: *World Regional Geography*. Nelson Education, 752 pp.
- Hoffmann, L., and Coauthors, 2019: From ERA-Interim to ERA5: The considerable impact of ECMWF's next-generation reanalysis on Lagrangian transport simulations. *Atmos. Chem. Phys.*, **19**, 3097–3124, <https://doi.org/10.5194/acp-19-3097-2019>.
- Jagannathan, K., A. D. Jones, and I. Ray, 2021: The making of a metric: Co-producing decision-relevant climate science. *Bull. Amer. Meteor. Soc.*, <https://doi.org/10.1175/BAMS-D-19-0296.1>, in press.
- 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>.
- Kharin, V. V., F. W. Zwiers, X. Zhang, and G. C. Hegerl, 2007: Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. *J. Climate*, **20**, 1419–1444, <https://doi.org/10.1175/JCLI4066.1>.
- Knutti, R., and J. Sedláček, 2013: Robustness and uncertainties in the new CMIP5 climate model projections. *Nat. Climate Change*, **3**, 369–373, <https://doi.org/10.1038/nclimate1716>.
- Kolb, T. E., C. J. Fettig, M. P. Ayres, B. J. Bentz, J. A. Hicke, R. Mathiasen, J. E. Stewart, and A. S. Weed, 2016: Observed and anticipated impacts of drought on forest insects and diseases in the United States. *For. Ecol. Manage.*, **380**, 321–334, <https://doi.org/10.1016/j.foreco.2016.04.051>.
- Koutroulis, A., M. Grillakis, I. Tsanis, and L. Papadimitriou, 2016: Evaluation of precipitation and temperature simulation performance of the CMIP3 and CMIP5 historical experiments. *Climate Dyn.*, **47**, 1881–1898, <https://doi.org/10.1007/s00382-015-2938-x>.
- Krakauer, N. Y., T. Lakhankar, and D. Hudson, 2019: Trends in drought over the Northeast United States. *Water*, **11**, 1834, <https://doi.org/10.3390/w11091834>.
- Lee, J., K. R. Sperber, P. J. Gleckler, C. J. Bonfils, and K. E. Taylor, 2018: Quantifying the agreement between observed and simulated extratropical modes of interannual variability. *Climate Dyn.*, **52**, 4057–4089, <https://doi.org/10.1007/s00382-018-4355-4>.
- Li, J.-L., and Coauthors, 2020: An overview of CMIP5 and CMIP6 simulated cloud ice, radiation fields, surface wind stress, sea surface temperatures, and precipitation over tropical and subtropical oceans. *J. Geophys. Res. Atmos.*, **125**, e2020JD032848, <https://doi.org/10.1029/2020JD032848>.
- Livneh, B., T. J. Bohn, D. W. Pierce, F. Munoz-Arriola, B. Nijssen, R. Vose, D. R. Cayan, and L. Brekke, 2015: A spatially comprehensive, hydrometeorological data set for Mexico, the US, and Southern Canada 1950–2013. *Sci. Data*, **2**, 150042, <https://doi.org/10.1038/sdata.2015.42>.
- Lower Mississippi River Conservation Committee, 2014: The economic profile of the Lower Mississippi River: An update. Final Rep., Lower Mississippi River Conservation Committee, 8 pp.,

- https://www.lmrcc.org/wp-content/uploads/2020/12/LMR_Economic_Profile_ExecutiveSummary.pdf.
- Lu, Y., I. Cohen, X. S. Zhou, and Q. Tian, 2007: Feature selection using principal feature analysis. *Proc. 15th ACM Int. Conf. on Multimedia*, Augsburg, Germany, Association for Computing Machinery, 301–304, <https://doi.org/10.1145/1291233.1291297>.
- Maraun, D., and M. Widmann, 2015: The representation of location by a regional climate model in complex terrain. *Hydrol. Earth Syst. Sci.*, **19**, 3449–3456, <https://doi.org/10.5194/hess-19-3449-2015>.
- McKee, T. B., and Coauthors, 1993: The relationship of drought frequency and duration to time scale. Preprints, *Eighth Conf. on Applied Climatology*, Anaheim, CA, Amer. Meteor. Soc., 179–184.
- Mearns, L. O., and Coauthors, 2017: The NA-CORDEX dataset, version 1.0. NCAR Climate Data Gateway, accessed 12 October 2019, <https://doi.org/10.5065/D6SJ1JCH>.
- Meher, J. K., L. Das, J. Akhter, R. E. Benestad, and A. Mezghani, 2017: Performance of CMIP3 and CMIP5 GCMs to simulate observed rainfall characteristics over the Western Himalayan region. *J. Climate*, **30**, 7777–7799, <https://doi.org/10.1175/JCLI-D-16-0774.1>.
- Melillo, J. M., T. Richmond, and G. Yohe, 2014: *Climate change impacts in the United States: The Third National Climate Assessment*. U.S. Global Change Research Program, 841 pp, <https://doi.org/10.7930/J0Z31WJ2>.
- Mishra, A. K., and V. P. Singh, 2010: A review of drought concepts. *J. Hydrol.*, **391**, 202–216, <https://doi.org/10.1016/j.jhydrol.2010.07.012>.
- Mo, K. C., 2011: Drought onset and recovery over the United States. *J. Geophys. Res.*, **116**, D20106, <https://doi.org/10.1029/2011JD016168>.
- Moon, H., L. Gudmundsson, and S. I. Seneviratne, 2018: Drought persistence errors in global climate models. *J. Geophys. Res. Atmos.*, **123**, 3483–3496, <https://doi.org/10.1002/2017JD027577>.
- Nasrollahi, N., A. AghaKouchak, L. Cheng, L. Damberg, T. J. Phillips, C. Miao, K. Hsu, and S. Sorooshian, 2015: How well do CMIP5 climate simulations replicate historical trends and patterns of meteorological droughts? *Water Resour. Res.*, **51**, 2847–2864, <https://doi.org/10.1002/2014WR016318>.
- Nguyen, P., and Coauthors, 2017: Evaluation of CMIP5 model precipitation using PERSIANN-CDR. *J. Hydrometeorol.*, **18**, 2313–2330, <https://doi.org/10.1175/JHM-D-16-0201.1>.
- NOAA/Physical Sciences Laboratory, 2020: CPC Unified Gauge-Based Analysis of Daily Precipitation over CONUS. NOAA Physical Sciences Laboratory, accessed 15 December 2019, <https://psl.noaa.gov/data/gridded/data.unified.daily.conus.html>.
- Orlowsky, B., and S. I. Seneviratne, 2013: Elusive drought: Uncertainty in observed trends and short- and long-term CMIP5 projections. *Hydrol. Earth Syst. Sci.*, **17**, 1765–1781, <https://doi.org/10.5194/hess-17-1765-2013>.
- Otkin, J. A., M. Svoboda, E. D. Hunt, T. W. Ford, M. C. Anderson, C. Hain, and J. B. Basara, 2018: Flash droughts: A review and assessment of the challenges imposed by rapid-onset droughts in the United States. *Bull. Amer. Meteor. Soc.*, **99**, 911–919, <https://doi.org/10.1175/BAMS-D-17-0149.1>.
- Pierce, D. W., D. R. Cayan, and B. L. Thrasher, 2014: Statistical downscaling using localized constructed analogs (LOCA). *J. Hydrometeorol.*, **15**, 2558–2585, <https://doi.org/10.1175/JHM-D-14-0082.1>.
- Pontoppidan, M., E. Kolstad, S. Sobolowski, and M. King, 2018: Improving the reliability and added value of dynamical downscaling via correction of large-scale errors: A Norwegian perspective. *J. Geophys. Res. Atmos.*, **123**, 11 875–11 888, <https://doi.org/10.1029/2018JD028372>.
- Prasanna, V., 2018: Statistical bias correction method applied on CMIP5 datasets over the Indian region during the summer monsoon season for climate change applications. *Theor. Appl. Climatol.*, **131**, 471–488, <https://doi.org/10.1007/s00704-016-1974-8>.
- Priestley, M. D., D. Ackerley, J. L. Catto, K. I. Hodges, R. E. McDonald, and R. W. Lee, 2020: An overview of the extra-tropical storm tracks in CMIP6 historical simulations. *J. Climate*, **33**, 6315–6343, <https://doi.org/10.1175/JCLI-D-19-0928.1>.
- Rummukainen, M., 2010: State-of-the-art with regional climate models. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 82–96, <https://doi.org/10.1002/wcc.8>.
- Seager, R., N. Pederson, Y. Kushnir, J. Nakamura, and S. Jurburg, 2012: The 1960s drought and the subsequent shift to a wetter climate in the Catskill Mountains region of the New York City watershed. *J. Climate*, **25**, 6721–6742, <https://doi.org/10.1175/JCLI-D-11-00518.1>.
- Song, F., Z. Guo, and D. Mei, 2010: Feature selection using principal component analysis. *2010 Int. Conf. on System Science, Engineering Design and Manufacturing Informatization*, Yichang, China, Institute of Electrical and Electronics Engineers, 27–30, <https://doi.org/10.1109/ICSEM.2010.14>.
- Srivastava, A., R. Grotjahn, and P. Ullrich, 2020: Evaluation of historical CMIP6 model simulations of extreme precipitation over contiguous US regions. *Wea. Climate Extremes*, **29**, 100268, <https://doi.org/10.1016/j.wace.2020.100268>.
- Stansfield, A. M., K. A. Reed, C. M. Zarzycki, P. A. Ullrich, and D. R. Chavas, 2020: Assessing tropical cyclones' contribution to precipitation over the eastern united states and sensitivity to the variable-resolution domain extent. *J. Hydrometeorol.*, **21**, 1425–1445, <https://doi.org/10.1175/JHM-D-19-0240.1>.
- Strzepek, K., G. Yohe, J. Neumann, and B. Boehlert, 2010: Characterizing changes in drought risk for the United States from climate change. *Environ. Res. Lett.*, **5**, 044012, <https://doi.org/10.1088/1748-9326/5/4/044012>.
- Svoboda, M., and B. Fuchs, 2016: Handbook of drought indicators and indices. WMO-1173, 52 pp., https://www.droughtmanagement.info/literature/GWP_Handbook_of_Drought_Indicators_and_Indices_2016.pdf.
- , M. Hayes, and D. Wood, 2012: Standardized Precipitation Index User Guide. WMO-1090, World Meteorological Organization, 24 pp., <https://public.wmo.int/en/resources/library/standardized-precipitation-index-user-guide>.
- Switanek, M. B., P. A. Troch, C. L. Castro, A. Leuprecht, H.-I. Chang, R. Mukherjee, and E. Demaria, 2017: Scaled distribution mapping: A bias correction method that preserves raw climate model projected changes. *Hydrol. Earth Syst. Sci.*, **21**, 2649–2666, <https://doi.org/10.5194/hess-21-2649-2017>.
- Taylor, I., E. Burke, L. McColl, P. Falloon, G. Harris, and D. McNeall, 2013: The impact of climate mitigation on projections of future drought. *Hydrol. Earth Syst. Sci.*, **17**, 2339–2358, <https://doi.org/10.5194/hess-17-2339-2013>.
- Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, <https://doi.org/10.1029/2000JD900719>.
- , R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, <https://doi.org/10.1175/BAMS-D-11-00094.1>.
- Teshome, A., and J. Zhang, 2019: Increase of extreme drought over Ethiopia under climate warming. *Adv. Meteorol.*, **2019**, 5235429, <https://doi.org/10.1155/2019/5235429>.
- The National Drought Mitigation Center, 2021: Drought Classification of U.S. Drought Monitor. The National Drought Mitigation

- Center, <https://droughtmonitor.unl.edu/About/AbouttheData/DroughtClassification.aspx>.
- Thomas, T., P. Nayak, and N. C. Ghosh, 2015: Spatiotemporal analysis of drought characteristics in the Bundelkhand region of central India using the standardized precipitation index. *J. Hydrol. Eng.*, **20**, 05015004, <https://doi.org/10.1061/%28ASCE%29HE.1943-5584.0001189>.
- Trenberth, K. E., and C. J. Guillemot, 1996: Physical processes involved in the 1988 drought and 1993 floods in North America. *J. Climate*, **9**, 1288–1298, [https://doi.org/10.1175/1520-0442\(1996\)009<1288:PPIITD>2.0.CO;2](https://doi.org/10.1175/1520-0442(1996)009<1288:PPIITD>2.0.CO;2).
- , G. W. Branstator, and P. A. Arkin, 1988: Origins of the 1988 North American drought. *Science*, **242**, 1640–1645, <https://doi.org/10.1126/science.242.4886.1640>.
- Ukkola, A., A. Pitman, M. De Kauwe, G. Abramowitz, N. Herger, J. Evans, and M. Decker, 2018: Evaluating CMIP5 model agreement for multiple drought metrics. *J. Hydrometeorol.*, **19**, 969–988, <https://doi.org/10.1175/JHM-D-17-0099.1>.
- Ullrich, P., Z. Xu, A. Rhoades, M. Dettinger, J. Mount, A. Jones, and P. Vahmani, 2018: California's drought of the future: A midcentury recreation of the exceptional conditions of 2012–2017. *Earth's Future*, **6**, 1568–1587, <https://doi.org/10.1029/2018EF001007>.
- U.S. Army Corps of Engineers, 2013: Conservation plan for the Interior Least Tern, Pallid Sturgeon, and Fat Pocketbook Mussel in the Lower Mississippi River. U.S. Army Corps of Engineers, 72 pp., <https://www.lmrcc.org/wp-content/uploads/2021/02/LMR-Conservation-Plan-Final-USACE-CIP-23-July-2013.pdf>.
- U.S. Bureau of Economic Analysis, 2016: Gross Domestic Product (GDP) by State (millions of current dollars). U.S. Bureau of Economic Analysis, <https://apps.bea.gov/regional/histdata/releases/1216qgsp/index.cfm>.
- USDA, 2012: Support for lower Mississippi River basin nomination for Long-Term Agroecosystem Research (LTAR) Network Membership. USDA, 31 pp., <https://www.ars.usda.gov/ARSUserFiles/np211/LMRBProposal.pdf>.
- , 2020: Long-Term Agroecosystem Research (LTAR) Report of the Lower Mississippi River Basin. USDA, 2 pp., <https://ltar.ars.usda.gov/sites/lmr/b/>.
- USGS and NRCS, 2013: Federal Standards and Procedures for the National Watershed Boundary Dataset (WBD). 4th ed. Techniques and Methods 11–A3, 63 pp., <https://pubs.usgs.gov/tm/11/a3/>.
- Velasquez, P., M. Messmer, and C. C. Raible, 2020: A new bias-correction method for precipitation over complex terrain suitable for different climate states: a case study using WRF (version 3.8.1). *Geosci. Model Dev.*, **13**, 5007–5027, <https://doi.org/10.5194/gmd-13-5007-2020>.
- Vicente-Serrano, S. M., J. J. Camarero, and C. Azorin-Molina, 2014: Diverse responses of forest growth to drought time-scales in the Northern Hemisphere. *Global Ecol. Biogeogr.*, **23**, 1019–1030, <https://doi.org/10.1111/geb.12183>.
- Wagner, T., and Coauthors, 2010: The future of hydrology: An evolving science for a changing world. *Water Resour. Res.*, **46**, W05301, <https://doi.org/10.1029/2009WR008906>.
- Wehner, M., P. Gleckler, and J. Lee, 2020: Characterization of long period return values of extreme daily temperature and precipitation in the CMIP6 models: Part 1, model evaluation. *Wea. Climate Extremes*, **30**, 100283, <https://doi.org/10.1016/j.wace.2020.100283>.
- Wilhite, D. A., 2005: *Drought and Water Crises: Science, Technology, and Management Issues*. CRC Press, 432 pp.
- , and M. H. Glantz, 1985: Understanding: The drought phenomenon: The role of definitions. *Water Int.*, **10**, 111–120, <https://doi.org/10.1080/02508068508686328>.
- Williams, A. P., R. Seager, J. T. Abatzoglou, B. I. Cook, J. E. Smerdon, and E. R. Cook, 2015: Contribution of anthropogenic warming to California drought during 2012–2014. *Geophys. Res. Lett.*, **42**, 6819–6828, <https://doi.org/10.1002/2015GL064924>.
- Xie, P., M. Chen, S. Yang, A. Yatagai, T. Hayasaka, Y. Fukushima, and C. Liu, 2007: A gauge-based analysis of daily precipitation over east asia. *J. Hydrometeorol.*, **8**, 607–626, <https://doi.org/10.1175/JHM583.1>.
- , —, and W. Shi, 2010: CPC unified gauge-based analysis of global daily precipitation. *24th Conf. on Hydrology*, Atlanta, GA, Amer. Meteor. Soc., 2.3A, https://ams.confex.com/ams/90annual/techprogram/paper_163676.htm.
- Yang, S., J. Feng, W. Dong, and J. Chou, 2014: Analyses of extreme climate events over China based on CMIP5 historical and future simulations. *Adv. Atmos. Sci.*, **31**, 1209–1220, <https://doi.org/10.1007/s00376-014-3119-2>.
- Yihdego, Y., B. Vaheddoost, and R. A. Al-Weshah, 2019: Drought indices and indicators revisited. *Arab. J. Geosci.*, **12**, 69, <https://doi.org/10.1007/s12517-019-4237-z>.
- Zargar, A., R. Sadiq, B. Naser, and F. I. Khan, 2011: A review of drought indices. *Environ. Rev.*, **19**, 333–349, <https://doi.org/10.1139/A11-013>.
- Zarzycki, C. M., P. A. Ullrich, and K. A. Reed, 2021: Metrics for evaluating tropical cyclones in climate data. *J. Appl. Meteor. Climatol.*, **60**, 643–660.
- Zhuang, J., R. Dussin, A. Jüling, and S. Rasp, 2020: JiaweiZhuang/xESMF: v0.3.0 Adding ESMF. LocStream capabilities. Zenodo, <https://doi.org/10.5281/zenodo.3700105>.