Quantifying and Classifying Streamflow Ensembles Using a Broad Range of Metrics for an Evidence-Based Analysis: Colorado River Case Study

Homa Salehabadi[1], David G. Tarboton[1], Kevin G. Wheeler[2,3], Rebecca Smith[4], Sarah Baker[4]

[1]Department of Civil and Environmental Engineering, Utah Water Research Laboratory, Utah State University, Logan, UT, USA.

[2]Environmental Change Institute, University of Oxford, Oxford, UK.

[3]Water Balance Consulting, Boulder, CO, USA.

[4]U.S. Bureau of Reclamation, Boulder, CO, USA.

Corresponding author: Homa Salehabadi (homa.salehabadi@gmail.com)

Key Points:

- Many ensembles representing plausible future streamflow are available for the Colorado River Basin.
- Metrics are presented to provide an evidence-based framework for evaluating these streamflow ensembles.
- A classification approach was developed to provide an analytical framework for grouping and assessing ensembles suitability.

Abstract

Stochastic hydrology produces ensembles of time series that represent plausible future streamflow to simulate and test the operation of water resource systems. A premise of stochastic hydrology is that ensembles should be statistically representative of what may occur in the future. In the past, the application of this premise has involved producing ensembles that are statistically equivalent to the observed or historical streamflow sequence. This requires a number of metrics or statistics that can be used to test statistical similarity. However, with climate change, the past may no longer be representative of the future. Ensembles to test future systems operations should recognize non-stationarity, and include time series representing expected changes. This poses challenges for their testing and validation. In this paper, we suggest an evidence-based analysis in which streamflow ensembles, whether statistically similar to and representative of the past or a changing future, should be characterized and assessed using an extensive set of statistical metrics. We have assembled a broad set of metrics and applied them to annual streamflow in the Colorado River at Lees Ferry to illustrate the approach. We have also developed a tree-based classification approach to categorize both ensembles and metrics. This approach provides a way to visualize and interpret differences between streamflow ensembles. The metrics presented and their classification provide an analytical framework for characterizing and assessing the suitability of future streamflow ensembles, recognizing the presence of non-stationarity. This contributes to better planning in large river basins, such as the Colorado, facing water supply shortages.

Plain Language Summary

Long-range water supply planning in many river basins requires an assessment of ensembles of plausible future streamflow time series used to simulate and test the operation of water resource systems. With climate change, and growing recognition that hydrologic processes are changing over time, the past may no longer be representative of the future. This poses challenges when using statistical metrics to test future streamflow ensembles. In this paper, we suggest an evidence-based approach in which streamflow ensembles, whether statistically similar to and representative of the past or a changing future, should be characterized using an extensive set of statistical metrics. We have assembled a broad set of metrics and applied them to annual streamflow in the Colorado River at Lees Ferry to illustrate the approach. We have also developed an approach to categorize both ensembles and metrics. The metrics presented and their classification provide an analytical framework for characterizing and assessing the suitability of future streamflow ensembles for water resources system planning. The metrics and classification developed advance and contribute to better planning in large river basins facing water supply shortages.

1. Introduction

In water resources planning in large river basins, such as the Colorado River in the southwestern U.S., ensembles of streamflow time series are commonly used to assess the performance of alternative policies and management strategies (Bonham et al., 2024; Wheeler et al., 2022). It is important that these ensembles have statistical properties representative of a wide range of plausible future streamflow conditions. Relying solely on historical flow records to generate data for water resource analyses limits the ability to test strategies and policies against the diverse range of sequences possible in the future. While the historical record holds valuable information for the future, given climate change (Milly et al., 2008; IPCC, 2021), we can reasonably assume that future flow sequences will not precisely mirror historical patterns. There is thus a need to have statistical metrics that characterize the properties of potential future streamflow ensembles and to use these metrics to assess the suitability of ensembles for use in future planning. This paper provides a broad set of metrics that can be used to characterize and classify streamflow ensembles, to address this need.

Stochastic streamflow models can generate a broad range of potential flow sequences for river basin planning and analyses. These models can use observed flow records, proxy data like tree-ring-reconstructed flows, and/or General Circulation Model (GCM) projections to generate ensembles of plausible future streamflow sequences. These ensembles serve as inputs to systems planning and operations models, allowing testing of their resilience against potential future scenarios. Most commonly, stochastic streamflow models generate ensembles of synthetic streamflow sequences primarily based on historical data, often assuming stationarity (Fiering, 1967; Matalas et al., 1982; Valencia & Schaake, 1973; Vogel, 2017; Yevjevich, 1963), although efforts have been made to adapt them for nonstationary hydrologic processes to capture changes due to climate and anthropogenic impacts (Borgomeo et al., 2014; Salas et al., 2018).

A suitable streamflow model should capture the fundamental characteristics expected during the planning period. For a particular river basin study, identifying which characteristics are essential is important, yet challenging. A premise of much prior stochastic hydrology is that the future will be different from, but statistically similar to, the past (Loucks et al., 2017). Statistical similarity is quantified using a number of statistics, or metrics, which ensemble sequences are expected to reproduce. The assumption of stationarity is not always plausible, especially in river basins where significant alterations in runoff characteristics have occurred due to changes in land cover, land use, climate, or groundwater utilization during the recorded flow period (Loucks et al., 2017). As a result, exact replication of past statistics is no longer directly applicable in such basins, especially in an era of climate change (Milly et al., 2008). Nevertheless, there remains a critical need to employ and further develop metrics that quantify attributes of stochastic ensembles as valuable evidence-based tools for interpreting streamflow model results. Furthermore, metrics provide objective and quantitative evidence to interpret and analyze representations of non-stationarity such as differences between past streamflows and ensembles that incorporate projected climate changes. Evidence-based analysis supports robust decision-making by offering clear, documented, and communicable information (Pezij et al., 2019). It helps prevent the adoption of ensembles without full information on their characteristics and solely because they have been used previously. Using a broad range of metrics to describe hydrologic characteristics associated with streamflow ensembles used in water resources planning provides evidence on how sufficient the ensembles are for their intended purposes.

Statistical attributes of the historical data provide quantitative context that plays a crucial role in analyzing streamflow ensembles and assessing their ability to replicate historical patterns or desired characteristics. Various common statistics, such as mean, standard deviation, skewness, minimum, maximum, probability distribution, and correlation are widely used in studies to either evaluate the model’s goodness-of-fit or compare different models (e.g. Koutsoyiannis et al., 2008; Lee & Ouarda, 2012, 2023; Lee et al., 2010; 2020; Prairie et al., 2006; 2007; 2008; Salas et al., 2005; Sharma et al., 1997; Srinivas & Srinivasan, 2000, 2005, 2006; Tarboton, 1994). In addition to these common statistics, a range of other metrics are available to capture various aspects of streamflow ensembles. The Hurst coefficient is used to quantify long-term memory or persistence beyond what is captured by correlation (Chaves &

Lorena, 2019; Hurst, 1951; Klemeš, 1974; Lee & Ouarda, 2023; Lee et al., 2020). Detecting trends is another useful approach to quantify non-stationarity in time series (Helsel et al., 2020; Kendall, 1955; Lee & Ouarda, 2023; Mann, 1945). Mutual information is a measure of dependence that, unlike correlation, accounts for both linear and nonlinear dependence present in the time series, offering a more comprehensive understanding of the relationships within the data (Gong et al., 2014; Harrold et al., 2001; Loritz et al., 2018; Pechlivanidis et al., 2016; 2018).

Hydrological droughts and surpluses are additional metrics that frequently draw significant interest and attention in hydrological studies. These metrics provide crucial insights for water resource management, especially in regions prone to water scarcity or excess. Understanding the occurrence, duration, and severity of hydrological droughts, as well as the frequency and magnitude of surpluses, is essential for making informed decisions regarding water allocation, reservoir management, and drought preparedness. Previous studies have commonly explored these statistics using the run-sum approach (Lee & Ouarda, 2023; Lee et al., 2020; Prairie et al., 2006; Salas et al., 2005; Srinivas & Srinivasan, 2006). However, a limitation of this method is that it defines a drought or surplus event as events when all consecutive years are above a below a threshold, without any breaking year during that period. Our earlier work offered duration-severity analysis as a more general approach to quantifying drought or surplus without this limitation (Salehabadi et al., 2022).

In addition to the above metrics, storage-related metrics quantify characteristics associated with the practical evaluation of the storage capacity needed in reservoirs to meet specific yields or to manage reservoirs to sustain desired demands (see for example Lee & Ouarda, 2023; Srinivas & Srinivasan, 2006). Storage metrics are thus directly meaningful to water resource management. For a given streamflow sequence, the storage required to support a specified yield can be estimated using sequent peak analyses (Loucks et al., 2017).

Overall, based on the literature, a diverse range of metrics are available to quantify and assess the characteristics of a streamflow ensemble. When there are multiple sources of streamflow ensembles, these metrics assist in informed decision-making regarding ensemble selection for various planning needs.

To facilitate the comparison of multiple ensembles, simplify the extraction of information from an extensive set of metrics, and classify the ensembles based on their characteristics, agglomerative hierarchical clustering analysis can be used (Hastie et al., 2009; Murtagh & Contreras, 2012). Clustering techniques employ a similarity or distance criterion to determine how and to what extent the objects (streamflow models in our case) are close/similar or far/dissimilar. Once a similarity criterion is selected, the algorithm begins by assigning each object to its own cluster. Then, it iteratively merges the two most similar clusters until all objects belong to a single cluster. Previous studies such as Papacharalampous et al. (2019) have suggested a comprehensive set of forecast quality metrics and used a clustering approach to compare the performance of various methods for forecasting hydrological processes. Some aspects of their approach are similar to ours, but our focus here is on the annual scale and longer-term storage and drought/surplus quantities important for watersheds such as the Colorado River Basin where there is reservoir capacity to support significant interannual storage. In another study, Ahmadalipour et al. (2015) employed a number of statistical metrics and a clustering approach to analyze, compare, and rank the performance of various global climate models from Climate Model Intercomparison Project 5 (CMIP5) dataset over the Columbia River Basin. Razavi et al. (2015) used a clustering analysis to cluster and assess the similarities or dissimilarities among various tree-ring chronology sites in the Saskatchewan River Basin. This literature suggest that such clustering techniques can be used to classify multiple streamflow ensembles based on their characteristics.

In this study, we employ an evidence-based approach to objectively analyze Colorado River Basin streamflow ensembles and quantify the differences between them. To do this, we identify and develop a comprehensive suite of metrics to quantitatively evaluate and describe streamflow ensembles, compare them with historical data, and explore their uncertainties. We use these metrics as evidence-based tools to assess whether an ensemble is sufficient for its intended purpose. The contribution is the comprehensive suite of metrics covering a broad class of statistical characteristics, with documented uncertainty and guidance on application and interpretation for the evaluation of a streamflow ensemble. Our metrics address limitations of drought statistics and also quantify the occurrence of high flows, which are important for filling reservoirs in some systems. We also developed a classification approach that groups similar ensembles based on the metrics and provides a classification of the metrics themselves. This classification offers opportunities for efficiency, since ensembles with like attributes may not need to be evaluated in full.

The paper is structured as follows: First, we describe the study area and the data used, encompassing 21 ensembles of streamflow sequences within the Colorado River Basin. Next, we provide an overview of the metrics employed for quantifying the streamflow ensembles. The results section provides ensemble-specific metrics utilized for individual ensemble interpretation, followed by comparative results and ensemble classification based on their attributes. Finally, we draw conclusions on the utilization of a diverse range of metrics to identify ensembles that closely align with the desired attributes essential for various planning purposes.

2. Study Area and Data Used

The Colorado River (Schmidt et al., 2022), often referred to as "America's Nile (LaRue, 1916)," is a vital water resource for the southwestern United States and northwestern Mexico (Figure 1). Originating in the Rocky Mountains, this river flows through arid landscapes, like the Colorado Plateau, before reaching northwestern Mexico. The river is managed by a set of agreements known as the Law of the River (MacDonnell, 2021) and provides water for millions of people, irrigated agriculture, and hydropower generation. It also holds cultural and ecological significance, with indigenous tribes relying on its waters and a set of protected areas, including National Wildlife Refuges, Recreation Areas, and National Parks, benefiting from its flow.

However, the basin faces significant challenges due to increasing water demand and climate change, which is expected to reduce water runoff and exacerbate droughts (Milly & Dunne, 2020; Schmidt et al., 2023; Udall & Overpeck, 2017; Williams et al., 2020; Xiao et al., 2018). These changes threaten the sustainability of water resources and call for innovative strategies to manage and adapt to evolving conditions in the basin (Rosenberg, 2022; Wheeler et al., 2021; 2022; Fleck & Castle, 2022). One of the primary inputs needed for addressing Colorado River management is projections of future streamflow, even though the precise characteristics of this future remain uncertain.

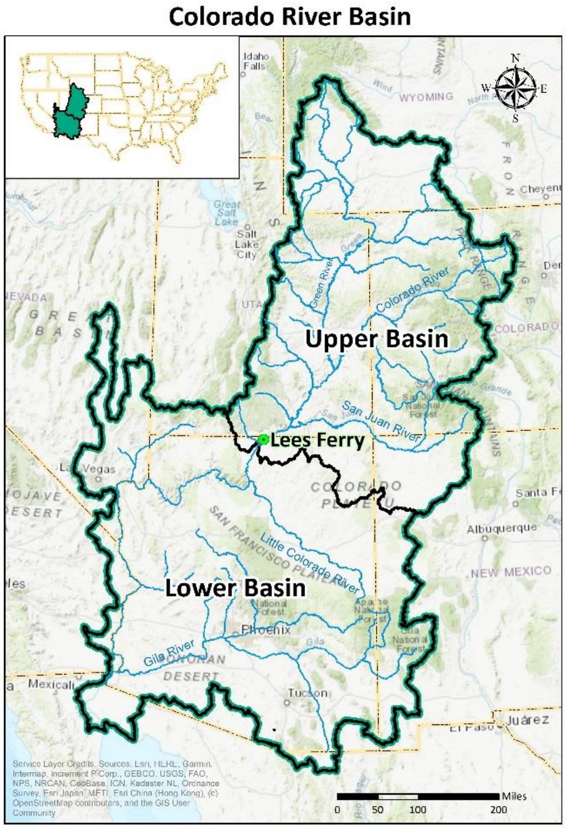


Figure 1. Study area, Colorado River Basin and Lees Ferry gage location

The Colorado River Basin splits into the Upper Basin and Lower Basin near the Lees Ferry gage, through which 85 to 90% of the river's flow passes (Figure 1). This makes the natural flow at Lees Ferry the main metric for quantifying runoff within the basin. Natural flow represents an estimate of what the flow would have been in the absence of consumptive uses, reservoir evaporation and dam operations. The U.S. Bureau of Reclamation (hereafter Reclamation) maintains a historical natural flow dataset derived from measurements and estimates of consumptive use and diversions (Prairie & Callejo, 2005). Reclamation updates this monthly dataset regularly. The most recent update, as of November 2023, includes historical data from 1906 through 2020, with provisional estimates for 2021 and 2022 (USBR, 2022). Additionally, tree-ring-reconstructed (or paleo-reconstructed) natural flow at Lees Ferry extends historical data beyond the 1906-2022 observed record. Meko et al. (2017) provides a tree-ring reconstruction for 1416 to 2015 at an annual water year timescale. These historical and paleo-reconstructed datasets were employed to compare their statistical attributes with future streamflow ensembles.

In the Colorado River Basin, there are multiple long-term streamflow ensembles developed by previous studies using different approaches (Prairie et al., 2006; 2007; 2008; Salehabadi et al., 2020; 2022; Tarboton, 1994; Udall, 2020; USBR, 2011, 2012, 2014; Vano et al., 2020; Woodhouse et al., 2021). Certain previously developed streamflow ensembles are based on either historical, paleo-reconstructed, or climate-change-informed flows, and some others are a combination of these sources. Each ensemble has particular statistical attributes and represents a set of assumptions about uncertain future hydrology. These streamflow ensembles have been developed to provide streamflow sequences as inputs to the Colorado River Simulation System (CRSS). CRSS, implemented in RiverWare (Zagona et al., 2001), is the major long-term water resources planning tool in the Colorado River Basin used by Reclamation to project future conditions in the basin for years and decades (Payton et al., 2020). The planning results are highly sensitive to the future streamflow used, and there is a need to characterize the ensembles to support scenario planning and robust decision-making under deep uncertainty (Smith et al., 2022). Additionlly, there is a planning effort ongoing in the basin called “Colorado River Post-2026 Operations” that will identify a range of water management alternatives for potentially decades into the future (USBR, 2023). The Post-2026 process will use specific streamflow ensembles and the findings of our study could help inform choices on adequate ensembles for various planning purposes.

The Colorado River streamflow ensembles we assessed in this study are listed in Table 1.

Table 1
Streamflow Ensembles in the Colorado River Basin.

	Ensemble name	**Ensemble identifier**	**Reference**	**Flow data source**	**Method**	**Number of traces**	**Length of planning period**	**Explanation**
1	Full hydrology	ISM_1906_2020	USBR (2012)	Observed natural flow, 1906-2020 (data from USBR, 2022)	Index Sequential Method (ISM)	115	50 years	ISM applied to the 1906 to 2020 period of the observed natural flow with the first 50 years of each ISM trace selected.
2	Pluvial-removed ISM	ISM_1931_2020		Observed natural flow, 1931-2020 (data from USBR, 2022)	Index Sequential Method (ISM)	90	50 years	ISM applied to the 1931 to 2020 period of the observed natural flow with the first 50 years of each ISM trace selected.
3	Stress test	ISM_1988_2020	USBR (2012)	Observed natural flow, 1988-2020 (data from USBR, 2022)	Index Sequential Method (ISM)	33	33 years	ISM applied to the 1988 to 2020 period of the observed natural flow.
4	Paleo ISM	ISM_1416_2015	USBR (2012)	Tree-ring-reconstructed flow, 1416-2015 (from Meko et al., 2017)	Index Sequential Method (ISM)	600	50 years	ISM applied to the 1416 to 2015 period of the tree-ring-reconstructed flow with the first 50 years of each ISM trace selected.
5	AR1	AR1	Salehabadi et al. (2022)	Observed natural flow, 1906-2020 (data from USBR, 2022)	Auto-Regressive order 1	100	50 years	Streamflow ensemble generated by Salehabadi et al. (2022)
6	Full record paleo conditioned	NPC_1906_2020	Prairie et al. (2008)	Observed natural flow, 1906-2020 (data from USBR, 2022); Tree-ring-reconstructed	Nonparametric Paleo-Conditioned (NPC)	100	50 years	NPC method described by Prairie et al. (2008) applied to the full record (1906-2020) of the observed natural flow

	Ensemble name	**Ensemble identifier**	**Reference**	**Flow data source**	**Method**	**Number of traces**	**Length of planning period**	**Explanation**
				flow, 1416-2015 (data from Meko et al., 2017)				
7	Stress test paleo conditioned	NPC_1988_2020	Prairie et al. (2008)	Observed natural flow, 1988-2020 (data from USBR, 2022); Tree-ring-reconstructed flow, 1416-2015 (data from Meko et al., 2017)	Nonparametric Paleo-Conditioned (NPC)	100	50 years	NPC method described by Prairie et al. (2008) applied to the stress test period (1988-2020) of the observed natural flow
8	Millennium drought paleo conditioned	NPC_2000_2020	Prairie et al. (2008)	Observed natural flow, 2000-2020 (data from USBR, 2022); Tree-ring-reconstructed flow, 1416-2015 (data from Meko et al., 2017)	Nonparametric Paleo-Conditioned (NPC)	100	50 years	NPC method described by Prairie et al. (2008) applied to the millennium drought period (2000-2020) of the observed natural flow
9	Millennium drought 5-yr block resampling	5YrBlockRes_2000_2018	Salehabadi et al. (2022)	Observed natural flow, 2000-2020 (data from USBR, 2022)	5-year Block Resampling	100	42 years	Streamflow ensemble generated by Salehabadi et al. (2022)
10	Millennium drought year resampling	DroughtYrRes_2000_2020	(Salehabadi et al., 2022)	Observed natural flow, 2000-2020 (data from USBR, 2022)	Drought scenario resampling (uncorrelated)	100	50 years	Streamflow ensemble generated by Salehabadi et al. (2022)
11	Mid-20th Century drought year resampling	DroughtYrRes_1953_1977	(Salehabadi et al., 2022)	Observed natural flow, 1953-1977	Drought scenario resampling (uncorrelated)	100	50 years	Streamflow ensemble generated by Salehabadi et al. (2022)

	Ensemble name	**Ensemble identifier**	**Reference**	**Flow data source**	**Method**	**Number of traces**	**Length of planning period**	**Explanation**
				(data from USBR, 2022)				
12	Paleo drought year resampling	DroughtYrRes_1576_1600	(Salehabadi et al., 2022)	Tree-ring-reconstructed flow, 1576-1600 (data from Meko et al., 2017)	Drought scenario resampling (uncorrelated)	100	50 years	Streamflow ensemble generated by Salehabadi et al. (2022)
13	CMIP3-BCSD hydrology projections	CMIP3_BCSD	USBR (2011)	Reclamation's flow projections, 1951-2099	CMIP3, BCSD, VIC	112	50 years (2027-2076)	Downscaled BCSD CMIP3 hydrology projections from USBR (2011)
14	CMIP5-BCSD hydrology projections	CMIP5_BCSD	USBR (2014)	Reclamation's flow projections, 1951-2099	CMIP5, BCSD, VIC	97	50 years (2027-2076)	Downscaled BCSD CMIP5 hydrology projections from USBR (2014)
15	CMIP5-LOCA hydrology projections	CMIP5_LOCA	Vano et al. (2020)	Reclamation's flow projections, 1951-2099	CMIP5, LOCA, VIC	64	50 years (2027-2076)	Downscaled LOCA CMIP5 hydrology projections from Vano et al. (2020)
16	Temperature-adjusted flow, RCP45-030	TempAdj_RCP4.5_3%	Udall (2020)	Observed natural flow, 1906-2017 (data from USBR, 2022)	Uniform proportional decreases in runoff. Future temperatures based on the RCP scenario and streamflow sensitivity to temperature set according to the percentage given	112	50 years (2027-2076)	Temperature-adjusted streamflow ensemble form Udall (2020). Emission scenario: RCP 4.5, Streamflow sensitivity to temperature: 3% per 1°C

	Ensemble name	**Ensemble identifier**	**Reference**	**Flow data source**	**Method**	**Number of traces**	**Length of planning period**	**Explanation**
17	Temperature-adjusted flow, RCP45-065	TempAdj_RCP4.5_6.5%	Udall (2020)	Observed natural flow, 1906-2017 (data from USBR, 2022)	Uniform proportional decreases in runoff	112	50 years (2027-2076)	Emission scenario: RCP 4.5, Streamflow sensitivity to temperature: 6.5% per 1°C
18	Temperature-adjusted flow, RCP45-100	TempAdj_RCP4.5_10%	Udall (2020)	Observed natural flow, 1906-2017 (data from USBR, 2022)	Uniform proportional decreases in runoff	112	50 years (2027-2076)	Emission scenario: RCP 4.5, Streamflow sensitivity to temperature: 10% per 1°C
19	Temperature adjusted flow, RCP85-030	TempAdj_RCP8.5_3%	Udall (2020)	Observed natural flow, 1906-2017 (data from USBR, 2022)	Uniform proportional decreases in runoff	112	50 years (2027-2076)	Emission scenario: RCP 8.5, Streamflow sensitivity to temperature: 3% per 1°C
20	Temperature-adjusted flow, RCP85-065	TempAdj_RCP8.5_6.5%	Udall (2020)	Observed natural flow, 1906-2017 (data from USBR, 2022)	Uniform proportional decreases in runoff	112	50 years (2027-2076)	Emission scenario: RCP 8.5, Streamflow sensitivity to temperature: 6.5% per 1°C
21	Temperature-adjusted flow, RCP85-100	TempAdj_RCP8.5_10%	Udall (2020)	Observed natural flow, 1906-2017 (data from USBR, 2022)	Uniform proportional decreases in runoff	112	50 years (2027-2076)	Emission scenario: RCP 8.5, Streamflow sensitivity to temperature: 10% per 1°C

3. Methodology

An extensive set of metrics was identified or developed to effectively describe hydrologic characteristics associated with streamflow ensembles. The metrics provide a framework to objectively test an ensembles ability to reproduce desired or historical attributes deemed important for the decision-making scenario being considered. Complete reproduction of all historical characteristics may not always be desired. For example, where the question involves managing for streamflow declining due to climate change, the historical mean is not expected to be reproduced. In this section, we provide an overview of these metrics, followed by a description of Ward's Agglomerative Hierarchical Clustering method, which we employed for ensemble classification based on the calculated metrics.

3.1. Common Metrics

There are well-known metrics such as mean, median, minimum, maximum, standard deviation, skewness, Auto Correlation Function (ACF), and trend that are commonly used in studies to either evaluate the goodness-of-fit of a model or compare different models (e.g. Koutsoyiannis et al., 2008; Lee & Ouarda, 2012, 2023; Lee et al., 2010; 2020; Prairie et al., 2006; 2007; 2008; Salas et al., 2005; Sharma et al., 1997; Srinivas & Srinivasan, 2000, 2005, 2006; Tarboton, 1994). Here they were evaluated from their readily available formulae using standard functions or libraries in R (R Core Team, 2023). The Mann-Kendall test (Kendall, 1955; Mann, 1945) was applied in this study to detect the occurrence of significant trend in streamflow ensembles. The full set of R scripts used in this paper have been published in HydroShare (Salehabadi & Tarboton, 2024).

3.2. Partial Autocorrelation Function (PACF)

The Partial Autocorrelation Function (PACF), like the Autocorrelation Function (ACF), provides information on the dependence structure of a time series (Bras & Rodriguez-Iturbe, 1985; Hipel & McLeod, 1994). This dependence structure indicates how each observation in the series is correlated with its lagged values, revealing how past observations influence present or future values. It is based on considerations of stationarity so is most meaningful for stationary processes but may also be helpful as a comparative statistic for non-stationary processes. While the ACF quantifies correlation across time lags, PACF is essentially the ACF adjusted for the intervening correlation and quantifies direct additional correlation at higher lags beyond those due to intervening correlation already represented by lower lag correlations. PACF is used to guide the selection of the order of an autoregressive (AR) model used in autoregressive moving average (ARMA) model development and is calculated using the Yule-Walker equations and implemented in R (Venables & Ripley, 2010). For an AR model, the PACF is zero beyond the order of AR model. In other words, the number of non-zero PACF values gives the number of lags that should be used in an AR model to capture historical dependence.

As a metric for quantifying and classifying streamflow ensembles, PACF provides information about dependence. Ensembles that intend to be representative of historical flows should have a similar dependence structure, and deviation from the historical dependence structure should be noted.

3.3. Drought Event Statistics: Length, Deficit, Intensity, Interarrival Time

Hydrologic drought is described as a deficiency in the water supply, which may include streamflow and reservoir storage (Wilhite & Buchanan-Smith, 2005). One way to quantify a hydrologic drought event is as a sequence of consecutive years during which the annual streamflow remains below a specified threshold level, which is typically taken to be the long-term average streamflow (Salas et al., 2005; Tarboton, 1994; Yevjevich, 1967). Alternatively, another definition of a hydrologic drought is consecutive years with streamflow below the long-term mean exceeded by no more than one above-average flow year (Woodhouse et al., 2021). In this framework, droughts may be quantified using metrics such as: (1) the duration of flow below a threshold, (2) magnitude, defined as the cumulative difference between actual flows and a defined threshold, (3) intensity, defined as the average of the below threshold deficit, and (4) the interarrival time. It should be noted that these drought characteristics depend on a specified threshold value and so it is important to consider an appropriate value as the threshold. Additionally, the number of acceptable above-threshold years within the drought duration should be specified. For instance, Woodhouse et al. (2021) allowed one above-average flow year in their drought definition.

For an annual streamflow time series denoted by x_t, $t=1, 2, \ldots, n$ and a constant threshold of x_0, these drought metrics are specified below (Salas et al., 2005) and illustrated in Figure 2.

- *Drought duration or length (L).* The period between the beginning and end of any drought event, i.e. the number of consecutive time intervals (e.g. years) in which $x_t < x_0$.
- *Cumulative deficit (D, drought magnitude).* The deficit that accumulates below the threshold during the drought duration (Equation 1).

$$D = \sum_{j=t}^{t+L-1} (x_0 - x_j) = \sum_{j=t}^{t+L-1} d_j \qquad (1)$$

- *Drought intensity (I)*. The average deficit over the drought duration, namely the ratio of the magnitude to duration of a drought, $I = D/L$.
- *Interarrival time (T)*. The time between the start of two successive droughts.

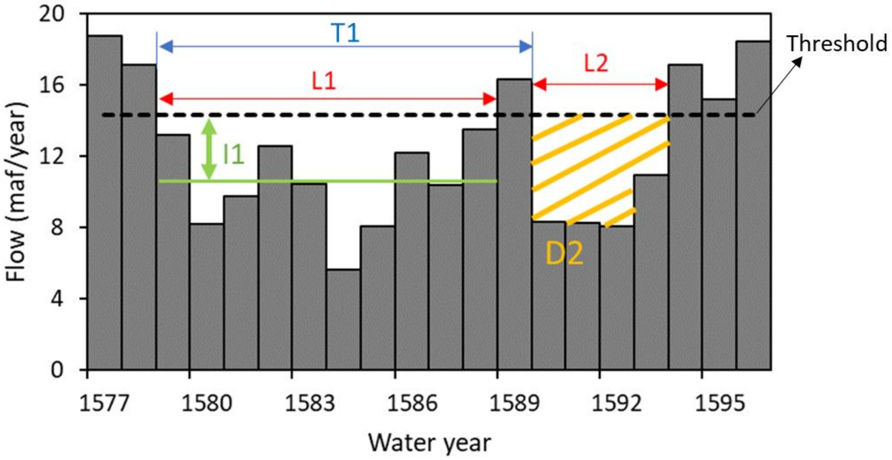


Figure 2. Schematic definition of drought characteristics. The black dashed line gives the threshold level. L1 and L2: length of the first and second drought, respectively. I1: intensity of

the first drought. T1: interarrival time of the first drought. D2: cumulative deficit of the second drought.

As metrics for quantifying streamflow ensembles and evaluating the sufficiency of them, averages, standard deviations, and distributions of these drought statistics provide information about the simulated droughts in a streamflow ensemble. For example, if an ensemble does not reproduce the drought metrics similar to the historical record, it is not representative of what has occurred in the past and this could be used to invalidate an ensemble intended to reproduce past statistics. These metrics also provide information about the characteristics of future droughts in an ensemble. A shortcoming of event statistics is that they break a sustained dry period into separate events when one year, or a selected number of years exceed the threshold. The duration-severity analysis described next are an effort to avoid this shortcoming.

3.4. Duration-Severity Analysis

The duration-severity approach, as introduced by (Salehabadi et al., 2020; 2022), provides a framework for analyzing streamflow data based on severity and duration in order to evaluate the severity and persistence of drought periods (and more generally wet extremes as well). In this approach, severity, which is different from the event definitions of magnitude and intensity discussed in the previous section, is quantified in terms of the mean flow over a specific duration. It considers all periods with that duration in the dataset, including both wet and dry years without separating specific drought events. The duration-severity analysis helps place droughts within the streamflow ensembles in a historical context by comparing these ensembles with either observed or paleo-reconstructed flows. In the context of extreme drought analysis, this approach sheds light on how the lowest mean flows within the ensemble may vary for different durations. It also reveals where the range of extreme droughts falls in relation to the historical flows.

As metrics for quantifying and evaluating streamflow ensembles, examining the position and spread of duration-severity within these ensembles in comparison to historical flows provides insights into the simulated events, such as droughts, present in the ensemble. If an ensemble is intended to be representative of past statistics, the extreme events need to be aligned with what has occurred in the past. This analysis also provides information about changes in the severity of extreme events, and whether an ensemble has more severe and sustained droughts than the historical or paleo-reconstructed record. Streamflow ensembles developed to consider a warmer future may exhibit droughts of greater severity (lower duration-severity values) compared to past data, and the duration-severity analysis provides a quantitative measure of this. Additionally, this analysis reveals the degree of variability within the simulated extreme events. Ensembles with lower variability in hydrologic events have a narrower spread of duration-severity values, while ensembles with higher variability display a broader spread. This variability information is valuable in understanding the range of simulated extreme events.

3.5. Cumulative Deviation

A recasting of the duration-severity analysis is the concept of cumulative deviation, which focuses on measuring the cumulative departure from a particular reference point, such as average conditions, over various durations (Salehabadi et al., 2020; 2022). The cumulative deviation for each n-year duration represents the total deficit or surplus accumulated relative to the reference over those n years. This metric differs from the cumulative deficit in drought event statistics discussed above as it is more general, not accumulating only values below the threshold during a drought duration. Like the duration-severity analysis and unlike the cumulative deficit in drought event statistics, the cumulative deviation includes all years within each duration, whether they are wet or dry years. In the context of drought analysis, this method gives insights on how cumulative deficits within an ensemble vary for various durations. Conversely, in the context of flood analysis, this approach illustrates the variations in cumulative surplus within an ensemble across various durations. Depending on the purpose of analysis, the duration-severity or cumulative deviation approach may be employed. It is important to note that the cumulative deviation calculation depends on a chosen reference mean, while duration-severity analysis is parameter-independent.

3.6. Count Below Threshold (CBT)

The count of periods (e.g. years) with flow below a threshold serves as a drought measure, similar to drought event statistics and duration-severity metrics. The “count below threshold (CBT)” for a specific duration represents the average number of years with flow below the threshold within that duration. CBT can be expressed as either a moving count or an overall average. The moving CBT metric is also a useful tool for visualizing changes (increase or decrease) in the occurrence of flows below the threshold. The difference between this metric and drought length in drought event statistics is that CBT counts the number of below-threshold years without requiring them to be consecutive under a specific drought definition.

3.7. Count Above Threshold (CAT)

The “count above threshold (CAT)” is a metric similar to CBT, but it quantifies the number of years with flow exceeding a specified threshold. It serves as a measure of high-flow occurrence. This metric is particularly valuable when assessing the occurrence of high flows, the occurrence of which is important for filling reservoirs in some systems.

3.8. Hurst Coefficient

The Hurst coefficient (Hurst, 1951) quantifies persistence or long memory in a time series beyond that quantified by correlation or a model that captures correlation. Hurst coefficient (H) can be used to explore the long-term persistence of streamflow, climate, and other geophysical records (Hurst, 1951; Montanari et al., 1997; Vogel et al., 1998). Range (R) is defined as the maximum minus minimum cumulative departure from the mean in a sequence of flows n years long. Rescaled range (R/S) is R divided by standard deviation (S). The Hurst coefficient is defined as the scaling exponent associated with the increase in rescaled range with sample size n. Given a streamflow time series $\{x_1, x_2, \dots, x_n\}$ with sample mean $\bar{x}$ and sample standard deviation S_x, the adjusted partial sums are (Equations 2-4):

$$Y_k = \sum_{t=1}^{k} (x_t - k\bar{x}) \qquad k = 1, \dots, n \tag{2}$$

and the range is

$$R_n = [max(Y_1, Y_2, \dots, Y_n) - min(Y_1, Y_2, \dots, Y_n)] \tag{3}$$

Hurst (1951) found that

$$E\left[\frac{R_n}{S_x}\right] \propto n^H \tag{4}$$

where the exponent H is the Hurst coefficient which varies between 0 and 1. Tarboton (1995) noted that this statistic is uncertain and depends on the length of record over which it is computed. Here, to have a consistent metric for comparison of ensembles we standardized on evaluating average R/S for durations of 8, 16, 32 and the full ensemble number of years and evaluated H from a regression of log(R/S) vs log(n).

A value of H less than or equal to 0.5 means absence of long memory. The occurrence of H > 0.5 is indicative of long-term structure in time series dependence and is referred to as the Hurst phenomenon. This may manifest as persistent droughts and wet periods. The Hurst phenomenon may also be caused by non-stationarity, where the mean of the time series changes with time. It is important to note that when working with short records, the data may be insufficient for a robust interpretation of the Hurst coefficient.

3.9. Mutual Information

Mutual Information (MI) is based on the concept of Shannon entropy (Shannon, 2001), first introduced in 1948, which is a measure of the uncertainty (or lack of information) of a random variable and provides a measure of the amount of information that one random variable contains about another (Cover & Thomas, 2006). In the context of time series, it quantifies the dependence between past and future values. It is similar to correlation in this respect, but while correlation quantifies linear dependence between two variables, mutual information quantifies dependence that may not necessarily be linear. Mathematically, for two continuous random variables X and Y, the mutual information MI(X,Y) is defined as in Equation 5 (Cover & Thomas, 2006).

$$MI(X,Y) = \iint p(x,y)\, log \frac{p(x,y)}{p(x)\, p(y)}\, dx\, dy \tag{5}$$

where p(x, y) is the joint probability density function and p(x) and p(y) are marginal probability density functions. In the time series context x and y may be the same variable at different lags. MI can be unbounded (infinite) and numerical estimation of mutual information from a sample involves discretization and binning, to approximate the probabilities and evaluate the integral above based on bin frequencies. Results depend on the chosen bin boundaries and thus comparison of numeric MI differences between ensembles should use consistent binning. Here, we used the optimal bin width suggested by (Scott, 2015), which depends on the standard deviation and the number of data values (see for example Gong et al., 2014). We then used the R *entropy* package (Hausser & Strimmer, 2021) to evaluate normalized MI, which is the MI standardized by the entropy of each variable. This metric helps quantify the nonlinear lagged dependence within streamflow ensembles.

Figure 3 illustrates how mutual information and correlation metrics quantify linear and nonlinear dependence between some hypothetical variables with dependence. In Figure 3a, there is a visible linear relationship between x and z so both MI and Cor quantify this relationship with high values. Variables x and t in Figure 3b, on the other hand, are two independent variables without any specific relationship between them so that MI and Cor are close to zero. In Figure 3c, there is an obvious relationship between x and y, however, this relationship is not linear and so the Cor is zero. In this case, the mutual information captures the nonlinear relationship between x and y. This example illustrates the value of including the mutual information metric where there may be nonlinear dependence.

With MI, there is no a-priori expectation that dependence should be linear, but with small sample sizes, as is typical for streamflow, the data may be insufficient to discern small nonlinear dependence robustly with statistical significance.

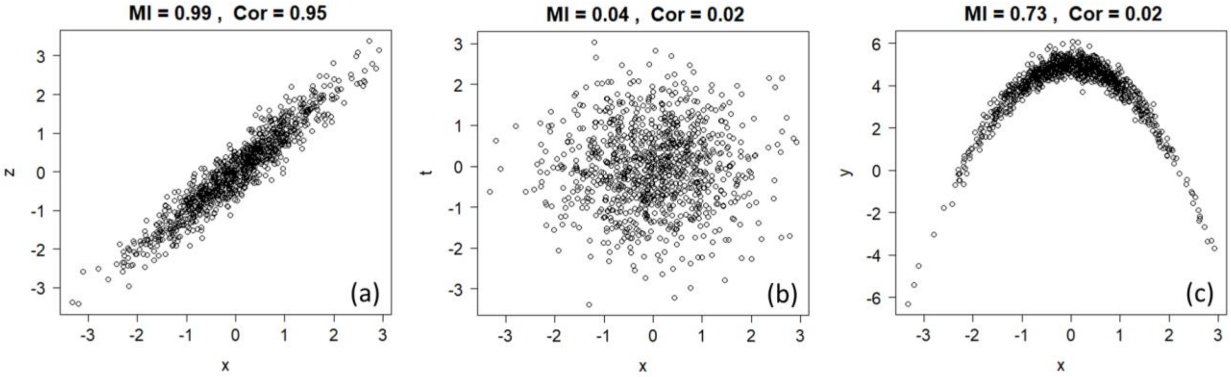


Figure 3. Mutual information (MI) and correlation (Cor) of some hypothetical variables of x, y, t, and z. (a) Two variables with a visible linear relationship. (b) Two independent variables. (c) Two variables with a visible but not linear relationship.

3.10. Reservoir Storage-Yield and Reliability

Reservoir storage-yield and reliability analysis illustrate responses of streamflow ensembles to a set of desired yields and reliabilities. This metric captures the storage attributes of the ensemble at an abstract level distinct from particular reservoir sizing or operation policies. Reservoir storage-yield analysis has traditionally been used to determine the minimum active storage capacity required for delivery of a constant yield rate with a given reliability or alternatively, the yield that can be supplied from a reservoir with a known storage capacity (Loucks et al., 2017). Here, the reliability indicates the probability that the reservoir yields are met. Given the natural variability of streamflow, which may increase due to climate change, it is unclear how well reservoirs are able to ensure the delivery of specified yields with the desired reliabilities (Kuria & Vogel, 2014). These metrics help quantify the variability of yields and reliabilities due to streamflow variability.

Given a time series of reservoir inflows, a computation based on mass balance may be used to determine the reservoir storage required to meet a certain specified yield or release. Let R_t denote the release volume at each time step t, Q_t denote the inflow volume at t, and K_t denote the storage needed at the end of t, with $K_0 = 0$. Then, K_t is calculated by Equation 6.

$$\begin{cases} K_t = K_{t-1} + R_t - Q_t & if\ positive, \\ K_t = 0 & otherwise \end{cases} \tag{6}$$

If K_t from this equation is negative, it indicates that inflow was higher than release plus available unfilled storage capacity. This means that release can be met with available inflow during that time step and there is no need for additional storage, and so K_t reset to 0. For a given series of inflows, the maximum of all K_t is the active storage capacity, S, required to sustain the specified releases or yield. A storage-yield curve is constructed by calculating S for a series of yields. After the storage-yield analysis, reservoir reliability can be evaluated. A reservoir reliability plot shows the probability that the storage required to meet a specified yield is less than a given value *S*.

3.11. Ward's Agglomerative Hierarchical Clustering method

Ward's Agglomerative Hierarchical Clustering method (hereafter Ward's method) was used to categorize the ensembles based on the metrics calculated (Hastie et al., 2009; Murtagh & Contreras, 2012). Ward's method is a bottom-up clustering (or classification) method in which each object (streamflow ensemble or metric in our case) is treated as a single cluster at the beginning of the algorithm. Then, pairs of clusters are merged (or agglomerated) until all clusters are merged into a single cluster containing all the objects. To choose the pair of clusters to merge at each step, Ward's method uses the minimum sum-of-squares as a distance (similarity) criterion that determines how close (similar) or far (dissimilar) the clusters are. The hierarchy of clusters can be shown as a tree (or dendrogram). In dendrograms, the X-axis represents the objects and the Y-axis represents the distance at which the clusters merge. The similar objects with minimum distance fall in the same cluster, and the dissimilar objects are placed farther in the hierarchy. We used the R package *pheatmap* to perform Ward's method (Kolde, 2019).

4. Results

We calculated all the metrics outlined in the preceding section for 21 streamflow ensembles available for the Colorado River Basin (Table 1). We employed these metrics for three primary purposes: 1) to provide a quantitative description of each individual ensemble, 2) to conduct comparisons among ensembles, identifying those that closely align with the desired attributes required for various planning purposes, and 3) to classify ensembles based on their characteristics.

In this section, we present and explain the metrics for one individual ensemble in detail, namely ISM_1906_2020. We selected this ensemble for a thorough explanation here because it is widely used in Colorado River Basin studies and because is easy to understand as it is a resampling of the full historical record, making it good for illustrating how the metrics work. The results for the remaining ensembles are available in the online Supporting Information and the code for generating these metrics is in HydroShare (Salehabadi & Tarboton, 2024). Then, we provide ensemble comparison results, where we have calculated a specific metric for all ensembles and presented them in a single plot. The metrics presented quantify the statistical characteristics of streamflow ensembles, providing a quantitative foundation for interpreting and analyzing their similarities and differences. As each ensemble comprises multiple time series, the metric ranges calculated for each ensemble are depicted using box plots. These ranges quantify the uncertainty in each metric, useful when comparing ensembles. Note that in this paper the box plots use R defaults (R Core Team, 2023), where boxes represent the central half of the data, with whiskers extending to 1.5 times the interquartile range, and outliers beyond the whiskers are displayed as individual dots.

4.1. Ensemble-Specific Metrics

Figure 4 through Figure 8 present the metrics calculated for the Full Hydrology Index Sequential Method ensemble labeled as "ISM_1906_2020". This ensemble comprises 115 time series, generated using the Index Sequential Method (ISM) as described by Ouarda et al. (1997)

and illustrated by Salehabadi et al. (2020). To generate this ensemble, ISM was applied to the full observed record from 1906 to 2020. The length of each time series within the ensemble is set by a designated planning period taken as 50 years here.

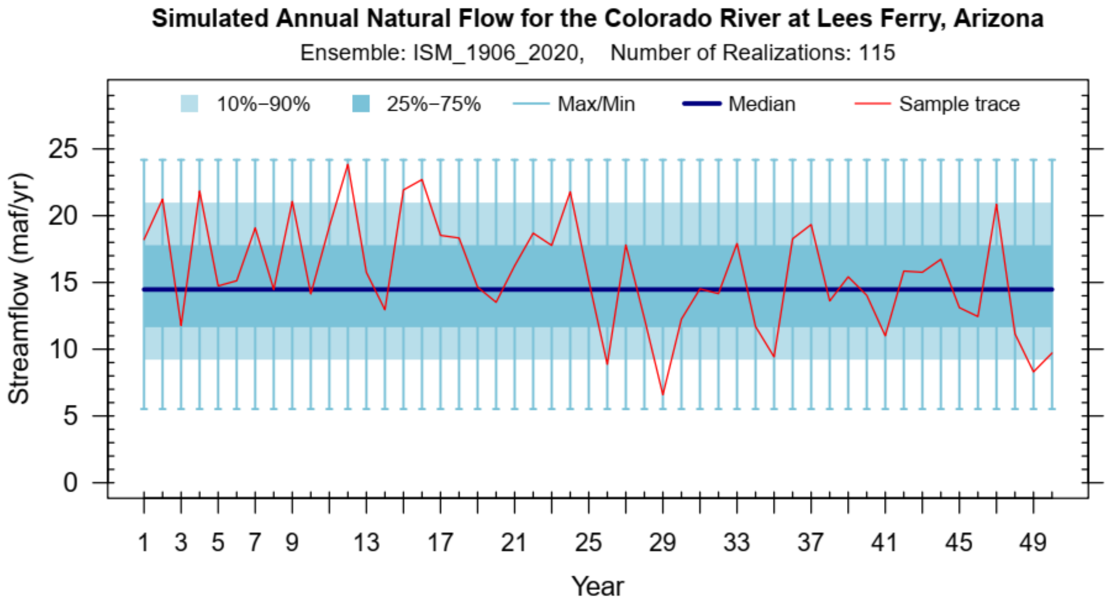


Figure 4. Time series of the simulated annual natural flow at Lees Ferry for the ISM_1906_2020 ensemble. This figure shows 10th to 90th percentiles (light blue area), and 25th to 75th percentiles (dark blue area), maximum and minimum (whiskers), median (navy line), and a sample sequence from the ensemble (red line).

The results show that simulated annual natural flows are in the range of 5 to 25 maf/yr and there is no trend or variability in the distribution of the annual flows during the planning period (Figure 4), as expected since ISM is a recycling of historical flow sequences. The ensemble has a mean of 14.5 maf/yr (Figure 5a) with a standard deviation of about one-third of the mean, similar to the observed record (Figure 5d). Minimum annual flows are bounded by the historical minimum annual flow of 5.5 maf/yr, showing that the ensemble does not have any years with flows less than what has previously been observed (Figure 5b). Maximum annual flows with a range from 21 to 24.2 maf/yr (Figure 5c) and the average count above threshold (1.3 years per decade, Figure 5l) indicate the frequency of high-flow years in the ensemble, which here is the same as the historical high-flow year frequency.

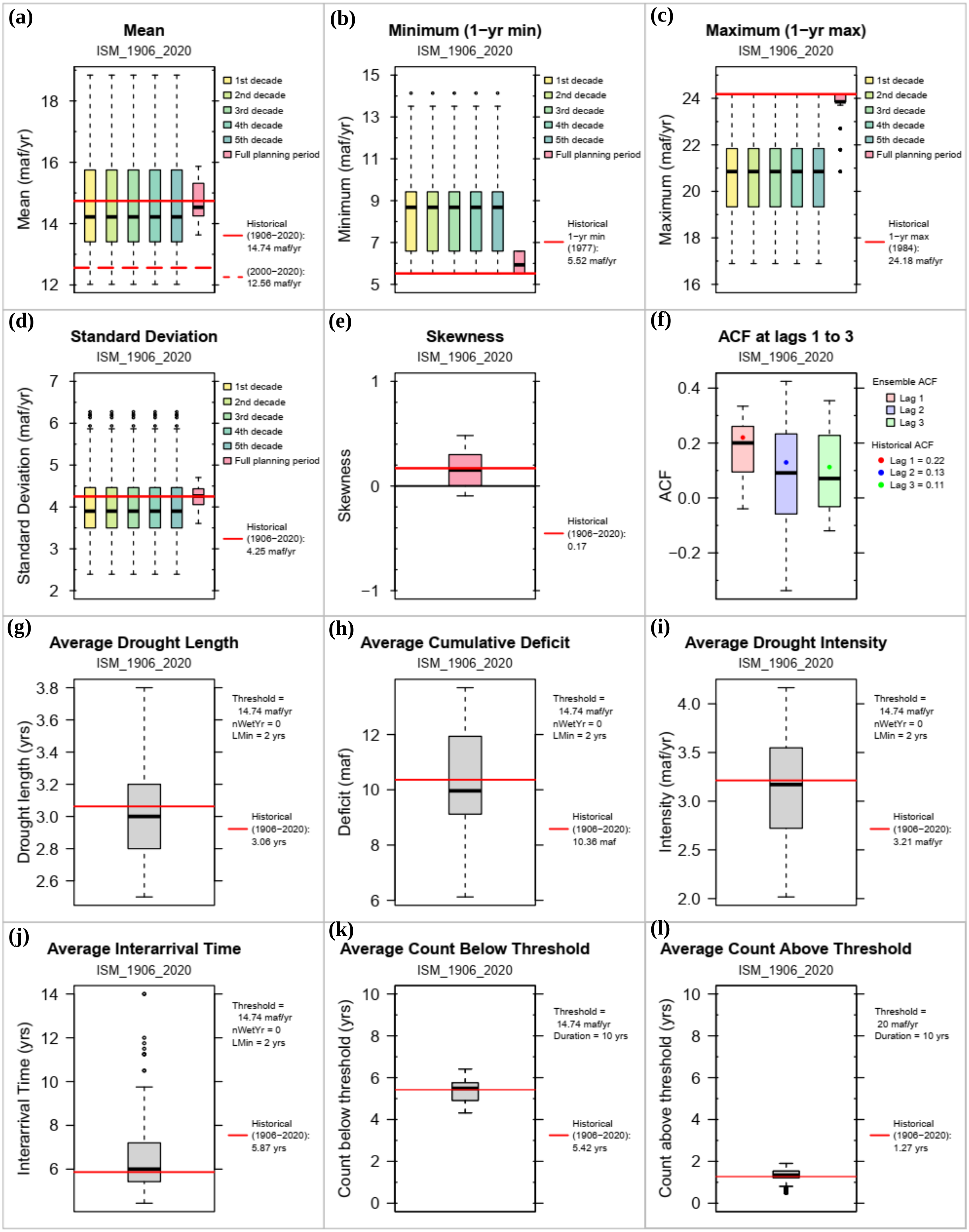


Figure 5. Summary metrics of simulated annual natural flow at Lees Ferry for the ISM_1906_2020 ensemble

The ensemble has a positive skewness of 0.15, equal to that of the historical record (Figure 5e). For a 50-year record, skewness needs to exceed a value of 0.66 to be statistically different from zero with a 95% confidence level. Thus, for this ensemble, the skewness is considered not significantly different from zero. Nevertheless, it is retained as a metric to provide historical context for other ensembles. Positive skewness means that, on average, there will be more flows below the mean than flows above the mean. This characteristic is also quantified using the count below threshold metric.

The ACF results show that the historical lag 1 to 3 correlation of the historical record are reproduced in this ensemble (Figure 5f). The lag-1 correlation of the ensemble is centered on the historical correlation value of 0.2. For a 115-year record, the threshold for statistical significance with 95% confidence is $1.96/\sqrt{n} = 0.18$, indicating that lag-1 correlation is statistically different from zero. For the ensemble members that have 50 years of data, the threshold for statistical significance with 95% confidence is $1.96/\sqrt{n} = 0.28$, indicating that we cannot discern this as being statistically different from zero. This is reflected in the range of the box whiskers crossing the zero axis, but from the pattern with historical dots within the box ranges we can see that historical correlations are reproduced.

Drought event statistics (drought length, cumulative deficit, intensity, and interarrival time) quantify characteristics of droughts, defined by consecutive years during which the annual flow remains below the historical long-term average (i.e. 14.74 maf/yr as the specified threshold). The results in Figure 5g-j indicate that, overall, drought event characteristics in the ensemble are very similar to droughts in the historical record. Therefore, this ensemble is representative of drought events that have occurred in the past. Note that these statistics break a sustained dry period into separate events when one (or a selected number) of years exceed the threshold.

Average count below/above threshold (Figure 5k and l) quantifies the average number of years in a decade with flows below/above a threshold. Below threshold years were counted using a threshold of 14.74 maf/yr, the long-term mean. Above threshold years were counted using a threshold of 20 maf/yr. This value is close to the highest flow occurring in the 21st century millennium drought period, which has been the worst 21-year drought that has occurred based on the observed record (Salehabadi et al., 2022), and by using this threshold, this metric helps evaluate whether an ensemble has occasional high flows at a higher or lower frequency than this period. Counts are reported as an average over 10-year durations. In this ensemble, on average, half of the years in each decade of the planning period are low-flow years (< 14.74 maf/yr) and one year in a decade is high-flow (> 20 maf/yr). These are similar to the number of low/high flow years in the full observed record. For this ISM-based ensemble, the moving count below/above threshold is flat, showing the lack of variability in the number of low/high flow years during various decades of the planning period (Supporting Information Figures S1 and S2).

Duration-severity analysis (Figure 6) was used as a more general approach to quantify droughts, regardless of the occurrence of wet years during the dry periods. Duration-severity analysis shows how the lowest mean flows may vary for different durations (from 1 to 25 years) and where the range of extreme droughts in the ensemble sit with respect to the observed and paleo-reconstructed flows. The results indicate that extreme droughts in the ensemble are aligned with those in the observed record, and the ensemble does not have droughts any more severe than previously observed in the last century. However, the paleo-reconstructed flow data (dates back to 1416) does contain droughts more severe than droughts in both the observed record and the ensemble across the full range of durations from 1 to 25 years depicted. The need to plan for potential recurrence of droughts as severe as in the paleo record, and potentially even more severe droughts associated with warming, suggests that this ISM_1906_2020 ensemble is not suitable for these planning purposes.

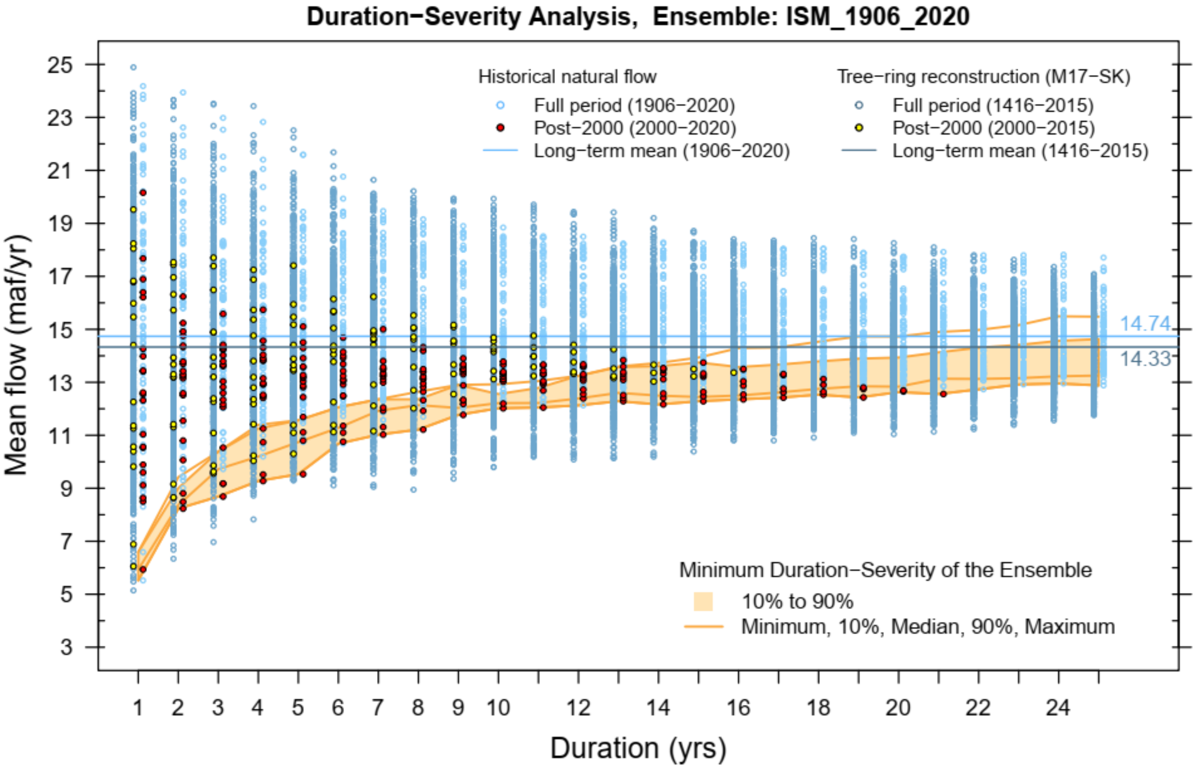


Figure 6. Duration-severity analysis; Overlaying the range of extreme droughts (quantified as the minimum duration-severity) within the ISM_1906_2020 ensemble (orange area) on the duration-severity plot of the observed (light dots) and tree-ring-reconstructed (dark dots) natural flows at Lees Ferry. The spread of the orange area illustrates how the ensemble's extreme droughts may vary across various durations, comparing them with the historical and tree-ring-reconstructed records. Each dot represents mean annual flow averaged over the duration on the x-axis. There is a dot for each duration (including overlaps) within the record.

Reservoir storage-yield and reliability results illustrate responses of the streamflow ensemble to a set of desired yields and reliabilities (Figure 7). The metric captures the storage attributes of the ensemble at an abstract level distinct from particular reservoir sizing or operation policies. The results show that under this streamflow ensemble, an active storage capacity of 60 maf (close to the combined storage capacity of all major reservoirs in the basin) is required to provide a yield of 15 maf/yr with 90% of reliability during 50 years of the planning period. The yield of 15 maf/yr is equal to the total water allocated by the Law of the River to the Upper and Lower Basins (7.5 maf to each basin, not including 1.5 maf to Mexico). This indicates that, even under the ISM_1906_2020 ensemble, which is based on the full observed record including the early 20th-century pluvial period of unusually high flows, a high storage capacity is needed to meet the Law of the River. In the case of meeting a yield of 13.5 maf/yr (which is the sum of Upper Basin’s average consumptive uses and losses of 4.4 maf/yr and 9 maf/yr of normal allocation in the Lower Basin and Mexico) with 90% of reliability, an active storage capacity of 20 maf is needed.

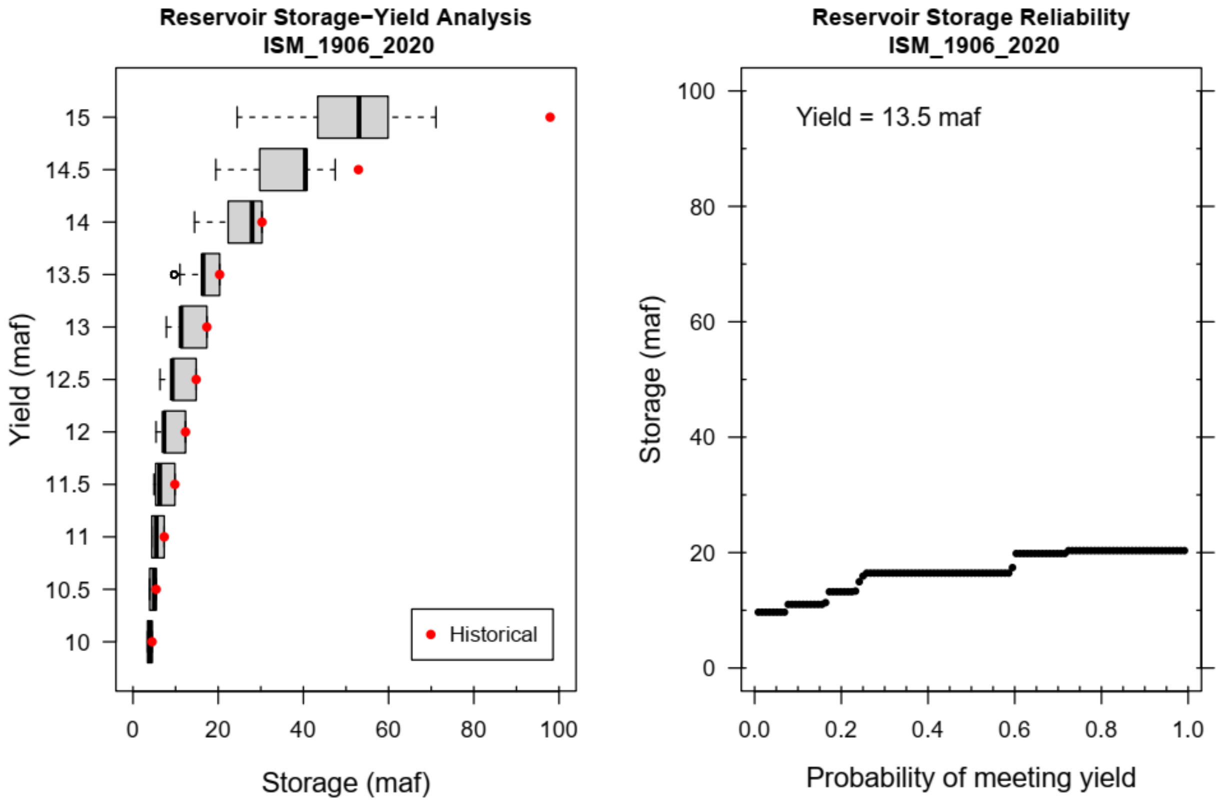


Figure 7. Reservoir storage-yield and reliability analysis for the ISM_1906_2020 ensemble. These plots illustrate the response of the streamflow ensemble to a set of desired yields and reliabilities. The metric captures the storage attributes of the streamflow ensemble at an abstract level distinct from particular reservoir sizing or operation policies. The plot on the left shows the storage needed for releasing the desired yields shown on the y axis. The plot on the right shows the storage needed for a specific yield and desired reliabilities.

The Hurst coefficient for this ensemble is centered around 0.77, denoting a long-term structure in its dependence. However, due to the short evaluation period (50 years), the uncertainty in this coefficient limits its interpretation. Nevertheless, when compared to the historical record, this ensemble shows similarity in long-term persistence quantified with the Hurst coefficient (Figure 8).

Overall, based on the metrics calculated, this ensemble will only test the system for flows already experienced. This was expected since this ISM-based ensemble is a recycling of historical flow sequences. This ensemble does not explore a sample space where the mean may have changed, or minima/maxima may go beyond the historical record, or droughts may be more severe or sustained than the historical record. Thus, based on this set of metrics, this ensemble is assessed to not provide enough variability to fulfill drought planning needs.

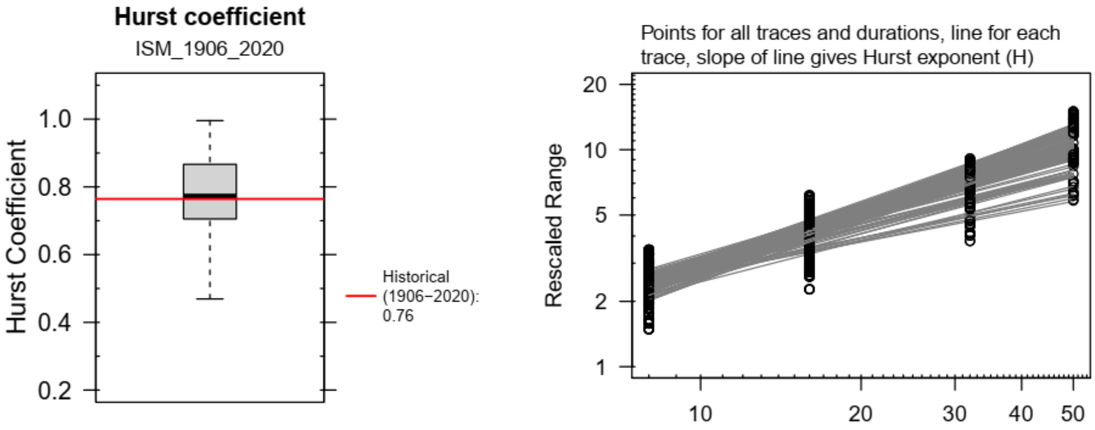


Figure 8. Hurst coefficient for the ISM_1906_2020 ensemble

4.2. Comparison Results

Figure 9 shows the ranges of decadal mean (yellow to green boxes) and full 50-year period mean (pink boxes) of the 21 ensembles. The mean ranges show how dry or wet the ensembles are, compared with each other and the historical long-term mean of 14.74 maf/yr (solid red line).

In the ISM_1906_2020, AR1, NPC_1906_2020, and CMIP5_BCSD ensembles, the medians of simulated means closely match the historical long-term mean (Figure 9). These ensembles are thus consistent with an assumption of stationarity of the mean, as the historical mean is preserved in the simulations. Note though that CMIP5_BCSD 10-year means have greater spread than the other ensembles, indicating that this ensemble has increased variability. The other ensembles, however, deviate from stationarity of the mean with means less than the historical mean, indicating drier conditions. Among these, TempAdj_RCP4.5_10% and TempAdj_RCP8.5_10% are the driest ensembles, with mean flows lower than even the millennium drought mean (as shown by dashed red line in Figure 9).

In the ISM-based ensembles, the stationarity of the simulated decadal mean values is clearly evident. These ensembles consistently provide similar mean flow ranges across various decades. On the other hand, in the temperature-adjusted flow ensembles (i.e. TempAdj_RCP), decadal mean values uniformly decrease, indicating a projected decrease.

Among the ensembles, those based on CMIP (i.e. climate change-informed hydrology including CMIP3_BCSD, CMIP5_BCSD, and CMIP5_LOCA) exhibit the widest mean ranges and uncertainties (Figure 9). One significant source of uncertainty in CMIP flow projections is the downscaling process, which involves adapting coarse-resolution GCM outputs for high-resolution hydrology models (Lukas et al., 2020). This downscaling-related uncertainty is evident when comparing the simulated mean values of the CMIP5_BCSD and CMIP5_LOCA ensembles. Interestingly, despite their common CMIP5 source, the choice of downscaling method (BCSD or LOCA) results in variations in the mean values, with CMIP5_BCSD showing a higher mean (closer to the full observed record mean) than CMIP5_LOCA (closer to the millennium drought mean). This is consistent with findings from other studies, such as Vano et al. (2020), which thoroughly compared downscaled LOCA and BCSD projections.

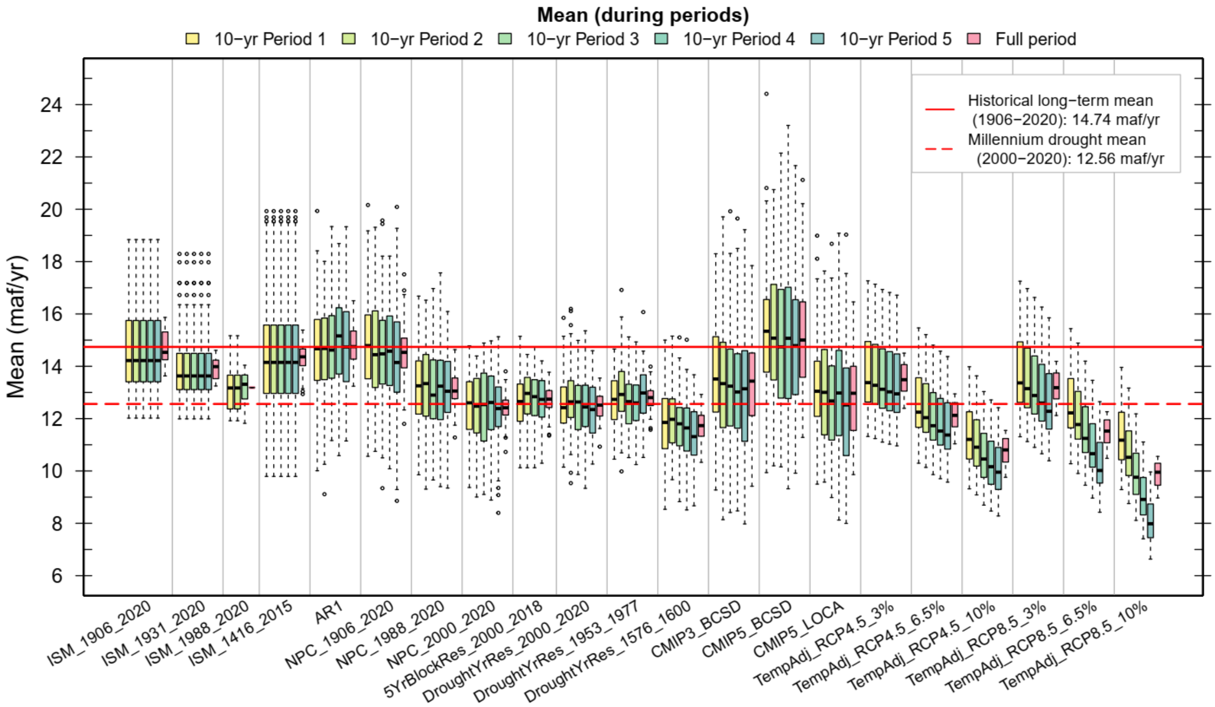


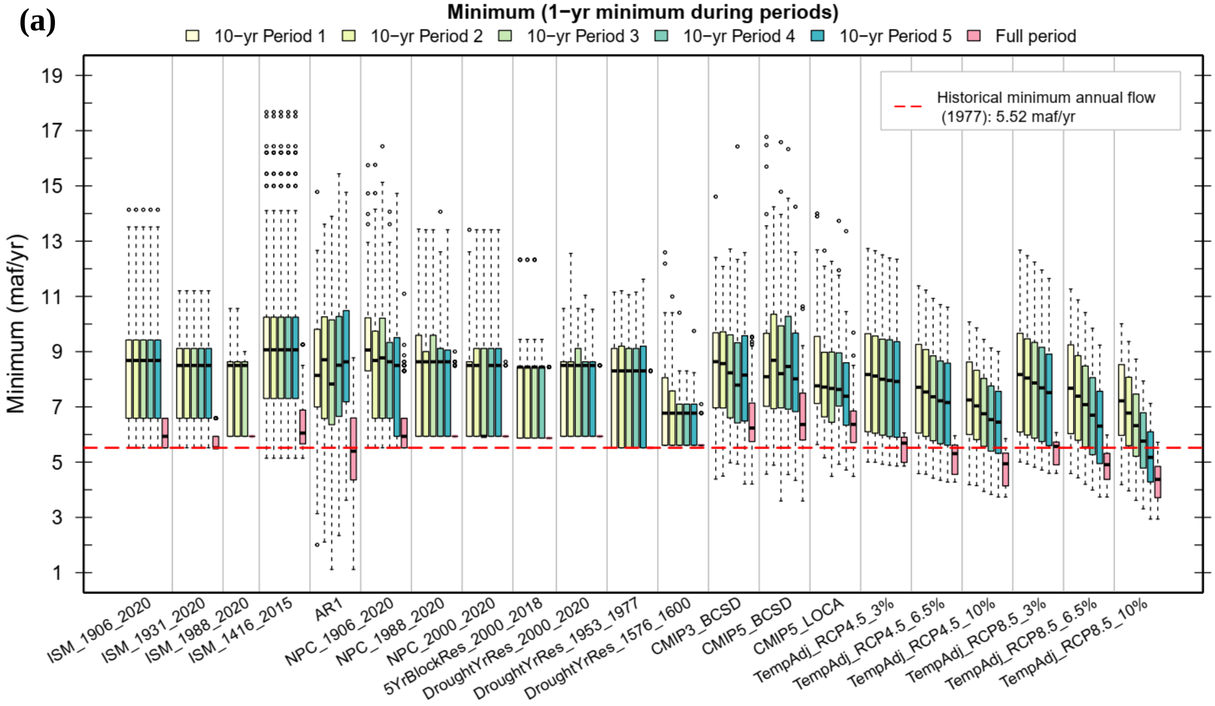
Figure 9. Mean of streamflow ensembles along with the long-term mean of the historical full record (1906-2020, solid red line) and the millennium drought mean (2000-2020, dashed red line). Yellow to green boxes of each ensemble show decadal mean and the pink boxes indicates the mean of full planning period.

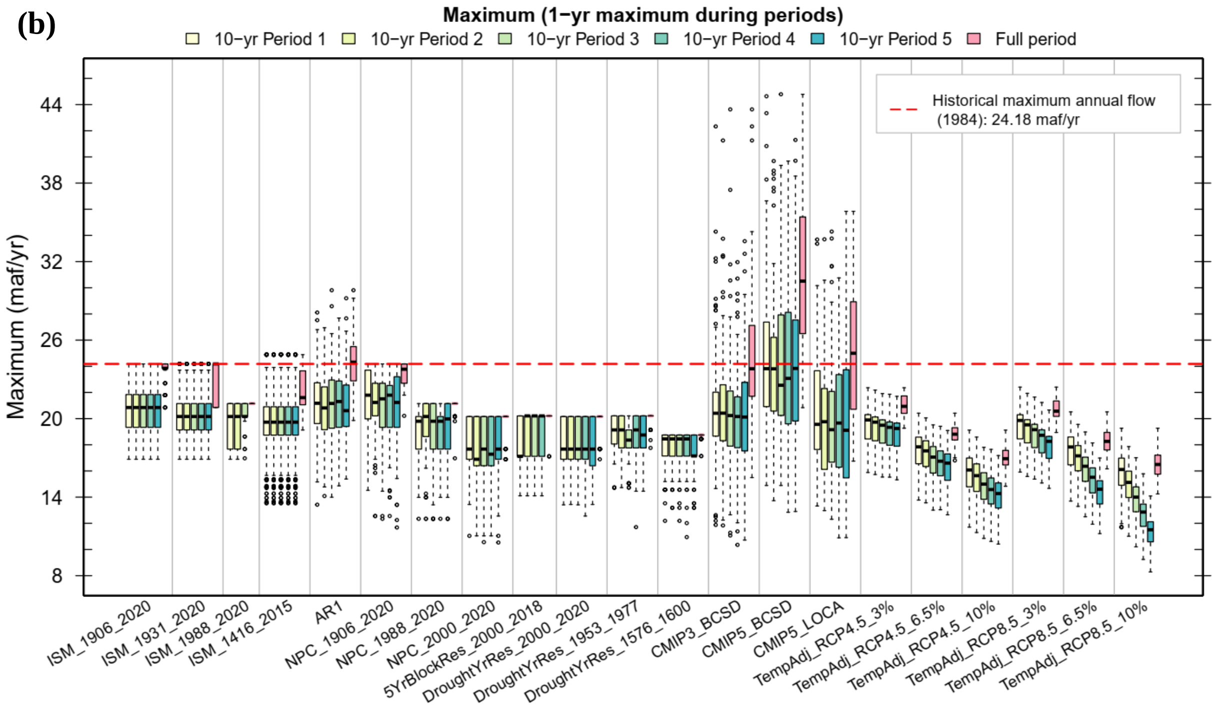
The minimum flow is a commonly used metric, particularly valuable when the purpose of using the streamflow ensembles is drought management. When the objective is to plan for future scenarios with low-flow years, the minimum flow serves as a crucial metric for quantifying and comparing ensembles, aiding in ensemble selection. Figure 10a shows the ranges of minimum one-year flow in decadal periods (yellow to green boxes) as well as during the full period (pink boxes). The results indicate that half of the ensembles (i.e., ISM-, NPC-, and Drought-based ensembles) are constrained to the historical minimum annual flow of 5.5 maf/yr (as shown by the red dashed line in Figure 10a). Furthermore, these ensembles exhibit limited variability in decadal minimum annual flows. Consequently, if the objective is to plan for or accommodate annual flows lower than historical records or to introduce some diversity in decadal minimum annual flows, these particular ensembles may not be the most suitable choices.

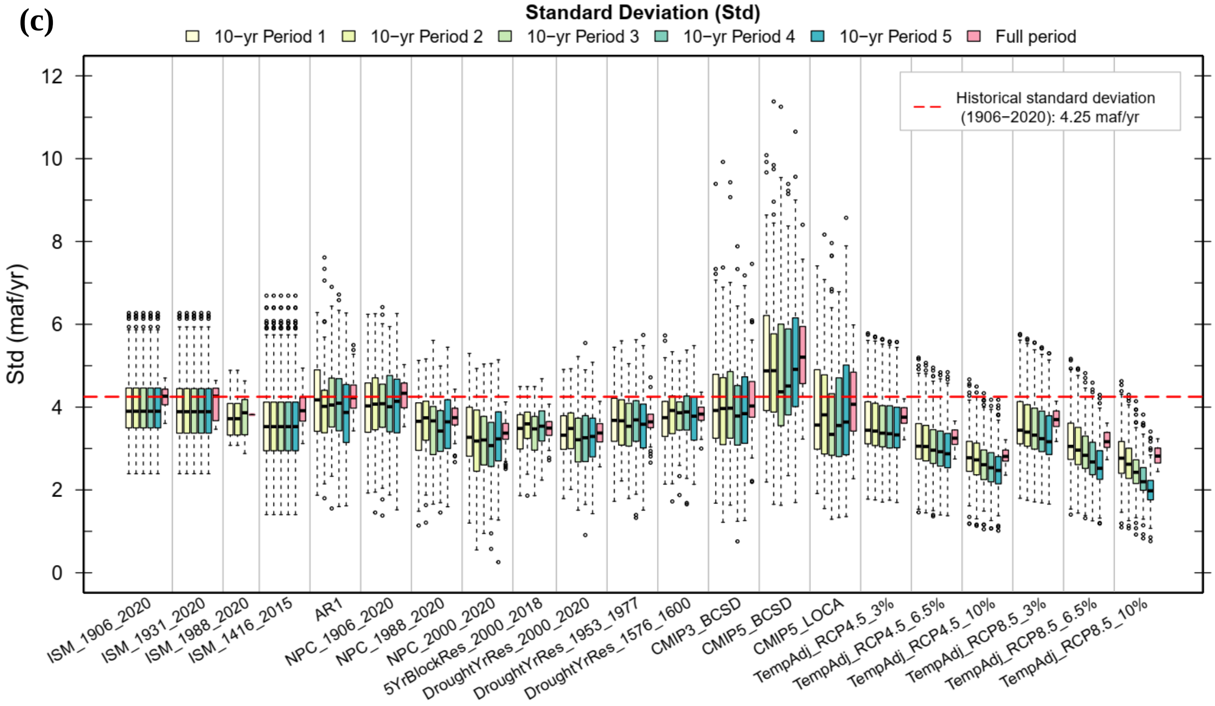
Maximum is another frequently used metric for assessing the upper boundaries of annual flows within the ensembles. This metric is particularly valuable when selecting ensembles for planning wet periods or comparing maximum annual flows among various dry ensembles. The results show that the majority of the ensembles have high flows lower than the historical maximum of 24.18 maf/yr (Figure 10b). In contract, the CMIP-based ensembles have the highest annual flows. There are significant differences in maximum annual flows within the CMIP5_BCSD and CMIP5_LOCA ensembles, highlighting the effect of downscaling-related uncertainty on these flow projections.

The standard deviation of the ensembles shows that the historical standard deviation of 4.25 maf/yr is preserved in those ensembles that use the full historical flow record to generate the flow sequences, except for the TempAdj ensembles (Figure 10c). Within the TempAdj ensembles, the proportionally reduction of historical natural flow in response to future temperature projections leads to a notable decline in standard deviations. This decreasing trend in variability over time may make these ensembles less suitable for planning purposes that require a broader range of variability when considering a changing future. In contrast, the CMIP5_BCSD ensemble has the highest standard deviation, higher than the variability provided by CMIP5_LOCA.

Figure 10d shows skewness calculated for the ensembles. The ISM_1906_2020 and ISM_1416_2015 results indicate the skewness of the historical and paleo data evaluated over 50-year intervals. The skewness values are mostly centered close to 0, indicating almost no skewness, but the range spanned by the boxes reveals the sampling variability in the skewness calculated within the 50-year intervals. Comparison between ensembles indicates that most of them have positive skewness (Figure 10d), showing that the simulated flows are more toward the values lower than the mean and median.




(b)
Maximum (1−yr maximum during periods)
10−yr Period 1
10−yr Period 2
10−yr Period 3
10−yr Period 4
10−yr Period 5
Full period
Historical maximum annual flow (1984): 24.18 maf/yr
Maximum (maf/yr)
44
38
32
26
20
14
8
ISM_1906_2020
ISM_1931_2020
ISM_1988_2020
ISM_1416_2015
AR1
NPC_1906_2020
NPC_1988_2020
NPC_2000_2020
5YrBlockRes_2000_2018
DroughtYrRes_2000_2020
DroughtYrRes_1953_1977
DroughtYrRes_1576_1600
CMIP3_BCSD
CMIP5_BCSD
CMIP5_LOCA
TempAdj_RCP4.5_3%
TempAdj_RCP4.5_6.5%
TempAdj_RCP4.5_10%
TempAdj_RCP8.5_3%
TempAdj_RCP8.5_6.5%
TempAdj_RCP8.5_10%


(c)
Standard Deviation (Std)
10−yr Period 1
10−yr Period 2
10−yr Period 3
10−yr Period 4
10−yr Period 5
Full period
Historical standard deviation (1906−2020): 4.25 maf/yr
Std (maf/yr)
12
10
8
6
4
2
0
ISM_1906_2020
ISM_1931_2020
ISM_1988_2020
ISM_1416_2015
AR1
NPC_1906_2020
NPC_1988_2020
NPC_2000_2020
5YrBlockRes_2000_2018
DroughtYrRes_2000_2020
DroughtYrRes_1953_1977
DroughtYrRes_1576_1600
CMIP3_BCSD
CMIP5_BCSD
CMIP5_LOCA
TempAdj_RCP4.5_3%
TempAdj_RCP4.5_6.5%
TempAdj_RCP4.5_10%
TempAdj_RCP8.5_3%
TempAdj_RCP8.5_6.5%
TempAdj_RCP8.5_10%

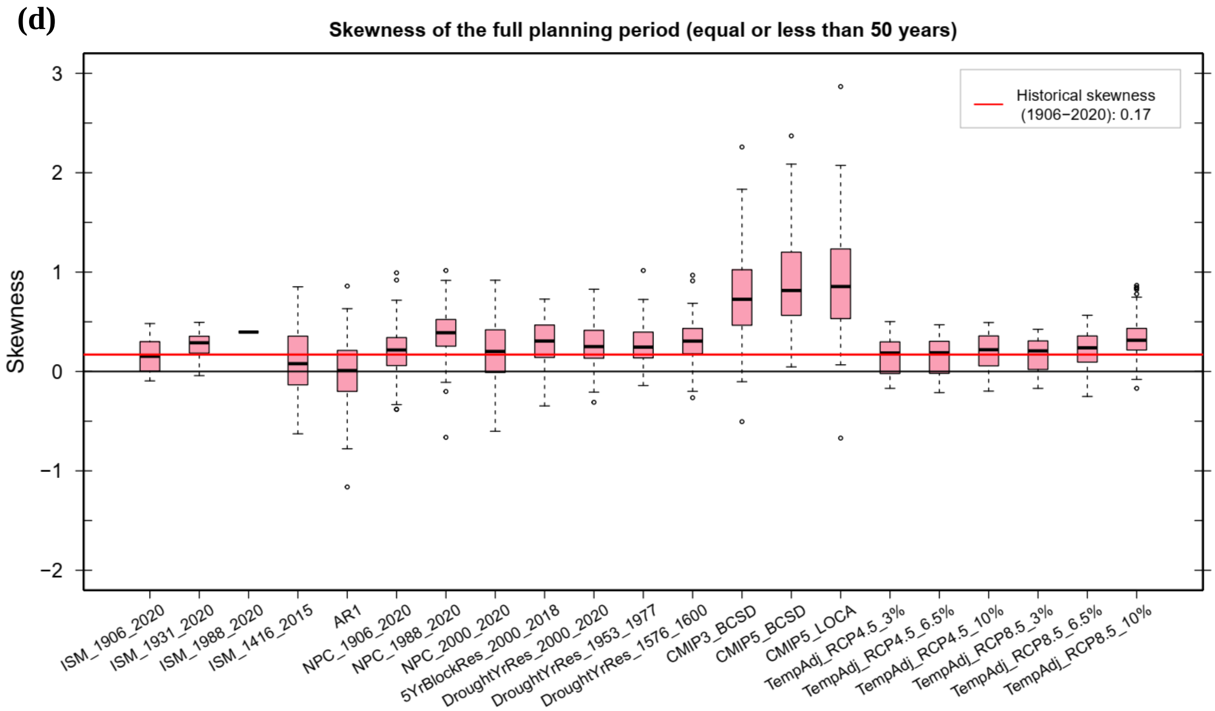


Figure 10. Common metrics for the streamflow ensembles: (a) minimum, (b) maximum, (c) standard deviation, and (d) skewness. Yellow to green boxes show decadal metric and the pink boxes are for the full planning period.

Figure 11 illustrates lags 1 to 3 correlation ranges of the ensembles, alongside the historical correlation. The results indicate that historical lag-1 correlation is not preserved the following ensembles: ISM_1988_2020, ISM_1416_2015, NPC_1988_2020, NPC_2000_2020, 5YrBlockRes_2000_2018, three DroughtYrRes ensembles, and TempAdj_RCP8.5_10%. While not reproducing lag 1 correlation may not disqualify the use of these ensembles, it does differentiate them. It should also be noted that, for a series length of 50 years, the significance level is 0.28, encompassing a wide-range of correlations to be considered significant. The PACF measures correlations at higher lags that are not directly influenced by lower lag correlations (Figure 12). Since lag-2 and higher correlations are generally low and rarely statistically significantly different from 0, the PACF higher lag values also tend to be low and lack significant deviations from 0, offering limited additional information beyond what is observed in the ACF.

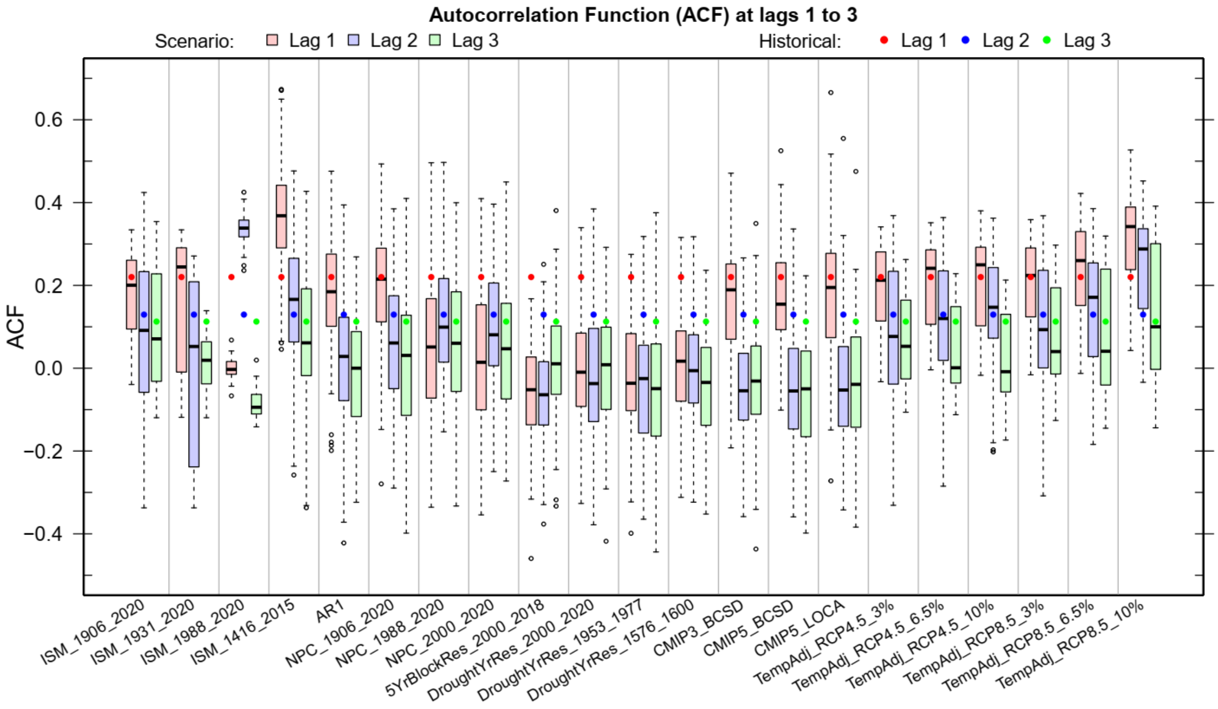


Figure 11. Autocorrelation function (ACF) at lags one to three for the streamflow ensembles

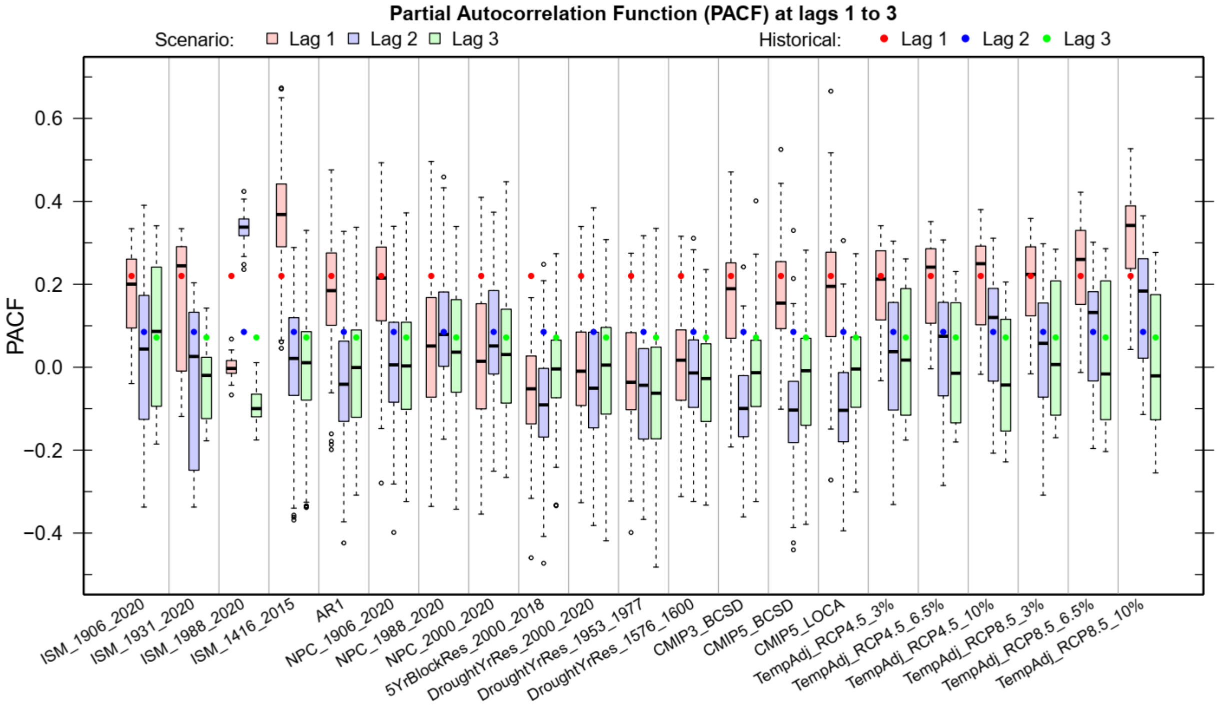


Figure 12. Partial Autocorrelation Function (PACF) at lags one to three for the ensembles

The Hurst coefficient for the ensembles we evaluated is shown in Figure 13. All ensembles have a length of 50 years, except ISM_1988_2020 and 5YrBlockRes_2000_2018, which span shorter periods of 33 and 42 years, respectively. Ideally, for accurate Hurst coefficient comparisons, the period should be consistent, as the computed value is dependent on the period length. The results show that the Hurst coefficient for ISM_1906_2020 effectively mirrors the Hurst coefficient for historical data assessed over 50-year periods, with the box range indicating uncertainty. Many of the evaluated ensembles exhibit box ranges lower than the historical Hurst coefficient, indicating that they are not preserving persistence. Ensembles that do maintain persistence include ISM_1906_2020, ISM_1416_2015, AR1, three NPC-based ensembles, CMIP5_LOCA, and six temperature-adjusted ensembles (identified by TempAdj_RCP at the beginning of their names on the plot).

Reservoir Storage-Yield and Reliability analysis was used to compare the streamflow variability in the ensembles. As discussed previously, Figure 7 shows reservoir storage-yield and reliability analysis for the ISM_1906_2020 ensemble. The results for the other ensembles are in the Supporting Information. When comparing ensembles representative of the full historical record (i.e., ISM_1906_2020, AR1, NPC_1906_2020), it becomes evident that the NPC_1906_2020 ensemble requires more storage to achieve a specific yield, suggesting that the NPC_1906_2020 ensemble is characterized by higher persistence (Figure 7, Figures S29, and S36 in Supporting Information).

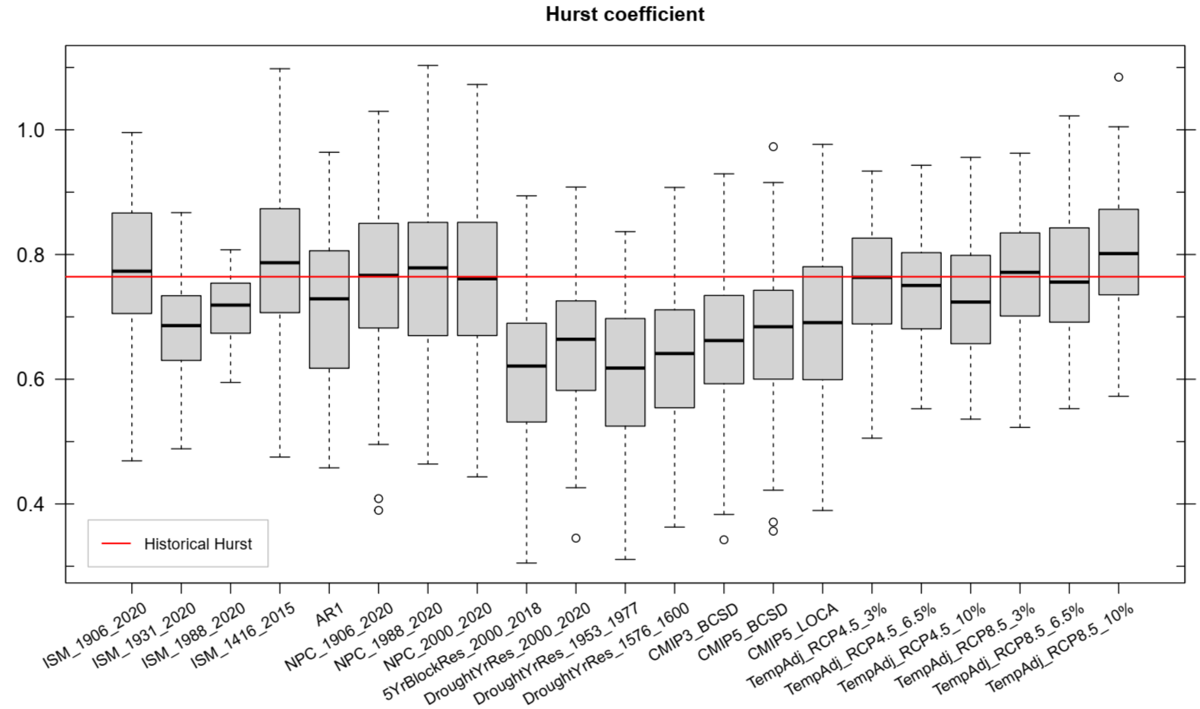
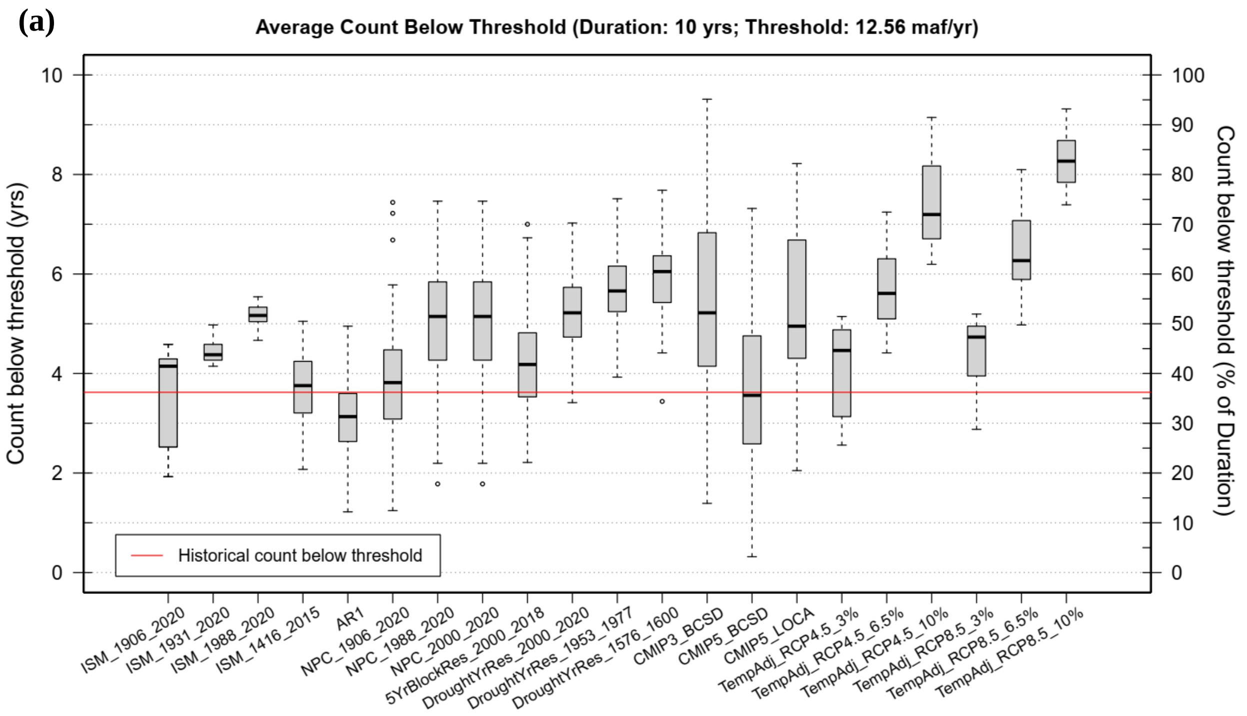


Figure 13. Hurst coefficient for the streamflow ensembles (box plots) along with the historical Hurst coefficient (red line)

The count below threshold metric, CBT, metric was calculated as the average number of years within 10-year durations with annual flows falling below a threshold of 12.56 maf/yr, representing the 21st-century average flow (Figure 14a). In general, ensembles with lower mean flow tend to have a higher CBT. However, there are exceptions to this pattern. Comparison of the millennium-drought-based ensembles (i.e. NPC_2000_2020, 5YrBlockRes_2000_2018, and DroughtYrRes_2000_2020) shows that, despite having similar mean values and other previously assessed metrics, the 5YrBlockRes_2000_2018 ensemble has fewer years below the threshold compared to the other two ensembles.

Similarly, the count above threshold, CAT, were calculated as the average number of years within 10-year durations with annual flows exceeding a threshold of 20 maf/yr, representing the 21st-century maximum annual flow (Figure 14b). The CAT results indicate that most ensembles have a lower frequency of high flows compared to the full observed record. A comparison between ISM_1906_2020 and ISM_1931_2020 shows that excluding the first 24 years of the observed record (i.e. 1906-1931, known as the unusual pluvial period) in the ISM_1931_2020 flow generation results in a 50% decrease in the number of high flows. The ISM_1931_2020 high-flow frequency is more similar to ISM_1416_2015, an ensemble based on paleo-reconstructed flows extending the historical data up to 1416. The results also highlight the limitation of some ensembles in simulating high flows. Ensembles like DroughtYrRes_1576_1600, TempAdj_RCP4.5_10%, and TempAdj_RCP8.5_10% fail to produce high flows at least as high as the maximum annual flow observed in the 21st century. Consequently, these ensembles may not be suitable for planning scenarios that need to account for occasional high flows.



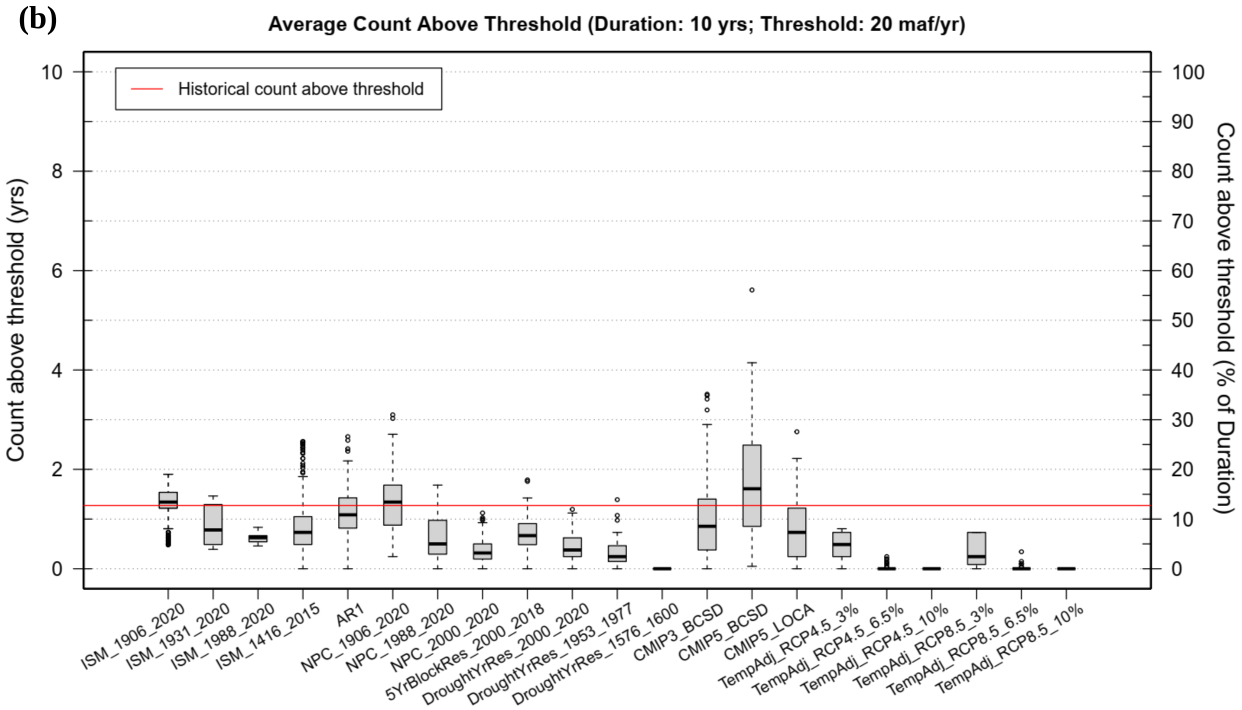


Figure 14. (a) Average count below a threshold of 12.56 maf/yr (21st-century mean flow at Lees Ferry) over 10-year durations. (b) Average count above a threshold of 20 maf/yr over 10-year durations.

Hydrologic drought event statistics were determined using a threshold of 14.74 maf/yr, which represents the historical long-term mean flow. This threshold was employed to identify consecutive years (with a length of two years or more) with flows below this value. Subsequently, we calculated the average drought length, magnitude (cumulative deficit), intensity, and interarrival time, as illustrated in Figure 15. As detailed in the methodology section, one limitation of drought event statistics is that they divide a sustained drought period into distinct events if there is a year that exceeds the threshold. To address this limitation and avoid dependency on a specific threshold, we conducted a duration-severity approach to quantify extreme droughts within the ensembles, regardless of the occasional occurrence of wet years during dry periods. Figure 6 shows duration-severity results for the ISM_1906_2020 ensemble. The results for the other ensembles are in Supporting Information.

Among the ensembles that closely resemble the observed record based on the previously accessed metrics, the ISM_1906_2020 ensemble stands out as the only one that replicates all the available drought event statistics from the observed record (Figure 15). The duration-severity results indicate that extreme droughts in this ensemble closely align with those in the observed record, and the ensemble does not exhibit droughts of greater severity than those observed in the last century (Figure 6). This characteristic makes the ensemble unsuitable for planning in a warmer future with declining flow.

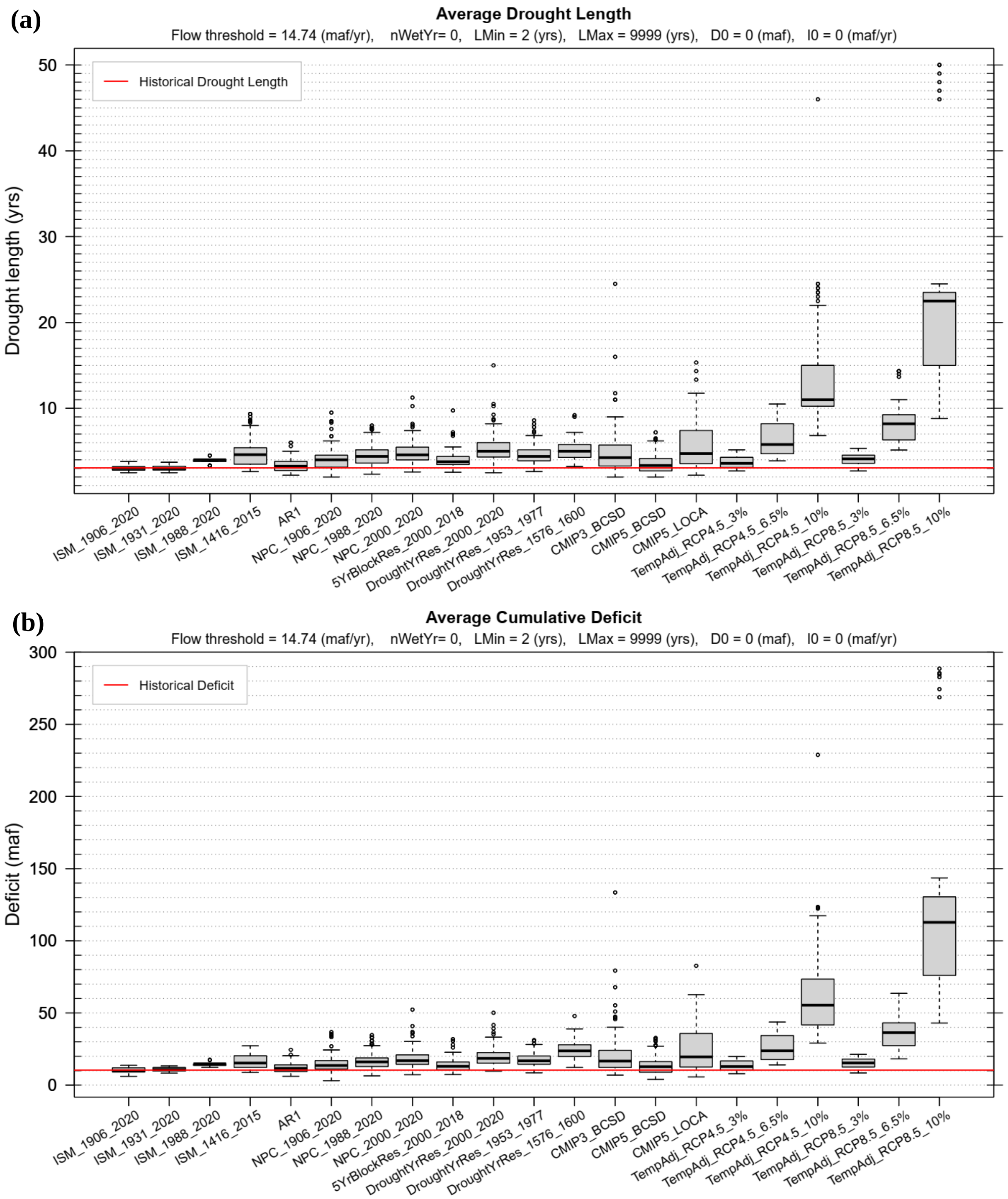
Drought event statistics for the AR1 ensemble indicate that, overall, drought characteristics in this ensemble are very similar to the ISM_1906_2020 ensemble (Figure 15). However, the duration-severity results indicate that extreme droughts more severe than the ISM_1906_2020 is present in the AR1 ensemble (Supporting Information Figure S28). The extreme droughts in the AR1 ensemble are mostly consistent with what has previously occurred in the observed and paleo-reconstructed records. In some short durations (1- and 2-year) however, the unrealistically low mean flows are also available in the AR1 ensemble (Supporting Information Figure S28).

The Paleo ISM ensemble (ISM_1416_2015) has drought length and magnitude higher than the ISM_1906_2020 ensemble, but drought intensity is similar, indicating a similar average deficit in dry years (Figure 15). The duration-severity results for the Paleo ISM ensemble show a wide range of variability for extreme droughts (Supporting Information Figure S21). Along with having extreme droughts similar to those in the observed record, the ensemble also includes more severe droughts similar to the extreme droughts in the paleo estimations. Therefore, this ensemble does provide extreme droughts that are more severe and sustained than what has been observed in the last century. However, there are not any droughts more severe or sustained than the paleo estimates. A warming future may add to the severity of the extreme paleo droughts and such droughts are needed to be considered in future drought planning.

The TempAdj_RCP8.5_10% exhibits the most severe and sustained droughts with the highest length and magnitude (Figure 15). Under this ensemble, there would be, on average, a 5 maf/yr deficit compared to the long-term mean during drought events. Looking at the duration-severity results (Supporting Information Figure S140) also indicates that extreme droughts in this ensemble are significantly more severe than what has previously occurred in the observed and paleo-reconstructed records. Overall, this ensemble stands out as the most extreme in terms of providing drought conditions.

Most of the metrics calculated for the NPC_1906_2020 ensemble are similar to the ISM_1906_2020 ensemble, with more variability in the metrics. The differences between these two ensembles are evident in the extreme droughts quantified by the duration-severity analysis (Figure 6 and Supporting Information Figure S35) and the reservoir storage-yield and reliability analysis (Figure 7 and Supporting Information Figure S36). The duration-severity results for the NPC_1906_2020 ensemble show a wide range of variability for the extreme droughts in which along with extreme droughts similar to those in the observed and paleo records, some more severe and sustained droughts are also available. This indicates that, even by only resampling from the full observed record, extreme droughts as severe and sustained as those in the paleo record can be created in an ensemble. While ISM is not able to produce such extreme droughts and thus is not a reasonable method to use. The extreme droughts available in the NPC_1906_2020 ensemble resulted in needs for higher storage than in the ISM_1906_2020 ensemble to provide yields with more reliability.

Looking at the millennium drought-based ensembles generated using NPC and drought resampling (i.e. NPC_2000_2020 and DroughtYrRes_2000_2020) indicates that these two ensembles are very similar in drought event statistics (Figure 15), but duration-severity analysis reveals the difference (Supporting Information Figures S49 and S63). The DroughtYrRes_2000_2020 ensemble does provide some extreme droughts (less than 10% of the extreme droughts in the ensemble) that are more severe and sustained than the past, but those are not as severe as the extreme droughts in the NPC_2000_2020 ensemble. This is despite these two ensembles being resampled from the same subset of the observed natural flow.


(a)
Average Drought Length
Flow threshold = 14.74 (maf/yr), nWetYr= 0, LMin = 2 (yrs), LMax = 9999 (yrs), D0 = 0 (maf), I0 = 0 (maf/yr)
Historical Drought Length
Drought length (yrs)
10
20
30
40
50
ISM_1906_2020
ISM_1931_2020
ISM_1988_2020
ISM_1416_2015
AR1
NPC_1906_2020
NPC_1988_2020
NPC_2000_2020
5YrBlockRes_2000_2018
DroughtYrRes_2000_2020
DroughtYrRes_1953_1977
DroughtYrRes_1576_1600
CMIP3_BCSD
CMIP5_BCSD
CMIP5_LOCA
TempAdj_RCP4.5_3%
TempAdj_RCP4.5_6.5%
TempAdj_RCP4.5_10%
TempAdj_RCP8.5_3%
TempAdj_RCP8.5_6.5%
TempAdj_RCP8.5_10%
(b)
Average Cumulative Deficit
Flow threshold = 14.74 (maf/yr), nWetYr= 0, LMin = 2 (yrs), LMax = 9999 (yrs), D0 = 0 (maf), I0 = 0 (maf/yr)
Historical Deficit
Deficit (maf)
0
50
100
150
200
250
300
ISM_1906_2020
ISM_1931_2020
ISM_1988_2020
ISM_1416_2015
AR1
NPC_1906_2020
NPC_1988_2020
NPC_2000_2020
5YrBlockRes_2000_2018
DroughtYrRes_2000_2020
DroughtYrRes_1953_1977
DroughtYrRes_1576_1600
CMIP3_BCSD
CMIP5_BCSD
CMIP5_LOCA
TempAdj_RCP4.5_3%
TempAdj_RCP4.5_6.5%
TempAdj_RCP4.5_10%
TempAdj_RCP8.5_3%
TempAdj_RCP8.5_6.5%
TempAdj_RCP8.5_10%

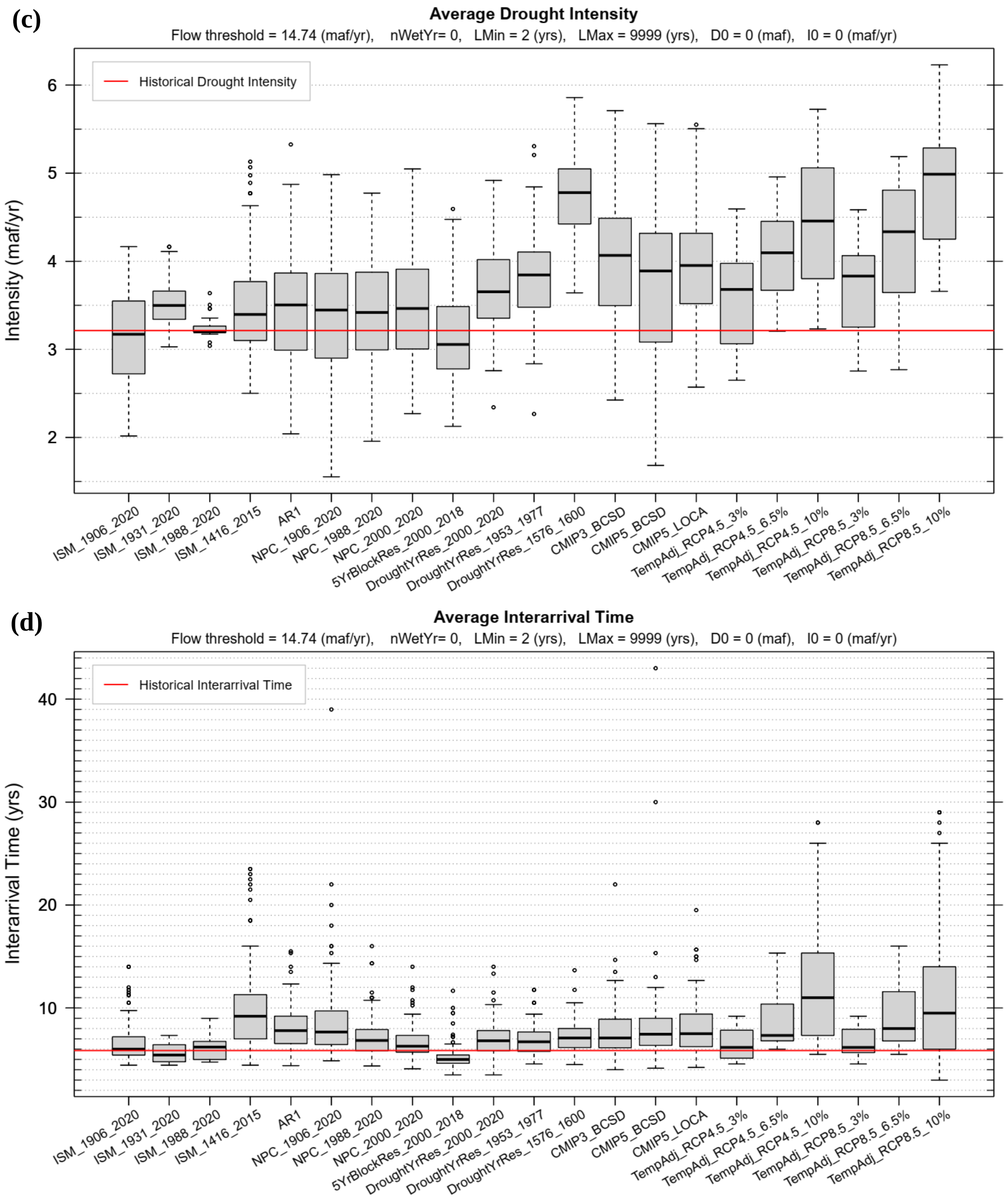


Figure 15. Drought event statistics: (a) drought length, (b) drought cumulative deficit, (c) drought intensity, and (d) drought interarrival time. The threshold is long-term average of the historical natural flow at Lees Ferry (14.7 maf/yr). All drought events with a length grater than 1 year (LMin=2 and LMax=9999) have been considered, without specific thresholds for drought magnitude and intensity (D0=0 and I0=0).

Lag-1 normalized Mutual Information (MI) was calculated for the ensembles and is shown in Figure 16. These results are highly sensitive to the chosen bin boundaries. Therefore, a consistent binning method was applied to ensure the comparability of MI values across ensembles. The findings show variations in the degree of nonlinear dependence among ensembles. Notably, NPC_2000_2020 exhibits a higher MI compared to DroughtYrRes_2000_2020, despite their lack of correlation in Figure 11. This suggests that although both the NPC and random resampling methods are unable to reproduce correlation when the sampling period is short (21 years from 2000 to 2020), the NPC method can generate more nonlinear dependence than a random resampling method.

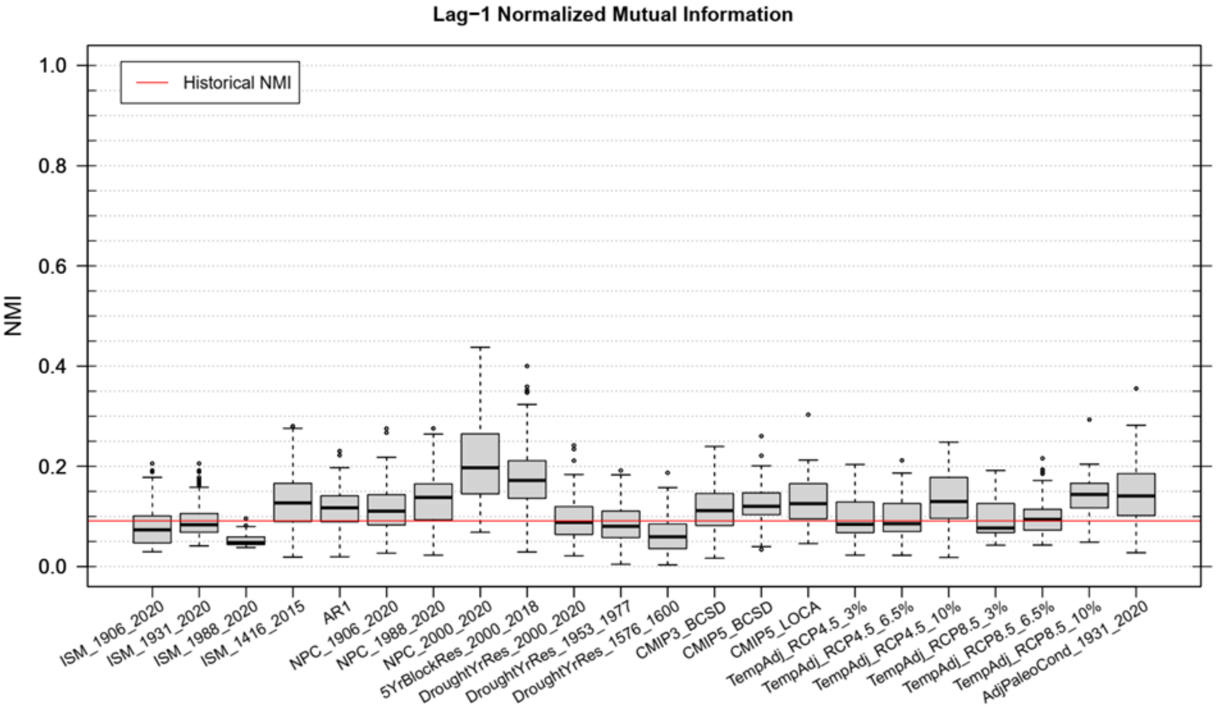


Figure 16. Lag-1 normalized Mutual Information (MI) of the streamflow ensembles (box plots) along with the historical normalized MI (red line)

4.3. Classifying Ensembles

After quantifying the characteristics of the ensembles, we applied Ward's method to classify ensembles based on the metric medians (Figure 17). To do this, we initially examined how sensitive the classification of streamflow ensembles was to metrics. Results indicated that when mutual information was in the set of metrics used for classification, ensembles tended to switch between groups for no apparent reason. Excluding mutual information from the set used for classification maintained the robustness of major ensemble classifications. Therefore, we excluded mutual information from our metric list used for classification.

The heatmap in Figure 17 summarizes the metric results for the ensembles and the historical values highlighted in red. In this figure, each row corresponds to a streamflow ensemble, and each column represents a metric, with each cell indicating a specific metric median for a given ensemble. The color scheme of the heatmap was standardized using subtraction of the metric mean divided by the metric standard deviation across all the ensembles.

The dendrograms on the left represent ensembles, with the X-axis as the ensembles and the Y-axis indicating the distance (as a similarity criterion) at which ensembles merge into the same category. Similar ensembles with minimum distance fall into the same category, while dissimilar ensembles are placed farther in the hierarchy.

The results indicate that some temperature-adjusted ensembles, characterized by a steep decline in flow, were grouped together with the paleo drought resampled ensemble, DroughYrRes_1576_1600 (group 1). This cluster of ensembles has the worst values for drought metrics, the lowest flow magnitudes, and no high flows. The dendrograms on the left show that the TempAdj_RCP8.5_10% ensemble in this group is the most distinct one, while the paleo resampled ensemble (DroughYrRes_1576_1600) is positioned in the middle of the group.

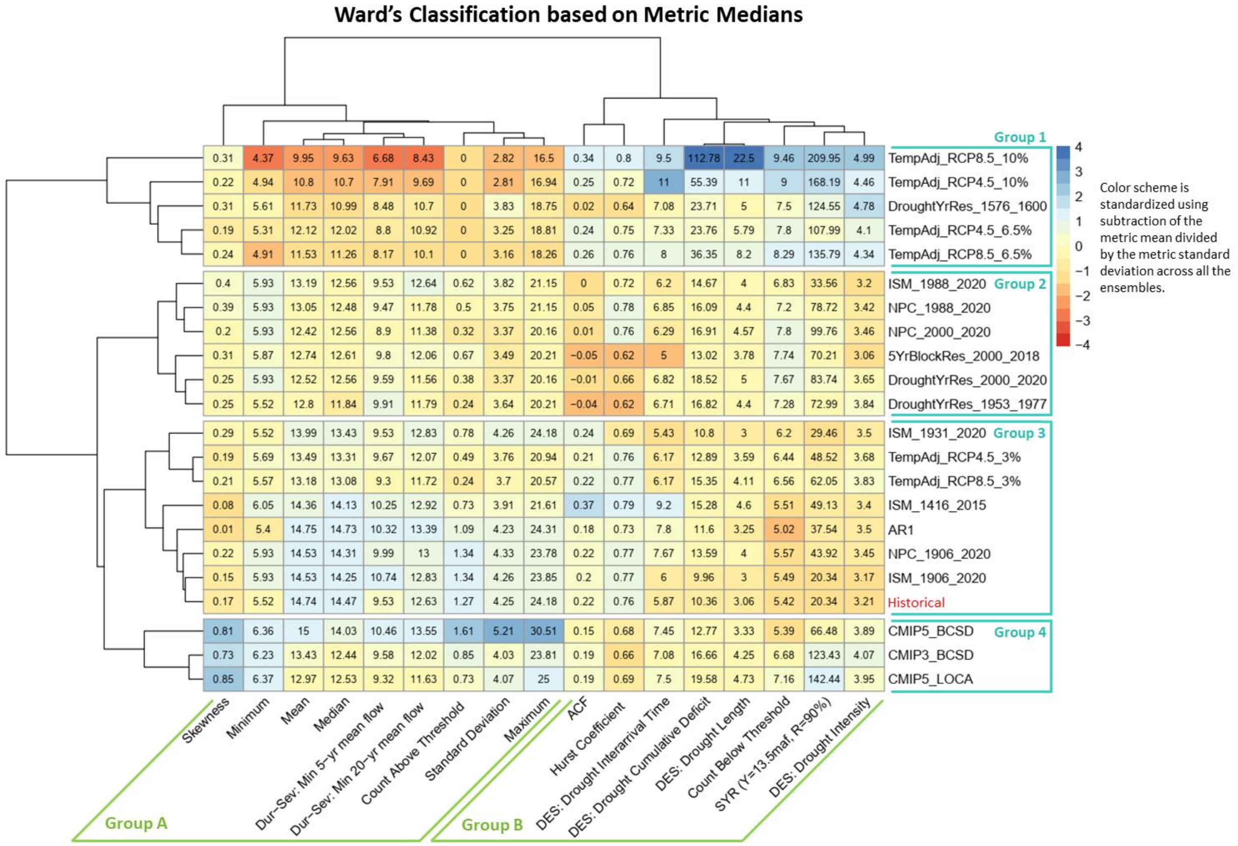
The ensembles based on resampling from specific drought periods are clustered together in group 2. In this group, it is interesting to note that the two millennium drought-based ensembles (NPC_2000_2020 and DroughtYrRes_2000_2020) are not the most similar ensembles despite them being resampled from the same drought period. A comparison of the two rows corresponding to these ensembles (Figure 17) shows that this dissimilarity is primarily due to the difference in the Hurst coefficient, which is higher in the NPC-based ensemble and is more similar to the historical Hurst coefficient. Therefore, when choosing between these two ensembles, the NPC-based one is preferred due to its preservation of historical persistence or long memory, as quantified by the historical Hurst coefficient.

Group 3 comprises ensembles that exhibit the highest similarity to the historical record. Among these ensembles, ISM_1906_2020 and NPC-1906_2020 are the most like the historical record. The paleo-based ensemble (ISM_1416_2015) within this group has the highest correlation (0.37) among all ensembles. The ISM_1931_2020 and two TempAdj ensembles stand out as the most distinct within this group, showing worse drought statistics and lower flows.

The CMIP-based ensembles also are clustered together (group 4). Based on the dendrograms on the left, the CMIP5-LOCA and CMIP3-BCSD are the most similar ensembles within this group. Interestingly, despite both CMIP5-LOCA and CMIP5-BCSD originating from the common CMIP5 source, the choice of downscaling method (BCSD or LOCA) introduces metric differences between these two ensembles. Nevertheless, they remain within the same group, representing a climate change-informed future.

This ensemble grouping provides an analytical framework for characterizing and assessing the ensembles suitability for planning under different future scenarios. Ensembles within the same category help evaluate the system’s response to the future scenario represented by that category. Planning based on ensembles within a single category results in similarities, but significant differences in the system's responses are expected across different ensemble groups. Robust planning should consider ensembles from all the major groups identified to have higher confidence that the sample space of ensembles represented by these groups has been covered.

Note that, in addition to classifying ensembles, Ward’s method also grouped metrics based on their median within each ensemble. This classification is indicated by the dendrograms at the top of Figure 17. Two major groupings emerge, Group A on the left and B on the right. Group A contains metrics largely related to flow magnitude, notably mean, minimum, median, maximum, and count above threshold. Here count above the threshold of 20 maf/yr serves as a proxy for flow magnitude so it is logical that it falls in this group. Standard deviation and skewness are not magnitude quantities, but evidently are more closely aligned with the magnitude metrics than those metrics in group B. Similarly, the minimum 5- and 20-year duration-severity metrics relate to both magnitude and persistence, but evidently, more so to magnitude, by falling in group A. Group B metrics appear to be largely related to drought persistence (ACF, Hurst coefficient, reservoir storage-yield-reliability, drought event statistics, and count below threshold). The count below threshold metric here, with threshold being the long-term mean, does relate to persistence of flows below this threshold and so appears to be logically placed in this group.



	Skewness	Minimum	Mean	Median	Dur–Sev: Min 5–yr mean flow	Dur–Sev: Min 20–yr mean flow	Count Above Threshold	Standard Deviation	Maximum	ACF	Hurst Coefficient	DES: Drought Interarrival Time	DES: Drought Cumulative Deficit	DES: Drought Length	Count Below Threshold	SYR (Y=13.5maf, R=90%)	DES: Drought Intensity
TempAdj_RCP8.5_10%	0.31	4.37	9.95	9.63	6.68	8.43	0	2.82	16.5	0.34	0.8	9.5	112.78	22.5	9.46	209.95	4.99
TempAdj_RCP4.5_10%	0.22	4.94	10.8	10.7	7.91	9.69	0	2.81	16.94	0.25	0.72	11	55.39	11	9	168.19	4.46
DroughtYrRes_1576_1600	0.31	5.61	11.73	10.99	8.48	10.7	0	3.83	18.75	0.02	0.64	7.08	23.71	5	7.5	124.55	4.78
TempAdj_RCP4.5_6.5%	0.19	5.31	12.12	12.02	8.8	10.92	0	3.25	18.81	0.24	0.75	7.33	23.76	5.79	7.8	107.99	4.1
TempAdj_RCP8.5_6.5%	0.24	4.91	11.53	11.26	8.17	10.1	0	3.16	18.26	0.26	0.76	8	36.35	8.2	8.29	135.79	4.34
ISM_1988_2020	0.4	5.93	13.19	12.56	9.53	12.64	0.62	3.82	21.15	0	0.72	6.2	14.67	4	6.83	33.56	3.2
NPC_1988_2020	0.39	5.93	13.05	12.48	9.47	11.78	0.5	3.75	21.15	0.05	0.78	6.85	16.09	4.4	7.2	78.72	3.42
NPC_2000_2020	0.2	5.93	12.42	12.56	8.9	11.38	0.32	3.37	20.16	0.01	0.76	6.29	16.91	4.57	7.8	99.76	3.46
5YrBlockRes_2000_2018	0.31	5.87	12.74	12.61	9.8	12.06	0.67	3.49	20.21	−0.05	0.62	5	13.02	3.78	7.74	70.21	3.06
DroughtYrRes_2000_2020	0.25	5.93	12.52	12.56	9.59	11.56	0.38	3.37	20.16	−0.01	0.66	6.82	18.52	5	7.67	83.74	3.65
DroughtYrRes_1953_1977	0.25	5.52	12.8	11.84	9.91	11.79	0.24	3.64	20.21	−0.04	0.62	6.71	16.82	4.4	7.28	72.99	3.84
ISM_1931_2020	0.29	5.52	13.99	13.43	9.53	12.83	0.78	4.26	24.18	0.24	0.69	5.43	10.8	3	6.2	29.46	3.5
TempAdj_RCP4.5_3%	0.19	5.69	13.49	13.31	9.67	12.07	0.49	3.76	20.94	0.21	0.76	6.17	12.89	3.59	6.44	48.52	3.68
TempAdj_RCP8.5_3%	0.21	5.57	13.18	13.08	9.3	11.72	0.24	3.7	20.57	0.22	0.77	6.17	15.35	4.11	6.56	62.05	3.83
ISM_1416_2015	0.08	6.05	14.36	14.13	10.25	12.92	0.73	3.91	21.61	0.37	0.79	9.2	15.28	4.6	5.51	49.13	3.4
AR1	0.01	5.4	14.75	14.73	10.32	13.39	1.09	4.23	24.31	0.18	0.73	7.8	11.6	3.25	5.02	37.54	3.5
NPC_1906_2020	0.22	5.93	14.53	14.31	9.99	13	1.34	4.33	23.78	0.22	0.77	7.67	13.59	4	5.57	43.92	3.45
ISM_1906_2020	0.15	5.93	14.53	14.25	10.74	12.83	1.34	4.26	23.85	0.2	0.77	6	9.96	3	5.49	20.34	3.17
Historical	0.17	5.52	14.74	14.47	9.53	12.63	1.27	4.25	24.18	0.22	0.76	5.87	10.36	3.06	5.42	20.34	3.21
CMIP5_BCSD	0.81	6.36	15	14.03	10.46	13.55	1.61	5.21	30.51	0.15	0.68	7.45	12.77	3.33	5.39	66.48	3.89
CMIP3_BCSD	0.73	6.23	13.43	12.44	9.58	12.02	0.85	4.03	23.81	0.19	0.66	7.08	16.66	4.25	6.68	123.43	4.07
CMIP5_LOCA	0.85	6.37	12.97	12.53	9.32	11.63	0.73	4.07	25	0.19	0.69	7.5	19.58	4.73	7.16	142.44	3.95

Figure 17. Classification of streamflow ensembles and metrics using Ward's method and based on metric medians. The heatmap summarizes the metric results for all ensembles. Each row corresponds to a streamflow ensemble, and each column represents a metric, with each cell indicating a specific metric median for a given ensemble. The color scheme is standardized using subtraction of the metric mean divided by the metric standard deviation across all the ensembles. The dendrograms on the left represent ensembles, with the X-axis as the ensembles and the Y-axis indicating the distance (as a similarity criterion) at which ensembles merge into the same category. Similar ensembles with minimum distance fall into the same category, while dissimilar ensembles are placed farther in the hierarchy. Dendrograms on the top represent metrics and show how similar the metrics are.

5. Conclusions

In this study, we suggested an evidence-based and structured framework for the quantification and comprehensive description of various streamflow ensembles, to assess their suitability for different planning purposes. Our approach offers objective and quantitative evidence to interpret and analyze differences among these ensembles based on their distinctive characteristics. We employed a broad range of statistical metrics to quantitatively assess a wide range of streamflow ensembles available in the Colorado River Basin and provided guidance on their application and uncertainty. Our metrics address limitations of previous drought statistics and also quantify high flows, the occurrence of which are important for filling reservoirs in some systems. We also developed a classification approach that grouped similar ensembles based on the metrics. The ensemble classification facilitated the comparison of multiple ensembles and provided an analytical framework for characterizing and assessing the ensembles suitability for planning under different future scenarios. It also offers opportunities for efficiency, since not all ensembles with similar attributes based on this classification need to be evaluated in a planning scenario. For robust planning, we suggest considering ensembles from all the major identified groups to have higher confidence that the sample space of ensembles represented by these groups has been covered.

This study's framework serves as a tool for evaluating the key attributes that define each streamflow ensemble, enabling a deeper understanding of ensembles' similarities and differences, which are critical for informed decision-making. Our evidence-based approach serves as a guiding tool for robust decision-making in operational water management, aiding in the selection of the ensembles to use for specific planning purposes such as Reclamation's ongoing Colorado River Post-2026 operations effort. By providing clear, documented, communicable, and evidence-based information, our findings help prevent the adoption of streamflow ensembles without full information on their characteristics.

In our upcoming studies, we plan to evaluate the characteristics of the streamflow ensembles from this study to associate each of them with a storyline that justifies their plausibility for future decision making in the face of uncertainty and non-stationarity. We also plan to investigate any gaps in the sample space represented by existing ensembles and to develop a new ensemble or ensembles as necessary to fill such gaps.

Acknowledgments

This work was supported by the U.S. Bureau of Reclamation grant R21AC10342 for Cataloguing and Generating Hydrology Scenarios in the Colorado River Basin. We are grateful for this support. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Bureau of Reclamation. Guidance, advice, feedback, and insightful conversation greatly improved the quality of this effort — thanks to James Prairie and Alan Butler, for their thoughtful contributions.

Open Research

The data and R Code used in this research is publicly available in HydroShare (Salehabadi & Tarboton, 2024).

References

Ahmadalipour, A., Rana, A., Moradkhani, H., & Sharma, A. (2015). Multi-criteria evaluation of CMIP5 GCMs for climate change impact analysis. *Theoretical and Applied Climatology*, *128*(1), 71-87. https://doi.org/10.1007/s00704-015-1695-4

Bonham, N., Kasprzyk, J., Zagona, E., & Rajagopalan, B. (2024). Subsampling and space-filling metrics to test ensemble size for robustness analysis with a demonstration in the Colorado River Basin. *Environmental Modelling & Software*, *172*, 105933. https://doi.org/10.1016/j.envsoft.2023.105933

Borgomeo, E., Hall, J. W., Fung, F., Watts, G., Colquhoun, K., & Lambert, C. (2014). Risk-based water resources planning: Incorporating probabilistic nonstationary climate uncertainties. *Water resources research*, *50*(8), 6850-6873. https://doi.org/10.1002/2014WR015558

Bras, R. L., & Rodriguez-Iturbe, I. (1985). *Random Functions and Hydrology*. Reading, MA: Addison-Wesley.

Chaves, H. M. L., & Lorena, D. R. (2019). Assessing reservoir reliability using classical and long-memory statistics. *Journal of Hydrology: Regional Studies*, *26*, 100641. https://doi.org/10.1016/j.ejrh.2019.100641

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*: Wiley.

Fiering, M. B. (1967). *Streamflow Synthesis*. Cambridge, MA: Harvard University Press.

Fleck, J., & Castle, A. (2022). Green Light for Adaptive Policies on the Colorado River. *Water*, *14*(1), 2. Retrieved from https://www.mdpi.com/2073-4441/14/1/2

Gong, W., Yang, D., Gupta, H. V., & Nearing, G. (2014). Estimating information entropy for hydrological data: One-dimensional case. *Water resources research*, *50*(6), 5003-5018. https://doi.org/10.1002/2014WR015874

Harrold, T. I., Sharma, A., & Sheather, S. (2001). Selection of a kernel bandwidth for measuring dependence in hydrologic time series using the mutual information criterion. *Stochastic Environmental Research and Risk Assessment*, *15*(4), 310-324. https://doi.org/10.1007/s004770100073

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second Edition ed.). New York, NY: Springer.

Hausser, J., & Strimmer, K. (2021). entropy: Estimation of Entropy, Mutual Information and Related Quantities (Version R package version 1.3.1). Retrieved from https://CRAN.R-project.org/package=entropy

Helsel, D. R., Hirsch, R. M., Ryberg, K. R., Archfield, S. A., & Gilroy, E. J. (2020). Statistical methods in water resources. In *Techniques and Methods* (pp. 484). Reston, VA: U.S. Geological Survey.

Hipel, K. W., & McLeod, A. I. (1994). *Time series modelling of water resources and environmental systems*. Amsterdam: Elsevier.

Hurst, H. E. (1951). Long-term Storage Capacity of Reservoirs. *Transactions American Society of Civil Engineers*, *116*, 770-799.

IPCC. (2021). Summary for Policymakers. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*: Cambridge University Press.

Kendall, M. G. (1955). *Rank correlation methods*. London: Charles Griffin.

Klemeš, V. (1974). The Hurst Phenomenon: A puzzle? *Water resources research*, *10*(4), 675-688. https://doi.org/10.1029/WR010i004p00675

Kolde, R. (2019). pheatmap: Pretty Heatmaps (Version R package version 1.0.12). Retrieved from https://CRAN.R-project.org/package=pheatmap

Koutsoyiannis, D., Yao, H., & Georgakakos, A. (2008). Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods / Prévision du débit du Nil à moyen terme: une comparaison de méthodes stochastiques et déterministes. *Hydrological sciences journal*, *53*(1), 142-164. https://doi.org/10.1623/hysj.53.1.142

Kuria, F. W., & Vogel, R. M. (2014). A global water supply reservoir yield model with uncertainty analysis. *Environmental Research Letters*, *9*(9), 095006. https://doi.org/10.1088/1748-9326/9/9/095006

LaRue, E. C. (1916). *Colorado River and its utilization*. Washington, D.C.: US Government Printing Office.

Lee, T., & Ouarda, T. B. M. J. (2012). Stochastic simulation of nonstationary oscillation hydroclimatic processes using empirical mode decomposition. *Water resources research*, *48*(2). https://doi.org/10.1029/2011WR010660

Lee, T., & Ouarda, T. B. M. J. (2023). Trends, Shifting, or Oscillations? Stochastic Modeling of Nonstationary Time Series for Future Water-Related Risk Management. *Earth's Future*, *11*(7), e2022EF003049. https://doi.org/10.1029/2022EF003049

Lee, T., Salas, J. D., & Prairie, J. (2010). An enhanced nonparametric streamflow disaggregation model with genetic algorithm. *Water resources research*, *46*(8). https://doi.org/10.1029/2009WR007761

Lee, T., Shin, J.-Y., Kim, J.-S., & Singh, V. P. (2020). Stochastic simulation on reproducing long-term memory of hydroclimatological variables using deep learning model. *Journal of Hydrology*, *582*, 124540. https://doi.org/10.1016/j.jhydrol.2019.124540

Loritz, R., Gupta, H., Jackisch, C., Westhoff, M., Kleidon, A., Ehret, U., & Zehe, E. (2018). On the dynamic nature of hydrological similarity. *Hydrology and Earth System Sciences, 22*(7), 3663-3684. https://doi.org/10.5194/hess-22-3663-2018

Loucks, D. P., van Beek, E., Stedinger, J. R., Dijkman, J. P. M., & Villars, M. T. (2017). *Water Resources Systems Planning and Management: An Introduction to Methods, Models and Applications*: Springer.

Lukas, J., Gutmann, E., Harding, B., & Lehner, F. (2020). Climate Change-Informed Hydrology. In J. Lukas & E. Payton (Eds.), *Colorado River Basin Climate and Hydrology: State of the Science* (pp. 384-449): Western Water Assessment, University of Colorado Boulder.

MacDonnell, L. (2021). Colorado River Basin. Waters and Water Rights, Lexis-Nexus, CORB-1. *SSRN*. Retrieved from https://ssrn.com/abstract=3780342

Mann, H. B. (1945). Nonparametric Tests Against Trend. *Econometrica*, *13*(3), 245-259. https://doi.org/10.2307/1907187

Matalas, N. C., Landwehr, J. M., & Wolman, M. G. (1982). Prediction in Water Management, Chapter 11. In *Scientific Basis of Water Management*. Washington D.C: Studies in Geophysics, National Academy Press.

Meko, D. M., Woodhouse, C. A., & Bigio, E. R. (2017). *Southern California Tree-Ring Study*. Retrieved from https://cwoodhouse.faculty.arizona.edu/content/california-department-water-resources-studies

Milly, P. C., & Dunne, K. A. (2020). Colorado River flow dwindles as warming-driven loss of reflective snow energizes evaporation. *Science, 367*(6483), 1252-1255. https://doi.org/10.1126/science.aay9187

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Stationarity Is Dead: Whither Water Management? *Science, 319*(5863), 573-574. http://doi.org/10.1126/science.1151915

Montanari, A., Rosso, R., & Taqqu, M. S. (1997). Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation. *Water resources research, 33*(5), 1035-1044. https://doi.org/10.1029/97WR00043

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *WIREs Data Mining and Knowledge Discovery*, *2*(1), 86-97. https://doi.org/10.1002/widm.53

Ouarda, T. B. M. J., Labadie, J. W., & Fontane, D. G. (1997). Indexed Sequential Hydrologic Modeling for Hydropower Capacity Estimation. *JAWRA Journal of the American Water Resources Association, 33*(6), 1337-1349. https://doi.org/10.1111/j.1752-1688.1997.tb03557.x

Papacharalampous, G., Tyralis, H., & Koutsoyiannis, D. (2019). Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Stochastic Environmental Research and Risk Assessment, 33*(2), 481-514. https://doi.org/10.1007/s00477-018-1638-6

Payton, E., Smith, R., Jerla, C., & Prairie, J. (2020). Primary Planning Tools. In J. Lukas & E. Payton (Eds.), *Colorado River Basin Climate and Hydrology: State of the Science* (pp. 82-111): Western Water Assessment, University of Colorado Boulder.

Pechlivanidis, I. G., Gupta, H., & Bosshard, T. (2018). An Information Theory Approach to Identifying a Representative Subset of Hydro-Climatic Simulations for Impact Modeling Studies. *Water resources research*, *54*(8), 5422-5435. https://doi.org/10.1029/2017WR022035

Pechlivanidis, I. G., Jackson, B., McMillan, H., & Gupta, H. V. (2016). Robust informational entropy-based descriptors of flow in catchment hydrology. *Hydrological sciences journal, 61*(1), 1-18. https://doi.org/10.1080/02626667.2014.983516

Pezij, M., Augustijn, D. C. M., Hendriks, D. M. D., & Hulscher, S. J. M. H. (2019). The role of evidence-based information in regional operational water management in the Netherlands. *Environmental Science & Policy, 93*, 75-82. https://doi.org/10.1016/j.envsci.2018.12.025

Prairie, J., & Callejo, R. (2005). *Natural Flow and Salt Computation Methods, Calendar Years 1971-1995*. Retrieved from Salt Lake City, Utah: https://digitalcommons.usu.edu/govdocs/135/

Prairie, J., Nowak, K., Rajagopalan, B., Lall, U., & Fulp, T. (2008). A stochastic nonparametric approach for streamflow generation combining observational and paleoreconstructed data. *Water Resour. Res., 44*(6), W06423. Retrieved from http://dx.doi.org/10.1029/2007WR006684

Prairie, J., Rajagopalan, B., Fulp, T., & Zagona, E. A. (2006). Modified K-NN Model for Stochastic Streamflow Simulation. *Journal of Hydrologic Engineering, 11*(4), 371-378. Retrieved from http://dx.doi.org/10.1061/(ASCE)1084-0699(2006)11:4(371)

Prairie, J., Rajagopalan, B., Lall, U., & Fulp, T. (2007). A stochastic nonparametric technique for space-time disaggregation of streamflows. *Water Resour. Res., 43*(3), W03432. Retrieved from http://dx.doi.org/10.1029/2005WR004721

R Core Team. (2023). R: A Language and Environment for Statistical Computing: R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Razavi, S., Elshorbagy, A., Wheater, H., & Sauchyn, D. (2015). Toward understanding nonstationarity in climate and hydrology through tree ring proxy records. *Water resources research, 51*(3), 1813-1830. https://doi.org/10.1002/2014WR015696

Rosenberg, D. E. (2022). Adapt Lake Mead Releases to Inflow to Give Managers More Flexibility to Slow Reservoir Drawdown. *Journal of Water Resources Planning and Management, 148*(10), 02522006. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001592

Salas, J., Obeysekera, J., & Vogel, R. (2018). Techniques for assessing water infrastructure for nonstationary extreme events: a review. *Hydrological sciences journal, 63*(3), 325-352. https://doi.org/10.1080/02626667.2018.1426858

Salas, J. D., Fu, C., Cancelliere, A., Dustin, D., Bode, D., Pineda, A., & Vincent, E. (2005). Characterizing the Severity and Risk of Drought in the Poudre River, Colorado. *Journal of Water Resources Planning and Management*, *131*(5), 383-393. https://doi.org/10.1061/(ASCE)0733-9496(2005)131:5(383)

Salehabadi, H., & Tarboton, D. G. (2024). *R Scripts for Evaluating Annual Streamflow Ensemble Metrics and Data and Results from their Application in the Colorado River Basin*. Retrieved from: http://www.hydroshare.org/resource/d7b65c91dda047e1969a9f9cd09b489f

Salehabadi, H., Tarboton, D. G., Kuhn, E., Udall, B., Wheeler, K. G., Rosenberg, D. E., . . . Schmidt, J. C. (2020). *The Future Hydrology of the Colorado River Basin* (White Paper 4). Retrieved from https://qcnr.usu.edu/coloradoriver/files/WhitePaper4.pdf

Salehabadi, H., Tarboton, D. G., Udall, B., Wheeler, K. G., & Schmidt, J. C. (2022). An Assessment of Potential Severe Droughts in the Colorado River Basin. *JAWRA Journal of the American Water Resources Association, 58*(6), 1053-1075. https://doi.org/10.1111/1752-1688.13061

Schmidt, J. C., Bruckerhoff, L., Salehabadi, H., & Wang, J. (2022). The Colorado River. In A. Gupta (Ed.), *Large Rivers: Geomorphology and Management* (Second ed., pp. 253-319): John Wiley & Sons, Ltd.

Schmidt, J. C., Yackulic, C. B., & Kuhn, E. (2023). The Colorado River water crisis: Its origin and the future. *WIREs Water, 10*(6), e1672. https://doi.org/10.1002/wat2.1672

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. N. Y.: Wiley.

Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev., 5*(1), 3–55. https://doi.org/10.1145/584091.584093

Sharma, A., Tarboton, D. G., & Lall, U. (1997). Streamflow Simulation: A Nonparametric Approach. *Water resources research, 33*(2), 291-308. Retrieved from http://dx.doi.org/10.1029/96WR02839

Smith, R., Zagona, E., Kasprzyk, J., Bonham, N., Alexander, E., Butler, A., . . . Jerla, C. (2022). Decision Science Can Help Address the Challenges of Long-Term Planning in the Colorado River Basin. *JAWRA Journal of the American Water Resources Association, 58*(5), 735-745. https://doi.org/10.1111/1752-1688.12985

Srinivas, V. V., & Srinivasan, K. (2000). Post-blackening approach for modeling dependent annual streamflows. *Journal of Hydrology, 230*, 86-126. https://doi.org/10.1016/S0022-1694(00)00168-2

Srinivas, V. V., & Srinivasan, K. (2005). Hybrid moving block bootstrap for stochastic simulation of multi-site multi-season streamflows. *Journal of Hydrology, 302*(1), 307-330. https://doi.org/10.1016/j.jhydrol.2004.07.011

Srinivas, V. V., & Srinivasan, K. (2006). Hybrid matched-block bootstrap for stochastic simulation of multiseason streamflows. *Journal of Hydrology, 329*(1), 1-15. https://doi.org/10.1016/j.jhydrol.2006.01.023

Tarboton, D. G. (1994). The Source Hydrology of Severe Sustained Drought in the Southwestern United States. *Journal of Hydrology, 161*, 31-69. Retrieved from http://dx.doi.org/10.1016/0022-1694(94)90120-1

Tarboton, D. G. (1995). Hydrologic Scenarios for Severe Sustained Drought in the Southwestern United States. *Water Resources Bulletin, 31*(5), 803-813. Retrieved from https://doi.org/10.1111/j.1752-1688.1995.tb03402.x

Udall, B. (2020). *CRSS-Ready Temperature-Adjusted Colorado River Inflows* (August 4, 2020). Retrieved from

Udall, B., & Overpeck, J. (2017). The twenty-first century Colorado River hot drought and implications for the future. *Water resources research, 53*(3), 2404-2418. Retrieved from https://doi.org/10.1002/2016WR019638

USBR. (2011). *West-Wide Climate Risk Assessments: Bias-Corrected and Spatially Downscaled Surface Water Projections* (Technical Memorandum No. 86-68210-2011-01). Retrieved from Technical Services Center, Denver, Colorado:

USBR. (2012). *Colorado River Basin Water Supply and Demand Study, Technical Report B – Water Supply assessment*. Retrieved from Available online at: http://www.usbr.gov/lc/region/programs/crbstudy.html

USBR. (2014). *Downscaled CMIP3 and CMIP5 climate and hydrology projections: Release of hydrology projections, comparison with preceding information, and summary of user needs*. Retrieved from Denver, Colorado: https://gdo-dcp.ucllnl.org/downscaled_cmip_projections/techmemo/BCSD5HydrologyMemo.pdf

USBR. (2022). Colorado River Basin Natural Flow and Salt Data. Retrieved from https://www.usbr.gov/lc/region/g4000/NaturalFlow/current.html

USBR. (2023, 12/7/2023). Colorado River Post 2026 Operations. Retrieved from https://www.usbr.gov/ColoradoRiverBasin/post2026/index.html

Valencia, D., & Schaake, J. C. (1973). Disaggregation processes in stochastic hydrology. *Water Resour. Res*, *9*(3), 580-585.

Vano, J., Hamman, J., Gutmann, E., Wood, A., Mizukami, N., Clark, M., . . . Arnold, J. (2020). *Comparing Downscaled LOCA and BCSD CMIP5 Climate and Hydrology Projections - Release of Downscaled LOCA CMIP5 Hydrology*. Retrieved from https://gdo-dcp.ucllnl.org/downscaled_cmip_projections/techmemo/LOCA_BCSD_hydrology_tech_memo.pdf

Venables, W. N., & Ripley, B. D. (2010). *Modern Applied Statistics with S*. New York, NY: Springer.

Vogel, R. M. (2017). Stochastic watershed models for hydrologic risk management. *Water Security*, *1*, 28-35. https://doi.org/10.1016/j.wasec.2017.06.001

Vogel, R. M., Tsai, Y., & Limbrunner, J. F. (1998). The regional persistence and variability of annual streamflow in the United States. *Water resources research*, *34*(12), 3445-3459. https://doi.org/10.1029/98WR02523

Wheeler, K. G., Kuhn, E., Bruckerhoff, L., Udall, B., Wang, J., Gilbert, L., . . . Schmidt, J. C. (2021). *Alternative Management Paradigms for the Future of the Colorado and Green Rivers* (White Paper 6). Retrieved from https://qcnr.usu.edu/coloradoriver/files/WhitePaper6.pdf

Wheeler, K. G., Udall, B., Wang, J., Kuhn, E., Salehabadi, H., & Schmidt, J. C. (2022). What will it take to stabilize the Colorado River? *Science*, *377*(6604), 373-375. http://doi.org/10.1126/science.abo4452

Wilhite, D. A., & Buchanan-Smith, M. (2005). Drought as hazard: understanding the natural and social context. In *Drought and water crises: science, technology, and management issues*.

Williams, A. P., Cook, E. R., Smerdon, J. E., Cook, B. I., Abatzoglou, J. T., Bolles, K., . . . Livneh, B. (2020). Large contribution from anthropogenic warming to an emerging North American megadrought. *Science*, *368*(6488), 314. https://doi.org/10.1126/science.aaz9600

Woodhouse, C. A., Smith, R. M., McAfee, S. A., Pederson, G. T., McCabe, G. J., Miller, W. P., & Csank, A. (2021). Upper Colorado River Basin 20th century droughts under 21st century warming: Plausible scenarios for the future. *Climate Services*, *21*, 100206. https://doi.org/10.1016/j.cliser.2020.100206

Xiao, M., Udall, B., & Lettenmaier, D. P. (2018). On the Causes of Declining Colorado River Streamflows. *Water resources research*, *54*(9), 6739-6756. 10.1029/2018wr023153

Yevjevich, V. M. (1963). Fluctuations of wet and dry years: research data assembly and mathematical models: part I. *Hydrology papers (Colorado State University); no. 1*.

Yevjevich, V. M. (1967). *Objective Approach to Definitions and Investigations of Continental Droughts* (Hydrology Paper 23). Retrieved from

Zagona, E. A., Fulp, T. J., Shane, R., Magee, T., & Goranflo, H. M. (2001). Riverware: A generalized tool for complex reservoir system modeling. *JAWRA Journal of the American Water Resources Association*, *37*(4), 913-929. Retrieved from https://doi.org/10.1111/j.1752-1688.2001.tb05522.x