Climatic Change
https://doi.org/10.1007/s10584-018-2347-3

A framework for testing dynamic classification of vulnerable scenarios in ensemble water supply projections

Bethany Robinson[1] • Jonathan D. Herman[1]

Received: 21 June 2018 / Accepted: 3 December 2018/Published online: 19 December 2018

Abstract

Recent water resources planning studies have proposed climate adaptation strategies in which infrastructure and policy actions are triggered by observed thresholds or "signposts." However, the success of such strategies depends on whether thresholds can be accurately linked to future vulnerabilities. This study presents a framework for testing the ability of adaptation thresholds to dynamically identify vulnerable scenarios within ensemble projections. Streamflow projections for 91 river sites predominantly in the western USA are used as a case study in which vulnerability is determined by the ensemble members with the lowest 10% of end-of-century mean annual flow. Illustrative planning thresholds are defined through time for each site based on the mean streamflow below which a specified fraction of scenarios is vulnerable. We perform a leave-one-out cross-validation to compute the frequency of incorrectly identifying or failing to identify a vulnerable scenario (false positives and false negatives, respectively). Results show that in general, this method of defining thresholds can identify vulnerable scenarios with low false positive rates (< 10%), but with false negative rates for many rivers remaining higher than random chance until roughly 2060. This finding highlights the tradeoff between frequently triggering unnecessary action and failing to identify potential vulnerabilities until later in the century, and suggests room for improvement in the threshold-setting technique that could be benchmarked with this approach. This testing framework could extend to thresholds defined with multivariate statistics, or to any application using thresholds and ensemble projections, such as long-term flood and drought risk, or sea level rise.

1 Introduction

Water resource planners routinely face uncertainty in supply and demand, including inter- and intra-annual weather variability as well as longer-term population changes (Fletcher et al.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10584-018-2347-3) contains supplementary material, which is available to authorized users.

✉ Bethany Robinson
bjrobins@ucdavis.edu

[1] Department of Civil and Environmental Engineering, University of California, Davis, CA, USA

2017). However, model projections of water security risks under climate change—for example, due to nonstationary mean and variance of annual streamflow, or changes in drought frequency and severity—remain highly uncertain by comparison (Ranger et al. 2013). This "deep uncertainty" prevents analysts from knowing or agreeing on the probabilities of future scenarios (Lempert 2002), which hinders the interpretation of ensemble projections from general circulation models (GCMs) (Hallegatte et al. 2012; Kwakkel and Pruyt 2013). Despite this uncertainty, decisions must be made regarding infrastructure and operating policies for water supply, flood control, hydropower generation, and environmental flows. Adapting to a changing climate could be done with exact knowledge of a future probability distribution, but planning based on a range of plausible scenarios, with some indicating significant departures from the historical record, remains challenging.

Multiple methodological frameworks have been proposed to address the challenge of climate adaptation under deep uncertainty, recognizing that probabilities may not be necessary to inform policy development (Dessai and Hulme 2004). Broadly, these frameworks focus on resilience, robustness, flexibility, or a combination of these (Adger et al. 2005; Walker et al. 2013). Resilience aims for quick recovery from extreme events (Dessai and Hulme 2004). Robustness requires preparing for the full range of plausible future scenarios, which includes changes to climate and other uncertain exogenous variables (Lempert and Collins 2007; Herman et al. 2015; Giuliani and Castelletti 2016). Because robust strategies are generally costly, plans may include improvements that would be beneficial in any future scenario, such as repairing leaks or developing drought-resistant crops (Hallegatte 2009). Similarly, flexible strategies are intended to keep options open as the climate changes, often using real options analysis, which allows water planners to alter or reverse strategies as needed while minimizing long-term costs (DiFrancesco and Tullos 2014; Jeuland and Whittington 2014; Buurman and Babovic 2016). These frameworks are termed bottom-up approaches to climate adaptation planning, as they identify static thresholds beyond which the current system would be vulnerable (e.g., thresholds of sea level rise, floods, or droughts) regardless of the likelihood of future climate projections (Herman et al. 2015). Specific examples include Scenario Discovery (Bryant and Lempert 2010) and Decision Scaling (Brown et al. 2012). This contrasts with top-down methods, which evaluate the impact of downscaled climate projections in simulation models to estimate system performance metrics of interest, typically expressed as a probability distribution (Wilby and Dessai 2010). Significant opportunities exist to combine insights from both approaches. For example, the vulnerability thresholds identified in bottom-up methods can be compared to an ensemble of top-down scenarios to approximate their relative likelihood (Taner et al. 2017; Ray et al. 2018), or—as in this study—to determine whether ensemble projections contain information that can dynamically identify future vulnerabilities. This study contributes a framework to evaluate the classification accuracy of planning thresholds using ensembles of top-down climate scenarios. This threshold testing framework could be applied to any adaptation problem using thresholds and ensembles of outcomes, such as water supply, sea level rise, flooding, or temperature changes.

Plans aiming for static robustness to all possible scenarios are likely to be costly. To avoid the risk of over-investment, one option is to monitor hydrologic variables over time and tie adaptation decisions to these observations. In other words, a dynamic planning process involves defining the conditions under which certain actions should be taken. Planning frameworks using this approach might generally be called "adaptive pathways," drawing from a range of published work, including Adaptation Tipping Points (Kwadijk et al. 2010) and Dynamic Adaptive Policy Pathways (Haasnoot et al. 2013; Kwakkel et al. 2015; Zeff et al.

2016). These methods work by identifying tipping points, triggers, or "signposts" as indicators to determine that the current management strategies of a system will no longer function, and then planning responses to those conditions. These responses can be set actions determined in advance, a reassessment of the plan itself, or a set of pathways to follow depending on which tipping points, triggers, or signposts are reached (Kwadijk et al. 2010; Haasnoot et al. 2013; Kwakkel et al. 2015). The common thread between these methods is the use of thresholds (i.e., tipping points or triggers) to determine the timing of policy changes, which is similar to control methods developed for short-term operational decisions (e.g., Zeff et al. 2014; Herman and Giuliani 2018). For long-term climate adaptation with greater uncertainty, an adaptive pathway can only be successful if its thresholds accurately identify vulnerable scenarios, ideally in advance of when vulnerability occurs to allow time to build infrastructure or change policies to prevent unwanted, costly outcomes.

For these threshold-based planning frameworks, analysis of GCM-derived scenario ensembles can inform how well a chosen threshold can classify vulnerable scenarios amid the uncertain range of projections. Ensemble projections of water availability under climate change come with several important limitations that affect the development of planning thresholds based on these simulations. First, some ensemble members may not adequately capture the internal variability of the climate system, represented by, for example, the historical mean, variance, and autocorrelation of streamflow at multiple timescales. Second, the uncertainty across scenarios in the ensemble does not represent the full range of possibilities (Hallegatte et al. 2012) and may only represent the lower bound of uncertainty (Stainforth et al. 2007). Third, if all simulations in an ensemble are created under the same model assumptions (e.g., model structure and parameterization), it is possible that they share the same errors (Hallegatte et al. 2012; Steinschneider et al. 2015). Lastly, many ensembles contain scenarios derived from multiple representative concentration pathways (RCPs), and although the different RCPs are not equally likely (Katz 2002; Kundzewicz et al. 2018), they are often treated as such due to the lack of an agreed-upon weighting system. Despite these limitations, GCM ensembles present potentially useful information for water resource planners. Therefore, a structured approach is needed to support adaptive planning while acknowledging the severe uncertainties that have been identified in long-term climate projections (Frigg et al. 2013).

This study focuses on a set of river basins predominantly in the western USA, a region with a host of water supply challenges due to climate change. These challenges, both predicted and observed, have been studied extensively. The population of the area continues to grow rapidly, while water resources face multiple competing demands such as flood control, hydropower production, irrigation supply, and environmental protection (Leung et al. 2004). Increasing temperatures across the region are predicted with high confidence (Leung et al. 2004), and spring temperatures have already risen by several degrees since the beginning of the twentieth century (Cayan et al. 2001; Donat et al. 2013). In addition, tree ring reconstructions indicate higher temperatures in the past half-century than any similar period in the last 900 years (Luckman 1998; Donat et al. 2013). As a result, hydrologic changes that depend on temperature increases are also predicted with high confidence (Cayan et al. 2001; Leung et al. 2004; Knowles et al. 2006; Pederson et al. 2011), and these temperature-related effects have already been observed across the region (Mote et al. 2005; Stewart et al. 2005; Barnett et al. 2008). Long-term trends in total annual precipitation are less certain, and many areas could become either wetter or drier. However, the frequency and magnitude of floods and droughts are expected to increase over the western USA (Anderson et al. 2008; US Global Change

Research Program 2009). With future water availability surrounded by severe uncertainty, designing adaptive pathways under climate change will require understanding how well thresholds or signposts can reliably indicate future vulnerabilities.

2 Methods

We propose a framework to test the classification accuracy of thresholds, defined in terms of long-term average streamflow, to identify vulnerable scenarios for rivers predominantly in the western USA. An illustrative approach is developed to define planning thresholds using ensembles of GCM-based streamflow projections, similar to an approach that might be taken by systems planners. Then, a leave-one-out cross-validation is used to determine how well those thresholds classify vulnerable scenarios over time. The methodology combines top-down and bottom-up approaches to adaptation strategies by incorporating both adaptive planning thresholds and climate projection information. Further spatial analysis considers the utility of planning thresholds across river basins in the region. Each river site represents a simplified case study, where water resource planners would in principle use these thresholds as vulnerability signals to trigger investments in infrastructure, or a revision of operating rules. Table 1 defines key terms used throughout this paper.

2.1 Data sources

Future streamflow projections are drawn from a 2014 U.S. Bureau of Reclamation (USBR) study in which climate projections (supplemental Table S1) from the Coupled Model Intercomparison Project (CMIP5) were downscaled using the Bias-Correction and Spatial Disaggregation (BCSD) technique described in Brekke et al. (2014). In the USBR study, these downscaled projections were then routed through the Variable Infiltration Capacity (VIC) hydrologic model (Liang et al. 1994) to create streamflow projections for sites predominantly across the western USA. For each river site, 97 different BCSD climate projections are available, representing combinations of 31 climate models and four RCPs (2.6, 4.5, 6.0, and 8.5). A total of 91 sites were selected from the set of 242 sites available in the USBR dataset. Sites were first eliminated to include only one point per river reach, eliminating sites whose downscaled streamflows were highly correlated. Next, sites with relatively low flows (less than

Table 1 Terms and definitions

Term	Definition
Scenario	One GCM-based streamflow projection. There are 97 for each river site in this study
Ensemble	A group of GCM-based streamflow projections
Vulnerability	A state in which a river is unlikely to have enough water volume (annually) to fulfill its needs, including water supply, hydropower production, and environmental flows
Threshold	A line defined through the time series that signals vulnerability if a scenario crosses below it at any time. The illustrative thresholds developed for this study are defined in Eq. 4
Model agreement	The fraction of vulnerable scenarios below the threshold at a given time. Model agreement defines how stringent a threshold is, with a higher value denoting a more stringent threshold
Missing point	A threshold value at a particular year that is undefined because a threshold value could not be found that met the model agreement (see Eq. 4), and interpolation was not possible
Error rates	The percent of false negatives and false positives for a given year, defining how well the chosen threshold is able to classify scenarios

roughly 1000 TAF/year or 1.23 km^3/year) were eliminated because they are considered less important for water supply, and their flows have high relative fluctuations when normalized against the historical baseline. The remaining 91 river sites selected are spread over the western USA (Fig. 1) and nine of the 19 continental US HUC-2 watersheds: California, Pacific Northwest, Colorado (Upper and Lower), Missouri, Arkansas-White-Red, Texas-Gulf, Upper Mississippi, and Lower Mississippi (Seaber et al. 1987).

The projections include statistically representative streamflow data for the historical period 1949–2000. Additionally, observed historical data through 2018 is drawn from multiple sources for comparison: the California Data Exchange Center (CDEC), the Natural Resources Conservation Service (NRCS), and the U.S. Geological Survey (USGS). Data sources for each river site are listed in supplemental Table S2, along with additional data including mean historical flow, location, and reservoir name where applicable.

Examples from the streamflow projection dataset are shown in Fig. 2 for four selected river sites. The data plotted in Fig. 2 includes three important transformations that will be used throughout the study, and these transformations are shown step by step in Fig. 3. First, the monthly streamflow values from the original dataset are summed annually (Fig. 3a). The inter-annual dynamics are considered more important for analyzing long-term water supply than monthly or seasonal dynamics, as many of these river locations have large reservoirs with carryover storage. Second, for each annual flow time series $Q(t)$, the 50-year moving average is taken: $Q_{MA(50)}(t)$ (Fig. 3b). Sensitivity experiments later in the study consider different moving window sizes. Finally, the moving average is normalized by the mean flow over the historical period (Fig. 3c) to allow comparison of rivers with different flow magnitudes:

$$\tilde{Q}_{\mathrm{MA}(50)} = \frac{Q_{\mathrm{MA}(50)}(t)}{Q_{\mathrm{MA}(50)}(t = 2000)} \tag{1}$$

The values of $\tilde{Q}_{\mathrm{MA}(50)}$, hereafter shortened to $\tilde{Q}_{\mathrm{MA}}$, reflect the percent change in average annual streamflow relative to the historical baseline.

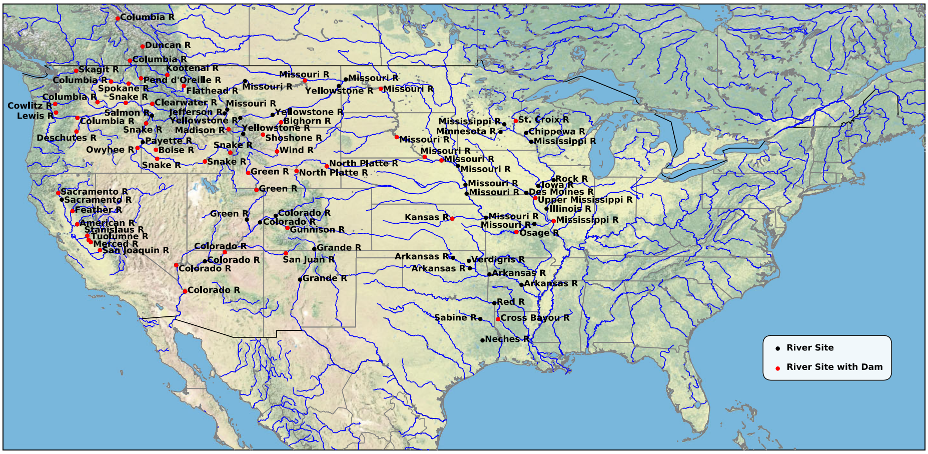


Fig. 1 River sites used for this study from the USBR streamflow projection dataset (Brekke et al. 2014). The sites marked with red dots have reservoirs, many of which provide water supply among other purposes

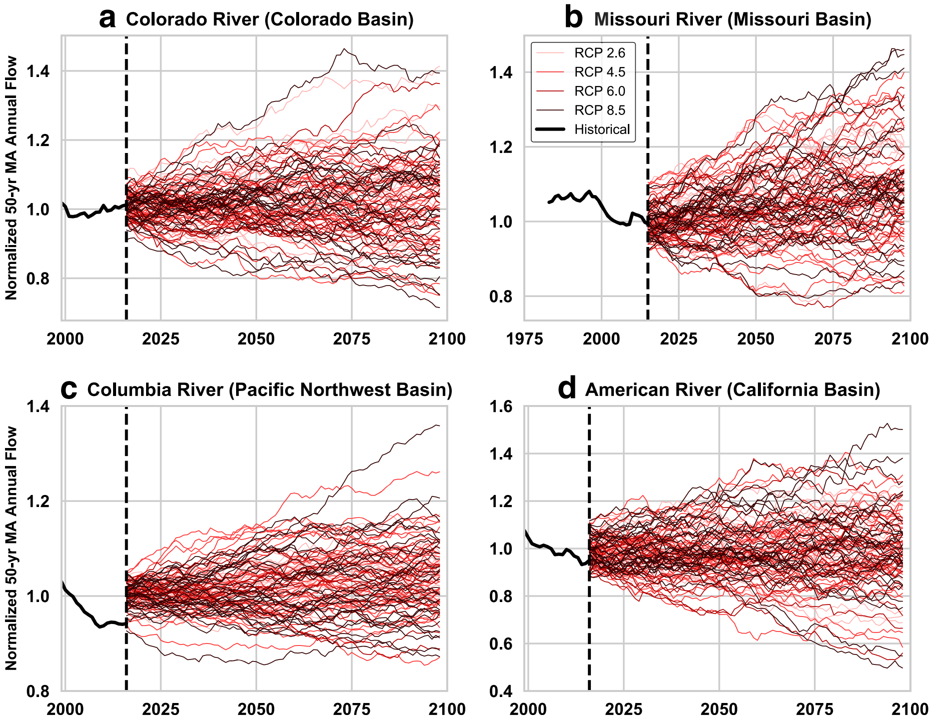


Fig. 2 Examples of the normalized 50-year moving average streamflow $(\tilde{Q}_{MA})$ showing the range of end-of-century uncertainty for selected rivers. (**a**) The Colorado River near Cameo. (**b**) The Missouri River at Fort Peck Dam. (**c**) The Columbia River at Priest Rapids Dam. (**d**) The American River at Folsom Dam

The same transformations are applied to the observed historical data (black lines in Fig. 2). The observed time series are included as visual references for these sites to show that in recent years, the long-term average streamflow has changed relatively little. However, the hydrologic projections vary much more in the coming decades, with some indicating end-of-century changes of ±20% or more. From a water supply perspective, the focus of this study, drier years are more concerning when attempting to determine how robust water supplies will be in the long-term. However, for the purpose of generality, the following methodology will also be applied to sites where the majority of scenarios are wetter than the historical average. Wetter years might suggest more flooding, although finer-timescale projections would be needed to analyze patterns in flood frequency, where downscaled GCM projections often show large disagreement (Hirabayashi et al. 2013; Dottori et al. 2018). Figure 2 also shows projections for the four RCPs in different colors, showing no clear distinction between them. Distributions for each of the RCPs at 2050 and 2100 are shown in supplemental Figs. S1 and S2, also indicating a lack of clear distinction between the RCPs.

2.2 Developing planning thresholds

Many approaches exist for developing planning thresholds, for example through a multi-objective optimization approach (Kwakkel et al. 2015) or a data assimilation approach (Hui et al. 2018). Any of these approaches could be tested using the methods proposed here. To define thresholds in this study based on an ensemble of streamflow projections, we employ an

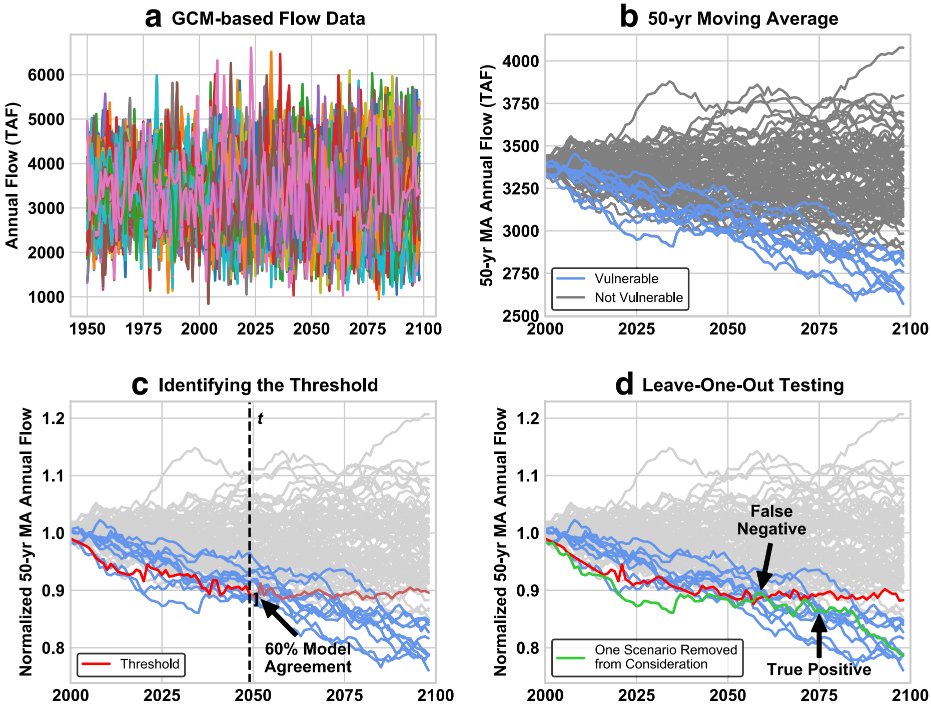


Fig. 3 Methods flowchart using the Lewis River (Pacific Northwest Basin) as an example. (**a**) The original projected annual streamflow data. (**b**) The projections transformed using a 50-year moving average, and classifying the lowest 10% of scenarios at end-of-century as vulnerable (blue). (**c**) The projections are normalized, and a vulnerability threshold (red) is drawn at the point for each time *t* where at least 60% of the scenarios below are classified as vulnerable. Validation of the method was performed using leave-one-out testing (**d**), in which each of the 97 scenarios was removed from defining the threshold one at a time to determine if that streamflow time series could be classified correctly in each year as either vulnerable or not

illustrative approach relying on the concept of vulnerability as might be defined by stakeholders. Specifically, we assume that the set of scenarios S_v with the lowest 10% of end-of-century streamflows in each of the ensembles are vulnerable:

$$S_\mathrm{v} = \left\{ \tilde{Q}_{\mathrm{MAi}}(t) : \tilde{Q}_{\mathrm{MAi}}(t = 2100) < p_{10}\left(\tilde{Q}_{\mathrm{MA}}(t = 2100) \right) \right\}, \;\; i \in \{1, \ldots, n\} \qquad (2)$$

where the subscript i denotes a member of the ensemble of streamflow projections (up to a total of n) and the operator p_{10} refers to the percentile of lowest end-of-century average streamflow values. Similarly, we define the set of scenarios that are not vulnerable, S_n:

$$S_\mathrm{n} = \left\{ \tilde{Q}_{\mathrm{MAi}}(t) : \tilde{Q}_{\mathrm{MAi}}(t = 2100) > p_{10}\left(\tilde{Q}_{\mathrm{MA}}(t = 2100) \right) \right\}, \;\; i \in \{1, \ldots, n\} \qquad (3)$$

The choice to use the 10th percentile of normalized moving-average streamflow to define vulnerability is an illustrative choice, as these are likely the scenarios that would cause most concern among water resources planners. However, this choice is not based on local water supply needs at each location, which would require a more detailed discussion with stakeholders to determine. Instead, this methodology is developed with the intention that it can be easily applied using different criteria to define vulnerable scenarios.

After defining the set of vulnerable scenarios, we assume that for each year *t*, there exists a threshold *T*(*t*) which will identify a scenario as "vulnerable" if the normalized streamflow $\tilde{Q}_{MA}$ falls below it. This threshold is determined by finding the maximum value of $\tilde{Q}_{\mathrm{MA}}$ in the ensemble below which a certain percent of streamflow projections are considered vulnerable by the end of the century (Fig. 3c). This percent is the model agreement, *m*.

$$T(t) = \max\left(\tilde{Q}_{\mathrm{MA}}(t) : \frac{\left|\left\{S_{\mathrm{v}} \leq \tilde{Q}_{\mathrm{MA}}(t)\right\}\right|}{\left|\left\{S_{\mathrm{v}} \leq \tilde{Q}_{\mathrm{MA}}(t)\right\}\right| + \left|\left\{S_{\mathrm{n}} \leq \tilde{Q}_{\mathrm{MA}}(t)\right\}\right|} \geq m\right) \tag{4}$$

where $\left|\left\{S_{\mathrm{v}} \leq \tilde{Q}_{\mathrm{MA}}(t)\right\}\right|$ denotes the count of the subset of vulnerable scenarios with streamflow less than $\tilde{Q}_{\mathrm{MA}}(t)$ for each year, and $\left|\left\{S_{\mathrm{n}} \leq \tilde{Q}_{\mathrm{MA}}(t)\right\}\right|$ is a count of the subset of scenarios that are not vulnerable with streamflow less than $\tilde{Q}_{\mathrm{MA}}(t)$ for each year. In this equation the model agreement, *m*, is expressed as a decimal value. The threshold *T*(*t*) cannot be solved for explicitly. Instead, numerical iteration is used to find the first (highest) value of the threshold where this condition is met. The process is repeated for each year, *t*. This study considers model agreement values ranging from 40 to 80%. The higher the model agreement, the more stringent the threshold criteria will be. The window size used for this process is also varied from 20 to 60 years to further explore the sensitivity of the approach to parameter assumptions, recognizing that the 60-year window size only covers 2010–2100 (10 years fewer than for other window sizes) because all climate projection data starts in 1949.

Two potential issues can occur with this approach to defining thresholds. First, for some rivers in some years, the threshold *T*(*t*) may be undefined because no values of $\tilde{Q}_{\mathrm{MA}}$ can be found where at least 60% of the scenarios below it are vulnerable. In general, a high model agreement makes it difficult to draw a threshold because the criteria cannot be met. In this case, interpolation is used to fill in threshold gaps. If gaps in the threshold occurred at the beginning or end of the century, no threshold was drawn for those years. Years without a threshold value are referred to as missing points, as defined in Table 1. The second issue that can occur, even when *T*(*t*) is defined, is a small sample size of scenarios below the threshold. In this study we do not consider sampling uncertainty in the definition of *T*(*t*), as the method of defining thresholds is intended to be illustrative. However, in future work, it may be possible to propagate this uncertainty into the testing framework, where the threshold could be defined as a range or a distribution instead of a single value.

2.3 Leave-one-out testing

We aim to determine the capability of a threshold-based planning approach to correctly identify vulnerable scenarios before they become vulnerable, i.e., well before the end of the century. For example, this warning signal would be especially useful for infrastructure investment. To test how accurately the threshold identifies vulnerable scenarios, a leave-one-out testing methodology is used. This approach tests the classification accuracy of the planning threshold for a scenario that was not used to determine the threshold initially. For each river site, one of the 97 streamflow projections is removed from the ensemble, and the threshold is identified using the data from the remaining 96 scenarios. Next, the removed scenario is tested against the threshold to see whether it can be correctly classified as vulnerable or not vulnerable in each year (Fig. 3d). This evaluation process is repeated for each of the 97 scenarios, and then for each of the 91 rivers.

For each river site in each year, the leave-one-out testing methodology classifies each scenario as either vulnerable or not. The classification is therefore either a true positive ($TP_i(t)$), false positive ($FP_i(t)$), false negative ($FN_i(t)$), or true negative ($TN_i(t)$). Figure 4 shows a confusion matrix defining the four possible outcomes of this classification. The following logical relationships were used to determine the classifications at each year for each scenario in each river:

$$\text{True positive}: \quad \mathrm{TP_i}(t) = \tilde{Q}_{\mathrm{MA_i}}(t) < T(t) \text{ and } \tilde{Q}_{\mathrm{MA_i}}(t) \in S_{\mathrm{v}} \tag{5}$$

$$\text{False positive}: \mathrm{FP_i}(t) = \tilde{Q}_{\mathrm{MA_i}}(t) < T(t) \text{ and } \tilde{Q}_{\mathrm{MA_i}}(t) \notin S_{\mathrm{v}} \tag{6}$$

$$\text{False negative}: \mathrm{FN_i}(t) = \tilde{Q}_{\mathrm{MA_i}}(t) > T(t) \text{ and } \tilde{Q}_{\mathrm{MA_i}}(t) \in S_{\mathrm{v}} \tag{7}$$

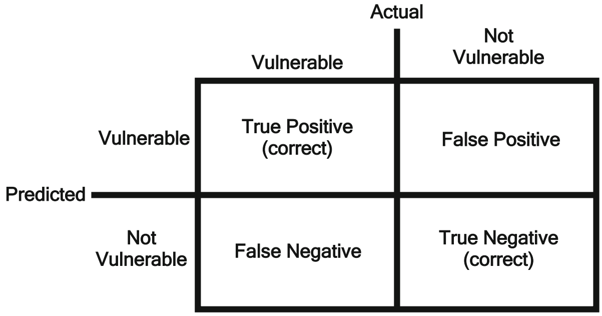
$$\text{True negative}: \mathrm{TN_i}(t) = \tilde{Q}_{\mathrm{MA_i}}(t) > T(t) \text{ and } \tilde{Q}_{\mathrm{MA_i}}(t) \notin S_{\mathrm{v}} \tag{8}$$

These expressions will result in binary values. The total error rates for a given river site in year t represent the fraction of leave-one-out scenarios ($n = 97$) in which that outcome occurred, e.g.:

$$TP(t) = \frac{\sum_{i=1}^{n} TP_i(t)}{n} \tag{9}$$

where i is the index of the ensemble member left out in validation. Years for which the threshold $T(t)$ is undefined (missing points) that cannot be interpolated are not considered in these error calculations.

Data Availability All code and data used in this study is available on Github (https://github.com/brobinson3/Testing_Thresholds).



		Actual	
		Vulnerable	Not Vulnerable
Predicted	Vulnerable	True Positive (correct)	False Positive
	Not Vulnerable	False Negative	True Negative (correct)

Fig. 4 Confusion matrix defining classification outcomes for the vulnerability threshold experiment. True positives and true negatives are classified correctly; false positives and false negatives are classified incorrectly

3 Results

3.1 Thresholds for Individual River sites

Figure 5 shows ensemble streamflow projections with vulnerability thresholds identified for four selected river sites. These time series represent the normalized moving average flow, $\tilde{Q}_{MA}$. The vulnerability threshold (red) represents the value of $\tilde{Q}_{MA}$ below which at least 60% of the streamflow projections are considered vulnerable by the end of the century. The black line shows the historical observed streamflow for comparison. Thresholds and historical data often show the same trends within the same basins. For instance, most rivers in the California Basin have missing threshold values for the first few decades of the century because their vulnerable scenarios cannot be separated reliably from non-vulnerable scenarios with the required model agreement (Fig. 5d). The ensemble spread by the end of the century is also similar across rivers within the same basin. For example, the American River (Fig. 5d) shows an ensemble spread of ±40% in the average streamflow by year 2100 compared to the historical mean, a substantial degree of uncertainty that is shared by other rivers in the California Basin.

Some rivers, such as the Columbia River (Fig. 5c), have historical flows that have already crossed below the vulnerability threshold. Following the logic of this threshold-based planning approach, this would indicate that long-term streamflow may be substantially reduced compared to historical streamflow, potentially resulting in water supply disruptions or less hydropower

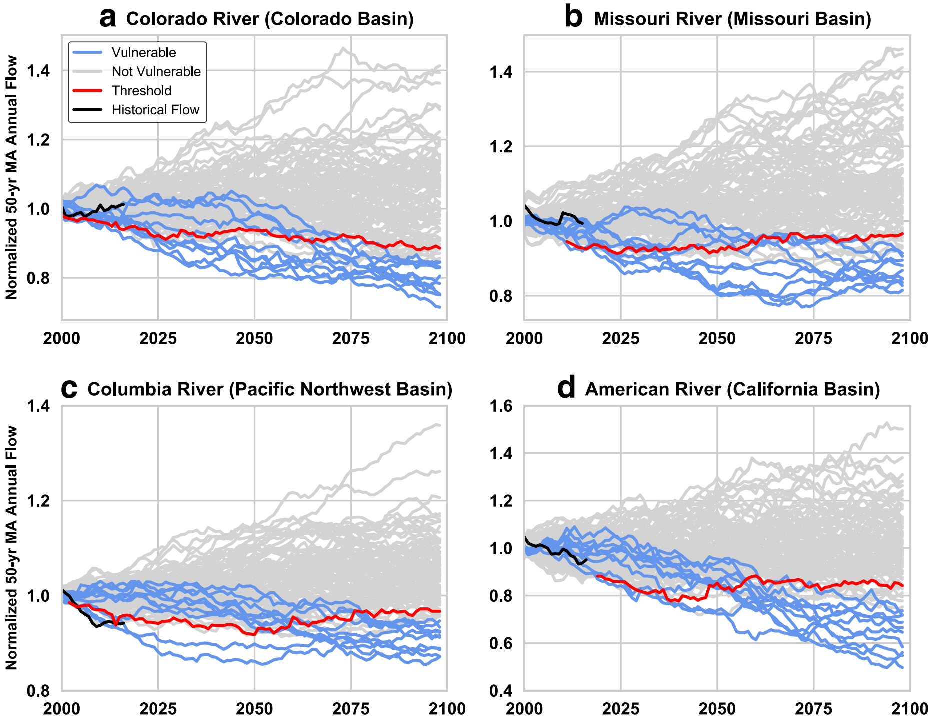


Fig. 5 Examples of the vulnerability threshold results for selected river sites. (**a**) The Colorado River near Cameo. (**b**) The Missouri River at Fort Peck Dam. (**c**) The Columbia River at Priest Rapids Dam. (**d**) The American River at Folsom Dam. The red line indicates the vulnerability threshold, and the black line shows the observed historical average for comparison

generation. However, without further statistical analysis, it is impossible to say whether this observed change in the average streamflow truly signals a long-term trend, or is only an anomaly that will eventually return to the long-term historical mean. In other words, a scenario crossing below the threshold could be a false positive. Similarly, a scenario remaining above the threshold could be a false negative—a scenario that should be classified as vulnerable, but is not. If such thresholds, or signposts, are to be used to trigger management actions, their reliability in correctly identifying vulnerable scenarios must be tested more rigorously.

3.2 Classification error rates

Figure 6 shows how the rates of classification errors change throughout the century using the leave-one-out testing methodology described in Sections 2.2–2.3. The error metrics are computed for each year and averaged across the 91 river sites. Figure 6a contains the true positive and false negative rates, which are the two possible outcomes for vulnerable scenarios and therefore sum to 100%. Similarly, Fig. 6b contains the true negative and false positive rates, which also sum to 100%.

False negative rates are high at the beginning of the century (roughly 80%) and remain high until the end of the century, dropping to 20% by 2075. However, the false positive rates are comparatively low, peaking at about 10%. High false negative rates mean that the classification of a river as “not vulnerable” does not provide much information, as it is likely a false negative. On the other hand, low false positive rates suggest that vulnerable classifications are likely correct. From a decision-making standpoint, the implications of both false positives and false negatives are important. If thresholds are used to trigger an irreversible decision, such as an infrastructure investment, it is important to be sure that it signifies a future vulnerability before taking action to secure water supply. High false positive rates would make it difficult to justify any action based on a vulnerable classification. High false negative rates, by contrast, indicate that many vulnerable scenarios are not being identified, which could lead to a water shortage situation for which planners have not prepared. Balancing the tradeoff between false negative and false positive rates remains an important decision for stakeholders.

The model agreement, *m*, was assigned a default value of 60% to demonstrate the threshold-based planning approach. The sensitivity of this choice was tested using model agreement values between 40 and 80%, the results of which are also shown in Fig. 6. The

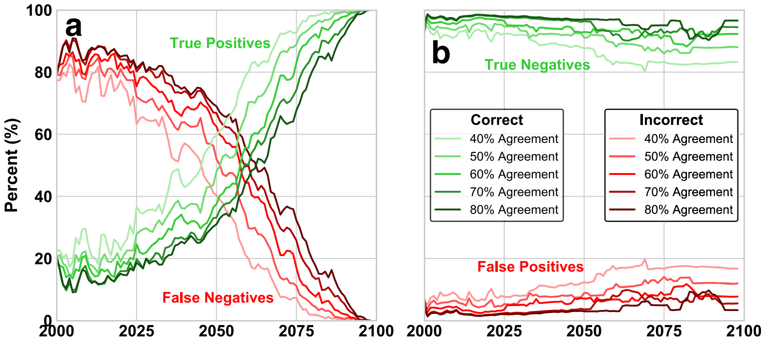


Fig. 6 Classification error rates from leave-one-out testing, averaged over all river sites for different values of the model agreement, *m*. Percentage of (**a**) true positives and false negatives, and (**b**) true negatives and false positives identified for multiple model agreements using a 50-year moving window. True positives and true negatives are correctly classified (green), while false positives and false negatives are incorrectly classified (red)

choice of model agreement reflects a tradeoff between false negative and false positive error rates; when one increases, the other decreases. The value of m essentially sets the end-of-century false positive rate, because it specifies the required fraction of vulnerable scenarios that must fall below the threshold. Increasing m decreases the false positive rates but increases the false negative rates, while decreasing m has the opposite effect. This choice offers a degree of freedom for decision makers to specify each of the error rates, depending whether false positives or false negatives are of greater concern for a particular planning problem. Figure 6 also suggests important implications for when during the century certain classifications can be trusted, according to this scenario ensemble. Positive (vulnerable) classifications are reliable throughout the century, while negative (not vulnerable) classifications only become reliable much later.

While the error rates associated with different levels of model agreement in Fig. 6 generally do not cross, one exception is the second half of the century for the false positive and true negative rates (Fig. 6b). The more stringent model agreements (70% and 80%) spike above others (50% and 60%) before returning to the original order at the very end of the century. This occurs because the more stringent model agreement values lead to many more missing points in the beginning of the century than other model agreements, which artificially deflates the false positive rates. In the second half of the century, there are fewer missing points and more error classification data, driving the false positive rates closer to their real values. The second half of the century would then be more accurately reflecting the error metrics of the more stringent model agreements (70% and 80%) than the beginning of the century when their error rates are artificially low. The supplemental Fig. S3 shows a graph of the missing points through time.

In addition to the model agreement, the second parameter that could affect the results is the window size for the moving average used in the threshold-setting approach. The moving window size represents how quickly the classification responds to new observations of annual streamflow: A classification based only on the most recent 5 years of data will likely be noisy and unreliable, while a very large window of 100 years might respond too slowly to identify an emerging trend. A default value of 50 years was assumed in the initial set of experiments. To test the sensitivity of this assumption, the window size was varied between 20 and 60 years using a model agreement of 60%. Figure 7 shows the classification error rates resulting from these experiments. The 20- and 30-year moving window sizes show much more noise in the

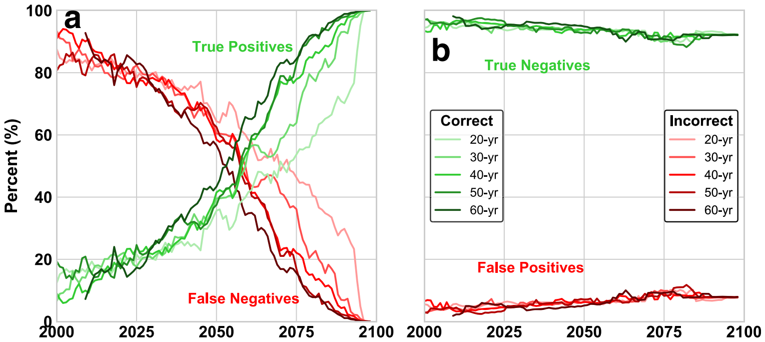


Fig. 7 Classification error rates for different moving average window sizes. Percentage of (**a**) true positives and false negatives, and (**b**) true negatives and false positives identified for multiple window sizes using a 60% model agreement

error rates than the larger window sizes because classifications will change more frequently through time when the moving average is taken over fewer years. Shorter window sizes also show a delayed decrease in the false negative rate, possibly due to the high amount of noise present in those window sizes. The 60-year window size has the lowest rate of false negatives for the second half of the century, suggesting that a long-term average that responds more slowly to new observations might be preferred if the rate of false negatives is to be reduced. The 50-year window size appears to have lower false negative rates than the 60-year window size in the first half of the century and provides consistently low false positive rates compared to the other window sizes, justifying the original choice of the 50-year moving average.

3.3 Spatially distributed error characteristics

Figures 6 and 7 show the results of this testing framework in terms of the dynamics of classification error rates through time. Because the dataset of climate projections used in this study includes many river sites, these results can be further explored to identify regional patterns in the error metrics. Figure 8 explores two characteristics of the classification errors at each site, chosen to reflect decision-relevant problems. First, Fig. 8a shows the year that the false negative rate drops below 50% for each river. In years preceding those shown in Fig. 8a, there is greater than a 50% chance that any negative (not vulnerable) classification is incorrect. False negative rates greater than 50% are particularly significant for a binary classification problem, because this indicates that the classification of negative (not vulnerable) scenarios is worse than random chance. After this point, negative classifications for each river are expected to become more accurate, a result observed in Figs. 6 and 7. Many of the river sites shown in Fig. 8a have false negative rates above 50% until at least 2050, which calls into question the utility of this particular threshold-based vulnerability classification scheme to confidently identify scenarios as not vulnerable in the near term.

While the testing framework has identified unacceptably high false negative rates for this experiment, the comparatively low false positive rates suggest one way that this approach to threshold-based vulnerability classification might be useful. Scenarios with normalized average streamflow crossing below the vulnerability threshold are highly likely to be those from the vulnerable set, S_v, with an error rate typically below 10% even in the near-term. Given these low error rates for positive (vulnerable) classifications, water planners may want to know how soon this threshold crossing may occur. Ideally, the threshold crossing would occur well in advance of when the normalized streamflow falls below the end-of-century vulnerability threshold (i.e., $\tilde{Q}_{\mathrm{MA}_\mathrm{i}}(t) < \tilde{Q}_{\mathrm{MA}0.10}$ for this study). This advance warning would allow time to build infrastructure or change policies to prevent unwanted, costly outcomes. Figure 8b shows the average year that the vulnerable scenarios first cross below the threshold for each river, resulting in correct positive vulnerability classifications. Similarities can be seen within regions, with the Colorado Basin having the earliest average crossing years. Combined with the generally low false positive rates observed in Figs. 6 and 7, this result suggests that reliable positive (vulnerable) classifications can be obtained for many of these river sites before 2040. This spatial analysis reveals the extent to which the classification performance identified with this framework is consistent within regions. In an idealized centralized planning context, this would enable investments in infrastructure or revised operating rules to be targeted spatially as well as in time.

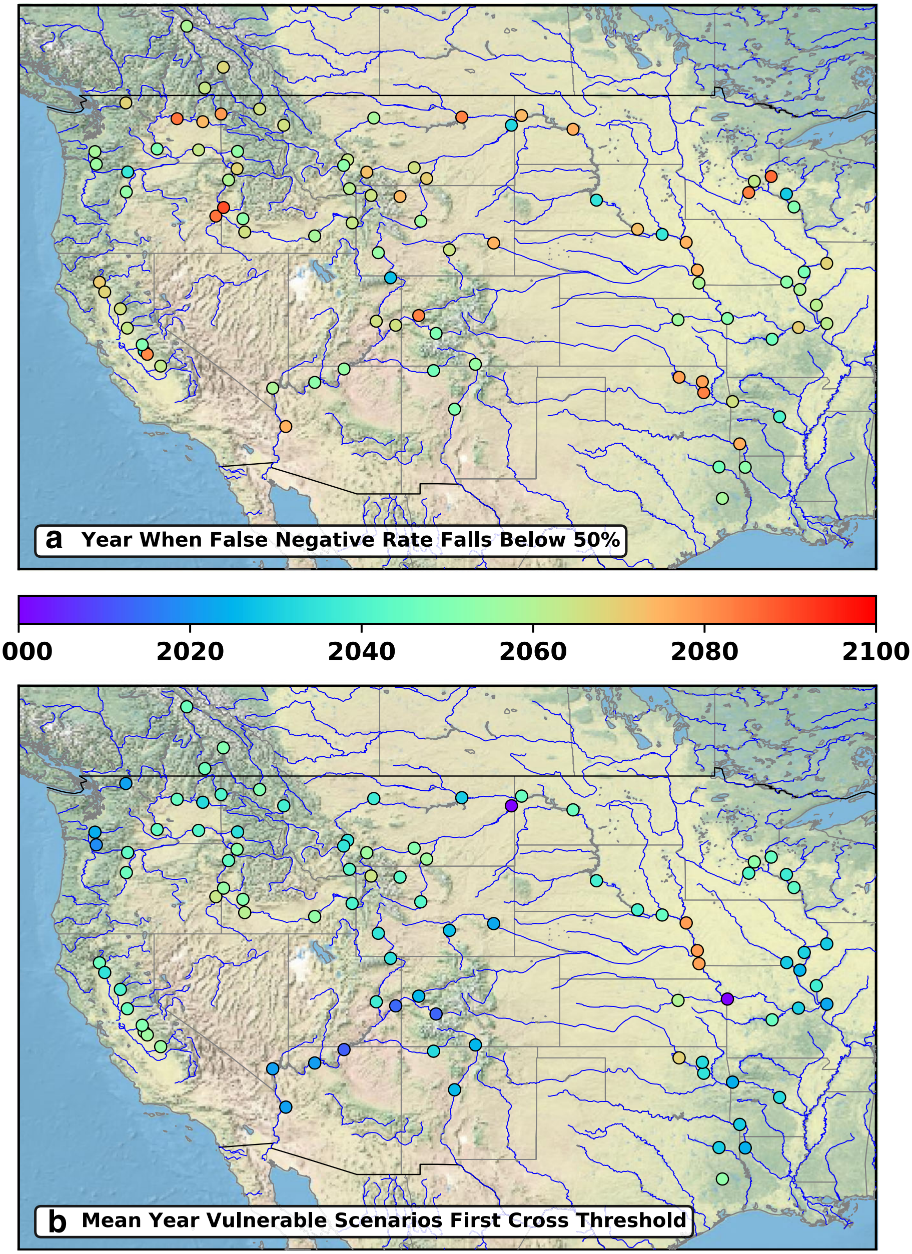

a Year When False Negative Rate Falls Below 50%
2000
2020
2040
2060
2080
2100
b Mean Year Vulnerable Scenarios First Cross Threshold

Fig. 8 (**a**) Year when false negative rate drops below 50% (i.e., random chance). (**b**) The average year of the first threshold crossing of vulnerable scenarios for each site, resulting in correct positive classifications

4 Discussion and conclusion

This study presents a framework for testing the ability of adaptation thresholds to dynamically identify vulnerable scenarios within ensemble projections. To demonstrate the testing method, an illustrative approach to defining dynamic thresholds has been developed, using ensemble projections of streamflow through the end of the century. The results of this experiment show that in general, these thresholds are able to classify vulnerable scenarios in advance, with low false positive rates (< 10%). However, the false negative rates remain high for many rivers until the second half of the century, becoming better than random chance after roughly 2060. This reduces confidence in these particular near-term adaptation thresholds and highlights the tradeoff observed in many threshold-based decision methods between triggering unnecessary action or failing to identify potential vulnerabilities until later in the century. Because false positive rates remain relatively low, positive classifications can reliably be used to trigger action to protect water supply. However, the uncertainty due to high false negative rates makes it difficult to define a river as not vulnerable until very late in the century. Given the low rate of false positives, a possible metric of interest for planning is the average year the vulnerable scenarios first cross below the threshold. These years ranged from as early as 2010 to as late as 2080 and showed a similar range of years within hydrologic regions.

Assigning thresholds for different rivers depends on the choice of model agreement, *m*, and the window size used to update the observed average streamflow. Testing the sensitivity of these parameter assumptions shows that when model agreements are more stringent, false negative rates are higher and false positive rates are lower. Therefore, choosing a value of the model agreement parameter controls the tradeoff between false positives and false negatives. Additionally, smaller window sizes (< 40 years) are more susceptible to noise in the climate projections and exhibit higher classification error rates than larger window sizes. The tradeoff between false negatives and false positives, and its relationship to parameter choices, would be of importance to decision makers tasked with determining infrastructure investments. In practice, this framework could be used to test planning thresholds developed with more advanced techniques, including optimization (Kwakkel et al. 2015) or data assimilation (e.g., Hui et al. 2018). By estimating the rates of incorrect classifications, decision makers would be able to make informed decisions regarding the choice of planning thresholds. False negatives could mean waiting too long to respond to observed trends; false positives might lead to unnecessary (and costly) actions. Though the thresholds developed in this study are only illustrative, some of the rivers analyzed in this paper exhibit long-term trends in observed mean flows that have already dropped below these thresholds, such as the Columbia River (Fig. 5). This crossing may be an indicator of a long-term reduction in flows, possibly causing a disruption in water supply or hydropower generation, though statistics suggest that it may be too early to tell.

The results of this paper show that combining ensemble projections with planning thresholds for climate adaptation can offer new and important insights to how thresholds can be created and tested. However, any application of climate model projections comes with limitations. GCM-based ensembles cannot represent the full range of climate possibilities; they only represent the lower bound of uncertainty (Stainforth et al. 2007). In the specific case of streamflow, this uncertainty is amplified by downscaling and hydrologic modeling (Wilby and Dessai 2010). Recognizing this uncertainty, this study treats the projections as the full range of possible futures and investigates the impacts of uncertainty on vulnerability classification. The leave-one-out cross-validation method also assumes that each scenario in the ensemble is independent, even though some scenarios are derived from the same RCPs or GCMs, and may be correlated. It may be possible to expand the ensemble of scenarios with synthetic streamflow generation designed to mimic the statistics of climate change projections (e.g., Borgomeo et al. 2015; Herman et al. 2016). As an additional limitation, the definition of "vulnerability" as the lowest 10% of mean annual streamflow is not based on the specific water supply needs of the river basins considered here. Instead, it is intended that water planners using this methodology would substitute system-specific definitions of vulnerability for the end of the century, potentially involving multivariate criteria. Finally, while this experiment focused on detecting changes in the long-term average streamflow, it could be expanded to account for changes in flood or drought frequency or changes in seasonal timing of runoff, effects of climate change that have already been observed and are projected to continue.

Testing the classification accuracy of adaptive planning thresholds using GCM ensembles has provided key insights into how to characterize the reliability of thresholds through time to trigger climate change adaptation. The leave-one-out cross-validation was instrumental in this approach, by treating individual ensemble members as potential realizations of the future. Positive and negative classifications show very different error rates for the illustrative thresholds used in this study, and the average timing of positive (i.e., vulnerable) classifications is region-dependent. Several alternate versions of this study could be pursued with the understanding that most adaptations would not require more than a decade to implement. One approach would be to identify vulnerable scenarios every 10 years and reset the threshold accordingly, or to compute classification accuracy 10 years before each scenario becomes vulnerable. Because this study only defined vulnerability based on end-of-century streamflows, the error rates at roughly 2090 could be emphasized. However, an analysis focused so far in the future would likely not be helpful to decision makers compared to an extended study that redefines vulnerable scenarios dynamically. Future research will consider the relationship between classification accuracy and other hydrologic variables in each river basin, such as annual flow, river classification, and regional climate to attempt to understand how these variables affect the assignment of planning thresholds and their error rates. This methodology could also be expanded to other climate adaptation problems such as sea level rise or flood risk, in which planned actions would be linked to dynamically updated observations of other variables.

Acknowledgements Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views or policies of the NSF. We further acknowledge the World Climate Research Program's Working Group on Coupled Modeling and the climate modeling groups listed in the Supplement of this paper for producing and making available their model output.

Funding information This work was partially supported by the U.S. National Science Foundation grants CNS-1639268 and CNH-1716130.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Adger WN, Arnell NW, Tompkins EL (2005) Successful adaptation to climate change across scales. Glob Environ Chang 15(2):77–86. https://doi.org/10.1016/j.gloenvcha.2004.12.005

Anderson J, Chung F, Anderson M, Brekke L, Easton D, Ejeta M, Peterson R, Snyder R (2008) Progress on incorporating climate change into management of California's water resources. Clim Chang 87(1 SUPPL). https://doi.org/10.1007/s10584-007-9353-1.

Barnett TP, Pierce DW, Hidalgo HG, Bonfils C, Santer BD, Das T, Bala G, Wood AW, Nozawa T, Mirin A a, Cayan DR, Dettinger MD (2008) Human-induced changes in the hydrology of the Western United States. Science 319:1080–1083. https://doi.org/10.1126/science.1152538

Borgomeo E, Farmer CL, Hall JW (2015) Numerical rivers: a synthetic streamflow generator for water resrouces vulnerability assessments. Water Resour Res 51:5382–5405. https://doi.org/10.1002/2014WR016827.Received

Brekke L, Wood A, Pruitt T (2014) Downscaled CMIP3 and CMIP5 hydrology climate projections: release of hydrology projections, comparison with preceding information, and summary of user needs, US Bureau of Reclamation. Available at: https://gdo-dcp.ucllnl.org/downscaled_cmip_projections/techmemo/BCSD5HydrologyMemo.pdf

Brown C, Ghile Y, Laverty M, Li K (2012) Decision scaling: linking bottom-up vulnerability analysis with climate projections in the water sector. Water Resour Res 48(9):1–12. https://doi.org/10.1029/2011WR011212

Bryant BP, Lempert RJ (2010) Thinking inside the box: a participatory, computer-assisted approach to scenario discovery. Technol Forecast Soc Chang 77(1):34–49. https://doi.org/10.1016/j.techfore.2009.08.002

Buurman J, Babovic V (2016) Adaptation pathways and real options analysis: an approach to deep uncertainty in climate change adaptation policies. Policy Soc 35(2):137–150. https://doi.org/10.1016/j.polsoc.2016.05.002

Cayan DR, Kammerdiener SA, Dettinger MD, Caprio JM, Peterson DH (2001) Changes in the onset of spring in the Western United States. Bull Am Meteorol Soc 82(3):399–416

Dessai S, Hulme M (2004) Does climate adaptation policy need probabilities? Clim Pol 4(2):107–128. https://doi.org/10.1080/14693062.2004.9685515

DiFrancesco KN, Tullos DD (2014) Flexibility in water resources management: review of concepts and development of assessment measures for flood management systems. J Am Water Resour Assoc 50(6): 1527–1539. https://doi.org/10.1111/jawr.12214

Donat MG et al (2013) Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: the HadEX2 dataset. J Geophys Res Atmos 118(5):2098–2118. https://doi.org/10.1002/jgrd.50150

Dottori F, Szewczyk W, Ciscar J-C, Zhao F, Alfieri L, Hirabayashi Y, Bianchi A, Mongelli I, Frieler K, Betts R, Feyen L (2018) Increased human and economic losses from river flooding with anthropogenic warming. Nat Clim Chang. https://doi.org/10.1038/s41558-018-0257-z.

Fletcher SM, Miotti M, Swaminathan J, Klemun MM, Strzepek K, Siddiqi A (2017) Water supply infrastructure planning: decision-making framework to classify multiple uncertainties and evaluate flexible design. J Water Resour Plan Manag 143(10):4017061. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000823.

Frigg R, Smith LA, Stainforth DA (2013) The Myopia of Imperfect Climate Models : The Case of UKCP09, 80(December):886–897

Giuliani M, Castelletti A (2016) Is robustness really robust? How different definitions of robustness impact decision-making under climate change. Clim Chang 135(3–4):409–424. https://doi.org/10.1007/s10584-015-1586-9.

Haasnoot M, Kwakkel JH, Walker WE, ter Maat J (2013) Dynamic adaptive policy pathways: a method for crafting robust decisions for a deeply uncertain world. Glob Environ Chang 23(2):485–498. https://doi.org/10.1016/j.gloenvcha.2012.12.006

Hallegatte S (2009) Strategies to adapt to an uncertain climate change. Glob Environ Chang 19(2):240–247. https://doi.org/10.1016/j.gloenvcha.2008.12.003

Hallegatte S, Shah A, Lempert R, Brown C, Gill S (2012) Investment decision making under deep uncertainty: application to climate change, Policy Research Working Paper, (6193), p 41. https://doi.org/10.1596/1813-9450-6193

Herman JD, Giuliani M (2018) Policy tree optimization for threshold-based water resources management over multiple timescales. Environ Model Softw 99:39–51. https://doi.org/10.1016/j.envsoft.2017.09.016

Herman JD, Reed PM, Zeff HB, Characklis GW (2015) How should robustness be defined for water systems planning under change? J Water Resour Plan Manag 141(10):4015012. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000509.

Herman JD, Zeff HB, Lamontagne JR, Reed PM, Characklis GW (2016) Synthetic drought scenario generation to support bottom-up water supply vulnerability assessments. J Water Resour Plan Manag 142(11):4016050. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000701.

Hirabayashi Y, Mahendran R, Koirala S, Konoshima L, Yamazaki D, Watanabe S, Kim H, Kanae S (2013) Global flood risk under climate change. Nat Clim Chang 3(9):816–821. https://doi.org/10.1038/nclimate1911.

Hui R, Herman J, Lund J, Madani K (2018) Adaptive water infrastructure planning for nonstationary hydrology. Adv Water Resour 118(May):83–94. https://doi.org/10.1016/j.advwatres.2018.05.009

Jeuland M, Whittington D (2014) Water resources planning under climate change: assessing the robustness of real options for the Blue Nile. Water Resour Res:2086–2107. https://doi.org/10.1002/2013WR013705.Received.

Katz RW (2002) Techniques for estimating uncertainty in climate change scenarios and impact studies. Clim Res 20:167–185

Knowles N, Dettinger MD, Cayan DR (2006) Trends in snowfall versus rainfall in the Western United States. J Clim 19(18):4545–4559. https://doi.org/10.1175/JCLI3850.1

Kundzewicz ZW, Krysanova V, Benestad RE, Hov O, Piniewski M, Otto IM (2018) Uncertainty in climate change impacts on water resources. Environ Sci Policy 79:1–8. https://doi.org/10.1016/j.envsci.2017.10.008

Kwadijk JCJ, Haasnoot M, Mulder JPM, Hoogvliet MMC, Jeuken ABM, van der Krogt RAA, van Oostrom NGC, Schelfhout HA, van Velzen EH, van Waveren H, de Wit MJM (2010) Using adaptation tipping points to prepare for climate change and sea level rise: a case study in the Netherlands. Wiley Interdiscip Rev Clim Chang 1(5):729–740. https://doi.org/10.1002/wcc.64

Kwakkel JH, Pruyt E (2013) Exploratory modeling and analysis, an approach for model-based foresight under deep uncertainty. Technol Forecast Soc Chang 80(3):419–431. https://doi.org/10.1016/j.techfore.2012.10.005

Kwakkel JH, Haasnoot M, Walker WE (2015) Developing dynamic adaptive policy pathways: a computer-assisted approach for developing adaptive strategies for a deeply uncertain world. Clim Chang 132(3):373–386. https://doi.org/10.1007/s10584-014-1210-4

Lempert RJ (2002) A new decision sciences for complex systems. Proc Natl Acad Sci 99(Supplement 3):7309–7313. https://doi.org/10.1073/pnas.082081699

Lempert RJ, Collins MT (2007) Managing the risk of uncertain threshold responses: comparison of robust, optimum, and precautionary approaches. Risk Anal 27(4):1009–1026. https://doi.org/10.1111/j.1539-6924.2007.00940.x

Leung LR, Qian Y, Bian X, Washington WM, Han J, Roads JO (2004) Mid-century ensemble regional climate change scenarios for the western United States. Clim Chang 62(1–3):75–113. https://doi.org/10.1023/B:CLIM.0000013692.50640.55.

Liang X, Lettenmaier DP, Wood EF, Burges SJ (1994) A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J Geophys Res Atmos 99(D7):14415–14428. https://doi.org/10.1029/94JD00483

Luckman BH (1998) Landscape and climate change in the Central Canadian Rockies during the 20th century. Can Geogr 42(4):319–336. https://doi.org/10.1111/j.1541-0064.1998.tb01349.x

Mote PW, Hamlet AF, Clark MP, Lettenmaier DP (2005) Declining mountain snowpack in Western North America, (January), pp. 39–49. https://doi.org/10.1175/BAMS-86-1-39

Pederson GT, Gray ST, Woodhouse CA, Betancourt JL, Fagre DB, Littell JS, Watson E, Luckman BH, Graumlich LJ (2011) The unusual nature of recent snowpack declines in the North American Cordillera 543(July): 332–336

Ranger N, Reeder T, Lowe J (2013) Addressing "deep" uncertainty over long-term climate in major infrastructure projects: four innovations of the Thames estuary 2100 project. EURO J Decis Process 1(3–4):233–262. https://doi.org/10.1007/s40070-013-0014-5

Ray PA, Bonzanigo L, Wi S, Yang YCE, Karki P, García LE, Rodriguez DJ, Brown CM (2018) Multidimensional stress test for hydropower investments facing climate, geophysical and financial uncertainty. Glob Environ Chang 48(January 2017):168–181. https://doi.org/10.1016/j.gloenvcha.2017.11.013

Seaber PR, Kapinos FP, Knapp GL (1987) Hydrologic unit maps: US Geological Survey Water Supply Paper 2294.

Stainforth DA, Allen MR, Tredger ER, Smith LA (2007) Confidence, uncertainty and decision-support relevance in climate predictions. Philos Trans R Soc A Math Phys Eng Sci 365(1857):2145–2161. https://doi.org/10.1098/rsta.2007.2074

Steinschneider S, McCrary R, Mearns LO, Brown C (2015) The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning. Geophys Res Lett 42(12):5014–5022. https://doi.org/10.1002/2015GL064529

Stewart IT, Cayan DR, Dettinger MD (2005) Changes toward earlier streamflow timing across Western North America. J Clim 18(8):1136–1155. https://doi.org/10.1175/JCLI3321.1

Taner MÜ, Ray P, Brown C (2017) Robustness-based evaluation of hydropower infrastructure design under climate change. Climat Risk Manag 18(July):34–50. https://doi.org/10.1016/j.crm.2017.08.002

US Global Change Research Program (2009) Global climate change impacts in the United States. Cambridge University Press

Walker WE, Haasnoot M, Kwakkel JH (2013) Adapt or perish: a review of planning approaches for adaptation under deep uncertainty. Sustainability (Switzerland) 5(3):955–979. https://doi.org/10.3390/su5030955

Wilby RL, Dessai S (2010) Robust adaptation to climate change. Weather 65(7):180–185. https://doi.org/10.1002/wea.504.

Zeff HB, Kasprzyk JR, Herman JD, Reed PM, Characklis GW (2014) Navigating financial and supply reliability tradeoffs in regional drought management portfolios. Water Resour Res:4906–4923. https://doi.org/10.1002/2013WR015126.Received.

Zeff HB, Herman JD, Reed PM, Characklis GW (2016) Cooperative drought adaptation: integrating infrastructure development, conservation, and water transfers into adaptive policy pathways. Water Resour Res 52: 7327–7346. https://doi.org/10.1002/2016WR018771