Earth’s Future

RESEARCH ARTICLE
10.1029/2020EF001650

Can Exploratory Modeling of Water Scarcity Vulnerabilities and Robustness Be Scenario Neutral?

J. D. Quinn1, A. Hadjimichael2, P. M. Reed2, and S. Steinschneider3

1Department of Engineering Systems and Environment, University of Virginia, Charlottesville, VA, USA, 2School of Civil and Environmental Engineering, Cornell University, Ithaca, NY, USA, 3Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, USA

Abstract Planning under deep uncertainty, when probabilistic characterizations of the future are unknown, is a major challenge in water resources management. Many planning frameworks advocate for “scenario-neutral” analyses in which alternative policies are evaluated over plausible future scenarios with no assessment of their likelihoods. Instead, these frameworks use sensitivity analysis to discover which uncertain factors have the greatest influence on performance. This knowledge can be used to design monitoring programs and adaptive policies that respond to changes in the critical uncertainties. However, scenario-neutral analyses make implicit assumptions about the range and independence of the uncertain factors that may not be consistent with the coupled human-hydrologic processes influencing the system. These assumptions could influence which factors are found to be most important and which policies are most robust, belying their neutrality; assuming uniformity and independence could have decision-relevant implications. This study illustrates these implications using a multistakeholder planning problem within the Colorado River Basin, where hundreds of rights holders vie for the river’s limited water under the law of prior appropriation. Variance-based sensitivity analyses are performed to assess users’ vulnerabilities to changing hydrologic conditions using four experimental designs: (1) scenario-neutral samples of hydrologic factors, centered on recent historical conditions, (2) scenarios informed by climate projections, (3) scenarios informed by paleohydrologic reconstructions, and (4) scenario-neutral samples of hydrologic factors spanning all previous experimental designs. Differences in sensitivities and user robustness rankings across the experiments illustrate the challenges of inferring the most consequential drivers of vulnerabilities to design effective monitoring programs and robust management policies.

Plain Language Summary How we should best manage our water resources depends on future water supply and demand, both of which are changing in uncertain ways as a consequence of climate change, population growth, and sectoral change. This makes it challenging to decide between alternative management plans that may be most favorable under different conditions. Given this uncertainty, many planning frameworks advocate for implementing “robust” management plans that perform reasonably well over a wide range of conditions. These plans can then be modified adaptively to cater to particular climate and socioeconomic conditions as our uncertainty about the future is reduced. Unfortunately, we find that the determination of which plan is most robust, and what conditions should trigger adaptation, is an additional challenge. In assessing the vulnerabilities of hundreds of water users in the Upper Colorado River Basin to different possible drought conditions, we find different users to be robust depending on which scenarios are considered. The sensitivities of these users’ water shortages to different climate conditions also depend on which scenarios are considered. Since we do not know which scenarios are most plausible, we recommend that climate vulnerability assessments consider multiple sets of scenarios when evaluating alternative water management plans and designing monitoring programs.

1. Introduction

Appropriately selecting, sizing, and operating water infrastructure to reduce the impacts of droughts and floods is an exercise in hedging uncertainty (Herman et al., 2020). Underbuild and you risk severe socioeconomic impacts from water scarcity or flooding; overbuild and you risk stranding assets for decades. Traditional planning mechanisms hedge against these risks by designing systems to be robust to historical
climate variability. Yet changing climate and socioeconomic conditions make these methods inappropriate for the “deeply” uncertain nature of the future (Brown et al., 2020). Deep uncertainty refers to conditions under which planners do not know, or cannot agree on, the probability distribution of the parameters describing a system model, its boundary conditions, or the model itself (Lempert & Collins, 2007).

Recent work on decision making under deep uncertainty seeks to find alternative infrastructure designs, or policies for managing that infrastructure, that are not only robust to historical climate variability but to possible future climate and socioeconomic conditions as well. Different sets of such possible conditions compose “scenarios” of the future. In “scenario-neutral” analyses, these scenarios are generated by sampling possible values of uncertain climatic and socioeconomic variables independently and uniformly over expansive ranges meant to encompass possible future values (Prudhomme et al., 2010; Wilby & Dessai, 2010). Examples of such variables are mean annual temperature, 100-yr rainfall volume, population, and the water intensity of human diets. Alternative infrastructure designs and policies are then evaluated over these scenarios with no assessment of their likelihoods, since they are unknown. Approaches applying this philosophy include Robust Decision Making (RDM) (Hadjimichael, Quinn, Wilson et al., 2020; Lempert et al., 2010; Moallemi, Elsawah & Ryan 2020; Shortridge & Guikema, 2016) and its Multi-Objective extension, MORDM (Hadjimichael, Reed, & Quinn, 2020; Herman et al., 2014; Kasprzyk et al., 2013; Quinn, Reed, & Keller, 2017), info-gap decision theory (Ben-Haim, 2006; Hine & Hall, 2010; Korteling et al., 2013), and decision scaling (Brown et al., 2012; Freeman et al., 2020; Knighton et al., 2017; Ray et al., 2018; Steinschneider et al., 2015).

The goal of scenario-neutral analyses is to use exploratory modeling (Bankes, 1993) to discover under what conditions (i.e., combinations of values of the uncertain factors defining the scenarios) alternative designs no longer meet satisfactory performance (Bryant & Lempert, 2010; Dittrich et al., 2016; Herman et al., 2015; Maier et al., 2016; Moallemi, Zare, et al., 2020). This process, also called “scenario discovery,” is a form of factor-mapping sensitivity analysis (Saltelli et al., 2008). From this mapping, one can learn when to adapt their current system design as the conditions migrate to regions of failure (Culley et al., 2016; Whateley et al., 2014).

However, determining exactly when and how to adapt in the face of changing conditions is an additional challenge. Subsequent research has moved toward using factor-ranking sensitivity analysis to determine what uncertain factors most control system performance (Herman et al., 2015; Whateley & Brown, 2016). This can help determine what uncertainties should be monitored to detect that the system is likely moving toward a region of failure (Haasnoot et al., 2018; Hermans et al., 2017; Raso et al., 2019). New actions can then be triggered dynamically to maintain satisfactory performance (Haasnoot et al., 2013). For example, in studying the vulnerability of the Metropolitan Water District of Southern California, Groves et al. (2015) found estimates of the utility’s future water supply reliability depended most on assumptions about their future demand and on water availability in the bay/delta region from where they import water, as opposed to changing local climate conditions or groundwater yield. They discuss how this knowledge could be used to design adaptive policies in which increases in demands or decreases of supply in the bay trigger additional infrastructure development. Implementing this plan would require monitoring these two factors but not local climate and groundwater conditions. Knowing this can therefore save money on monitoring investments in groundwater observation wells since this factor is not as important.

After using scenario-neutral analyses to determine factors to monitor, optimal control methods can be applied to optimize values of these factors that should “trigger” adaptive management decisions such as expanding reservoir capacity, building a desalination plant, or raising levee heights (Fletcher, Lickley, & Strzepek, 2019; Fletcher, Strzepek, et al., 2019; Groves et al., 2015; Kwakkel et al., 2015, 2016; Mortazavi-Naeini et al., 2015; Trindade et al., 2017, 2019; Woodward et al., 2014; Zeff et al., 2016). Applying such optimization approaches for dynamic adaptation requires that performance across the scenarios be aggregated into an objective function (Herman et al., 2020). Consequently, a probability distribution must be specified. A “neutral” approach typically assumes all scenarios are equally likely over a prespecified range, that is, that they have a uniform probability density function (pdf); scenarios outside of that range are implicitly deemed impossible (given zero probability).

Some have considered these assumptions unrealistic, choosing instead to optimize alternative designs to be robust to probability distributions informed by climate projections (Fletcher, Lickley, & Strzepek, 2019;
Herman & Giuliani, 2018). However, there are known shortcomings of this approach as well: Climate projections underrepresent climate variability and persistence (Brown & Wilby, 2012), are not independent (Knutti et al., 2013; Steinschneider et al., 2015), have known biases that cannot be corrected via downscaling (Ehret et al., 2012), and are also implicitly bounded by forcing scenarios that only provide a lower bound on the true range of future uncertainty (Lamontagne et al., 2018; Stainforth et al., 2007).

Clearly neither of these approaches is entirely appropriate. Policies may be overdesigned if the scenarios are too broad and unrealistic or underdesigned if the range is too narrow. Herman et al. (2020) argue that policies should therefore be optimized over multiple assumed probability distributions to test the sensitivity of the optimal solutions to these assumptions. Bartholomew and Kwakkel (2020) illustrate such a sensitivity analysis, optimizing lake management plans to be robust to deep uncertainties in the lake's phosphorus recycling and loss parameters under alternative robust optimization frameworks. Each of these optimizations can be considered a “rival framing” that might reveal unintended consequences or unforeseen benefits of optimizing to different assumed probability distributions (Quinn, Reed, Giuliani & Castelletti 2017). We argue that performing such sensitivity analyses may be equally important in the vulnerability analysis, so planners should assess the sensitivity and robustness of alternative water management policies using multiple scenario designs and assumed probability distributions. As noted by Saltelli et al. (2020), “the technique is never neutral”; rather “the choice of the methodology conditions the narrative produced by an analysis.”

In this study, we explore how vulnerability assessments performed over competing hypotheses of how future hydrology might evolve dictate which uncertainties are found to most control water shortages for different users in an institutionally complex, multiactor system and, subsequently, which users are found to be most robust. Several studies have compared how robustness ranks of alternative management strategies or multiple water users (i.e., policies and objectives) differ under alternative definitions of robustness (Hadjimichael, Quinn, Wilson, et al., 2020; Herman et al., 2015; Giuliani & Castelletti, 2016; McPhail et al., 2018; Spence & Brown, 2018) or under alternative assumptions about the range and joint distribution of uncertain factors (i.e., the experimental design) (Moody & Brown, 2013; Reis & Shortridge, 2019; Taner et al., 2019). Yet none of these studies has explored if and how the importance of uncertain factors differs under alternative experimental designs. This could have decision-relevant implications since such sensitivity analyses are often an advised first step for designing monitoring programs (Kwakkel et al., 2016) and optimizing triggers for adaptive management policies (Groves et al., 2015). Furthermore, differences in sensitivities across designs could explain why we see differences in robustness ranks across them.

To clarify these concerns, the next section presents a stylized example of how the choice of experimental design itself might influence robustness analyses. We then investigate how this problem manifests in the Upper Colorado River Basin within the state of Colorado, where hundreds of water users vie for the region’s limited water under the doctrine of prior appropriation. We assess each of these user’s sensitivities to different hydrologic parameters under rival framings of how the future might evolve. In essence, we perform a sensitivity analysis of our sensitivity analysis (Noacco et al., 2019; Paleari & Confalonieri, 2016; Puy et al., 2020; Shin et al., 2013) to see if our conclusions change under alternative assumptions about the range and correlation of uncertain hydrologic parameters.

2. Conceptualization of the Problem

Figure 1 presents a stylistic example of how the experimental design for a robustness analysis might influence which climate uncertainties are found to be most important to monitor and which policies most robust to these uncertainties. The numerical details of this illustrative example are provided in the supporting information (SI). In Experimental Design 1 on the left, two policies are evaluated over a range of changes in precipitation (x axis of Figures 1a and 1b) and temperature (y axis of Figures 1a and 1b) from current conditions (black point). These ranges are chosen to span a set of precipitation and temperature observations from different periods of the paleorecord (green points). The regions in which each policy does or does not satisfy some minimum performance criterion are shown in blue and red, respectively.

A scenario-neutral risk assessment would find the policy with the greater blue region to be more robust, which in this case is Policy 1. A paleo-informed risk assessment might compute the probability of achieving different performance levels using a probability distribution fit to the Paleo points (Figure 1e). One could integrate the area under this pdf within the blue success region to determine which policy is most robust.
(Figure 1f), again concluding it is Policy 1. To determine which factor is most important for monitoring, one could use the scenario-neutral experiment to decompose which factors most explain variability in each policy’s performance, here reliability. This would conclude that Policy 1 is most influenced by changes in temperature (Figure 1i), while Policy 2 is nearly equally influenced by changes in temperature and precipitation (Figure 1j).

These analyses and conclusions, while reasonable, are strongly dependent on the use of the paleodata to define the ranges of precipitation and temperature, as well as their joint distribution used to calculate robustness. Another analyst might have a different perspective on what are plausible future scenarios. For example, they might also consider climate projections from the Coupled Model Intercomparison Project (CMIP) in their analysis. These projections, shown in yellow in Figures 1c and 1d, might span beyond the range of temperature and precipitation explored in Experimental Design 1 (outlined in black in Figures 1c and 1d). Designing a second scenario-neutral experimental design to encompass both sets of points (Experimental Design 2) would arrive at different conclusions. Under this design, the success region for Policy 2 is now larger, as well as its probability of success when integrated over a probability distribution fit to both the Paleo and CMIP points (Figures 1g and 1h). Not only that, but conclusions about which factors are most important to monitor also change, as precipitation and temperature now equally in
fluence performance under Policy 1 (Figure 1k), while precipitation now dominates performance under Policy 2 (Figure 1l).

Due to the deep uncertainty in future conditions, it is not clear which, if any, of these experimental designs is “right.” What is more concerning is that the two experimental designs lead to different conclusions on which policy is more robust and therefore should be implemented. Similarly, under the two designs, the analysis
identifies different uncertainties that most influence the robustness of each policy, resulting in different conclusions about how to allocate monitoring investments (e.g., in tipping buckets to track changes in precipitation or thermometers to track changes in temperature). In the following sections, we explore whether we see such consequences in the real-world setting of the Upper Colorado River Basin.

3. Study Area and Model

The headwaters of the Upper Basin of the Colorado River (UCRB) originate at the Continental Divide and flow southwest to the Colorado-Utah state line (Figure 2), draining an area of 25,682 km² (9,915 mi²). Within and outside of the UCRB, hundreds of stakeholders own rights to the river’s flow, which annually averages 6.9 billion m³ (5.6 million acre-ft). These users include municipalities, industries, irrigation districts, and power plants, among others. The primary consumptive water use within the UCRB is for irrigation, with diversions for this purpose irrigating 930 km² (230,000 acre-ft) (State of Colorado, 2015). Yet many of the basin’s demands actually lie east of the Continental Divide, where 80% of Colorado’s population resides. To meet these users’ needs, approximately 569 million m³ of water (461,000 acre-ft) are annually diverted eastward outside the basin through tunnels in the Colorado Rockies (State of Colorado, 2015), pictured in Figure 2. These users include the cities of Denver and Colorado Springs. Other demands include municipal and industrial uses on the west slope, recreational uses, and fisheries. The remaining flows are either stored in reservoirs with capacity totaling just under 1.8 billion m³ (1.5 million acre-ft), diverted for power generation and returned downstream (1.3 billion m³/yr, or 1 million acre-ft/yr), or left for the environment. The flows at the outlet contribute downstream deliveries to Lake Powell that are required by the Colorado River Compact.

Concern is growing over whether these deliveries can be met in the future under climate change without curtailing upstream users. Since the turn of the century, inflows to Lake Powell have exceed the mean only five times. Paleohydrologic reconstructions in the basin suggest that decadal and multidecadal droughts are not uncommon in the UCRB (Woodhouse et al., 2006; Ault et al., 2013, 2014), but modeling suggests anthropogenic warming has exacerbated the emerging megadrought (Williams et al., 2020). Furthermore, rising temperatures are expected to increase agricultural water demands and result in earlier snowmelt, decreasing summer flows in the growing season (Christensen & Lettenmaier, 2006; Christensen et al., 2004; Rasmussen et al., 2014), a trend that is already being observed (Xiao et al., 2018; Milly & Dunne, 2020). These recent

Figure 2. Map of the Upper Basin of the Colorado River within the state of Colorado. Major transbasin diversions are indicated with large black circles, while all other diversions (primarily for irrigation) are indicated with small olive circles. Figure adapted from Hadjimichael, Quinn, Wilson, et al. (2020).
climatic trends, compounded by population growth and development in the basin, suggest UCRB users could be greatly impacted by climate change.

In this study, we investigate the vulnerability of UCRB rights holders to potential changes in hydrologic conditions using the State of Colorado's Stream Simulation Model, StateMod. StateMod was jointly developed by the Colorado Water Conservation Board (CWCB) and the Division of Water Resources (DWR) to aid water resources planning in each of the State’s major basins (Malers et al., 2001). StateMod simulates streamflows, diversions, environmental flow demands, and reservoir operations in the basin according to federal operating rules and the “law of the river” specifying how water is allocated by priority. It relies on detailed historical demand and operation records, including individual water right information for all consumptive use and diversions from all water structures (wells, ditches, reservoirs, and tunnels). Irrigation demands in StateMod are typically computed by a separate model, StateCU, based on historical soil moisture, crop type, irrigated acreage, and conveyance and application efficiencies for each individual irrigation unit.

StateMod simulations take as input all natural flows (flows that would occur without human diversions) and water demands and use the law of the river to compute the volume of diversions for each user. All consumptive use in the basin is modeled, although some small structures with decrees less than 0.3 m³/s (11 ft³/s) (25% of them) are aggregated into larger structures for model simplicity. This results in nearly 350 key diversion structures (CWCB & CDWR, 2016), each of which we consider a different user. Similarly, only reservoirs whose capacities exceed 4.9 million m³ (4,000 acre-ft) are modeled explicitly (accounting for 94% of total storage in the system), with the remaining storage aggregated into 10 reservoirs and one stock pond (CWCB & CDWR, 2016). The model runs at a monthly time step, and reports the volumes of water diverted to each structure and their demand, as well as the flows along all reaches. The model can be run with historical flow and demand data, or synthetic data.

In this study, we use historical demand data but generate synthetic flows to assess the effects of changing hydrologic conditions on the basin’s rights holders in absence of demand growth or conservation. While we do not change mean irrigation demands, we do ensure their historical correlation with streamflow is preserved through the synthetic generator.

4. Methods

4.1. Synthetic Streamflow Generator

The synthetic streamflow generator used in this study is based on a two-state Hidden Markov Model (HMM) that has been shown to accurately capture the extreme hydrologic variability and persistence observed in the basin and projected for the future (Bracken et al., 2014). The two states in the model represent wet and dry years, which tend to cluster throughout the historical record due to persistence in large-scale climate phenomena, such as the Pacific Decadal Oscillations (PDO) and El Niño–Southern Oscillation (ENSO). Time series of wet and dry years are generated from a Markov model, and flow volumes in those years are then generated from a distribution conditioned on the state. Since only the flow volume is observed, the states are “hidden” and can only be inferred. We assume the distribution of annual flows under each state is log-normal.

The model is fit to annual flows at the Colorado-Utah state line, the last stream node in StateMod. This requires estimation of six parameters: the mean and standard deviation of the dry-state ($\mu_d$ and $\sigma_d$, respectively) and wet-state ($\mu_w$ and $\sigma_w$, respectively) Gaussian distributions, as well as the probabilities of

Figure 3. Two-state Gaussian HMM fit to the historical record from 1944–2013. (a) The log space historical distribution of annual flows in gray, with the fitted dry- and wet-state distributions overlain in red and blue, respectively, and their combined distribution in black. (b) State identification of historical years as dry in red or wet in blue.
transitioning from a dry state in Year $t$ to a dry state in Year $t + 1$ ($p_{d,d}$) and from a wet state in Year $t$ to a wet state in Year $t + 1$ ($p_{w,w}$). Note that $p_{d,w}$ and $p_{w,d}$ are immediately given by $1 - p_{d,d}$ and $1 - p_{w,w}$ respectively. For an exploratory analysis, we can change these six parameters to determine which hydrologic conditions most influence users’ shortage and to map what combinations of parameters lead to unsatisfactory performance.

Figure 3 shows the fit of the two-state Gaussian HMM to historical flows at the Colorado-Utah state line from 1944–2013, and Table 1 lists the parameter values. The model was fit to this period, the final two thirds of the 105-yr record, to capture recent conditions in the nonstationary system, and to approximately match the length of the CMIP projections (64 yr). Parameter estimation was performed using Expectation-Maximization with Python’s hmmlearn package (Lebedev, 2015). The fitted dry-state and wet-state distributions exhibit nontrivial overlap and together provide a strong fit to the observed distribution (Figure 3a). As seen in Table 1, there is also strong persistence in the underlying states, with $p_{d,d} = 0.68$ and $p_{w,w} = 0.65$. This can also be seen from the state identification (Figure 3b), performed using the Viterbi algorithm in the hmmlearn package. More details on the fitting method and validation of the generator are provided in the SI (Figures S1 and S2).

For our exploratory modeling experiment, we modify the parameters of the historical two-state Gaussian HMM using delta shifts of the transition probabilities and multipliers of the means and standard deviations. These modifications are determined using four different experimental designs, described below. For each parameterization, 10 realizations of 105-yr-long time series of log space annual flows at the Colorado-Utah state line are synthetically generated from the two-state Gaussian HMM. The log space annual flows are then converted to real space and temporally downscaled to monthly flows using a modification of the proportional scaling method used by Nowak et al. (2010).

First, a historical year is probabilistically selected based on its “nearness” to the synthetically generated flow in terms of annual total. The proportions of the annual flow delivered each month of the historical year are then applied to the synthetic annual flow to generate synthetic monthly flows. Similarly, the synthetic monthly flows at all upstream StateMod nodes are generated by applying the historical year’s ratios between the monthly flows at the upstream nodes and the monthly flow at the Colorado-Utah state line. Validation of the generator’s ability to capture spatial correlation with this approach is provided in the SI (Figure S3), with links to the code for the synthetic streamflow generator. These monthly time series are provided as input to StateMod for the vulnerability assessment.

Demand time series for the experiment use the maximum historical transbasin diversion demands, recent historical municipal and industrial demands, and synthetically generated irrigation demands. For irrigation demands, it is assumed their mean and correlation with annual streamflows is unchanged. To ensure this, we use a regression between historical annual flow anomalies and annual irrigation demand anomalies, totaled across all users in the basin. Details of the regression and its performance are provided in the SI (Figure S4). Based on the synthetically generated annual flow anomaly, a total annual irrigation anomaly is generated from this regression, with added noise to preserve variance. The time series of total annual irrigation anomalies is then added to the mean and distributed to the irrigation structures using their average historical proportion of the total demand. While the mean demands across all sectors will likely change in the future in deeply uncertain ways, we focus this analysis exclusively on hydrologic changes for simplicity. However, based on our prior work in the basin using one experimental design (Hadjimichael, Quinn, Wilson, et al., 2020), we expect the influence of changing demands to be significant. Future work can explore how these effects differ depending on the assumed correlation structure between mean demand and streamflow.

### 4.2. Alternative Experimental Designs

Determining what scenarios to include in a vulnerability assessment is not straightforward. A key first step of exploratory modeling and robustness-focused methods is to work collaboratively with system experts and stakeholders to identify major uncertainties exogenous to the system that might influence its performance, such as climatic, demographic, and political conditions (Marchau et al., 2019). Scenarios are then generally designed to sample these uncertainties broadly enough to encompass possible future conditions but not so far as to be unreasonable. Of course, these initial modeling steps can create path dependencies in how model-based inferences are made. This is particularly true for contentious, high stakes analyses where
different experts and stakeholders may have different opinions of what future conditions are possible versus unreasonable (see discussions in Moallemi, Zare, et al., 2020). Consider hydrologic conditions as an example. One expert might generate future streamflow scenarios from paleohydrologic reconstructions because these are conditions that have been observed in the past. Another expert might believe the past is not representative of the future and design streamflow scenarios from climate model projections. Other experts might agree the past is not representative of the future but also think climate models are insufficient in capturing future climate uncertainty. They might instead design scenario-neutral streamflow scenarios but assuming different ranges of streamflow statistics are plausible.

In this study, we illustrate four alternative sets of hydrologic scenarios framed around these possible schools of thought. While not considered here, similar tensions arise in systems where the relative influences of internal variability, model structural uncertainties, and longer-term trends are dynamic and region-specific (Lehner et al., 2020). In practice, alternative experimental designs should be elicited from stakeholders and experts based on their perception of the system’s key uncertainties, their plausible ranges, and their

### Table 1

| Parameter | Description                        | Maximum likelihood estimate |
|-----------|------------------------------------|-----------------------------|
| $\mu_d$   | log space dry-state mean, m$^3$ (acre-ft) | 22.38 (15.26)               |
| $\sigma_d$| log space dry-state standard deviation | 0.26                        |
| $\mu_w$   | log space wet-state mean, m$^3$ (acre-ft) | 22.78 (15.66)               |
| $\sigma_w$| log space wet-state standard deviation | 0.25                        |
| $P_{d,d}$ | dry-to-dry transition probability   | 0.68                        |
| $P_{w,w}$ | wet-to-wet transition probability   | 0.65                        |

Note. Using Maximum Likelihood Estimation with Expectation-Maximization, a two-state Gaussian HMM is fit to log space annual flows at the Colorado-Utah state line, the last node in StateMod.

### Table 2

| Parameter | Experiment      | Current value | Lower bound | Upper bound |
|-----------|-----------------|---------------|-------------|-------------|
| $\mu_d$   | multiplier      | 1.0           | 0.98        | 1.02        |
|           | CMIP            | 1.0           | 0.97        | 1.03        |
|           | Paleo           | 1.0           | 0.90        | 1.01        |
|           | All-Encompassing| 1.0           | 0.90        | 1.03        |
| $\sigma_d$| multiplier      | 1.0           | 0.75        | 1.25        |
|           | CMIP            | 1.0           | 1.14        | 1.38        |
|           | Paleo           | 1.0           | 0.80        | 2.63        |
|           | All-Encompassing| 1.0           | 0.75        | 2.63        |
| $\mu_w$   | multiplier      | 1.0           | 0.98        | 1.02        |
|           | CMIP            | 1.0           | 0.98        | 1.03        |
|           | Paleo           | 1.0           | 0.98        | 1.01        |
|           | All-Encompassing| 1.0           | 0.97        | 1.03        |
| $\sigma_w$| multiplier      | 1.0           | 0.75        | 1.25        |
|           | CMIP            | 1.0           | 0.81        | 1.12        |
|           | Paleo           | 1.0           | 0.69        | 1.22        |
|           | All-Encompassing| 1.0           | 0.39        | 1.25        |
| $P_{d,d}$ | delta           | 0.0           | $-0.30$     | 0.30        |
|           | CMIP            | 0.0           | $-0.01$     | 0.10        |
|           | Paleo           | 0.0           | $-0.65$     | 0.07        |
|           | All-Encompassing| 0.0           | $-0.65$     | 0.30        |
| $P_{w,w}$ | delta           | 0.0           | $-0.30$     | 0.30        |
|           | CMIP            | 0.0           | $-0.07$     | 0.06        |
|           | Paleo           | 0.0           | $-0.33$     | 0.33        |
|           | All-Encompassing| 0.0           | $-0.33$     | 0.33        |

Note. Parameters from the Box Around Historical and All-Encompassing experiments are sampled uniformly and independently over these ranges. Parameters for the CMIP and Paleo experiments are estimated from data, and the resulting ranges are shown. The CMIP and Paleo samples are not uniform or independent. Multipliers are applied to the log space parameter values, estimated from the annual flows in acre-ft.
correlation structure. These designs will likely include far more uncertainties than hydrologic parameters and differ across experts, highlighting the importance of analyzing how different design assumptions might lead to different conclusions about future monitoring needs and management decisions.

4.2.1. Box Around Historical Experiment

We first consider a “scenario-neutral” experimental design commonly used in RDM analyses, in which the parameters of the two-state Gaussian HMM listed in Table 1 are varied independently and uniformly over prespecified ranges (Lempert et al., 2010; Prudhomme et al., 2010). These ranges are simply meant to be expansive and enable the discovery of failure boundaries, that is, combinations of the parameters under which different users no longer meet satisfactory performance levels. We call this experiment the “Box Around Historical” experiment because the parameter ranges were chosen to expand above and below the historical values, creating a hypercube around the historical parameters. The ranges sampled for each parameter are provided in Table 2. The real-space equivalent across all of the samples generated from this experiment are provided in the SI (Table S2).

For this experiment, 1,000 samples were generated using Latin hypercube sampling over the ranges in Table 2. In some of these samples, the dry-state mean was greater than the wet-state mean. In these cases, the wet- and dry-state parameter labels were swapped. After relabeling, some points were then outside of the sampled ranges in Table 2, so these points were removed, leaving 985 samples. The remaining sample points are shown in salmon in Figure 4. One can see they form a hypercube (“box”) around the historical parameter values, shown in blue, with the exception of the lower right corner of the \( \mu_w \) versus \( \mu_d \) panel where the dry-state mean would exceed the wet-state mean. Generating 10 realizations of 105-yr time series of annual flows at the Colorado-Utah state line from each of these samples results in the range of annual flows shown in salmon in Figure 5. This range extends beyond that of historical annual flows (shown in blue), as well as annual flows from the CMIP5 experiment, shown in yellow and described next.

Finally, since some of the experimental designs in this study contained far less than 1,000 samples, we also repeated the Box Around Historical experiment with a Latin hypercube sample size of 100 for the SI. After relabeling and removing samples outside the parameter ranges in Table 2, this resulted in 96 samples. By repeating the sensitivity analysis with a smaller sample, we test the stability of our findings and ensure differences between experiments cannot be ascribed to the different sample sizes but rather their different correlation structures and ranges.

4.2.2. CMIP5 Experiment

The second experimental design is informed by climate projections from the CMIP5 data set. Several iterations of CMIP projections have been used in numerous studies assessing potential impacts of climate change in the UCRB (Christensen & Lettenmaier, 2006; Christensen et al., 2004; Rasmussen et al., 2014). On average, these projections suggest future deliveries to Lake Powell will decrease. However, there is great variability across these projections as well as within the time series of each projection (Harding et al., 2012).

This study makes use of 97 CMIP5 projections used in the Colorado River Water Availability Study (CWCB, 2012). In each of these projections, monthly precipitation factor changes and temperature delta changes were computed between mean projected 2035–2065 climate statistics and mean historical climate statistics from 1950–2013. These 97 different combinations of 12 monthly precipitation multipliers and 12 monthly temperature delta shifts were applied to historical precipitation and temperature time series from 1950–2013. The resulting climate time series were run through a Variable Infiltration Capacity (VIC) model of the UCRB, resulting in 97 time series of projected future streamflows at the Colorado-Utah state line.

Since these projections rely on VIC simulations that will underestimate variability relative to observations (Farmer & Vogel, 2016), we add errors to the projected streamflows using a model of the error between historical VIC streamflow simulations and observations. Details of the historical error model and how it is applied to the CMIP projections are provided in the SI (see Figure S5). After adding noise to the CMIP projections, we then fit a two-state Gaussian HMM to the resulting time series. We repeat this with 100 realizations of added noise and use the mean Gaussian HMM parameter estimates as our sample points. Using this approach on the historical VIC simulations, we see that the mean HMM parameter estimates across the VIC + noise simulations more closely match the observed record’s HMM parameter estimates than fitting the HMM directly to the VIC outputs (see SI Figure S6). Across the 97 CMIP-forced VIC models, this method results in the 97 parameter combinations shown in yellow in Figure 4.
For each of these CMIP samples, we generate 10 realizations of 105-yr streamflow inputs to StateMod. The range of annual flows generated across these simulations is similar to that of the Box Around Historical Experiment (see Figure 5), even though it is generated using an entirely different parameter set. The CMIP simulations have a near-perfect correlation between $\mu_d$ and $\mu_w$. If this correlation structure better describes the true joint distribution of plausible futures, sensitivity analysis using the Box Around Historical analysis could provide misleading conclusions. Conversely, the CMIP experiment could negatively impact conclusions if its correlation structure is an artifact of applying multipliers to historical precipitation and deltas to historical temperature to generate inputs to the VIC simulations. Another way to gain insight into potential correlation structures between parameters is to consider paleohydrologic reconstructions. This is our third experiment.

**4.2.3. Paleohydrologic Experiment**

The Paleohydrologic experiment (Paleo) relies on reconstructed Colorado River streamflows at Cisco, Utah, from Woodhouse et al. (2006). Since Cisco, Utah, is a little downstream of the Colorado-Utah state line, these flows are bias corrected by multiplying the reconstructed Cisco flows by the average ratio of natural flows at the Colorado-Utah state line to natural Cisco flows from 1909–1995. During the period of overlap between...
the scaled, reconstructed flows and observed flows at the state line (1909–1997), these reconstructions are unbiased but less variable than the observed record (see SI Figure S7). Like the VIC simulations of the CMIP projections, we build a model of the residuals between reconstructed and observed flows to ensure the paleoreconstructions do not underestimate variability. Details of the error model and how it is applied to the paleoreconstructions before the observed record are provided in the SI (Figure S7).

Again, similar to the CMIP projections, we use this error model to add noise to 64-yr moving windows of the paleoreconstructions from 1569–1997. Shifting this window every year results in 366 sixty-four-year windows. A window of 64 yr was chosen since this is the same length as the CMIP projections. After adding noise to the annual flows in these windows, we then fit a two-state Gaussian HMM to the resulting time series. We repeat this with 100 realizations of added noise and use the mean Gaussian HMM parameter estimates as our sample points. The performance of this approach over the overlapping reconstruction and observed record is shown in SI Figure S8, where it can be seen that the mean HMM parameter estimates are closer to the historical HMM parameter estimates than if the HMM were fit directly to the reconstructed flows.

The HMM parameter estimates over the paleohydrologic 64-yr moving windows are shown in green in Figure 4. The corresponding range of synthetically generated annual flows from 10 realizations of 105-yr time series under each parameterization is shown in Figure 5. Much drier years are synthetically generated by the Paleo HMM parameters than under either the CMIP or Box Around Historical experiments, consistent with past paleohydrological studies in the basin (Woodhouse et al., 2006). In addition, the HMM parameters over the paleorecord have a completely different correlation structure than the CMIP projections. While there is still a positive correlation between \( \mu_d \) and \( \mu_w \), the relationship is different. There are also several interesting nonlinear relationships, for example, between \( \mu_d \) and \( p_{w,w} \) and between \( \mu_d \) and \( \sigma_w \). There is also a bifurcation in the relationship between \( \mu_d \) and \( \sigma_d \), with a different correlation structure across the two clusters than within them. This suggests the correlation structure could be complex. However, these samples are not independent, since they were generated from fits to overlapping windows to ensure a sufficient number of samples for sensitivity analysis. Similar to the CMIP experiment, this may lead to artifacts in parameter correlations that influence conclusions from the sensitivity analysis.

### 4.2.4. All-Encompassing Experiment

The Paleo experiment illustrates that the Box Around Historical simulations far underrepresent potential droughts in the basin (Figure 5), as well as potential Gaussian HMM parameters (Figure 4). Consequently, we consider one more experimental design that spans the parameter ranges explored across all experimental designs. This All-Encompassing experiment also assumes uniformity and independence across all parameters but over wider ranges than the Box Around Historical experiment (see Table 2; note the ranges extend beyond the minimum and maximum of all others to encompass a fifth experiment using the CMIP projections without added noise, which we are not discussing but which is described in the SI and illustrated in SI Figures S9, S10, and S12). We employ the same sampling strategy for the All-Encompassing experiment as for the Box Around Historical experiment, ultimately resulting in 932 parameter samples, shown in lavender in Figure 4. Their corresponding annual flow distribution extends far beyond all other experiments (Figure 5). Finally, to test the stability of the variance decomposition to sample size, we again repeated this experiment with only 100 Latin hypercube samples for the SI. After relabeling and removing points outside their sampled ranges, this sample size was reduced to 92.

### 4.3. Sensitivity Analysis

We investigate the sensitivity of water rights holders in the UCRB to changing hydrologic conditions under each of the above experiments using both factor ranking and factor-mapping approaches (Saltelli et al., 2008). Factor ranking methods are used to rank uncertain parameters from most influential to least influential.
This is useful for prioritizing which factors to monitor, a potentially expensive investment but one that is necessary to detect failure. Factor-mapping approaches are used to predict a performance metric, such as water shortage or the probability of satisfactory performance, as a function of uncertain parameters. This is useful for mapping the joint influence of different uncertain factors on performance to determine if the system is moving toward an unacceptable region under which new actions may be needed (e.g., building more infrastructure or increasing water efficiency). We use Sobol variance decomposition for factor ranking and linear and logistic regression for factor mapping.

4.3.1. Sobol Variance Decomposition

Sobol sensitivity analysis decomposes the variability in a response variable, \( Y \), into amounts contributed by each of \( n \) independent variables, individually and jointly (Sobol, 1993). In this study, we use Sobol variance decomposition to estimate how much of the variance in a UCRB user’s annual shortage (\( Y \)) can be explained by each of the Gaussian HMM parameters, where the \( i \)th parameter is denoted \( X_i \) and \( n = 6 \) for the six HMM parameters. Building off of Hadjimichael, Quinn and Reed (2020), we perform this decomposition at different percentiles of annual shortage. That is, for each StateMod simulation, each user experiences a different time series of annual shortages that form a distribution. Which of these shortages is of most concern may vary by user, depending on whether frequent or severe shortages are most impactful (Hadjimichael, Quinn, Wilson, et al., 2020; Hadjimichael, Quinn, & Reed, 2020). Consequently, we use Sobol sensitivity analysis to compute how much of the variability in a user's shortages at each percentile is explained by each Gaussian HMM parameter. These first-order contributions, \( V_{fi} \), represent the amount of variability in the output explained by each parameter, \( X_i \) individually. Higher-order contributions represent additional variability caused by the interaction of multiple variables.

In this study, we use Sobol variance decomposition with Python’s SALib package (Herman & Usher, 2017) to estimate first-order contributions only of each Gaussian HMM parameter in explaining a particular percentile of shortage for each UCRB user. We report each parameter’s first-order contribution, \( S_i \), as a fraction of the total variance in \( Y: S_i = V_{fi}/\text{Var}(Y) \). These are called first-order Sobol sensitivity indices. Note \( 1 - \sum_{i=1}^{n} S_i \) represents the portion of the variability in \( Y \) explained by interactions. This term can be positive or negative. Positive interactions indicate some of the variability in shortage can only be explained by simultaneous changes in more than one HMM parameter. This means the relationship between the HMM parameters and shortage is nonlinear. Negative interactions indicate some of the variability in shortage explained by each of the HMM parameters is redundant. This can only occur when the HMM parameters are correlated with one another. If parameters are redundant, we should be able to detect changes in expected shortages by monitoring only one of them. If parameters have positive, nonlinear interactions, we may need to monitor both to detect these changes.

4.3.2. Response Surface Modeling

In addition to determining which Gaussian HMM parameters are most important in explaining a particular percentile of shortage, we also use linear and logistic regression to create a response surface (Brown et al., 2012; Moody & Brown, 2013) that either predicts a particular percentile of shortage as a function of the parameters, or the probability of keeping a particular percentile of shortage below some threshold as a function of the parameters. We display these response surfaces as two-dimensional contour plots of the shortage/probability estimates given values of the two most predictive HMM parameters. The two most predictive parameters in explaining shortage are the two parameters with the greatest first-order Sobol sensitivity indices. Denoting these two variables \( X_1 \) and \( X_2 \), our contour plots display the predicted value of shortage, \( \hat{Y} \) according to Equation 1:

\[
\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2
\]  

(1)

The \( \beta \) coefficients are estimated using ordinary least squares regression using Python’s statsmodels package (Seabold & Perktold, 2010).

For a given percentile of shortage, we not only display an estimated value of shortage given the two most predictive HMM parameters but also an estimated probability that shortage is below different thresholds, \( T \), given the two most predictive HMM parameters. This is useful if users consider shortages above some threshold to be intolerable, allowing us to determine under what conditions those failures occur. Not only that, but the probabilistic rather than binary classification of these failures allows users to define regions...
of success based on their level of risk aversion (Lamontagne et al., 2019; Quinn et al., 2018). For example, they may define a region of success as staying below the threshold with 95% probability if they are highly risk averse, or only 50% probability if they are risk neutral.

The two HMM parameters that are most predictive of this probability are estimated by first fitting univariate logistic regression models that estimate the log odds that shortage is below some threshold $T$, as a function of each of the individual HMM parameters, $X_i$:

$$\ln \left( \frac{P(Y < T)}{1 - P(Y < T)} \right) = \beta_0 + \beta_1 X_i$$

where $\ln \left( \frac{P(Y < T)}{1 - P(Y < T)} \right)$ is the log odds. For each of these univariate models, we compute the McFadden’s pseudo-$R^2$:

$$R^2_{\text{McFadden}} = 1 - \frac{\ln L(M_{\text{full}})}{\ln L(M_{\text{intercept}})}$$

where $\ln L(M_{\text{full}})$ is the log-likelihood of the full model (model with $X_0$ and $X_i$ as predictors) and $\ln L(M_{\text{intercept}})$ is the log-likelihood of the intercept model (model with just $X_0$ as a predictor). McFadden’s pseudo-$R^2$ is therefore a measure of the improvement of predictor $X_i$ in estimating $P(Y < T|X_i)$ compared to always predicting the average probability of success. The two HMM parameters resulting in the largest $R^2_{\text{McFadden}}$ are therefore considered the two most predictive. After determining these two predictors, $X_1$ and $X_2$, we then make contour plots showing the estimated probability of success using a logistic regression model with both parameters and their interaction:

$$\ln \left( \frac{P(Y < T)}{1 - P(Y < T)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

or

$$P(Y < T) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2)}$$

Estimation of the logistic regression models was performed using maximum likelihood estimation with Python’s statsmodels package (Seabold & Perktold, 2010).

### 4.4. Robustness Analysis

Finally, robustness analyses not only use sensitivity analysis to determine to which uncertainties users are most sensitive, and under what conditions they fail, but also to rank users or management plans by their “robustness” across the possible future worlds investigated. Many definitions have been proposed to define how robust users are to these changes, and it has been noted that different definitions result in different conclusions about what plans or users are most robust (Giuliani & Castelletti, 2016; Herman et al., 2015; McPhail et al., 2018, 2019; Spence & Brown, 2018). This study is concerned with a slightly different question of whether or not the ranking of user robustness under a single metric is consistent across alternative experimental designs. For this investigation, we consider the domain satisfying criterion (Starr, 1969) denoted as the fraction of realizations in which a particular criterion is met. We choose hypothetical satisfying criteria, for example, “Shortage greater than 20% of demand occurs no more than 20% of the time,” and determine the percent of realizations in which each UCRB user meets these criteria under each experimental design. We then rank all of the users from most to least robust and compare the user rankings across experiments. We repeat this for several hypothetical satisfying criteria to see how sensitive the robustness rankings are to the experimental design when using different definitions of success for the satisfying metric.

### 5. Results

#### 5.1. Impacts of Hydrologic Change Across Experiments

Hundreds of UCRB water rights holders are modeled in StateMod, and we can examine the vulnerabilities of each of them to changing hydrologic conditions using our four experimental designs. For brevity, we focus
our results on comparing three users’ sensitivity differences across the four experiments, two in the main text and one in the SI. User 1 is an aggregation of irrigation users with a fairly senior right of moderately sized decree. However, User 1 is positioned upstream in the basin on a tributary to the Colorado River, making them potentially vulnerable to water shortages in the headwaters despite their right seniority. User 2 is another irrigation user with a much larger decree but less senior (midrank) priority. This user is located further downstream off the mainstem Colorado, though, where larger flows may be available to meet their demand if they are in priority. Finally, User 3 is a midstream, median right transbasin diversion serving mostly municipal uses on the eastern slope. This user is analyzed in the SI.

Figure 6 displays the distribution of shortages experienced by the two agricultural users in each of our four experimental designs, with User 1 on the top row (Figures 6a–6d) and User 2 on the bottom row (Figures 6e–6h). The shortage distributions for the transbasin diversion (User 3) are shown in SI Figure S11. In each panel of these figures, the black line indicates the historical cumulative frequency of different shortage levels. Shades of purple indicate the percent of realizations in each experimental design below different shortage levels at each percentile.

As highlighted previously by Hadjimichael, Quinn, Wilson, et al. (2020), despite modest changes in right seniority and no change in sector, two users in the same basin can experience starkly different shortage distributions across historical and alternative hydrologic conditions. Here we see that these differences also depend on the experimental design. Looking first at User 1 (Figures 6a–6d), we can see that across all experiments, small magnitude shortages become less severe relative to historical, while large magnitude shortages become more severe, increasing the variability of this user’s shortage across years. However, as would be expected, the range of shortages across the four experiments varies, with the All-Encompassing experiment experiencing the widest range (Figure 6d) due to its expansive sampling. It is somewhat surprising that the differences in the ranges of shortage experienced in the CMIP experiment (Figure 6b) and the Paleo experiment (Figure 6c) are not greater given their extremely different parameterizations (Figure 4) and annual...
flow ranges (Figure 5). Therefore, similar impacts can be achieved in different ways, suggesting the parameter sensitivities for User 1 may be different in these two experiments.

The implications are similar for User 2 (Figures 6e–6h). Across experiments, this user’s shortages tend to become more severe relative to historical, but the impacts are not very significant under the Box Around Historical (Figure 6e) and CMIP (Figure 6f) experiments. Once again, the ranges of shortage experienced across these two experiments is very similar despite their different parameterizations, suggesting sensitivities may be different. Under the Paleo experiment (Figure 6g), this user’s most severe shortages increase significantly, while their smaller shortages are not greatly impacted. Under the All-Encompassing experiment (Figure 6h), their shortage magnitude increases at all percentiles.

5.2. Sobol Variance Decomposition

Figure 7 shows how the sensitivities of these two users differ across experiments based on the Sobol sensitivity analysis for each percentile of shortage, while SI Figure S11 shows the same for the transbasin diversion. To illustrate the stability of the variance decomposition with smaller sample sizes, we also compare the Box Around Historical and All-Encompassing decompositions with 100 versus 1,000 initial samples in SI Figures S13 and S14, respectively. In both figures, the x axis in each plot represents the percentile of shortage and the y axis represents the portion of the variability in shortage at that percentile that is explained by each HMM parameter (or their interactions).

For all users, there are clearly strong differences in sensitivities across experiments. Under the Box Around Historical experiment, User 1’s low percentiles of shortage are most explained by the wet-state mean with smaller, relatively equal contributions from the dry-state mean and transition probabilities (Figure 7a). At higher percentiles of shortage, the influence of the wet-state mean decreases to the level of the transition probabilities, while that of the dry-state mean increases. The influences of the dry- and wet-state standard deviations are negligible across percentiles, but there are significant positive interactions between parameters, as seen by the lavender color filling at least 20% of the decomposition across percentiles. This indicates nonlinearity in the shortage response to the HMM parameters. The All-Encompassing experiment, which makes the same assumptions of uniformity and independence in the HMM parameters but over
different ranges, follows a similar shape (Figure 7d). The strength of the parameters' first-order sensitivities just become larger in the All-Encompassing experiment, with the importance of positive interactions decreasing to about 10% across percentiles.

However, User 1’s sensitivities under the CMIP and Paleo experiments are completely different from the other two and each other. Under the CMIP experiment (Figure 7b), shortage at almost all percentiles is explained most by the wet- and dry-state means in relatively equal magnitudes. The dry-state standard deviation is also more influential across percentiles in this design, with only minor contributions from the other parameters. Except for the most extreme shortage percentiles, the first-order sensitivity indices sum to more than 1, meaning the parameters explain redundant information, resulting in negative interactions. This is due to the near-perfect correlation (and consequently, nearly equal influence) of the wet- and dry-state mean parameters, which is likely a consequence of using the delta change method to generate precipitation and temperature time series for the CMIP projections. If only the CMIP projections were used for this vulnerability assessment, one would conclude that either the dry-state mean or wet-state mean could be monitored to detect changes in shortage for User 1. Yet if the true correlation structure in the future is more similar to the past, User 1’s sensitivities under the Paleo experiment suggest the opposite could be true.

Under the Paleo experiment (Figure 7c), for most of the shortage distribution, none of the parameters alone explain much of the variability across realizations. Instead, complex nonlinear relationships between them are needed to understand the cause of this variation. Even at the most extreme end of the distribution where individual parameters gain influence, their contributions are all fairly comparable, indicating they should all be monitored.

Clearly one would come to different conclusions about how to design a monitoring program for User 1 depending on the experimental design used for the vulnerability assessment. Monitoring in this example consists of tracking changes in parameter estimates of a basin-scale HMM to understand when User 1’s shortage will likely increase. In this case, using the Box Around Historical experiment, User 1 might choose to simply track changes in estimates of the dry state mean to understand when their 90th percentile shortage might increase, suggesting they should adapt by changing what crop they grow. However, this might not be sufficient in triggering adaptation if multiple parameters should be considered, as implied by the Paleo design. Furthermore, while monitoring multiple HMM parameters is not more expensive than monitoring one, in other systems, needing to monitor multiple uncertain factors (e.g., well depths and population) may be. This highlights the importance of considering multiple experimental designs to see if some assumptions imply monitoring expenses could be reduced if correlations suggest redundancy (as in the CMIP experiment), or if some adaptation plans may be underinformed if they underestimate interactions (e.g., the Box Around Historical experiment relative to the Paleo experiment). One can then assess which of these experiments to use for monitoring and design based on their confidence in each of the design's underlying assumptions and their level of risk aversion.

These conclusions are reaffirmed by the differing sensitivities observed for Users 2 and 3 across experiments. For User 2, little variability in shortage is experienced in the Box Around Historical (Figure 6e) and CMIP (Figure 6f) experiments, making it hard to determine which parameters are most influential (Figure 7e and 7f). This may not be a problem if this user is truly not vulnerable to climate change, but the highest shortages experienced by this user under the Paleo experiment (Figure 6g) suggest they would have certainly been vulnerable to conditions observed in the past. The sensitivities at these high percentiles under the Paleo experiment have strong negative interactions suggesting a number of parameters could be monitored and explain the same information. Yet unlike the redundancy in User 1’s CMIP experiment, there are not two predominant factors that are obviously explaining the same information. Rather, the influence of all of the parameters is relatively equal, so one would need to compute higher-order interactions to determine which are redundant (those with negative interactions) before designing a monitoring program. This is different than the conclusions one would draw from the All-Encompassing experiment (Figure 7h), which suggests the dry-state mean explains most of the variability in User 2’s extreme shortages.

Interestingly, User 3 experiences similar sensitivities under each experiment to User 1 despite being a median rather than senior rights holder in a different sector (SI Figure S11). However, the conclusion remains that these sensitivities vary across experiments, challenging the design of monitoring programs for both users.
5.3. Response Surface Modeling

While Figure 7 shows that the experimental design can clearly influence which factors one concludes are most important for monitoring a user’s shortage, it does not explain why that is. Building response surfaces of shortage as a function of the uncertain parameters can help explain those differences. By mapping shortage as a function of the HMM parameters, we can see how they interact and how the detection of those interactions differs depending on where the sample points lie.

Figure 8 shows the response surfaces for User 1’s 50th and 90th percentiles of shortage, while response surfaces for Users 2 and 3 are shown in SI Figure S15 and S17. Black lines in Figure 8a show the ranges of shortage experienced at the 50th and 90th percentiles across the All-Encompassing realizations, while black lines in Figure 8d show the portion of that variability explained by the different uncertain factors. Dots in Figure 8b represent two-dimensional projections of the different sample points from the All-Encompassing design, colored by the average 50th percentile shortage experienced across the 10 realizations of that parameterization. The x axis is the value of the HMM parameter with the greatest first-order Sobol index at that percentile, and the y axis is the value of the HMM parameter with the second greatest first-order Sobol index. The contours represent the predicted 50th percentile shortage from a second order linear regression model estimating shortage as a function of the two most important HMM parameters and their interaction. Figure 8e shows the same contours but with the two-dimensional projections of the CMIP sample points in yellow and the Paleo sample points in green instead of the All-Encompassing sample points. Figures 8c and 8f show the same thing as Figures 8b and 8e but for the 90th percentile of shortage instead of the 50th.

At the 50th percentile, the wet- and dry-state means are most predictive of shortage (Figure 8d). The relationship of these three variables is shown in Figure 8b. The upper left corner is left blank as the dry-state mean exceeds the wet-state mean in this region. From this plot, one can see that shortage increases as both the wet and dry-state mean decrease. The shortage contours are linear, meaning the two parameters do not interact. From Figure 8e, one can see that the CMIP sample points are perpendicular to the shortage contours. Consequently, as either the wet-state mean or dry-state mean decreases in the CMIP sample points, they
experience the same increase in shortage. Hence, either one or the other could be monitored to detect shortage changes, explaining their negative interactions in Figure 7b. The Paleo points in green are not perpendicular to these contours. Rather, as the wet state mean decreases (x axis), shortages increase much more dramatically than as the dry state mean increases (y axis), making the wet state mean more influential at this percentile under the Paleo experiment.

At the 90th percentile, a different set of factors are most important in explaining shortage: the dry state mean and wet-wet transition probability (Figure 8d). Shortage increases across the All-Encompassing experiment as these two parameters decrease (Figure 8c). This relationship has some curvature, indicating positive interactions between the two variables. Looking at Figure 8f, one can see that the Paleo points in green nearly following these nonlinear shortage contours. This makes it hard to detect changes in shortage from only one of the parameters alone, explaining why there are strong positive interactions in explaining User 1’s shortage under the Paleo experiment (Figure 7c). The CMIP points, however, experience very little variability in the wet-wet transition probability across their experimental design. Consequently, they miss this interaction and only detect changes in 90th percentile shortage as a consequence of changes in the dry-state mean. They also attribute these to changes in the wet-state mean because of their near-perfect correlation in the CMIP experiment. This could be a misattribution if that correlation structure is not representative of the distribution of plausible futures.

While Figure 8 is useful for understanding how a UCRB user’s shortage will change at different percentiles as a function of the basin’s hydrologic parameters, water users are often more concerned with detecting changes in the probability of observing a critical, intolerable value of shortage rather than its whole range, that is, a “failure.” In addition to exploring how the most important factors in explaining variability depend on the experimental design, another important question is, are the most important factors in explaining the probability of failure the same as the most important factors in explaining variability? We investigate this question by building logistic regression models that map the probability that shortage at the 50th and 90th percentiles stays below different thresholds of acceptability as a function of the HMM parameters. These response surfaces are shown in Figure 9 for User 1 and in SI Figures S16 and S18 for Users 2 and 3, respectively.

Figure 9. User 1’s probability of success response surfaces under different definitions of success. (a) User 1’s shortage distribution under the All-Encompassing experiment with a sample definition of success. (b–d) Predicted probability (contours) of different shortage magnitudes occurring less than 50% of the time, with the historical magnitude printed in gold. The x axis indicates the most important factor, and the y axis the second most important. All-Encompassing sample points are shaded blue if at least 5/10 realizations are successes and red otherwise. (e) User 1’s satisficing surface under the All-Encompassing experiment. The color of each box represents the percent of realizations in which shortages of varying magnitudes on the y axis are experienced not more than x percentile of the time. Boxes highlighted in yellow indicate the historical shortage magnitude experienced at each frequency. (f–h) Same as (b–d) but with different shortage magnitudes occurring less than 10% of the time.
Figure 9a illustrates User 1’s shortage distribution across the All-Encompassing realizations with an example of how success and failure could be defined for the 50th percentile of shortage. In this example, a success is defined as the 50th percentile shortage being at or below the historical level shown in gold. This shortage level is displayed on the y axis as a fraction of the user’s demand. Figure 9e displays a satisficing surface introduced by Hadjimichael, Quinn, Wilson, et al. (2020) illustrating the percent of realizations in the All-Encompassing experiment meeting a range of such possible success definitions (see SI Figures S19 and S20 for how this surface differs depending on the experimental design and number of Latin hypercube samples). In this figure, the color in each box represents the percent of realizations in which the percent of demand that is short (x axis) is experienced different percentages of time in the simulation (y axis). The gold boxes in Figure 9e represent the historical frequency at which different shortage magnitudes are experienced. For example, Box (c) corresponds to the example success threshold in Figure 9a in which shortages of 50% of demand are experienced no more than 50% of the time. At the 90th percentile, shortages of 70% of demand are experienced 10% of the time historically (Box (g)).

We explore which HMM parameters are most predictive of these two possible success definitions, as well as more and less conservative definitions at the same percentile. In addition to being successful at the 50th percentile if shortages of no more than 50% of demand are experienced (Figure 9c), we also consider a stricter definition of no more than 30% of demand if User 1 finds even historical conditions unacceptable (Figure 9b), and a more lenient definition of no more than 70% of demand if User 1 is willing to tolerate an increase in shortage (Figure 9d). Similarly, at the 90th percentile, we consider a smaller ratio of no more than 50% of demand being experienced 10% of the time (Figure 9f), a historical ratio of no more than 70% of demand (Figure 9g), and a larger ratio of no more than 90% of demand (Figure 9h).

The response surfaces for the 50th percentile definitions are shown in the top row of Figures 9b–9d, while the response surfaces for the 90th percentile definitions are shown in the bottom row (Figures 9f–9h). In each response surface, the most predictive HMM parameter is shown on the x axis and the second most predictive parameter on the y axis. The dots in these figures represent two-dimensional projections of the All-Encompassing sample points and are shaded light blue if 5 or more of the 10 realizations of that parameterization met the success definition and red otherwise. The contours display the predicted probability of success from a logistic regression model with the two most predictive parameters and their interaction. For this user, the most predictive parameters when using the historical shortage magnitude for the failure definition (printed in gold above Figures 9c and 9g) is the same as the most predictive parameters of variability at that percentile. However, interactions between $\mu$ and $\rho_{w,w}$ at the 90th percentile are weaker when predicting success of staying below the historical shortage level than when predicting shortage itself. In fact, interactions for this user are minimal across all failure definitions.

While interactions are negligible across all failure definitions, the most predictive parameters are not the same across them. This suggests that designing monitoring programs is not only complicated by the fact that sensitivities depend on the experimental design but also by the fact that sensitivities depend on whether one is monitoring for any changes in shortage, or for changes around a particular level. These sensitivity differences are not simply because we are only displaying the top two at each threshold and there is always a consistent close third. Table 3 reports the McFadden’s pseudo-$R^2$ values for each HMM parameter under all possible failure definitions in 10% increments for the 50th percentile definitions, while Table 4 reports the same for the 90th percentile definitions. As you can see, the strength of each parameter’s predictability varies strongly across the possible failure definitions. Note columns of these tables with no reported McFadden’s pseudo-$R^2$ values either experienced no failures or no successes under that success definition.

### 5.4. Robustness Analysis

The variance decomposition and response surface modeling clearly reveal that using sensitivity analysis to design monitoring programs to detect changes in water users’ vulnerabilities is challenging in contexts of deep uncertainty. The other common use of sensitivity analysis for decision making under deep uncertainty is to evaluate alternative management plans to see which are most robust to potential futures. It would be concerning if different experimental designs led to different conclusions about policy robustness, as it would influence which policy is chosen to be implemented. While we do not investigate alternative policies in this study, we can instead see how the ranking of different water users’ robustness changes depending on the experimental design.
Table 3
McFadden's Pseudo-$R^2$ Values in Predicting User 1's Success in 50th Percentile Shortage Staying Below Different Magnitudes Under the All-Encompassing Experiment

| Parameter | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|-----------|----|----|----|----|----|----|----|----|----|-----|
| $\mu_d$   | 0.162 | 0.019 | 0.025 | 0.059 | 0.103 | 0.149 | 0.188 | 0.309 | 0.402 | — |
| $\sigma_d$ | 0.001 | 0.002 | 0.002 | 0.001 | 0.000 | 0.002 | 0.005 | 0.005 | 0.000 | — |
| $\mu_w$   | 0.343 | 0.361 | 0.400 | 0.352 | 0.232 | 0.062 | 0.014 | 0.008 | 0.000 | — |
| $\sigma_w$ | 0.000 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 | — |
| $p_{d,d}$  | 0.006 | 0.013 | 0.020 | 0.025 | 0.043 | 0.141 | 0.225 | 0.234 | 0.253 | — |
| $p_{w,w}$  | 0.117 | 0.100 | 0.054 | 0.041 | 0.056 | 0.071 | 0.061 | 0.037 | 0.049 | — |

Note. Values of each HMM parameter when fitting a univariate logistic regression are reported for models predicting the probability that User 1's 50th percentile shortage is below different shortage levels, reported as a fraction of demand. The McFadden's pseudo-$R^2$ values of the two most predictive parameters at each shortage threshold are shown in bold.

Figure 10 illustrates the variability in this ranking across experiments under different satisficing definitions of robustness. These satisficing definitions again correspond to combinations of shortage magnitudes and frequencies. Figure 10f indicates which satisficing definitions are used here with black boxes around User 1's satisficing surface for the All-Encompassing design. The three black boxes in the top row correspond to increasing one's tolerance of 90th percentile shortages from 10% of demand, to 50% of demand to 100% of demand. The criteria thus become easier to meet as one moves from left to right. The three black boxes in the first column correspond to increasing one's tolerance for the frequency at which shortages of 10% of demand are experienced from 10% of the time to 50% of the time to 100% of the time. The criteria thus become easier to meet as one moves from top to bottom.

Figures 10a–10c show the robustness ranks of 342 water users in StateMod under the satisficing definitions of increasing magnitudes of 90th percentile shortage. Figures 10a, 10d, and 10e show the robustness ranks under satisficing definitions of increasing frequencies for shortages of 10% of demand. The x axis in each panel is the rank of each user in the All-Encompassing experiment, where 1 corresponds to the most robust user. The y axis in each panel corresponds to the same user's rank in each of the other experiments, with the Box Around Historical ranks shown in salmon, the CMIP in yellow, the Paleo in green, and the All-Encompassing in lavender. If all of the experiments were consistent in their robustness ranks, all points would lie on top of each other on a one-to-one line. This clearly does not always happen. Horizontal lines of points with a constant rank on the y axis represent ties in user robustness. Ties are less common in the All-Encompassing experiment, indicating its wider sampling is able to better distinguish users in terms of robustness. Deviations along the vertical dimension indicate a given user's rank varies significantly across experiments. The variability in ranks along the y axis decreases as the shortage magnitude or frequency in the satisficing definition increases. This suggests that, at least for this problem, robustness rankings are more consistent for easier-to-meet criteria. Conversely, the more conservative one wants to be in finding robust

Table 4
McFadden's Pseudo-$R^2$ Values in Predicting User 1's Success in 90th Percentile Shortage Staying Below Different Magnitudes Under the All-Encompassing Experiment

| Parameter | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|-----------|----|----|----|----|----|----|----|----|----|-----|
| $\mu_d$   | — | — | 0.276 | 0.069 | 0.064 | 0.248 | 0.327 | 0.478 | 0.546 | 0.261 |
| $\sigma_d$ | — | — | 0.143 | 0.028 | 0.002 | 0.007 | 0.008 | 0.006 | 0.008 | 0.018 |
| $\mu_w$   | — | — | 0.303 | 0.217 | 0.203 | 0.139 | 0.050 | 0.010 | 0.006 | 0.002 |
| $\sigma_w$ | — | — | 0.065 | 0.039 | 0.021 | 0.003 | 0.001 | 0.001 | 0.000 | 0.006 |
| $p_{d,d}$  | — | — | 0.009 | 0.000 | 0.014 | 0.003 | 0.005 | 0.010 | 0.013 | 0.148 |
| $p_{w,w}$  | — | — | 0.003 | 0.260 | 0.217 | 0.099 | 0.081 | 0.039 | 0.027 | 0.049 |

Note. Values of each HMM parameter when fitting a univariate logistic regression are reported for models predicting the probability that User 1's 90th percentile shortage is below different shortage levels, reported as a fraction of demand. The McFadden's pseudo-$R^2$ values of the two most predictive parameters at each shortage threshold are shown in bold.
policies, the harder it is to choose this consistently across experimental designs. The same conclusions hold when only using 100 samples for the Box Around Historical and All-Encompassing experiments, indicating these differences are not due to the different sample sizes of the experiments but their different designs (see SI Figure S21).

6. Conclusions and Future Work

The results of this work illustrate the challenges of designing scenarios for vulnerability assessments under deep uncertainty. The diverging parameter ranges and correlation structures across past and projected climate conditions indicate that the appropriate experimental design for such analyses is itself deeply uncertain. This would not be concerning if the alternative potential designs led to similar conclusions about policies’ or users’ sensitivities and robustness ranks. However, the results presented here show that alternative ranges and correlation structures have decision-relevant implications about the needed monitoring complexity to detect change or success, with the potential to promote insufficient, underdesigned monitoring programs or costly, overdesigned programs. For this reason, we recommend that vulnerability assessments under deep uncertainty be performed using competing hypotheses of how the future might evolve. That is, to be “robustly robust,” analysts should perform robustness analyses over alternative possible experiment designs as done here to find policies/users that perform consistently well across competing designs.

Since the goal of robustness analyses is to find policies that are less sensitive to design assumptions about critical uncertainties, one should extend this philosophy not only to the range of such critical uncertainties but also to their correlation structure. Investigating these rival framings of experimental designs could reveal the potential consequences of each, enabling water managers to guard against the potential misidentification of important factors or their interactions in designing a monitoring program to detect changing or failure conditions. The design of such alternative experiments should be informed by system stakeholders and experts who may have different views on what uncertainties are critical to consider, what their likely ranges are, and how they are related.

While this study only considered hydrologic uncertainty for illustrative purposes, analysts should also ask stakeholders and experts to consider uncertainties in climate-human feedbacks to inform how they design
scenarios for vulnerability assessments. Many past bottom-up vulnerability assessments have found changes in human demands to equal or exceed climatic influences on vulnerability (Herman et al., 2014; Hadjimichael, Quinn, Wilson, et al., 2020). Yet such assessments rarely consider the correlation between human decisions and the climate, when correlations were shown in this study to significantly influence which factors and interactions were most important. Furthermore, past research on sociohydrologic systems has shown that there can be strong positive and negative correlations between human decisions and water availability but that these relationships are location dependent. For example, Worland et al. (2018) found that in the water-abundant Northeastern U.S., water use is more sensitive to social variables like persons per household and the Cook Partisan Voting Index, while in the drier Southwest, water use is more sensitive to environmental variables like precipitation and temperature. While humans generally decrease water consumption in drier areas (Worland et al., 2018), they also build reservoirs to mitigate the severity of droughts. However, this can actually exacerbate the problem if humans then increase water consumption due to the perceived increased availability from the reservoir (Di Baldassarre et al., 2018). Consequently, water system analysts should also consider how alternative policies themselves interact with model uncertainties before drawing conclusions about their robsoness. Incorporating these possible climate-human feedbacks into alternative correlation structures for climate vulnerability assessments under deep uncertainty will be important for capturing interactions between the two systems that may influence perceived user/policy sensitivity and robustness. More broadly, as system complexity grows, it is increasingly important to design vulnerability assessments under deep uncertainty that test the sensitivity of the assessment to its underlying assumptions. As eloquently stated in Saltelli and Funtowicz (2015), we need to “Find sensitive assumptions before these find [us].”

Data Availability Statement

All the scripts to replicate the analysis performed in this paper and to generate the figures can be found online (at https://github.com/julianneq/UCRB_analysis). StateMod is available at https://github.com/OpenCDSS and its input files for the UCRB online (at https://www.colorado.gov/pacifc/cdss/surface-water-statemod).

Acknowledgments

This study was partially supported by the National Oceanic and Atmospheric Administration (NOAA) through the Sectoral Applications Research Program (SARP). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding entities. We would like to thank Page Weil for sharing VIC simulations of the Upper Colorado and Erin Wilson for guidance on running StateMod. We also appreciate the helpful feedback from two anonymous reviewers.

References

Ault, T. R., Cole, J. E., Overpeck, J. T., Pederson, G. T., & Meko, D. M. (2014). Assessing the risk of persistent drought using climate model simulations and paleoclimatic data. Journal of Climate, 27(20), 7529–7549.
Ault, T. R., Cole, J. E., Overpeck, J. T., Pederson, G. T., St. George, S., Otto-Bliesner, B., & Deser, C. (2013). The continuum of hydroclimate variability in western North America during the last millennium. Journal of Climate, 26(16), 5863–5878.
Bankes, S. (1993). Exploratory modeling for policy analysis. Operations research, 41(3), 435–449.
Bartholomew, E., & Kwakkel, J. H. (2020). On considering robustness in the search phase of Robust Decision Making: A comparison of Many-Objective Robust Decision Making, multi-scenario Many-Objective Robust Decision Making, and Many Objective Robust Optimization. Environmental Modelling & Software, 127, 104,699.
Ben-Haim, Y. (2006). Info-gap decision theory: Decisions under severe uncertainty. Amsterdam, The Netherlands: Elsevier.
Bracken, C., Rajagopalan, B., & Zagoura, E. (2014). A hidden Markov model combined with climate indices for multidecadal streamflow simulation. Water Resources Research, 50, 7836–7846. https://doi.org/10.1002/2014WR015567
Brown, C., Boltz, F., Freeman, S., Tront, J., & Rodriguez, D. (2020). Resilience by design: A deep uncertainty approach for water systems in a changing world. Water Security, 9, 100,051.
Brown, C., Chihle, Y., Laverty, M., & Li, K. (2012). Decision scaling: Linking bottom-up vulnerability analysis with climate projections in the water sector. Water Resources Research, 48, W09537. https://doi.org/10.1029/2011WR011212
Brown, C., & Wilby, R. L. (2012). An alternate approach to assessing climate risks. Eos. Transactions American Geophysical Union, 93(41), 401–402.
Bryant, B. P., & Lempert, R. J. (2010). Thinking inside the box: A participatory, computer-assisted approach to scenario discovery. Technological Forecasting and Social Change, 77, 34–49.
CWCB (2012). Colorado River Water Availability Study Phase I Report. Colorado Water Conservation Board.
CWCB, & CDWR (2016). Upper Colorado River Basin Water Resources Planning Model User-Manual Model. Retrieved 2019-10-02, from https://www.colorado.gov/pacifc/cdss/modeling-dataset-documentation
Christensen, N., & Lettenmaier, D. (2006). A multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the Colorado River Basin. Hydrology and Earth System Sciences Discussions, 3(6), 3727–3770.
Christensen, N., Wood, A., Voisin, N., Lettenmaier, D., & Palmer, R. (2004). The effects of climate change on the hydrology and water resources of the Colorado River basin. Climatic change, 62(1–3), 337–363.
Culley, S., Noble, S., Yates, A., Timbs, M., Westra, S., Maier, H., & Castelletti, A. (2016). A bottom-up approach to identifying the maximum operational adaptive capacity of water resource systems to a changing climate. Water Resources Research, 52, 6751–6768. https://doi.org/10.1002/2015WR018253
Di Baldassarre, G., Wanders, N., AghaKouchak, A., Kuil, L., Rancecroft, S., Veldkamp, T. I., & Van Loon, A. F. (2018). Water shortages worsened by reservoir effects. Nature Sustainability, 1(11), 617–622.
Dittman, R., Wicen, A., & Moran, D. (2016). A survey of decision-making approaches for climate change adaptation: Are robust methods the way forward? Ecological Economics, 122, 79–89.

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., & Liebert, J. (2012). HISS Opinions Should we apply bias correction to global and regional climate model data?. Hydrology & Earth System Sciences Discussions, 9, 4.

Farmer, W. H., & Vogel, R. M. (2016). On the deterministic and stochastic use of hydrologic models. Water Resources Research, 52, 5619–5633. https://doi.org/10.1002/2016WR019129

Fletcher, S. M., Lickley, M., & Strzepek, K. (2019). Learning about climate change uncertainty enables flexible water infrastructure planning. Nature communications, 10(1), 1–11.

Fletcher, S. M., Strzepek, K., Alsdorf, A., & de Weeck, O. (2019). Learning and flexibility for water supply infrastructure planning under groundwater resource uncertainty. Environmental Research Letters, 14(11), 114022.

Freeman, S. S. G., Brown, C., Cañada, H., Martinez, V., Nava, A. P., Ray, P., et al. (2020). Resilience by design in Mexico City: A participatory human-hydrologic systems approach. Water Security, 9, 100,053.

Giuliani, M., & Castelletti, A. (2016). Is robustness really robust? How different definitions of robustness impact decision-making under climate change. Climatic Change, 135(3–4), 409–424.

Groves, D. G., Bloom, E., Lempert, R. J., Fischbach, J. R., Nevills, J., & Goshi, B. (2015). Developing key indicators for adaptive water planning. Journal of Water Resources Planning and Management, 141(7), 501408.

Haasnoot, M., Kwakkel, J. H., Walker, W. E., & ter Maat, J. (2013). Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. Global Environmental Change, 23(2), 485–498. https://doi.org/10.1016/j.gloenvcha.2012.12.006

Herman, J. D., & Giuliani, M. (2018). Policy tree optimization for threshold-based water resources management over multiple timescales. Environmental modelling & software, 99, 39–51.

Herman, J. D., Quinn, J. D., Steinschneider, S., Giuliani, M., & Fletcher, S. M. (2020). Climate adaptation as a control problem: Review and perspectives on dynamic water resources planning under uncertainty. Water Resources Research, 56, e24389. https://doi.org/10.1029/2019WR025502

Herzog, K. J., Quinn, J., & Reed, P. (2020). Developing a system distributed model to evaluate better understand water shortage mechanisms in institutionally complex river basins. Earth and Space Science Open Archive. https://doi.org/10.5194/essoar.10503255.1

Hadjimichael, A., Quinn, J., & Reed, P. (2020). Developing adaptive model evaluation to better understand water shortage mechanisms in institutionally complex river basins. Earth and Space Science Open Archive. https://doi.org/10.5194/essoar.10503255.1

Hadjimichael, A., Quinn, J., Wilson, E., Reed, P., Basdekas, L., Yates, A., & Garrison, M. (2020). Defining robustness, vulnerabilities, and consequential scenarios for diverse stakeholder interests in institutionally complex river basins. Earth’s Future, 8, e2020EF001503. https://doi.org/10.1029/2020EF001503

Hadjimichael, A., Reed, P., & Quinn, J. (2020). Navigating deeply uncertain tradeoffs in harvested predator-prey systems. Complexity, 2020, 1–8.

Hading, B., Wood, A., & Prairie, J. (2012). The implications of climate change scenario selection for future streamflow projection in the Upper Colorado River Basin. Hydrology and Earth System Sciences, 16(11), 3989–4007.

Herger, J., & Giuliani, M. (2018). Policy tree optimization for threshold-based water resources management over multiple timescales. Environmental modelling & software, 99, 39–51.

Herman, J. D., Quinn, J. D., Steinschneider, S., Giuliani, M., & Fletcher, S. M. (2020). Climate adaptation as a control problem: Review and perspectives on dynamic water resources planning under uncertainty. Water Resources Research, 56, e24389. https://doi.org/10.1029/2019WR025502

Herman, J. D., Reed, P. M., Zeff, H. B., & Characklis, G. W. (2015). How should robustness be defined for water systems planning under change? Journal of Water Resources Planning and Management, 141, 4015012.

Herman, J. D., & Usher, W. (2017). SAaLib: An open-source Python library for sensitivity analysis. Journal of Open Source Software, 2(9), 27.

Herman, J. D., Zeff, H. B., Reed, P. M., & Characklis, G. W. (2014). Beyond optimality: Multistakeholder robustness tradeoffs for regional water portfolio planning under deep uncertainty. Water Resources Research, 50, 7692–7713. https://doi.org/10.1002/2014WR015338

Hermans, L. M., Haasnoot, M., ter Maat, J., & Kwakkel, J. H. (2017). Designing monitoring arrangements for collaborative learning about adaptation pathways. Environmental Science & Policy, 69, 29–38.

Hine, D., & Hall, J. W. (2010). Information gap analysis of flood model uncertainties and regional frequency analysis. Water Resources Research, 46, W01514. https://doi.org/10.1029/2009WR007620

Kasprzyk, J. R., Nataraj, S., Reed, P. M., & Lempert, R. J. (2013). Many objective Robust Decision Making for complex environmental systems undergoing change. Environmental Modelling & Software, 42, 55–71.

Kinnlick, J., Steinschneider, S., & Walter, M. T. (2017). A vulnerability-based, bottom-up assessment of future riverine flood risk using a modified peaks-over-threshold approach and a physically based hydrologic model. Water Resources Research, 53, 10,043–10,064. https://doi.org/10.1002/2017WR021036

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. Geophysical Research Letters, 40, 1194–1199. https://doi.org/10.1002/2013GL052506

Korteling, B., Dessler, S., & Kapelan, Z. (2013). Using information-gap decision theory for water resources planning under severe uncertainty. Water Resources Management, 27, 1149–1172.

Kwakkel, J. H., Haasnoot, M., & Walker, W. E. (2015). Developing dynamic adaptive policy pathways: A computer-assisted approach for developing adaptive strategies for a deeply uncertain world. Climatic Change, 132(3), 373–386.

Kwakkel, J. H., Haasnoot, M., & Walker, W. E. (2016). Comparing robust decision-making and dynamic adaptive policy pathways for model-based decision support under deep uncertainty. Environmental Modelling & Software, 86, 168–183.

Laamanet, J. R., Reed, P. M., Link, R., Calvin, K. V., Clarke, L. E., & Edmonds, J. A. (2018). Large ensemble analytic framework for consequence-driven discovery of climate change scenarios. Earth’s Future, 6, 488–504. https://doi.org/10.1002/2017EF001701

Laamanet, J. R., Reed, P., Marzangoni, G., Keller, K., & Garner, G. (2019). Robust assessment pathways to tolerable climate futures require immediate global action. Nature Climate Change, 9(4), 290–294.

Lebedev, S. (2015). hmlmlearn. Retrieved from https://hmlmlearn.readthedocs.io/

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., & Hawkins, E. (2020). Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. Earth System Dynamics, 11(2), 491–508.

Lempert, R. J., & Collins, M. (2007). Managing the risk of an uncertain threshold response: Comparison of robust, minimunum, and precautionary approaches. Risk Analysis, 27(4), 1099–1026.

Lempert, R. J., Popper, S. W., & Bankes, S. C. (2010). Robust Decision Making: Coping with uncertainty. The Futurist, 44(1), 47.

Maier, H. R., Guillaume, J. H., van Deelen, H., Riddell, G. A., Haasnoot, M., & Kwakkel, J. H. (2016). An uncertain future, deep uncertainty, scenarios, robustness and adaptation: How do they fit together? Environmental Modelling & Software, 81, 154–164.

Malers, S. A., Ray, R. B., & Catherine, N. L. (2001). Colorado’s decision support systems: Data-centered water resources planning and administration. Watershed Management and Operations Management 2000, 1–9. https://doi.org/10.1016/0449-0000(00)153

Marchau, V. A., Walker, W. E., Bloemen, P. J., & Popper, S. W. (2019). Decision making under deep uncertainty: From theory to practice. Springer Nature.
McPhail, C., Maier, H., Kwakkel, J., Giuliani, M., Castelletti, A., & Westra, S. (2018). Robustness metrics: How are they calculated, when should they be used and why do they give different results? Earth’s Future, 6, 169–191. https://doi.org/10.1002/2017EF000649

Milly, P., & Dunne, K. (2020). Colorado River flow dwindles as warming-driven loss of effective snow energies evaporation. Science, 367, 1252–1255.

Moallemi, E. A., Elsawah, S., & Ryan, M. J. (2020). Robust Decision Making and Epoch–Era analysis: A comparison of two robustness frameworks for decision-making under uncertainty. Technological Forecasting and Social Change, 151, 119-797.

Moallemi, E. A., Zare, F., Reed, P. M., Elsawah, S., Ryan, M. J., & Bryan, B. A. (2020). Structuring and evaluating decision support processes to enhance the robustness of complex human-natural systems. Environmental Modelling & Software, 123, 104551. https://doi.org/10.1016/j.envsoft.2019.104551

Moody, P., & Brown, C. (2013). Robustness indicators for evaluation under climate change: Application to the upper Great Lakes. Water Resources Research, 49, 3576–3588. https://doi.org/10.1002/wrcr.20228

Mortazavi-Naeini, M., Kuczera, G., Kiem, A. S., Cui, I., Henley, B., Berghout, B., & Turner, E. (2015). Robust optimization to secure urban bulk water supply against extreme drought and uncertain climate change. Environmental Modelling & Software, 69, 437–451.

Noacco, V., Sarrazin, F., Pianosi, F., & Wagener, T. (2019). Matlab/R workflows to assess critical choices in Global Sensitivity Analysis using the SAFE toolbox. MethodsX, 6, 2258–2280.

Nowak, K., Prairie, J., Rajagopalan, B., & Lall, U. (2010). A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow. Water Resources Research, 46, W08529. https://doi.org/10.1029/2009WR008530

Paleari, L., & Confalonieri, R. (2016). Sensitivity analysis of a sensitivity analysis: We are likely overlooking the impact of distributional assumptions. Ecological Modelling, 340, 57–63.

Prudhomme, C., Wilby, R. L., Crooks, S., Kay, A. L., & Reynard, N. S. (2010). Scenario-neutral approach to climate change impact studies: Application to flood risk. Journal of Hydrology, 380(3–4), 198–209.

Puy, A., Plano, S. L., & Saltelli, A. (2020). A sensitivity analysis of the PAWN sensitivity index. Environmental Modelling & Software, 127, 104679.

Quinn, J., Reed, P., Giuliani, M., & Castelletti, A. (2017). Rival framings: A framework for discovering how problem formulation uncertainties shape risk management trade-offs in water resources systems. Water Resources Research, 53, 7208–7233. https://doi.org/10.1002/2017WR020524

Quinn, J., Reed, P., Giuliani, M., Castelletti, A., Oyler, J., & Nicholas, R. (2018). Exploring how changing monsoonal dynamics and human pressures challenge multi-reservoir management for flood protection, hydropower production and agricultural water supply. Water Resources Research, 54, 4638–4662. https://doi.org/10.1029/2018WR022743

Quinn, J., Reed, P. M., & Keller, K. (2017). Direct policy search for robust multi-objective management of deeply uncertain socio-ecological tipping points. Environmental modelling & software, 92, 25–141.

Rasmussen, R., Ikeda, K., Liu, C., Gochis, D., Clark, M., Dai, A., et al. (2014). Climate change impacts on the water balance of the Colorado headwaters: High-resolution regional climate model simulations. Journal of Hydrometeorology, 15, 1091–1116.

Raso, L., Kwakkel, J., Timmermans, J., & Panthou, G. (2019). How to evaluate a monitoring system for adaptive policies: Criteria for signposts selection and their model-based evaluation. Climatic change, 153(1–2), 267–283.

Ray, P. A., Bonzanigo, L., Wi, S., Yang, Y. C. E., Karki, P., Garcia, L. E., & Brown, C. M. (2018). Multidimensional stress test for hydropower investments facing climate, geophysical and financial uncertainty. Global Environmental Change, 48, 168–181.

Reis, J., & Shortridge, J. (2019). Impact of uncertainty parameter distribution on Robust Decision Making outcomes for climate change adaptation under deep uncertainty. Risk Analysis.

Saltelli, A., Benini, L., Funstonowicz, S., Giampietro, M., Kaiser, M., Reintert, E., & van der Sluijs, J. P. (2020). The technique is never neutral. How methodological choices condition the generation of narratives for sustainability. Environmental Science & Policy, 106, 87–98.

Saltelli, A., & Funstonowicz, S. (2015). 9 Evidence-based policy at the end of the Cartesian dream. Science, Philosophy and Sustainability: The End of the Cartesian dream, 147.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., & Tarantola, S. (2008). Global sensitivity analysis: The primer. West Sussex, England: John Wiley & Sons.

Seaborne, S., & Perkold, J. (2010). statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference. Austin, TX.

Shin, M. J., Guillaume, J. H., Croke, B. F., & Jakeman, A. J. (2013). Addressing ten questions about conceptual rainfall–runoff models with global sensitivity analyses in R. Journal of Hydrology, 503, 135–152.

Shortridge, J. E., & Guikema, S. D. (2016). Scenario discovery with multiple criteria: An evaluation of the robust decision-making framework for climate change adaptation. Risk Analysis, 36(12), 2298–2312.

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. Mathematical modelling and computational experiments, 1(4), 407–414.

Spence, C. M., & Brown, C. M. (2018). Decision analytic approaches to resolving divergent climate assumptions in water resources planning. Journal of Water Resources Planning and Management, 144(9), 4018054.

Stainforth, D. A., Downing, T. E., Washington, R., Lopez, A., & New, M. (2007). Issues in the interpretation of climate model ensembles to inform decisions. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365, 1263–1277.

Starr, C. (1969). Social benefit versus technological risk. Science, 165, 1232–1238.

State of Colorado (2015). State of Colorado’s water plan. Denver, Colorado.

Steinschneider, S., McCrory, R., Means, L. O., & Brown, C. (2015). The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning. Geophysical Research Letters, 42, 5014–5044. https://doi.org/10.1002/2015GL064529

Taner, M. U., Ray, P., & Brown, C. (2019). Incorporating multidimensional probabilistic information into robustness-based water systems planning. Water Resources Research, 55, 3659–3679. https://doi.org/10.1029/2018WR022909

Trindade, B., Reed, P., & Characklis, G. (2019). Deeply uncertain pathways: Integrated multi-city regional water supply infrastructure investment and portfolio management. Advances in Water Resources, 134, 103442.

Trindade, B., Reed, P., Herman, J., Zeff, H., & Characklis, G. (2017). Reducing regional drought vulnerabilities and multi-city robustness conflicts using many-objective optimization under deep uncertainty. Advances in Water Resources, 104, 195–209.

Whateley, S., & Brown, C. (2016). Assessing the relative effects of emissions, climate means, and variability on large water supply systems. Geophysical Research Letters, 43, 31–329. https://doi.org/10.1002/2016GL070241

Whateley, S., Steinschneider, S., & Brown, C. (2014). A climate change range-based method for estimating robustness for water resources supply. Water Resources Research, 50, 8944–8961. https://doi.org/10.1002/2014WR015956
Wilby, R. L., & Dessai, S. (2010). Robust adaptation to climate change. Weather, 65(7), 180–185.
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), 314–318.
Woodhouse, C. A., Gray, S. T., & Meko, D. M. (2006). Updated streamflow reconstructions for the Upper Colorado River basin. Water Resources Research, 42, W05415. https://doi.org/10.1029/2005WR004455
Woodward, M., Kapelan, Z., & Gouldby, B. (2014). Adaptive flood risk management under climate change uncertainty using real options and optimization. Risk Analysis, 34(1), 75–92.
Worland, S. C., Steinschneider, S., & Hornberger, G. M. (2018). Drivers of variability in public-supply water use across the contiguous United States. Water Resources Research, 54, 1368–1389. https://doi.org/10.1002/2017WR021268
Xiao, M., Udall, B., & Lettenmaier, D. P. (2018). On the causes of declining Colorado River streamflows. Water Resources Research, 54, 6739–6756. https://doi.org/10.1002/2018WR023153
Zeff, H. B., Herman, J. D., Reed, P. M., & Characklis, G. W. (2016). Cooperative drought adaptation: Integrating infrastructure development, conservation, and water transfers into adaptive policy pathways. Water Resources Research, 52, 7327–7346. https://doi.org/10.1002/2016WR018771