Can exploratory modeling of water scarcity vulnerabilities and robustness be scenario neutral?

Julianne Quinn[1], Antonia Hadjimichael[2], Patrick Reed[2], and Scott Steinschneider[2]

[1]University of Virginia
[2]Cornell University

November 24, 2022

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 most robust. Consequently, the assumptions of uniformity and independence could have decision-relevant implications. This study illustrates these implications using a multi-stakeholder planning problem within the Colorado River Basin, where hundreds of rights-holders vie for the river's limited water under the law of prior appropriations. 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 paleo-hydrologic 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.

Can exploratory modeling of water scarcity vulnerabilities and robustness be scenario neutral?

J. D. Quinn[1], A. Hadjimichael[2], P. M. Reed[2] and S. Steinschneider[3]

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

Key Points:

- Alternative experimental designs for climate vulnerability assessments lead to different inferences about important factors to monitor
- These differences have implications for our ability to detect failure conditions and to rank user robustness
- Inconsistencies in robustness ranks across experimental designs increase with more conservative performance criteria

Corresponding author: Julianne Quinn, julianne.quinn@virginia.edu

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 most robust. Consequently, the assumptions of uniformity and independence could have decision-relevant implications. This study illustrates these implications using a multi-stakeholder planning problem within the Colorado River Basin, where hundreds of rights-holders vie for the river's limited water under the law of prior appropriations. 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 paleo-hydrologic 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.

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). Under-build and you risk severe socioeconomic impacts from water scarcity or flooding; over-build and you risk stranding assets for decades. Traditional planning mechanisms hedge against these risks by designing systems to be robust to historical 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 has advocated for "scenario-neutral" analyses in which alternative designs or policies are evaluated across a number of possible future scenarios, with no assessment of their likelihoods, since they are unknown (Prudhomme et al., 2010; Wilby & Dessai, 2010). Approaches applying this philosophy include Robust Decision Making (RDM) (Lempert et al., 2010; Shortridge & Guikema, 2016; Hadjimichael et al., Accepted; Moallemi, Elsawah, & Ryan, 2020) and its Multi-Objective extension, MORDM (Kasprzyk et al., 2013; Herman et al., 2014; Quinn, Reed, & Keller, 2017; Hadjimichael et al., 2020), info-gap decision theory (Ben-Haim, 2006; Hine & Hall, 2010; Korteling et al., 2013), and decision scaling (Brown et al., 2012; Steinschneider et al., 2015; Knighton et al., 2017; Ray et al., 2018; Freeman et al., 2020). The goal of scenario-neutral analyses is to use exploratory modeling (Bankes, 1993) to discover under what conditions, or scenarios, alternative designs no longer meet satisfactory performance (Bryant & Lempert, 2010; Herman et al., 2015; Maier et al., 2016; Dittrich 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 (Whateley et al., 2014; Culley et al., 2016).

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). Optimal control methods have been applied to map changes in these trigger values to adaptive management decisions such as expanding reservoir capacity, building a desalination plant, or raising levee heights (Woodward et al., 2014; Kwakkel et al., 2015; Mortazavi-Naeini et al., 2015; Groves et al., 2015; Kwakkel et al., 2016; Zeff et al., 2016; Trindade et al., 2017; Fletcher, Lickley, & Strzepek, 2019; Fletcher, Strzepek, et al., 2019; Trindade et al., 2019). Optimization has also been used to define at what points in time options should be triggered, as opposed to under what climate or demand conditions (Jeuland & Whittington, 2014; Beh et al., 2014, 2015; Borgomeo et al., 2016; Beh et al., 2017; Fletcher et al., 2017).

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 pre-specified range, i.e., 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; Borgomeo et al., 2014). However, there are known shortcomings of this approach as well: climate projections under-represent 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 (Stainforth et al., 2007; J. R. Lamontagne et al., 2018).

Clearly neither of these approaches is entirely appropriate. Policies may be over-designed if the scenarios are too broad and unrealistic, or under-designed 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 step, 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, multi-actor 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 (Herman et al., 2015; Giuliani & Castelletti, 2016; Spence & Brown, 2018; McPhail et al., 2018; Hadjimichael et al., Accepted), or under alternative assumptions about the range and joint distribution of uncertain factors (i.e., the experimental design) (Moody & Brown, 2013; Taner et al., 2019; Reis & Shortridge, 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 (Shin et al., 2013; Paleari & Confalonieri, 2016; Noacco et al., 2019; Puy et al., 2020) to see if our conclusions change under alternative assumptions about the range and correlation of uncertain hydrologic parameters.

2 Conceptualization of the Problem

Fig. 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 Fig. 1(a-b)) and temperature (y axis of Fig. 1(a-b)) from current conditions (black point). These ranges are chosen to span a set of precipitation and temperature observations from different periods of the paleo-record (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 (Fig. 1(e)). One could integrate the area under this pdf within the blue success region to determine which policy is most robust (Fig. 1(f)), 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 (Fig. 1(i)), while Policy 2 is nearly equally influenced by changes in temperature and precipitation (Fig. 1(j)).

These analyses and conclusions, while reasonable, are strongly dependent on the use of the paleo-data to define the ranges of precipitation and temperature, as well as their joint distribution used to calculate robustness. Another analyst might also consider climate projections from the Coupled Model Intercomparison Project (CMIP) in their analysis. These projections, shown in yellow in Fig. 1(c-d), might span beyond the range of temperature and precipitation explored in Experimental Design 1 (outlined in black in Fig. 1(c-d)). 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 (Fig. 1(g-h)). Not only that, but conclusions about which factors are most important to monitor also change, as precipitation and temperature now equally influence

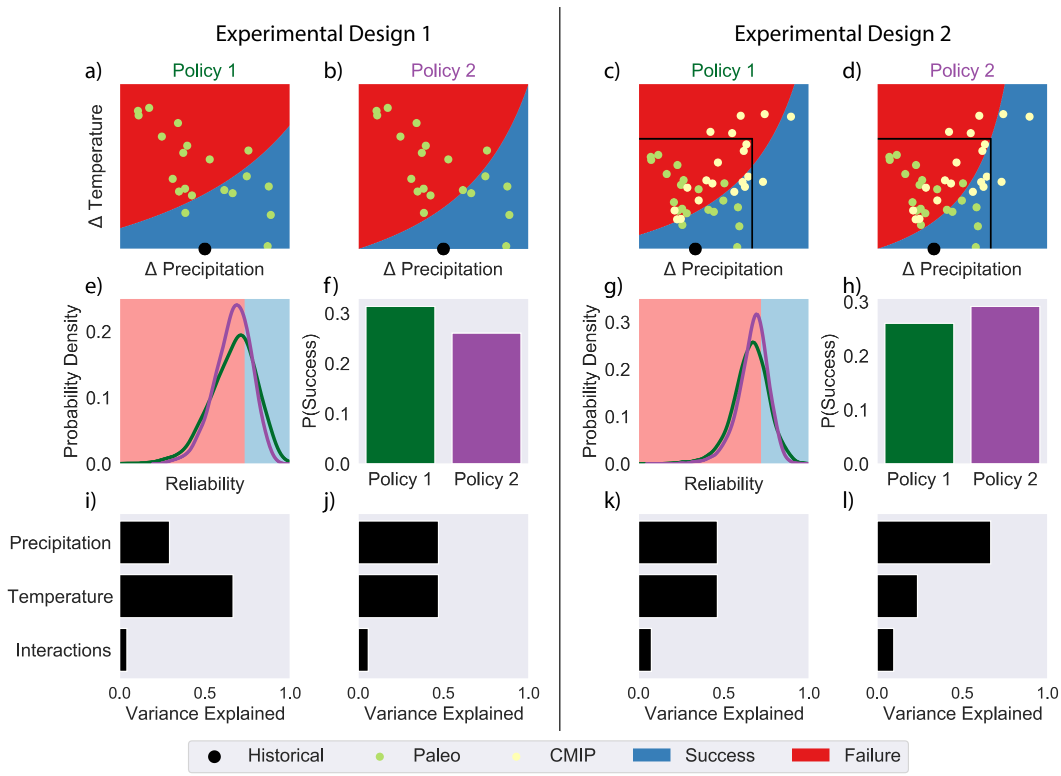

Figure 1. Stylistic example of the influence of the experimental design on sensitivity and robustness analyses. (a)-(d) Regions of success (blue) and failure (red) in meeting an acceptable reliability threshold for two different policies sampled over two different size domains of precipitation and temperature changes, informed by paleo-hydrologic reconstructions (green) and CMIP climate projections (yellow). (e,g) Estimated probability distributions of performance of Policies 1 (green) and 2 (purple) over (e) just paleo-reconstructions and (g) both paleo-reconstructions and CMIP climate projections. Acceptable reliability levels are shaded blue and unacceptable red. (f,h) Corresponding probabilities of success from integrating (e,g) over the blue success region. (i-l) Decomposition of which factors most explain the variability in performance of Policies 1 and 2 under each experimental design.

performance under Policy 1 (Fig. 1(k)), while precipitation now dominates performance under Policy 2 (Fig. 1(l)).

Due to the deep uncertainty in future conditions, it is difficult to determine which 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. 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 (Fig. 2), draining an area of 25,682 km^2 (9,915 mi^2). Within and outside of the UCRB, hundreds of stakeholders own rights to the river's flow, which annually averages 6.9 billion m^3 (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 km^2 (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 m^3 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 Fig. 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 m^3 (1.5 million acre-feet), diverted for power generation and returned downstream (1.3 billion m^3/year, or 1 million acre-ft/year), or left for the environment. The flows at the outlet contribute downstream deliveries to Lake Powell that are required by the Colorado River Compact.

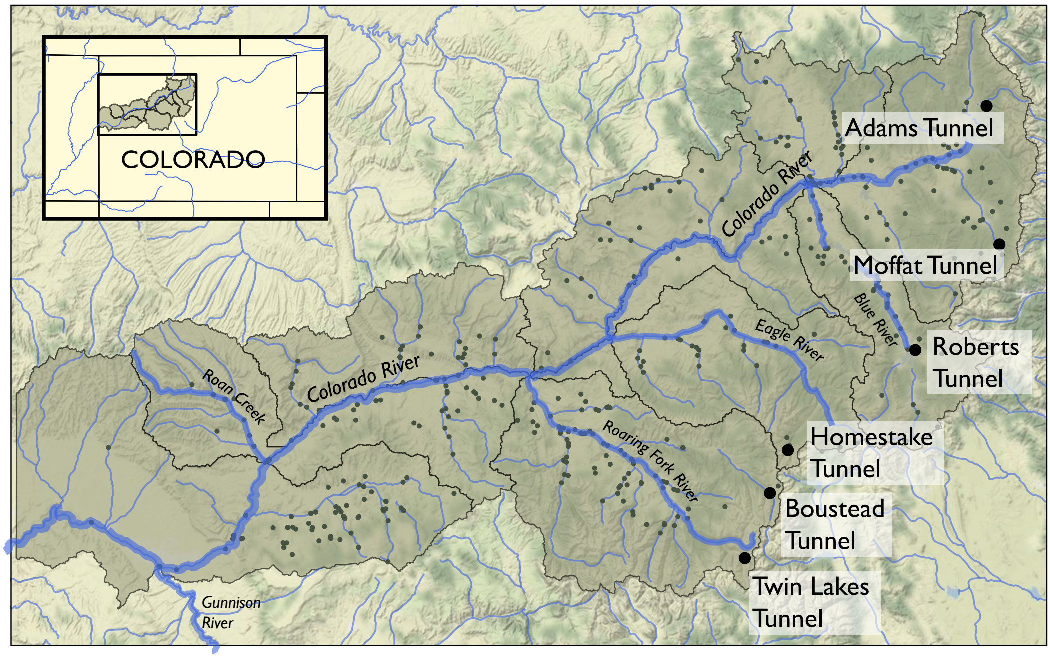

Figure 2. Map of the Upper Basin of the Colorado River within the state of Colorado. Major transbasin diversions are indicated with large black circles while all other diversions (primarily for irrigation), are indicated with small olive circles. Figure adapted from Hadjimichael et al. (Accepted).

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 exceeded the mean only five times. Paleo-hydrologic 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 et al., 2004; Christensen & Lettenmaier, 2006; Rasmussen et al., 2014), a trend that is already being observed (Xiao et al., 2018; Milly & Dunne, 2020). These recent 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 $0.3\text{m}^3/\text{s}$ ($11\ \text{ft}^3/\text{s}$) (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 m^3 (4,000 acre-feet) are modeled explicitly (accounting for 94% of total storage in the system), with the remaining storage aggregated into ten 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 Oscillations (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 (μ_d and σ_d, respectively) and wet state (μ_w and σ_w, respectively) Gaussian distributions, as well as the probabilities of 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.

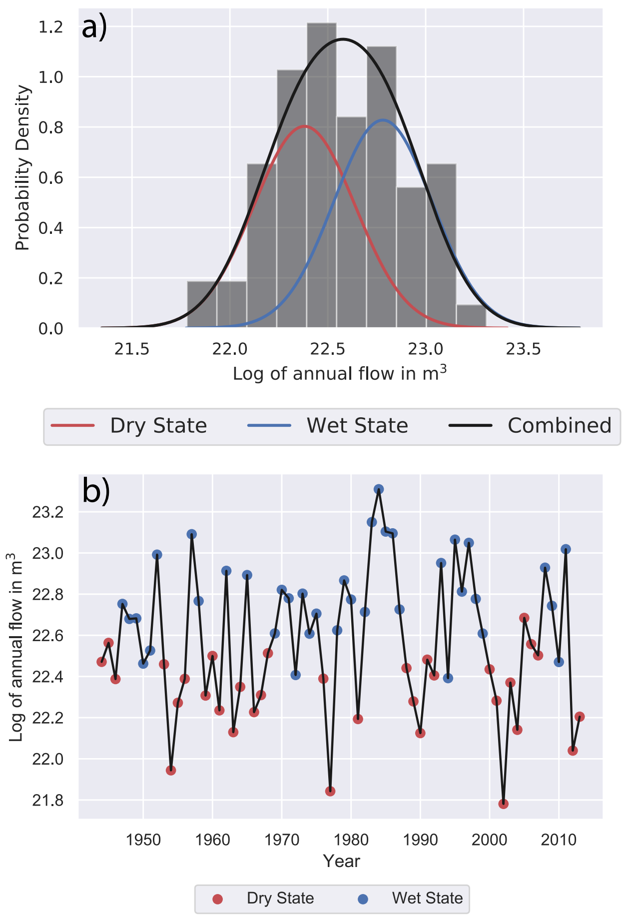

Figure 3. Two-state Gaussian HMM fit to the historical record from 1944-2013. (a) The log-space historical distribution of annual flows in gray, with the fitted dry- and wet-state distributions overlain in red and blue, respectively, and their combined distribution in black. (b) State identification of historical years as dry in red or wet in blue.

Fig. 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. Parameter estimation was performed using Expectation-Maximization with Python's hmmlearn package (Lebedev, 2015). The fitted dry-state and wet-state distributions exhibit non-trivial overlap and together provide a strong fit to the observed distribution (Fig. 3(a)). 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 (Fig. 3(b)), 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 (Figs. S1-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, ten

Table 1. Two-State Gaussian HMM Parameter Estimates. 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.

Parameter	Description	Maximum Likelihood Estimate
μ_d	log-space dry-state mean, m^3 (acre-ft)	22.38 (15.26)
σ_d	log-space dry-state standard deviation	0.26
μ_w	log-space wet-state mean, m^3 (acre-ft)	22.78 (15.66)
σ_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

realizations of 105-year-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 (Fig. 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 (Fig. 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 et al., Accepted), 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

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 pre-specified ranges (Prudhomme et al., 2010; Lempert et al., 2010). These ranges are simply meant to be expansive and enable the discovery of failure boundaries, i.e., 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).

Table 2. Uncertain factors and log-space sampling ranges. 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 below. These samples are not uniform or independent. Multipliers are applied to the log-space parameter values, estimated from the annual flows in acre-ft.

Parameter	**Experiment**	**Current value**	**Lower bound**	**Upper bound**
μ_d multiplier	Box Around Historical	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
σ_d multiplier	Box Around Historical	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
μ_w multiplier	Box Around Historical	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
σ_w multiplier	Box Around Historical	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	Box Around Historical	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	Box Around Historical	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

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 re-labeling, 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 Fig. 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 in the μ_w vs. μ_d panel where the dry-state mean would exceed the wet-state mean. Generating 10 realizations of 105-year 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 Fig. 5. This range extends beyond that of historical annual flows (shown in blue), as well as annual flows from the Coupled Model Intercomparison Project 5 (CMIP) 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 re-labeling 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.

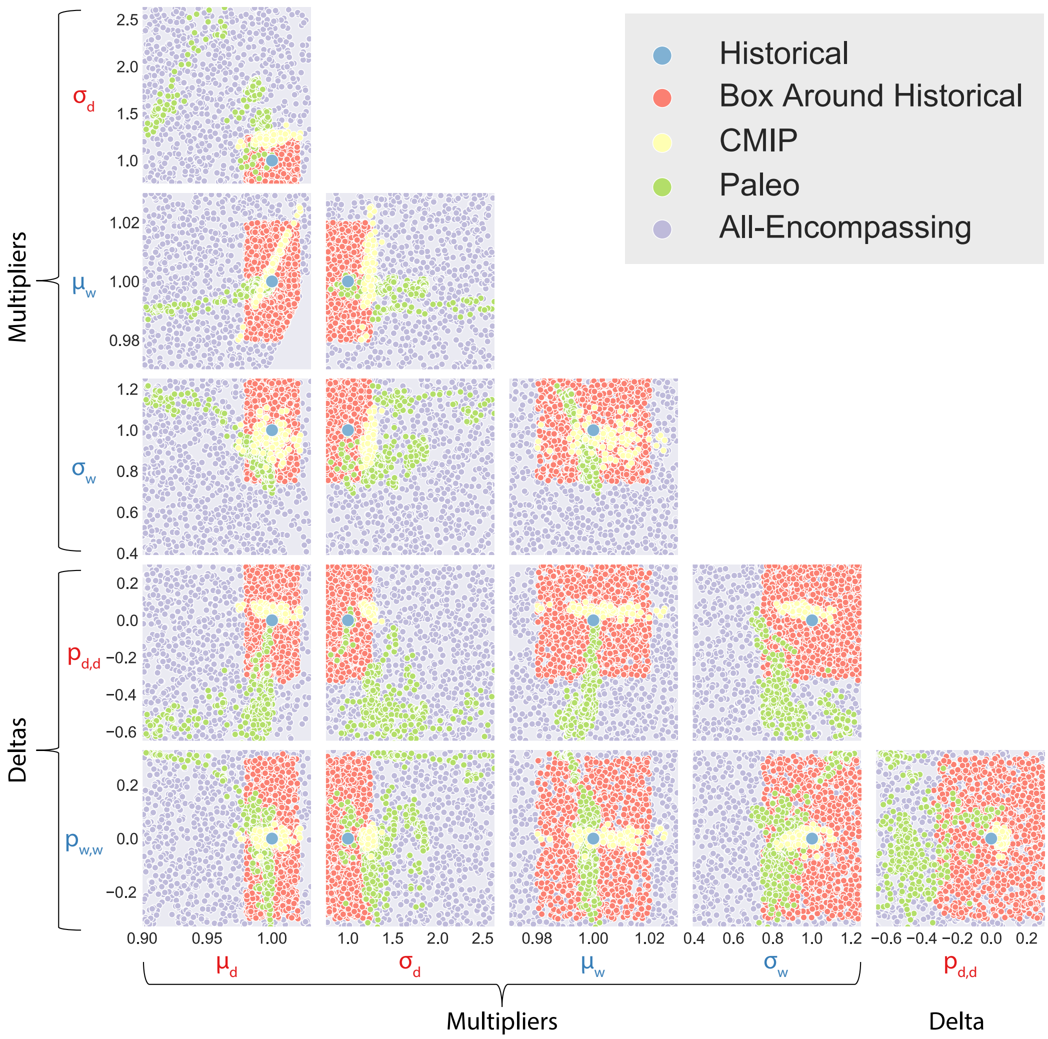

Figure 4. Gaussian HMM parameter samples across four experimental designs. The historical parameter values are signified by a large, blue circle in each panel. Box Around Historical samples are shown in salmon, CMIP samples in yellow, Paleo samples in green and All-Encompassing samples in lavender. Box Around Historical and All-Encompassing experiments assume uniformity and independence of all HMM parameters over different ranges, while CMIP sample points are estimated from climate projections and Paleo sample points from paleohydrologic reconstructions.

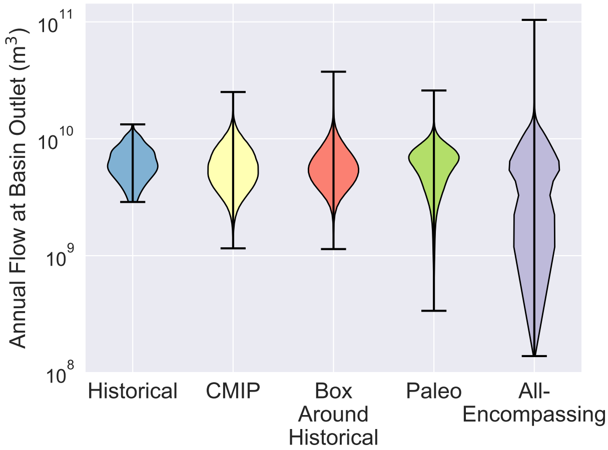

Figure 5. Range of historical and synthetically-generated annual flows from each experimental design. CMIP flows in yellow and Box Around Historical flows in salmon experience similar ranges despite their different parameterizations. Both expand upon the range of historical flows in blue, but do not experience any dry years as severe as the Paleo flows in green. The All-Encompassing experiment in lavender generates the widest range of flows.

4.2.2 Coupled Model Intercomparison Project 5 Experiment (CMIP)

The second experimental design is informed by climate projections from the Coupled Model Intercomparison Project 5 (CMIP5) dataset. Several iterations of CMIP projections have been used in numerous studies assessing potential impacts of climate change in the UCRB (Christensen et al., 2004; Christensen & Lettenmaier, 2006; 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 Fig. 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 Fig. S6). Across the 97 CMIP-forced VIC models, this method results in the 97 parameter combinations shown in yellow in Fig. 4.

For each of these CMIP samples, we generate 10 realizations of 105-year streamflow inputs to StateMod. The range of annual flows generated across these simulations is similar to that of the Box Around Historical Experiment (see Fig. 5), even though it is generated using an entirely different parameter set. The CMIP simulations have a near-perfect correlation between μ_d and μ_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 paleo-reconstructions of streamflow. This is our third experiment.

4.2.3 Paleo-hydrologic Experiment (Paleo)

The Paleo experiment 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 Fig. S7). Like the VIC simulations of the CMIP projections, we build a model of the residuals between reconstructed and observed flows to ensure the paleo-reconstructions do not underestimate variability. Details of the error model and how it is applied to the paleo-reconstructions before the observed record are provided in the SI (Fig. S7).

Again, similar to the CMIP projections, we use this error model to add noise to 64-year moving windows of the paleo-reconstructions from 1569-1997. Shifting this window every year results in 366 64-year windows. A window of 64 years 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 Fig. S8. Similar to the CMIP estimation, 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 paleo-hydrologic 64-year moving windows are shown in green in Fig. 4. The corresponding range of synthetically-generated annual flows from 10 realizations of 105-year time series under each parameterization is shown in Fig. 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 paleo-hydrological studies in the basin (Woodhouse et al., 2006). In addition, the HMM parameters over the paleo-record have a completely different correlation structure than the CMIP projections. While there is still a positive correlation between μ_d and μ_w, the relationship is different. There are also several interesting non-linear relationships, e.g., between μ_d and $p_{w,w}$, and between μ_d and σ_w. There is also a bifurcation in the relationship between μ_d and σ_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 under-represent potential droughts in the basin (Fig. 5), as well as potential Gaussian HMM parameters (Fig. 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). 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 Fig. 4. Their corresponding annual flow distribution extends far beyond all other experiments (Fig. 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 re-labeling 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 6 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 et al., Accepted; 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_i, 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: $S_i = V_i/\text{Var}(Y)$. 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, non-linear 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:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \tag{1}$$

The β 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 (Quinn et al., 2018; J. Lamontagne et al., 2019). 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:

$$\ln\left(\frac{P(Y<T)}{1-P(Y<T)}\right) = \beta_0 + \beta_1 X_i \tag{2}$$

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:

$$R^2_{McFadden} = 1 - \frac{\ln \hat{L}(M_{full})}{\ln \hat{L}(M_{intercept})} \tag{3}$$

where $\ln \hat{L}(M_{full})$ is the log-likelihood of the full model (model with X_0 and X_i as predictors) and $\ln \hat{L}(M_{intercept})$ 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 $R^2_{McFadden}$ 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:

$$\ln\left(\frac{P(Y < T)}{1 - P(Y < T)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \tag{4}$$

$$P(Y < T) = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2\right)} \tag{5}$$

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 (McPhail et al., 2018; Herman et al., 2015; Giuliani & Castelletti, 2016; 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 satisficing criterion (Starr, 1969) denoted as the fraction of realizations in which a particular criterion is met. We choose hypothetical satisficing criteria, e.g., "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 satisficing criteria to see how sensitive the robustness rankings are to the experimental design when using different definitions of success for the satisficing 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 two users' sensitivity differences across the four experiments. 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 (mid-rank) 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.

Fig. 6 displays the distribution of shortages experienced by these two users in each of our four experimental designs, with User 1 on the top row (Fig. 6(a)-(d)) and User 2 on the bottom row (Fig. 6(e)-(h)). In each panel, the black line shows the inverse cumulative distribution function of the respective user's shortage over the historical simulation from 1909-2013. For a given shortage magnitude on the y-axis, the corresponding value on the x-axis represents the percent of years in which shortages were below that magnitude. Each realization of each experiment has a different inverse CDF. In purple, we show the percentage of those realizations whose shortage is below different magnitudes at each percentile, ranging from 90-100% realizations in light purple to only 0-10% in dark purple. The effects of the different experimental designs can therefore be seen by the differences in the ranges of these purple regions across panels.

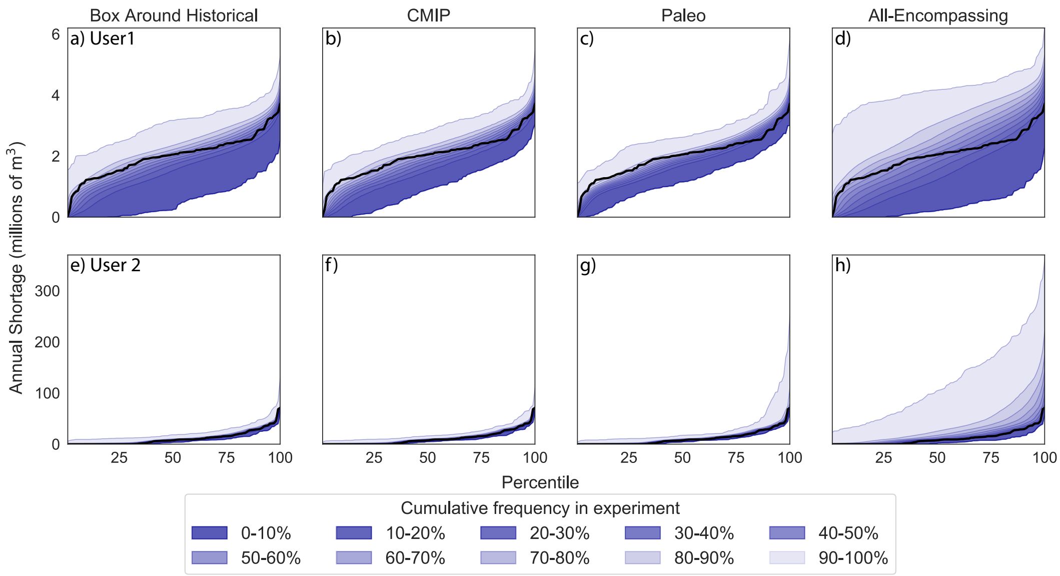

Figure 6. Shortage distributions across experimental designs for two UCRB users. (a)-(d) User 1's shortage distributions across experiments. (e)-(h) User 2's shortage distributions. User 1 is a small, senior irrigation user located in an upstream tributary. User 2 is a large, mid-rank irrigation user located downstream on the Colorado mainstem. Black lines in each panel indicate 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 et al. (Accepted), 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 depending on the experimental design. Looking first at User 1 (Fig. 6(a)-(d)), 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 (Fig. 6(d)) due to its expansive sampling. It is somewhat surprising that the differences in the ranges of shortage experienced in the CMIP experiment (Fig. 6(b)) and the Paleo experiment (Fig. 6(c)) are not greater given their extremely different parameterizations (Fig. 4) and annual flow ranges (Fig. 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 (Fig. 6(e)-(h)). 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 (Fig. 6(e)) and CMIP (Fig. 6(f)) 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 (Fig. 6(g)), this user's most severe shortages increase significantly, while their smaller shortages are not greatly impacted. Under the All-Encompassing experiment (Fig. 6(h)), their shortage magnitude increases at all percentiles.

5.2 Sobol Variance Decomposition

Fig. 7 shows how the sensitivities of these two users differ across experiments based on the Sobol sensitivity analysis for each percentile of shortage. To illustrate the stability of the variance decomposition with smaller sample sizes, this figure is also shown in SI Fig. S9 using only 100 initial samples in the Box Around Historical and All-Encompassing designs instead of 1000. 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).

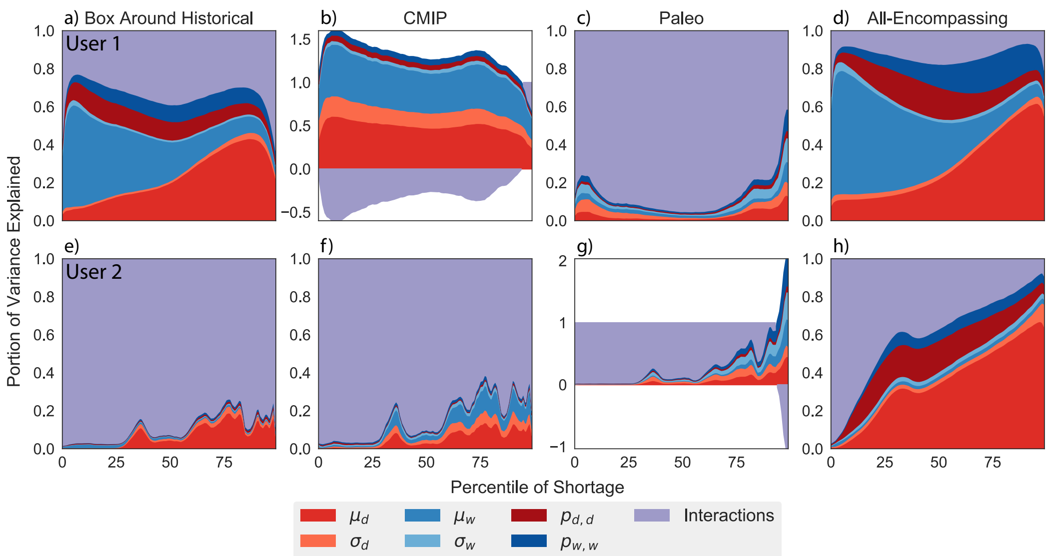

Figure 7. Variance decomposition of two users' shortage distributions across experimental designs. (a)-(d) Variance decomposition of User 1's shortage distribution. (e)-(h) Variance decomposition of User 2's shortage distribution. The y-axis in each panel indicates the portion of the variability in shortage at each percentile that is explained by each of the HMM parameters. The influence of dry-state HMM parameters is shown in shades of red and of wet-state parameters in shades of blue, with interactions shown in lavender. Note these interactions can be negative, which occurs when the first order indices sum to more than 1, indicating they explain redundant information.

For both 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 (Fig. 7(a)). The influences of the dry- and wet-state standard deviations are negligible, but there are significant positive interactions across parameters, indicating non-linearity in the shortage response to the HMM parameters. Moving to higher percentiles of shortage, the influence of the wet-state mean decreases, while that of the dry-state mean increases. They are nearly equally influential at the median, but then their influences swap at higher shortage percentiles, where the dry-state mean becomes more important. 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 (Fig. 7(a)). The strength of the parameters' first order sensitivities just become larger in the All-Encompassing experiment, with the importance of positive interactions decreasing.

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 (Fig. 7(b)), 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 (Fig. 7(c)), 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.

Similar conclusions are drawn for User 2 (Fig. 7(e)-(h)). The lack of variability in shortage experienced by this user in the Box Around Historical (Fig. 6(e)) and CMIP (Fig. 6(f)) experiments makes it hard to determine which parameters are controlling the variability (Fig. 7(e)-(f)). 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 (Fig. 6(g)) 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 (Fig. 7(h)), which suggests the dry-state mean explains most of the variability in User 2's extreme shortages.

5.3 Response Surface Modeling

While Fig. 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.

Fig. 8 shows the response surfaces for User 1's 50th and 90th percentiles of shortage, while response surfaces for User 2's 50th and 90th percentiles of shortage are shown in SI Fig. S10. Black lines in Fig. 8(a) show the ranges of shortage experienced at the 50th and 90th percentiles across the All-Encompassing realizations, while black lines in 8(d) show the portion of that variability explained by the different uncertain factors. Dots in Fig. 8(b) 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. Fig. 8(e) 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. Fig. 8(c) and (f) show the same thing as Fig. 8(b) and (e), but for the 90th percentile of shortage instead of the 50th.

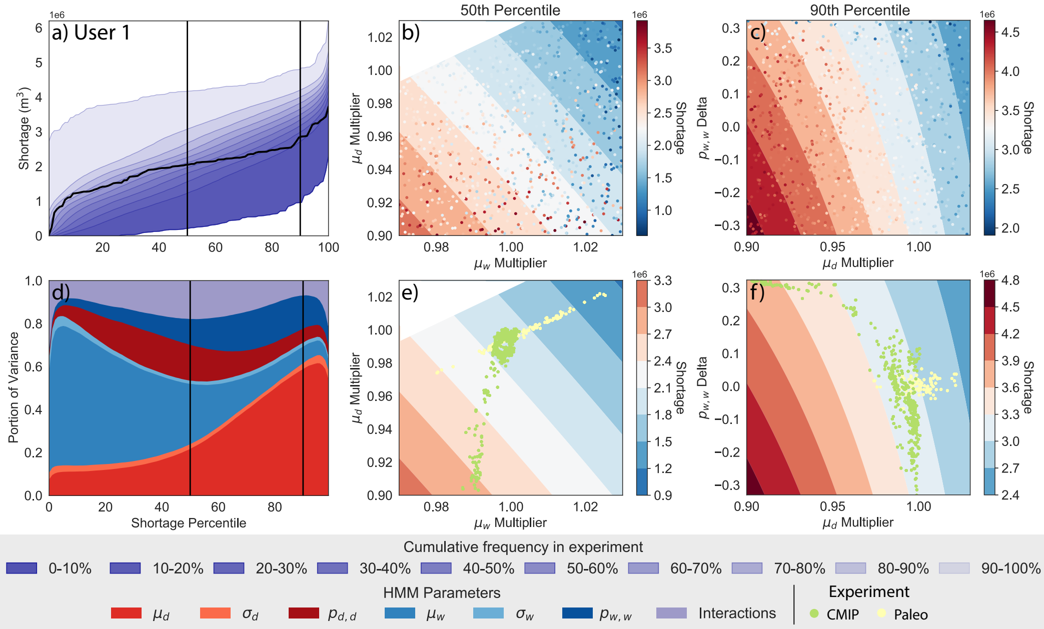

Figure 8. User 1's 50th and 90th percentile shortage response surfaces. (a) User 1's shortage distribution across the All-Encompassing experiment. (b) Average simulated 50th percentile shortage (points) in each All-Encompassing sample and the predicted shortage (contours) from a linear regression with the two most predictive parameters and their interaction, shown on the x and y axes. (c) Same as (b) but for the 90th percentile of shortage. (d) Variance decomposition of User 1's shortage under the All-Encompassing experiment. (e)-(f) Same as (b)-(c) but with CMIP samples in yellow and Paleo samples in green.

At the 50th percentile, the wet- and dry-state means are most predictive of shortage (Fig. 8(d)). The relationship of these three variables is shown in Fig. 8(b). 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 Fig. 8(e), 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 Fig. 7(b). 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 (Fig. 8(d)). Shortage increases across the All-Encompassing experiment as these two parameters decrease (Fig. 8(c)). This relationship has some curvature, indicating positive interactions between the two variables. Looking at Fig. 8(f), 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 (Fig. 7(c)). 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 mis-attribution if that correlation structure is not representative of the distribution of plausible futures.

While Fig. 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, i.e., 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 Fig. 9 for User 1 and in SI Fig. S11 for User 2.

Fig. 9(a) 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. Fig. 9(e) displays a satisficing surface introduced by Hadjimichael et al. (Accepted) illustrating the percent of realizations in the All-Encompassing experiment meeting a range of such possible success definitions (see SI Fig. S12-S13 for how this surface differs depending on the experimental design). 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 Fig. 9(e) represent the historical frequency at which different shortage magnitudes are experienced. For example, Box (c) corresponds to the example success threshold in Fig. 9(a) 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)).

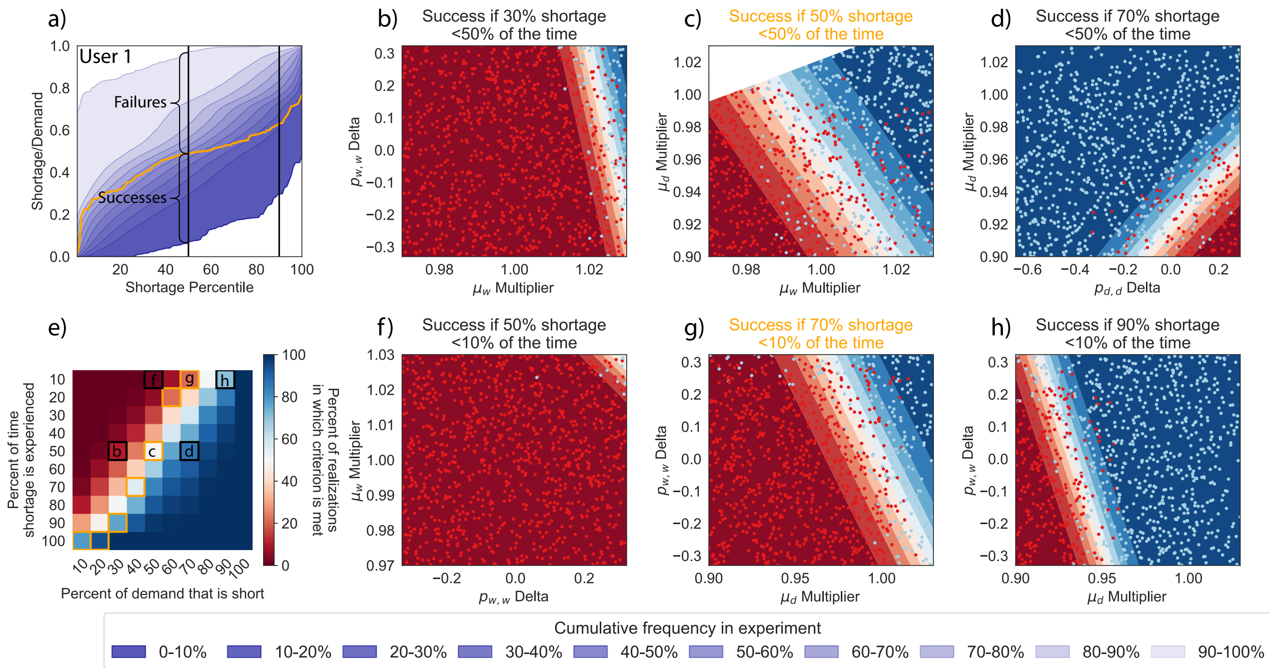

Figure 9. User 1's probability of success response surfaces under different definitions of success. (a) User 1's shortage distribution under the All-Encompassing experiment with a sample definition of success. (b)-(d) Predicted probability (contours) of different shortage magnitudes occurring less than 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 percent 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.

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 (Fig. 9(c)), we also consider a stricter definition of no more than 30% of demand if User 1 finds even historical conditions unacceptable (Fig. 9(b)), and a more lenient definition of no more than 70% of demand if User 1 is willing to tolerate an increase in shortage (Fig. 9(d)). Similarly, at the 90th percentile, we consider a smaller ratio of no more than 50% of demand being experienced 10% of the time (Fig. 9(f)), a historical ratio of no more than 70% of demand (Fig. 9(g)), and a larger ratio of no more than 90% of demand (Fig. 9(h)).

The response surfaces for the 50th percentile definitions are shown in the top row of Fig. 9 ((b)-(d)), while the response surfaces for the 90th percentile definitions are shown in the bottom row ((f)-(h)). 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 Fig. 9(c) and (g)) is the same as the most predictive parameters of variability at that percentile. However, interactions between μ_d and $p_{w,w}$ 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.

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. 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.

	Percent of Demand that is Short 50% of the Time									
Parameter	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
μ_d	**0.162**	0.019	0.025	**0.059**	**0.103**	**0.149**	**0.188**	**0.309**	**0.402**	-
σ_d	0.001	0.002	0.002	0.001	0.000	0.002	0.005	0.005	0.000	-
μ_w	**0.343**	**0.361**	**0.400**	**0.352**	**0.232**	0.062	0.014	0.008	0.000	-
σ_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	-

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.

Fig. 10 illustrates the variability in this ranking across experiments under different satisficing definitions of robustness. These satisficing definitions again correspond

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. 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.

	Percent of Demand that is Short 10% of the Time									
Parameter	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
μ_d	-	-	**0.276**	0.069	0.064	**0.248**	**0.327**	**0.478**	**0.546**	**0.261**
σ_d	-	-	0.143	0.028	0.002	0.007	0.008	0.006	0.008	0.018
μ_w	-	-	**0.303**	**0.217**	**0.203**	**0.139**	0.050	0.010	0.006	0.002
σ_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

to combinations of shortage magnitudes and frequencies. Fig. 10(f) 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.

Fig. 10(a)-(c) show the robustness ranks of 342 water users in StateMod under the satisficing definitions of increasing magnitudes of 90th percentile shortage. Fig. 10(a),(d) and (e) 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 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 Fig. S14).

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 corre-

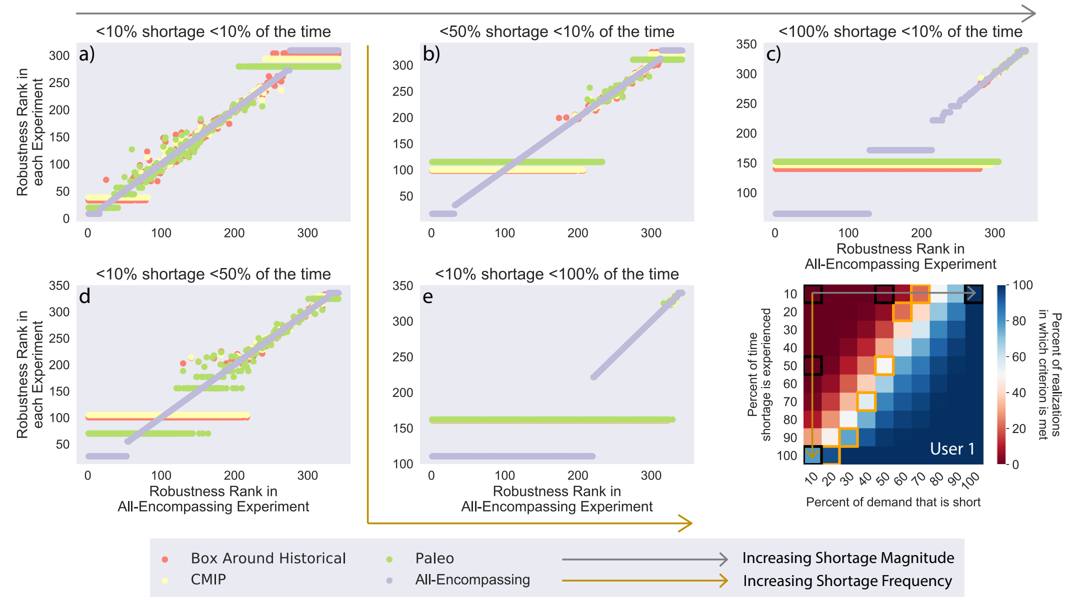

Figure 10. Consistency of robustness ranks across experiments. (a)-(c) Robustness ranks across experiments with increasing shortage magnitudes experienced no more than 10% of the time. (d)-(e) Robustness ranks across experiments with increasing frequencies of observing shortages of no more than 10% of demand. (f) User 1's satisficing surface under the All-Encompassing experiment. Black boxes indicate the satisficing thresholds under which robustness rankings are compared in the other panels.

lation 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, under-designed monitoring programs or costly, over-designed 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 mis-identification of important factors or their interactions in designing a monitoring program to detect changing or failure conditions.

While this study only considered hydrologic uncertainty for illustrative purposes, vulnerability assessments under deep uncertainty should also include uncertainties in climate-human feedbacks as part of the alternative experimental designs. 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 et al., Accepted). 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 socio-hydrologic 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 robustness. 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]."

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. StateMod is available at `https://github.com/OpenCDSS`. The input files to run StateMod for the UCRB can be found at `https://www.colorado.gov/pacific/cdss/surface-water-statemod`. All the scripts to replicate the analysis performed in this paper and regenerate the figures can be found at `https://github.com/julianneq/cdss-app-statemod-fortran/tree/2063d13/UCRB_analysis`.

References

Akintug, B., & Rasmussen, P. (2005). A markov switching model for annual hydrologic time series. *Water resources research*, *41*(9).

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 paleoclimate 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.

Ault, T. R., Mankin, J. S., Cook, B. I., & Smerdon, J. E. (2016). Relative impacts of mitigation, temperature, and precipitation on 21st-century megadrought risk in the american southwest. *Science Advances*, *2*(10), e1600873.

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*, 104699.

Beh, E. H., Dandy, G. C., Maier, H. R., & Paton, F. L. (2014). Optimal sequencing of water supply options at the regional scale incorporating alternative water supply sources and multiple objectives. *Environmental Modelling & Software*, *53*, 137–153.

Beh, E. H., Maier, H. R., & Dandy, G. C. (2015). Adaptive, multiobjective optimal sequencing approach for urban water supply augmentation under deep uncertainty. *Water Resources Research*, *51*(3), 1529–1551.

Beh, E. H., Zheng, F., Dandy, G. C., Maier, H. R., & Kapelan, Z. (2017). Robust optimization of water infrastructure planning under deep uncertainty using metamodels. *Environmental modelling & software*, *93*, 92–105.

Ben-Haim, Y. (2006). *Info-gap decision theory: decisions under severe uncertainty.* Elsevier.

Borgomeo, E., Hall, J. W., Fung, F., Watts, G., Colquhoun, K., & Lambert, C. (2014). Risk-based water resources planning: Incorporating probabilistic nonstationary climate uncertainties. *Water Resources Research*, *50*(8), 6850–6873.

Borgomeo, E., Mortazavi-Naeini, M., Hall, J. W., O'Sullivan, M. J., & Watson, T. (2016). Trading-off tolerable risk with climate change adaptation costs in water supply systems. *Water Resources Research*, *52*(2), 622–643.

Bracken, C., Rajagopalan, B., & Zagona, E. (2014). A hidden m arkov model combined with climate indices for multidecadal streamflow simulation. *Water Resources Research*, *50*(10), 7836–7846.

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*, 100051.

Brown, C., Ghile, Y., Laverty, M., & Li, K. (2012). Decision scaling: Linking bottom-up vulnerability analysis with climate projections in the water sector. *Water Resources Research*, *48*(9).

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.

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*(9), 6751–6768.

CWCB. (2012). *Colorado River Water Availability Study Phase I Report* (Tech. Rep.). Colorado Water Conservation Board.

CWCB, & CDWR. (2016). *Upper Colorado River Basin Water Resources Planning Model User's Manual* (Tech. Rep.). Colorado Water Conservation Board and Colorado Division of Water Resources. Retrieved 2019-10-02, from `https://www.colorado.gov/pacific/cdss/modeling-dataset-documentation`

Di Baldassarre, G., Wanders, N., AghaKouchak, A., Kuil, L., Rangecroft, S., Veldkamp, T. I., … Van Loon, A. F. (2018). Water shortages worsened by reservoir effects. *Nature Sustainability*, *1*(11), 617–622.

Dittrich, R., Wreford, 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). Hess 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*(7), 5619–5633.
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., Miotti, M., Swaminathan, J., Klemun, M. M., Strzepek, K., & Siddiqi, A. (2017). Water supply infrastructure planning: decision-making framework to classify multiple uncertainties and evaluate flexible design. *Journal of Water Resources Planning and Management*, *143*(10), 04017061.
Fletcher, S. M., Strzepek, K., Alsaati, A., & de Weck, 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., ... others (2020). Resilience by design in mexico city: A participatory human-hydrologic systems approach. *Water Security*, *9*, 100053.
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), 05014008.
Haasnoot, M., Kwakkel, J. H., Walker, W. E., & ter Maat, J. (2013, April). Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. *Global Environmental Change*, *23*(2), 485–498. Retrieved 2014-09-16, from `http://www.sciencedirect.com/science/article/pii/S095937801200146X` doi: 10.1016/j.gloenvcha.2012.12.006
Haasnoot, M., van't Klooster, S., & Van Alphen, J. (2018). Designing a monitoring system to detect signals to adapt to uncertain climate change. *Global environmental change*, *52*, 273–285.
Hadjimichael, A., Quinn, J., & Reed, P. (2020). Advancing diagnostic model evaluation to better understand water shortage mechanisms in institutionally complex river basins. *Earth and Space Science Open Archive*. doi: 10.1002/essoar.10503255.1
Hadjimichael, A., Quinn, J., Wilson, E., Reed, P., Basdekas, L., Yates, D., & Garrison, M. (Accepted). Defining robustness, vulnerabilities, and consequential scenarios for diverse stakeholder interests within the upper colorado river basin. *Earth's Future*.
Hadjimichael, A., Reed, P., & Quinn, J. (2020). Navigating deeply uncertain tradeoffs in harvested predator-prey systems. *Complexity*, *2020*.
Harding, 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.
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.
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*(10), 04015012.
Herman, J. D., & Usher, W. (2017). Salib: an open-source python library for sensitivity analysis. *Journal of Open Source Software*, *2*(9), 97.
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*(10), 7692–

7713.

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*(1).

Jeuland, M., & Whittington, D. (2014). Water resources planning under climate change: Assessing the robustness of real options for the blue nile. *Water Resources Research*, *50*(3), 2086–2107.

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.

Knighton, 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*(12), 10043–10064.

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation cmip5 and how we got there. *Geophysical Research Letters*, *40*(6), 1194–1199.

Korteling, B., Dessai, S., & Kapelan, Z. (2013). Using information-gap decision theory for water resources planning under severe uncertainty. *Water resources management*, *27*(4), 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.

Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, *32*(3), 679–693.

Lamontagne, J., Reed, P., Marangoni, G., Keller, K., & Garner, G. (2019). Robust abatement pathways to tolerable climate futures require immediate global action. *Nature Climate Change*, *9*(4), 290–294.

Lamontagne, 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*(3), 488–504.

Lebedev, S. (2015). *hmmlearn.* Retrieved from `https://hmmlearn.readthedocs.io/`

Lempert, R. J., & Collins, M. (2007). Managing the risk of an uncertain threshold response: Comparison of robust, optimimum, and precautionary approaches. *Risk Analysis*, *27*(4), 1009–1026.

Lempert, R. J., Popper, S. W., & Bankes, S. C. (2010). Robust decision making: coping with uncertainty. *The Futurist*, *44*(1), 47. Retrieved from `https://search.proquest.com/docview/218561458?accountid=10267`

Maier, H. R., Guillaume, J. H., van Delden, 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. Bennett, & Catherine, N.-L. (2001). Colorado's Decision Support Systems: Data-Centered Water Resources Planning and Administration. *Watershed Management and Operations Management 2000*, 1–9. Retrieved 2019-12-04, from `https://ascelibrary.org/doi/abs/10.1061/40499(2000)153` doi: 10.1061/40499(2000)153

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*(2), 169–191.

Milly, P., & Dunne, K. (2020). Colorado river flow dwindles as warming-driven loss of reflective snow energizes evaporation. *Science*.

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*, 119797.

Moallemi, E. A., Zare, F., Reed, P. M., Elsawah, S., Ryan, M. J., & Bryan, B. A. (2020, January). Structuring and evaluating decision support processes to enhance the robustness of complex human–natural systems. *Environmental Modelling & Software*, *123*, 104551. Retrieved 2019-12-03, from `http://www.sciencedirect.com/science/article/pii/S1364815219306905` doi: 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*(6), 3576–3588.

Mortazavi-Naeini, M., Kuczera, G., Kiem, A. S., Cui, L., 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*(8).

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*, *390*(3-4), 198–209.

Puy, A., Piano, S. L., & Saltelli, A. (2020). A sensitivity analysis of the pawn sensitivity index. *Environmental Modelling & Software*, 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*(8), 7208–7233.

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*.

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*, 125–141.

Rasmussen, R., Ikeda, K., Liu, C., Gochis, D., Clark, M., Dai, A., … others (2014). Climate change impacts on the water balance of the colorado headwaters: high-resolution regional climate model simulations. *Journal of Hydrometeorology*, *15*(3), 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., Funtowicz, S., Giampietro, M., Kaiser, M., Reinert, 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., & Funtowicz, 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.* John Wiley & Sons.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th python in science conference.*

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 approach to resolving divergent climate assumptions in water resources planning. *Journal of Water Resources Planning and Management*, *144*(9), 04018054.

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*(1857), 2163–2177.

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

State of Colorado. (2015). *Colorado's water plan* (Tech. Rep.). Denver, Colorado.

Steinschneider, S., McCrary, R., Mearns, 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*(12), 5014–5044.

Taner, M. Ü., Ray, P., & Brown, C. (2019). Incorporating multidimensional probabilistic information into robustness-based water systems planning. *Water Resources Research*, *55*(5), 3659–3679.

Thyer, M., & Kuczera, G. (2003). A hidden markov model for modelling long-term persistence in multi-site rainfall time series 1. model calibration using a bayesian approach. *Journal of Hydrology*, *275*(1-2), 12–26.

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*(21), 11–329.

Whateley, S., Steinschneider, S., & Brown, C. (2014). A climate change range-based

method for estimating robustness for water resources supply. *Water Resources Research*, *50*(11), 8944–8961.
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*(5).
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*(3), 1868–1889.
Xiao, M., Udall, B., & Lettenmaier, D. P. (2018). On the causes of declining colorado river streamflows. *Water Resources Research*, *54*(9), 6739–6756.
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*(9), 7327–7346.