Supporting Information for "Defining robustness, vulnerabilities, and consequential scenarios for diverse stakeholder interests within the Upper Colorado River Basin"

Antonia Hadjimichael[1], Julianne Quinn[2], Erin Wilson[3], Patrick Reed[1], Leon Basdekas[4], David Yates[5], Michelle Garrison[6]

[1]School of Civil and Environmental Engineering, Cornell University, Ithaca, NY, USA

[2]Department of Engineering Systems and Environment, University of Virginia, Charlottesville, VA, USA

[3]Wilson Water Group LLC, Lakewood, Colorado, USA

[4]Black and Veatch, Colorado Springs, Colorado, USA

[5]National Center for Atmospheric Research, Boulder, Colorado, USA

[6]Colorado Water Conservation Board, Denver, Colorado, USA

ah986@cornell.edu

Contents of this file

Introduction

This supporting information document contains additional information on how the two-state Gaussian Hidden Markov Model was fit and validated using historic observations, and on how irrigation time series were correlated with synthetic streamflow time series. Six figures and two tables complementing results in the main manuscript are also included.

Section S1: Synthetic Streamflow Model

Generation of Annual Flows at the Basin Outlet
Hydrologic variables often exhibit long-term persistence caused by regime-shifting behavior in the climate, such as the El Niño-Southern Oscillations (ENSO). One popular way of modeling this long-term persistence is with hidden Markov models (HMMs) (Thyer & Kuczera, 2003; Akintug & Rasmussen, 2005; Bracken et al., 2014). In a hidden Markov model, the climate "state" (e.g., wet or dry) at a particular time step depends only on the state from the previous time step, but the state in this case is "hidden," i.e. not observable. All that is observed is a random variable (discrete or continuous) that was generated under a particular (unknown) state.

HMMs are useful for describing hydrologic variables that may exhibit great persistence in the state (e.g. wet or dry) without being highly correlated themselves, making simple auto-regressive models unfit (Bracken et al., 2014). Furthermore, paleodata suggests that greater persistence (e.g. megadroughts) in precipitation is often observed than would be predicted by autoregressive models (Ault et al., 2013, 2016). HMMs are used in this study to capture the persistence of wet and dry years known to exist in the Colorado River Basin, while also investigating how shortage and over-year drought might be impacted by changes in the region's future climate regimes, as modeled through changing parameters of the HMM.

In this study we fit a two-state Gaussian HMM to the log-space annual streamflow, Y_t, observed in year t at the Colorado-Utah state line, the last node in the StateMod model. The distribution of Y_t depends on the state at time t, X_t (in this case, wet or dry). We find that the log-space annual flows in each state X_t at the Colorado-Utah state line can be modeled by Gaussian distributions, i.e. $f(Y_t|X_t = 0) \sim N(\mu_0, \sigma_0^2)$ and $f(Y_t|X_t = 1) \sim N(\mu_1, \sigma_1^2)$. The state at time t, X_t, depends on the state at the previous time step, X_{t-1}. The probability of switching between states is represented by the state transition matrix $\mathbf{P}$, where each element $p_{i,j}$ represents the probability of transitioning from state i at time t to state j at time $t+1$, i.e. $p_{i,j} = P(X_{t+1} = j|X_t = i)$. P is a $n \times n$ matrix where n is the number of states, here 2 for wet and dry. In all Markov models, the unconditional probability of being in each state, π can be modeled by the equation $\pi\mathbf{P}$, where π is a $1 \times n$ vector in which each element π_i represents the unconditional probability of being in state i, i.e. $\pi_i = P(X_t = i)$. π is also called the stationary distribution and its elements can be calculated from $\pi_i = \frac{e_{i,1}}{\sum_{j=1}^{n} e_{j,1}}$ where $e_{j,1}$ is the j-th element of the eigenvector of $\mathbf{P^T}$ corresponding to an eigenvalue of 1. Since we have no prior information on which to condition the first set of observations, we assume the initial probability of being in each state is the stationary distribution.

In fitting a two-state Gaussian HMM to the log-space annual flows at the CO-UT state line, we need to estimate the following vector of parameters: $\theta = [\mu_0, \sigma_0, \mu_1, \sigma_1, p_{0,0}, p_{1,1}]$. Note $p_{0,1} = 1 - p_{0,0}$ and $p_{1,0} = 1 - p_{1,1}$. We do this with Python's hmmlearn package, which uses the Baum-Welch algorithm, an application of Expectation-Maximization built off of the forward-backward algorithm. The first step of this process is to set initial estimates for each of the parameters, which was done using hmmlearn default methods. The means of the Gaussian distributions are initiated at the means of $k = n$ clusters found by k-means clustering of the data (i.e. clustering the log-space annual flows into as many clusters as there are states). The variances of these distributions are each initiated at the sample variance of the data. Finally, the transition probabilities in $\mathbf{P}$ and the probabilities of being in each state at time $t = 0$ are all initiated at $1/n$.

After setting initial parameter estimates, the forward-backward algorithm is implemented. The forward step computes the joint probability of observing the first t observations and ending up in state i at time t, given the initial parameter estimates: $P(X_t = i, Y_1 = y_1, Y_2 = y_2, ..., Y_t = y_t, \theta)$. This is computed for all $t \in \{1, ..., T\}$. Then in the backward step, the conditional probability of observing the remaining observations after time t given the state at time t is computed: $P(Y_{t+1} = y_{t+1}, ..., Y_T = y_T | X_t, \theta)$. Using Bayes' theorem, it can shown that the product of the forward and backward probabilities is proportional to the probability of ending up in state i at time t given all of the observations, i.e. $P(X_t = i | Y_1 = y_1, ..., Y_T = y_T, \theta)$ (see Equations 1-4 at the end of the text). In fitting the HMM, we find the set of parameters, θ, that maximize this probability, i.e. the likelihood function of the state trajectories given our observations. As shown in equation 4, this is equivalent to maximizing the product of the probability estimates from the forward-backward algorithm, which is done using Expectation-Maximization.

Expectation-Maximization is a two-step process for maximum likelihood estimation when the likelihood function cannot be computed directly, for example, because its observations are hidden as in an HMM. The first step is to calculate the expected value of the log likelihood function with respect to the conditional distribution of X given Y and θ (the left hand side of equation 1, or proportionally, the numerator of the right hand side). The second step is to find the parameters that maximize this function. These parameter estimates are then used to re-implement the forward-backward algorithm and the process repeats iteratively until convergence or some specified number of iterations.

We also predict which states the observations were likely to have come from given the estimated parameters using the Viterbi algorithm. The Viterbi algorithm employs dynamic programming (DP) to find the most likely state trajectory. In this case, the "decision variables" of the DP problem are the states at each time step, X_t, and the "future value function" being optimized is the probability of observing the true trajectory, $(Y_1, ..., Y_T)$, given those alternative possible state trajectories. For example, let the probability that the first state was k be $V_{1,k}$. Then $V_{1,k} = P(X_1 = k) = P(Y_1 = y_1 | X_1 = k)\pi_k$. For future time steps, $V_{t,k} = P(Y_t = y_t | X_t = k)\pi_k V_{t-1,i}$ where i is the state in the previous time step. Thus, the Viterbi algorithm finds the state trajectory $(X_1, ..., X_T)$ maximizing $V_{T,k}$.

The goodness of fit of the two-state Gaussian HMM to log-space annual flows at the CO-UT state line is shown in Figure S1. Panel a shows a quantile-quantile plot of the mixed Gaussian distribution across both states, while panel b shows a quantile-quantile plot of the fitted wet-state and dry-state distributions using the years identified by the Viterbi algorithm to fall under each state. Panels c and d show the same fits using histograms and overlaid PDFs. The state identification of the historical years is shown in Figure S2(a). For the Markov property to hold, the auto-correlation of the states should decay exponentially, as is confirmed in Figure S2(b). The partial auto-correlation should be significant at one lag, as the state at time $t+1$ depends on the state at time t, but for no other lags, as all dependency is captured by the first lag, i.e. $P(X_{t+1} = x_{t+1} | X_t = x_t, ..., X_1 = x_1) = P(X_{t+1} = x_{t+1} | X_t = x_t)$. This is confirmed in Figure S2(c).

Finally, the parameters of the two-state Gaussian HMM are modified across a range of values in our scenario discovery experiment. An illustration of the real-space effect of the upper and lower bound of these changes on the annual flow distribution at the CO-UT state line is shown in Figure S3.

Disaggregation of Annual Flows in Time and Space

After generating log-space annual flows at the CO-UT state line from the HMM, these flows were 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 was probabilistically selected based on its "nearness" to the synthetic flow at the last node in terms of annual total. The daily hydrograph observed at that site in the selected year was then altered to model earlier, dissipated peaks. This shift was applied to the observed daily flows at the CO-UT state line by re-computing monthly sums over moving windows of the historical record shifted by 1-60 days. While this will simply shift, not reshape the hydrograph at the daily time step, when re-computed to monthly values, it captures the earlier rise, lower peak and protracted decline expected from earlier snowmelt (see Fig. S4(b)).

After shifting the observed monthly hydrograph at the last node, the flows were then converted to the natural flows that would have been observed without human influences from diversions and reservoir operations. This conversion was performed by first computing the difference between the cumulative flow observed each month at the gauge and the cumulative natural flow each month in the model, then applying that same difference to the shifted observed flows (see Fig. S4(a) for an example). We used these estimates of natural monthly flows under shifts of 1-60 days to temporally downscale our annual synthetic flows by applying the same proportion of flows each month in the shifted historical time series to the synthetic time series. Finally, the monthly flows at all upstream nodes were spatially downscaled using the same approach in which the selected historical year's ratios of monthly flows at each upstream node to monthly flows at the last node were also used for the synthetic year. The ability of this method to reproduce the spatial correlation structure of monthly flows in the basin is shown in Fig. S5.

Section S2: Generation of Total Annual Irrigation Anomalies

Since irrigation demands are higher when precipitation is low, while streamflow is lower when precipitation is low, there is a strong negative correlation between irrigation demands and streamflows in the UCRB. To make sure our synthetic streamflow and irrigation time series reproduce this historical correlation, we used the historical record to build a regression model of the anomalies of total irrigation demands across irrigation users in the basin as a function of annual flow anomalies at the CO-UT state line. This model is shown in Figure S6(a). The residuals from this model are independent in time (Figure S6(b)), normally distributed (Figure S6(c)) and homoscedastic (Figure S6(d)). As such, we use this model to estimate the total irrigation demand anomaly each synthetic year as a function of the annual flow anomaly in that synthetic year. We then randomly sample a residual from the fitted Gaussian distribution in Figure S6(c) to add to the prediction from the regression model. This ensures we preserve the variance in total irrigation demand anomalies. We add this time series of irrigation demand anomalies to the mean irrigation demand for that state of the world. These total irrigation demands are then downscaled to all of the irrigation structures using their average historical proportions of the total demand.

The code for the synthetic streamflow generator and irrigation model are available at `https://github.com/antonia-had/cdss-app-statemod-fortran/tree/2063d13/UCRB_analysis/Qgen`

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., 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.

Bracken, C., Rajagopalan, B., & Zagona, E. (2014). A hidden Markov model combined with climate indices for multidecadal streamflow simulation. *Water Resources Research*, *50*(10), 7836–7846. Retrieved 2019-03-26, from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014WR015567` doi: 10.1002/2014WR015567

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).

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

$$
\begin{aligned}
P(X_t = i | Y_1 = y_1, ..., Y_T = y_T, \theta) &= \frac{P(Y_1 = y_1, ..., Y_T = y_T | X_t = i, \theta) P(X_t = i|\theta)}{P(Y_1 = y_1, ..., Y_T = y_t|\theta)} && (1)\\
&= \frac{P(X_t = i, Y_1 = y_1, ..., Y_t = y_t|\theta) P(Y_{t+1} = y_{t+1}, ..., Y_T = y_T | X_t = i, \theta)}{P(Y_1 = y_1, ..., Y_T = y_t|\theta)} && (2)\\
&= \frac{P(X_t = i, Y_1 = y_1, ..., Y_t = y_t|\theta) P(Y_{t+1} = y_{t+1}, ..., Y_T = y_T | X_t = i, \theta)}{P(Y_1 = y_1, ..., Y_T = y_t|\theta)} && (3)\\
&\propto P(X_t = i, Y_1 = y_1, ..., Y_t = y_t|\theta) P(Y_{t+1} = y_{t+1}, ..., Y_T = y_T | X_t = i, \theta) && (4)
\end{aligned}
$$

Equations 1-4 (1) Bayes' Theorem. (2) conditional independence of the observations up to time t $(Y_1, Y_2, ..., Y_t)$ and the observations after time t $({}_{t+1}, Y_{t+2}, ..., Y_T)$, given the state at time t (X_t). (3) definition of conditional probability. (4) recognizing denominator as a normalizing constant.

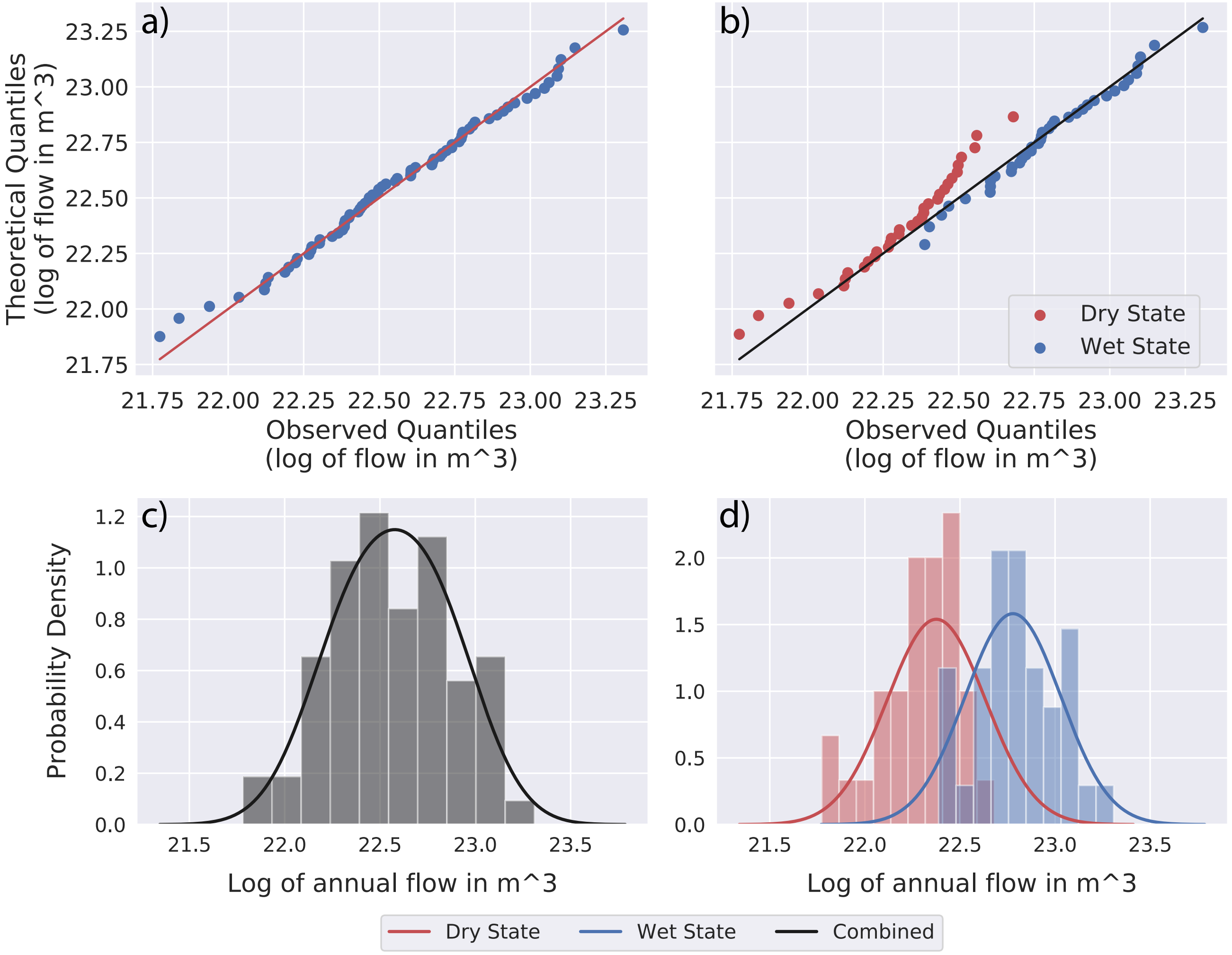

Figure S1. Goodness of Fit Assessment: Gaussian distributions. (a) QQ-plot of fitted mixed Gaussian distribution of log-space annual flows. (b) QQ-plot of Gaussian distribution from estimated wet state (blue) and dry state (red) log-space annual flows. (c) Same as a, but displayed as a histogram. (d) Same as b, but displayed as a histogram.

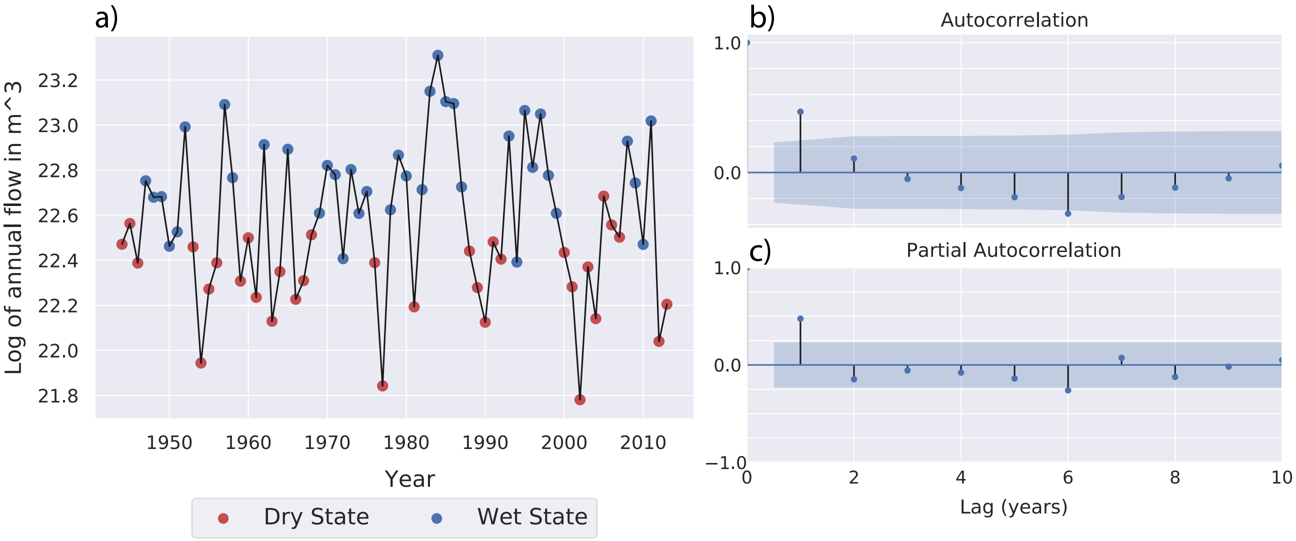

Figure S2. Goodness of Fit Assessment: State identification and Markov property. (a) Classification of last 70 historical annual flows as being generated from a wet or dry state according to the Viterbi algorithm. (b) Auto-correlation of identified states illustrate exponential decay. (c) Partial auto-correlation of identified states is only significant for one lag, confirming the Markov property.

Table S1. 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 last node of the basin.

Two-state Gaussian HMM Parameter	**Maximum Likelihood Estimate**
Log-space dry flow mean (m^3), μ_0	22.38
Log-space dry flow standard deviation (m^3), σ_0	0.26
Log-space wet flow mean (m^3), μ_1	22.78
Log-space wet flow standard deviation (m^3), σ_1	0.25
Dry-to-dry transition probability, $p_{0,0}$	0.68
Wet-to-wet transition probability, $p_{1,1}$	0.65

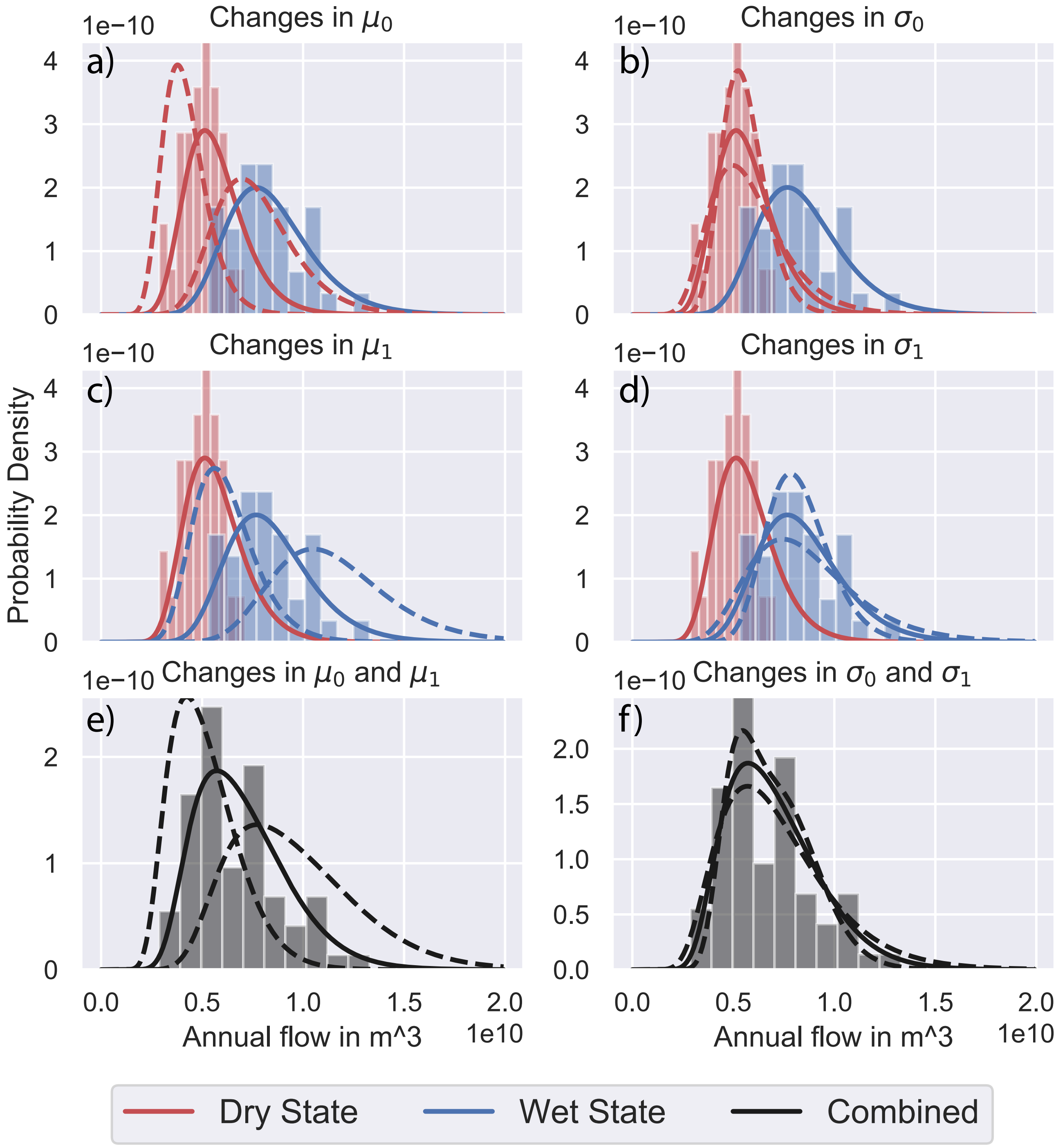

Figure S3. Illustration of Parameter Changes. (a) Changes in μ_0. (b) Changes in σ_0. (c) Changes in μ_1. (d) Changes in σ_1. (e) Changes in μ_0 and $\mu 1$. (f) Changes in σ_0 and σ_1. Base-case distributions shown in solid lines, increases and decreases in dashed lines. Dry state distribution shown in red, wet state distribution in blue, and combined distribution in black.

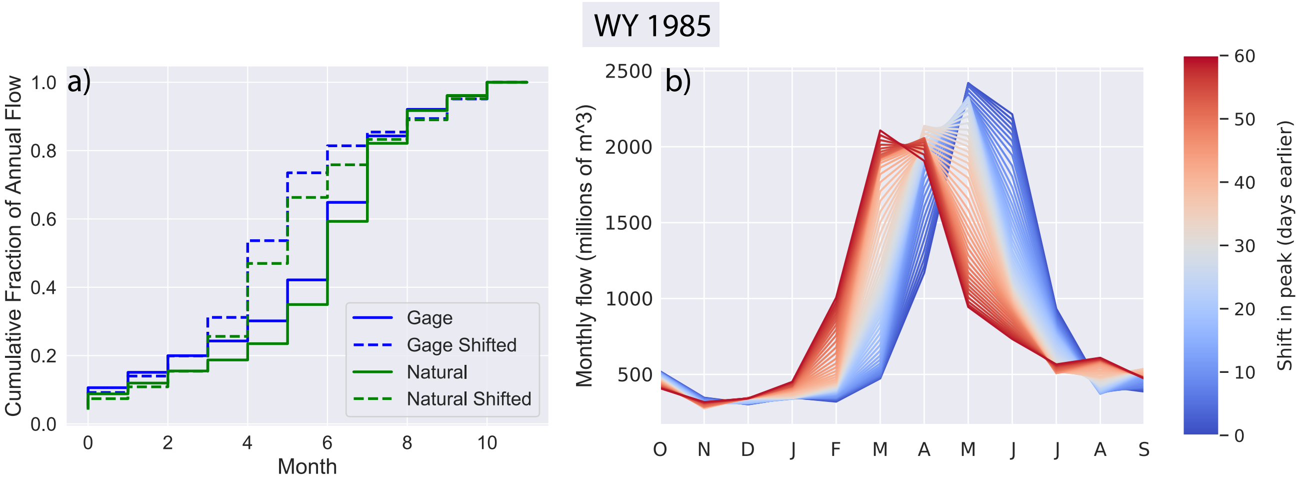

Figure S4. Illustration of seasonal shift model. (a) Example CDF correction to convert gauged flows to natural flows using a 60-day shift for water year 1973. (b) Effect of peak runoff shift for water year 1973.

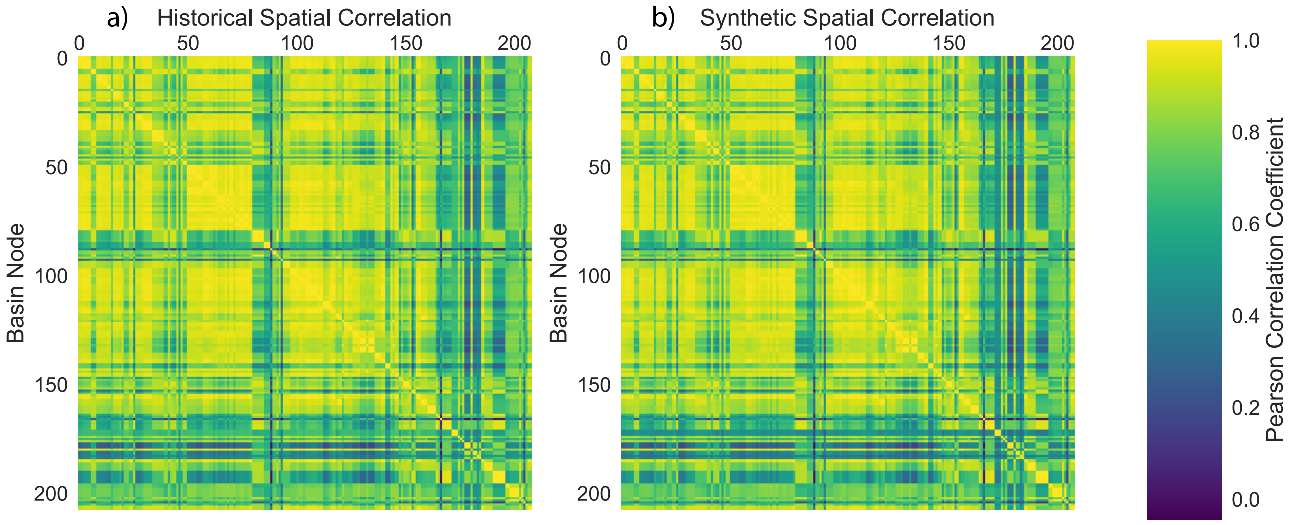

Figure S5. Validation of the spatial correlation of synthetic monthly flows at the 208 streamflows nodes in StateMod. (a) Historical spatial correlation. (b) Synthetic spatial correlation.

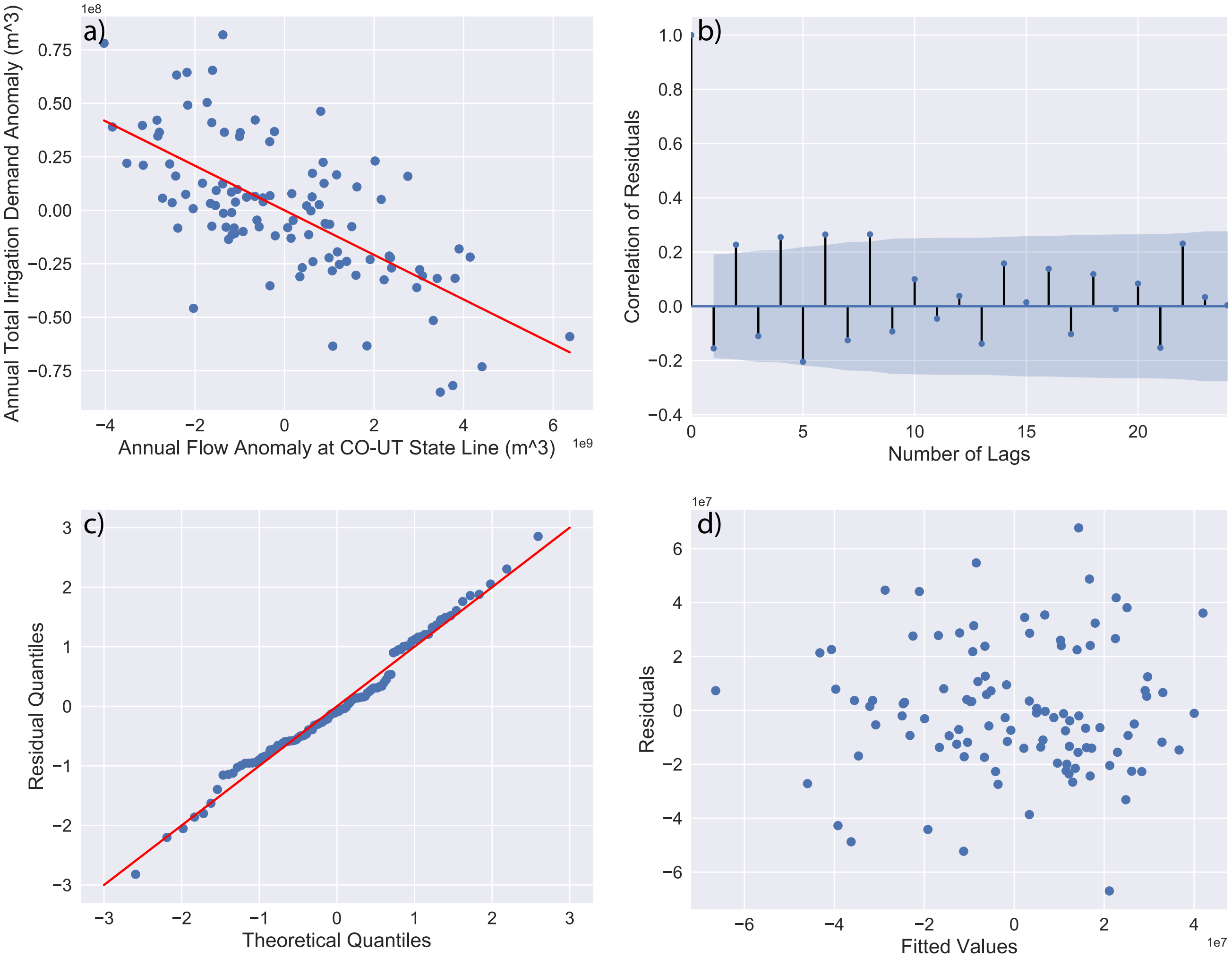

Figure S6. Irrigation anomaly model (a) Total annual irrigation anomalies as a function of annual streamflow anomalies. (b) Autocorrelation function of residuals shows they are independent. (c) Normal QQ-plot of residuals shows they are normally distributed. (d) Residuals vs. fitted plot shows residuals are homoscedastic.

Table S2. McFadden Pseudo R^2 values for predictor variables For each user and performance threshold in Fig. 7, we iteratively build regression models by using one paramater at a time. The two with the highest McFadden Pseudo R^2 values in each case are used to build the factor maps. Each table column corresponds to a panel in Fig. 7: (a) *Robust if 10% shortage occurs no more than 80% of the time* (b) *Robust if 80% shortage occurs no more than 20% of the time* (c) *Robust if 10% shortage occurs no more than 80% of the time* (d) *Robust if 40% shortage occurs no more than 10% of the time* (e) *Robust if 70% shortage occurs no more than 30% of the time* (f) *Robust if 80% shortage occurs no more than 20% of the time*

	McFadden Pseudo R^2 values					
	Median-right user		**Junior-right user**		**Transbasin diversion**	
Parameter	*(a)*	*(b)*	*(c)*	*(d)*	*(e)*	*(f)*
Change in evaporation (cm/month)	0.00	0.00	0.00	0.00	0.00	0.00
Shift in timing of snowmelt (days earlier)	0.20	0.16	0.02	0.01	0.03	0.06
Log-space dry flow mean (m^3) multiplier	0.02	0.20	0.12	0.31	0.02	0.07
Log-space dry flow standard deviation (m^3) multiplier	0.00	0.00	0.00	0.01	0.00	0.00
Log-space wet flow mean (m^3) multiplier	0.16	0.02	0.16	0.02	0.01	0.01
Log-space wet flow standard deviation (m^3) multiplier	0.00	0.00	0.00	0.00	0.00	0.00
Change in dry-to-dry transition probability	0.02	0.02	0.04	0.03	0.00	0.00
Change in wet-to-wet transition probability	0.01	0.01	0.02	0.01	0.01	0.01
Irrigation demand multiplier	0.04	0.02	0.00	0.07	0.00	0.00
Transbasin demand multiplier	0.00	0.00	0.00	0.00	0.52	0.33
Municipal and industrial demand multiplier	0.00	0.00	0.00	0.00	0.00	0.00
Reservoir storage	0.00	0.00	0.00	0.00	0.00	0.00
Operation of Shoshone Power Plant	0.00	0.00	0.00	0.00	0.00	0.01
Seniority of environmental flows	0.00	0.00	0.00	0.00	0.00	0.00