Supporting Information for A Statistical Graphical Model of the California Reservoir System

A. Taeb,[1] J.T. Reager,[2], M. Turmon,[2], V. Chandrasekaran,[1]

Contents of this file

1. Preprocessing the Reservoir/Covariates Data
2. Checking for Gaussianity
3. Sensitivity of Graphical Model to λ
4. Correlating Covariates to the Latent Space
5. Identifying Reservoirs Most at Risk of Exhaustion

Preprocessing the Reservoir/Covariates Data: Let $\{\bar{y}^{(i)}\}_{i=1}^{n_{\text{train}}} \subset \mathbb{R}^{55}$ and $\{\bar{y}^{(i)}\}_{i=1}^{n_{\text{test}}} \subset \mathbb{R}^{55}$ be the averaged monthly reservoir volumes in the training and validation set respectively. Focusing on a reservoir r and the month of January, let $\mu_{\bar{y}_r}$ be the average reservoir level during January (obtained only from training observations). For each observation i in January, we apply the transformation:

$$\tilde{y}_r^{(i)} = \bar{y}_r^{(i)} - \mu_{\bar{y}_r}.$$

[1]California Institute of Technology.

[2]Jet Propulsion Laboratory

We repeat the same steps for all months. Furthermore, letting σ_r be the sample standard deviation of the training observations $\{\tilde{y}_r^{(i)}\}_{i=1}^{n_{\text{train}}}$, we produce unit variance observations with the transformation,

$$y_r^{(i)} = \frac{1}{\sigma_r^{1/2}} \tilde{y}_r^{(i)}.$$

We repeat the same steps for all reservoirs to obtain the preprocessed reservoir observations $\{y^{(i)}\}_{i=1}^{n_{\text{train}}}$ and $\{y^{(i)}\}_{i=1}^{n_{\text{test}}}$. Finally, the same steps are repeated to preprocess the covariates data.

Checking Gaussianity: We verify that the joint reservoir anomalies (after preprocessing steps) can be well-approximated by a multivariate Gaussian distribution. To check for the Gaussianity assumption, we use a commonly employed method known as Q-Q plot. This is a graphical procedure for comparing two probability distribution by plotting their quantiles against each other. In particular, we compare the quantiles of the reservoir observations with a multivariate normal distribution. Figure 1(a) shows the Q-Q plot for the 55 reservoirs. We notice that by removing the Farmington reservoir, the Q-Q plot shown in Figure 1(b) exhibits a strong linear relationship, suggesting that these 54 reservoirs are well-approximately jointly by a multivariate Gaussian distribution.

Sensitivity of Graphical Model to λ: As described in the main text, the regularization parameter λ is varied from 0 to 1 to identify a collection of graphical models. For each graphical model, we measure the training and validation log-likelihood performances. Figure 2 illustrates the training and validation performances for different values of λ. Recall that $\lambda = 0$ corresponds to an unregularized maximum likelihood estimate and $\lambda = 1$ corresponds to independent reservoir model. We chose $\lambda = 0.23$ to obtain a graphical

model with the best validation performance. The training and validation performances of these models are summarized in Table 1.

To demonstrate that the graphical model estimate does not vary significantly under small perturbations to λ, we also obtain graphical model estimates with $\lambda = 0.26$ and $\lambda = 0.20$ (Recall that the edge strengths in a graphical model contain the relevant information of the model). Figure 3(a) compares the edge strengths of the model with $\lambda = 0.23$ and the model with $\lambda = 0.20$. Furthermore, Figure 3(b) compares the edge strengths of the model with $\lambda = 0.23$ and the model with $\lambda = 0.26$. Evidently, strong edges persist across all models, with a few weak edges removed or added as λ is varied. The total number of edges in the graphical model when $\lambda = 0.20$, $\lambda = 0.23$, and $\lambda = 0.26$ is 295, 285, and 279 respectively. Furthermore, the quantity κ (defined in equation (4) of main paper) is $0.852, 0.859$, and 0.862 for $\lambda = 0.20$, $\lambda = 0.23$, and $\lambda = 0.26$. These results suggest that our conclusions are not particularly sensitive to the choice of the regularization parameter, although we chose $\lambda = 0.23$ as it leads to the best validation performance.

Correlating Covariates to the Latent Space: Latent variable graphical modeling identifies a summarization of external phenomena influencing the reservoir network; these influences are summarized by global latent variables. In the main paper, we introduced the *latent space*, a space of all possible configurations of the latent variable time series. Here, we describe the manner in which compute the correlation of a candidate covariate with the latent space. Let $\mathcal{T} \subset \mathbb{R}^n$ with $\dim(\mathcal{T}) = k$ denote the latent space. Let $X_1 \in \mathbb{R}^{n_{\text{train}}}$ be the n_{train} observations of the covariate x_1 (normalized to have unit variance). The

correlation of this covariate with the latent space is given by:

$$\text{corr}(x_1) = \left\| \mathcal{P}_{\mathcal{T}}(X_1) \right\|_{\ell_2},$$

where $\mathcal{P}_{\mathcal{T}}$ denotes the projection matrix onto the subspace $\mathcal{T}$. By definition, the quantity $\text{corr}(x_1)$ is between 0 and 1 with large values indicating that the covariate x_1 has a strong influence over the entire reservoir network.

Suppose we have identified a covariate x_1 that is highly correlated with the latent space. We can modify our technique to identify other covariates that are correlated with the latent space after taking away the effect of the covariate x_1.

Let $U_1 D_1 V_1'$ be the reduced SVD of X_1 where $U_1 \in \mathbb{R}^{n_{\text{train}}}$, $D_1 \in \mathbb{R}$ and $V_1 \in \mathbb{R}$. Let $X_2 \in \mathbb{R}^{n_{\text{train}}}$ be the n_{train} observations of the covariate x_2. The correlation of a covariate x_2 with the latent space after taking away the effect of x_1 is given by:

$$\text{corr}_{x_1}(x_2) = \left\| (I - U_1 U_1') \mathcal{P}_{\mathcal{T}} (I - U_1 U_1')(X_2) \right\|_{\ell_2}.$$

If the quantity $\text{corr}_{x_1}(x_2)$ is large, then the covariate x_2 is strongly correlated to the second global statewide variable. We can once again take away the effect of the covariates x_1 and x_2 from the latent space, and find its correlation with another covariate x_3. Let $U_2 D_2 V_2'$ be the reduced SVD of $[X_1, X_2] \in \mathbb{R}^{n_{\text{train}} \times 2}$ where $U_2 \in \mathbb{R}^{n_{\text{train}} \times 2}$, $D_2 \in \mathbb{R}^{2 \times 2}$ and $V_2 \in \mathbb{R}^{2 \times 2}$. Let $X_3 \in \mathbb{R}^{n_{\text{train}}}$ be the n_{train} observations of the covariate x_3. The correlation of a covariate x_3 with the latent space after taking away the effect of x_1 and x_2 is given by:

$$\text{corr}_{x_1,x_2}(x_3) = \left\| (I - U_2 U_2') \mathcal{P}_{\mathcal{T}} (I - U_2 U_2')(X_3) \right\|_{\ell_2}.$$

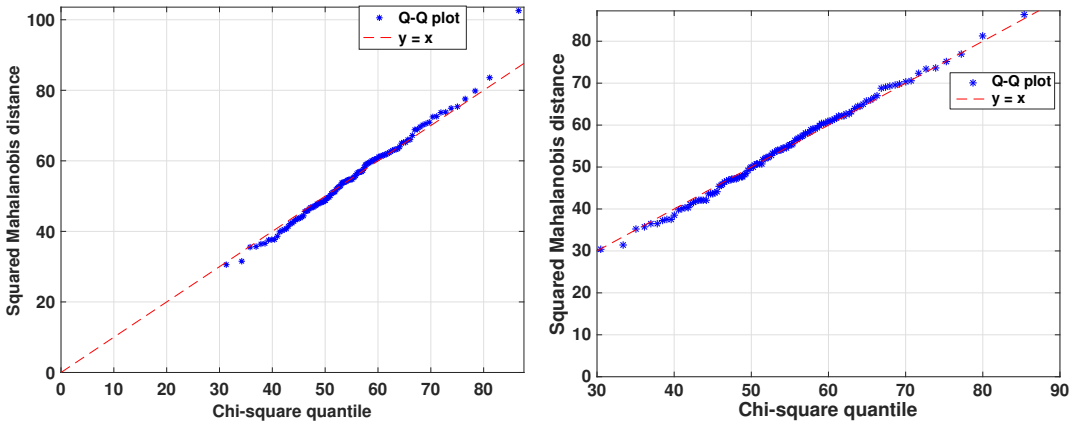
Similarly, if the quantity $\text{corr}_{x_1,x_2}(x_3)$ is large, then the covariate x_3 is strongly correlated to the third global driver. We can repeat this procedure to identify all the k global drivers

influencing the reservoir network.

The latent variable graphical model identified two global drivers influencing the reservoir network. As described in the preceding paragraphs, this yields a two dimensional latent space corresponding to all possible observations of the global drivers. To obtain real-world representation of these two global drivers, we link the two dimensional *latent space* to the 7 covariates described in Section 2.2 (main paper). The correlation values of each covariate with the latent space are shown in the second column of Table 2. We then take the effect of PDSI away from the latent space to find the correlation of the modified latent space with the remaining 6 covariates. These correlation values are shown in the third column of Table 2.

Identifying Reservoirs Most at Risk of Exhaustion: As described in the main text, our modeling framework serves a powerful tool to identify reservoirs that are high risk of exhaustion so that appropriate preventive management practices could be employed. For each reservoir, we sweep over a range of PDSI and use (11) (main text) to compute probabilities of exhaustion. Figure 4 shows those reservoirs (among 31 large reservoirs with capacity greater than $10^8 m^3$) that were highly sensitive to PDSI. Evidently, these reservoirs are at high risk of exhaustion, and additionally, some have a greater sensitivity to small PDSI changes than others. We focus on two reservoirs with highest risk of exhaustion: Buchanan and Hidden Dam reservoir. We consider Figure 5 which demonstrates the historical volumes of these reservoirs in response to PDSI. Notice that as expected, there is a positive correlation between PDSI and reservoir volumes: smaller values of PDSI generally result in a lower volume. An interesting phenomenon seems to occur for very small values of PDSI (e.g. less than 3 corresponding to drought period 2014-2015). In

this range, changes to PDSI do not appear to substantially impact the reservoir volumes. In other words, the correlation between PDSI and reservoir volumes is significantly reduced as compared to the correlation during normal and wet periods. To provide concrete numbers on the reduction in this correlation, we focus on November volumes of Buchanan and Hidden Dam reservoirs and the corresponding September PDSI values. We further restrict to observations where PDSI is less than 3. We compute the Pearson Correlation Coefficient between PDSI and each reservoir during this period. This correlation for the Buchanan reservoir is a factor of $\approx 6/100$ of the value estimated by our model. Similarly, the correlation for the Hidden Dam is a factor $\approx 2/5$ of the correlation estimated by our model. As described in the main paper, the large drops in correlations are due to strict management. Figure 6 demonstrates the amount of water from precipitation into the Hidden Dam and Buchanan reservoirs, the total inflow, and the outflow as a consequence of the stringent management efforts. Examining Figure 6, notice that there was little to no outflow of water, which keeps the reservoir volumes mostly constant and prevents them from running dry.



(a) (b)

Figure 1: (a): Q-Q plot of the entire set of 55 reservoirs, (b): Q-Q plot of 54 reservoirs (excluding the Farmington reservoir). The Q-Q plots are against a multivariate Gaussian distribution. Notice that $y = x$ is a close approximation to the Q-Q plot in (b) implying that 54 reservoirs (excluding Farmington reservoir) is well approximated by a multivariate Gaussian distribution.

Model	Training performance	Validation performance
unregularized ML estimate ($\lambda = 0$)	-23.91	-1140.4
independent reservoir model ($\lambda = 1$)	-83.23	-101.95
graphical model ($\lambda = 0.23$)	-63.52	$-$ **85.54**

Table 1: Training and validation performances of unregularized maximum likelihood (ML) estimate, independent reservoir model, and graphical model. As larger values of log-likelihood are indicative of better performance, the graphical model is the superior model.

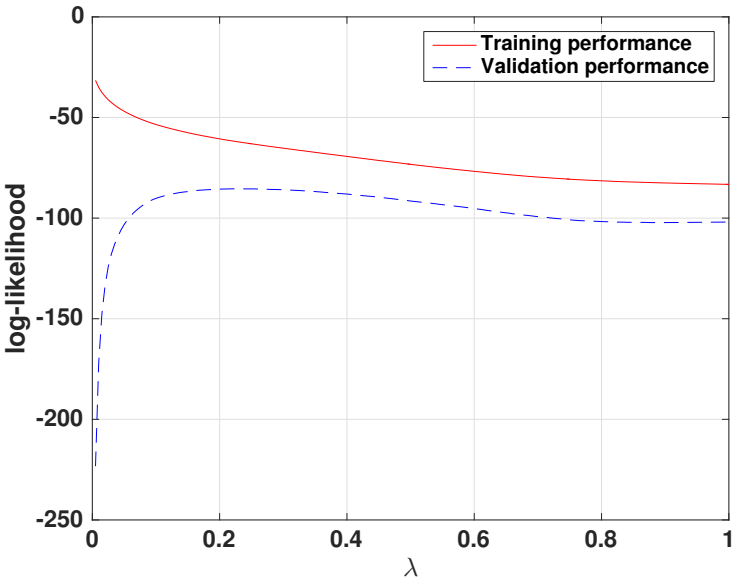
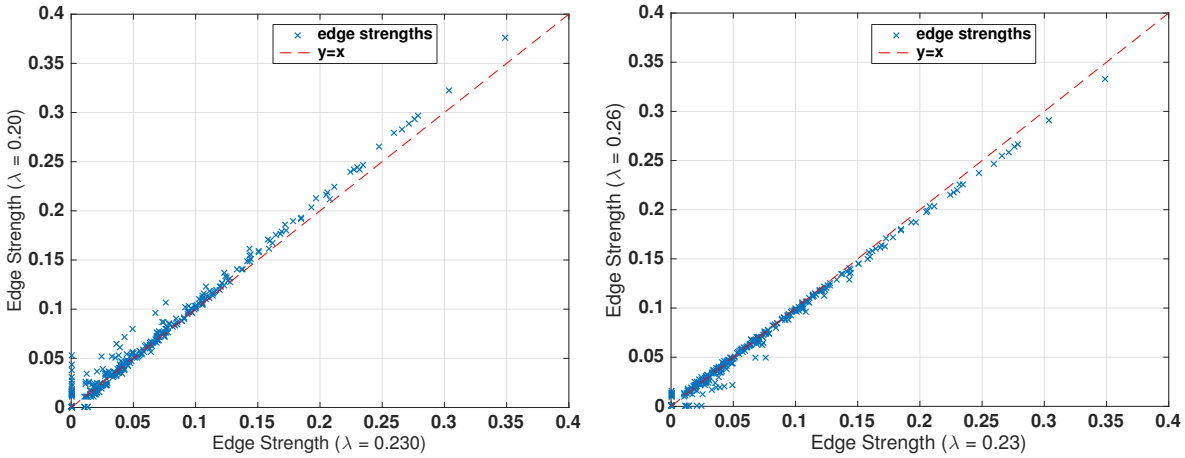


Figure 2: Training and validation performance of graphical modeling for different values of the regularization parameter λ. The training performance is computed as the average log-likelihood of training samples and the validation performance is computed as the average log-likelihood of validation samples.



(a) $\lambda = 0.23$ vs $\lambda = 0.2$

(b) $\lambda = 0.23$ vs $\lambda = 0.26$

Figure 3: Sensitivity of the graphical model estimate to perturbations of λ around the optimal value $\lambda = 0.23$ (this choice of λ leads to optimal validation performance): we observe that strong edges in the original model are strong edges in the perturbed model (i.e. with perturbed λ) with approximately the same strength.

Covariate	Correlation	Correlation after removing PDSI
Palmer Drought Severity Index (PDSI)	0.88	N/A
Hydroelectric power	0.80	0.09
Sierra Nevada snow pack	0.50	0.32
Consumer Price Index (CPI)	0.33	0.25
Colorado river discharge	0.29	0.23
Number of agricultural workers	0.17	0.03
Temperature	0.10	0.04

Table 2: Covariates and correlations with the latent space before and after removing PDSI

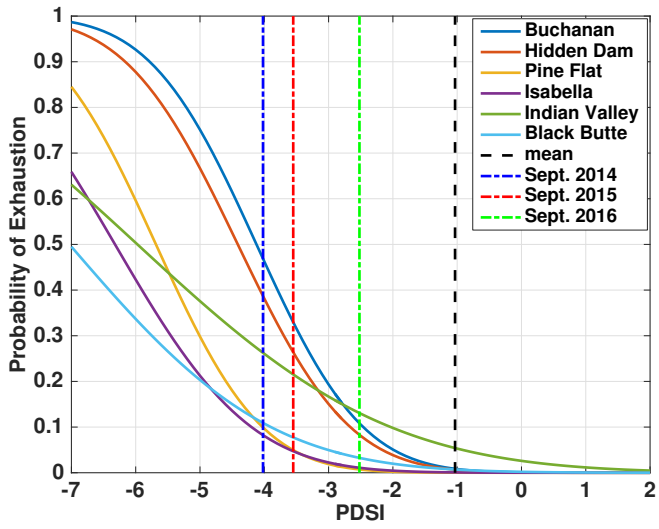
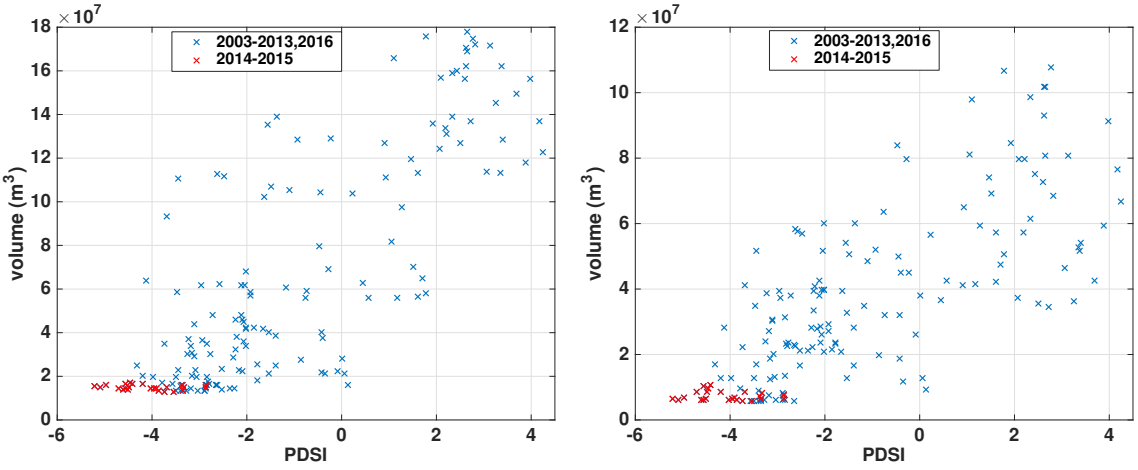


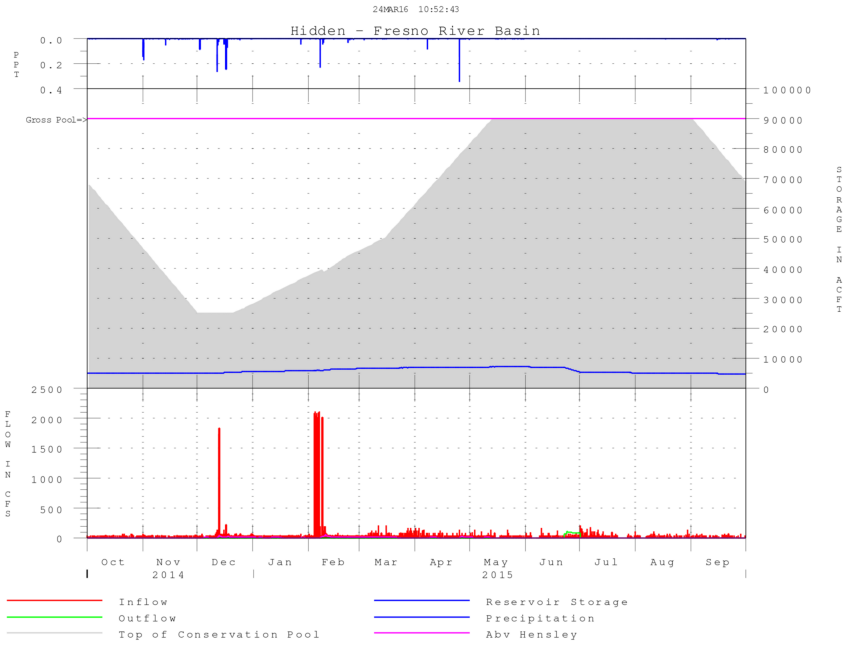
Figure 4: Individual reservoir responses to drought in a conditional latent variable graphical model: probability that six most-at-risk reservoirs out of 31 large reservoirs (with capacity $\geq 10^8 m^3$) will have volume drop below zero; Dashed black line: average September PDSI (September 2004-September 2015). Dashed blue line: September 2014 PDSI. Dashed red line: September 2015 PDSI. Dashed green line: September 2016 PDSI.



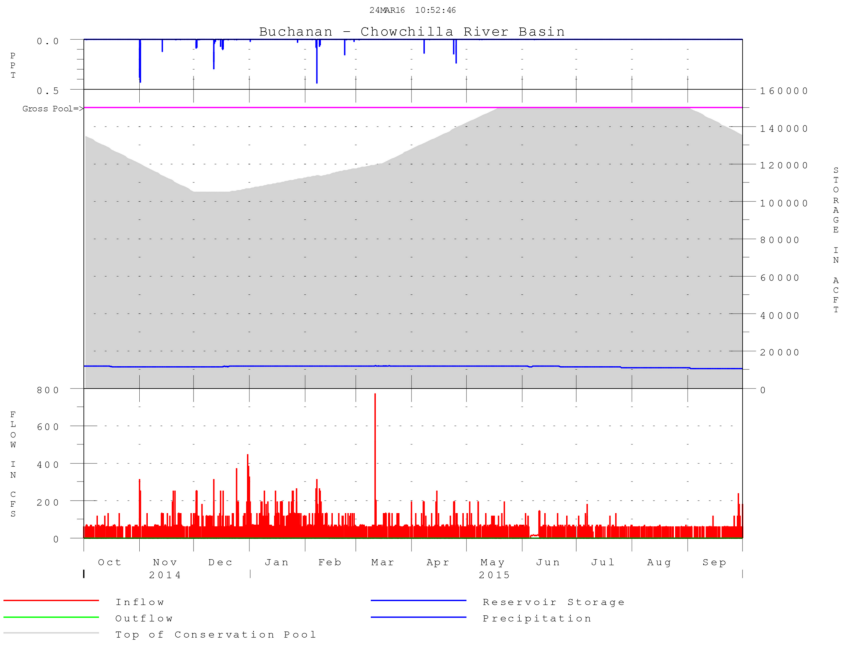
(a) Buchanan

(b) Hidden Dam

Figure 5: PDSI vs reservoir levels for the Buchanan and Hidden Dam reservoirs during the period of study (i.e. January 2003 to November 2016). Notice a positive correlation between PDSI and the reservoir volumes: smaller values of PDSI generally lead to lower reservoir volumes. During the 2014-2015 drought period (shown in red), the correlation is substantially reduced as a result of stringent management efforts.



(a) Hidden Dam, 2014-2015



(b) Buchanan, 2014-2015

Figure 6: Inflows, outflows, precipitation, and water levels for the Buchanan and Hidden Dam reservoirs during the extreme drought period of 2014-2015. Notice that there was little precipitation, leading to marginal inflow of water into each reservoirs. Due to heavy management, there was little to no outflow of water from these reservoirs, preventing them from running dry. These figures are obtained from the Sacramento District Water Control Data System at `http://www.spk-wc.usace.army.mil/plots/california.html`.