A Data-Driven Method for Robust Water Allocation under Uncertainty

Fellow:
David LOVE
Interdisciplinary Program in Applied Mathematics
University of Arizona

Adviser:
Dr. Güzin BAYRAKSAN
Integrated Systems Engineering
The Ohio State University, and Systems & Industrial Engineering
University of Arizona

May 10, 2013

University of Arizona, Technology and Research Initiative Fund 2012/2013
Water Sustainability Graduate Student Fellowship Program
1. Arizona water issue addressed

More than 25 million people in the southwestern United States depend on the water supplied by the Lower Colorado River Basin for their livelihood. More than half of Tucson’s water, for instance, comes from this source. The Colorado River Basin has experienced a sustained period of drought in recent years, which has led to questions about the adequacy of the Colorado to meet future demands, especially as the population of Arizona (and of other states that depend on this water source) increases. Thus, the problem of allocating Colorado water is of critical importance.

In this work, we model, solve, and analyze a robust water allocation problem to optimally allocate Colorado River water to different users over 40 years in the southeastern part of Tucson shown in Figure 1. There are many uncertainties in this water allocation problem—two of most important ones are water supply and water demands. On the water demand side, the southeastern portion of Tucson is expected to grow during the time horizon considered; therefore, the water demand is expected to increase but by how much is uncertain. On the water supply side, it is well recognized that the reliability of the Colorado River system under future climate variability is critically important to the long-term well-being of Arizona and the other six states that depend on this water supply [7]. The current approach of estimating water supply using climate models is to use Global Circulation Models (GCM) to generate regional rainfall and temperature scenarios by the so-called statistical downscaling techniques [2, 5]. Even though we have some idea about these uncertainties, there is considerable ambiguity as to how to model these uncertainties and how to make long-term decisions under such ambiguous uncertainties. The work presented here aims to tackle this important question applies our solution to a water allocation problem in Tucson, AZ. This would help authorities with future water plans in this area of Tucson while being robust to uncertainties in water supplies and demands.

Our approach uses a generalized network model of Colorado River water allocation in Tucson, Arizona motivated by the CALVIN (CALifornia Value Integrated Network) optimal water allocation model of California created by Draper et al. [6]. More general models of Colorado River water distribution have also been studied, such as the Colorado River Reservoir Model [3] and the Colorado River Budget Model [1]. Our model is modified to incorporate ambiguous future uncertainty by using the Likelihood Robust Optimization (LRO) approach of Wang, Glynn and Ye [9].

LRO is a data-driven method that uses observations of random or unknown parameters to account not only for inherent stochasticity of a problem, but the uncertainty of the probabilistic model itself. This is especially important in our application as we are looking 40 years into the future and there is considerable ambiguity in the uncertainties. The LRO is especially attractive because only those scenarios of interest, obtained either through observation or simulation, are used directly in the calculations. The size of the problem is polynomial in the number of scenarios used, making it computationally tractable. Furthermore, the scenarios used in the LRO can represent select ones that are of interest to the authorities. For instance, the US Bureau of Reclamation uses an approach called the scenario planning that examines water demand and supply scenarios over the next 50 years [8]. Similar studies are being conducted by Tucson Water [4].

In this report, we first discuss the methods used in Section 2, and point out some key results in Section 3. Any figures and tables mentioned are included in the appendices. We end with a short summary and future work.
2. Methods

2.1. Generalized Network Water Model. We begin with a multi-period generalized network flow model of Colorado River water allocation in southeastern Tucson, shown in Figure 1. The generalized network model is defined by a set of nodes and directed arcs \((N, A)\). The nodes \(N\) represent available water supply from the Colorado River, water treatment plants, reservoirs, and water demand sites. The arcs \(A\) represent the conveyance system (pipes, etc.) that carry water between the nodes. Figures 2, 3, and 4 show the layout of the water distribution model. Water can be stored in between time periods in reservoirs to meet future demands. The model aims to find the minimal cost water flows considering energy, treatment, storage, and transportation costs over the planning period. In addition, the model contains penalty costs for unmet demands. Generalized network water allocation models have been used to find water allocations and delivery reliabilities and to assess values of different water use operations; see, e.g., [6].

Water flows on arc \((i, j)\) during time period \(t = 1, \ldots, P\) are represented by decisions \(x_{ijt}\). Each arc \((i, j)\) and time period \(t\) has a unit cost \(c_{ijt}\), loss coefficient \(0 \leq a_{ijt} \leq 1\) to account for evaporation, leakage from the pipes, etc., and bounds on the flow \(l_{ijt} \leq x_{ijt} \leq u_{ijt}\). Each node \(j \in N\) has a supply/demand for time period \(t\), denoted by \(b_{jt}\). Nodes representing reservoirs are able to store water between time periods. Stored water available at node \(j\) at the beginning of time period \(t\) is \(s_{jt}\), with associated storage cost \(c_s\) and bounds \(l_s \leq s_{jt} \leq u_s\). Finally, water released into the environment from node \(j\) in period \(t\) is given by \(r_{jt}\), with bounds \(l_r \leq r_{jt} \leq u_r\). The deterministic model is a multi-period generalized network flow model of the form

\[
\min_{x, s, r} \sum_{(i,j) \in A} \sum_{t=1}^{P} c_{ijt}^x x_{ijt} + \sum_{j \in N} \sum_{t=1}^{P} c_{jt}^s s_{jt}
\]

\[
\text{s.t. } \sum_{i \in N} x_{ijt} + s_{jt+1} + r_{jt} = \sum_{i \in N} a_{ijt} x_{ijt} + s_{jt} + b_{jt}, \quad \forall j, t
\]

\[
l_{ijt}^x \leq x_{ijt} \leq u_{ijt}^x, \quad \forall i, j, t
\]

\[
l_s^r \leq s_{jt} \leq u_s^r, \quad \forall j, t
\]

\[
l_r^r \leq r_{jt} \leq u_r^r, \quad \forall j, t.
\]

In practice, many parameters are unknown, especially as we further look into the future. This is particularly true of the water supply and demands, denoted by \(b_{jt}\) in the above model. To capture this uncertainty, we first make this model stochastic by rewriting it as a two-stage linear program. Later, we will modify this model to be robust against all future scenarios considered in the next subsection. The model has a total of \(P\) time periods, which is split into two stages of \(P_1\) and \(P_2 = P - P_1\) years each. The water supplies and demands \(b_{jt}\) are assumed to be known during the first stage of \(P_1\) years and in the remaining \(P_2\) years, they are uncertain and modeled using a discrete set of scenarios with a specific probability distribution. Note that we also assume some lower and upper bounds to be stochastic in the second stage.

The two-stage stochastic model allocates water at a minimal cost in the first stage and in the second stage, it tries to minimize the total expected cost given the first-stage decisions.
The model thus becomes

\[ \min_{(x,s,r) \in L^1} \sum_{(i,j) \in A} c^r_{ij} x_{ijt} + \sum_{j \in N} \sum_{t=1}^{P_1} c^s_{jt} s_{jt} + \sum_{\omega=1}^{n} p_{\omega} h^1_{\omega}(s) \]

s.t. \[ \sum_{i \in N} x_{jit} + s_{j,t+1} + r_{jt} = \sum_{i \in N} a_{ijt} x_{ijt} + s_{jt} + b_{jt}, \quad \forall j, 1 \leq t \leq P_1, \]

where the \( \omega \) index indicates the \( n \) second-stage scenarios with probabilities \( p_{\omega} \), the last term in (1) denotes the expected second-stage costs, which are found by solving the following problems over each scenario \( \omega \)

\[ h^1_{\omega}(s) = \min_{(x,s,r) \in L^2} \sum_{(i,j) \in A} c^r_{ij} x_{ijt} + \sum_{j \in N} \sum_{t=P_1+1}^{P} c^s_{jt} s_{jt} \]

s.t. \[ \sum_{i \in N} x_{jit} + s_{j,t+1} + r_{jt} = \sum_{i \in N} a_{ijt} x_{ijt} + s_{jt} + b^\omega_{jt}, \quad \forall j, P_1 + 1 \leq t \leq P, \]

and \( L^1 \) and \( L^2 \) represent the feasible regions defined by the lower and upper variable bounds.

In the rest of the report, we simplify the notation for the first-stage (1) and second-stage (2) problems as follows. In the first stage, decision variables \( \{x_{ijt}\}, \{s_{jt}\} \) and \( \{r_{jt}\} \) become the vector \( x \), costs \( \{c^r_{ijt}\} \) and \( \{c^s_{jt}\} \) are written as the row vector \( c \), the supply/demand parameters \( b_{jt} \) become the vector \( b \) and the constraint matrix is written as \( A \). In the second stage, we denote the decisions as \( y^\omega \), the costs as \( q^\omega \), the supply/demands as \( d^\omega \), and the constraint matrices multiplying \( y^\omega \) and \( x \) as \( D^\omega \) and \( B^\omega \), respectively. This notation puts the generalized network model in the form of a two-stage stochastic linear program with recourse (SLP-2), formulated as

\[ \min_{x} cx + \sum_{\omega=1}^{n} p_{\omega} h^1(x) \]

s.t. \[ Ax = b \]
\[ x \geq 0 \]

where

\[ h^1_{\omega}(x) = \min_{y^\omega} q^\omega y^\omega \]

s.t. \[ D^\omega y^\omega = B^\omega x + d^\omega \]
\[ y^\omega \geq 0. \]

For simplicity we assume relatively complete recourse; i.e., the second-stage problems \( h^1_{\omega}(x) \) are feasible for every feasible solution \( x \) of the first-stage problem. In our application, there are penalty costs when demand is not met; therefore, our model has relatively complete recourse.

2.2. Robust Water Allocation Formulation. The two-stage stochastic formulation, compactly written as SLP-2 above, assumes that the probability of each scenario \( \{p_{\omega}\}_{\omega=1}^{n} \) is known. However, in many applications, including our water planning, these probabilities are unknown; in other words, the governing probability distribution is unknown. One technique to deal with this type of uncertainty is to replace the assumed known distribution with an ambiguity set of distributions; i.e., a set of distributions which is believed to contain the
true distribution and to be robust against this set. In the likelihood robust formulation, we assume scenario $\omega$ has been observed $N_\omega$ times, with $N = \sum_{\omega=1}^n N_\omega$ total observations. Note that these observations could represent expert opinions, different simulations of the future, etc. If not much is known, each scenario can be taken to be equally likely ($N_\omega = 1$ or equal for all $\omega$) and the robustness level can be adjusted. In the stochastic formulation SLP-2, having $N_\omega$ observations for each scenario $\omega$ would correspond to probability of scenario $\omega$ to be $p_\omega = N_\omega/N$, which is the maximum likelihood distribution. The ambiguity set, however, is defined by the set of distributions with sufficiently high empirical likelihood $\prod_{\omega=1}^n p_\omega^{N_\omega}$. By replacing the specific distribution in SLP-2 with a set of distributions with high empirical likelihood, we create a model that we refer to as two-stage likelihood robust linear program with recourse (LRLP-2).

To derive the LRLP-2, we begin by writing SLP-2 given in (3)–(4) in extensive form by putting the first and second stages together

$$
\begin{align*}
\min_{x,y,\omega} & \quad cx + \sum_{\omega} p_\omega q_\omega y_\omega \\
\text{s.t.} & \quad Ax = b \\
& \quad -B_\omega x + D_\omega y_\omega = d_\omega, \; \forall \omega \\
& \quad x \geq 0 \\
& \quad y_\omega \geq 0, \; \forall \omega.
\end{align*}
$$

The SLP-2 formulation is then augmented by the set of distributions with sufficiently high likelihood. To be robust against all these possible distributions, the distribution that results in the maximum expected cost is considered. Then, the objective function is minimized with respect to this worst-case distribution selected from the ambiguity set of distributions. The resulting minimax formulation of LRLP-2 is

$$
\begin{align*}
\min_{x,y,\omega} & \quad \max_p \left( cx + \sum_{\omega} p_\omega q_\omega y_\omega \right) \\
\text{s.t.} & \quad Ax = b \\
& \quad -B_\omega x + D_\omega y_\omega = d_\omega, \; \forall \omega \\
& \quad \sum_{\omega} N_\omega \log p_\omega \geq \gamma \\
& \quad \sum_{\omega} p_\omega = 1 \\
& \quad x \geq 0 \\
& \quad y_\omega, p_\omega \geq 0, \; \forall \omega.
\end{align*}
$$

Following Wang et al. [9], we have introduced the likelihood parameter $\gamma$, and used it to construct the ambiguity set of distributions $\{p_\omega\}_{\omega=1}^n$ satisfying constraints (6), (7), and (8). Note the likelihood constraint (6) is equivalent to $\prod_{\omega=1}^n p_\omega^{N_\omega} \geq e^\gamma$, which explicitly states that the empirical likelihood should be above a certain desired level dictated by $e^\gamma$. Constraint (7), along with nonnegativity constraints on $p_\omega$ given in (8), simply ensure that $\{p_\omega\}_{\omega=1}^n$ constitutes a probability distribution. Let $0 \leq \gamma' \leq 1$ be the relative likelihood parameter that expresses $\gamma$ as a proportion of the maximum likelihood; i.e., $\gamma = \log(\gamma' \prod_{\omega} (N_\omega/N_\omega^{N_\omega})).$

When discussing our application and key results in the next section, we will be referring to $\gamma'$.
3. Key Findings

3.1. Application. Using the generalized network model (1) of Colorado River water allocation in the southeastern portion of Tucson, we created a likelihood robust water allocation problem of the form LRLP-2. The model has a total of $P = 41$ time periods, representing years 2010–2050. For each time period, the network has 62 nodes representing demand for potable and nonpotable (reclaimed) water, pumps, water treatment plants, and the available water supply from the Colorado River. The network in each time period has 102 arcs, representing the pipe network carrying the water between the nodes physically and connecting the network to the five reservoirs that connect the time stages in the model. We use $P_1 = 5$ time periods for the first stage. Uncertainty in the second stage takes the form of uncertain population (thus, demand for water) and supply of water.

There are four scenarios considered in this model: (i) high population, high supply, (ii) high population, low supply, (iii) low population, high supply, and (iv) low population, low supply. Specific numbers for the population and water availability are given in Table 1 in Appendix A. Each scenario is assumed to have one observation. The high population scenarios are more costly as the system needs to meet demand or pay for unmet demand. The low population scenarios, on the other hand, are not as costly.

3.2. Results. Figure 6 shows how the worst-case distribution changes with $\gamma'$. When $\gamma'$ is close to 1, the ambiguity set is very small and we essentially use the maximum likelihood distribution, which has equal $\frac{1}{4}$ probabilities on each of the four scenarios. As $\gamma'$ is decreased, the size of the ambiguity set increases and the worst-case distribution used by LRLP-2 changes. For values of $\gamma'$ larger than about 0.15, the worst-case distribution assigns probability larger than $\frac{1}{4}$ to each of the high population scenarios, and probability smaller than $\frac{1}{4}$ to the low population scenarios. As $\gamma'$ becomes smaller, the worst-case distribution continues moving toward the high population-low supply scenario, the costliest of the four. We can see that as $\gamma'$ becomes zero, the worst-case distribution will focus only on this most expensive scenario. A closer look at the optimal solutions reveals that as $\gamma'$ is decreased, or as robustness is increased, the solution uses more and more reclaimed water (represented by the purple color in Figures 2–5) in an effort to meet demands in a least-costly way.

It is natural to ask what value of $\gamma'$ should be picked in the robust formulation. We can determine this because the likelihood ambiguity region is related to the Kullback-Leibler (KL) divergence, a well studied concept in statistics. We use the fact that, under an appropriate normalization, the KL divergence has a $\chi^2$ distribution asymptotically. Using this fact, we find that choosing $\gamma' = \exp(-\chi^2_{n-1,1-\alpha}/2)$ for a problem with $n$ scenarios will result in an ambiguity region that contains the true distribution with probability $1 - \alpha$. Setting $n = 4$ and $\alpha = 0.05$ yields $\gamma' \approx 0.02$.

The Tucson water model was originally designed to help determine the usefulness of installing decentralized water treatment facilities in the southeastern area. One possible change is illustrated in Figure 5, where a decentralized wastewater treatment plant has been built in elevation zone GS. We investigated the cost savings of installing an additional decentralized facility, as shown in Figure 7. The facility itself has an estimated cost of $40$ million, indicated by the black line in Figure 7, which needs to be exceeded by the 41-year projected cost savings. Our results, using the recommended ambiguity level given by $\gamma' = 0.02$, show that the additional plant would generate savings over its initial cost over the time period of the model.
4. Conclusion and Future Work

Our major contributions in this work are two-fold. First, we proposed an extension of the Likelihood Robust Optimization (LRO) method of Wang et al. [9] to general two-stage stochastic programs with recourse, creating a two-stage likelihood robust program with recourse, denoted LRLP-2. These new models use the empirical likelihood function to define an ambiguity set of probability distributions using observed data and optimize the worst-case expected cost with respect to this likelihood ambiguity set. Secondly, we have applied this method to plan future water distribution in southeast area of Tucson, Arizona. Our results indicate that (i) as we aim to be more robust, we need to use more reclaimed water, and (ii) a decentralized wastewater treatment plant in the considered area could result in an overall cost saving over the 40 years.

Our future work includes the following. We plan to augment the existing model first with a richer set of scenarios. In addition to more varied estimates for future population, we will integrate climate change predictions into the model to generate scenarios for future water demand and supply from the Colorado River. Another important extension is to augment the water allocation model with a facility location problem to determine the best places for an additional waste water treatment plants to increase the use of reclaimed water in the most cost-efficient manner.

Acknowledgements. We are grateful to the University of Arizona’s Technology and Research Initiative Fund and the Water Sustainability Program. The data of the model considered in this work have been generated by many participants. In particular, we are grateful to Tucson Water and Alicia Forrester and Dr. Kevin Lansey of Civil Engineering and Engineering Mechanics for sharing their data. This work has also been partially supported by the National Science Foundation through grant CMMI-1151226.

References

This appendix provides a graphical representation of the water model of the southeastern part of Tucson. It also lists the population/CAP supply scenarios used in the model.

**Figure 1.** The southeastern region of Tucson considered in the water model, marked with a purple border.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Population 2010</th>
<th>Population 2050</th>
<th>Supply (CAP/yr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>33,304</td>
<td>371,804</td>
<td>130,000</td>
</tr>
<tr>
<td>High</td>
<td>40,705</td>
<td>762,427</td>
<td>140,000</td>
</tr>
</tbody>
</table>

**Table 1.** Population of the southeastern region shown in Figure 1 in the low and high population estimates, along with the low and high estimates for amount of CAP water available per year. Note that the estimates for population and CAP availability are independent.
Figure 2. (This figure is continued in Figures 3 and 4, and Figure 5 provides a version with a decentralized treatment plant.) Network layout of the centralized water treatment system in Tucson. Nodes represent treatment facilities, reservoirs, etc. and arcs (arrows) represent transportation systems between nodes.

Color coding: blue represents potable water, purple shows nonpotable reclaimed water, and black shows wastewater.

Figure 3. Network layout of demand in elevation zones C–E of the southeastern region shown in Figure 1.
Figure 4. Network layout of demand in elevation zones F–H of the southeastern region shown in Figure 1. Highway I-10 splits these zones into north and south sections, with no water infrastructure crossing the highway.

*N = North of I-10  *S = South of I-10

Figure 5. Layout of south elevation zones F–H with a decentralized wastewater treatment plant in zone GS.
Appendix B. Figures for key findings

The figures in this appendix provide a summary of the key findings.

**Figure 6.** The probability of each of the four scenarios of the worst-case distribution for different values of the likelihood parameter $\gamma'$.

**Figure 7.** Total savings on operating costs over 40 years from installing an additional wastewater treatment plant, depicted in Figure 5. The cost of the plant is $40 million, indicated by the black line. At the recommended likelihood parameter of $\gamma' = 0.02$, indicated by the red line, the additional wastewater treatment plant results in approximately $1$ million cost savings.