Data-Driven Water Allocation under Climate Uncertainty: A Distributionally Robust Approach

David K. Love,[1] Jangho Park,[2] Güzin Bayraksan[2]

[1]American Express, New York, NY, USA.
[2]Department of Integrated Systems Engineering, Ohio State University, Columbus, Ohio, USA.

Abstract

This paper investigates the application of techniques from distributionally robust optimization (DRO) to water allocation under future uncertainty. Specifically, we look at a rapidly-developing area of Tucson, Arizona. Tucson, like many arid and semi-arid regions around the world, faces considerable uncertainty in its ability to provide water for its citizens in the future. The main sources of uncertainty in the Tucson region include (1) the unpredictable future population growth, (2) the availability of water from the Colorado River in light of competing claims from other states and municipalities, and (3) the effects of climate variability and how this relates to water consumption. This paper presents a new data-driven approach for integrating forecasts for all these sources of uncertainty in a single optimization model for robust and sustainable water allocation. We use this model to analyze the value of constructing additional treatment facilities to reduce future water shortages. The results indicate that DRO can provide water resource managers important insights to minimize their risks and, in revealing critical uncertainties in their systems, plan for the future.

1 Introduction

More than 60% of the water in Tucson is provided by the Colorado River. Without this water source, citizens of Tucson—as well as millions in California, Nevada, and Mexico—are threatened. The Colorado River has been facing extreme water shortages in recent years. In 2015 and 2016, Lake Mead water elevation hit back-to-back record low, and, in June 2016, it reached its lowest level of 1071.64 feet for the first time in its 80-year history [*U.S. Bureau of Reclamation*, 2016]. As the Colorado River runs dry, the imbalance between supply and demand widens. And, as climate variability threatens the Colorado River availability [*Udall and Overpeck*, 2017], it is imperative to sustainably manage this water resource by taking into account the many complex uncertainties it faces.

This paper presents a new modeling approach for sustainable water allocation that handles uncertainties from various sources in a robust manner. This approach is called the *Distributionally Robust Optimization* (DRO) with ϕ-divergences, and it has been recently proposed by *Ben-Tal et al.* [2013]; see also further investigations [*Bayraksan and Love*, 2015; *Jiang and Guan*, 2016]. DRO acknowledges that uncertainties—like the long-term, multi-period, and complex ones considered in this paper on climate, population, and the Colorado River basin's hydrology—are not known fully. However, there is often historical data, sophisticated simulations, and detailed forecasts available from government and utility sources. So, it is possible to build approximate future scenarios with an approximate distribution for these uncertainties. DRO then considers all distributions that are sufficiently close to this "nominal" distribution and optimizes a worst-case expected objective based on the all the distributions considered. The appeal of DRO is that it is more realistic because it explicitly considers existing data and forecasts, while acknowledging that these forecasts and simulations may contain prediction errors.

Corresponding author: Güzin Bayraksan, bayraksan.1@osu.edu

Given the many uncertainties in water resources problems, Stochastic Optimization (SO) models have long been used [*Hall and Howell*, 1970; *Escudero*, 2000; *Higgins et al.*, 2008; *Rosenberg and Lund*, 2009]. While there are many SO models and methods, e.g. [*Lee and Labadie*, 2007; *Singh*, 2012; *Vicuna et al.*, 2010], the "classical" SO approach considers (i) a number of future scenarios (or infinitely many scenarios in the case of continuous distributions) and assumes (ii) a *known* probability distribution on these uncertainties. SO then (iii) optimizes the *expected* costs or benefits. The shortcomings of this approach are multifold. First, in real-world applications, the probabilities and distributions are rarely known. If an incorrect distribution is used, it may lead to significantly suboptimal decisions. Also, the risk-neutral 'expected' objective function is rarely appropriate for these problems. Many water managers are risk averse because they would like to avoid the large consequences of rare but powerful events like extreme droughts. SO can also become computationally burdensome for large-scale problems.

Given the above deficiencies with classical SO, several researchers used different modeling approaches. To overcome the issue of *unknown* probability distributions, for instance, interval programming—which develops bounds on decision variables for given bounds in uncertain parameters—have been used [*Li and Huang*, 2008; *Li et al.*, 2009]. Another popular approach is Robust Optimization (RO). Classical RO (i) assumes the uncertain parameters lie in a set (called the *uncertainty set*) but (ii) ignores any information on the likelihood of realizations. It then (iii) optimizes a worst-case objective from the uncertainty set. Polyhedral [*Chung et al.*, 2009], ellipsoid [*Housh et al.*, 2011; *Perelman et al.*, 2013], or convex hull of some realizations [*Lan et al.*, 2015] have been used as uncertainty sets for water problems. The pitfall with RO is that, in reality, more is known about the uncertainties than some set it lies in. And, RO can lead to overly conservative solutions because the worst-case among *all* distributions that lie in that uncertainty set are considered. This results in very robust, but also very costly decisions.

DRO lies between the classical SO and RO approaches. In fact, SO and RO are two special cases of DRO. To see this, recall that DRO considers a "nominal" distribution given the current data. Because the nominal distribution may not be correct, it considers a set that includes all distributions sufficiently close to the nominal distribution. This is called the *ambiguity set of distributions*. If this set contains *only one* distribution, then DRO is equivalent to the classical SO. Clearly, the danger is that this distribution might not be the correct one. If, on the other hand, the ambiguity set contains *all* distributions with, e.g., the same support, then DRO is equivalent to the classical RO. RO typically considers too many distributions; so we obtain very costly solutions. By adjusting the ambiguity set with respect to the available data (see Section 3 for details), DRO becomes more appealing than both SO and RO.

Controlling risk is crucial for long-term water decisions. SO uses unrealistic risk-neutral expected-value objectives, and RO is extremely risk averse. To appropriately address the issue of risk averseness of water managers, probabilistic constraints [*Jia and Culver*, 2006] or objective terms [*Borgomeo et al.*, 2016] have been added. Alternatively, the objective has been changed to a mean-variance functional [*Watkins Jr and McKinney*, 1997; *Mulvey et al.*, 1995]. These models are sometimes referred to as "probabilistic robust" [*Pan et al.*, 2015]. Probabilistic constraints and objectives often lead to nonconvex optimization problems that are difficult to solve. DRO with ϕ-divergences, on the other hand, results in convex optimization problems that can be efficiently solved. Using the Mean-Variance risk measure can also be problematic. Because mean-variance may lack desired properties like monotonicity, it is not a coherent risk measure. As a result, it might lead to irrational decisions; see, e.g., Example 6.38 of *Shapiro et al.* [2009] for what can go wrong with such an objective.

Coherent risk measures—a concept pioneered by *Artzner et al.* [1999]—have nice properties that make them amenable to optimization and provide rational solutions. The popular coherent risk measure Conditional-Value-at-Risk has been used for water allocation problems [*Shao et al.*, 2011; *Zhang et al.*, 2016]. Under mild conditions (e.g., real-valued costs, convex closed ambiguity set of distributions), DRO is equivalent to optimizing a coherent risk measure. Our setting satisfies this equivalence relation. Consequently, our DRO approach automatically induces risk-averse behavior in water allocation, helping water managers with their sustainability goals.

To the best of our knowledge, this is the first paper to apply DRO with ϕ-divergences to urban water systems for long-term sustainability. Close to our approach are the models of *Pan et al.* [2015]; *Gauvin et al.* [2017], where inflow distributions with same support, mean, and covariance matrix are considered. They then use Decision Rules (DR) to approximately solve this problem. Our approach differs in several important aspects. First, our model uses significantly more complicated uncertainties—some developed by experts in their domains such as hydrologists and climate scientists—rather than just inflows. Thus, the data of our model is substantially more realistic. Second, DR are often suboptimal because they consider solutions of only one type. Our model, on the other hand, can be solved efficiently to optimality. Finally, we focus on urban water systems, not energy generation. We note that the models in *Pan et al.* [2015]; *Gauvin et al.* [2017]; *Zhang et al.* [2016] are multistage, while ours is two-stage (even though each stage consists of several years). Extensions of this work to the dynamic multistage setting is ongoing. This and other future/ongoing work will be discussed in Section 7.

One of the unique features of our model is that it combines various sources of data into a single integrated projection model. We incorporate bias corrected and spatially downscaled global circulation climate models (formed via different organizations around the world), greenhouse concentration paths (as adopted by the Intergovernmental Panel on Climate Change (IPCC)), population forecasts (developed by governing agencies in the area), water-use trends, as well as hydrological simulations of the Colorado River (conducted by the U.S. Bureau of Reclamation). We will explain our data-driven methodology in Section 4. It is important to emphasize that DRO can be formed in various ways—e.g., other climate models can be easily added, data can be quickly updated, etc.—and the uncertain data put into the optimization model can be changed readily.

Climate is one of the most important sources of uncertainty for long-term sustainability of water resources. Extensive research analyze the sensitivity of mitigation plans to uncertainties in climate, e.g., [*Brown et al.*, 2012; *Singh et al.*, 2014; *Harou et al.*, 2010; *O'Hara and Georgakakos*, 2008]. Some of this research provide a feedback to optimization, e.g. [*Kasprzyk et al.*, 2013; *Borgomeo et al.*, 2016], and some do not, e.g. [*Yan et al.*, 2016]. The CALVIN (CALifornia Value Integrated Network) model—a multi-period generalized network flow model like our underlying model—has been used to evaluate different climate scenarios for California's water system. Some of these studies use only few climate scenarios. For instance, *Tanaka et al.* [2006] consider one dry (run B06.06 from the PCM model) and one wet (HadCM2) scenario, and *Connell-Buck et al.* [2011] consider two warm-dry scenarios based on the GFDL-CM3 climate model. *Vicuna et al.* [2010], similar to this paper, consider many global circulation models and acknowledge the uncertainty of the climate scenarios by embedding it into optimization. None of the existing work, however, explicitly considers the *ambiguities* in these uncertainties like the DRO model. We believe this is critically important for models that have ambiguous uncertainties—like those that incorporate climate models.

Apart from the (mostly uncontrollable) natural environment, water supplies can be increased by *reusing* water. For many arid and semi-arid areas—like our study area—reclaimed water (treated wastewater) is the only remaining water source [*Woods et al.*, 2012; *Lan et al.*, 2016]. We consider constructing decentralized water treatment facilities to increase water reuse. New infrastructures cost hundreds of millions of dollars, and they should be evaluated carefully. We use our DRO model to investigate two mitigation strategies: The first strategy uses treated wastewater for nonpotable needs, saving freshwater for potable demands. The second alternative considers Indirect Potable Reuse (IPR)—treating the wastewater to a very high quality and blending with other potable sources.

In summary, this paper presents a first distributionally robust approach for sustainable water allocation in urban water systems. It applies this model to allocate Colorado River water through mid-century to a developing portion of Tucson, incorporating various ambiguous uncertainties on climate, population, water-use trends, and the Colorado River water availability. This DRO model is then used to assess water reuse strategies by evaluating the value of constructing additional water treatment facilities. Our results show that these facilities can significantly reduce water shortages, but their economic value depends on the cost of water shortage. This analysis prompts further investigations on long-term health consequences and social acceptance of IPR, especially as water shortages become more frequent and costly.

We believe DRO has the potential to be very important in water resources problems. It is expected to be most beneficial for problems with (i) ambiguous uncertainties (e.g., climate, hydrological variability), (ii) risk aversion (e.g., avoid water shortages as much as possible), and (iii) robust decision making (e.g., be able to provide water even under extreme droughts). While this paper applies DRO within a two-stage stochastic linear optimization context, given the explosion of DRO models and methods in the operations research literature in recent years, DRO has the potential to be applied to many important and distinct water resources problems within chance-constrained, nonlinear, and dynamic multi-period contexts, e.g., [*Jiang and Guan*, 2016; *Philpott et al.*, 2017; *Rahimian et al.*, 2018].

The remainder of this paper is organized as follows. Section 2 introduces the water allocation model, and Section 3 describes the DRO approach with ϕ-divergences. Section 4 describes the application to Tucson, including the data-driven method used in this application. Section 5 presents the results, and Section 6 discusses the water allocation and infrastructure recommendations as a consequence of this study. We end in Section 7 with a summary and future research directions. An earlier version of this paper appeared in *Love and Bayraksan* [2013]. That paper only considers one ϕ-divergence, whereas this paper presents a general model and tests it on several ϕ-divergences. Furthermore, that paper has only four generic scenarios. The scenarios here are significantly more complicated. They are realistic and based a variety of uncertain factors. Finally, the analysis here is substantially more involved. For instance, additional infrastructures, shortage levels, and price of data are examined.

2 Water Allocation Problem

We first introduce the water allocation problem as a two-stage stochastic program because our DRO model is constructed on this type of model. (Text S1 of the supporting information *drwa-si.pdf* presents a detailed DRO formulation of our application.) In two-stage stochastic programs, the first-stage decisions are made before knowing what will happen later. In the second stage, the uncertainty becomes known, and decisions for each scenario is determined. These are called the second-stage (or recourse) decisions. The uncertain future is modeled by a *known* probability distribution. We will change this later.

Even though our problem is two-staged, each stage contains several time periods. For instance, our application has yearly time periods till 2050. In general, we assume the model has P total time periods with P_1 periods in the first stage and $P - P_1$ periods in the second stage. The water system is represented as a network $(\mathcal{N}, \mathcal{A})$, where $\mathcal{N}$ denotes the set of nodes and $\mathcal{A}$ denotes the set of arcs. The general properties of the model are as follows.

Nodes are categorized into four:

1. **Supply nodes:** e.g., surface water, groundwater, and a node for external water purchased in case of water shortage;
2. **Demand nodes:** e.g., potable/nonpotable users;
3. **Intermediate nodes:** e.g., water treatment plant interconnection points, potable/nonpotable interceptors, return interceptors;
4. **Infrastructures:** e.g., wastewater treatment plant.

Arcs represent the pipe network and canals carrying the water between nodes.

Constraints can be grouped into three:

1. **Balance/Satisfaction Constraints.** (a) **Flow Balance:** At pumps/water treatment plants /reservoir/interconnection points, water flow should be balanced (these include any water loss through the pipes or in treatment) (b) **Demand Satisfaction:** Meet users' demands (can be satisfied by expensive external supply if in shortage) (c) **Storage Balance:** At aquifer

recharge facilities, storage/inventory constraints tie different time periods (infiltration into ground needs a one-year lag).

2. **Water-Reuse Constraints.** A portion of the potable water used is returned to a wastewater treatment plant. Treated wastewater (not IPR) can only be used for nonpotable demands in later years.
3. **Capacity Constraints.** (a) **On Nodes:** Water supply availability depending on the scenario; capacities on the inflow/outflow of recharge facilities and treatment plants based on physical limitations. (b) **On Arcs:** Bounds on the water flows based on physical characteristics of the pipe network.

The model determines the water flows between the nodes and storage levels at reservoirs that minimize the expected total cost of conveyance, storage, and water shortages.

The water flows on arc $(i, j) \in \mathcal{A}$ during time period $t = 1, \ldots, P$ are represented by decisions x_{ijt}. They have conveyance costs c^x_{ijt} and bounds on flow $l^x_{ijt} \le x_{ijt} \le u^x_{ijt}$. The loss coefficient $0 \le a_{ijt} \le 1$ accounts for evaporation and leakage from the pipes. Each node $j \in \mathcal{N}$ has a supply/demand for period t, b_{jt} (for flow balance, $b_{jt} = 0$). Nodes capable of storing water are members of the set S. Stored water available at node $j \in S$ at the end of period t is s_{jt}, with storage cost c^s_{jt} and bounds on storage $l^s_{jt} \le s_{jt} \le u^s_{jt}$. Finally, water released into the environment from node j in period t is given by r_{jt}, with bounds on release $l^r_{jt} \le r_{jt} \le u^r_{jt}$. We assume initial storage levels s_{j0} are given and $x_{ij0} \equiv 0$.

In the general model, the water supplies and demands are uncertain, as well as available water treatment capacities (as an effect of relative population growth). The uncertainty is modeled through a finite number of scenarios $\omega \in \Omega$. Without loss of generality, we set $\Omega = \{1, 2, \ldots, n\}$ so that there are n distinct scenarios. Below, $\{q_\omega\}^n_{\omega=1}$ represent the "assumed-known" probability of each scenario ω. To differentiate the second-stage decisions and parameters that depend on a particular scenario, we use superscript ω: e.g., x^ω_{ijt}, s^ω_{jt}, b^ω_{it}. The second-stage constraints are similar, but differ with scenario.

The two-stage stochastic model is

$$
\begin{aligned}
\min_{\substack{x_{ijt}, s_{jt}, r_{jt}, \\ \forall i,j,t}} \quad & \sum_{(i,j)\in\mathcal{A}} \sum_{t=1}^{P_1} c^x_{ijt} x_{ijt} + \sum_{j\in\mathbf{N}} \sum_{t=1}^{P_1} c^s_{jt} s_{jt} + \sum_{\omega=1}^{n} q_\omega h_\omega(\mathbf{s}) \qquad (1)\\
\text{s.t.} \quad & \sum_{j:(j,i)\in A} a_{jit} x_{jit} = \sum_{j:(i,j)\in\mathcal{A}} x_{ijt} + r_{it} + b_{it} \quad \forall i \notin S, 1 \le t \le P_1 \\
& \sum_{j:(j,i)\in\mathcal{A}} a_{ji,t-1} x_{ji,t-1} + s_{i,t-1} = \sum_{j:(i,j)\in\mathcal{A}} x_{ijt} + s_{it}, \quad \forall i \in S, 1 \le t \le P_1, \\
& l^x_{ijt} \le x_{ijt} \le u^x_{ijt},\ l^s_{jt} \le s_{jt} \le u^s_{sjt},\ l^r_{jt} \le r_{jt} \le u^r_{sjt},\ \forall i, j, t,
\end{aligned}
$$

where $\mathbf{s} = \{s_{jt},\ \forall j, t\}$ and $h_\omega(\mathbf{s})$ denotes the optimal second-stage cost for scenario ω as

$$
\begin{aligned}
h_\omega(\mathbf{s}) = \min_{\substack{x^\omega_{ijt}, s^\omega_{jt}, r^\omega_{jt}, \\ \forall i,j,t}} \quad & \sum_{(i,j)\in\mathcal{A}} \sum_{t=P_1+1}^{P} c^x_{ijt} x^\omega_{ijt} + \sum_{j\in\mathcal{N}} \sum_{t=P_1+1}^{P} c^s_{jt} s^\omega_{jt} \qquad (2)\\
\text{s.t.} \quad & \sum_{j:(j,i)\in\mathcal{A}} a_{jit} x^\omega_{jit} = \sum_{j:(i,j)\in\mathcal{A}} x^\omega_{ijt} + r^\omega_{it} + b^\omega_{it},\ \forall i \notin S, P_1 + 1 \le t \le P, \\
& \sum_{j:(j,i)\in\mathcal{A}} a_{ji,t-1} x^\omega_{ji,t-1} + s_{i,t-1} = \sum_{j:(i,j)\in\mathcal{A}} x^\omega_{ijt} + s^\omega_{it},\ \forall i \in S, t = P_1 + 1, \\
& \sum_{j:(j,i)\in\mathcal{A}} a_{ji,t-1} x^\omega_{ji,t-1} + s^\omega_{i,t-1} = \sum_{j:(i,j)\in\mathcal{A}} x^\omega_{ijt} + s^\omega_{it},\ \forall i \in S, P_1 < t \le P, \\
& l^{x,\omega}_{ijt} \le x^\omega_{ijt} \le u^{x,\omega}_{ijt},\ l^{s,\omega}_{jt} \le s^\omega_{jt} \le u^{s,\omega}_{sjt},\ l^{r,\omega}_{jt} \le r^\omega_{jt} \le u^{r,\omega}_{sjt},\ \forall i, j, t.
\end{aligned}
$$

Although the stochastic model acknowledged the existence of uncertainties in the problem, it introduced a distribution through $q_1, \ldots, q_n$. These probabilities need to be estimated and might

contain estimation errors. DRO helps to counteract this problem. In DRO, $\mathbf{q} = (q_1, \ldots, q_n)^T$ becomes the "nominal" distribution that is extracted directly from the data, and the optimization takes place over a set of distributions that are "similar" to the nominal distribution. The advantage is that both the nominal distribution and the level of similarity are extracted from the data. Thus, there is no need to assume the distribution is known.

Before we introduce DRO, to ease subsequent presentation, let us rewrite the stochastic problem in a compact form. The first- and second-stage problems (1)–(2) become

$$\min_{\mathbf{x}\in X} \left\{ \mathbf{cx} + \sum_{\omega=1}^{n} q_\omega h_\omega(\mathbf{x}) \right\}, \tag{3}$$

where $X = \{\mathbf{x} : \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq 0\}$, and $h_\omega(\mathbf{x}) = \min_{\mathbf{y}^\omega} \{\mathbf{k}^\omega \mathbf{y}^\omega : \mathbf{D}^\omega \mathbf{y}^\omega = \mathbf{B}^\omega \mathbf{x} + \mathbf{d}^\omega, \mathbf{y}^\omega \geq 0\}$. In the first stage, $\mathbf{x}$ represents the decision variables $\{x_{ijt}\}$, $\{s_{jt}\}$ and $\{r_{jt}\}$, $\mathbf{c}$ denotes the costs $\{c^x_{ijt}\}$, the supply/demand parameters $\{b_{jt}\}$ become the vector $\mathbf{b}$, and the constraint matrix is written as $\mathbf{A}$. To clearly differentiate from the first stage, in the second stage, $\mathbf{y}^\omega$ denotes the decisions, $\mathbf{k}^\omega$ represents the costs, $\mathbf{d}^\omega$ the supply/demands, and $\mathbf{D}^\omega$ and $\mathbf{B}^\omega$ denote the constraint matrices multiplying $\mathbf{y}^\omega$ and $\mathbf{x}$, respectively. We assume *relatively complete recourse*; i.e., the second-stage problems are feasible for every feasible solution $\mathbf{x}$ of the first-stage problem. We also assume dual feasibility of the second-stage problems. Our application satisfies these assumptions because a contract is in place to purchase external water. We now review DRO with ϕ-divergences.

3 Distributionally Robust Optimization Using ϕ-Divergences

The background below is based on *Pardo* [2005]; *Ben-Tal et al.* [2013]; *Bayraksan and Love* [2015].

3.1 ϕ-Divergences

ϕ-divergences are used in statistics to measure a "distance" between two distributions. Let $\mathbf{q} = (q_1, \ldots, q_n)^T$ and $\mathbf{p} = (p_1, \ldots, p_n)^T$ be two probability vectors—i.e., satisfying $q_\omega, p_\omega \geq 0$, $\forall \omega$ and $\sum_{\omega=1}^n p_\omega = \sum_{\omega=1}^n q_\omega = 1$. Recall $\mathbf{q}$ denotes the nominal distribution. The ϕ-divergence from $\mathbf{p}$ to $\mathbf{q}$ is defined by

$$I_\phi(\mathbf{p}, \mathbf{q}) = \sum_{\omega=1}^{n} q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right), \tag{4}$$

where $\phi(t)$—called the *ϕ-divergence function*—is a convex function on $t \geq 0$ such that $\phi(t) \geq 0$ and $\phi(1) = 0$, $0\phi(a/0) = a \lim_{t\to\infty} \frac{\phi(t)}{t}$, and $0\phi(0/0) = 0$. The right-hand side of (4) is the expectation of the ϕ-divergence function with respect to the nominal distribution $\mathbf{q}$, evaluated at the ratios $\frac{p_\omega}{q_\omega}$. For instance, if the probabilities are same at scenario ω ($p_\omega = q_\omega > 0$), that scenario's contribution to the divergence is zero.

Table 1 lists ϕ-divergences used in this paper. The Modified χ^2 distance is related to the famous χ^2 goodness-of-fit test. Kullback-Leibler (KL) divergence is commonly used in probability and information theory. It is the expected log-scale loss $\sum p_\omega \left(\log(p_\omega) - \log(q_\omega)\right)$ with respect to $\mathbf{p}$. Burg Entropy changes the order of $\mathbf{p}$ and $\mathbf{q}$ in KL divergence; so it is the expected log-scale loss with respect to $\mathbf{q}$. We will soon discuss why we picked these ϕ-divergences in Section 3.5.

3.2 DRO Formulation

DRO minimizes the worst-case expectation from a set of distributions that are similar to the nominal distribution $\mathbf{q}$. The resulting minimax formulation is

$$\min_{\mathbf{x}\in X} \max_{p\in\mathcal{P}} \left\{ \mathbf{cx} + \sum_{\omega=1}^{n} p_\omega h_\omega(\mathbf{x}) \right\}, \tag{5}$$

Table 1: ϕ-divergences used in this study.

Divergence	$\phi(t)$	$\phi(t), t \geq 0$	$\lim_{t\to\infty} \frac{\phi(t)}{t}$	$I_\phi(p,q)$
Modified χ^2 Distance	$\phi_{m\chi^2}$	$(t-1)^2$	∞	$\sum \frac{(p_\omega - q_\omega)^2}{q_\omega}$
Kullback-Leibler Divergence	ϕ_{kl}	$t\log t - t + 1$	∞	$\sum p_\omega \log\left(\frac{p_\omega}{q_\omega}\right)$
Burg Entropy	ϕ_b	$-\log t + t - 1$	1	$\sum q_\omega \log\left(\frac{q_\omega}{p_\omega}\right)$

where the ambiguity set of distributions is given by

$$\mathcal{P} = \left\{ \mathbf{p} : I_\phi(\mathbf{p},\mathbf{q}) \leq \rho,\ \sum_{\omega=1}^{n} p_\omega = 1,\ p_\omega \geq 0,\ \forall \omega \right\}. \tag{6}$$

The first constraint in (6) ensures that only distributions sufficiently close—in terms of a given ϕ-divergence—to $\mathbf{q}$ are selected. The remaining constraints in (6) simply ensure that $\mathbf{p}$ itself is a probability vector.

Consider each scenario ω observed N_ω times with $N = \sum_{\omega=1}^{n} N_\omega$ total observations. The nominal probabilities are set to $q_\omega = \frac{N_\omega}{N}$. Then, when ϕ is twice continuously differentiable around 1 with $\phi''(1) > 0$ (like our choices in Table 1), the value

$$\rho = \frac{\phi''(1)}{2N} \chi^2_{n-1,1-\alpha}, \tag{7}$$

where $\chi^2_{n-1,1-\alpha}$ is the $1-\alpha$ percentile of a chi-squared distribution with $n-1$ degrees of freedom, produces an approximate $1-\alpha$ confidence region on the true distribution [*Pardo*, 2005; *Ben-Tal et al.*, 2013].

3.3 Solution Method

DRO can be formulated in alternate ways that make it computationally tractable. By taking the Lagrangian dual of the inner maximization problem—with Lagrange multipliers λ and μ corresponding the to first two constraints in (6), respectively—and combining the two minimization problems, an equivalent reformulation of DRO is:

$$\min_{\mathbf{x},\lambda,\mu} \quad \mathbf{cx} + \mu + \rho\lambda + \sum_{\omega=1}^{n} q_\omega \left(\lambda \phi^* \left(\frac{h_\omega(\mathbf{x}) - \mu}{\lambda} \right) \right) \tag{8}$$

$$\text{s.t.} \quad \mathbf{x} \in X,\ \ \lambda \geq 0,\ \ h_\omega(\mathbf{x}) - \mu \leq \lambda \lim_{t\to 0} \frac{\phi(t)}{t},\ \ \omega = 1, 2, \ldots, n. \tag{9}$$

Above, ϕ^* is the conjugate of ϕ defined as $\phi^*(s) = \sup_{t\geq 0}\{st - \phi(t)\}$, $s \in \mathbb{R}$. The last set of constraints in (9) result from an implicit feasibility consideration. If this limit is ∞ (Table 1), these constraints are redundant and can be ignored.

The above DRO is a convex program: With Modified χ^2 distance, it is a second order cone program, whereas with KL and Burg, it has a self-concordant barrier [*Ben-Tal et al.*, 2013]. Unfortunately, state-of-the-art optimization software fail to solve our problem due to its large size. So, we employ a decomposition algorithm. Observe that (8)–(9) is a two-stage stochastic convex program. The first stage has variables $\mathbf{x}$, λ, and μ. The expectation of convex function $h^\dagger_\omega(\mathbf{x},\lambda,\mu) = \lambda\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{\lambda}\right)$ is taken with respect to the nominal distribution.

The decomposition method replaces the expectation of $h^\dagger_\omega(\mathbf{x},\lambda,\mu)$ by its lower approximation through affine cutting planes using (sub)gradients. It is easy to generate (sub)gradients of $h^\dagger_\omega(\mathbf{x},\lambda,\mu)$

by translating the (sub)gradients $h_\omega(\mathbf{x})$ through the chain rule. Also, for Burg entropy, the last set of constraints in (9) is removed and replaced by feasibility cuts when violated. This means that, in our application, only linear problems need to be solved, leading to computational efficiency [*Love and Bayraksan*, 2016].

In the rest of the paper, $\{\mathbf{x}^*, \mathbf{p}^*\}$ denotes an optimal solution of the primal DRO (5). Similarly, $\{\mathbf{x}^*, \lambda^*, \mu^*\}$ denotes an optimal solution of the dual DRO (8).

3.4 Price and Value of Data

It is natural to ask whether gathering additional data is worthwhile or not. We can make this decision by considering the *Price of Data (PoD)*. It is the difference between the current optimal value and expected optimal values with one additional data [*Bertsimas et al.*, 2017]. The exact PoD might be computationally expensive to calculate because n problems must be solved. *Love and Bayraksan* [2016] provide two simple bounds on PoD, without needing to solve any additional optimization problems. For any ϕ-divergence and any given nominal distribution

$$\text{PoD} \geq \max\{\text{L}_1, \text{L}_2\}, \tag{10}$$

where $\text{L}_1 = \frac{\rho\lambda^*}{N+1}$ and $\text{L}_2 = \lambda^* \sum_\omega q_\omega \left[\phi^*\left(\frac{h_\omega(\mathbf{x}^*)-\mu^*}{\lambda^*}\right) - \frac{N+1}{N}\phi^*\left(\frac{N}{N+1}\frac{h_\omega(\mathbf{x}^*)-\mu^*}{\lambda^*}\right)\right]$. If the cost of data collection is less than $\max\{\text{L}_1, \text{L}_2\}$, it would be worthwhile to do so.

After we decide to collect additional data based on PoD, it is natural to ask whether a specific additional data changes the optimal value—the *Value of Data (VoD)*. The below inequalities provide sufficient conditions for an additional observation of scenario $\hat{\omega}$ to decrease the optimal value of DRO for

Modified χ^2: $2\sum_\omega p^*_\omega \frac{p^*_{\hat{\omega}}}{q_\omega} > \left(\frac{p^*_{\hat{\omega}}}{q_{\hat{\omega}}}\right)^2 + \left(\frac{N+1}{N}\right)^2$,

Kullback-Leibler: $\sum_\omega q_\omega \left(\frac{p^*_\omega}{q_\omega}\right)^{\frac{N}{N+1}} \log\left(\frac{p^*_\omega}{q_\omega}\right)^{\frac{N}{N+1}} + 1 > \left(\frac{p^*_{\hat{\omega}}}{q_{\hat{\omega}}}\right)^{\frac{N}{N+1}}$,

Burg entropy: $\frac{p^*_{\hat{\omega}}}{q_{\hat{\omega}}} < \frac{N}{N+1}$.

These conditions provide a simple test for determining which scenarios would decrease the expected cost. Such scenarios can be considered to be "low-cost."

3.5 Properties of ϕ-Divergences Selected for this Study

The data in this study comes in the form of projections of possible future scenarios. The decision maker may be interested in knowing whether some scenarios should be given zero probability $p^*_\omega = 0$ in the optimal solution. That is, some scenarios may be overly optimistic, and the decision maker may want a formulation capable of making this distinction. We refer to having a zero optimal probability $p^*_\omega = 0$ when the nominal probability is positive $q^*_\omega > 0$ as *suppressing* scenario ω. In essence, such a scenario is excluded from the final optimal expectation.

Both the Modified χ^2 and the KL divergence are capable of suppressing scenarios, but they do so in different ways. Problem (5) formulated with the Modified χ^2 distance is capable of choosing whether to suppress any scenario individually, thus generating a wide variety of possible model output. In contrast, when KL divergence is used, the only possible results are: (a) no scenarios will be suppressed (i.e., $p_\omega > 0$ for every scenario ω), or (b) all but the most costly scenarios will be suppressed (i.e., only the most costly scenarios will have $p^*_\omega > 0$, while all others will have zero probability $p^*_\omega = 0$).

Unlike the two ϕ-divergences discussed above, the Burg entropy is not capable of suppressing scenarios. Thus, the solution will always have $p^*_\omega > 0$ for every scenario. With Burg entropy, the decision maker ensures that all scenarios put into the model are represented with a positive probability in the optimal solution. This may be especially desired if the data comes from highly trusted sources. We refer the readers to *Love and Bayraksan* [2016] for a derivation of these behaviors.

4 Application to Tucson, AZ

4.1 Application-Specific Network

We applied the above techniques to allocate water in a developing region of Tucson, AZ. A schematic view of the study area's water system is shown in Figure 1. Majority of Tucson's water comes from the Colorado River, brought in by the Central Arizona Project (CAP) canal. This water is then treated and sent to customers, or seeped into underground to be saved for future use. These are represented as "CAP" and a few other white nodes in top-left corner of Figure 1.

The southeastern portion of Tucson is being increasingly developed. This area is split into different elevation zones: C, D, E, FS, FN, . . ., I. The elevation zones also split the region into several demand zones. Given the capacity of the existing treatment plants and the energy cost of pumping water through a series of zones, the governing agencies in Tucson are interested in building additional treatment facilities in this area, hereafter known as the RESIN (REsilient and Sustainable INfrastructures) area.

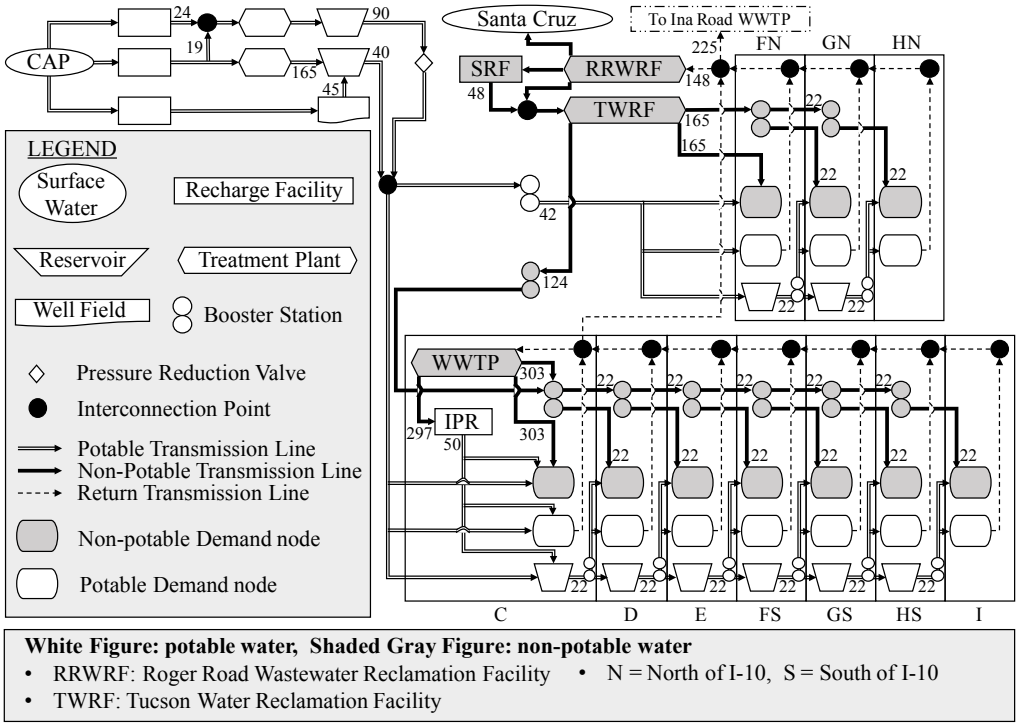


Figure 1: A schematic of the water system in the RESIN area.

Figure 1 shows both existing infrastructures (e.g., CAP) and proposed new infrastructures—a satellite wastewater treatment plant (WWTP) and indirect potable reuse facility—in Zone C. The schematic also shows both the potable water system (through white nodes and double-lined acrs) and the reclaimed water system (through gray nodes and solid black arcs). Each zone C–I contains potable and non-potable demand nodes, a reservoir and booster station for transporting the potable and non-potable water, and wastewater return pipes. Potable water, being of higher quality, can be used to meet either type of demand. We assume potable demand makes up 80% of the total demand and nonpotable demand is the remaining 20%. Figure 1 provides cost ($/acre-foot (af)) on each arc if it is not negligible.

A dummy node capable of supplying water in the event of a water shortage is also included in the model (but not shown in figure). The cost of this extra supply is set at \$800/af by a fixed contract. In other words, there exits a water market at a constant exogenous price, which is common in the literature [*Murali et al.*, 2015; *Calatrava and Garrido*, 2005; *Weinberg et al.*, 1993; *Dinar and Letey*, 1991].

The resulting model has a total of $P = 37$ time periods, representing years 2014–2050. We use $P_1 = 3$ for the first stage. So, the first-stage problem considers years 2014–2016, and the uncertain second stage considers 2017–2050. For each year, the network has $|\mathcal{N}| = 60$ nodes representing demands for potable and nonpotable water, pumps, water treatment plants, reservoirs, and the available water supply from the Colorado River. The network in each year has $|\mathcal{A}| = 100$ arcs, representing the pipes carrying water between the nodes and connecting the network to the five reservoirs for storage. Costs for all time periods are brought into present value by applying a 4% discount rate per year. See *drwa-si.pdf* for a detailed formulation.

4.2 Data-Driven Methodology to Predict Water Demands and Supplies

Our optimization model requires two primary uncertain data: annual water demand by zone and annual water supply. Figure 2 provides a summary of our data-driven methodology to predict these inputs. Throughout the paper, a number beginning with 'h' denotes historical data, and a number beginning with 'p' denotes predicted data. The arrows in Figure 2 indicate flow from the input data to the output data. Data in solid square boxes represent calculated, predicted, and formed output with input data, whereas data without solid square boxes are obtained from various existing sources. The bold boxes highlight important steps of the procedure. Before we discuss the details of our methodology, let us first summarize our data and its sources.

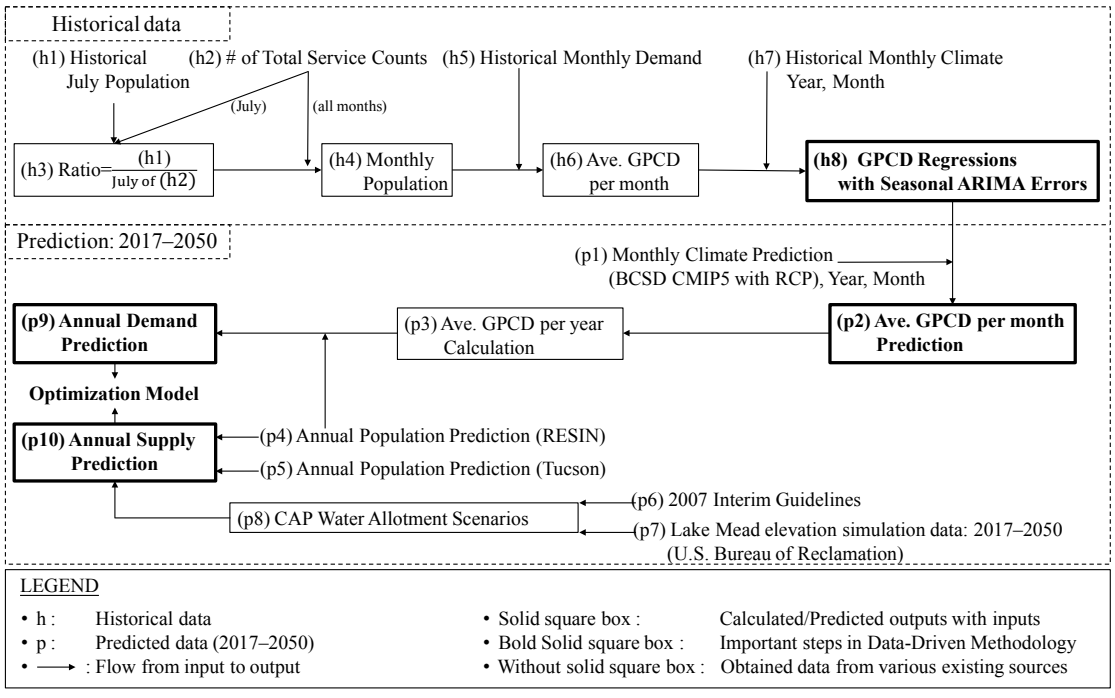


Figure 2: Flowchart of data-driven methodology to predict annual water demands and supplies.

4.2.1 Data

The supporting information and datasets provide most of the raw data and all of the intermediate and final input data. These, along with data sources are as follows:

- *drwa-si.pdf*: Intermediate/regression results (h8);
- *drwa-ds1.xlsx*: Raw datasets {(h1), (h2), (h5), (h7)} [*Tucson Water*, 2013], Intermediate results {(h3), (h4), (h6), (p2), (p3), (p8)}, Predicted data from various sources {(p1) [*Brekke et al.*, 2013], (p4) [*WISP*, 2009; *Pima Association of Governments*, 2012; *U.S. Census Bureau*, 2010; *Scott et al.*, 2012], (p5) [*City of Tucson*, 2008], (p6) [*Johnson and Kempthorne*, 2007], (p7) [*Nowak*, 2014; *U.S. Department of the Interior Bureau of Reclamation*, 2012]};
- *drwa-ds2.xlsx*: Final inputs to optimization. For each scenario, nominal probabilities, and for each year 2014–2050, water supplies (p10) and water demands by zones (p9).

4.2.2 Population Estimates

Because population affects both water demands and supply, we discuss it first. Water demand in each zone is proportional to the population of each zone. The CAP water allocation (=water supply), on the other hand, depends on the *ratio* of the study area's population to the overall Tucson population. Therefore, we need population estimates for both the RESIN and Tucson areas.

For the RESIN area (p4), we have two 2050 population estimates: (i) A low-population estimate of 413,115 from the Traffic Analysis Zone (TAZ) study [*Pima Association of Governments*, 2012], and (ii) a high-population estimate of 693,116 from the Water & Wastewater Infrastructure, Supply & Planning Study [*WISP*, 2009]. To predict the population in intermediate years, we used the 2010 U.S. Census number of 37,005 [*U.S. Census Bureau*, 2010] and assumed a linear increase. The annual population is then broken down to demand zones, based on the population model developed by Tucson Water for the RESIN area. Zones D and E have the highest population increases during the study period.

For the Tucson population (p5), we used the U.S. Census number of 750,000 in 2010 and a 2050 population estimate of 1,300,000 from [*City of Tucson*, 2008] and again assumed a linear increase. Detailed population numbers are available in *drwa-ds1.xlsx* and Table S5 of *drwa-si.pdf*.

4.2.3 Water Demand Prediction

We first use historical data (upper part of Figure 2) to determine how Gallons Per Capita per Day (GPCD) is affected by climate variables like temperature and precipitation as well as water-use trends. We produce two GPCD regressions with seasonal autoregressive errors. These statistical models are passed to prediction (lower part of Figure 2) to estimate water demand.

4.2.3.1 ***Preparation of Historical Data*** Our optimization model has annual time periods. But, in order to capture the seasonal weather effects more accurately, we first work with monthly data. The historical demand per month is decomposed into "population in a given month × average GPCD that month × number of days in that month". Because only July historical population (h1) is available, our first task is to estimate populations of all other months. We define monthly population service count ratio for each year (h3) as $\frac{\text{July historical population (h1)}}{\text{July total service counts (h2)}}$. This ratio decreases from 3.44 in 1991 to 3.14 in 2011. Using these ratios, we estimate monthly historical population (h4) simply as "ratio of the year (h1)× total service counts of each month of that year (h2)".

4.2.3.2 ***Two GPCD Regressions with Seasonal ARIMA Errors*** Given the monthly population estimates (h4) by the above calculation and already available historical data on monthly water demand (h5), average GPCD for every month (h6) are calculated. We directly use this data (h6) to investigate water-use trends and relate GPCD to climate. Because both GPCD and the climate variables temperature and precipitation (h7) are time-series data, residuals of the ordinary least squares are autocorrelated. In order to handle this problem, we use generalized least squares (GLS) with seasonal autoregressive integrated moving average (ARIMA) errors. Both models below are GLS

with ARIMA $(1,0,0) \times (1,0,0)_{11}$ errors. The regressors of GLS include temperature, precipitation, year, and twelve monthly indicator variables. As a result, intercepts are not included. Residuals of both models satisfy all assumptions based on sample (partial) autocorrelation function, residual plot, Ljung-Box test, and Kolmogorov-Smirnov normality test.

The average GPCD began dropping near the beginning of the 21st century, from over 170 GPCD in 1996 to 140 GPCD in 2011. This analysis led to two GPCD regression functions: The first projects a continual increase in water efficiency out to 2050, and the other has a cap on this drop. The first (lower-GPCD) fit might be appropriate if technological advances and water conservation efforts lead to significantly lower water consumption. The second (higher-GPCD) fit assumes that people cannot decrease water consumption indefinitely. Let us briefly discuss them; details can be found in *drwa-si.pdf*.

Lower-GPCD Fit. This fit has an adjusted $R^2 = 0.8939$ (Table S1 of *drwa-si.pdf*). Both temperature and precipitation are statistically significant, and precipitation seems to explain more of the variation in the data. This regression projects water demand to decrease from 140–146 GPCD in 2014 to 94–98 GPCD by 2050. In contrast, Tucson Water uses an estimate of 120–145 GPCD in their projections.

Higher-GPCD Fit. The second model captures the decrease in water usage but does not extrapolate this trend. To achieve this, we conducted the GLS on a bounded value of the 'year' by choosing the value of 'year' in a bounded set $\{Y_l, Y_l + 1, \ldots, Y_u\}$. The highest year $Y_u = 2011$ was chosen so that decrease in water demand would not be projected beyond the available historical data. The earliest year $Y_l = 2004$ was then selected because it gives the best fit, with an adjusted $R^2 = 0.8770$ (Table S2 of *drwa-si.pdf*). The higher-GPCD fit typically predicts 133–143 GPCD, which is in-line with Tucson Water.

Figure S1 of *drwa-si.pdf* contrasts the projections from the two fits.

4.2.3.3 ***Estimating Future Demands*** By the above analysis, we now have two functions to estimate average GPCD in a future month. These functions take as input temperature and precipitation predictions of climate models with a given greenhouse concentration pathway. Then, the predicted average GPCDs in future months (p2) are turned into average GPCDs in future years (p3) by simply considering the number of days in a month and year. This way, these predictions not only include seasonal weather patterns but also climate variability by mid-century. Finally, we estimate the demands in a zone as "annual water demand in a zone (p9) = average GPCD in a given year (p3) $\times$ population estimate in that zone of that year (p4)" (*drwa-ds2.xlsx*). The population predictions were discussed in Section 4.2.2. We now discuss the climate models and greenhouse concentration paths used in our study (p1).

Climate Models Used in the Study. Table 2 lists the climate models used. Bias-Corrected and Spatially Downscaled (BCSD) data from Coupled Model Intercomparison Project: Phase 5 (CMIP5) was obtained from *Brekke et al.* [2013]; see also *Reclamation* [2013]. Each model includes predictions of "`tasmax`" and "`pr`", the average daily high temperature ($^\circ C$) and the average precipitation rate (mm/day) in each month, respectively. These are direct inputs to our lower- and higher-GPCD regression functions to forecast GPCDs. We picked these climate models to have a good representation without overly increasing the problem size. Additional climate models can be easily added to the study.

Representative Concentration Pathways Used in the Study. Each climate model has output associated with a given path for future greenhouse gas concentration, called the Representative Concentration Pathway (RCP). Our analysis includes the four paths RCP2.6, RCP4.5, RCP6.0 and RCP8.5 adopted by IPCC [*Pachauri et al.*, 2014]. The greenhouse concentration path RCP2.6 is an optimistic case where concentrations are drastically reduced by mid-century. The paths RCP4.5 and RCP6.0 show stabilization of concentrations before and after 2100, respectively. Finally, RCP8.5 is the case where concentration continue to grow quickly throughout the remainder of the century.

Table 2: A list of climate models used in this analysis.

Institution	Model
Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia	CSIRO-mk-3-6-0
Geophysical Fluid Dynamics Laboratory	GFDL-CM3 GFDL-ESM2M
Met Office Hadley Centre	HadGEM2-ES
Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies and Japan Agency for Marine-Earth Science and Technology	MIROC5 MIROC-ESM

With the high-concentration path RCP 8.5, all models forecast the average daily high temperature to be consistently 1–3°C higher than the historical record in every month after 2040, with smaller variations between the models (Table S4 of *drwa-si.pdf*). We note that the models typically predict more rain than the historic record as well, especially during the summer monsoon in July–September.

4.2.4 Water Supply Prediction

Annual water supply to the study area (p10) is calculated by "CAP allocation to Tucson $\times \frac{\text{RESIN Population}}{\text{Tucson Population}}$." Population estimates were discussed in Section 4.2.2. Below, we explain how we estimated future CAP water allocation to Tucson.

4.2.4.1 ***CAP Water Allocation*** In this study, the conditions described in the Colorado River Compact 2007 Interim Guidelines (p6) [*Johnson and Kempthorne*, 2007] dictate the CAP water allocation to Tucson. Under Normal conditions, Tucson Water has an annual CAP water allocation of 144,000 af. According to the compact, there are three drought conditions: Tiers 1, 2, and 3. Tier 1 drought is declared if Lake Mead elevation is between 1,050–1,075 feet. If so, CAP allocation is reduced by 11.43%. Tier 2 water shortage happens when Lake Mead elevation belongs in the range $[1,025,\ 1,050)$. Then, the CAP allocation is reduced by 14.29%. Finally, under extreme water shortage of Tier 3 (Lake Mead elevation below 1,025 feet), only 119,318 af is allocated to Tucson—a 17.14% reduction.

4.2.4.2 ***Water Shortage Prediction*** To predict the future water allocations, we used the Lake Mead elevation simulations (p7) by the *U.S. Department of the Interior Bureau of Reclamation* [2012]; *Nowak* [2014]. We divided the second stage of our problem into three periods: 2017–2025, 2026–2035, and 2036–2050. Then, we estimated the *nominal* probability of each condition—Normal, Tiers 1, 2, and 3—as the fraction of all end-of-December Lake Mead elevation simulations that satisfy a specific condition at least once during a given period. Table 3 summarizes the results. These simulations indicate that the chance of Normal condition decreases and the chance of extreme shortage increases over the years. Overall, the four conditions for three time periods provide $4^3 = 64$ different water allocation paths (p8) for our study. This calculation assumes the water allocation remains same for a time period and each time period is independent. We note that these assumptions can be easily changed to include other allotment paths into our model.

4.3 Scenarios and Infrastructure Configurations

Putting this all together, we consider the following uncertain elements:

- 2 population projections (p4),

Table 3: Estimated nominal probabilities of CAP water allotment conditions.

Stage	Period	Year	Probability of Conditions			
			Normal (144,000 af)	Tier 1 (127,541 af)	Tier 2 (123,422 af)	Tier 3 (119,318 af)
First		2014–2016	1.0000	0.0000	0.0000	0.0000
Second	First	2017–2025	0.6232	0.0891	0.0749	0.2128
	Second	2026–2035	0.4757	0.0960	0.0777	0.3506
	Third	2036–2050	0.3787	0.0831	0.0601	0.4781

- 2 per-capita demand models (h8), (higher-GPCD, lower-GPCD)
- 6 climate models (p1),
- 4 greenhouse gas concentration (p1),
- 64 water allotment scenarios (p8).

These considerations yield 6,144 ($= 2 \times 2 \times 6 \times 4 \times 64$) future scenarios for our study. To apply these scenarios to DRO, we assume $N = 6,144$ total observations, and change the weighting of each scenario only according to its water allotment. All other uncertainties are assumed to be equally likely because we do not have a preference for climate models, population models, etc. For example, scenario ω with Normal condition for all three periods in the second stage has number of observations $N_\omega = 64 \times 0.6232 \times 0.4757 \times 0.3787 = 7.1852$. These numbers are chosen to have a total of 64 allotment observations while maintaining the desired nominal probabilities. The nominal probabilities of DRO (listed in *drwa-ds2.xlsx*) can be easily changed based on expert opinions or other considerations.

In addition to the scenarios outlined, we evaluate three infrastructure options in the RESIN area:

1. **NI:** No additional infrastructure is constructed.
2. **WWTP:** A satellite wastewater treatment plant is constructed, capable of treating wastewater up to a non-potable quality, for satisfying demands in its own zone and and higher zones.
3. **IPR:** In addition to the WWTP, an indirect potable reuse facility can be constructed, which further treats water from the WWTP up to potable quality.

Figure 1 illustrates an additional WWTP and IPR constructed in Zone C.

5 Results

The water allocation model was solved using the ϕ-divergences outlined in Section 3.5 and Table 1. For each ϕ-divergence, we considered values of ρ in (7) corresponding to the asymptotic confidence regions of 90%, 95%, and 99%. The decomposition method outlined in Section 3.3 was used to solve the model, which was implemented in MATLAB using the CPLEX optimization software.

5.1 Results with No Additional Infrastructure

5.1.1 Costs

Table 4 shows the optimal expected operating costs, including \$800/af water-shortage cost, by infrastructure type, ϕ-divergence, and confidence level. Here, we examine the first part of this table. Modified χ^2 generates the lowest costs, followed by the KL divergence. Burg entropy produces similar results to KL, but has a slightly lower cost. The major difference between these

ϕ-divergences is in the scenarios they suppressed: The Modified χ^2 distance suppressed scenarios for every confidence level tested, and it consistently suppressed the low-population, lower-GPCD scenarios—the scenarios with the smallest water shortages. KL divergence maintains an "all-or-nothing" approach to suppressing scenarios, but a confidence level of 99% is not high enough to induce the suppressing behavior for this problem. In contrast, Burg entropy is incapable of suppressing scenarios and yields slightly higher costs than KL.

Table 4: Optimal expected operating costs through 2050 for each infrastructure configuration and break-even shortage costs that balance construction costs with savings.

Infrastructure	ϕ–divergence	Operating Cost ($ Million) 90%	95%	99%	Break-even Shortage Cost
NI	Modified χ^2	326.62	326.79	327.11	—
	Kullback-Leibler	329.14	329.33	329.69	
	Burg	329.81	329.99	330.33	
WWTP	Modified χ^2	311.47	311.63	311.93	$4,550/af rel. to **NI**
	Kullback-Leibler	313.83	314.01	314.35	
	Burg	314.61	314.78	315.10	
IPR	Modified χ^2	285.32	285.48	285.77	$1,866/af rel. to **NI**
	Kullback-Leibler	287.52	287.69	288.03	
	Burg	287.91	288.09	288.39	$1,419/af rel. to **WWTP**

5.1.2 Comparison of Climate Models and Greenhouse Gas Concentration Paths

We computed the total optimal probability assigned to each climate model and concentration path by DRO. Table 5 presents the results for the KL divergence at 95% confidence. For other ϕ-divergences and confidence levels, the optimal probabilities are similar. DRO assigns the highest probability to the highest concentration path RCP8.5. This way, DRO induces a risk-averse behavior, protecting against the more frequent water shortages associated with this path. The second highest probability is given to RCP4.5—a somewhat lower concentration path than RCP6.0. In fact, RCP6.0 has the lowest overall cost during the study period. This path has lower greenhouse gas concentrations in earlier years, allowing the system to store water, and thereby reducing water shortages later on.

Among the climate models, HadGEM2-ES results in the highest operating cost, followed by MIROC-ESM-CHEM. GFDL-ESM2M tends to generate lower temperatures (see Table S4 in *drwa-si.pdf*) and thus lowers the demands. As a result, DRO gives it the lowest probability. Considering both the climate models and concentration paths together, we find that HadGEM2-ES with RCP8.5 results in the highest operating cost (and probability), while GFDL-ESM2M with RCP2.6 has the lowest cost (and probability).

5.1.3 Price and Value of Data

We applied the price and value of data from Section 3.4 to the results of the **NI** model with all three ϕ-divergences. We present the results at 95% confidence level. PoD lower bound, given in (10), yielded the following result: For an additional observation with Modified χ^2, KL, and Burg at 95% confidence, one is willing to pay at least $4,267.06, $4,849.05, and $4,534.06, respectively. Note that the Modified χ^2 achieved the maximum bound with L_1, and KL and Burg achieved the maximum bound with L_2. Reliable data collection/prediction costing less than $4,000 could be beneficial for this study.

With the VoD calculations, several themes immediately emerged for this problem: (i) every low-population (3,072 scenarios), and (ii) every high-population, lower-GPCD scenarios (1,536 sce-

Table 5: Optimal probabilities for each climate model and concentration path (KL, 95%).

	RCP2.6	RCP4.5	RCP6.0	RCP8.5	(all)
CSIRO	0.0408	0.0442	0.0395	0.0406	0.1651
GFDL-CM3	0.0414	0.0412	0.0412	0.0443	0.1682
GFDL-ESM2M	0.0389	0.0402	0.0402	0.0393	0.1586
HadGEM2-ES	0.0415	0.0437	0.0396	0.0457	0.1705
MIROC5	0.0420	0.0421	0.0409	0.0432	0.1683
MIROC-ESM-CHEM	0.0410	0.0429	0.0424	0.0440	0.1703
(all)	0.2457	0.2543	0.2438	0.2570	1

narios) satisfied the condition. These scenarios *guarantee* a decrease in the overall cost. In contrast, none of the high-population, higher-GPCD scenarios (1,536 scenarios) satisfied this condition. These results suggest that higher demands—as a product of population × GPCD—are the most important factor that increase the frequency of water shortages, and hence the overall cost.

5.2 Results with Additional Decentralized Infrastructure

5.2.1 Water Shortage

One of the most important aspects of decentralized water treatment is that it increases water reuse. With this additional water supply, water shortage is decreased. We examine the effect of additional infrastructure on water shortage first. Figure 3a depicts the empirical Cumulative Distribution Function (CDF) of the total shortage (in af) for each infrastructure configuration using the nominal distribution. The CDFs of **WWTP** and **IPR** are always above that of **NI**. This means that they have benefits over **NI** regarding shortage. For example, the nominal probability that total shortage is less than or equal to 200,000 af is 0.47, 0.52 and 0.77 for **NI**, **WWTP** and **IPR**, respectively. Looking at the highest CDFs, **IPR** provides the most substantial reduction. **WWTP** is almost always better than **NI**, but always worse than **IPR**.

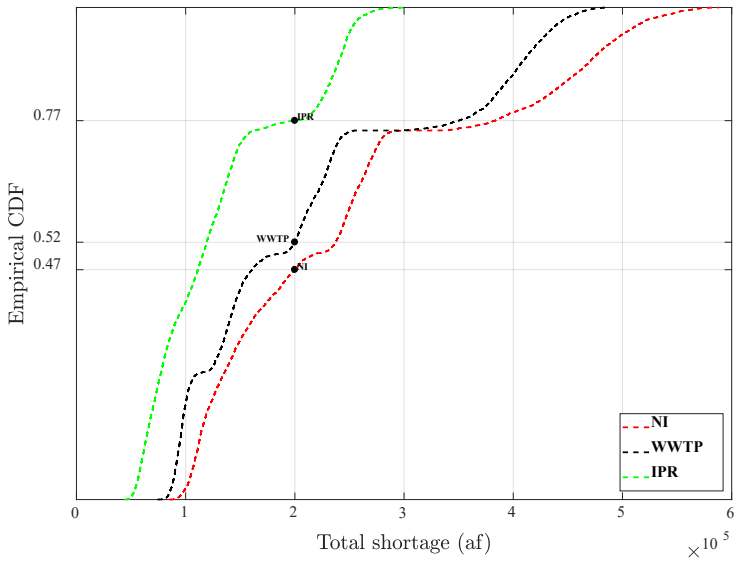
Figure 3b depicts the total shortage amount over the 37-year study period, broken down by (i) infrastructure, and (ii) four population/GPCD-demand categories. A satellite **WWTP** provides a substantial reduction in shortage severity—especially in the lower-GPCD scenarios. But, the **IPR** facility decreases the extreme shortages, which occur in high-population, higher-GPCD scenarios. These shortage amounts are computed with the KL divergence at 95% confidence, but they are dictated largely by the physical constraints of the system and do not change substantially with different ϕ or ρ.

The analysis clearly shows the value of **IPR**—and to a lesser extent the value of **WWTP**—in reducing water shortages. We examine their economic value next.

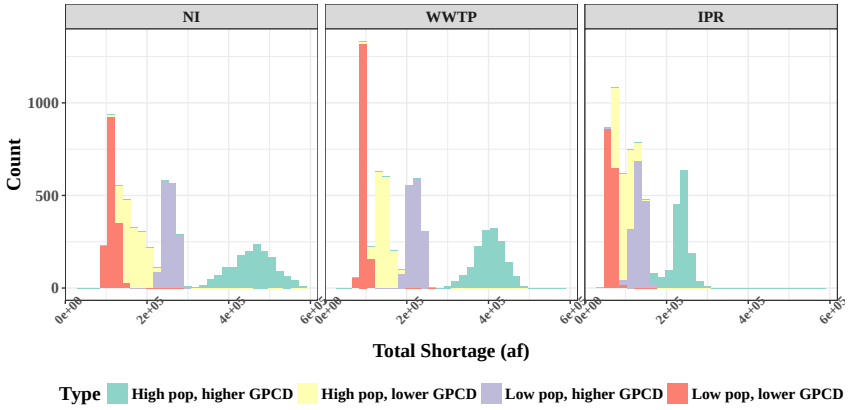
5.2.2 Cost-Benefit Analysis

Table 4 reveals that **WWTP** consistently decreases the operating cost by about $15 million, and the **IPR** facility reduces the cost by an additional $26 (or total $42) million over the 37-year time span. This is mainly due to the reduced water shortages. Our earlier analysis indicates that the **WWTP** and the **IPR** facility, if constructed, would need a capacity of 10 million gallons per day. At that capacity, the estimated cost of these facilities is $55 and $119(=55+64) million, respectively. As a result, the additional facilities will *not* pay for themselves over the planning period.

So far we used a shortage cost of $800/af. We now examine break-even shortage costs that balance the construction costs with operational savings. These break-even costs can provide guidelines on the appropriate level of satellite infrastructure. Last column of Table 4 summarizes the results.



(a) Empirical CDF of total shortage.



(b) Histogram of total shortage.

Figure 3: Total shortage over the 37-year study period for each infrastructure configuration (KL, 95%).

We find that the break-even cost for **IPR** is lower because it drastically lowers extreme shortages. These imply that the increased operation cost of the **IPR** facility plus its higher construction cost is significantly lower than the benefits it provides—especially at higher shortage costs. Figure S2 in *drwa-si.pdf* further depicts the break-even costs.

6 Discussion

Although results show that **IPR** is the best option in terms of shortages—total shortage amounts and break-even shortage costs—public opinions and long-term health effects should be considered because a planned IPR facility has not yet been fully supported by the public [*Ormerod and Scott*, 2013; *Scott et al.*, 2012; *Martin*, 2013].

This study reveals that GPCD is the main driver of water shortages among the categorized uncertainties (Figure 3b). The two highest shortages in Figure 3b occur at higher-GPCD scenarios. The implication of this result is twofold. First, water conservation efforts and technologies that lower GPCD could have a drastic effect in the area. Second, next to water-shortage costs, a decision on infrastructure largely depends on GPCD. It should not be overlooked that GPCD in itself depends on climate models and greenhouse concentration paths. Therefore, the final construction decision needs to consider the impact of climate scenarios, as in the discussion of Table 5.

We used the distributionally robust approach with the ambiguity set derived from ϕ-divergences. This approach is preferable when there is little data or not much trust in the data. If we have enough data with trust, gain from this approach might not be significant compared to the traditional two-stage stochastic model.

Nevertheless, the principle advantage of this method is clear. It produces a measure of the relative importance (i.e., optimal probability p^*_ω) of each scenario, which we used to estimate the expected cost of operation in the RESIN area. We also showed that this optimal distribution and the resulting robust solution, along with VoD, can be used to evaluate the impact of climate and other scenarios directly.

7 Conclusion

The summary of the paper is as follows:

(i) To the best of our knowledge, this is the first application of DRO with ϕ-divergences to a water allocation problem. We believe this modeling approach is very promising for long-term water resources management problems because the future uncertainties (e.g., climate, population) are not known and their uncertainty quantification is difficult.

(ii) We applied DRO to a developing area in Tucson, Arizona by integrating future climate and population predictions, water-use trends, and hydrological simulations. This model aims to find solutions that are robust to these uncertainties. We evaluated the effect of different uncertainties on this problem by examining the severity of future water shortages and expected operation costs of the system.

(iii) A cost-benefit analysis of additional water treatment infrastructures is conducted by considering both the construction costs and reduction in water shortages. Our results indicate IPR can significantly reduce shortages, and it is economically feasible to build if the shortage cost increases.

A worthwhile extension of this study is a multistage model—which is subject of ongoing work—where the uncertainty is revealed over time. This will allow for a more refined treatment of uncertainty and infrastructure decisions at multiple points. If more data is available, the model can be improved on several fronts. First, there is not sufficient data to model the dependence of water shortage cost to climate events. Such analysis would further reveal the role of climate in infrastructure decisions. Additional data analysis on dependencies between water demands and supplies in the RESIN area would also be valuable. Another future extension could incorporate health impacts—especially for IPR—into the model. Finally, increasing the size of the model to include more of Tucson and the state of Arizona would permit a more detailed water planning.

Acknowledgments
All data used in the paper is summarized in Section 4.2. The World Climate Research Programme's Working Group on Coupled Modeling is responsible for CMIP, and we thank the climate modeling groups (listed in Table 2) for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. The authors are grateful to Tucson Water, Kevin Lansey, Gwen Woods, and Alicia Forrester for the water allocation model details, and to Kenneth

Nowak for sharing the U.S. Bureau of Reclamation simulation data on Lake Mead elevation. An earlier version of this work has been presented at ISERC [*Love and Bayraksan*, 2014]. The first author acknowledges support provided by a Water Sustainability Fellowship from the University of Arizona. This research has been supported in part by the National Science Foundation [Grants CMMI-1345626 and CMMI-1563504].

References

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999), Coherent measures of risk, *Mathematical Finance*, *9*(3), 203–228.

Bayraksan, G., and D. K. Love (2015), Data-driven stochastic programming using phi-divergences, in *Tutorials in Operations Research*, pp. 1–19, INFORMS, Hanover, MD.

Ben-Tal, A., D. D. Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen (2013), Robust solutions of optimization problems affected by uncertain probabilities, *Management Science*, *59*, 341–357.

Bertsimas, D., V. Gupta, and N. Kallus (2017), Robust sample average approximation, *Mathematical Programming*, published online before print, URL https://doi.org/10.1007/s10107-017-1174-z.

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

Brekke, L., B. L. Thrasher, E. P. Maurer, and T. Pruitt (2013), Downscaled CMIP3 and CMIP5 climate projections: Release of downscaled CMIP5 climate projections, comparison with preceding information, and summary of user needs, U.S. Department of the Interior, Bureau of Reclamation, Technical Services Center, Denver, CO, URL https://gdo-dcp.ucllnl.org/downscaled_cmip_projections/, Last accessed: March 15, 2018.

Brown, C., Y. Ghile, M. Laverty, and K. Li (2012), Decision scaling: Linking bottom-up vulnerability analysis with climate projections in the water sector, *Water Resources Research*, *48*(9), W09537.

Calatrava, J., and A. Garrido (2005), Spot water markets and risk in water supply, *Agricultural Economics*, *33*(2), 131–143.

Chung, G., K. Lansey, and G. Bayraksan (2009), Reliable water supply system design under uncertainty, *Environmental Modelling & Software*, *24*(4), 449 – 462.

City of Tucson (2008), Update to water plan: 2000-2050, URL https://www.tucsonaz.gov/files/water/docs/wp08-update.pdf, Last accessed: March 15, 2018.

Connell-Buck, C., J. Medellín-Azuara, J. Lund, and K. Madani (2011), Adapting California's water system to warm vs. dry climates, *Climatic Change*, *109*(1), 133–149.

Dinar, A., and J. Letey (1991), Agricultural water marketing, allocative efficiency, and drainage reduction, *Journal of Environmental Economics and Management*, *20*(3), 210 – 223.

Escudero, L. (2000), WARSYP: a robust modeling approach for water resources system planning under uncertainty, *Annals of Operations Research*, *95*(1), 313–339.

Gauvin, C., E. Delage, and M. Gendreau (2017), Decision rule approximations for the risk averse reservoir management problem, *European Journal of Operational Research*, *261*(1), 317–336.

Hall, W., and D. Howell (1970), Optimal allocation of stochastic water supply, *Journal of the Irrigation and Drainage Division*, *96*(4), 395–402.

Harou, J., J. Medellín-Azuara, T. Zhu, S. Tanaka, J. Lund, S. Stine, M. Olivares, and M. Jenkins (2010), Economic consequences of optimized water management for a prolonged, severe drought in California, *Water Resources Research*, *46*(5).

Higgins, A., A. Archer, and S. Hajkowicz (2008), A stochastic non-linear programming model for a multi-period water resource allocation with multiple objectives, *Water Resources Management*, *22*(10), 1445–1460.

Housh, M., A. Ostfeld, and U. Shamir (2011), Optimal multiyear management of a water supply system under uncertainty: Robust counterpart approach, *Water Resources Research*, *47*(10).

Jia, Y., and T. B. Culver (2006), Robust optimization for total maximum daily load allocations, *Water Resources Research*, *42*(2).

Jiang, R., and Y. Guan (2016), Data-driven chance constrained stochastic program, *Mathematical Programming*, *158*(1-2), 291–327.

Johnson, R., and D. Kempthorne (2007), Record of Decision–Colorado River Interim Guidelines for Lower Basin Shortages and the Coordinated Operations for Lake Powell and Lake Mead, U.S. Department of the Interior, Bureau of Reclamation, Washington D.C.

Kasprzyk, J. R., S. Nataraj, P. M. Reed, and R. J. Lempert (2013), Many objective robust decision making for complex environmental systems undergoing change, *Environmental Modelling & Software*, *42*, 55–71.

Lan, F., W. H. Lin, and K. Lansey (2015), Scenario-based robust optimization of a water supply system under risk of facility failure, *Environmental Modelling & Software*, *67*, 160 – 172.

Lan, F., G. Bayraksan, and K. Lansey (2016), Reformulation linearization technique based branch-and-reduce approach applied to regional water supply system planning, *Engineering Optimization*, *48*(3), 454–475.

Lee, J.-H., and J. W. Labadie (2007), Stochastic optimization of multireservoir systems via reinforcement learning, *Water Resources Research*, *43*(11), W11408.

Li, Y., G. Huang, and X. Chen (2009), Multistage scenario-based interval-stochastic programming for planning water resources allocation, *Stochastic Environmental Research and Risk Assessment*, *23*(6), 781–792.

Li, Y. P., and G. H. Huang (2008), Interval-parameter two-stage stochastic nonlinear programming for water resources management under uncertainty, *Water Resources Management*, *22*(6), 681–698.

Love, D., and G. Bayraksan (2016), Phi-divergence constrained ambiguous stochastic programs for data-driven optimization, *Optimization Online, http://www.optimization-online.org/DB_FILE/2016/03/5350.pdf*.

Love, D. K., and G. Bayraksan (2013), Two-stage likelihood robust linear program with application to water allocation under uncertainty, in *Proceedings of the Winter Simulation Conference*.

Love, D. K., and G. Bayraksan (2014), A data-driven method for robust water allocation under uncertainty, in *Proceedings of the Industrial & Systems Engineering Research Conference (ISERC)*, edited by Y. Guan and H. Liao.

Martin, L. (2013), Direct potable reuse vs. indirect: Weighing the pros and cons, *Water Online*, URL https://www.wateronline.com/doc/direct-potable-reuse-vs-indirect-weighing-the-pros-and-cons-0001, Last accessed: March 15, 2018.

Mulvey, J. M., R. J. Vanderbei, and S. A. Zenios (1995), Robust optimization of large-scale systems, *Operations Research*, *43*(2), 264–281.

Murali, K., M. K. Lim, and N. C. Petruzzi (2015), Municipal groundwater management: Optimal allocation and control of a renewable natural resource, *Production and Operations Management*, *24*(9), 1453–1472.

Nowak, K. (2014), Hydrologic Engineer, U.S. Bureau of Reclamation, Lower Colorado Region, personal communication.

O'Hara, J. K., and K. P. Georgakakos (2008), Quantifying the urban water supply impacts of climate change, *Water Resources Management*, *22*(10), 1477–1497.

Ormerod, K. J., and C. A. Scott (2013), Drinking wastewater, *Science, Technology, & Human Values*, *38*(3), 351–373.

Pachauri, R. K., M. R. Allen, V. R. Barros, J. Broome, W. Cramer, R. Christ, J. A. Church, L. Clarke, Q. Dahe, P. Dasgupta, et al. (2014), *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*, IPCC.

Pan, L., M. Housh, P. Liu, X. Cai, and X. Chen (2015), Robust stochastic optimization for reservoir operation, *Water Resources Research*, *51*(1), 409–429.

Pardo, L. (2005), *Statistical Inference Based On Divergence Measures*, Chapman and Hall/CRC.

Perelman, L., M. Housh, and A. Ostfeld (2013), Robust optimization for water distribution systems least cost design, *Water Resources Research*, *49*(10), 6795–6809.

Philpott, A., V. de Matos, and L. Kapelevich (2017), Distributionally robust sddp, *Tech. rep.*, URL http://www.epoc.org.nz/papers/DROPaperv52.pdf.

Pima Association of Governments (2012), Annual traffic count program, URL http://www.pagnet.org/RegionalData/Maps/MapsandGISDownloads/tabid/902/Default.aspx, Last accessed: March 15, 2018.

Rahimian, H., G. Bayraksan, and T. H. de Mello (2018), Identifying effective scenarios in distributionally robust stochastic programs with total variation distance, *Mathematical Programming*, published online before print https://doi.org/10.1007/s10107-017-1224-6.

Reclamation (2013), Downscaled CMIP3 and CMIP5 climate and hydrology projections: Release of downscaled cmip5 climate projections, comparison with preceding information, and summary of user needs, *Tech. rep.*, prepared by the U.S. Department of the Interior, Bureau of Reclamation, Technical Services Center, Denver, Colorado.

Rosenberg, D. E., and J. R. Lund (2009), Modeling integrated decisions for a municipal water system with recourse and uncertainties: Amman, Jordan, *Water Resources Management*, *23*(1), 85–115.

Scott, C., C. Bailey, R. Marra, G. Woods, K. Ormerod, and K. Lansey (2012), Scenario planning to address critical uncertainties for robust and resilient water–wastewater infrastructures under conditions of water scarcity and rapid development, *Water*, *4*(4), 848–868.

Shao, L., X. Qin, and Y. Xu (2011), A conditional value-at-risk based inexact water allocation model, *Water Resources Management*, *25*(9), 2125–2145.

Shapiro, A., D. Dentcheva, and A. Ruszczyński (2009), *Lectures on stochastic programming: modeling and theory*, MPS-SIAM series on optimization, Society for Industrial and Applied Mathematics, Philadelphia, USA.

Singh, A. (2012), An overview of the optimization modelling applications, *Journal of Hydrology*, *466*(Supplement C), 167 – 182.

Singh, R., T. Wagener, R. Crane, M. E. Mann, and L. Ning (2014), A vulnerability driven approach to identify adverse climate and land use change combinations for critical hydrologic indicator thresholds: Application to a watershed in Pennsylvania, USA, *Water Resources Research*, *50*(4), 3409–3427.

Tanaka, S. K., T. Zhu, J. R. Lund, R. E. Howitt, M. W. Jenkins, M. A. Pulido, M. Tauber, R. S. Ritzema, and I. C. Ferreira (2006), Climate warming and water management adaptation for California, *Climatic Change*, *76*(3), 361–387.

Tucson Water (2013), personal communication.

Udall, B., and J. Overpeck (2017), The twenty-first century Colorado River hot drought and implications for the future, *Water Resources Research*, *53*(3), 2404–2418.

U.S. Bureau of Reclamation (2016), Lake Mead at Hoover Dam, elevation (feet), URL http://www.usbr.gov/lc/region/g4000/hourly/mead-elv.html, Last accessed: March 15, 2018.

U.S. Census Bureau (2010), The 2010 U.S. Census., URL https://www.census.gov/2010census/popmap, Last accessed: March 15, 2018.

U.S. Department of the Interior Bureau of Reclamation (2012), Colorado river basin water supply and demand study: Technical report G — system reliability analysis and evaluation of options and strategies, URL https://www.usbr.gov/lc/region/programs/crbstudy/finalreport/techrptG.html.

Vicuna, S., J. A. Dracup, J. R. Lund, L. L. Dale, and E. P. Maurer (2010), Basin-scale water system operations with uncertain future climate conditions: Methodology and case studies, *Water Resources Research*, *46*(4), w04505.

Watkins Jr, D. W., and D. C. McKinney (1997), Finding robust solutions to water resources problems, *Journal of Water Resources Planning and Management*, *123*(1), 49–58.

Weinberg, M., C. L. Kling, and J. E. Wilen (1993), Water markets and water quality, *American Journal of Agricultural Economics*, *75*(2), 278–291.

WISP (2009), Location of growth, urban form, and cost of infrastructure, URL https://webcms.pima.gov/UserFiles/Servers/Server_6/File/Government/Wastewater%20Reclamation/Water%20Resources/WISP/062509-Growth.pdf, Last accessed: March 15, 2018.

Woods, G., D. Kang, D. Quintanar, E. Curley, S. Davis, K. Lansey, and R. Arnold (2012), Centralized versus decentralized wastewater reclamation in the Houghton area of Tucson, Arizona, *Journal of Water Resources Planning and Management*, *139*(3), 313–324.

Yan, D., S. E. Werners, H. Q. Huang, and F. Ludwig (2016), Identifying and assessing robust water allocation plans for deltas under climate change, *Water Resources Management*, *30*(14), 5421–5435.

Zhang, W., H. Rahimian, and G. Bayraksan (2016), Decomposition algorithms for risk-averse multistage stochastic programs with application to water allocation under uncertainty, *INFORMS Journal on Computing*, *28*(3), 385–404.

Supporting Information for "Data-Driven Water Allocation under Climate Uncertainty: A Distributionally Robust Approach"

David K. Love,[1] Jangho Park,[2] Güzin Bayraksan[2]

Contents of this file

1. **Text S1.** Detailed Mathematical Formulation of Distributionally Robust Water Allocation Model for Our Application

2. **Text S2.** Detailed Regression with Time Series Error

3. **Figure S1.** The lower- and higher-GPCD demand projection for the climate model CSIRO-mk-3-6-0, with greenhouse concentration pathway RCP2.6.

4. **Figure S2.** Total operating cost with shortage cost $\in \{1,300; 1,400; \ldots; 4,250; 4,900\}$ (Kullback-Leibler, 95%).

Corresponding author: Güzin Bayraksan, Department of Integrated Systems Engineering, 246 Baker Systems Engineering, 1971 Neil Avenue Columbus, OH 43210, USA. (bayraksan.1@osu.edu)

[1]American Express, New York, NY, USA.

[2]Integrated Systems Engineering, the Ohio State University, Columbus, OH, USA.

5. **Table S1.** Results of the lower-GPCD water demand regression.

6. **Table S2.** Results of the higher-GPCD water demand regression.

7. **Table S3.** Projected average daily high temperatures and average precipitation of different climate models.

8. **Table S4.** Population estimates for the RESIN and Tucson Water service areas.

Additional Supporting Information (Files uploaded separately)

1. **Dataset S1.** [*drwa-ds1.xlsx*] Raw and intermediate data.

2. **Dataset S2.** [*drwa-ds2.xlsx*] Final dataset used in optimization. For every scenario, provides (i) nominal probability of scenario, and per each year 2014–2050 for every scenario provides (ii) water supply to RESIN area, and (iii) water demands by zones.

Introduction

This supporting information lists a detailed mathematical formulation of the distributionally robust water allocation application (**Text S1**), a detailed regression model with time series error (**Text S2**), and provides additional information regarding the data-driven methodology used in the paper.

In particular, **Tables S1 and S2** of the supporting information tabulate the final regression coefficients with seasonal ARIMA errors for lower-GPCD and higher-GPCD water demands, respectively. **Figure S1** shows the predicted per-capita demands from both regressions through 2050 for one climate model, CSIRO-mk-3-6-0, with one greenhouse concentration pathway, RCP2.6.

These regression functions take input the results from the climate models and greenhouse concentration pathways. **Table S3** lists projected (2040–2050) average daily high

temperatures ($^{\circ}C$) and average precipitation per month (mm) from all the climate models used in this study and compares these projections with the historical record (1991–2011). All forecasts in this table use greenhouse concentration pathway RCP8.5, and all values are for the RESIN area.

Table S4 shows the population estimates for the RESIN and Tucson Water service areas. These population numbers contain the U.S. Census population of 2010 as well as projections by mid-century (2050). **Table S4** also provides the data sources for the population studies in the area.

Finally, **Figure S2** depicts total operating cost with shortage costs of 1,300, 1,400, ..., 4,250, 4,900 for all three considered options, **NI**, **WWTP**, and **IPR**.

Dataset S2 provides the 6,144 scenarios that are final inputs to the distributionally robust optimization model with detailed scenario characteristics. **Dataset S2** is predicted from raw and intermediate data sets available in **Dataset S1**.

All in all, this supporting information's tables and datasets—together with Section 4.2 of the paper—completely cover Figure 3 (Flowchart of data-driven methodology to predict annual demand and supply to RESIN area) of the main paper. It also provides detailed problem formulation and analysis results.

Text S1. Detailed Mathematical Formulation of Distributionally Robust Water Allocation Model for Our Application

We categorize all nodes of the optimization model into five sets—pumps and reservoir (PR), water treatment plant (TP), potable municipal user (PU), non-potable municipal user (NU), recharge facility (RF) and Central Arizona Project (CAP). Also, we define

capacity values of node i at t year—recharge facility ($U_{i,t}^{RF}$), pumping capacity at treatment plan or recharge facility ($U_{i,t}^{RFTP}$), and treatment storage capacity ($U_{i,t}^{TP}$). The value U_t^{CAP} is capacity on the CAP water allotment node at time t. Finally, $d_{i,t}$ is a demand of user node i at time t, $S_{i,0}$ is an initial storage at i recharge facility, and l is a portion of potable municipal user demand returned to a wastewater treatment facility. For a scenario ω in the second stage, $U_t^{CAP,\omega}$ and $d_{i,t}^{\omega}$ indicate each a CAP water allotment at time t and a demand of user node i at time t. The minimax formulation is

$$\min_{(\mathbf{x},\mathbf{s})\in X} \max_{p\in\mathcal{P}} \left\{ \sum_{(i,j)\in A} \sum_{t=1}^{|P_1|} c_{ij,t}^{x} x_{ij,t} + \sum_{j\in N} \sum_{t=1}^{|P_1|} c_{j,t}^{s} s_{j,t} + \sum_{\omega=1}^{n} p_\omega h_\omega(\mathbf{s}) \right\},$$

where the set of distribution $\mathcal{P}$ is given by

$$\mathcal{P} = \left\{ p : \sum_{\omega=1}^{n} q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right) \le \rho, \sum_{\omega=1}^{n} p_\omega = 1, p_\omega \ge 0, \ \forall\omega \right\}.$$

The set X is given by $(\mathbf{x}, \mathbf{s})$ that satisfy:

$$\sum_{j:(j,i)\in A} a_{ji,t} x_{ji,t} = \sum_{j:(i,j)\in A} x_{ij,t} \quad \forall i \in PR \cup TP, 1 \le t \le |P_1| \tag{1}$$

$$\sum_{j:(j,i)\in A} a_{ji,t} x_{ji,t} = d_{i,t} \quad \forall i \in PU \cup NU, 1 \le t \le |P_1| \tag{2}$$

$$\sum_{j:(j,i)\in A} a_{ji,1} x_{ji,1} + S_{i,0} = \sum_{j:(i,j)\in A} x_{ij,1} + s_{i,1} \quad \forall i \in RF, \tag{3}$$

$$\sum_{j:(j,i)\in A} a_{ji,t} x_{ji,t} + s_{i,t-1} = \sum_{j:(i,j)\in A} x_{ij,t} + s_{i,t} \quad \forall i \in RF, 2 \le t \le |P_1| \tag{4}$$

$$\sum_{j:(i,j)\in A} x_{ij,1} \le S_{i,0} \quad \forall i \in RF, \tag{5}$$

$$\sum_{j:(i,j)\in A} x_{ij,t} \le s_{i,t-1} \quad \forall i \in RF, 2 \le t \le |P_1| \tag{6}$$

$$\sum_{j:(i,j)\in A} x_{ij,t} \le U_t^{CAP} \quad \forall i = CAP, 1 \le t \le |P_1| \tag{7}$$

$$\sum_{j:(i,j)\in A} x_{ij,t} \le U_{i,t}^{RFTP} \quad \forall i \in RF \cup TP, 1 \le t \le |P_1| \tag{8}$$

$$\sum_{j:(i,j)\in A} x_{ji,t} \le U_{i,t}^{TP} \quad \forall i \in TP, 1 \le t \le |P_1| \tag{9}$$

$$x_{ij,t} = l \cdot d_{i,t} \quad \forall i \in PU, (i,j) \in A, 1 \le t \le |P_1| \tag{10}$$

$$s_{i,t} \leq U_{i,t}^{RF} \quad \forall i \in RF, 1 \leq t \leq |P_1| \tag{11}$$

$$s_{i,t} \geq 0 \quad \forall i \in RF, 1 \leq t \leq |P_1| \tag{12}$$

$$x_{ij,t} \geq 0, \quad \forall j : (i,j) \in A, 1 \leq t \leq |P_1|. \tag{13}$$

The uncertain second-stage optimal value $h_\omega(\mathbf{s})$ is given by

$$\begin{aligned}
h_\omega(\mathbf{s}) = \quad & \min_{\mathbf{x}^\omega, \mathbf{s}^\omega} \sum_{(i,j)\in A} \sum_{t=|P_1|+1}^{P} c_{ij,t}^x x_{ij,t}^\omega + \sum_{j\in N} \sum_{t=|P_1|+1}^{P} c_{j,t}^s s_{j,t}^\omega \\
\text{s.t.} \quad & \sum_{j:(j,i)\in A} a_{ji,t} x_{ji,t}^\omega = \sum_{j:(i,j)\in A} x_{ij,t}^\omega \quad \forall i \in PR \cup TP, |P_1|+1 \leq t \leq |P| && (14)\\
& \sum_{j:(j,i)\in A} a_{ji,t} x_{ji,t}^\omega = d_{i,t}^\omega \quad \forall i \in PU \cup NU, |P_1|+1 \leq t \leq |P| && (15)\\
& \sum_{j:(j,i)\in A} a_{ji,|P_1|+1} x_{ji,|P_1|+1}^\omega + s_{i,|P_1|} = \sum_{j:(i,j)\in A} x_{ij,|P_1|+1}^\omega + s_{i,|P_1|+1}^\omega \quad \forall i \in RF, && (16)\\
& \sum_{j:(j,i)\in A} a_{ji,t} x_{ji,t}^\omega + s_{i,t-1}^\omega = \sum_{j:(i,j)\in A} x_{ijt}^\omega + s_{it}^\omega \quad \forall i \in RF, |P_1|+2 \leq t \leq |P| && (17)\\
& \sum_{j:(i,j)\in A} x_{ij,|P_1|+1}^\omega \leq s_{i,|P_1|} \quad \forall i \in RF, && (18)\\
& \sum_{j:(i,j)\in A} x_{ij,t}^\omega \leq s_{i,t-1}^\omega \quad \forall i \in RF, |P_1|+2 \leq t \leq |P| && (19)\\
& \sum_{j:(i,j)\in A} x_{ij,t}^\omega \leq U_t^{CAP,\omega} \quad \forall i = CAP, |P_1|+1 \leq t \leq |P| && (20)\\
& \sum_{j:(i,j)\in A} x_{ij,t}^\omega \leq U_{i,t}^{RFTP} \quad \forall i \in RF \cup TP, |P_1|+1 \leq t \leq |P| && (21)\\
& \sum_{j:(i,j)\in A} x_{ji,t}^\omega \leq U_{i,t}^{TP} \quad \forall i \in TP, |P_1|+1 \leq t \leq |P| && (22)\\
& x_{ij,t}^\omega = l \cdot d_{i,t}^\omega \quad \forall i \in PU, (i,j) \in A, |P_1|+1 \leq t \leq |P| && (23)\\
& s_{i,t}^\omega \leq U_{i,t}^{RF} \quad \forall i \in RF, |P_1|+1 \leq t \leq |P| && (24)\\
& s_{i,t}^\omega \geq 0 \quad \forall i \in RF, |P_1|+1 \leq t \leq |P| && (25)\\
& x_{ij,t}^\omega \geq 0, \quad \forall j : (i,j) \in A, |P_1|+1 \leq t \leq |P| && (26)
\end{aligned}$$

The constraints are: (1) flow balance constraints on nodes include (i) water flow balance at pumps/water treatment plant/reservoir/interconnection point {(1), (14)}; (ii) demand satisfaction at potable/non-potable municipal user {(2), (15)}; (iii) water storage balance

at aquifer recharge facilities {(3), (4), (16), (17)}, infiltration needs a one-year {(5), (6), (18), (19)}; and (2) maximum capacity constraints on nodes (i) bounds on the Colorado River water supply depending on scenario {(7), (20)}; (ii) bounds on the in/out-flow of recharge facilities and treatment plants {(8), (9), (21), (22)}; and finally (3) constraints regarding arcs that entail (i) a fixed portion of the potable used water is returned to a wastewater treatment plant {(10), (23)}—the treated water can be used only for non-potable municipal user demand for later years—(ii) upper bound and non-negativity constraints on the water flows {(11), (12), (13), (24), (25), (26)}.

Text S2. Detailed Regression with Time Series Error

In the below regressions, t indicates the monthly time period. From 12 years of historical data from 1991 to 2011, we have $t = \{1, \ldots, 252\}$ monthly data to conduct the regressions. For each time period t, we define the dependent variable GPCD_t and regressors Temperature_t, Precipitation_t, Year_t, and binary indicator variables for each month, $I_1, \ldots, I_{12}$. For example, $\{I_1, \ldots, I_{12}\} = \{1, 0, \ldots, 0\}$ represents January. We first regress GPCD on the above predictor variables. However, the resulting residuals do not pass the autocorrelation tests and still show seasonality. To correct this issue, we use seasonal autocorrelation models on the residuals.

Define residual at time t as

$$r_t = \text{GPCD}_t - \mathbf{X}_t' \vec{\boldsymbol{\beta}},$$

where $\mathbf{X}_t = [\text{Temperature}_t, \text{Precipitation}_t, \text{Year}_t, I_1, \ldots, I_{12}]$ and $\vec{\boldsymbol{\beta}} = \{\beta_1, \ldots, \beta_{15}\}$ are estimated parameters of Temperature, Precipitation, Year, monthly indicator variables. Let ψ_1 and Ψ_1 be estimated parameters—for autoregressive AR(1) and seasonal autoregressive

SAR(1), respectively—of ARIMA $(1,\ 0\ ,0)\times(1,\ 0\ ,0)_{11}$ on the residual. This time-series model with lag operator B is

$$(1-\psi_1 B)(1-\Psi_1 B^{11})r_t = u_t, \tag{27}$$

where the random noise u_t follows a Normal distribution with mean zero and constant variance (per usual assumptions on errors). Equation (27) is equivalent to

$$r_t = \psi_1 r_{t-1} + \Psi_1 r_{t-11} - \psi_1\Psi_1 r_{t-12} + u_t. \tag{28}$$

Putting this all together, for the out-of-sample prediction, we use

$$\begin{aligned}\mathrm{GPCD}_t \ = \ & \mathbf{X}_t'\vec{\boldsymbol{\beta}} + \psi_1\left(\mathrm{GPCD}_{t-1} - \mathbf{X}_{t-1}'\vec{\boldsymbol{\beta}}\right) + \Psi_1\left(\mathrm{GPCD}_{t-11} - \mathbf{X}_{t-11}'\vec{\boldsymbol{\beta}}\right) \\ & -\psi_1\Psi_1\left(\mathrm{GPCD}_{t-12} - \mathbf{X}_{t-12}'\vec{\boldsymbol{\beta}}\right).\end{aligned} \tag{29}$$

The estimated parameters $\vec{\boldsymbol{\beta}}, \psi_1, \Psi_1$ are shown in **Table S1** and **Table S2**. The results of these predictions for one climate model, CSIRO-mk-3-6-0, with one greenhouse concentration pathway, RCP2.6, are shown in **Figure S1**.

Dataset S1. [*drwa-ds1.xlsx*] Raw and intermediate data to predict future water supply and demand.

Dataset S2. [*drwa-ds2.xlsx*] Final dataset used in optimization. For every scenario, provides (i) nominal probability of scenario, and per each year 2014–2050 for every scenario provides (ii) water supply to RESIN area, and (iii) water demands by zones.

References

City of Tucson (2008), Update to water plan: 2000–2050, `https://www.tucsonaz.gov/files/water/docs/wp08-update.pdf`, Last accessed: March 15, 2018.

Pima Association of Governments (2012), Annual traffic count program, http://www.pagnet.org/RegionalData/Maps/MapsandGISDownloads/tabid/902/Default.aspx, Last accessed: March 15, 2018.

WISP (2009), Location of growth, urban form, and cost of infrastructure, http://webcms.pima.gov/government/wastewaterreclamation/waterresources/wisp/, Last accessed: March 15, 2018.

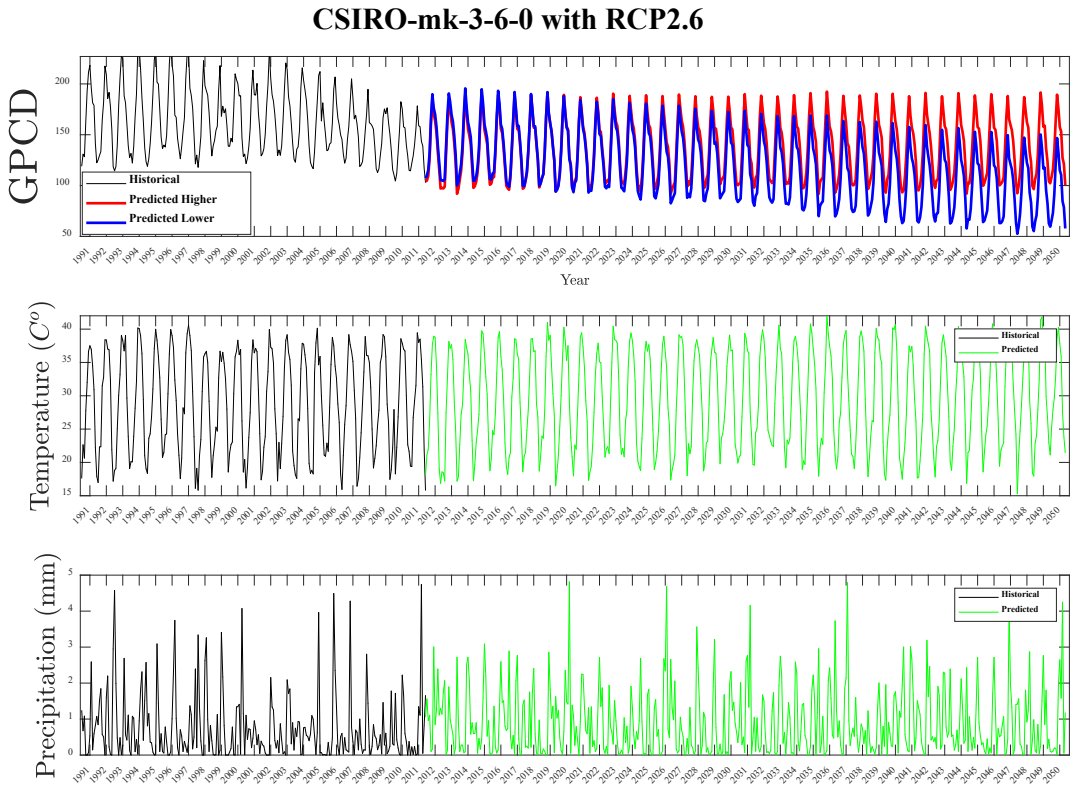


Figure S1. The lower- and higher-GPCD demand projection for the climate model CSIRO-mk-3-6-0, with greenhouse concentration pathway RCP2.6.

Table S1. Results of the lower-GPCD water demand regression[a]

Parameter	[Variable]	Estimate	p-value
ψ_1	[AR(1)]	0.5924	0.00e-00
Ψ_1	[SAR(1)]	0.2848	1.08e-05
β_1	[Temperature]	1.0419	2.70e-08
β_2	[Precipitation]	-2.7854	6.56e-08
β_3	[Year]	-1.3783	3.11e-08

Parameter	[Variable]	Estimate	p-value
β_4	$[I_1]$	2863.3435	9.20e-09
β_5	$[I_2]$	2865.2054	8.98e-09
β_6	$[I_3]$	2869.9794	8.48e-09
β_7	$[I_4]$	2886.6418	6.95e-09
β_8	$[I_5]$	2912.7050	5.09e-09
β_9	$[I_6]$	2931.9489	4.04e-09
β_{10}	$[I_7]$	2921.7047	4.57e-09
β_{11}	$[I_8]$	2908.4331	5.36e-09
β_{12}	$[I_9]$	2903.0661	5.71e-09
β_{13}	$[I_{10}]$	2893.3416	6.40e-09
β_{14}	$[I_{11}]$	2879.4088	7.54e-09
β_{15}	$[I_{12}]$	2864.2649	9.00e-09

[a] Small p-values indicate the estimated variables to be significant.

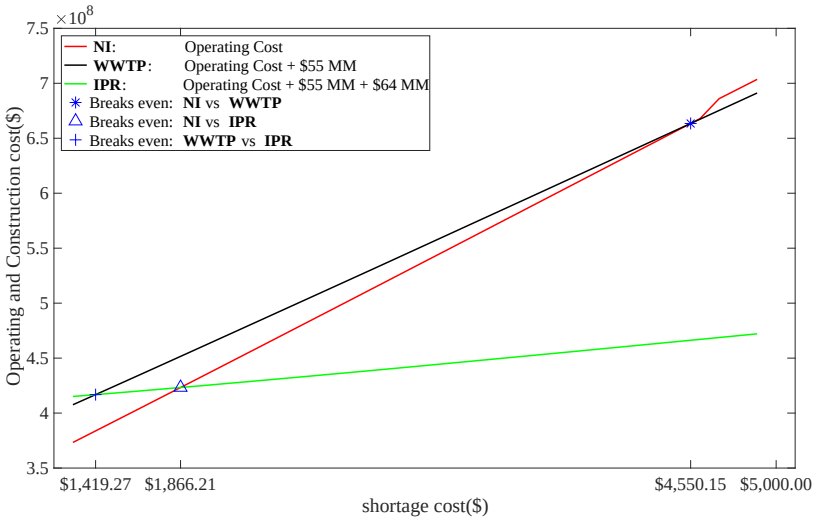


Figure S2. Total operating cost with shortage cost $\in \{1,300; 1,400; \ldots; 4,250; 4,900\}$ (Kullback-Leibler, 95%).

Table S2. Results of the higher-GPCD water demand regression[a]

Parameter	[Variable]	Estimate	p-value
ψ_1	[AR(1)]	0.4852	0.00e-00
Ψ_1	[SAR(1)]	0.2292	2.69e-04
β_1	[Temperature]	1.0601	2.04e-09
β_2	[Precipitation]	-2.8020	4.59e-08
β_3	[Year]	-4.4732	0.00e-00

Parameter	[Variable]	Estimate	p-value
β_4	$[I_1]$	9075.8635	0.00e-00
β_5	$[I_2]$	9077.6571	0.00e-00
β_6	$[I_3]$	9082.3316	0.00e-00
β_7	$[I_4]$	9098.9419	0.00e-00
β_8	$[I_5]$	9124.9206	0.00e-00
β_9	$[I_6]$	9144.0846	0.00e-00
β_{10}	$[I_7]$	9133.8402	0.00e-00
β_{11}	$[I_8]$	9120.5717	0.00e-00
β_{12}	$[I_9]$	9115.2403	0.00e-00
β_{13}	$[I_{10}]$	9105.6550	0.00e-00
β_{14}	$[I_{11}]$	9091.8141	0.00e-00
β_{15}	$[I_{12}]$	9076.7594	0.00e-00

[a] Small p-values indicate the estimated parameters to be significant.

Table S3. A comparison of projected (2040–2050) average daily high temperatures ($°C$) and average precipitation per month (mm) with the historical record (1991–2011) in the RESIN area. All values come from models using concentration path RCP8.5.

Month	Historic		GFDL-CM3		GFDL-ESM2M		CSIRO		HadGEM2-ES		MIROC-ESM-CHEM		MIROC5	
	($°C$)	(mm)	($°C$)	(mm)	($°C$)	(mm)	($°C$)	(mm)	($°C$)	(mm)	($°C$)	(mm)	($°C$)	(mm)
Jan	19.1	22	21.1	19	19.5	31	21.2	13	19.6	37	21.8	12	20.6	25
Feb	20.7	24	23.3	23	22.5	41	21.8	23	22.5	23	23.7	13	22.0	26
Mar	24.0	17	26.1	21	23.6	22	24.9	22	24.5	17	25.8	20	25.2	19
Apr	28.0	8	31.7	2	30.0	12	30.0	9	31.0	3	31.9	4	31.1	5
May	33.5	5	36.0	3	33.9	7	34.2	4	35.3	3	35.8	4	35.9	3
Jun	38.2	4	41.5	4	39.7	6	40.1	8	40.5	7	40.1	8	39.9	6
Jul	38.1	52	41.0	75	39.6	58	39.8	85	40.0	57	40.2	71	39.7	58
Aug	37.0	58	40.3	55	38.4	74	39.0	66	39.6	54	39.0	66	38.8	55
Sep	35.5	33	38.6	32	36.5	35	36.6	50	37.3	32	37.4	45	37.6	35
Oct	30.1	14	33.5	47	32.6	26	32.1	42	34.1	14	31.8	53	32.6	26
Nov	23.7	14	26.4	27	24.3	20	25.5	10	26.2	20	25.4	15	26.0	9
Dec	18.6	23	20.6	20	19.4	25	21.1	31	21.0	30	21.0	25	21.1	25

Table S4. Population estimates for the RESIN and Tucson Water service areas.

		population		
Region	Data Source	2010	2014	2050
RESIN	TAZ [*Pima Association of Governments*, 2012]	37,005	67,549	413,115
	WISP [2009]	37,005	90,129	693,116
Tucson	*City of Tucson* [2008]	750,000	805,000	1,300,000