

Risk Aversion to Parameter Uncertainty in Markov Decision Processes with an Application to Slow-Onset Disaster Relief

Merve Meraklı, Simge Küçükyavuz

Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL, U.S.A.

merakli@northwestern.edu, simge@northwestern.edu

June 21, 2019

ABSTRACT: In classical Markov Decision Processes (MDPs), action costs and transition probabilities are assumed to be known, although an accurate estimation of these parameters is often not possible in practice. This study addresses MDPs under cost and transition probability uncertainty and aims to provide a mathematical framework to obtain policies minimizing the risk of high long-term losses due to not knowing the true system parameters. To this end, we utilize the risk measure value-at-risk associated with the expected performance of an MDP model with respect to parameter uncertainty. We provide mixed-integer linear and nonlinear programming formulations and heuristic algorithms for such risk-averse models of MDPs under a finite distribution of the uncertain parameters. Our proposed models and solution methods are illustrated on an inventory management problem for humanitarian relief operations during a slow-onset disaster. The results demonstrate the potential of our risk-averse modeling approach for reducing the risk of highly undesirable outcomes in uncertain/risky environments.

Keywords: Markov decision processes, parameter uncertainty, value-at-risk, chance constraints, humanitarian supply chains, disaster relief

1. Introduction Markov Decision Processes (MDPs) are effectively used in many applications of sequential decision making in uncertain environments including inventory management, manufacturing, robotics, communication systems, and healthcare, e.g., [Puterman \(2014\)](#); [Altman \(1999\)](#); [Boucherie and van Dijk \(2017\)](#). In an MDP model, the decision makers take an action at specified points in time considering the current state of the system with the aim of minimizing their expected loss (resp., maximizing their expected utility), and depending on the action taken, the system transitions to another state. The evolution of the underlying process is mainly characterized by the action costs (resp., rewards) and transition probabilities between the system states, inducing two types of uncertainty. The *internal uncertainty* stems from the probabilistic behaviour of transitions between states, costs and actions (see, e.g., [Ruszczyński, 2010](#); [Bäuerle and Ott, 2011](#); [Xu and Mannor, 2011](#); [Fan and Ruszczyński, 2018](#) for studies addressing the risk arising from internal uncertainty). The *parameter uncertainty*, on the other hand, is due to the ambiguities in the parameters representing the costs and transition probabilities. In classical MDPs, these parameters are assumed to be known; they are usually estimated from historical data or learned from previous experiences. However, in practice, it is usually not possible to obtain a single estimate that fully captures the nature of the uncertainties. The actual performance of the system may significantly differ from the anticipated performance of the MDP model due to the inherent varia-

tion in the parameters (Mannor et al., 2007). Our focus in this study is on the parameter uncertainty in MDPs—decision makers are assumed to be sensitive to the risk associated with the fluctuations in parameters while being risk-neutral to the internal randomness due to state transitions, costs and randomized actions. This setting is especially suitable for applications in which the objective function is aggregated over a number of problem instances, e.g., total inventory cost over various types of supply items. In such cases, aggregation across multiple instances mitigates the variation due to internal uncertainty, and parameter uncertainty becomes the main source of variation.

A widely used approach to incorporate parameter uncertainty into MDPs is robust optimization. In the robust modeling framework, the objective is to optimize the worst-case performance over all possible realizations in a given uncertainty set. This approach is appealing in the sense that it requires no prior information on the distribution of costs or transition probabilities and it gives rise to computationally efficient solution algorithms. However, it often leads to conservative results because the focus is on the worst-case system performance, which may be rarely encountered in practice. In the initial studies on robust MDPs, uncertainty is usually described using a polyhedral set, because it leads to tractable solution algorithms (Satia and Lave Jr, 1973; White III and Eldeib, 1994; Givan et al., 2000; Tewari and Bartlett, 2007; Bagnell et al., 2001). The uncertainty sets are later extended for more general definitions with the aim of balancing the conservatism of the solutions and the tractability of the solution algorithms. Nilim and El Ghaoui (2005) model the uncertainty in transition probabilities using a set of stochastic matrices satisfying *rectangularity* property, i.e., when there are no correlations between transition probabilities for different states and actions. The authors devise an efficient dynamic programming algorithm for this case. Similarly, Iyengar (2005) studies robust MDPs under transition probability uncertainty with rectangularity assumption and provides robust value and policy iteration algorithms for *finite-horizon* nonstationary and *infinite-horizon* stationary problem settings. Sinha and Ghate (2016) propose a policy iteration algorithm for robust infinite-horizon nonstationary MDPs following the rectangularity assumption. Wiesemann et al. (2013) relax the rectangularity assumption of Nilim and El Ghaoui (2005) and consider a more general class of uncertainty sets in which the assumption of no correlation between transition probabilities is only made for states, not for actions. Mannor et al. (2016) define a tractable subclass of nonrectangular uncertainty sets, namely k -rectangular uncertainty sets, such that the number of possible conditional projections of the uncertainty set is at most k . Another alternative modeling approach to balance the conservatism of the solutions is to consider the distributional robustness, where the uncertain parameters are assumed to follow the worst-case distribution from a set of possible distributions described by some general properties such as expectations or moments. Unlike robust MDPs, distributionally robust MDPs incorporate the available—but incomplete—information on the a priori distribution of the uncertain parameters. For relevant studies on distributionally robust MDPs, we refer the reader to Xu and Mannor (2012); Yu and Xu (2016).

Bayesian approaches to address parameter uncertainty have been receiving increasing attention in the recent literature. This uncertainty model treats unknown parameters as random variables with corresponding probability distributions. Hence, it provides the means to incorporate complete distributional information about the unknown parameters and the attitude of decision makers towards uncertainty and risk, oftentimes at the cost of increasing problem complexity. [Steimle et al. \(2018b\)](#) consider a multi-model MDP, where the aim is to find a policy maximizing the weighted sum of expected total rewards over a finite horizon associated with different sets of parameters obtained by different estimation methods. This modeling framework is analogous to an expected value problem considering a finite number of scenarios in the context of stochastic programming, because each set of parameters can be treated as a scenario in which the corresponding scenario probability is set as the normalized weight value. The authors prove the existence of a deterministic optimal policy, show that the problem is NP-hard, and provide a mixed-integer linear programming (MILP) formulation and a heuristic algorithm. In a subsequent work, [Steimle et al. \(2018a\)](#) propose a customized branch-and-bound algorithm to solve the multi-model MDPs. [Buchholz and Scheftelowitsch \(2018\)](#) study multi-model MDPs, where each model corresponds to an infinite-horizon MDP. Unlike the finite-horizon case, it is demonstrated that infinite-horizon variant may not have a deterministic optimal policy. The authors propose two nonlinear programming formulations and a heuristic algorithm for the case where randomized policies are allowed, and an MILP model utilizing the dual linear programming formulation of MDPs for the deterministic policy case.

A potential drawback of previously mentioned modeling frameworks for multi-model MDPs is that the expected value objective ignores the risk arising from the parameter uncertainty. Considering this issue, [Xu and Mannor \(2009\)](#) address reward uncertainty in MDPs with respect to parametric regret. The MDP of interest is a finite-horizon, discounted model with finite state and action spaces. The authors consider two different objectives: minimax regret based on the robust approach and mean-variance trade-off of the regret based on the Bayesian approach. They propose a nonconvex quadratic program for the former objective and a convex quadratic program for the latter. [Delage and Mannor \(2010\)](#) consider reward and transition probability uncertainty separately and propose a chance-constrained model in the form of percentile optimization, which corresponds to the risk measure, *value-at-risk* (VaR). The authors give a formulation for infinite-horizon MDPs with finite state and action spaces, and stationary policies. They show that the problem is intractable in the general case but can be efficiently solved when the rewards follow a Gaussian distribution or transition probabilities are modeled using independent Dirichlet priors satisfying rectangularity property. Alternatively, [Chen and Bowling \(2012\)](#) investigate a class of percentile-based objective functions that are easy to approximate for any probability distribution and [Adulyasak et al. \(2015\)](#) focus on finding objective functions that are separable over realizations of uncertain parameters under a sampling framework. Our modeling approach for incorporating parameter uncertainty into MDPs is along the same line with [Delage and Mannor \(2010\)](#), however we allow general

probability distributions with finite support. Throughout the manuscript, we use the terms percentile, quantile and VaR interchangeably.

In this study, we study MDPs under cost and transition probability uncertainty. Our aim is to obtain a stationary policy that optimizes the quantile function value, VaR_α , at a certain confidence level α with respect to parameter uncertainty. The VaR_α assesses an estimate of the largest potential loss excluding the worst outcomes with at most $1 - \alpha$ probability. Conservatism of solutions in optimization problems involving VaR_α can be adjusted using different confidence levels α , reflecting decision makers' risk aversion—quantile optimization is equivalent to a robust optimization approach when $\alpha = 1$. In addition, quantile-based performance measures, such as VaR, are used in many applications in the service industry because of their clear interpretation and correspondence with the service-level requirements, e.g., minimum investment to guarantee 100 α % service level (DeCandia et al., 2007; Benoit and Van den Poel, 2009; Atakan et al., 2017). Although VaR is nonconvex in general, the main challenge in our case is the combinatorial nature of policies independent from the choice of the risk measure, which makes the problem NP-hard even for the expected value objective (Steimle et al., 2018b).

Unlike Delage and Mannor (2010), who also consider the VaR objective for cost and transition probability uncertainty cases separately under certain rectangularity assumptions, we assume that action costs and transition probabilities are both uncertain and follow a finite joint distribution without imposing any independence requirements on system parameters. This approach directly represents the cases in which a finite set of possible parameter realizations can be obtained based on historical data and/or multiple estimation tools (Bertsimas and Mišić, 2016; Steimle et al., 2018b). Moreover, it provides a general framework that can be used on sample approximations of a wide range of probability distributions. Finite representation of uncertainty also facilitates competitive solution methods for optimization problems incorporating VaR. Since VaR corresponds to a quantile function, optimization problems incorporating VaR can be formulated as chance-constrained programs (CCPs), which are known to be NP-hard in general and usually require multidimensional integration. When uncertain parameters follow a finite distribution, the challenges of working with multivariate distributions can be circumvented, and the CCP formulations can be stated as mixed-integer programming (MIP) models by employing big- M inequalities and additional binary variables for each possible realization of uncertainty (Luedtke and Ahmed, 2008) (see, e.g., Küçükyavuz, 2012; Liu et al., 2017a; Zhao et al., 2017 for further strengthenings). Majority of the studies on CCPs considers a single-state (i.e., static) decision-making framework, where the uncertainty is revealed only after all required decisions are made. Luedtke (2014) and Liu et al. (2016) extend the literature for a two-stage decision-making framework such that recourse decisions are allowed in the second stage and provide branch-and-cut algorithms employing mixing inequalities to ensure feasibility/optimality of the second-stage problems. Along the same lines, Zhang et al. (2014) consider a multi-stage (finite-horizon) setting and propose a branch-and-cut algorithm using continuous

mixing inequalities. For an overview of CCPs and related approaches, we refer the reader to [Küçükyavuz and Sen \(2017\)](#). Motivated by these advances in the CCP literature, [Feng et al. \(2015\)](#) formulate the VaR portfolio optimization problem as a single-stage CCP and provide an MILP reformulation and a branch-and-cut algorithm utilizing the mixing inequalities of [Luedtke \(2014\)](#). Their results suggest superiority of the MIP approaches over branch-and-cut based algorithms and emphasize the significance of the big- M terms on computational performance of the MIP formulations for VaR optimization. Based on these conclusions, [Pavlikov et al. \(2018\)](#) provide a bounding scheme that produces tighter big- M values for the MILP formulation.

Optimizing VaR associated with parameter uncertainty in MDPs, on the other hand, brings out additional challenges due to the combinatorial nature of the decisions and the underlying Markovian system dynamics in an MDP. In this problem setting, the aim is to obtain a single optimal policy (selected at the beginning) minimizing the VaR associated with parameter uncertainty in an MDP, which is to be implemented over the entire planning horizon under any realization of uncertainty. Different from the previously mentioned studies on CCPs, the underlying process is assumed to be Markovian and possible actions in each state belong to a finite set. For this purpose, we provide a two-stage CCP formulation capable of modeling the dynamics of a Markov chain for any selected policy considering possible values of uncertain parameters. Note that here the second stage represents the performance of the MDP for a given policy. We additionally propose relaxations and heuristic solution algorithms that can be used for obtaining lower and upper bounds on the optimal objective function value. Although we focus on infinite-horizon MDPs, our results and algorithms can be easily extended for finite-horizon MDPs after small adjustments.

We test our modeling framework and solution algorithms on a humanitarian inventory management problem for relief items required to sustain basic needs of a population affected by a slow-onset disaster, e.g., war, political insurgence, extreme poverty, famine, or drought. Since the progress and impact of a slow-onset disaster generally depend on unpredictable political and/or natural events, the demand for the relief items is highly variable. The supply amounts are also exposed to uncertainty as they mainly rely on voluntary donations. Another critical issue in humanitarian inventory management is the perishability of many relief items such as food and medication. At the beginning of each time period, based on the current inventory level, the decision makers need to determine an additional order quantity to minimize the expected total inventory holding, stock-out and disposal costs considering the expiration dates and the uncertainty in supply and demand. This problem can be modeled as an MDP, where the current inventory level represents the state of the system, and the uncertainty in supply, demand, and inventory is captured by the transition probabilities between different states ([Ferreira et al., 2018](#)). However, the cost and transition probability parameters of the MDP model are subject to high level of uncertainty because demand and supply rates and shelf life of perishable relief items used in the estimation of these

parameters may widely fluctuate. The VaR_α objective in this setting has a natural interpretation: to find a replenishment strategy that minimizes the budget required to cover the expected total costs considering all possible parameter realizations with at least α probability.

The rest of this paper is organized as follows. In Section 2, we formulate the MDP problem minimizing α -quantile under cost and transition probability uncertainty, provide MIP models and explore characteristics of the optimal policies. Section 3 presents preprocessing procedures, which can be used for initializing auxiliary parameters of the proposed mathematical models as well as reducing the problem size, and a heuristic solution algorithm. We describe a stochastic inventory management problem for slow-onset disasters in Section 4, which is later used for demonstrating effectiveness of the quantile-optimizing modeling approach and proposed solution methods in Section 5. We summarize our contributions and future research directions in Section 6.

2. Problem Formulation and Structural Properties Consider a discrete-time infinite-horizon MDP model with finite state space \mathcal{H} and finite action space \mathcal{A} . We define $\tilde{c}_i(a)$ as the immediate expected cost of taking action $a \in \mathcal{A}$ in state $i \in \mathcal{H}$ and $\tilde{P}_{ij}(a)$ as the probability of transitioning from state $i \in \mathcal{H}$ to state $j \in \mathcal{H}$ under action $a \in \mathcal{A}$. The future costs are discounted by $\gamma \in [0, 1)$ and the distribution of the initial state is given as $|\mathcal{H}|$ -dimensional vector \mathbf{q} . A stationary policy $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{|\mathcal{H}|})$ refers to a sequence of decision rules π_i describing the action strategy for each state $i \in \mathcal{H}$. When the policy is *randomized*, each element of $|\mathcal{A}|$ -dimensional vector π_i denotes the probability of taking the respective action at each time state $i \in \mathcal{H}$ is encountered. For a *deterministic* policy $\boldsymbol{\pi}$, on the other hand, π_i refers to a unit vector in which only the element corresponding to the action selected in state $i \in \mathcal{H}$ is one. Assuming that the cost and transition probability parameters are nonnegative, stationary and bounded, the expected total discounted cost of the underlying Markov chain for a given policy $\boldsymbol{\pi}$ and known system parameters $(\tilde{\mathbf{c}}, \tilde{\mathbf{P}})$ can be stated as

$$C(\boldsymbol{\pi}, \tilde{\mathbf{c}}, \tilde{\mathbf{P}}) = \mathbb{E}_{\mathbf{x} \in \mathcal{H}} \left(\sum_{t=0}^{\infty} \gamma^t \tilde{c}_{x_t}(\pi_{x_t}) | x_0 \propto \mathbf{q}, \boldsymbol{\pi} \right),$$

where x_0 and x_t denote the initial state and the state of the system at decision epoch $t > 0$, respectively. The term $\tilde{c}_{x_t}(\pi_{x_t})$ corresponds to the immediate expected cost of taking action π_{x_t} in state x_t at time period $t = 0, 1, \dots$. The expectation is taken based on the given policy $\boldsymbol{\pi}$ and the initial state x_0 following probability distribution \mathbf{q} . Let Π be the set of stationary Markov policies. In an MDP model, the aim is to find a policy $\boldsymbol{\pi} \in \Pi$ minimizing the expected total discounted cost, i.e.,

$$\min_{\boldsymbol{\pi} \in \Pi} C(\boldsymbol{\pi}, \tilde{\mathbf{c}}, \tilde{\mathbf{P}}). \quad (1)$$

This problem is known to have a stationary and deterministic optimal solution over the set of all policies, and it can be solved efficiently using several well-known methods such as the value iteration algorithm, the policy iteration algorithm, and linear programming (Puterman, 2014). Using the Bellman equation (Bellman, 2013), problem (1) can be alternatively stated as

$$\min_{\pi \in \Pi, \mathbf{v} \in \mathbb{R}^{|\mathcal{H}|}} \sum_{i \in \mathcal{H}} q_i v_i \quad (2a)$$

$$\text{s.t. } v_i = \tilde{c}_i(\pi_i) + \gamma \sum_{j \in \mathcal{H}} \tilde{P}_{ij}(\pi_i) v_j, \quad i \in \mathcal{H}, \quad (2b)$$

where v_i is the expected sum of discounted costs under the selected policy π when starting from state $i \in \mathcal{H}$, which satisfies Bellman optimality condition

$$v_i = \min_{a \in \mathcal{A}} \left\{ \tilde{c}_i(a) + \gamma \sum_{j \in \mathcal{H}} \tilde{P}_{ij}(a) v_j \right\}, \quad i \in \mathcal{H}. \quad (3)$$

The cost vector $\tilde{\mathbf{c}}$ and transition probability matrix $\tilde{\mathbf{P}}$ in MDP model (1) are assumed to be known. However, in practice, it is usually difficult to predict exact values of these parameters. For example, the prices or the weekly demand rate for relief items during a slow-onset disaster (e.g., war) are subject to a high level of uncertainty due to various factors. In addition, for cases where rare events may have a tremendous impact—as in the case of disaster management—the decision makers may prefer to incorporate their aversion towards risk into the decision support tools. Motivated by these arguments, in this study, we consider the setting where the elements of the cost vector $\tilde{\mathbf{c}}$ and transition probability matrix $\tilde{\mathbf{P}}$ are assumed to be random variables instead of known parameters. The decision makers need to determine a policy $\pi \in \Pi$ in this uncertain environment with the aim of minimizing their risk of realizing a large amount of expected total discounted cost with respect to the uncertainty in parameters. In terms of stochastic programming terminology, policy decisions can be thought of as *non-anticipative*, that is, a policy minimizing the risk is selected at the beginning in the presence of uncertainty, and it will be implemented throughout the planning horizon for any realization of parameters $\tilde{\mathbf{c}}$ and $\tilde{\mathbf{P}}$, the values of which are not revealed. We seek a policy $\pi \in \Pi$ minimizing the α -quantile of the expected total discounted cost with respect to parameter uncertainty, which corresponds to the optimal policy of the quantile-minimizing MDP problem

$$\begin{aligned} (\text{QMDP}) \quad & \min_{y \in \mathbb{R}, \pi \in \Pi} y \\ & \text{s.t. } \mathbb{P}_{\tilde{\mathbf{c}}, \tilde{\mathbf{P}}} \left(C(\pi, \tilde{\mathbf{c}}, \tilde{\mathbf{P}}) \leq y \right) \geq \alpha. \end{aligned}$$

Formulation QMDP ensures that the expected total discounted cost for the optimal policy π^* is less than

or equal to the optimal objective function value y^* with probability at least α under the distributions of $\tilde{\mathbf{c}}$ and $\tilde{\mathbf{P}}$. Note that the optimal value y^* corresponds to the VaR of the expected total discounted cost for the optimal policy π^* at confidence level α . Such formulations optimizing VaR (α -quantile) are also referred to as quantile or percentile optimization in the literature. [Delage and Mannor \(2010\)](#) consider special cases of the quantile optimization problem QMDP in which cost and transition probability uncertainty are treated separately. The authors assume that the cost parameters follow a Gaussian distribution in the former case, and the transition probabilities in the latter are modeled using Dirichlet priors. Different than their model, we consider the cost and transition probability uncertainty simultaneously without making any assumptions on the distributions of uncertain parameters other than that it can be represented/approximated with a finite discrete distribution.

Using the nominal MDP model in (2), we obtain an alternative CCP formulation for problem QMDP,

$$\min_{y \in \mathbb{R}, \pi \in \Pi, \mathbf{v} \in \mathbb{R}^{|\mathcal{H}|}} y \quad (5a)$$

$$s.t. \quad \mathbb{P}_{\tilde{\mathbf{c}}, \tilde{\mathbf{P}}} \left(\sum_{i \in \mathcal{H}} q_i v_i^{(\tilde{\mathbf{c}}, \tilde{\mathbf{P}})} \leq y \right) \geq \alpha. \quad (5b)$$

$$v_i^{(\tilde{\mathbf{c}}, \tilde{\mathbf{P}})} = \tilde{c}_i(\pi_i) + \gamma \sum_{j \in \mathcal{H}} \tilde{P}_{ij}(\pi_i) v_j^{(\tilde{\mathbf{c}}, \tilde{\mathbf{P}})} \quad i \in \mathcal{H}. \quad (5c)$$

The $|\mathcal{H}|$ -dimensional random vector $\mathbf{v}^{(\tilde{\mathbf{c}}, \tilde{\mathbf{P}})}$ is composed of random variables $v_i^{(\tilde{\mathbf{c}}, \tilde{\mathbf{P}})}$, whose values depend on the joint realization of random parameters $\tilde{\mathbf{c}}$ and $\tilde{\mathbf{P}}$ and the selected policy π . In general, such problems with chance constraints are highly challenging to solve since they require computation of the joint distribution function, which usually involves numerical integration in multidimensional spaces ([Deák, 1988](#)). On the other hand, using a discrete representation of the distribution function obtained by a sampling method significantly reduces the computational complexity and provides reliable approximations to CCPs for a sufficiently large sample size ([Calafiore and Campi, 2006](#); [Luedtke and Ahmed, 2008](#)). In stochastic optimization, each sample of parameters is referred to as a *scenario*. Note that this approach yields optimal solutions for multi-model MDPs in which parameter uncertainty can be finitely discretized. For example in medical applications for designing optimal treatment and screening protocols, the system state usually represents patient health status and transition probabilities between states can be computed using multiple tools from the clinical literature, which often produce different sets of parameters ([Steimle et al., 2018b](#)). In this context, the parameters computed using each tool can be treated as a scenario. Similarly, for the humanitarian inventory management problem, each scenario may correspond to a set of parameters computed using a different demand/supply forecasting method. In this case, the VaR objective can be interpreted as minimizing the worst-case expected total cost over α fraction of the possible MDP parameters (scenarios).

Another challenge in solving problem (5) is the possibility that none of the optimal policies is deter-

ministic. In such cases, finding an optimal policy requires searching over the larger set of randomized policies and the corresponding mathematical models may involve additional nonlinearities. Despite these computational challenges, randomized policies are successfully implemented in various applications of constrained MDPs, such as power management (Benini et al., 1999), wireless transmission systems (Djonin and Krishnamurthy, 2007), infrastructure maintenance and rehabilitation (Smilowitz and Madanat, 2000) and queuing systems with reservation (Feinberg and Reiman, 1994). As mentioned before, for unconstrained infinite-horizon discounted MDPs with discrete state space, finite action space and known parameters, an optimal policy (deterministic or randomized) can be efficiently obtained by solving linear programs and it is ensured that there always exists a deterministic stationary optimal policy. However, this is not necessarily true for MDPs under parameter uncertainty or additional constraints. Even for the expected value objective, the infinite-horizon problem may not have any deterministic optimal policy (Buchholz and Scheftelowitsch, 2018), while the existence of a deterministic optimal policy is guaranteed for the finite-horizon problem (Steimle et al., 2018b). The problem QMDP, on the other hand, does not necessarily have a deterministic optimal policy for either infinite or finite-horizon cases. The following example demonstrates this observation for the infinite-horizon case, and it can be easily adjusted for a finite-horizon problem with a single period and zero termination costs.

EXAMPLE 2.1 Consider a single-state infinite-horizon MDP with two actions, a and b , under two scenarios with equal probabilities. Under scenario 1, $\tilde{c}(a) = 0$ and $\tilde{c}(b) = 2$, and under scenario 2, $\tilde{c}(a) = 2$ and $\tilde{c}(b) = 0$. In case a stationary deterministic policy is applied, i.e., either action a or b is chosen, the optimal objective function value of the problem at confidence level $\alpha = 0.9$ is $2/(1 - \gamma)$. However, if a randomized policy of selecting action a or b with equal probabilities is applied, then the optimal objective function value is $1/(1 - \gamma)$, proving that no stationary deterministic policy is optimal.

For classical MDPs with known parameters, the Bellman optimality condition (3) leads to a linear programming formulation, where policy decisions are implied by the constraints that are satisfied as strict equalities. However, the quantile optimization problem QMDP cannot utilize such implicit representation of the policies because Bellman optimality condition does not necessarily hold when the same policy should be imposed across all scenarios due to non-anticipativity. Even though the Bellman optimality condition is no longer valid, value functions in each scenario still need to obey the Bellman equation (2b) to correctly represent the dynamics of the underlying Markov chain for any given policy. Using this property, we propose a mixed-integer nonlinear programming (MINLP) formulation for problem (5) assuming the existence of a discrete representation for parameter uncertainty as in the following statement:

- A1. Joint distribution of $\tilde{\mathbf{c}}$ and $\tilde{\mathbf{P}}$ is represented as a finite set of scenarios $\mathcal{S} = \{1, \dots, n\}$ with corresponding probabilities p^1, \dots, p^n .

Note that our mathematical models and solution algorithms can be easily adjusted for the finite-horizon MDPs with nonstationary/stationary policies by introducing a time dimension on the decisions and/or parameters. Here, we focus on the infinite-horizon case, in which the stationarity assumption is desired for practicality and tractability, and ignore the time indices for brevity.

PROPOSITION 2.1 *Assuming a finite representation of parameter uncertainty as described in A1, problem (5) can be restated as the following deterministic equivalent formulation,*

$$\text{(QMDP-R)} \quad \min y \tag{6a}$$

$$s.t. \quad \sum_{a \in \mathcal{A}} w_{ia} = 1, \quad i \in \mathcal{H}, \tag{6b}$$

$$\sum_{s \in \mathcal{S}} z^s p^s \geq \alpha, \tag{6c}$$

$$\sum_{i \in \mathcal{H}} q_i v_i^s \leq y + (1 - z^s)M, \quad s \in \mathcal{S}, \tag{6d}$$

$$v_i^s \geq \sum_{a \in \mathcal{A}} \tilde{c}_i^s(a) w_{ia} + \gamma \sum_{a \in \mathcal{A}} \sum_{j \in \mathcal{H}} \tilde{P}_{ij}^s(a) v_j^s w_{ia}, \quad i \in \mathcal{H}, s \in \mathcal{S}, \tag{6e}$$

$$z^s \in \{0, 1\}, \quad s \in \mathcal{S}, \tag{6f}$$

$$w_{ia} \in [0, 1], \quad i \in \mathcal{H}, a \in \mathcal{A}, \tag{6g}$$

where M represents a large number such that constraint (6d) for $s \in \mathcal{S}$ is redundant if $z^s = 0$.

PROOF. Constraint (6b) combined with the domain constraint (6g) ensures a probability distribution on the action space for each state, which is effective for all stages. Hence, w_{ia} corresponds to the probability that action a is taken in state i under the optimal policy. Note that the variables \mathbf{w} are required for enforcing a single stationary policy across all scenarios. Constraints (6b) and (6g) also guarantee that constraint (6e) is equivalent to the Bellman equations in (5c) for the policy determined by \mathbf{w} , and variable v_i^s corresponds to the value of $v_i^{(\tilde{c}, \tilde{P})}$ in (5) for parameter realizations in scenario $s \in \mathcal{S}$. Finally, by definition of M , z^s denotes a binary variable equal to 1 if scenario s satisfies $\sum_{i \in \mathcal{H}} q_i v_i^s \leq y$ and consequently, constraints (6c)–(6d) and (6f) are equivalent to the chance constraint (5b). \square

The formulation QMDP-R is nonlinear and nonconvex in general due to constraints (6e), which contain the bilinear terms $v_j^s w_{ia}$ for $s \in \mathcal{S}$, $i \in \mathcal{H}$, $a \in \mathcal{A}$. For such MINLP models, most of the existing solution algorithms only guarantee local optimality of the resulting solutions. To obtain a lower bound on the globally optimal objective function value, we approximate each bilinear term $x_{ija}^s = v_j^s w_{ia}$ for $i, j \in \mathcal{H}$, $a \in \mathcal{A}$, $s \in \mathcal{S}$ in (6e) by the following linear inequalities using McCormick envelopes (McCormick,

1976)

$$\ell_j^s w_{ia} \leq x_{ija}^s \leq u_j^s w_{ia}, \quad (7a)$$

$$v_j^s - (1 - w_{ia})u_j^s \leq x_{ija}^s \leq v_j^s - (1 - w_{ia})\ell_j^s, \quad (7b)$$

$$\ell_j^s \leq v_j^s \leq u_j^s, \quad (7c)$$

where ℓ_j^s and u_j^s are lower and upper bounds on the value of v_j^s , respectively. Then, nonlinear constraint (6e) can be replaced by

$$v_i^s \geq \sum_{a \in \mathcal{A}} \tilde{c}_i^s(a) w_{ia} + \gamma \sum_{a \in \mathcal{A}} \sum_{j \in \mathcal{H}} \tilde{P}_{ij}^s(a) x_{ija}^s, \quad i \in \mathcal{H}, s \in \mathcal{S}, \quad (8)$$

which provides a linear relaxation of QMDP-R in a lifted space.

Despite the possibility that all optimal policies are randomized, implementation of deterministic policies may be preferred over randomized policies in certain application areas. These include the cases in which the decision makers are prone to making errors (Chen and Blankenship, 2002), and the cases that making randomized decisions raises ethical concerns as in the health care systems (Steimle et al., 2018b) and humanitarian relief operations. Considering these cases, we make an additional assumption that

- A2. The policy space is restricted to the set of stationary deterministic policies, denoted as Π_D , i.e., set Π in QMDP is replaced by Π_D ,

and propose an MINLP formulation that aims to minimize the α -quantile value over all stationary deterministic policies in set Π_D .

PROPOSITION 2.2 *Under assumptions A1–A2, problem (5) can be restated as the following deterministic equivalent formulation,*

$$\begin{aligned} \text{(QMDP-D)} \quad & \min y \\ & \text{s.t. (6b) – (6f),} \\ & w_{ia} \in \{0, 1\}, \quad i \in \mathcal{H}, a \in \mathcal{A}, \end{aligned}$$

where M represents a large number such that constraint (6d) for $s \in \mathcal{S}$ is redundant if $z^s = 0$.

The proof easily follows from Proposition 2.1 and the characterization of deterministic policies.

Similar to QMDP-R, Formulation QMDP-D is also nonlinear and nonconvex due to the bilinear terms in constraint (6e), but in this case, binary representation of deterministic policies can be further utilized to

obtain MILP reformulations. When policy vector \mathbf{w} is binary, the McCormick envelopes (7) correspond to an exact linearization of the bilinear terms $x_{ija}^s = v_j^s w_{ia}$ for $i, j \in \mathcal{H}$, $a \in \mathcal{A}$, $s \in \mathcal{S}$. Hence, the MILP obtained by replacing (6e) with (8) and adding constraints (7) is equivalent to QMDP-D. This McCormick reformulation requires the addition of $\mathcal{O}(|\mathcal{S}||\mathcal{A}||\mathcal{H}|^2)$ variables and constraints. An alternative reformulation of QMDP-D as an MILP of smaller size by a factor of $\mathcal{O}(|\mathcal{H}|)$ can be achieved by replacing (6e) in QMDP-D with linear inequalities

$$v_i^s \geq \tilde{c}_i^s(a) + \gamma \sum_{j \in \mathcal{H}} \tilde{P}_{ij}^s(a) v_j^s - (1 - w_{ia}) M_{is}, \quad i \in \mathcal{H}, a \in \mathcal{A}, s \in \mathcal{S}. \quad (10)$$

Constraint (10) assures that the multi-scenario Bellman equations (5c) in the CCP model (5) are satisfied for the selected scenarios, i.e., $z^s = 1$, and the policy represented by the state-action pairs (i, a) such that $w_{ia} = 1$. The big- M term M_{is} for any $i \in \mathcal{H}$, $s \in \mathcal{S}$ denotes a large number sufficient to make constraint (10) redundant when $w_{ia} = 0$ for $a \in \mathcal{A}$. Note that replacing (6e) with (10) does not provide an exact reformulation of QMDP-R.

3. Implementation Details In this section, we propose preprocessing methods to initialize problem parameters, which may also be used for reducing the problem size. We also provide a heuristic solution algorithm. Note that the proposed methods are applicable for both QMDP-R and QMDP-D, but here we consider on QMDP-R due to its generality.

3.1 Preprocessing The formulations for problems QMDP-R and QMDP-D presented in the previous section require determination of the auxiliary terms M , M_{is} , ℓ_i^s and u_i^s for $i \in \mathcal{H}$, $s \in \mathcal{S}$. In what follows, we provide preprocessing procedures to prespecify these terms and to narrow down the solution space before solving the original problem. Existing solution algorithms for MDPs usually suffer from the size of state and action spaces, often referred to as the curse of dimensionality. In our case, the computational complexity is additionally amplified with the number of scenarios and the combinatorial nature of the quantile calculation. Hence it is worthwhile to search for methods that reduce the size of the problem.

Let y^* be the optimal objective function value of problem QMDP-R. First, we relax the requirement that the same policy should be selected over all scenarios, and use monotonicity property of the VaR function to find bounds on the value of y^* . We denote \bar{b} as a random variable representing the maximum expected total discounted cost of the relaxed MDP model. The realization of \bar{b} under scenario $s \in \mathcal{S}$ can be computed by solving the following linear programming problem

$$\bar{b}^s = \min \sum_{i \in \mathcal{H}} q_i v_i \quad (11a)$$

$$s.t. \quad v_i \geq \tilde{c}_i^s(a) + \gamma \sum_{j \in \mathcal{H}} \tilde{P}_{ij}^s(a) v_j, \quad i \in \mathcal{H}, a \in \mathcal{A}. \quad (11b)$$

Then we order the realizations of \bar{b} under each scenario as $\bar{b}^{s_1} \leq \bar{b}^{s_2} \leq \dots \leq \bar{b}^{s_n}$, and let $k_1 \in \{1, \dots, n\}$ be the order of the scenario such that $\sum_{t=1}^{k_1} p^{s_t} \geq \alpha$ and $\sum_{t=1}^{k_1-1} p^{s_t} < \alpha$, and $b_u = \bar{b}^{s_{k_1}}$. Note that b_u corresponds to the VaR of random variable \bar{b} at confidence level α . Hence b_u provides an upper bound on y^* , i.e., $b_u \geq y^*$, since $\bar{b}^s \geq C(\boldsymbol{\pi}, \tilde{\mathbf{c}}^s, \tilde{\mathbf{P}}^s)$ for all $s \in \mathcal{S}$, $\boldsymbol{\pi} \in \Pi$.

Similarly, let \underline{b} be a random variable representing the minimum expected total discounted cost of the MDP with relaxed policy selection requirements, where realizations of \underline{b} under each scenario $s \in \mathcal{S}$ can be computed by solving the linear programs

$$\underline{b}^s = \max \sum_{i \in \mathcal{H}} q_i v_i \quad (12a)$$

$$s.t. \quad v_i \leq \tilde{c}_i^s(a) + \gamma \sum_{j \in \mathcal{H}} \tilde{P}_{ij}^s(a) v_j, \quad i \in \mathcal{H}, a \in \mathcal{A}. \quad (12b)$$

Let b_l correspond to the VaR of random variable \underline{b} at confidence level α . The value of b_l provides a lower bound on y^* , i.e., $b_l \leq y^*$ as $\underline{b}^s \leq C(\boldsymbol{\pi}, \tilde{\mathbf{c}}^s, \tilde{\mathbf{P}}^s)$ for all $s \in \mathcal{S}$, $\boldsymbol{\pi} \in \Pi$ by definition.

Problems (11) and (12) can be efficiently solved using a policy iteration or value iteration algorithm in polynomial time. Using these bounds, we can conclude that any scenario s , whose lower bound is greater than the upper bound on the quantile value, i.e., $\underline{b}^s > b_u$, cannot satisfy the term inside chance constraint (5b) in the optimal policy, hence we can set $z^s = 0$. Similarly, any scenario $s \in \mathcal{S}$ with an upper bound value smaller than the lower bound on the quantile function value, i.e., $\bar{b}^s < b_l$ must satisfy the term inside chance constraint (5b) in the optimal solution. Therefore, the corresponding scenario variable can be set as $z^s = 1$. This result can be used for reducing the number of \mathbf{z} variables. Furthermore, we can add the constraints $b_l \leq y \leq b_u$ into our mathematical model. As previously mentioned, validity of these bounds follows from monotonicity property of the VaR function. Additionally, the inequalities

$$y \geq \underline{b}^s z^s, \quad s \in \mathcal{S} \quad (13)$$

are valid due to constraints (6d) and the fact that $\underline{b}^s \leq C(\boldsymbol{\pi}, \tilde{\mathbf{c}}^s, \tilde{\mathbf{P}}^s)$ for all $s \in \mathcal{S}$, $\boldsymbol{\pi} \in \Pi$. Note that Küçükyavuz and Noyan (2016) propose similar bounding and scenario elimination ideas and demonstrate their effectiveness in the context of multivariate CVaR-constrained optimization (see also, Liu et al., 2017b; Noyan et al., 2019).

Let \bar{v}_i^s and \underline{v}_i^s be the optimal values of variable v_i in problem (11) and (12), respectively, for state $i \in \mathcal{H}$ and scenario $s \in \mathcal{S}$. Then, we can set the lower and upper bounds on variable v_i^s in (7) as $\ell_i^s = \underline{v}_i^s$ and $u_i^s = \bar{v}_i^s$, respectively. This result follows from the fixed-point and contraction properties of the

value functions for classical MDPs. The information obtained by solving problems (11) and (12) can also be used to obtain tighter values of the big- M terms in constraints (6d) and (10). Clearly, we can set $M = b_u - b_l$ in (6d) because its value is bounded by $M \geq \max_{\mathbf{v}: (6b)-(6g)} \sum_{i \in \mathcal{H}} q_i v_i^s - \min_{\mathbf{y}: (6b)-(6g)} y$ for $s \in \mathcal{S}$ such that $z^s = 1$. Moreover, Steimle et al. (2018b) show that the big- M term in constraint (10) can be set as $M_{is} = \bar{v}_i^s - \underline{v}_i^s$ for all $i \in \mathcal{H}$, $s \in \mathcal{S}$.

3.2 Obtaining a Feasible Solution Our preliminary computational experiments suggest that the upper bound b_u described in the previous section can be further improved by finding a feasible solution to problem QMDP-R. Here, we propose a polynomial time heuristic algorithm (Algorithm 1), which benefits from the connection between a substructure of our problem with the robust MDPs to attain a feasible policy effectively.

Algorithm 1: *initialSolution()*

- 1 Given distinct cost and transition matrices $\{\tilde{\mathbf{c}}, \tilde{\mathbf{P}}\}_{s \in \mathcal{S}}$ with corresponding probabilities $\{p\}_{s \in \mathcal{S}}$, set $\hat{\mathbf{z}} \leftarrow \mathbf{0}$;
 - 2 **for** each scenario $s \in \mathcal{S}$ **do**
 - 3 \lfloor Solve problem (12) to obtain its optimal objective function value \underline{b}^s ;
 - 4 Compute $\text{VaR}_\alpha(\underline{b})$ and find a maximal subset of scenarios $\bar{\mathcal{S}} \subseteq \mathcal{S}$ such that $\underline{b}^s \leq \text{VaR}_\alpha(\underline{b})$ for each scenario $s \in \bar{\mathcal{S}}$ and $\sum_{s \in \bar{\mathcal{S}} \setminus \{s'\}} p^s < \alpha$ for all $s' \in \bar{\mathcal{S}}$;
 - 5 **for** each scenario $s \in \bar{\mathcal{S}}$ **do**
 - 6 \lfloor Set $\hat{z}^s \leftarrow 1$;
 - 7 Return *robustPolicySelection*($\hat{\mathbf{z}}$);
-

Algorithm 1 follows two sequential phases: scenario selection and policy selection. In the scenario selection phase, we decide on which scenarios will be enforced to satisfy the chance constraint, i.e., $\hat{z}^s = 1$. As in the previous subsection, we relax the requirement that the selected policy should be the same over all scenarios, and solve problems (12) independently to obtain the optimal objective function value \underline{b}^s corresponding to each scenario $s \in \mathcal{S}$. Then, we set $\hat{z}^s = 1$ for each scenario s in a maximal subset $\bar{\mathcal{S}} \subseteq \mathcal{S}$ such that $\underline{b}^s \leq \text{VaR}_\alpha(\underline{b})$ for each scenario $s \in \bar{\mathcal{S}}$ and $\sum_{s \in \bar{\mathcal{S}} \setminus \{s'\}} p^s < \alpha$ for all $s' \in \bar{\mathcal{S}}$.

In the policy selection phase provided in Algorithm 2, we use the scenarios selected in the first phase to obtain a feasible policy and the corresponding quantile value. In other words, we let $\mathcal{S}(\mathbf{z}) = \bar{\mathcal{S}}$ in Line 1 of Algorithm 2 and solve a relaxation of the robust MDP problem

$$\text{rMDP}(\mathbf{z}): \min_{\boldsymbol{\pi} \in \Pi} \max_{s \in \mathcal{S}(\mathbf{z})} C(\boldsymbol{\pi}, \tilde{\mathbf{c}}^s, \tilde{\mathbf{P}}^s),$$

where $\mathcal{S}(\mathbf{z}) := \{s \in \mathcal{S} \mid z^s = 1\}$, to find a feasible policy for QMDP-R. For any scenario vector \mathbf{z} , $\text{rMDP}(\mathbf{z})$ can be seen as an adversarial game, where decision maker selects a stationary policy at the very beginning, and the system evolves based on the worst possible scenario for the selected policy afterwards.

An important observation is that for any vector \mathbf{z} satisfying constraints (6c) and (6f), the optimal policy obtained by solving $\text{rMDP}(\mathbf{z})$ and its optimal objective function value correspond to a feasible solution of QMDP-R for policy variables $\boldsymbol{\pi}$ and quantile variable y , respectively. To see this, we reformulate QMDP-R using the definition of $C(\boldsymbol{\pi}, \tilde{\mathbf{c}}^s, \tilde{\mathbf{P}}^s)$ as

$$\begin{aligned} \min_{y \in \mathbb{R}, \boldsymbol{\pi} \in \Pi} \quad & y \\ \text{s.t.} \quad & y \geq C(\boldsymbol{\pi}, \tilde{\mathbf{c}}^s, \tilde{\mathbf{P}}^s) - (1 - z^s)M, \quad s \in \mathcal{S}, \\ & \text{(6c), (6f),} \end{aligned}$$

or equivalently $\min_{\mathbf{z}: \text{(6c), (6f)}, \boldsymbol{\pi} \in \Pi} \max_{s \in \mathcal{S}(\mathbf{z})} C(\boldsymbol{\pi}, \tilde{\mathbf{c}}^s, \tilde{\mathbf{P}}^s)$.

Note that the substructure $\text{rMDP}(\mathbf{z})$ is a robust MDP, where the uncertainty in the transition matrix is coupled across the time horizon and the state space, i.e., a single realization of the cost and transition matrices is selected randomly at the beginning and it holds for all decision epochs and states of the system. Problem $\text{rMDP}(\mathbf{z})$ is known to be NP-hard even when the parameters are allowed to follow a different scenario for each state and only the scenario considered at each time a state is encountered should be consistent over time (Iyengar, 2005). Therefore, in Algorithm 2, using the scenario vector $\hat{\mathbf{z}}$ obtained in the scenario selection phase, we find a policy by solving a relaxed version of $\text{rMDP}(\hat{\mathbf{z}})$, which we describe next.

Algorithm 2: *robustPolicySelection*(\mathbf{z})

- 1 Given a small tolerance parameter $\epsilon > 0$, set some positive $\hat{\mathbf{v}}_1 \in \mathbb{R}^{|\mathcal{H}|}$, $\mathcal{S}(\mathbf{z}) \leftarrow \{s \in \mathcal{S} \mid z^s = 1\}$ and $t \leftarrow 1$;
 - 2 **for** each state $i \in \mathcal{H}$ **do**
 - 3 **for** each action $a \in \mathcal{A}$ **do**
 - 4 Set $\sigma_{ia} \leftarrow \max_{s \in \mathcal{S}(\mathbf{z})} \left\{ c_i^s(a) + \gamma \sum_{j \in \mathcal{H}} \tilde{P}_{ij}^s(a) \hat{v}_t(j) \right\}$;
 - 5 Compute $\hat{\mathbf{v}}_{t+1}(i) \leftarrow \min_{a \in \mathcal{A}} \sigma_{ia}$;
 - 6 **if** $\|\hat{\mathbf{v}}_{t+1} - \hat{\mathbf{v}}_t\| < \frac{(1-\gamma)\epsilon}{\gamma}$ **then**
 - 7 Go to line 10;
 - 8 **else**
 - 9 Set $t \leftarrow t + 1$ and go to line 2;
 - 10 Return $\boldsymbol{\pi}$ such that $\pi_i \in \arg \max_{a \in \mathcal{A}} \sigma_{ia}$;
-

Similar to Nilim and El Ghaoui (2005), we consider a relaxed variant of $\text{rMDP}(\hat{\mathbf{z}})$ by assuming that the costs and transition probabilities are independent over different states and actions, and these parameters are allowed to be time-varying. This setting can be seen as a sequential game, where at each time step, an action is taken by the decision maker and accordingly the system generates a cost and makes a transition based on the worst possible parameter realization in scenario set $\mathcal{S}(\hat{\mathbf{z}})$ for the current state and action,

in an iterative fashion. Note that Nilim and El Ghaoui (2005) consider the case of transition matrix uncertainty only and prove that the optimal policy in this setting obeys a set of optimality conditions, which can be solved using a robust value iteration algorithm. Here, we further generalize their algorithm for the case of uncertainty in both cost and transition matrices to find a policy that performs well in terms of the VaR objective, but is not necessarily optimal. For the vector $\hat{\mathbf{z}}$ computed in the first phase, we construct the set $\mathcal{S}(\hat{\mathbf{z}}) := \{s \in \mathcal{S} \mid \hat{z}^s = 1\}$ and solve the following equations

$$v_i = \min_{a \in \mathcal{A}} \max_{s \in \mathcal{S}(\hat{\mathbf{z}})} \left\{ c_i^s(a) + \gamma \sum_{j \in \mathcal{H}} \tilde{P}_{ij}^s(a) v_j \right\}, \quad i \in \mathcal{H}$$

through a variant of the value iteration algorithm as presented in Algorithm 2. Note that the policy obtained by Algorithm 2 and the corresponding value functions provide upper bounds on the results of the robust problem $\text{rMDP}(\hat{\mathbf{z}})$. Moreover, due to line 10, Algorithm 2 is guaranteed to terminate with a deterministic policy, hence the obtained solution is also feasible for problem QMDP-D.

4. An Application to Inventory Management in Long-Term Humanitarian Relief Operations Long-term humanitarian relief operations, alternatively referred to as continuous aid operations, are vital to sustain daily basic needs of a population affected by a slow-onset disaster including war (Syria, Afghanistan, Iraq), political insurrection (Syria), famine (Yemen, South Sudan, Somalia), drought (Ethiopia) and extreme poverty (Niger, Liberia). Unlike the sudden-onset disasters (e.g., earthquakes, hurricanes, terrorist attacks), they require delivery of materials such as food, water and medical supplies to satisfy a chronic need over a long period of time. Since the progress of a slow-onset disaster usually presents irregularities in terms of its scale and location, the demand is highly uncertain. In addition, the supply levels are also exposed to uncertainty as they mainly depend on donations from multiple resources. More than 90% of the people affected by slow-onset disasters lives in developing countries, hence the required relief items are usually outsourced from resources in various locations around the world (Rottkemper et al., 2012). Another important consideration in long-term humanitarian operations is the perishability of the relief items. Many items needed during and after a disaster, e.g., food and medication, have a limited shelf life. In addition to the possible uncertainty in the initial shelf life, the remaining shelf life upon arrival is also affected by the uncertainty in the lead times due to unknown location of the donations. Hence, high level of uncertainty inherent in the supply chain makes the system prone to unwanted shortages and disposals. To prevent interruptions, the policy makers may interfere through different actions such as campaigns and advertisements that provide additional relief items.

In this section, we consider an inventory management model for a single perishable item in long-term humanitarian relief operations proposed by Ferreira et al. (2018). The authors formulate the problem as an infinite-horizon MDP with finite state and action spaces, where the states represent possible levels

of inventory. The aim is to minimize the long-term average expected cost. Their model assumes that both demand and supply is uncertain following Poisson distributions with known demand and supply rates, respectively, and the expiration time of the supply items is deterministic. These parameters are generally obtained using various forecasting methods. However, due to multiple sources of uncertainty in the context of long-term humanitarian operations, the forecasted values may be erroneous, which may consequently affect the performance of the resulting policy. To handle the parameter uncertainty, different from Ferreira et al. (2018), we assume that demand and supply rates and the expiration time probabilistically take value from a finite set of scenarios \mathcal{S} and the objective in our model is to minimize the α -quantile of the total discounted expected total cost with respect to the uncertainty in these parameters.

In what follows, we elaborate on the components and assumptions of the MDP model based on Ferreira et al. (2018).

States: The state of the system is represented by the number of available items in the inventory at the beginning of each decision epoch that will not expire before delivery. It is assumed that the inventory has a capacity K , e.g., the state space $\mathcal{H} = \{0, 1, \dots, K\}$.

Actions: At the beginning of each decision epoch, the decision maker takes an action a from a finite set \mathcal{A} that provides an additional n_a number of relief items. The set of actions taken for each possible state of the system that minimizes the objective function is referred to as the optimal policy.

Transition Probabilities: Under each scenario $s \in \mathcal{S}$, the demand (D) and supply (U) for the item follow Poisson distribution with probability mass functions f_d^s and f_u^s of rates μ_d^s and μ_u^s , respectively, and the expiration time takes value t_e^s . Let Δ_{\min} and Δ_{\max} be the minimum and maximum possible difference between the supply and demand at any decision epoch considering a $100(1 - \epsilon)\%$ confidence level for a small $\epsilon > 0$. Note that the minimum possible demand and supply amounts are assumed as zero so that Δ_{\min} and Δ_{\max} represent the negative of maximum possible demand and the maximum possible supply, respectively. Then, ignoring the perishability of the item and the actions taken by the decision maker, the probability of transitioning from inventory level $i \in \mathcal{H}$ to $j \in \mathcal{H}$ under scenario $s \in \mathcal{S}$ is

$$\hat{P}_{ij}^s = \begin{cases} \sum_{\Delta=\Delta_{\min}}^{j-i} \sum_{D=0}^{-\Delta_{\min}} f_d^s(D) f_u^s(D + \Delta), & \text{if } j = 0, \\ \sum_{\Delta=j-i}^{\Delta_{\max}} \sum_{D=0}^{-\Delta_{\min}} f_d^s(D) f_u^s(D + \Delta), & \text{if } j = K, \\ \sum_{D=0}^{-\Delta_{\min}} f_d^s(D) f_u^s(D + j - i), & \text{otherwise.} \end{cases}$$

Instead of keeping track of the number of items that expire in each period, the model is simplified by assuming that the items procured in each decision epoch have the same expiration time. As a result, at the beginning of each decision epoch, the expiration probability for the whole batch of newly arriving

items is computed as the probability of not being able to consume all available items before the expiration time. Note that since the demand follows Poisson distribution with rate μ_d^s , the probability of consuming k items before time t_e^s follows Erlang distribution with parameters μ_d^s and k , which can be stated as $(1 - f_e^s(t_e^s, k))$. Incorporating perishability of the items based on this simplification, the probability of a transition from inventory level $i \in \mathcal{H}$ to inventory level $j \in \mathcal{H}$ under scenario $s \in \mathcal{S}$ when action $a \in \mathcal{A}$ is taken can be computed as

$$\tilde{P}_{ij}^s(a) = \begin{cases} (1 - f_e^s(t_e^s, j)) \hat{P}_{i+n_a, j}^s, & \text{if } j > i + n_a, \\ \sum_{j'=i+n_a+1}^{\min(K, i+n_a+\Delta_{\max})} f_e^s(t_e^s, j') \hat{P}_{i+n_a, j'}^s + \hat{P}_{i+n_a, j}^s, & \text{if } j = i + n_a, \\ \hat{P}_{i+n_a, j}^s, & \text{otherwise.} \end{cases} \quad (15)$$

Equation (15) follows from the fact that if an arriving batch of items is known to expire, the new arrivals are immediately used to fulfill the demand with priority over the existing inventory (the existing inventory can be used later since it is guaranteed not to expire by the previous assumption), and the remaining items in the batch are disposed. Hence, the case $j > i + n_a$ can occur only if the new items do not expire. Similarly, $j = i + n_a$ implies either that the incoming supply is at least as much as the demand in the current period and the supply surplus is disposed, or demand is equal to the sum of supply and additional items acquired by the action taken.

Cost function: The expected cost of taking action $a \in \mathcal{A}$ at inventory level $i \in \mathcal{H}$ under scenario $s \in \mathcal{S}$, stated as $\tilde{c}_i^s(a)$, consists of three main components: the inventory holding cost, the expected shortage cost and the expected disposal cost. Assuming a unit disposal cost of \tilde{d} , the total expected disposal cost when action $a \in \mathcal{A}$ is taken at inventory level $i \in \mathcal{H}$ under scenario $s \in \mathcal{S}$ is

$$DC(i, a, s) = \sum_{\Delta=1}^{\min(K-i-n_a, \Delta_{\max})} f_e^s(t_e^s, i+n_a+\Delta) \hat{P}_{i+n_a, i+n_a+\Delta}^s \Delta \tilde{d} \\ + \sum_{j=K}^{K+n_a+\Delta_{\max}} \sum_{D=0}^{-\Delta_{\min}} f_d^s(D) f_u^s(D+j-i-n_a) (j-K) \tilde{d}.$$

The first summation term is due to the expired items, while the second term is because the arrivals that cannot be stored in inventory due to capacity limit K are disposed. Similarly, the shortage costs are stated as

$$SC(i, a, s) = \sum_{j=i+n_a+\Delta_{\min}}^{-1} \sum_{D=0}^{-\Delta_{\min}} f_d^s(D) f_u^s(D+j-i-n_a) (-j) \tilde{u}, \quad i \in \mathcal{H}, a \in \mathcal{A}, s \in \mathcal{S},$$

where \tilde{u} is the unit shortage cost. Assuming that the unit inventory holding cost is h , the total expected cost of taking action $a \in \mathcal{A}$ at inventory level $i \in \mathcal{H}$ under scenario $s \in \mathcal{S}$ can be computed using

$\tilde{c}_i^s(a) = hi + DC(i, a, s) + SC(i, a, s)$, $i \in \mathcal{H}$, $a \in \mathcal{A}$, $s \in \mathcal{S}$. Note that the cost structure may be different depending on the characteristics of the environment in consideration. The MDP formulation provides decision makers the flexibility to use even nonconvex cost functions. Our methodology only requires the cost terms to be bounded.

5. Computational Study In this section, we conduct computational experiments on the long-term humanitarian relief operations inventory management problem described in Section 4 to examine the effects of incorporating risk aversion towards parameter uncertainty into MDP models and to compare the efficiency of different solution approaches. The problem instances are generated based on the experiments provided in Ferreira et al. (2018) considering the inventory of a blood center, which collects and distributes blood packs to support humanitarian relief operations. We suppose that the inventory replenishment decisions are made on a weekly basis, and the weekly demand and supply rates for the blood packs take values uniformly in the intervals $[30, 130]$ and $[20, 80]$, respectively. After donation, each blood pack has a shelf life of six weeks, however unknown lead times may affect the remaining shelf life at the time of arrival to the blood center. Hence, the shelf life is assumed to be uniform on the interval $[1, 6]$. In case of need, the center may procure additional supply of blood packs by sending up to $V \in \{2, 3, 4\}$ blood collection vehicles to distant areas so that $\mathcal{A} = \{0, 1, \dots, V\}$. Each vehicle collects additional 20 blood packs at the expense of incurring a certain cost. We use the cost parameters for additional procurement, inventory holding costs, disposal costs and shortage costs given in Ferreira et al. (2018). Note that state action costs are scaled by 1000 in our computations. We additionally assume that the blood center has a capacity of $K \in \{50, 100, 150\}$ units and the blood packs are in batches of 10 units so that $\mathcal{H} = \{0, 1, \dots, \lfloor \frac{K}{10} \rfloor\}$. Different than Ferreira et al. (2018), we generate random instances with equiprobable scenarios $\mathcal{S} = \{1, \dots, n\}$ for $n \in \{50, 100, 250, 500, 1000\}$, where the parameters of shelf life and demand and supply rates for each scenario randomly take value on their respective intervals stated above. The distribution on the initial state is assumed to be uniform as well.

All experiments are performed using single thread of a Linux server with Intel Haswell E5-2680 processor at 2.5 GHz and 128 GB of RAM using Python 3.6 combined with Gurobi Optimizer 7.5.1 for MILP formulations, and Pyomo 5.5 (Hart et al., 2012) and Bonmin 1.8 (Bonmin) for MINLP formulations. The time limit for each instance on Gurobi is set to 3600 seconds and we use the default settings for the MIP gap and feasibility tolerances. The results are obtained for the instances with $\alpha \in \{0.90, 0.95, 1\}$ and $\gamma = 0.99$. Unless otherwise stated, each reported value corresponds to the average of five replications.

5.1 The impact of deterministic policies We first investigate how the quantile value and the computational performance of solution methods are affected by the choice of narrowing down the policy space to deterministic policies by comparing the solutions obtained by solving QMDP-R, which includes randomized policies, with the solutions of QMDP-D considering only deterministic policies. Note that

implementation of a randomized policy in the context of humanitarian inventory management may require ordering a random quantity of critical humanitarian supplies (e.g., blood packs) even if the current inventory level is extremely low. Deterministic policies, on the other hand, lead to consistent order quantities at each inventory level and therefore are more desirable in such humanitarian settings due to their interpretability—in the sense that they are easier to explain and justify. Hence, our goal in this section is mainly to determine this cost of interpretability and also to investigate computational aspects of obtaining optimal randomized policies for broader contexts, where implementation of randomized policies does not raise ethical/practical issues, such as power management, wireless transmission systems, infrastructure maintenance and rehabilitation and queuing systems with reservation as previously mentioned.

Because finding a globally optimal solution for the nonconvex formulation QMDP-R is computationally challenging, we use the open-source MINLP solver Bonmin, which provides local optima to nonconvex problems, in combination with Python-based, open-source optimization modeling language Pyomo. The solver is warm-started with the policy obtained by solving Algorithm 1. Since this approach only guarantees local optimality, we additionally obtain lower bounds by solving the McCormick relaxation of QMDP-R as described in Section 2. Our preliminary experiments demonstrate that the McCormick relaxation provides up to 1.26% tighter lower bound values than the lower bound computed by considering each scenario separately. Hence, we opt for the McCormick relaxation for the analysis in this section, where our main concern is to have as tight a lower bound as possible. On the other hand, we use the lower bounds obtained by considering each scenario separately in the preprocessing phase, because this is much more efficient than the McCormick relaxation of QMDP-R, which is a mixed-integer program (due to z^s variables) requiring significant solution times when the sample size is large (see Table 1). The McCormick relaxation of QMDP-R and deterministic policy problem QMDP-D are solved under the best settings as detailed in Section 5.4.

Table 1 compares the results at confidence level $\alpha = 0.95$ for deterministic policies (QMDP-D) reported under column “Deterministic” with the results of the case with randomized policies: the locally optimal solution obtained by solving QMDP-R using the MINLP solver Bonmin under column “Randomized – NL” and its lower bound provided by the McCormick relaxation under the column “Lower bound – MC”. Each value under column “Time (s)” refers to the solution time of the corresponding problem. It can be seen that Gurobi terminates within the time limit of one hour for the instances of QMDP-D and the McCormick relaxation of QMDP-R, whereas Bonmin takes longer than the time limit to terminate for some QMDP-R instances due to its non-preemptive internal processes. The blue (resp., red) superscripts correspond to the number of instances that the solver exceeded the time limit with (resp., without) a feasible solution. We additionally report the optimality gap at the time of termination for linear formulations QMDP-D and the McCormick relaxation under the columns “Gap (%)”. Note that the nonlinear solver does not provide an optimality gap for QMDP-R because it can only prove

Table 1: Comparison of deterministic and randomized policies

Instance			Deterministic		Randomized – NL			Lower bound – MC		
$ \mathcal{S} $	$ \mathcal{H} $	$ \mathcal{A} $	Time (s)	Gap (%)	Time (s)	Diff. (%)	Time (s)	Gap (%)	Diff. (%)	
50	6	3	0.53	0.00	9.42	0.11	0.17	0.00	1.43	
50	6	4	1.54	0.00	10.31	0.46	0.21	0.00	6.74	
50	6	5	6.96	0.00	75.52	0.40	0.26	0.00	14.80	
50	11	3	5.01	0.00	31.38	0.00	0.45	0.00	0.41	
50	11	4	48.97	0.00	52.73	0.10	0.64	0.00	1.90	
50	11	5	1371.54	0.00	432.63	0.21	2.31	0.00	6.85	
50	16	3	48.41	0.00	146.35	0.03	0.84	0.00	0.30	
50	16	4	773.00 ¹	0.26	307.02	0.05	1.22	0.00	0.77	
50	16	5	2883.24 ⁴	2.72	2502.51 ¹	-0.47	6.32	0.00	3.33	
100	6	3	1.11	0.00	18.29	0.13	0.91	0.00	0.69	
100	6	4	6.76	0.00	83.42	0.30	4.86	0.00	8.11	
100	6	5	109.59	0.00	130.38	0.65	2.02	0.00	19.91	
100	11	3	1.75	0.00	71.51	0.02	5.61	0.00	0.12	
100	11	4	331.41	0.00	246.10	0.18	22.41	0.00	2.57	
100	11	5	3483.96 ⁴	7.61	4940.98	0.07	8.34	0.00	10.60	
100	16	3	4.07	0.00	296.91	0.02	35.50	0.00	0.07	
100	16	4	1762.18 ²	1.04	1789.07	0.07	20.23	0.00	1.70	
100	16	5	3600.01 ⁵	5.07	2359.51	0.05	23.11	0.00	5.53	
250	6	3	3.40	0.00	134.41	0.16	3.70	0.00	1.38	
250	6	4	26.24	0.00	3636.85	0.10	36.90	0.00	6.89	
250	6	5	317.29	0.00	5177.90 ¹	-0.73	51.30	0.00	13.01	
250	11	3	13.97	0.00	588.75	0.02	182.71	0.00	0.26	
250	11	4	1558.46 ¹	0.49	10329.68 ²	-0.03	74.69	0.00	2.56	
250	11	5	3600.03 ⁵	5.19	3713.75 ⁴	-1.11	126.33	0.00	6.26	
250	16	3	42.39	0.01	1700.00 ¹	-0.02	5.03	0.00	0.15	
250	16	4	3035.12 ⁴	1.42	3913.63 ⁵	-0.87	2532.96 ²	1.64	0.70	
250	16	5	3600.01 ⁵	5.34	3887.01 ⁴	-1.56	3587.00 ⁴	14.49	1.37	
500	6	3	10.04	0.00	884.64	0.09	1.21	0.00	1.04	
500	6	4	228.39	0.00	2123.58 ²	-0.28	200.24	0.00	7.41	
500	6	5	3137.55 ²	2.42	3767.50 ⁵	-1.18	116.72	0.00	16.40	
500	11	3	30.83	0.00	1305.78	0.02	4.42	0.00	0.24	
500	11	4	3077.46 ³	0.74	3692.29 ⁵	-0.15	2150.46 ¹	0.33	2.41	
500	11	5	3600.02 ⁵	7.48	4166.01 ⁵	-1.12	3600.08 ⁵	36.17	-0.57	
500	16	3	264.96	0.00	2872.86 ²	-0.03	3600.19 ⁵	0.27	-0.07	
500	16	4	3600.02 ⁵	1.36	6368.03 ⁵	-0.86	3600.13 ⁵	37.67	-0.97	
500	16	5	3600.04 ⁵	6.72	5627.73 ⁵	-1.26	3600.17 ⁵	81.17	-1.61	
1000	6	3	74.47	0.00	1864.43 ¹	-0.07	1496.29	0.00	1.56	
1000	6	4	1930.16 ¹	0.50	3691.58 ⁵	-0.45	2163.89 ²	5.35	7.10	
1000	6	5	3600.07 ⁵	8.44	3682.36 ⁵	-2.22	3600.05 ⁵	30.16	8.18	
1000	11	3	210.63	0.00	2960.04 ²	0.02	3600.23 ⁵	14.97	-0.20	
1000	11	4	3600.03 ⁵	3.07	3807.47 ⁵	0.04	3600.12 ⁵	81.79	-1.18	
1000	11	5	3600.02 ⁵	6.95	4296.76 ⁵	-0.30	3600.19 ⁵	100.00	-2.71	
1000	16	3	2003.15 ²	0.24	17223.97 ⁵	-0.04	3600.24 ⁵	28.95	-0.09	
1000	16	4	3600.09 ⁵	3.58	9645.38 ⁵	-0.09	3600.35 ⁵	85.56	-0.25	
1000	16	5	3600.11 ⁵	6.30	11432.46 ⁵	-0.06	3600.37 ⁵	100.00	0.42	
Average			1459.67 ^{1.76}	1.71	3022.20 ^{1.87,0.02}	-0.21	1166.03 ^{1.42}	13.75	3.46	
Maximum			3600.11 ⁵	8.44	17223.97 ^{5,1}	0.65	3600.37 ⁵	100.00	19.91	

local optimality. The percentage difference between the optimal objective value considering only the deterministic policies, denoted by “ o_d ”, and the objective value of each solution approach allowing for randomized policies, denoted by “ o_r ”, is reported as $\text{Diff. } (\%) = 100 \times \frac{o_d - o_r}{o_d}$ under the column of the corresponding solution method.

The results in Table 1 show that solving the nonlinear formulation QMDP-R provides solutions which perform at least as well as the optimal deterministic policies for all instances that terminate with a feasible solution within one hour. The gain achieved by permitting randomized policies is at most 0.65%. However, considering the time limit, the best deterministic policy outperforms the best randomized policy by 0.21% on average. Additionally, the results for the McCormick relaxation, reported under column “Lower bound – MC”, indicate that, for our problem instances, the maximum possible gain in the optimal quantile value obtained by allowing for randomized policies is bounded by 3.46% on average and 19.91% at maximum. We emphasize that, unlike “Randomized – NL” that produces feasible solutions to QMDP-R, the solutions of “Lower bound – MC” provide only theoretical upper bounds on the actual improvement that can be gained by allowing randomized policies. The results of “Lower bound – MC” may not be possible to achieve in practice because the reformulation using McCormick envelopes corresponds to a relaxation of the original problem QMDP-R. The lower bounds provided by “Lower bound – MC” also indicate that for 149 of 225 problem instances, the reduction in risk (VaR_α) due to policy randomization is guaranteed not to be more than 3% with respect to the objective value of the optimal deterministic policy. Moreover, a deterministic policy of good quality can usually be obtained in reasonable solution times (1459.67 seconds on average), comparable to the solution times of the QMDP-R (3022.20 seconds on average) and its McCormick relaxation (1166.03 seconds on average). Furthermore, “Lower bound – MC” fails to provide valid lower bounds for the problem for large instances that reach the time limit as evidenced by the negative entries under the Diff column.

Motivated by these results and given our humanitarian problem context, in the following sections, we focus on the problem QMDP-D, which only considers deterministic policies.

5.2 The impact of parameter uncertainty Next we examine the effect of incorporating parameter uncertainty into the MDP model in terms of the value gained by using the available stochastic information and the potential loss due to not having perfect information on the true realization of random parameters.

In Table 2, we compare the optimal objective function value of the quantile optimization problem QMDP-D, denoted as OPT, with two benchmark cases for $\alpha \in \{0.90, 0.95, 1\}$. Each blue superscript represents the number of replications that terminate with a feasible solution due to time limit for the corresponding α -quantile minimization problem. The first benchmark assumes that the decision maker waits until observing the actual parameter realizations and makes a decision for each scenario indepen-

dently. This approach does not provide a feasible solution to the original problem because it may produce distinct policies for each scenario. Clearly the quantile value in this case corresponds to a lower bound on the OPT since it can be stated as $LB = \text{VaR}_\alpha(\underline{b})$, where the realization under scenario $s \in \mathcal{S}$ is $\underline{b}^s = \min_{\pi \in \Pi_D} C(\pi, \tilde{c}^s, \tilde{P}^s)$. Using this value, we compute the value of perfect information on MDP parameters as $VPI = \text{OPT} - LB$ and its percentage value as $\%VPI = 100 \times \frac{\text{OPT} - LB}{\text{OPT}}$. Column $\%VPI$ presented in Table 2 indicates that the losses in the quantile function value due to not knowing the true parameter realizations are 0.58%, 4.77% and 8.95% on average, and at most 4.19%, 20.03% and 25.16% for $\alpha = 1, 0.95, 0.90$, respectively, for our problem instances. Therefore, it is worthwhile to use additional information on the uncertain parameters whenever possible. Incorporating additional information is especially valuable for decision makers whose main concern is to ensure a robust system performance against parameter uncertainty at lower confidence levels.

The second benchmark considers the quantile function value corresponding to a policy obtained by solving a single MDP with expected parameter values, referred as the mean value problem in stochastic programming. This provides a heuristic approach to solve QMDP, and the corresponding α -quantile, denoted as MV, can be computed by treating the expected total cost for the policy obtained by solving the mean value problem under each scenario as a realization of the corresponding random variable, and it provides an upper bound on OPT. Based on MV, the value of incorporating stochasticity of parameters into the MDP model can be measured as $VSS = MV - \text{OPT}$ and $\%VSS = 100 \times \frac{MV - \text{OPT}}{MV}$. The negative $\%VSS$ values in Table 2 are due to the instances that terminate with a suboptimal OPT value because of the time limit. The results show that by incorporating uncertainty in the parameters, the α -quantile value can be improved by 2.26%, 3.77% and 3.84% on average with a maximum improvement of 14.40%, 22.94% and 23.86% at $\alpha = 1, 0.95, 0.90$, respectively. Hence, it is possible to achieve significant reductions in risk by incorporating parameter uncertainty into MDPs, and thereby reducing the possibility of undesirable outcomes. Additionally, just as in the case of VPI, the value of considering parameter uncertainty is usually higher for risk-averse decision making at lower confidence levels.

5.3 The impact of incorporating risk aversion Given the advantages of incorporating parameter uncertainty into MDPs, next, we analyze how the performance of a policy would be affected by decision makers' attitude towards risk. For rarely occurring events such as disasters, it is important to ensure that the selected policy performs well even under undesirable realizations of random parameters. Our goal in this section is to demonstrate potential reductions in risk achieved by solving the proposed risk-averse model instead of the expected value problem introduced by Steimle et al. (2018b) for such cases where the decision makers are concerned about their performance with respect to the worst α -quantile. We point out that an appropriate modeling approach and confidence level α , which reflect decision makers' preferences and attitude towards parameter uncertainty, can be determined by performing a sensitivity

analysis using our proposed model at different α values as well as the expected value problem.

Table 3 reports the α -quantile value for the policy obtained by solving the expected value problem given in Steimle et al. (2018b) under a time limit of one hour, namely “E-VaR”, and compares it to the best α -quantile value obtained within one hour, denoted as “VaR*” at confidence levels $\alpha = 0.90, 0.95, 1$ for small instances with 50–100 scenarios and large instances with 250–500 scenarios. The percentage deviation of the quantile value for the expected value policy from the best quantile value is given under the corresponding column “E-VaR – VaR* (%)” as $100 \times \frac{\text{E-VaR} - \text{VaR}^*}{\text{E-VaR}}$. Each blue superscript denotes the number of replications for which the corresponding expected value problem or the α -quantile problem terminates with a feasible solution due to time limit. Note that negative values are reported for the instances with larger number of scenarios for which the expected value policy performs better than the VaR-minimizing policy due to the increased computational complexity of the quantile optimization problem. The results in Table 3 show that the policy obtained by solving the expected value problem may perform worse than the VaR-optimizing policy by up to 3.96%, 5.30% and 5.79% for smaller problem instances and up to 3.54%, 3.05% and 3.02% for large instances in terms of the α -quantile value, for $\alpha = 1, 0.95, 0.90$, respectively. Thus, minimizing the VaR objective instead of the expected value may be beneficial for taking extreme outcomes into consideration and the proposed risk-averse modeling framework may provide significant savings in the quantile value.

We additionally analyze the optimal policies obtained for the quantile optimization problem QMDP-D at different confidence levels, the associated expected value problem (EV) and the mean value problem (MV) in comparison to the optimal policies of distinct scenarios. Figure 1 illustrates the behaviour of the optimal policies under different settings for a particular instance with five scenarios, an inventory capacity of 50 units and at most four vehicles, i.e., $|\mathcal{S}| = 5$, $|\mathcal{H}| = 6$, $|\mathcal{A}| = 5$. Figure 1a demonstrates the savings in the quantile function value when the policy minimizing α -quantile is used instead of the optimal policies of the expected value problem and the mean value problem at different α values. While the EV policy preserves a consistent gap with the optimal quantile value, the performance of the MV policy fluctuates significantly. Figure 1a also illustrates the nonconvex nature of the quantile function. The performances of the quantile optimizing policy, EV policy and the MV policy converge as α gets closer to 0.80, whereas the performance differences grow for α values larger than 0.80. Additionally, Figure 1b reports the weekly supply and demand rates in each scenario and Figure 1c presents the optimal policies considering these scenarios under different settings. In Figure 1c, the dashed lines represent the optimal policies for each scenario independently. The solid line marked with diamonds corresponds to the quantile-optimizing policy at $\alpha = 1$, where four vehicles are dispatched whenever the available inventory level drops to 10, three vehicles if the inventory level is in the interval (10, 20], and so on. On the other hand, the policy minimizing 0.60-quantile, depicted with the solid line marked with squares, is the least conservative: three vehicles are dispatched whenever there is no stocked items, a single vehicle is used for any inventory level

in the interval $(0,20]$ and none otherwise. It can be seen that the optimal policies for scenarios 2, 4 and 5, which have a total probability of 0.60, are similar to each other, while scenarios 1 and 3 also have similar optimal policies and balanced supply and demand rates. As expected, the optimal policy for the quantile optimization problem at $\alpha = 1$ is more aligned with the scenarios 2, 4, and 5, while the optimal policy for the quantile optimization problem at $\alpha = 0.60$ is closer to the optimal policies of scenarios 1 and 3. The EV and the MV policies (depicted with solid lines marked with triangles and circles, respectively) are more moderate and stay in between the quantile-minimizing policies at $\alpha = 0.60$ and $\alpha = 1$.

Furthermore, we observe, empirically, that the optimal policies for quantile-optimizing humanitarian inventory management problem are monotone policies, where the number of vehicles to be dispatched decreases as the inventory level increases, as illustrated in Figure 1c. Note that the provable existence of a monotone optimal policy for a classical MDP is usually contingent upon certain properties of the cost and transition probability functions such as monotonicity and subadditivity (see, e.g., Puterman, 2014). However, the nonconvexity of the VaR objective poses significant challenges for proving the existence of optimal monotone policies even when these properties are satisfied. Although it is difficult to make theoretical conclusions about the preservation of monotonicity over a nonconvex quantile function even if there is a monotone optimal policy for each scenario considered independently, we believe that a monotone policy of ordering nondecreasing quantities of supply as the inventory level decreases is an intuitive (and interpretable) policy that may be appealing to the decision makers in a humanitarian context. Using such information on the characteristics of the desired policy, it is possible to add the following constraint into our model to enforce a monotone policy structure

$$w_{ia} \leq \sum_{a' \in \mathcal{A}: a' \leq a} w_{i'a'}, \quad a \in \mathcal{A}, i, i' \in \mathcal{H} : i' > i. \quad (16)$$

Henceforth, we refer to the problem QMDP-D with additional monotonicity constraint (16) as QMDP-M. Constraint (16) ensures that if n_a vehicles are dispatched at the inventory level i , then for any higher inventory level $i' > i$, the number of vehicles dispatched can be at most n_a (assuming $n_a > n'_a$ for $a, a' \in \mathcal{A}$ such that $a > a'$, as given in the problem statement). Incorporating more information on the characteristics of the desired policy may also provide computational advantages as it reduces the solution space.

5.4 Comparison of Solution Approaches In this section, we evaluate the computational performance of the proposed heuristic methods and the MILP reformulations of QMDP-D in terms of solution times and optimality gaps achieved within the time limit of one hour. Our preliminary experiments using a branch-and-cut algorithm with mixing inequalities indicate poor computational performance because of the difficulty with balancing the strength of bounds obtained from the mixing inequalities with the

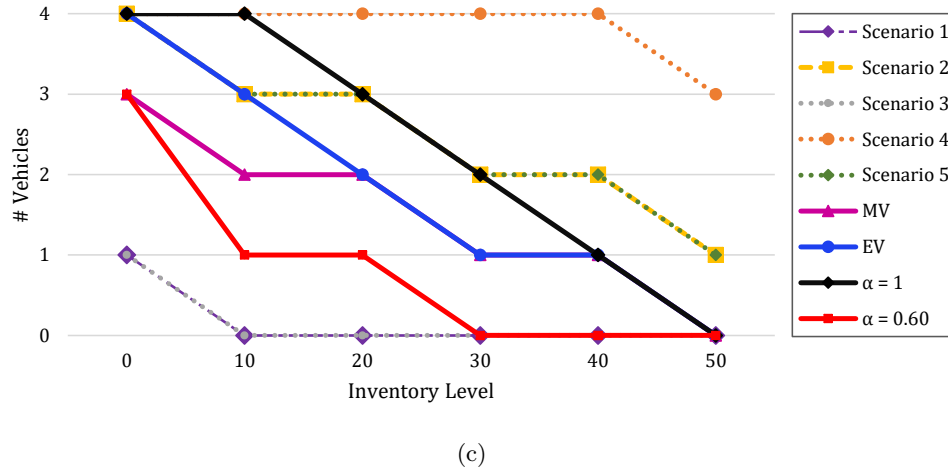
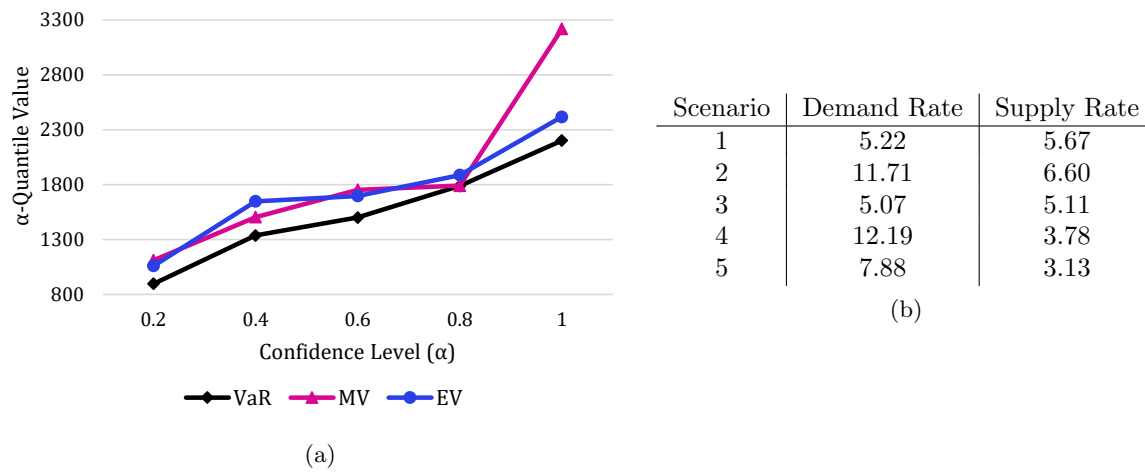


Figure 1: Comparison of the (a) quantile values for policies minimizing the VaR, expected value and mean value, (b) supply and demand rates for each scenario and (c) policies obtained under different settings.

computational effort required to solve the subproblems in each iteration of the algorithm. Hence, the implementation details and results of the branch-and-cut algorithm are omitted here to conserve space; see Meraklı (2018) for more information and our preliminary results on problem instances with fewer scenarios.

We first examine the quality of the feasible solutions generated by the proposed heuristic methods: the mean value (MV) problem, Algorithm 1, and the problem with monotone policies (QMDP-M). It bears emphasizing that QMDP-M is considered as a heuristic method because it produces feasible (not necessarily optimal) solutions to the original problem QMDP-D due to the additional constraint that restricts the optimal policies to be monotone in the sense that the number of vehicles to be dispatched decreases as the inventory level increases. QMDP-M is formulated by adding constraint (16) into the big- M reformulation of QMDP-D where constraint (6e) is replaced by constraints (10), and the big- M terms in (6d) and (10) are set to $M = b_u - b_l$ and $M_{is} = \bar{v}_i^s - \underline{v}_i^s$ for all $i \in \mathcal{H}$, $s \in \mathcal{S}$. Moreover, we add constraints (13) and (7c), and apply scenario elimination using the best feasible solution as an upper bound as described in Section 3.1. Column “Time” of Table 4 reports the computation time in seconds and column “%Diff” reports the percentage difference with respect to the best solution of QMDP-D for these three heuristic solution methods at confidence levels $\alpha = 0.95, 1.00$. We additionally report the number of replications that cannot be solved to optimality due to time limit and terminate with a feasible solution using blue superscripts. The percentage difference for a policy with quantile function value obj is computed as $100 \times \frac{(obj - obj^*)}{obj}$, where obj^* refers to the best objective function value of the quantile minimization problem obtained within the time limit. The results show that the MV problem and Algorithm 1 produce feasible policies of similar quality with respect to the average percentage differences. For $\alpha = 1.00$, the average percentage difference for Algorithm 1 is 1.65%, while the solutions provided by MV problem are on average within 2.26% of the best deterministic policy. The maximum differences for both algorithms are 14.40%. Similarly, for $\alpha = 0.95$, the average percentage difference is 3.77% for the MV solution and 3.90% for Algorithm 1. However, the MV problem produces feasible policies in at most 15.09 seconds while Algorithm 1 may take more than one hour for large instances. Note that the performance of Algorithm 1 could be potentially improved by performing a local search on the scenario decision vector \mathbf{z} . On the other hand, due to the reduced search space of monotone policies, the heuristic QMDP-M terminates relatively quickly (in 47.23 and 635.51 seconds for $\alpha = 1.00$ and 0.95, respectively), and the monotone policy it generates turns out to be at least as good as the best solution of the original problem for almost all of our instances.

Next we evaluate the computational performance of the proposed MILP reformulations of QMDP-D at confidence level $\alpha = 0.95$ under the settings detailed below. Note that these MILP reformulations provide an exact representation of problem QMDP-D.

- *BM – Basic*: It corresponds to formulation QMDP-D where constraint (6e) is replaced by constraint (10) with additional big- M terms. All big- M terms are fixed as 10^6 .
- *BM*: In addition to the setting in *BM – Basic*, the big- M term in (6d) is set to $M = b_u - b_l$ and the big- M terms $M_{is} = \bar{v}_i^s - \underline{v}_i^s$, $i \in \mathcal{H}, s \in \mathcal{S}$ in constraint (10). Moreover, constraints (13) and (7c) are added and the scenario elimination procedure described in Section 3.1 is applied using the optimal monotone policy solution of QMDP-M as an upper bound. Gurobi solver is provided with the optimal monotone policy solution as an initial feasible solution.
- *MC – Basic*: It corresponds to the McCormick reformulation of QMDP-D with constraints (7) and (8) replacing (6e). All big- M terms are fixed as 10^6 and lower bounds are set to zero.
- *MC*: The big- M term in (6d) is set to $M = b_u - b_l$ in formulation *MC – Basic*. Additionally, the lower and upper bounds on variable v_i^s in (7) are $\ell_j^s = \underline{v}_i^s$ and $u_j^s = \bar{v}_i^s$ for $i \in \mathcal{H}, s \in \mathcal{S}$, respectively. The optimal monotone policy solution of QMDP-M is embedded into the solver as an initial feasible solution. Constraint (13) is added and scenario elimination is applied using the optimal monotone policy solution as an upper bound as described in Section 3.1.

In Table 5, we report the total solution time in seconds (“Time (s)”), the best optimality gap achieved within the time limit (“Gap (%)”), and the number of nodes in the branch-and-bound tree (“Nodes”) for the MILP reformulations detailed above. The optimality gap values are computed as $100 \times \frac{ub-lb}{ub}$, where ub and lb correspond to the best upper and lower bounds on the optimal objective function value achieved at termination, respectively. The reported solution times do not include the time to obtain bounds and initial solutions. Each row corresponds to the average of five replications and each blue (resp., red) superscript in columns “Time (s)” indicates the number of replications that exceeded the time limit with (resp., without) a feasible solution. The results in Table 5 show that the MILP models with additional big- M terms (*BM* and *BM – Basic*) outperform the corresponding MILP models with McCormick reformulation (*MC* and *MC – Basic*). Out of 225 instances *BM – Basic* can solve 60 instances to optimality within the time limit, whereas *MC – Basic* can achieve optimality within the time limit for only 11 instances and fails to obtain a feasible solution for six instances. Our results also demonstrate the effectiveness of preprocessing methods for both formulations. The solution times are reduced by 48.79% and 24.87% for *BM* and *MC*, respectively. Out of 225, the number of instances that can be solved to optimality increases from 60 instances to 146 for *BM* and from 11 instances to 76 for *MC*. Similarly, the percentage optimality gaps and the number of nodes in the branch-and-bound tree usually decrease after the preprocessing methods are applied. Overall, formulation *BM* with preprocessing demonstrates the best performance with respect to both the solution times and the number of instances that can be solved to optimality.

Table 2: Value of perfect information and value of stochastic solution of the quantile optimization problem

Instance			%VPI			%VSS		
$ \mathcal{S} $	$ \mathcal{H} $	$ \mathcal{A} $	$\alpha = 1$	$\alpha = 0.95$	$\alpha = 0.90$	$\alpha = 1$	$\alpha = 0.95$	$\alpha = 0.90$
50	6	3	0.48	1.43	2.92	0.91	4.30	3.08
50	6	4	1.59	6.74	12.91	7.66	9.87	13.59
50	6	5	2.09	14.80	17.30	13.07	22.94	23.86
50	11	3	0.08	0.41	1.11	0.41	0.97	1.67
50	11	4	0.28	1.90	6.18	1.43	3.27	4.61
50	11	5	0.45	6.92	8.89 ⁴	2.34	5.97	5.67 ⁴
50	16	3	0.05	0.30	1.20	0.25	0.84	1.50
50	16	4	0.18	0.77 ¹	4.07 ⁴	1.04	1.78 ¹	2.31 ⁴
50	16	5	0.30	3.44 ⁴	7.27 ⁵	1.99	3.02 ⁴	4.90 ⁵
100	6	3	0.18	0.69	3.61	0.44	0.98	1.17
100	6	4	2.22	8.11	13.77	3.93	4.71	3.89
100	6	5	4.19	20.03	25.16	14.40	15.93	14.63
100	11	3	0.04	0.12	1.36	0.20	0.38	0.92
100	11	4	0.35	2.61	8.00 ²	1.72	2.60	4.57 ²
100	11	5	0.86	10.70 ⁴	17.25 ⁵	2.84	7.10 ⁴	5.82 ⁵
100	16	3	0.03	0.07	1.19 ¹	0.17	0.23	0.84 ¹
100	16	4	0.19	1.79 ²	5.23 ⁵	1.01	1.72 ²	5.00 ⁵
100	16	5	0.44	5.72 ⁵	11.44 ⁵	2.52	6.05 ⁵	3.81 ⁵
250	6	3	0.08	1.38	3.35	0.32	0.96	1.46
250	6	4	1.61	6.98	11.75	3.42	3.41	3.47
250	6	5	1.04	13.13	24.77	11.42	15.81	15.03
250	11	3	0.03	0.27	0.89	0.15	0.45	1.74
250	11	4	0.37	2.71 ¹	5.90 ⁴	1.03	1.28 ¹	2.29 ⁴
250	11	5	0.30	6.50 ⁵	13.55 ⁵	1.70	4.00 ⁵	6.08 ⁵
250	16	3	0.02	0.15	0.70 ¹	0.16	0.34	1.12 ¹
250	16	4	0.38 ¹	1.92 ⁴	4.04 ⁵	0.52 ¹	0.99 ⁴	2.14 ⁵
250	16	5	0.23	5.81 ⁵	9.69 ⁵	1.96	2.62 ⁵	4.32 ⁵
500	6	3	0.10	1.04	3.00	0.31	0.65	1.14
500	6	4	1.84	7.61	11.99	2.86	2.72	2.92
500	6	5	0.87	16.97 ²	23.45 ⁴	7.12	14.78 ²	15.70 ⁴
500	11	3	0.04	0.24	1.10	0.16	0.21	0.48
500	11	4	0.38	2.64 ³	5.71 ⁵	1.03	1.72 ³	2.09 ⁵
500	11	5	0.29	7.79 ⁵	16.86 ⁵	1.44	3.74 ⁵	3.02 ⁵
500	16	3	0.02	0.20	0.44 ¹	0.16	0.20	0.39 ¹
500	16	4	0.25	1.60 ⁵	6.20 ⁵	0.49	1.13 ⁵	-0.44 ⁵
500	16	5	0.22 ¹	6.84 ⁵	15.60 ⁵	0.91 ¹	1.51 ⁵	-1.53 ⁵
1000	6	3	0.15	1.63	3.50	0.26	0.45	0.68
1000	6	4	1.57	7.66 ¹	12.92	1.89	2.44 ¹	2.35
1000	6	5	0.67	14.17 ⁵	23.70 ⁵	3.12	13.44 ⁵	13.19 ⁵
1000	11	3	0.05	0.57	1.26 ³	0.14	0.20	0.32 ³
1000	11	4	0.55	3.19 ⁵	7.27 ⁵	0.93	1.17 ⁵	1.31 ⁵
1000	11	5	0.24	6.89 ⁵	18.49 ⁵	1.31	2.45 ⁵	-0.08 ⁵
1000	16	3	0.04	0.44 ²	2.54 ⁵	0.15	0.26 ²	-1.01 ⁵
1000	16	4	0.50 ²	3.36 ⁵	7.70 ⁵	1.06 ²	0.13 ⁵	-2.26 ⁵
1000	16	5	0.19 ¹	6.34 ⁵	17.44 ⁵	1.36 ¹	-0.07 ⁵	-5.03 ⁵
Average			0.58 ^{0.11}	4.77 ^{1.76}	8.95 ^{2.42}	2.26 ^{0.11}	3.77 ^{1.76}	3.84 ^{2.42}
Maximum			4.19 ²	20.03 ⁵	25.16 ⁵	14.40 ²	22.94 ⁵	23.86 ⁵

Table 3: Performance of the optimal policy of the expected value problem at the α -quantile

Small Instances						Large Instances					
Instance			E-VaR – VaR* (%)			Instance			E-VaR – VaR* (%)		
$ \mathcal{S} $	$ \mathcal{H} $	$ \mathcal{A} $	$\alpha = 1$	$\alpha = 0.95$	$\alpha = 0.90$	$ \mathcal{S} $	$ \mathcal{H} $	$ \mathcal{A} $	$\alpha = 1$	$\alpha = 0.95$	$\alpha = 0.90$
50	6	3	1.04	1.47	1.83	250	6	3	0.41	0.98	1.00
50	6	4	2.26	4.08	3.46	250	6	4	2.70	2.28	2.07
50	6	5	3.96	5.30	5.63	250	6	5	3.54	3.05	2.29
50	11	3	0.57	0.99	0.60	250	11	3	0.27	0.59	0.80
50	11	4	1.22 ²	2.43 ²	3.58 ²	250	11	4	0.82 ⁵	0.66 ⁵	1.14 ⁵
50	11	5	2.12 ⁵	3.34 ⁵	5.79 ⁵	250	11	5	1.52 ⁵	2.66 ⁵	2.03 ⁵
50	16	3	0.36 ⁵	0.70 ⁵	0.85 ⁵	250	16	3	0.28 ⁵	0.44 ⁵	0.71 ⁵
50	16	4	0.94 ⁵	1.63 ⁵	2.31 ⁵	250	16	4	0.55 ⁵	0.78 ⁵	1.82 ⁵
50	16	5	1.84 ⁵	2.45 ⁵	4.25 ⁵	250	16	5	1.81 ⁵	2.19 ⁵	3.02 ⁵
100	6	3	0.67	1.48	1.86	500	6	3	0.41	0.98	0.94
100	6	4	3.42	3.22	2.82	500	6	4	1.92	1.09	0.76
100	6	5	2.86	2.25	3.63	500	6	5	3.16	2.02 ²	2.26 ⁴
100	11	3	0.34	0.75	1.53	500	11	3	0.24	0.49	0.67
100	11	4	1.67 ⁵	2.56 ⁵	2.68 ⁵	500	11	4	0.75 ⁵	0.88 ⁵	1.08 ⁵
100	11	5	2.25 ⁵	4.68 ⁵	1.76 ⁵	500	11	5	1.20 ⁵	2.11 ⁵	0.11 ⁵
100	16	3	0.30 ⁵	0.48 ⁵	1.09 ⁵	500	16	3	0.28 ⁵	0.34 ⁵	0.50 ⁵
100	16	4	0.86 ⁵	1.33 ⁵	3.73 ⁵	500	16	4	0.57 ⁵	0.64 ⁵	–0.98 ⁵
100	16	5	2.33 ⁵	4.14 ⁵	3.32 ⁵	500	16	5	0.97 ⁵	1.11 ⁵	–3.31 ⁵
Average			1.61 ^{2.61}	2.40 ^{2.61}	2.82 ^{2.61}				1.19 ^{2.78}	1.29 ^{2.89}	0.94 ³
Maximum			3.96 ⁵	5.30 ⁵	5.79 ⁵				3.54 ⁵	3.05 ⁵	3.02 ⁵

Table 4: Evaluation of the heuristic algorithms in terms of solution times and optimality gaps

Instance			$\alpha = 1.00$						$\alpha = 0.95$					
			MV		Algorithm 1		QMDP-M		MV		Algorithm 1		QMDP-M	
$ S $	$ \mathcal{H} $	$ \mathcal{A} $	Time	%Diff	Time	%Diff	Time	%Diff	Time	%Diff	Time	%Diff	Time	%Diff
50	6	3	0.30	0.91	13.27	0.46	0.42	0.00	0.35	4.29	15.17	0.79	0.51	0.00
50	6	4	0.35	7.66	17.46	3.72	0.61	0.00	0.40	9.87	20.18	4.13	1.37	0.00
50	6	5	0.41	13.07	22.17	9.61	0.85	0.00	0.46	22.94	25.19	20.47	1.89	0.00
50	11	3	0.77	0.41	36.71	0.06	1.31	0.00	0.90	0.97	42.28	0.22	1.79	0.00
50	11	4	1.04	1.43	49.44	0.41	2.42	0.00	1.05	3.27	55.41	0.92	2.90	0.00
50	11	5	1.11	2.34	63.66	0.96	2.38	0.00	1.21	5.97	69.71	4.14	9.45	0.00
50	16	3	1.51	0.25	72.66	0.04	2.18	0.00	1.69	0.84	82.05	0.15	3.81	0.00
50	16	4	1.75	1.04	95.09	0.16	4.47	0.00	1.98	1.78	108.25	0.34	7.19	0.00
50	16	5	2.10	2.00	125.16	0.57	7.12	0.00	2.34	3.02	137.1	1.73	28.40	0.00
100	6	3	0.42	0.43	26.85	0.16	0.64	0.00	0.46	0.98	32.32	0.98	1.24	0.00
100	6	4	0.48	3.93	35.90	3.93	1.28	0.00	0.52	4.72	43.14	4.72	2.61	0.00
100	6	5	0.53	14.40	44.78	14.40	2.10	0.00	0.58	15.93	53.82	22.95	10.55	0.00
100	11	3	1.08	0.20	75.39	0.03	1.86	0.00	1.44	0.38	86.3	0.12	5.01	0.00
100	11	4	1.25	1.72	101.98	0.47	4.03	0.00	1.36	2.60	115.39	1.37	10.68	0.00
100	11	5	1.42	2.84	128.15	1.42	8.61	0.00	1.54	7.10	146.25	4.67	56.22	0.00
100	16	3	2.10	0.17	151.24	0.02	4.14	0.00	2.37	0.23	175.87	0.06	9.60	0.00
100	16	4	2.43	1.01	203.29	0.15	10.21	0.00	2.72	1.72	235.28	0.74	28.98	0.00
100	16	5	2.72	2.52	252.33	0.82	19.17	0.00	2.93	6.05	274.84	2.19	439.71	0.00
250	6	3	0.76	0.32	68.77	0.00	0.95	0.00	0.83	0.96	94.41	0.86	4.28	0.00
250	6	4	0.82	3.42	90.81	3.42	3.17	0.00	0.88	3.41	127.28	3.41	10.97	0.00
250	6	5	0.88	11.42	113.84	11.42	4.53	0.00	0.94	15.81	156.05	22.77	36.51	0.00
250	11	3	2.01	0.15	193.21	0.01	5.34	0.00	2.07	0.45	241.07	0.18	16.07	0.00
250	11	4	2.16	1.03	257.62	0.32	11.78	0.00	2.24	1.28	321.58	1.16	44.86	0.00
250	11	5	2.30	1.70	315.43	1.21	20.72	0.00	2.46	4.00	397.67	4.73	205.50	0.00
250	16	3	3.84	0.16	374.71	0.02	10.20	0.00	4.17	0.34	427.55	0.09	34.22	0.00
250	16	4	4.34	0.52	508.76	0.17	31.86	0.00	4.22	0.99	580.6	1.12	105.47	0.00
250	16	5	4.64	1.96	630.10	0.97	73.32	0.00	4.53	2.62	718.95	2.29	1176.84	0.00
500	6	3	1.34	0.31	138.24	0.06	4.20	0.00	1.51	0.65	245.97	0.65	23.23	0.00
500	6	4	1.41	2.87	187.98	2.87	8.06	0.00	1.60	2.72	330.05	2.72	114.54	0.00
500	6	5	1.48	7.12	234.11	7.12	8.67	0.00	1.65	14.78	413.66	29.01	1413.45	0.00
500	11	3	3.56	0.16	395.40	0.02	12.16	0.00	3.85	0.21	582.49	0.11	69.80	0.00
500	11	4	3.81	1.03	542.51	0.35	24.67	0.00	3.89	1.72	741.39	1.72	233.66	0.00
500	11	5	3.78	1.43	647.11	0.88	47.36	0.00	4.33	3.74	1028.56	3.82	3221.29 ⁴	0.00
500	16	3	6.82	0.16	768.25	0.02	31.17	0.00	7.55	0.21	1092.92	0.07	255.53	0.00
500	16	4	7.20	0.49	1046.28	0.15	75.58	0.00	7.83	1.13	1448.83	0.95	881.26	0.00
500	16	5	7.58	0.91	1277.61	0.52	238.15	0.00	7.89	1.51	1723.18	1.65	3046.32 ⁴	0.00
1000	6	3	2.20	0.26	243.07	0.22	8.32	0.00	2.80	0.45	707.36	0.71	93.71	0.00
1000	6	4	2.28	1.89	326.49	1.89	20.36	0.00	2.90	2.44	945.14	2.44	603.58	0.00
1000	6	5	2.47	3.12	422.04	3.12	18.71	0.00	2.69	13.44	1109.15	20.43	2112.05	0.00
1000	11	3	6.40	0.14	768.36	0.06	35.94	0.00	7.42	0.20	1643.97	0.20	306.78	0.00
1000	11	4	7.18	0.93	1117.65	0.35	122.17	0.00	6.64	1.17	1950.06	1.17	2586.68 ¹	0.00
1000	11	5	7.17	1.31	1371.73	0.69	168.70	0.00	6.97	2.45	2455.05	2.63	3600.56 ⁵	0.00
1000	16	3	12.68	0.15	1550.63	0.04	218.16	0.00	11.08	0.26	2376.31	0.09	710.92	0.00
1000	16	4	13.37	1.06	2080.92	0.31	378.42	0.00	13.85	0.13	3607.18 ¹	0.24	3467.09 ⁴	0.20
1000	16	5	15.09	1.36	2876.66	0.59	469.08	0.00	12.61	-0.07	3912.40 ²	-0.45	3601.10 ⁵	0.00
Average			3.32	2.26	445.86	1.65	47.23	0.00	3.42	3.77	691.05 ^{0.07}	3.90	635.51 ^{0.51}	0.00
Maximum			15.09	14.40	2876.66	14.40	469.08	0.00	13.85	22.94	3912.40 ²	29.01	3601.10 ⁵	0.20

Table 5: Comparison of MILP reformulations of QMDP-D

Instance			<i>BM</i>			<i>BM – basic</i>			<i>MC</i>			<i>MC – basic</i>		
$ S $	$ H $	$ A $	Time (s)	Gap (%)	Nodes	Time (s)	Gap (%)	Nodes	Time (s)	Gap (%)	Nodes	Time (s)	Gap (%)	Nodes
50	6	3	0.53	0.00	74.0	3.71	0.00	1576.4	8.81	0.00	94.40	68.97	0.00	2524.0
50	6	4	1.54	0.00	477.6	123.76	0.00	9177.4	39.69	0.00	889.80	1284.51	0.00	17011.8
50	6	5	6.96	0.00	2702.2	188.00	0.00	15325.0	340.62	0.00	7263.60	3569.76 ⁴	74.79	28518.6
50	11	3	5.01	0.00	1115.4	595.36	0.00	39926.8	151.53	0.00	558.80	3600.01 ⁵	94.17	9179.4
50	11	4	48.97	0.00	9489.2	3482.07 ⁴	62.91	260029.0	1605.48 ¹	1.09	4852.20	3600.01 ⁵	98.60	5664.2
50	11	5	1371.54	0.00	180041.6	3600.00 ⁵	90.91	134324.6	2930.94 ⁴	6.19	9891.00	3600.01 ⁵	98.52	5257.2
50	16	3	48.41	0.00	4955.2	3600.00 ⁵	47.15	116557.8	996.60 ¹	0.15	1100.00	3600.06 ⁵	99.16	1449.0
50	16	4	773.00 ¹	0.26	30838.0	3600.00 ⁵	75.61	74818.2	2264.10 ³	0.55	1815.20	3600.04 ⁵	99.17	1281.2
50	16	5	2883.24 ⁴	2.72	66421.6	3600.00 ⁵	82.87	23468.2	3077.71 ⁴	3.03	2822.00	3600.04 ⁵	99.16	1041.4
100	6	3	1.11	0.00	60.0	14.21	0.00	2597.4	22.18	0.00	154.80	2260.04 ³	50.70	9626.6
100	6	4	6.76	0.00	1284.8	348.74	0.00	13624.4	338.6	0.00	3089.20	3600.04 ⁵	98.37	8594.2
100	6	5	109.59	0.00	9182.8	949.59	0.00	32120.0	3300.32 ⁴	11.39	11485.00	3600.01 ⁵	98.60	6170.4
100	11	3	1.75	0.00	23.2	2554.87 ¹	13.88	68821.4	279.21	0.00	361.20	3600.03 ⁵	99.00	2259.2
100	11	4	331.41	0.00	14263.8	3600.00 ⁵	91.92	40103.6	3102.97 ⁴	2.00	4304.20	3600.03 ⁵	98.95	1000.4
100	11	5	3483.97 ⁴	7.61	153772.0	3600.00 ⁵	97.98	30367.0	3600.02 ⁵	10.44	3403.60	3600.04 ⁵	99.07	720.0
100	16	3	4.07	0.00	7.0	3600.02 ⁵	69.47	21071.6	263.18	0.00	33.00	3600.05 ⁵	99.61	292.8
100	16	4	1762.18 ²	1.04	26697.6	3600.03 ⁵	89.07	6343.2	3600.02 ⁵	5.03	1107.20	3600.04 ⁵	99.49	261.0
100	16	5	3600.01 ⁵	5.07	38582.4	3600.12 ⁵	98.09	9030.0	3600.05 ⁵	10.41	1014.60	3600.08 ⁵	99.30	309.4
250	6	3	3.40	0.00	97.6	307.79	0.00	5921.6	88.21	0.00	142.60	3600.03 ⁵	98.91	3180.2
250	6	4	26.24	0.00	2023.4	1692.88	0.00	16706.6	1890.50 ¹	0.91	3756.60	3600.03 ⁵	99.14	1493.8
250	6	5	317.29	0.00	8098.8	3233.03 ²	34.70	38015.6	3600.11 ⁵	10.81	2443.00	3600.05 ⁵	99.13	1046.4
250	11	3	13.97	0.00	154.8	3600.00 ⁵	93.90	17657.2	1108.67 ¹	0.10	312.80	3600.55 ⁵	99.48	307.0
250	11	4	1558.46 ¹	0.49	13295.4	3600.04 ⁵	99.17	8746.2	3600.16 ⁵	7.79	698.80	3600.06 ⁵	99.47	238.4
250	11	5	3600.03 ⁵	5.19	28274.2	3600.06 ⁵	99.60	2696.2	3600.10 ⁵	10.92	334.00	3600.11 ⁵	99.48	125.4
250	16	3	42.39	0.01	602.0	3600.02 ⁵	97.76	4886.2	3117.60 ³	0.06	316.00	3600.12 ⁵	99.71	81.2
250	16	4	3035.12 ⁴	1.42	12251.6	3600.05 ⁵	99.67	1665.0	3600.30 ⁵	20.99	134.00	3600.27 ⁵	99.68	47.2
250	16	5	3600.01 ⁵	5.34	10113.2	3600.03 ⁵	99.93	4998.0	3600.32 ⁵	15.47	37.80	3600.26 ⁵	99.59	43.0
500	6	3	10.04	0.00	87.8	1219.99	0.00	8865.8	2299.07 ¹	1.00	326.20	3601.08 ⁵	99.37	728.2
500	6	4	228.39	0.00	3696.8	3479.89 ⁴	78.56	10500.6	3600.41 ⁵	41.51	0.80	3600.97 ⁵	99.40	486.8
500	6	5	3137.55 ²	2.42	25889.2	3600.01 ⁵	99.27	8176.0	3601.14 ⁵	68.80	0.60	3600.85 ⁵	99.46	462.4
500	11	3	30.83	0.00	116.2	3600.01 ⁵	99.16	5640.8	2659.70 ²	0.06	114.00	3600.05 ⁵	99.59	140.8
500	11	4	3077.46 ³	0.74	7469.6	3600.02 ⁵	99.69	1658.4	3600.25 ⁵	19.43	1.00	3600.10 ⁵	99.56	72.4
500	11	5	3600.02 ⁵	7.48	4394.4	3600.27 ⁵	99.75	1527.2	3600.24 ⁵	28.53	1.00	3600.13 ⁵	99.57	62.8
500	16	3	264.96	0.00	1215.0	3600.01 ⁵	99.63	1867.8	3600.14 ⁵	0.27	0.00	3600.26 ^{4,1}	99.75	10.5
500	16	4	3600.02 ⁵	1.36	2104.8	3600.03 ⁵	99.88	1170.6	3600.14 ⁵	37.67	0.20	3600.14 ^{4,1}	99.72	10.3
500	16	5	3600.04 ⁵	6.72	1725.2	3600.02 ⁵	100.00	1702.8	3600.16 ⁵	81.72	0.00	3600.64 ⁵	99.77	6.0
1000	6	3	74.47	0.00	638.0	3472.34 ⁴	78.97	6475.6	3600.08 ⁵	6.16	1.00	3600.10 ⁵	99.48	364.0
1000	6	4	1930.16 ¹	0.50	6128.6	3600.07 ⁵	99.37	4296.4	3600.16 ⁵	17.86	1.00	3600.05 ⁵	99.55	251.8
1000	6	5	3600.07 ⁵	8.44	11428.0	3600.39 ⁵	99.44	3478.2	3600.12 ⁵	43.53	1.00	3600.04 ⁵	99.53	170.0
1000	11	3	210.63	0.00	563.4	3600.09 ⁵	99.48	2491.0	3600.18 ⁵	14.97	0.00	3600.60 ⁵	99.75	1.0
1000	11	4	3600.03 ⁵	3.07	1293.8	3600.32 ⁵	99.81	1444.0	3600.18 ⁵	81.88	0.00	3600.39 ^{2,3}	99.86	1.0
1000	11	5	3600.02 ⁵	6.95	1353.0	3600.43 ⁵	99.89	1036.8	3600.16 ⁵	100.00	0.00	3600.28 ⁵	99.90	1.0
1000	16	3	2003.15 ²	0.24	1155.4	3600.23 ⁵	99.82	977.0	3600.21 ⁵	28.95	0.00	3600.43 ⁵	99.82	1.0
1000	16	4	3600.09 ⁵	3.58	461.2	3600.04 ⁵	99.97	809.0	3600.41 ⁵	85.56	0.00	3600.62 ^{4,1}	99.80	1.0
1000	16	5	3600.11 ⁵	6.30	533.0	3600.04 ⁵	99.99	696.6	3600.45 ⁵	100.00	0.00	3601.29 ⁵	99.96	1.0
Average			1475.67 ^{1,76}	1.71	15225.6	2881.52 ^{3,67}	68.83	23630.9	2584.25 ^{3,31}	19.45	1396.9	3439.85 ^{4,57,0,13}	93.22	2455.5
Maximum			3600.11 ⁵	8.44	180041.6	3600.43 ⁵	100.00	260029.0	3601.14 ⁵	100.00	11485.0	3601.29 ^{5,3}	99.96	28518.6

6. Conclusions In this study, we investigate the risk associated with parameter uncertainty in MDPs. We formulate the problem with the objective of minimizing the VaR of the expected total discounted cost of an MDP at a prespecified confidence level α and explore characteristics of the optimal policies. Assuming a discrete representation of uncertainty, we provide MINLP and MILP formulations considering randomized and deterministic policies, and propose preprocessing methods and heuristic algorithms that can be applied for both cases. The proposed modeling approach and solution algorithms are tested on an inventory management problem in the long term humanitarian relief operations context and our results show that the proposed modeling approach may provide significant reductions in risk arising from parameter uncertainty compared with solving an MDP with average system parameters or solving an expected value problem as proposed in [Steimle et al. \(2018b\)](#).

In contrast to classical MDPs in which system parameters are assumed to be known, all optimal policies in an MDP with uncertain parameters minimizing the associated VaR may be randomized. The results on our instances of the humanitarian inventory management problem indicate that policy randomization may provide additional savings in the expected discounted total cost, but it also amplifies computational complexity of the problem due to additional nonlinearities. Hence, we recommend the use of deterministic policies especially in humanitarian settings, where implementation of randomized policies may raise ethical concerns. On the other hand, our mixed-integer nonlinear programming formulation QMDP-R can be used for application areas such as power management, infrastructure maintenance and rehabilitation and queuing systems, where randomized policies are effectively implemented.

Comparing the solution methods for the deterministic policy case, formulation *BM* with additional big-*M* terms outperforms the formulations based on McCormick envelopes. Our preprocessing procedures also provide significant improvements on the computational performance of both formulations. In addition, we provide heuristic methods and show that QMDP-M, enforcing monotonicity constraints on the optimal policy, performs best since it reduces solution times remarkably and provides feasible solutions that are also optimal for the original problem on our problem instances.

It is worthy to note that, just as policy randomization, it may also be possible to gain from utilizing nonstationary policies in a setting where the uncertain parameters are learned over time. However, computing a nonstationary optimal policy that minimizes the risk arising from parameter uncertainty in an infinite-horizon setting is possible if the objective is to optimize a dynamic risk measure. On the other hand, the risk measure of interest in our paper, namely VaR, is not a dynamic risk measure and hence it is not amenable to dynamic updates. [Zhang et al. \(2014\)](#) discuss the challenges in using chance constraints in a dynamic framework and show that solving static optimization problems at each period in a rolling horizon fashion may result in violations in the chance constraint representing a service level requirement of the performance of the system over the entire planning horizon. Moreover, [Steimle](#)

et al. (2018b) demonstrate that learning the uncertain parameters over time significantly increases the computational complexity even in the case of the expected value criterion while providing low level of gains in the objective function. Hence, we believe that the use of dynamic risk measures for addressing parameter uncertainty in MDPs would be an interesting and nontrivial extension of this study. We refer interested readers to a recent paper by Dentcheva and Ruszczyński (2018), which provides a foundation for dynamic risk measures under distributional uncertainty.

Acknowledgment We thank the AE and the three anonymous referees whose comments improved the paper.

References

- Adulyasak, Y., Varakantham, P., Ahmed, A., and Jaillet, P. (2015). Solving uncertain MDPs with objectives that are separable over instantiations of model uncertainty. In *AAAI*, pages 3454–3460.
- Altman, E. (1999). *Constrained Markov decision processes*, volume 7. CRC Press.
- Atakan, S., Bülbül, K., and Noyan, N. (2017). Minimizing value-at-risk in single-machine scheduling. *Annals of Operations Research*, 248(1-2):25–73.
- Bagnell, J., Ng, A., and Schneider, J. (2001). Solving uncertain Markov decision problems. *Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-01-25*.
- Bäuerle, N. and Ott, J. (2011). Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379.
- Bellman, R. (2013). *Dynamic Programming*. Courier Corporation.
- Benini, L., Bogliolo, A., Paleologo, G., and De Micheli, G. (1999). Policy optimization for dynamic power management. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 18(ARTICLE):813–833.
- Benoit, D. and Van den Poel, D. (2009). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. *Expert Systems with Applications*, 36(7):10475–10484.
- Bertsimas, D. and Mišić, V. V. (2016). Robust product line design. *Operations Research*, 65(1):19–37.
- Bonmin. Basic Open-source Mixed INteger programming. <https://projects.coin-or.org/Bonmin>. Accessed: 2019-01-02.
- Boucherie, R. and van Dijk, N. (2017). *Markov Decision Processes in Practice*. Springer.
- Buchholz, P. and Scheftelowitsch, D. (2018). Computation of weighted sums of rewards for concurrent MDPs. *Mathematical Methods of Operations Research*, pages 1–42.
- Calafiore, G. and Campi, M. (2006). The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753.
- Chen, K. and Bowling, M. (2012). Tractable objectives for robust policy optimization. In *Advances in Neural Information Processing Systems*, pages 2069–2077.
- Chen, R. and Blankenship, G. (2002). Dynamic programming equations for constrained stochastic control. In *Proceedings of the 2002 American Control Conference*, volume 3, pages 2014–2022. IEEE.

- Deák, I. (1988). Multidimensional integration and stochastic programming. *Numerical techniques for stochastic optimization*, pages 187–200.
- DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., and Vogels, W. (2007). Dynamo: Amazon’s highly available key-value store. *ACM SIGOPS Operating Systems Review*, 41(6):205–220.
- Delage, E. and Mannor, S. (2010). Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213.
- Dentcheva, D. and Ruszczyński, A. (2018). Risk forms: representation, disintegration, and application to partially observable two-stage systems. *Mathematical Programming*, pages 1–21.
- Djonin, D. V. and Krishnamurthy, V. (2007). MIMO transmission control in fading channels—a constrained Markov decision process formulation with monotone randomized policies. *IEEE Transactions on Signal Processing*, 55(10):5069–5083.
- Fan, J. and Ruszczyński, A. (2018). Risk measurement and risk-averse control of partially observable discrete-time Markov systems. *Mathematical Methods of Operations Research*, pages 1–24.
- Feinberg, E. A. and Reiman, M. I. (1994). Optimality of randomized trunk reservation. *Probability in the Engineering and Informational Sciences*, 8(4):463–489.
- Feng, M., Wächter, A., and Staum, J. (2015). Practical algorithms for value-at-risk portfolio optimization problems. *Quantitative Finance Letters*, 3(1):1–9.
- Ferreira, G., Arruda, E., and Marujo, L. (2018). Inventory management of perishable items in long-term humanitarian operations using Markov decision processes. *International Journal of Disaster Risk Reduction*, 31:460–469.
- Givan, R., Leach, S., and Dean, T. (2000). Bounded-parameter Markov decision processes. *Artificial Intelligence*, 122:71–109.
- Hart, W. E., Laird, C. D., Watson, J.-P., Woodruff, D. L., Hackebeil, G. A., Nicholson, B. L., and Sirola, J. D. (2012). *Pyomo-optimization modeling in python*, volume 67. Springer.
- Iyengar, G. (2005). Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280.
- Küçükyavuz, S. (2012). On mixing sets arising in chance-constrained programming. *Mathematical Programming*, 132(1-2):31–56.
- Küçükyavuz, S. and Noyan, N. (2016). Cut generation for optimization problems with multivariate risk constraints. *Mathematical Programming*, 159(1-2):165–199.
- Küçükyavuz, S. and Sen, S. (2017). An introduction to two-stage stochastic mixed-integer programming. In *Leading Developments from INFORMS Communities*, pages 1–27. INFORMS.
- Liu, X., Kılınç-Karzan, F., and Küçükyavuz, S. (2017a). On intersection of two mixing sets with applications to joint chance-constrained programs. *Mathematical Programming*, pages 1–40.
- Liu, X., Küçükyavuz, S., and Luedtke, J. (2016). Decomposition algorithms for two-stage chance-constrained programs. *Mathematical Programming*, 157(1):219–243.
- Liu, X., Küçükyavuz, S., and Noyan, N. (2017b). Robust multicriteria risk-averse stochastic programming models. *Annals of Operations Research*, 259(1):259–294.
- Luedtke, J. (2014). A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support. *Mathematical Programming*, 146(1-2):219–244.
- Luedtke, J. and Ahmed, S. (2008). A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699.

- Mannor, S., Mebel, O., and Xu, H. (2016). Robust MDPs with k -rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509.
- Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. (2007). Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322.
- McCormick, G. (1976). Computability of global solutions to factorable nonconvex programs: Part I - Convex underestimating problems. *Mathematical Programming*, 10(1):147–175.
- Meraklı, M. (2018). *Risk-Averse Optimization in Multicriteria and Multistage Decision Making*. PhD thesis, University of Washington.
- Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798.
- Noyan, N., Meraklı, M., and Küçükyavuz, S. (2019). Two-stage stochastic programming under multivariate risk constraints with an application to humanitarian relief network design. *Mathematical Programming*, pages 1–39.
- Pavlikov, K., Veremyev, A., and Pasiliao, E. L. (2018). Optimization of Value-at-Risk: computational aspects of MIP formulations. *Journal of the Operational Research Society*, 69(5):676–690.
- Puterman, M. (2014). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons.
- Rottkemper, B., Fischer, K., and Blecken, A. (2012). A transshipment model for distribution and inventory relocation under uncertainty in humanitarian operations. *Socio-Economic Planning Sciences*, 46(1):98–109.
- Ruszczynski, A. (2010). Risk-averse dynamic programming for Markov decision processes. *Mathematical programming*, 125(2):235–261.
- Satia, J. and Lave Jr, R. (1973). Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740.
- Sinha, S. and Ghatge, A. (2016). Policy iteration for robust nonstationary Markov decision processes. *Optimization Letters*, 10(8):1613–1628.
- Smilowitz, K. and Madanat, S. (2000). Optimal inspection and maintenance policies for infrastructure networks. *Computer-Aided Civil and Infrastructure Engineering*, 15(1):5–13.
- Steimle, L., Ahluwalia, V., Kamdar, C., and Denton, B. (2018a). Decomposition methods for solving multi-model Markov decision processes. Optimization Online http://www.optimization-online.org/DB_FILE/2018/11/6958.pdf.
- Steimle, L., Kaufman, D., and Denton, B. (2018b). Multi-model Markov decision processes: A new method for mitigating parameter ambiguity. Optimization Online http://www.optimization-online.org/DB_FILE/2018/01/6434.pdf.
- Tewari, A. and Bartlett, P. (2007). Bounded parameter Markov decision processes with average reward criterion. In *International Conference on Computational Learning Theory*, pages 263–277. Springer.
- White III, C. and Eldeib, H. (1994). Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749.
- Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183.
- Xu, H. and Mannor, S. (2009). Parametric regret in uncertain Markov decision processes. In *Proceedings of the 48th IEEE Conference on Decision and Control*, pages 3606–3613. IEEE.
- Xu, H. and Mannor, S. (2011). Probabilistic goal Markov decision processes. In *Proceedings of 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*, pages 2046–2052.

- Xu, H. and Mannor, S. (2012). Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 37(2):288–300.
- Yu, P. and Xu, H. (2016). Distributionally robust counterpart in Markov decision processes. *IEEE Transactions on Automatic Control*, 61(9):2538–2543.
- Zhang, M., Küçükyavuz, S., and Goel, S. (2014). A branch-and-cut method for dynamic decision making under joint chance constraints. *Management Science*, 60(5):1317–1333.
- Zhao, M., Huang, K., and Zeng, B. (2017). A polyhedral study on chance constrained program with random right-hand side. *Mathematical Programming*, 166(1-2):19–64.