We thank the reviewers for their valuable feedback. We are glad that they found our approach "very interesting" (R4), that it "sets a ... high bar for alternative[s]" (R2); our results "competitive" (R1) and "very promising" (R3); and our experiments "refreshing"(R4). At the same time, the reviewers requested clarifications and improvements focused on the novelty of our approach, and the discussion of key ideas. We respond in detail below to major questions/comments. A final version of this paper will fully address all concerns raised by the reviewing team, including these major comments and other minor suggestions.

Novelty: We address an "important and understudied problem in the [BO] literature" (R1), that "is very interesting and can certainly find applications in important areas" (R4), using an "approach [that] brings together essentially all of the ingredients of a real-world approach to [BO]" (R2). Although many of the ideas come from the existing literature, "there is novelty in the combination of all of these ideas, which are numerous, ... and non-trivial." (R1).

Solving $\min_x \rho[F(x, W)]$ with BO, while leveraging the ability to choose x *and* w at query time is novel. As established by Janusevskis & Le Roche (2013), and Toscano-Palmerin & Frazier (2018), the ability to select w is crucial while optimizing $\mathbb{E}[F(x, W)]$. Risk measures differ significantly from expectation, as the posterior risk-measure objective is non-Gaussian and cannot be treated using a simple GP model. This introduces significant methodological challenges. BO with risk measures, with ability to choose w, requires overcoming these challenges and provides significant value.

Selecting w is realistic and dramatically improves performance : "select[ing] ... w to query may not be realistic" (R4). We left this implicit, but our setting is designed for making decisions about the real world using a *simulation*. For example, in our COVID-19 experiment, $F(x, W)$ is the output of a simulation that models the real world. In the simulation, we can set any W we want. Then, once we find an approximate test allocation x, we would institute that in the real world without knowing W. It is only at this point, after we choose a solution, that we cannot select W.

"the benchmark[s] ... select x, while you ... select both x and w"(R4). The existing literature cannot choose w while optimizing a risk measure. This and the corresponding statistical model are a major contribution of our paper and a significant source of novelty, as evidenced by the superior performance of random sampling under our statistical model.

Literature: R1 points out two recent papers that we were not aware of. Gong, et al. (2019) is actually not relevant: it considers parallel BO in a classical setting, and quantiles are used to induce diversity in the evaluation batches. Torossian, et al. (2020) studies $\min_x \text{VaR}[F(x, W)]$. Our approach differs in that i) while they only pick x, we choose w as well, which is critical when $\mathcal{W}$ is large; ii) we allow for noisy observations of $F(x, w)$, which would introduce additional bias in their method. We requested the code from the authors but they said that they were not yet ready to release it, preventing inclusion of a comparison at the moment.

Other comments: "techniques to approximate VaR and CVaR ... seem to only hold for uniform distributions" (R4). This is a simple misunderstanding, as we only use the uniform distribution for ease of exposition (see lines 142-5). In fact, the distributions in two of our numerical experiments are non-uniform (Branin Williams and COVID).

"the estimators ... were studied in [42] and [43]" (R3). The estimators in [42] and [43] are for a given function F where the only randomness is in W, and require conditions on distribution of $F(\cdot, W)$. In our case F is also random, and the gradient is propagated through multiple operators (see §4 of supplement). This ties into "use ... [of] envelope theorem ... popularized by Wu, et al. 2017" (R1) as the risk measures require more involved treatment than the posterior mean.

"QMC ... would offer faster convergence than simple [MC]" (R1). We use MC only for exposition. As discussed in the supplement, we sample using (qMC) Sobol sequences in practice. "[G-H] quadrature for integrating over f" (R1): The integrals are over ρ, which is non-Gaussian. "... "mild regularity" ... condition ... isn't continuity enough?" (R1) The detailed conditions are discussed in §4 of supplement. They include differentiability of the GP sample paths and conditions on the distribution of W.

"... experimental results ... comparing the run time... "(R4) An iteration of ρKG^{apx} takes between 1-3 minutes, running on 4 cores of an Intel® Xeon® Platinum 8124M processor. The computational cost is justified by the sampling efficiency, particularly when the function evaluations are expensive (hours). We will include relevant plots in the final version.

R1 and R4 ask for more discussion on the approximate method. ρKG^{apx} is inspired by the EI and KGCP algorithms, which each consider only the previously evaluated points and show good numerical performance. A low-level intuition is that the underlying GP model is an extrapolation of the data. Thus, these points carry an immense amount of information on the GP model and the posterior objective, which makes them an ideal set of candidates to consider.

"... high-level discussion or motivation wrt alternative ideas" (R1) "The details ... in ... 5.1 and 5.2 ... difficult to follow ..." (R3) We will expand the discussion of key ideas, and clarify the details of our approach.

"LBFGS ... used ... gradient is ... with noise" (R1) We use SAA to trade the biased stochastic problem with a biased and deterministic one. This allows the use of LBFGS which offers computational improvements over stochastic alternatives.