**Responses to Referee 2:**

We greatly thank for the valuable efforts of Dr. Caitlin Spence to improve our manuscript. The constructive comments led us to carefully review our results. For validation of the logistic models, we used another internal scheme that reuses the resamples for regression analyses. This would improve the validity of the models in this study. We generally accepted Dr. Spence's suggestions in the second round of revision process. Specific responses are following as per comment.

Page 1, Lines 15-20: I think the "should" in "stakeholders should accept the risk…" should be changed to "could." This implies a possible choice in a more neutral way.
→ We revised it as suggested (P1L16).

Page 2, line 5: Do the authors really mean "underutilization" here? This implies that there are good reasons to use GCM projections more, but practitioners choose not to use them despite this. I suggest rephrasing to "less utilization of GCM-led strategies" unless I have misunderstood.
→ We revised it as "hindering utilization of the GCM-led strategies" (P2L6).

Introduction is much clearer and makes sense. I'm still not sold on whether logistic regression is actually makes sense given their argument that a stochastic metric would be too computationally expensive to evaluate across the response surface.
→ We exemplified the Monte-Carlo approach to gauge the uncertainty of response surfaces. When applying the typical stochastic uncertainty analysis for the stress test, the computational costs could be multiplied manyfold (P2L5-L11).

Page 6: Is "instreamflows" the correct spelling here? Perhaps "instream flows"?
→ We globally corrected it throughout the manuscript.

Page 6, line 14-17: Could also be interesting to perturb demand across a wider range of uncertainty, but I understand there are limits.
→ We understand the necessity of perturbing the water demand, because socioeconomic conditions could change in various directions. However, it is beyond the scope of this work, and may require a different framework for assessments. We will consider this valuable comment in further studies.

Page 7, line 15 - Page 8, line 6: So only one stochastic series was really generated, then it was perturbed in 539 flavors of precipitation CV, mean, and mean temperature? The explanation of the logistic model on Page 8 lines 4-6 is jarringly brief- perhaps mention that the model is described in more detail further on. This section also needs clearer explanation of exactly what the weather generator simulations were here (i.e., there was only one sampled time series) for those who are not already familiar with it.

→ We explained more clearly about the weather generations (P8L6-L15). It is emphasized that **all the stochastic weather series are of different internal variability,** and the 539 climatic perturbations were more clearly explained. Yet, we did not include more descriptions of the WG, because they are available in the given references. The core components of the WG (i.e., Wavelet AR, Markov chain, and Quantile Mapping) are summarized in P7L26-P8L5. I think too much detail on this tool would rather distract potential readers to catch the focus of this work.

Page 7, line 29-Page 8, line 6: Please say from the beginning how long each simulated weather series is. Based on Page 9, it looks like each weather series is 20 years long. The phrasing on page 7- "three bidecadal properties"- is confusing. The reader may interpret that to mean that the time series is longer than a decade, and each of the properties is calculated twice a decade- or the series is even longer and the property is calculated every 20 years, or for 20 random years within a longer time series. Please state clearly that the stochastic simulation was 20 years long, and that the properties are calculated based on the entire perturbed series.
→ We clearly indicated it as "the 20-year-long precipitation and temperature series" in P8L6. Whenever necessary, we rephrased "bi-decadal" to "20 years" across the manuscript. We are now saying, "…, each weather series were represented with the mean annual precipitation (Pavg), the CV of daily precipitation (Pcv), and the mean annual temperature (Tavg) over the 20-year time horizon." (P8L14). This may be clearer.

Page 7, line 29-Page 8, line 6: The range of climate change evaluated here is VERY wide, to the point that it might be preferable to refine the change increments (e.g. 5% change instead of 10% change) to get a better idea of the response surface.
→ Seemingly, the interval appears to be wide. However, as shown in Figure 3, it was sufficiently narrow to develop the regression models. There is a trade-off in narrowing the interval. If we chose a smaller interval, the computational costs would become more burdensome for the stress test. We think it was efficient to build the logistic models.

Page 8, line 29: Please change "539 stochastic weather sets" to "539 climate-altered versions of the stochastic weather simulation" if that accurately describes the set.
→We corrected the sentence as suggested (P9L5).

Page 9, line 30: "20-year" is much clearer than "bidecadal" (Page 7)
→ As relied, we corrected "bi-decadal" throughout the manuscript.

Page 14, lines 1-19: This strengthens the paper, but belongs in the "Results" section with more detail about the analysis. For example, were there really 300 stochastic simulations, or 100 stochastic simulations that were perturbed to match each of the three climate perturbation examples? What was the range of deviation between the logistic model's prediction and risk evaluated through the 100 simulations? Did the logistic model's skill vary substantially among

the three climate perturbation examples? This analysis is what convinces the reader that the paper's main technical contribution could be worthwhile, so I suggest the authors elaborate on the results of this analysis as suggested above.
I also suggest that the authors replace this space with a discussion of the analysis. What are the implications of the logistic model's predictive skill relative to that evaluated through the stochastic simulations for water management? What would a water manager think about this- when would a computationally efficient logistic model be worthwhile despite the less rigorous exploration of internal climate variability? What would the authors say to a water manager about applying the results of this analysis to the water manager's work? How would the authors advise other practitioners applying this logistic model- would they suggest a similar validation of the logistic results against an in-depth exploration of a few scenarios?
→We moved the validation to the result section (P13L15-L31). Although the validation was from only three chosen stresses, they are out of the samples used in the logistic models. So, it could be regarded as an "external validation". The median and the highest differences between the two estimates were 0.004 and 0.15, respectively. However, due to expensive computational costs, it was difficult to increase the outside samples. The 300 Monte Carlo simulations are already burdensome when comparing to the 539 samples used for the logistic models. Hence, we added an internal validation (the bootstrap prognoses), providing an indication that the logistic models were of acceptable predictive performance. The validations were discussed in the section 5.1. As advised, we addressed advantages of the logistic models in practice too.

Figure 9: There were 100 simulations at three locations- shouldn't there be 300 points on this plot? It does not look like there are. Please add more information to the caption- the readers should understand what they are being shown here. The caption should include the three climate perturbations the points come from and list the number of points.
→ Since we compared the probabilities of success at the sub-basins and the instream flow locations, the total number of points should be 3 perturbations * (14 sub-basins + 7 instream flow locations) = 63. We added this in the caption of the Figure 8 as advised.