Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2018-221-AC1, 2018


Hydrology and Earth System Sciences
Discussions

# *Interactive comment on* "Assessing water supply capacity in a complex river basin under climate change using the logistic eco-engineering decision scaling framework" *by* Daeha Kim et al.

**Daeha Kim et al.**

sjchoi@kict.re.kr

Received and published: 29 August 2018

We greatly appreciate valuable efforts of the referee 2. All the comments are sound and constructive, and we believe that they will improve our texts and assessments in the revision process. Specific responses are following as per comment.

General Comments: This manuscript describes a method of extending a bottom-up climate risk assessment by using logistic regression to estimate the probability that a water system will meet minimum performance criteria over a planning horizon based on the values of climate variables. The method is demonstrated through a case study of water management in the Geum River Basin in South Korea. The Geum River is host to



two dams which are managed for water supply, flood control, and environmental flows. The case study analyzes two alternative operating policies' ability to meet both water supply goals and instream flow requirements under a broad range of potential changes in average temperature, average precipitation, and precipitation variability. It is interesting to see the framework applied for multiple sub-basins within a larger system, and important to acknowledge uncertainty that an operating policy will meet a performance goal within specific climate scenarios. The text is poorly written and organized, with many strangely used words that inhibit understanding. Key examples include "successive", "sub-component", and "risk of system failure," which are applied in ways that are not standard in the literature and never clearly defined. Many crucial details related to the methods and motivation do not become clear until carefully examining the results section. For example, I believed the logistic model was simply modeling the water supply/environmental flow reliability as a function of climate variables rather than the risk of falling short of the reliability threshold until carefully examining the figures and results. This was the main point of the manuscript, so it is critically important that it is immediately apparent upon reading the abstract and within every part of the manuscript. The text requires substantial rewording and re-organization to clearly summarize the methodological contribution and motivation earlier in the text, better define scientific notation, and ensure new words and concepts are defined clearly the first time they are introduced. While the goal of the logistic model is a worthy one, it is not clear that the framework has been well executed in the case study or that the novel technical contribution bears sufficient relationship to the EEDS framework to be named for it. This lack of clarity may be a symptom of the confused text. However, based on my understanding of the case study, the methods used to execute the case study are flawed in several important ways. Further, the interpretation of results relies on questionable assumptions related to the fitness of GCM projections for water system risk assessment. Both the manuscript and analysis require major revisions.

–> The insufficient readability commented by the referee may be because we inadequately addressed the scientific meaning of the proposed combined framework in

the introduction and methodology sections. We will more clearly address the EEDS framework and potential challenges in stochastic uncertainty analyses. We will globally review the terminology used in our manuscript, and will improve its readability. Still, the contribution of our work is to combine the logistic regression with the bottom-up approach (not limited to the EEDS) so that users can efficiently quantify the risk of system failure. The value of the logistic regression in this work is to enable analyzers to efficiently gauge the probability of system failure without a large number of realizations for evaluating performance of a complex hydrologic system. We agree that internal climatic variability over a time horizon can significantly contribute to performance of hydrologic systems as shown in Whateley and Brown (2016). However, if we quantified the uncertainty from climatic variability using many (or long) stochastic generations, computational costs would not be small. When 100 random samplings are applied for a single climatic perturbation, computational costs increase by 100 times (in our case, it will become $343 \times 100$ tests). Although it is a simplified approach, the logistic regression makes it possible to gauge the risk of system failure in a collective manner using a single weather series for each climatic perturbation. This is a main contribution of our work. The bottom-right panel of Figure 4 shows the advantage of the logistic regression clearly. With -25% or larger changes in Pavg approximately, the river system is unlikely to satisfy the performance threshold (i.e., 100% of system failure). Between -25% and 10% changes in Pavg, the system can be either satisfactory or unsatisfactory (i.e., the risks of failure are between 100% and 0%). With +10% or greater changes in Pavg, the system seems to perfectly satisfy the threshold (0% of system failures). The number of failure (zero) decreases with increasing Pavg, while the number of success (one) increases. Importantly, variations of the stochastic weathers are different from one perturbation to another because of randomness given by the weather generator. In other words, even under different climatic variabilities across the 343 perturbations, it can be indicated that the risk of system failure declines with increasing Pavg. This can be modeled by the theoretical logistic equation that draws a smooth line between 100% failure (zero) and 100% success (one). We are not arguing this simplification



is perfect, but it can be efficient when quantifying the probability of system failures (or success). If one uses two explanatory variables (e.g., Pavg and Tavg), he or she can obtain the logistic surface of the probability of success such as the top-right panel of Figure 4 without additional stress tests. This approach is theoretically similar to the simple linear regression that approximates the variance of predicted values using the residuals from diverse data points. A single data point has only one residual and thus the variance of residuals cannot be obtained from it. However, a collection of residuals from many data points enables to quantify the variance of the residuals. Likewise, one outcome (i.e. 0 or 1) under a single climate perturbation does not allow us to quantify the risk of failure, but a collection of outcomes under various climate conditions enables it. We agree that many or long weather generations for one perturbation may quantify the risk of failure too, but it is not the only approach for assessing associated uncertainty. In revision, we can explain more clearly this contribution of our work. We will address some disadvantages and challenges in a typical stochastic uncertainty analysis and the bottom-up framework in the introduction. Then, we will describe the EEDS more clearly and how to incorporate the logistic regression into the framework in the methodology section. Thereafter, we will show the case study for Geum River Basin. This reorganization may improve readability of the manuscript. And, discussion and conclusions will be revised accordingly to highlight the scientific contribution of this work.

Specific Comments: Logistic regression model: (1) Limited calibration set: It is my understanding that the logistic model was calibrated from 434 binary values that correspond to either water supply reliability or environmental flow reliability meeting a threshold under 434 unique combinations of three climate variables. If my understanding is correct, this would mean that there is one response (binary performance metric) per climate scenario (this should be clarified in the manuscript if that is incorrect). This is a very limited data set for analyzing risk of failure resulting from internal climate variability, especially given that each scenario-specific stochastic trace was (a) only 20 years long, and (b) initially identical to every other weather sequence in the analysis

that had then been perturbed from the original trace to match a unique combination of average precipitation, average temperature, and precipitation coefficient of variation using quantile mapping. To characterize the effects of internal variability on risk of failure over a planning period, it would be preferable to use the binary reliability outcomes from many more stochastic realizations of weather sequences within each combination of climate variables. With a single stochastic trace perturbed into many climate scenarios, the modelled risk of failure is likely to be driven entirely by the climate scenario rather than the actual risk of missing a performance target under internal climate variability, and furthermore heavily biased across the climate response function by the single stochastic realization used to generate all climate scenarios. This seems to be the opposite of the intentions described in the introduction.

–> We agree that the internal variability significantly affect variation of the performance metrics. However, we disagree that multiple weather generations for each perturbation are required for quantifying the risk of failure. The theory of linear regressions unnecessarily requires many Y values for a single X value. Rather, many pairs of (X, Y) are needed to quantify the variance of residuals and the prediction intervals. Since the weather generator bootstraps the observed weathers for each perturbation, the 343 sets of the 20-year-long weather series contain different internal variability each other. In other words, the risk of failure can be obtained from diverse variability of the 343 weather series in a collective manner, not from a fixed one. It is true that long or many weather generations can quantify the internal variability of a single climatic perturbation rigorously. However, it is not the only approach for quantifying the risk of failure. Indeed, it can be time-consuming. We rather argue that there is no clear evidence to confirm that the risk estimates from the logistic regression were biased. We did not fix the variation of weather series for the 343 perturbations. The length of time horizon in this work (20 years) was determined following the definition of climate change given in the IPCC 4th assessment report. The IPCC defined it as the statistical changes during a decadal or a longer period, and thus we set the 20 years for generating weather series for each climatic perturbation. This length is not for capturing climatic variability



by random weather generations. Though one weather generation has a length of 20 years, the uncertainty from internal variability may be captured by collecting the 343 sets all together. In this work, the risk of system failure did not come from a single perturbation, but from integration of the 343 stress tests that contain different internal variabilities.

(2) I do not see any part of the manuscript that assesses the performance of the logistic regression model using out-of-sample data. This is critical to the manuscript's success because it would provide evidence that the loss of information from modelling the risk of failing a satisficing criterion rather than evaluating the risk of failure through many simulations at each combination of climate variables could be worth the savings in computation time.

–> We agree. It seems necessary to validate the risk estimate from the logistic regression. A possible approach is to compare one risk estimates at a selected perturbation from the logistic regression to the risk estimate from a number of random generations (e.g., 100 times) for the selected perturbation. This will add this validation in revision.

(3) It is not clear whether there are separate logistic regression models for each sub-basin, performance metric, etc. How many logistic regression models are there in this case study? One per sub-basin, to model simultaneously meeting water supply reliability and environmental flow requirements? Two per sub-basin, each modelling risk of failing one of the objectives' minimum performance criterion? One, with sub-basins represented through dummy variables? If the model is used to predict risk of failing mutual satisficing rather than risk of failing one performance threshold, would the model structure work if the two objectives were in tension (as in the Poff et al. 2015 case study) rather than aligned (as they are in this case study)? This section needs to clearly list the explanatory variables and document the dependent variables much more clearly.

–> Figure 5 shows the climatic bounds for 95% probability of success for each demand

node. So, the number of lines is same as the number of demand nodes. Figure 7 shows the climatic bounds for 95% of success for each instream flows location (seven in total). The logistic regression applied for each node and each instreamflow location. And, the highest bound for water supply and that for instreamflow were combined. In other words, if the most vulnerable demand node and the most vulnerable instreamflow location can have 95% probability of success under a certain climate stress, the other nodes and locations will have 95% or more probabilities automatically. The mutual zone made by the two highest bounds is the key information. We will explain more clearly in revision.

Water system modelling framework: (1) Synthetic weather generator and streamflow temporal resolution: A daily weather generator is used to generate perturbed weather sequences and run them through a runoff model to generate streamflow. After simulating climate-changed streamflow using the runoff model, daily streamflows are aggregated to monthly flow. Why aggregate ex post rather than using a computationally cheaper weather generator and/or runoff model that is designed to operate at the monthly temporal resolution?

–> This is due to validity of the hydrologic model. We needed a method for ungauged basins for each sub-basin, and already had a validated model. GR4J was validated by the LOOCV across South Korea by Kim et al. (2017). Though it is true that a monthly model is computationally efficient than daily models, another validation for ungauged basins will be required. Aggregating daily simulations was not very time-consuming, but the main computational cost in this work was the time required for 20-year-long sequential optimizations.

(2) Temporal aggregation and precipitation coefficient of variation (cv): Perhaps the monthly streamflow resolution is the reason precipitation coefficient of variation was not a strong predictor of performance metrics? The authors should consider this possibility and potentially discard precipitation cv from their analysis, which might be better served by more stochastic realizations in each climate scenario rather than more cli-



mate variables.

–> It is unlikely. Even with the temporal aggregation from daily to monthly values, the monthly flows were affected by Pcv. A higher Pcv resulted in larger streamflow, because precipitated water would reside in the soil for a shorter length due to more frequent high-intensity rainfall events, leading to less evapotranspiration. We found that Pcv was one of significant factors that explains the variation of total streamflow. However, it was not significant to explain the variation of the water supply reliability given by the 343 perturbations. We believe that the storage capacities of the sub-basin and the dams are likely factors that nullified the influence of Pcv on water availability. We will explain this more clearly in revision.

(3) Climate response surface: The sampling of average precipitation and precipitation coefficient of variation (cv) is coarse (20% increments). I suggest sampling these factors at tighter increments. (4) The computational expense of conducting bottom-up climate risk assessment is mentioned several times in the text. How computationally intense is the Geum water system model to evaluate?

–> Even with the interval we applied, changes in water supply performance seems to be sufficiently captured as shown in Figure 3a. However, in revision, the range and interval may be adjusted to zoom in the range between 500 and 1500 mm of Pavg (e.g., -60% to +60% at a 10% interval), because 1500+ mm of Pavg in Figure 3a mostly resulted in the maximum reliability (i.e., 1). We guess one week will be adequate to update the stress tests with the two scenarios, because all the models were readily available now. We believe that this would be possible during the revision process, and will improve this work.

Role of GCM projections in the case study: (1) GCMs are limited in their ability to simulate land/ocean/atmospheric mechanisms, especially those that take place at sub-grid scale resolution. This limits the information that can be credibly derived from projections for water resources planning. Precipitation coefficient of variation (CV), one of

the climate variables used in the case study, is not well represented in GCMs so it is questionable to infer precipitation CV from GCM projections. This is why GCM projections are not shown on some of the response surfaces in Poff et al. 2015 (in response to page 3, Line 18-19 of this manuscript).

–> We agree that the GCMs have limitations, and it is true that all the projections can be subject to significant uncertainty. It is not limited to Pcv. Pavg and Tavg may not be well captured by GCMs either. However, because of that reason, the bottom-up frameworks emerged by employing the stochastic tests imposing arbitrary climatic stresses on the hydrologic systems. And, overlaying GCMs projections on the response surfaces is a common approach to gauge climate change risks in most bottom-up assessments. Should we neglect Pcv values from GCMs because of the limitations in GCMs, even though the bottom-up framework was intended to consider the uncertainty in GCMs for practical decision makings (e.g., Brown et al., 2012)? We disagree that Pcv should not be overlaid on the response surfaces due to that uncertainty. Rather, we believe that Pcv values need to be overlaid as many as possible to combine the knowledge from the stress tests (i.e., response surface) and the climate sciences (i.e. GCMs). Without any reference points, the response surfaces can provide information of system sensitivity to climate stresses only. What if future climatic stresses are out of the range in which we can withstand? Poff et al. (2016) is a very innovative approach that allows quantifying multi-faceted system robustness to climate change. However, if any predictions are not combined with it, its usability may be limited in practice.

(2) This manuscript repeatedly mentions GCM counts as though GCM count in the feasible region on the climate response surface could be a decision criterion (e.g. page 3 line 19), and perhaps to some stakeholders it would be. However, this could also imply an attempt to quantify risk across the entire sampled climate space. Uncertainty quantification via ensembles of GCM projections is a challenging research question in its own right and would not be well treated by simply counting GCM projections from an arbitrary ensemble. Indeed, the point of bottom-up decision frameworks for climate risk



management is avoiding this type of reliance on GCM projections with little scientific basis. Since this manuscript is designed to build on a bottom-up risk assessment framework, it is strange that so much emphasis is put on understanding performance under GCM projections in the text and figures. Titling the framework: As mentioned above, it is not clear whether the logistic model is designed to model the risk of failing to mutually satisfice the eco-engineering performance thresholds or the risk of failing to meet one performance threshold. If the latter, the main technical contribution seems as appropriate for any single-objective climate response surface type risk assessment as for multi-objective climate response surface analyses, though it is applied here in a multi-objective climate response surface analysis. I would suggest the authors re-frame the analysis and revise the title to put the focus on the manuscript's main technical contribution, which is analyzing and communicating probabilistic information through a climate response surface (with an eco-engineering case study) rather than presenting a novel decision framework.

–> Perhaps, we put too much emphasis on the GCM counts in the text, though it seems to be intended by the original decision scaling framework (Brown et al., 2012). We will tone down in revision. Our point is that while the response surfaces of system performance is developed to consider risks (or uncertainty) associated with climate change, there is no quantified risk estimates in there interestingly. How do we get lessons from a response surface and climate projections? One of implications is where the locations of climate projections are on the response surface. It is natural for potential users to check if the projections are beyond the performance threshold or not. Thus, it was strange to us that there were no projections on the response surfaces in Poff et al. (2016). So, we guessed some GCMs projections might be located out of acceptable ranges in Poff et al. (2016). By its nature, a multi-purpose response surface should have a narrower acceptable range than a single-purpose one. For better indications from the multi-purpose response surfaces, we just suggest to convert them to the logistic surfaces directly indicating the risk of system failure and then overlay climate projections. In revision, we will better frame the manuscript and retitle it to improve

readability, as responded earlier.

Technical corrections (typing errors, etc.) Word choice: The meaning of the terms “successive”, “risk of system failure”, and “sub-components” in the context of this analysis is not clear from the text.

–> We will provide clear definitions for them.

Page 2, line 31: Whatley et al. 2014 should be Whateley et al. 2014

–> We will globally check mistypos.

Page 3, section 5: “However, all assessments using the response surfaces have focused on the “expected performance” rather than risk of system failure” Is this true? I thought many decision scaling papers evaluated reliability, which is risk of failure.

–> We will tone down. However, to our knowledge, many studies have usually developed the response surfaces in terms of the expected performance rather than the risk of failures even in the case that assessing uncertainty was a main objective (e.g., Kay et al., 2014).

Figure 9: Labels on X axis would be clearer in words. Also, isolating the results of the analysis to GCM projections is totally counter-intuitive here. The point of bottom-up climate response surface analyses is to avoid relying on GCMs in climate risk management.

–> We will describe the label more clearly. However, we disagree that the risk estimates from the GCMs are counter-intuitive. Then, how can we assess future climatic risks in practice? The response surface itself does not have predictions. If GCMs are discarded, only information from the response surfaces is system robustness. The probability of successive outcomes estimated from GCMs can be important information for decision-makers. Figure 9 shows the trade-off in future risks of system failures when changing the human-demand-only management policy.



Figure 2: It is not clear where and how the logistic model comes into this framework based on Figure 2. Figures: None of the response surface figures include precipitation CV as one of the axes, though this is one of the sampled climate variables. The reasoning behind this should be clarified in the text.

–> We will improve relevance of this figure to bring a better implication.

References

Brown, C., Ghile, Y., Laverty, M., and Li, K.: Decision scaling: linking bottom-up vulnerability analysis with climate projections in the water sector. Water Resour. Res., W09537, https://doi.org/10.1029/2011WR011212, 2012.

Kay, A. L., Crooks, S. M., and Reynard, N. S.: Using response surfaces to estimate impacts of climate change on flood peaks: assessment of uncertainty, Hydrol. Process., 28, 5273–5287, https://doi.org/ 10.1002/hyp.10000, 2014.

Poff, N. L., Brown, C. M., Grantham, T. E., Matthews, J. H., Palmer, M. A., Spence, C. M., Wilby, R. L., Haasnoot, M., Mendoza, G. F., Dominique, K. C., and Baeza, A.: Sustainable water management under future uncertainty with ecoengineering decision making, Nature Clim. Change, 6, 25-34. https://doi.org/10.1038/nclimate2765, 2016.

Whateley, S., and Brown C.: Assessing the relative effects of emissions, climate means, and variability on large water supply systems, Geophys. Res. Lett., 43, 11,329–11,338, http://doi.org/ 10.1002/2016GL070241, 2016.

---

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-221, 2018.