Impact of Scenario Selection on Robustness

McPhail, C.; Maier, H. R.; Westra, S.; Kwakkel, J. H.; van der Linden, L.

DOI
10.1029/2019WR026515

Publication date
2020

Document Version
Final published version

Published in
Water Resources Research

Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR026515

Impact of Scenario Selection on Robustness

C. McPhail[1], H. R. Maier[1], S. Westra[1], J. H. Kwakkel[2], and L. van der Linden[3]

[1]School of Civil, Environmental, and Mining Engineering, University of Adelaide, Adelaide, South Australia, Australia, [2]Faculty of Technology Policy and Management, Delft University of Technology, Delft, The Netherlands, [3]South Australia Water Corporation, Adelaide, South Australia, Australia

Key Points:
- We present the first quantification of the effects of scenario selection on robustness
- Through a case study we show that scenario selection has a significant effect on the robustness of solutions but a small influence on their rankings
- This effect is due to complex interactions of scenario space coverage, behavior of the system performance metric, and characteristics of the robustness metric

Supporting Information:
- Supporting Information S1

Correspondence to:
C. McPhail,
cameron.mcphail@adelaide.edu.au

Citation:
McPhail, C., Maier, H. R., Westra, S., Kwakkel, J. H., & van der Linden L. (2020). Impact of scenario selection on robustness. *Water Resources Research*, *56*, e2019WR026515. https://doi.org/10.1029/2019WR026515

Received 8 OCT 2019
Accepted 8 AUG 2020
Accepted article online 14 AUG 2020

Abstract Multiple plausible future scenarios are being used increasingly in preference to a single deterministic or probabilistic prediction of the future in the long-term planning of water resources systems. These scenarios enable the determination of the robustness of a system—the consideration of performance across a range of plausible futures—and allow an assessment of which possible future system configurations result in a greater level of robustness. There are many approaches to selecting scenarios, and previous studies have observed that the choice of scenarios might affect the estimated robustness of the system. However, these observations have been anecdotal and qualitative. This paper develops a systematic, quantitative methodology for exploring the influence of scenario selection on the robustness and the ranking of decision alternatives. The methodology is illustrated on the Lake Problem. The quantitative results obtained confirm the qualitative observations of previous works, showing that the selection of scenarios is important, as it has a large influence on the robustness value calculated for each decision alternative. However, we show that it has a relatively small influence on how those decision alternatives are ranked. This implies that despite the difference in robustness values, similar decision outcomes will be reached in this case study, regardless of the basis on which the scenarios are obtained. It is also revealed that the impact of the scenarios on the robustness values is due to complex interactions with the system model and robustness metrics.

1. Introduction

Traditionally, model-based assessments of different water resources decision alternatives (i.e., plans and policies) have been based on a single "expected" future (Giuliani et al., 2016; Hall & Harvey, 2009; Kwakkel & van der Pas, 2011; Morgan et al., 1990). However, this does not consider the significant uncertainties associated with drivers of change such as climate, technology, economy, and society (Döll & Romero-Lankao, 2016; Maier et al., 2016; Shepherd et al., 2018), potentially resulting in a range of negative consequences when conditions occur that are different from those expected future conditions (Lempert & Trujillo, 2018; McInerney et al., 2012; Raso et al., 2019).

In response to the recognition that many future changes are "deeply uncertain" (Kwakkel et al., 2010; Lempert, 2003), the relative merits of potential decision alternatives are now commonly assessed under a range of plausible future conditions (scenarios) (Herman et al., 2014; Kwakkel et al., 2010; Lempert, 2003; Little et al., 2018; Maier et al., 2016; Varum & Melo, 2010; Walker, Lempert, et al., 2013). As part of model-based assessment, such scenarios correspond to different points in the hyperspace of plausible ranges of model inputs. However, how these points are distributed in this hyperspace for different scenarios can be highly variable, depending on scenario type and number.

Scenarios are generally classified into three different types: predictive ("what is likely to happen"), explorative ("what could happen"?), or normative ("how can a specific future be realized"?) (Maier et al., 2016). A number of water resources studies have generated explorative scenarios by considering the impact of plausible changes in atmospheric carbon concentrations (Anghileri et al., 2018; Beh et al., 2014, 2015a, 2015b; Giuliani & Castelletti, 2016; Giuliani et al., 2016; Haasnoot et al., 2012, 2013; Herman & Giuliani, 2018; Huskova et al., 2016; McPhail et al., 2018), as well as plausible changes in regional socioeconomic conditions (Haasnoot et al., 2013; Wada et al., 2019). In contrast, normative scenarios consider conditions that represent interesting outcomes, as is the case with scenario discovery (e.g., Bryant & Lempert, 2010; Groves & Lempert, 2007; Hadka et al., 2015; Kasprzyk et al., 2013; Kwakkel, 2017; Kwakkel, Walker, et al., 2016; Matrosov et al., 2013; Trindade et al., 2017); conditions that result in one decision alternative being

preferable to another, as is the case with MORE (Ravalico et al., 2010), POMORE (Ravalico et al., 2009) and decision scaling (e.g., Brown et al., 2012); or conditions under which certain decision alternatives no longer perform adequately, as is the case with adaptive tipping point approaches (e.g., Haasnoot et al., 2013; Kwadijk et al., 2010; Kwakkel et al., 2015; Kwakkel, Haasnoot, et al., 2016; Vervoort et al., 2014; Walker, Haasnoot, et al., 2013).

How many scenarios are generated is generally linked to the philosophy that underpins scenario generation. When scenarios correspond to coherent descriptions of alternative hypothetical futures (e.g., van Notten et al., 2005), the number of scenarios considered is generally small (~3–9, see Table S1 in the supporting information), and scenarios are generally identified using some type of human input, such as the use of participatory approaches involving a variety of stakeholders (e.g., Wada et al., 2019). In contrast, when scenarios are designed to represent a broad range of combined changes in future conditions, the number of scenarios considered is generally large (~100–15,000, see Table S1 in the supporting information), and scenarios are generated using numerical modeling and/or sampling- or optimization-based approaches, with minimal stakeholder input (e.g., Culley et al., 2016, 2019; Hadka et al., 2015; Hall et al., 2012; Herman et al., 2014, 2015; Kasprzyk et al., 2013; Kwakkel et al., 2015; Kwakkel, 2017; Kwakkel, Walker, et al., 2016; McPhail et al., 2018; Quinn et al., 2017, 2018; Singh et al., 2015; Trindade et al., 2017; Watson & Kasprzyk, 2017; Weaver et al., 2013; Zeff et al., 2014).

In order to enable the performance of different decision alternatives to be compared across scenarios, robustness metrics are commonly used (Maier et al., 2016; McPhail et al., 2018; Walker, Lempert, et al., 2013). Different robustness metrics combine values of performance metrics obtained for individual scenarios, such as cost, reliability (frequency of failure), vulnerability (magnitude of failure), and resilience (time to recover from failure) (Burn et al., 1991; Hashimoto et al., 1982; Maier et al., 2001; Zongxue et al., 1998) in different ways, depending on decision-maker preferences and decision context (McPhail et al., 2018). Previous studies have shown that the relative robustness of different decision alternatives can vary depending on which robustness metric is used (Borgomeo et al., 2018; Drouet et al., 2015; Giuliani & Castelletti, 2016; Hall et al., 2012; Herman et al., 2015; Kwakkel, Eker, et al., 2016; Lempert & Collins, 2007; Roach et al., 2016), highlighting the importance of choosing robustness metrics that are appropriate for the decision context considered (McPhail et al., 2018). However, robustness values are also a function of which scenarios are considered.

Given the diversity of scenario types and generation methods adopted in the water resources literature, as discussed above, there is a need to assess the impact of the choice of scenarios on robustness values, and the resulting ranking of decision alternatives, in addition to the impact of the choice of the robustness metric itself, as has been done in previous studies (Borgomeo et al., 2018; Drouet et al., 2015; Giuliani & Castelletti, 2016; Hall et al., 2012; Herman et al., 2015; Kwakkel, Eker, et al., 2016; Lempert & Collins, 2007; McPhail et al., 2018; Roach et al., 2016). While the potential impact of the choice of plausible futures via different approaches to creating scenarios has been recognized in qualitative or anecdotal terms (Kwakkel et al., 2012; Phadnis, 2019), there is a lack of a systematic methodology for assessing this in a quantitative fashion. Kwakkel et al. (2012) describe an experiment in airport strategic planning where they show that if the set of scenarios represents a narrow range of future airport demands rather than a wide range, then a static plan will outperform an adaptive plan. However, if the set of scenarios represents a wider range of future airport demands, then the adaptive plan outperforms the static plan. Phadnis (2019) compares four different decision-making approaches for competitive businesses and shows that no single decision-making approach outperforms all others under all sets of future conditions. Specifically, it is shown that different decision-making approaches are superior depending on whether a narrow or wide set of future conditions is considered. However, as was the case in Kwakkel et al. (2012), this analysis was case specific.

As discussed above, there is a lack of a generalized, quantitative method for assessing the impact of different sets of scenarios on the absolute and relative (i.e., ranking) robustness values of different decision alternatives under conditions of deep uncertainty, especially in the water resources domain. In order to address this shortcoming, the objectives of this paper are as follows:

1. To develop a methodology to quantitatively analyze how different sets of scenarios can influence both (a) robustness and (b) the ranking of decision alternatives based on robustness values (i.e., the relative robustness of different decision alternatives) and

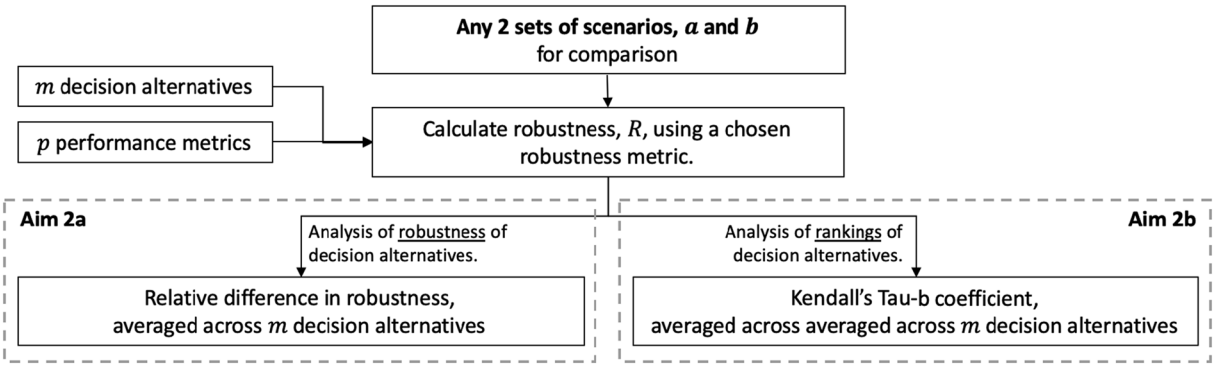

Figure 1. Approach for the quantitative analysis of the influence of any two sets of scenarios on the robustness and ranking of decision alternatives.

2. To illustrate the methodology in (1) on the Lake Model, which is a stylized, hypothetical water resources case study that is well represented in the literature (Carpenter et al., 1999; Eker & Kwakkel, 2018; Hadka et al., 2015; Kwakkel, 2017; Lempert & Collins, 2007; Quinn et al., 2017; Singh et al., 2015; Ward et al., 2015). As part of the case study analysis, a number of issues are explored, including the influence of (a) the number and distribution of scenarios, (b) the behavior of the robustness metric, and (c) the behavior of the system performance metric on absolute and relative robustness.

The remainder of the paper is organized as follows: Section 2 introduces the methodology for quantifying and visualizing the effect of the selection of different sets of scenarios on robustness and the ranking of decision alternatives; section 3 describes how this methodology was applied to the Lake Model; and section 4 shows the results of this analysis, along with a discussion of the effects of different sets of scenarios on robustness and on the rankings of decision alternatives. This is followed by a summary and conclusions in section 5.

2. Generic Approach for Assessing the Influence of Scenario Selection on Robustness

To quantify the impact of scenario selection/creation on robustness (Aim 2a) and on the rankings of decision alternatives (Aim 2b), we propose the approach presented in Figure 1. The approach compares outcomes from applying two distinct sets of scenarios and thus provides insight into the sensitivity of those outcomes on the method of scenario selection. Thus, the proposed approach is generic, as it can cater to and is independent of the approach used to create the sets of scenarios—including aspects such as the number of scenarios considered, the distribution of scenarios over the scenario space, and the method used to generate the scenarios (e.g., sampling or using stakeholder input) (see section 1).

The two sets of scenarios to be compared are denoted by a and b, which comprise some number of distinct scenarios (possibly a different number of scenarios in each set). These scenarios form inputs to a system model, which is run for all m decision alternatives, with the model outputting values of the p possible measures of system performance. Considering each of the p performance metrics one at a time, and for a single robustness metric, the robustness value R is calculated for each of the decision alternatives for each of the two scenario sets via some form of aggregation of the system performance values (see McPhail et al., 2018). These calculations can be repeated for each of the p performance metrics and any number of other robustness metrics to enable exploration of the effect of metric choice on the study objectives. The final part of the approach is the quantification and visualization of the influence of the selected scenarios on the robustness and the rankings of the decision alternatives.

The methodology used for assessing the impact of two sets of scenarios on the robustness values is shown in Figure 2. For a single decision alternative and single robustness metric, the two different sets of scenarios produce one robustness value each, and these two robustness values are compared. This difference is then averaged across all m decision alternatives.

The methodology used for assessing the impact of two different sets of scenarios on ranking similarity (i.e., relative robustness) is shown in Figure 3. We begin with the robustness of all m decision alternatives

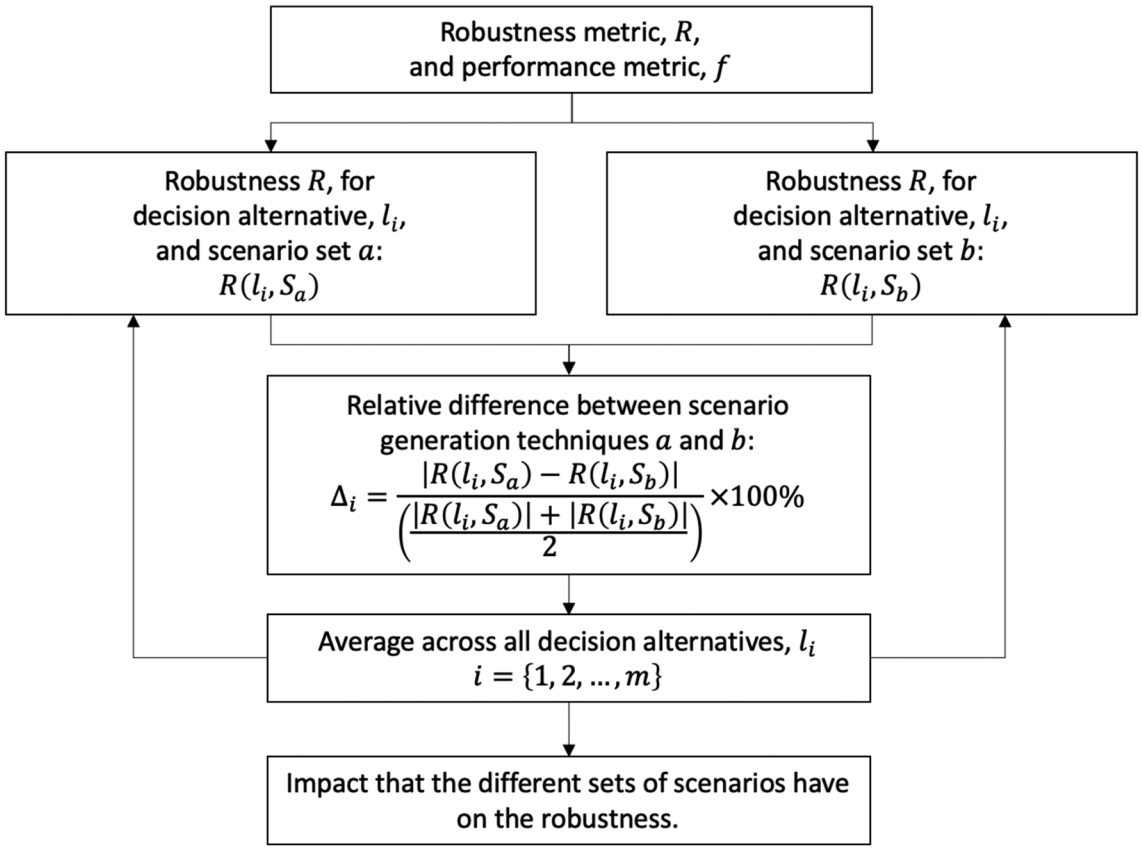

Figure 2. Calculation of the sensitivity of robustness to different sets of scenarios.

when using one set of scenarios and compare this to the robustness of the same m decision alternatives evaluated with a different set of scenarios. These two sets of robustness values are compared using Kendall's rank correlation. This statistical metric tests how similarly two quantities are ranked. In this case, we are testing how the decision alternatives are ranked when robustness is calculated twice, each time with a different set of scenarios.

In other words, there are two sets of robustness values, R, for each of the m decision alternatives: $\{R(l_1, S_a), R(l_2, S_a), \ldots, R(l_m, S_a)\}$ for scenario set a, S_a, and $\{R(l_1, S_b), R(l_2, S_b), \ldots, R(l_m, S_b)\}$ for scenario set b, S_b. If two decision alternatives, l_i and l_j, are ranked the same way regardless of whether robustness is calculated using S_a or S_b, then the ranking is considered "similar" or "concordant." More explicitly, concordance is defined as one of the following two conditions being true:

$$R(l_i, S_a) > R(l_j, S_a) \text{ and } (l_i, S_b) > R(l_j, S_b), \tag{1}$$

$$\text{or } R(l_i, S_a) < R(l_j, S_a) \text{ and } R(l_i, S_b) < R(l_j, S_b) \tag{2}$$

If the two scenario sets lead to a different ranking of decision alternatives, then the rankings of the decision alternatives are considered "dissimilar" or "discordant." Discordance occurs under either of the following two conditions:

$$R(l_i, S_a) > R(l_j, S_a) \text{ and } (l_i, S_b) < R(l_j, S_b), \tag{3}$$

$$\text{or } R(l_i, S_a) < R(l_j, S_a) \text{ and } R(l_i, S_b) > R(l_j, S_b) \tag{4}$$

In the case that either set of scenarios produces a tie in ranking, then it is considered neither similar (concordant) nor dissimilar (discordant). This occurs during either of the following two conditions:

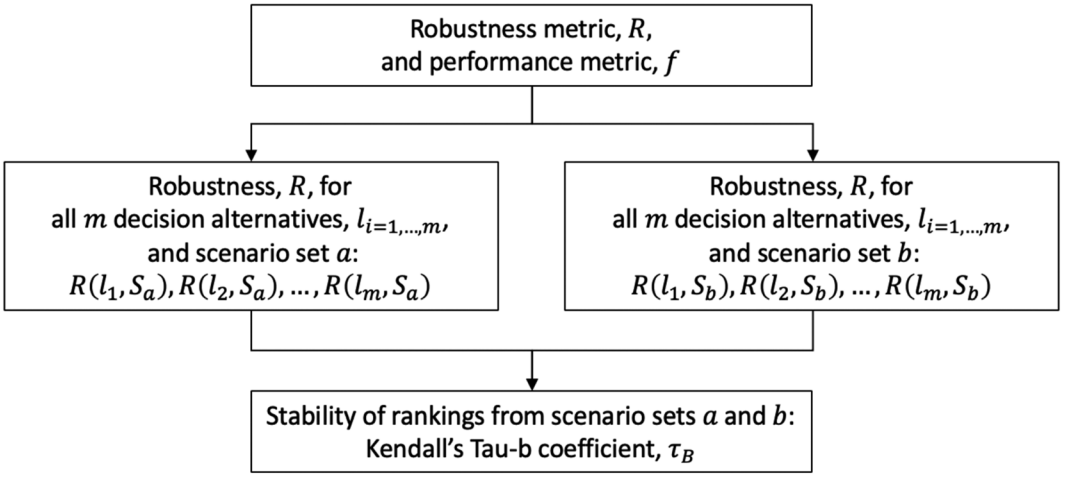

Figure 3. Methodology used to determine the similarity of the rankings of decision alternatives.

$$R(l_i, S_a) = R(l_j, S_a), \tag{5}$$

$$\text{or } R(l_i, S_b) = R(l_j, S_b) \tag{6}$$

Kendall's rank correlation compares all pairs of decision alternatives, l_i and l_j, to obtain a measure of the agreement in ranking under the two sets of scenarios. We use Kendall's Tau-b metric because it makes adjustments for ties in rankings to ensure that the values of Tau-b, τ, range between -1 (opposite rankings/complete disagreement) and $+1$ (same rankings/complete agreement). This gives a high-level view of how scenario selection impacts the rankings of the decision alternatives, providing confidence to decision makers that a particular decision alternative is more robust than another irrespective of the choice of scenario sets if there is a high degree of ranking similarity across the scenarios. Conversely, a high degree of disagreement in the ranking of the decision alternatives across the different scenario sets indicates that it is difficult to identify the most robust decision alternative and that the scenarios considered might have to be examined more carefully.

3. Case Study

3.1. Background

In order to illustrate the generic approach for assessing the impact of scenario selection on absolute and relative robustness, we use the intertemporal Lake Problem as a case study. It is a stylistic, hypothetical problem that has been used in many previous studies (Carpenter et al., 1999; Eker & Kwakkel, 2018; Hadka et al., 2015; Kwakkel, 2017; Lempert & Collins, 2007; Quinn et al., 2017; Singh et al., 2015; Ward et al., 2015). It is based on the idea of a town that releases pollution into a lake and has many of the characteristics commonly encountered by decision makers dealing with real water resources problems, such as (1) environmental thresholds, (2) deep uncertainty in future conditions, (3) deep uncertainty associated with identifying environmental thresholds, and (4) conflicting objectives (e.g., economic vs. environmental) (Lempert & Collins, 2007; Lenton, 2013; Quinn et al., 2017). The specific details of the Lake Problem are contained in the studies mentioned above, and an overview of the performance metrics, decision alternatives, and scenarios is given below.

There are environmental consequences of the release of pollution into the lake, which are measured by two of the performance metrics: maximum phosphorus concentration (to be minimized) and the frequency of time where the pollution is below a critical threshold (i.e., reliability) (to be maximized). Competing against these environmental metrics is a third performance metric, the economic utility (to be maximized), which is decreased when action is taken to reduce pollution.

The performance metric values are influenced by the decision alternatives and scenarios. The decision alternatives represent the annual pollution control strategies that the inhabitants of the town implement (i.e.,

Table 1
Deeply Uncertain Scenario Variables (Model Inputs) and Associated Ranges of Values for the Lake Problem

Variable	Range	Description
μ	0.01–0.05	Mean of the lognormal distribution of natural pollution inflows
σ	0.001–0.005	Standard deviation of the lognormal distribution of natural pollution inflows
b	0.1–0.45	Natural removal rate of pollution
q	2–4.5	Natural recycling rate of pollution
δ	0.93–0.99	Discount rate (for economic utility)

they define the annual quantity of industrial pollution that is allowed to enter the lake for each year in the 100 year planning horizon). A reduction in annual pollution improves reliability and maximum phosphorous (by increasing the number of years the system is below the pollution threshold and minimizing the maximum level of phosphorus). However, this decreases economic utility (because it costs money).

3.2. Scenario Set Generation

In principle, the Lake Problem can be represented using a range of qualitative and quantitative approaches, with important choices related to system model boundaries, process representations, and other key modeling considerations. In the particular case considered in this paper, a system model (referred to as the "Lake Model") is used, as it represents a trusted numerical representation of the system that has reasonable fidelity in simulating key system processes (Carpenter et al., 1999; Lempert & Collins, 2007). System model selection represents a key consideration in model-based assessments, and the system model boundaries effectively delineate the scenarios that are required as model inputs. These inputs are described in Table 1, with the set of valid combinations of scenarios depicted as a five-dimensional hypercube with plausible bounds selected based on previous studies (Eker & Kwakkel, 2018; Kwakkel, 2017; Quinn et al., 2017), as given in Table 1.

The objectives of the case study are to emulate the impact of the diversity of scenario selection approaches used in the water resources literature, as summarized in section 1, on absolute and relative robustness values. However, regardless of which scenario selection approach is used, for a quantitative study such as the Lake Problem, the outcome of the scenario selection step needs to be the quantitative specification of inputs to the system model (i.e., points in the five-dimensional hypercube that represents the input parameter space for the Lake Model). Thus, several sampling strategies are used to generate the requisite Lake Model inputs, which encapsulate key features of alternative scenario generation techniques, including

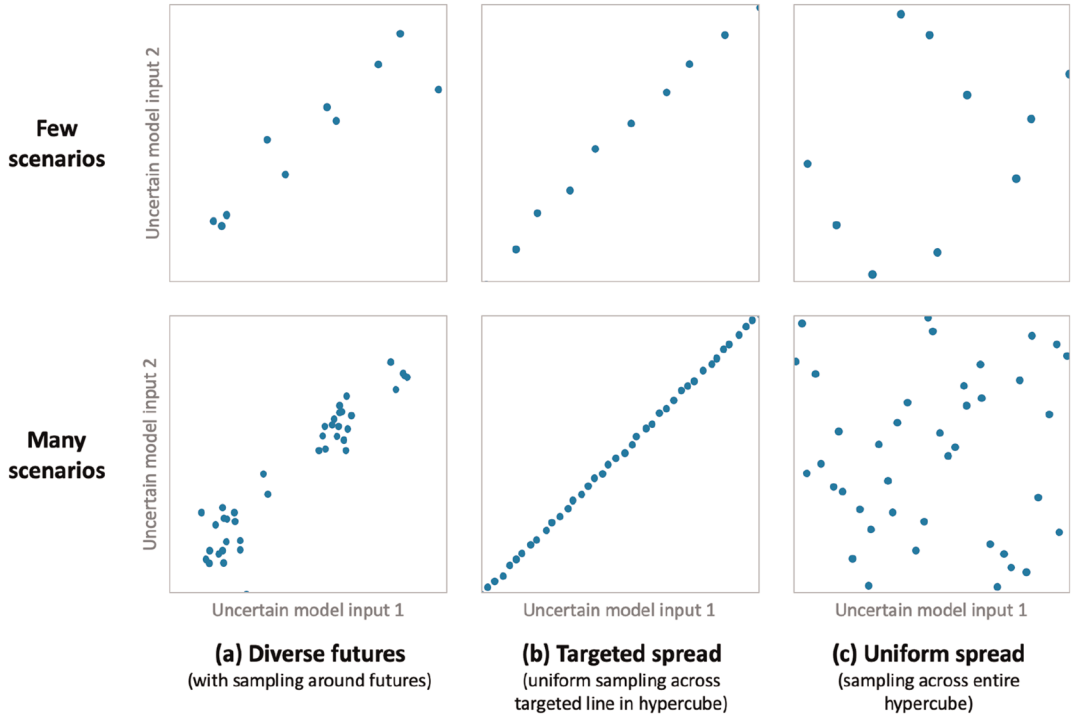

Figure 4. Two-dimensional illustration of how (a–c) the three distributions of scenarios are implemented for the case study, with examples for both a small and large number of scenarios, emulating the diversity in scenarios that could be obtained by using different scenarios selection approaches.

Table 2
Robustness Metrics Used in Analysis

Metric name	Brief description
Maximin	Worst-case performance (high level of risk aversion)
Maximax	Best-case performance (low level of risk aversion)
Hurwicz's optimism-pessimism rule	Weighted sum of the best and worst cases
Laplace's principle of insufficient reason	Mean performance
Minimax regret	The worst-case cost of making a wrong decision in any given scenario (high level of risk aversion)
90th percentile Minimax regret	The 90th percentile cost of making a wrong decision (high level of risk aversion) (percentile-based calculation)
Mean-variance	A function of the mean and variance in performance
Undesirable deviations	The sum of performance below the median performance
Percentile-based skewness	The skew of performance (toward high or low performance) (percentile-based calculation)
Percentile-based peakedness	The kurtosis (peakedness) of performance (percentile-based calculation)
Starr's domain criterion	Calculates the proportion of scenarios with acceptable levels of performance

1. how the space is covered (i.e., whether the focus is on evenly covering the space or on identifying regions of the space that are more or less likely) and
2. the number of scenarios considered.

To ensure the generality of our findings, we have analyzed 300 different potential scenario sets for each distribution of scenarios, consisting of a total of 18,000 individual scenarios in sets of size 20, 40, 60, 80, and 100 scenarios per set. These are distributed in different ways throughout the scenario space, including uniform coverage of the space, sparse coverage of diverse regions of the space, and a targeted spread over certain regions of the space (see the supporting information for details on how the different scenarios were generated).

Illustrative examples of the resulting differences in the distributions of the scenarios obtained are shown in Figure 6.

- *Diverse*: Figure 4a depicts the situation where four diverse futures are first identified (analogous to Representative Concentration Pathways [RCPs]) with many samples taken around each of these four points (analogous to the use of multiple global and regional climate models to create multiple downscaled realizations of each of the RCPs) of which there are many examples in the water resources literature (Anghileri et al., 2018; Giuliani & Castelletti, 2016; Giuliani, Castelletti, et al., 2016; Haasnoot et al., 2012, 2013; Herman & Giuliani, 2018; Huskova et al., 2016; McPhail et al., 2018).
- *Targeted*: Figure 4b depicts a targeted approach to identifying samples that cover "interesting" regions of the system model space, for the situation where the model performance responds monotonically to each input (i.e., an increase in one variable always results in increased or decreased performance). This can occur when two model inputs (e.g., water supply and water demand) are lined up from worst to best, and the two worst values (e.g., lowest water supply and highest water demand) are paired, and so forth, leading to a clear set of worst to best points in the hypercube (Beh et al., 2014, 2015a, 2015b).
- *Uniform*: Figure 4c depicts a uniform sampling of the entire hypercube to consider a wide range of plausible futures, as is often done in the water resources literature (Culley et al., 2016, 2019; Hadka et al., 2015; Hall et al., 2012; Herman et al., 2015; Kasprzyk et al., 2013; Kwakkel, 2017; Kwakkel et al., 2015; Kwakkel, Walker, et al., 2016; McPhail et al., 2018; Quinn et al., 2017, 2018; Singh et al., 2015; Trindade et al., 2017; Watson & Kasprzyk, 2017; Weaver et al., 2013; Zeff et al., 2014).

3.3. Decision Alternatives and Performance Values

Robustness values are determined relative to potential decision alternatives, and in this analysis we consider 4,611 such alternatives. These were obtained using a many-objective evolutionary algorithm to identify a set of Pareto optimal decision alternatives for a reference scenario, as is recommended in many-objective robust decision making (Kasprzyk et al., 2013). Specifically, we used a generational version of the BORG algorithm (Hadka & Reed, 2013), to allow for easy parallelization to reduce run times. The generational version of BORG uses autoadaptive operator selection, restarts for stalled search, and adaptive population sizing from BORG (Hadka & Reed, 2013), within the generational e-NSGA2 structure. As a stopping condition, we used 500,000 function evaluations, while convergence was assessed using hypervolume and epsilon progress

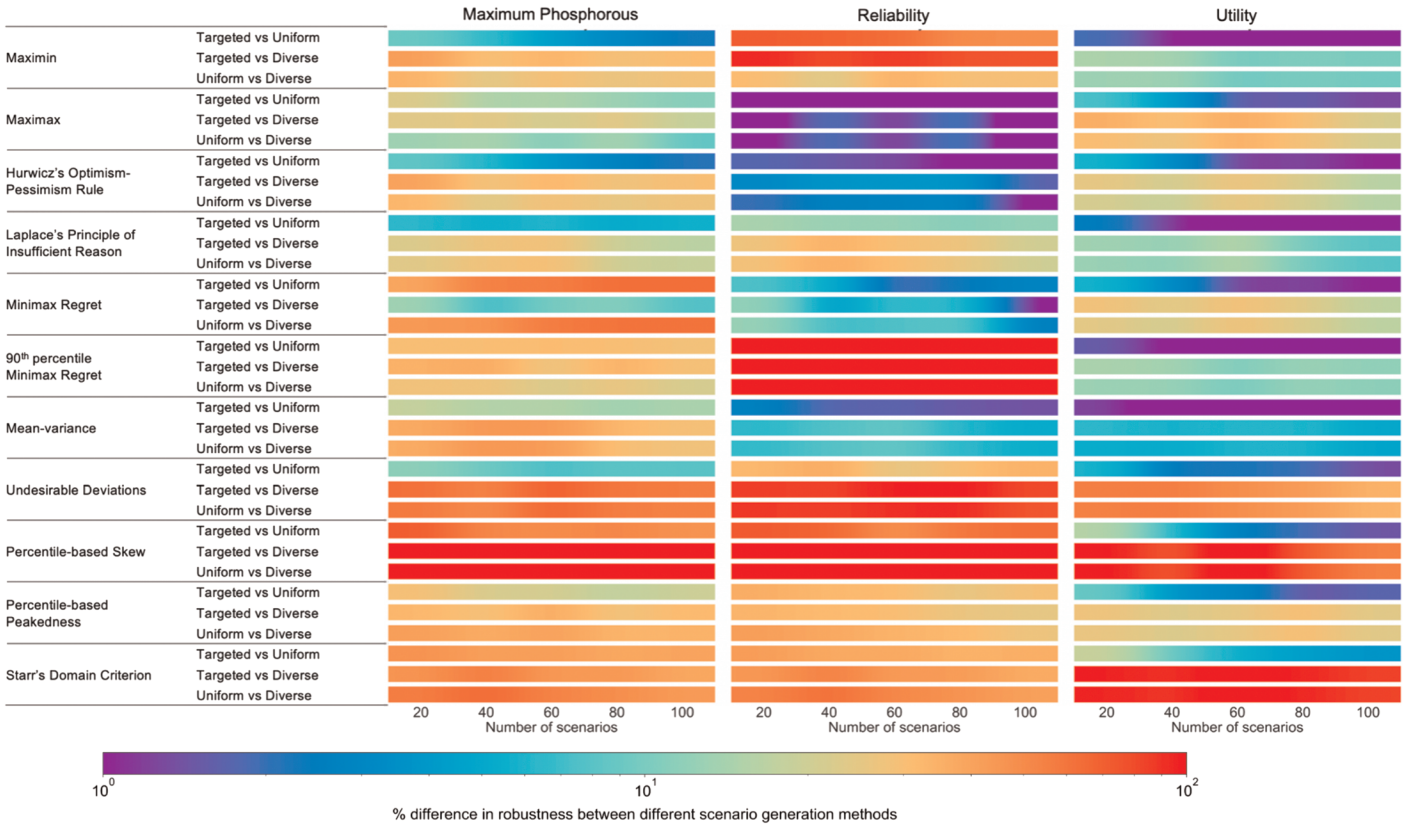

Figure 5. Sensitivity of the robustness metrics (as measured by the percentage difference), for each of the case study performance metrics (maximum phosphorous, utility, and reliability) for each distribution of scenarios in the scenario space (diverse futures, uniform spread, and targeted spread). Red represents high sensitivity, and purple represents low sensitivity of the robustness metric to the set of scenarios.

(Reed et al., 2013; Ward et al., 2015). We repeated this for 50 different initial random seeds and merged the final results into one large set of final decision alternatives. For each decision alternative, three performance values are produced per simulation (described in more detail in section 4.1) and per scenario, leading to a total of 248,994,000 performance values (i.e., the product of 18,000 scenarios that were grouped into 300 sets of scenarios, 4,611 decision alternatives, and three performance metrics).

3.4. Robustness Metrics

Robustness values were calculated using 10 different robustness metrics (see Table 2), also used by McPhail et al. (2018), and chosen because they assess global robustness, rather than local robustness (i.e., no "reference" or "best estimate" scenario needs to be selected) (Matrosov et al., 2013; Roach et al., 2016). The consideration of global robustness, rather than local robustness, is important, due to the ability for global robustness to better analyze and manage nonprobabilistic uncertainty (Sniedovich, 2010). The aggregation of performance values across each set of scenarios for the robustness metrics involved the manipulation of the 248,994,000 performance values into 45,648,900 robustness values (i.e., the product of 300 sets of scenarios, 4,611 decision alternatives, 3 performance metrics, and 11 robustness metrics). These robustness values were then used to assess the impact of different scenario sets on (a) the robustness of decision alternatives and (b) the ranking of decision alternatives (methodology explained in more detail in sections 3.2 and 3.3, respectively).

4. Results and Discussion

4.1. Robustness Values

Following the methodology outlined in Figure 2, the sensitivity of each robustness metric to the different distributions of scenarios is shown in Figure 5. The sensitivity is the percentage difference between the

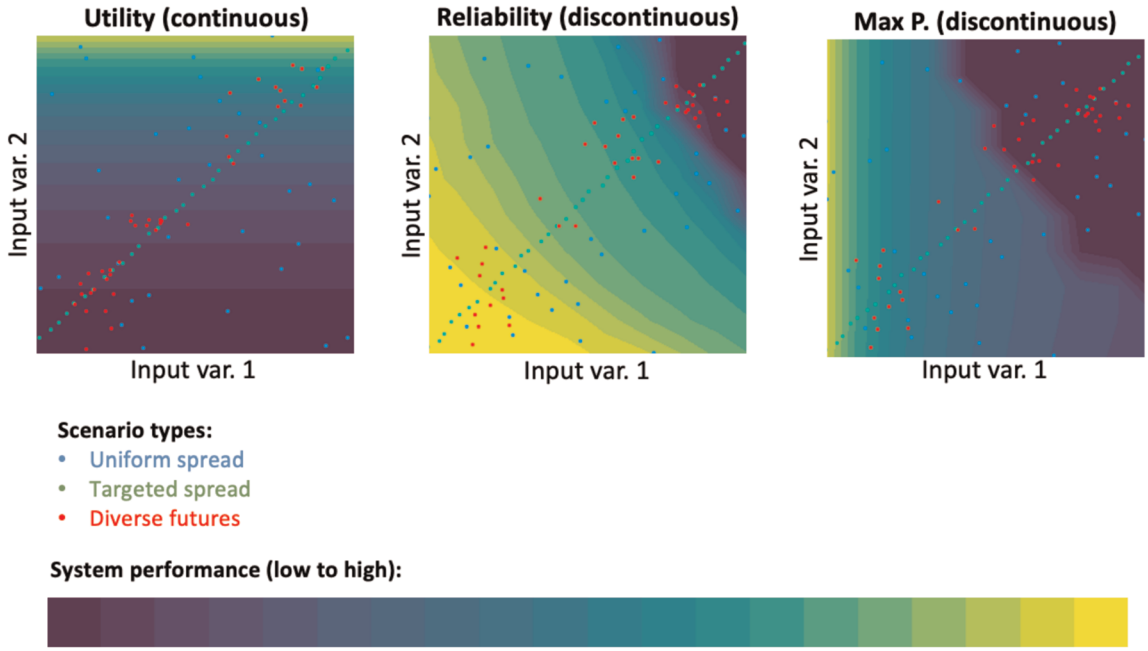

Figure 6. Illustration of how different sets of scenarios will sample different points in the space of system performance values for the Lake Problem. Robustness is calculated by the sampled system performance values and therefore affected by the distribution of scenarios and system performance values.

robustness calculated for two different sets of scenarios, and this is averaged across all of the Pareto-optimal decision alternatives (as described in Figure 2) from each of the 50 optimization runs. Orange and red represents high sensitivity (i.e., >10% difference in robustness for the two different sets of scenarios) and purple and blue represents low sensitivity (i.e., <10% difference in robustness for the two different sets of scenarios). The robustness values (and therefore the sensitivity of the robustness values) are calculated using the distribution of scenarios in the scenario space (e.g., diverse futures or uniform spread), the decision alternatives, the performance metric (e.g., reliability), and the robustness metric (e.g., Maximin). The decision alternatives are purely case specific, while the other three factors are more general; therefore, we have presented the results in Figure 5 in a way that allows the scenario distribution, performance metric, and robustness metric to be compared one by one, or in combination.

Overall, Figure 5 indicates that scenario selection has a large impact on robustness values. This is evidenced by the fact that the bars are generally green, orange, or red when comparing the robustness values obtained when different scenario sets are used (indicating a difference in robustness values in excess of 10%) (Figure 5). This is most likely because the different sets of scenarios are covering very different areas of the scenario space (Figure 6), and therefore, different input variables are being used by the model to determine system performance and robustness.

The results also show that differences in robustness methods between scenarios that represent a uniform spread and scenarios that represent a targeted spread are smaller than those between the other two combinations of distributions of scenarios (Figure 5), particularly for the reliability performance metric. To explain why this occurs, Figure 6 shows one set of scenarios for each of the different scenario distributions, overlaid on the performance values for a 2-D subspace of the scenario space for a single decision alternative. Figure 6 indicates that the scenario space is covered very differently by scenarios that represent diverse futures, a uniform spread and a targeted spread. In particular, scenarios are spread across all levels of performance when the set represents a uniform spread or targeted spread (all colors in Figure 6); however, some levels of performance (some colors) will be missed when there is a clustering of scenarios, as happens when the scenarios are representative of diverse futures, particularly if there are thresholds in performance (e.g., for reliability and maximum phosphorous). The similarity in coverage of the performance values by the distribution of scenarios representing a uniform spread and targeted spread leads them to produce more similar values of robustness relative to the distribution of scenarios that is representative of diverse futures.

Example #	Coverage of scenario space	Behavior of performance metric	Behavior of robustness metric	Degree of similarity in robustness	
A	Dissimilar E.g. Diverse futures vs Uniform spread	Discontinuous, unbounded E.g. Max. Phosphorous	All E.g. Mean-variance	Very dissimilar	
B	Dissimilar E.g. Diverse futures vs Uniform spread	Discontinuous, bounded E.g. Reliability	Most metrics E.g. Laplace's Principle	Very dissimilar	to similar
C	Dissimilar E.g. Diverse futures vs Uniform spread	Discontinuous, bounded E.g. Reliability	Extreme low or high risk averseness E.g. Maximin, maximax	Very dissimilar	to very similar
D	Dissimilar E.g. Diverse futures vs Uniform spread	Continuous E.g. Utility	Percentile-based E.g. Skewness	Very dissimilar	
E	Dissimilar E.g. Diverse futures vs Uniform spread	Continuous E.g. Utility	Most metrics E.g. Maximin, Mean-variance	Dissimilar	to similar
F	Similar E.g. Uniform spread vs Targeted spread	Discontinuous, unbounded E.g. Max. Phosphorous	All E.g. Mean-variance	Very dissimilar	to dissimilar
G	Similar E.g. Uniform spread vs Targeted spread	Discontinuous, bounded E.g. Reliability	Most metrics E.g. Minimax regret	Very dissimilar	to dissimilar
H	Similar E.g. Uniform spread vs Targeted spread	Discontinuous, bounded E.g. Reliability	Extreme low or high risk averseness E.g. Maximin, Maximax	Very dissimilar	to very similar
I	Similar E.g. Uniform spread vs Targeted spread	Continuous E.g. Utility	Percentile-based E.g. Skewness	Neutral	
J	Similar E.g. Uniform spread vs Targeted spread	Continuous E.g. Utility	Most metrics E.g. Mean-variance	Very similar	

Figure 7. General indication of how different distributions of scenarios, different performance metrics, and different robustness metrics all affect the robustness of decision alternatives in for the Lake Problem.

As mentioned above, the degree of similarity in robustness values can be affected by the distribution of the performance values. For example, when considering the utility metric column in Figure 5, it can be seen that there are significantly fewer orange and red bars, which indicate high similarity in robustness values. The utility metric shows slightly more similarity in robustness values when the distribution of scenarios is representative of diverse futures but much greater similarity when the scenarios correspond to a uniform spread

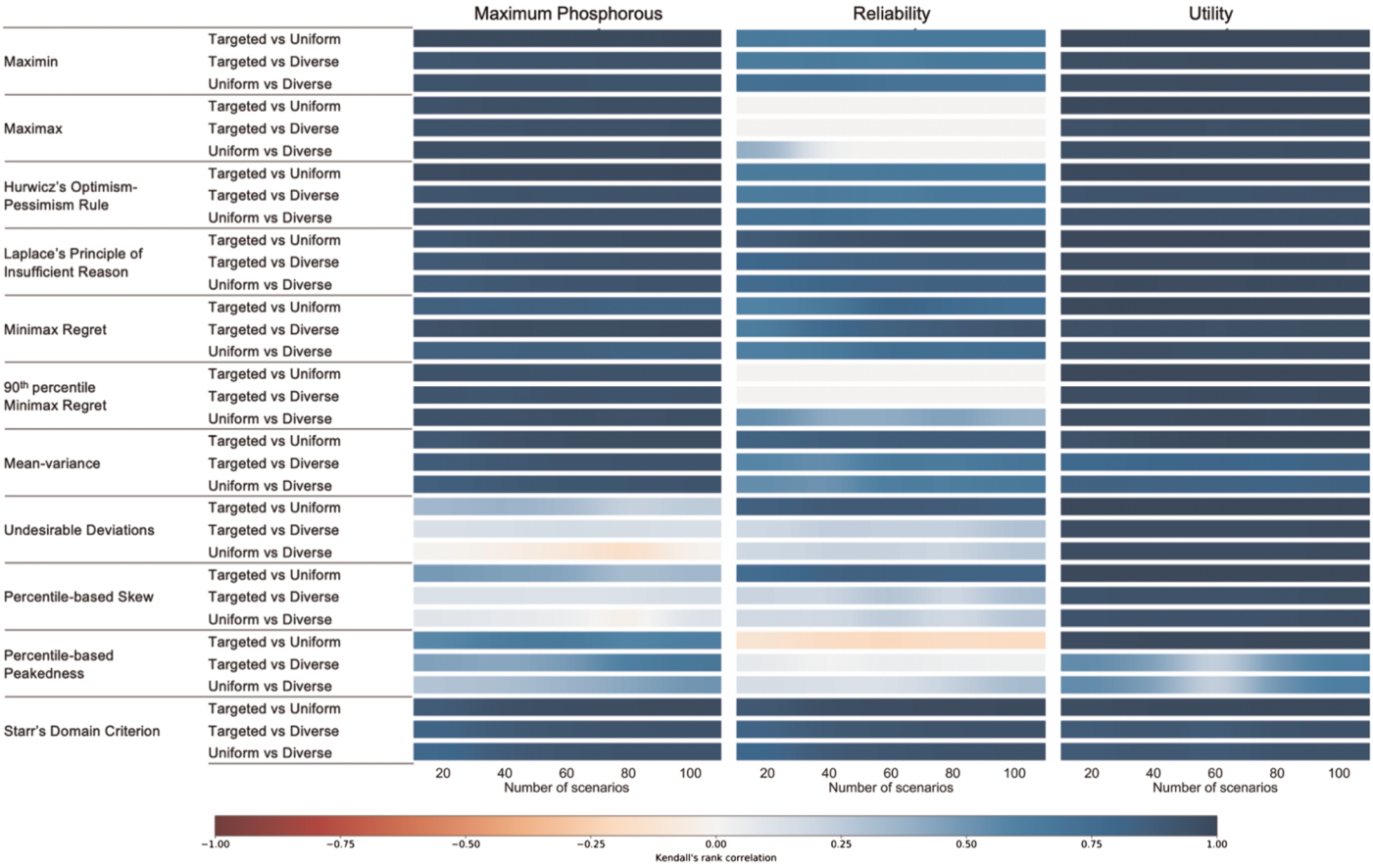

Figure 8. Similarity of the rankings of decision alternatives (as measured by Kendall's Tau-b) for each of the case study objectives (maximum phosphorous, utility, and reliability) for each pair of distributions of scenarios (diverse futures, uniform spread, and targeted spread). Red or white represents low and blue represents high similarity (decision alternatives have the same rankings).

or targeted spread. Figure 6 illustrates that the performance values for the utility metric form a smooth and continuous space, relative to the reliability and maximum phosphorous metrics, which have sharp gradients and nonlinearities (due to tipping points in the environmental dynamics of the Lake Problem). This leads to correspondingly higher dissimilarity in robustness values for the latter metrics.

In most instances, the number of scenarios considered does not have a significant effect on the relative similarities or differences in the robustness values obtained using the different distributions of scenarios throughout the scenario space (Figure 5). This indicates that the way that the scenarios cover the scenario space (i.e., diverse futures, uniform spread, or targeted spread) plays a greater role in determining robustness values than the number of scenarios used for each approach. For situations where there is a gradient in robustness values with an increase in the number of scenarios, the level of agreement in robustness values increases as the number of scenarios increases.

Figure 7 summarizes the trends in similarity in robustness values from Figure 5, with reference to the factors affecting this similarity (Figure 6). Figure 7 highlights (using examples) that, in general, a dissimilar coverage of the scenario space (e.g., the diverse futures vs. uniform spread scenarios, as discussed previously) will lead to a lower degree of similarity in robustness (Example A in Figure 7), and a more similar coverage of the scenario space (e.g., targeted spread vs. uniform spread scenarios) leads to a higher degree of similarity in robustness (Example F in Figure 7). However, the interactions of the distribution of scenarios, the behavior of the system performance metrics, and the behavior of the robustness metrics (Figure 6) are complex, and so there are exceptions to these findings.

An exception to the general findings is that the value of the Maximax robustness metric is insensitive to the distribution of scenarios used for the reliability system performance metric, especially if a sufficiently large

number of scenarios is used. This is because almost any decision alternative will achieve 100% reliability if the uncertain model inputs affecting the pollution levels (e.g., the mean natural pollution inflow) are favorable. In other words, for almost any decision alternative, there is some favorable region of the scenario space where the decision alternative can achieve 100% reliability. Due to the Maximax metric selecting the scenario with the best performance, the robustness will always be 100%, regardless of the distribution of scenarios or the decision alternative. This highlights how a performance metric with bounds (e.g., reliability is bounded between 0% and 100%) can interact with some robustness metrics (e.g., Maximax and Maximin, which use the best- and worst-case performances, respectively), as highlighted in Examples C and H in Figure 7. Note that this effect is not seen for the Maximin metric in this case study, because the starting conditions for the lake do not allow for the possibility of 0% reliability and thus the reliability is always greater than 0% in practice.

Robustness metrics that use percentiles (e.g., the undesirable deviations metric, percentile-based skewness, and percentile-based peakedness) are sensitive to the distribution of scenarios in the scenario space because they are dependent on the higher-order moments of the distribution of performance values, which can vary more significantly than mean performance (e.g., Laplace's principle of insufficient reason) and also vary more significantly than bounded maximum and minimum performance (Maximax and Maximin metrics respectively) (see Examples D and I, Figure 7). Metrics that use percentiles were an exception to the generalized findings, and it should be noted that these metrics were also found to behave very differently to the other metrics in McPhail et al. (2018).

The above results indicate that the similarity of robustness metrics when comparing the robustness calculated from different distributions of scenarios is a function of the complex interactions between

1. The similarity/dissimilarity of the coverage of the space of plausible values of the model inputs that are represented by scenarios.
2. The behavior (e.g., smoothness and discontinuities) of the system performance metric over the space of plausible model input values.
3. The number of scenarios used in the calculation of robustness (when comparing the distributions of scenarios corresponding to a uniform spread and a targeted spread).

4.2. Ranking Similarity

Following the methodology outlined in Figure 3, the correlation of the performance values (i.e., similarity in how the decision alternatives are ranked) is shown in Figure 8. The similarity of the rankings of the decision alternatives is given by Kendall's Tau-b for two different sets of scenarios, and this is averaged across all decision alternatives and all random seeds (as described in Figure 3). A value of −1 (red) indicates that the two distributions of scenarios give perfectly opposite rankings for the decision alternatives and a value of 1 (blue) represents the case where the rankings are the same (regardless of how different the robustness values are). A value of 0 represents the case where there is no correlation between the two methods, and therefore this represents a low similarity in rankings. Figure 9 summarizes the results from Figure 8, highlighting that in general, the coverage of the scenario space has little to no impact on the ranking of decision alternatives, which are almost always ranked the same way. However, as with the analysis of robustness values (section 4.1), there are some exceptions to the above findings for the rankings, which are due to the interactions between the distribution of scenarios, the behavior of the system performance metrics, and the behavior of the robustness metrics (see Figure 6).

Overall, Figure 8 indicates that for the majority of robustness values, the distribution of scenarios in the scenario space has a minor impact on the rankings of decision alternatives (with a few exceptions explained in more detail below). This is evidenced by the fact that much of Figure 8 is shaded dark blue, representing a positive correlation in the rankings of the decision alternatives when different distributions of scenarios are used to calculate robustness. This is likely because a high dissimilarity in robustness values (evidenced by much of Figure 5) does not necessarily mean a high dissimilarity in rankings. Therefore, although the robustness values may be very dissimilar when different scenario selection methods are used, the values are not changing relative to each other so that the same decision alternative would be selected as the most robust in both cases (i.e., the relative robustness of different decision alternatives is the same).

Example #	Coverage of scenario space	Behavior of performance metric	Behavior of robustness metric	Degree of similarity in rankings
A	Dissimilar E.g. Diverse futures vs Uniform spread	Discontinuous, unbounded E.g. Max. Phosphorous	Multiple percentiles or undesirable deviations E.g. Percentile-based skewness	Dissimilar to similar
B	Dissimilar E.g. Diverse futures vs Uniform spread	Discontinuous, unbounded E.g. Max. Phosphorous	Most metrics E.g. Minimax regret	Very similar
C	Dissimilar E.g. Diverse futures vs Uniform spread	Discontinuous, bounded E.g. Reliability	Multiple percentiles or undesirable deviations E.g. Percentile-based skewness	Dissimilar to similar
D	Dissimilar E.g. Diverse futures vs Uniform spread	Discontinuous, bounded E.g. Reliability	All E.g. Maximin	Similar to very similar
E	Dissimilar E.g. Diverse futures vs Uniform spread	Continuous E.g. Utility	All E.g. Maximin	Similar to very similar
F	Similar E.g. Uniform spread vs Targeted spread	Discontinuous, unbounded E.g. Max. Phosphorous	Multiple percentiles or undesirable deviations E.g. Percentile-based skewness	Dissimilar to similar
G	Similar E.g. Uniform spread vs Targeted spread	Discontinuous, unbounded E.g. Max. Phosphorous	Most metrics E.g. Minimax regret	Very similar
H	Similar E.g. Uniform spread vs Targeted spread	Discontinuous, bounded E.g. Reliability	Multiple percentiles or undesirable deviations E.g. Percentile-based skewness	Dissimilar to similar
I	Similar E.g. Uniform spread vs Targeted spread	Discontinuous, bounded E.g. Reliability	All E.g. Maximin	Similar to very similar
J	Similar E.g. Uniform spread vs Targeted spread	Continuous E.g. Utility	All E.g. Maximin	Very similar

Figure 9. General indication of how different distributions of scenarios, different performance metrics, and different robustness metrics all affect the rankings of decision alternatives in for the Lake Problem.

The number of scenarios does not have a significant effect on the rankings of the decision alternatives when comparing two sets of scenarios obtained by different methods. The reason for this is that, as described above, a high level of dissimilarity in robustness values calculated for two different distributions of scenarios does not necessarily lead to a change in the rankings of the decision alternatives, and therefore, the rankings have high similarity even as the number of scenarios increases.

Some examples of exceptions to the above findings include that the metrics that consist of multiple percentiles (percentile-based skewness and peakedness) and the undesirable deviations metric can lead to dissimilar rankings in some cases (Examples A, C, F, and G in Figure 9), whereas most other metrics rank decision

alternatives very similarly (see Figure 9). It should also be noted that McPhail et al. (2018) showed these same three robustness metrics to produce very dissimilar rankings when compared to other robustness metrics, even when the same scenarios were used in all robustness calculations.

Another exception is that there is relatively high dissimilarity in the rankings of the Maximax metric when robustness is calculated using the reliability metric. This is because, as mentioned previously, the region of the scenario space where any decision alternative can achieve 100% reliability is very large for this case study, and therefore, when using the Maximax metric, most decision alternatives have the same robustness value (100%). Kendall's Tau-b metric (used to determine similarity in ranking) becomes highly sensitive when there are many decision alternatives with the same rankings, because a change in robustness to 99% for a single decision alternative will cause Kendall's Tau-b metric to see the two distributions of scenarios as having a high dissimilarity.

To summarize Figures 8 and 9, the rankings of decision alternatives are generally not strongly affected by scenario selection. However, there are some exceptions, based on the complex interactions between the behavior (e.g., smoothness vs. discontinuities) of the system performance metric (e.g., economic utility vs. reliability and maximum phosphorous) over the space of plausible model input values. The multifaceted nature of the interactions between different aspects of the analysis means that while the overall methodology of assessing the impact of scenarios on the robustness analysis is generalizable, the specific results presented here are likely to be case study specific.

5. Summary and Conclusions

As part of model-based assessment of decision alternatives under deep uncertainty, the performance of the different alternatives is assessed under a range of plausible future conditions (scenarios). However, while each of these scenarios corresponds to a different combination of values of model inputs, there is a diversity of approaches for generating these values in the water resources literature. For example, some studies have determined plausible future conditions by considering changes in atmospheric carbon concentrations and/or socioeconomic conditions, whereas other studies have generated normative scenarios using techniques such as scenario discovery, decision scaling, or adaptive pathways approaches. These scenarios can also be generated in different ways, including qualitative, participatory approaches, or purely quantitative methods. Given this diversity of scenario creation approaches, it is important to determine the impact this has on the robustness values and rankings of decision alternatives.

This paper proposes a methodology for quantitatively assessing the impact of different sets of scenarios on the robustness and rankings (relative robustness) of decision alternatives. The methodology for comparing two sets of scenarios begins by first simulating the decision alternatives across the different sets of scenarios and then calculating the robustness of those decision alternatives using a variety of robustness metrics. The robustness values are analyzed by looking at the relative difference in robustness and by looking at the correlation in the rankings of the decision alternatives (based on robustness) when different distributions of scenarios are used.

As a simplified example of how to apply this methodology, it was used to analyze the effect of three conceptually different distributions of scenarios (Figure 4). The methodology was applied to the Lake problem, using a variety of robustness metrics (Table 2). The results show that the distribution of scenarios has a significant effect on the robustness values calculated (Figure 5) but a small effect on how decision alternatives are ranked (i.e., relative robustness) (Figure 8). With regard to the degree of similarity of robustness values, the results indicated that dissimilar coverage of the scenario space (e.g., a diverse set of futures compared to a uniform spread) generally led to a lower degree of similarity in robustness values, in contrast to a similar coverage of the scenario space (e.g., a uniform spread and a targeted spread), which led to a higher degree in similarity of robustness values. Similarity of the robustness values is also affected by complex interactions of scenario selection with the number of scenarios, the behavior (e.g., smoothness and discontinuities) of the system performance metric over the space of plausible model input values, and the robustness metric itself (Figure 6). In contrast to the robustness values, it was found that the rankings of the decision alternatives based on robustness values often had a moderate to high degree of similarity when different sets of scenarios are used. Again, exceptions to this were caused by certain combinations of the behavior of the system performance metric and the characteristics of the robustness metric used.

The effects of several distributions of scenarios have been assessed using both theoretical and computational evidence, but the results presented are by no means representative of all combinations of scenario selections, robustness metrics, case studies, and so forth. This study used many stochastic simulations to highlight that scenarios can have an effect, but in order to see the effect on real-life decision making, further investigation is warranted. One way to explore the effects on decision making could be through simulation gaming workshops with students, followed by workshops with decision makers, case studies of successful long-term infrastructure plans, and the creation of carefully designed pilot studies to compare these approaches, as recommended by Kwakkel and van der Pas (2011). Further exploration would also be required to understand the impact that the decision alternatives have on this analysis. Here, we used a large set of decision alternatives built from multiple Pareto fronts. Using the generic methodology presented here, it would be possible to see whether scenarios have the same impact when the set of decision alternatives is smaller or is composed of a single Pareto front.

The application of the generic methodology presented in this paper to a simple case study (the Lake Model) allowed this paper to explore the effect of a variety of sets of scenarios, emulating different approaches to creating scenarios used in practice, on the robustness of a system, something that has not been explored previously. Without this method, there is no approach in the literature to understanding the impact of scenario selection on the absolute and relative robustness values of different decision alternatives. We highlighted several examples of how different distributions of scenarios could affect the robustness of decision alternatives in different ways, which shows the utility of the generic methodology. Interestingly, in the case study considered, the number of scenarios seemed to have relatively little impact, and the results also showed that despite the significant effect of the distribution of scenarios on robustness values, the effect on the rankings of the decision alternatives was relatively small (and in many cases negligible).

Data Availability Statement

Descriptions of scenario generation methods across the water resources literature and a detailed methodology to calculate diverse futures and targeted spread scenarios are available in the supporting information. The Lake Model is widely available on GitHub in multiple repositories, including in the EMAworkbench (https://github.com/quaquel/EMAworkbench).

Acknowledgments
Thanks is given to SA Water Corporation (Australia) who support the research of Cameron McPhail through Water Research Australia, and thanks is also given to Water Research Australia. The authors would also like to thank Andrea Castelletti and Matteo Giuliani (both from Politecnico di Milano) for their important conceptual contributions to this research.

References

Anghileri, D., Botter, M., Castelletti, A., Weigt, H., & Burlando, P. (2018). A comparative assessment of the impact of climate change and energy policies on Alpine hydropower. *Water Resources Research*, *54*, 9144–9161. https://doi.org/10.1029/2017WR022289

Beh, E., Dandy, G., Maier, H. R., & Paton, F. L. (2014). Optimal sequencing of water supply options at the regional scale incorporating alternative water supply sources and multiple objectives. *Environmental Modelling and Software*, *53*, 137–153. https://doi.org/10.1016/j.envsoft.2013.11.004

Beh, E., Maier, H. R., & Dandy, G. C. (2015a). Adaptive, multiobjective optimal sequencing approach for urban water supply augmentation under deep uncertainty. *Water Resources Research*, *51*, 1529–1551. https://doi.org/10.1002/2014WR016254

Beh, E., Maier, H. R., & Dandy, G. C. (2015b). Scenario driven optimal sequencing under deep uncertainty. *Environmental Modelling and Software*, *68*, 181–195. https://doi.org/10.1016/j.envsoft.2015.02.006

Borgomeo, E., Mortazavi-Naeini, M., Hall, J. W., & Guillod, B. P. (2018). Risk, robustness and water resources planning under uncertainty. *Earth's Future*, *6*(3), 468–487. https://doi.org/10.1002/2017EF000730

Brown, C., Ghile, Y., Laverty, M., & Li, K. (2012). Decision scaling: Linking bottom-up vulnerability analysis with climate projections in the water sector. *Water Resources Research*, *48*, W09537. https://doi.org/10.1029/2011WR011212

Bryant, B. P., & Lempert, R. J. (2010). Thinking inside the box: A participatory, computer-assisted approach to scenario discovery. *Technological Forecasting and Social Change*, *77*(1), 34–49. https://doi.org/10.1016/j.techfore.2009.08.002

Burn, D. H., Venema, H. D., & Simonovic, S. P. (1991). Risk-based performance criteria for real-time reservoir operation. *Canadian Journal of Civil Engineering*, *18*(1), 36–42. https://doi.org/10.1139/l91-005

Carpenter, S. R., Ludwig, D., & Brock, W. A. (1999). Management of eutrophication for lakes subject to potentially irreversible change. *Ecological Applications*, *9*(3), 751–771. https://doi.org/10.1890/1051-0761(1999)009[0751:MOEFLS]2.0.CO;2

Culley, S., Bennett, B., Westra, S., & Maier, H. R. (2019). Generating realistic perturbed hydrometeorological time series to inform scenario-neutral climate impact assessments. *Journal of Hydrology*, *576*, 111–122. https://doi.org/10.1016/j.jhydrol.2019.06.005

Culley, S., Noble, S., Yates, A., Timbs, M., Westra, S., Maier, H. R., et al. (2016). A bottom-up approach to identifying themaximum operational adaptive capacity of water resource systems to a changing climate. *Water Resources Research*, *52*, 6751–6768. https://doi.org/10.1002/2015WR018253

Döll, P., & Romero-Lankao, P. (2016). How to embrace uncertainty in participatory climate change risk management—A roadmap. *Earth's Future*, *5*(1), 18–36. https://doi.org/10.1002/eft2.161

Drouet, L., Bosetti, V., & Tavoni, M. (2015). Selection of climate policies under the uncertainties in the fifth assessment report of the IPCC. *Nature Climate Change*, *5*(10), 937–940. Retrieved from. https://doi.org/10.1038/nclimate2721%5Cn

Eker, S., & Kwakkel, J. H. (2018). Including robustness considerations in the search phase of many-objective robust decision making. *Environmental Modelling and Software*, *105*, 201–216. https://doi.org/10.1016/j.envsoft.2018.03.029

Giuliani, M., Anghileri, D., Castelletti, A., Vu, P. N., & Soncini-Sessa, R. (2016). Large storage operations under climate change: Expanding uncertainties and evolving tradeoffs. *Environmental Research Letters*, *11*(3), 35009. https://doi.org/10.1088/1748-9326/11/3/035009

Giuliani, M., & Castelletti, A. (2016). Is robustness really robust? How different definitions of robustness impact decision-making under climate change. *Climatic Change*, *135*(3–4), 409–424. https://doi.org/10.1007/s10584-015-1586-9

Giuliani, M., Castelletti, A., Pianosi, F., Mason, E., & Reed, P. M. (2016). Curses, tradeoffs, and scalable management: Advancing evolutionary multiobjective direct policy search to improve water reservoir operations. *Journal of Water Resources Planning and Management*, *142*(2), 04015050. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000570

Groves, D. G., & Lempert, R. J. (2007). A new analytic method for finding policy-relevant scenarios. *Global Environmental Change*, *17*(1), 73–85. https://doi.org/10.1016/j.gloenvcha.2006.11.006

Haasnoot, M., Kwakkel, J. H., Walker, W. E., & ter Maat, J. (2013). Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. *Global Environmental Change*, *23*(2), 485–498. https://doi.org/10.1016/j.gloenvcha.2012.12.006

Haasnoot, M., Middelkoop, H., Offermans, A., van Beek, E., & van Deursen, W. P. A. (2012). Exploring pathways for sustainable water management in river deltas in a changing environment. *Climatic Change*, *115*(3–4), 795–819. https://doi.org/10.1007/s10584-012-0444-2

Hadka, D., Herman, J., Reed, P., & Keller, K. (2015). An open source framework for many-objective robust decision making. *Environmental Modelling and Software*, *74*, 114–129. https://doi.org/10.1016/j.envsoft.2015.07.014

Hadka, D., & Reed, P. (2013). Borg: An auto-adaptive many-objective evolutionary computing framework. *Evolutionary Computation*, *21*(2), 231–259. https://doi.org/10.1162/EVCO_a_00075

Hall, J. W., & Harvey, H. (2009). *Decision making under severe uncertainties for flood risk management: A case study of info-gap robustness analysis*. Paper presented at Proceedings of 8th International Conference on Hydroinformatics.

Hall, J. W., Lempert, R. J., Keller, K., Hackbarth, A., Mijere, C., & Mcinerney, D. J. (2012). Robust climate policies under uncertainty: A comparison of robust decision making and info-gap methods. *Risk Analysis*, *32*(10), 1657–1672. https://doi.org/10.1111/j.1539-6924.2012.01802.x

Hashimoto, T., Stedinger, J. R., & Loucks, D. P. (1982). Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. *Water Resources Research*, *18*(1), 14–20. https://doi.org/10.1029/WR018i001p00014

Herman, J. D., & Giuliani, M. (2018). Policy tree optimization for threshold-based water resources management over multiple timescales. *Environmental Modelling and Software*, *99*, 39–51. https://doi.org/10.1016/j.envsoft.2017.09.016

Herman, J. D., Reed, P. M., Zeff, H. B., & Characklis, G. W. (2015). How should robustness be defined for water systems planning under change? *Journal of Water Resources Planning and Management*, *141*(10), 04015012. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000509

Herman, J. D., Zeff, H. B., Reed, P. M., & Characklis, G. W. (2014). Beyond optimality: Multistakeholder robustness tradeoffs for regional water portfolio planning under deep uncertainty. *Water Resources Research*, *50*, 7692–7713. https://doi.org/10.1002/2014WR015338

Huskova, I., Matrosov, E. S., Harou, J. J., Kasprzyk, J. R., & Lambert, C. (2016). Screening robust water infrastructure investments and their trade-offs under global change: A London example. *Global Environmental Change*, *41*, 216–227. https://doi.org/10.1016/j.gloenvcha.2016.10.007

Kasprzyk, J. R., Nataraj, S., Reed, P. M., & Lempert, R. J. (2013). Many objective robust decision making for complex environmental systems undergoing change. *Environmental Modelling and Software*, *42*, 55–71. https://doi.org/10.1016/j.envsoft.2012.12.007

Kwadijk, J. C. J., Haasnoot, M., Mulder, J. P. M., Hoogvliet, M. M. C., Jeuken, A. B. M., van der Krogt, R. A. A., et al. (2010). Using adaptation tipping points to prepare for climate change and sea level rise: A case study in the Netherlands. *Wiley Interdisciplinary Reviews: Climate Change*, *1*(5), 729–740. https://doi.org/10.1002/wcc.64

Kwakkel, J. H. (2017). The Exploratory Modeling Workbench: An open source toolkit for exploratory modeling, scenario discovery, and (multi-objective) robust decision making. *Environmental Modelling and Software*, *96*, 239–250. https://doi.org/10.1016/j.envsoft.2017.06.054

Kwakkel, J. H., Eker, S., & Pruyt, E. (2016). How robust is a robust policy? Comparing alternative robustness metrics for robust decision-making. In *International series in operations research and management science* (Vol. 241, pp. 221–237). New York: Springer. https://doi.org/10.1007/978-3-319-33121-8_10

Kwakkel, J. H., Haasnoot, M., & Walker, W. E. (2015). Developing dynamic adaptive policy pathways: A computer-assisted approach for developing adaptive strategies for a deeply uncertain world. *Climatic Change*, *132*(3), 373–386. https://doi.org/10.1007/s10584-014-1210-4

Kwakkel, J. H., Haasnoot, M., & Walker, W. E. (2016). Comparing robust decision-making and dynamic adaptive policy pathways for model-based decision support under deep uncertainty. *Environmental Modelling and Software*, *86*, 168–183. https://doi.org/10.1016/j.envsoft.2016.09.017

Kwakkel, J. H., & van der Pas, J. W. G. M. (2011). Evaluation of infrastructure planning approaches: An analogy with medicine. *Futures*, *43*(9), 934–946. https://doi.org/10.1016/j.futures.2011.06.003

Kwakkel, J. H., Walker, W., & Marchau, V. (2012). Assessing the efficacy of adaptive airport strategic planning: Results from computational experiments. *Environment and Planning. B, Planning & Design*, *39*(3), 533–550. https://doi.org/10.1068/b37151

Kwakkel, J. H., Walker, W. E., & Haasnoot, M. (2016). *Coping with the wickedness of public policy problems: Approaches for decision making under deep uncertainty*. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000626

Kwakkel, J. H., Walker, W. E., & Marchau, V. A. W. J. (2010). Classifying and communicating uncertainties in model-based policy analysis. *International Journal of Technology, Policy and Management*, *10*(4), 299. https://doi.org/10.1504/IJTPM.2010.036918

Lempert, R. J. (2003). *Shaping the next one hundred years: New methods for quantitative, long-term policy analysis*. Rand Corporation. https://doi.org/10.1016/j.techfore.2003.09.006

Lempert, R. J., & Collins, M. T. (2007). Managing the risk of uncertain threshold responses: Comparison of robust, optimum, and precautionary approaches. *Risk Analysis*, *27*(4), 1009–1026. https://doi.org/10.1111/j.1539-6924.2007.00940.x

Lempert, R. J., & Trujillo, H. R. (2018). *Deep decarbonization as a risk management challenge*. https://doi.org/10.7249/PE303

Lenton, T. M. (2013). Environmental tipping points. *Annual Review of Environment and Resources*, *38*(1), 1–29. https://doi.org/10.1146/annurev-environ-102511-084654

Little, J. C., Hester, E. T., Elsawah, S., Filz, G. M., Sandu, A., Carey, C. C., et al. (2018). A tiered, system-of-systems modeling framework for resolving complex socio-environmental policy issues. *Environmental Modelling & Software*, *112*, 82–94.

Maier, H. R., Guillaume, J. H. A., van Delden, H., Riddell, G. A., Haasnoot, M., & Kwakkel, J. H. (2016). An uncertain future, deep uncertainty, scenarios, robustness and adaptation: How do they fit together? *Environmental Modelling and Software*, *81*, 154–164. https://doi.org/10.1016/j.envsoft.2016.03.014

Maier, H. R., Lence, B. J., Tolson, B. A., & Foschi, R. O. (2001). First order reliability method for estimating reliability, vulnerability, and resilience. *Water Resources Research*, *37*(3), 779–790. https://doi.org/10.1029/2000WR900329

Matrosov, E. S., Padula, S., & Harou, J. J. (2013). Selecting portfolios of water supply and demand management strategies under uncertainty—Contrasting economic optimisation and "robust decision making" approaches. *Water Resources Management*, *27*, 1123–1148. https://doi.org/10.1007/s11269-012-0118-x

McInerney, D., Lempert, R., & Keller, K. (2012). What are robust strategies in the face of uncertain climate threshold responses? *Climatic Change*, *112*(3–4), 547–568. https://doi.org/10.1007/s10584-011-0377-1

McPhail, C., Maier, H. R., Kwakkel, J. H., Giuliani, M., Castelletti, A., & Westra, S. (2018). Robustness metrics: How are they calculated, when should they be used and why do they give different results? *Earth's Future*, *6*(2), 169–191. https://doi.org/10.1002/2017EF000649

Morgan, M. G., Henrion, M., & Small, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511840609

Phadnis, S. (2019). Effectiveness of Delphi-and scenario planning-like processes in enabling organizational adaptation: A simulation-based comparison. *Futures & Foresight Science*, *1*(2), e9.

Quinn, J. D., Reed, P. M., Giuliani, M., Castelletti, A., Oyler, J. W., & Nicholas, R. E. (2018). Exploring how changing monsoonal dynamics and human pressures challenge multireservoir management for flood protection, hydropower production, and agricultural water supply. *Water Resources Research*, *54*, 4638–4662. https://doi.org/10.1029/2018WR022743

Quinn, J. D., Reed, P. M., & Keller, K. (2017). Direct policy search for robust multi-objective management of deeply uncertain socio-ecological tipping points. *Environmental Modelling & Software*, *92*, 125–141. https://doi.org/10.1016/j.envsoft.2017.02.017

Raso, L., Kwakkel, J., Timmermans, J., & Panthou, G. (2019). How to evaluate a monitoring system for adaptive policies: criteria for signposts selection and their model-based evaluation. *Climatic Change*, *153*(1–2), 267–283. https://doi.org/10.1007/s10584-018-2355-3

Ravalico, J. K., Dandy, G. C., & Maier, H. R. (2010). Environmental Modelling & Software Management Option Rank Equivalence (MORE)—A new method of sensitivity analysis for decision-making. *Environmental Modelling and Software*, *25*(2), 171–181. https://doi.org/10.1016/j.envsoft.2009.06.012

Ravalico, J. K., Maier, H. R., & Dandy, G. C. (2009). Sensitivity analysis for decision-making using the MORE method—A Pareto approach. *Reliability Engineering and System Safety*, *94*(7), 1229–1237. https://doi.org/10.1016/j.ress.2009.01.009

Reed, P. M., Hadka, D., Herman, J. D., Kasprzyk, J. R., & Kollat, J. B. (2013). Evolutionary multiobjective optimization in water resources: The past, present, and future. *Advances in Water Resources*, *51*, 438–456. https://doi.org/10.1016/j.advwatres.2012.01.005

Roach, T., Kapelan, Z., Ledbetter, R., & Ledbetter, M. (2016). Comparison of robust optimization and info-gap methods for water resource management under deep uncertainty. *Journal of Water Resources Planning and Management*, *142*(9), 04016028. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000660

Shepherd, T. G., Boyd, E., Calel, R. A., Chapman, S. C., Dessai, S., Dima-West, I. M., et al. (2018). Storylines: An alternative approach to representing uncertainty in physical aspects of climate change. *Climatic Change*, *151*(3-4), 555–571.

Singh, R., Reed, P. M., & Keller, K. (2015). Many-objective robust decision making for managing an ecosystem with a deeply uncertain threshold response. *Ecology and Society*, *20*(3), art12. https://doi.org/10.5751/ES-07687-200312

Sniedovich, M. (2010). A bird's view of info-gap decision theory. *The Journal of Risk Finance*, *11*(3), 268–283. https://doi.org/10.1108/15265941011043648

Trindade, B. C., Reed, P. M., Herman, J. D., Zeff, H. B., & Characklis, G. W. (2017). Reducing regional drought vulnerabilities and multi-city robustness conflicts using many-objective optimization under deep uncertainty. *Advances in Water Resources*, *104*, 195–209. https://doi.org/10.1016/j.advwatres.2017.03.023

van Notten, P. W. F., Sleegers, A. M., & van Asselt, M. B. A. (2005). The future shocks: On discontinuity and scenario development. *Technological Forecasting and Social Change*, *72*(2), 175–194. https://doi.org/10.1016/j.techfore.2003.12.003

Varum, C. A., & Melo, C. (2010). Directions in scenario planning literature—A review of the past decades. *Futures*, *42*(4), 355–369. https://doi.org/10.1016/j.futures.2009.11.021

Vervoort, J. M., Thornton, P. K., Kristjanson, P., Förch, W., Ericksen, P. J., Kok, K., et al. (2014). Challenges to scenario-guided adaptive action on food security under climate change. *Global Environmental Change*, *28*, 383–394. https://doi.org/10.1016/j.gloenvcha.2014.03.001

Wada, Y., Vinca, A., Parkinson, S., Willaarts, B. A., Magnuszewski, P., Mochizuki, J., et al. (2019). Co-designing Indus water-energy-land futures. *One Earth*, *1*(2), 185–194. https://doi.org/10.1016/j.oneear.2019.10.006

Walker, W. E., Haasnoot, M., & Kwakkel, J. H. (2013). Adapt or perish: A review of planning approaches for adaptation under deep uncertainty. *Sustainability (Switzerland)*, *5*(3), 955–979. https://doi.org/10.3390/su5030955

Walker, W. E., Lempert, R., & Kwakkel, J. (2013). Deep uncertainty. In *Encyclopedia of operations research and management science* (pp. 395–402). New York: Springer. https://doi.org/10.1007/978-1-4419-1153-7_1140

Ward, V. L., Singh, R., Reed, P. M., & Keller, K. (2015). Confronting tipping points: Can multi-objective evolutionary algorithms discover pollution control tradeoffs given environmental thresholds? *Environmental Modelling & Software*, *73*, 27–43. https://doi.org/10.1016/j.envsoft.2015.07.020

Watson, A. A., & Kasprzyk, J. R. (2017). Incorporating deeply uncertain factors into the many objective search process. *Environmental Modelling and Software*, *89*, 159–171. https://doi.org/10.1016/j.envsoft.2016.12.001

Weaver, C. P., Lempert, R. J., Brown, C., Hall, J. A., Revell, D., & Sarewitz, D. (2013). Improving the contribution of climate model information to decision making: The value and demands of robust decision frameworks. *Wiley Interdisciplinary Reviews: Climate Change*, *4*(1), 39–60.

Zeff, H. B., Kasprzyk, J. R., Herman, J. D., Reed, P. M., & Characklis, G. W. (2014). Navigating financial and supply reliability tradeoffs in regional drought management portfolios. *Water Resources Research*, *50*, 4906–4923. https://doi.org/10.1002/2013WR015126

Zongxue, X., Jinno, K., Kawamura, A., Takesaki, S., & Ito, K. (1998). Performance risk analysis for Fukuoka water supply system. *Water Resources Management*, *21*(1), 49–62. https://doi.org/10.1007/s11269-006-9040-4