AGU PUBLICATIONS

Earth's Future

RESEARCH ARTICLE
10.1002/2017EF000649

Special Section:
Resilient Decision-making for a Riskier World

Robustness Metrics: How Are They Calculated, When Should They Be Used and Why Do They Give Different Results?

C. McPhail[1], H. R. Maier[1], J. H. Kwakkel[2], M. Giuliani[3], A. Castelletti[3], and S. Westra[1]

[1]School of Civil, Environmental, and Mining Engineering, University of Adelaide, Adelaide, SA, Australia, [2]Faculty of Technology Policy and Management, Delft University of Technology, Delft, The Netherlands, [3]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

Abstract Robustness is being used increasingly for decision analysis in relation to deep uncertainty and many metrics have been proposed for its quantification. Recent studies have shown that the application of different robustness metrics can result in different rankings of decision alternatives, but there has been little discussion of what potential causes for this might be. To shed some light on this issue, we present a unifying framework for the calculation of robustness metrics, which assists with understanding how robustness metrics work, when they should be used, and why they sometimes disagree. The framework categorizes the suitability of metrics to a decision-maker based on (1) the decision-context (i.e., the suitability of using absolute performance or regret), (2) the decision-maker's preferred level of risk aversion, and (3) the decision-maker's preference toward maximizing performance, minimizing variance, or some higher-order moment. This article also introduces a conceptual framework describing when relative robustness values of decision alternatives obtained using different metrics are likely to agree and disagree. This is used as a measure of how "stable" the ranking of decision alternatives is when determined using different robustness metrics. The framework is tested on three case studies, including water supply augmentation in Adelaide, Australia, the operation of a multipurpose regulated lake in Italy, and flood protection for a hypothetical river based on a reach of the river Rhine in the Netherlands. The proposed conceptual framework is confirmed by the case study results, providing insight into the reasons for disagreements between rankings obtained using different robustness metrics.

Key Points:
- A unifying framework for the calculation of robustness metrics is presented, providing guidance on the selection of appropriate metrics
- A conceptual framework for conditions under which the relative robustness from different metrics agree and disagree is introduced
- The above frameworks are tested on three diverse case studies from Australia, Italy and the Netherlands

Supporting Information:
- Supporting Information S1.

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

Citation:
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*, 169–191, https://doi.org/10.1002/2017EF000649

Received 1 AUG 2017
Accepted 19 DEC 2017
Accepted article online 8 JAN 2018
Published online 6 FEB 2018

1. Introduction

Uncertainty has long been considered an important facet of environmental decision-making. This uncertainty arises from natural variability, as well as changes in system conditions over time (Maier et al., 2016). In the past, the latter have generally been represented by a "best guess" or "expected" future (Lempert et al., 2006). Consequently, much of the consideration of uncertainty was concerned with the impact of localized uncertainty surrounding expected future conditions (Giuliani et al., 2016c; Morgan et al., 1990) and a realization of the value of information for reducing this localized uncertainty (Howard, 1966; Howard & Matheson, 2005). The consideration of localized uncertainty is reflected in the wide-spread usage of performance metrics such as reliability, vulnerability, and resilience (Burn et al., 1991; Hashimoto et al., 1982a; Maier et al., 2001; Zongxue et al., 1998). However, as a result of climatic, technological, economic and sociopolitical changes, there has been a realization that it is no longer possible to determine a single best-guess of how future conditions might change, especially when considering longer planning horizons (e.g., on the order of 70–100 years) (Döll & Romero-Lankao, 2017; Grafton et al., 2016b; Guo et al., 2017; Maier et al., 2016).

In response, there has been increased focus on deep uncertainty, which is defined as the situation in which parties to a decision do not know, or cannot agree on, how the system under consideration, or parts thereof, work, how important the various outcomes of interest are, and/or what the relevant exogenous inputs to the system are and how they might change in the future (Kwakkel et al., 2010; Lempert, 2003; Maier et al., 2016; Walker et al., 2013). In such a situation, one can enumerate multiple plausible possibilities without being able to rank them in terms of likelihood (Döll & Romero-Lankao, 2017; Kwakkel et al., 2010). This inability can be due to a lack of knowledge or data about the mechanism or functional relationships being studied. However, it can also arise because the various parties involved in the decision cannot come to an agreement.

That is, under deep uncertainty, there is a variety of uncertain factors that jointly affect the consequences of a decision. These uncertain factors define different possible states of the world in a deterministic and set-based manner (Ben-Tal et al., 2009).

As pointed out by Maier et al. (2016), when dealing with deep uncertainty, system performance is generally measured using metrics that preference systems that perform well under a range of plausible conditions, which fall under the umbrella of robustness. It should be noted that while robustness metrics have been considered in different problem domains, such as water resources planning (Hashimoto et al., 1982b), dynamic chemical reaction models (Samsatli et al., 1998), timetable scheduling (Canon & Jeannot, 2007), and data center network service levels (Bilal et al., 2013) for some time, this has generally been in the context of perturbations centered on expected conditions, or local uncertainty, rather than deep uncertainty. In contrast, consideration of robustness metrics for quantifying system performance under deep uncertainty, which is the focus of this article, has only occurred relatively recently.

A number of robustness metrics have been used to measure system performance under deep uncertainty, such as:

- Expected value metrics (Wald, 1950), which indicate an expected level of performance across a range of scenarios.
- Metrics of higher-order moments, such as variance and skew (e.g., Kwakkel et al., 2016b), which provide information on how the expected level of performance varies across multiple scenarios.
- Regret-based metrics (Savage, 1951), where the regret of a decision alternative is defined as the difference between the performance of the selected option for a particular plausible condition and the performance of the best possible option for that condition.
- Satisficing metrics (Simon, 1956), which calculate the range of scenarios that have acceptable performance relative to a threshold.

However, although these metrics all measure system performance over a set of future states of the world, they do so in different ways, making it difficult to assess how robust the performance of a system actually is. For example, these metrics reflect varying levels of risk aversion, and differences about what is meant by robustness. Is robustness about ensuring insensitivity to future developments, reducing regret, or avoiding very negative outcomes? This meta-problem of deciding how to decide (Schneller & Sphicas, 1983) raises the following question: how robust is a robust solution?

Studies in environmental literature discussing this question have been receiving some attention in recent years. Lempert and Collins (2007) compared optimal expected utility, the precautionary principle, and robust decision making using a regret based measure of robustness. They found that the three approaches generated similar results, although some approaches may be more appropriate for different audiences and under different circumstances. Herman et al. (2015) compared two regret-based metrics and two satisficing metrics, showing how the choice of metric could significantly affect the choice of decision alternative. However, they found that the two regret-based metrics tended to agree with each other.

Drouet et al. (2015) contrasted maximin, subjective expected utility, and maxmin expected utility, while Roach et al. (2016) compared two satisficing metrics (info-gap decision theory and Starr's domain criterion). Both studies found that the choice of metric can greatly influence the trade-offs for decision-makers. The former highlighted the importance of understanding the preferences of the decision-maker, while the latter acknowledged the need for studies on more complex systems and the need to compare and combine metrics. Giuliani and Castelletti (2016) compared the classic decision theoretic metrics maximin, maximax, Hurwicz optimism-pessimism rule, minimax regret, and Laplace's principle of insufficient reason, further showing that it is very important to select a metric that is appropriate for the decision-maker's preferences to avoid underestimation of system performance. Kwakkel et al. (2016b) compared five robustness metrics and highlighted the importance of using a combination of metrics to see not just the expected value of performance, but also the dispersion of performance around the mean.

A common conclusion across this work is that different robustness metrics reflect different aspects of what makes a choice robust. This not only makes it difficult to assess the absolute "robustness" of an alternative, but also makes it difficult to determine whether a particular alternative is more robust than another. This leads to confusion for decision-makers, as they have no means of comparing the robustness values

and rankings of different decision alternatives obtained using different robustness metrics in an objective fashion.

To address this shortcoming, the objectives of this article are to (1) introduce a unified framework for the calculation of a wide range of robustness metrics, enabling the robustness values obtained from different metrics to be compared in an objective fashion, (2) introduce a taxonomy of robustness metrics and discuss how this can be used to assist with deciding which robustness metric is most appropriate, providing guidance for decision makers as to which robustness metric should be used in their particular context, (3) introduce a conceptual framework for conditions under which different robustness metrics result in different decisions, or how stable ("robust") the ranking of an alternative is when different robustness metrics are used, providing further guidance to decision-makers, and (4) test the conceptual framework from (3) on three case studies that provide a variety of decision contexts, objectives, scenario types and decision alternatives. The selected case studies are: the water supply augmentation in the southern Adelaide region in Australia (Paton et al., 2013), the operation of Lake Como in Italy for flood protection and water supply purposes (Giuliani & Castelletti, 2016), and flood protection for a hypothetical river called the Waas, which is based on a river reach of the Rhine delta in the Netherlands (Haasnoot et al., 2012).

The remainder of this article is organized as follows. In Section 2, the unified framework for the calculation of robustness metrics is introduced and a variety of robustness metrics are categorized according to this framework. A taxonomy based on these categories is provided in Section 3, as well as a summary of how the robustness metrics are classified in accordance with this taxonomy, the way they consider future uncertainties and the relative level of risk aversion they exhibit. In Section 4 an analysis of the conditions under which robustness metrics agree or disagree with other robustness metrics is given, as well as a conceptual framework categorizing the relative degree of agreement of the rankings of decision alternatives obtained using different robustness metrics based on the properties of the metric and the performance of the system under consideration. The three case studies are introduced in Section 5, as well as a summary of the similarities and differences between them. The robustness of different decision alternatives for the three case studies is calculated in Section 6 using a range of robustness metrics and the results are presented and discussed in terms of the stability of the ranking of different decision alternatives when different robustness metrics are used. Finally, conclusions are presented in Section 7.

2. How Are Robustness Metrics Calculated?

Even though there are many different robustness metrics, irrespective of which metric is used, their calculation generally requires the specification of (1) the decision alternatives (e.g., policy options, designs, solutions, management plans) for which robustness is to be calculated, (2) the outcome of interest (performance metric) of the decision alternatives (e.g., cost, reliability), and (3) the plausible future conditions (scenarios) over which the outcomes of interest/performance of the decision alternatives is to be evaluated. These three components of robustness are illustrated in Figure 1.

Robustness is generally calculated for a given decision alternative, x_i, across a given set of future scenarios $S = \{s_1, s_2, \ldots, s_n\}$ using a particular performance metric $f(\cdot)$. Consequently, the calculation of robustness using a particular metric corresponds to the transformation of the performance of a set of decision alternatives over different scenarios, $f(x_i, S) = \{f(x_i, s_1), f(x_i, s_2), \ldots, f(x_i, s_n)\}$ to the robustness $R(x_i, S)$ of these decision alternatives over this set of scenarios. Although different robustness metrics achieve this transformation in different ways, a unifying framework for the calculation of different robustness metrics can be introduced by representing the overall transformation of $f(x_i, S)$ into $R(x_i, S)$ by three separate transformations: performance value transformation (T_1), scenario subset selection (T_2), and robustness metric calculation (T_3), as shown in Figure 2. Details of these transformations for a range of commonly used robustness metrics are given in Table 1 and their mathematical implementations are given in Supporting Information S1.

The performance value transformation (T_1) converts the performance values $f(x_i, S)$ into the type of information $f'(x_i, S)$ used in the calculation of the robustness metric $R(x_i, S)$. For some robustness metrics, the absolute performance values (e.g., cost, reliability) are used, in which case T_1 corresponds to the identity transform (i.e., the performance values are not changed). For other robustness metrics, the absolute system performance values are transformed to values that either measure the regret that results from selecting

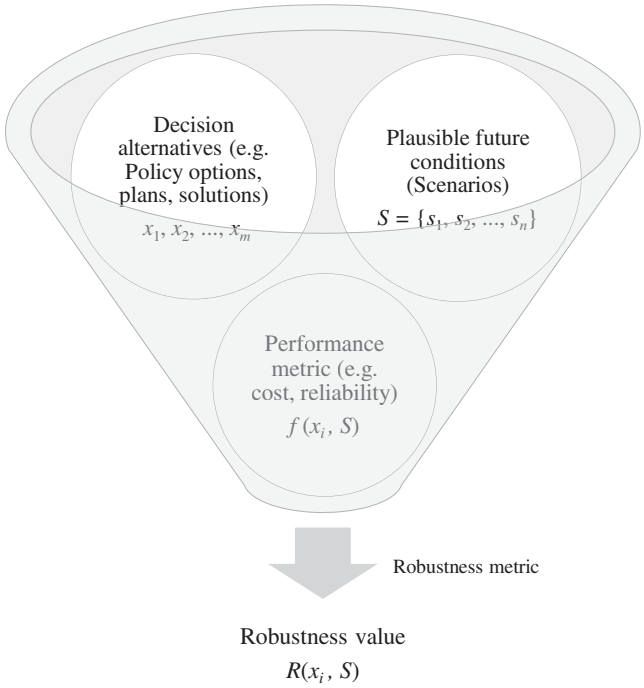

Figure 1. Common components contributing to the calculation of robustness.

a particular decision alternative rather than the one that performs best had a particular future actually occurred or indicate whether the selection of a decision alternative results in satisfactory system performance or not (i.e., whether required system constraints have been satisfied or not).

The scenario subset selection transformation (T_2) involves determining which values of $f'(x_i, S)$ to use in the robustness metric calculation (T_3) (i.e., $f'(x_i, S') \subseteq f'(x_i, S)$), which is akin to selecting a subset of the available scenarios over which system performance is to be assessed. This reflects a particular degree of risk aversion, where consideration of more extreme scenarios in the calculation of a robustness metric corresponds to a higher degree of risk aversion and vice versa. As can be seen from Table 1, which scenarios are considered in the robustness calculation is highly variable between different metrics.

The third transformation (T_3) involves the calculation of the actual robustness metric based on transformed system performance values (T_1) for the selected scenarios (T_2), which corresponds to the transformation of $f'(x_i, S')$ to a single robustness value, $R(x_i, S)$. This equates to an identity transform in cases where only a single scenario is selected in T_2, as there is only a single transformed performance value, which automatically becomes the robustness value. However, in cases where there are transformed performance values for multiple scenarios, these have to be transformed into a single value by means of calculating statistical moments of these values, such as the mean, standard deviation, skewness or kurtosis.

3. When Should Different Robustness Metrics Be Used?

In this section, a taxonomy of different robustness metrics is given in accordance with the three transformations introduced in Section 2. A summary of the three transformations, as well as the relative level of risk aversion, is provided in Section 3.4.

3.1. Transformation 1 (T_1): Performance Value Transformation

A categorization of different robustness metrics in accordance with the performance value transformation (T_1) is given in Table 2. As can be seen, the categorization is based on (1) whether calculation of a robustness metric is based on the absolute performance of a particular decision alternative or the performance of a decision alternative relative to that of another decision alternative or a benchmark; and (2) whether a robustness metric provides an indication of actual system performance or whether system performance is satisfactory compared with a pre-specified performance threshold.

Many of the classic decision analytic robustness metrics belong to the bottom-right hand quadrant of Table 2, including the maximax and maximin criteria, Hurwicz's optimism-pessimism rule and Laplace's principle of insufficient reason, as well as well more recently developed metrics such as the mean-variance criterion, percentile based skewness and percentile-based peakedness. These metrics utilize information about the absolute performance (e.g., cost, reliability) of a particular decision alternative in a particular scenario. Consequently, values of $f(x_i, S')$ consist of these performance values, and robust decision alternatives are those that maximize system performance across the scenarios. The difference

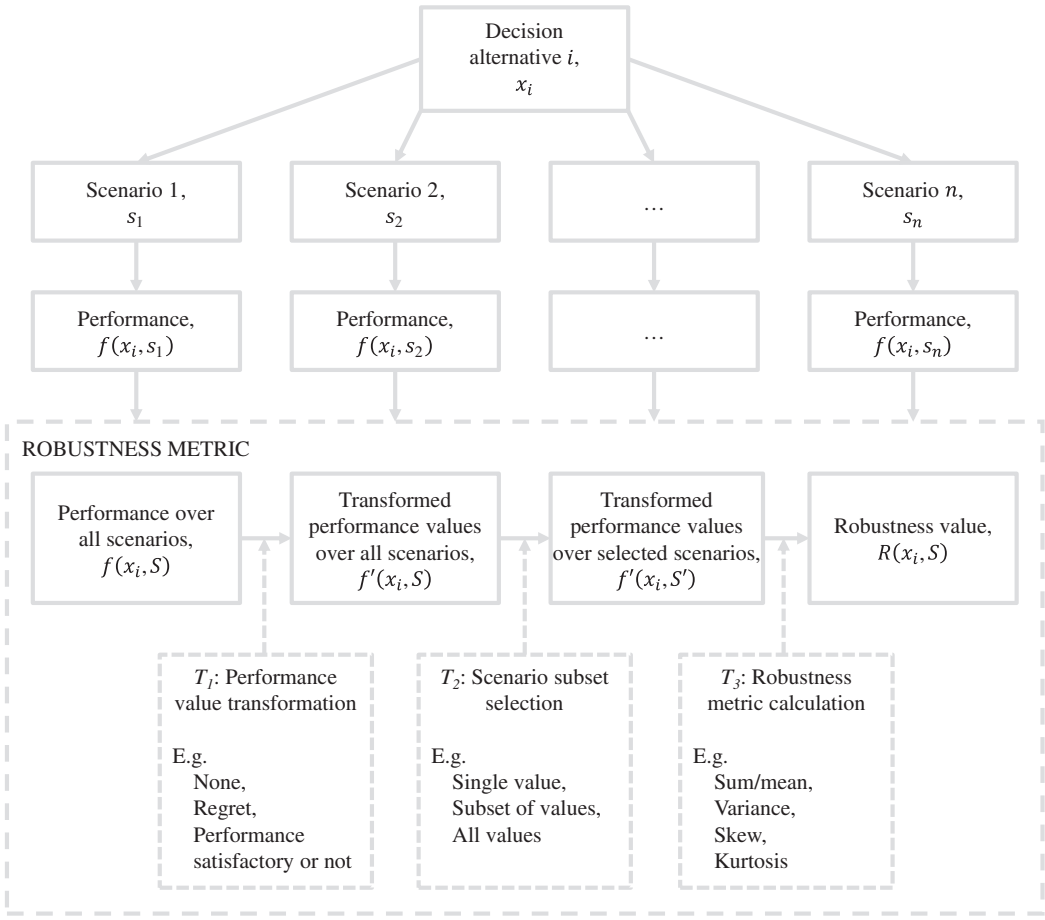

Figure 2. Unifying framework of components and transformations in the calculation of commonly used robustness metrics.

between these metrics is which values of the distribution of performance values over the different scenarios $f(x_i, S)$ they use in the robustness calculation (i.e., scenario subset selection (T_2)) and how these values are combined into a single value of R (i.e., robustness metric calculation (T_3)), as discussed in Sections 3.2 and 3.3.

Metrics in the bottom-left quadrant of Table 2 are calculated in a similar manner to those in the bottom-right quadrant, except that they use information about the performance of a decision alternative *relative* to that of other decision alternatives or a benchmark, and therefore generally express robustness in the form of regret or other measures of deviation. Consequently, the resulting values of $f'(x_i, S)$ consist of the differences between the actual performance of a decision alternative (e.g., cost, reliability) and that of another decision alternative or a benchmark. A robust decision alternative is the one that minimizes the maximum regret across scenarios (e.g., Herman et al., 2015). Alternative metrics that are based on the relative performance of decision alternatives use some type of baseline performance for a given scenario instead of the performance of the best decision alternative (Herman et al., 2015; Kasprzyk et al., 2013; Kwakkel et al., 2016b; Lempert & Collins, 2007; Popper et al., 2009).

Metrics in the top right quadrant of Table 2 measure robustness relative to a threshold or constraint in order to determine whether a decision alternative performs satisfactorily under different scenarios, and are commonly referred to as satisficing metrics. These metrics build on the work of Simon (1956), who pointed out that decision makers often look for a decision that meets one or more requirements (i.e., performance constraints) under a range of scenarios, rather than determining optimal system performance. Therefore, values of $f'(x_i, S)$ consist of information on the scenarios for which the decision alternatives under consideration meet a minimum performance threshold and the larger the number

Table 1.
A Summary of the Three Transformations that are Used by Each Robustness Metric Considered in This Article

Metric	Original reference	T_1: Performance value transformation	T_2: Scenario subset selection	T_3: Robustness metric calculation
Maximin	Wald (1950)	Identity	Worst-case	Identity
Maximax	Wald (1950)	Identity	Best-case	Identity
Hurwicz optimism-pessimism rule	Hurwicz (1953)	Identity	Worst- and best-cases	Weighted mean
Laplace's principle of insufficient reason	Laplace and Simon (1951)	Identity	All	Mean
Minimax regret	Savage (1951) and Giuliani and Castelletti (2016)	Regret from best decision alternative	Worst-case	Identity
90th percentile minimax regret	Savage (1951)	Regret from best decision alternative	90th percentile	Identity
Mean-variance	Hamarat et al. (2014)	Identity	All	Mean-variance
Undesirable deviations	Kwakkel et al. (2016b)	Regret from median performance	Worst-half	Sum
Percentile-based skewness	Voudouris et al. (2014) and Kwakkel et al. (2016b)[a]	Identity	10th, 50th, and 90th percentiles	Skew
Percentile-based peakedness	Voudouris et al. (2014) and Kwakkel et al. (2016b)[a]	Identity	10th, 25th, 75th and 90th percentiles	Kurtosis
Starr's domain criterion	Starr (1963) and Schneller and Sphicas (1983)	Satisfaction of constraints	All	Mean

[a]Kwakkel et al. (2016b) adapted some metrics from Voudouris et al. (2014).

of these scenarios, the more robust a decision alternative. A well-known example of this is the domain criterion, which focuses on the volume of the total space of plausible futures where a given performance threshold is met; the larger this space, the more robust the decision alternative. Often, this is simplified to looking at the fraction of scenarios where the performance threshold is met (e.g., Beh et al. 2015a; Herman et al., 2015; Culley et al., 2016), as scenarios provide a discrete representation of the space of plausible futures.

Satisficing metrics can also be based on the idea of a radius of stability, which has made a recent resurgence under the label of info-gap decision theory (Ben-Haim, 2004; Herman et al., 2015). Here, one identifies the uncertainty horizon over which a given decision alternative performs satisfactorily. The uncertainty horizon α is the distance from a pre-specified reference scenario to the first scenario in which the pre-specified performance threshold is no longer met (Hall et al., 2012; Korteling et al., 2012). However, as these metrics are based on deviations from an expected future scenario, they only assess robustness locally and are therefore not suited to dealing with deep uncertainty (Maier et al., 2016). These metrics also assume that the uncertainty increases at the same rate for all uncertain factors when calculating the uncertainty horizon on a set of axes. Consequently, they are shown in parentheses in Table 2 and will not be considered further in this article.

Metrics in the top-left quadrant of Table 2 base robustness calculation on relative performance values and indicate whether these values result in satisfactory system performance or not. Methods belonging to this category are generally based on the concept of stability. However, in contrast to the stability-based methods in the top-right quadrant of Table 2, these methods assess stability of a decision alternative *relative*

to that of another by identifying crossover points (Guillaume et al., 2016) at which the performance of one decision alternative becomes preferable to that of another and identifying the regions of the scenario space in which a given decision alternative is preferred over another. Methods belonging to this category include the management option rank equivalence (MORE) (Ravalico et al., 2010) and Pareto optimal management option rank equivalence (POMORE) (Ravalico et al., 2009) methods, as well as decision scaling (Brown et al., 2012; Poff et al., 2015). However, as these methods do not quantify robustness explicitly, they are shown in parentheses in Table 2 and will not be considered further in this article.

3.2. Transformation 2 (T_2): Scenario Subset Selection

A categorization of different robustness metrics in accordance with the scenario subset selection transformation (T_2) is given in Table 3. As can be seen, the categorization is based on whether all or a subset of the values of $f'(x_i, S)$ are used in the calculation of the robustness metric. If a subset of values is used, this can consist of a single value or a number of values. As shown in Table 3, Laplace's principle of insufficient reason, the mean-variance metric and Starr's domain criterion use the full set of scenarios S and thus $S' = S$. In contrast, the maximin, maximax, minimax regret and 90th percentile minimax regret metrics only use a single value from S to form S'. The metrics that use a number of selected scenarios S' in the calculation of R include Hurwicz's optimism-pessimism rule, undesirable deviations, percentile-based skewness and percentile-based peakedness.

Which scenarios from S are selected has a significant influence on the relative level of inherent risk aversion of a robustness metric, as shown in Figure 3. For example, the maximax metric has a very low inherent level of risk aversion, as its calculation is only based on the best performance over all scenarios considered (Table 3). In contrast, the maximin metric has a very high level of intrinsic risk aversion, as its calculation is only based on the worst performance over all scenarios considered (Table 3), leading to a very conservative solution (Bertsimas & Sim, 2004). Similarly, the minimax regret metric assumes that the selected decision alternative will have the largest regret possible, as its calculation is based on the worst-case relative performance (Table 3). The other metrics fit somewhere in-between these extremes of low and high levels of intrinsic risk aversion, as shown in Figure 3 and explained below.

Calculation of the metrics in the middle of Figure 3 is based on S' that covers all regions of S, thereby providing a balanced perspective, corresponding to neither a low or high level of intrinsic risk aversion. Some of these metrics use all scenarios (S), such as Laplace's principle of insufficient reason and the mean-variance metric, whereas others are based on a subset of percentiles S' that sample the distribution of S in a balanced way, such as percentile-based skewness, which uses the 10th, 50th and 90th percentiles,

Table 2.
Classification of Robustness Metrics Based on the Performance Value Transformation (T_1)

	Robustness calculation based on *relative* performance values	Robustness calculation based on *absolute* performance values
Indication of whether *system performance is satisfactory* or not	• (Management option rank equivalence (MORE)) • (Pareto optimal MORE (POMORE)[b]) • (Decision Scaling[b])	• Starr's domain criterion • (Info Gap [a])
Indication of *actual system performance*	• Minimax regret • 90th percentile minimax regret • Undesirable deviations	• Maximin (minimax) • Maximax • Hurwicz's optimism-pessimism rule • Laplace's principle of insufficient reason • Mean-variance • Percentile-based skewness • Percentile-based peakedness

Note that brackets around a metric indicate that the metrics is considered unsuitable and is not considered in the following analysis.
[a]Robustness calculated explicitly, but based on deviations from an expected scenario.
[b]Robustness not calculated explicitly.

Table 3.
Classification of Robustness Metrics in Terms of Scenario Subset Selection (T_2)

	Scenarios from S used to form the subset S'		
	Subset		
Robustness metric	Single	Number	All
Maximin	Worst-case		
Maximax	Best-case		
Hurwicz optimism-pessimism rule		Best- and worst-case	
Laplace's principle of insufficient reason			All
Minimax regret	Worst-case		
90th percentile minimax regret	90th percentile		
Mean-variance			All
Undesirable deviations		All performance values worse than the 50th percentile	
Percentile-based skewness		10th, 50th and 90th percentiles	
Percentile-based peakedness		10th, 25th, 75th and 90th percentiles	
Starr's domain criterion			All

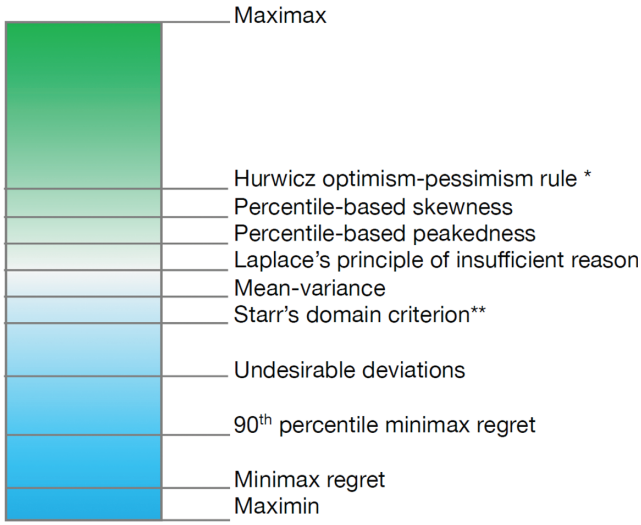

Figure 3. Classification of robustness metrics in terms of relative level of risk aversion from a low level of risk aversion (green) to highly risk averse (blue). *Hurwicz optimism-pessimism rule is a weighted average between the minimax and maximax metrics, where the weighting is chosen by the decision-maker (see Section 3.3). Hence this metric could be placed anywhere on the scale. **As Starr's domain criterion is based on a user-selected threshold, which scenarios are considered in the robustness calculation is variable (see Table 2). Consequently, this metric could be placed anywhere on the scale. It should be noted that the relative level of risk aversion is subjective and is included for illustrative purposes only.

and percentile-based peakedness, which uses the 10th, 25th, 75th, and 90th percentiles (Table 3). Intuitively, Hurwicz's optimism-pessimism rule should also belong to this category, as it utilizes both the best and worst values of $f(x_i, S)$. However, as these values are weighted in the calculation of R using user-defined values (see Section 3.3), the resulting robustness values can correspond to either low to high levels of intrinsic risk aversion, depending on the selected weightings, as indicated in Figure 3. Similarly, robustness values obtained using Starr's domain criterion could range from low to high, depending on the value of the user-selected minimum performance threshold. For example, if this threshold corresponds to a very high level of performance, the resultant robustness value will correspond to a very high level of intrinsic risk aversion and vice versa.

The undesirable deviations and 90th percentile minimax metrics also use a subset S', however, these scenarios do not cover all regions of this S and are therefore less balanced. The undesirable deviations metric considers regret from the median for scenarios for which values of $f(x_i, S)$ are less than the median, resulting in robustness values that have a higher level of intrinsic risk aversion than those obtained using metrics that used information from all regions of

Table 4.
Robustness Metric Calculation (T_3) Used to Transform the Sampled Performance Information into the Value of Robustness

	Robustness metric calculation						
Robustness metric	None	Sum	Mean	Weighted mean	Variance	Skew	Kurtosis
Maximin	√						
Maximax	√						
Hurwicz optimism-pessimism rule				√			
Laplace's principle of insufficient reason			√				
Minimax regret	√						
90th percentile minimax regret	√						
Mean-variance			√		√		
Undesirable deviations		√					
Percentile-based skewness						√	
Percentile-based peakedness							√
Starr's domain criterion							

the distribution (Table 3). The 90th percentile minimax regret metric corresponds to an even greater level of intrinsic risk aversion, as it is based on a single value that is close to the worst case (90th percentile—see Table 3).

3.3. Transformation 3 (T_3): Robustness Metric Calculation

A categorization of different robustness metrics in accordance with the final robustness metric calculation (T_3) is given in Table 4. As can be seen, for some metrics, such as the maximin, maximax, minimax regret and 90th percentile minimax regret metrics, $f'(x_i, S')$ and $R(x_i, S)$ are identical (i.e., the robustness metric calculation corresponds to the identity transformation). This is because for these metrics, S' consists of a single scenario and there is no need to combine a number of values in order to arrive at a single value of robustness. However, for the remaining metrics, for which S' contains at least two values, some sort of transformation is required. Metrics that are based on the mean or sum of $f'(x_i, S')$, such as Laplace's principle of insufficient reason, mean-variance and undesirable deviations, effectively assign an equal weighting to different scenarios and then suggest that the best decision is the one with the best mean performance, producing an expected value of performance. In contrast, in Hurwicz's optimism-pessimism rule, the user can select the relative weighting of the two scenarios (low and high levels of risk aversion) considered, as mentioned in Section 3.2.

Alternatively, some metrics consider aspects of the variability of $f'(x_i, S')$. For example, the mean-variance metric attempts to balance the mean and variability of the performance of a decision alternative over different scenarios. However, a disadvantage of considering a combination of the mean and variance is that the resultant metric is not always monotonically increasing (Ray et al., 2013). Moreover, when considering variance, good and bad deviations from the mean are treated equally (Takriti & Ahmed, 2004). The undesirable deviations metric overcomes this limitation, while still providing a measure of variability. Other metrics are focused on different attributes of $f'(x_i, S')$, such as the skewness and kurtosis.

3.4. Summary of Categorization of Robustness Metrics

The complete categorization of the commonly used robustness metrics considered in this article in accordance with the three transformations (performance value transformation (T_1) (Table 2), scenario subset selection (T_2) (Table 3) and robustness metric calculation (T_3) (Table 4)), as well as the relative level of risk aversion that is associated with T_2 (Figure 3), is given in Table 5. It is hoped that this can provide some guidance to decision-makers in relation to which robustness metric is appropriate for their decision context.

In relation to the performance value transformation (T_1), which robustness metric is most appropriate depends on whether the performance value in question relates to the satisfaction of a system constraint or not, and is therefore a function of the properties of the system under consideration. For example, if the

AGU Earth's Future

Table 5.
Summary of Categorizations of Commonly Used Robustness Metrics in Accordance with Performance Value Transformation, Scenario Subset Selection, Calculation of the Robustness Metric, and the Relative Level of Risk Aversion. See the Supporting Information for equations

	T_1: Performance value transformation				T_2: Scenario subset selection				T_3: Robustness metric calculation				
	Optimize system performance		Satisfy constraints										
Robustness metric	Absolute values (no transform)	Relative values	Absolute values (performance meets constraints)	Relative values	Single value	Subset of values	All values	Low (☆) to high (☆☆☆☆☆) level of risk aversion	None, sum, or mean	Weighted mean	Variance	Skew	Kurtosis (peakedness)
Maximin	√				√			☆☆☆☆☆	√				
Maximax	√				√			☆	√				
Hurwicz optimism-pessimism rule	√					2		☆ to ☆☆☆☆☆[b]		√			
Laplace's principle of insufficient reason	√						√	☆☆☆	√				
Minimax regret		√			√			☆☆☆☆☆	√				
90th percentile minimax regret		√			√			☆☆☆☆	√				
Mean-variance	√						√	☆☆☆	√		√		
Undesirable deviations		√				V[a]		☆☆☆☆	√				
Percentile-based skewness	√					3		☆☆☆				√	
Percentile-based peakedness	√					4		☆☆☆					√
Starr's domain criterion			√			V[a]		☆ to ☆☆☆☆☆[c]	√				

[a]V = variable.
[b]Hurwicz optimism-pessimism rule has a parameter (selected by the decision-maker) to determine the relative level of risk aversion.
[c]This is dependent on the minimum performance threshold selected by the decision-maker.

10.1002/2017EF000649

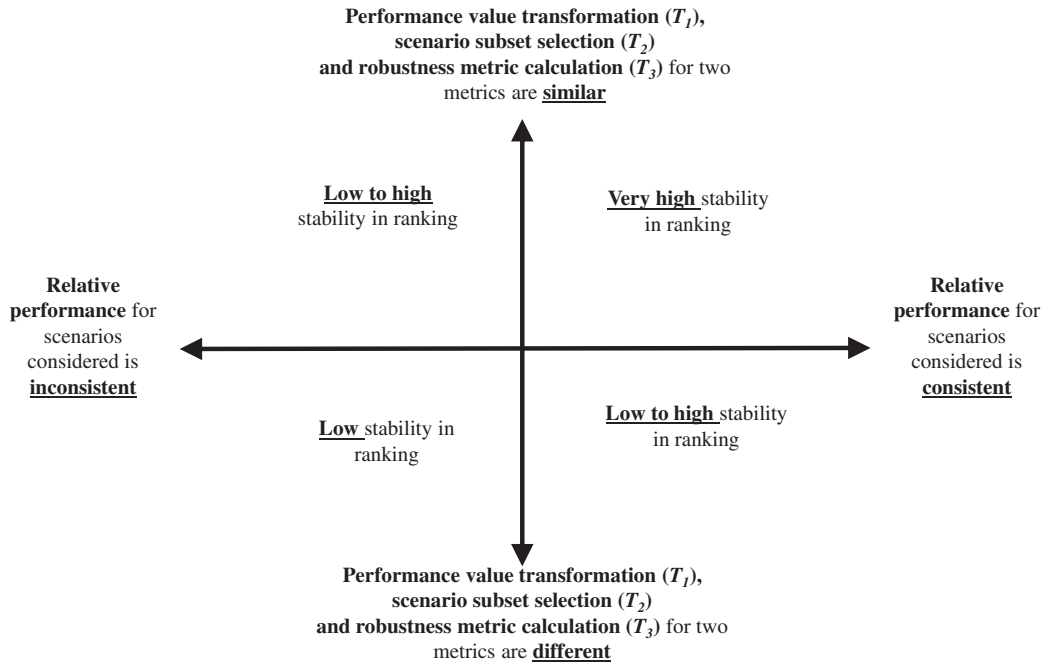

Figure 4. Conceptual representation of conditions affecting ranking stability. A high stability in ranking indicates that two metrics will rank the decision alternatives the same, whereas a low stability indicates that two metrics will rank the decision alternatives differently.

system is concerned with supplying water to a city, there is generally a hard constraint in terms of supply having to meet or exceeding demand, so that the city does not run out of water (Beh et al., 2017). The system performs satisfactorily if this demand is met and that is the primary concern of the decision-maker. Alternatively, there might be a fixed budget for stream restoration activities, which also provides a constraint. In this case, a solution alternative performs satisfactorily if its cost does not exceed the budget. For the above examples, where performance values correspond to determining whether constraints have been met or not, satisficing metrics, such as Starr's domain criterion, are most appropriate.

In contrast, if the performance value in question relates to optimizing system performance, metrics that use the identity or regret transforms would be most suitable. For example, for the water supply security case mentioned above, the objective might be to identify the cheapest solution alternative that enables supply to satisfy demand. However, there might also be concern in over-investment in expensive water supply infrastructure that is not needed, in which case robustness metrics that apply a regret transformation might be most appropriate, as this would enable the degree of over-investment to be minimized when applied to the cost performance value. For the stream restoration example, however, decision-makers might simply be interested in maximizing ecological response for the given budget. In this case, robustness metrics that use the identity transform might be most appropriate when considering performance values related to ecological response.

In relation to scenario subset selection (T_2), which robustness metric is most appropriate depends on a combination of the likely impact of system failure and the degree of risk aversion of the decision-maker. In general, if the consequences of system failure are more severe, the degree of risk-aversion adopted would be higher, resulting in the selection of robustness metrics that consider scenarios that are likely to have a more deleterious impact on system performance. For example, in the water supply security case, it is likely that robustness metrics that consider more extreme scenarios would be considered, as a city running out of water would most likely have severe consequences. In contrast, as the potential negative impacts for the stream restoration example are arguably less severe, robustness metrics that use a wider range or less severe scenarios might be considered. However, this also depends on the values and degree of risk aversion of the decision maker.

As far as the robustness value calculation (T_3) goes, this is only applicable to metrics that consider more than one scenario, as discussed previously, and relates to the way performance values over the different

scenarios are summarized. For example, if there is interest in the average performance of the system under consideration over the different scenarios selected in T_2, such as the average cost for the water supply security example or the average ecological response for the stream restoration example, a robustness metric that sums or calculates the mean of these values should be considered. However, decision-makers might also be interested in (1) the variability of system performance (e.g., cost, ecological response) over the selected scenarios, in which case robustness metrics based on variance should be used, (2) the degree to which the relative performance of different decision alternatives is different under more extreme scenarios, in which case robustness metrics based on skewness should be used, and/or (3) the degree of consistency in the performance of different decision alternatives over the scenarios considered, in which case robustness metrics based on kurtosis should be used.

4. When Do Robustness Metrics Disagree?

As mentioned previously, robustness metrics have been shown to disagree in certain cases (Giuliani & Castelletti, 2016; Herman et al., 2015; Kwakkel et al., 2016b). As these metrics are used to make decisions on outcomes, it is important to obtain greater insight into the conditions under which different robustness metrics result in different decisions. It is important to note that the relative ranking of two decision alternatives (x_1 and x_2), when assessed using two robustness metrics (R_a and R_b), will be the same, or stable, if the following three conditions hold:

$$R_a\left(x_1\right) > R_a\left(x_2\right) \text{ and } R_b\left(x_1\right) > R_b\left(x_2\right), \tag{1}$$

$$\text{or } R_a\left(x_1\right) < R_a\left(x_2\right) \text{ and } R_b\left(x_1\right) < R_b\left(x_2\right), \tag{2}$$

$$\text{or } R_a\left(x_1\right) = R_a\left(x_2\right) \text{ and } R_b\left(x_1\right) = R_b\left(x_2\right) \tag{3}$$

The relative rankings will be different or "flipped" if the following two conditions hold:

$$R_a\left(x_1\right) > R_a\left(x_2\right) \text{ and } R_b\left(x_1\right) < R_b\left(x_2\right), \tag{4}$$

$$\text{or } R_a\left(x_1\right) < R_a\left(x_2\right) \text{ and } R_b\left(x_1\right) > R_b\left(x_2\right). \tag{5}$$

Consequently, relative differences in robustness values obtained when different robustness metrics are used are a function of (1) the differences in the transformations (i.e., performance value transformation (T_1), scenario subset selection (T_2), robustness metric calculation (T_3)) used in the calculation of R_a and R_b and (2) differences in the relative performance of decision alternatives x_1 and x_2 over the different scenarios considered. In general, ranking stability is greater if there is greater similarity in the three transformations for R_a and R_b and if there is greater consistency in the relative performance of x_1 and x_2 for the scenarios considered in the calculation of R_a and R_b, as shown in the conceptual representation in Figure 4. In fact, if the relative performance of two decision alternatives is the same under all scenarios, the relative ranking of these decision alternatives is stable, irrespective of which robustness metric is used.

4.1. Similar Transformations and Consistent Relative Performance

If the transformations used in the calculation of the robustness metrics are similar and the performance of the two decision alternatives considered is consistent across the scenarios, one would expect ranking stability to be very high (top-right quadrant, Figure 4). For example, when minimax regret and 90th percentile minimax regret correspond to R_a and R_b, there is a high degree of similarity in the performance value transformation (T_1), scenario subset selection (T_2), and robustness metric calculation (T_3) (y-axis). For both metrics, the performance values are transformed to regret, S' corresponds to a single scenario and there is no need to combine any values as part of the robustness metric calculation (T_3), as there is only a single value of regret (Table 5). Similarly, there is a high degree of consistency in the relative performance values used for the calculation of R_a and R_b (x-axis), as minimax regret uses the worst-case scenario and 90th percentile minimax regret uses a scenario that almost corresponds to the worst case (Table 3). Consequently, one would expect the ranking of decision alternatives to be very stable when these two metrics are used.

4.2. Different Transformations and Inconsistent Relative Performance

Ranking stability is generally low if there are marked differences in the three transformations for R_a and R_b and if there is greater inconsistency in the relative performance of x_1 and x_2 for the scenarios considered in the calculation of R_a and R_b. Consequently, if both of these conditions are met, one would expect ranking stability to be low (bottom-left quadrant, Figure 4). For example, when R_a and R_b correspond to minimax regret and percentile based peakedness, there is a high degree of difference in performance value transformation (T_1), scenario subset selection (T_2) and robustness metric calculation (T_3) (y-axis). For the former, performance values are transformed to regret, S' consists of one scenario (worst-case scenario) and there is no need to combine any values as part of the robustness metric calculation (T_3). For the latter, the actual performance values are used, S' consists of four scenarios (10th, 25th, 75th, and 90th percentiles) and the robustness metric calculation is the kurtosis of the four regret values (see Tables 3 and 5). Similarly, there is a potentially high degree of inconsistency in the relative performance values used for the calculation of R_a and R_b (x-axis), as minimax regret uses the worst-case scenario, whereas percentile-based peakedness uses four scenarios spread evenly across the distribution of S (Table 3). Consequently, one would expect the ranking of decision alternatives to be generally unstable when these two metrics are used.

4.3. Different Transformations and Consistent Relative Performance

In cases where there are marked differences in the three transformations for R_a and R_b but consistency in the relative performance of x_1 and x_2 over the scenarios considered in the calculation of R_a and R_b (bottom-right quadrant, Figure 4), ranking stability can range from high to low. This is because the interactions between various drivers of ranking stability are complex and difficult to predict a priori. For example, when maximax and maximin correspond to R_a and R_b, there is a high degree of similarity in the three transformations (y-axis). For both metrics, the actual performance values are used (T_1 is the identity transform), S' corresponds to a single scenario and there is no need to combine any values as part of the robustness metric calculation (T_3), as there is only a single value of performance (Table 5). However, there is a potentially low degree of consistency in the relative performance values used in the robustness calculations (x-axis), as the single performance values used in the calculations of these two robustness metrics come from different ends of the distribution of performance values (i.e., one corresponds to the best-case and one to the worst-case). Consequently, this case belongs to the top-left quadrant in Figure 4, where ranking stability can vary from low to high, depending on the consistency in relative performance of x_1 and x_2 for the best- and worst-case scenarios.

4.4. Similar Transformations and Inconsistent Relative Performance

In cases where the three transformations for R_a and R_b are similar but the relative performance of x_1 and x_2 is inconsistent over the scenarios considered in the calculation of R_a and R_b (top-left quadrant, Figure 4), ranking stability can also range from high to low due to the complex interactions between the different drivers affecting ranking stability. For example, when Laplace's principle of insufficient reason and percentile based skewness correspond to R_a and R_b, there is a moderate degree of difference in the three transformations (y-axis). Both use actual performance values, but the former uses values from all scenarios and averages them, whereas the latter uses the 10th, 50th, and 90th percentiles and calculates their skewness (see Tables 3 and 5). However, as both use values from similar regions of the performance distribution, it is likely that there is a high degree of consistency in the relative performance values used in the robustness calculation (x-axis). Consequently, this case belongs to the bottom-right quadrant in Figure 4, where ranking stability can vary from low to high, depending on the relative impact of using the average and skewness of performance values for the robustness metric calculation (T_3).

5. Case Studies

Three case studies with different properties are used to test the conceptual model of ranking stability introduced in Section 4, as shown in Table 6. As can be seen, the case studies represent water supply systems and flood prevention systems, with decision variables including changes to existing infrastructure, construction of new infrastructure, and changes to operational rules or policies. The number of scenarios

Table 6.
Summary of the Characteristics of the Southern Adelaide, Lake Como and Waas Case Studies

Name	Location	Decision variables, components of x_i	Selected objectives and performance metrics, $f(x_i, S)$	Number of scenarios, n, where $S = \{s_1, \ldots, s_n\}$	Number of decision alternatives, m where $X = \{x_1, \ldots, x_m\}$
Southern Adelaide water supply system	Adelaide, Australia	Construction of new water supply infrastructure (e.g., desalination plants, rainwater tanks, stormwater harvesting) and time of implementation	Reliability (water supply)	125	72
Lake Como	Como, Italy	Parameterization of policies to determine releases based on day of year, current lake storage and previous day inflow.	Reliability (flood prevention) Reliability (water supply)	28	19
Waas	Rhine delta, The Netherlands (hypothetical model based on the real River Waal)	Changes to existing infrastructure for flood reduction and flood damage reduction, and changes to operations (e.g., limits to upstream maximum discharge).	Flood damage Casualties	3000	11

varies greatly in each case study (28–3000), as does the number of optimal decision alternatives considered (11–72).

5.1. Southern Adelaide

This urban water supply augmentation case study models the southern region of the Adelaide water supply system, as it existed in 2010 (Beh et al., 2014, 2015a, 2015b, 2017; Paton et al., 2013, 2014a, 2014b). Adelaide has a population of approximately 1.3 million people and is the capital city of the state of South Australia. Characterized by a Mediterranean climate and an annual rainfall of between 257 and 882 mm (average of 552 mm) over the period from 1889 to 2010 (Paton et al., 2013), Adelaide is one of the driest capital cities in the world (Wittholz et al., 2008). The southern Adelaide system supplies approximately 50% of the water mains consumption (168 GL in 2008) (Beh et al., 2014).

In 2010, the southern Adelaide system consisted of three reservoirs to supply water, as illustrated in Figure 5: Myponga Reservoir collects water from local catchments; Mt. Bold Reservoir collects water both from local catchments and water pumped from the River Murray via the Murray Bridge—Onkaparinga pipeline; Happy Valley reservoir is a service reservoir storing water that has been transferred from the Mt. Bold Reservoir. Water from the River Murray is limited to a maximum of 650 GL over a 5-year rolling period and it is assumed that half of this water is available to the southern Adelaide system.

Due to projected increases in demand and a changing climate there is a need to augment the water supply system (Paton et al., 2013). In particular, the River Murray will be greatly affected by climate change (Grafton et al., 2016a). This article considers 125 scenarios corresponding to various combinations of representative concentration pathways (RCPs) and global circulation models (GCMs) to project changes for future rainfall for the Adelaide system.

There are a number of options for augmentation including the construction of desalination plants, stormwater harvesting schemes, and household rainwater tanks. A previous study (Beh et al., 2015b) generated 72 optimal decision alternatives for this case study using a multiobjective evolutionary algorithm, which will be used in this article. Greenhouse gas emissions and cost were used as objectives, and the vulnerability and reliability of each decision alternative was used to further analyze each optimal decision alternative.

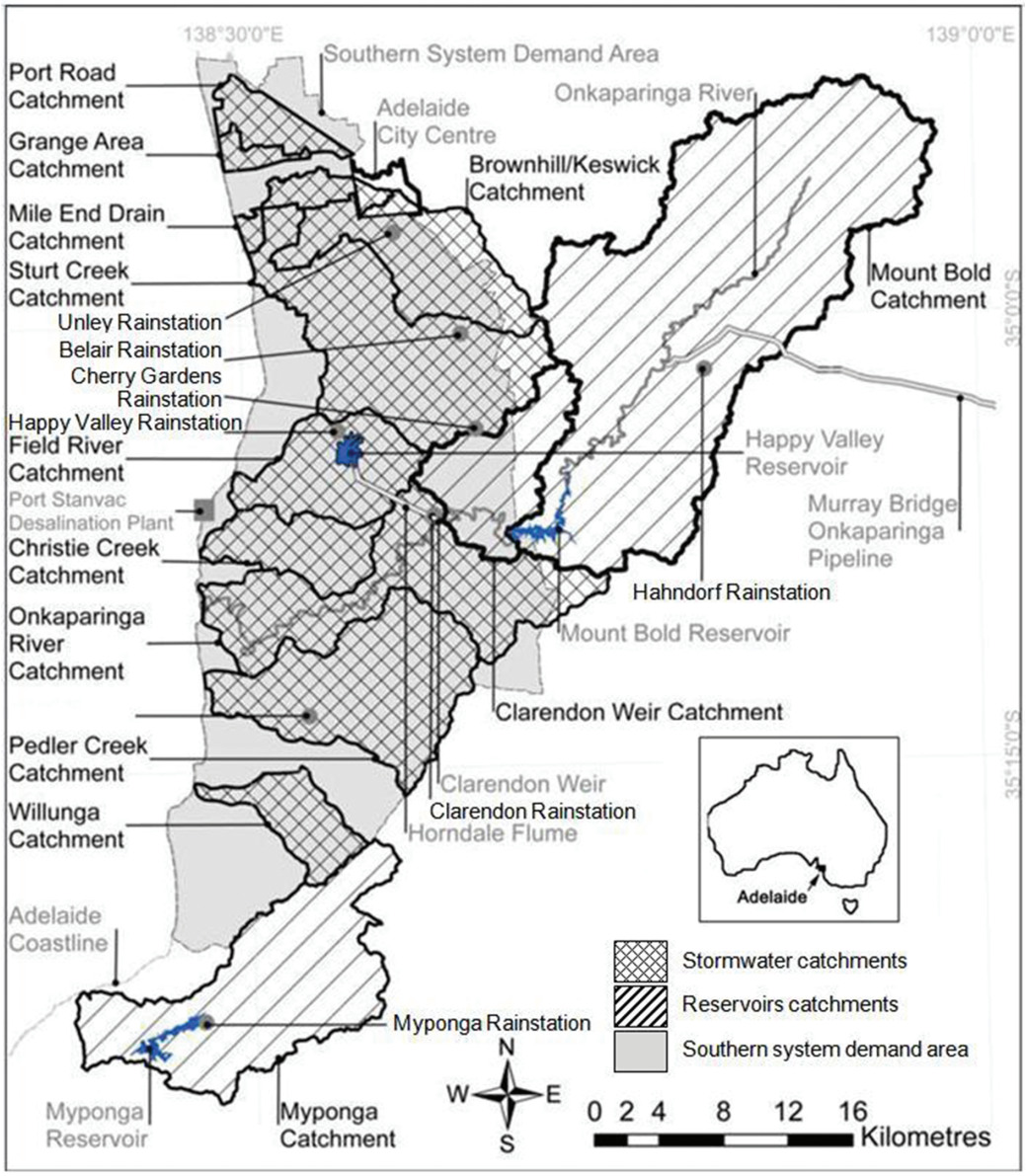

Figure 5. The southern Adelaide water supply system as it existed in 2010.

The reliability of the water supply was calculated over a range of future climate and demand scenarios. Reliability was calculated in the following manner:

$$\text{Reliability} = \frac{T_s}{T} \tag{6}$$

where T_s is the number of years that supply meets demand and T is the total number of years in the planning horizon. A higher reliability implies that the supply meets demand in more years and hence a higher reliability is more desirable than a lower reliability.

5.2. Lake Como

Lake Como is the third largest Italian lake with a total volume of 23.4 km^3. The lake is fed by a 4552 km^2 watershed (see Figure 6) characterized by a mixed snow-rain dominated hydrological regime with relatively dry winters and summers, and higher flows in spring and autumn due to snow-melt and precipitation, respectively. The lake releases are controlled since 1946 with the twofold purpose of flood protection along

the lake shores, particularly in the city of Como, and water supply to the downstream users, including eight run-of-the-river hydropower plants and a dense network of irrigation canals, which distribute the water to four agricultural districts with a total surface of 1400 km^2 mostly cultivated with maize (Giuliani et al., 2016a; Guariso et al., 1985, 1986).

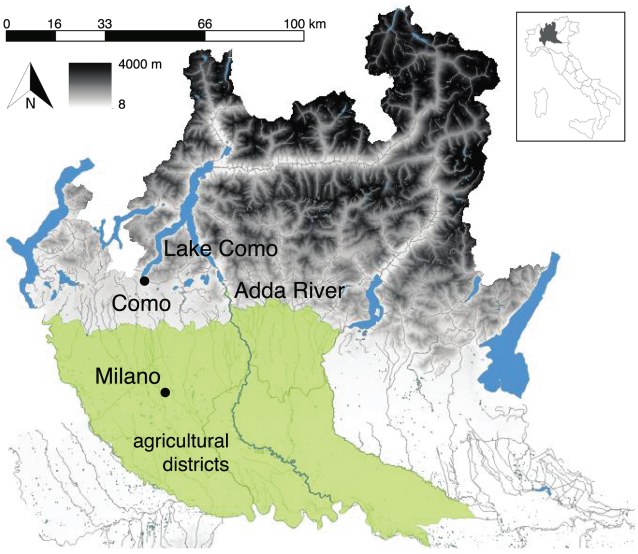

Figure 6. Map of the Lake Como system.

To satisfy the summer water demand peak, the current regulation operates the lake to store a large fraction of the snowmelt in order to be, approximately, at full capacity between June and July (Denaro et al., 2017). The projected anticipation of the snow melt caused by increasing temperature, coupled with the predicted decrease of water availability in the summer period, would require storing additional water and for longer periods, ultimately increasing the flood risk. The optimal flood protection would be instead obtained by drawing down the lake level as much as possible (Giuliani & Castelletti, 2016).

Due to a changing climate and thus a changing flood risk (Giuliani & Castelletti, 2016; McDowell et al., 2014) and availability of water (Iglesias & Garrote, 2015), a climate ensemble of 28 scenarios was used for analysis by Giuliani and Castelletti (2016) and in the following analysis. These scenarios are combinations of different RCPs, and Global, and Regional Climate Models. The resulting trajectories of temperature and precipitation are then statistically downscaled by means of quantile mapping and used as inputs to a hydrological model to generate projections of the Lake Como inflows over the time-period 2096–2100.

There are two primary conflicting operating strategies: maximizing water availability versus reducing flood risk. Consistent with previous works (Castelletti et al., 2010; Culley et al., 2016; Giuliani et al., 2016d; Giuliani & Castelletti, 2016), the trade-offs between these two strategies are modeled using the following two objectives:

- *Flooding*: the storage reliability (to be maximized), defined as

$$\text{st\_rel} = 1 - \frac{n_F}{H} \tag{7}$$

where n_F is the number of days during which the lake level is higher than the flooding threshold of 1.24 m and H is the evaluation horizon.

- *Irrigation*: the daily average volumetric reliability (to be maximized), defined as

$$\text{vol\_rel} = \frac{1}{H}\sum_{t=1}^{H}\frac{Y_t}{D_t} \tag{8}$$

where Y_t is the daily water supply and D_t the corresponding water demand.

A previous study (Giuliani & Castelletti, 2016) generated 19 Pareto optimal decision alternatives by optimizing the flooding and irrigation objectives over historical climate conditions via evolutionary multiobjective direct policy search, a simulation-based optimization approach that combines direct policy search, nonlinear approximating networks, and multiobjective evolutionary algorithms (Giuliani et al., 2016b). These optimal reservoir operation policies are used in the following analysis.

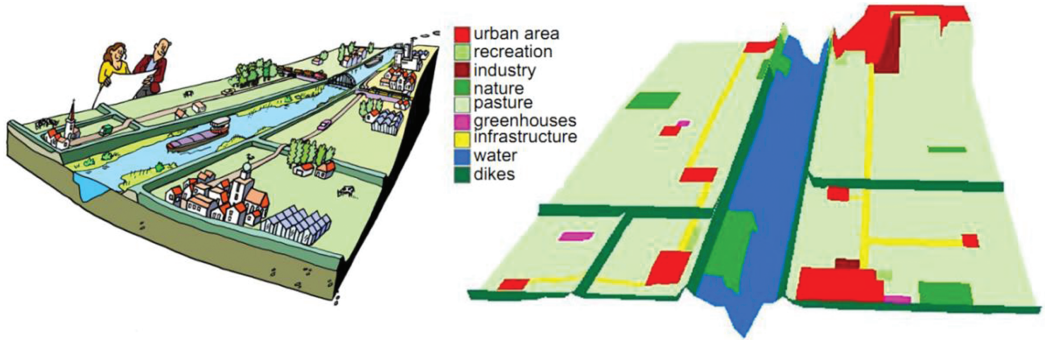

Figure 7. The Waas case study area (left) is heavily schematized (right) into a three-dimensional image of the floodplain presenting the land use and elevations (exaggerated vertically). The flow direction is from back to front (Haasnoot et al., 2012).

5.3. Waas

The Waas case study is a hypothetical case, based on a river reach in the Rhine delta of the Netherlands (the river Waal). An Integrated Assessment Meta Model is used (Haasnoot et al., 2012), which is theory motivated (Haasnoot et al., 2014) and has been derived from more detailed, validated models of the Waal area. The river and floodplain are highly schematized, but have realistic characteristics (see Figure 7), with the river being bound by embankments and the floodplain composed of five dike rings. In the southeast, a large city is situated on higher ground, while smaller villages exist in the remaining area. Other forms of land use include greenhouses, industry, conservation areas, and pastures. In the recent past, two large flood events occurred in the Waal area, on which this hypothetical case study is based, resulting in considerable damage to houses and agriculture (Haasnoot et al., 2009). In the future, changes in land use and climate, as well as socioeconomic developments, may further increase the risk of damage, so action is needed.

There is a wide range of uncertainties that are considered, including climate change and its impact on river discharge (see Haasnoot et al. (2012) for details) and land use change through seven transient land use scenarios. Uncertainty with respect to the fragility of dikes and economic damage functions is taken into account by putting a bandwidth of plus and minus 10% around the default values. Finally, some aspects of policy uncertainty are included both through the uncertainty of the fragility function and by letting the impact of the action vary (Kwakkel et al., 2015). These drivers of change are combined to form a total of 3000 scenarios.

Damage due to the flooding of dike rings is calculated from water depth and damage relations (De Bruijn, 2008; Haasnoot et al., 2009). Using these relations, the model calculates the flood impacts per hectare for each land use to obtain the total damage for sectors such as agriculture, industry, and housing. Casualties are assessed using water depth, land use, and flood alarms triggered by the probability of dike failure. These performance measures form the three objectives that are considered in the original studies (Kwakkel et al., 2015, 2016a): costs, loss of life, and economic damages. However, due to the fact that the costs were rarely effected by the scenario, this objective was not included in this study. In previous studies, a many-objective robust optimization approach was used to design robust adaptation pathways (Kwakkel et al., 2015, 2016a) and 11 distinct adaptation pathways were identified. These optimal adaptation pathways are used in the following analysis.

6. Results and Discussion

To assess if the rankings of decision alternatives are likely to be similar between two metrics for the different case studies and objectives considered, the percentage of pairs of decision alternatives where the ranking is stable is used. A stable pair of decision alternatives is one where one of these decision alternatives is always ranked higher than another, regardless of the robustness metric used, as described in Section 4. The ranking stability for each pair of metrics is displayed in Figure 8. A ranking stability of 100% indicates that the metrics agreed on the rankings for every pair of decision alternatives, while 0% indicates that one metric ranked the decision alternatives in reverse to the other metric. The robustness values for each case study are included in Supporting Information S1. Figure 8 also provides basic information about the three

transformations used in the calculation of each robustness metric in an effort to assess how well the results agree with the conceptual model presented in Figure 4.

6.1. Impact of Transformations

Figure 8 indicates that the pairs of metrics with high stability (lower portion of the figure, shaded mostly green), tend to share the same robustness metric calculation transformation (T_3). For example, in cases where both metrics use the identity transformation, sums or averages of $f'(x_i, S)$ (all indicated by "M" in the T_3 columns), rankings are generally stable. In contrast, the metrics with low stability (upper portion of Figure 8, shaded mostly red and yellow) tend to have different robustness metric calculation transformations. An example is the percentile-based peakedness metric, being the only metric to use kurtosis. Every other metric uses a different robustness metric calculation transformation and hence when percentile-based peakedness is used as one of the two robustness metrics considered, rankings are generally unstable. This can be explained by the fact that when different types of calculations from $f'(x_i, S)$ to $R(x_i, S)$ are used, different attributes of the distribution of $f'(x_i, S)$ result in "similarity," as discussed in Section 4. For example, as can be seen in Figure 4, two metrics that use different robustness metric calculation transformations (T_3) will result in low stability unless there are consistent differences between two decision alternatives over the different scenarios.

In general, a pair of metrics with the same robustness metric calculation transformation (T_3) almost always has high ranking stability, while a pair with a different T_3 almost always has low ranking stability. However, Figure 8 indicates the same is not always true of the other two transformations (i.e., performance value transformation (T_1) and scenario subset selection (T_2)), although in some cases, they can have an impact. For example, the maximax and maximin metrics share the same robustness metric calculation transformation (T_3). However, their ranking stability is markedly lower than that for other metrics that share the same T_3, particularly for the Adelaide and Lake Como case studies. In this case, the primary cause of ranking stability is associated with scenario subset selection (T_2). The selected scenarios S' for the maximin and maximax criteria correspond to different extremes of the distribution of S and hence these two metrics show high levels of disagreement. This puts the comparison of these two metrics in the middle or lower region of Figure 4 and explains the large variance in the ranking stability of the maximin and maximax metrics in Figure 8. This variance in ranking stability is particularly clear when there is not a large consistent difference in performance between decision alternatives. The maximax metric is also different from most other metrics, although to a lesser extent than the difference with the maximin metric, and it can be seen in Figure 8 that this results in variable levels of agreement between the maximax metric and the other metrics in each case study.

Similarly, the undesirable deviations metric uses the sum of $f'(x_i, S)$ and is hence categorized with many other metrics when considering the robustness metric calculation transformation (T_3). Like the maximin and maximax comparison, the undesirable deviations metric shows varying ranking stability depending on the case study. The complex effects of the performance value transformation (T_1) explain this. Regret of a decision alternative in each scenario is used by the undesirable deviations metric, compared to most metrics, which use the actual performance values. This calculation of regret is also different from that of the other regret metrics (minimax regret and 90th percentile minimax regret) because it is considering regret relative to the median performance of that decision alternative, rather than regret relative to the absolute best performance across all decision alternatives.

A relatively low level of agreement is seen when comparing the maximax and undesirable deviations (Figure 8). Similar to the above discussion, this variability is due to the different sampling methods for the scenario subset selection (T_2) and different performance value transformations (T_1). Maximax samples a single value from the left-hand side of the distribution, whereas the undesirable deviations metric samples the 50% of values from the right-hand side of the distribution. In addition, there is also a difference in the initial performance value transformation (T_1), with the maximax metric using the raw performance values, while the undesirable deviations metric uses the regret of a decision alternative relative to the median performance.

6.2. Impact of Relative Performance

As can be seen in Figure 8, although there is generally a high degree of consistency in ranking stability based on the similarity between the three transformations, this does not hold for certain combinations

Metrics		T1		T2		T3		% of times that metrics agree on relative rankings				
								Adelaide	Lake Como		Waas	
1	2	1	2	1	2	1	2	Supply	Flooding	Irrigation	Flood damage	Casualties
Maximax	Percentile-based peakedness	I	I	Si	Su	M	K	11%	12%	42%	40%	56%
Laplace	Percentile-based peakedness	I	I	A	Su	M	K	9%	45%	24%	40%	47%
Mean-variance	Percentile-based peakedness	I	I	A	Su	M+V	K	8%	49%	23%	38%	47%
Maximin	Percentile-based peakedness	I	I	Si	Su	M	K	8%	50%	23%	38%	47%
Minimax regret	Percentile-based peakedness	R	I	Si	Su	M	K	11%	50%	23%	40%	51%
Hurwicz	Percentile-based peakedness	I	I	Su	Su	WM	K	10%	50%	29%	40%	45%
90th percentile minimax regret	Percentile-based peakedness	R	I	Si	Su	M	K	12%	50%	23%	42%	55%
Undesirable deviations	Percentile-based peakedness	I	I	Su	Su	M	K	38%	51%	75%	53%	51%
Percentile-based skewness	Percentile-based peakedness	I	I	Su	Su	S	K	16%	58%	54%	27%	36%
Maximax	Percentile-based skewness	I	I	Si	Su	M	S	69%	18%	65%	80%	65%
Maximin	Percentile-based skewness	I	I	Si	Su	M	S	38%	44%	50%	84%	75%
Hurwicz	Percentile-based skewness	I	I	Su	Su	WM	S	71%	44%	60%	84%	76%
Mean-variance	Percentile-based skewness	I	I	A	Su	M+V	S	73%	45%	50%	85%	75%
Laplace	Percentile-based skewness	I	I	A	Su	M	S	73%	43%	51%	84%	75%
Minimax regret	Percentile-based skewness	R	I	Si	Su	M	S	71%	44%	53%	84%	71%
90th percentile minimax regret	Percentile-based skewness	R	I	Si	Su	M	S	71%	44%	53%	82%	67%
Undesirable deviations	Percentile-based skewness	R	I	Su	Su	M	S	63%	45%	49%	71%	60%
Maximin	Undesirable deviations	I	R	Si	Su	M	M	41%	98%	15%	85%	78%
Laplace	Undesirable deviations	I	R	A	Su	M	M	67%	92%	11%	87%	78%
Mean-variance	Undesirable deviations	I	R	A	Su	M+V	M	67%	96%	10%	85%	78%
Hurwicz	Undesirable deviations	I	R	Su	Su	WM	M	61%	98%	18%	87%	76%
Minimax regret	Undesirable deviations	R	R	Si	Su	M	M	63%	97%	6%	87%	89%
90th percentile minimax regret	Undesirable deviations	R	R	Si	Su	M	M	64%	97%	9%	89%	85%
Maximax	Undesirable deviations	I	R	Si	Su	M	M	60%	48%	22%	84%	60%
Maximax	90th percentile minimax regret	I	R	Si	Si	M	M	91%	49%	75%	95%	75%
Maximax	Mean-variance	I	I	Si	A	M	M+V	88%	50%	72%	95%	82%
Maximax	Minimax regret	I	R	Si	Si	M	M	92%	50%	78%	96%	71%
Maximax	Laplace	I	I	Si	A	M	M	88%	53%	77%	96%	82%
Maximax	Hurwicz	I	I	Si	Su	M	WM	98%	49%	84%	96%	84%
Maximin	Maximax	I	I	Si	Si	M	M	49%	49%	68%	95%	82%
Maximin	Minimax regret	I	R	Si	Si	M	M	46%	98%	90%	98%	89%
Maximin	90th percentile minimax regret	I	R	Si	Si	M	M	44%	98%	91%	96%	93%
Maximin	Laplace	I	I	Si	A	M	M	41%	95%	90%	98%	100%
Maximin	Mean-variance	I	I	Si	A	M	M+V	41%	98%	92%	98%	100%
Maximin	Hurwicz	I	I	Si	Su	M	WM	51%	100%	84%	98%	98%
Hurwicz	90th percentile minimax regret	I	R	Su	Si	WM	M	93%	98%	88%	98%	91%
Hurwicz	Minimax regret	I	R	Su	Si	WM	M	93%	98%	89%	100%	87%
Hurwicz	Mean-variance	I	I	Su	A	WM	M+V	90%	98%	84%	98%	98%
Hurwicz	Laplace	I	I	Su	A	WM	M	90%	95%	89%	100%	98%
Laplace	Minimax regret	I	R	A	Si	M	M	95%	95%	94%	100%	89%
Laplace	90th percentile minimax regret	I	R	A	Si	M	M	96%	95%	94%	98%	93%
Minimax regret	Mean-variance	R	I	Si	A	M	M+V	95%	99%	94%	98%	89%
90th percentile minimax regret	Mean-variance	R	I	Si	A	M	M+V	96%	96%	96%	96%	93%
Minimax regret	90th percentile minimax regret	R	R	Si	Si	M	M	97%	97%	96%	98%	96%
Laplace	Mean-variance	I	I	A	A	M	M+V	100%	96%	95%	98%	100%

Figure 8. Agreement in relative rankings when considering all pairwise combinations of metrics for all case studies. For performance value transformation (T_1): I = identity; R = regret; for scenario subset selection (T_2): Si = single decision alternative; Su = subset of decision alternatives; A = all decision alternatives; for robustness metric calculation (T_3): M = none, sum or mean; WM = weighted mean; V = variance; S = skew; K = kurtosis. The rows are ordered approximately from least stable combinations (red) to most stable (green), although some rows have been moved to aid the illustration of concepts in the following discussion.

of robustness metrics and case studies/objectives. This is because ranking stability is not only affected by the similarities in/differences between robustness metrics, but also the similarities/differences in the relative performance of two decision alternatives under the different scenarios considered (see Figure 4). For example, as can be seen in Figure 8, ranking stability for the Adelaide case study is low when the maximin metric is paired with other metrics that also used the same type of robustness metric calculation transformation (T_3), while this is not the case for the other case studies. In this case, this is because many of the decision alternatives have a reliability of 0% in the worst-case scenario, and due to the scenario subset selection (T_2), the maximin metric only considers this worst-case scenario and thus ranks many of the decision alternatives as equal. Other metrics with different scenario subset selection methods use different scenarios (which vary depending on the decision alternative) or use more scenarios and thus rank the decision alternatives differently.

It is also worth noting the high level of disagreement obtained in the Lake Como case for the undesirable deviations when considering the reliability of water supply for irrigation. This effect does not appear when considering the reliability against flooding. This asymmetry can be explained by the fact that the IPCC projections in the alpine region consistently suggest a decrease of water availability in the summer period

due to a change in the snow accumulation/melting dynamics. In fact, the impacts of global warming are expected to reduce the precipitation that falls as snow in winter and, at the same time, to reduce snow melt. The combined effect of this reduction of snow accumulation and reduction of the snow melt strongly challenges the possibility of filling up the lake to provide irrigation during the summer period. Yet, the temporal distribution of such effects can be different due to the variability in the considered scenarios, ultimately producing variable impacts on the performance of different operating policies, which implement different hedging strategies over time. The variable and asymmetric distribution of the resulting performance (toward degradation) is then captured by the metrics relying on a subset of values in the scenario subset selection transformation (T_2) (i.e., undesirable deviations and the metrics relying on multiple percentiles), while other metrics do not recognize this effect and produce inconsistent rankings.

7. Summary and Conclusions

Metrics that consider local uncertainty (i.e., reliability, vulnerability, and resilience) have long been considered in environmental decision-making. Due to deeply uncertain drivers of change including climate, technological and sociopolitical changes, decision-makers have begun to consider multiple scenarios (plausible futures) and robustness metrics to quantify the performance of decision alternatives under deep uncertainty. A large variety of robustness metrics has been considered in recent research with little discussion of the implications of using each metric, and little understanding of the way the metrics are similar or different. However, it has become clear that the choice of robustness metric can have a large effect, with metrics sometimes showing disagreement with regard to which decision alternative is more robust.

This article presents a unifying framework for the calculation of robustness metrics derived from three major transformations (performance value transformation (T_1), scenario subset selection (T_2) and robustness metric calculation (T_3)) used to convert system performance values (e.g., reliability) into the final value of robustness that can be used to rank decision alternatives. The performance value transformation (T_1) converts the original performance values into the information that the decision-maker is interested in. The second transformation (T_2) corresponds to the selection of which scenarios (and associated system performance values) the metric will use. The final transformation (T_3) involves the conversion of transformed performance values over the selected scenarios into a single value of robustness.

This article also presents a conceptual framework for assessing the stability of the ranking of different decision alternatives when different robustness metrics are used. The framework indicates that the greater the similarity in the three transformations for robustness metrics, the more stable the ranking of decision alternatives that use these metrics is and vice versa. Ranking stability is also affected by the degree of consistency of the relative performance of different decision alternatives across the scenarios, where ranking stability is increased if one decision alternative always outperforms the other and vice versa. In order to test this conceptual understanding of ranking stability when different robustness metrics are used, the stability of any two metrics was determined for five objectives in three case studies, which confirmed the proposed conceptual model. The robustness metric calculation (T_3) was found to be the most influential of the three transformations in determining ranking stability, however, the other two transformations also contributed.

In conclusion, robustness metrics can be split into three transformations, which provides a unifying framework for the calculation of robustness. This framework helps decision-makers understand when different robustness metrics should be used by considering (1) the information the decision context relates to most (e.g., absolute performance, regret, or the satisfaction of constraints) (performance value transformation (T_1)), (2) the preference of a decision-maker toward a high or low level of risk aversion for the case study under consideration through scenario subset selection (T_2), and (3) the decision-maker's preference toward maximizing average performance, minimizing variance, or some other higher-order moment, as described by the robustness metric calculation (T_3). These three transformations and the properties of the case studies are useful in describing why rankings of decision alternatives obtained using different robustness metrics sometimes disagree.

Acknowledgments
Thanks is given to Leon van der Linden for his guidance on behalf of SA Water Corporation (Australia) who support the research of Cameron McPhail through Water Research Australia, as well as the comments from the anonymous reviewers, which have improved the quality of this article significantly. The case study robustness data is included in Supporting Information S1.

References

Beh, E. H. Y., Dandy, G. C., 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 & Software*, *53*, 137–153. https://doi.org/10.1016/j.envsoft.2013.11.004

Beh, E. H. Y., Maier, H. R., & Dandy, G. C. (2015a). Adaptive, multiobjective optimal sequencing approach for urban water supply augmentation under deep uncertainty. *Water Resources Research*, *51*(3), 1529–1551. https://doi.org/10.1002/2014WR016254

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

Beh, E. H. Y., Zheng, F., Dandy, G. C., Maier, H. R., & Kapelan, Z. (2017). Robust optimization of water infrastructure planning under deep uncertainty using metamodels. *Environmental Modelling & Software*, *93*, 92–105. https://doi.org/10.1016/j.envsoft.2017.03.013

Ben-Haim, Y. (2004). Uncertainty, probability and information-gaps. *Reliability Engineering and System Safety*, *85*(1), 249–266. https://doi.org/10.1016/j.ress.2004.03.015

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust Optimization*. Princeton and Oxford: Princeton University Press.

Bertsimas, D., & Sim, M. (2004). The price of robustness. *Operations Research*, *52*(1), 35–53. https://doi.org/10.1287/opre.1030.0065

Bilal, K., Manzano, M., Khan, S. U., Calle, E., Li, K., & Zomaya, A. Y. (2013). On the characterization of the structural robustness of data center networks. *IEEE Transactions on Cloud Computing*, *1*(1), 1. https://doi.org/10.1109/TCC.2013.6

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*(9). https://doi.org/10.1029/2011WR011212

De Bruijn, K. M. (2008). Bepalen van schade ten gevolge van overstromingen. Voor verschillende scenario's en bij verschillende beleidsopties (Determining flood damage for different scenarios and policy options). *Deltares Report*, Q4345.

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

Canon, L.-C., & Jeannot, E. (2007). A comparison of robustness metrics for scheduling dags on heterogeneous systems. In *2007 IEEE International Conference on Cluster Computing* (pp. 558–567). IEEE.

Castelletti, A., Galelli, S., Restelli, M., & Soncini-Sessa, R. (2010). Tree-based reinforcement learning for optimal water reservoir operation. *Water Resources Research*, *46*(9). https://doi.org/10.1029/2009WR008898

Culley, S., Noble, S., Yates, A., Timbs, M., Westra, S., Maier, H. R. R., … Castelletti, A. (2016). A bottom up approach to identifying the maximum operational adaptive capacity of water resource systems to a changing climate. *Water Resources Research*, *52*(9), 6751–6768. https://doi.org/10.1002/2015WR018253

Denaro, S., Anghileri, D., Giuliani, M., & Castelletti, A. (2017). Informing the operations of water reservoirs over multiple temporal scales by direct use of hydro-meteorological data. *Advances in Water Resources*, *103*, 51–63. https://doi.org/10.1016/j.advwatres.2017.02.012

Döll, P., & Romero-Lankao, P. (2017). How to embrace uncertainty in participatory climate change risk management—A roadmap. *Earth's Future*, *5*(1), 18–36. https://doi.org/10.1002/2016EF000411

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. https://doi.org/10.1038/nclimate2721

Giuliani, M., & Castelletti, A. (2016). Is robustness really robust? How different definitions of robustness impact decision-making under climate change. *Climatic Change*, 1–16.

Giuliani, M., Li, Y., Castelletti, A., & Gandolfi, C. (2016a). A coupled human-natural systems analysis of irrigated agriculture under changing climate. *Water Resources Research*, *52*(9), 6928–6947. https://doi.org/10.1002/2016WR019363

Giuliani, M., Castelletti, A., Pianosi, F., Mason, E., & Reed, P. M. (2016b). 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), 4015050. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000570

Giuliani, M., Anghileri, D., Castelletti, A., Vu, P. N., & Soncini-Sessa, R. (2016c). 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., Fedorov, R., & Fraternali, P. (2016d). Using crowdsourced web content for informing water systems operations in snow-dominated catchments. *Hydrology and Earth System Sciences*, *20*(12), 5049–5062. https://doi.org/10.5194/hess-20-5049-2016

Grafton, R. Q., Horne, J., & Wheeler, S. A. (2016a). On the marketisation of water: Evidence from the Murray-Darling Basin, Australia. *Water Resources Management*, *30*(3), 913–926. https://doi.org/10.1007/s11269-015-1199-0

Grafton, R. Q., McLindin, M., Hussey, K., Wyrwoll, P., Wichelns, D., Ringler, C., … Orr, S. (2016b). Responding to global challenges in food, energy, environment and water: Risks and options assessment for decision making. *Asia Pacific Policy Study*, *3*(2), 275–299. https://doi.org/10.1002/app5.128

Guariso, G., Rinaldi, S., & Soncini-Sessa, R. (1985). Decision support systems for water management: The Lake Como case study. *European Journal of Operational Research*, *21*(3), 295–306. https://doi.org/10.1016/0377-2217(85)90150-X

Guariso, G., Rinaldi, S., & Soncini-Sessa, R. (1986). The management of Lake Como: A multiobjective analysis. *Water Resources Research*, *22*(2), 109–120. https://doi.org/10.1029/WR022i002p00109

Guillaume, J. H. A., Arshad, M., Jakeman, A. J., Jalava, M., & Kummu, M. (2016). Robust discrimination between uncertain management alternatives by iterative reflection on crossover point scenarios: Principles, design and implementations. *Environmental Modelling & Software*, *83*, 326–343. https://doi.org/10.1016/j.envsoft.2016.04.005

Guo, J., Huang, G., Wang, X., Li, Y., & Lin, Q. (2017). Investigating future precipitation changes over China through a high-resolution regional climate model ensemble. *Earth's Future*, *5*(3), 285–303. https://doi.org/10.1002/2016EF000433

Haasnoot, M., Middelkoop, H., Van Beek, E., & Van Deursen, W. P. A. (2009). A method to develop sustainable water management strategies for an uncertain future. *Sustainable Development*, *19*(6), 369–381.

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

Haasnoot, M., Van Deursen, W. P. A., Guillaume, J. H. A., Kwakkel, J. H., Van Beek, E., & Middelkoop, H. (2014). Fit for purpose? Building and evaluating a fast, integrated model for exploring water policy pathways. *Environmental Modelling & Software*, *60*, 99–120. https://doi.org/10.1016/j.envsoft.2014.05.020

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

Hamarat, C., Kwakkel, J. H., Pruyt, E., & Loonen, E. T. (2014). An exploratory approach for adaptive policymaking by using multi-objective robust optimization. *Simulation Modelling Practice and Theory*, *46*, 25–39. https://doi.org/10.1016/j.simpat.2014.02.008

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

Hashimoto, T., Loucks, D. P., & Stedinger, J. R. (1982b). Robustness of water resources systems. *Water Resources Research*, *18*(1), 21–26. https://doi.org/10.1029/WR018i001p00021

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), 4015012. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000509

Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, *2*(1), 22–26. https://doi.org/10.1109/TSSC.1966.300074

Howard, R. A., & Matheson, J. E. (2005). Influence diagrams. *Decision Analysis*, *2*(3), 127–143. https://doi.org/10.1287/deca.1050.0020

Hurwicz, L. (1953). Optimality criterion for decision making under ignorance. In *Uncertain. Expect. Econ. Essays Honour GLS Shackle*.

Iglesias, A., & Garrote, L. (2015). Adaptation strategies for agricultural water management under climate change in Europe. *Agricultural Water Management*, *155*, 113–124. https://doi.org/10.1016/j.agwat.2015.03.014

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 & Software*, *42*, 55–71. https://doi.org/10.1016/j.envsoft.2012.12.007

Korteling, B., Dessai, S., & Kapelan, Z. (2012). Using information-gap decision theory for water resources planning under severe uncertainty. *Water Resources Management*, *27*(4), 1149–1172.

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), 299–315. https://doi.org/10.1504/IJTPM.2010.036918

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

Kwakkel, J. H., Eker, S., & Pruyt, E. (2016b). How robust is a robust policy? Comparing alternative robustness metrics for robust decision-making. In *Robustness Analysis in Decision Aiding, Optimization, and Analytics* (pp. 221–237). Singapore: Springer.

Laplace, P. S., & Simon, P. (1951). A philosophical essay on probabilities. *Translated from the 6th French edition by Frederick Wilson Truscott and Frederick Lincoln Emory*.

Lempert, R. J. (2003). *Shaping the Next one hundred Years: New Methods for Quantitative, Long-Term Policy Analysis*. Santa Monica, CA: Rand Corporation.

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., Groves, D. G., Popper, S. W., & Bankes, S. C. (2006). A general, analytic method for generating robust strategies and narrative scenarios. *Management Science*, *52*(4), 514–528. https://doi.org/10.1287/mnsc.1050.0472

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

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 & Software*, *81*, 154–164. https://doi.org/10.1016/j.envsoft.2016.03.014

McDowell, G., Stephenson, E., & Ford, J. (2014). Adaptation to climate change in glaciated mountain regions. *Climatic Change*, *126*(1–2), 77–91. https://doi.org/10.1007/s10584-014-1215-z

Morgan, M. G., Henrion, M., & Small, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511840609

Paton, F. L., Maier, H. R., & Dandy, G. C. (2013). Relative magnitudes of sources of uncertainty in assessing climate change impacts on water supply security for the southern Adelaide water supply system. *Water Resources Research*, *49*(3), 1643–1667. https://doi.org/10.1002/wrcr.20153

Paton, F. L., Maier, H. R., & Dandy, G. C. (2014a). Including adaptation and mitigation responses to climate change in a multiobjective evolutionary algorithm framework for urban water supply systems incorporating GHG emissions. *Water Resources Research*, *50*(8), 6285–6304. https://doi.org/10.1002/2013WR015195

Paton, F. L., Dandy, G. C., & Maier, H. R. (2014b). Integrated framework for assessing urban water supply security of systems with non-traditional sources under climate change. *Environmental Modelling & Software*, *60*, 302–319. https://doi.org/10.1016/j.envsoft.2014.06.018

Poff, N. L., Brown, C. M., Grantham, T. E., Matthews, J. H., Palmer, M. A., Spence, C. M., … Dominique, K. C. (2015). Sustainable water management under future uncertainty with eco-engineering decision scaling. *Nature Climate Change*. https://doi:10.1038/nclimate2765

Popper, S. W., Berrebi, C., Griffin, J., Light, T., & Min, E. Y. (2009). *Natural Gas and Israel's Energy Future: Near-Term Decisions from a Strategic Perspective*. Rand Corporation.

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

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

Ray, P. A., Watkins Jr, D. W., Vogel, R. M., & Kirshen, P. H. (2013). Performance-based evaluation of an improved robust optimization formulation. *Journal of Water Resources Planning and Management*, *140*(6), 4014006.

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*, 4016028.

Samsatli, N. J., Papageorgiou, L. G., & Shah, N. (1998). Robustness metrics for dynamic optimization models under parameter uncertainty. *AIChE Journal*, *44*(9), 1993–2006. https://doi.org/10.1002/aic.690440907

Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical Association*, *46*(253), 55–67. https://doi.org/10.1080/01621459.1951.10500768

Schneller, G. O., & Sphicas, G. P. (1983). Decision making under uncertainty: Starr's domain criterion. *Theory and Decision*, *15*(4), 321–336. https://doi.org/10.1007/BF00162111

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

Starr, M. K. (1963). *Product Design and Decision Theory*. Prentice-Hall.
Takriti, S., & Ahmed, S. (2004). On robust optimization of two-stage systems. *Mathematical Programming*, *99*(1), 109–126. https://doi.org/10.1007/s10107-003-0373-y
Voudouris, V., Matsumoto, K., Sedgwick, J., Rigby, R., Stasinopoulos, D., & Jefferson, M. (2014). Exploring the production of natural gas through the lenses of the ACEGES model. *Energy Policy*, *64*, 124–133. https://doi.org/10.1016/j.enpol.2013.08.053
Wald, A. (1950). *Statistical Decision Functions*. London/New York: Chapman & Hall.
Walker, W. E., Lempert, R. J., & Kwakkel, J. H. (2013). Deep uncertainty. In *Encyclopedia of Operations Research and Management Science* (pp. 395–402). Springer.
Wittholz, M. K., O'Neill, B. K., Colby, C. B., & Lewis, D. (2008). Estimating the cost of desalination plants using a cost database. *Desalination*, *229*(1–3), 10–20. https://doi.org/10.1016/j.desal.2007.07.023
Zongxue, X., Jinno, K., Kawamura, A., Takesaki, S., & Ito, K. (1998). Performance risk analysis for Fukuoka water supply system. *Water Resources Management*, *12*(1), 13–30. https://doi.org/10.1023/A:1007951806144