Towards a better understanding of scenarios and robustness for the long-term planning of water and environmental systems

Cameron McPhail

B. Eng. (Civil & Structural) (Honours)

Thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

The University of Adelaide
Faculty of Engineering, Computer, and Mathematical Sciences
School of Civil, Environmental, and Mining Engineering

August 2020

Table of Contents

Abstract

The long-term planning of water and environmental systems presents major challenges to decision-makers, requiring them to make decisions despite a significant degree of uncertainty in the future state of the world. Frequently, decision-makers are operating at the level of deep uncertainty, which refers to when deterministic and stochastic processes are insufficient for representing the future state of the world, and the consideration of multiple plausible futures (scenarios) is required. Further complicating this, probabilities cannot be placed on the scenarios, and therefore traditional performance metrics such as reliability, vulnerability, resilience, or expected value do not apply. Rather, deep uncertainty requires robustness metrics, which aim to determine the level of system performance and how that performance varies across all scenarios.

The specific aims of this research are (i) to introduce a unified framework for the calculation of a wide range of robustness metrics, enabling the robustness values and rankings obtained from different metrics to be compared in an objective fashion; (ii) to develop a deeper understanding of how different selections of scenarios can affect the absolute and relative robustness and rankings of decision alternatives of interest; and (iii) to create a generic guidance framework and software tool to assist with the identification of the most robust decision alternative for a given problem.

For the first aim, this research presents 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 the decision-context, the decision-maker's preferred level of risk aversion, and the decision-maker's preference towards maximizing performance or minimizing variance. This research also introduces a conceptual framework describing when different robustness metrics are likely to agree and disagree.

For the second aim, the research describes how scenarios are generally represented in model-based assessments, and develops a systematic, quantitative methodology for exploring the influence of different sets of scenarios on the absolute and relative robustness of different decision alternatives, which is then applied to the Lake Problem.

Case study results show that despite different sets of scenarios causing a significant difference in robustness values, there is little difference in the corresponding rankings, and therefore similar decision outcomes will be reached regardless of how the scenarios are selected. It is also revealed that the impact of the scenarios on the robustness values is due to complex interactions with the system model and robustness metrics.

For the third aim, the research considers the knowledge developed in the first two aims and builds a guidance framework for decision-makers on how to identify the most robust decision alternative for a given problem. The guidance caters to a variety of situations where the scenarios and/or robustness metrics are known or not known and also includes guidance on how to create a custom robustness metric for the problem at hand. An open-source software package is introduced, the RAPID package, to assist in the consistency and ease-of-use of implementing the guidance framework.

Statement of originality

I, Cameron McPhail, certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

26/02/2020

Cameron McPhail Date

Acknowledgements

Firstly, I would like to thank my supervisors, Holger Maier, Seth Westra, and Leon van der Linden. Their guidance throughout the last four years has been invaluable, teaching me many skills and ways of thinking that I now cannot imagine living without. In addition to my supervisors, I also thank the other co-authors of this research, Jan Kwakkel, Andrea Castelletti, and Matteo Giuliani. They brought many different and useful perspectives and contributions to the papers and were a great pleasure to work with. Thank you to SA Water for their sponsorship of this research through Water Research Australia, and a big thank you to Water Research Australia for their support, in particular to Carolyn Bellamy who always went out of her way to ensure that the PhD went as smoothly as possible.

Thank you to my friend and colleague, Sam Culley, for his contributions. During our PhDs, the many coffees, lunches, and drinks were a very important part of maintaining a good life balance.

I would also like to thank all of those who offered advice on my PhD and career. In particular, thank you to Josh Cantone, Andy Roberts, and Michael Storey, who have always offered very valuable guidance to me for many years.

A very big thank you must go to my partner, David Kartika, who always encouraged me to persevere and to make an impact. His patience and support through the ups and downs of this journey made the journey much better than it would otherwise have been.

Lastly, but perhaps most importantly, thank you to my family, in particular to my parents. Their contributions are too many to count and without their continuous support, encouragement, and patience throughout my life, I would not have been capable of getting an opportunity to start a PhD, let alone to finish one.

List of figures

List of tables

Chapter 1

Around the world, decision-makers are required to make long-term plans for complex water and environmental systems. These plans and decisions may result in severe and long-lasting consequences as the future unfolds, including environmental (e.g. pollution, damage to ecosystems), economic (e.g. unexpectedly high costs), social (e.g. corporate or governmental reputation) and/or technological (e.g. unacceptable levels of service, infrastructure failure) (Ascough et al., 2008; Grafton et al., 2016; Walker et al., 2013a). As decision-makers seek to avoid these consequences, they face major challenges including the question of how to plan under significant uncertainty. Often, decision-makers are operating in conditions of "deep" uncertainty, which refers to when deterministic and stochastic processes are insufficient for representing the future state of the world, and the consideration of multiple plausible futures (scenarios) is required (Bradfield et al., 2005; Herman et al., 2014; Kwakkel et al., 2010; Kwakkel and Haasnoot, 2019; Lempert, 2003; Little et al., 2018; Maier et al., 2016; Schwarz, 1991; van der Heijden, 1996; Varum and Melo, 2010; Walker et al., 2013b; Wright and Cairns, 2011).

When representing uncertain future conditions with the aid of scenarios, traditional performance metrics are used for each individual scenario, including metrics such as reliability (frequency of failure), vulnerability (severity of failure), and resilience (time to recover from failure) (Burn et al., 1991; Hashimoto et al., 1982; Maier et al., 2001; Zongxue et al., 1998). However, since scenarios have no probabilities or likelihoods attached to them, these metrics cannot be used to determine the robustness of the system (the system performance across multiple plausible futures) (Maier et al., 2016).

1.1. Background on scenarios and robustness

Decision-makers have used a wide variety of metrics to quantify the robustness of a system (the system performance across multiple plausible futures). These include:

- Expected value metrics (Wald, 1951), 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. (2016)), 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.

Despite the wide variety of robustness metrics available to decision-makers, there has been no guidance on which robustness metrics are most appropriate for any given problem or set of decision-maker preferences. This is problematic, because a common conclusion from recent research is that different robustness metrics can sometimes lead to different decisions being made (Borgomeo et al., 2018; Drouet et al., 2015; Giuliani and Castelletti, 2016; Hall et al., 2012; Herman et al., 2015; Kwakkel et al., 2016; Lempert and Collins, 2007; Roach et al., 2016). The impact that the choice of robustness metric has on decision-making has never been quantified in a generalisable manner in the literature.

Included in the calculation of robustness is the set of scenarios under consideration. And just as there is a variety of robustness metrics, there is also a variety of factors that can contribute to different plausible futures (scenarios) being selected. This can include which school of thought is used to create the scenarios, with the three main schools being *La Prospective* (prospective thinking) (using experts and computational models to create scenarios and assign probabilities to these scenarios) (Berger, 1964; Bradfield et al., 2005), Probabilistic Modified Trends method (determining a "most-likely" future, as well as upper- and lower-quartiles of expected futures) (Bishop et al., 2007; Bradfield et al., 2005), and the Intuitive Logics school (using domain experts to develop flexible, generalizable scenarios, without assigning probabilities) (Bryant and Lempert, 2010; Kwakkel et al., 2013). Generally, in the literature for water and environmental systems, the Intuitive Logics school is most commonly followed because it is the only one that does not assign probabilities to the scenarios. Even within the Intuitive Logics school, there are several factors that influence which scenarios are to be used to calculate robustness.

Generally, scenarios can be divided into three categories, including predictive, explorative and normative (Börjeson et al., 2006). Predictive scenarios aim to answer the question “what will happen?” and can either represent a particular trajectory in future conditions or changes in this trajectory in response to particular events, whereas explorative scenarios aim to answer the question “what could happen?”, and can be framed on uncertainties in conditions that are known to affect system performance or can be completely unframed (Maier et al., 2016). In contrast, normative scenarios aim to answer the question “how can a specific future be realized”, and can represent conditions that result in interesting outcomes, or conditions under which certain decision alternatives no longer perform adequately. Different approaches may also result in different numbers of scenarios being used in the analysis. Since scenarios are used as inputs to the calculation of robustness, and there are many methods and approaches to calculating scenarios, it follows that it is important to understand the impact of the selection of scenarios on robustness. However, the only analyses of this have been qualitative and anecdotal (Kwakkel et al., 2012; Phadnis, 2019). Consequently (and similar to the robustness metrics), there are many methods for the selection of scenarios, and a need to better understand the impact that these methods have on the robustness of a system.

Given that the robustness of decision alternatives is impacted by scenario selection and choice of robustness metric, it can be difficult to determine which decision alternatives are the most robust. For this reason, there is a need to provide a systematic approach to assist decision-makers with selecting the most appropriate robustness metric for their decision context (i.e. the attributes of the system) and decision-maker preferences. There is also a need for a method of quantifying the impact that the selection of scenarios and the robustness metric have on the robustness of decision alternatives.

1.2. Research objectives

This thesis addresses several of the key needs that arise from the gaps in the literature on the use of scenarios and robustness metrics. Specifically, three aims have been developed, and the link between these and the calculation of robustness is shown in Figure 1-1. The aims are:

1. To introduce a unified framework for the calculation of a wide range of robustness metrics, enabling the robustness values and rankings obtained from different metrics to be compared in an objective fashion;
2. To develop a deeper understanding of how different selections of scenarios can affect the absolute and relative robustness of the decision alternatives of interest; and
3. To create a generic guidance framework and software tool to assist with the identification of the most robust decision alternative for a given problem.

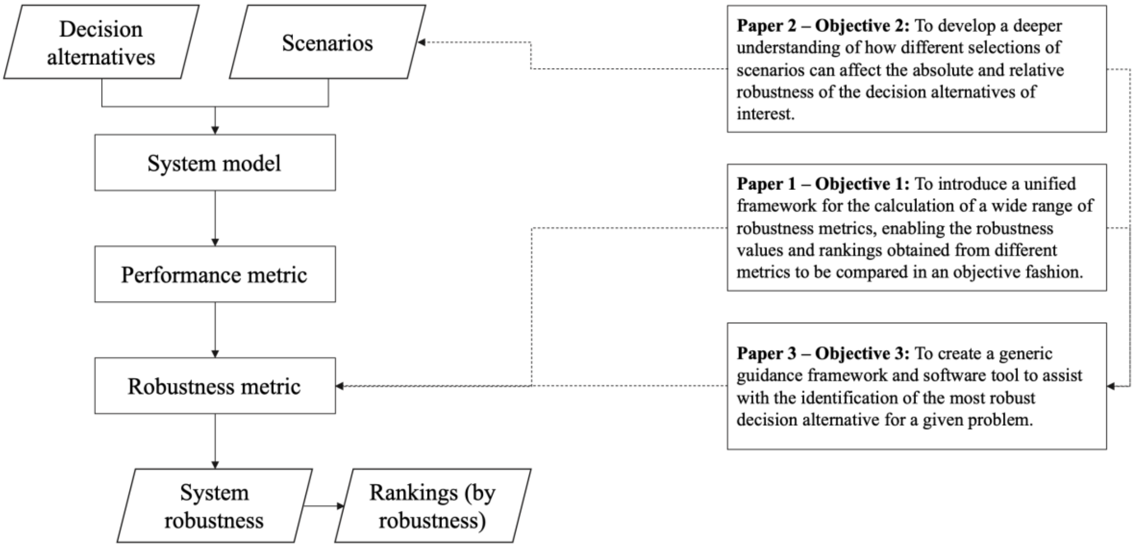

Decision alternatives
Scenarios
System model
Performance metric
Robustness metric
System robustness
Rankings (by robustness)
Paper 2 – Objective 2: To develop a deeper understanding of how different selections of scenarios can affect the absolute and relative robustness of the decision alternatives of interest.
Paper 1 – Objective 1: To introduce a unified framework for the calculation of a wide range of robustness metrics, enabling the robustness values and rankings obtained from different metrics to be compared in an objective fashion.
Paper 3 – Objective 3: To create a generic guidance framework and software tool to assist with the identification of the most robust decision alternative for a given problem.

Figure 1-1 Links between each research objective and the process of calculating the system robustness.

1.3. Thesis organisation

This thesis is comprised of five chapters. The bulk of the research is contained in Chapters 2 to 4. These chapters are the three papers (Figure 1-1): Chapter 2 (Objective 1) has been published in Earth's Future; Chapter 3 (Objective 2) has been submitted to Water Resources Research and is soon to be re-submitted after major revisions, which are included in the chapter presented in this thesis; and Chapter 4 (Objective 3) is to be submitted to Environmental Modelling and Software. The section, figure and table numbers have been modified in line with University guidelines but the manuscript material is otherwise unchanged.

Chapter 2 presents 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. This chapter also introduces a conceptual framework describing when relative robustness values of decision alternatives obtained using different metrics are likely to agree and disagree. 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.

Chapter 3 looks closely at scenarios, describing three conceptually different distributions of scenarios in the scenario space, followed by the development of a systematic, quantitative methodology for exploring the influence of these distributions on the robustness and the ranking of decision alternatives. The influence of the distribution of scenarios is illustrated on The Lake Problem, a hypothetical case study commonly used in the literature. It is revealed that the impact of the scenarios on the robustness values is due to complex interactions with the system model and robustness metrics.

Chapter 4 considers the knowledge developed in the first two aims (regarding the influence of robustness metrics and the selection of scenarios on the absolute and relative robustness of decision alternatives of interest), and from this, builds guidance for decision-makers on the identification of the most robust decision alternative for the problem at hand. This guidance includes the consideration of a variety of situations where the scenarios and/or robustness metrics are known or not known, and also includes guidance on how to create a custom robustness metric based on problem attributes and decision-maker preferences. This chapter also presents an open-source software package, the RAPID (Robustness Analysis Producing Intelligent Decisions) package, to assist in the consistency and ease-of-use of implementing the guidance framework. The guidance framework and software package are illustrated on The Lake Problem, where the guidance is used to create custom robustness metrics best suited to the case study, assess the impact of different candidate scenario sets and robustness metrics, and determine the most robust decision alternatives.

Conclusions are provided in Chapter 5, which also summarises the research contributions, limitations, and recommended future research.

1.4. References

Ascough, J.C., Maier, H.R., Ravalico, J.K., Strudley, M.W., 2008. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. Ecol. Modell. 219, 383–399. https://doi.org/10.1016/j.ecolmodel.2008.07.015

Berger, G., 1964. L'attitude prospective. Manag. Int. 43–46.

Bishop, P., Hines, A., Collins, T., 2007. The current state of scenario development: an overview of techniques. foresight 9, 5–25. https://doi.org/10.1108/14636680710727516

Borgomeo, E., Mortazavi-Naeini, M., Hall, J.W., Guillod, B.P., 2018. Risk, Robustness and Water Resources Planning Under Uncertainty. Earth's Futur. 6, 468–487.

Börjeson, L., Höjer, M., Dreborg, K.H., Ekvall, T., Finnveden, G., 2006. Scenario types and techniques: Towards a user's guide. Futures 38, 723–739. https://doi.org/10.1016/j.futures.2005.12.002

Bradfield, R., Wright, G., Burt, G., Cairns, G., Van Der Heijden, K., 2005. The origins and evolution of scenario techniques in long range business planning. Futures 37, 795–812. https://doi.org/10.1016/j.futures.2005.01.003

Bryant, B.P., Lempert, R.J., 2010. Thinking inside the box: A participatory, computer-assisted approach to scenario discovery. Technol. Forecast. Soc. Change 77, 34–49. https://doi.org/10.1016/j.techfore.2009.08.002

Burn, D.H., Venema, H.D., Simonovic, S.P., 1991. Risk-Based Performance Criteria for Real-Time Reservoir Operation. Can. J. Civ. Eng. 18, 36–42. https://doi.org/10.1139/l91-005

Drouet, L., Bosetti, V., Tavoni, M., 2015. Selection of climate policies under the uncertainties in the Fifth Assessment Report of the IPCC. Nat. Clim. Chang. 5, 937–940.

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

Grafton, Q., Horne, J., Wheeler, S.A., 2016. On the Marketisation of Water: Evidence from the Murray-Darling Basin, Australia. Water Resour. Manag. 30, 913–926. https://doi.org/10.1007/s11269-015-1199-0

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 Anal. 32, 1657–1672. https://doi.org/10.1111/j.1539-6924.2012.01802.x

Hashimoto, T., Stedinger, J.R., Loucks, D.P., 1982. Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. Water Resour. Res. 18, 14–20. https://doi.org/10.1029/WR018i001p00014

Herman, J.D., Reed, P.M., Zeff, H.B., Characklis, G.W., 2015. How Should Robustness Be Defined for Water Systems Planning under Change? J. Water Resour. Plan. Manag. 141, 04015012. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000509

Herman, J.D., Zeff, H.B., Reed, P.M., Characklis, G.W., 2014. Beyond optimality: Multistakeholder robustness tradeoffs for regional water portfolio planning under deep uncertainty. Water Resour. Res. 50, 7692–7713.

Kwakkel, J.H., Auping, W.L., Pruyt, E., 2013. Dynamic scenario discovery under deep uncertainty: The future of copper. Technol. Forecast. Soc. Change 80, 789–800. https://doi.org/10.1016/j.techfore.2012.09.012

Kwakkel, J.H., Eker, S., Pruyt, E., 2016. How robust is a robust policy? Comparing alternative robustness metrics for robust decision-making, in: International Series in Operations Research and Management Science. Springer, pp. 221–237. https://doi.org/10.1007/978-3-319-33121-8_10

Kwakkel, J.H., Haasnoot, M., 2019. Supporting DMDU: A taxonomy of approaches and tools, in: Decision Making under Deep Uncertainty. Springer, pp. 355–374.

Kwakkel, J.H., van der Pas, J.W.G.M., 2011. Evaluation of infrastructure planning approaches: an analogy with medicine. Futures 43, 934–946.

Kwakkel, J.H., Walker, W., Marchau, V., 2012. Assessing the efficacy of adaptive airport strategic planning: Results from computational experiments. Environ. Plan. B Plan. Des. 39, 533–550.

Kwakkel, J.H., Walker, W.E., Marchau, V.A.W.J., 2010. Classifying and communicating uncertainties in model-based policy analysis. Int. J. Technol. Policy Manag. 10, 299. https://doi.org/10.1504/IJTPM.2010.036918

Lempert, R.J., 2003. Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis. Rand Corporation. https://doi.org/10.1016/j.techfore.2003.09.006

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

Little, J.C., Hester, E.T., Elsawah, S., Filz, G.M., Sandu, A., Carey, C.C., Iwanaga, T., Jakeman, A.J., 2018. A tiered, system-of-systems modeling framework for resolving complex socio-environmental policy issues. Environ. Model. Softw.

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? Environ. Model. Softw. 81, 154–164. https://doi.org/10.1016/j.envsoft.2016.03.014

Maier, H.R., Lence, B.J., Tolson, B.A., Foschi, R.O., 2001. First order reliability method for estimating reliability, vulnerability, and resilience. Water Resour. Res. 37, 779–790.

Phadnis, S., 2019. Effectiveness of Delphi-and scenario planning-like processes in enabling organizational adaptation: A simulation-based comparison. Futur. Foresight Sci. e9.

Roach, T., Kapelan, Z., Ledbetter, R., Ledbetter, M., 2016. Comparison of Robust Optimization and Info-Gap Methods for Water Resource Management under Deep Uncertainty. J. Water Resour. Plan. Manag. 142, 04016028. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000660

Savage, L.J., 1951. The theory of statistical decision. J. Am. Stat. Assoc. 46, 55–67. https://doi.org/10.1080/01621459.1951.10500768

Schwarz, P., 1991. The art of the long view: planning for the future in an uncertain world. John Wiley & Sons, Chichester, England.

Simon, H.A., 1956. Rational choice and the structure of the environment. Psychol. Rev. 63, 129–138. https://doi.org/10.1037/h0042769

van der Heijden, K., 1996. Scenarios: the art of strategic conversation. John Wiley & Sons.

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

Wald, A., 1951. Statistical decision functions, Nature. New York; Chapman & Hall: London. https://doi.org/10.1038/1671044b0

Walker, W.E., Haasnoot, M., Kwakkel, J.H., 2013a. Adapt or perish: A review of planning approaches for adaptation under deep uncertainty. Sustain. 5, 955–979. https://doi.org/10.3390/su5030955

Walker, W.E., Lempert, R., Kwakkel, J., 2013b. Deep Uncertainty, in: Encyclopedia of

Operations Research and Management Science. Springer, pp. 395–402. https://doi.org/10.1007/978-1-4419-1153-7_1140

Wright, G., Cairns, G., 2011. Scenario thinking: Practical approaches to the future. Springer.

Zongxue, X., Jinno, K., Kawamura, A., Takesaki, S., Ito, K., 1998. Performance risk analysis for Fukuoka water supply system. Water Resour. Manag. 12, 13–30. https://doi.org/10.1007/s11269-006-9040-4

Chapter 2

Robustness metrics: How are they calculated, when should they be used, and why do they give different results?

Cameron McPhail, Holger R. Maier, Jan Kwakkel, Matteo Giuliani, Andrea Castelletti, and Seth Westra

Published in Earth's Future

Statement of Authorship

Title of Paper	Robustness metrics: How are they calculated, when should they be used, and why do they give different results?
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published in Earth's Future

Principal Author

Name of Principal Author (Candidate)	Cameron McPhail		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote manuscript, made edits to manuscript, and was corresponding author.		
Overall percentage (%)	60%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	24/02/2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Holger Maier		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature		Date	17/08/20

Name of Co-Author	Jan Kwakkel		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature		Date	25-02-2020

Name of Co-Author	Matteo Giuliani		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature		Date	24/02/2020

Name of Co-Author	Andrea Castelletti		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature		Date	24/02/2020

Name of Co-Author	Seth Westra		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, and made edits to manuscript.		
Signature		Date	18/08/2020

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 (i) the decision-context (i.e. the suitability of using absolute performance or regret), (ii) the decision-maker's preferred level of risk aversion, and (iii) the decision-maker's preference towards maximizing performance, minimizing variance, or some higher-order moment. This paper 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.

2.1. 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 (Matteo Giuliani et al., 2016a; Monaco, 1992) and a realization of the value of information for reducing this localized uncertainty (Howard Matheson, J. E., 1984; Howard, 1966). 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., 1982b; Maier et al., 2001; Zongxue et al., 1998). However, as a result of climatic, technological, economic and socio-political 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 and Romero-Lankao, 2016; 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 and Romero-Lankao, 2016; 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 (Bandi, 2012).

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., 1982a), dynamic chemical reaction models (Samsatli et al., 1998), timetable scheduling (Canon and 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 paper, 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, 1951), 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. (2016a)), 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 and 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* (2008) 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.* (2016a) 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 paper are to (i) 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, (ii) 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, (iii) 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 (iv) test the conceptual framework from (iii) 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 and 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 paper is organized as follows. In Section 2.2, 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 2.3, 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 2.4 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 2.5, 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 2.6 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 2.7.

2.2. 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 (i) the decision alternatives (e.g. policy options, designs, solutions, management plans) for which robustness is to be calculated, (ii) the outcome of interest (performance metric) of the decision alternatives (e.g. cost, reliability) and (iii) 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 2-1.

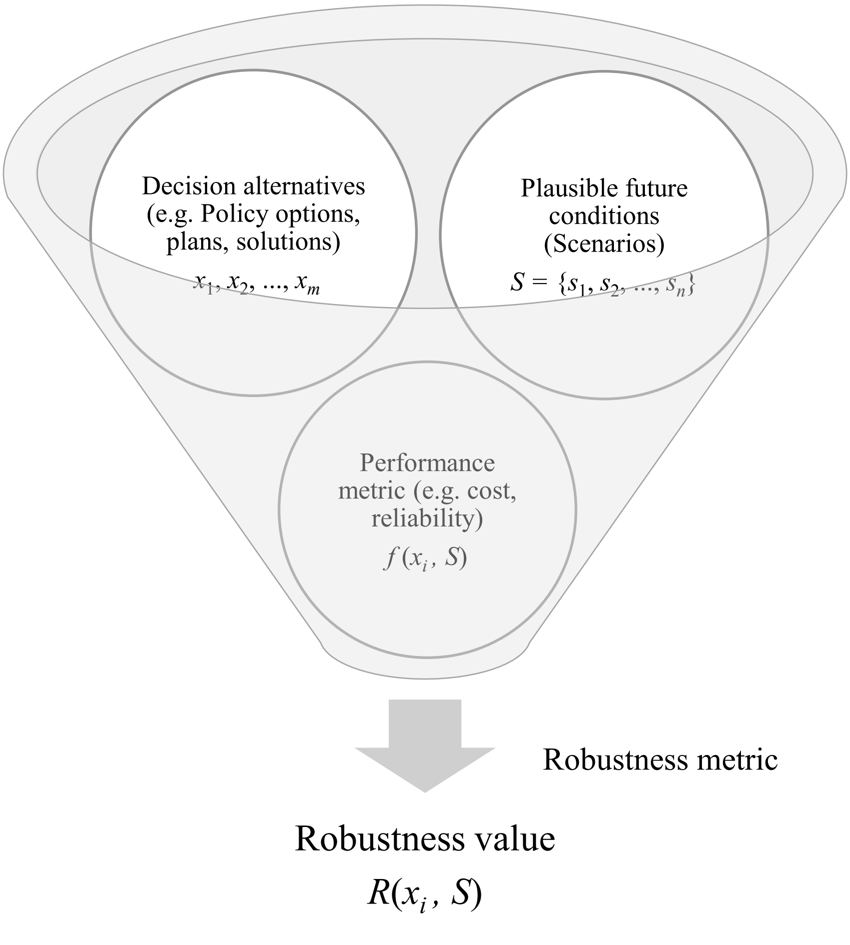

Decision alternatives (e.g. Policy options, plans, solutions)
$x_1, x_2, \ldots, x_m$
Plausible future conditions (Scenarios)
$S = \{s_1, s_2, \ldots, s_n\}$
Performance metric (e.g. cost, reliability)
$f(x_i, S)$
Robustness metric
Robustness value
$R(x_i, S)$

Figure 2-1. Common components contributing to the calculation of robustness.

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 2-2. Details of these transformations for a range of commonly used robustness metrics are given in Table 2-1 and their mathematical implementations are given in the Supporting Information.

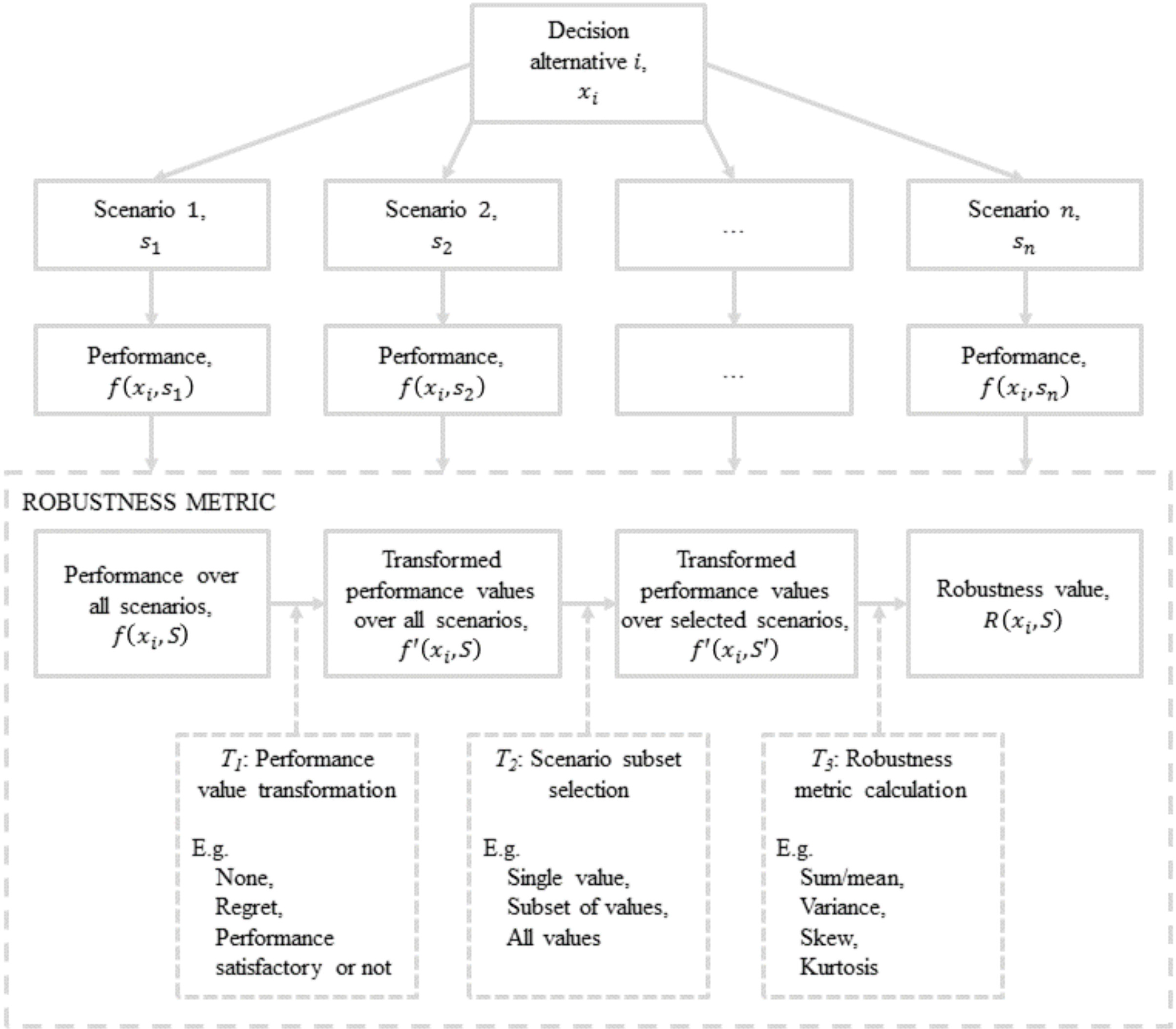

Decision alternative i, x_i
Scenario 1, s_1
Scenario 2, s_2
…
Scenario n, s_n
Performance, $f(x_i,s_1)$
Performance, $f(x_i,s_2)$
…
Performance, $f(x_i,s_n)$
ROBUSTNESS METRIC
Performance over all scenarios, $f(x_i,S)$
Transformed performance values over all scenarios, $f'(x_i,S)$
Transformed performance values over selected scenarios, $f'(x_i,S')$
Robustness value, $R(x_i,S)$
T_1: Performance value transformation
E.g.
None,
Regret,
Performance satisfactory or not
T_2: Scenario subset selection
E.g.
Single value,
Subset of values,
All values
T_3: Robustness metric calculation
E.g.
Sum/mean,
Variance,
Skew,
Kurtosis

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

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 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 2-1, which scenarios are considered in the robustness calculation is highly variable between different metrics.

Table 2-1. A summary of the three transformations that are used by each robustness metric considered in this paper.

Metric	Original reference	*T_1*: Performance value transformation	*T_2*: Scenario subset selection	*T_3*: Robustness metric calculation
Maximin	Wald (1951)	Identity	Worst-case	Identity
Maximax	Wald (1951)	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); 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. (2016a)	Regret from median performance	Worst-half	Sum
Percentile-based skewness	Voudouris et al. (2014); Kwakkel et al. (2016a)*	Identity	10th, 50th and 90th percentiles	Skew
Percentile-based peakedness	Voudouris et al. (2014); Kwakkel et al. (2016a)*	Identity	10th, 25th, 75th and 90th percentiles	Kurtosis
Starr's domain criterion	Starr (1962); Schneller and Sphicas (1983)	Satisfaction of constraints	All	Mean

* Kwakkel et al. (2016a) adapted some metrics from Voudouris et al. (2014).

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 of kurtosis.

2.3. 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 2.2. A summary of the three transformations, as well as the relative level of risk aversion, is provided in Section 2.3.4.

2.3.1. 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 2-2. As can be seen, the categorization is based on (i) 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 (ii) whether a robustness metric provides an indication of actual system performance or whether system performance is satisfactory compared with a pre-specified performance threshold.

Table 2-2. 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)**) - (Decision Scaling**)	- Starr's domain criterion - (Info Gap*)
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

* Robustness calculated explicitly, but based on deviations from an expected scenario

** Robustness not calculated explicitly

Note that brackets around a metric indicate that the metrics is considered unsuitable and is not considered in the following analysis.

Many of the classic decision analytic robustness metrics belong to the bottom-right hand quadrant of Table 2-1, 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 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 (*T₂*)) and how these values are combined into a single value of R (i.e. robustness metric calculation (*T₃*)), as discussed in Sections 2.3.2 and 2.3.3.

Metrics in the bottom-left quadrant of Table 2-2 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., 2016a; Lempert et al., 2008; Popper et al., 2009).

Metrics in the top right quadrant of Table 2-2 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 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.* (2015), *Herman et al.* (2015) and *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., 2013). 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 2-2 and will not be considered further in this paper.

Metrics in the top-left quadrant of Table 2-2 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 2-2, 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., 2016). However, as these methods do not quantify robustness explicitly, they are shown in parentheses in Table 2-2 and will not be considered further in this paper.

2.3.2. 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 2-3. 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 2-3, 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.

Table 2-3. Classification of robustness metrics in terms of scenario subset selection (T_2).

Robustness metric	Scenarios from S used to form the subset S'		
	Subset		All
	Single	Number	
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

Which scenarios from S are selected has a significant effect on the relative level of inherent risk aversion of a robustness metric, as shown in Figure 2-3. 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 2-3). 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 2-3), leading to a very conservative solution (Bertsimas and 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 2-3). The other metrics fit somewhere in-between these extremes of low and high levels of intrinsic risk aversion, as shown in Figure 2-3 and explained below.

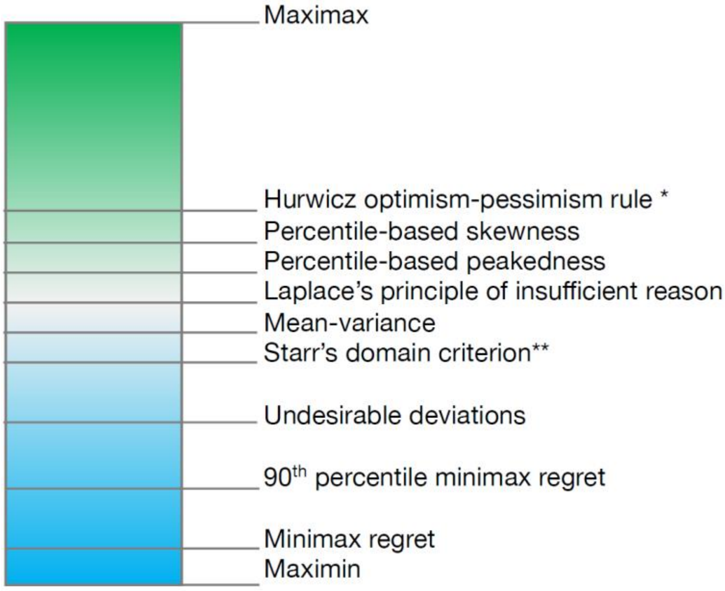

Maximax
Hurwicz optimism-pessimism rule *
Percentile-based skewness
Percentile-based peakedness
Laplace's principle of insufficient reason
Mean-variance
Starr's domain criterion**
Undesirable deviations
90th percentile minimax regret
Minimax regret
Maximin

Figure 2-3. 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 2.3.3). 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 2-2). 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.

Calculation of the metrics in the middle of Figure 2-3 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, and percentile-based peakedness, which uses the 10th, 25th, 75th and 90th percentiles (Table 2-3). 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 2.3.3), 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 2-3. 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 the distribution (Table 2-3). 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 2-3).

2.3.3. 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 2-4. As can be seen, for some metrics, such as the maximin, maximin, 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 2.3.2.

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 and 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.

Table 2-4. Robustness metric calculation (*T*$_3$) used to transform the sampled performance information into the value of robustness.

Robustness metric	Robustness metric calculation						
	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			✓				

2.3.4. Summary of categorization of robustness metrics

The complete categorization of the commonly used robustness metrics considered in this paper in accordance with the three transformations (performance value transformation (*T*$_1$) (Table 2-2), scenario subset selection (*T*$_2$) (Table 2-3) and robustness metric calculation (*T*$_3$) (Table 2-4)), as well as the relative level of risk aversion that is associated with *T*$_2$ (Figure 2-3), is given in Table 2-5. 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 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 (i) 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, (ii) 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 (iii) 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.

Table 2-5. 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 Materials for equations.

Robustness metric	*T*₁: Performance value transformation				*T*₂: Scenario subset selection				*T*₃: Robustness metric calculation				
	Optimize system performance		Satisfy constraints										
	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 ☆**		✓			
Laplace's principle of insufficient reason	✓						✓	☆☆☆	✓				
Minimax regret		✓			✓			☆☆☆☆☆	✓				
90th percentile minimax regret		✓			✓			☆☆☆☆	✓				
Mean-variance	✓						✓	☆☆☆	✓		✓		
Undesirable deviations		✓				V*		☆☆☆☆	✓				
Percentile-based skewness	✓					3		☆☆☆				✓	
Percentile-based peakedness	✓					4		☆☆☆					✓
Starr's domain criterion			✓			V*		☆ to ☆***	✓				

* V = variable

** Hurwicz optimism-pessimism rule has a parameter (selected by the decision-maker) to determine the relative level of risk aversion.

*** This is dependent on the minimum performance threshold selected by the decision-maker.

2.4. When do robustness metrics disagree?

As mentioned previously, robustness metrics have been shown to disagree in certain cases (Giuliani and Castelletti, 2016; Herman et al., 2015; Kwakkel et al., 2016a). 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:

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

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

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

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

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

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

Consequently, relative differences in robustness values obtained when different robustness metrics are used are a function of (i) 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 (ii) 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 2-4. 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.

2.4.1. 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 2-4). 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 2-5). 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 2-3). Consequently, one would expect the ranking of decision alternatives to be very stable when these two metrics are used.

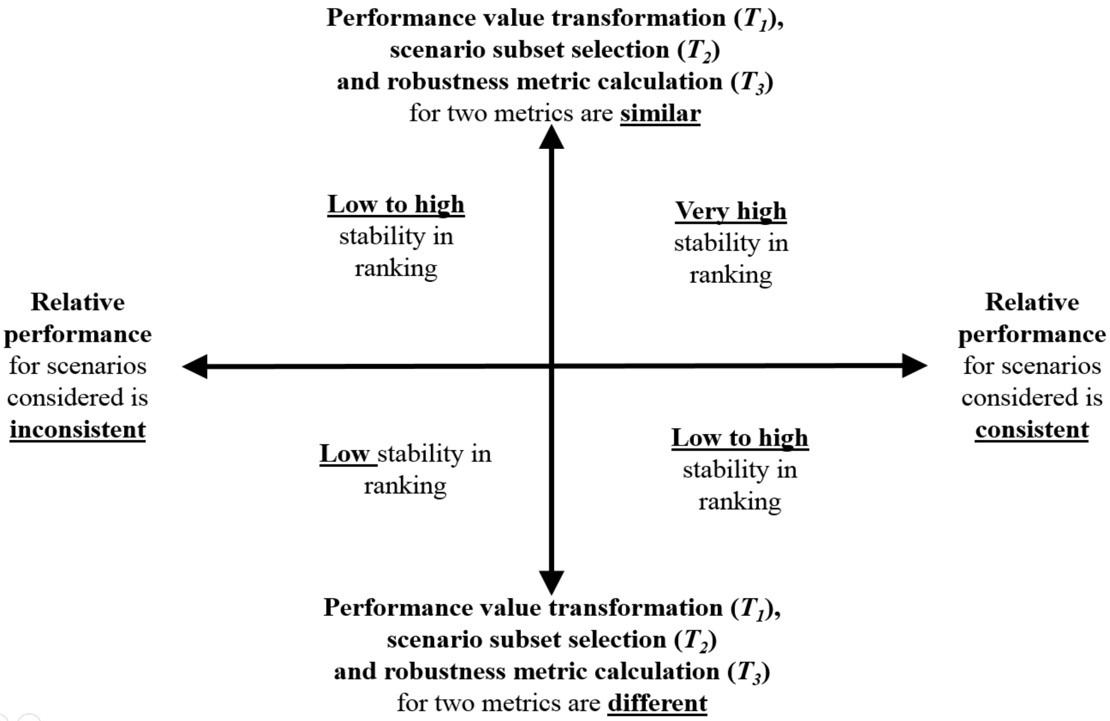

Performance value transformation (T_1), scenario subset selection (T_2) and robustness metric calculation (T_3) for two metrics are similar
Low to high stability in ranking
Very high stability in ranking
Relative performance for scenarios considered is inconsistent
Relative performance for scenarios considered is consistent
Low stability in ranking
Low to high stability in ranking
Performance value transformation (T_1), scenario subset selection (T_2) and robustness metric calculation (T_3) for two metrics are different

Figure 2-4. 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.

2.4.2. 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 2-4). 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 Table 2-3 and Table 2-5). Similarly, there is a potentially high degree of inconsistency in the relative performance values used for 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 2-3). Consequently, one would expect the ranking of decision alternatives to be generally unstable when these two metrics are used.

2.4.3. 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 2-4), ranking stability can range from high to low. 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 Table 2-3 and Table 2-5). 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 2-4, 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).

2.4.4. 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 2-4), ranking stability can also range from high to low due to the complex interactions between the different drivers affecting ranking stability. 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 2-5). 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 2-4, 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.

2.5. Case studies

Three case studies with different properties are used to test the conceptual model of ranking stability introduced in Section 2.4, as shown in Table 2-6. 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 varies greatly in each case study (28 to 3000), as does the number of optimal decision alternatives considered (11 to 72).

Table 2-6. Summary of the characteristics of the Southern Adelaide, Lake Como and Waas case studies.

Name	**Location**	**Decision variables, components of $\boldsymbol{x_i}$**	**Selected objectives and performance metrics, $\boldsymbol{f(x_i, S)}$**	**Number of scenarios, $\boldsymbol{n}$, where $\boldsymbol{S = \{s_1, \dots, s_n\}}$**	**Number of decision alternatives, $\boldsymbol{m}$ where $\boldsymbol{X = \{x_1, \dots, x_m\}}$**
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

2.5.1. 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., 2017, 2015, 2014; Clark et al., 2015; Paton et al., 2014b, 2014a, 2013). 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 2-5: 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.

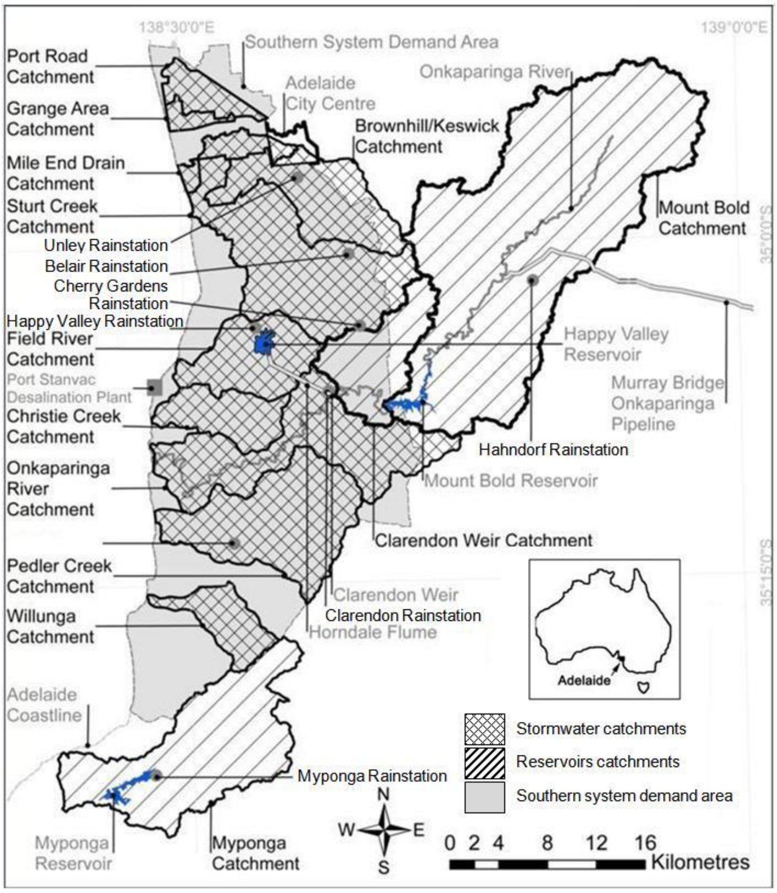


Figure 2-5. The southern Adelaide water supply system as it existed in 2010.

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 paper 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., 2015) generated 72 optimal decision alternatives for this case study using a multi-objective evolutionary algorithm, which will be used in this paper. 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. 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.

2.5.2. 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 4,552 km^2 watershed (see Figure 2-6) 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 1,400 km^2 mostly cultivated with maize (Giuliani et al., 2016; Guariso et al., 1986, 1985).

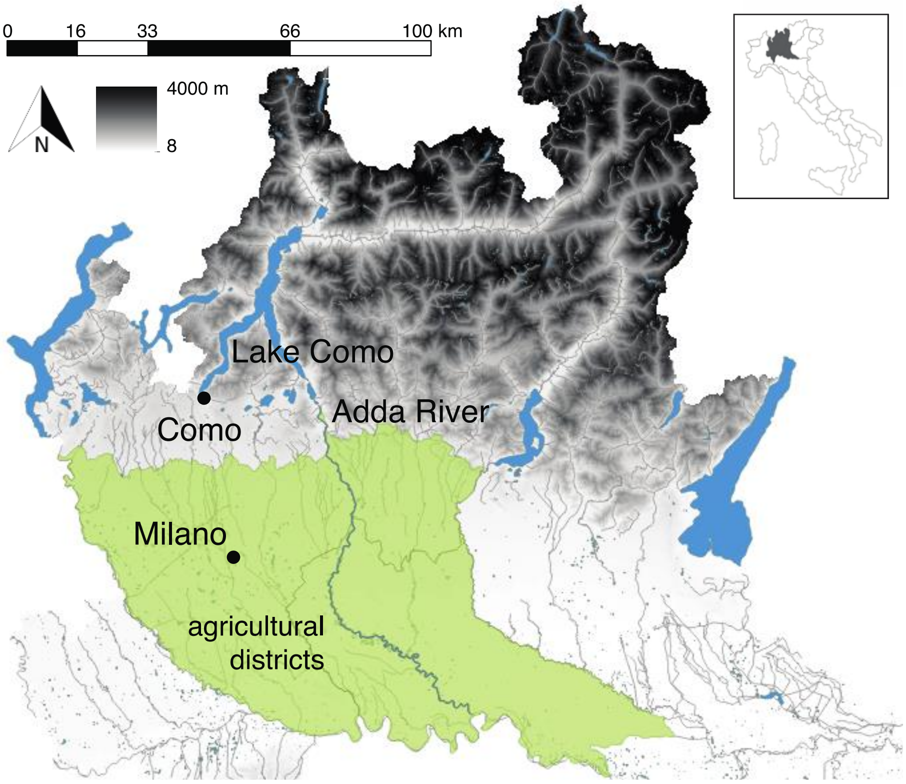


Figure 2-6. Map of the Lake Como system.

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 and Castelletti, 2016).

Due to a changing climate and thus a changing flood risk (Giuliani and Castelletti, 2016; McDowell et al., 2014) and availability of water (Iglesias and 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 Representative Concentration Pathways, 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; Matteo Giuliani et al., 2016b; Giuliani and 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 and Castelletti, 2016) generated 19 Pareto optimal decision alternatives by optimizing the Flooding and Irrigation objectives over historical climate conditions via Evolutionary Multi-Objective Direct Policy Search, a simulation-based optimization approach that combines direct policy search, nonlinear approximating networks, and multi-objective evolutionary algorithms (Matteo Giuliani et al., 2016c). These optimal reservoir operation policies are used in the following analysis.

2.5.3. 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 2-7), 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., 2011). In the future, changes in land use and climate, as well as socio-economic developments, may further increase the risk of damage, so action is needed.

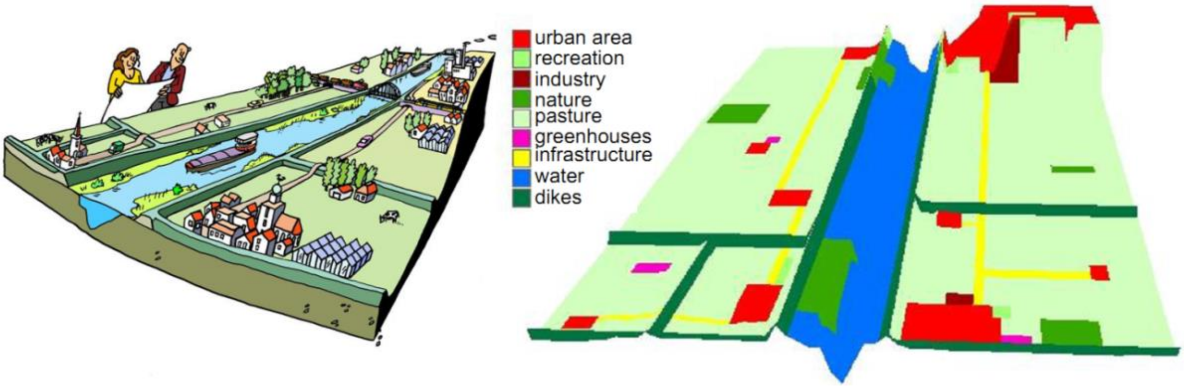


Figure 2-7. The Waas case study area (left) is heavily schematized (right) into a three-dimensional image of the floodplain presenting the land use and elevations (exaggerated vertically). The flow direction is from back to front (Haasnoot et al., 2012).

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 ten percent 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., 2011). 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., 2016b, 2015): costs, loss of life, and economic damages. However, due to the fact that the costs were rarely affected 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., 2016b, 2015) and 11 distinct adaptation pathways were identified. These optimal adaptation pathways are used in the following analysis.

2.6. 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 2.4. The ranking stability for each pair of metrics is displayed in Figure 2-8. 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 the Supporting Information. Figure 2-8 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 2-4.

Metrics		T_1		T_2		T_3		% 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 2-8. 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.

2.6.1. Impact of transformations

Figure 2-8 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 2-8, 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 2.4. For example, as can be seen in Figure 2-4, 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 2-8 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 2-4 and explains the large variance in the ranking stability of the maximin and maximax metrics in Figure 2-8. 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 2-8 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 categorised 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 2-8). 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.

2.6.2. Impact of relative performance

As can be seen in Figure 2-8, 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 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 2-4). For example, as can be seen in Figure 2-8, 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 (towards 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.

2.7. 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 socio-political 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 paper 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 paper 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 (i) the information the decision context relates to most (e.g. absolute performance, regret, or the satisfaction of constraints) (performance value transformation (*T_1*)), (ii) the preference of a decision-maker towards a high or low level of risk aversion for the case study under consideration through scenario subset selection (*T_2*), and (iii) the decision-maker's preference towards 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.

Acknowledgements

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 paper significantly. The case study robustness data is included in the Supporting Information.

References

Bandi, C., 2012. Robust Optimization. Princeton University Press.

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

Beh, E., Maier, H.R., Dandy, G.C., 2015. Scenario driven optimal sequencing under deep uncertainty. Environ. Model. Softw. 68, 181–195. https://doi.org/10.1016/j.envsoft.2015.02.006

Beh, E., Zheng, F., Dandy, G.C., Maier, H.R., Kapelan, Z., 2017. Robust optimization of water infrastructure planning under deep uncertainty using metamodels. Environ. Model. Softw. 93, 92–105. https://doi.org/10.1016/j.envsoft.2017.03.013

Ben-Haim, Y., 2004. Uncertainty, probability and information-gaps. Reliab. Eng. Syst. Saf. 85, 249–266. https://doi.org/10.1016/j.ress.2004.03.015

Bertsimas, D., Sim, M., 2004. The Price of Robustness. Oper. Res. 52, 35–53. https://doi.org/10.1287/opre.1030.0065

Bilal, K., Manzano, M., Khan, S., Calle, E., Li, K., Zomaya, a, 2013. On the Characterization of the Structural Robustness of Data Center Networks. IEEE Trans. Cloud Comput. 1, 64–77.

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

Burn, D.H., Venema, H.D., Simonovic, S.P., 1991. Risk-Based Performance Criteria for Real-Time Reservoir Operation. Can. J. Civ. Eng. 18, 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: Proceedings - IEEE International Conference on Cluster Computing, ICCC. IEEE, pp. 558–567. https://doi.org/10.1109/CLUSTR.2007.4629283

Castelletti, A., Galelli, S., Restelli, M., Soncini-Sessa, R., 2010. Tree-based reinforcement learning for optimal water reservoir operation. Water Resour. Res. 46. https://doi.org/10.1029/2009WR008898

Clark, M.P., Fan, Y., Lawrence, D.M., Adam, J.C., Bolster, D., Gochis, D.J., Hooper, R.P., Kumar, M., Leung, L.R., Mackay, D.S., Maxwell, R.M., 2015. Water Resources Research. Water Resour. Res. 51, 1–27. https://doi.org/10.1002/2015WR017096.Received

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

De Bruijn, K.M., 2008. Bepalen van schade ten gevolge van overstromingen. Voor verschillende scenario's en bij verschillende beleidsopties. Deltares Rep. Q 4345.

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. Adv. Water Resour. 103, 51–63. https://doi.org/10.1016/j.advwatres.2017.02.012

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

Drouet, L., Bosetti, V., Tavoni, M., 2015. Selection of climate policies under the uncertainties in the Fifth Assessment Report of the IPCC. Nat. Clim. Chang. 5, 937–940.

Giuliani, M., Anghileri, D., Castelletti, A., Vu, P.N., Soncini-Sessa, R., 2016a. Large storage operations under climate change: Expanding uncertainties and evolving tradeoffs. Environ. Res. Lett. 11, 35009. https://doi.org/10.1088/1748-9326/11/3/035009

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

Giuliani, M., Castelletti, A., Fedorov, R., Fraternali, P., 2016b. Using crowdsourced web content for informing water systems operations in snow-dominated catchments. Hydrol. Earth Syst. Sci. 20, 5049–5062. https://doi.org/10.5194/hess-20-5049-2016

Giuliani, M., Castelletti, A., Pianosi, F., Mason, E., Reed, P.M., 2016c. Curses, Tradeoffs, and Scalable Management: Advancing Evolutionary Multiobjective Direct Policy Search to Improve Water Reservoir Operations. J. Water Resour. Plan. Manag. 142, 04015050. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000570

Giuliani, M, Li, Y., Castelletti, A., Gandolfi, C., 2016. A coupled human-natural systems analysis of irrigated agriculture under changing climate. Water Resour. Res. 52, 6928–6947. https://doi.org/10.1002/2016WR019363

Grafton, Q., Horne, J., Wheeler, S.A., 2016a. On the Marketisation of Water: Evidence from the Murray-Darling Basin, Australia. Water Resour. Manag. 30, 913–926. https://doi.org/10.1007/s11269-015-1199-0

Grafton, Q., Mclindin, M., Hussey, K., Wyrwoll, P., Wichelns, D., Ringler, C., Garrick, D., Pittock, J., Wheeler, S., Orr, S., Matthews, N., Ansink, E., Aureli, A., Connell, D., De Stefano, L., Dowsley, K., Farolfi, S., Hall, J., Katic, P., Lankford, B., Leckie, H., Mccartney, M., Pohlner, H., Ratna, N., Rubarenzya, M.H., Sai Raman, S.N., Wheeler, K., Williams, J., 2016b. Responding to Global Challenges in Food, Energy, Environment and Water: Risks and Options Assessment for Decision-Making. Asia Pacific Policy Stud. 3, 275–299. https://doi.org/10.1002/app5.128

Guariso, G., Rinaldi, S., Sessa, R.S., 1985. Decision support systems for water management: the Lake Como case study. Eur. J. Oper. Res. 21, 295–306.

Guariso, G., Rinaldi, S., Soncini-Sessa, R., 1986. The Management of Lake Como: A Multiobjective Analysis. Water Resour. Res. 22, 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. Environ. Model. Softw. 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 Futur. 5, 285–303. https://doi.org/10.1002/2016EF000433

Haasnoot, M., Middelkoop, H., Beek, E. Van, Deursen, W.P.A. Van, 2011. A method to develop management strategies for an uncertain future. Sustain. Dev. 381, 369–381. https://doi.org/10.1002/sd

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. Clim. Change 115, 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. Environ. Model. Softw. 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 Anal. 32, 1657–1672. https://doi.org/10.1111/j.1539-6924.2012.01802.x

Hamarat, C., Kwakkel, J., Pruyt, E., 2014. An exploratory approach for adaptive policymaking by using multi-objective robust optimization. Simul. Model. Pract. Theory 46, 25–39. https://doi.org/10.1016/j.simpat.2014.02.008

Hashimoto, T., Loucks, D.P., Stedinger, J.R., 1982a. Robustness of water resources systems. Water Resour. Res. 18, 21–26. https://doi.org/10.1029/WR018i001p00021

Hashimoto, T., Stedinger, J.R., Loucks, D.P., 1982b. Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. Water Resour. Res. 18, 14–20. https://doi.org/10.1029/WR018i001p00014

Herman, J.D., Reed, P.M., Zeff, H.B., Characklis, G.W., 2015. How Should Robustness Be Defined for Water Systems Planning under Change? J. Water Resour. Plan. Manag. 141, 04015012. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000509

Howard M., J. E., R.A., 1984. Influence diagrams. Readings Princ. Appl. Decis. Anal. Vol. 2, Strateg. Decis. Gr. 2, 719–762.

Howard, R.A., 1966. Information Value Theory. Syst. Sci. 2, 22–26. https://doi.org/10.1109/TSSC.1966.300074

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

Iglesias, A., Garrote, L., 2015. Adaptation strategies for agricultural water management under climate change in Europe. Agric. Water Manag. 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. Environ. Model. Softw. 42, 55–71. https://doi.org/10.1016/j.envsoft.2012.12.007

Korteling, B., Dessai, S., Kapelan, Z., 2013. Using Information-Gap Decision Theory for Water Resources Planning Under Severe Uncertainty. Water Resour. Manag. 27, 1149–1172. https://doi.org/10.1007/s11269-012-0164-4

Kwakkel, J.H., Eker, S., Pruyt, E., 2016a. How robust is a robust policy? Comparing alternative robustness metrics for robust decision-making, in: International Series in Operations Research and Management Science. Springer, pp. 221–237.

https://doi.org/10.1007/978-3-319-33121-8_10

Kwakkel, J.H., Haasnoot, M., Walker, W.E., 2016b. Comparing Robust Decision-Making and Dynamic Adaptive Policy Pathways for model-based decision support under deep uncertainty. Environ. Model. Softw. 86, 168–183. https://doi.org/10.1016/j.envsoft.2016.09.017

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. Clim. Change 132, 373–386. https://doi.org/10.1007/s10584-014-1210-4

Kwakkel, J.H., Walker, W.E., Marchau, V.A.W.J., 2010. Classifying and communicating uncertainties in model-based policy analysis. Int. J. Technol. Policy Manag. 10, 299. https://doi.org/10.1504/IJTPM.2010.036918

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. Rand Corporation. https://doi.org/10.1016/j.techfore.2003.09.006

Lempert, R.J., Bryant, B.P., Bankes, S.C., 2008. Comparing Algorithms for Scenario Discovery. RAND Infrastructure, Saf. Environ. 27, 1–35.

Lempert, R.J., Groves, D.G., Popper, S.W., Bankes, S.C., 2006. A General, Analytic Method for Generating Robust Strategies and Narrative Scenarios. Manage. Sci. 52, 514–528. https://doi.org/10.1287/mnsc.1050.0472

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? Environ. Model. Softw. 81, 154–164. https://doi.org/10.1016/j.envsoft.2016.03.014

Maier, H.R., Lence, B.J., Tolson, B.A., Foschi, R.O., 2001. First order reliability method for estimating reliability, vulnerability, and resilience. Water Resour. Res. 37, 779–790.

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

Monaco, R.M., 1992. Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis, International Journal of Forecasting. Cambridge University

Press, Cambridge, UK. https://doi.org/10.1016/0169-2070(92)90021-Z

Paton, F.L., Dandy, G.C., Maier, H.R., 2014a. Integrated framework for assessing urban water supply security of systems with non-traditional sources under climate change. Environ. Model. Softw. 60, 302–319. https://doi.org/10.1016/j.envsoft.2014.06.018

Paton, F.L., Maier, H.R., Dandy, G.C., 2014b. Including adaptation and mitigation responses to climate change in a multiobjective evolutionary algorithm framework for urban water supply systems incorporating GHG emissions. Water Resour. Res. 50, 6285–6304. https://doi.org/10.1002/2013wr015195

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 Resour. Res. 49, 1643–1667. https://doi.org/10.1002/wrcr.20153

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

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

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

Ravalico, J.K., Maier, H.R., Dandy, G.C., 2009. Sensitivity analysis for decision-making using the MORE method-A Pareto approach. Reliab. Eng. Syst. Saf. 94, 1229–1237. https://doi.org/10.1016/j.ress.2009.01.009

Ray, P.A., Watkins Jr, D.W., Vogel, R.M., Kirshen, P.H., 2013. Performance-based evaluation of an improved robust optimization formulation. J. Water Resour. Plan. Manag. 140, 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. J. Water Resour. Plan. Manag. 142, 04016028. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000660

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

Savage, L.J., 1951. The theory of statistical decision. J. Am. Stat. Assoc. 46, 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 Decis. 15, 321–336. https://doi.org/10.1007/BF00162111

Simon, H.A., 1956. Rational choice and the structure of the environment. Psychol. Rev. 63, 129–138. https://doi.org/10.1037/h0042769

Starr, M.K., 1962. Product Design and Decision Theory. Prentice-Hall.

Takriti, S., Ahmed, S., 2004. On robust optimization of two-stage systems. Math. Program. 99, 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., 1951. Statistical decision functions, Nature. New York; Chapman & Hall: London. https://doi.org/10.1038/1671044b0

Walker, W.E., Lempert, R., Kwakkel, J., 2013. Deep Uncertainty, in: Encyclopedia of Operations Research and Management Science. Springer, pp. 395–402. https://doi.org/10.1007/978-1-4419-1153-7_1140

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, 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 Resour. Manag. 12, 13–30. https://doi.org/10.1007/s11269-006-9040-4

Chapter 3

Impact of scenario selection on robustness

Cameron McPhail, Holger R. Maier, Seth Westra, Jan Kwakkel, and Leon van der Linden

Accepted for publication in Water Resources Research

Statement of Authorship

Title of Paper	Impact of scenario selection on robustness
Publication Status	☐ Published ☑ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Accepted by Water Resources Research

Principal Author

Name of Principal Author (Candidate)	Cameron McPhail		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote manuscript, made edits to manuscript, and was corresponding author.		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	16/08/2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Holger Maier		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature		Date	17/08/20

Name of Co-Author	Seth Westra		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature		Date	18/08/2020

Name of Co-Author	Jan Kwakkel		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature		Date	18-08-2020

Name of Co-Author	Leon van der Linden		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature		Date	19/08/2020

Abstract

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

3.1. Introduction

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

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

Scenarios are generally classified into three different types: predictive (“what is likely to happen”), explorative (“what could happen”?), or normative (“how can a specific future be realized”?) (Maier et al., 2016). A number of water resources studies have generated explorative scenarios by considering the impact of plausible changes in atmospheric carbon concentrations (Anghileri et al., 2018; Beh et al., 2014, 2015b, 2015a; Giuliani, Castelletti, et al., 2016; Giuliani & Castelletti, 2016; Haasnoot et al., 2012, 2013; Herman & Giuliani, 2018; Huskova et al., 2016; McPhail et al., 2018), as well as plausible changes in regional socio-economic conditions (Haasnoot et al., 2013; Wada et al., 2019). In contrast, normative scenarios consider conditions that represent interesting outcomes, as is the case with scenario discovery (e.g. Bryant and Lempert. (2010); Groves & Lempert (2007); Hadka et al. (2015); Kasprzyk et al. (2013); Kwakkel (2017); Kwakkel, Haasnoot, et al. (2016); Matrosov et al. (2013); Trindade et al. (2017)); conditions that result in one decision alternative being preferable to another, as is the case with MORE (Ravalico et al., 2010), POMORE (Ravalico et al., 2009) and decision scaling (e.g. Brown et al.

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

How many scenarios are generated is generally linked to the philosophy that underpins scenario generation. When scenarios correspond to coherent descriptions of alternative hypothetical futures (e.g. van Notten (2005)), the number of scenarios considered is generally small (~3-9, see Table 1 – Supporting Information) and scenarios are generally identified using some type of human input, such as the use of participatory approaches involving a variety of stakeholders (e.g. Wada et al. (2019)). In contrast, when scenarios are designed to represent a broad range of combined changes in future conditions, the number of scenarios considered is generally large (~100-15,000, see Table 1 – Supporting Information) and scenarios are generated using numerical modelling and/or sampling- or optimization-based approaches, with minimal stakeholder input (e.g. Trindade et al. (2017); Culley et al. (2016, 2019); Hadka et al. (2015); Hall et al. (2012); Herman et al. (2014, 2015); Kasprzyk et al. (2013); Kwakkel (2017); Kwakkel et al. (2015); Kwakkel, Walker, et al. (2016); McPhail et al. (2018); Quinn et al. (2017, 2018); Singh et al. (2015); Trindade et al. (2017); Watson & Kasprzyk (2017); Weaver et al. (2013); Zeff et al. (2014)).

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

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

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

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

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

The remainder of the paper is organized as follows: Section 3.2 introduces the methodology for quantifying and visualizing the effect of the selection of different sets of scenarios on robustness and the ranking of decision alternatives; Section 3.3 describes how this methodology was applied to the Lake Model; and Section 3.4 shows the results of this analysis, along with a discussion of the effects of different sets of scenarios on robustness and on the rankings of decision alternatives. This is followed by a summary and conclusions in Section 3.5.

3.2. Generic approach for assessing the influence of scenario selection on robustness

To quantify the impact of scenario selection/creation on robustness (Aim 2a) and on the rankings of decision alternatives (Aim 2b), we propose the approach presented in Figure 3-1. The approach compares outcomes from applying two distinct sets of scenarios, and thus provides insight into the sensitivity of those outcomes on the method of scenario selection. Thus, the proposed approach is generic, as it can cater to and is independent of the approach used to create the sets of scenarios—including aspects such as the number of scenarios considered, the distribution of scenarios over the scenario space, the method used to generate the scenarios (e.g. sampling or using stakeholder input) etc. (see Section 3.1).

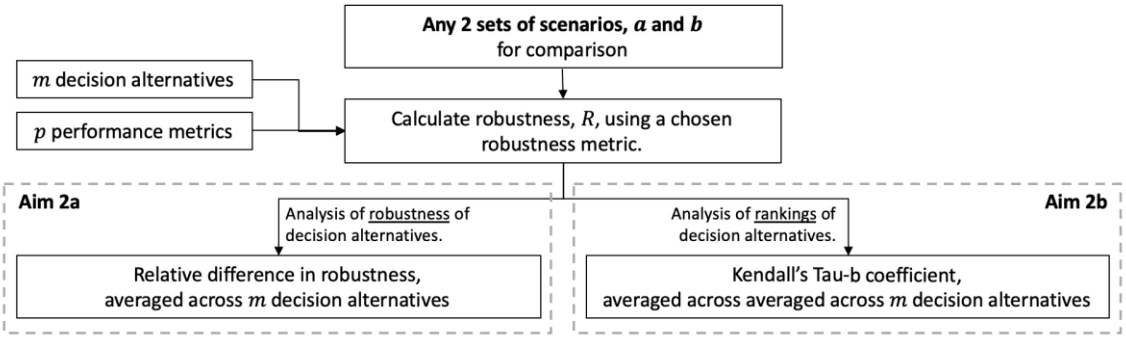

Any 2 sets of scenarios, a and b for comparison
m decision alternatives
p performance metrics
Calculate robustness, R, using a chosen robustness metric.
Aim 2a
Analysis of robustness of decision alternatives.
Relative difference in robustness, averaged across m decision alternatives
Aim 2b
Analysis of rankings of decision alternatives.
Kendall's Tau-b coefficient, averaged across averaged across m decision alternatives

Figure 3-1. Approach for the quantitative analysis of the influence of any two sets of scenarios on the robustness and ranking of decision alternatives.

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

The methodology used for assessing the impact of two sets of scenarios on the robustness values is shown in Figure 3-2. For a single decision alternative and single robustness metric, the two different sets of scenarios produce one robustness value each, and these two robustness values are compared. This difference is then averaged across all *m* decision alternatives.

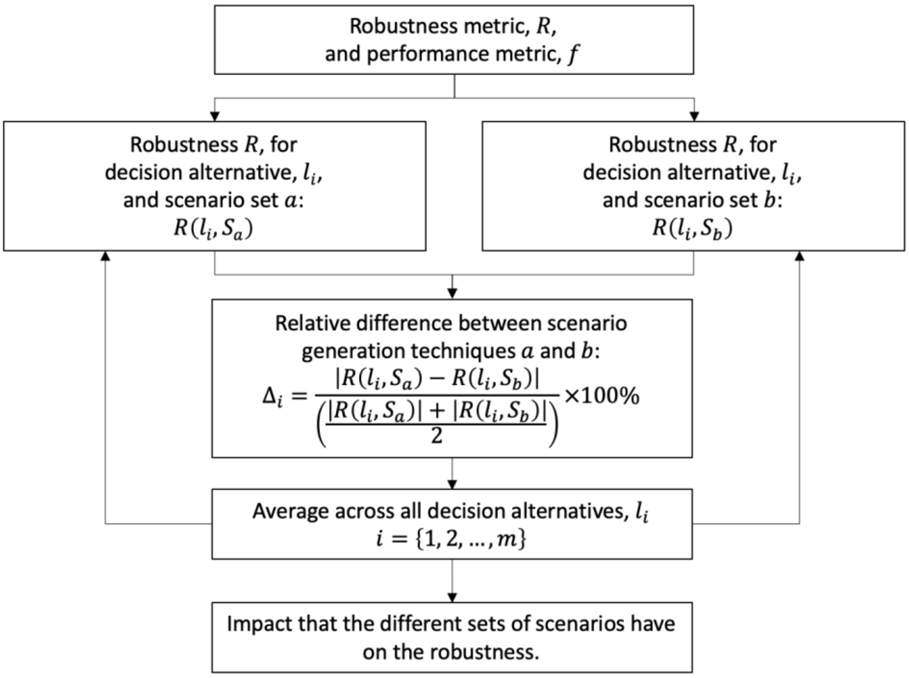

Robustness metric, R, and performance metric, f
Robustness R, for decision alternative, l_i, and scenario set a: $R(l_i, S_a)$
Robustness R, for decision alternative, l_i, and scenario set b: $R(l_i, S_b)$
Relative difference between scenario generation techniques a and b: $\Delta_i = \frac{|R(l_i,S_a) - R(l_i,S_b)|}{\left(\frac{|R(l_i,S_a)| + |R(l_i,S_b)|}{2}\right)} \times 100\%$
Average across all decision alternatives, l_i $i = \{1, 2, \ldots, m\}$
Impact that the different sets of scenarios have on the robustness.

Figure 3-2. Calculation of the sensitivity of robustness to different sets of scenarios.

The methodology used for assessing the impact of two different sets of scenarios on ranking similarity (i.e. relative robustness) is shown in Figure 3-3. We begin with the robustness of all *m* decision alternatives when using one set of scenarios and compare this to the robustness of the same *m* decision alternatives evaluated with a different set of scenarios. These two sets of robustness values are compared using Kendall's rank correlation. This statistical metric tests how similarly two quantities are ranked. In this case, we are testing how the decision alternatives are ranked when robustness is calculated twice, each time with a different set of scenarios.

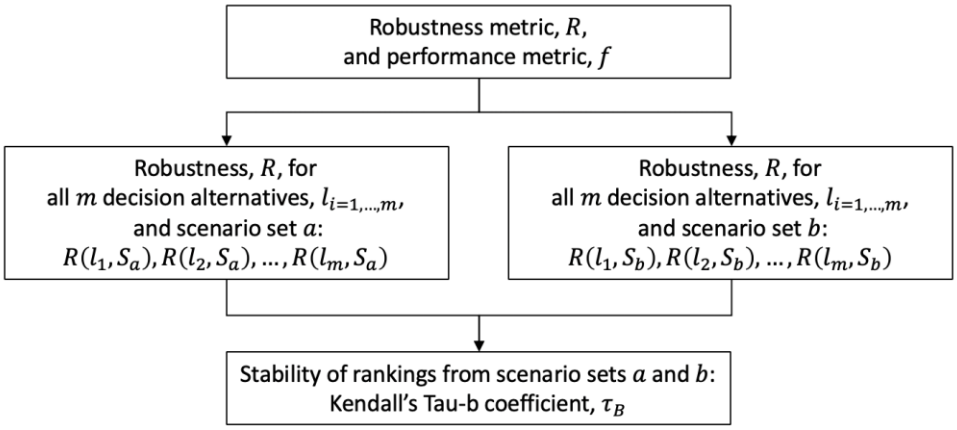

Robustness metric, R, and performance metric, f
Robustness, R, for all m decision alternatives, $l_{i=1,\dots,m}$, and scenario set a: $R(l_1,S_a), R(l_2,S_a), \dots, R(l_m,S_a)$
Robustness, R, for all m decision alternatives, $l_{i=1,\dots,m}$, and scenario set b: $R(l_1,S_b), R(l_2,S_b), \dots, R(l_m,S_b)$
Stability of rankings from scenario sets a and b: Kendall's Tau-b coefficient, τ_B

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

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

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

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

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

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

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

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

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

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

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

3.3. Case study

3.3.1. Background

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

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

The performance metric values are influenced by the decision alternatives and scenarios. The decision alternatives represent the annual pollution control strategies that the inhabitants of the town implement (i.e. they define the annual quantity of industrial pollution that is allowed to enter the lake for each year in the 100 year planning horizon). A reduction in annual pollution improves reliability and maximum phosphorous (by increasing the number of years the system is below the pollution threshold and minimizing the maximum level of phosphorus). However, this decreases economic utility (because it costs money).

3.3.2. Scenario set generation

In principle, the Lake Problem can be represented using a range of qualitative and quantitative approaches, with important choices related to system model boundaries, process representations and other key modelling considerations. In the particular case considered in this paper, a system model (referred to as the "Lake Model") is used, as it represents a trusted numerical representation of the system that has reasonable fidelity in simulating key system processes (Carpenter et al., 1999; Lempert & Collins, 2007). System model selection represents a key consideration in model-based assessments, and the system model boundaries effectively delineate the scenarios that are required as model inputs. These inputs are described in Table 3-1, with the set of valid combinations of scenarios depicted as a five-dimensional hypercube with plausible bounds selected based on previous studies (Kwakkel (2017); Quinn et al. (2017); Eker & Kwakkel (2018)), as given in Table 3-1.

Table 3-1. Deeply uncertain scenario variables (model inputs) and associated ranges of values for the Lake Problem.

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

The objectives of the case study are to emulate the impact of the diversity of scenario selection approaches used in the water resources literature, as summarized in Section 3.1, on absolute and relative robustness values. However, regardless of which scenario selection approach is used, for a quantitative study such as the Lake Problem, the outcome of the scenario selection step needs to be the quantitative specification of inputs to the system model (i.e. points in the five-dimensional hypercube that represents the input parameter space for the Lake Model). Thus, several sampling strategies are used to generate the requisite Lake Model inputs, which encapsulate key features of alternative scenario generation techniques, including:

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

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

Illustrative examples of the resulting differences in the distributions of the scenarios obtained are shown in Figure 3-4.

- **Diverse:** Figure 3-4 (a) depicts the situation where four diverse futures are first identified (analogous to RCPs) with many samples taken around each of these four points (analogous to the use of multiple global and regional climate models to create multiple downscaled realizations of each of the RCPs) of which there are many examples in the water resources literature (Anghileri et al., 2018; Giuliani, Castelletti, et al., 2016; Giuliani & Castelletti, 2016; Haasnoot et al., 2012, 2013; Herman & Giuliani, 2018; Huskova et al., 2016; McPhail et al., 2018).
- **Targeted:** Figure 3-4 (b) depicts a targeted approach to identifying samples that cover “interesting” regions of the system model space, for the situation where the model performance responds monotonically to each input (i.e. an increase in one variable always results in increased or decreased performance). This can occur when two model inputs (e.g. water supply and water demand) are lined up from worst to best, and the two worst values (e.g. lowest water supply and highest water demand) are paired, etc., leading to a clear set of worst to best points in the hypercube (Beh et al., 2014, 2015a, 2015b).
- **Uniform:** Figure 3-4 (c) depicts a uniform sampling of the entire hypercube to consider a wide range of plausible futures, as is often done in the water resources literature (Culley et al., 2016, 2019; Hadka et al., 2015; Hall et al., 2012; Herman et al., 2015; Kasprzyk et al., 2013; Kwakkel, 2017; Kwakkel et al., 2015; Kwakkel, Walker, et al., 2016; McPhail et al., 2018; Quinn et al., 2017, 2018; Singh et al., 2015; Trindade et al., 2017; Watson & Kasprzyk, 2017; Weaver et al., 2013; Zeff et al., 2014).

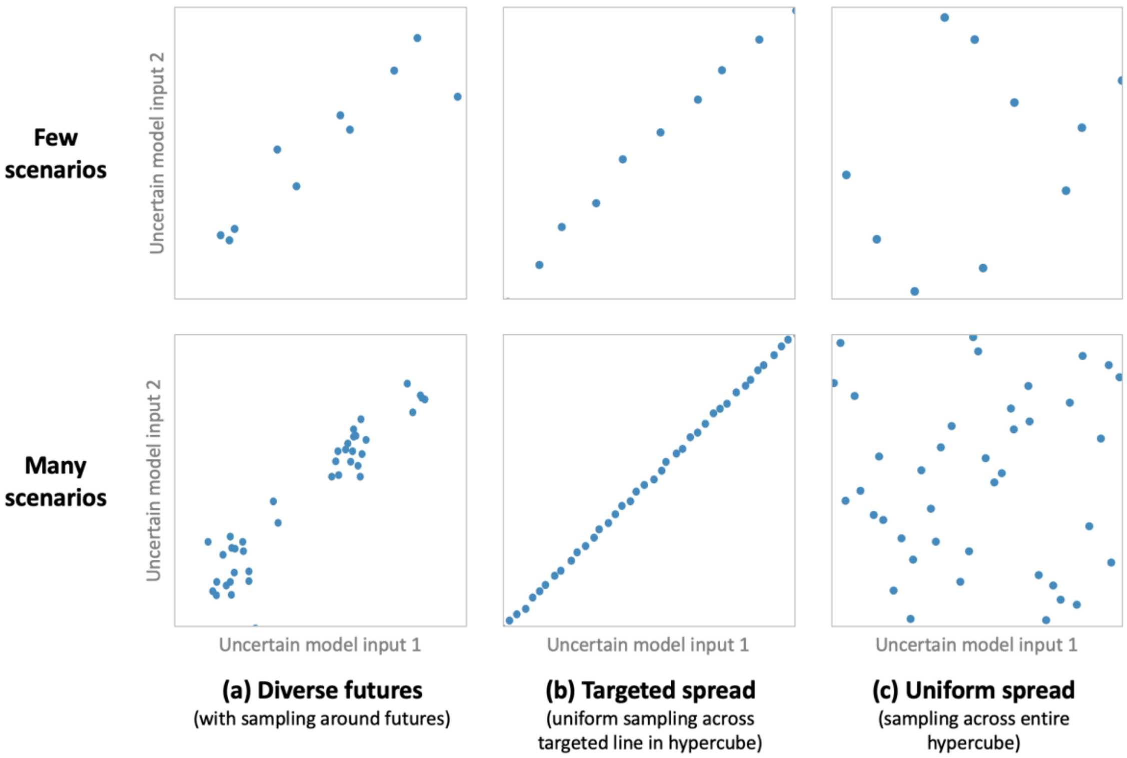


Figure 3-4. Two-dimensional illustration of how the three distributions of scenarios are implemented for the case study, with examples for both a small and large number of scenarios, emulating the diversity in scenarios that could be obtained by using different scenarios selection approaches.

3.3.3. Decision alternatives and performance values

Robustness values are determined relative to potential decision alternatives, and in this analysis we consider 4,611 such alternatives. These were obtained using a many-objective evolutionary algorithm to identify a set of Pareto optimal decision alternatives for a reference scenario, as is recommended in many-objective robust decision making (Kasprzyk et al., 2013). Specifically, we used a generational version of the BORG algorithm (Hadka & Reed, 2013), to allow for easy parallelization to reduce run times. The generational version of BORG uses autoadaptive operator selection, restarts for stalled search, and adaptive population sizing from BORG (Hadka & Reed, 2013), within the generational e-NSGA2 structure. As a stopping condition, we used 500,000 function evaluations, while convergence was assessed using hypervolume and epsilon progress (Reed et al., 2013; Ward et al., 2015). We repeated this for 50 different initial random seeds and merged the final results into one large set of final decision alternatives. For each decision alternative, three performance values are produced per simulation (described in more detail in Section 3.4.1) and per scenario, leading to a total of 248,994,000 performance values (i.e. the product of 18,000 scenarios that were grouped into 300 sets of scenarios, 4,611 decision alternatives and three performance metrics).

3.3.4. Robustness metrics

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

Table 3-2.Robustness metrics used in analysis

Metric name	**Brief description**
Maximin	Worst-case performance (high level of risk aversion).
Maximax	Best-case performance (low level of risk aversion).
Hurwicz's Optimism-Pessimism Rule	Weighted sum of the best- and worst-cases.
Laplace's Principle of Insufficient Reason	Mean performance.
Minimax Regret	The worst-case cost of making a wrong decision in any given scenario (high level of risk aversion).
90th Percentile Minimax Regret	The 90th percentile cost of making a wrong decision (high level of risk aversion) (percentile-based calculation).
Mean-variance	A function of the mean and variance in performance.
Undesirable Deviations	The sum of performance below the median performance.

Percentile-based Skewness	The skew of performance (towards high- or low-performance) (percentile-based calculation).
Percentile-based Peakedness	The kurtosis (peakedness) of performance (percentile-based calculation).
Starr's Domain Criterion	Calculates the proportion of scenarios with acceptable levels of performance.

3.4. Results and discussion

3.4.1. Robustness values

Following the methodology outlined in Figure 3-2, the sensitivity of each robustness metric to the different distributions of scenarios is shown in Figure 3-5. The sensitivity is the percentage difference between the robustness calculated for two different sets of scenarios, and this is averaged across all of the Pareto-optimal decision alternatives (as described in Figure 3-2) from each of the 50 optimization runs. Orange and red represents high sensitivity (i.e. >10% difference in robustness for the two different sets of scenarios) and purple and blue represents low sensitivity (i.e. <10% difference in robustness for the two different sets of scenarios). The robustness values (and therefore the sensitivity of the robustness values) is calculated using the distribution of scenarios in the scenario space (e.g. diverse futures or uniform spread), the decision alternatives, the performance metric (e.g. reliability), and the robustness metric (e.g. Maximin). The decision alternatives are purely case-specific while the other three factors are more general; therefore we have presented the results in Figure 3-5 in a way that allows the scenario distribution, performance metric, and robustness metric to be compared one-by-one, or in combination.

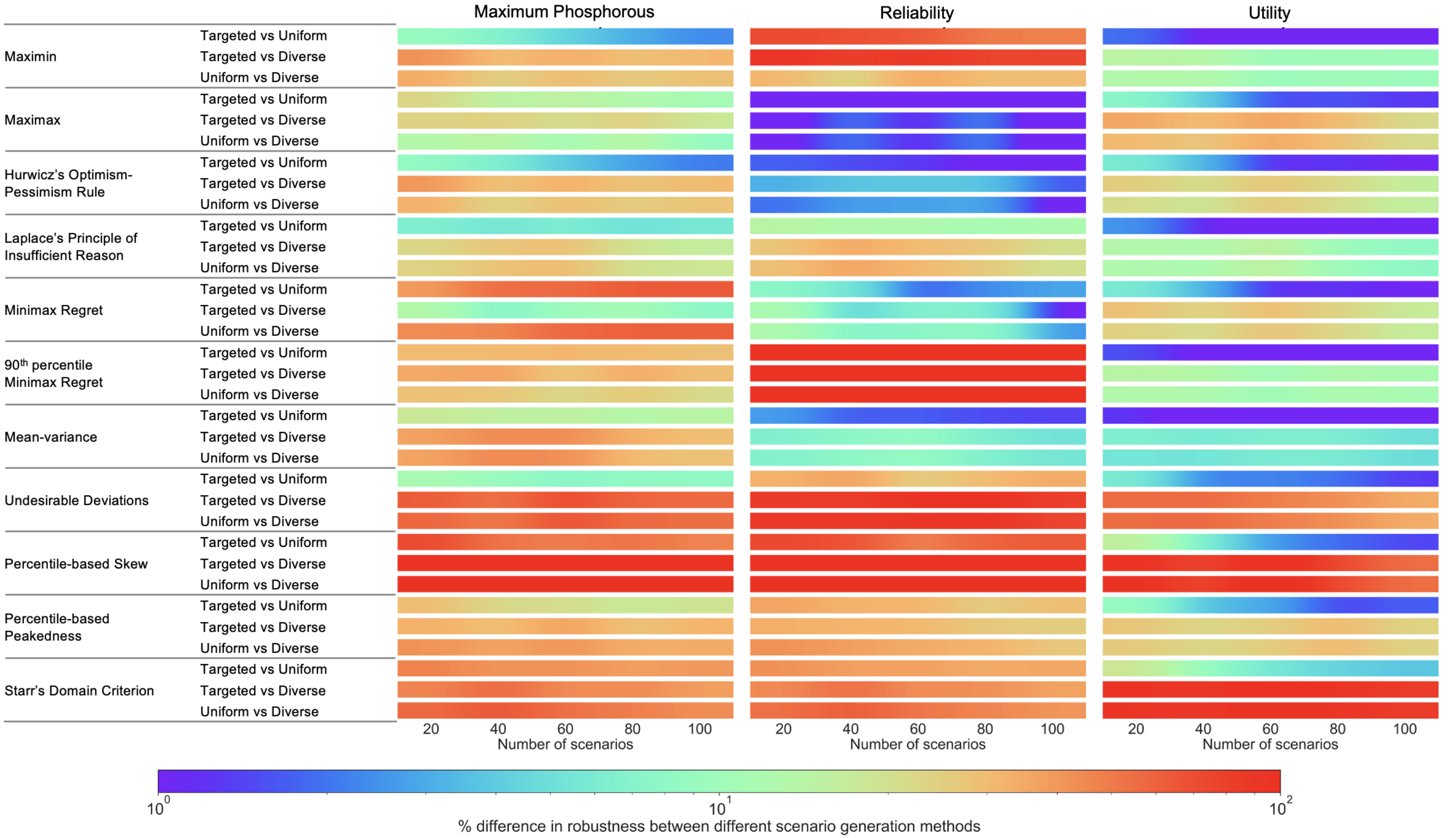


Figure 3-5. Sensitivity of the robustness metrics (as measured by the percentage difference), for each of the case study performance metrics (maximum phosphorous, utility, and reliability) for each distribution of scenarios in the scenario space (diverse futures, uniform spread, targeted spread). Red represents high sensitivity and purple represents low sensitivity of the robustness metric to the set of scenarios.

Overall, Figure 3-5 indicates that scenario selection has a large impact on robustness values. This is evidenced by the fact that the bars are generally green, orange, or red when comparing the robustness values obtained when different scenario sets are used (indicating a difference in robustness values in excess of 10%) (Figure 3-5). This is most likely because the different sets of scenarios are covering very different areas of the scenario space (Figure 3-4), and therefore different input variables are being used by the model to determine system performance and robustness.

The results also show that differences in robustness methods between scenarios that represent a uniform spread and scenarios that represent a targeted spread are smaller than those between the other two combinations of distributions of scenarios (Figure 3-5), particularly for the reliability performance metric. To explain why this occurs, Figure 3-6 shows one set of scenarios for each of the different scenario distributions, overlaid on the performance values for a 2D subspace of the scenario space for a single decision alternative. Figure 3-6 indicates that the scenario space is covered very differently by scenarios that represent diverse futures, a uniform spread and a targeted spread. In particular, scenarios are spread across all levels of performance when the set represents a uniform spread or targeted spread (all colors in Figure 3-6), however, some levels of performance (some colors) will be missed when there is a clustering of scenarios, as happens when the scenarios are representative of diverse futures, particularly if there are thresholds in performance (e.g. for reliability and maximum phosphorous). The similarity in coverage of the performance values by the distribution of scenarios representing a uniform spread and targeted spread leads them to produce more similar values of robustness relative to the distribution of scenarios that is representative of diverse futures.

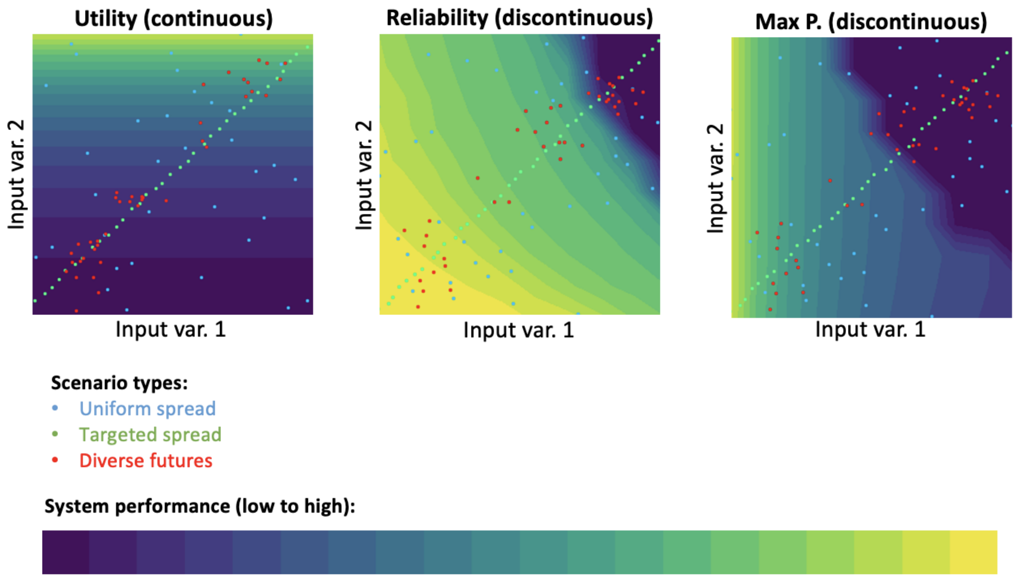


Figure 3-6. Illustration of how different sets of scenarios will sample different points in the space of system performance values for The Lake Problem. Robustness is calculated by the sampled system performance values and therefore affected by the distribution of scenarios and system performance values.

As mentioned above, the degree of similarity in robustness values can be affected by the distribution of the performance values. For example, when considering the utility metric column in Figure 3-5, it can be seen that there are significantly fewer orange and red bars, which indicates high similarity in robustness values. The utility metric shows slightly more similarity in robustness values when the distribution of scenarios is representative of diverse futures, but much greater similarity when the scenarios correspond to a uniform spread or targeted spread. Figure 3-6 illustrates that the performance values for the utility metric form a smooth and continuous space, relative to the reliability and maximum phosphorous metrics, which have sharp gradients and non-linearities (due to tipping points in the environmental dynamics of the Lake Problem). This leads to correspondingly higher dissimilarity in robustness values for the latter metrics.

In most instances, the number of scenarios considered does not have a significant effect on the relative similarities or differences in the robustness values obtained using the different distributions of scenarios throughout the scenario space (Figure 3-5). This indicates that the way that the scenarios cover the scenario space (i.e. diverse futures, uniform spread or targeted spread) plays a greater role in determining robustness values than the number of scenarios used for each approach. For situations where there is a gradient in robustness values with an increase in the number of scenarios, the level of agreement in robustness values increases as the number of scenarios increases.

Figure 3-7 summarizes the trends in similarity in robustness values from Figure 3-5, with reference to the factors affecting this similarity (Figure 3-6). Figure 3-7 highlights (using examples) that, in general, a dissimilar coverage of the scenario space (e.g. the diverse futures vs. uniform spread scenarios, as discussed previously) will lead to a lower degree of similarity in robustness (Example A in Figure 3-7), and a more similar coverage of the scenario space (e.g. targeted spread vs. uniform spread scenarios) leads to a higher degree of similarity in robustness (Example F in Figure 3-7). However, the interactions of the distribution of scenarios, the behavior of the system performance metrics, and the behavior of the robustness metrics (Figure 3-6) are complex, and so there are exceptions to these findings.

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

	E.g. Uniform spread vs Targeted spread			🟥	🟧
H	Similar E.g. Uniform spread vs Targeted spread	Discontinuous, bounded E.g. Reliability	Extreme low or high risk averseness E.g. Maximin, Maximax	Very dissimilar	to very similar
I	Similar E.g. Uniform spread vs Targeted spread	Continuous E.g. Utility	Percentile-based E.g. Skewness	Neutral	
J	Similar E.g. Uniform spread vs Targeted spread	Continuous E.g. Utility	Most metrics E.g. Mean-variance	Very similar	

Figure 3-7. General indication of how different distributions of scenarios, different performance metrics, and different robustness metrics all affect the robustness of decision alternatives in for the Lake Problem.

An exception to the general findings is that the value of the Maximax robustness metric is insensitive to the distribution of scenarios used for the reliability system performance metric, especially if a sufficiently large number of scenarios is used. This is because almost any decision alternative will achieve 100% reliability if the uncertain model inputs affecting the pollution levels (e.g. the mean natural pollution inflow) are favorable. In other words, for almost any decision alternative, there is some favorable region of the scenario space where the decision alternative can achieve 100% reliability. Due to the Maximax metric selecting the scenario with the best performance, the robustness will always be 100%, regardless of the distribution of scenarios or the decision alternative. This highlights how a performance metric with bounds (e.g. reliability is bounded between 0% and 100%) can interact with some robustness metrics (e.g. Maximax and Maximin, which use the best- and worst-case performance, respectively), as highlighted in Examples C and H in Figure 3-7. Note that this effect is not seen for the Maximin metric in this case study, because the starting conditions for the lake do not allow for the possibility of 0% reliability and thus the reliability is always greater than 0% in practice.

Robustness metrics that use percentiles (e.g. the Undesirable Deviations metric, Percentile-based Skewness, and Percentile-based Peakedness) are sensitive to the distribution of scenarios in the scenario space because they are dependent on the higher order moments of the distribution of performance values, which can vary more significantly than mean performance (e.g. Laplace's Principle of Insufficient Reason), and also vary more significantly than bounded maximum and minimum performance (Maximax and Maximin metrics respectively) (see Examples D and I, Figure 3-7). Metrics that use percentiles were an exception to the generalized findings and it should be noted that these metrics were also found to behave very differently to the other metrics in McPhail et al. (2018).

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

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

3.4.2. Ranking similarity

Following the methodology outlined in Figure 3-3, the correlation of the performance values (i.e. similarity in how the decision alternatives are ranked) is shown in Figure 3-8. The similarity of the rankings of the decision alternatives is given by Kendall's Tau-b for two different sets of scenarios, and this is averaged across all decision alternatives and all random seeds (as described in Figure 3-3). A value of -1 (red) indicates that the two distributions of scenarios give perfectly opposite rankings for the decision alternatives and a value of 1 (blue) represents the case where the rankings are the same (regardless of how different the robustness values are). A value of 0 represents the case where there is no correlation between the two methods, and therefore this represents a low similarity in rankings. Figure 3-9 summarizes the results from Figure 3-8, highlighting that in general, the coverage of the scenario space has little to no impact on the ranking of decision alternatives, which are almost always ranked the same way. However, as with the analysis of robustness values (Section 3.5.1), there are some exceptions to the above findings for the rankings, which are due to the interactions between the distribution of scenarios, the behavior of the system performance metrics, and the behavior of the robustness metrics (see Figure 3-6).

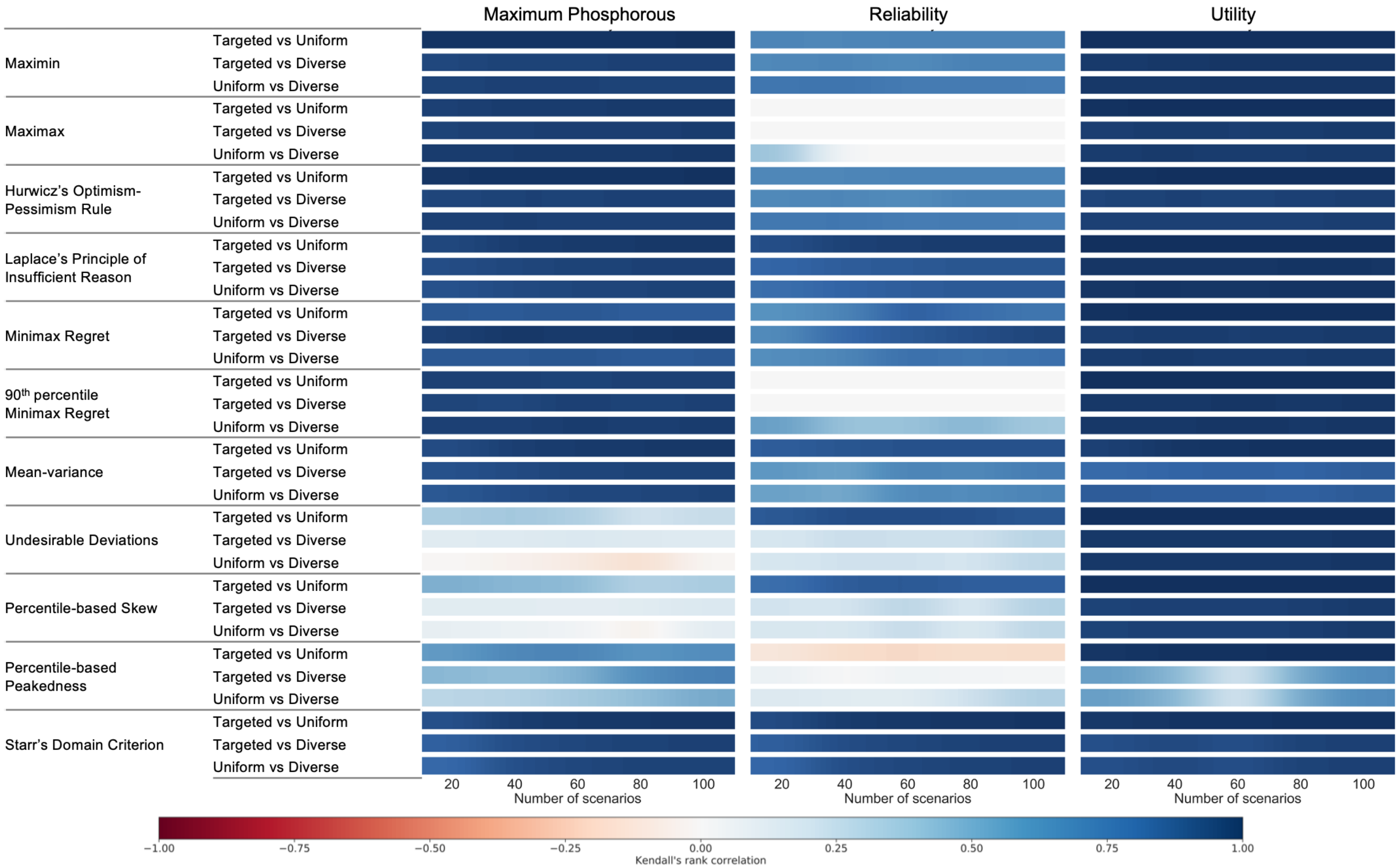


Figure 3-8. Similarity of the rankings of decision alternatives (as measured by Kendall's Tau-b) for each of the case study objectives (maximum phosphorous, utility, and reliability) for each pair of distributions of scenarios (diverse futures, uniform spread, targeted spread). Red or white represents low and blue represents high similarity (decision alternatives have the same rankings).

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

Figure 3-9. General indication of how different distributions of scenarios, different performance metrics, and different robustness metrics all affect the rankings of decision alternatives in for the Lake Problem.

Overall, Figure 3-8 indicates that for the majority of robustness values, the distribution of scenarios in the scenario space has a minor impact on the rankings of decision alternatives (with a few exceptions explained in more detail below). This is evidenced by the fact that much of Figure 3-8 is shaded dark blue, representing a positive correlation in the rankings of the decision alternatives when different distributions of scenarios are used to calculate robustness. This is likely because a high dissimilarity in robustness values (evidenced by much of Figure 3-5) does not necessarily mean a high dissimilarity in rankings. Therefore, although the robustness values may be very dissimilar when different scenario selection methods are used, the values are not changing relative to each other so that the same decision alternative would be selected as the most robust in both cases (i.e. the relative robustness of different decision alternatives is the same).

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

Some examples of exceptions to the above findings include that the metrics that consist of multiple percentiles (percentile-based skewness, peakedness) and the undesirable deviations metric can lead to dissimilar rankings in some cases (Examples A, C, F, and G in Figure 3-9) whereas most other metrics rank decision alternatives very similarly (see Figure 3-9). It should also be noted that McPhail et al. (2018) showed these same three robustness metrics to produce very dissimilar rankings when compared to other robustness metrics, even when the same scenarios were used in all robustness calculations.

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

To summarize Figure 3-8 and Figure 3-9, the rankings of decision alternatives is generally not strongly affected by scenario selection. However, there are some exceptions, based on the complex interactions between the behavior (e.g. smoothness vs. discontinuities) of the system performance metric (e.g. economic utility vs. reliability and maximum phosphorous) over the space of plausible model input values. The multi-faceted nature of the interactions between different aspects of the analysis means that while the overall methodology of assessing the impact of scenarios on the robustness analysis is generalizable, the specific results presented here are likely to be case-study specific.

3.5. Summary and conclusions

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

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

As a simplified example of how to apply this methodology, it was used to analyze the effect of three conceptually different distributions of scenarios (Figure 3-4). The methodology was applied to the Lake problem, using a variety of robustness metrics (Table 3-2). The results show that the distribution of scenarios has a significant effect on the robustness values calculated (Figure 3-5), but a small effect on how decision alternatives are ranked (i.e. relative robustness) (Figure 3-8). With regard to the degree of similarity of robustness values, the results indicated that dissimilar coverage of the scenario space (e.g. a diverse set of futures compared to a uniform spread) generally led to a lower degree of similarity in robustness values, in contrast to a similar coverage of the scenario space (e.g. a uniform spread and a targeted spread), which led to a higher degree in similarity of robustness values. Similarity of the robustness values is also affected by complex interactions of scenario selection with the number of scenarios, the behavior (e.g. smoothness, discontinuities) of the system performance metric over the space of plausible model input values, and the robustness metric itself (Figure 3-6). In contrast to the robustness values, it was found that the rankings of the decision alternatives based on robustness values often had a moderate to high degree of similarity when different sets of scenarios are used. Again, exceptions to this were caused by certain combinations of the behavior of the system performance metric and the characteristics of the robustness metric used.

The effects of several distributions of scenarios have been assessed using both theoretical and computational evidence, but the results presented are by no means representative of all combinations of scenario selections, robustness metrics, case studies etc. This study used many stochastic simulations to highlight that scenarios can have an effect, but in order to see the effect on real-life decision-making, further investigation is warranted.

One way to explore the effects on decision-making could be through simulation gaming workshops with students, followed by workshops with decision-makers, case-studies of successful long-term infrastructure plans, and the creation of carefully designed pilot studies to compare these approaches, as recommended by Kwakkel & van der Pas (2011). Further exploration would also be required to understand the impact that the decision alternatives have on this analysis. Here, we used a large set of decision alternatives built from multiple Pareto fronts. Using the generic methodology presented here it would be possible to see whether scenarios have the same impact when the set of decision alternatives is smaller or is comprised of a single Pareto front.

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

Acknowledgments, Samples, and Data

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

Descriptions of scenario generation methods across the water resources literature, and a detailed methodology to calculate diverse futures and targeted spread scenarios is available in the Supporting Information.

The Lake Model is widely available on GitHub in multiple repositories, including in the EMAworkbench: https://github.com/quaquel/EMAworkbench

References

Anghileri, D., Botter, M., Castelletti, A., Weigt, H., & Burlando, P. (2018). A comparative assessment of the impact of climate change and energy policies on Alpine hydropower. Water Resources Research, 54(11), 9144–9161.

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

Beh, E., Maier, H. R., & Dandy, G. C. (2015a). Adaptive, multiobjective optimal sequencing approach for urban water supply augmentation under deep uncertainty. Water Resources Research, 51(3), 1529–1551.

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

Borgomeo, E., Mortazavi-Naeini, M., Hall, J. W., & Guillod, B. P. (2018). Risk, Robustness and Water Resources Planning Under Uncertainty. Earth's Future, 6(3), 468–487.

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

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

Burn, D. H., Venema, H. D., & Simonovic, S. P. (1991). Risk-Based Performance Criteria for Real-Time Reservoir Operation. Canadian Journal of Civil Engineering, 18(1), 36–42. https://doi.org/10.1139/l91-005

Carpenter, S. R., Ludwig, D., & Brock, W. A. (1999). Management of eutrophication for lakes subject to potentially irreversible change. Ecological Applications, 9(3), 751–771.

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

Culley, S., Bennett, B., Westra, S., & Maier, H. R. (2019). Generating realistic perturbed hydrometeorological time series to inform scenario-neutral climate impact assessments. Journal of Hydrology.

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

Drouet, L., Bosetti, V., & Tavoni, M. (2015). Selection of climate policies under the uncertainties in the Fifth Assessment Report of the IPCC. Nature Clim. Change, 5(10), 937–940. Retrieved from http://dx.doi.org/10.1038/nclimate2721%5Cn10.1038/nclimate2721%5Cnhttp://www.nature.com/nclimate/journal/v5/n10/abs/nclimate2721.html#supplementary-information

Eker, S., & Kwakkel, J. H. (2018). Including robustness considerations in the search phase of Many-Objective Robust Decision Making. Environmental Modelling and Software, 105, 201–216. https://doi.org/10.1016/j.envsoft.2018.03.029

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

Giuliani, M., Castelletti, A., Pianosi, F., Mason, E., & Reed, P. M. (2016). Curses, Tradeoffs, and Scalable Management: Advancing Evolutionary Multiobjective Direct Policy Search to Improve Water Reservoir Operations. Journal of Water Resources Planning and Management, 142(2), 04015050. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000570

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

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

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

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

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

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

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

Hall, J. W., Lempert, R. J., Keller, K., Hackbarth, A., Mijere, C., & Mcinerney, D. J. (2012). Robust Climate Policies Under Uncertainty: A Comparison of Robust Decision Making and Info-Gap Methods. Risk Analysis, 32(10), 1657–1672. https://doi.org/10.1111/j.1539-6924.2012.01802.x

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

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

Herman, J. D., Zeff, H. B., Reed, P. M., & Characklis, G. W. (2014). Beyond optimality: Multistakeholder robustness tradeoffs for regional water portfolio planning under deep uncertainty. Water Resources Research, 50(10), 7692–7713.

Herman, J. D., Reed, P. M., Zeff, H. B., & Characklis, G. W. (2015). How Should Robustness Be Defined for Water Systems Planning under Change? Journal of Water Resources Planning and Management, 141(10), 04015012. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000509

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

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

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

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

Kwakkel, J. H., & van der Pas, J. W. G. M. (2011). Evaluation of infrastructure planning approaches: an analogy with medicine. Futures, 43(9), 934–946.

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

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

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

Kwakkel, J. H., Haasnoot, M., & Walker, W. E. (2016). Comparing Robust Decision-Making and Dynamic Adaptive Policy Pathways for model-based decision support under deep uncertainty. Environmental Modelling and Software, 86, 168–183. https://doi.org/10.1016/j.envsoft.2016.09.017

Kwakkel, J. H., Walker, W. E., & Haasnoot, M. (2016). Coping with the Wickedness of Public Policy Problems: Approaches for Decision Making under Deep Uncertainty. Journal of Water Resources Planning and Management, 142(3), 01816001. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000626

Kwakkel, J. H., Eker, S., & Pruyt, E. (2016). How robust is a robust policy? Comparing alternative robustness metrics for robust decision-making. In International Series in Operations Research and Management Science (Vol. 241, pp. 221–237). Springer. https://doi.org/10.1007/978-3-319-33121-8_10

Lempert, R. J. (2003). Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis. Rand Corporation. https://doi.org/10.1016/j.techfore.2003.09.006

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

Lempert, R. J., & Trujillo, H. R. (2018). Deep Decarbonization as a Risk Management Challenge.

Lenton, T. M. (2013). Environmental tipping points. Annual Review of Environment and Resources, 38, 1–29.

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

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.

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

Matrosov, E. S., Padula, S., & Harou, J. J. (2013). Selecting portfolios of water supply and demand management strategies under uncertainty—contrasting economic optimisation and “robust decision making” approaches. Water Resources Management, 27(4), 1123–1148.

McInerney, D., Lempert, R., & Keller, K. (2012). What are robust strategies in the face of uncertain climate threshold responses? Climatic Change, 112(3–4), 547–568.

McPhail, C., Maier, H. R., Kwakkel, J. H., Giuliani, M., Castelletti, A., & Westra, S. (2018). Robustness Metrics: How Are They Calculated, When Should They Be Used and Why Do They Give Different Results? Earth’s Future, 6(2), 169–191. https://doi.org/10.1002/2017EF000649

Morgan, M. G., Henrion, M., & Small, M. (1990). Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press.

Van Notten, P. W. F., Sleegers, A. M., & van Asselt, M. B. A. (2005). The future shocks: on discontinuity and scenario development. Technological Forecasting and Social Change, 72(2), 175–194.

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

Quinn, J. D., Reed, P. M., & Keller, K. (2017). Direct policy search for robust multi-objective management of deeply uncertain socio-ecological tipping points. Environmental Modelling & Software, 92, 125–141.

Quinn, J. D., Reed, P. M., Giuliani, M., Castelletti, A., Oyler, J. W., & Nicholas, R. E. (2018). Exploring how changing monsoonal dynamics and human pressures challenge multireservoir management for flood protection, hydropower production, and agricultural water supply. Water Resources Research, 54(7), 4638–4662.

Raso, L., Kwakkel, J., Timmermans, J., & Panthou, G. (2019). How to evaluate a monitoring system for adaptive policies: criteria for signposts selection and their model-based evaluation. Climatic Change, 1–17. https://doi.org/10.1007/s10584-018-2355-3

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). Environmental Modelling & Software Management Option Rank Equivalence (MORE) – A new method of sensitivity analysis for decision-making. Environmental Modelling and Software, 25(2), 171–181. https://doi.org/10.1016/j.envsoft.2009.06.012

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

Roach, T., Kapelan, Z., Ledbetter, R., & Ledbetter, M. (2016). Comparison of Robust Optimization and Info-Gap Methods for Water Resource Management under Deep Uncertainty. Journal of Water Resources Planning and Management, 142(9), 04016028. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000660

Shepherd, T. G., Boyd, E., Calel, R. A., Chapman, S. C., Dessai, S., Dima-West, I. M., et al. (2018). Storylines: an alternative approach to representing uncertainty in physical aspects of climate change. Climatic Change, 1–17.

Singh, R., Reed, P. M., & Keller, K. (2015). Many-objective robust decision making for managing an ecosystem with a deeply uncertain threshold response. Ecology and Society, 20(3).

Sniedovich, M. (2010). A bird's view of info-gap decision theory. The Journal of Risk Finance, 11(3), 268–283.

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

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

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

Wada, Y., Vinca, A., Parkinson, S., Willaarts, B. A., Magnuszewski, P., Mochizuki, J., et al. (2019). Co-designing Indus Water-Energy-Land Futures. One Earth, 1(2), 185–194.

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

Walker, W. E., Lempert, R., & Kwakkel, J. (2013). Deep Uncertainty. In Encyclopedia of Operations Research and Management Science (pp. 395–402). Springer. https://doi.org/10.1007/978-1-4419-1153-7_1140

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

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

Weaver, C. P., Lempert, R. J., Brown, C., Hall, J. A., Revell, D., & Sarewitz, D. (2013).

Improving the contribution of climate model information to decision making: the value and demands of robust decision frameworks. Wiley Interdisciplinary Reviews: Climate Change, 4(1), 39–60.

Zeff, H. B., Kasprzyk, J. R., Herman, J. D., Reed, P. M., & Characklis, G. W. (2014). Navigating financial and supply reliability tradeoffs in regional drought management portfolios. Water Resources Research, 50(6), 4906–4923. https://doi.org/10.1002/2013WR015126

Zongxue, X., Jinno, K., Kawamura, A., Takesaki, S., & Ito, K. (1998). Performance risk analysis for Fukuoka water supply system. Water Resources Management, 12(1), 13–30. https://doi.org/10.1007/s11269-006-9040-4

Chapter 4

Guidance framework and software for understanding and achieving system robustness

Cameron McPhail, Holger R. Maier, Seth Westra, Leon van der Linden, and Jan Kwakkel

To be submitted to Environmental Modelling and Software

Statement of Authorship

Title of Paper	Guidance framework and software for understanding and achieving system robustness
Publication Status	□ Published □ Accepted for Publication □ Submitted for Publication ☑ Unpublished and Unsubmitted work written in manuscript style
Publication Details	To be submitted in the near future to Environmental Modelling and Software

Principal Author

Name of Principal Author (Candidate)	Cameron McPhail		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote manuscript, made edits to manuscript, and was corresponding author.		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	16/08/2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Holger Maier		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature		Date	17/8/20

Name of Co-Author	Seth Westra		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature		Date	18/08/2020

Name of Co-Author	Leon van der Linden		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature	[redacted]	Date	19-08-2020

Name of Co-Author	Jan Kwakkel		
Contribution to the Paper	Contributed to the conception and design of the project, analysis, interpretation of results, wrote parts of manuscript, and made edits to manuscript.		
Signature	[redacted]	Date	18-08-2020

Abstract

The long-term planning of environmental systems presents several challenges to decision-makers including the question of how to make decisions when model inputs cannot be represented by deterministic or stochastic processes (i.e. when the future is deeply uncertain) and must instead be represented as multiple plausible futures (scenarios). A variety of metrics exist for determining the robustness (performance across the scenarios) and recent research shows that different metrics can lead to different decisions being made. Similarly, a variety of approaches exist for the selection or generation of scenarios and recent research has shown that because they are an input to the calculation of robustness, these different approaches can lead to different decisions being made. Despite the uncertainty in which robustness metric or scenario selection approach should be used to determine which decision alternative is most robust, no guidance for decision-makers exists for how to conduct a holistic robustness analysis for the problem at hand. In this paper, we develop a generic guidance framework to assist with the identification of the most robust decision alternative. To ensure consistency and ease-of-use for the proposed guidance framework, this paper also introduces a software package that assists in the implementation of this framework. We illustrate the guidance framework and software package on a hypothetical lake pollution problem, known in the literature as The Lake Problem, showing how the guidance and software package applies to several situations where the decision-makers may or may not know or which scenarios or robustness metrics to use.

4.1. Introduction

The long-term planning of water and environmental systems presents major challenges to decision-makers, requiring them to make decisions despite a significant degree of uncertainty in the future state of the world. Frequently, decision-makers are operating at the level of deep uncertainty, which refers to when deterministic and stochastic processes are insufficient for representing the future state of the world, and the consideration of multiple plausible futures (scenarios) is required (Bradfield et al., 2005; Herman et al., 2014; Kwakkel et al., 2010; Kwakkel and Haasnoot, 2019; Lempert, 2003; Little et al., 2018; Maier et al., 2016; Schwarz, 1991; van der Heijden, 1996; Varum and Melo, 2010; Walker et al., 2013; Wright and Cairns, 2011). Further complicating this, probabilities cannot be placed on the scenarios, and therefore traditional performance metrics such as reliability (frequency of acceptable performance), vulnerability (magnitude of failure), resilience (duration of failure), or expected value (expected level of performance) do not apply because a traditional Monte Carlo analysis would require the probabilities to be known (Maier et al., 2016). Rather, deep uncertainty requires robustness metrics, which aim to determine the level of system performance and how that performance varies across all scenarios (Herman et al., 2015; Kwakkel and Haasnoot, 2019; Lempert, 2003; Maier et al., 2016; McPhail et al., 2018).

However, the literature contains a multitude of approaches to quantify the performance across a range of deeply uncertain futures, including: (i) expected value metrics (Wald, 1951), which indicate an expected level of performance across a range of scenarios; (ii) metrics of higher-order moments, such as variance and skew (e.g. Kwakkel et al. (2016a)), which provide information on how the expected level of performance varies across multiple scenarios; (iii) 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; and (iv) satisficing metrics (Simon, 1956), which calculate the range of scenarios that have acceptable performance relative to a threshold. A common conclusion from recent research is that different robustness metrics can sometimes lead to decision alternatives being ranked differently, making it difficult to determine which decision alternatives are the most robust (Borgomeo et al., 2018; Drouet et al., 2015; Giuliani and Castelletti, 2016; Hall et al., 2012; Herman et al., 2015; Kwakkel et al., 2016a; Lempert and Collins, 2007; McPhail et al., 2018; Roach et al., 2016). For example, one of the conclusions of the Lake Como study by Giuliani and Castelletti (2016) was that "solutions obtained with misdefined robustness metrics generally underestimate the system performance with respect to the one achievable with a correctly defined metric, with the degradation of performance that is larger in the case of the more pessimistic metrics." Similarly, a Kwakkel et al. (2016a) case study on the transition of the European energy system towards a more sustainable future concluded that "there is no clearly superior single robustness metric. Case specific consideration and system characteristics affect the merits of the various robustness measures. This implies that an analyst has to choose carefully which robustness measure is being used and assess its appropriateness."

Due to the scenarios being an input for the calculation of robustness, the scenarios also have an impact on the robustness of a decision alternative (McPhail et al., 2020, 2018). However, just as there is a diversity of approaches to calculating robustness, there is also a diversity of approaches for selecting or creating scenarios. A common categorization of approaches to scenario generation is given by Börjeson et al. (2006), where scenarios are split into three types:

- Predictive scenarios – where the aim is to determine "what will happen?" For example, the future state of the world could be based on some future trajectory or change in trajectory due to some event;
- Explorative scenarios – where the aim is to determine "what could happen?" Generally, this is done by framing the future in terms of the uncertainties that have the largest effects on system performance, but the future can also be unframed (Maier et al., 2016); and
- Normative scenarios – where the aim is to determine "how can a specific future be realized?" This is generally focused on interesting future outcomes or failure points for decision alternatives.

In addition to this, scenarios can be created in very different ways. For example, a set of scenarios for a particular problem could be created in a largely qualitative manner through a participatory process with stakeholders with the aim of producing generalizable scenarios (e.g. Wada et al. (2019)), while a different set of scenarios for the same problem could be created through a largely quantitative process by varying the inputs to the system model of interest (e.g. using an approach such as Latin hypercube sampling (LHS))

(Culley et al., 2019, 2016; Hadka et al., 2015; Hall et al., 2012; Herman et al., 2015; Kasprzyk et al., 2013; Kwakkel, 2017; Kwakkel et al., 2016b, 2015; McPhail et al., 2018; Quinn et al., 2018, 2017; Singh et al., 2015; Trindade et al., 2017; Watson and Kasprzyk, 2017; Weaver et al., 2013; Zeff et al., 2014). However, each of these approaches can lead to vastly different scenarios being produced (Shepherd et al., 2018), for example, a participatory approach will generally result in a small number of scenarios in targeted regions of the uncertain variable space, while the latter example (LHS of scenarios) would lead to a large number of scenarios with even coverage of the space. Recent studies have shown that, as is the case for the use of different robustness metrics, the use of different sets of scenarios can also result in different relative robustness values of decision alternatives (McPhail et al., 2020), further inhibiting our ability to identify which decision alternative is most robust.

In order to assist analysist and decision makers to better understand the sensitivity of the absolute and relative robustness of decision alternatives (e.g. designs, policies) of interest to the choice of robustness metrics and scenarios, McPhail et al. (2018) developed a generalizable, quantitative approach to assess the impact of the choice of different robustness metrics on the absolute and relative robustness of decision alternatives, and McPhail et al. (2020) did the same for the impact of the selection of different scenario sets. However, there is still a lack of a holistic procedure that provides guidance to analysts on the best way to identify which of the available decision alternatives is likely to be the most robust.

Given the uncertainty in the choice of the most appropriate robustness metric and the variability in the relative robustness of different decision alternatives when different robustness metrics and/or scenarios are used, the overarching aim of this paper is to develop a generic guidance framework to assist with the identification of the most robust decision alternative. It should be noted that the focus is on the relative robustness of different decision alternatives, rather than their absolute robustness values, as the selection of the most robust solution is the generally the primary objective from a decision-making context, rather that the calculation of robustness per se. In order to enable to the proposed guidance framework to be implemented in a consistent and user-friendly manner, this paper also introduces a software package that enables the most robust decision alternatives to be identified for a given problem. We illustrate the guidance framework and software package on a hypothetical lake pollution problem, known in the literature as The Lake Problem, as it is a simple and well-represented case study in the literature (Carpenter et al., 1999; Eker and Kwakkel, 2018; Hadka et al., 2015; Kwakkel, 2017; Lempert and Collins, 2007; Quinn et al., 2017; Singh et al., 2015; Ward et al., 2015).

The specific objectives of this paper are:

1. To develop a generic guidance framework to assist with the identification of the most robust decision alternative for a given decision context.
2. To develop and describe a software package that enables the guidance framework to be implemented in a consistent and user-friendly manner.
3. To illustrate the application of the framework and software package on the Lake Problem.

The remainder of this paper is organized as follows: Section 4.2 introduces the guidance framework for analyzing the robustness of a set of decision alternatives, including how to create a custom robustness metric and how to assess the impact of the selection of scenarios and choice of robustness metric; Section 4.3 introduces a software package that can be used to implement this guidance and quantitatively and visually assess the impact of the choice of scenarios and robustness metric on the robustness values and rankings of the decision alternatives; Section 4.4 introduces the Lake Problem and provides a simple illustration of how the guidance and software package can be applied to an environmental model; and conclusions are presented in Section 4.5.

4.2. Guidance framework for identifying the most robust decision alternative

At the heart of the proposed framework for the assisting with the identification of the most robust decision alternatives is the calculation of different robustness metrics. The calculation of these metrics requires scenarios, decision alternatives (i.e. plans, policies, solutions), and one or more quantitative metrics (e.g. reliability or vulnerability), which can be used to determine the level of performance of each decision alternative in each scenario (Herman et al., 2015; McPhail et al., 2018). Figure 4-1 shows the processes through which these three inputs are used to calculate the robustness of each decision alternative (i.e. the system performance across all scenarios). It consists of two main steps, including the use of the system model to calculate each decision alternative's performance in each scenario, followed by the combination of these values in order to calculate a single robustness metric. While these steps are identical for each robustness metric, different robustness metrics correspond to the selection of different options at each of one of three transformations: (1) performance value transformation; (2) scenario subset selection; and (3) aggregation of performance values (McPhail et al., 2018) (Figure 4-1). At the first transformation, the options are whether to use the raw values of system performance or whether to transform these values using regret or satisficing transforms. At the second transformation, the choice is which subset of the available scenarios to use in the calculation of the robustness metric. At the third transformation, the options are whether to combine the transformed performance values over the selected scenarios using a measure of the level of performance, such as the mean, or a measure of variability in performance, such as the standard deviation.

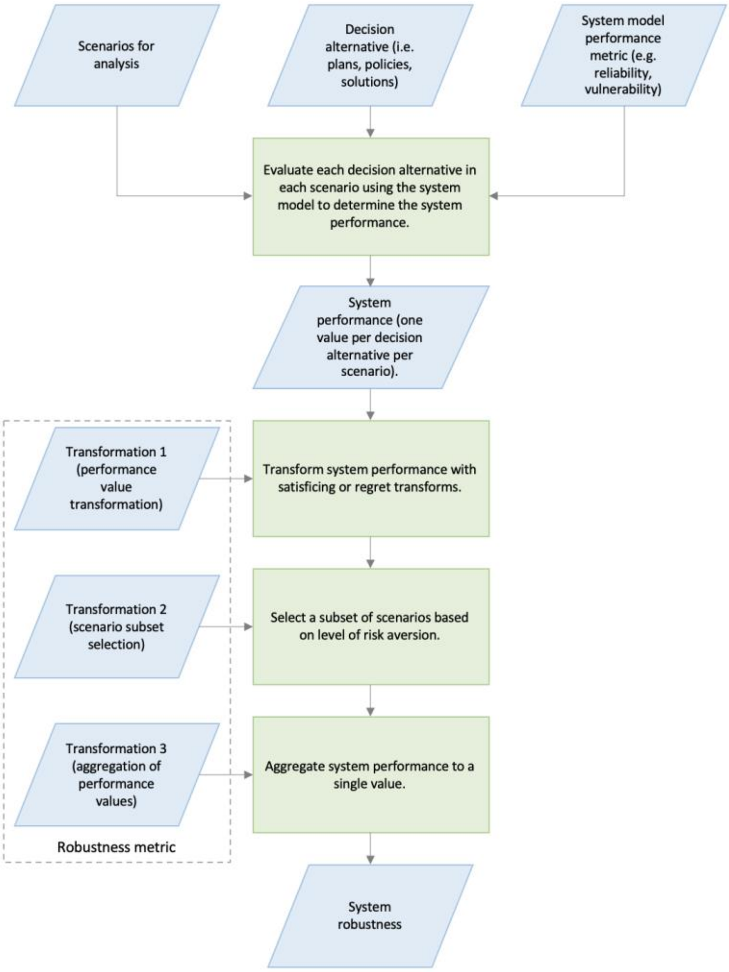

Scenarios for analysis
Decision alternative (i.e. plans, policies, solutions)
System model performance metric (e.g. reliability, vulnerability)
Evaluate each decision alternative in each scenario using the system model to determine the system performance.
System performance (one value per decision alternative per scenario).
Transformation 1 (performance value transformation)
Transform system performance with satisficing or regret transforms.
Transformation 2 (scenario subset selection)
Select a subset of scenarios based on level of risk aversion.
Transformation 3 (aggregation of performance values)
Aggregate system performance to a single value.
Robustness metric
System robustness

Figure 4-1: Inputs and processes for calculating system performance and robustness. Transformations 1-3 are the components of the robustness metric, explained further in the guidance below.

The proposed guidance framework for assisting with the identification of the most robust decision alternatives is given in Figure 4-2. The framework is designed to be as generic as possible, catering to situations where the robustness metric to be used has been pre-determined, where a range of robustness metrics are to be considered or where the most appropriate robustness metric is to be determined based on the different attributes of the decision context (the properties of the problem) and the preferences of the decision-maker. The framework also caters to situations where the scenarios under which system performance is to be calculated are known and situations where the influence of different sets of scenarios on the identification of the most robust solution is to be considered. It should be noted that the proposed framework assumes that the decision alternatives to be considered have already been selected and that the relevant performance metrics for these decision alternatives have been calculated.

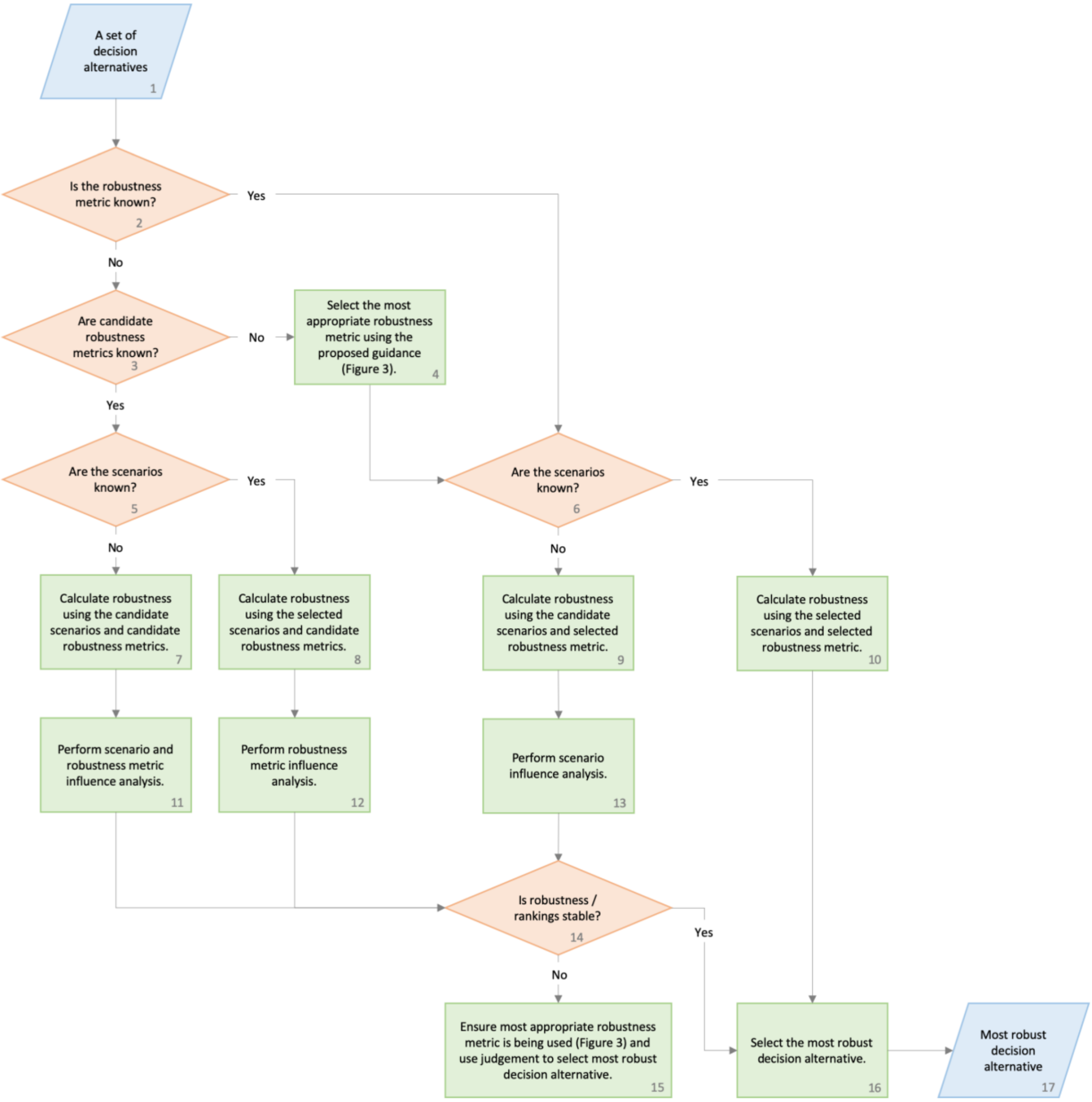

A set of decision alternatives 1
Is the robustness metric known? 2
Yes
No
Are candidate robustness metrics known? 3
No
Select the most appropriate robustness metric using the proposed guidance (Figure 3). 4
Yes
Are the scenarios known? 5
Yes
Are the scenarios known? 6
Yes
No
No
Calculate robustness using the candidate scenarios and candidate robustness metrics. 7
Calculate robustness using the selected scenarios and candidate robustness metrics. 8
Calculate robustness using the candidate scenarios and selected robustness metric. 9
Calculate robustness using the selected scenarios and selected robustness metric. 10
Perform scenario and robustness metric influence analysis. 11
Perform robustness metric influence analysis. 12
Perform scenario influence analysis. 13
Is robustness / rankings stable? 14
Yes
No
Ensure most appropriate robustness metric is being used (Figure 3) and use judgement to select most robust decision alternative. 15
Select the most robust decision alternative. 16
Most robust decision alternative 17

Figure 4-2: Proposed generic guidance framework for assisting with the identification of the most robust decision alternative.

The process of identifying the decision alternative that has the highest relative robustness commences with the candidate set of decision alternatives for which the relative robustness is to be calculated. The first decision point in this process is whether the robustness metric to be used in the assessment is known (Figure 4-2, Box 2). If an appropriate metric has already been selected, the next decision point in determining the most robust decision alternative is whether the scenarios to be used to determine the performance of the decision alternatives under consideration are known or not (Figure 4-2, Box 6). If this is the case, the robustness of each decision alternative can be calculated by combining its performance over the selected scenarios with the aid of the selected robustness metric. Then the alternative with the highest robustness value can be selected (Figure 4-2, Boxes 10 and 16). If it is not clear which scenarios should be used for the robustness calculation, the sensitivity of the relative robustness values of the different decision alternatives can be determined for different user-defined scenario sets, using the approach of McPhail et al. (2020) (Figure 4-2, Box 13). Visualizations of the relative ranking of the decision alternatives can be used to determine (using human judgement) whether the choice of candidate scenario set matters (Figure 4-2, Box 14), as illustrated in McPhail et al. (2020). If the choice of candidate scenarios does not matter because the visualizations indicate that the decision alternatives are ranked similarly regardless of the scenarios, then the decision alternative that is considered most robust can be easily selected (Figure 4-2, Box 16). However, if the choice of scenarios does affect the relative robustness of the decision alternatives of interest, then depending on the degree of sensitivity of the relative robustness of the different decision alternatives to the selected scenario sets, some degree of judgement will be required to determine which decision alternative is considered most robust (Figure 4-2, Box 15), or it might be concluded that it is not possible to identify which decision alternative is most robust. Note that in the situation where a robustness metric is known or pre-selected, it may still be useful to consider the pathways through Figure 4-2 where the robustness metric is not known. This would provide extra information about the system and the impact of the selected robustness metric on the robustness and rankings, as described below.

If the robustness metric to be used is not known, the key decision point is whether a set of alternative robustness metrics to be considered in the analysis is known or not (Figure 4-2, Box 3). If this is known, the next decision point is whether the scenarios are known or not (Figure 4-2, Box 5). If there is a known set of scenarios, the stability of the relative robustness of the decision alternatives under consideration can be calculated for the selected robustness metrics over the selected scenarios, using the approach of McPhail et al. (2018) (Figure 4-2, Box 12). Visualizations of the relative ranking of the decision alternatives can be used to determine whether the choice of candidate robustness metrics matters (Figure 4-2, Box 14), as illustrated in McPhail et al. (2020) and further discussed in Sections 4.3 and 4.4. If the choice of robustness metrics does not matter because the visualizations indicate that the decision alternatives are ranked similarly regardless of the robustness metric, then the decision alternative that is considered most robust can be easily selected (Figure 4-2, Box 16). However, if the robustness metric does affect the relative robustness of the decision alternatives of interest, then depending on the degree to which it has an effect, some degree of judgement will be required to determine which is most robust, and it is recommended that the process for identifying the most appropriate robustness metric for the decision context under consideration introduced in Figure 4-3 and discussed below be applied and that the analysis be repeated for the selected robustness metric (Figure 4-2, Box 15).

If the scenarios to be used are not known, the sensitivity of the relative robustness values of the different decision alternatives to the different user-defined scenario sets and robustness values can be determined using the approach of McPhail et al. (2020) (Figure 4-2, Box 11). Again, the visualizations (as illustrated in McPhail et al. (2020) and further discussed in Sections 4.3 and 4.4) allow the decision-maker to see whether the candidate scenarios and candidate robustness metrics have a significant effect on the relative robustness (Figure 4-2, Box 14). If the selection of scenarios and robustness metrics has an insignificant effect on the rankings, the most robust decision alternative can be easily selected (Figure 4-2, Box 16). However, if there is an effect on the relative robustness, then depending on the degree to which this is the case, some degree of judgement will be required to determine which decision alternative is most robust, and it is recommended that the most appropriate robustness metric is used to help determine this (Figure 4-2, Box 15).

If the set of alternative robustness metrics to be considered in the analysis is unknown (Figure 4-2, Box 3), the most appropriate robustness metric to be used for each individual performance metric can be determined by selecting the most appropriate options at each of the three transformations in Figure 4-1 with the aid of the guidance in Figure 4-3 and the corresponding equations in Table 4-1 (Figure 4-2, Box 4). The first step in this process is to determine whether there is a meaningful threshold in the problem under consideration. For example, in a water supply system, supply must be greater than demand and thus the required demand becomes a constraint for the problem. In this case, the question then becomes whether solutions can be assessed using a pass or fail criterion, or whether the magnitude of the failure is important. In the previous example, a water supply system would be deemed to fail if demand was greater than the supply, so all decision alternatives could be classed as passing or failing in each scenario. Alternatively, a decision-maker looking at a water supply system could choose to set a threshold as the point where supply is low enough to cause water restrictions, in which case the magnitude of failure does matter, since less water would mean greater water restrictions.

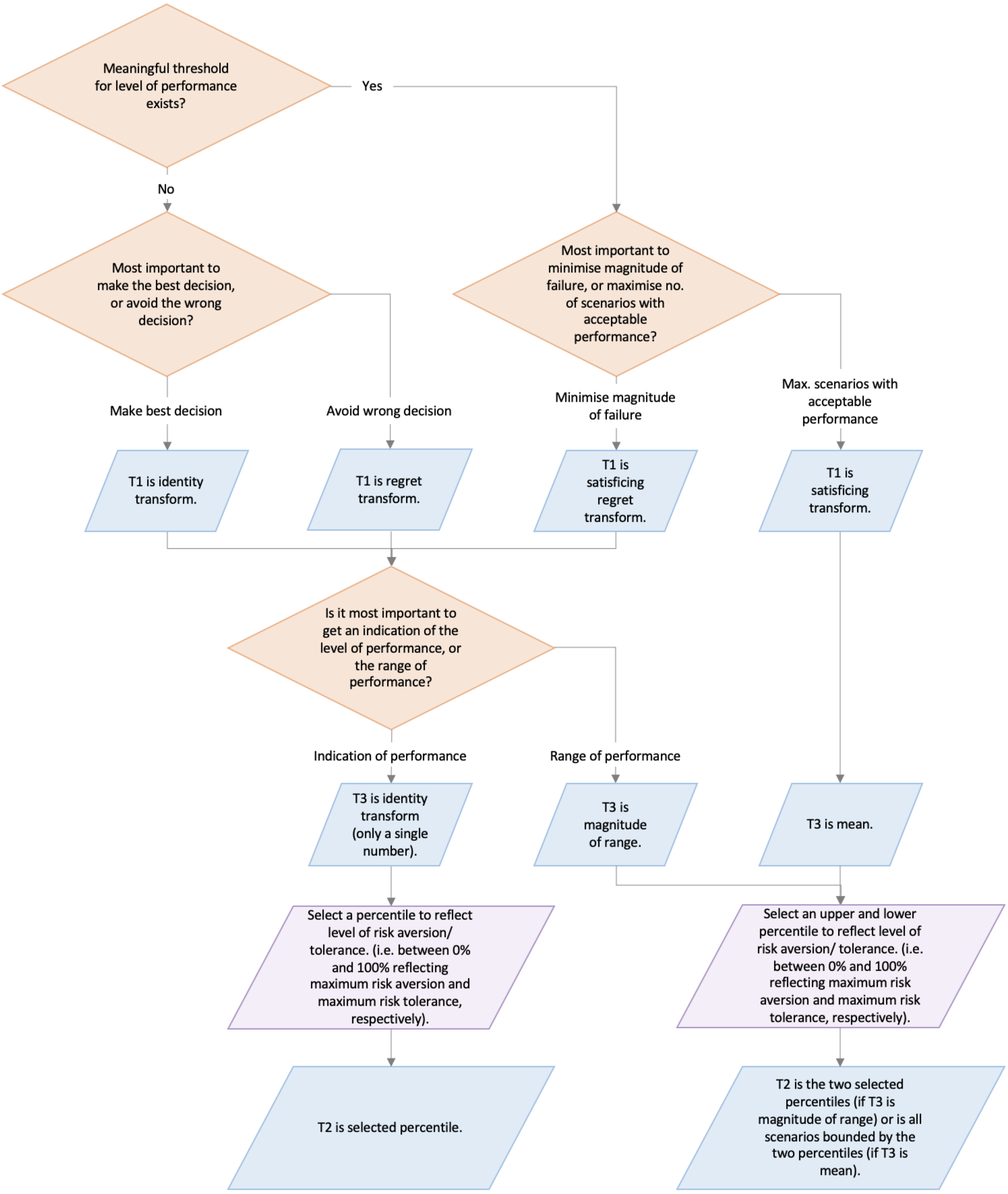

Meaningful threshold for level of performance exists?
Yes
No
Most important to make the best decision, or avoid the wrong decision?
Most important to minimise magnitude of failure, or maximise no. of scenarios with acceptable performance?
Make best decision
Avoid wrong decision
Minimise magnitude of failure
Max. scenarios with acceptable performance
T1 is identity transform.
T1 is regret transform.
T1 is satisficing regret transform.
T1 is satisficing transform.
Is it most important to get an indication of the level of performance, or the range of performance?
Indication of performance
Range of performance
T3 is identity transform (only a single number).
T3 is magnitude of range.
T3 is mean.
Select a percentile to reflect level of risk aversion/ tolerance. (i.e. between 0% and 100% reflecting maximum risk aversion and maximum risk tolerance, respectively).
Select an upper and lower percentile to reflect level of risk aversion/ tolerance. (i.e. between 0% and 100% reflecting maximum risk aversion and maximum risk tolerance, respectively).
T2 is selected percentile.
T2 is the two selected percentiles (if T3 is magnitude of range) or is all scenarios bounded by the two percentiles (if T3 is mean).

Figure 4-3: Guidance for the creation of a robustness metric for each performance metric according to the problem being analyzed and the preferences of the decision-maker. (Note that the equations assume the objective here is to maximize system performance).

Table 4-1: Equations for the robustness metric transformations (assuming the aim is to maximize performance).

T1 (performance value transformation)	
Identity transform	$f_1(l_a, s_i) = f(l_a, s_i)$ (performance metric f; decision alternative a, l_a; scenario i, s_i)
Regret transform	$f_1(l_a, s_i) = \underset{j}{\text{argmax}}\, f(l_a, s_j) - f(l_a, s_i)$
Satisficing regret transform	$f_1(l_a, s_i) = \begin{cases} 0, & f(l_a, s_i) \geq c \\ c - f(l_a, s_i), & f(l_a, s_i) \leq c \end{cases}$ (constraint of c)
Satisficing transform	$f_1(l_a, s_i) = \begin{cases} 1, f(l_a, s_i) \geq c \\ 0, f(l_a, s_i) \leq c \end{cases}$
T2 (scenario subset selection)	
Select a single percentile	$f_2(l_a, S) = f_1(l_a, s_p)$ (pth percentile; S is full set of scenarios)
Select bounds of range	$f_2(l_a, S) = \{f_1(l_a, s_{up}), f_1(l_a, s_{lp})\}$ (where T3 is magnitude of range) (up is the upper percentile, lp is the lower percentile)
Select range of scenarios	$f_2(l_a, S) = \{f_1(l_a, s_i) \; \forall \, i \; : \; f_1(l_a, s_{lp}) \leq f_1(l_a, s_i) \leq f_1(l_a, s_{up})\}$
T3 (performance value aggregation)	
Identity transform	$f_3(l_a, S) = f_2(l_a, S)$
Magnitude of range	$f_3(l_a, S) = f_2(l_a, s_{up}) - f_2(l_a, s_{lp})$
Mean	$f_3(l_a, S) = \left(\sum_{i=1}^{n} f_2(l_a, s_i)\right)/n$

If there is no threshold, then the question is whether the aim is to maximize performance or avoid making the "wrong" decision. By avoiding making the "wrong" decision, we are referring to some decision-makers who may have a desire to avoid selecting decision alternatives if there is a potential that, with hindsight, the decision-maker could be criticized for having made the wrong decision, even if at the time of making the decision, it appeared to be a reasonable option with the available information. For example, many publicly owned water authorities face intense public scrutiny, and for that reason some decision-makers may want to avoid making decisions (e.g. large capital expenditure projects, such as a desalination plant for water security) that could be perceived to be "wrong" after the fact (e.g. an unnecessary expenditure because climate change or population growth eventuates to be less than expected). Decision-makers in this situation may prefer to choose a decision alternative that is not the best in any single scenario but is never far from the best decision alternative in extreme good or bad scenarios.

The next step in Figure 4-3 is to determine whether it is most important to get an indication of the *level* of performance, or the *range* of performance across the multiple plausible futures. Generally, the former is of greatest importance, but the latter may also be important as an additional robustness metric. In other words, if the range of performance is considered important, it would generally be considered as a secondary metric to be used in addition to a robustness metric that indicates the range of performance. For example, in a water supply system, it would be most important for decision makers to have an indication of how much water each decision alternative will supply. But, as an additional metric, the decision makers may opt to choose a decision alternative with a slightly lower performance if the range of performance values is smaller across the different scenarios. In this case, the decision makers could consider both robustness metrics in their decision-making.

In the case where an indication of the *level* of performance is chosen as being most important, this is based on the level of risk tolerance or risk aversion required for the problem or preferred by the decision-maker. Often, a high level of risk aversion is warranted when the consequences of failure are very high. For example, the design of a water supply system would require a high level of risk aversion. In contrast to this, the remediation of an environmental stream may allow a high level of risk tolerance, depending on the preference of the decision maker. Alternatively, the level of risk aversion may also be a matter of personal preference, with some decision-makers being more tolerant of risk than others. This scale of risk aversion and risk tolerance can be represented in the robustness metric by selecting a percentile between 0% and 100%, with 0% reflecting the worst-case scenario (extreme risk aversion) for each decision alternative (i.e. 0% of scenarios have worse performance) and 100% reflecting the best-case scenario (extreme risk tolerance). It must be noted that unlike a probabilistic assessment of level of performance, percentiles that are used for robustness metrics are reflective of relative (not absolute) risk. For example, the 50th percentile does not reflect the median level of performance that can be expected in future, however, it does represent a level of performance that is worse than the 90th percentile and therefore is more risk averse than selecting the 90th percentile.

Once the most appropriate “custom” robustness metric has been determined based on the attributes of the decision context (the properties of the problem) and the preferences of the decision-maker with the aid of the process in Figure 3, the next decision point is whether the scenarios under which the performance of the decision alternatives under consideration should be evaluated are known or not (Figure 4-2, Box 5). From here, the same process is followed as if the robustness metric was known in advance (as described above), leading to a scenario analysis (Figure 4-2, Box 13) if the scenarios are unknown, and the selection of the most robust decision alternative if the scenarios are known (Figure 4-2, Boxes 10 and 16).

4.3. The RAPID software package

The RAPID (Robustness Analysis Producing Intelligent Decisions) Python software package enables the generic guidance framework introduced in Figure 4-2 to be implemented in a user-friendly and consistent manner, including functionality to guide the user through the process of creating a custom robustness metric as described in Figure 4-3. RAPID is written in Python, which is increasingly being used for scientific modelling because it is a high-level, general-purpose, and open source programming language with an emphasis on code readability. It also has a very large standard library, and a significant repository of third-party Python packages. The fact that the RAPID package is written in Python also makes it easier for it interact with many other software packages, including the Exploratory Modeling (EM) Workbench (Kwakkel, 2017), which is also written in Python. As the EM Workbench includes functionality for the generation of decision alternatives (i.e. policy options, solutions, etc.), the generation of scenarios (i.e. states of the world, plausible futures) and vulnerability analyses (including scenario discovery, feature scoring, and sensitivity analyses), the EM Workbench can be used for the creation of all of the inputs needed for the generic guidance framework (Figure 4-2) which the RAPID package implements.

As shown in Figure 4-4, the processes from the guidance framework are implemented across two sub-packages, *metrics* and *analysis* (colored purple and green respectively in Figure 4-4). The sub-package *metrics* contains functions that enable each of the three transformations required for the calculation of robustness metrics (Figure 4-1) to be implemented (see Table 4-2 for available options at each of the three transformations). This enables user-defined “custom” robustness metrics to be implemented (see Table 4-2 for available options at each of the three transformations), including those obtained by following the process outlined in Figure 4-3 (either by manually selecting the transformations and combining them using the *custom_R_metric* function, or by interacting with the guidance helper function, *guidance_to_R*, which steps through the process in Figure 4-3). A number of commonly used robustness metrics have also been pre-programmed (see Table 4-3 for these metrics as well as the corresponding choices at each of the three transformations). These robustness metrics can then be used to calculate the robustness values for given decision alternatives, scenarios, and performance metrics, as highlighted in Figure 4-1.

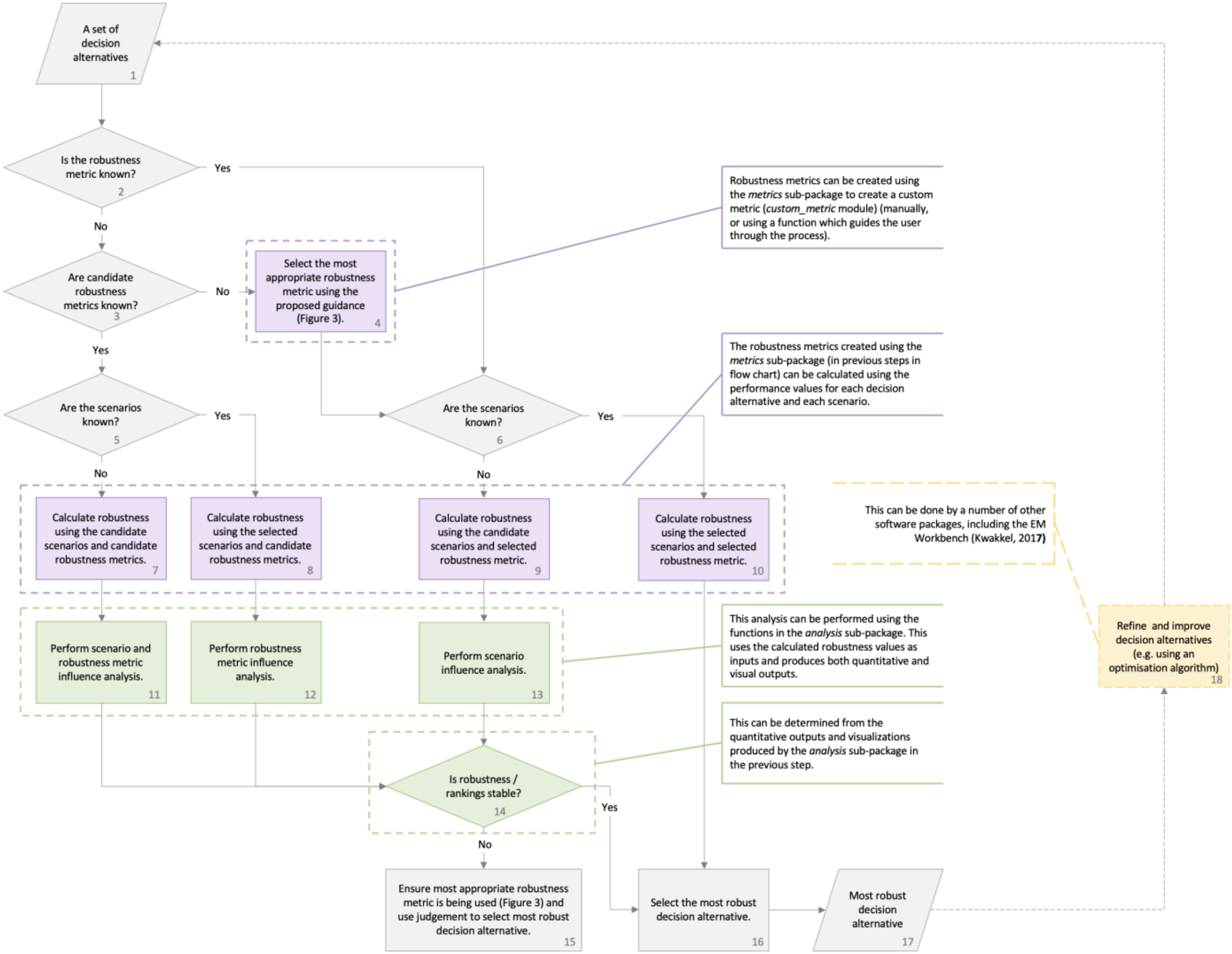

A set of decision alternatives
1
Is the robustness metric known?
2
Yes
No
Are candidate robustness metrics known?
3
No
Select the most appropriate robustness metric using the proposed guidance (Figure 3).
4
Robustness metrics can be created using the *metrics* sub-package to create a custom metric (*custom_metric* module) (manually, or using a function which guides the user through the process).
Yes
Are the scenarios known?
5
Yes
Are the scenarios known?
6
Yes
The robustness metrics created using the *metrics* sub-package (in previous steps in flow chart) can be calculated using the performance values for each decision alternative and each scenario.
No
No
Calculate robustness using the candidate scenarios and candidate robustness metrics.
7
Calculate robustness using the selected scenarios and candidate robustness metrics.
8
Calculate robustness using the candidate scenarios and selected robustness metric.
9
Calculate robustness using the selected scenarios and selected robustness metric.
10
This can be done by a number of other software packages, including the EM Workbench (Kwakkel, 2017)
Perform scenario and robustness metric influence analysis.
11
Perform robustness metric influence analysis.
12
Perform scenario influence analysis.
13
This analysis can be performed using the functions in the *analysis* sub-package. This uses the calculated robustness values as inputs and produces both quantitative and visual outputs.
Refine and improve decision alternatives (e.g. using an optimisation algorithm)
18
This can be determined from the quantitative outputs and visualizations produced by the *analysis* sub-package in the previous step.
Is robustness / rankings stable?
14
Yes
No
Ensure most appropriate robustness metric is being used (Figure 3) and use judgement to select most robust decision alternative.
15
Select the most robust decision alternative.
16
Most robust decision alternative
17

Figure 4-4: The general guidance framework introduced in Figure 4-2, with an explanation of how the RAPID software package assists in the implementation of this guidance and one way that it can interact with the EM Workbench package.

Table 4-2: Options for each of the three robustness metric calculation transformations included in the software package

Transformation number	**Transformation name**	**Used in traditional metrics**	**Used in proposed guidance**	**Software package function**
T1 (performance value transformation)	Identity	✓	✓	*t1.identity*
	Regret	✓	✓	*t1.regret_from_best_da* (regret from best decision alternative)
	Satisficing regret	✓	✓	*t1.satisficing_regret*
	Regret from median	✓		*t1.regret_from_median*
	Regret from value	✓	✓	*t1.regret_from_value* (used to calculate the other regret metrics (which are all calculating regret with respect to different values)
	Satisficing	✓	✓	*t1.satisfice*
T2 (scenario subset selection)	Select a single percentile	✓	✓	*t2.select_percentiles*, *t2.worst_case* (for 0th percentile), or *t2.best_case* (for 100th percentile)
	Worst- and best-case scenarios	✓		*t2.worst_and_best_cases*
	Select bounds of range		✓	*t2.select_percentiles*
	Select range of scenarios	✓	✓	*t2.range*, *t2.worst_half*, or *t2.all_scenarios*
(continues on next page…)				

(…continues from previous page)

Transformation number	**Transformation name**	**Used in traditional metrics**	**Used in proposed guidance**	**Software package function**
T3 (performance value aggregation)	Identity transform	✓	✓	*t3.f_identity*
	Magnitude of range		✓	*t3.f_range*
	Mean	✓	✓	*t3.f_mean*
	Sum	✓		*t3.f_sum*
	Weighted sum	✓		*t3.f_w_sum*
	Variance	✓		*t3.f_variance*
	Mean-variance	✓		*t3.f_mean_vairance*
	Skew	✓		*t3.f_skew*
	Kurtosis	✓		*t3.f_kurtosis*

Table 4-3: Commonly used robustness metrics included in the software package, as well as corresponding choices at each of the three transformations

Metric name	**T1 (performance value transformation)**	**T2 (scenario subset selection)**	**T3 (performance value aggregation)**	**Software package function**
Maximin	Identity	Worst-case	Identity	*maximin*
Maximax	Identity	Best-case	Identity	*maximax*
Hurwicz's Optimism-Pessimism Rule	Identity	Worst- and best-cases	Mean	*hurwicz*
Laplace's Principle of Insufficient Reason	Identity	All scenarios	Mean	*laplace*
Minimax Regret	Regret	Worst-case	Identity	*minimax_regret*
Percentile Regret (e.g. 90th percentile regret)	Regret	Percentile	Identity	*percentile_regret*
Mean-variance	Identity	All scenarios	Mean-variance	*mean_variance*
Undesirable deviations	Regret from median	Worst-half	Sum	*undesirable_deviations*
Percentile-based skew	Identity	10th, 50th, and 90th percentiles	Skew	*percentile_skew*
Percentile-based kurtosis	Identity	10th, 25th, 75th, and 90th percentiles	Kurtosis	*percentile_kurtosis*
Starr's Domain Criterion	Satisfice	All scenarios	Mean	*starrs_domain*

The *analysis* sub-package (colored green in Figure 4-4) contains the quantitative methods and visualizations for assessing the sensitivity of the relative robustness values of different decision alternatives to the choice of robustness metrics and / or scenario sets. For the assessment of the impact of scenario selection on the robustness values, the software package uses the approach outlined by McPhail et al. (2020). That is, the software package calculates the difference in robustness values when the robustness is calculated using two different sets of scenarios. First, for each decision alternative, l_i, one can calculate robustness, R, using one set of scenarios, S_a, then calculate the robustness again with a second set of scenarios, S_b, and compare the relative difference between the two robustness values. We use the average relative difference, Δ, across all n decision alternatives:

$$\Delta = \sum_{i=1}^{n} \frac{|R(l_i, S_a) - R(l_i, S_b)|}{\left(\frac{|R(l_i, S_a)| + |R(l_i, S_b)|}{2}\right)} / n \times 100\%$$

Similarly, for the assessment of the impact that scenario selection has on the rankings of the decision alternatives, we follow McPhail et al. (2020), using Kendall's Tau-b ranking correlation to determine the difference in rankings when robustness is calculated using two different sets of scenarios. Kendall's Tau-b ranking has a range between -1 and +1 (inclusive), where -1 indicates that all decision alternatives have opposite rankings, +1 indicates that the rankings are exactly the same, and 0 implies that there is no correlation between the rankings. Specifically, Kendall's Tau-b metric is used to compare two sets of robustness values, one calculated using a set of scenarios, S_a, and the other calculated using a different set of scenarios, S_b:

$$\{R(l_1, S_a), R(l_2, S_a), \dots, R(l_n, S_a)\}$$
$$\{R(l_1, S_b), R(l_2, S_b), \dots, R(l_n, S_b)\}$$

Similarly, Kendall's Tau-b ranking can be used to assess the difference in rankings when robustness is calculated using two different *robustness metrics* (rather than two different *sets of scenarios*, considered above), as recommended by McPhail et al. (2018). Specifically, Kendall's Tau-b metric is used to compare two sets of robustness values, one calculated using a robustness metric, R_1, and the other calculated using a different robustness metric, R_2:

$$\{R_1(l_1, S), R_1(l_2, S), \ldots, R_1(l_n, S)\}$$
$$\{R_2(l_1, S), R_2(l_2, S), \ldots, R_2(l_n, S)\}$$

Note that since we are comparing different robustness metrics, they can be in different scales or units. Therefore, the relative difference in robustness values cannot be calculated, unlike when assessing the impact of scenario selections on the robustness values, where a single robustness metric is used and therefore the values can be compared directly.

The structure of the two sub-packages mentioned above (i.e. *metrics* and *analysis*) is as follows:

- *metrics*; a sub-package containing functions for each of the three robustness metric transformations, common metrics from the literature, functions to help build custom robustness metrics, and a helper function which asks the user the questions from the guidance provided in Section 4.2. This sub-package is structured as:
    - *transforms*; a sub-package, split into the three transformations (T1, T2, T3) as three separate modules (the *t1*, *t2*, and *t3* sub-packages), which implement the transformations listed in Table 4-2. Note that if the aim is to minimize the performance value (e.g. if cost is the measure of performance), the sign of the performance values is inverted in all T1 functions, because this ensures that the value of all robustness metrics is maximized.
    - *common_metrics*; a sub-package for calculating a number of the following 11 commonly used robustness metrics (McPhail et al., 2018): Maximin, Maximax, Hurwicz's Optimism-Pessimism Rule, Laplace's Principle of Insufficient Reason, Minimax Regret, Percentile Minimax Regret, Mean-Variance, Undesirable Deviations, Percentile-based Skew, Percentile-based Kurtosis, and Starr's Domain Criterion, implementing the three transformations from the *transforms* sub-package (as listed in Table 4-3).
    - *custom_metrics*; a module that includes a function (*custom_R_metric*) for creating a custom robustness metric composed of three transformations (from the *transforms* sub-package), and also provides a helper function for stepping users through the flowchart in Figure 4-3 to create a custom

robustness metric that is most appropriate for the decision context under consideration (the *guidance_to_R* function). This helper function asks questions of the user and uses the responses to create the resulting custom robustness metric (using the *custom_R_metric* function).

- *analysis;* a sub-package that enables the influence of different sets of scenarios and robustness metrics on the robustness values and rankings to be determined (the *scenarios_similarity* and *robustness_similarity* functions, respectively). This module also produces plots to visualize the influence that the scenarios and robustness metrics have, including (i) the *delta_plot* function for plotting the relative difference in robustness values (i.e. the deltas) caused by different scenario selections or robustness metrics and (ii) the *tau_plot* function for plotting the ranking similarity (i.e. the Kendall's Tau-b correlation) from different robustness metrics (both functions explained in more detail above).

A number of examples using the software package are also contained within the package, including a multi-objective robust optimization of the Lake Problem (also explored in Section 4.4); a common, hypothetical environmental modelling problem used in the environmental systems modelling literature.

4.4. The Lake Problem

4.4.1. Background

The *examples* directory in the RAPID package includes the Lake Problem as an example of common usage of the package. The Lake Problem is a hypothetical, stylized model which is well-represented in the literature (Carpenter et al., 1999; Eker and Kwakkel, 2018; Hadka et al., 2015; Kwakkel, 2017; Lempert and Collins, 2007; McPhail et al., 2020; Quinn et al., 2017; Singh et al., 2015; Ward et al., 2015), and represents a city that must decide the amount of pollution that it releases into a lake. There are four competing objectives: (1) the average concentration of phosphorous in the lake; (2) the frequency of pollution levels exceeding a critical threshold (i.e. the reliability); (3) the economic benefit (i.e. economic utility) of polluting the lake; and (4) a penalty for if the change in level of pollution is too high from year to year (i.e. a measure of inertia of the pollution) to help achieve more realistic and appropriate solutions. Both deep and stochastic uncertainties are present for the natural inflows of pollution into the lake, the natural removal and recycling rates of pollution in the lake, and the discount rate for the economic benefits. To illustrate the generic guidance framework on the Lake Problem, we will follow several different pathways through the framework (Figure 4-2), including the situations where:

1. Section 4.4.2 – The robustness metric is unknown, and there are no candidate robustness metrics under consideration. The method for generating the scenarios is known.
2. Section 4.4.3 – The robustness metric is unknown and there are no candidate robustness metrics under consideration. There are multiple candidate sets of scenarios.
3. Section 4.4.4 – The robustness metric is unknown, however there are multiple candidate robustness metrics. The method for generating the scenarios is known.

4.4.2. No candidate robustness metrics but scenario generation method known

Following the guidance framework, we consider a situation in which we aim to use an optimization process (Figure 4-4, Box 18) to determine a set of robust decision alternatives. In this situation, we also assume that the robustness metric is unknown (Figure 4-4, Box 2) and that there are no candidate robustness metrics (Figure 4-4, Box 3), leading to Box 4 in Figure 4-4. Here, we deviate from the EM Workbench (Kwakkel, 2017) example of the Lake Problem which used standard robustness metrics for each of the objectives. In our example, we create a custom robustness metric by following the guidelines in Figure 4-3. Note that the creation of these custom robustness metrics is illustrative of how to follow the guidance and uses many assumptions about decision maker preferences that are not present in previous formulations of the Lake Problem. Also note that we have created one robustness metric for each of the four Lake Problem performance metrics, but this need not be the case.

First, for the average concentration of the phosphorous in the lake, we decide that there is no meaningful threshold (note that some studies have created a threshold for this objective), and that we are most interested in making the best decision, which gives us the identity transform for T1. We are looking for an indication of the level of performance, leading to the identity transform for T3, and are relatively risk averse, so the 25th percentile is used for T2 (also see summary in Table 4-4).

Table 4-4: Custom robustness metrics created for the Lake Problem.

Performance metric	**T1**	**T2**	**T3**
Average phosphorous	Identity	25th percentile	Identity
Reliability	Satisfice (threshold 80%)	All scenarios	Mean
Economic utility	Magnitude below threshold of 0.75	50th percentile	Identity
Inertia	Identity	50th percentile	Identity

For the reliability, we assume a situation where a requirement for the project is a minimum of 80% reliability for whichever decision alternative is selected, and that this requirement should be met in as many scenarios as possible. Thus, the T1 transformation is the satisficing transform and the T3 transformation is the mean. It is also decided that the aim is to understand what percentage of all scenarios under consideration have acceptable performance, and so all scenarios are selected for T2.

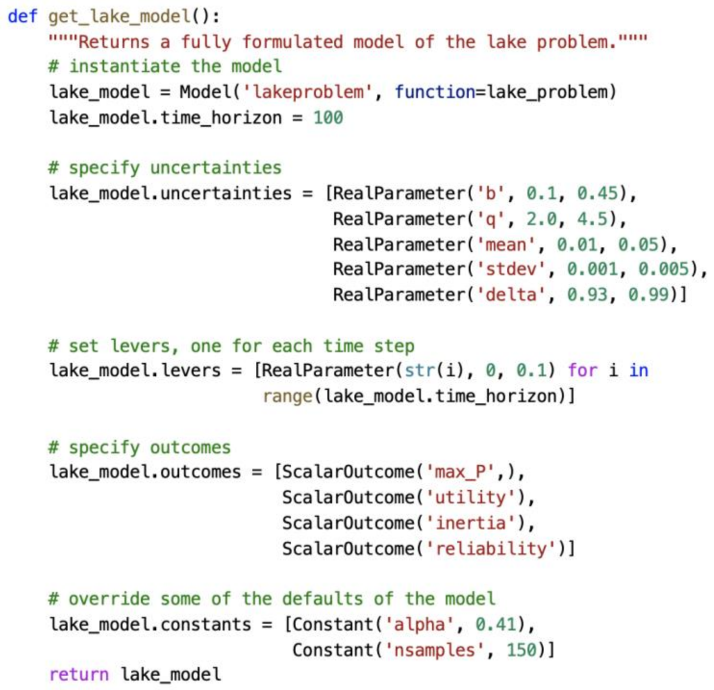
For the economic utility, it is assumed that a level of 0.75 is required, and that any level lower than this will have significant consequences. Therefore, the satisficing regret transform is used, since that includes the threshold of 0.75, but also penalizes decision alternatives in each scenario that fail to achieve this. The level of performance (i.e. the level of potential regret) is most important, and therefore the identity transform is used for T3. It is also assumed that the decision-maker has a moderate level of risk aversion for this objective, and T2 is the 50th percentile of performance (i.e. regret).

The inertia is a measure of how much the decision alternative options vary from year to year (it is preferred that there are no significant changes in the level of pollution from one year to the next). We are not using a specific threshold for this (although some other studies have), and the objective of the decision-maker is to make the best decision regarding the level of performance (level of inertia). Therefore, the identity transform is chosen for T1 and T3. Again, the level of risk aversion is moderate for this objective, and thus the 50th percentile is chosen for T2.

Returning to the overarching guidance framework for robustness analysis (Figure 4-2 and Figure 4-4), now that we have the robustness metrics (Figure 4-4, Box 4) and the scenarios are known (Figure 4-4, Box 6), we can calculate the robustness using the selected scenarios and selected (custom) robustness metrics (Figure 4-4, Box 10). To illustrate this with the RAPID software package, we build upon an example of the Lake Problem that is included in the EM Workbench (Kwakkel, 2017), with the following methodology:

1. Using the EM Workbench, we formulate the model (e.g. uncertain parameters, objectives, etc.).
2. Using the RAPID package, we create the custom robustness metrics defined above in Table 4-4.
3. Using the EM Workbench, we formulate an optimization problem with the formulated model (from Step 1) and custom robustness metrics (from Step 2).
4. Using the EM Workbench, we run the optimization to determine the most robust decision alternatives.

For Step 1, the Lake Problem was specified in the same manner as in the EM Workbench example (i.e. the uncertain parameters, options for the decision alternatives, and the performance objectives were defined in the same way) using the EM Workbench functionality for defining a model (Figure 4-5).

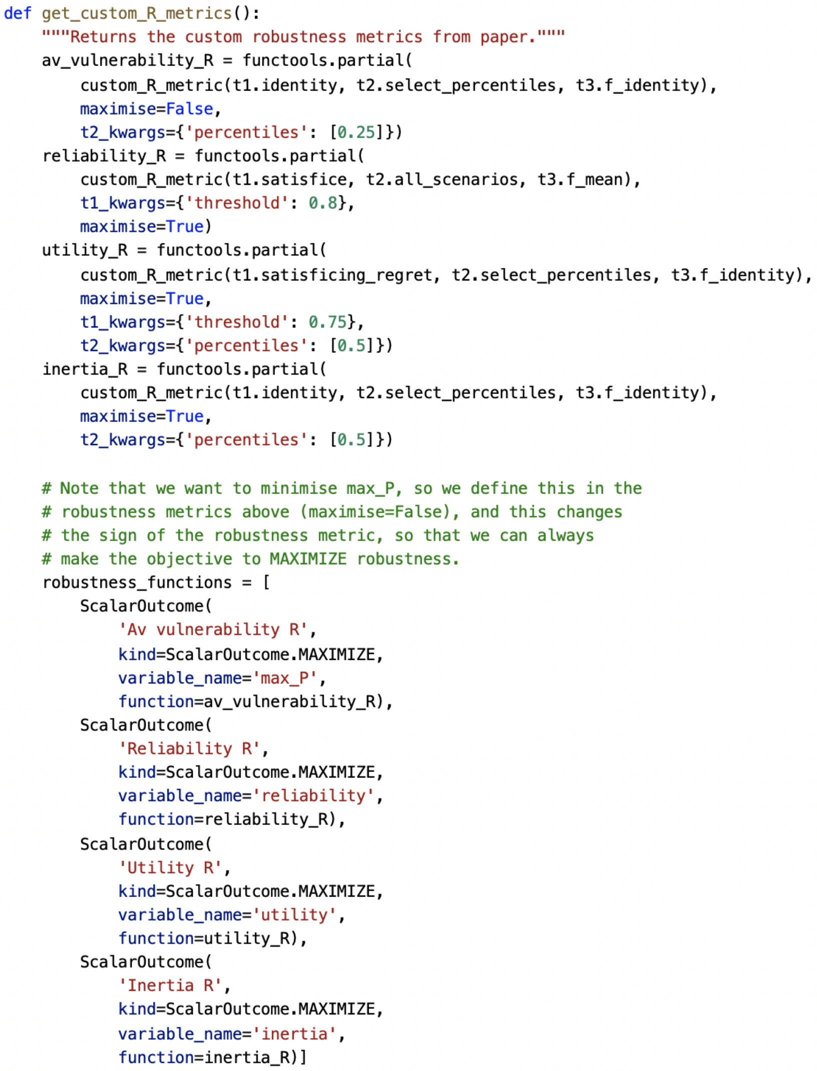


```
def get_lake_model():
    """Returns a fully formulated model of the lake problem."""
    # instantiate the model
    lake_model = Model('lakeproblem', function=lake_problem)
    lake_model.time_horizon = 100

    # specify uncertainties
    lake_model.uncertainties = [RealParameter('b', 0.1, 0.45),
                                RealParameter('q', 2.0, 4.5),
                                RealParameter('mean', 0.01, 0.05),
                                RealParameter('stdev', 0.001, 0.005),
                                RealParameter('delta', 0.93, 0.99)]

    # set levers, one for each time step
    lake_model.levers = [RealParameter(str(i), 0, 0.1) for i in
                         range(lake_model.time_horizon)]

    # specify outcomes
    lake_model.outcomes = [ScalarOutcome('max_P',),
                           ScalarOutcome('utility'),
                           ScalarOutcome('inertia'),
                           ScalarOutcome('reliability')]

    # override some of the defaults of the model
    lake_model.constants = [Constant('alpha', 0.41),
                            Constant('nsamples', 150)]
    return lake_model
```

Figure 4-5: Code snippet - formulation of the lake model.

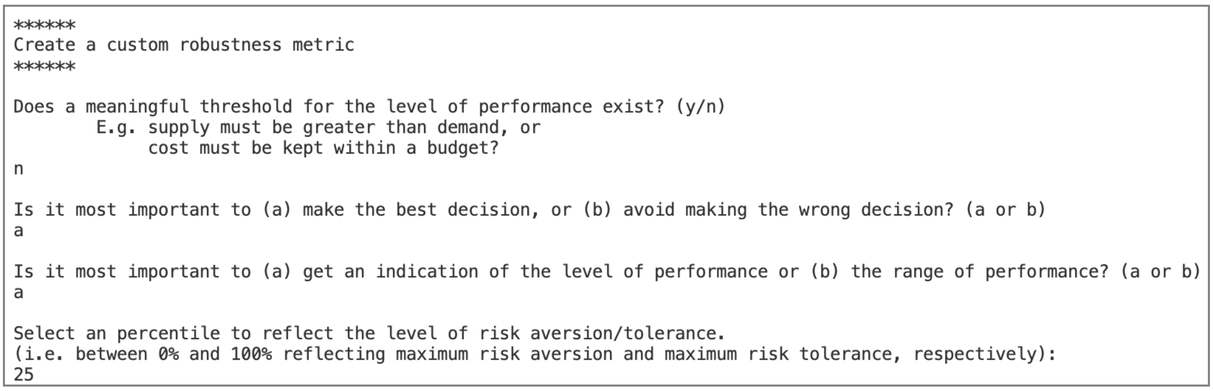
For Step 2, the custom robustness metrics defined in Table 4-4 were first specified using the RAPID package and then put into the form required for the EM Workbench (Figure

4-6). Note that when defining these custom metrics, it was possible to use any combination of the three robustness metric transformations (from the guidance for decision-makers Figure 4-3, and defined in Table 4-1). These metrics can be defined using code as shown or can also be created using the *metrics.guidance_to_R* function. This function asks the user the questions from the flow chart in Figure 4-3, guiding them to the creation of the robustness metric best suited for the problem that they can then use in proceeding analyses (as shown in Figure 4-7). The output from the *metrics.guidance_to_R* function is the same as the output from the *metrics.custom_R_metric* function in the example code.



```python
def get_custom_R_metrics():
    """Returns the custom robustness metrics from paper."""
    av_vulnerability_R = functools.partial(
        custom_R_metric(t1.identity, t2.select_percentiles, t3.f_identity),
        maximise=False,
        t2_kwargs={'percentiles': [0.25]})
    reliability_R = functools.partial(
        custom_R_metric(t1.satisfice, t2.all_scenarios, t3.f_mean),
        t1_kwargs={'threshold': 0.8},
        maximise=True)
    utility_R = functools.partial(
        custom_R_metric(t1.satisficing_regret, t2.select_percentiles, t3.f_identity),
        maximise=True,
        t1_kwargs={'threshold': 0.75},
        t2_kwargs={'percentiles': [0.5]})
    inertia_R = functools.partial(
        custom_R_metric(t1.identity, t2.select_percentiles, t3.f_identity),
        maximise=True,
        t2_kwargs={'percentiles': [0.5]})

    # Note that we want to minimise max_P, so we define this in the
    # robustness metrics above (maximise=False), and this changes
    # the sign of the robustness metric, so that we can always
    # make the objective to MAXIMIZE robustness.
    robustness_functions = [
        ScalarOutcome(
            'Av vulnerability R',
            kind=ScalarOutcome.MAXIMIZE,
            variable_name='max_P',
            function=av_vulnerability_R),
        ScalarOutcome(
            'Reliability R',
            kind=ScalarOutcome.MAXIMIZE,
            variable_name='reliability',
            function=reliability_R),
        ScalarOutcome(
            'Utility R',
            kind=ScalarOutcome.MAXIMIZE,
            variable_name='utility',
            function=utility_R),
        ScalarOutcome(
            'Inertia R',
            kind=ScalarOutcome.MAXIMIZE,
            variable_name='inertia',
            function=inertia_R)]
```

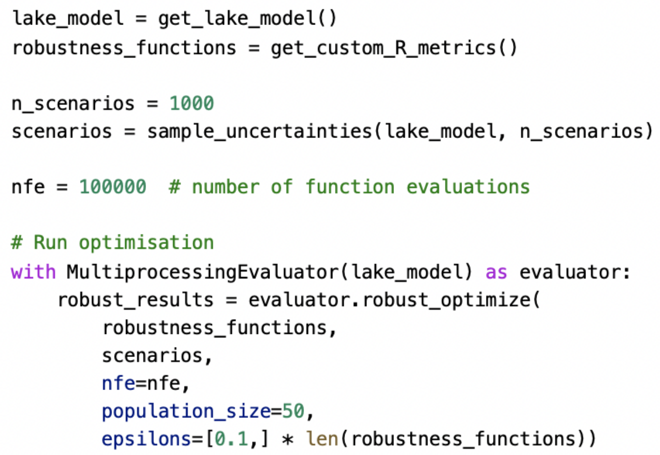
Figure 4-6: Code snippet - creation of custom robustness metrics using the RAPID package, followed by putting this in the form required for the EM Workbench.



```
******
Create a custom robustness metric
******

Does a meaningful threshold for the level of performance exist? (y/n)
        E.g. supply must be greater than demand, or
             cost must be kept within a budget?
n

Is it most important to (a) make the best decision, or (b) avoid making the wrong decision? (a or b)
a

Is it most important to (a) get an indication of the level of performance or (b) the range of performance? (a or b)
a

Select an percentile to reflect the level of risk aversion/tolerance.
(i.e. between 0% and 100% reflecting maximum risk aversion and maximum risk tolerance, respectively):
25
```

Figure 4-7: Example of the dialogue provided by the *metrics.guidance_to_R* function in the RAPID package.

As per the EM Workbench example, once the model has been formulated and the robustness metrics have been defined, the next step is to use the EM Workbench to create a set of scenarios, formulate an optimization problem, and then run that optimization problem to find optimally robust decision alternatives (Figure 4-8). This corresponds to the loop formed by Box 18 in Figure 4-4. The results found from this process are shown in Figure 4-9. Note again that the robustness metric transformations from the RAPID software package ensure that a higher robustness value is always better (e.g. we seek to minimize vulnerability, but the sign for the robustness metric for vulnerability is switched so that we are aiming to maximize the robustness value). The Pareto front (Figure 4-9) shows expected relationships between objectives. For example, better vulnerability also results in better reliability but a worse result for the economic utility. The relationship between the inertia and the other three objectives is weaker.



```
lake_model = get_lake_model()
robustness_functions = get_custom_R_metrics()

n_scenarios = 1000
scenarios = sample_uncertainties(lake_model, n_scenarios)

nfe = 100000  # number of function evaluations

# Run optimisation
with MultiprocessingEvaluator(lake_model) as evaluator:
    robust_results = evaluator.robust_optimize(
        robustness_functions,
        scenarios,
        nfe=nfe,
        population_size=50,
        epsilons=[0.1,] * len(robustness_functions))
```

Figure 4-8: Code snippet - formulation and execution of the optimization of The Lake Problem using robustness metrics from the RAPID software package and the optimization functionality from the EM Workbench.

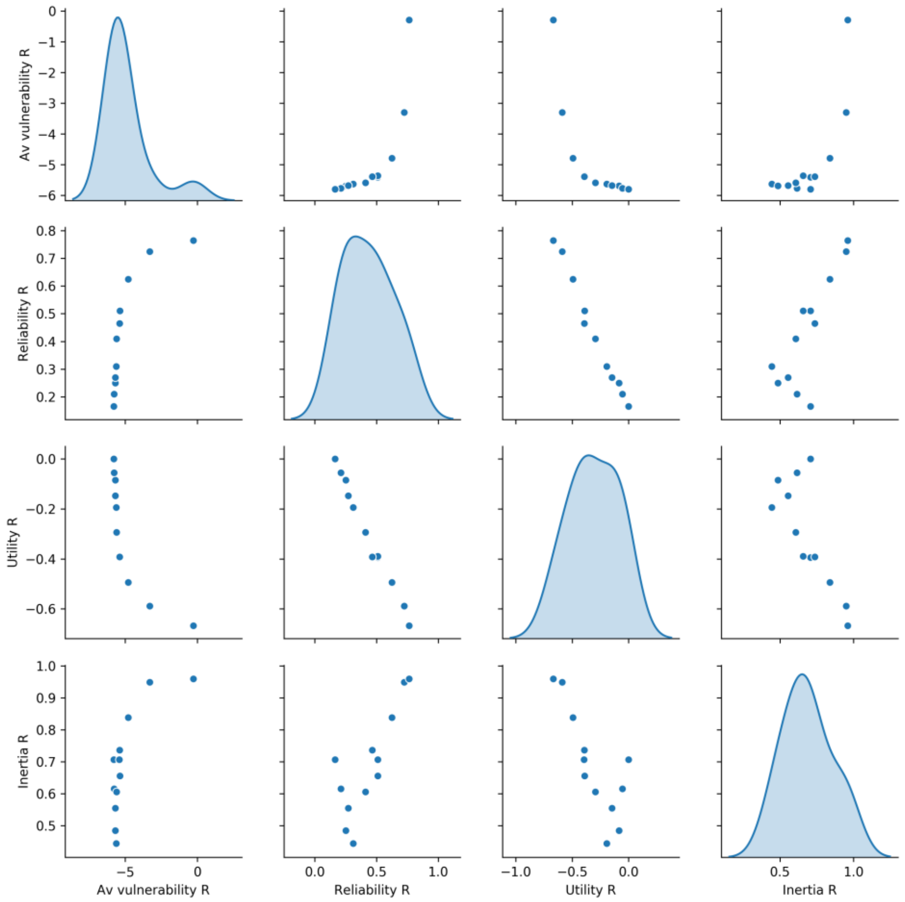


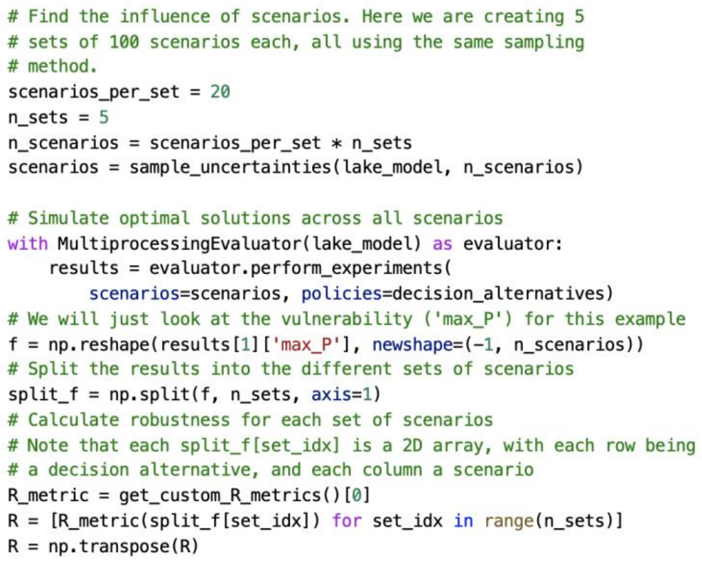
Figure 4-9: Example results that can be produced using custom robustness metrics from the RAPID package and multi-objective optimisation functionality from the EM Workbench package. The axes are the robustness metrics and each point represents the robustness of a single solution from the 4-dimensional Pareto front.

In this example of following the guidance framework (Figure 4-2 and Figure 4-4), we showed that with no known robustness metric or set of candidate robustness metrics we could create a set of custom robustness metrics that were best suited to the problem (Table 4-4) using the guidance for creating a custom robustness metric (Figure 4-3) to determine the appropriate robustness metric transformations from Table 4-2. We then created these robustness metrics in a systematic manner using the RAPID software package and used these newly created robustness metrics in conjunction with another software package, the EM Workbench, to run a robust optimization and develop a Pareto front of optimal decision alternatives.

4.4.3. No candidate robustness metrics and multiple candidate scenario sets

Again, following the guidance framework, we use the optimal decision alternatives from the previous section and we assume a situation in which the robustness metric is unknown (Figure 4-4, Box 2) and there are no candidate robustness metrics (Figure 4-4, Box 3), leading to Box 4 in Figure 4-4. Here, we create custom robustness metrics as per Section 4.4.2, leading to the robustness metrics in Table 4-4. Unlike Section 4.4.2, in this section we consider a situation where there are multiple candidate sets of scenarios (Figure 4-4, Box 6).

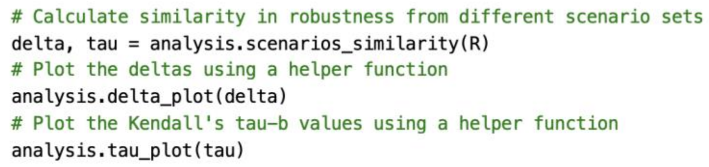
Different sets of scenarios correspond to different sets of points within the space of uncertain model inputs (McPhail et al., 2020). Because these points are inputs to the calculation of robustness (see Figure 4-1), different sets of scenarios can lead to differences in robustness. As a simplified illustration of this, we create five candidate sets of 20 scenarios, where each set is sampled from the uncertain variable space using the EM Workbench package with Latin hypercube sampling (Figure 4-10). We then evaluate the optimal decision alternatives (from Section 4.4.2) in all 100 scenarios using the EM Workbench package and calculate the robustness for all 5 scenario sets and all decision alternatives using the custom robustness metrics created in Section 4.4.2 using the RAPID package (Figure 4-4, Box 9). Note that for simplicity, we only focus on the vulnerability objective from here on. The same analysis could be applied to each of the four objectives.



```
# Find the influence of scenarios. Here we are creating 5
# sets of 100 scenarios each, all using the same sampling
# method.
scenarios_per_set = 20
n_sets = 5
n_scenarios = scenarios_per_set * n_sets
scenarios = sample_uncertainties(lake_model, n_scenarios)

# Simulate optimal solutions across all scenarios
with MultiprocessingEvaluator(lake_model) as evaluator:
    results = evaluator.perform_experiments(
        scenarios=scenarios, policies=decision_alternatives)
# We will just look at the vulnerability ('max_P') for this example
f = np.reshape(results[1]['max_P'], newshape=(-1, n_scenarios))
# Split the results into the different sets of scenarios
split_f = np.split(f, n_sets, axis=1)
# Calculate robustness for each set of scenarios
# Note that each split_f[set_idx] is a 2D array, with each row being
# a decision alternative, and each column a scenario
R_metric = get_custom_R_metrics()[0]
R = [R_metric(split_f[set_idx]) for set_idx in range(n_sets)]
R = np.transpose(R)
```

Figure 4-10: Code snippet - Creating 5 candidate scenario sets of 20 scenarios each, evaluating them, and calculating robustness for each of these 5 sets.

Returning to the robustness analysis guidance framework, this brings us to Box 13 in Figure 4-4, where we use the *analysis* module of the RAPID package to evaluate the relative difference in robustness values and the Kendall's Tau-b rank correlation (for determining the ranking similarity) (Figure 4-11), as described in Section 4.3. The *analysis* module also enables us to visualize the influence of the scenarios by creating heatmaps that show all combinations of candidate sets of scenarios (see Figure 4-12 (a) and (b)). The diagonal of the heatmaps is each candidate scenario set compared to itself, and therefore the relative difference is 0% (indicated by purple in Figure 4-12 (a)) and the ranking correlation is 1 (indicated by blue in Figure 4-12 (b)) as expected. From Figure 4-12 (a) we can see that for the other comparisons of the scenario sets, the relative difference in robustness values is very high in general (indicated by mostly orange squares, ~30% difference in robustness values), however there are some cases (e.g. scenario sets 1 and 5, and scenario sets 4 and 5) that are more similar than the rest (indicated by the green). Note that despite a high difference in robustness values, Figure 4-12 (b) indicates that the rankings of the decision alternatives are very stable (consistent with McPhail et al. (2020)).



```
# Calculate similarity in robustness from different scenario sets
delta, tau = analysis.scenarios_similarity(R)
# Plot the deltas using a helper function
analysis.delta_plot(delta)
# Plot the Kendall's tau-b values using a helper function
analysis.tau_plot(tau)
```

Figure 4-11: Code snippet - Calculation and visualization of the impact of scenario selection on robustness and robustness rankings of the decision alternatives.

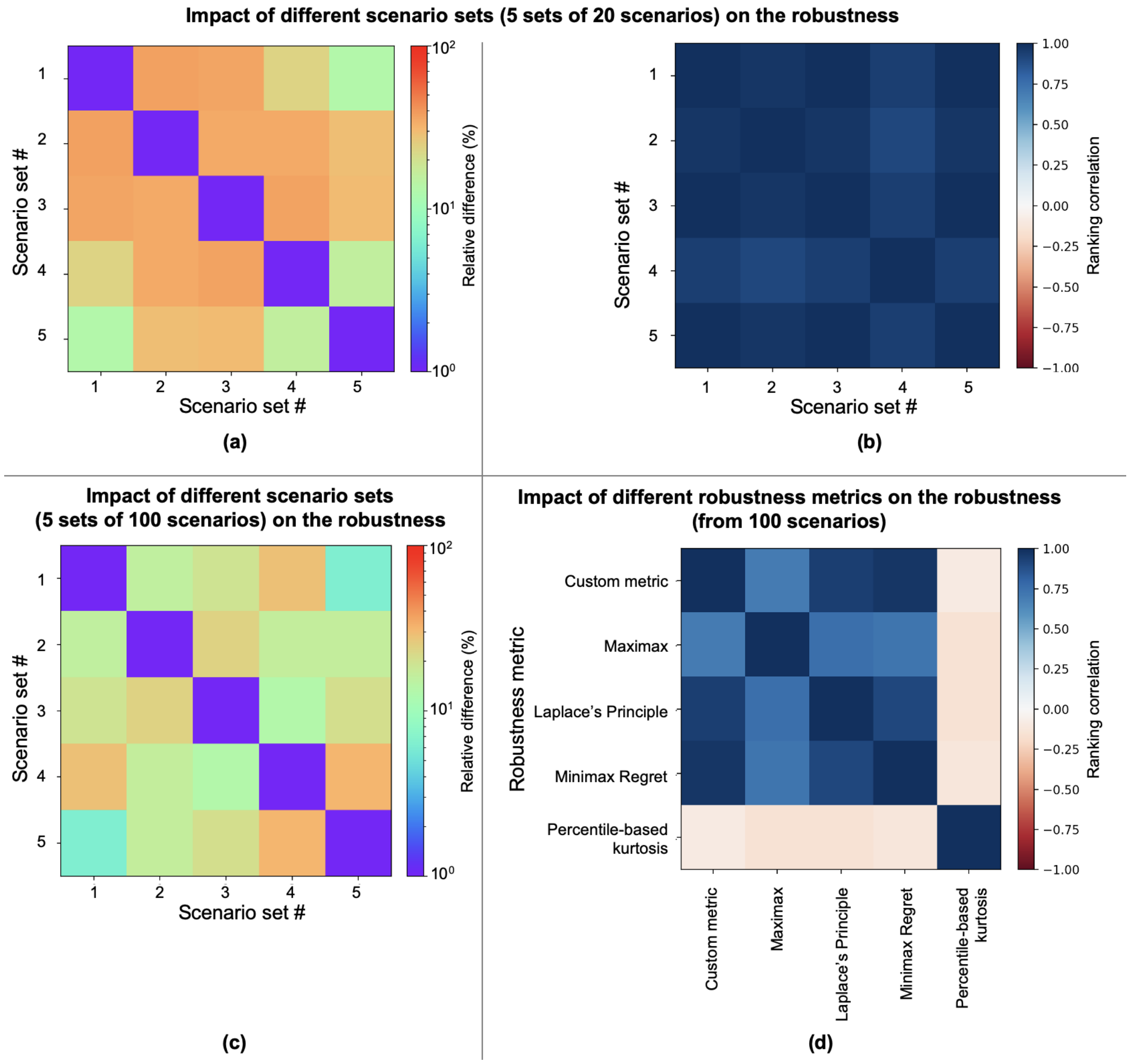


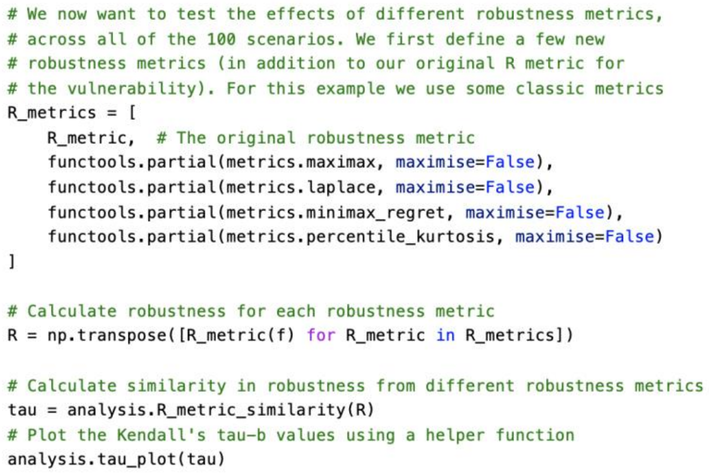
Figure 4-12: Example of outputs produced by the RAPID package. For the Lake Problem analysed as described above: (a) relative difference in robustness for pairs of scenario sets (5 sets of 20 scenarios); (b) ranking similarity for pairs of scenario sets (5 sets of 20 scenarios); (c) relative difference in robustness for pairs of scenario sets (5 sets of 100 scenarios); (d) ranking similarity for pairs of robustness metrics (one set of 100 scenarios).

Given that all five candidate sets of scenarios were sampled using Latin hypercube sampling, it is interesting that the relative difference in robustness is so high in Figure 4-12 (a). If the robustness values were important for the decision-making process, it would be difficult to be sure of the actual robustness values because the values would depend on which set of scenarios is being considered (leading to Figure 4-4, Box 15). There are many reasons why the relative difference could be high, including dissimilarity in the coverage of the scenario space, and discontinuities in performance space (McPhail et al., 2020). In this example, we can use judgement to estimate it is the former of these reasons, because the number of scenarios in each set is small. Running the same code as above but with a larger number of scenarios (100 scenarios per set, rather than 20 scenarios per set in Figure 4-12 (a)), we produce the heatmap shown in Figure 4-12 (c). With the larger number of scenarios, the relative difference is significantly lower in general (likely due to a more similar coverage of the scenario space), indicated by the greater number of blue and green squares and smaller number of orange squares. In this case, we move from Box 14 to Boxes 16 and 17 in Figure 4-4, being able to accurately determine the robustness of the decision alternatives. Alternatively, if we are simply interested in the rankings of the solutions (see Figure 4-12 (b)), then we would be able to move from Box 14 to Boxes 16 and 17 without increasing the number of scenarios (assuming that we judge the Kendall's Tau-b values (approximately in the range between 0.7 and 1.0) to be sufficiently high for our purposes.

In this second example of following the guidance framework (Figure 4-2 and Figure 4-4), we showed that with multiple candidate sets of scenarios, we could use the RAPID software package to evaluate the influence these candidate sets of scenarios had on both the robustness values and rankings. Using the visualizations produced by the software package, we were then able to determine that the relative robustness values of different decision alternatives was not substantially affected by the different scenario sets (Figure 4-6b), giving confidence to decision makers and enabling the most robust decision alternative to be identified.

4.4.4. Multiple candidate robustness metrics and a known set of scenarios

In this situation, we assume that the robustness metric is unknown (Figure 4-4, Box 2), but that there are multiple candidate robustness metrics (Figure 4-4, Box 5) and that the set of scenarios is known, leading to Box 8 in Figure 4-4. Note that if there were multiple candidate sets of scenarios, the analysis would be a combination of the following method and the method in Section 4.4.3. We create the candidate robustness metrics using the RAPID software package, retaining the original robustness metric for the vulnerability determined in Section 4.4.2 (Table 4-4) using the *metrics.custom_R_metric* module, and four traditional robustness metrics as the other candidate metrics, including the Maximax, Laplace's Principle of Insufficient Reason, Minimax Regret, and Percentile-Based Kurtosis robustness metrics (all included in the *metrics.common_metrics* module). As with the previous examples, these metrics were calculated, evaluated (this time across a known set of 100 scenarios, sampled using Latin hypercube sampling), and visualized using the RAPID package (see Figure 4-12 (d)).



```
# We now want to test the effects of different robustness metrics,
# across all of the 100 scenarios. We first define a few new
# robustness metrics (in addition to our original R metric for
# the vulnerability). For this example we use some classic metrics
R_metrics = [
    R_metric,  # The original robustness metric
    functools.partial(metrics.maximax, maximise=False),
    functools.partial(metrics.laplace, maximise=False),
    functools.partial(metrics.minimax_regret, maximise=False),
    functools.partial(metrics.percentile_kurtosis, maximise=False)
]

# Calculate robustness for each robustness metric
R = np.transpose([R_metric(f) for R_metric in R_metrics])

# Calculate similarity in robustness from different robustness metrics
tau = analysis.R_metric_similarity(R)
# Plot the Kendall's tau-b values using a helper function
analysis.tau_plot(tau)
```

Figure 4-13: Code snippet – Creation of robustness metrics, and calculation and visualization of the impact of the robustness metrics on robustness rankings of the decision alternatives.

In the visualization of the similarity in rankings (Figure 4-12 (d)), the diagonal shows full ranking similarity (a value of 1, indicated by blue) because that is where each robustness metric is being compared to itself. Most of the metrics also show high levels of ranking similarity with each other, with the exception of the percentile-based kurtosis metric, which shows a slight negative correlation with all other metrics (indicated by the slightly red squares). This potentially leads us from Box 14 to Box 15 in Figure 4-4, because it is unknown which ranking is the one that we should follow: the rankings provided by the percentile-based kurtosis or the rankings provided by the rest of the metrics. Again, using our judgement, we decide that the percentile-based kurtosis does not reflect the needs of the decision-makers as much as the other robustness metrics do, because the T3 transformation does not reflect the need to get an indication of the level of performance (as explained Figure 4-3 and by McPhail et al. (2018)). Also, since all of the other candidate solutions generally agree with the custom robustness metric, it follows that we can rely on this custom metric to determine which decision alternative is most robust (Figure 4-4, Box 16).

In this final illustration of using the guidance framework (Figure 4-2 and Figure 4-4) and RAPID software package, we showed that with multiple candidate robustness metrics, we could use the software package to evaluate the influence these robustness metrics had on the rankings of the decision alternatives. Using the visualizations produced by the software package, we were then able to determine whether or not the influence was great enough to affect these rankings.

All three of the simple examples considered show that the RAPID package is easy to use and can be used in conjunction with other related software packages, such as the EM Workbench. They also show that the RAPID package is a practical tool for systematically following the guidance framework in Figure 4-2 and Figure 4-4, the guidance for creating robustness metrics in Figure 4-3 (shown in Section 4.4.2), assessing the influence of candidate sets of scenarios on the robustness values and rankings (shown in Section 4.4.3), and assessing the influence of candidate robustness metrics on the robustness rankings of decision alternatives (shown in Section 4.4.4).

4.5. Summary and conclusions

Robustness is important in the long-term planning of environmental systems. However, there is a variety of metrics that can be used to calculate the robustness of a set of decision alternatives, and recent research has shown that the choice of metric can affect the ranking of decision alternatives. Similarly, there is a variety of approaches to selecting or generating scenarios (which are an input to the calculation of robustness), and the chosen approach has also been shown to have an effect on the robustness values and rankings of decision alternatives. Despite the uncertainty in which selection of scenarios or which robustness metric to use to determine the rankings of decision alternatives, no guidance exists for decision-makers on which choices to make.

As a response to this need for guidance, this paper proposes a generic guidance framework to assist decision-makers in the identification of robust decision alternatives (Figure 4-2). This framework caters to a variety of situations where the scenarios and/or robustness metrics are known or not known. The framework includes guidance on how to create a custom robustness metric for the problem at hand (Figure 4-3), based on the attributes of the problem (e.g. the presence of performance thresholds / tipping points, or the objectives of the problem) as well as the preferences of the decision-maker (e.g. the level of risk-aversion). The output from the guidance for the creation of a custom robustness metric is three robustness metric transformations (Table 4-1), which form the robustness metric when combined (Figure 4-1). The overarching guidance framework also identifies situations where quantitative analyses can be used to determine the influence that the selection of scenarios and/or the choice of robustness metric has on the rankings of decision alternatives.

This paper also introduces an open-source software package, the RAPID (Robustness Analysis Producing Intelligent Decisions) package, to assist in the consistency and ease-of-use of implementing the guidance framework (see Figure 4-4). The software package includes a module for the creation of custom robustness metrics using a wide range of robustness metric transformations (Table 4-2), including a function that leads the user through the guidance of how to create the robustness metric most suited for the problem at hand (Figure 4-3). It also includes a variety of traditional robustness metrics from the literature (Table 4-3), commonly used in the absence of the guidance introduced in this paper. The software package also contains a module for the calculation and visualization of the impact of the selection of scenarios and choice of robustness metric on robustness values and rankings.

To illustrate the implementation of the guidance framework and RAPID software package, we consider the Lake Problem, a hypothetical lake pollution problem, commonly used in the literature. We use the guidance in Figure 4-3 to create custom robustness metrics for The Lake Problem, based on hypothetical problem attributes and decision-maker preferences (Table 4-4). In conjunction with the EM Workbench (Kwakkel, 2017), we use these robustness metrics as objectives in a robust optimization to create a set of robust decision alternatives. As an example of the utility of the guidance framework and software package, we use these optimal decision alternatives to consider a situation where there are multiple sets of scenarios under consideration. Using the RAPID software package, we visualize the impact of these different sets of scenarios, showing that the robustness values are affected (Figure 4-12 (a)), but rankings of the decision alternatives are not (Figure 4-12 (b)), providing confidence to decision makers that the most robust decision alternative has been identified. We also show that when using a larger set of scenarios, the impact of the set of scenarios on the robustness values is greatly decreased (Figure 4-12 (c)). In another example to highlight the utility of the guidance framework and software package, we consider a situation where there is a variety of candidate robustness metrics. We use the framework and software package to visualize the impact of the choice of robustness metric (Figure 4-12 (d)), showing that most of the metrics agree on the rankings of the decision alternatives, again providing confidence to decision makers that the most robust solution has been identified.

This guidance framework and software package assist decision-makers in the identification of robust decision alternatives. It does so in a systematic way, and the software package increases the consistency and ease-of-use of implementing the guidance. The guidance framework and software package are generic and cater to a wide variety of circumstances where the robustness metrics and/or scenarios may or may not be known, greatly increasing the accessibility of robustness analyses and techniques to decision-makers.

Acknowledgements

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

The Lake Model is widely available on GitHub in multiple repositories, including in the EMAworkbench: https://github.com/quaquel/EMAworkbench

The RAPID (Robustness Analysis Producing Intelligent Decisions) software package is available on GitHub (https://github.com/cameronmcphail/RAPID) and in the Python Package Index (PyPI) (https://pypi.org/project/rapidrobustness/).

References

Borgomeo, E., Mortazavi-Naeini, M., Hall, J.W., Guillod, B.P., 2018. Risk, Robustness and Water Resources Planning Under Uncertainty. Earth's Futur. 6, 468–487.

Börjeson, L., Höjer, M., Dreborg, K.H., Ekvall, T., Finnveden, G., 2006. Scenario types and techniques: Towards a user's guide. Futures 38, 723–739. https://doi.org/10.1016/j.futures.2005.12.002

Bradfield, R., Wright, G., Burt, G., Cairns, G., Van Der Heijden, K., 2005. The origins and evolution of scenario techniques in long range business planning. Futures 37, 795–812. https://doi.org/10.1016/j.futures.2005.01.003

Carpenter, S.R., Ludwig, D., Brock, W.A., 1999. Management of eutrophication for lakes subject to potentially irreversible change. Ecol. Appl. 9, 751–771.

Culley, S., Bennett, B., Westra, S., Maier, H.R., 2019. Generating realistic perturbed hydrometeorological time series to inform scenario-neutral climate impact assessments. J. Hydrol.

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

Drouet, L., Bosetti, V., Tavoni, M., 2015. Selection of climate policies under the uncertainties in the Fifth Assessment Report of the IPCC. Nat. Clim. Chang. 5, 937–940.

Eker, S., Kwakkel, J.H., 2018. Including robustness considerations in the search phase of Many-Objective Robust Decision Making. Environ. Model. Softw. 105, 201–216. https://doi.org/10.1016/j.envsoft.2018.03.029

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

Hadka, D., Herman, J., Reed, P., Keller, K., 2015. An open source framework for many-objective robust decision making. Environ. Model. Softw. 74, 114–129. https://doi.org/10.1016/j.envsoft.2015.07.014

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 Anal. 32, 1657–1672. https://doi.org/10.1111/j.1539-6924.2012.01802.x

Herman, J.D., Reed, P.M., Zeff, H.B., Characklis, G.W., 2015. How Should Robustness Be Defined for Water Systems Planning under Change? J. Water Resour. Plan. Manag. 141, 04015012. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000509

Herman, J.D., Zeff, H.B., Reed, P.M., Characklis, G.W., 2014. Beyond optimality: Multistakeholder robustness tradeoffs for regional water portfolio planning under deep uncertainty. Water Resour. Res. 50, 7692–7713.

Kasprzyk, J.R., Nataraj, S., Reed, P.M., Lempert, R.J., 2013. Many objective robust decision making for complex environmental systems undergoing change. Environ. Model. Softw. 42, 55–71. https://doi.org/10.1016/j.envsoft.2012.12.007

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

Kwakkel, J.H., Eker, S., Pruyt, E., 2016a. How robust is a robust policy? Comparing alternative robustness metrics for robust decision-making, in: International Series in Operations Research and Management Science. Springer, pp. 221–237.

https://doi.org/10.1007/978-3-319-33121-8_10

Kwakkel, J.H., Haasnoot, M., 2019. Supporting DMDU: A taxonomy of approaches and tools, in: Decision Making under Deep Uncertainty. Springer, pp. 355–374.

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. Clim. Change 132, 373–386. https://doi.org/10.1007/s10584-014-1210-4

Kwakkel, J.H., Walker, W., Marchau, V., 2012. Assessing the efficacy of adaptive airport strategic planning: Results from computational experiments. Environ. Plan. B Plan. Des. 39, 533–550.

Kwakkel, J.H., Walker, W.E., Haasnoot, M., 2016b. Coping with the Wickedness of Public Policy Problems: Approaches for Decision Making under Deep Uncertainty. J. Water Resour. Plan. Manag. 142, 01816001. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000626

Kwakkel, J.H., Walker, W.E., Marchau, V.A.W.J., 2010. Classifying and communicating uncertainties in model-based policy analysis. Int. J. Technol. Policy Manag. 10, 299. https://doi.org/10.1504/IJTPM.2010.036918

Lempert, R.J., 2003. Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis. Rand Corporation. https://doi.org/10.1016/j.techfore.2003.09.006

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

Little, J.C., Hester, E.T., Elsawah, S., Filz, G.M., Sandu, A., Carey, C.C., Iwanaga, T., Jakeman, A.J., 2018. A tiered, system-of-systems modeling framework for resolving complex socio-environmental policy issues. Environ. Model. Softw.

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? Environ. Model. Softw. 81, 154–164. https://doi.org/10.1016/j.envsoft.2016.03.014

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 Futur. 6, 169–191. https://doi.org/10.1002/2017EF000649

McPhail, C., Maier, H.R., Westra, S., Kwakkel, J.H., van der Linden, L., 2020. Impact of scenario selection on robustness. Water Resour. Res.

Phadnis, S., 2019. Effectiveness of Delphi-and scenario planning-like processes in enabling organizational adaptation: A simulation-based comparison. Futur. Foresight Sci. e9.

Quinn, J.D., Reed, P.M., Giuliani, M., Castelletti, A., Oyler, J.W., Nicholas, R.E., 2018. Exploring how changing monsoonal dynamics and human pressures challenge multireservoir management for flood protection, hydropower production, and agricultural water supply. Water Resour. Res. 54, 4638–4662.

Quinn, J.D., Reed, P.M., Keller, K., 2017. Direct policy search for robust multi-objective management of deeply uncertain socio-ecological tipping points. Environ. Model. Softw. 92, 125–141.

Roach, T., Kapelan, Z., Ledbetter, R., Ledbetter, M., 2016. Comparison of Robust Optimization and Info-Gap Methods for Water Resource Management under Deep Uncertainty. J. Water Resour. Plan. Manag. 142, 04016028. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000660

Savage, L.J., 1951. The theory of statistical decision. J. Am. Stat. Assoc. 46, 55–67. https://doi.org/10.1080/01621459.1951.10500768

Schwarz, P., 1991. The art of the long view: planning for the future in an uncertain world. John Wiley & Sons, Chichester, England.

Shepherd, T.G., Boyd, E., Calel, R.A., Chapman, S.C., Dessai, S., Dima-West, I.M., Fowler, H.J., James, R., Maraun, D., Martius, O., 2018. Storylines: an alternative approach to representing uncertainty in physical aspects of climate change. Clim. Change 1–17.

Simon, H.A., 1956. Rational choice and the structure of the environment. Psychol. Rev. 63, 129–138. https://doi.org/10.1037/h0042769

Singh, R., Reed, P.M., Keller, K., 2015. Many-objective robust decision making for managing an ecosystem with a deeply uncertain threshold response. Ecol. Soc. 20.

Trindade, B.C., Reed, P.M., Herman, J.D., Zeff, H.B., Characklis, G.W., 2017. Reducing regional drought vulnerabilities and multi-city robustness conflicts using many-objective optimization under deep uncertainty. Adv. Water Resour. 104, 195–209.

van der Heijden, K., 1996. Scenarios: the art of strategic conversation. John Wiley & Sons.

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

Wada, Y., Vinca, A., Parkinson, S., Willaarts, B.A., Magnuszewski, P., Mochizuki, J., Mayor, B., Wang, Y., Burek, P., Byers, E., 2019. Co-designing Indus Water-Energy-Land Futures. One Earth 1, 185–194.

Wald, A., 1951. Statistical decision functions, Nature. New York; Chapman & Hall: London. https://doi.org/10.1038/1671044b0

Walker, W.E., Lempert, R., Kwakkel, J., 2013. Deep Uncertainty, in: Encyclopedia of

Operations Research and Management Science. Springer, pp. 395–402. https://doi.org/10.1007/978-1-4419-1153-7_1140

Ward, V.L., Singh, R., Reed, P.M., Keller, K., 2015. Confronting tipping points: Can multi-objective evolutionary algorithms discover pollution control tradeoffs given environmental thresholds? Environ. Model. Softw. 73, 27–43.

Watson, A.A., Kasprzyk, J.R., 2017. Incorporating deeply uncertain factors into the many objective search process. Environ. Model. Softw. 89, 159–171. https://doi.org/10.1016/j.envsoft.2016.12.001

Weaver, C.P., Lempert, R.J., Brown, C., Hall, J.A., Revell, D., Sarewitz, D., 2013. Improving the contribution of climate model information to decision making: the value and demands of robust decision frameworks. Wiley Interdiscip. Rev. Clim. Chang. 4, 39–60.

Wright, G., Cairns, G., 2011. Scenario thinking: Practical approaches to the future. Springer.

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

Chapter 5

Long-term decision making for water and environmental systems presents significant challenges to decision-makers. Prominent among these challenges is how to make decisions under conditions of deep uncertainty, where deterministic and stochastic processes are insufficient for representing the future state of the world, and the consideration of multiple plausible futures is required. Recent research into decision-making under deep uncertainty has highlighted a number of challenges and questions including (i) how do different robustness metrics affect the robustness of the system, (ii) how do different selections of scenarios affect the robustness of a system, and (iii) how do decision-makers decide which robustness metrics should be used for any given problem and which of the decision alternatives under consideration are most robust?

This research contributes to the field of long-term decision-making for water and environmental systems by achieving the aims set out at the beginning of this thesis: (i) to introduce a unified framework for the calculation of a wide range of robustness metrics, enabling the robustness values and rankings obtained from different metrics to be compared in an objective fashion; (ii) to develop a deeper understanding of how different selections of scenarios can affect the absolute and relative robustness of the decision alternatives of interest; and (iii) to create a generic guidance framework and software tool to assist with the identification of the most robust decision alternative for a given problem.

5.1. Research contributions

The overall contribution of this research is that it provides a better understanding of robustness for the long-term planning of water and environmental systems and how to identify the most robust decision alternatives. This provides decision-makers with better information, understanding guidance, and confidence for making decisions on these complex systems. More specifically, the contributions of this thesis are:

1. In Chapter 2, we contribute to the field a better understanding of how robustness metrics work. We show that the wide variety of robustness metrics in the literature can be split into a set of three transformations, which provides a unifying framework for the calculation of robustness. This chapter also provides a

conceptual framework for assessing the impact different robustness metrics have on robustness. 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. This framework and the properties of the case studies are useful in explaining why the robustness and the rankings of decision alternatives obtained using different robustness metrics sometimes disagree. Previously, this had only been observed but never explained. We illustrate this framework on three water and environmental case studies.

2. Chapter 3 explores how the selection of scenarios can affect the robustness of a system. The literature only contained qualitative or anecdotal evidence of the effects of scenario selection. This chapter contributes to the field by providing the first generalisable, quantitative methodology for assessing the impact of different selections of scenarios on the absolute and relative robustness of decision alternatives of interest. Without this method, there is no approach in the literature to understanding if different sets of scenarios have an impact on robustness, and what that impact might be. As an illustration of this generalisable methodology, it was applied to the Lake Problem. Within this case study, several examples were highlighted of how different scenario selections could affect the absolute and relative robustness of decision alternatives in different ways, which demonstrates the utility of the generic methodology.

3. Building upon the knowledge developed in Chapters 2 and 3, Chapter 4 contributes to this field of research by providing a guidance framework and software package. Previously, no guidance has existed for decision-makers on how to determine the most robust decision alternative for their problem, and this chapter provides guidance (including flow charts) that leads users through several situations including those situations where they know or do not know which scenarios and/or robustness metrics to use. The guidance framework also identifies situations where quantitative analyses can be used to determine the influence that the selection of scenarios and/or the choice of robustness metric has on the rankings of decision alternatives, and these analyses can also be implemented by the software package. The guidance and software package

increase the consistency, ease-of-use, and accessibility of robustness analyses for decision-makers.

5.2. Limitations and recommendations for future research

Below is a discussion of the limitations of this research, as well as recommended future research, with the aim of further improving the long-term planning of water environmental systems.

5.2.1. Further development of software package

The RAPID software package developed in Chapter 4 assists decision-makers in identifying the most robust decision alternative for a given problem. However, this software is still in a beta mode, and not yet widely tested on a variety of applications. Further development of this software is recommended to ensure that it meets the needs of practitioners. This software package should also increase in scope to include other useful analyses or integrations with other software packages, as needed by practitioners.

5.2.2. Extension of guidance framework to make recommendations on scenarios

The guidance framework in Chapter 4 (built using knowledge obtained in Chapter 3) cannot recommend one scenario selection approach over another. Rather, it provides quantitative techniques to assess whether two or more scenario selection approaches agree or disagree on the robustness values and rankings. It is recommended that further research is done on scenario selection approaches with the aim of developing guidance for decision-makers to select the best scenarios for a given problem. This would be an interesting avenue of research, particularly given the recent research in this area for scenario generation techniques that focus on areas of interest in the space of possible scenarios.

5.2.3. Increased applicability of guidance framework to additional problem types

5.2.3.1. Increased testing of guidance framework

The guidance framework was built from knowledge in the literature, as well as the knowledge developed through Chapters 2 and 3. The guidance is a compelling conceptual contribution and has been explored through a case study. As highlighted by Kwakkel and van der Pas (2011), conceptual contributions and exploratory modelling are important foundational steps for improving the long-term planning of water and environmental systems, but further work is required for it to become a commonly used best practice. It is recommended that these concepts are further tested, refined, and improved through simulation gaming workshops with students, followed by workshops with decision-makers, case-studies of successful long-term infrastructure plans, and the creation of carefully designed pilot studies to compare these approaches, as recommended by Kwakkel & van der Pas (2011).

5.2.3.2. Extension of framework with additional considerations

It is also recommended that the guidance framework is tested more widely across a variety of problem types to further improve its effectiveness and applicability. As it is tested more widely, it may be realised that the framework does not consider particular types of problems that decision-makers come across. If this happens, it may be appropriate to extend the framework to include these additional considerations. For example, the guidance framework does not consider some robustness metric transformations (such as kurtosis), since it was unclear when these transformations would be useful in a decision-making context. However, if a use for it was found, then there is no reason why this cannot be included in the guidance framework.

5.2.3.3. Further awareness and education of guidance framework

Additionally, the use of this guidance framework across pilot studies in a variety of problem types will increase awareness of the utility of considering deep uncertainty and robustness metrics. At present, the consideration of deep uncertainty is confined to a relatively small number of specialists in a small number of fields. To support the widespread adoption of these techniques, practitioners in water and environmental systems modelling would benefit from wider recognition and additional education for how to achieve the greatest benefits from these techniques.

Appendix A

Supplementary Material (Paper 1): Robustness metrics: How are they calculated, when should they be used and why do they give different results?

A.1 Robustness metric Transformation 1

The performance value transformation (*T₁*) converts $f(x_i, S)$ to $f'(x_i, S)$ where $f'(x_i, S)$ is the relevant information about the performance values, which may be the performance values themselves, the regret, or the satisfaction of a constraint.

Some of the common performance value transformations are given in the table below to show the transformation of $f(x_i, s_j)$ to $f'(x_i, s_j)$ for $j = \{1, 2, \dots, n\}$ where n is the total number of scenarios in S. Note that some of the transformations depend on whether the aim is to maximise or minimise the performance values.

Description	**Equation**
Identity transformation	$f'(x_i, s_j) = f(x_i, s_j)$
Regret from best decision alternative	$f'(x_i, s_j) = \begin{cases} \max\limits_x f(x, s_j) - f(x_i, s_j), & \text{maximisation} \\ f(x_i, s_j) - \min\limits_x f(x, s_j), & \text{minimisation} \end{cases}$
Regret from median	$f'(x_i, s_j) = \begin{cases} q_{50} - f(x_i, s_j), & \text{maximisation} \\ f(x_i, s_j) - q_{50}, & \text{minimisation} \end{cases}$ where q_{50} is the median performance for decision alternative x_i. i.e. $P(f(x_i, S) \leq q_{50}) = \frac{1}{2}$
Satisfaction of constraints	$f'(x_i, s_j) = \begin{cases} \begin{cases} 1 & \text{if } f(x_i, s_j) \geq c \\ 0 & \text{if } f(x_i, s_j) < c \end{cases}, & \text{maximisation} \\ \begin{cases} 1 & \text{if } f(x_i, s_j) \leq c \\ 0 & \text{if } f(x_i, s_j) > c \end{cases}, & \text{minimisation} \end{cases}$ where c is a constraint

A.2 Robustness metric Transformation 2

The scenario subset selection transformation (*T*$_2$) is the process of selecting the values from $f'(x_i, S)$ to be used in the calculation of robustness $R(x_i, S)$. This is done by choosing a subset $S' \subseteq S$ to transform $f'(x_i, S)$ to $f'(x_i, S')$. The table below describes how S' is found for an individual decision alternative (x_i):

Description	**Equation**
Worst-case	$S' = \begin{cases} \left\{\arg\min\limits_{s} f'(x_i, s)\right\}, & \text{maximisation} \\ \left\{\arg\max\limits_{s} f'(x_i, s)\right\}, & \text{minimisation} \end{cases}$
Best-case	$S' = \begin{cases} \left\{\arg\max\limits_{s} f'(x_i, s)\right\}, & \text{maximisation} \\ \left\{\arg\min\limits_{s} f'(x_i, s)\right\}, & \text{minimisation} \end{cases}$
Worst- and best-cases	$S' = \left\{\arg\max\limits_{s} f'(x_i, s), \arg\min\limits_{s} f'(x_i, s)\right\}$
All	$S' = S$
Worst-half	$S' = \begin{cases} \{s \in S: f'(x_i, s) \leq q_{50}\}, & \text{maximisation} \\ \{s \in S: f'(x_i, s) \geq q_{50}\}, & \text{minimisation} \end{cases}$ where q_{50} is the 50th percentile (median) value of $f'(x_i, S)$
Percentile	$S' = \{f'(x_i, s) = q_k\}$ where q_k is the kth percentile value of $f'(x_i, S)$ Note that the scenario s that produces the value of $f'(x_i, s)$ closest to q_k is the scenario that is used.

A.3 Robustness metric Transformation 3

The robustness metric calculation (*T₃*) transforms the set $f'(x_i, S')$ into a single value of robustness, $R(x_i, S)$. Common methods for the robustness metric calculation are given in the table below. Note that if the set contains a single value then the robustness metric calculation will be an identity transformation.

Description	Equation
Identity transformation	$$R(x_i, S) = f'(x_i, S')$$
Mean	$$R(x_i, S) = \frac{1}{n'} \sum_{j=1}^{n'} f'(x_i, s_j)$$ where n' is the number of scenarios in S'
Sum	$$R(x_i, S) = \sum_{j=1}^{n'} f'(x_i, s_j)$$
Weighted mean (two scenarios)	$$R(x_i, S) = \alpha f'(x_i, s_a) + (1-\alpha) f'(x_i, s_b)$$ where s_a and s_b are two scenarios and α is the preference of the decision maker towards using s_a and $0 < \alpha < 1$
Variance-based (i.e. the standard deviation)	$$R(x_i, S) = \sqrt{\frac{1}{n'-1} \sum_{j=1}^{n'} \left(f'(x_i, s_j) - \mu\right)^2}$$ where μ is the mean (see the equation earlier in this table)
Mean-variance	$$R(x_i, S) = \begin{cases} (\mu+1)/(\sigma+1), & \text{maximisation} \\ -(\mu+1)(\sigma+1), & \text{minimisation} \end{cases}$$ where μ is the mean and σ is the standard deviation (given by equations above)
Skew	$$R(x_i, S) = \begin{cases} \frac{\left(f'(x_i, s_{90}) + f'(x_i, s_{10})\right)/2 - f'(x_i, s_{50})}{\left(f'(x_i, s_{90}) - f'(x_i, s_{10})\right)/2}, & \text{maximisation} \\ -\frac{\left(f'(x_i, s_{90}) + f'(x_i, s_{10})\right)/2 - f'(x_i, s_{50})}{\left(f'(x_i, s_{90}) - f'(x_i, s_{10})\right)/2}, & \text{minimisation} \end{cases}$$ where s_{10}, s_{50} and s_{90} are scenarios that represent the 10th, 50th and 90th percentiles for $f'(x_i, S)$

Kurtosis	$R(x_i, S) = \frac{f'(x_i, s_{90}) - f'(x_i, s_{10})}{f'(x_i, s_{75}) - f'(x_i, s_{25})}$ where s_{10}, s_{25}, s_{75} and s_{90} are scenarios that represent the 10th, 25th, 75th and 90th percentiles for $f'(x_i, S)$

A.4 Description of robustness metrics

Maximin

The maximin (minimax) metric was first used by Wald (1950). It is a very risk averse metric that assumes that the scenario that will occur is the scenario under which the performance is lowest.

Maximax

Maximax is the opposite of the maximin metric. It is a metric with a low level of risk aversion that looks for the best possible performance that is possible in a decision alternative.

Hurwicz's optimism-pessimism rule

Hurwicz's optimism-pessimism rule (Hurwicz, 1953) uses a weighted sum of the maximin and maximax metrics that have previously been discussed. Like the previous metrics, the Hurwicz's optimism-pessimism rule uses the distribution of performances for an individual decision alternative (i.e. it does not compare the performances of multiple decision alternatives). It has a parameter α that determines the relative degree of intrinsic risk aversion of the metric where $0 < \alpha < 1$ is the weighting of the maximin (high level of risk aversion) metric. In other words, α may be described as the proportion of high to low risk aversion for the decision-maker. Being composed of both the maximin and maximax metrics brings many of the characteristics of these metrics.

Laplace's principle of insufficient reason

Laplace's principle of insufficient reason (Laplace and Simon, 1951) states that in the absence of information on the relative probabilities of the scenarios, each scenario should be treated as equally likely. This is equivalent to assuming the mean performance across the distributions represents the expected value of robustness. Unlike the previously discussed metrics, Laplace's principle of insufficient reason uses the performance values from every scenario rather than just using one or two performance values.

Minimax regret

Rather than looking at individual decision alternatives, regret metrics including minimax regret (Savage, 1951) look for the regret of choosing a particular option. Specifically, minimax regret calculates the maximum regret that can be expected in any scenario. The regret for a decision alternative x_i in scenario s_j is calculated by comparing the performance $f(x_i, s_j)$ to the best possible performance of any decision alternative in scenario s_j. For decision alternative x_i, the robustness value is the regret from the scenario with the greatest level of regret. In this case, the objective is to minimize the regret. Unlike other metrics which consider an individual decision alternative, the minimax regret metric is sensitive to the distributions of performance of two or more decision alternatives. However, it is only sensitive to the largest difference between the distributions.

90th percentile minimax regret

The 90th percentile minimax regret metric (Herman et al., 2015) is a variant of the minimax metric (Savage, 1951) that was discussed previously. Regret is calculated using the same transformation as the minimax regret metric, and thus this metric also is used to compare two or more decision alternatives rather than only looking at an individual decision alternative. The expected amount of regret for decision alternative x_i is calculated using the 90^{th} percentile of regret rather than the maximum possible regret.

This metric is thus more sensitive to the overall distribution of the performance when compared to the minimax regret metric. However, it is still most sensitive to only a small number of scenarios when compared to a metric such as Laplace's principle of insufficient reason which uses the average of every scenario.

Mean-variance

The mean-variance metric (Kwakkel et al., 2016b) is similar to Laplace's principle of insufficient reason in that it uses the mean to determine the expected value of the distribution of performances for an individual decision alternative. Unlike Laplace's principle of insufficient reason, the mean-variance metric also considers the variability in the distribution of performances by using the standard deviation of performance values. This metric does face several challenges including that the influence of the mean and standard deviation will depend on their relative magnitude and thus the trade-off between mean and standard deviation is unknown (Kwakkel et al., 2016b).

Undesirable deviations

The undesirable deviations metric (Kwakkel et al., 2016b) is a variation on the approach used by Takriti & Ahmed (2004). This metric only considers undesirable deviations (regret) away from the median performance value (which is considered the expected value).

Percentile-based skewness

The percentile-based skewness metric (Voudouris et al., 2014) considers the skewness of the distribution of performance values. This metric gives preference to decision alternatives where the performance values are skewed towards better performance values. It uses the 10th, 50th and 90th percentile values.

Percentile-based peakedness

A variation of Kurtosis was applied by Voudouris et al. (2014) to determine robustness. This metric indicates the "peakedness" of the distribution (Kwakkel et al., 2016b). It uses the 10th, 25th, 75th and 90th percentile performance values respectively for each decision alternative. Unlike the percentile-based skewness metric, this metric does not consider whether the distribution is skewed towards higher or lower performance values. A higher value implies that the performance values are more peaked around the median value.

Starr's Domain Criterion

Unlike previous metrics, Starr's domain criterion (Starr, 1963; Schneller and Sphicas, 1983) compares the distribution of performance values to a threshold value. This metric is most useful when the threshold is selected such that the level of performance above or below the threshold does not matter, but preferably the decision alternative will meet this threshold. For example, a system may have a threshold such that any decision alternative with a performance below the threshold is a fail and any performance above the threshold is a pass.

A.5 Robustness values for the three case studies

Robustness values for each robustness metric for the Southern Adelaide case study (reliability of water supply under 125 scenarios).

Solution #	Robustness									
	Maximin	Maximax	Hurwicz	Laplace	Minimax regret	90th percentile minimax regret	Mean-variance	Undesirable deviations	Percentile-based skewness	Percentile-based peakedness
1	0.000	0.780	0.390	0.341	0.920	0.850	1.099	3.804	0.119	1.513
2	0.000	0.800	0.400	0.413	0.910	0.802	1.147	5.324	0.297	1.565
3	0.000	0.800	0.400	0.425	0.910	0.780	1.157	5.648	0.340	1.590
4	0.000	0.830	0.415	0.436	0.900	0.776	1.160	5.853	0.315	1.580
5	0.000	0.830	0.415	0.460	0.870	0.760	1.188	5.893	0.403	1.674
6	0.000	0.830	0.415	0.451	0.890	0.766	1.171	6.127	0.368	1.630
7	0.000	0.860	0.430	0.483	0.860	0.740	1.200	6.116	0.390	1.726
8	0.000	0.840	0.420	0.452	0.890	0.760	1.174	6.049	0.368	1.630
9	0.000	0.860	0.430	0.494	0.850	0.726	1.208	6.842	0.463	1.773
10	0.000	0.830	0.415	0.436	0.900	0.776	1.160	5.853	0.315	1.580
11	0.000	0.860	0.430	0.486	0.860	0.736	1.202	6.356	0.421	1.726
12	0.000	0.860	0.430	0.483	0.860	0.740	1.200	6.116	0.390	1.726
13	0.000	0.860	0.430	0.494	0.850	0.726	1.208	6.842	0.463	1.773
14	0.000	0.860	0.430	0.494	0.850	0.726	1.208	6.842	0.463	1.773
15	0.000	0.870	0.435	0.514	0.850	0.716	1.223	6.928	0.486	1.846
16	0.000	0.860	0.430	0.494	0.850	0.726	1.207	6.862	0.463	1.773
17	0.000	0.860	0.430	0.494	0.850	0.726	1.208	6.842	0.463	1.773
18	0.000	0.860	0.430	0.494	0.850	0.726	1.208	6.842	0.463	1.773
19	0.000	0.860	0.430	0.494	0.850	0.726	1.207	6.862	0.463	1.773
20	0.000	0.860	0.430	0.494	0.850	0.726	1.207	6.862	0.463	1.773
21	0.000	0.790	0.395	0.365	0.920	0.830	1.115	4.195	0.155	1.480
22	0.000	0.800	0.400	0.413	0.910	0.802	1.147	5.324	0.297	1.565
23	0.000	0.830	0.415	0.436	0.900	0.776	1.160	5.853	0.315	1.580
24	0.000	0.860	0.430	0.494	0.850	0.726	1.207	6.862	0.463	1.773
25	0.000	0.830	0.415	0.436	0.900	0.776	1.160	5.853	0.315	1.580
26	0.000	0.860	0.430	0.494	0.850	0.726	1.208	6.842	0.463	1.773
27	0.000	0.860	0.430	0.494	0.850	0.726	1.207	6.862	0.463	1.773
28	0.000	0.860	0.430	0.494	0.850	0.726	1.208	6.842	0.463	1.773
29	0.000	0.860	0.430	0.494	0.850	0.726	1.207	6.862	0.463	1.773
30	0.000	0.860	0.430	0.494	0.850	0.726	1.208	6.842	0.463	1.773
31	0.000	0.860	0.430	0.494	0.850	0.726	1.207	6.862	0.463	1.773
32	0.000	0.860	0.430	0.494	0.850	0.726	1.207	6.862	0.463	1.773
33	0.000	0.860	0.430	0.494	0.850	0.726	1.207	6.862	0.463	1.773
34	0.000	0.900	0.450	0.622	0.770	0.580	1.321	6.479	0.529	2.214
35	0.000	0.900	0.450	0.630	0.750	0.576	1.331	6.091	0.508	2.259

36	0.000	0.910	0.455	0.644	0.710	0.568	1.346	5.829	0.519	2.156
37	0.000	0.910	0.455	0.644	0.710	0.568	1.346	5.829	0.519	2.156
38	0.000	0.900	0.450	0.617	0.770	0.586	1.318	6.253	0.515	2.289
39	0.000	0.900	0.450	0.630	0.750	0.576	1.331	6.091	0.508	2.259
40	0.000	0.900	0.450	0.637	0.720	0.568	1.339	5.787	0.502	2.150
41	0.000	0.900	0.450	0.630	0.750	0.576	1.331	6.091	0.508	2.259
42	0.000	0.900	0.450	0.645	0.710	0.560	1.344	6.228	0.547	2.143
43	0.000	0.900	0.450	0.608	0.790	0.596	1.310	6.574	0.556	2.333
44	0.000	0.900	0.450	0.630	0.750	0.576	1.331	6.091	0.508	2.259
45	0.000	0.900	0.450	0.638	0.740	0.564	1.336	6.237	0.535	2.076
46	0.000	0.910	0.455	0.656	0.700	0.556	1.354	6.175	0.529	2.121
47	0.000	0.930	0.465	0.677	0.710	0.548	1.367	6.401	0.552	2.148
48	0.000	0.950	0.475	0.689	0.690	0.548	1.380	6.544	0.562	2.304
49	0.000	0.930	0.465	0.677	0.710	0.548	1.367	6.401	0.552	2.148
50	0.000	0.950	0.475	0.689	0.690	0.548	1.380	6.544	0.562	2.304
51	0.000	0.950	0.475	0.689	0.690	0.548	1.380	6.544	0.562	2.304
52	0.000	0.950	0.475	0.695	0.690	0.536	1.387	6.264	0.532	2.317
53	0.000	0.950	0.475	0.689	0.690	0.548	1.380	6.544	0.562	2.304
54	0.000	0.950	0.475	0.712	0.690	0.512	1.406	6.192	0.579	2.270
55	0.000	0.950	0.475	0.695	0.690	0.536	1.387	6.264	0.532	2.317
56	0.000	0.950	0.475	0.695	0.690	0.536	1.387	6.264	0.532	2.317
57	0.430	0.990	0.710	0.928	0.110	0.030	1.775	0.948	0.524	2.520
58	0.500	0.990	0.745	0.932	0.040	0.020	1.795	0.723	0.524	2.520
59	0.500	0.990	0.745	0.937	0.020	0.000	1.802	0.745	0.552	2.900
60	0.490	0.990	0.740	0.938	0.010	0.000	1.806	0.700	0.552	2.900
61	0.500	0.990	0.745	0.937	0.020	0.000	1.802	0.745	0.552	2.900
62	0.490	0.990	0.740	0.938	0.010	0.000	1.806	0.700	0.552	2.900
63	0.500	0.990	0.745	0.931	0.040	0.020	1.794	0.729	0.524	2.520
64	0.500	0.990	0.745	0.931	0.040	0.020	1.794	0.729	0.524	2.520
65	0.500	0.990	0.745	0.937	0.020	0.000	1.802	0.745	0.552	2.900
66	0.430	0.990	0.710	0.928	0.110	0.030	1.775	0.948	0.524	2.520
67	0.500	0.990	0.745	0.932	0.040	0.020	1.795	0.723	0.524	2.520
68	0.500	0.990	0.745	0.937	0.020	0.000	1.802	0.745	0.552	2.900
69	0.500	0.990	0.745	0.931	0.040	0.020	1.794	0.729	0.524	2.520
70	0.430	0.990	0.710	0.930	0.070	0.020	1.784	0.843	0.524	2.520
71	0.500	0.990	0.745	0.931	0.040	0.020	1.794	0.729	0.524	2.520
72	0.500	0.990	0.745	0.931	0.040	0.020	1.794	0.729	0.524	2.520

Robustness values for each robustness metric for the Lake Como case study (reliability against flooding under 28 scenarios).

Solution #	Robustness									
	Maximin	Maximax	Hurwicz	Laplace	Minimax regret	90th percentile minimax regret	Mean-variance	Undesirable deviations	Percentile-based skewness	Percentile-based peakedness
1	0.946	1.000	0.973	0.987	0.034	0.015	1.964	0.004	0.382	2.027
2	0.941	1.000	0.970	0.985	0.039	0.019	1.959	0.004	0.176	1.761
3	0.961	1.000	0.981	0.991	0.019	0.010	1.973	0.002	0.323	1.691
4	0.969	1.000	0.984	0.993	0.015	0.007	1.977	0.002	0.425	1.763
5	0.969	1.000	0.985	0.993	0.013	0.007	1.978	0.001	0.462	1.731
6	0.962	1.000	0.981	0.991	0.018	0.009	1.973	0.002	0.273	1.619
7	0.927	0.999	0.963	0.976	0.058	0.033	1.943	0.006	0.214	1.533
8	0.898	0.996	0.947	0.961	0.096	0.058	1.916	0.012	0.245	1.851
9	0.933	0.999	0.966	0.979	0.049	0.029	1.949	0.005	0.228	1.475
10	0.948	1.000	0.974	0.987	0.032	0.015	1.965	0.003	0.240	1.755
11	0.980	1.000	0.990	0.996	0.000	0.000	1.985	0.001	0.538	1.733
12	0.931	1.000	0.965	0.981	0.049	0.030	1.949	0.007	0.408	1.918
13	0.928	0.999	0.963	0.977	0.054	0.030	1.946	0.007	0.339	1.507
14	0.918	1.000	0.959	0.979	0.061	0.033	1.943	0.008	0.392	2.416
15	0.937	1.000	0.968	0.983	0.043	0.022	1.956	0.004	0.195	1.831
16	0.706	0.973	0.840	0.860	0.288	0.209	1.747	0.064	0.019	1.659
17	0.964	1.000	0.982	0.991	0.017	0.009	1.974	0.002	0.385	1.763
18	0.933	0.999	0.966	0.979	0.049	0.029	1.949	0.006	0.264	1.537
19	0.949	1.000	0.975	0.987	0.031	0.015	1.965	0.003	0.240	1.755

Robustness values for each robustness metric for the Lake Como case study (reliability of irrigation supply under 28 scenarios).

Solution #	Robustness									
	Maximin	Maximax	Hurwicz	Laplace	Minimax regret	90th percentile minimax regret	Mean-variance	Undesirable deviations	Percentile-based skewness	Percentile-based peakedness
1	0.528	0.934	0.731	0.770	0.036	0.028	1.634	0.117	0.066	1.971
2	0.534	0.934	0.734	0.779	0.019	0.015	1.643	0.125	0.117	1.976
3	0.526	0.920	0.723	0.763	0.047	0.036	1.630	0.112	0.069	1.932
4	0.524	0.910	0.717	0.755	0.058	0.046	1.624	0.104	0.049	1.947
5	0.523	0.911	0.717	0.755	0.057	0.046	1.625	0.105	0.051	1.949
6	0.529	0.920	0.724	0.765	0.042	0.034	1.632	0.108	0.052	1.982
7	0.542	0.929	0.735	0.785	0.014	0.008	1.651	0.124	0.131	2.195
8	0.540	0.934	0.737	0.786	0.010	0.006	1.651	0.131	0.157	2.032
9	0.542	0.934	0.738	0.785	0.011	0.007	1.649	0.129	0.132	2.143
10	0.535	0.930	0.733	0.778	0.016	0.013	1.643	0.125	0.120	2.091
11	0.499	0.856	0.677	0.699	0.131	0.110	1.580	0.081	-0.031	1.954
12	0.532	0.943	0.737	0.779	0.026	0.014	1.640	0.119	0.046	2.003
13	0.543	0.932	0.737	0.784	0.012	0.007	1.649	0.127	0.129	2.143
14	0.532	0.941	0.737	0.775	0.031	0.021	1.637	0.115	0.036	1.952
15	0.540	0.929	0.735	0.782	0.013	0.010	1.648	0.127	0.144	2.163
16	0.536	0.921	0.729	0.776	0.040	0.026	1.642	0.109	0.017	1.720
17	0.526	0.922	0.724	0.763	0.046	0.035	1.629	0.111	0.054	1.906
18	0.541	0.933	0.737	0.783	0.013	0.009	1.648	0.128	0.124	2.126
19	0.536	0.931	0.733	0.778	0.015	0.013	1.643	0.125	0.115	2.170

Robustness values for each robustness metric for the Waas case study (reduction in flood impacts under 3000 scenarios).

Solution #	Robustness									
	Maximin	Maximax	Hurwicz	Laplace	Minimax regret	90th percentile minimax regret	Mean-variance	Undesirable deviations	Percentile-based skewness	Percentile-based peakedness
1	-7.67E+03	-6.14E+02	-4.14E+03	-3.19E+03	1.26E+03	1.40E+01	-7.85E+06	6.77E+08	-0.368	1.448
2	-7.67E+03	-3.75E+02	-4.02E+03	-3.03E+03	0.00E+00	0.00E+00	-7.93E+06	5.78E+08	-0.306	1.398
3	-9.50E+03	-2.55E+03	-6.02E+03	-5.19E+03	1.22E+04	7.04E+03	-1.07E+07	7.45E+09	-0.525	1.845
4	-9.50E+03	-2.55E+03	-6.02E+03	-5.19E+03	1.22E+04	7.04E+03	-1.07E+07	7.45E+09	-0.525	1.845
5	-4.35E+04	-1.46E+04	-2.90E+04	-2.91E+04	5.97E+04	3.18E+04	-2.51E+08	3.15E+10	0.195	1.292
6	-3.66E+04	-1.03E+04	-2.34E+04	-2.35E+04	5.45E+04	3.34E+04	-1.58E+08	7.73E+10	0.062	1.592
7	-1.61E+04	-8.25E+03	-1.22E+04	-1.19E+04	1.97E+04	1.35E+04	-2.23E+07	4.58E+09	0.010	1.497
8	-7.24E+04	-2.47E+04	-4.86E+04	-4.76E+04	8.92E+04	5.90E+04	-6.53E+08	9.10E+10	0.237	1.449
9	-8.95E+03	-2.55E+03	-5.75E+03	-4.97E+03	9.48E+03	6.06E+03	-9.39E+06	5.21E+09	-0.565	1.949
10	-4.77E+04	-1.65E+04	-3.21E+04	-3.22E+04	6.60E+04	4.24E+04	-2.82E+08	7.64E+10	0.238	1.499
11	-5.43E+04	-1.89E+04	-3.66E+04	-3.71E+04	7.13E+04	4.43E+04	-3.92E+08	5.43E+10	0.282	1.303

Robustness values for each robustness metric for the Waas case study (reduction in casualties under 3000 scenarios).

Solution #	Robustness									
	Maximin	Maximax	Hurwicz	Laplace	Minimax regret	90th percentile minimax regret	Mean-variance	Undesirable deviations	Percentile-based skewness	Percentile-based peakedness
1	-1.40E+02	-2.05E+01	-8.03E+01	-7.48E+01	3.63E+02	5.36E+01	-3.30E+03	7.16E+06	-0.334	1.318
2	-8.82E+01	-2.05E+01	-5.43E+01	-5.76E+01	2.90E+02	5.34E+01	-1.44E+03	4.89E+06	0.105	1.558
3	-2.38E+02	-1.70E+01	-1.28E+02	-1.18E+02	7.61E+02	2.45E+02	-1.07E+04	2.59E+07	0.193	1.126
4	-2.38E+02	-1.70E+01	-1.28E+02	-1.18E+02	7.61E+02	2.45E+02	-1.07E+04	2.59E+07	0.193	1.126
5	-4.32E+02	-1.05E+02	-2.68E+02	-2.64E+02	7.44E+02	3.55E+02	-2.78E+04	2.21E+07	-0.036	1.519
6	-7.80E+02	-1.59E+02	-4.70E+02	-4.54E+02	1.49E+03	8.94E+02	-8.45E+04	1.81E+08	-0.220	1.534
7	-1.38E+03	-3.88E+02	-8.82E+02	-9.22E+02	1.89E+03	1.23E+03	-2.93E+05	5.28E+07	0.348	1.253
8	-9.37E+02	-2.46E+02	-5.91E+02	-5.75E+02	1.23E+03	7.11E+02	-1.11E+05	1.53E+07	0.305	1.239
9	-1.54E+02	-1.70E+01	-8.57E+01	-8.06E+01	4.42E+02	2.07E+02	-4.41E+03	2.04E+07	0.056	1.215
10	-8.70E+02	-2.02E+02	-5.36E+02	-5.30E+02	1.79E+03	1.07E+03	-9.83E+04	1.98E+08	0.110	1.531
11	-9.38E+02	-2.36E+02	-5.87E+02	-5.93E+02	1.24E+03	8.01E+02	-1.31E+05	3.21E+07	0.202	1.276

Appendix B

Supplementary Material (Paper 2):
Impact of scenario selection on robustness

B.1 Examples in literature of how scenarios were selected

Reference	Case Study	Number of Scenarios	Number of Axes/ Dimensions	Method to generate final scenarios, *S*
Herman & Giuliani (2018)	Folsom Reservoir, Northern California, USA	97	1	CMIP5 ensemble → downscaling → variable infiltration capacity (VIC) model → streamflow projections.
Giuliani & Castelletti (2016), McPhail et al. (2018)	Lake Como, Italy	28	2	3 RCPs → particular RCPs applied to particular GCMs (8 GCMs) (17 combinations) → particular combinations applied to particular RCMs (5 RCMs) (28 total combinations) → statistically downscaled using quantile mapping for the Lake Como site → rainfall and temperature data extracted → HBV hydrological model ➔ inflow projections
Haasnoot et al. (2013)	Lower Rhine Delta, the Netherlands	4	2	2 climate change scenarios (affecting temperature, seasonal precipitation, sea level, salt intrusion); 2 socio-economic scenarios (affecting population, land use, water demand, economic growth). These form the 2 axes to create a set of 4 scenarios.
McPhail et al. (2018)	Southern Adelaide Water Supply System, Australia	125	1	4 RCPs → particular RCPs applied to particular GCMs (40 GCMs) (128 combinations) → statistically downscaled for the Southern Adelaide region → rainfall data extracted → perturbed historical rainfall (end scenarios).
Haasnoot et al. (2012)	Waas River, the Netherlands (hypothetical)	3	2	3 climate scenarios (Royal Dutch Meteorological Institute (KNMI)) (includes no climate change as a scenario) → affects temperature and seasonal precipitation
Giuliani et al. (2016)	Red River, Vietnam	5	??	1 SRES scenario ➔ 5 parameterization of the same GCM (perturbed physics ensemble) ➔ 1 RCM ➔ statistically downscaled using quantile mapping for the Red River basin ➔ HBV hydrological model ➔ inflow projections
Anghileri et al. (2018)	Mattmark, Switzerland	100	2	1 emission scenario (A1B) simulated by ECHAM5r3, RegCM3, REMO models, ➔ 100 realizations generated via stochastic downscaling; 1 electricity price
Hall et al. (2012)	Dynamic Integrated model of Climate and Economy (DICE)	2,662	4	4 uncertain parameters: (1) climate sensitivity, (2) initial growth rate of carbon intensity, (3) economic damages associated with collapse of the North Atlantic Meridional Overturning Circulation (MOC), (4) binary parameter to indicate whether MOC will actually collapse if critical emissions thresholds are reached. First three

				parameters split into 11 values and final parameter has 2 values. All combinations were used.
Kwakkel et al. (2016)	EU Energy Model	500	46	46 parametric uncertainties identified and LHS is used to select 500 combinations of values. Uncertainties can be classed as: (1) economic lifetime of technologies; (2) learning curve for technology; (3) economic growth; (4) electrification rate of economy; (5) physical limits of technology penetration; (6) preferences of investors; (7) battery storage; (8) timing of a ban on nuclear energy; (9) price demand elasticity.
Weaver et al. (2013)	Inland Empire Utilities Agency (IEUA) water supply, Southern California, USA	450	6	6 key uncertainties identified: (1) 450 climate scenarios (monthly temperature and precipitation changes), (2) water demand (increases in efficiency), (3) declines in imported water supply, (4) change in groundwater infiltration, (5) delay in achievement of groundwater replenishment goals, (6) changes in costs of imported supply. Sampled using unknown approach.
Culley et al. (2016)	Lake Como, Italy	861	2	Range of 21 changes in temperature and 41 changes in precipitation (including present-day conditions) were used to form a grid of 861 changes in climate. These changes were used to perturb historical temperature and precipitation records.
Hadka et al. (2015)	Lake Model (hypothetical)	1000	5	The 5 uncertain inputs are parameters associated with pollution inputs, pollution removal and an economic discount rate. These are sampled from using LHS.
Singh et al. (2015)		9	2	2 parameters associated with log-normal pollution inputs are used to create a 3x3 grid of log-normal pollution inputs (i.e. 3 different means and 3 different variances).
Kwakkel (2017)		150	5	Same 5 uncertainties as Hadka et al. (2015). Sampled using LHS.
Quinn et al. (2017)		1000	6	Same 5 uncertainties as Hadka et al. (2015) plus uncertainty in the initial concentration of pollution in the lake. Sampled using LHS.
Kasprzyk et al. (2013), Watson and Kasprzyk (2017)	Lower Rio Grande Valley (LRGV) Water Resources System, Texas, USA	10,000	3	3 model inputs: initial rights for water supply for city, demand growth rate, and initial reservoir level. Sampled using Latin Hypercube Sampling.
Herman et al. (2015, 2014)	North Carolina Research Triangle, USA	10,000	13	13 uncertain factors were determined to form axes for 13-dimensional LHS (10,000 scenarios). The 13 dimensions can be categorized as climate, demand, capacity, or costs changes.

Trindade et al. (2017)		10,000	13	Same as Herman et al. (2015, 2014), but also includes a *posteriori* scenario discovery step to narrow the number of scenarios to those that are most of interest.
Kwakkel et al. (2015)	Waas River, the Netherlands (hypothetical)	150	5	5 primary uncertainties determined and used to form axes for Latin Hypercube Sampling (LHS) (150 scenarios used): (1) 3 climate change scenarios; (2) 7 land use scenarios; (3) bandwidth of ±10% fragility of the dikes (10% chosen for illustrative purposes); (4) ±10% in parametric uncertainty for flood damage functions (chosen for illustrative purposes); (5) effect of policy actions on upstream collaboration.
McPhail et al. (2018)		3,000	5	Same as Kwakkel et al. (2015) above but with a larger sample size (3,000) during the LHS.
Quinn et al., (2018)	Red River, Vietnam	1,000	11	6 parameters related to inflow hydrograph (i.e., log-space mean, log-space standard deviation, log-space amplitudes of annual and semiannual monsoonal cycle, log-space shifts of annual and semiannual cycle), 1 parameter related to evaporation, 4 parameters related to water demands
Beh et al. (2015a, 2015b) (using method recommended by Paton et al. (2013))	Southern Adelaide Water Supply System, Australia	7	3	5 SRES scenarios (IPCC, 2000) → each applied to 7 GCMs (35 combinations) → 7 combinations handpicked → rainfall and temperature data extracted → perturbed historical rainfall and evaporation data → ranked worst to best and paired up with 7 population scenarios (also worst to best) → end scenarios were a rainfall, evaporation and population timeseries.
Roach et al. (2016)	Sussex North Water Resource Zone (SNWRZ), UK	288	2	11 flow projections from UK Climate Impact Programme at upstream site → changes in upstream site mapped onto downstream site of interest onto different 30 year historical timeseries with different seasonal flow factors (produces 72 scenarios). Demand scenarios from Southern Water's WRMP (Southern Water, 2009) for up to 2035 (4 scenarios) → extrapolated to 2060 using 2030-2035 data → monthly demand factors applied to create daily time steps for a 50-year period (4 scenarios). 72 supply and 4 demand scenarios combined to form 288 scenarios.

Matrosov et al. (2013)	Thames Water Resource System, UK	15,554	3	101 hydrological scenarios (including the historical hydrological observations) derived from previous climate change studies; 14 uniformly spaced energy cost estimates based on the assumption energy costs will increase; 3 demand scenarios from previous studies → a normal distribution for each scenario → 9 deciles from the 3 distribution and the 0.01 and 0.99 percentiles → 11 demand scenarios; All combinations of 101 hydrological, 14 energy cost, and 11 demand scenarios used.
Matrosov et al. (2013)	Thames Water Resource System, UK	3850	3	Same as Matrosov et al. (2013) but with 25 hydrological scenarios. All combinations of 25 hydrological, 14 energy cost, and 11 demand scenarios used.
Huskova et al. (2016)		88	4	SRES A1B (medium emissions) scenario → HadRM3-PPE Regional Climate Model → 11 hydrological timeseries; A future demand projection from a previous study is used, and a second demand scenario is produced using a 10% increase on this; Energy price scenarios (13p/kWh and 35p/kWh) are very different and are taken from 2 independent previous studies; 1 sustainability scenario is that current water extractions are continued and the other sustainability scenario is that extractions decrease. All combinations of 11 hydrological, 2 water demand, 2 sustainability, and 2 energy scenarios used.
Wada et al. (2019)	Indus River Basin	NA	4	One "BAU" scenario, created by assuming SSP2 ("Middle of the road"). Through a participatory process, it was determined how things would look different from the BAU scenario if the economy, society, or environmental domains were prioritized, creating 3 new scenarios. This paper highlighted the ability to create new scenarios relative to the BAU (e.g. by prioritizing particular Sustainable Development Goals), by using a participatory process, and then quantifying each scenario using large integrated models.

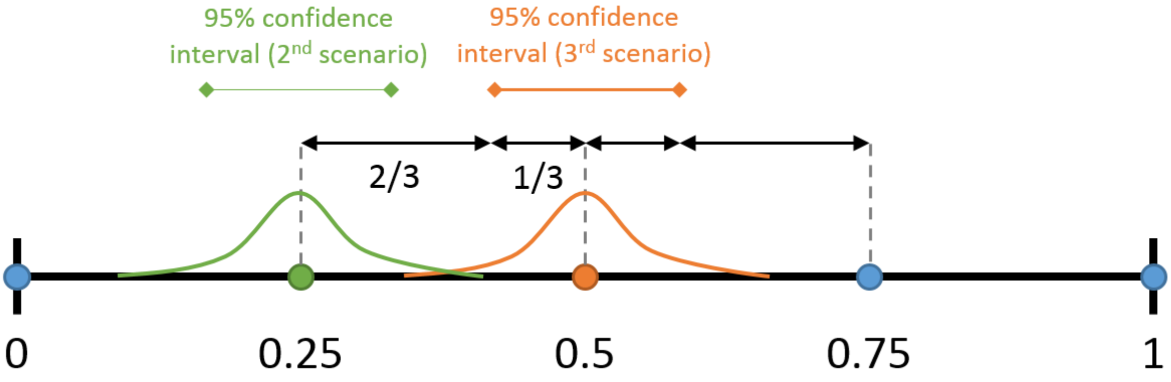
B.2 Generation of targeted spread scenarios

Let n be the number of scenarios and d be the number of dimensions (number of variables) for each scenario (the Lake Problem has 5 variables). Let the full set of scenarios be $S = \{s_1, s_2, \dots, s_n\}$.

The mean value of s_i for variable $j = \{1,2,\dots,d\}$ is given by

$$\mu_{i,j} = \begin{cases} \dfrac{i-1}{n-1}, & \text{a higher value for variable } j \text{ implies better performance} \\ 1 - \dfrac{i-1}{n-1}, & \text{a lower value for variable } j \text{ implies better performance} \end{cases}$$

In the real example where $n = 3$, the "medium" value is not necessarily halfway between 0 and 1. It could depend on the chain of processes that led to the creation of the scenarios. Therefore, some Gaussian noise is added to these mean values. To preserve the monotonicity, the standard deviation is set such that the 95% bounds of the distribution fall within 1/3 of the distance to either side of the point. See the diagram below for the 1D case with 5 scenarios:


95% confidence interval (2nd scenario)
95% confidence interval (3rd scenario)
2/3
1/3
0
0.25
0.5
0.75
1

This standard deviation, σ, is the same for all scenarios in all dimensions (since we are treating all dimensions as being between 0 and 1 at this point):

$$\sigma = \left(\frac{1}{3} \times \frac{1}{n-1}\right) / 1.96$$

where 1.96 is used to achieve the 95% confidence interval.

Thus each scenario, s_i, (with d dimensions) can be sampled a series of normal distributions, $\mathcal{N}$ such that:

$$s_i = \{\mathcal{N}(\mu_{i,1}, \sigma), \mathcal{N}(\mu_{i,2}, \sigma), \dots, \mathcal{N}(\mu_{i,d}, \sigma)\}$$

B.3 Generation of diverse scenarios

Let n be the number of scenarios and d be the number of dimensions (number of variables) for each scenario (the Lake Problem has 5 variables). Let the full set of scenarios be $S = \{s_1, s_2, \dots, s_n\}$.

For simplicity we assume 4 clusters of scenarios. Each cluster, c, is given a random weighting, where this weighting is the likelihood of any given scenario being placed in that cluster.

The mean values in for each cluster, c, in each of the d dimensions is calculated in the same way as the monotonic scenarios above. For $j = \{1,2, \dots, d\}$ and $c = \{1, 2, 3, 4\}$ (4 clusters), the mean is given by:

$$\mu_{c,j} = \begin{cases} \dfrac{c-1}{n-1}, & \text{a higher value for variable } j \text{ implies better performance} \\ 1 - \dfrac{c-1}{n-1}, & \text{a lower value for variable } j \text{ implies better performance} \end{cases}$$

Also similar to the hybrid scenarios above, the standard deviation was calculated such that a 95% confidence interval around each cluster mean was restricted. To reflect real situations, we wanted the spread of scenarios around a mean to be different for each cluster. I.e. we want some clusters to have a high spread of scenarios, and others to have a low spread. So we allow the standard deviation for each cluster and each dimension to be different. We let the standard deviation, $\sigma_{c,j}$, of cluster $c = \{1, 2, 3, 4\}$ in dimension $j = \{1,2, \dots, d\}$ have a 95% confidence interval between 0.2 and 0.35 of the gap between the clusters. This is similar to the hybrid scenarios but rather than fix the standard deviation at 1/3 of the distance between clusters (see diagram above), we allow it to vary within 0.2 and 0.35 of the distance between the clusters.

$$\sigma_{c,j} = \mathcal{N}(\mu, \sigma)$$

where μ is in the middle of 0.2 and 0.35 of the gap between the cluster centres (i.e. 0.275 of the gap), and σ is the standard deviation such that the 95% bounds fit between 0.2 and 0.35 of the gap.

Then for each scenario, s_i, (with d dimensions), the scenario is selected to go into cluster c according to the random weightings for each cluster.

$$s_i = \{\mathcal{N}(\mu_{c,1}, \sigma_{c,1}), \mathcal{N}(\mu_{c,2}, \sigma_{c,2}), \dots, \mathcal{N}(\mu_{c,d}, \sigma_{c,d})\}$$