Understanding the Problem Structure of Optimisation Problems in Water Resources

Siwei Zhu
B.Eng., M.Eng

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

Copyright© 2021

This page is intentionally blank

Abstract

Optimisation algorithms are widely used in water resources to identify the optimal solutions for problems with multiple possible solutions. Many studies in this field focus on the development and application of advanced optimisation algorithms, making significant contributions in improving optimisation performance. On the other hand, the performance of optimisation algorithms is also related to the features of the problems being solved, therefore, selecting appropriate algorithms for corresponding problems is also a key to the success of optimisation. Although a number of metrics have been developed to assess these features, they have not been applied to problems in the water resources field. The primary reason for this is that the computational cost associated with the calculation of many of these metrics increases significantly with problem size, making them unsuitable for problems in water resources. Consequently, there is a lack of knowledge about the features of problems in the water resources field. This PhD thesis aims to understand the features of problems in water resources, and the process can be split into two stages. The first stage is to identify metrics that can be applied within an affordable computational cost. This is addressed in the first content chapter (Paper 1). The second stage is to apply metrics identified in the first stage to understand the features of problems in the water resources field, including the calibration of artificial neural network models (Paper 2) and conceptual rainfall runoff models (Paper 3). This includes the understanding of optimisation difficulty of these problems

according to their features, and how their features change through the change of their problem structure and the types of problems to which they are applied.

In the first paper, the computational cost of fitness landscape metrics (explanatory landscape analysis (ELA) metrics) used in computer science is tested and metrics that are suitable for application to water resources problems are identified. Each metric used to understand the features of problems requires a given number of samples, which usually increases with an increase in problem size (dimensionality). Consequently, metrics which require a big increase in sample size through the increase of problem size are not suitable for real-world water resources problems. To identify ELA metrics that have low dependence on problem size, 110 metrics in total are tested on a range of benchmark functions and a number of environmental modelling problems, and 28 are identified to be able to be applied to complex problems without significant increase in computational cost. This finding provides us a new approach to better understand the problem structure of optimisation problems in water resources and has the potential to provide guidance in optimisation algorithm selection for problems in the water resources field.

In the second paper, metrics identified to have low dependence on problem size in the first paper are applied to Artificial Neural Network (ANN) model calibration problems. ANN models for different environmental problems with different number of inputs and hidden nodes are used in the test. The environmental problems considered include Kentucky River Catchment

Rainfall-Runoff Data (USA), Murray River Salinity Data (Australia), Myponga Water Distribution System Chlorine Data (Australia), and South Australian Surface Water Turbidity Data (Australia). It is demonstrated that ELA metrics can be used successfully to characterize the features of the error surfaces of ANN models, thereby helping to explain the reasons for an increase or decrease in calibration difficulty, and in doing so, shedding new light on findings in existing literature. Results show that the error surfaces of ANNs with relatively simple structures have a more well-defined overall shape and have fewer local optima, while the error surfaces of ANNs with more complex structures are flatter and have many distributed, deep local optima. Consequently, ANNs with simpler structures can be calibrated successfully using gradient-based methods, such as the back-propagation algorithm, whereas ANNs with more complex structures are best calibrated using a hybrid approach combining metaheuristics, such as genetic algorithms, with gradient-based methods.

In the third paper, the ELA metrics identified to have low dependence on problem size in the first paper are applied to Conceptual Rainfall Runoff (CRR) model calibration problems. Different CRRs with different model types, error functions, catchment conditions and data lengths are tested to identify how they affect the features of problem structure, which are related to their model calibration and parameter identification difficulty. It is suggested that ELA metrics can be used to quantify key features of the error surfaces of CRR models, including their roughness and flatness, as well as their degree of optima dispersion. This enables key error surface features to be compared for CRR

models with different combinations of attributes (e.g. model structure, catchment climate conditions, error metrics and calibration data lengths and composition) in a consistent, efficient and easily communicable fashion. Results from the application of these metrics to the error surfaces of 420 CRR models with different combinations of the above attributes indicate that model structure differences result in the differences in surface roughness and relative optima dispersion. Additionally, increasing catchment wetness increases the relative roughness of error surfaces, it also decreases optima dispersion. This suggests that model structure and catchment climate conditions can be key issues in affecting the calibration difficulty, efficiency and parameter uniqueness. The experiments conducted in this study also encourage further tests on further CRR models and catchments to identify general patterns between calibration performance, model structure and catchment characteristics.

This page is intentionally blank.

Statement of Originality

I 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.

Signed: …… ……Date: 30/11/2021

This page is intentionally blank.

Acknowledgements

I would like to express my greatest gratitude to my supervisor Professor Holger Maier and Dr Aaron Zecchin. Thanks to Holger for his creative mind and thought that lead the way of my research through the whole of my PhD. Through the project with Holger, I was impressed by his solid, comprehensive and critical thought. Besides, Holger's strong sense of responsibility and high level of professionalism were also qualities I am and will keep learning in my rest of life. Thanks to Aaron for his continuous supports and guidance especially in technique and analysis. His talents and rich knowledge in mathematics and computer science are no doubt one of the key reasons for the successful completion of this PhD program. I appreciate for my supervisors for their substantial comments and feedback for the whole process of my PhD, such as literature review, experiment design, data analysis and manuscript drafting. I appreciate for the continuous weekly meeting with them, which even kept through Zoom meeting when I was stuck in China due to the outbreak of COVID-19.

I would like to thank the University of Adelaide that supported me with scholarship, especially for the great help during the hard time of COVID-19 outbreak. I also would like to thank my supervisor Professor Holger Maier for the extra short-term scholarship he provided, which helped me finalised the

PhD program.

I would like to thank my parents for their love and supports. They are always my powerful and secure backing in my life, and nothing I can achieve and make without their assists.

Table of Contents

Chapter 1 Introduction 1

1 Background 1

1.1 Metrics to understand the features of fitness landscapes of optimisation problems 3

1.2 Applications of ELA metrics to water resource optimisation problems 4

1.3 Current knowledge about fitness landscapes of environmental optimisation problems 5

1.4 Limitations of understanding the problem structure for optimisation problems in water resources 8

2 Research aims 10

3 Organisation of Thesis 15

Chapter 2 Identification of Metrics Suitable for Determining the Features of Real-World Optimisation Problems 16

1 Introduction 20

2 Methodology 24

2.1 Overview 24

2.2 Sampling of fitness landscapes 31

2.3 Fitness landscape metrics 34

2.4 Calculation of degree of dependence of metrics on dimensionality and sample size 38

2.5 Categorisation of fitness landscape metrics 41

3 Results and Discussion 43

3.1 Categorisation of fitness landscape metrics 43

3.2 Validation of categorisation of fitness landscape metrics 49

3.3 Interpretation of metrics with low dependence on sample size and dimensionality 52

4 Summary and Conclusions 55

Chapter 3 Impact of Model Structure on the Difficulty of Calibrating Artificial Neural Network Models 59

1 Introduction 63

2 Methodology 67

2.1 Overview 67

2.2 Case Studies 70

2.3 ANN Models and Data Transformations 71

2.4 Fitness Landscape Metrics 72

3 Results and Discussion 76

3.1 Mean pairwise convexity deviation 76

3.2 Maximum entropy of information content *(H_{max})* 78

3.3 Epsilon of information content 81

3.4 Median basin centroidal distance 83

3.5 Median search function evaluations 85

3.6 Discussion 87

4 Summary and Conclusions 91

Chapter 4 Improved Understanding of Calibration Efficiency, Difficulty and Parameter Uniqueness of Conceptual Rainfall Runoff Models using Fitness Landscape Metrics 94

1 Introduction 99

2 Methodology 103

2.1 Overview 103

2.2 CRR models 107

2.3 Catchments 110

2.4 Error metrics 112

2.5 Fitness landscape metrics 114

3 Results and Discussion 118

3.1 Overview 118

3.2 Impact of model structure and complexity 122

3.3 Impact of catchment climate condition 130

4 Summary and Conclusions 134

Chapter 5 Conclusions 137

1 Research Contributions 137

2 Scope of Future Work 141

References 144

Appendices of Chapter 2 (Paper 1) 177

Appendix A: Detailed Outline of Methodology 177

Appendix B: Details of Benchmark Functions 178

Appendix C: Details of ELA Metrics 179

Appendix D: Slopes and Coefficients of Determination (r^2) of Benchmark Test for All Tested Metrics 195

Appendices of Chapter 4 (Paper 3) 199

Appendix A – Details of CRR Models 199

Appendix B – Results of Clustering of ELA Metrics for the 420 Error Surfaces Considered 203

Appendix C – Influence of model structure / complexity on error surface features (Raw Results) 218

Supplementary Materials of Chapter 2 (Paper 1) 219

S1: Example R-code for calculating ELA metric results 219

S2: Example R-Code for dependency analysis 224

List of Tables

Table 2.1 Descriptions of High-Level Fitness Landscape Features.....31

Table 2.2 Case Study Information 34

Table 2.3 Relationship between ELA Metric Classes (in the columns) and Fitness Landscape Features (in the rows) 38

Table 2.4 Metrics in Each Class in Different Clusters.....49

Table 2.5 Features Represented by Suitable Metrics.....54

Table 3.1 Case Study Information 71

Table 3.2 Selected Inputs and Outputs of Four Data Sets 72

Table 4.1 Catchment characteristics 111

List of Figures

Figure 1.1 Structure of Thesis 14

Figure 2.1 Outline of methodology 29

Figure 2.2 Metric Categorisation based on Impacts on Sample Size and Dimension. The numbers in the figure refer to the metric number, details of which are given in Appendix C. 45

Figure 2.3 Typical Rate of Rejection Plots of Different Clusters 46

Figure 2.4 No. of Simulations until Values are Within 10% of the Values at 50,000 Simulations 52

Figure 3.1 Outline of methodology 69

Figure 3.2 Results of Mean Pairwise Convexity Deviation: (a) Change through No. of Hidden Nodes; (b) Change through No. of Parameters; (c) Change of the Global Structure 78

Figure 3.3 Results of the Hmax: (a) Change through No. of Hidden Nodes; (b) Change through No. of Parameters; (c) Change of the Multimodality / Roughness 80

Figure 3.4 Results of the Epsilon of Information Content: (a) Change through No. of Hidden Nodes; (b) Change through No. of Parameters; (c) Change of the Prevalence of Plateaus..... 82

Figure 3.5 Results of Median Basin Centroidal Distance: (a) Change through No. of Hidden Nodes; (b) Change through No. of Parameters; (c) Change of the Distribution of Local Basins 84

Figure 3.6 Results of Median Search Function Evaluations: (a) Change through No. of Hidden Nodes; (b) Change through No. of Parameters (c) Change of the Error Surface Nearby the Local Optima 86

Figure 3.7 Illustration of How a 1-D Slice of the Error Surface Changes due to an Increase in the Number of Hidden Nodes / Parameters: (a) Small MLPs; (b) Large MLPs..... **Error! Bookmark not defined.**

Figure 4.1 Outline of methodology 104

Figure 4.2 Model framework (adapted from Andrew et al., 2011) 108

Figure 4.3 Locations of rain gauges, catchment outlets, and weather stations from which data were obtained for calibration of the rainfall-runoff models for the five catchments (adapted from Guo et al., (2017)) 111

Figure 4.4 Impact of relative values of selected fitness landscape metrics on features of error surfaces of CRR models 115

Figure 4.5 Relative impact of Roughness (a), Flatness (b) and Optima Dispersion (c). Note that the catchments are ordered from left to right by decreasing wetness (wet, mild and dry catchments are coloured as red, green and blue respectively). 121

Figure 4.6 Influence of model structure / complexity on error surface features: Relative Roughness (a); Relative Flatness (b); Relative Optima Dispersion (c). The complexity of the CRR models increases from left to right: AWBM has 2 parameters, GR4J has 4 parameters, IHACRES has 6 parameters and Sacramento has 13 parameters. 124

Figure 4.7 Impact of degree of model complexity on optimisation efficiency, optimisation difficulty and parameter uniqueness based on the result of the application of the proposed metrics for quantifying error surface roughness and optima dispersion, as well as relevant previous studies confirming these findings.....126

Figure 4.8 Change in features of error surface from dry to wet catchments: (a) Change in the cluster category of Relative Roughness; (b) Change in degree of Relative Flatness; (c) Change in degree of Relative Optima Dispersion. 131

Figure 4.9 Change in features of error surface from dry to wet catchments: (a) Change in the cluster category of Relative Roughness; (b) Change in degree of Relative Flatness; (c) Change in degree of Relative Optima Dispersion. 132

Figure 4.10 Impact of model complexity and catchment climate condition on model calibration difficulty, efficiency and parameter uniqueness, based on the results in this study 136

Chapter 1 Introduction

1 Background

Optimisation methods are being used extensively for assisting with the identification of the most appropriate solutions for a range of environmental problems (Maier et al., 2014; 2019), such as stormwater management (Liu et al., 2016; Di Matteo et al., 2019), wastewater treatment (Hamed et al., 2004), land use management (Emirhüseyinoğlu and Ryan, 2020; Newman et al., 2020), environmental management (Kasprzyk et al., 2013), water-energy system design (Guidici et al., 2019), water distribution system design (Zecchin et al., 2006) and irrigation scheduling (Nguyen et al., 2017; Sedighkia et al., 2021), as well as the development of environmental models, including input variable selection (Grivas and Chaloulakou, 2006; Galelli et al., 2014) and model calibration (Pelletier et al., 2006; Burton et al., 2008). Existing research in this field has primarily focused on the development of improved optimisation algorithms, such as GALAXY (Wang et al., 2020), DREAM (Vrugt, 2016), Borg (Hadka and Reed, 2015), particle swarm optimisation (Chau, 2007), NSGA-II (Fu et al., 2008), ant colony optimisation (Emami Skardi et al., 2015) and policy tree optimisation (Herman and Giuliani, 2018), as well as the comparison of the performance of different algorithms on different problems

(e.g. Tikhamarine et al., 2020; Piotrowski and Napiorkowski, 2011; Kisi et al., 2012; Bullinaria and AiYahya, 2014; Wang et al., 2020). However, in accordance with the "No Free Lunch" theorem (Wolpert and Macready, 1997), no optimisation algorithm can outperform all others across every single problem. Consequently, there is a need to better understand the features of different optimisation problems so that algorithms that are better suited to particular problem types can be selected (Maier et al., 2014).

The features of optimisation problems can be represented geometrically by considering the "fitness landscape", which depicts the shape of the fitness function (otherwise termed objective function) for a particular objective with respect to the decision variables (e.g. model error as a function of different values of model parameters for model calibration problems) (see Maier et al., 2019). As the aim of the optimisation process is to find the highest or lowest points in this landscape, depending on whether the aim is to maximise or minimise the objective function, the ease or difficulty with which this can be done is a function of the features of this landscape. For example, if the landscape is smooth with a single, well-defined high- or low-point (global optimum), this point is relatively easy to find. Conversely, if the landscape is rough, with many minima or maxima of similar or equal value (local optima), the overall best solution (global optimum) is more difficult to find. Similarly, the presence of flat regions or plateaus in the fitness landscape generally makes it more difficult to guide the search towards the highest or lowest point in the landscape.

1.1 Metrics to understand the features of fitness landscapes of optimisation problems

In order to enable a better understanding of the features of optimisation problems to be obtained, a number of Exploratory Landscape Analysis (ELA) metrics have been developed (Mersmann et al., 2010; Munoz et al., 2015a). For example, such metrics can provide an indication of the global structure of the fitness landscape (e.g. its curvature), its degree of multi-modality (e.g. the prevalence of local optima) or the presence of plateaus (Mersmann et al., 2011). These metrics have been demonstrated to provide useful results for Black-Box Optimisation Benchmarks (BBOB) (Hansen et al., 2009), and they can successfully distinguish the differences of features between different benchmarks by using a given number of samples (Mersmann et al., 2011; Munoz et al., 2015a; Munoz and Smith-Miles, 2017). Furthermore, machine learning frameworks have also been successfully used to link the metric results with optimisation algorithm performance, so that the framework can be used to predict the performance of selected algorithms on different benchmark problems without trial and error (Smith-Miles et al., 2014; Munoz and Smith-Miles, 2017).

1.2 Applications of ELA metrics to water resource optimisation problems

Application of these metrics to environmental optimisation problems has been extremely limited (e.g. Gibbs et al., 2011; Bi et al., 2015). Instead, an empirical "brute force" approach is often used to determine which algorithm or parameterisation to use on a case study-by-case study basis (Maier et al., 2014). One potential reason for this is that there are different metrics for different landscape features (Mersmann et al., 2010; Malan and Engelbrecht, 2013; Maier et al., 2014; Munoz et al., 2015a), as well as different metrics for the same features, all with particular biases (Munoz et al., 2015a), making it difficult to know which metrics to use. However, the main reason for the lack of adoption of ELA metrics is likely to be related to the computational effort required to calculate them. As these metrics are calculated based on samples from the fitness landscape (Pitzer and Affenzeller, 2012), the number of samples required to obtain meaningful metric values can increase significantly with the size of the search space (Munoz et al., 2015a). When addressing real-world environmental optimisation problems, which are often characterised by large search spaces, this can either lead to computational intractability or the case where the computational effort associated with calculating the metrics is greater than that required as part of the "brute-force" approach of applying different algorithms or algorithm parameterisations to determine which works best.

1.3 Current knowledge about fitness landscapes of environmental optimisation problems

For environmental optimisation problems, previous studies mainly focus on the results or performance of optimisation. This includes 1) whether the best solutions can be found for a given optimisation problem; 2) whether there is only one best solution or a group of solutions with the same fitness value.

The first point was analysed by different kinds of optimisation problems listed in Section 1 in this chapter, and is particularly concerned with the application of data-driven models such as artificial neural networks (ANNs) to a wide range of hydrological modelling problems (e.g. Maier and Dandy, 1999; Zounemat-Kermani et al., 2016; Sivakumar et al., 2002; Piotrowski and Napiorkowski, 2013; Tan et al., 2018; Xie et al., 2021). As with all models, calibration is a critical component of the development of ANN models (termed "training" in the ANN literature). The unknown structure of the selected ANN models (e.g. the different settings of hidden nodes, hidden layers and transfer functions) (Maier et al., 2010; Wu et al., 2014) and corresponding parameterisation (Kingston et al., 2006; Mount et al., 2013; Humphrey et al., 2017) can lead to inaccurate information if the calibration process is not successful.

In order to increase the calibration performance of ANNs, an area of particular focus has been the comparison of optimisation algorithms for calibrating ANN

models (Kingston et al., 2005; Wu et al., 2013), with many studies reaching contradictory conclusions about which optimisation approaches perform better, with limited insight into why this might be the case. For example, Piotrowski and Napiokowski (2011) compared the performance of differential evolutionary algorithm (DE), particle swarm optimisation (PSO), differential evolution with global and local neighborhoods (DEGL) and Levenberg–Marquardt (LM) algorithms for calibrating ANN models, where they found that DE had very poor performance, and LM, a second-order gradient algorithm, outperformed all metaheuristics. In contrast, Maroufpoor et al. (2020) found that LM often prematurely converged to local optima, and grey wolf optimisation (GWO), a metaheuristic, could obtain better solutions than LM. Second-order gradient algorithms were also found to perform worse than first-order gradient methods by Maier and Dandy (1999).

The second point is also known as the parameter identification problem, which is a very typical problem for conceptual rainfall runoff (CRR) model calibration (Duan et al., 1992; Guillaume et al., 2019). Due to different sources of modelling non-identifiability in CRR models, such as model structural non-identifiability and observation error non-identifiability (for details, see Guillaume et al., 2019), there is usually more than one parameter set that leads to the minimum calibration errors when calibrating a CRR model. Different runs/initializations of an optimisation algorithm on the given problem can return different values of parameters (e.g. Shin et al., 2015), consequently, the physical meaning of the parameters are hard to identify and the calibrated

model may be not appropriate to use for prediction as the nonidentifiable parameters failed to capture the behaviours of the system.

In order to minimize the impact of non-identifiability of CRR models, most previous studies have focused on different ways of quantifying the difference between modelled and corresponding measured outputs (e.g. which error metric to use, which data to use against which to compare model performance (e.g. data length, data splitting, missing data, types of catchments)) (Gan et al., 1997; van Griensven, 2006; Vaze et al., 2010; Fowler et al., 2016), different approaches to identifying the best set of model parameter values (e.g. different optimisation methods) (Duan et al., 1992; Shin et al., 2015), different model structures (Andréassian et al, 2001; Shin et al., 2015; García-Romero et al, 2019) or how to best understand and quantify uncertainties associated with the calibration process (Beven 2016; Jackman et al., 2006; Kavetski et al., 2006; 2010). While the above papers are based on an implicit understanding that model errors change with values of model parameters, and that automated calibration using optimisation methods corresponds to the process of finding the lowest point in this "error surface" (i.e. fitness landscape), explicit assessments of how the characteristics of this surface change as a function of different model structures and the way errors are calculated, as well as the influence this has on the computational efficiency and difficulty of the calibration process and the uniqueness of the calibrated model parameters, have received less attention. Although a number of studies have demonstrated that knowledge of the features of the error surface is important for explaining and

interpreting the results of CRR model calibration trials (Sorooshian and Gupta, 1983; Iorgulescu and Jordan, 1994; Thyer et al., 1999; Suliman et al., 2016), the selection of appropriate model structures (e.g. Kavetski and Kuczera, 2007) and the choice of suitable optimisation algorithms (e.g. Duan et al., 1992; Kuczera, 1997; Kavetski et al., 2007), the above studies used ad-hoc methods for obtaining visualizations of lower-dimensional components of the error surface, and there are still no appropriate ways to understand the features of the entire error surface of high dimensional problems.

1.4 Limitations of understanding the problem structure for optimisation problems in water resources

The key limitation of understanding the problem structure for optimisation problems in water resources is that previous studies have primarily focused on the results of optimisation, including random/manual/automatic selection/comparison of different optimisation algorithms, models, and other factors to find the combination that leads to the best results among all trials. However, there has been a lack of understanding about why one kind of problem can be easier/harder to optimise than the other. While the visualization of lower-dimensional components of the fitness landscape has been used to interpret the results of optimisation problems (Razavi and Gupta, 2016a; 2016b), there is lack of methods to interpret the fitness landscape of high dimensional problems.

1.4.1 Lack of methods for fitness landscape analysis

As mentioned in Section 1.2 in this chapter, application of ELA metrics to environmental optimisation problems has been extremely limited, due to the complexity and high-dimensionality of real-world environmental optimisation problems, despite their success in the application to benchmark problems. To enable ELA metrics to be used for developing a better understanding of fitness landscape features and selecting the most appropriate optimisation algorithms for real-world optimisation problems, there is a need to determine (i) which ELA metrics, if any, have low dependence on problem dimensionality and sample size, so that they can be applied to real-world problems in a computationally efficient manner, and (ii) what information about the features of the fitness landscape can be ascertained from these metrics. This is the basis for the application of ELA metrics to real-world problems. However, there is a lack of knowledge about these metrics and their performance on real-world problems. Consequently, there is a need to determine which are suitable to be used to understand the fitness landscape of real-world problems.

1.4.2 Lack of understanding of fitness landscape of real-world optimisation problems

Due to the lack of knowledge of the features of fitness landscapes of real-world optimisation problems, previous studies only focused on optimisation results, including finding/developing optimisation algorithms that can outperform other algorithms for the case studies considered, and assessing the influence of different aspects of the problem investigated on optimisation performance to

select the combinations that maximize performance (e.g. finding the model structure and parameterisation of ANN models that can find the minimum model error efficiently, and finding the appropriate model structure, data length, error metric or other factors that can lead to the most identifiable parameter sets of CRR models). However, the settings determined by these studies can only be suitable for their particular cases, and the optimal settings are likely to change when applied to other case studies, due to the differences in fitness landscapes of these case studies. Therefore, it is worth identifying the reasons for the differences in selections between different case studies by understanding the features of fitness landscape of these problems. This can help to ensure the reliability of the selected combinations, as the reason why a given combination can result in good performance can be interpreted. Additionally, understanding the features of the fitness landscape can also provide prior knowledge to the selection of different combinations of problem characteristics, which can help to increase the quality of combinations for comparison by eliminating combinations that do not have suitable features of fitness landscapes.

2 Research aims

This thesis has two main aims. The first aim is to test the current widely used ELA metrics on benchmark problems to check their applicability to real-world environmental optimisation problems. The second aim is to apply the

applicable ELA metrics to understand the features of fitness landscapes of real-world optimisation problems, including ANN and CRR model calibration problems. These are typical data-driven and process-driven models in the water resources field, respectively, so that it is good to see the applicability of ELA metrics to both kinds of models. Additionally, the focuses of calibrating these two models are different. The focus of calibrating ANN models is mainly on finding the model structure and parameterisation which can obtain best solutions (i.e. minimum error) efficiently. As more complex ANN models can theoretically perform at least as well as simpler ones due to their higher degree of freedom, the failure of more complex ANNs to find better solutions is therefore likely to be caused by fitness landscapes / error surfaces that are complex, which makes it more difficult to find the better solutions. As a result, it is worth knowing how fitness landscape features change through the increase of ANN complexity. On the other hand, the focus of calibrating CRR models is not only about finding the best solutions, but also about the identifiability of model parameters, as they generally have a physical interpretation.

In order to meet these two aims, this thesis has three objectives. The first objective is related to the first aim discussed above. The second and third objectives are related to the application of ELA metrics that are found to be suitable (Objective 1) to ANN model calibration (Objective 2) and CRR model calibration (Objective 3), respectively. The general relationship between aims and objectives is shown in Figure 1.1, and detailed descriptions of the objectives are given below.

Objective 1. To check the applicability of ELA metrics to real-world environmental optimisation problems by identifying ELA metrics with low dependence on problem dimensionality and sample size.

Objective 1.1. To identify which ELA metrics have low dependence on problem dimensionality and sample size for a range of benchmark functions with a wide variety of known fitness landscape properties. This indicates which ELA metrics can be applied to real-world environmental optimisation problems from the perspective of computational tractability. It also opens the door to assessing the potential practical value of using ELA metrics to assist with determining which optimisation algorithm or settings might be most appropriate from a computational efficiency perspective, as the computational effort associated with the calculation of ELA statistics should be less than that associated with the "brute force" approach for determining which optimisation algorithm performs best.

Objective 1.2. To check whether the ELA metrics identified as having low dependence on problem dimensionality and sample size for benchmark functions also have low dependence on these factors for a number of real-life environmental modelling problems.

Objective 1.3. To map the ELA metrics that have low dependence on problem dimensionality and sample size to the fitness landscape features they are designed to provide information on, thereby providing a desktop assessment of the potential usefulness of the ELA metrics that are suitable to determining the features of real-world optimisation problems.

Objective 2. To assess whether ELA metrics that have low dependence on problem dimensionality and sample size are able to provide meaningful information on the potential calibration difficulty of ANN models of different complexity.

Objective 2.1. To assess the impact of model structure on different features of fitness landscape of ANN model calibration problems.

Objective 2.2. To identify the rules and trends of how different features of fitness landscapes of ANN model calibration problems change with changes in model complexity.

Objective 3. To assess whether ELA metrics that have low dependence on problem dimensionality and sample size are able to quantify the key features of the error surfaces (fitness landscapes) of CRR models.

Objective 3.1. To use ELA metrics to identify key error surface features for different combinations of model structures, catchments, error metrics and calibration data lengths.

Objective 3.2. To assess the corresponding implications for calibration (optimisation) efficiency, calibration (optimisation) difficulty and parameter uniqueness for different combinations of model structures with different key error surface features.

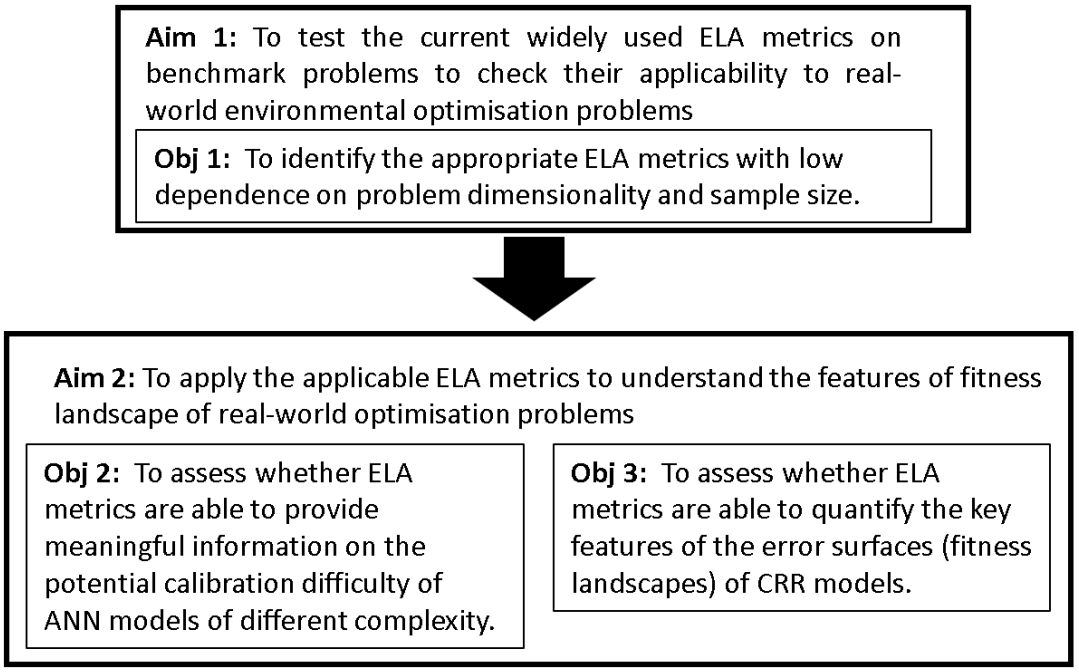


Figure 1.1 Structure of Thesis

3 Organisation of Thesis

The main body of this thesis (Chapters 2 to 4) comprises of three journal articles produced within this research. A summary of the thesis chapters is given below.

Chapter 2 (Journal paper 1) tests different ELA metrics and identifies the ones that have performance with low dependence on problem dimensionality and sample size, which can be suitable for applying to real-world environmental optimisation problems.

Chapter 3 (Journal paper 2) applies the metrics identified in Chapter 2 to ANN model calibration problems. ANN models with different complexity are tested, in order to identify the trends between the change of features of fitness landscapes of ANN model calibration problems and corresponding change of model complexity.

Chapter 4 (Journal paper 3) applies the metrics identified in Chapter 2 to CRR model calibration problems. Different CRR models for different catchments, error metrics and data length are tested. The chapter evaluates the level of impacts of these components on the calibration efficiency, difficulty and parameter uniqueness of CRR models through quantifying their key error surface features.

Chapter 5 summarises the contributions of the research. Future work is also discussed.

Chapter 2 Identification of Metrics Suitable for Determining the Features of Real-World Optimisation Problems

S. Zhu[1], H. R. Maier[1] and A. C. Zecchin[1].

[1]School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, SA, Australia.

(Revision 1 submitted to Environmental Modelling and Software after addressing reviewer comments)

Statement of Authorship

Title of Paper	Identification of Metrics Suitable for Determining the Features of Real-World Optimisation Problems
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Siwei Zhu
Contribution to the Paper	Primary innovator, analyst and author Experiment design and data analysis Manuscript draft and revise
Overall percentage (%)	80
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: 22/11/2021

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	Conception, Knowledge, Analysis, Drafting
Signature	Date: 22/11/2021

Name of Co-Author	Aaron C. Zecchin
Contribution to the Paper	Conception, Knowledge, Analysis, Drafting
Signature	Date: 29/11/2021

Please cut and paste additional co-author panels here as required.

Abstract

Optimisation methods are applied increasingly to environmental problems. Much research in this area is concerned with the behaviour of optimisation algorithms, however, the effectiveness of these algorithms is also a function of features of the problem being solved. Although a number of metrics have been developed to quantify these features, they have not been applied to environmental problems. The primary reason for this is that the computational cost associated with the calculation of many of these metrics increases significantly with problem size, making them unsuitable for real-world problems. In this chapter, 28 fitness landscape metrics that have low dependence on problem size are identified through extensive computational experiments on a range of benchmark functions and testing on a number of environmental modelling problems. These metrics can be applied to real-world optimisation problems in a computationally efficient manner to better understand their features and determine which optimisation algorithms are most suitable.

Highlights:

- Formal optimisation methods are used increasingly for a range of environmental problems.
- Fitness landscape metrics can be used to better understand the characteristics of optimisation problems.
- Many fitness landscape metrics are unsuitable for application to

real-world problems due to their high computational cost.

- 28 fitness landscape metrics that can be applied to real-world environmental optimisation problems are identified.
- These metrics provide insight into the identifiability of models and the selection of optimisation algorithms and parameters.

Keywords:

Optimisation, Calibration, Fitness Landscape, Error Function, Exploratory Landscape Analysis (ELA), Evolutionary Algorithms

1 Introduction

Optimisation methods are being used extensively to assist with the identification of the most appropriate solutions for a range of environmental problems (Maier et al., 2014; 2019), such as stormwater management (Liu et al., 2016; Di Matteo et al., 2019), wastewater treatment (Hamed et al., 2004), land use management (Emirhüseyinoğlu and Ryan, 2020; Newman et al., 2020), environmental management (Kasprzyk et al., 2013), water-energy system design (Guidici et al., 2019), water distribution system design (Zecchin et al., 2006) and irrigation scheduling (Nguyen et al., 2017; Sedighkia et al., 2021), as well as the development of environmental models, including input variable selection (Grivas and Chaloulakou, 2006; Galelli et al., 2014) and model calibration (Pelletier et al., 2006; Burton et al., 2008). Existing research in this field has primarily focused on the development of improved optimisation algorithms, such as GALAXY (Wang et al., 2020), DREAM (Vrugt, 2016), Borg (Hadka and Reed, 2015), particle swarm optimisation (Chau, 2007), NSGA-II (Fu et al., 2008), ant colony optimisation (Emami Skardi et al., 2015) and policy tree optimisation (Herman and Giuliani, 2018), as well as the comparison of the performance of different algorithms on different problems (e.g. Tikhamarine et al., 2020; Piotrowski and Napiorkowski, 2011; Kisi et al., 2012; Bullinaria and AiYahya, 2014; Wang et al., 2020). However, in accordance with the "No Free Lunch" theorem (Wolpert and Macready, 1997), no optimisation algorithm can outperform all others across every single problem. Consequently, there is a need to better understand the features of different optimisation problems so that algorithms that are better suited to

particular problem types can be selected (Maier et al., 2014).

The features of optimisation problems can be represented geometrically by considering the "fitness landscape", which depicts the shape of the fitness function (otherwise termed objective function) for a particular objective with respect to the decision variables (e.g. model error as a function of different values of model parameters for model calibration problems) (see Maier et al., 2019). As the aim of the optimisation process is to find the highest or lowest points in this landscape, depending on whether the aim is to maximise or minimise the objective function, the ease or difficulty with which this can be done is a function of the features of this landscape. For example, if the landscape is smooth with a single, well-defined high- or low-point (global optimum), this point is relatively easy to find. Conversely, if the landscape is rough, with many minima or maxima of similar or equal value (local optima), the overall best solution (global optimum) is more difficult to find. Similarly, the presence of flat regions or plateaus in the fitness landscape makes it more difficult to guide the search towards the highest or lowest point in the landscape. It should be noted that for multi-objective optimisation problems, each objective has its own fitness landscape, as variations in objective values with changes in decision variable values are likely to be different for different objectives (see Maier et al., 2019).

In order to enable a better understanding of the features of optimisation problems to be obtained, a number of Exploratory Landscape Analysis (ELA)

metrics have been developed (Mersmann et al., 2010; Munoz et al., 2015a). For example, such metrics can provide an indication of the global structure of the fitness landscape (e.g. its curvature), its degree of multi-modality (e.g. the prevalence of local optima) or the presence of plateaus (Mersmann et al., 2011). However, application of these metrics to environmental optimisation problems has been extremely limited (e.g. Gibbs et al., 2011; Bi et al., 2015). Instead, an empirical "brute force" approach is often used to determine which algorithm or parameterisation to use on a case study-by-case study basis (Maier et al., 2014). One potential reason for this is that there are different metrics for different landscape features (Mersmann et al., 2010; Malan and Engelbrecht, 2013; Maier et al., 2014; Munoz et al., 2015a), as well as different metrics for the same features, all with particular biases (Munoz et al., 2015a), making it difficult to know which metrics to use. However, the main reason for the lack of adoption of ELA metrics is likely to be related to the computational effort required to calculate them. As these metrics are calculated based on samples from the fitness landscape (Pitzer and Affenzeller, 2012), the number of samples required to obtain meaningful metric values can increase significantly with the size of the search space (Munoz et al., 2015a). When addressing real-world environmental optimisation problems, which are often characterised by large search spaces, this can either lead to computational intractability or the case where the computational effort associated with calculating the metrics is greater than that required as part of the "brute-force" approach of applying different algorithms or algorithm parameterisations to determine which works best. Consequently, to enable ELA metrics to be used for developing a better

understanding of fitness landscape features and selecting the most appropriate optimisation algorithms for real-world problems, there is a need to determine (i) which ELA metrics, if any, have low dependence on problem dimensionality and sample size, so that they can be applied to real-world problems in a computationally efficient manner, and (ii) what information about the features of the fitness landscape can be ascertained from these metrics.

In order to address these shortcomings, the objectives of this chapter are:

1. To identify which ELA metrics have low dependence on problem dimensionality and sample size for a range of benchmark functions with a wide variety of known fitness landscape properties. This indicates which ELA metrics can be applied to real-world environmental optimisation problems from the perspective of computational tractability. It also opens the door to assessing the potential practical value of using ELA metrics to assist with determining which optimisation algorithm or settings might be most appropriate from a computational efficiency perspective, as the computational effort associated with the calculation of ELA statistics should be less than that associated with the "brute force" approach determining which optimisation algorithm performs best.

2. To check whether the ELA metrics identified as having low dependence on problem dimensionality and sample size for benchmark functions also have low dependence on these factors for a number of real-life environmental modelling problems.

3. To map the ELA metrics that have low dependence on problem dimensionality and sample size to the fitness landscape features they are designed to provide information on, thereby providing a desktop assessment of the potential usefulness of the ELA metrics that are suitable to determining the features of real-world optimisation problems.

The remainder of this chapter is organised as follows. Details of the methodology used to achieve the above objectives are given in Section 2, followed by the results and discussion in Section 3. Summary and conclusions are provided in Section 4.

2 Methodology

2.1 Overview

An overview of the methodology used to achieve the three objectives stated in the Introduction is given in Figure 2.1 (with further details provided in Appendix A). As can be seen, the first three steps of the identification of ELA metrics with low dependence on problem dimensionality and sample size with the aid of benchmark functions (objective 1), and checking whether these metrics also have low dependence on problem dimensionality and sample size for real-life environmental modelling problems (objective 2), are the same. The first of these steps includes the sampling of fitness landscapes with different features and dimensionality using a range of samples sizes, as these samples are required for the calculation of the different ELA metrics in Step 2.

For objective 1, fitness landscapes with different features are represented by the noiseless BBOB suite of 24 benchmark functions (Hansen et al., 2009), as

these contain a wide range of known landscape features (see Section 2.2.1 for details), can be scaled to different dimensionalities (Hansen et al., 2019) and have been used in a number of fitness landscape studies (Mersmann et al., 2010, 2011; Shirakawa and Nagao, 2014, 2016; Munoz et al., 2015b; Munoz and Smith-Miles, 2017; He et al., 2018). Twenty replicates are generated for five different dimensionalities (2, 5, 10, 20, 30) for each of the 24 benchmark functions, resulting in 2,400 (24 functions x 5 dimensionalities x 20 replicates) fitness landscapes. Each of these is sampled 30 times with different sampling lengths ranging from 100 to 120,000, resulting in 72,000 (2,400 fitness landscapes x 30 sample lengths) sets of fitness landscape samples. A maximum dimensionality of 30 is selected, as this corresponds to the upper end of dimensionalities used in previous studies using these benchmark functions (Mersmann et al., 2011; Munoz et al., 2015b; Munoz and Smith-Miles, 2017; Shirakawa and Nagao, 2014, 2016; Kerschke et al., 2015; Garden and Engelbrecht, 2014). A maximum sample length of 120,000 is used, as this has been found to be sufficient for the convergence of ELA metric values in preliminary analyses and is significantly greater than sample sizes used in previous ELA studies, which are generally on the order of 6,000 or 1000 × Dimension (Mersmann et al., 2011; Munoz et al., 2015b; Munoz and Smith-Miles, 2017; Shirakawa and Nagao, 2014, 2016; Kerschke and Preuss, 2015; Garden and Engelbrecht, 2014).

For objective 2, the fitness landscapes with different features correspond to those for the calibration (training) of artificial neural network (ANN) models

used to predict a number of environmental variables (runoff, turbidity, salinity) (see Section 2.2.2 for details). These environmental modelling problems have been selected as (i) model calibration is a common environmental optimisation problem (Maier et al., 2019), (ii) ANNs have been used extensively for environmental modelling (see Maier et al., 2010; Wu et al., 2014; Cabaneros et al., 2019), (iii) the parametric dimensionality of ANNs can be changed within a single model framework by increasing the number of hidden nodes (Maier et al., 2010) and (iv) the fitness landscapes associated with the calibration of ANN models have been shown to vary in complexity (e.g. Kingston et al., 2005; Samarasinghe, 2006). Ten replicates are generated for 11 model structures (0-10 hidden nodes, corresponding to problem dimensionalities ranging from 1 to 70 – see Section 2.2.2 for details) for each of the 3 modelling problems considered, resulting in 330 (3 problems x 11 dimensionalities x 10 replicates) fitness landscapes. Each of these is sampled 23 times with different sampling lengths ranging from 100 to 50,000, resulting in 7,590 (330 fitness landscapes x 23 sample lengths) sets of fitness landscape samples. A maximum sample length of 50,000 is selected as this has been found to be sufficient for convergence of ELA metric values in preliminary analyses.

The second of these three steps involves the calculation of the desired ELA metrics for each of the sets of fitness landscape samples generated in the previous step. For objective 1, 89 ELA metrics are considered, which constitute the full set of metrics used in previous fitness landscape analysis studies (Mersmann et al., 2011; Munoz et al., 2015b; Munoz and Smith-Miles, 2017),

to maximise the chances of identifying metrics with low dependence on problem dimensionality and sample size that also provide useful information about a range of fitness landscape features. For the environmental modelling problems (objective 2), only the ELA metrics that are found to have low dependency on problem dimensionality and sample size for the benchmark functions (see Figure 2.1, step 4) are used in order to check whether the findings from the benchmark functions apply in real-life environmental modelling contexts.

The third and last of these steps involves the calculation of the degree of dependence of the ELA metrics considered in the previous step on both problem dimensionality and sample size. If a metric has low dependence on both, it is likely to be able to be applied to real-world environmental optimisation problems. If this is not the case, the computational effort required to calculate the metric is likely to be too large for practical purposes.

For the benchmark functions (objective 1), dependence is represented by the relationship between sample size, problem dimensionality and the reject rate, which is the fraction of the 24 test functions for which the hypothesis that a particular sample size gives the "true" value of a particular fitness landscape metric does not hold based on the Wilcoxon Rank Sum Test (calculated using 20 replicates) (see Section 2.4.1 for details). In this context, the "true" value is taken as the value obtained for the largest number of samples considered (i.e. 120,000). If the reject rate for a particular metric is low, and remains so with

increasing dimensionality (over the 5 dimensionalities considered) and sample size (over the 30 sample lengths considered), calculation of the "true" value of this metric can be considered to have low dependence on problem dimensionality and sample size, making it a suitable candidate for application to real-world environmental optimisation problems. The Wilcoxon Rank Sum Test is used for this purpose as it provides a statistically rigorous approach to testing dependence for a desired confidence level.

For the environmental modelling problems (objective 2), whether the low degree of dependence on problem dimensionality and sample size holds for the selected metrics is checked by calculating the number of samples required for a particular ELA metric value to be within 10% of the "true" value of this metric for different problem dimensionalities (averaged over the three case studies), as represented by ANN models with different numbers of hidden nodes (and hence model parameters) (See Section 2.4.2 for details). In this context, the "true" value is taken as the value obtained for the largest number of samples considered (i.e. 50,000). If the number of samples required to achieve "accurate" results is relatively small (e.g. less than 5,000), then the low dependence of the metric under consideration on sample size is confirmed. If the number of required samples is small for ANN models with different numbers of hidden nodes (i.e. different problem dimensionalities), then the low dependence of the metric under consideration on problem dimensionality is also confirmed.

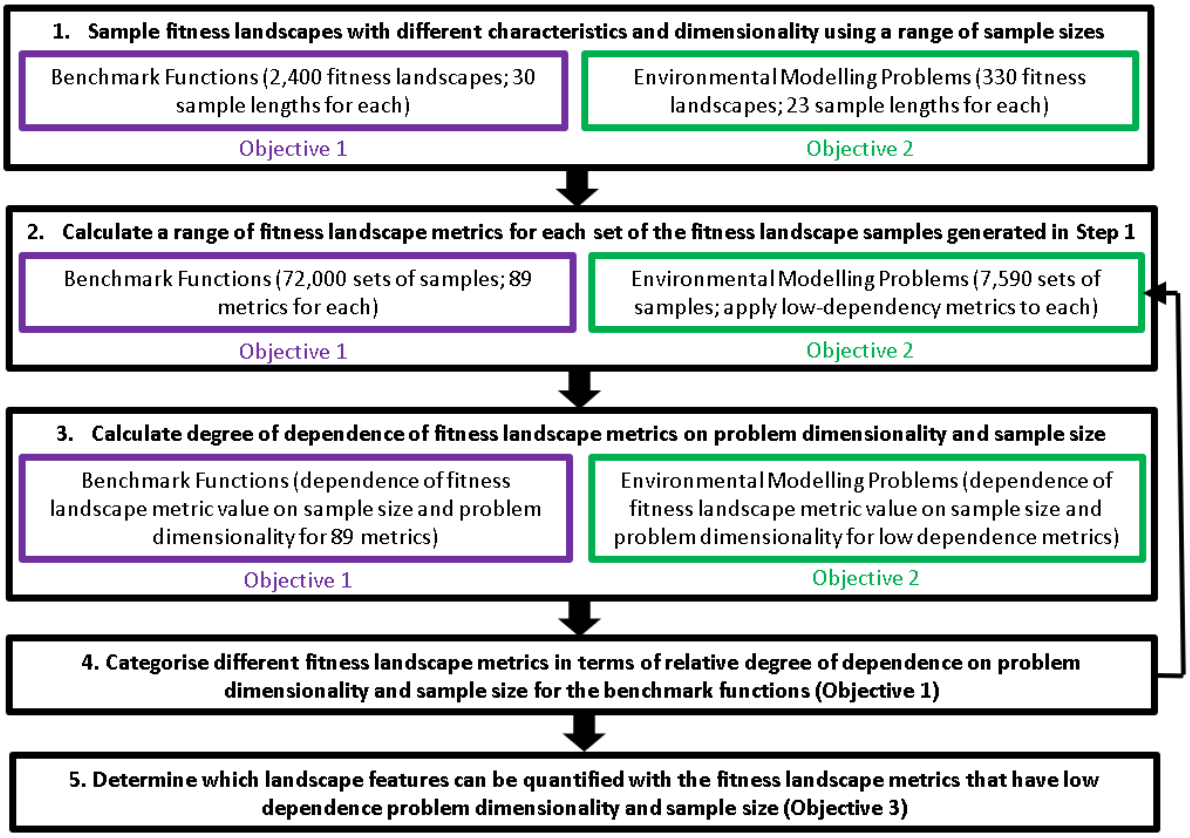


Figure 2.1 Outline of methodology

The fourth step involves the categorisation of the different ELA metrics in terms of their relative degree of dependence on problem dimensionality and sample size for the benchmark functions (objective 1). This is achieved by grouping the 89 ELA metrics considered based on the centroids of the Euclidean distances of the slopes of the linear regression relationships (with logarithm transformation) between reject rate (see Step 4) and ((sample size) or (problem dimensionality)) using a hierarchical clustering approach (Nielsen, 2016) (see Section 2.5 for details). This provides an indication of whether there are natural groupings of metrics with differing degrees of dependence on problem dimensionality and sample size and potential reasons for this, as well as which metrics, if any, have low dependence on both problem dimensionality

and sample size, and are hence suitable for application to real-world optimisation problems.

As part of the fifth and final step, the landscape features that can be quantified with the fitness landscape metrics that have low dependence on problem dimensionality and sample size, identified as part of Objective 1 and validated in Objective 2, are identified by mapping these metrics to the landscape features they are designed to provide information on via a desktop assessment (objective 3). This provides information on the types of landscape features that can be obtained from ELA metrics that can be applied to real-world problems, which can then be used to gain a better understanding of the characteristics of different environmental optimisation problems and which optimisation methods and parameterisations might be most appropriate for these. However, performing such assessments is beyond the scope of this chapter.

The above analyses are conducted using the University of Adelaide's supercomputing facilities, which consist of 48 skylake nodes, with 80 cpus and 377GB of memory per node. Samples from the 24 noiseless BBOB benchmark functions are generated using the R Package FLACCO (Kerschke and Trautmann, 2016), the R package ValidANN (Humphrey et al., 2017) is used for ANN development, the R package fastcluster (Müllner , 2013) is used for hierarchical clustering, the R codes wilcox.test() and lm() are used for the Wilcoxon Rank Sum test and regression analysis, respectively, and the MATLAB code PLHS (Sheikholeslami and Razavi, 2017) is used for sample

generation. R-code for how to calculate the ELA metrics for a given set of fitness landscape samples is provided as supplementary material.

2.2 Sampling of fitness landscapes

2.2.1 Benchmark functions

As mentioned in Section 2.1, the noiseless BBOB suite of 24 benchmark functions (Hansen et al., 2009) has been used extensively for a number of fitness landscape studies. The functions have been designed specifically to represent fitness landscapes with a wide range of features, such as the degree of multi-modality, global structure, variable scaling in different directions (i.e. variable sensitivity), degree of similarity in different areas of the search space, size of basins of attraction (optima), and differences in the magnitudes between global and local optima and degrees of flatness of the search space (plateaus) (see Table 2.1). The degree to which these different features are represented in the 24 functions is summarised in Table B.1 in Appendix B.

Table 2.1 Descriptions of High-Level Fitness Landscape Features

Feature	Description
Multi-Modality	Problems with higher multi-modality have a higher density of local optima in the search space and are therefore more difficult to solve. In contrast, problems with lower multi-modality have a lower density of local optima, making it easier to identify the global optimum.
Global to Local Optima Contrast	Fitness landscapes with a greater global to local optima contrast are easier to search, as the differences between global and local optima are larger, making it is easier to distinguish global optima from local ones, avoiding convergence to local optima. In contrast, fitness landscapes with a smaller global to local optima contrast are harder to search, as it is easier for algorithms to become trapped in local optima, leading to premature convergence.
Basin Size Homogeneity	Problems with greater basin size homogeneity are easier to solve, as this enables algorithms with particular parameterisations, and hence searching

	behaviour, to effectively search the entire fitness landscape. In contrast, problems with lower basin size homogeneity are more difficult to solve, as the algorithm parameterisation that results in optimal searching behaviour in one basin is unlikely to be optimal in another basin. Therefore, only some basins can be fully searched, resulting in a higher probability that the global optimum will be missed.
Search Space Homogeneity	Problems with a greater search space homogeneity are easier to solve, as their features are more similar throughout the entire fitness landscape and hence only a single algorithm/algorithm parameterisation is likely to be required for the search to be successful. In contrast, lower search space homogeneity is an indication that fitness landscape features are likely to be different in different areas of the search space, making it more difficult to find the global optimum with a single algorithm/algorithm parameterisation.
Global Structure	Problems with a more well-defined global structure (e.g. a "big-bowl" shape) are easier to solve as they are able to guide optimisation algorithms into promising regions of the search space. In contrast, the global optimum is more difficult to find for problems with a less well-defined global structure (e.g. a "flat" fitness landscape), as there is little information to guide optimisation algorithms towards this optimum.
Plateaus	Fitness landscapes with more plateaus generally result in slower convergence during the search, as they contain regions where there are minimal differences in fitness function values, providing less useful information to guide optimisation algorithms into promising areas of the search space in these regions. In contrast, fitness landscapes with fewer plateaus generally result in faster convergence, as they provide more useful information in guiding the algorithm search more consistently throughout the fitness landscape.
Separability	Problems with a higher degree of separability are easier to solve, as they enable the problem to be split into several lower-dimensional sub-problems that can be solved independently, making the overall problem easier to solve. In contrast, problems with a lower degree of separability are more difficult to solve, as the problem is more difficult to be split into lower-dimensional sub-problems, requiring higher-dimensional problems to be solved.
Variable Scaling	Problems with a higher degree of variable scaling introduce potential challenges during the optimisation process, as there are larger differences in the contribution of different decision variables to fitness values. Consequently, changes in values of variables with small contributions during the optimisation process do not result in significant changes in fitness function values, requiring variables with larger contributions to converge to good values before the other decision variables have an influence, which may have converged to poor values in the meantime (Gibbs et al., 2011; Maier et al., 2014). In contrast, problems with a lower degree of variable scaling do not present these challenges (unless all variables have no contribution to fitness function values) as changes in the values in any of the decision variables during the optimisation process result in relatively significant changes in the fitness function, thereby enabling all decision variables to converge to good values simultaneously.

2.2.2 Environmental modelling problems

As mentioned in Section 2.1, three different environmental modelling problems are considered, including: rainfall-runoff modelling in the Kentucky River, USA; the prediction of filtered water turbidity from a range of raw water quality parameters and the added Alum dose (treatment) for surface waters in South Australia; and the forecasting of salinity in the River Murray at Murray Bridge, South Australia, based on values of upstream salinities and flows. These are selected as they represent a diversity of environmental problems that have been used in a number of previous benchmarking studies (e.g. Wu et al., 2013; Humphrey et al., 2017). As was the case in these studies, the selected ANN model architecture is a multi-layer perceptron (MLP), as this type of model architecture has been used widely and successfully in practice and enables fitness landscapes (i.e. the calibration error functions) with different dimensionalities (in terms of number of model parameters) to be generated within the same model structural framework simply by changing the number of hidden nodes (Maier et al., 2010; Wu et al., 2014). Details of the model inputs and outputs, as well as the available data, are given in Table 2.2, which are identical to those used in previous studies. As mentioned in Section 2.1, the number of hidden nodes for each ANN is varied between 0 and 10 to ensure that fitness landscapes with different features are obtained (see Table 2.1). The root mean square error (RMSE) is used as the objective function for model calibration, as was the case in previous studies (Wu et al., 2013; Humphrey et al., 2017). It should be noted that as the aim of this study is to identify the characteristics of the fitness landscapes of different ANN models, model

calibration (training) is not required; only samples generated from the fitness landscapes are needed.

Table 2.2 Case Study Information

	Case Study		
	Kentucky Rainfall Runoff	River Murray Salinity	South Australia Turbidity
References	Jain and Srinivasulu (2006), Wu et al. (2013), Humphrey et al. (2017)	Maier and Dandy (1996), Bowden et al. (2002), Wu et al. (2013), Humphrey et al. (2017)	Maier et al. (2004), Wu et al. (2013), Humphrey et al. (2017)
No. of Inputs	2	2	5
No. of Hidden Nodes	0, 1, 2, …, 10		
No. of Parameters	3 to 41	3 to 41	6 to 71
Calibration Data Points	2842	1215	120
Inputs (Lags)	Flow (t, t-1)	Mannum salinity (t), Waikerie Salinity (t)	Raw Water Turbidity, Raw Water pH, Raw Water Color, Raw Water UVA, Alum dose
Outputs (Lags)	Flow (t+1)	Murray Bridge salinity (t+14)	Filtered water turbidity

2.3 Fitness landscape metrics

As mentioned in Section 2.1, a total of 89 ELA metrics are considered. These consist of the six low-level groups of metrics developed by Mersmann et al. (2011), including convexity, y-distribution, level set, meta model, local search and curvature, and the information content of fitness sequences (ICoFS) metrics as shown in Table 2.3. Each of these groups of metrics is designed to provide information on different combinations of fitness landscape features, including global structure, multimodality, separability, global to local optima

contrast, search space homogeneity, plateaus, variable scaling and basin size homogeneity (Tables 2.1 and 2.3). Brief outlines of these metrics are given below. More detailed information, such as the statistics used to summarise the results of these metrics, and the mapping between the metrics and features, are summarised in Appendix C. Unfortunately, there is no direct correlation between particular metrics and individual fitness landscape features, making it difficult to provide a more intuitive understanding of the metrics.

Convexity Metrics: convexity metrics use the deviation between linear regressed fitness values y' and the true fitness value y to analyse the shape of fitness landscapes. Random pairs of points (x_i, x_j) are selected from the total sample pool X, and a third point is selected from the line between these two. At this new point, the actual fitness landscape value is compared to the linear interpolation of fitness values from the original points. Whether the landscape is positively or negatively convex can provide information about the overall shape of the fitness landscape.

y-Distribution Metrics: y-distribution metrics use the probability density function (PDF) of fitness values of samples to provide information on the scaling and distribution of fitness landscapes in terms of the fitness values. The PDF is estimated based on the frequency of fitness values identified by selected samples on the search space. The distribution shown by the PDF can provide information about how easy it is to identify solutions with better fitness values, as if more samples that have good fitness values are shown to have a higher probability to be identified, it should be relatively easier to find the globally

optimal solution, unless the fitness landscape is “deceptive”, in which case the global optimum is not in the vicinity of good local optima, making it more difficult to locate (e.g. needle in the haystack problems) (Deb and Goldberg, 1994; Maier et al., 2014).

Level Set Metrics: level set metrics use discriminant analysis to check the complexity of fitness landscapes. Samples are assigned to high- and low-quality groups based on their fitness values. Next, different predictive models are used to check whether they can re-classify the samples accurately.

Meta Model Metrics: meta model metrics involve the building of regression models based on sampling points and checking how well these fit to the fitness landscapes. This is in order to show the similarity between the regression models and the corresponding problem’s fitness landscape. Different regression models, including both independent (simple) and cross-term parameters, are used to check the separability of a fitness landscape. Separable fitness landscapes should be more easily fitted to simple models, whereas non-separable fitness landscapes should be better fitted by the cross-term models.

Local Search Metrics: local search metrics use the information provided by a set of local optima (obtained by a gradient algorithm using random starting points) to assess the properties of the distribution of optimal solutions across a fitness landscape. Of importance here is estimating the size of the basin of attraction for local optima. Local optima with short distances between each other are clustered within a common basin. The size of these basins and how they are distributed across the fitness landscape are related to the features the local optima.

Curvature Metrics: curvature metrics assess the information provided by first order derivatives and Hessian matrices of sample points on the fitness landscape. This information can help to assess whether each variable has the same influence in guiding the searching process, and whether the fitness landscape provides enough information to guide the searching to find good solutions, especially for derivative/perturbation based algorithms.

ICoFS Metrics: The Information Content of Fitness Sequences (ICoFS) metrics use a set of samples to construct a sequence of fitness landscape values (based on either nearest neighbour in the parameter space or a random ordering). These sequences are then used to create an indicator sequence of values from {-1,0,1} depending on the comparison of sequential values in the original sequence. That is, for consecutive sequence values y_n and y_{n+1}, a value of -1 is assigned if $y_n > y_{n+1}$, 0 for $y_n = y_{n+1}$, and 1 for $y_n < y_{n+1}$ (note that a threshold is used for the inequality comparisons). A smooth, near monotonic, fitness landscape (e.g. single modal one) would be expected to have a sequence without frequent signal change (e.g. [1, 1, 1, 1, 1] or [-1, -1, -1, -1, -1]). A multi-model fitness landscape, on the other hand, would be expected to have a sequence with frequent changes in the signal (e.g. [1, -1, 1, -1, 0, -1, 0, 1]). Finally, a flat fitness landscape would have a sequence of zeros due to its only slight difference in fitness values. The information contained in the sequence is processed, and used to provide characterisations of the level of roughness in the fitness landscape, which is highly related to multi-modality.

Table 2.3 Relationship between ELA Metric Classes (in the columns) and Fitness Landscape Features (in the rows)

		Fitness Landscape Features							
		Global Structure	Multimodality	Separability	Global to Local Optima Contrast	Search Space Homogeneity	Plateaus	Variable Scaling	Basin Size Homogeneity
ELA Metric Classes	Convexity	√	√			√			
	Y-Distribution	√	√				√		
	Level Set	√	√				√		
	Meta Model	√	√	√			√	√	
	Local Search		√		√	√			√
	Curvature						√	√	
	ICoFS	√	√				√		

2.4 Calculation of degree of dependence of metrics on dimensionality and sample size

As mentioned in the Introduction, the objective of this chapter is to identify which ELA metrics have low dependence on problem dimensionality and sample size so that they can be applied to real-world environmental optimisation problems. It should be noted that although the primary features of the benchmark functions are known, assessing the degree to which the ELA metrics under consideration are able to correctly assess these features is beyond the scope of this chapter. In addition, there is no direct one-to-one correspondence between different metrics and different features, as mentioned in Section 2.3.

2.4.1 Benchmark functions

As mentioned in Section 2.1, the degree of dependence of the benchmark

functions on problem dimensionality and sample size is assessed with the aid of the Two-Tailed Wilcoxon Rank Sum Test, which is a nonparametric test with a null hypothesis that the probability of the selected populations arises from the same underlying distribution. In this case, the two populations under consideration consist of values of fitness landscape metrics calculated using a given number of samples (e.g. 100 to 120,000 – see Figure 2.1) and those calculated using the largest number of 120,000 samples. The reasoning behind this is that the metric value at the largest number of samples is taken as the most accurate computation of this metric. That is, the hypothesis test is given as:

$$\begin{aligned} H_0&: ELA_{i,j,d,k} = ELA_{i,j,d.120{,}000} \\ H_1&: ELA_{i,j,d,k} \neq ELA_{i,j,d,120{,}000} \end{aligned} \tag{2.1}$$

where $ELA_{i,j,d,k}$ and $ELA_{i,j,d,120{,}000}$ represent the ELA metric results of the ith metric, the jth test function and the dth dimension, with k samples and 120,000 samples, respectively. Consequently, if the null hypothesis is satisfied for relatively small sample sizes for a particular fitness landscape metric, this metric can be considered to have low dependence on sample size.

In this study, a 95% confidence level is used to test the hypotheses. To enable the results of the different computational experiments to be compared more easily, they are represented in terms of the reject rate, symbolized by R, which is the percentage of experiments for which the above null hypothesis is rejected, and is calculated as follows:

$$R_{i,d,k} = \sum_{j=1}^{N} \frac{I\{p_{i,j,d,k} \leq 0.05\}}{N} \tag{2.2}$$

where $R_{i,d,k}$ refers to the reject rate of metric i among the 24 test functions for problems with d dimensions and k samples, $I\{x\}$ is an indicator function which is equal to 1 if the Boolean statement x is true and zero otherwise, and N is the total number of test functions, which is 24 in this study. Consequently, lower values of the reject rate indicate that a metric is more independent of sample size and dimensionality and hence more suited to being applied to real-world environmental optimisation problems.

2.4.2 Environmental modelling problems

As mentioned in Section 2.1, fitness landscape metrics that are found to have low dependence on both problem dimensionality and sample size (i.e. a low reject rate) for the test functions (see Section 2.3) are applied to the real-world environmental modelling problems to check if the low dependence of these metrics on sample size and dimensionality holds for the real-world environmental optimisation problems considered. As also mentioned in Section 2.1, this check is achieved by calculating the number of samples required for a particular ELA metric value to be within 10% of the "true" value of this metric for different problem dimensionalities, which is considered reasonable for practical purposes. The percentage difference is calculated using the normalized difference between the "true" value of a given ELA metric, obtained for a sample size of 50,000 (see Section 2.1), and the corresponding value for a smaller sample size, k (e.g. 100, …, 50,000 – see Figure 2.1), as follows:

$$Err_{i,h,c,k} = \left|Med_{i,h,c,k} - Med_{i,h,c,50,000}\right| / Med_{i,h,c,50,000} \times 100\% \quad (2.3)$$

where $Err_{i,h,c,k}$ is the normalised error, which is calculated by using the corresponding medians of Metric i, Case c and Number of Hidden Nodes h (corresponding to different problem dimensionalities – see Table 2.2), and $Med_{i,h,c,k}$ is the median of the metric values in the data set as identified by the subscripts (the subscripts have the same meaning as in $Err_{i,h,c,k}$). In addition to metrics that were considered unsuitable based on the computational criteria, five ELA metrics that were not considered to provide useful information on the real-world environmental case studies considered were also excluded. These include metrics that apply to linear and non-continuous relationships, which is not the case for ANNs, as they are highly non-linear and continuous. Consequently, values of these metrics are equal to zero for the case studies considered, therefore not providing any useful information.

2.5 Categorisation of fitness landscape metrics

As mentioned in Section 2.1, using the results from the analysis on the benchmark functions, the assessed ELA metrics are categorized based on their degree of dependence on problem dimensionality and sample size, in order to identify metrics with low dependence on both. This is achieved via a two-step process.

The first step involves the quantification of the degree of dependence of metric values on sample size and problem dimensionality. This is achieved by developing a regression model that relates the reject rates calculated in Eq. (4)

to sample size and problem dimension as follows:

$$R = a \cdot \ln(Dim) + b \cdot \ln(SS) + c \qquad (2.4)$$

where R represents the reject rate, Dim and SS represent dimension and sample size, respectively, and a , b and c represent the slope of dimension and sample size, and intercept, respectively. By way of interpretation, for example, a large value of a implies that the reject rate is highly influenced by the dimension. Example R-code for performing these calculations is provided as supplementary material.

This form of the relationship was considered most appropriate based on visual inspection of the plots of reject rate versus sample size and problem dimensionality, with a logarithm transformation used to scale the magnitude of the coefficients. As shown in Appendix D, the r^2 values of these relationships generally range between 0.3 and 0.86, indicating the ability to discriminate between relationships of different strengths. For a small number (10) of relationships, r^2 values were less than 0.2. However, as shown in Appendix D, this was for relationships with very low dependence on sample size and problem dimensionality, where the fluctuations in the relationship (i.e. noise) had a significant impact on the r^2 values. However, this did not affect the correct quantification of the relative impact of sample size and problem dimensionality on ELA metric values, which is the primary objective.

The second step involves hierarchical clustering (Nielsen, 2016) of the values of the slopes for dimensionality (i.e. values of a in Eq. (6)) and sample size (i.e. values of b in Eq. (6)) for different ELA metrics based on the centroids of their

Euclidean distances to identify groups of ELA metrics with different degrees of dependence on sample size and problem dimensionality.

It should be noted that in order for the above results to be meaningful, the absolute values of R also need to be checked. While low dependence on sample size and problem dimensionality are pre-conditions for the application of ELA metrics to real-world environmental problems, metrics belonging to this category only provide useful information if the values of R are consistently low, rather than consistently high. In this study, this check is performed by visual inspection of the plots of R versus sample size and problem dimensionality.

3 Results and Discussion

3.1 Categorisation of fitness landscape metrics

3.1.1 Cluster Location

As can be seen in Figure 2.2, the ELA metrics considered form five distinct clusters with different degrees of dependence on problem dimensionality and sample size. Typical relationships between reject rate, dimensionality and sample sizes for metrics in these clusters are shown in Figure 2.3.

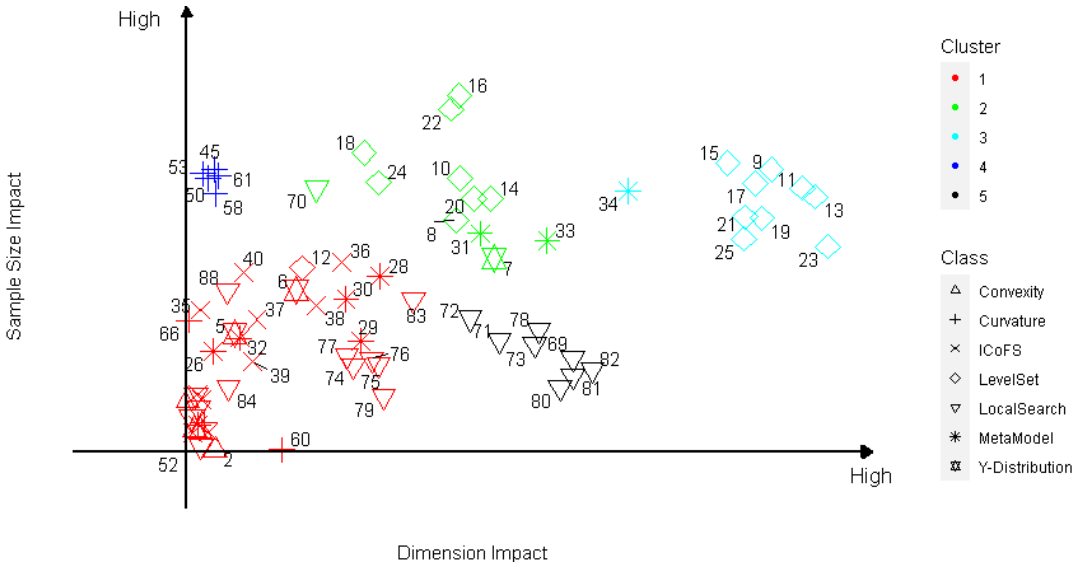
The 39 metrics belonging to cluster 1 have a low dependence on problem dimensionality and sample size (Figure 2.2), as evidenced by the flat slopes of the relationships between reject rate and both problem dimensionality and sample size, as seen in Figure 2.3(a). The fact that the reject rate for metrics belonging to this cluster is very low across the full range of sample sizes and

problem dimensionalities investigated (Figure 2.3(a)) suggests that metrics belonging to this cluster are suitable for application to real-world environmental optimisation problems. In contrast, this does not appear to be the case for the metrics belonging to the remaining clusters. For example, the 23 metrics belonging to Clusters 2 and 3 have high dependence on sample size and medium-high dependence on problem dimensionality, with typical plots of the relationships between reject rate, problem dimensionality and sample size for metrics belonging to these clusters shown in Figures 2.3(b) to 3(d). This is likely to make the application of these metrics to real-world environmental optimisation problems computationally intractable.

The 5 metrics belonging to cluster 4 have low dependence on dimensionality, but high dependence on sample size, as evidenced by a typical plot of reject rate versus these two factors in Figure 2.3(e). This makes metrics belonging to this cluster difficult to apply in practice, as large sample sizes are required for even relatively simple problems. The 7 metrics belonging to cluster 5 have low dependence on sample size, but medium dependence on problem dimensionality (see Figure 2.3(f) for a typical plot of the relationship of reject rate versus sample size and dimensionality). This makes them applicable to relatively simple real-world problems, but computational tractability is likely to become an issue for higher-dimensional problems.

It should be noted that 15 metrics are excluded from clusters 1 to 5 in Figure 2.2, as their reject rates are very high (see Figures 2.3(g) and 3(h)), making them unsuitable for application to real-world optimisation problems, as

discussed in Section 2.5. As can be seen in Figure 2.3(g), there are some cases where the reject rate is low when the sample size is small, but this increases rapidly when the sample size increases to a given level. This is because when the sample size is small, the values of these metrics are highly variable, providing greater opportunities for the median values to be close to the "true" value. However, this variability decreases with an increase in sample size, reducing the chance that the median values are close to the "true" value of the metric, indicating that the actual reject rates for these metrics are very high and therefore not suitable for application to real-world environmental problems.



Cluster	Dimension Impact	Sample Size Impact
1	Low	Low
2	Median	High
3	High	High
4	Low	High
5	Median	Low

Figure 2.2 Metric Categorisation based on Impacts on Sample Size and Dimension. The numbers in the figure refer to the metric number, details of which are given in Appendix C.

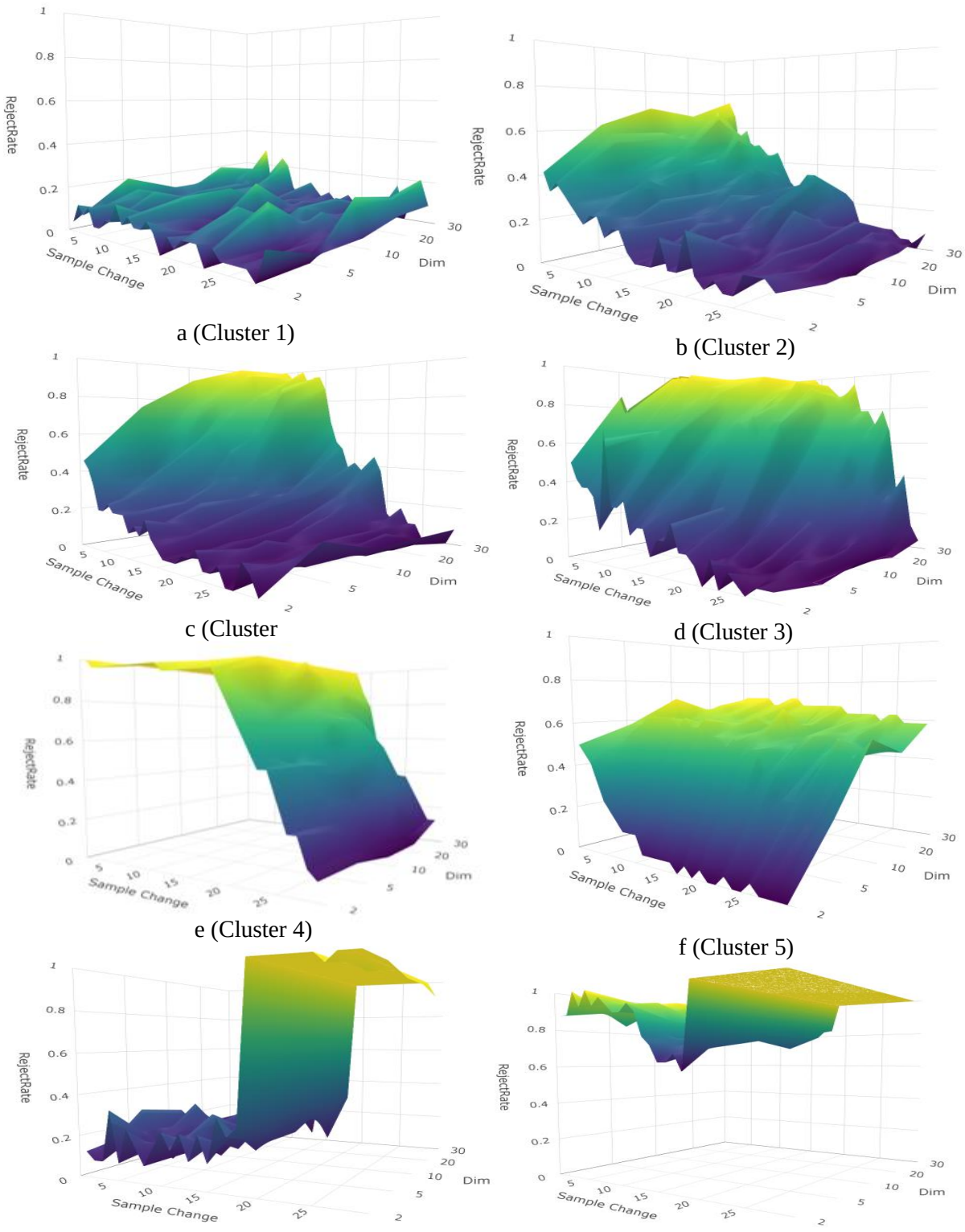


Figure 2.3 Typical Rate of Rejection Plots of Different Clusters

3.1.2 Cluster Composition

A summary of the composition of each of the five clusters in Figure 2.3 in terms of metric class is given in Table 2.4. As can be seen, cluster 1 contains at least one metric from each class, with all of the 4 convexity and 10 ICoFS metrics belonging to this cluster. The majority of the 3 Y-Distribution (66.7%), the 9 Meta Model (66.7%) and 21 Local Search (57.1%) metrics also belong to cluster 1, while only one out of the 18 Level Set (5.6%) and 4 out of the 24 Curvature (16.7%) metrics fall into this cluster.

A common feature of all metrics belonging to cluster 1, irrespective of which metric class they are part of, is that their calculation only requires fitness values and the relative distance between samples, without knowledge of the location of each sample in the search space, as is the case with many of the other metrics. This is a likely cause for the low dependence of the calculation of these metrics on sample size and problem dimensionality.

The majority of the metrics belonging to clusters 2 and 3 are part of the Level Set and Meta Model classes and have relatively high levels of dependence on both sample size and problem dimensionality. The likely reason for this is that calculation of metrics in these two classes requires the development of regression models using the available samples. Consequently, the values of the metrics obtained are a function of sample size and dimensionality, as the development of representative regression models generally requires a larger

number of samples for higher-dimensional problems. However, the degree to which this is the case is a function of the complexity and non-linearity of the required regression models. For example, as the calculation of some of these metrics is based on simple linear regression models, some of the metrics belonging to these classes have low dependence on sample size and problem dimensionality and hence belong to Cluster 1 (Table 2.4).

All of the metrics belonging to Cluster 4 are part of the curvature metric class (Table 2.4) and have a high dependence on sample size, but a low dependence on dimensionality. This is because the values of the curvature metrics are based on the first-order derivative and Hessian matrix of each sample point. Consequently, calculation of these metric values is a function of individual samples without considering the spatial dependencies of their relationships, and is therefore the likely reason they are not affected significantly by dimensionality. In contrast, sample size has a significant impact on curvature metric values. This is because sample points in different regions of the search space are likely to provide different gradient information, resulting in high variability unless the sample size is sufficient. This issue is likely to be exacerbated for some curvature metrics (metrics related to C_G and C_H in Eq. (C9) and (C12), respectively) that rely on information about relative gradients, especially in flat regions of the search space, as this is likely to result in infinite values. Consequently, these curvature metrics are the ones that result in high reject rates, even for low-dimensional problems and large sample sizes (e.g. Figure 2.3(g) and 2.3(h)), and have therefore been excluded from the clusters

in Figure 2.2.

Cluster 5 consists of metrics belonging to the local search class, which have low dependence on sample size but high dependence on problem dimensionality. This is because these metrics are related to the size of the local basins within the search space, which is calculated based on the number of local optima identified in each basin. As problem dimensionality increases, the number of basins grows dramatically, making it virtually impossible to identify more than one local optimum in each basin. As a result, the values of the metrics become meaningless, even for relatively large sample sizes.

Table 2.4 Metrics in Each Class in Different Clusters

Metric Class (total size)	No. of Metrics in the Cluster						% of Metric Class in Cluster 1
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Excluded Metrics	
Convexity (4)	4	0	0	0	0	0	100
Y-Distribution (3)	2	1	0	0	0	0	66.7
Level Set (18)	1	8	9	0	0	0	5.6
Meta Model (9)	6	2	1	0	0	0	66.7
ICoFS (10)	10	0	0	0	0	0	100
Curvature (24)	4	0	0	5	0	15	16.7
Local Search (21)	12	1	1	0	7	0	57.1

3.2 Validation of categorisation of fitness landscape metrics

As mentioned in Section 2.4.2, five of the 39 Cluster 1 metrics are not suitable for application to real-world environmental problems. As a result, only the 34 remaining metrics are validated using the real-world environmental modelling problems. To summarise these results, Figure 2.4 shows the number of samples

required for each of these metrics to achieve convergence for the ANNs with different numbers of hidden nodes. As discussed in Section 2.4.2, convergence was taken as the number of samples required for metric values to be within 10% of the "true" value obtained for the maximum number of samples considered.

As can be seen from Figure 2.4, 28 of these 34 metrics (82.4%) converge within 2,000 samples, which is typically within 4% of the number of samples used to generate the "true" metric values (i.e. 50,000 for most metrics – see Figure A.1). For the vast majority of these metrics (22), converge occurs within 500 samples, which is typically within 1% of the number of samples used to generate the "true" metric values. The results in Figure 2.4 clearly illustrate that there is no increase in the number of samples required for convergence with an increase in problem dimensionality (i.e. the number of hidden nodes). Consequently, these 28 metrics can be considered to have low dependence on sample size and problem dimensionality from a practical perspective, indicating that they are likely to provide a computationally efficient means for better understanding the fitness landscapes of a range of complex, highly-dimensional, real-world environmental optimisation problems. It should be noted that for some of the metrics, there is a slight increase in the number of samples required for convergence for lower problem dimensionalities (i.e. smaller number of hidden nodes). However, these variations are very small (i.e. on the order of hundreds of samples), compared with the 50,000 samples used to obtain the "true" metric values.

The six metrics that showed low dependence on sample size and problem dimensionality for the benchmark problems, but not for the real-world environmental problems, include four metrics belonging to the meta model class, one belonging to the ICoFS class and one belonging to the y-distribution class. These four meta model metrics all utilise single regression models that do not consider interactions between parameters. As interactions between parameters are likely to be a feature of real-world environmental modelling problems, the single regression models used in the four metrics in question are unlikely to represent the fitness landscapes of the environmental modelling problems considered. Consequently, these metric values are likely to become non-informative, requiring a larger number of samples for accurate calculation. In contrast, the lack of rapid convergence of the y-distribution (i.e. kurtosis) and ICoFS (i.e. ϵ_s) metrics that appear to not be suitable for real-world environmental modelling problems is likely to be related to the scaling of fitness values for the problems considered, suggesting that the scaling of fitness values is more difficult to recognise by using samples from real-world problems than test functions. It is likely that different samples provide fitness values with different scaling, making these two metric values unstable for environmental modelling problems.

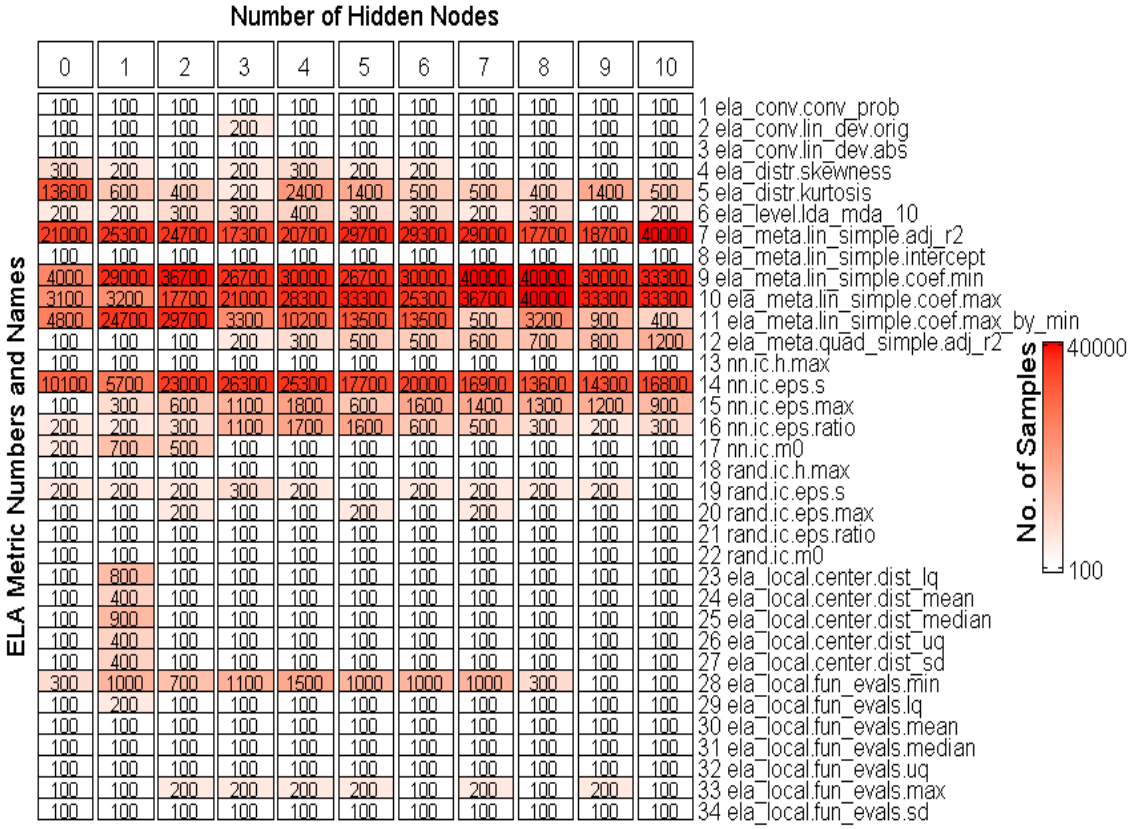


Figure 2.4 No. of Simulations until Values are Within 10% of the Values at 50,000 Simulations

3.3 Interpretation of metrics with low dependence on sample size and dimensionality

Based on the results presented in Sections 3.1 and 3.2, there are 28 ELA metrics that appear to be suitable for application to real-world environmental optimisation problems, as they have been shown to have low dependence on sample size and problem dimensionality for a wide range of benchmark and real-world problems. However, in addition to their computational tractability, the usefulness of these ELA metrics is also a function of the type of information they can provide about the different features of the fitness landscapes of environmental optimisation problems.

As shown in Table 2.5, the 28 suitable metrics cover six metric classes, excluding curvature metrics, and provide information on six of the eight major fitness landscape features. This provides the opportunity to obtain a better understanding of different attributes of a range of environmental optimisation problems in a computationally efficient manner. For example, application of metrics such as *{nn, rand}.ic.eps.max* and *{nn, rand}.ic.eps.ratio* can provide information on the potential identifiability of environmental models (Shin et al., 2013; Gupta et al., 2006) that can complement information provided by more commonly used sensitivity analysis approaches (e.g. Razavi and Gupta, 2015; Guillaume et al., 2019; Razavi et al., 2021). Specifically, information about the magnitude of multimodality and plateaus of fitness landscapes can provide information on the size of regions with non-unique parameters and information about search space homogeneity can provide insight into the distribution and location of these regions.

Alternatively, ELA metrics can provide insight into which optimisation algorithm or optimisation algorithm parameterisation are most appropriate for a given problem (Maier et al., 2014; Gibbs et al., 2011; 2015). For example, *{nn, rand}.ic.h.max*, *ela_local.center.dist_mean* and *ela_local.fun_evals.median* can be used to obtain information about the degree of multimodality, distribution of optima regions and overall depth of optima region of the fitness landscape in a computationally efficient manner prior to the optimisation process. If the fitness landscape is found to have low multimodality and the optima regions are shallow and converged to a small

area on the fitness landscape, use of a gradient-based optimisation approach might be most appropriate (Maier et al., 2019). In contrast, if the fitness landscape is found to have high multimodality, with deep and widely distributed optima regions, use of global search evolutionary algorithms might be preferred (Maier et al., 2019). The degree of multimodality and the presence of plateaus is also able to inform which values of the parameters that control the searching behaviour of evolutionary algorithms are most appropriate (see Munoz and Smith-Miles, 2017; Wang et al., 2020; Zecchin et al., 2005; 2012) and the degree of homogeneity of the search space is able to assist with determining whether there is value is adapting the values of the parameters that control the searching behaviour of evolutionary algorithms (e.g. Zheng et al., 2017).

Table 2.5 Features Represented by Suitable Metrics

Metric	Metric Class	Required No. of Samples	Feature Number (* Feature Names See Footnotes)							
			1	2	3	4	5	6	7	8
ela_conv.conv_prob	Convexity	<500	√	√			√			
ela_conv.lin_dev.orig	Convexity	<500	√							
ela_conv.lin_dev.abs	Convexity	<500	√							
ela_distr.skewness	Y-Distribution	<500	√							
ela_level.lda_mda_10	Level Set	<500	√	√						
ela_meta.lin_simple.intercept	Meta Model	<500						√		
ela_meta.quad_simple.adj_r2	Meta Model	<2000			√					
nn.ic.h.max	ICoFS	<500		√						
nn.ic.eps.max	ICoFS	<2000	√					√		
nn.ic.eps.ratio	ICoFS	<2000	√					√		
nn.ic.m0	ICoFS	<500		√						
rand.ic.h.max	ICoFS	<500		√						
rand.ic.eps.s	ICoFS	<500	√					√		
rand.ic.eps.max	ICoFS	<500	√					√		
rand.ic.eps.ratio	ICoFS	<500	√					√		
rand.ic.m0	ICoFS	<500		√						

ela_local.center.dist_lq	Local Search	<1000		√			√			
ela_local.center.dist_mean	Local Search	<500		√			√			
ela_local.center.dist_median	Local Search	<1000		√			√			
ela_local.center.dist_uq	Local Search	<500		√			√			
ela_local.center.dist_sd	Local Search	<500		√			√			
ela_local.fun_evals.min	Local Search	<2000								√
ela_local.fun_evals.lq	Local Search	<500								√
ela_local.fun_evals.mean	Local Search	<500								√
ela_local.fun_evals.median	Local Search	<500								√
ela_local.fun_evals.uq	Local Search	<500								√
ela_local.fun_evals.max	Local Search	<500								√
ela_local.fun_evals.sd	Local Search	<500								√

***1 - Global Structure 2 – Multimodality 3 – Separability 4 – Global to Local Optima Contrast 5 – Search Space Homogeneity 6 – Plateaus 7 – Variable Scaling 8 – Basin Size Homogeneity**

4 Summary and Conclusions

Optimisation algorithms are used extensively for the development of environmental models and the identification of solutions to environmental problems. How well a particular algorithm performs on a given problem is a function of both algorithm behaviour and the characteristics of the problem being solved, as represented by the fitness landscape. While significant attention has been given to the development of algorithms with different behaviours, little effort has been devoted to better understanding problem characteristics, generally resulting in a brute-force approach to identifying algorithms and parameterisations that perform acceptably for a particular problem. This is despite the fact that a number of metrics have been developed to assist with identifying features of fitness landscapes, such as their global structure, their degree of multimodality and the presence of plateaus, the identification of which would assist in the selection of appropriate optimisation algorithms and parameterisations without the need for a brute-force approach.

The primary reason for the lack of adoption of fitness landscape metrics in practice is that the calculation of these metrics is based on samples from the fitness landscape, which can be computationally expensive for real-world environmental problems, as they are often based on complex and highly-dimensional simulation models. In order to test whether this is the case, the degree of dependence on problem dimensionality and sample size of 89 fitness landscape metrics was assessed. Each metric was calculated for 72,000 different sets of fitness landscape samples obtained from 2,400 fitness landscapes derived from commonly used benchmark functions, and their degree of dependence on problem dimensionality and sample size was assessed. Results show that 39 of the 89 metrics have low dependence on dimensionality and sample size, 34 of which are considered suitable for application to environmental problems.

The low degree of dependence on problem dimensionality and sample size of these 34 metrics was tested on a number of real-world environmental modelling problems, corresponding to 7,590 sets of fitness landscape samples from 390 fitness landscapes. Results indicate that 28 of the 34 aforementioned fitness landscape metrics also have low dependence on problem dimensionality and sample size for the real-world environmental modelling problems, often requiring fewer than 500 fitness landscape samples for convergence. These 28 metrics cover a wide range of fitness landscape features, including their global structure, multimodality, separability, search space and basin size homogeneity and the presence of plateaus.

A limitation of this study is that although ELA metrics that have low dependence on problem dimensionality and sample size were identified using a large number of test functions with different dimensionalities and a wide variety of fitness landscape features, these mathematical functions are unlikely to represent all of the complexities and features of fitness landscapes associated with real-world optimisation problems. However, the fact that the majority of the ELA metrics that were found to have low dependence on sample size and problem dimensionality for the test functions also had low dependence on sample size and problem dimensionality for the real-world problems considered provides confidence in the generality of the findings presented. Another limitation of this study is that the ELA metrics can only be calculated for continuous optimisation problems, which excludes certain types of problems encountered in practice, such as the optimisation of water distribution systems using discrete pipe sizes (e.g. Zheng et al., 2017; Wang et al., 2020) and the optimisation of best-practice stormwater management options (e.g. Di Matteo et al., 2019).

The findings that there are 28 fitness landscape metrics that are able to provide insight on a range of fitness landscape characteristics that appear to be suitable for application to real-world environmental optimisation problems opens the door to gaining greater insights and improving the efficiency of a range of environmental optimisation problems. For example, these metrics can provide insight into the potential identifiability of the parameters of different environmental models, as well as information on the suitability of different

optimisation algorithms and parameterisations for particular environmental optimisation problems. Consequently, future research efforts should focus on testing the applicability of the identified metrics to a wide range of real-world optimisation problems in order to better understand the features of their fitness landscapes and to check whether the features of these landscapes identified with the aid of the ELA metrics align with those identified in previous studies, providing further confidence in the usefulness of the metrics. In addition, there would be value in applying the metrics to the fitness landscapes of the individual objective functions for multi- and many-objective optimisation problems, and to better understand the extent to which knowledge of the features of fitness landscapes can inform the selection of appropriate optimisation algorithms and their parameterisations.

Chapter 3 Impact of Model Structure on the Difficulty of Calibrating Artificial Neural Network Models

S. Zhu[1], A. C. Zecchin[1] and H. R. Maier[1].

[1]School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, SA, Australia.

(Submitted to Journal of Hydrology)

Statement of Authorship

Title of Paper	Impact of Model Structure on the Difficulty of Calibrating Artificial Neural Network Models
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Siwei Zhu
Contribution to the Paper	Primary innovator, analyst and author Experiment design and data analysis Manuscript draft and revise
Overall percentage (%)	80
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: 22/11/2021

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	Aaron C. Zecchin
Contribution to the Paper	Conception, Knowledge, Analysis, Drafting
Signature	Date: 29/11/2021

Name of Co-Author	Holger Maier
Contribution to the Paper	Conception, Knowledge, Analysis, Drafting
Signature	Date: 22/11/2021

Please cut and paste additional co-author panels here as required.

Abstract

Artificial Neural Network (ANN) models have been used for hydrological and water resources modelling for several decades, where the problem of the calibration of ANN models has drawn much attention. The corresponding literature has largely focused on obtaining an optimal calibrated model by using appropriate inputs and ANN structure and optimisation algorithms. However, the selection of appropriate models and calibration procedures is traditionally undertaken through a trial-and-error process, with little insight as to the links between the problem properties, ANN model structure and calibration difficulty. Recognizing that calibration is the process of finding the minima on an error surface, calibration difficulty can be considered as a function of the so-called *fitness landscape* properties of this surface. In this chapter, a set of fitness landscape metrics are adopted to characterise the features of the error surface of ANN models, including the global convexity structure, the surface roughness and multi-modality, and separability. A large-scale computational study is undertaken, involving the application of multi-layer perceptrons (MLPs) to a range of environmental problems, where the model structure is systematically varied in order to understand the impact of a changing MLP model structure on the properties of the calibration error surface. This work helps to explain the reasons for an increase or decrease in calibration difficulty, and in doing so, sheds new light on findings in past literature.

Highlights:

- Exploratory fitness landscape metrics can be used to characterise the features of the error surfaces of ANNs
- The error surfaces of simple ANNs have a more well-defined overall shape and have fewer local optima
- The error surfaces of complex ANNs are flatter and have many distributed, deep local optima
- Simple ANNs can be calibrated successfully using gradient-based methods
- Complex ANNs are best calibrated using a hybrid approach combining metaheuristics with gradient-based methods

Keywords

Artificial neural networks (ANNs), multi-layer perceptrons (MLPs), calibration (training), optimization, fitness landscape, error surface, exploratory landscape analysis (ELA)

1 Introduction

Artificial neural networks (ANNs) have been used extensively for hydrological modelling over the last several decades, including water quality forecasting (Maier and Dandy, 1999; Lischeid, 2001; Bowden et al., 2005; Cigizoglu and Kisi, 2006; Kisi, 2010; Bayram et al., 2012; Lafdani et al., 2013; Olyaie et al., 2015; Zounemat-Kermani et al., 2016; Amamra et al., 2018; Meral et al., 2018; Banadkooki et al., 2020; Kim et al., 2021), water quantity forecasting (Sajikumar and Thandaveswara, 1999; Zealand et al., 1999; Gautam et al., 2000; Kim and Barros, 2001; Sivakumar et al., 2002; Zhang and Govindaraju, 2003; Rajurkar et al., 2004; Coulibaly and Baldwin, 2005; Kingston et al., 2005; Wang et al., 2006; Yu and Liong, 2007; Chua et al., 2008; Wu et al., 2009; Adamowski and Sun, 2010; Khatibi et al., 2011; Jothiprakash and Magar, 2012; Piotrowski and Napiorkowski, 2013; He et al., 2015; Humphrey et al., 2016; Tan et al., 2018; Fathian et al., 2019; Cheng et al., 2020), water level forecasting (See and Openshaw, 1999; Phien and Kha, 2003; Pereira Filho and Dos Santos, 2006; Chau, 2006; 2007; Leahy et al., 2008; Tiwari and Chatterjee, 2010; Hajji et al., 2012; Pan et al., 2013; Nourani and Mousavi, 2016; Mukherjee and Ramachandran, 2018; Kurian et al., 2020; Xie et al., 2021), evaporation modelling (Kişi, 2006; 2013; Cobaner, 2011; Chaudhari et al., 2012; Kişi and Tombul, 2013; Feng et al., 2018; Ferreira et al., 2019; Maroufpoor et al., 2020; Nourani et al., 2020a; 2020b), the prediction of soil properties (Elshorbagyand Parasuraman, 2008; Parchami-Araghi et al., 2013; Trenouth and Gharabaghi, 2015; Zanetti et al., 2015; Zhuo et al., 2016;

Rahmati, 2017; Patrignani and Ochsner, 2018; Li et al., 2020; Jian et al., 2021), water temperature forecasting (Sahoo et al., 2009; Sabouri et al., 2013; 2016; Cole et al., 2014; DeWeber and Wagner, 2014; Graf et al., 2019), and some other applications (Xu et al., 2017; Nourani et al., 2017; Zubaidi et al., 2018; Nguyen-ky et al., 2018; Rezaali et al., 2021). As with all models, calibration is a critical component of the development of ANN models (termed "training" in the ANN literature). However, for ANN models, calibration takes on additional importance. Firstly, compared with process-driven models (see Mount et al., 2016), the structure of ANN models is generally unknown and is often determined via a trial-and-error process - models with different structures are calibrated and the model structure that performs best on the calibration data (or test data if cross-validation is used) is selected (Maier et al., 2010; Wu et al., 2014). Consequently, the selected model structure is a function of the success of the calibration process, in that, if the calibration does not identify the combination of model parameter values that corresponds to the lowest error for a given model structure, the conclusions about which model structure is most appropriate can be incorrect. Secondly, given the black-box nature of ANNs, the only way to extract meaningful information about the system being modelled is via analysis of the calibrated model structure and parameters (Dimopoulos et al., 1995; Lek et al., 1996; Maier and Dandy, 1997; Kingston et al., 2006; Mount et al., 2013; Humphrey et al., 2017). Consequently, the model information obtained is a function of the success of the calibration process, and for reasons outlined above, a poorly performing calibration process can lead to inaccurate information.

Given the importance of the calibration of ANN models, it is not surprising that this issue has received significant attention in literature. An area of particular focus has been the comparison of optimisation algorithms for calibrating ANN models (Kingston et al., 2005; Wu et al., 2013), with many studies reaching contradictory conclusions about which optimisation approaches perform better, while providing limited insight into why this might be the case. For example, Piotrowski and Napiokowski (2011) compared the performance of differential evolutionary algorithm (DE), particle swarm optimisation (PSO), and differential evolution with that of global and local neighborhoods (DEGL) and Levenberg–Marquardt (LM) algorithms for calibrating ANN models, where they found that DE had very poor performance, and LM, a second-order gradient algorithm, outperformed all metaheuristics. In contrast, Maroufpoor et al. (2020) found that LM often prematurely converged to local optima, and grey wolf optimisation (GWO), a metaheuristic, could obtain better solutions than LM. Second-order gradient algorithms were also found to perform worse than first-order gradient methods by Maier and Dandy (1999).

In order to better understand potential reasons for the contrasting findings in the studies outlined above, it is important to recognize that the success of a particular calibration approach is not only a function which algorithm is used, but also a function of the difficulty of the calibration problem itself (Maier et al., 2014). Calibration problems can be represented geometrically by the error surface (otherwise known as the error function, fitness function, fitness landscape or response surface), which consists of the relationship between

different values of model parameters and the corresponding calibration errors (Maier et al., 2019). As the purpose of calibration is to identify the set of model parameters that result in the smallest calibration error (akin to finding the lowest point in the error function), calibration difficulty is a function of the features of the error surface. For example, if the error surface is smooth with a single, well-defined minimum, this optimal point is relatively easy to find. Conversely, if the error surface is "rough", that is, it possesses many minima of similar or equal value (i.e. local minima), the overall global minimum is difficult to find (Guillaume et al., 2019). Similarly, the presence of flat regions, or plateaus, in the error function generally make it less computationally efficient to guide the search towards regions with lower error.

Given that many ANN models contain a relatively large number of parameters that need to be calibrated, it is difficult to visualize their error surfaces (Kingston et al., 2005). An alternative approach to gaining insight into the features of the error surfaces of ANNs with different model structures is with the aid of exploratory landscape analysis (ELA) metrics (Mersmann et al., 2010). While these metrics have been shown to be able to identify a range of fitness landscape (error surface) features on a number of mathematical benchmark problems (Mersamann et al., 2011; Munoz et al., 2015b; Munoz and Smith-Miles, 2017), their application to high-dimensional real-world problems has been limited as their computational requirements are generally considered prohibitive. This is because these metrics are calculated based on samples from the parameter space and the number of samples required to obtain

meaningful metric values has been shown to increase exponentially with problem dimensionality, at least for some metrics (Munoz et al., 2015a). However, recently Zhu et al. (2021) identified a number of ELA metrics that have low dependence on problem dimensionality and sample size, making them computationally tractable for application to higher-dimensional problems. This opens the door to using these metrics to gain a better understanding of the features of the error surface, and hence how the calibration difficulty of ANNs is likely to change with model structure.

Consequently, the overall aim of this chapter is to assess whether ELA metrics that have low dependence on problem dimensionality and sample size are able to provide meaningful information on the potential calibration difficulty of ANN models of different complexity. The remainder of this chapter is organised as follows. Details of the methodology used to achieve the above objectives are given in Section 2, followed by the results and discussion in Section 3. A summary and conclusions are provided in Section 4.

2 Methodology

2.1 Overview

In order to better understand the impact of ANN model structure on calibration difficulty, the features of error surfaces of ANNs for a range of diverse case studies and model structures are determined with the aid of a number of ELA metrics (Figure 3.1). As can be seen, four case studies are used, including Kentucky River Catchment Rainfall-Runoff Data (USA) (hereon refer to as the

Kentucky Runoff case), Murray River Salinity Data (Australia) (Murray Salinity case), Myponga Water Distribution System Chlorine Data (Australia) (Myponga Chlorine case), and South Australian Surface Water Turbidity Data (Australia) (SA Turbidity case) (see Section 2.2 for details). These are selected as they represent a variety of hydrological modelling problems that have different numbers of inputs and data series lengths, therefore representing a diversity of error surfaces. These case studies have also been used as benchmark problems in a number of previous studies investigating the impact of different ANN model development practices (e.g. Wu et al., 2013; Humphrey et al., 2017).

Multi-layer perceptrons (MLPs) are used as the ANN model architecture, as this has been by far the dominant architecture used in previous ANN applications in hydrology and water resources (see Maier et al., 2010; Wu et al., 2014). For each case study, a single layer of hidden nodes is used. The number of hidden nodes is varied between 0 and 10, resulting in numbers of model parameters (connection and bias weights) ranging from 1 to 120. Ten stochastic replicates of each MLP model with a given number of hidden nodes (0, 1, 2, …, 10) for each of the four case studies are generated, resulting in a total of 440 error surfaces (Figure 3.1). Further details of the MLP models are given in Section 2.3.

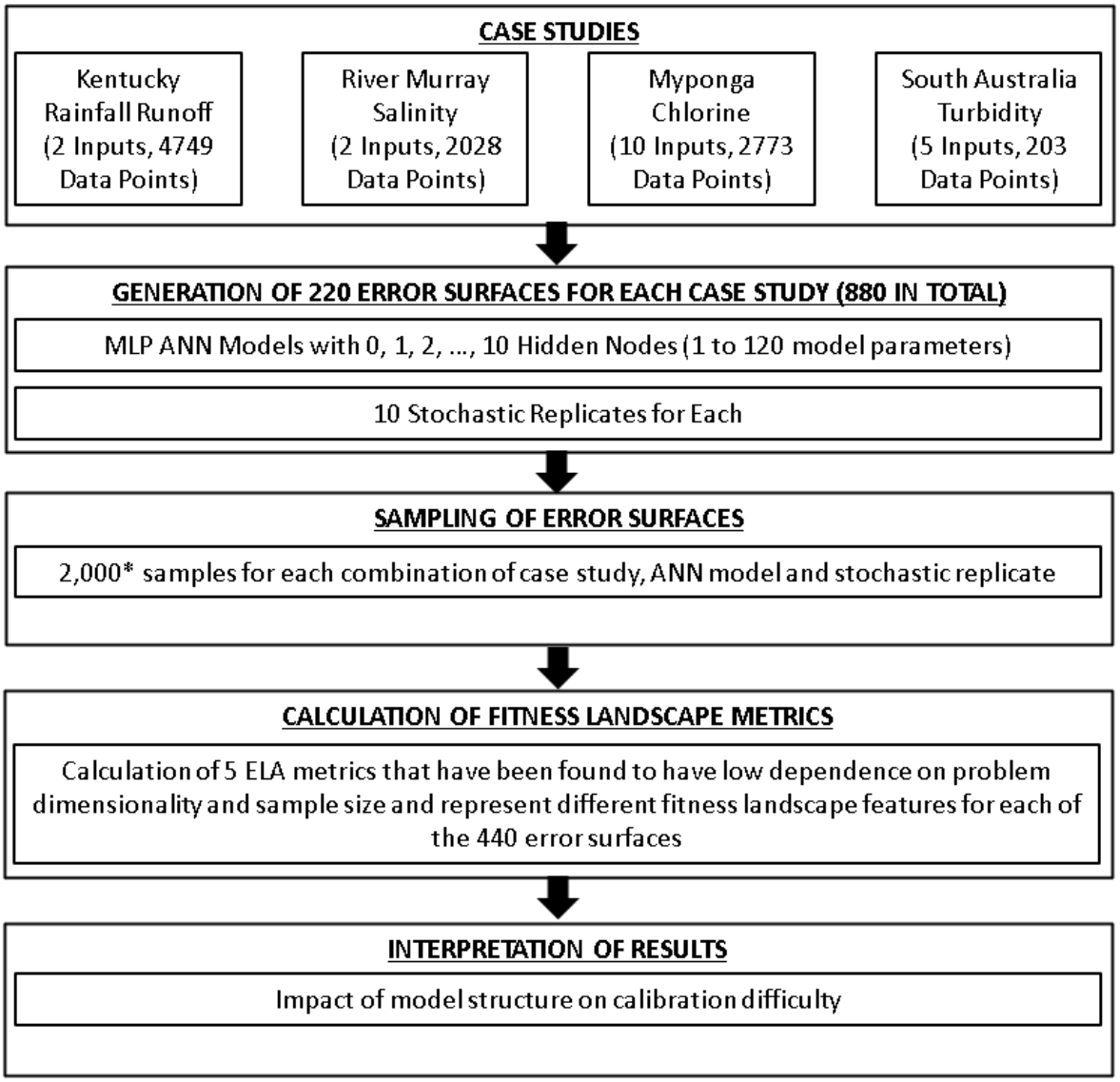


Figure 3.2 Outline of methodology

In order to enable the ELA metrics to be calculated, 2,000 samples of the error surfaces are generated for each of the 11 model structures and four case studies (Mersmann et al., 2010). This is considered adequate based on the findings of Zhu et al. (2021). These sets of samples for each of the 440 error surfaces are used to calculate a corresponding set of values for 5 ELA metrics. These metrics are selected as (i) they have been found to have low dependence on problem dimensionality and sample size (Zhu et al. 2021), thereby enabling

them to be applied to real-world problems and (ii) because they are able to identify a wide range of features of an error surface, enabling a comprehensive assessment of the impact of model structure on calibration difficulty to be performed (see Section 2.4 for details).

The above analyses were conducted using the University of Adelaide's supercomputing facilities, which consist of 48 Skylake nodes, with 80 CPUs and 377GB of memory per node. The R package ValidANN (Humphrey et al., 2017) is used for MLP development, Latin Hypercube Sampling (generated using the lhs package in R) is used for sample generation and the R package FLACOO (Kerschke and Trautmann, 2016) is used for the calculation of the 5 fitness landscape metrics.

2.2 Case Studies

As outlined above, the case studies used in this chapter are the Kentucky Runoff, Murray Salinity, Myponga Chlorine, and SA Turbidity cases. Further information on the data sets used for these case studies is provided in Table 3.1. The input selection and data splitting undertaken by Wu et al., (2013) is used, with the calibration datasets used as the basis for ELA metric calculation.

Table 3.1 Case Study Information

Case Study	Kentucky Runoff	Murray Salinity	Myponga Chlorine	South Australia (SA) Turbidity
Previous Studies	Jain and Srinivasulu (2006); Wu et al. (2013)	Maier and Dandy (1996); Bowden et al. (2002); Wu et al. (2013); Humphrey et al. (2017)	Bowden et al. (2006); May et al. (2008); Wu et al. (2013)	Maier et al. (2004); Wu et al. (2013); Humphrey et al. (2017)
Total Available Data Points	4749	2028	2773	203
Potential No. of Inputs	20	416	384	6
Selected No. of Inputs	2	2	10	5
Model Output	Flow (t+1)	Salinity (t+14)	Chlorine (t+24)	Filtered Water Turbidity
Calibration Data Points	2842	1215	1662	120

2.3 ANN Models and Data Transformations

The inputs and outputs of all cases used in this chapter are summarised in Table 3.2. Inputs X are transformed to the standard normal distribution ($X' \sim N(0,1)$) and outputs are scaled linearly between 0.1 and 0.9, in order to be aligned with Humphrey et al., (2017). The hyperbolic tangent function is used as the MLP transfer function between input nodes and hidden nodes, and a linear function is used between hidden nodes and output nodes.

Table 3.2 Selected Inputs and Outputs of Four Data Sets

Case Study	Inputs	Input Lags	Output(s)	Output Lags
Kentucky Runoff	Flow	t, t - 1	Flow	t + 1
Murray Salinity	Mannum salinity	t	Murray Bridge salinity	t + 14
	Waikerie salinity	t		
Myponga Chlorine	Myponga WTP chlorine	t	Aldinga chlorine	t + 24
	Myponga tank chlorine	t, t - 17		
	Cactus Canyon Temperature	t - 13		
	Aldinga chlorine	t, t – 1, t – 3, t – 24, t – 27, t - 47		
South Australian Turbidity	Raw Water Turbidity	-	Filtered water turbidity	-
	Raw Water pH	-		
	Raw Water Colour	-		
	Raw Water UVA	-		
	Alum dose	-		

2.4 Fitness Landscape Metrics

As shown in Figure 3.1, 5 ELA metrics, found to have low dependence on dimensionality and sample size by Zhu et al., (2021), are used to analyze the features of error surfaces. A total of 5 different features described in previous studies (Mersmann et al., 2010; Munoz et al., 2015a) are determined by these metrics. Details of these metrics and features are presented below.

Global Structure, and the mean pairwise convexity deviation

The global structure refers to the general shape of the error surface. Global optima of error surfaces with a more well-defined global structure (e.g. a “big-bowl” shape) are easier to find, as such error surfaces are able to guide

optimisation algorithms into promising regions of the search space. In contrast, the global optimum is more difficult to find for error surfaces with a less well-defined global structure (e.g. a "flat" landscape or a landscape with slopes that change direction frequently or lead to different regions), as there is little consistent information to guide optimisation algorithms towards this optimum. The *mean pairwise convexity deviation* metric (Mersmann et al., 2010) calculates the convexity of the error surface based on pairs of samples, where convexity is measured with respect to the line between the two points. Convexity is related to the global shape of the error surface. Typically, a surface with positive convexity refers to a well-defined global structure, which makes it easier to find global optima as outlined above. Non-positive convexity, on the other hand, can make the search process more difficult, as in this case the gradient information provides little guidance to the search.

Multimodality and the maximum entropy of information content

The landscape feature multimodality refers to the number of local optima on the error surface, which is also highly correlated with the degree of "roughness" of the error surface. Error surfaces with higher degrees of multimodality have a higher density of local optima, making it more difficult to find the global optimum. In contrast, error surfaces with a lower multimodality have a lower density of local optima, making it easier to identify the global optimum.

Multi-modality can be measured using the *maximum entropy of information content* metric (H_{max}) (Munoz et al., 2015a). This metric builds a ternary sequence based on the fitness values of a sequence of samples, where values of

"1", "-1" and "0" are used in the sequence to refer to fitness values of a sample that is bigger, smaller and equal to that of the following sample. The sequence of a rough surface (a high multi-modal surface) will involve frequent changes in number. In contrast, a smooth surface (an error surface with low multimodality) will have a relatively consistent sequence. The maximum entropy of the sequence is calculated to characterise the frequency of change in the sequences.

Plateaus and the epsilon of information content

Plateaus refer to regions of flatness in an error surface. Searching on error surfaces with more plateaus is generally less computationally efficient, as such error surfaces contain regions where there are minimal differences in function values, providing less distinct information to guide optimisation algorithms into promising areas. In contrast, error surfaces with fewer plateaus generally make searching more computationally efficient, as they provide useful information more consistently throughout the landscape.

The *epsilon of information content* metric (Munoz et al., 2015a) utilizes the same sample sequence as H_{max} to characterise the plateaus. A tolerance value (ε) is assigned for comparison of whether the fitness values of two neighboring samples are to be considered as equal. The corresponding ternary sequence is generated as for H_{max} but where the strict equality for label "0" is replaced by the ε interval about the given sample value. The epsilon of information content metric value is the value of ε that returns a sequence completely of the label

"0". A relatively flat surface will return a very small ε value, whereas a highly variable surface will return a large ε value. The logarithm of the ε values is used for result presentation.

Basin size homogeneity and associated metrics

Basin Size homogeneity is associated with the properties of local optima on the error surface. This feature refers to both the distribution of local optima and the difficulty in finding local optima. Firstly, with regard to the optima distribution, if local optima are contained within a small sub-region of the entire parameter space, it is easier for an algorithm to find the global optimum, or the near global optimum region. If local optima are spread throughout the error surface, it can be more difficult for an algorithm to find the global optimal region. Secondly, difficulty in finding local optima refers to the number evaluations required by a gradient based heuristic to find the local optima from random initial start points. This is related to the efficiency of finding local optima. Error surfaces that require a large number of evaluations to find the local optima are considered hard to calibrate.

The *median basin centroidal distance* is a metric that assesses the distribution of local basins (containing local optima) on the error surface. The metric finds a large pre-specified number of local optima using a gradient algorithm, and uses hierarchical clustering to collate local optima within a very small distance in the same local basin. It calculates the pairwise distance between the

identified local basins and uses the median to summarise the average distance. The second metric to characterise basin size homogeneity is the *median search function evaluations* metric, which assesses the difficulty in finding local optima. This metric refers to the median number of function evaluations required to identify each local optimum. It can indicate how extensive and complex the basin is for a given local optimum.

3 Results and Discussion

The results for each of the 5 ELA metrics are given in Sections 3.1 to 3.5, followed by a summary and discussion of the results in Section 3.6.

3.1 Mean pairwise convexity deviation

The results of the *mean pairwise convexity deviation* are shown in Figure 3.2. The results represent the deviation between the fitness value on the error surface from that of a linear regression line, so that negative results indicate that the error surface is positively convex. The results show that the error surfaces of the MLP models become more convex as the number of hidden nodes increases, except for the zero hidden node cases, where MLPs have a linear structure (i.e. the hyperbolic tangent transformation from the hidden layer is not utilized). This increase in convexity indicates a change of error surface structure as illustrated in Figure 3.2(c).

This is a key result, as it shows that models that have more hidden nodes have an advantage over models with fewer hidden nodes, as a more convex structure

can provide clearer gradient information to guide the search through the calibration process. However, the decreased calibration difficulty of models with more hidden nodes due to this increase in convexity is potentially counteracted by the increased calibration difficulty resulting from an increase in the dimension of the error surface. Nevertheless, the increase in convexity with an increase in the number of hidden nodes could explain why optimisation algorithms can still find good solutions for models with a large number of hidden nodes (and consequently parameters), even though the problems are more highly dimensional.

As the number of hidden nodes increases, so does the number of parameters, and so a similar pattern is observed in Figure 3.2(b) as in Figure 3.2(a). In interpreting this figure, it is important to note that the Kentucky Runoff and Murray Salinity cases represent smaller MLPs (with only two inputs) in comparison to the larger SA Turbidity case (5 inputs) and the largest Myponga Chlorine case (10 inputs). It is seen in Figure 3.2(b), that for a given number of parameters (i.e. error surface dimension), the smaller MLPs are more convex than the larger ones, implying that increasing the number of inputs can serve to reduce the convexity of an MLP's error surface, making it more difficult to calibrate. This highlights the potential importance of using formal input variable selection (IVS) algorithms for identifying the smallest number of inputs that have a significant impact on model performance (e.g. Galelli et al., 2014).

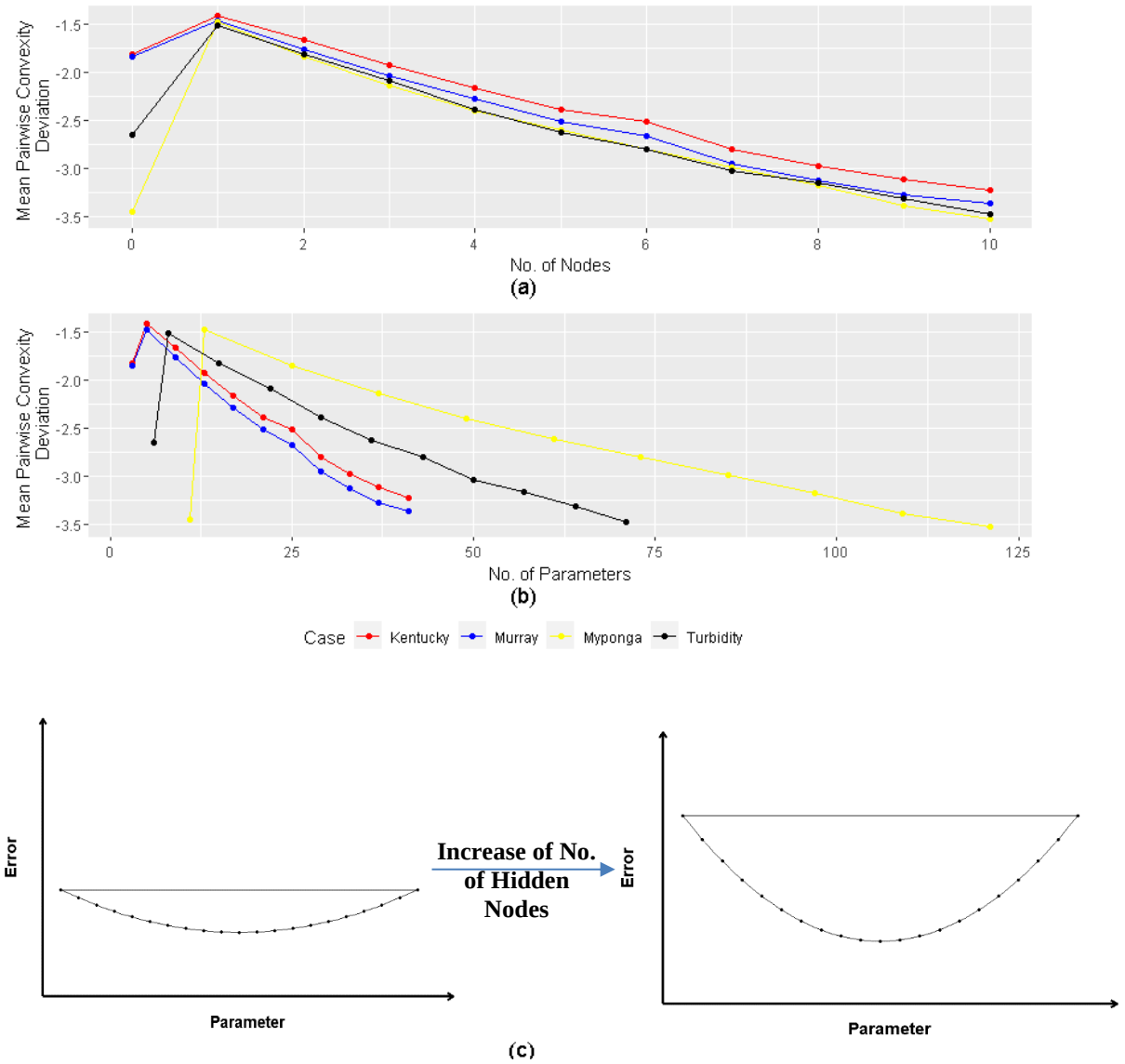


Figure 3.2 Results of *Mean Pairwise Convexity Deviation*: (a) Change through No. of Hidden Nodes; (b) Change through No. of Parameters; (c) Change of the Global Structure

3.2 Maximum entropy of information content (H_{max})

H_{max} represents how rough or how multi-modal the error surface is, where a higher H_{max} refers to a rougher error surface, and vice versa. Figure 3.3 presents the results of H_{max} versus the number of hidden nodes (Figure 3.3(a)) and the

associated number of parameters (Figure 3.3(b)) for all case studies. As shown Figure 3.3(a), H_{max} grows dramatically as the number of hidden nodes increases to larger than 3 (and the number of parameters is greater than 15). This increase in H_{max} indicates an increase in the number of oscillations in, and hence the roughness of, the error surface, as the number of hidden nodes increases, as illustrated in Figure 3.3(c). However, after this initial increase, H_{max} reaches a plateau for numbers of hidden nodes ranging from 3 to 10. This is because the error surfaces of these ANNs is already extremely rough, with H_{max} values close to their theoretical maximum of 1.0.

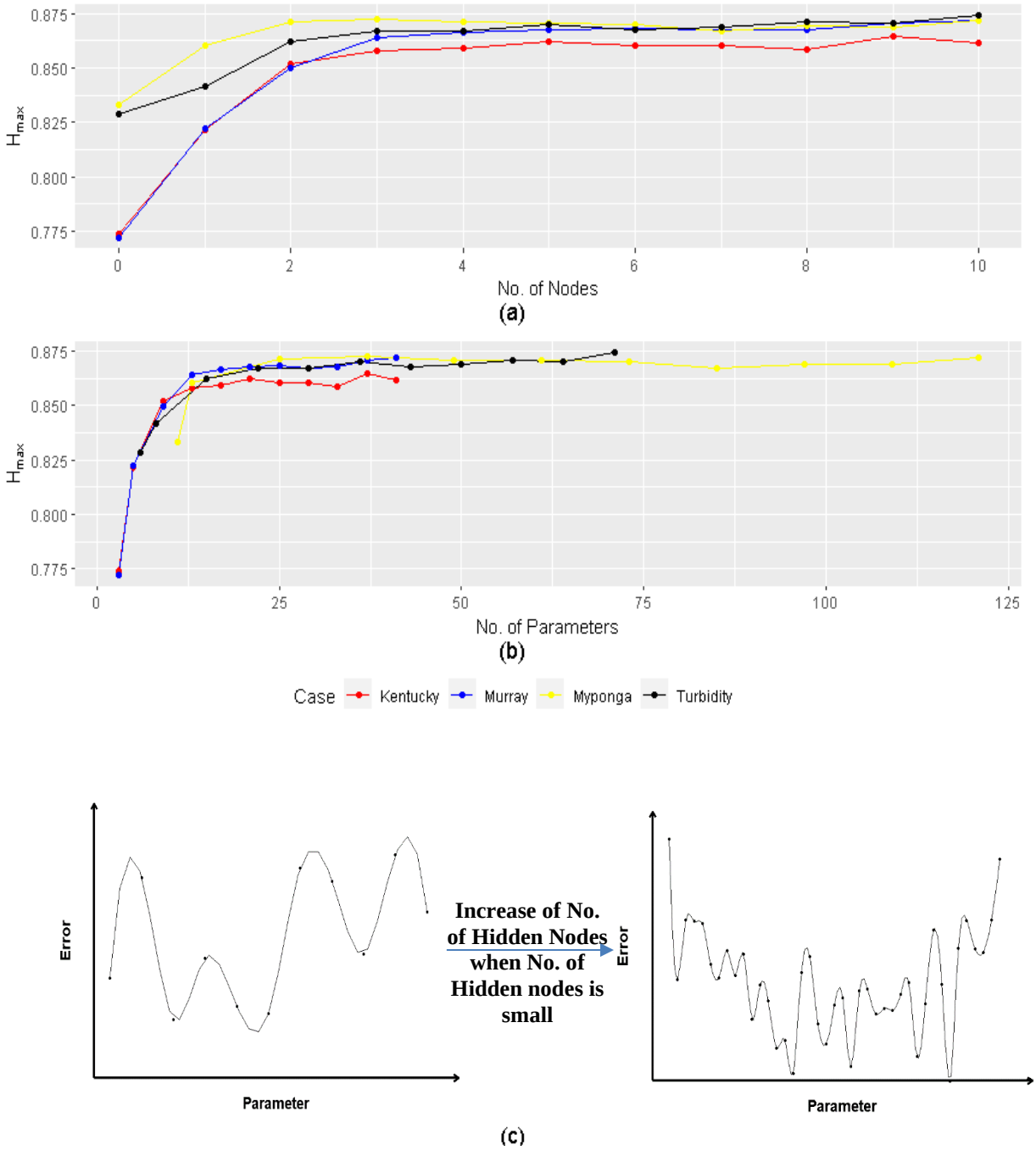


Figure 3.3 Results of the H_{max}: (a) Change through No. of Hidden Nodes; (b) Change through No. of Parameters; (c) Change of the Multimodality / Roughness

3.3 Epsilon of information content

The *epsilon of information content* represents the range of the fitness values of the error surface, where a big ε indicates a big range in fitness values and a small ε refers to a flatter error surface. Figure 3.4 shows the result of the *epsilon of information content* for all case studies. These cases show a clear trend of an increase in surface flatness with an increasing number of hidden nodes (evidenced in Figure 3.4(a)), as illustrated in Figure 3.4(c)). In this figure, it is also seen that models with more inputs have a flatter structure than models with fewer inputs (that is, compare the high input case of Myponga Chlorine with the low input cases of Kentucky Runoff and Murray Salinity). Even when the number of parameters is the same for models with a different number of inputs (i.e. consider points intersecting a vertical line in Figure 3.4(b)), models with more inputs still show a flatter surface than those with a fewer number of inputs. This again highlights the importance of the use of formal IVS algorithms, as discussed in Section 3.1.

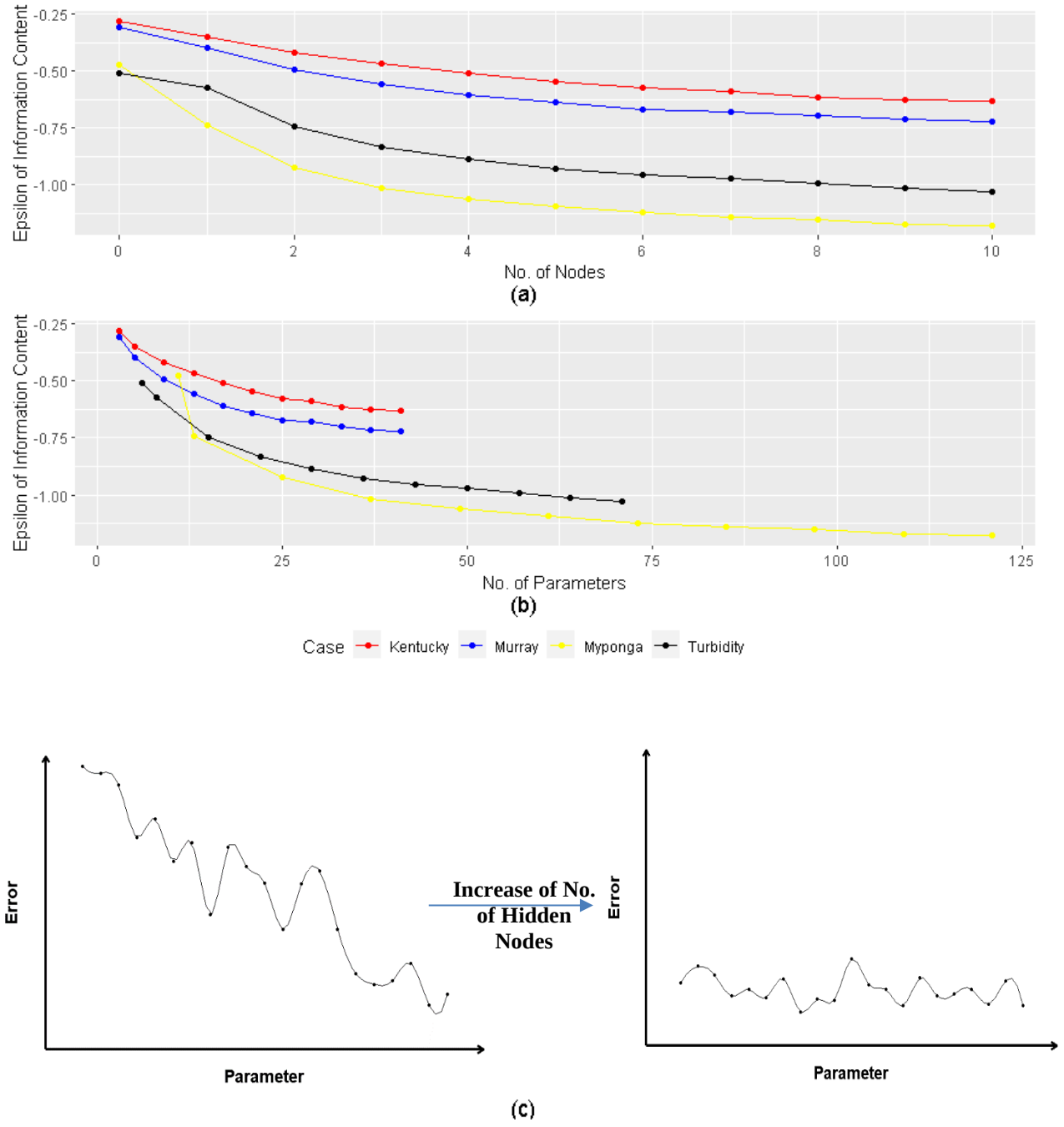


Figure 3.4 Results of the *Epsilon of Information Content*: (a) Change through No. of Hidden Nodes; (b) Change through No. of Parameters; (c) Change of the Prevalence of Plateaus

3.4 Median basin centroidal distance

Median basin centroidal distance characterises the distribution of local basins (and associated local optima) across the error surface. A bigger distance refers to a greater spread of basins, and a small distance refers to basins clustered in a small region of the error surface. Figure 3.5 presents the results of this metric for all cases. The plots for all cases start at a 0 distance, as only one optimum can be found for ANNs with no hidden nodes. However, for one or more hidden nodes, the distance increases almost linearly with an increase in the number of hidden nodes. As seen in Figure 3.5(b), there is a strong consistency across the cases of basin distance for a given number of parameters, indicating the distance is more dependent on the number of parameters, regardless of the number of hidden nodes. This aids in the interpretation of Figure 3.5(a), where it is seen that, for a given number of hidden nodes, the larger cases (Myponga Chlorine and SA Turbidity) possess a greater basin distance (i.e. more inputs increase the basin distance).

An increase in the distance between local optima is considered a disadvantage for optimisation, as this means algorithms have to explore a larger area on the error surface in order to identify the global optimum from the distributed local optima, compared with the cases where the local optima are clustered in a relatively small region (see Figure 3.5(c)). Therefore, models with many parameters require the use of optimisation algorithms with a strong exploration capacity, in order to be able to search through the entire space without missing any local optima or pre-maturely converging to sub-optimal sub-regions (see

Maier et al., 2019).

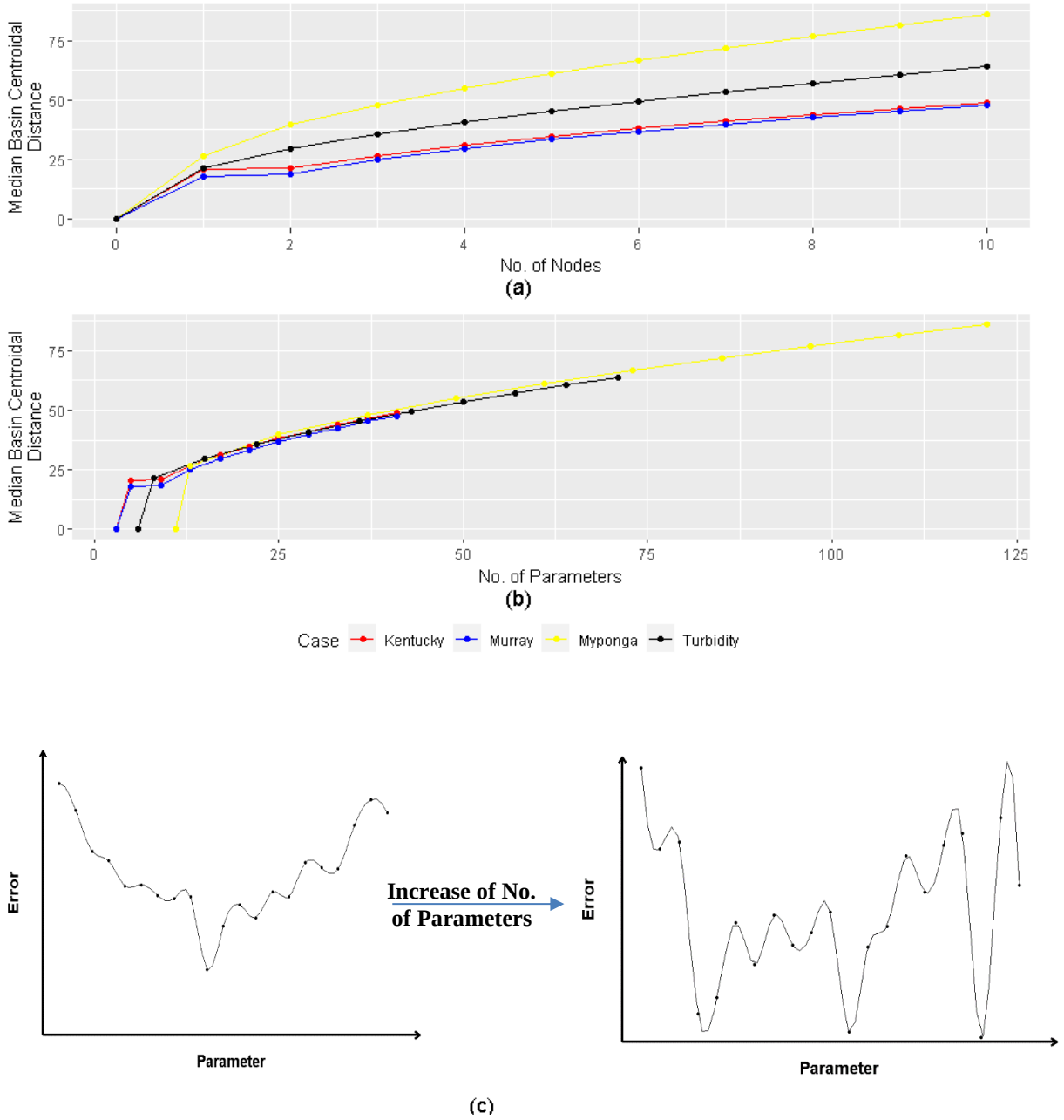


Figure 3.5 Results of *Median Basin Centroidal Distance*: (a) Change through No. of Hidden Nodes; (b) Change through No. of Parameters; (c) Change of the Distribution of Local Basins

3.5 Median search function evaluations

The *median search function evaluations* metric measures the difficulty of identifying local optima through a gradient-based local search. A larger number of evaluations indicates that the error surface is more difficult to optimise. As shown in Figure 3.6(a), as with the *median basin centroidal distance*, the number of evaluations increases near linearly for an increasing number of hidden nodes (with the larger cases requiring more evaluations). This expected result can be explained by considering Figure 3.6(b), where the function evaluations increase near linearly with the MLP parameter number, where the increase can be directly attributed to the increases in error surface dimension (i.e. higher dimensional surfaces require more evaluations). This is illustrated in Figure 3.6(c), where basins with local optima of MLPs with a smaller number of parameters are relatively small and easy to search, whereas those of MLPs with larger numbers of parameters are relatively large and spreading, resulting in difficulty in exploiting the local optima in the basin.

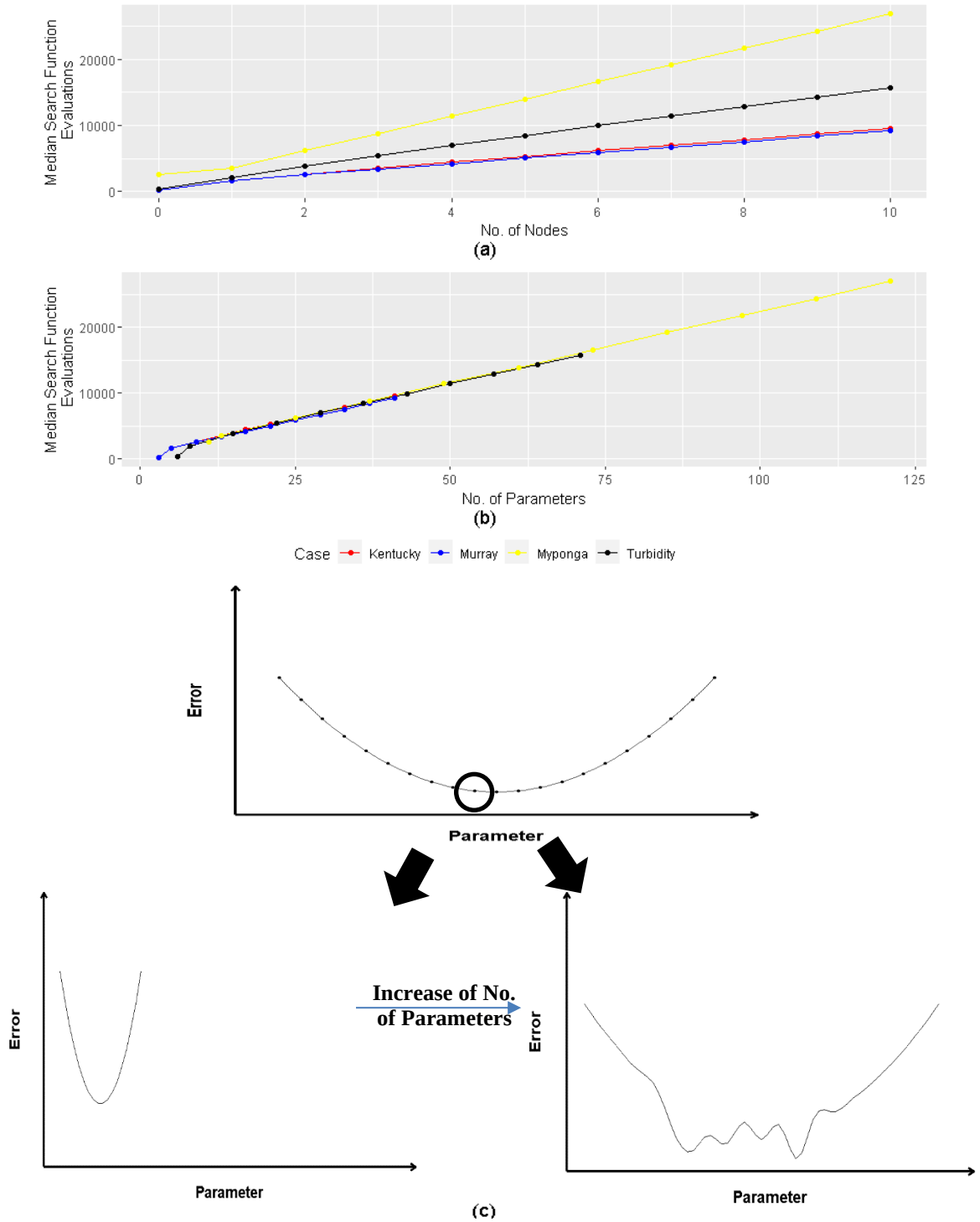


Figure 3.6 Results of *Median Search Function Evaluations*: (a) Change through No. of Hidden Nodes; (b) Change through No. of Parameters (c) Change of the Error Surface Nearby the Local Optima

3.6 Discussion

Based on the results presented in Sections 3.1 to 3.5, the error surfaces of MLP models become more complex as the number of hidden nodes increases, irrespective of case study, as illustrated conceptually in Figure 3.7. The key change is that as the number of hidden nodes increases, the general shape of the error surface structure becomes flatter overall, whilst also becoming rougher, with an increasing number of local optima with smaller and deeper basins of attraction widely spread across the entire surface. Overall, the global structure of the error surface was found to be more related to the number of hidden nodes, while the features related to multimodality and local basin distribution were found to be related to the number of parameters.

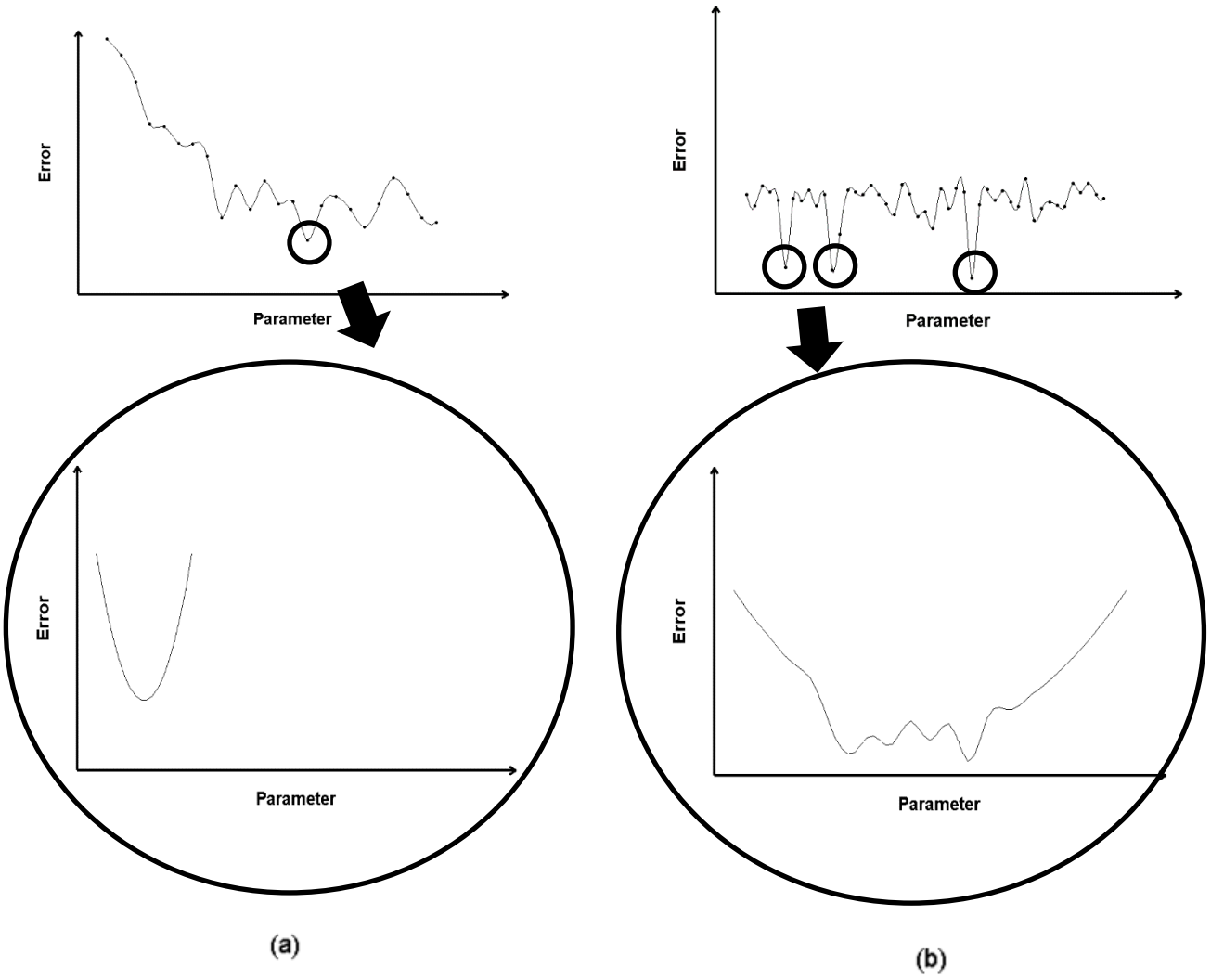


Figure 3.7 Illustration of How a 1-D Slice of the Error Surface Changes due to an Increase in the Number of Hidden Nodes / Parameters: (a) Small MLPs; (b) Large MLPs

The above findings are in agreement with the conclusions of Maier and Dandy (1998a, b), who found that, based on the results of extensive calibration trials on the Murray Salinity case study using MLPs with different numbers of inputs and hidden nodes, the error surfaces of complex MLP ANNs have large plateaus with many local optima that are deep and have steep slopes. The existence of many local optima for complex MLPs was also demonstrated by Kingston et al. (2005), who produced 3-D error surface plots of an ANN

rainfall-runoff model by fixing all connection weights to their optimal values while altering two weights at a time between a set of pre-determined limits.

The above results suggest that it is more difficult to calibrate complex ANN models when calibration methods are used that do not have the ability to explore different regions of the error surface widely. This is because more complex models have error surfaces with a larger number of local optima, in which calibration approaches with low exploratory capability, such as gradient-based optimisation algorithms, can become trapped (see Maier et al., 2019). This is why the commonly used back-propagation (BP) and Levenberg-Marquardt (LM) algorithms have been found to perform relatively poorly on more complex MLP ANNs in previous studies. For example, Maier and Dandy (1996) found that model performance decreased with an increase in the number of model parameters when the BP algorithm was used for calibrating models with different numbers of inputs and hidden nodes for the Murray Salinity case study. Similarly, Piotrowski and Napiorkowski (2011) found that the variability in calibration performance increased for calibration trials from different starting positions in model parameter space when a LM algorithm was applied to more complex ANN models, suggesting increased difficulty in finding better solutions for more complex models.

Conversely, the above results also explain why the relative performance of calibration approaches with a greater ability to explore the search space has been shown to increase for more complex ANN models, as they have a greater

ability to escape local optima and find better regions in the error surface. For example, for the Murray Salinity case study, Maier and Dandy (1998a) found that the performance of the BP algorithm increased for complex models when the exploratory ability of the algorithm was improved by increasing the step size used to explore the error surface. In addition, a number of studies have found that metaheuristics, which are known for their increased exploration ability (see Maier et al., 2019), are able to achieve better calibration performance than gradient-based methods for more complex ANNs. For example, Kingston et al. (2005) found that a Genetic Algorithm (GA) and the Complex Shuffled Complex Evolution algorithm outperformed the BP algorithm for a complex rainfall-runoff MLP, and Maroufpoor (2020) found that a Grey Wolf Optimisation algorithm outperformed the LM algorithm for calibrating complex MLPs ANNs for estimating reference evapotranspiration. However, although the greater exploratory ability of metaheuristics enables them to find better regions in complex error surfaces, because of their decreased exploitative capability, they generally have difficulty in finding the bottom of the deep, narrow local optima that are a feature of the error surfaces of complex MLP models. This explains why Piotrowski and Napiorkowski (2011) found that while the average performance of evolutionary algorithms over a number of calibration trials (from different starting positions in model parameter space) was better than that of the LM algorithm, the LM algorithm was able to find the best solutions in individual trials, provided the number of starting positions was sufficiently large. This also explains why a number of studies have found that a hybrid approach (as part of which a metaheuristic is used to find good

regions in the error surface that are then used as starting positions for gradient-based approaches) have been found to result in improved calibration performance of complex ANNs. For example, Alavi and Gandomi (2011) used simulated annealing (SA) coupled with a LM algorithm, Bahrami et al., (2016) coupled SA and GAs with a LM algorithm, and Chau (2007) used a split-step particle swarm optimisation (PSO) algorithm, which coupled standard PSO and LM algorithms. Consequently, the use of such hybrid algorithms is recommended for the calibration of complex MLP ANN models.

4 Summary and Conclusions

Calibration is an important component of the development of any model, but is especially critical for ANNs, as the quality of the calibration not only determines values of the unknown model parameters, but also the structure of the model and the degree to which underlying system knowledge can be elicited from the calibrated model. The success of model calibration is a function of how well suited the optimisation algorithm used is to exploring the error surface under consideration. While there have been many studies comparing the performance of different optimisation algorithms, existing literature has been largely silent on the properties of the error surface of ANNs with different structures, making it difficult to understand and explain why certain optimisation algorithms perform better than others, and which optimisation approaches are preferred, under particular circumstances.

This chapter has addressed this shortcoming by demonstrating that five exploratory landscape analysis (ELA) metrics that have been shown to have low dependence on problem dimensionality and sample size in previous studies can be used to better understand the features of the error surfaces of ANNs of varying complexity. Based on the results of four water quantity and quality case studies from the literature (Kentucky Runoff, Murray Salinity, Myponga Chlorine, SA Turbidity), it has been demonstrated that MLPs with a smaller number of hidden nodes and parameters are easier to calibrate, as they have a more well-defined overall shape that is able to guide optimisation algorithms to better regions in the error surface more easily. Additionally, the error surface of smaller MLPs is smoother, so that it is harder for algorithms to be trapped in local optima. In contrast, the generally flatter error surface of MLPs with more parameters and hidden nodes provides limited information to guide the search to better regions in the error surface. In addition, the higher level of multimodality / roughness of larger MLPs can also make it more difficult to identify the global optimum, especially for optimisation algorithms with limited exploration capacity, such as gradient-based methods.

On the other hand, as error surfaces of larger MLPs are more convex than those of smaller MLPs, which results in better-defined gradient information in local regions, it should be easier to converge to the local optima of larger MLPs. However, this is also likely to lead to premature convergence to local optima, rather than the identification of the global optimum. In addition, the presence of these widely distributed, narrow and deep local optima in the error surfaces

of more complex MLPs means that hybrid approaches to calibration are likely to result in better performance. This is because such approaches use algorithms with higher degrees of exploration, such as metaheuristics, in the initial stages of the calibration to find good regions in the error surface, followed by algorithms with a higher degree of exploitation, such as gradient methods, in the latter stages of calibration, to enable good locally optimal, or globally optimal, solutions to be identified.

While the findings of this study highlight the potential of using ELA metrics for better understanding the error surfaces of MLPs of different complexity for a range of case studies, thereby enabling light to be shed on the findings of previous studies, further analysis is needed to generalize the results more broadly. This would include application of the metrics to a broader range of case studies and types of ANNs. In addition, the findings of this research open the door to developing evidence-based approaches to tailoring optimisation methods and parameterisations (see Wang et al., 2020; Zheng et al., 2017) for calibrating ANN models of different types and complexity based on the knowledge of error surface features, rather than relying on a brute-force approach to using a range of optimisation approaches and picking the one that performs best for the problem at hand.

Chapter 4 Improved Understanding of Calibration Efficiency, Difficulty and Parameter Uniqueness of Conceptual Rainfall Runoff Models using Fitness Landscape Metrics

S. Zhu[1], H. R. Maier[1], A. C. Zecchin[1]. and M. A. Thyer[1]

[1]School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, SA, Australia

Statement of Authorship

Title of Paper	Improved Understanding of Calibration Efficiency, Difficulty and Parameter Uniqueness of Conceptual Rainfall Runoff Models using Fitness Landscape Metrics
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Siwei Zhu		
Contribution to the Paper	Primary innovator, analyst and author Experiment design and data analysis Manuscript draft and revise		
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	22/11/2021

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	Conception, Knowledge, Analysis, Drafting		
Signature		Date	22/11/2021

Name of Co-Author	Aaron C. Zecchin		
Contribution to the Paper	Conception, Knowledge, Analysis, Drafting		
Signature		Date	29/11/2021

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Mark Thyer
Contribution to the Paper	Framing, Conceptualisation, Drafting
Signature	Date 29/11/2021

Please cut and paste additional co-author panels here as required.

.

Abstract

The ease and efficiency with which conceptual rainfall runoff (CRR) models can be calibrated, as well as issues related to the uniqueness of their parameters, has received significant attention in literature. While several studies have tried to gain a better understanding of the underlying factors affecting these issues by examining the features of the model error surfaces, this has generally been done in an ad-hoc fashion using lower-dimensional representations of higher-dimensional surfaces. In this chapter, it is suggested that exploratory landscape analysis (ELA) metrics can be used to quantify key features of the error surfaces of CRR models, including their roughness and flatness, as well as their degree of optima dispersion. This enables key error surface features to be compared for CRR models with different combinations of attributes (e.g. model structure, catchment climate conditions, error metrics and calibration data lengths and composition) in a consistent, efficient and easily communicable fashion. Results from the application of these metrics to the error surfaces of 420 CRR models with different combinations of the above attributes indicate that increasing model complexity results in an increase in relative error surface roughness and relative optima dispersion and that, while increasing catchment wetness increases the relative roughness of error surfaces, it also decreases optima dispersion. This suggests that in this particular study, optimisation efficiency might decrease with model complexity and catchment wetness, while optimisation difficulty may increase and parameter uniqueness might

decrease with model complexity and catchment dryness. This finding does not implicate the desirability of using simpler models, but highlights the potential value of the novel way to understand the optimisation difficulty of particular hydrological models.

Highlights

- Exploratory landscape analysis (ELA) metrics are used to quantify key features of the error surfaces of conceptual rainfall runoff (CRR) models
- The features quantified include roughness, flatness and optima dispersion
- Results show increased model complexity increases error surface roughness and optima dispersion
- Results show that increasing catchment wetness increases error surface roughness and decreases optima dispersion
- Results suggest that optimisation efficiency decreases with model complexity and catchment wetness
- Results suggest that optimisation difficulty increases and parameter uniqueness decrease with model complexity and catchment dryness

Keywords:

Conceptual rainfall runoff (CRR) models, calibration, error surface, calibration efficiency, calibration difficulty, parameter uniqueness, optimisation, exploratory fitness analysis (ELA) metrics

1 Introduction

The calibration of conceptual rainfall runoff (CRR) models involves the identification of values of model parameters that enable model outputs to best match a set of measured data. While this is a conceptually simple process, it has many practical challenges, leading to the publication of a large number of papers on the topic. The vast majority of these have focused on different ways of quantifying the difference between modelled and corresponding measured outputs (e.g. which error metric to use, what model and observed data properties to use to compare model performance (e.g. data length, data splitting, missing data, types of catchments)) (Gan et al., 1997; van Griensven, 2006; Vaze et al., 2010; Fowler et al., 2016; Gibbs et al., 2018; Guo et al., 2020), different approaches to identifying the best set of model parameter values (e.g. different optimisation methods) (Duan et al., 1992; Shin et al., 2015), different model structures (Andréassian et al, 2001; Gibbs et al., 2018; Shin et al., 2015; García-Romero et al, 2019) or how to best understand and quantify uncertainties associated with the calibration process (Beven 2006; 2016; Guo et al., 2017; Kavetski et al., 2006; Renard et al., 2010).

While the above papers are based on an implicit understanding that model errors change with values of model parameters, and that automated calibration using optimisation methods corresponds to the process of finding the lowest point in this "error surface" (i.e. the *n*-dimensional surface comprised of the calibration error metric as a function of the *n* model parameters), explicit assessments of how the characteristics of this surface change as a function of

different model structures and error metrics, as well as the influence this has on the computational efficiency and difficulty of the calibration process and the uniqueness of the calibrated model parameters, have received less attention. However, explicit knowledge of the features of the error surface is required to fully interpret and understand the results of calibration trials using different model structures, error metrics, optimisation algorithms and calibration data (see Duan et al., 1992; Kavetski et al., 2007; Kavetski and Kuczera, 2007; Maier et al., 2019; Guillaume et al., 2019).

For example, if the error surface of a particular model calibration problem (e.g. combination of model structure, data and error metric) is smooth, has a clearly defined global optimum and informative gradients, the calibration process is easy (i.e. a wide range of optimisation algorithms would perform well), with a unique optimal parameter set (Maier et al., 2019). In contrast, if the error surface of a particular model calibration problem is flat at a large scale but rough at a finer scale, with many local optima that are widely dispersed, the calibration process is difficult, as it would not be easy to find the lowest point on the error surface (Maier et al., 2019). Additionally, if a number of the local optima have similar error values (or optima that are continuously distributed throughout the parameter space, as with ridges for maximisation problems), the problem of parameter non-uniqueness arises, where it is not possible to identify which set of parameter values is "best" based on the calibration error alone.
On the other hand, for different calibration problems, explicit knowledge of the relative degree of roughness and flatness of the error surfaces, as well as the

relative degree of the dispersion of the local optima across this surface, would enable generalisations to be drawn between the impact of the aforementioned attributes and the corresponding calibration performance. This would also enable the selection of the most appropriate optimisation approaches and parameterisations for the calibration problem under consideration (Maier et al., 2014; Gibbs et al., 2015).

While the vast majority of CRR calibration studies have not considered the features of the error surface, a number of studies have demonstrated that knowledge of the features of the error surface is important for explaining and interpreting the results of CRR model calibration trials (Sorooshian and Gupta, 1983; Iorgulescu and Jordan, 1994; Thyer et al., 1999; Suliman et al., 2016), the selection of appropriate model structures (e.g. Kavetski and Kuczera, 2007) and the choice of suitable optimisation algorithms (e.g. Duan et al., 1992; Kuczera, 1997; Kavetski et al., 2007). However, the above studies used ad-hoc methods for obtaining visualisations of lower-dimensional components of the error surface. In order to address this shortcoming, a number of more formal methods of visualising the error surface have been suggested. For example, Xiong and O'Connor (2000) proposed a graphical approach to describe error surfaces for high-dimensional problems, whereas Shin et al. (2015) proposed the use of dotty and eigenvalue plots and Razavi and Gupta (2015) suggest the use of sensitivity analysis to identify the features of different cross sections of the error surface.

The major shortcomings of the above approaches to identifying the features of

error surface are that they are only able to obtain lower-dimensional "slices" of higher-dimensional error surfaces and that they rely on graphical means to communicate the relevant information. This makes it difficult to compare different error surface features in an objective fashion, as it requires interpretation of graphical outputs of lower-dimensional sub-problems. In addition, it makes the quantitative comparison of the features of the error surfaces under a range of case-study attributes almost impossible.

In order to overcome the shortcomings of existing methods of characterising the features of the error surfaces of CRR models, the objectives of this chapter are to (i) propose three exploratory landscape analysis (ELA) metrics (Mersmann et al., 2010; Munoz et al., 2015b) as a means to objectively quantify key features of the error surfaces of CRR models, including relative roughness, relative flatness and relative optima dispersion, and (ii) use these three metrics to identify (a) key error surface features for different combinations of model structures, catchments, error metrics and calibration data lengths, and (b) the corresponding implications for calibration (optimisation) efficiency and difficulty, and parameter uniqueness. This opens the door to better understand the calibration performance of different CRR models under a range of conditions and to provide guidance on the selection of appropriate model structures, error metrics and optimisation algorithms.

The remainder of this chapter is organised as follows. The three metrics for characterising the relative roughness, flatness and optima dispersion of error

surfaces are introduced in Section 2, along with details of the computational experiments in which they are used to quantify these features for combinations across all case-study attributes consisting of four CRR models, five catchments in different climate zones, three error metrics and three different calibration data lengths. The results of the computational experiments are presented and discussed in Section 3, followed by summary and conclusions in Section 4.

2 Methodology

2.1 Overview

The methodology for using the proposed ELA metrics to characterise different features of the error surfaces for CRR models with different case-study attributes is shown in Figure 4.1. As can be seen, four CRR models with different structures and of varying complexity are considered. These include the Australian Water Balance Model (AWBM), which has two parameters, IHACRES-CMD (abbreviated as IHACRES in the remainder of this chapter), which has six parameters, GR4J, which has four parameters and the Sacramento model, which has 13 parameters, and were selected due to their wide use in rainfall-runoff modelling (e.g. Gichamo and Tarboton, 2019; Guo et al., 2017; Perrin et al., 2003). The models are all soil moisture accounting models with varying complexity in terms of the physical processes incorporated (See Section 2.2).

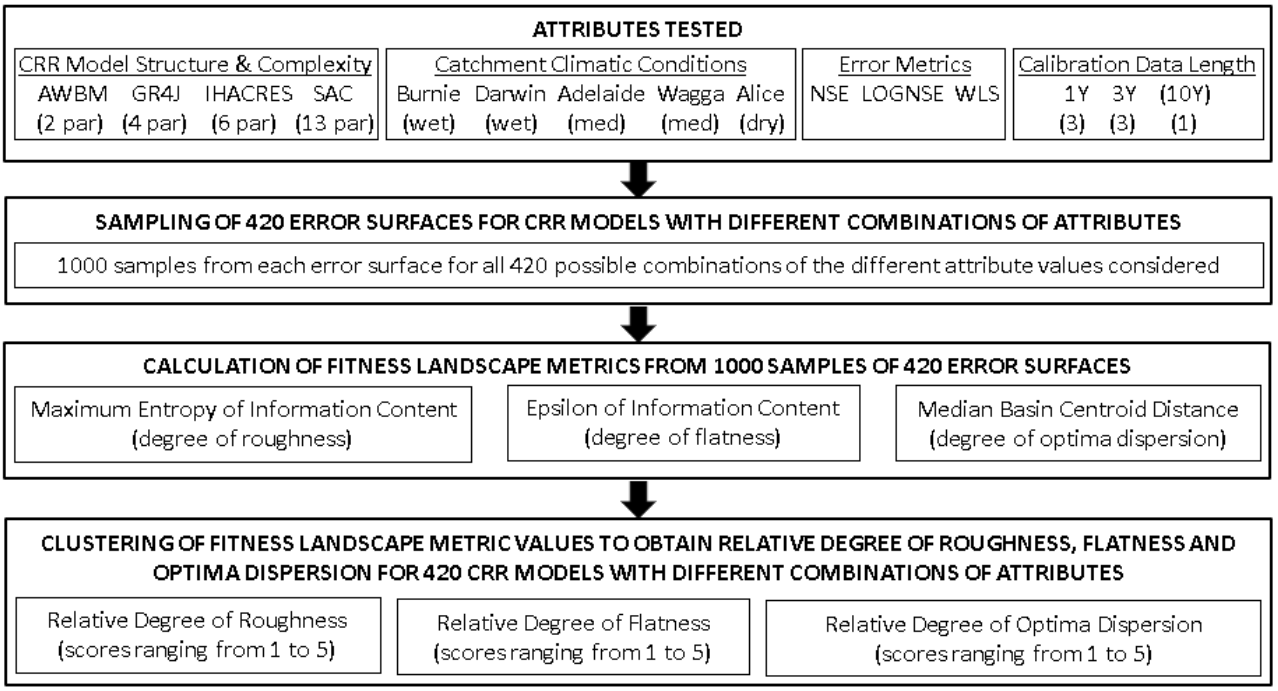


Figure 4.3 Outline of methodology

To cover a range of climate conditions, a total of five Australian catchments are considered, which are located in different climate zones and have different levels of rainfall, evapotranspiration and streamflow. The catchments are Black River (Burnie, Tasmania), which is considered a wet catchment, Elizabeth River (Darwin, Northern Territory), which is also considered a wet catchment, Scott Creek (Adelaide, South Australia), which is considered a catchment of medium wetness / dryness, Adelong Creek (Wagga Wagga, New South Wales), which is also considered a catchment of medium wetness / dryness, and Hugh River (Alice Springs, Northern Territory), which is considered a dry catchment.

With regard to the error metric attributes, three metrics are considered, namely, the Nash-Sutcliffe coefficient (NS), the NS of log-transformed flows (LOGNS) and the weighted least squared (WLS) errors of Bayesian inference (Kavetski et al., 2006). The first two metrics are commonly used in rainfall-runoff studies

for comparison in terms of their difference in sensitivity to peak and low flows (Shin et al., 2013; 2015). The WLS metric assumes errors between observed and simulated flows are normally distributed with zero mean but different variance, which is controlled by the magnitude of flows. As a result, the WLS metric contains two extra parameters to characterise heteroscedasticity (see Section 2.4), which increases the dimensionality of the error surface compared with that for the NS and LOGNS metrics.

Concerning the attribute of data length, a total of three different lengths are considered (1 year, 5 years and 10 years), with three combinations of subsets of 10 years of data used for each of the 1- and 5-year data lengths, resulting in the consideration of 7 different calibration data sets. Data length is considered an important issue affecting calibration results (see e.g. Iorgulescu and Jordan, 1994; Gan et al., 1997). However, there are also studies which have shown that data length has a very limited impact on calibration results (e.g. Zhang et al., 2015). Given this discrepancy in the literature, it is worth assessing how data length affects the features of error surfaces.

In order to enable ELA metrics to be calculated for the error surfaces resulting from the 420 CRR models with the different combinations of attributes tested (i.e. 4 CRR models × 5 catchments × 3 error metrics × 7 calibration data sets), Latin Hypercube sampling is used to generate 1000 samples for each of the 420 error surfaces based on the findings of Zhu et al. (2021) (Figure 4.1).

The generated samples are used to calculate values of three ELA metrics (Figure 4.1). These metrics relate to three features that have been shown to have an impact on CRR model calibration problems (see Duan et al., 1992; Kavetski et al., 2007; Kavetski and Kuczera, 2007; Maier et al., 2019; Guillaume et al., 2019), including roughness (fine-scale, non-smooth surface features), flatness (the distribution of error surface values) and optima dispersion (the extent of clustering of the local optima). All three metrics have been shown to be effective in identifying the above fitness landscape features for a number of benchmark functions (Mersmann et al., 2010; Munoz et al., 2015b; Munoz and Smith-Miles, 2017). Furthermore, the selected metrics have been found to have low dependence on problem dimensionality and sample size (Zhu et al., 2021), enabling them to be applied to real-world problems.

In order to enable the impact of the different attributes on the features of the error surfaces to be assessed in an easy-to-understand fashion for the large number of combinations of attributes considered, the values of the three ELA metrics for each of the 420 error surfaces are represented as five categorical values, with 1 corresponding to the most desirable value of each of the error surface features (e.g. very low roughness, very low flatness, very low optima dispersion) and 5 corresponding to the least desirable value of each of the error surface features (e.g. very high roughness, very high flatness, very high optima dispersion). This was done as the raw values of the metrics span different ranges that do not have an intrinsic meaning (see Appendix C). Consequently, the use of categorical values enables the relative values of the different metrics

to be compared in a manner that is intuitive and easy to interpret (i.e. same scale, interpretable scale where higher values are less desirable). As shown in Figure 4.1, the categorization of the raw values of the metrics was achieved by applying *K*-means clustering to values of each individual metric separately, where the value of *K* was set to 5 (i.e. the number of desired categories).

The numerical study was conducted using the University of Adelaide's supercomputing facilities, which consist of 48 Skylake nodes, with 80 CPUs and 377GB of memory per node. The R package *Hydromad* (Andrews et al., 2011) was used for model simulation, Latin Hypercube Sampling (generated using the *lhs* package in R) was used for sample generation and the R package *FLACCO* (Kerschke and Trautmann, 2016) was used for the calculation of the three ELA metrics.

Details of the CRR models, catchments, error metrics and ELA metrics used are given in Sections 2.2. 2.3. 2.4 and 2.5, respectively.

2.2 CRR models

All of the four models (AWBM, GR4J, IHACRES and Sacramento) are based on Unit Hydrograph theory (Andrew et al., 2011), and consist of a two-component structure including a soil-moisture accounting (SMA) module and a routing hydrograph module (Shown in Figure 4.2). These models require a daily time series of rainfall and potential evapotranspiration (PET) as inputs and predict daily stream flow as an output. The differences between the models

in this chapter are differences in the SMA model and which routing model is used.

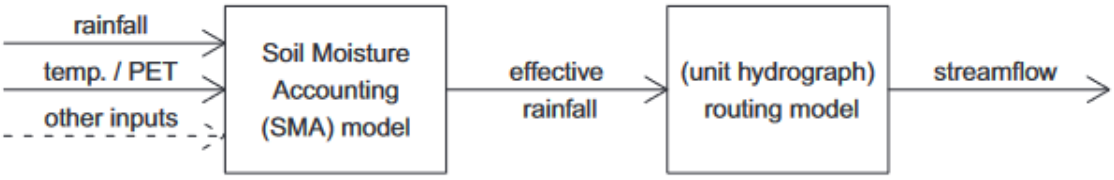


Figure 4.2 Model framework (adapted from Andrew et al., 2011)

The first model, AWBM, contains three production stores (S_1, S_2 and S_3) with different capacities (see Appendix A). On each day, precipitation P is added to each store, and effective rainfall P_r is produced when the storage exceeds the corresponding store capacity. To reduce the parameter size for the storages in AWBM from three to one, Boughton (2004) proposed an average capacity parameter to define the capacities of all three stores, at no loss of model performance. In addition to the single storage parameter, a parameter related to the multiplier of input PET is also considered. As a result, two parameters require calibration in the AWBM model, making it the simplest model considered in this study.

The second model, GR4J, contains two stores (production store and routing store) and the store water exchange in the production store is related to the degree of catchment wetness (see Appendix A). On wet days, a net precipitation P_n is produced, and a portion of P_n, denoted as P_s, fills the production store. On dry days, on the other hand, a net evaporation E_n is produced, and a portion of E_n, denoted as E_s, is extracted from the production store. The remaining component, P_r , is produced based on the updated water level in the production

store, and undergoes a further routing process based on quick flow and slow flow components. In addition to the surface water, groundwater exchange is also considered in each component, where the final streamflow is the sum of each component. A total of four parameters require calibration in the GR4J model.

In IHACRES, the level of storage is represented by the catchment moisture deficit (CMD), which is the difference between the store capacity and the current water level (see Appendix A). The CMD is used to define whether the day is dry or wet, which is related to the proportion of PET transferring to AET. In addition, CMD is also a function of the production of effective rainfall P_r from the store. The produced P_r is subjected to a routing process, which splits P_r into quick and slow flows and analyses with two different unit hydrographs with different time base. The final streamflow of IHACRES is equal to the sum of the quick and slow flows. In total, IHACRES has six parameters that require calibration.

The structure of the Sacramento model can be split into three zones, including the surface, upper and lower zones (see Appendix A). The surface zone is split into permeable, additional impermeable and impermeable areas. The net precipitation in the permeable area drains to the upper zone, while that in the impermeable area generates direct flow. The additional impermeable area generates direct flow when the catchment tension water requirements in the upper zone are met, otherwise the net precipitation also drains to the upper

zone. The upper zone consists of tension and free water. Inflow exceeding the tension water store drains as surface flow. A proportion of free water percolates to the lower zones, which also consists of tension and free water. A proportion of free water in the lower zone can also produce baseflow. As a result, the total flow at each time step produced by the Sacramento model is the sum of the flows produced in the three zones. The Sacramento model is the most complex of the models considered in this chapter, with 13 parameters that require calibration.

2.3 Catchments

Figure 4.3 shows the names and locations of the five catchments considered. As can be seen, the catchments are located in climatologically different regions in Australia, as defined in the Australian Köppen climate classifications of Stern et al. (2000). The historical data required for CRR model calibration are average daily rainfall, PET and runoff at each location. Daily rainfall data were obtained from a single rain gauge within each of the Scott Creek, Black River, Elizabeth River, and Adelong Creek catchments, while the rainfall data for the Huge River catchment were obtained by spatially averaging values from three rainfall gauges using the Thiessen polygon method. Daily runoff data were obtained from gauging stations at the outlet of each catchment. PET data were calculated using the Penman-Monteith model, which requires data on temperature, relative humidity, solar radiation and wind speed (Guo et al., 2017). These data were obtained from a weather station nearby each catchment. Table 4.1 summarises the climate characteristics of the five catchments. As can be seen, the catchments were categorized into wet, mild and dry catchments

based on their annual rainfall, PET and runoff.

The data for each catchment consist of a 10-year period from 01/01/1995 to 31/12/2004. In order to assess the impact of data length on the features of the error surface, three groups of 1-year and 5-year data (with different start and end dates) are used as inputs, in addition to the entire 10-year period.

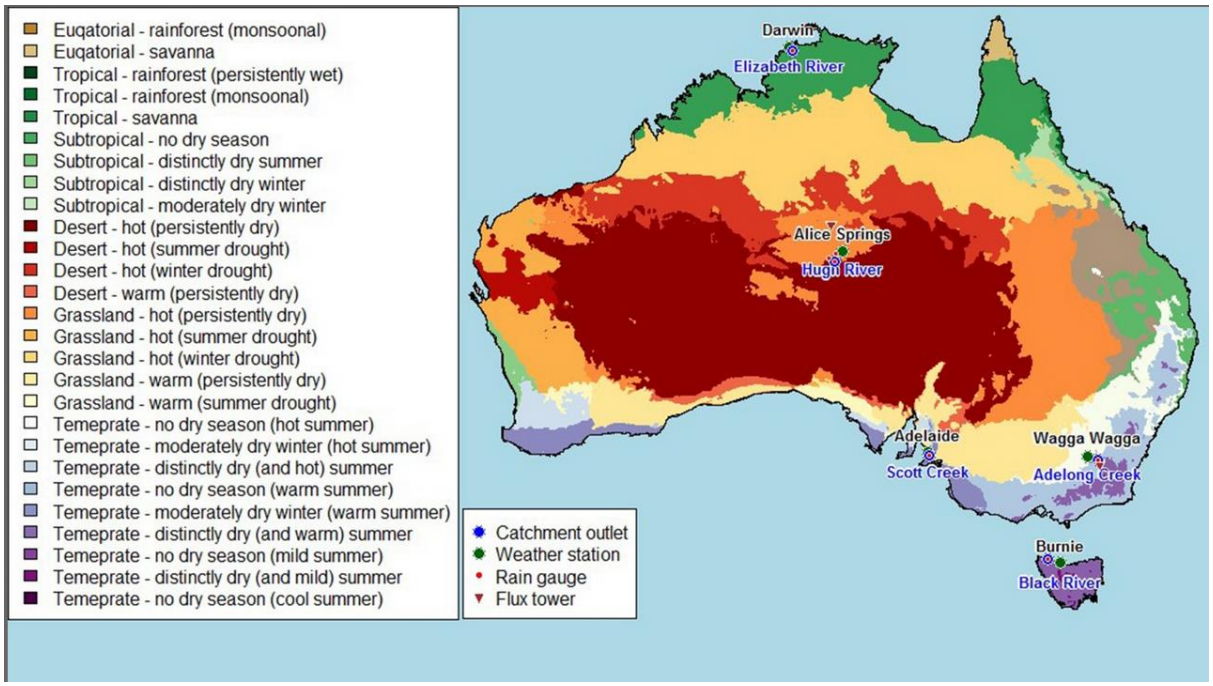


Figure 4.3 Locations of rain gauges, catchment outlets, and weather stations from which data were obtained for calibration of the rainfall-runoff models for the five catchments (adapted from Guo et al., (2017))

Table 4.1 Catchment characteristics

Catchment (and Area)	Location	Annual P (mm)	Annual Q (mm)	Annual PET (mm)	PET/P	Property
Black River (318.5 km^2)	Burnie	1182	550	958	0.81	Wet
Elizabeth River (95.6 km^2)	Darwin	1979	777	1864	0.94	Wet
Scott Creek (29 km^2)	Adelaide	892	133	1372	1.54	Mild

Adelong Creek (146.1 km^2)	Wagga Wagga	799	195	1436	1.80	Mild
Hugh River (3324 km^2)	Alice Springs	344	56.2	1822	5.29	Dry

2.4 Error metrics

As mentioned above, the NS coefficients of flows and log-transformed flows were used as two of the three error metrics considered in this study. The equations of NS and LOGNS are given as:

$$NS = 1 - \frac{\sum_{t=1}^{T}(Q_s^t - Q_o^t)^2}{\sum_{t=1}^{T}(Q_o^t - \overline{Q_o})^2} \quad (1)$$

$$LOGNS = 1 - \frac{\sum_{t=1}^{T}(\ln(Q_s^t + 1) - \ln(Q_o^t + 1))^2}{\sum_{t=1}^{T}(\ln(Q_o^t + 1) - \ln(\overline{Q_o} + 1))^2} \quad (2)$$

where Q_s^t and Q_o^t are the simulated and observed flow at time step *t*; $\overline{Q_o}$ is the average observed flow, *T* is the total number of time steps, and ln(·) is the natural logarithm, and the constant one is added to each log-transformed flow to avoid the negative infinite values caused by small and zero flows. Both error metrics have a range from negative infinity to 1, where 1 indicates a perfect fit. It can be seen that the only difference between NS and LOGNS is the log-transformation of the flows. This difference changes the error of flows from homoscedastic to heteroscedastic based on flow volume in different time steps. The impact of this transformation is to increase the influence of low flows on the metric value. For example, Shin et al., (2013) demonstrated that parameters related to the slow flow store of IHACRES became insensitive when using NS. In contrast, parameter related to the unit hydrograph in GR4J became more

sensitive in NS, but less sensitive in LOGNS, as the time base of the unit hydrograph is more sensitive to capturing the flow peak. As parameter sensitivity is highly related to features of the error surface, it is very likely that features of the error surface change when the error metric is changed.

In addition to NS and LOGNS, WLS is also included as an error metric due to its novel treatment of model errors. Unlike other error metrics, WLS also considers the heteroscedasticity of flow errors, but uses two additional parameters to estimate the variance of errors, which depends on the flow volume in each time step. The WLS metric assumes errors are normally distributed as

$$\varepsilon_t \sim N(0, a + b \cdot Q_o^t) \quad (3)$$

where ε_t is the error between observed and simulated flow at time step *t*; Q_o^t is the observed flow at time step *t*; a and b are constant parameters to describe the linear relationship between the variance of ε_t and Q_o^t. It can be seen that higher flows have a bigger variation in errors, while this variation is smaller for lower flows. As a result of this heteroscedasticity, the error surface of models using WLS as an error metric are likely to be very different in comparison to NS. Additionally, because of the two additional parameters (*a* and *b*)), it is also likely to be different in comparison to LOGNS, due to the difference with which heteroscedasticity is treated. The likelihood function is the product of the likelihood of ε_t at each time step, and is given by:

$$p(\boldsymbol{\varepsilon}) = \prod_{t=1}^{T} \boldsymbol{N}(\varepsilon_t | a, b, Q_o^t) \quad (4)$$

where *T* is the total number of time steps; and $p(\boldsymbol{\varepsilon})$ is the likelihood of errors at all time steps. For the interested reader, a detailed discussion of Bayesian inference and the WLS can be found in Renard et al. (2010).

2.5 Fitness landscape metrics

As mentioned above, three typical features (roughness, flatness and optima dispersion) are considered relevant in CRR model calibration problems (see Duan et al., 1992; Kavetski et al., 2007; Kavetski and Kuczera, 2007; Maier et al., 2019; Guillaume et al., 2019). An illustration of the relative impact of these features on the error surface of CRR models is shown in Figure 4.4. As can be seen, eight different conceptual error surfaces based on different combinations of the relative values of the features are shown.

The error surfaces on the right side ((e) to (h)) are considered to be less efficient to calibrate than those on the left side ((a) to (d)). This is generally due to the increased relative roughness of the error surfaces on the right side, which may lead to a greater number of evaluations required to converge to the global optimum, or even to a failure to identify the global optimum due to pre-mature convergence. This problem can be severe even for flat error surfaces, such as those in (f) and (h), as there is lack of information to guide the search out of the rough areas. In contrast, the calibration of error surfaces on the left side is considered relatively more efficient, as there is no roughness on the error surfaces to slow down or mislead the search. Consequently, even though error surfaces (b) and (d) are relatively flat, they are still likely to be more efficient

to optimise, unless they are completely flat, in which case there would be no gradient information to guide the search at all.

The error surfaces in (c), (d), (g) and (h) are expected to have parameter uniqueness problems, regardless of whether calibration efficiency is hindered due to the roughness of the error surface or not. This is because these error surfaces contain multiple local basins or regions of attraction with a wide distance between them, so that parameter non-uniqueness is a structural problem, rather than a problem of roughness. This is illustrated by error surfaces (c) and (d), which are smooth and are therefore expected to be more efficient to calibrate. However, they still have a problem with parameter uniqueness, as two different parameter sets have the same error metric value.

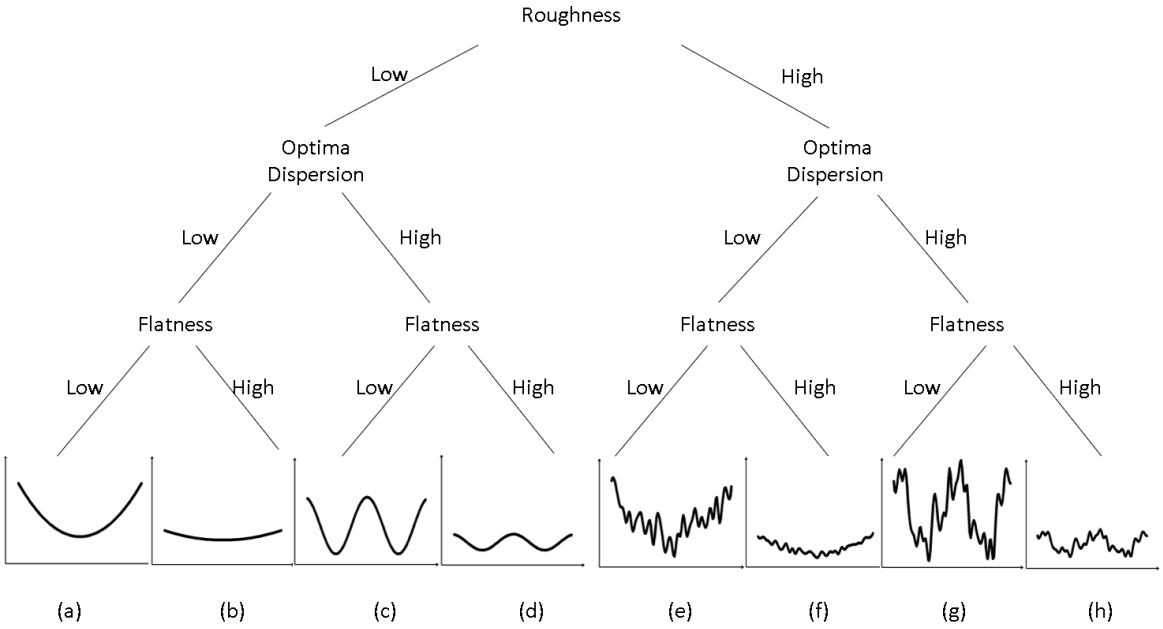


Figure 4.4 Impact of relative values of selected fitness landscape metrics on features of error surfaces of CRR models

To enable the above features (i.e. roughness, flatness and optima dispersion) to be quantified for the error surface of the models with different attributes (Figure

4.1), three ELA metrics are utilised in this chapter. These metrics have been used in previous studies for classifying the features of benchmark problems (Mersmann et al., 2010; Munoz et al., 2015b; Munoz and Smith-Miles, 2017) and have been shown to have low dependence on sample size and problem dimensionality (Zhu et al., 2021). Details of these three metrics are presented in the following subsections.

Maximum entropy of information content (Degree of Roughness)

The *maximum entropy of information content* metric (H_{max}) (Munoz et al., 2015b) is used to measure the multimodality of the error surface. The feature multimodality (Mersmann et al., 2010) refers to the number of local optima on the error surface, which is highly correlated to the roughness of the error surface. Error surfaces with higher degrees of multimodality have a higher density of local optima, which manifest themselves as rough surfaces. In contrast, error surfaces with a lower multimodality have a lower density of local optima, which are manifest as smooth surfaces.

This metric builds a ternary sequence based on the fitness values of a sequence of samples, where values of "1", "-1" and "0" are used in the sequence to refer to fitness values of a sample that is bigger, smaller and equal to that of the following sample. The sequence of a rough surface will involve frequent changes in number. In contrast, a smooth surface will have a relatively consistent sequence. The maximum entropy of the sequence is calculated to characterise the frequency of change in the sequences. The theoretical range of

this metric is [0, 1].

Epsilon of information content (Degree of Flatness)

The *epsilon of information content* metric (Munoz et al., 2015b) characterises the plateaus (Mersmann et al., 2010) of error surfaces, which refer to regions of flatness. Error surfaces with more plateaus are generally flatter, and can cause parameter identification problems and slow down the calibration process. In contrast, error surfaces with fewer plateaus have a better shape, which corresponds to more sensitive and identifiable parameters.

The metric utilises the same sample sequence as H_{max}. A tolerance value (ε) is assigned for comparison of whether the fitness values of two neighbouring samples are to be considered as equal. The corresponding ternary sequence is generated as for H_{max}, but where the strict equality for label "0" is replaced by the ε interval about the given sample value. The epsilon of information content metric value is the value of ε that returns a sequence completely of the label "0". A relatively flat surface will return a very small ε value, whereas a highly variable surface will return a large ε value. The logarithm of the ε values is used for result presentation. The theoretical range of this metric is $[-\infty, \infty]$.

Median basin centroidal distance (Degree of Optima Dispersion)

The *median basin centroidal distance* is a metric that assesses the distribution of local basins (Mersmann et al., 2010). A local basin is defined as a region on the error surface that contains multiple local optima with a very short distance

between them. This local basin conception is similar to multiple optima and regions of attraction as defined by Kavetski and Kuczera (2007) and Duan et al., (1992), respectively, for error surfaces of CRR models.

The metric finds a large pre-specified number of local optima using a gradient algorithm, and uses hierarchical clustering to collate local optima within a very small distance in the same local basin. It calculates the pairwise distance between the identified local basins and uses the median to summarize the average distance. Therefore, a long distance represents local basins or regions of attraction that are dispersed on the error surface, so that values of model parameters are more difficult to identify. In contrast, a short distance means that local basins or regions of attraction are contained within a small region on the error surface, making the parameter identification process less difficult. The theoretical range of this metric is [0, ∞].

3 Results and Discussion

3.1 Overview

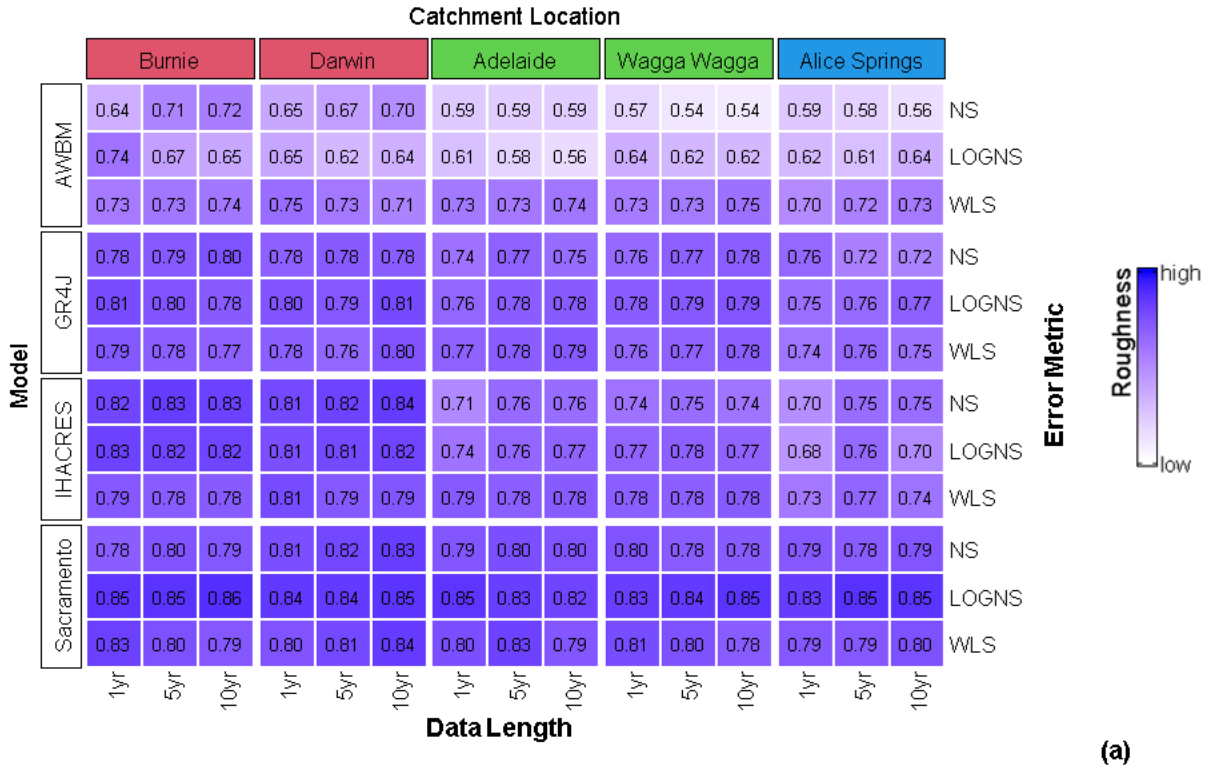
The heatmaps in Figure 4.5 indicate the relative influence of the four attributes investigated (i.e. model structure and complexity, catchment climate condition, error metric and calibration data length) on the three error surface metrics considered (i.e. roughness, flatness and optima dispersion). It should be noted that the results for the different combinations of calibration data for data lengths of 1 and 3 years have been combined, resulting in a comparison of the features of 180 (4 models × 5 catchments × 3 error functions × 3 data lengths) error functions. The colour of the heatmaps indicates the relative desirability of the

error surface characteristics based on the five categorical values for each fitness landscape metric (see Appendix B), with lighter colours (i.e. white) indicating more desirable characteristics (i.e. a lower value) and darker colours (i.e. dark blue) indicating less desirable characteristics (i.e. a higher value) (see Section 2.1). There are some cells of optima dispersion of the Sacramento model that are shown as "NAN", which indicates a failure in finding local optima for these cases. This is most probably due to the broadly acknowledged complex and poorly-behaved structure of the Sacramento model, which potentially contains numeric errors that are overwhelmed by uncertainties in the data and governing equations (Clark and Kavetski, 2010).

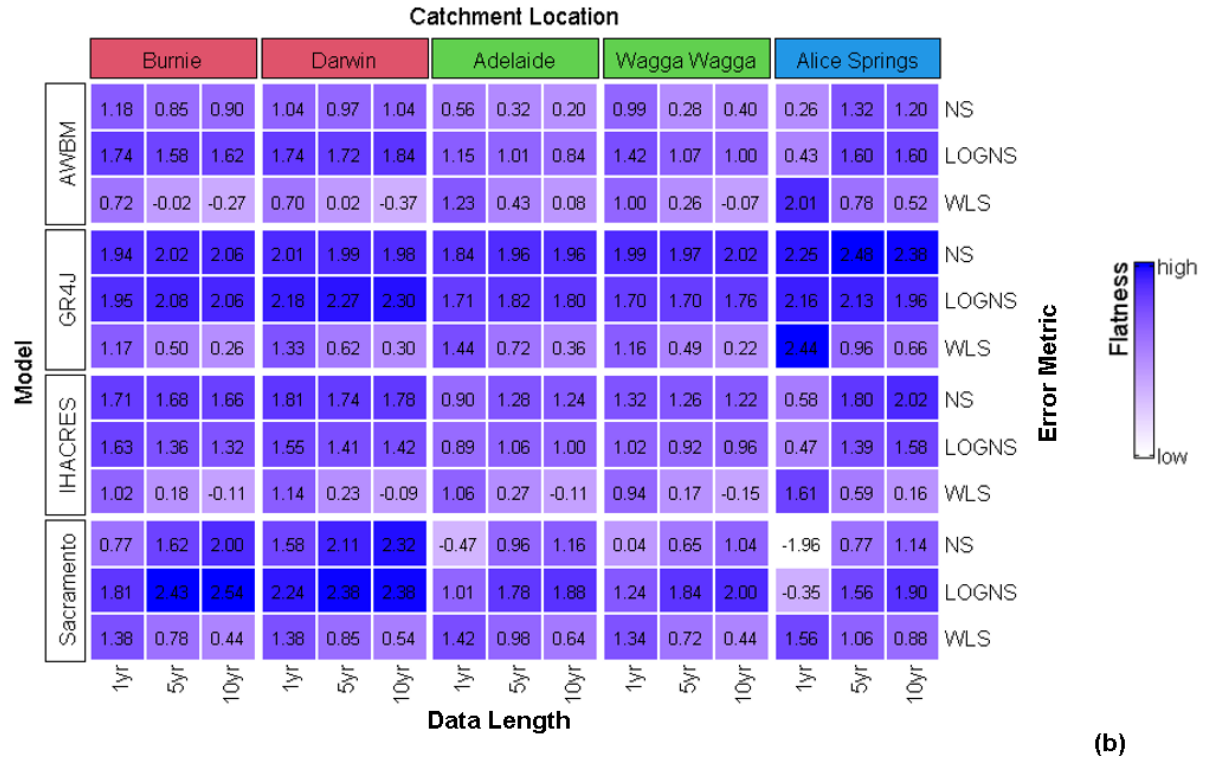
The numbers in the cells of the heatmap are the calculated ELA metric values for the given features (see Section 2.5). It should be noted that as lower values of the *epsilon of information content* indicates a higher level of relative flatness, which is opposite to H_{max} and the *median basin centroidal distance*, where larger values represent higher levels of relative roughness and optima dispersion, respectively, so that larger values of each of the three metrics correspond to less desirable fitness landscape features. It should also be noted that these results only refer to the features of the error surfaces, and not calibration performance in terms of absolute values of the error metrics. For example, a problem with higher calibration difficulty could contain lower absolute error values and vice versa.

As can be seen from a visual inspection of Figure 4.5, based on the results of the experiments presented here, model structure and complexity appear to have

the largest influence on the metrics, with an increase in relative roughness and optima dispersion with an increase in model complexity. Catchment climate condition also appears to have an influence on relative roughness and optima dispersion, with wetter catchments resulting in rougher, less dispersed error surfaces. In contrast, the impact of different error metrics on the error surface characteristics appears to be confined to particular ELA metrics, model structures and catchment climate condition. The most pronounced influence of error metrics is associated with the application of the WLS metric to simpler models, especially for wetter catchments. As can be seen from Figure 4.5, error surface roughness appears to increase when the WLS metric is used for models with a small number of parameters, such as AWBM, as use of this metric results in a significant relative increase in the number of parameters (see Section 2.4). In contrast, use of the WLS metric appears to reduce error surface flatness and optima dispersion for simpler models, especially for wetter catchments, resulting in more well-defined regions of attraction in parameter space. The results in Figure 4.5 also indicate that there appears to be no pronounced impact of calibration data length on error surface characteristics for the experiments conducted. Given that model structure and complexity, and catchment climate condition, appear to have the biggest impact on the error surface metrics considered, these are discussed in detail in the subsequent sections.


Catchment Location
Burnie
Darwin
Adelaide
Wagga Wagga
Alice Springs
Model
AWBM
GR4J
IHACRES
Sacramento
NS
LOGNS
WLS
1yr
5yr
10yr
Data Length
Error Metric
Roughness
high
low

(a)


Catchment Location
Burnie
Darwin
Adelaide
Wagga Wagga
Alice Springs
Model
AWBM
GR4J
IHACRES
Sacramento
NS
LOGNS
WLS
1yr
5yr
10yr
Data Length
Error Metric
Flatness
high
low

(b)

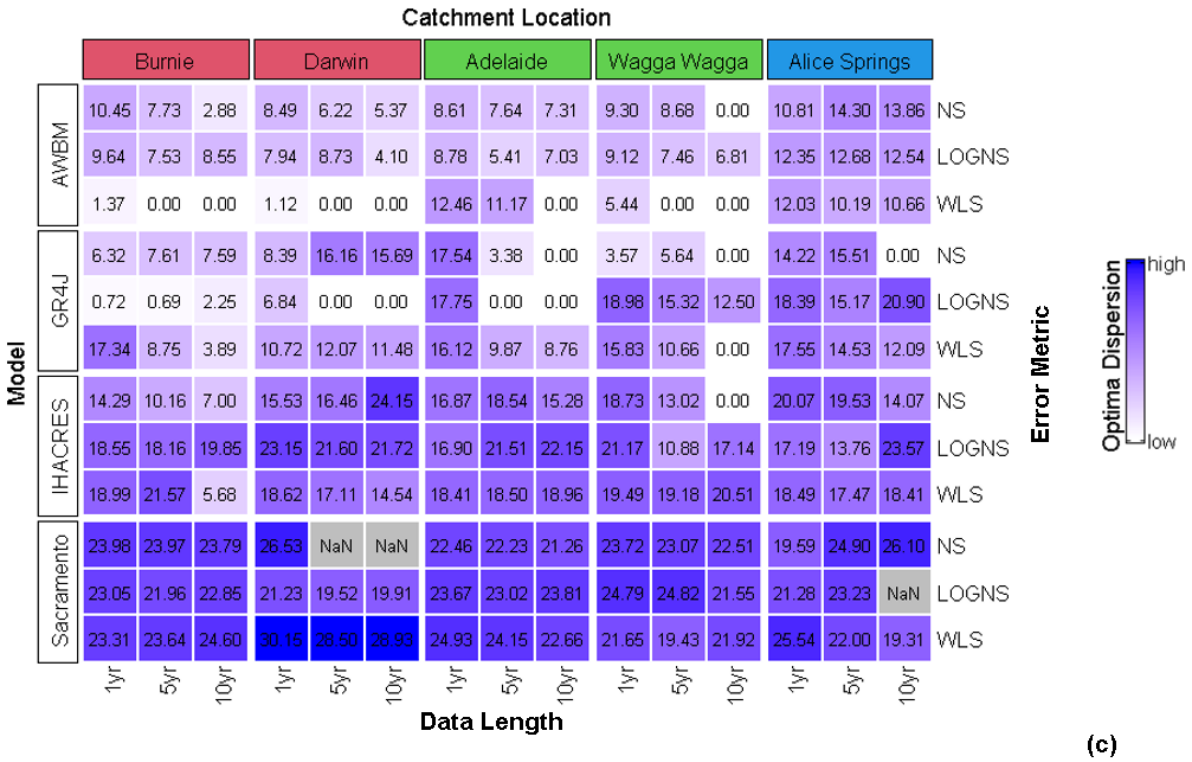


Figure 4.5 Relative impact of Roughness (a), Flatness (b) and Optima Dispersion (c). Note that the catchments are ordered from left to right by decreasing wetness (wet, mild and dry catchments are coloured as red, green and blue respectively).

3.2 Impact of model structure and complexity

The impact of model structure and complexity on relative roughness, flatness and optima dispersion for the experiments conducted is shown in Figure 4.6, where boxplots of the percentage of the categorical values (i.e. 1 to 5) for each model and each of the metrics is presented. As can be seen, there is a general increase in the relative roughness in the error surface as model complexity increases, with AWBM (2 parameters) clearly the smoothest and Sacramento (13 parameters) clearly the roughest. The roughness of GR4J (4 parameters) and IHACRES (6 parameters) is very similar and clearly in-between the roughness of AWBM and Sacramento.. The rate in relative increase in roughness tends to decrease for more complex models, most likely because the

underlying roughness of more complex models is already very high. These trends can also be observed in the raw values of the metrics (Appendix C).

The increase in the relative degree of optima dispersion with an increase in model complexity is even more pronounced than that for relative roughness, with a clear increase in the degree of dispersion of optima across the error surface as the number of model parameters increases (Figure 4.6, Appendix C). However, based on the results obtained, there appears to be no clear relationship between model complexity and the relative flatness of the error surface. Instead, this seems to be more affected by particular model structures. For example, for the experiments conducted, the relative flatness of the error surface of GR4J is much higher than that of the other models.

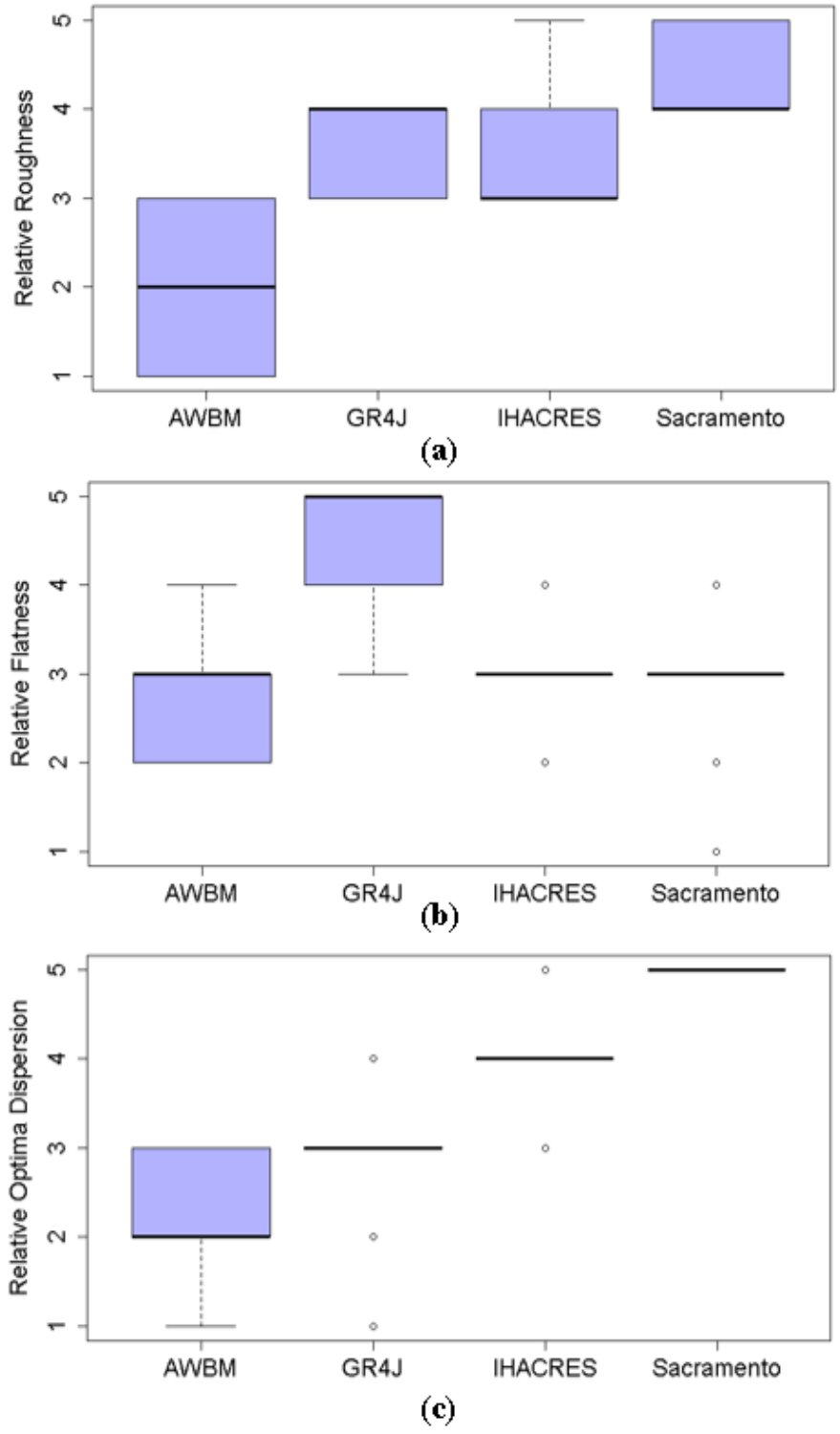


Figure 4.6 Influence of model structure / complexity on error surface features: Relative Roughness (a); Relative Flatness (b); Relative Optima Dispersion (c). The complexity of the CRR models increases from left to right: AWBM has 2 parameters, GR4J has 4 parameters, IHACRES has 6 parameters and Sacramento has 13 parameters.

The fact that, based on the results of the experiments conducted, more complex models have rougher error surfaces (i.e. Figure 4.7) suggests that, by inference, they should have decreased optimisation efficiency (i.e. they should require a larger number of function evaluations to find the globally optimal solution in the error surface). This is because all optimisation algorithms search the error surface in an iterative fashion, generally starting at a random location and using various mechanisms (e.g. gradient information, evolutionary operators) to move to regions with lower errors in subsequent iterations based on information derived from the error surface (Maier et al., 2019). If the error surface is rougher, the information about the error surface used to guide optimisation algorithms is "noisier" and potentially misleading at times, thereby slowing down the optimisation process.

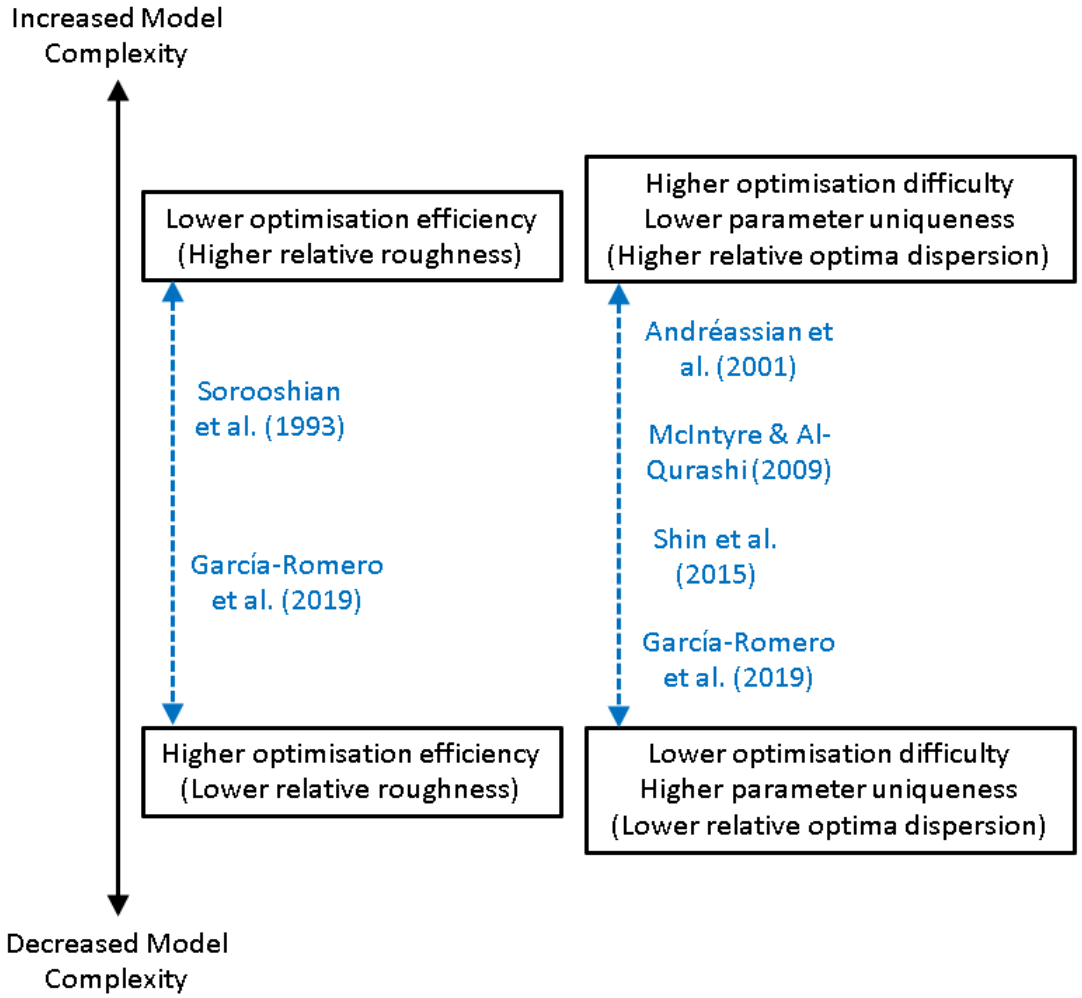


Figure 4.7 Impact of degree of model complexity on optimisation efficiency, optimisation difficulty and parameter uniqueness based on the result of the application of the proposed metrics for quantifying error surface roughness and optima dispersion, as well as relevant previous studies confirming these findings

The above inferences align with the findings of a number of previous studies (see Figure 4.7). For example, Garcia-Romero et al. (2019) calibrated 3 CRR models of different complexity (GR4J (4 parameters), HBV (10 parameters), Sacramento (16 parameters, the three extra parameters associated with surface vegetation area and groundwater transferring)) on 9 catchments using the SCE-UA algorithm and found that the simplest model (i.e. GR4J) only required

around 1,000 iterations for convergence, whereas the most complex model (i.e. Sacramento) required around 6,000. While Sorooshian et al. (1993) did not vary model complexity, they found that when calibrating the complex Sacramento model (13 parameters) over a number of trials using the Multistart Simplex algorithm, a very large number of function evaluations (an average of 45,887) was required for convergence, indicating low optimisation efficiency for this complex model.

By inference, the results in Figure 4.6 also suggest that optimisation difficulty (i.e. how difficult it is to find the global minimum in the error surface as part of the optimisation process) is likely to increase with increasing model complexity as a result of a corresponding increase in the dispersion of the optima over the error surface (Figure 4.7). This is because the global minimum in the error surface is more difficult to find if there are multiple regions of attraction at a greater distance from one another (i.e. disparate regions with low errors), as the feedback from the error surface is equally likely to direct the search to locally optimal solutions than the globally optimum solution. However, the degree to which this is the case is also likely to be affected by the explorative capability of the optimisation algorithm used (Maier et al., 2019). If this degree is low, as is the case for gradient methods, whether a local or the global optimum in the error surface is identified is generally highly dependent on the starting position of the search, as such algorithms are unable to escape local optima (Maier et al., 2019). However, if the exploratory capability of the algorithm is increased (e.g. by using multi-starts or evolutionary algorithms), a

wider area of the error surface is explored, including the different regions of attraction, making it more likely that the globally optimal solution is identified.

The finding that the optima of more complex models are likely to be dispersed over a greater region of the error surface also implies that the parameters of more complex models identified as part of the optimisation process should be to be less unique (Figure 4.7). This is because if optima are distributed over a smaller region of the error surface, the calibrated parameter values that are likely to be identified as part of the optimisation process are likely to be concentrated in a smaller region of the parameter space. As a result, even if there are different local optima, the resulting parameters values are similar to each other, ensuring that the values of the calibrated parameters are more well-defined.

The above inferences align with the findings of a number of previous studies (see Figure 4.7). For example, McIntyre and Al-Qurashi (2009) found that when they calibrated IHACRES models with 3, 4 and 5 parameters, respectively, the model with 5 parameters performed worst, even though it had the greatest number of degrees of freedom, suggesting that it was more difficult to find the global optimum in the error surface as model complexity increased. Garcia-Romero et al. (2019) found that while the calibrated parameters of the simpler GR4J models (4 parameters) generally had unique values, this was not the case for the more complex Sacramento model (16 parameters), where calibrated values for 10 of the 16 parameters corresponded to a large range of

values. Each of these was found approximately the same number of times as part of the calibration trials conducted, suggesting the existence of optima that are more widely distributed over the error surface, increasing the difficulty of the optimisation problem and decreasing the uniqueness of the parameter values identified.

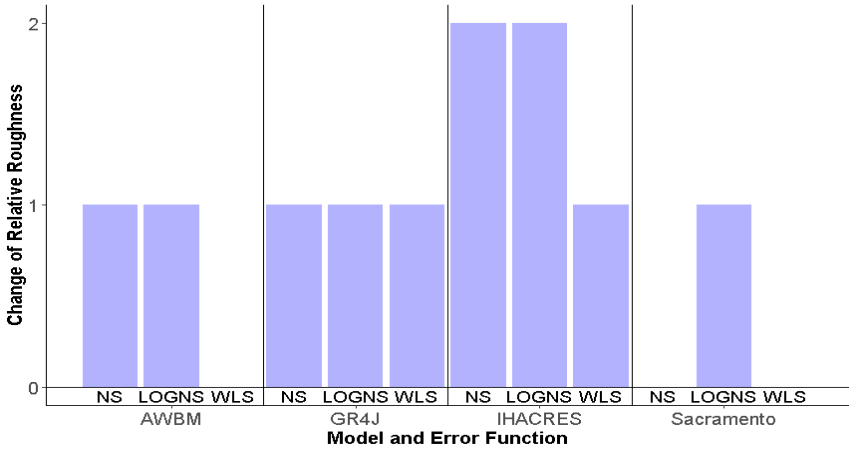
Andreassian et al. (2001) and Shin et al. (2015) had similar findings, where unique sets of parameter values were able to be obtained for simpler models, but not for more complex models. In the case of Andreassian et al. (2001), the parameter values obtained for a GR3J model (an earlier 3 parameter version of GR4J) were unique, while those obtained for TOPMOD (6 parameters) and IHACRES (6 parameters) were not. In the case of Shin et al. (2015), unique parameter sets were able to be identified for both GR4J (4 parameters) and IHACRES (4 parameters), whereas this was not the case for SIMHYD (9 parameters) and Sacramento (13 parameters). It is interesting to note that Andreassian et al. (2001) were not able to identify unique parameter sets for IHACRES, while Shin et al. (2015) were. One potential reason for this is that the former used a gradient method for calibration, which has limited exploratory capability, whereas the latter used 5 different global search algorithms (SCE, NSGA2, DREAM, DE, CMA-ES), which have greater exploratory ability and are therefore more likely to find global minima in error surfaces that have more widely distributed optima, as discussed above.

It should be noted that the ability to identify unique sets of calibrated

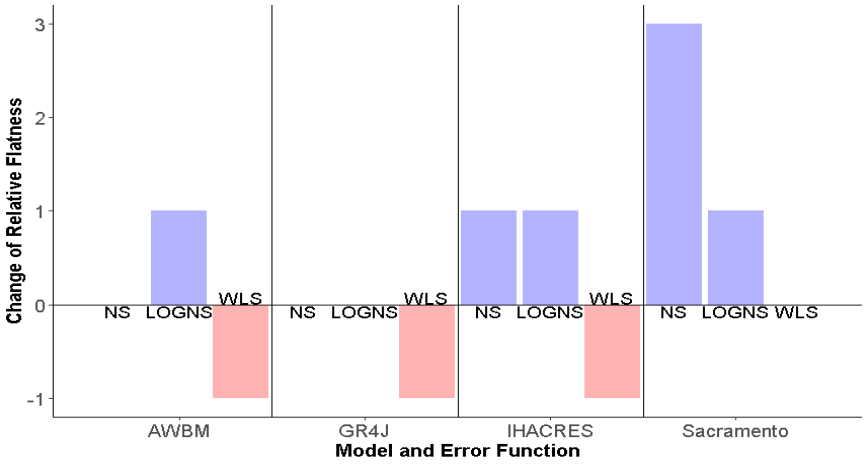
parameters by finding the global minimum in the error surface as part of an optimisation process is related to the well-known issue of parameter identifiability (Guillaume et al., 2019). However, parameter identifiability is also related to the quality of the calibration data (e.g. how well the calibration data represent the underlying physical processes being modelled). If the quality of the calibration data is low, the global minimum in the error surface found with the aid of the optimisation algorithm is not necessarily the "true" optimum (Maier et al., 2019). This was demonstrated by Andreassian et al. (2001) and Shin et al. (2015), who found that parameter uniqueness could be improved by improving the quality of the calibration data.

3.3 Impact of catchment climate condition

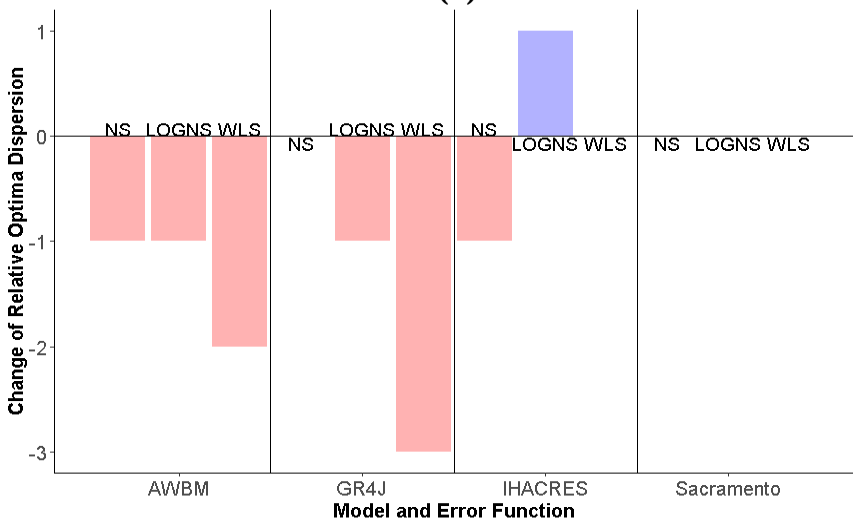
The impact of catchment climate condition on relative roughness, flatness and optima dispersion obtained for the experiments conducted is shown in Figure 4.8, which shows a column chart with the primary category as the model type, and the secondary category as the error metric. As can be seen, there is a general increase in the relative roughness of the error surface as catchment wetness increases. In contrast, increasing catchment wetness generally results in a decrease in the relative degree of optima dispersion. As was the case for the influence of model complexity, there appear to be no clear trends for the impact of catchment climate condition on the relative flatness of the error surface, with increases in flatness for some models and error metrics, and decreases for the three simplest models when the WLS error metric is used.



(a)



(b)



(c)

Figure 4.8 Change in features of error surface from dry to wet catchments: (a) Change in Relative Roughness; (b) Change in Relative Flatness; (c) Change in Relative Optima Dispersion.

Based on inference, the above results suggest that optimisation efficiency is likely to decrease with increasing catchment wetness (Figure 4.9). This is in agreement with Thyer et al. (1999), who found that the number of function evaluations required for an SCE-UA algorithm to converge when calibrating a modified SFB model for the wetter Allyn River catchment was 7,000, while this number was only 4,000 for the dryer Scott Creek catchment.

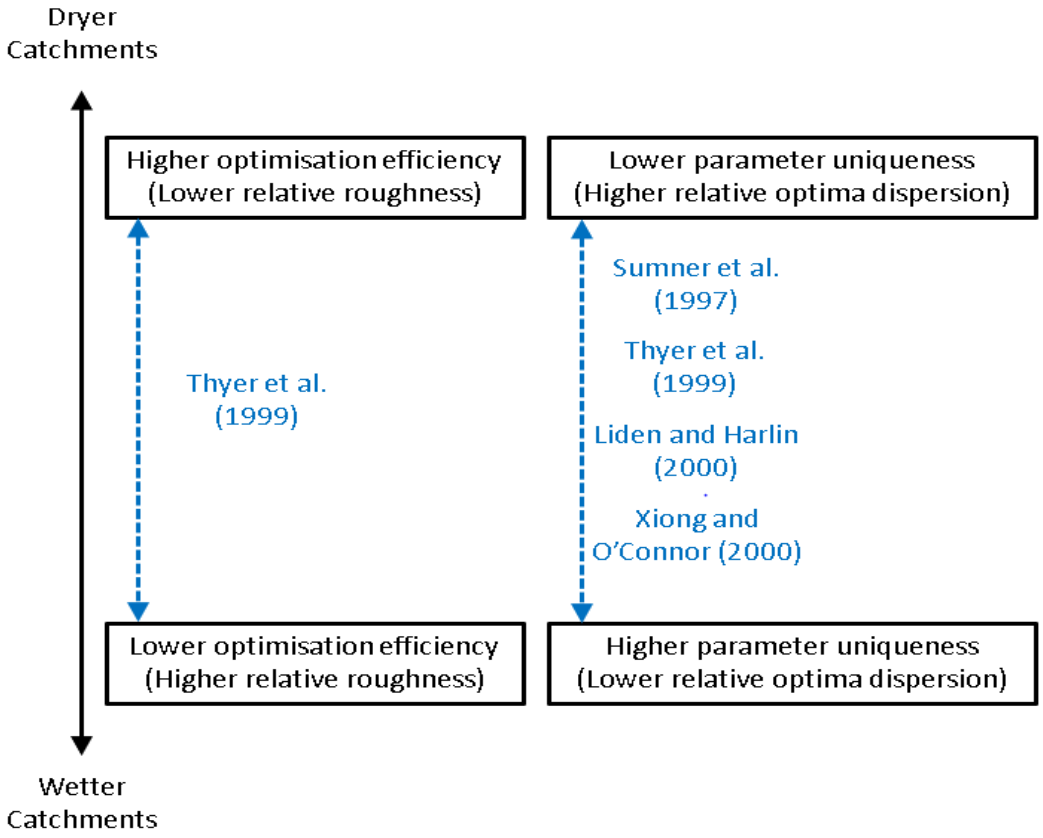


Figure 4.9 Impact of degree of catchment wetness on optimisation efficiency and parameter uniqueness based on the result of the application of the proposed metrics for quantifying error surface roughness and optima dispersion, as well as relevant previous studies confirming these findings

As the results in Figure 4.8(c) indicate that, for the experiments conducted, parameter dispersion decreases for wetter catchments, it can be inferred that parameter wetness should increase accordingly (Figure 4.9), which is in

agreement with the findings of a number of previous studies. For example, Liden and Harlin (2000) found that for dry catchments, it was not possible to identify unique values of parameters related to baseflow, but that unique parameter values could be identified for a wet catchment. When calibrating a modified SFB model for twenty-five Australian catchments with different climates, Sumner et al. (1997) found that unique parameters sets could be identified for wet catchments. However, this was not the case for semi-arid catchments. When examining plots of the error surface for wet and dry catchments for the same CRR model, Thyer et al. (1999) found that the error surface for the dry catchment had multiple local optima, thereby providing an explanation for the difficulty in finding unique parameter values for such catchments. Similar results were found by Xiong and O'Connor (2000), who visualized error surfaces for catchments in a humid region in Japan and a semi-arid region in Tanzania, where the error surfaces for the semi-arid catchment had more optima that were more widely dispersed. The reason why so many studies have similar findings (i.e. that catchment wetness results in an increase in parameter uniqueness) is likely because many commonly-used rainfall-runoff models demonstrate a trade-off between fitting high and low flows (i.e. there is insufficient flexibility in the model structure or error model). In a dry catchment, the ability to model behaviour associated with baseflow is much more important than in wet catchments. In wet catchments, the inability to reproduce this behaviour might manifest itself as "roughness" rather than dispersion of optima.

4 Summary and Conclusions

The calibration of conceptual rainfall-runoff (CRR) models has been the subject of a large number of research papers. While the impact of attributes such as model structure and complexity, catchment climate conditions, error metrics and calibration data length (and properties), as well as the optimisation method used, has received significant attention in these studies, there has been a lack of focus on systematic approaches to quantifying how the features of the error surface are affected by these attributes. This makes it difficult to provide a transparent, objective and repeatable assessment of the impact model structures, catchment climate conditions, error metrics and calibration data properties have on calibration efficiency, difficulty and parameter uniqueness. In order to address this shortcoming, three exploratory landscape analysis (ELA) metrics are proposed in this chapter as a way to quantify key features of the error surfaces of CRR models. The metrics were applied to 420 error surfaces of CRR models consisting of different combinations of model structures of varying complexity, catchments with different climate conditions, error metrics and calibration data sets of different length and composition.

Results show that, in the experiments conducted, there are clear differences in error surface roughness and optima dispersion between different models and catchments with different levels of wetness. More complex models are shown to have higher roughness and optima dispersion, while wetter catchments have higher roughness but lower optima dispersion. However, the generality of these findings needs to be tested using a wider range of models and conditions. The primary contribution of this chapter is the presentation of an alternative

approach to identify useful information to support decisions that have to be made during the development of CRR models, such as the selection of an appropriate model structure, the selection of an appropriate optimisation algorithm and the selection of appropriate values that control the searching behaviour (e.g. relative degree of exploration and exploitation) of the selected optimisation algorithm.

The results for the experiments conducted also provide valuable insight into how optimisation efficiency, optimisation difficulty and parameter uniqueness can be affected by different models and the degree of catchment wetness (Figure 4.10). For the experiments conducted, the results suggested that more complex models applied to wet catchments have low optimisation efficiency, medium optimisation difficulty and medium parameter uniqueness. However, when such models were applied to dry catchments, there was an increase in optimisation efficiency, an increase in optimisation difficulty and a decrease in parameter uniqueness. Irrespective of whether catchments were wet or dry, the results indicated that reducing model complexity for the models used in this study resulted in an improvement in all three performance measures. However, as mentioned above, the generality of these results need to be tested using a larger number of experiments.

The fact that the three ELA metrics proposed in this chapter are able to identify key features of the error surfaces of CRR models, including relative roughness, relative flatness and relative optima dispersion, creates opportunities for determining and comparing the features of the error surfaces

of a broad range of CRR models under a wide range of conditions, in an effective and efficient manner, providing insight into how these features change in response to different attributes. Consequently, as mentioned above, future studies should apply the approach introduced in this chapter to a larger number of combinations of CRR models, catchments and error metrics in order to test the generality of the findings of this study and to potentially provide high-level guidance on the selection of CRR models and optimisation approaches for different types of catchments. It could also guide the development of novel CRR models that can be calibrated more easily and efficiently and have more well-defined parameters.

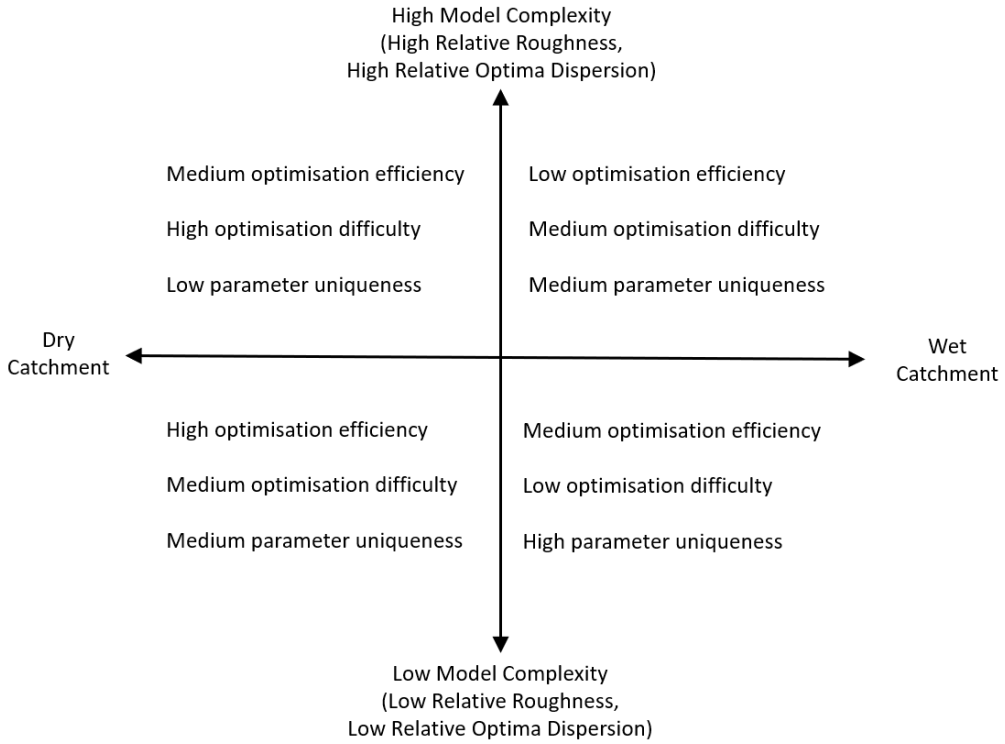


Figure 4.10 Impact of model complexity and catchment climate condition on model calibration difficulty, efficiency and parameter uniqueness, based on the results in this study

Chapter 5 Conclusions

1 Research Contributions

Optimisation algorithms are used extensively for the calibration of environmental models and the identification of solutions to environmental problems. How well a particular algorithm performs on a given problem is a function of both algorithm behaviour and the characteristics of the problem being solved, as represented by the fitness landscape. While significant attention has been given to the development of algorithms with different behaviours, little effort has been devoted to better understanding problem characteristics, generally resulting in a brute-force approach to identifying algorithms and parameterisations that perform acceptably for a particular problem. This is despite the fact that a number of metrics have been developed to assist with identifying features of fitness landscapes, such as their global structure, their degree of multimodality and the presence of plateaus, the identification of which would assist in the selection of appropriate optimisation algorithms and parameterisations without the need for a brute-force approach.

The most likely reason for the lack of adoption of fitness landscape metrics in practice is that the calculation of these metrics is based on samples from the

fitness landscape, which can be computationally expensive to generate for real-world environmental problems, as they are often require the running of complex and highly-dimensional simulation models. However, this research shows that not all of these metrics have high dependence on problem dimensionality and sample size. In this thesis, these metrics are identified through a large number of computational experiments and are shown to assist with the identification of optimisation difficulty and efficiency for two typical environmental modelling problems.

In Chapter 2, the degree of dependence on problem dimensionality and sample size of 110 fitness landscape metrics was assessed. Each metric was calculated for 72,000 different sets of fitness landscape samples obtained from 2,400 fitness landscapes derived from commonly used benchmark functions, and their degree of dependence on problem dimensionality and sample size was assessed. Results show that 39 of the 110 metrics have low dependence on dimensionality and sample size, 34 of which are considered suitable for application to environmental problems.

The low degree of dependence on problem dimensionality and sample size of these 34 metrics was tested on a number of real-world environmental modelling problems, corresponding to 7,590 sets of fitness landscape samples from 390 fitness landscapes. Results indicate that 28 of the 34 aforementioned fitness landscape metrics also have low dependence on problem dimensionality and sample size for the real-world environmental modelling problems, often

requiring fewer than 500 fitness landscape samples for convergence. These 28 metrics cover a wide range of fitness landscape features, including their global structure, multimodality, separability, search space and basin size homogeneity and the presence of plateaus.

In Chapter 3, 5 ELA metrics shown to have low dependence on problem dimensionality and sample size in Chapter 2 are used to better understand the features of the error surfaces of multi-layer perceptron (MLP) ANNs of varying complexity. Based on the results of four water quantity and quality case studies from the literature (Kentucky Runoff, Murray Salinity, Myponga Chlorine, SA Turbidity), it has been demonstrated that MLPs with a smaller number of hidden nodes and parameters are easier to calibrate, as they have a more well-defined overall shape that is able to guide optimisation algorithms to better regions in the error surface more easily. Additionally, the error surface of smaller MLPs is smoother, so that it is harder for algorithms to become trapped in local optima. In contrast, the generally flatter error surface of MLPs with more parameters and hidden nodes provides limited information to guide the search to better regions in the error surface. In addition, the higher level of multimodality / roughness of larger MLPs can also make it more difficult to identify the global optimum, especially for optimisation algorithms with limited exploration capacity, such as gradient-based methods.

On the other hand, as error surfaces of larger MLPs are more convex than those of smaller MLPs, which results in better-defined gradient information in local

regions, it should be easier to converge to the local optima of larger MLPs. However, this is also likely to lead to premature convergence to local optima, rather than the identification of the global optimum. In addition, the presence of these widely distributed, narrow and deep local optima in the error surfaces of more complex MLPs means that hybrid approaches to calibration are likely to result in better performance. This is because such approaches use algorithms with higher degrees of exploration, such as metaheuristics, in the initial stages of the calibration to find good regions in the error surface, followed by algorithms with a higher degree of exploitation, such as gradient methods, in the latter stages of calibration, enabling good locally optimal, or globally optimal, solutions to be identified.

In Chapter 4, 3 ELA metrics with low dependence on problem dimensionality and sample size, and related to roughness, flatness and optima dispersion are used to understand the features of error surfaces of different CRR models.

According to the results for error surfaces for fitness landscapes for different combinations of 4 models of different complexity (AWBM, GR4J, IHACRES and Sacramento), 5 catchments in different regions (Burnie, Darwin, Adelaide, Wagga Wagga, and Alice Springs) in Australia with climate condition varying from wet to dry, 3 different error metrics (NS, LOGNS and WLS) and 7 different groups of data with different lengths (3 groups of 1-year data, 3 groups of 5-year data, 1 group on 10-year data), which corresponds to total of 420 different error surfaces, model structure have the largest influence on the

features of error surface, with a clear difference shown in relative roughness and optima dispersion for different models (in this study, complex models are shown to have higher relative roughness and higher optima dispersion).

Climate conditions of catchments also appear to have a marked influence on relative roughness and optima dispersion, with wetter catchments resulting in rougher error surfaces, but with optima that are less dispersed. In contrast, the impact of different error metrics on error surface characteristics is confined to particular metrics, model structures and catchment climate conditions. The most pronounced influence is associated with the application of the WLS metric to simpler models, especially for wetter catchments. Error surface roughness increases when the WLS metric is used for models with a small number of parameters, such as AWBM, as use of this metric results in a significant relative increase in the number of parameters. In contrast, use of the WLS metric reduces the error surface flatness and optima dispersion for simpler models, especially for wetter catchments, resulting in more well-defined regions of attraction in parameter space. On the other hand, the results show that there is no pronounced impact of calibration data length on error surface characteristics.

2 Scope of Future Work

The recommendations for future work related to understanding the problem structure of real-world environmental optimisation problems are given below.

In this thesis, features of fitness landscapes of two typical environmental optimisation problems are assessed and it is shown that ELA metrics can successfully interpret the calibration difficulty and efficiency of the two kinds of models. As a result, it is worth investigating if ELA metrics can also be successfully applied to other environmental models. In this thesis, only CRR and one kind of ANN (i.e. MLP) models are considered. However, there are many different kinds of models, including data-driven models such as geomorphology-based artificial neural network (GANN) (Zhang and Govindaraju, 2003) and wavelet neural network (WNN) (Feng et al., 2016), and process-driven models for other cases such as groundwater simulation (Belmans et al., 1983) and water quality estimation (Abbaspour et al., 2007). Apart from model calibration problems, a range of other environmental optimisation problems, such as land use management (Emirhüseyinoğlu and Ryan, 2020; Newman et al., 2020), wastewater treatment (Hamed et al., 2004) and irrigation scheduling (Nguyen et al., 2017; Sedighkia et al., 2021) are also worth assessing, as long as the fitness landscapes of these problems are continuous. The application to multi- or many-objective optimisation problems would also be of interest.

The other promising future work related to this thesis should be the prediction of algorithm performance on different kinds of environmental optimisation problems. As the prediction of algorithm performance has been successful for benchmark problems through a machine learning framework (Munoz and Smith-Miles, 2017), it is worth duplicating the success in environmental

optimisation problems. This opens the door to developing evidence-based approaches to tailoring optimisation methods and parameterisations (see Wang et al., 2020; Zheng et al., 2017) for a range of optimisation problems, such as for calibrating ANN models of different types and complexity based on the knowledge of error surface features, rather than relying on a brute-force approach to using a range of optimisation approaches and picking the one that performs best for the problem at hand.

References

Abbaspour, K. C., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., . . . Srinivasan, R. (2007). Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT. Journal of Hydrology, 333(2-4), 413-430. doi:10.1016/j.jhydrol.2006.09.014

Adamowski, J., & Sun, K. (2010). Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds. Journal of Hydrology, 390(1-2), 85-91. doi:10.1016/j.jhydrol.2010.06.033

Ahmad, S. K., and Hossain, F. (2019). A generic data-driven technique for forecasting of reservoir inflow: Application for hydropower maximization. Environmental Modelling and Software, 119, 147-165. doi:10.1016/j.envsoft.2019.06.008.

Alavi, A. H., & Gandomi, A. H. (2011). Prediction of principal ground-motion parameters using a hybrid method coupling artificial neural networks and simulated annealing. Computers and Structures, 89(23-24), 2176-2194. doi:10.1016/j.compstruc.2011.08.019

Amamra, A., Khanchoul, K., Eslamian, S., & Zobir, S. H. (2018). Suspended sediment estimation using regression and artificial neural network models: Kebir watershed, northeast of Algeria, North Africa. International Journal of Hydrology Science and Technology, 8(4), 352-371. doi:10.1504/IJHST.2018.095526

Andréassian, V., Perrin, C., Michel, C., Usart-Sanchez, I., & Lavabre, J.

(2001). Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models. Journal of Hydrology, 250(1-4), 206-223. doi:10.1016/S0022-1694(01)00437-1

Andrews, F. T., Croke, B. F. W., & Jakeman, A. J. (2011). An open software environment for hydrological model assessment and development. Environmental Modelling and Software, 26(10), 1171-1185. doi:10.1016/j.envsoft.2011.04.006

Araujo, L. N., Belotti, J. T., Alves, T. A., Tadano, Y. D. S., and Siqueira, H. (2020). Ensemble method based on Artificial Neural Networks to estimate air pollution health risks. Environmental Modelling and Software, 123. doi:10.1016/j.envsoft.2019.104567.

Bahrami, S., Doulati Ardejani, F., & Baafi, E. (2016). Application of artificial neural network coupled with genetic algorithm and simulated annealing to solve groundwater inflow problem to an advancing open pit mine. Journal of Hydrology, 536, 471-484. doi:10.1016/j.jhydrol.2016.03.002

Banadkooki, F. B., Ehteram, M., Panahi, F., Sammen, S. S., Othman, F. B., & El-Shafie, A. (2020). Estimation of total dissolved solids (TDS) using new hybrid machine learning models. Journal of Hydrology, 587. doi:10.1016/j.jhydrol.2020.124989

Bayram, A., Kankal, M., & Önsoy, H. (2012). Estimation of suspended sediment concentration from turbidity measurements using artificial neural networks. Environmental Monitoring and Assessment, 184(7), 4355-4365. doi:10.1007/s10661-011-2269-2

Belmans, C., Wesseling, J. G., & Feddes, R. A. (1983). Simulation model of

the water balance of a cropped soil: SWATRE. Journal of Hydrology, 63(3-4), 271-286. doi:10.1016/0022-1694(83)90045-8

Beven, K. (2006). A manifesto for the equifinality thesis. Journal of Hydrology, 320(1-2), 18-36. doi:10.1016/j.jhydrol.2005.07.007

Beven, K. (2016). Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. Hydrological Sciences Journal, 61(9), 1652-1665. doi:10.1080/02626667.2015.1031761

Bi, W., Dandy, G. and Maier, H. (2016). Use of domain knowledge to increase the convergence rate of evolutionary algorithms for optimizing the cost and resilience of water distribution systems. J. Water Resour. Plan. Manag. 142(9): 04016027.

Boughton, W. (2004). The australian water balance model. Environmental Modelling & Software 19 (10), 943-956. http://dx.doi.org/10.1016/j.envsoft.2003.10.007

Bowden, G. J., Maier, H. R. and Dandy, G. C. (2002), Optimal division of data for neural network models in water resources applications, Water Resour. Res., 38(2), 1010, doi:10.1029/2001WR000266.

Bowden, G. J., Maier, H. R., & Dandy, G. C. (2005). Input determination for neural network models in water resources applications. Part 2. Case study: Forecasting salinity in a river. Journal of Hydrology, 301(1-4), 93-107. doi:10.1016/j.jhydrol.2004.06.020

Bowden, G. J., Nixon, J. B., Dandy, G. C., Maier, H. R., & Holmes, M. (2006). Forecasting chlorine residuals in a water distribution system using a general regression neural network. Mathematical and Computer Modelling, 44(5-6),

469-484. doi:10.1016/j.mcm.2006.01.006

Bullinaria, J. A., and AlYahya, K. (2014). Artificial Bee Colony training of neural networks: Comparison with back-propagation. Memetic Computing, 6(3), 171-182. doi:10.1007/s12293-014-0137-7

Burton, A., Kilsby, C. G., Fowler, H. J., Cowpertwait, P. S. P., and O'Connell, P. E. (2008). RainSim: A spatial-temporal stochastic rainfall modelling system. Environmental Modelling and Software, 23(12), 1356-1369. doi:10.1016/j.envsoft.2008.04.003.

Cabaneros, S. M., Calautit, J. K., and Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. Environmental Modelling and Software, 119, 285-304. doi:10.1016/j.envsoft.2019.06.014

Chau, K. W. (2006). Particle swarm optimization training algorithm for ANNs in stage prediction of Shing Mun River. Journal of Hydrology, 329(3-4), 363-367. doi:10.1016/j.jhydrol.2006.02.025

Chau, K. W. (2007). A split-step particle swarm optimization algorithm in river stage forecasting. Journal of Hydrology, 346(3-4), 131-135. doi:10.1016/j.jhydrol.2007.09.004

Chaudhari, N., Londhe, S., & Khare, K. (2012). Estimation of pan evaporation using soft computing tools. International Journal of Hydrology Science and Technology, 2(4), 373-390. doi:10.1504/IJHST.2012.052375

Cheng, M., Fang, F., Kinouchi, T., Navon, I. M., & Pain, C. C. (2020). Long lead-time daily and monthly streamflow forecasting using machine learning methods. Journal of Hydrology, 590. doi:10.1016/j.jhydrol.2020.125376

Chua, L. H. C., Wong, T. S. W., & Sriramula, L. K. (2008). Comparison between kinematic wave and artificial neural network models in event-based runoff simulation for an overland plane. Journal of Hydrology, 357(3-4), 337-348. doi:10.1016/j.jhydrol.2008.05.015

Clark, M. P., & Kavetski, D. (2010). Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes. 46(10). doi:https://doi.org/10.1029/2009WR008894

Cobaner, M. (2011). Evapotranspiration estimation by two different neuro-fuzzy inference systems. Journal of Hydrology, 398(3-4), 292-302. doi:10.1016/j.jhydrol.2010.12.030

Cole, J. C., Maloney, K. O., Schmid, M., & McKenna, J. E. (2014). Developing and testing temperature models for regulated systems: A case study on the Upper Delaware River. Journal of Hydrology, 519(PA), 588-598. doi:10.1016/j.jhydrol.2014.07.058

Coulibaly, P., & Baldwin, C. K. (2005). Nonstationary hydrological time series forecasting using nonlinear dynamic methods. Journal of Hydrology, 307(1-4), 164-174. doi:10.1016/j.jhydrol.2004.10.008

Croke, B.F.W., and A.J. Jakeman (2004) A catchment moisture deficit module for the IHACRES rainfall-runoff model. Environmental Modelling and Software 19:1-5.

Deb, K., & Goldberg, D. E. (1994). Sufficient conditions for deceptive and easy binary functions. Annals of Mathematics and Artificial Intelligence, 10(4), 385-408. doi:10.1007/BF01531277

DeWeber, J. T., & Wagner, T. (2014). A regional neural network ensemble for

predicting mean daily river water temperature. Journal of Hydrology, 517, 187-200. doi:10.1016/j.jhydrol.2014.05.035

Di Matteo, M., Maier, H. R. and Dandy, G. C. (2019). Many-objective portfolio optimization approach for stormwater management project selection encouraging decision maker buy-in, Environmental Modelling and Software, 111, 340-355, doi: 10.1016/j.envsoft.2018.09.008

Dimopoulos, Y., Bourret, P., & Lek, S. (1995). Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Processing Letters, 2(6), 1-4. doi:10.1007/BF02309007

Duan, Q., Sorooshian, S., & Gupta, V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resources Research, 28(4), 1015-1031. doi:10.1029/91WR02985

Elshorbagy, A., & Parasuraman, K. (2008). On the relevance of using artificial neural networks for estimating soil moisture content. Journal of Hydrology, 362(1-2), 1-18. doi:10.1016/j.jhydrol.2008.08.012

Emami Skardi, M. J., Afshar, A., Saadatpour, M., and Sandoval Solis, S. (2015). Hybrid ACO–ANN-Based Multi-objective Simulation–Optimization Model for Pollutant Load Control at Basin Scale. Environmental Modeling and Assessment, 20(1), 29-39. doi:10.1007/s10666-014-9413-7

Emirhüseyinoğlu, G., and Ryan, S. M. (2020). Land use optimization for nutrient reduction under stochastic precipitation rates. Environmental Modelling and Software, 123. doi:10.1016/j.envsoft.2019.104527

Fathian, F., Mehdizadeh, S., Kozekalani Sales, A., & Safari, M. J. S. (2019). Hybrid models to improve the monthly river flow prediction: Integrating

artificial intelligence and non-linear time series models. Journal of Hydrology, 575, 1200-1213. doi:10.1016/j.jhydrol.2019.06.025

Feng, Y., Jia, Y., Zhang, Q., Gong, D., & Cui, N. (2018). National-scale assessment of pan evaporation models across different climatic zones of China. Journal of Hydrology, 564, 314-328. doi:10.1016/j.jhydrol.2018.07.013

Feng, Y., Cui, N., Zhao, L., Hu, X., & Gong, D. (2016). Comparison of ELM, GANN, WNN and empirical models for estimating reference evapotranspiration in humid region of Southwest China. Journal of Hydrology, 536, 376-383. doi:10.1016/j.jhydrol.2016.02.053

Ferreira, L. B., da Cunha, F. F., de Oliveira, R. A., & Fernandes Filho, E. I. (2019). Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. Journal of Hydrology, 572, 556-570. doi:10.1016/j.jhydrol.2019.03.028

Fowler, K. J. A., Peel, M. C., Western, A. W., Zhang, L., & Peterson, T. J. (2016). Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models. Water Resources Research, 52(3), 1820-1846. doi:10.1002/2015WR018068

Fu, G., Butler, D., and Khu, S. T. (2008). Multiple objective optimal control of integrated urban wastewater systems. Environmental Modelling and Software, 23(2), 225-234. doi:10.1016/j.envsoft.2007.06.003

Galelli S., Humphrey G. B. , Maier H. R., Castelletti A., Dandy G. C. and Gibbs M. S. (2014). An evaluation framework for input variable selection algorithms for environmental data-driven models, Environmental Modelling and Software, 62, 33-51, doi: 10.1016/j.envsoft.2014.08.015

Gan, T. Y., Dlamini, E. M., & Biftu, G. F. (1997). Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. Journal of Hydrology, 192(1-4), 81-103. doi:10.1016/S0022-1694(96)03114-9

García-Romero, L., Paredes-Arquiola, J., Solera, A., Belda, E., Andreu, J., & Sánchez-Quispe, S. T. (2019). Optimization of the multi-start strategy of a direct- search algorithm for the calibration of rainfall- runoff models for water-resource assessment. Water (Switzerland), 11(9). doi:10.3390/w11091876

Garden, R. W., & Engelbrecht, A. P. (2014). Analysis and classification of optimisation benchmark functions and benchmark suites. Paper presented at the Proceedings of the 2014 IEEE Congress on Evolutionary Computation, CEC 2014.

Gautam, M. R., Watanabe, K., & Saegusa, H. (2000). Runoff analysis in humid forest catchment with artificial neural network. Journal of Hydrology, 235(1-2), 117-136. doi:10.1016/S0022-1694(00)00268-7

Gibbs, M. S., Maier, H. R. and Dandy, G. C. (2015). Using characteristics of the optimisation problem to determine genetic algorithm population size when evaluation number is limited, Environmental Modelling and Software, 69, 226-239, doi:10.1016/j.envsoft.2014.08.023

Gibbs, M. S., Maier, H. R., and Dandy, G. C. (2011). Relationship between problem characteristics and the optimal number of genetic algorithm generations. Engineering Optimization, 43(4), 349-376. doi:10.1080/0305215X.2010.491547.

Gibbs, M. S., McInerney, D., Humphrey, G., Thyer, M. A., Maier, H. R., Dandy, G. C., & Kavetski, D. (2018). State updating and calibration period

selection to improve dynamic monthly streamflow forecasts for an environmental flow management application. Hydrology and Earth System Sciences, 22(1), 871-887. doi:10.5194/hess-22-871-2018

Gichamo, T. Z., & Tarboton, D. G. (2019). Ensemble Streamflow Forecasting Using an Energy Balance Snowmelt Model Coupled to a Distributed Hydrologic Model with Assimilation of Snow and Streamflow Observations. Water Resources Research, 55(12), 10813-10838. doi:10.1029/2019WR025472

Graf, R., Zhu, S., & Sivakumar, B. (2019). Forecasting river water temperature time series using a wavelet–neural network hybrid modelling approach. Journal of Hydrology, 578. doi:10.1016/j.jhydrol.2019.124115

Grivas, G., and Chaloulakou, A. (2006). Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. Atmospheric Environment, 40(7), 1216-1229. doi:10.1016/j.atmosenv.2005.10.036

Guidici, F., Castelletti, A., Garofalo, E., Giuliani, M. and Maier, H. R. (2019). Dynamic, multi-objective optimal design and operation of water-energy systems for small, off-grid islands , Applied Energy, 250, 605-616, doi: 10.1016/j.apenergy.2019.05.084

Guillaume, J. H. A., Jakeman, J. D., Marsili-Libelli, S., Asher, M., Brunner, P., Croke, B., . . . Stigter, J. D. (2019). Introductory overview of identifiability analysis: A guide to evaluating whether you have the right type of data for your modeling purpose. Environmental Modelling and Software, 119, 418-432. doi:10.1016/j.envsoft.2019.07.007

Guo, D., Westra, S., & Maier, H. R. (2017). Impact of evapotranspiration process representation on runoff projections from conceptual rainfall-runoff models. Water Resources Research, 53(1), 435-454. doi:10.1002/2016WR019627

Guo, D., Zheng, F., Gupta, H., & Maier, H. R. (2020). On the Robustness of Conceptual Rainfall-Runoff Models to Calibration and Evaluation Data Set Splits Selection: A Large Sample Investigation. 56(3), e2019WR026752. doi:https://doi.org/10.1029/2019WR026752

Hadka, D., and Reed, P. (2015). Large-scale parallelization of the Borg multiobjective evolutionary algorithm to enhance the management of complex environmental systems. Environmental Modelling and Software, 69, 353-369. doi:10.1016/j.envsoft.2014.10.014.

Hajji, S., Hachicha, W., Bouri, S., & Dhia, H. B. (2012). Spatiotemporal groundwater level forecasting and monitoring using a neural network-based approach in a semi arid zone. International Journal of Hydrology Science and Technology, 2(4), 342-361. doi:10.1504/IJHST.2012.052366

Hamed, M. M., Khalafallah, M. G., and Hassanien, E. A. (2004). Prediction of wastewater treatment plant performance using artificial neural networks. Environmental Modelling and Software, 19(10), 919-928. doi:10.1016/j.envsoft.2003.10.005

Hansen, N., Finck, S., Ros, R., Auger, A. (2009). Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions Technical Report RR-6829, INRIA.

He, J., Reeves, C., Witt, C. and Yao, X. (2007). A note on problem difficulty

measures in black-box optimization: classification, realizations and predictability, Evol. Comput. 15 (4) pp.435–443.

He, X., Guan, H., & Qin, J. (2015). A hybrid wavelet neural network model with mutual information and particle swarm optimization for forecasting monthly rainfall. Journal of Hydrology, 527, 88-100. doi:10.1016/j.jhydrol.2015.04.047

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

Houle, M., Kriegel, H., Kroger, P., Schubert, E. and Zimek, A. (2010). Can shared-neighbour distances defeat the curse of dimensionality?, in: M. Gertz, B. Ludascher (Eds.), Scientific and Statistical Database Management, Lect. Notes Comput. Sci., vol. 6187, Springer, pp. 482–500.

Humphrey, G. B., Gibbs, M. S., Dandy, G. C., & Maier, H. R. (2016). A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network. Journal of Hydrology, 540, 623-640. doi:10.1016/j.jhydrol.2016.06.026

Humphrey, G. B., Maier, H. R., Wu, W., Mount, N. J., Dandy, G. C., Abrahart, R. J., & Dawson, C. W. (2017). Improved validation framework and R-package for artificial neural network models. Environmental Modelling and Software, 92, 82-106. doi:10.1016/j.envsoft.2017.01.023

Iorgulescu, I., & Jordan, J. P. (1994). Validation of TOPMODEL on a small Swiss catchment. Journal of Hydrology, 159(1-4), 255-273. doi:10.1016/0022-1694(94)90260-7

Jain, A., & Srinivasulu, S. (2006). Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques. Journal of Hydrology, 317(3-4), 291-306. doi:10.1016/j.jhydrol.2005.05.022

Jakeman, A.J., I.G. Littlewood and P.G. Whitehead (1990). Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments.Journal of Hydrology 117: 275-300.

Jian, J., Shiklomanov, A., Shuster, W. D., & Stewart, R. D. (2021). Predicting near-saturated hydraulic conductivity in urban soils. Journal of Hydrology, 595. doi:10.1016/j.jhydrol.2021.126051

Jothiprakash, V., & Magar, R. B. (2012). Multi-time-step ahead daily and hourly intermittent reservoir inflow prediction by artificial intelligent techniques using lumped and distributed data. Journal of Hydrology, 450-451, 293-307. doi:10.1016/j.jhydrol.2012.04.045

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

Kavetski, D., & Kuczera, G. (2007). Model smoothing strategies to remove microscale discontinuities and spurious secondary optima im objective functions in hydrological calibration. Water Resources Research, 43(3). doi:10.1029/2006WR005195

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resources Research,

42(3). doi:10.1029/2005WR004368

Kavetski, D., Kuczera, G., Thyer, M., & Renard, B. (2007). Multistart Newton-type optimisation methods for the calibration of conceptual hydrological models. Paper presented at the MODSIM07 - Land, Water and Environmental Management: Integrated Systems for Sustainability, Proceedings.

Kerschke, P. and Trautmann, H. (2016). The R-Package FLACCO for exploratory landscape analysis with applications to multi-objective optimization problems, 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, pp. 5262-5269.

Kerschke, P., Preuss, M., Wessing, S., & Trautmann, H. (2015). Detecting funnel structures by means of exploratory landscape analysis. Paper presented at the GECCO 2015 - Proceedings of the 2015 Genetic and Evolutionary Computation Conference.

Khatibi, R., Ghorbani, M. A., Kashani, M. H., & Kisi, O. (2011). Comparison of three artificial intelligence techniques for discharge routing. Journal of Hydrology, 403(3-4), 201-212. doi:10.1016/j.jhydrol.2011.03.007

Kim, G., & Barros, A. P. (2001). Quantitative flood forecasting using multisensor data and neural networks. Journal of Hydrology, 246(1-4), 45-62. doi:10.1016/S0022-1694(01)00353-5

Kim, J., Seo, D., Jang, M., & Kim, J. (2021). Augmentation of limited input data using an artificial neural network method to improve the accuracy of water quality modeling in a large lake. Journal of Hydrology, 602. doi:10.1016/j.jhydrol.2021.126817

Kingston, G. B., Maier, H. R., & Lambert, M. F. (2005). Calibration and

validation of neural networks to ensure physically plausible hydrological modeling. Journal of Hydrology, 314(1-4), 158-176. doi:10.1016/j.jhydrol.2005.03.013

Kingston, G. B., Maier, H. R., & Lambert, M. F. (2006). A probabilistic method for assisting knowledge extraction from artificial neural networks used for hydrological prediction. Mathematical and Computer Modelling, 44(5-6), 499-512. doi:10.1016/j.mcm.2006.01.008

Kişi, O. (2006). Daily pan evaporation modelling using a neuro-fuzzy computing technique. Journal of Hydrology, 329(3-4), 636-646. doi:10.1016/j.jhydrol.2006.03.015

Kişi, Ö. (2013). Evolutionary neural networks for monthly pan evaporation modeling. Journal of Hydrology, 498, 36-45. doi:10.1016/j.jhydrol.2013.06.011

Kişi, Ö., & Tombul, M. (2013). Modeling monthly pan evaporations using fuzzy genetic approach. Journal of Hydrology, 477, 203-212. doi:10.1016/j.jhydrol.2012.11.030

Kisi, O., Ozkan, C., and Akay, B. (2012). Modeling discharge-sediment relationship using neural networks with artificial bee colony algorithm. Journal of Hydrology, 428-429, 94-103. doi:10.1016/j.jhydrol.2012.01.026

Kuczera, G. (1997). Efficient subspace probabilistic parameter optimization for catchment models. Water Resources Research, 33(1), 177-185. doi:10.1029/96WR02671

Kurian, C., Sudheer, K. P., Vema, V. K., & Sahoo, D. (2020). Effective flood forecasting at higher lead times through hybrid modelling framework. Journal

of Hydrology, 587. doi:10.1016/j.jhydrol.2020.124945

Leahy, P., Kiely, G., & Corcoran, G. (2008). Structural optimisation and input selection of an artificial neural network for river level prediction. Journal of Hydrology, 355(1-4), 192-201. doi:10.1016/j.jhydrol.2008.03.017

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling, 90(1), 39-52. doi:10.1016/0304-3800(95)00142-5

Li, P., Zha, Y., Shi, L., Tso, C. H. M., Zhang, Y., & Zeng, W. (2020). Comparison of the use of a physical-based model with data assimilation and machine learning methods for simulating soil water dynamics. Journal of Hydrology, 584. doi:10.1016/j.jhydrol.2020.124692

Lidén, R., & Harlin, J. (2000). Analysis of conceptual rainfall-runoff modelling performance in different climates. Journal of Hydrology, 238(3-4), 231-247. doi:10.1016/S0022-1694(00)00330-9

Lischeid, G. (2001). Investigating short-term dynamics and long-term trends of SO4 in the runoff of a forested catchment using artificial neural networks. Journal of Hydrology, 243(1-2), 31-42. doi:10.1016/S0022-1694(00)00399-1

Liu, Y., Cibin, R., Bralts, V. F., Chaubey, I., Bowling, L. C., and Engel, B. A. (2016). Optimal selection and placement of BMPs and LID practices with a rainfall-runoff model. Environmental Modelling and Software, 80, 281-296. doi:10.1016/j.envsoft.2016.03.005

Lunacek, M. and Whitley, D. (2006). The dispersion metric and the CMA evolution strategy, in: Proceedings of the 8th Annual Conference on Genetic

and Evolutionary Computation, ACM, New York, NY, USA, pp. 477–484.

Maier, H. R. , Kapelan, Z., Kasprzyk, J., Kollat, J., Matott, L. S., Cunha, M. C., Dandy, G. C., Gibbs, M. S., Keedwell, E., Marchi, A., Ostfeld, A., Savic, D., Solomatine, D. P., Vrugt, J. A., Zecchin, A. C., Minsker, B. S., Barbour, E. J., Kuczera, G., Pasha, F., Castelletti, A., Giuliani, M. and Reed, P. M. (2014). Evolutionary algorithms and other metaheuristics in water resources: Current status, research challenges and future directions , Environmental Modelling and Software, 62, 271-299, doi: 10.1016/j.envsoft.2014.09.013.

Maier, H. R., & Dandy, G. C. (1996). The use of artificial neural networks for the prediction of water quality parameters. Water Resources Research, 32(4), 1013-1022. doi:10.1029/96wr03529

Maier, H. R., & Dandy, G. C. (1997). Determining inputs for neural network models of multivariate time series. Microcomputers in Civil Engineering, 12(5), 353-368. doi:10.1111/0885-9507.00069

Maier, H. R., & Dandy, G. C. (1998a). The effect of internal parameters and geometry on the performance of back-propagation neural networks: An empirical study. Environmental Modelling and Software, 13(2), 193-209. doi:10.1016/S1364-8152(98)00020-6

Maier, H. R., & Dandy, G. C. (1998b). Understanding the behaviour and optimising the performance of back-propagation neural networks: An empirical study. Environmental Modelling and Software, 13(2), 179-191. doi:10.1016/S1364-8152(98)00019-X

Maier, H. R., & Dandy, G. C. (1999). Empirical comparison of various methods for training feed-forward neural networks for salinity forecasting. Water

Resources Research, 35(8), 2591-2596. doi:10.1029/1999WR900150

Maier, H. R., Jain, A., Dandy, G. C., & Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environmental Modelling and Software, 25(8), 891-909. doi:10.1016/j.envsoft.2010.02.003

Maier, H. R., Jain, A., Dandy, G. C., and Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environmental Modelling and Software, 25(8), 891-909. doi:10.1016/j.envsoft.2010.02.003

Maier, H. R., Morgan, N. and Chow C. W. K. (2004). Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters, Environ. Modell. Software, 19(5), 485–494, doi:10.1016/ s1364-8152(03)00163-4.

Maier, H. R., Razavi, S., Kapelan, Z., Matott, L. S., Kasprzyk, J., and Tolson, B. A. (2019). Introductory overview: Optimization using evolutionary algorithms and other metaheuristics. Environmental Modelling and Software, 114, 195-213. doi:10.1016/j.envsoft.2018.11.018

Malan, K. and Engelbrecht, A. (2013). A survey of techniques for characterising fitness landscapes and some possible ways forward, Inform. Sci. 241 (0), pp. 148–163.

Maroufpoor, S., Bozorg-Haddad, O., & Maroufpoor, E. (2020). Reference evapotranspiration estimating based on optimal input combination and hybrid artificial intelligent model: Hybridization of artificial neural network with grey wolf optimizer algorithm. Journal of Hydrology, 588.

doi:10.1016/j.jhydrol.2020.125060

May, D. B., and Sivakumar, M. (2009). Prediction of urban stormwater quality using artificial neural networks. Environmental Modelling and Software, 24(2), 296-302. doi:10.1016/j.envsoft.2008.07.004.

May, R. J., Dandy, G. C., Maier, H. R., & Nixon, J. B. (2008). Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. Environmental Modelling and Software, 23(10-11), 1289-1299. doi:10.1016/j.envsoft.2008.03.008

McIntyre, N., & Al-Qurashi, A. (2009). Performance of ten rainfall-runoff models applied to an arid catchment in Oman. Environmental Modelling and Software, 24(6), 726-738. doi:10.1016/j.envsoft.2008.11.001

Meral, R., Dogan Demir, A., & Cemek, B. (2018). Analyses of turbidity and acoustic backscatter signal with artificial neural network for estimation of suspended sediment concentration. Applied Ecology and Environmental Research, 16(1), 697-708. doi:10.15666/aeer/1601_697708

Mersmann, O., Bischl, B., Trautmann, H., Preuß, M., Weihs, C., and Rudolph G. (2011). Exploratory landscape analysis, in: Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation, GECCO '11, ACM, New York, NY, USA, pp. 829–836.

Mersmann, O., Preuss, M., and Trautmann, H. (2010). Benchmarking evolutionary algorithms: Towards exploratory landscape analysis. In PPSN XI: Proceedings of the 11th International Conference on Parallel Problem Solving from Nature, R. Schaefer et al., Eds. Lecture Notes in Computer Science 6238. Springer, 71–80.

Mount, N. J., Dawson, C. W., & Abrahart, R. J. (2013). Legitimising data-driven models: Exemplification of a new data-driven mechanistic modelling framework. Hydrology and Earth System Sciences, 17(7), 2827-2843. doi:10.5194/hess-17-2827-2013

Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F. J., & Abrahart, R. J. (2016). Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan. Hydrological Sciences Journal, 61(7), 1192-1208. doi:10.1080/02626667.2016.1159683

Mukherjee, A., & Ramachandran, P. (2018). Prediction of GWL with the help of GRACE TWS for unevenly spaced time series data in India : Analysis of comparative performances of SVR, ANN and LRM. Journal of Hydrology, 558, 647-658. doi:10.1016/j.jhydrol.2018.02.005

Müller, C. and Sbalzarini, I. (2011). Global characterization of the CEC 2005 fitness landscapes using fitness distance analysis, in: Applications of Evolutionary Computation, Lect. Notes Comput. Sci., vol. 6624, Springer, pp. 294–303.

Müllner, D. (2013). Fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. Journal of Statistical Software, 53(9), 1-18. doi:10.18637/jss.v053.i09.

Munoz, M. and Smith-Miles, A. (2017). Performance analysis of continuous black-box optimization algorithms via footprints in instance space. Evolutionary Computation, 25(4), 529–554.

Munoz, M., Kirley, M., and Halgamuge, S. (2015b). Exploratory landscape

analysis of continuous space optimization problems using information content. IEEE Transactions on Evolutionary Computation, 19(1):74–87.

Munoz, M., Sun, Y., Kirley, M., and Halgamuge, S. (2015a). Algorithm selection for black-box continuous optimization problems: A survey on methods and challenges. Information Sciences, 317:224–245.

Newland, C. P., van Delden, H., Zecchin, A. C., Newman, J. P. and Maier, H. R. (2020). A hybrid (semi) automatic calibration method for Cellular Automata land-use models: Combining evolutionary algorithms with process understanding, Environmental Modelling and Software, 134, 104830, doi: 10.1016/j.envsoft.2020.104830.

Nguyen-ky, T., Mushtaq, S., Loch, A., Reardon-Smith, K., An-Vo, D. A., Ngo-Cong, D., & Tran-Cong, T. (2018). Predicting water allocation trade prices using a hybrid Artificial Neural Network-Bayesian modelling approach. Journal of Hydrology, 567, 781-791. doi:10.1016/j.jhydrol.2017.11.049

Nielsen, J. H. (2016). Do Group Decision Rules Affect Trust? A Laboratory Experiment on Group Decision Rules and Trust. Scandinavian Political Studies, 39(2), 115-137. doi:10.1111/1467-9477.12058

Nourani, V., & Mousavi, S. (2016). Spatiotemporal groundwater level modeling using hybrid artificial intelligence-meshless method. Journal of Hydrology, 536, 10-25. doi:10.1016/j.jhydrol.2016.02.030

Nourani, V., Elkiran, G., & Abdullahi, J. (2020a). Multi-step ahead modeling of reference evapotranspiration using a multi-model approach. Journal of Hydrology, 581. doi:10.1016/j.jhydrol.2019.124434

Nourani, V., Mousavi, S., Sadikoglu, F., & Singh, V. P. (2017). Experimental

and AI-based numerical modeling of contaminant transport in porous media. Journal of Contaminant Hydrology, 205, 78-95. doi:10.1016/j.jconhyd.2017.09.006

Nourani, V., Sayyah-Fard, M., Alami, M. T., & Sharghi, E. (2020b). Data pre-processing effect on ANN-based prediction intervals construction of the evaporation process at different climate regions in Iran. Journal of Hydrology, 588. doi:10.1016/j.jhydrol.2020.125078

Olyaie, E., Banejad, H., Chau, K. W., & Melesse, A. M. (2015). A comparison of various artificial intelligence approaches performance for estimating suspended sediment load of river systems: a case study in United States. Environmental Monitoring and Assessment, 187(4). doi:10.1007/s10661-015-4381-1

Pan, T. Y., Yang, Y. T., Kuo, H. C., Tan, Y. C., Lai, J. S., Chang, T. J., . . . Hsu, K. H. (2013). Improvement of watershed flood forecasting by typhoon rainfall climate model with an ANN-based southwest monsoon rainfall enhancement. Journal of Hydrology, 506, 90-100. doi:10.1016/j.jhydrol.2013.08.018

Parchami-Araghi, F., Mirlatifi, S. M., Ghorbani Dashtaki, S., & Mahdian, M. H. (2013). Point estimation of soil water infiltration process using Artificial Neural Networks for some calcareous soils. Journal of Hydrology, 481, 35-47. doi:10.1016/j.jhydrol.2012.12.007

Patrignani, A., & Ochsner, T. E. (2018). Modeling transient soil moisture dichotomies in landscapes with intermixed land covers. Journal of Hydrology, 566, 783-794. doi:10.1016/j.jhydrol.2018.09.049

Pelletier, G. J., Chapra, S. C., and Tao, H. (2006). QUAL2Kw - A framework for modeling water quality in streams and rivers using a genetic algorithm for calibration. Environmental Modelling and Software, 21(3), 419-425. doi:10.1016/j.envsoft.2005.07.002.

Pereira Filho, A. J., & Dos Santos, C. C. (2006). Modeling a densely urbanized watershed with an artificial neural network, weather radar and telemetric data. Journal of Hydrology, 317(1-2), 31-48. doi:10.1016/j.jhydrol.2005.05.007

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology, 279(1-4), 275-289. doi:10.1016/S0022-1694(03)00225-7

Phien, H. N., & Kha, N. D. A. (2003). Flood forecasting for the upper reach of the Red River Basin, North Vietnam. Water SA, 29(3), 267-272.

Pijanowski, B. C., Tayyebi, A., Doucette, J., Pekin, B. K., Braun, D., and Plourde, J. (2014). A big data urban growth simulation at a national scale: Configuring the GIS and neural network based Land Transformation Model to run in a High Performance Computing (HPC) environment. Environmental Modelling and Software, 51, 250-268. doi:10.1016/j.envsoft.2013.09.015

Piotrowski, A. P., & Napiorkowski, J. J. (2011). Optimizing neural networks for river flow forecasting - Evolutionary Computation methods versus the Levenberg-Marquardt approach. Journal of Hydrology, 407(1-4), 12-27. doi:10.1016/j.jhydrol.2011.06.019

Piotrowski, A. P., & Napiorkowski, J. J. (2013). A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling. Journal of Hydrology, 476, 97-111.

doi:10.1016/j.jhydrol.2012.10.019

Pitzer, E., and Affenzeller, M. (2012). A comprehensive survey on fitness landscape analysis. In: Vol. 378. Studies in Computational Intelligence (pp. 161-191).

Ráduly, B., Gernaey, K. V., Capodaglio, A. G., Mikkelsen, P. S., and Henze, M. (2007). Artificial neural networks for rapid WWTP performance evaluation: Methodology and case study. Environmental Modelling and Software, 22(8), 1208-1216. doi:10.1016/j.envsoft.2006.07.003.

Rahmati, M. (2017). Reliable and accurate point-based prediction of cumulative infiltration using soil readily available characteristics: A comparison between GMDH, ANN, and MLR. Journal of Hydrology, 551, 81-91. doi:10.1016/j.jhydrol.2017.05.046

Rajurkar, M. P., Kothyari, U. C., & Chaube, U. C. (2004). Modeling of the daily rainfall-runoff relationship with artificial neural network. Journal of Hydrology, 285(1-4), 96-113. doi:10.1016/j.jhydrol.2003.08.011

Razavi, S., and Gupta, H.V. (2016). A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory. Water Resources Research, 52(1), 423-439. doi:10.1002/2015WR017558.1111/1467-9477.12058

Razavi, S., & Gupta, H. V. (2015). What do we mean by sensitivity analysis? the need for comprehensive characterisation of "global" sensitivity in Earth and Environmental systems models. Water Resources Research, 51(5), 3070-3092. doi:10.1002/2014WR016527

Razavi, S., & Gupta, H. V. (2015). What do we mean by sensitivity analysis?

the need for comprehensive characterisation of "global" sensitivity in Earth and Environmental systems models. Water Resources Research, 51(5), 3070-3092. doi:10.1002/2014WR016527

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Lo Piano, S., Iwanaga, T., Becker, W., Tarantola, S., Guillaume, J. H. A., Jakeman, J., Gupta, H., Melillo, N., Rabitti, G., Chabridon, V., Duan, Q., Sun, X., Smith, S., Sheikholeslami, R., Hosseini, N., Asadzadeh, M., Puy, A., Sergei Kucherenko, S., Maier, H. R. (2021). The future of sensitivity analysis: An essential discipline for systems modeling and policy support Environmental Modelling and Software, 137, 104954, doi: 10.1016/j.envsoft.2020.104954.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resources Research, 46(5). doi:10.1029/2009WR008328

Rezaali, M., Quilty, J., & Karimi, A. (2021). Probabilistic urban water demand forecasting using wavelet-based machine learning models. Journal of Hydrology, 600. doi:10.1016/j.jhydrol.2021.126358

Rice, J. (1976). The algorithm selection problem, Advances in Computers, vol. 15, Elsevier, pp. 65–118.

Sabouri, F., Gharabaghi, B., Mahboubi, A. A., & McBean, E. A. (2013). Impervious surfaces and sewer pipe effects on stormwater runoff temperature. Journal of Hydrology, 502, 10-17. doi:10.1016/j.jhydrol.2013.08.016

Sabouri, F., Gharabaghi, B., Sattar, A. M. A., & Thompson, A. M. (2016).

Event-based stormwater management pond runoff temperature model. Journal of Hydrology, 540, 306-316. doi:10.1016/j.jhydrol.2016.06.017

Sahoo, G. B., Schladow, S. G., & Reuter, J. E. (2009). Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models. Journal of Hydrology, 378(3-4), 325-342. doi:10.1016/j.jhydrol.2009.09.037

Sajikumar, N., & Thandaveswara, B. S. (1999). A non-linear rainfall-runoff model using an artificial neural network. Journal of Hydrology, 216(1-2), 32-55. doi:10.1016/S0022-1694(98)00273-X

Sedighkia, M., Datta, B., and Abdoli, A. (2021). Minimizing physical habitat impacts at downstream of diversion dams by a multiobjective optimization of environmental flow regime. Environmental Modelling and Software, 140. doi:10.1016/j.envsoft.2021.105029

See, L., & Openshaw, S. (1999). Applying soft computing approaches to river level forecasting. Hydrological Sciences Journal, 44(5), 763-778. doi:10.1080/02626669909492272

Sheikholeslami, R. and Razavi, S. (2017). Progressive Latin Hypercube Sampling: An efficient approach for robust sampling-based analysis of environmental models. Environmental modelling and software, 93, 109-126. doi: 10.1016/j.envsoft.2017.03.010

Shin, M. J., Guillaume, J. H. A., Croke, B. F. W., & Jakeman, A. J. (2015). A review of foundational methods for checking the structural identifiability of models: Results for rainfall-runoff. Journal of Hydrology, 520, 1-16. doi:10.1016/j.jhydrol.2014.11.040

Shin, M. J., Guillaume, J. H. A., Croke, B. F. W., & Jakeman, A. J. (2013). Addressing ten questions about conceptual rainfall-runoff models with global sensitivity analyses in R. Journal of Hydrology, 503, 135-152. doi:10.1016/j.jhydrol.2013.08.047

Shirakawa S., Nagao T. (2014) Local Landscape Patterns for Fitness Landscape Analysis. In: Dick G. et al. (eds) Simulated Evolution and Learning. SEAL 2014. Lecture Notes in Computer Science, vol 8886. Springer, Cham.

Shirakawa, S., Nagao, T. (2016). Bag of local landscape features for fitness landscape analysis. Soft Comput 20, 3787–3802. https://doi.org/10.1007/s00500-016-2091-4.

Sivakumar, B., Jayawardena, A. W., & Fernando, T. M. K. G. (2002). River flow forecasting: Use of phase-space reconstruction and artificial neural networks approaches. Journal of Hydrology, 265(1-4), 225-245. doi:10.1016/S0022-1694(02)00112-9

Smith-Miles, K., Baatar, D., Wreford, B., and Lewis, R. (2014). Towards objective measures of algorithm performance across instance space. Computers & Operations Research, 45:12–24.

Sorooshian, S., & Gupta, V. K. (1983). Automatic calibration of conceptual rainfall-runoff models: The question of parameter observability and uniqueness. Water Resources Research, 19(1), 260-268. doi:10.1029/WR019i001p00260

Sorooshian, S., Duan, Q., & Gupta, V. K. (1993). Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting Model. Water Resources Research, 29(4), 1185-1194.

doi:10.1029/92WR02617

Steer, K., Wirth, A. and Halgamuge, S. (2008). Information theoretic classification of problems for metaheuristics, in: Proceedings of Simulated Evolution and Learning 2008, Lect. Notes Comput. Sci., vol. 5361, Springer, pp. 319–328.

Stern, H., De Hoedt, G., & Ernst, J. (2000). Objective classification of Australian climates. Australian Meteorological Magazine, 49(2), 87-96.

Suliman, A. H. A., Katimon, A., Darus, I. Z. M., & Shahid, S. (2016). TOPMODEL for Streamflow Simulation of a Tropical Catchment Using Different Resolutions of ASTER DEM: Optimization Through Response Surface Methodology. Water Resources Management, 30(9), 3159-3173. doi:10.1007/s11269-016-1338-2

Sumner, N. R., Fleming, P. M., & Bates, B. C. (1997). Calibration of a modified SFB model for twenty-five Australian catchments using simulated annealing. Journal of Hydrology, 197(1-4), 166-188. doi:10.1016/S0022-1694(96)03277-5

Tan, Q. F., Lei, X. H., Wang, X., Wang, H., Wen, X., Ji, Y., & Kang, A. Q. (2018). An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach. Journal of Hydrology, 567, 767-780. doi:10.1016/j.jhydrol.2018.01.015

Thyer, M., Kuczera, G., & Bates, B. C. (1999). Probabilistic optimization for conceptual rainfall-runoff models: A comparison of the shuffled complex evolution and simulated annealing algorithms. Water Resources Research, 35(3), 767-773. doi:10.1029/1998WR900058

Tikhamarine, Y., Souag-Gamane, D., Ahmed, A. N., Sammen, S. S., Kisi, O., Huang, Y. F., and El-Shafie, A. (2020). Rainfall-runoff modelling using improved machine learning methods: Harris hawks optimizer vs. particle swarm optimization. Journal of Hydrology, 589. doi:10.1016/j.jhydrol.2020.125133.

Tiwari, M. K., & Chatterjee, C. (2010). Development of an accurate and reliable hourly flood forecasting model using wavelet-bootstrap-ANN (WBANN) hybrid approach. Journal of Hydrology, 394(3-4), 458-470. doi:10.1016/j.jhydrol.2010.10.001

Tomassini, M., Vanneschi, L., Collard, P. and Clergue, M. (2005). A study of fitness distance correlation as a difficulty measure in genetic programming, Evol. Comput. 13 (2) pp. 213–239.

Trenouth, W. R., & Gharabaghi, B. (2015). Event-based soil loss models for construction sites. Journal of Hydrology, 524, 780-788. doi:10.1016/j.jhydrol.2015.03.010

Uliana, E. M., da Silva, D. D., Moreira, M. C., & Pereira, D. R. (2019). Global sensitivity analysis methods applied to hydrologic modeling with the SAC-SMA model. Engenharia Agricola, 39(1), 65-74. doi:10.1590/1809-4430-Eng.Agric.v39n1p65-74/2019

van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, M., & Srinivasan, R. (2006). A global sensitivity analysis tool for the parameters of multi-variable catchment models. Journal of Hydrology, 324(1-4), 10-23. doi:10.1016/j.jhydrol.2005.09.008

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., & Teng, J.

(2010). Climate non-stationarity - Validity of calibrated rainfall-runoff models for use in climate change studies. Journal of Hydrology, 394(3-4), 447-457. doi:10.1016/j.jhydrol.2010.09.018

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environmental Modelling and Software, 75, 273-316. doi:10.1016/j.envsoft.2015.08.013.

Wagena, M. B., Goering, D., Collick, A. S., Bock, E., Fuka, D. R., Buda, A., and Easton, Z. M. (2020). Comparison of short-term streamflow forecasting using stochastic time series, neural networks, process-based, and Bayesian models. Environmental Modelling and Software, 126. doi:10.1016/j.envsoft.2020.104669

Wang, P., Zecchin, A. C., Maier, H. R., Zheng, F., & Newman, J. P. (2020). Do Existing Multiobjective Evolutionary Algorithms Use a Sufficient Number of Operators? An Empirical Investigation for Water Distribution Design Problems. Water Resources Research, 56(5). doi:10.1029/2019WR026031

Wang, W., Van Gelder, P., Vrijling, J. K., & Ma, J. (2006). Forecasting daily streamflow using hybrid ANN models. Journal of Hydrology, 324(1-4), 383-399. doi:10.1016/j.jhydrol.2005.09.032

Wolpert, D., and Macready, W. (1997). No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, 1(1):67–82.

Wu, C. L., Chau, K. W., & Li, Y. S. (2009). Methods to improve neural network performance in daily flows prediction. Journal of Hydrology, 372(1-4), 80-93. doi:10.1016/j.jhydrol.2009.03.038

Wu, W., Dandy, G. C., & Maier, H. R. (2014). Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. Environmental Modelling and Software, 54, 108-127. doi:10.1016/j.envsoft.2013.12.016

Wu, W., May, R. J., Maier, H. R., & Dandy, G. C. (2013). A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks. Water Resources Research, 49(11), 7598-7614. doi:10.1002/2012WR012713

Xie, S., Wu, W., Mooser, S., Wang, Q. J., Nathan, R., & Huang, Y. (2021). Artificial neural network based hybrid modeling approach for flood inundation modeling. Journal of Hydrology, 592. doi:10.1016/j.jhydrol.2020.125605

Xiong, L., & O'Connor, K. M. (2000). Analysis of the response surface of the objective function by the optimum parameter curve: How good can the optimum parameter values be? Journal of Hydrology, 234(3-4), 187-207. doi:10.1016/S0022-1694(00)00250-X

Xu, S., Yu, Z., Ji, X., & Sudicky, E. A. (2017). Comparing three models to estimate transpiration of desert shrubs. Journal of Hydrology, 550, 603-615. doi:10.1016/j.jhydrol.2017.05.027

Yazdi, J., and Salehi Neyshabouri, S. A. A. (2014). Identifying low impact development strategies for flood mitigation using a fuzzy-probabilistic approach. Environmental Modelling and Software, 60, 31-44. doi:10.1016/j.envsoft.2014.06.004

Yu, X. Y., & Liong, S. Y. (2007). Forecasting of hydrologic time series with ridge regression in feature space. Journal of Hydrology, 332(3-4), 290-302.

doi:10.1016/j.jhydrol.2006.07.003

Zanetti, S. S., Cecílio, R. A., Silva, V. H., & Alves, E. G. (2015). General calibration of TDR to assess the moisture of tropical soils using artificial neural networks. Journal of Hydrology, 530, 657-666. doi:10.1016/j.jhydrol.2015.10.037

Zealand, C. M., Burn, D. H., & Simonovic, S. P. (1999). Short term streamflow forecasting using artificial neural networks. Journal of Hydrology, 214(1-4), 32-48. doi:10.1016/S0022-1694(98)00242-X

Zecchin, A. C., Simpson, A. R., Maier, H. R. and Nixon, J. B. (2005). Parametric study for an ant algorithm applied to water distribution system optimisation. IEEE Transactions on Evolutionary Computation, 9(2), 175-191, doi: 10.1109/TEVC.2005.844168

Zecchin, A. C., Simpson, A. R., Maier, H. R., Leonard, M., Roberts, A. J., and Berrisford, M. J. (2006). Application of two ant colony optimisation algorithms to water distribution system optimisation. Mathematical and Computer Modelling, 44(5-6), 451-468. doi:10.1016/j.mcm.2006.01.005

Zecchin, A. C., Simpson, A. R., Maier, H. R., Marchi, A. and Nixon, J. B. (2012). Improved understanding of the searching behaviour of ant colony optimization algorithms applied to the water distribution design problem, Water Resources Research, 48(9), doi:10.1029/2011WR011652

Zhang, B., & Govindaraju, R. S. (2003). Geomorphology-based artificial neural networks (GANNs) for estimation of direct runoff over watersheds. Journal of Hydrology, 273(1-4), 18-34. doi:10.1016/s0022-1694(02)00313-x

Zhang, C., Wang, R. B., & Meng, Q. X. (2015). Calibration of conceptual

rainfall-runoff models using global optimization. Advances in Meteorology, 2015. doi:10.1155/2015/545376

Zheng, F. F., Zecchin, A. C., Newman, J. P., Maier, H. R., & Dandy, G. C. (2017). An Adaptive Convergence-Trajectory Controlled Ant Colony Optimization Algorithm With Application to Water Distribution System Design Problems. Ieee Transactions on Evolutionary Computation, 21(5), 773-791. doi:10.1109/tevc.2017.2682899

Zheng, F., Maier, H. R., Wu, W., Dandy, G. C., Gupta, H. V. and Zhang, T. (2018). On lack Of robustness In hydrological model development due to absence of guidelines for selecting calibration and evaluation data: Demonstration for data driven models , Water Resources Research, 54(2), 1013-1030, doi:10.1002/2017WR021470.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. ACM Transactions on Mathematical Software, 23(4), 550-560. doi:10.1145/279232.279236

Zhu, S., Maier, H. R., Zecchin, A.C. (2021). Identification of Metrics Suitable for Determining the Features of Real-World Optimisation Problems. Environmental Modelling and Software. (Submitted Manuscript).

Zhuo, L., Han, D., & Dai, Q. (2016). Soil moisture deficit estimation using satellite multi-angle brightness temperature. Journal of Hydrology, 539, 392-405. doi:10.1016/j.jhydrol.2016.05.052

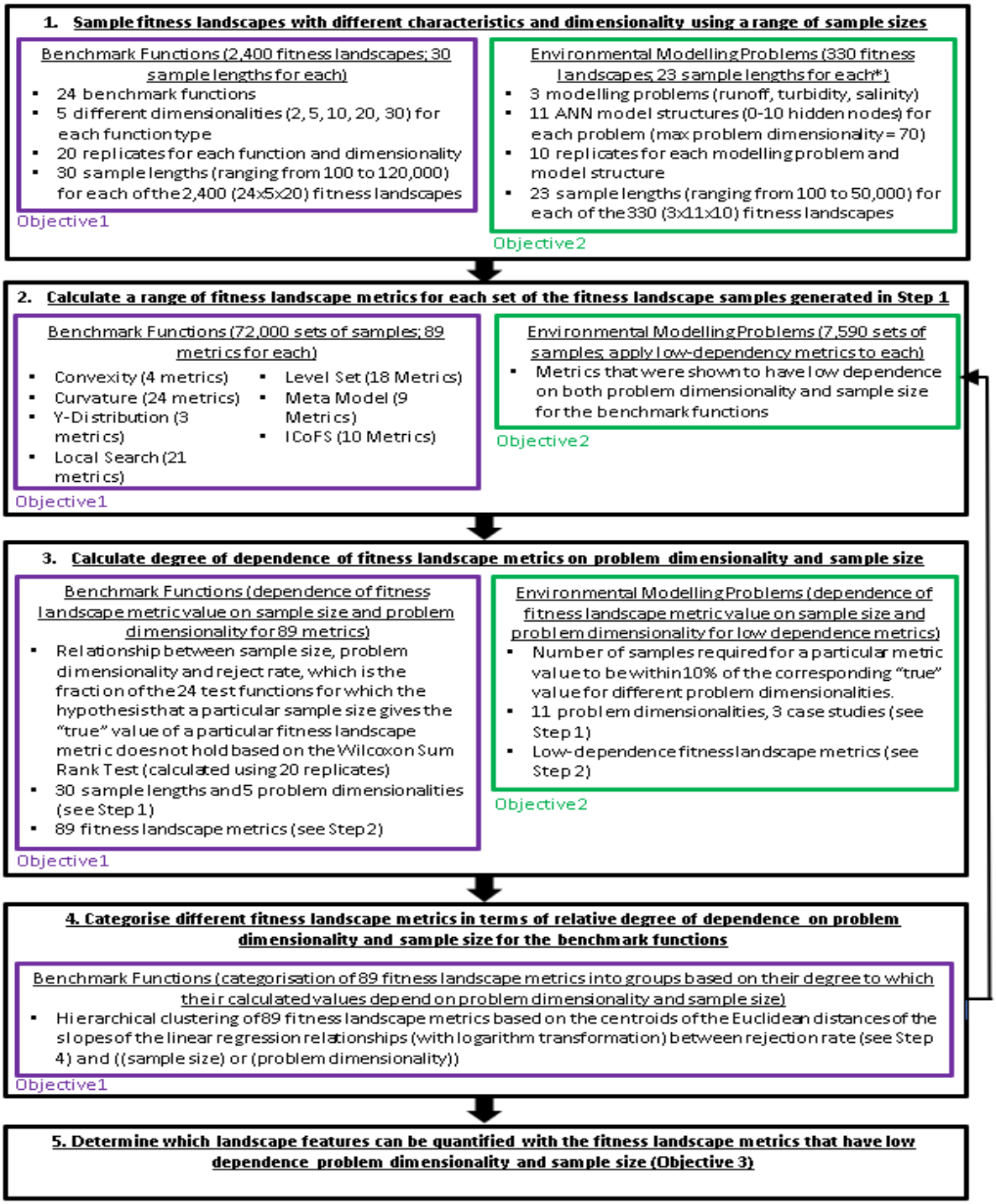
Zounemat-Kermani, M., Kişi, O., Adamowski, J., & Ramezani-Charmahineh, A. (2016). Evaluation of data driven models for river suspended sediment

concentration modeling. Journal of Hydrology, 535, 457-472. doi:10.1016/j.jhydrol.2016.02.012

Zubaidi, S. L., Dooley, J., Alkhaddar, R. M., Abdellatif, M., Al-Bugharbee, H., & Ortega-Martorell, S. (2018). A Novel approach for predicting monthly water demand by combining singular spectrum analysis with neural networks. Journal of Hydrology, 561, 136-145. doi:10.1016/j.jhydrol.2018.0

Appendices of Chapter 2 (Paper 1)

Appendix A: Detailed Outline of Methodology



*** Maximum sample length for local search class metrics is restricted to 3,000 due to high computational effort.**

Figure A.1 Detailed Outline of Methodology

Appendix B: Details of Benchmark Functions

Table B.1 Detailed Features of Benchmark Functions (adapted from Mersmann et al., 2010)

Function	Multi-Modality	Global Structure	Separa-bility	Variable Scaling	Space Homoge-neity	Basin Homoge-neity	Global to local Contrast
1	none	none	high	none	high	none	none
2	none	none	high	high	high	none	none
3	high	strong	none	low	high	low	low
4	high	strong	high	low	high	med	low
5	none	none	high	none	high	none	none
6	none	none	high	low	med	none	none
7	none	none	high	low	high	none	none
8	low	none	none	none	med	low	low
9	low	none	none	none	med	low	low
10	none	none	none	high	high	none	none
11	none	none	none	high	high	none	none
12	none	none	none	high	high	none	none
13	none	none	none	low	med	none	none
14	none	none	none	low	med	none	none
15	high	strong	none	low	high	low	low
16	High	med	none	med	high	med	low
17	high	med	none	low	med	med	high
18	high	med	none	high	med	med	high
19	high	strong	none	none	high	low	low
20	med	weak	none	none	high	low	low
21	med	none	none	med	high	med	low
22	low	none	none	med	high	med	med
23	high	none	none	none	high	low	low
24	high	weak	none	low	high	low	low

***As plateaus is not specified in BBOB suite, it is not included in this table.**

Appendix C: Details of ELA Metrics

A summary of the ELA metrics used is given in Table C.1 and details of how different groups of metrics are calculated and how they assist with the characterisation of different fitness landscape features are given below.

Table C.1 Summary of ELA Metrics used

No.	Metric	Cluster*	Class
1	ela_conv.conv_prob	1	Convexity
2	ela_conv.lin_prob	1	Convexity
3	ela_conv.lin_dev.orig	1	Convexity
4	ela_conv.lin_dev.abs	1	Convexity
5	ela_distr.skewness	1	Y-Distribution
6	ela_distr.kurtosis	1	Y-Distribution
7	ela_distr.number_of_peaks	2	Y-Distribution
8	ela_level.mmce_lda_10	2	Level Set
9	ela_level.mmce_qda_10	3	Level Set
10	ela_level.mmce_mda_10	2	Level Set
11	ela_level.lda_qda_10	3	Level Set
12	ela_level.lda_mda_10	1	Level Set
13	ela_level.qda_mda_10	3	Level Set
14	ela_level.mmce_lda_25	2	Level Set
15	ela_level.mmce_qda_25	3	Level Set
16	ela_level.mmce_mda_25	2	Level Set
17	ela_level.lda_qda_25	3	Level Set
18	ela_level.lda_mda_25	2	Level Set
19	ela_level.qda_mda_25	3	Level Set
20	ela_level.mmce_lda_50	2	Level Set
21	ela_level.mmce_qda_50	3	Level Set
22	ela_level.mmce_mda_50	2	Level Set
23	ela_level.lda_qda_50	3	Level Set
24	ela_level.lda_mda_50	2	Level Set
25	ela_level.qda_mda_50	3	Level Set
26	ela_meta.lin_simple.adj_r2	1	Meta Model
27	ela_meta.lin_simple.intercept	1	Meta Model
28	ela_meta.lin_simple.coef.min	1	Meta Model
29	ela_meta.lin_simple.coef.max	1	Meta Model
30	ela_meta.lin_simple.coef.max_by_min	1	Meta Model
31	ela_meta.lin_w_interact.adj_r2	2	Meta Model

32	ela_meta.quad_simple.adj_r2	1	Meta Model
33	ela_meta.quad_simple.cond	2	Meta Model
34	ela_meta.quad_w_interact.adj_r2	3	Meta Model
35	nn.ic.h.max	1	ICoFS
36	nn.ic.eps.s	1	ICoFS
37	nn.ic.eps.max	1	ICoFS
38	nn.ic.eps.ratio	1	ICoFS
39	nn.ic.m0	1	ICoFS
40	rand.ic.h.max	1	ICoFS
41	rand.ic.eps.s	1	ICoFS
42	rand.ic.eps.max	1	ICoFS
43	rand.ic.eps.ratio	1	ICoFS
44	rand.ic.m0	1	ICoFS
45	ela_curv.grad_norm.min	4	Curvature
46	ela_curv.grad_norm.lq	UC	Curvature
47	ela_curv.grad_norm.mean	UC	Curvature
48	ela_curv.grad_norm.med	UC	Curvature
49	ela_curv.grad_norm.uq	UC	Curvature
50	ela_curv.grad_norm.max	4	Curvature
51	ela_curv.grad_norm.sd	UC	Curvature
52	ela_curv.grad_norm.nas	1	Curvature
53	ela_curv.grad_scale.min	4	Curvature
54	ela_curv.grad_scale.lq	UC	Curvature
55	ela_curv.grad_scale.mean	UC	Curvature
56	ela_curv.grad_scale.med	UC	Curvature
57	ela_curv.grad_scale.uq	UC	Curvature
58	ela_curv.grad_scale.max	4	Curvature
59	ela_curv.grad_scale.sd	UC	Curvature
60	ela_curv.grad_scale.nas	1	Curvature
61	ela_curv.hessian_cond.min	4	Curvature
62	ela_curv.hessian_cond.lq	UC	Curvature
63	ela_curv.hessian_cond.mean	UC	Curvature
64	ela_curv.hessian_cond.med	UC	Curvature
65	ela_curv.hessian_cond.muq	UC	Curvature
66	ela_curv.hessian_cond.max	1	Curvature
67	ela_curv.hessian_cond.sd	UC	Curvature
68	ela_curv.hessian_cond.nas	1	Curvature
69	ela_local.n_loc_opt.abs	5	Local Search
70	ela_local.n_loc_opt.rel	2	Local Search
71	ela_local.best2mean_contr.orig	5	Local Search
72	ela_local.best2mean_contr.ratio	5	Local Search
73	ela_local.center.dist_min	5	Local Search

74	ela_local.center.dist_lq	1	Local Search
75	ela_local.center.dist_mean	1	Local Search
76	ela_local.center.dist_median	1	Local Search
77	ela_local.center.dist_uq	1	Local Search
78	ela_local.center.dist_max	5	Local Search
79	ela_local.center.dist_sd	1	Local Search
80	ela_local.basin_sizes.avg_best	5	Local Search
81	ela_local.basin_sizes.avg_non_best	5	Local Search
82	ela_local.basin_sizes.avg_worst	5	Local Search
83	ela_local.fun_evals.min	1	Local Search
84	ela_local.fun_evals.lq	1	Local Search
85	ela_local.fun_evals.mean	1	Local Search
86	ela_local.fun_evals.median	1	Local Search
87	ela_local.fun_evals.uq	1	Local Search
88	ela_local.fun_evals.max	1	Local Search
89	ela_local.fun_evals.sd	1	Local Search

***UC in Cluster column represents metrics not classified in this study.**

C1. Convexity Metrics:

As can be seen in Table 2.3, convexity metrics are able to provide information on a number of fitness landscape features, including global structure, multimodality and search space homogeneity. Their calculation requires implementation of the following general steps:

(i) Select random pairs of points (x_i, x_j) from the total number of samples of the fitness landscape considered (i.e. 100 to 120,000 samples, see Figure 2.1). To ensure most of the samples are included in the calculation, we use *n* random pairs of points in this study, where *n* is the number of initial samples.

(ii) Calculate a linear combination of (x_i, x_j) to select a new point x_n between the two points, where

$$x_n = w \cdot x_i + (1 - w) \cdot x_j \quad (C1)$$

where w is a random number between 0 and 1. Calculate the fitness value y_n of x_n based on the corresponding test function.

(iii) Calculate the fitness values (y_i, y_j) of (x_i, x_j) based on the corresponding test function. Use linear regression to calculate the approximated linear fitness value y_n' at x_n,

$$y_n' = w \cdot y_i + (1 - w) \cdot y_j \quad (C2)$$

where w is the same w as in (ii).

(iv) Calculate the difference (Δ) between y_n and y_n' by $\Delta = y_n - y_n'$.

a. If Δ is negative, the landscape between the selected two points is convex, providing good gradient information to guide the search in this region of the fitness landscape.
b. If Δ is positive, the landscape between the selected two points is not convex, providing poor gradient information to guide the search in this region of the fitness landscape.

(v) In total, 4 convexity metrics are considered, which differ in terms of statistics methods to summarise the results of Δ obtained from n pairs of samples. They are the probability of convexity (*ela_conv.conv_prob*), which relates to the probability of negative Δ; probability of linearity (*ela_conv.lin_prob*), which relates to the probability of $\Delta = 0$; mean original deviation (*ela_conv.lin_dev.orig*),

which relates to the mean value of Δ from n pairs of samples; mean absolute deviation (*ela_conv.lin_dev.abs*), which relates to the mean value of $|\Delta|$ from n pairs of samples.

Convexity metrics are able to provide information on the global structure of fitness landscapes as they present information about the general shape of fitness landscapes and can therefore provide information on whether fitness landscapes have a clear structure to guide searching or not. They are also able to provide information on search space homogeneity and multimodality as they take the probability of convexity into account. A high or low convexity rate indicates fitness landscapes maintain the same trend and shape in most areas of the fitness landscape, which is representative of greater homogeneity and reduced multimodality. In contrast, middle-range values of the convexity rate indicate that fitness landscapes have different trends and shapes in different areas, increasing changes of inhomogeneity and multi-modality.

B2. y-Distribution Metrics:

As can be seen in Table 2.3, y-distribution metrics are able to provide information on a number of fitness landscape features, including global structure, multimodality and search space homogeneity. Their calculation requires implementation of the general following steps:

(i) Generate the PDF of the fitness values Y^s of samples X^s.

(ii) In total, 3 y-distribution metrics were considered, which relates to the properties of the PDF of Y^s including skewness (*ela_distr.skewness*),

kurtosis (*ela_distr.kurtosis*), and the number of peaks (*ela_distr.number_of_peaks*).

Y-distribution metrics are able to provide information on the global structure of fitness landscapes, for example, if the skewness of the PDF is negative, most of the obtained y are small, indicating that the bottom region of a fitness landscape is bigger than the top region, referring to a bigger "bowl" bottom than for a fitness landscapes with a positive skewness. They are also able to provide information on multimodality, as a multi-modal fitness landscapes are likely to have several peaks in their PDFs, which refers to different bottom regions of the fitness landscape. Additionally, y-distribution metrics are able to provide information on the prevalence of plateaus within the landscape, as plateau-like landscapes contain region(s) with the same fitness values, as a result, they would tend to have high kurtosis values, which indicates that most of the fitness values have no significant difference.

C3. Level Set Metrics:

As can be seen in Table 2.3, level set metrics are able to provide information concerning a number of fitness landscape features, including global structure, multimodality and plateaus. Their calculation requires implementation of the following steps:

(i) Split all the samples to high-quality and low-quality ones based on a given quantile threshold of their fitness values Y^s. In this study, 10%, 25% and 50% quantiles are used as thresholds.

(ii) Linear (LDA), quadratic (QDA) and mixture (MDA) discriminant analysis are used to predict whether the fitness values Y^s are high or

low-quality. The number of Y^s which are classified to a wrong quality group are recorded.

(iii) Calculate the mean misclassification error (MMCE), which refers to the probability of misclassification of Y^s by using corresponding discriminant analysis methods.

(iv) In total, 18 level set metrics are considered, which differ in terms of quantile thresholds and discriminant analysis methods (i.e. *ela_level.mmce_*{*lda, qda, mda*}*_*{*10, 25, 50*}). The quotient of MMCE of LDA divided by MMCE of QDA (*ela_level.lda_qda*}*_*{*10, 25, 50*}), the quotient of MMCE of LDA divided by MMCE of MDA (*ela_level.lda_mda*}*_*{*10, 25, 50*}), and the quotient of MMCE of QDA divided by MMCE of MDA (*ela_level.qda_mda*}*_*{*10, 25, 50*}) are also included in the metrics, as they can show the differences of MMCE between simple models (LDA and QDA) and complex models (MDA).

Level set metrics are able to provide information on global structure and multimodality, as through the MMCE of different discriminant analysis, the distribution of fitness values can be determined. For example, if MMCEs of LDA and QDA are low and MMCEs of MDA are high, fitness values on fitness landscapes can be easily classified by these two simple models, indicating high-quality values and low-quality values are not located in the same region, so that high-quality values can be easily identified, as finding these values will not be interrupted by low-quality values in this case. As differences between fitness values are not big in the small region, the landscapes should not contain multiple optima. In contrast, if MMCEs of LDA and QDA are high, the global

structure of a fitness landscape is likely to be complex, as this indicates that there are no clear "top" and "bottom" regions of the fitness landscapes, but very frequent variation in fitness values. This can also result in a high level of multimodality of fitness landscapes. Furthermore, if MMCEs of MDA are also high, the structure of fitness landscapes can be quite complex and multi-modal. Level set metrics are also able to provide information on plateaus. As plateau-like landscapes have many similar fitness values, the threshold of high and low-quality values is not clear for such problems. As a result, plateau-like landscapes are more likely to have high MMCEs for all discriminant analysis methods.

C4. Meta Model Metrics:

As can be seen in Table 2.3, meta model metrics are able to provide information on a number of fitness landscape features, including global structure, multimodality, plateaus and variable scaling. Their calculation requires implementation of the general following steps:

(i) Build the corresponding regression models between samples X^s and corresponding fitness values Y^s. Four regression models are built in this study, which are:

Simple linear regression:

$$\bar{y} = \sum_{i=1}^{n} a_i v_i \qquad (C3)$$

where a_i is the coefficient of corresponding variable v_i.

Interacted linear regression:

$$\bar{y} = \sum_{i=1}^{n} a_i v_i + \sum_{i=1}^{n} \sum_{j=1}^{n} b_k v_i \, v_j \qquad (C4)$$

where $b_k$is the coefficient of corresponding variable $v_i v_j$.

Simple quadratic regression:

$$\bar{y} = \sum_{i=1}^{n} a_i v_i + \sum_{i=1}^{n} c_i v_i^2 \quad (C5)$$

where c_i is the coefficient of corresponding variable v_i^2, which is the square of v_i.

Interacted quadratic regression:

$$\bar{y} = \sum_{i=1}^{n} a_i v_i + \sum_{i=1}^{n} c_i v_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} b_k v_i v_j + \sum_{i=1}^{n} \sum_{j=1}^{n} t_k v_i v_j^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} l_k v_i^2 v_j^2 \quad (C6)$$

where $t_k$and l_k are the coefficient of corresponding variable $v_i v_j^2$ and $v_i^2 v_j^2$, respectively.

(ii) In total, 9 meta model metrics are considered, which are related to adjusted coefficients of determination R^2 (*ela_meta*.{*lin,quad*}_{*simple,w_interact*}.*adj_r2*) of four regression models, maximum, minimum and intercept coefficients of simple linear regression models (*ela_meta.lin_simple*.{*coef.max*, *coef.min*, *intercept*}) and the quotient between maximum and minimum coefficients (*ela_meta*.{*lin,quad*}*_simple.cond*) of simple linear and simple quadratic regression models.

Meta model metrics are able to provide information on the global structure and multimodality of fitness landscapes as adjusted R^2 can show how well global structure matches the corresponding models, and a goodness-of-fit shown by R^2 also indicates models with a low-level of multimodality, as all these regression models are not multi-modal. They are also able to provide information on separability as separate fitness landscapes are likely to have

higher adjusted R^2 as they are likely to be represented more easily by simpler models. Additionally, meta model metrics are able to provide information on variable scaling, as shown by the maximum and minimum of coefficients of the models. Low-degree variable scaling fitness landscapes should have models with maximum and minimum coefficients the values of which are close to each other, indicating that all variables make similar contributions to the fitness values. On the other hand, High-degree variable scaling fitness landscapes should have models with significant different maximum and minimum coefficients, indicating that the contributions of different variables to the fitness values are not the same.

C5. Local Search Metrics:

As can be seen in Table 2.3, local search metrics are able to provide information on a number of fitness landscape features including multimodality, global to local optima contrast, basin size homogeneity and search space homogeneity. Their calculation requires implementation of the general following steps:

(i) Use a gradient algorithm to find local optima starting from initial samples X^s. In this study, the L-BFGS-B algorithm (Zhu et al., 1997) was used due to its capacity to setup the range of calculation to avoid the identified local optima being beyond the range of the fitness landscape.

(ii) Use of hierarchical clustering to cluster identified local optima in (i). Local optima within a given Euclidean distance e are included in the same cluster, which refers to a corresponding local basin. In this study, e is 5% of total Euclidean distance length of the whole fitness

landscape, as this distance performs well in distinguishing different clusters without resulting in a computational burden that results in intractability (if e is too small, a larger number of clusters is likely to be generated by hierarchical clustering, which increases complexity and the computational requirements of subsequent calculations).

(iii) Calculate the centroid X_c of each local basin identified in (ii), based on the local optima in the corresponding basin. Calculate the fitness value Y_c of all centroids.

(iv) In total, 21 local search metrics are considered, which relates to the number of identified local basins (*ela_local.n_loc_opt.*{*abs*, *rep*}), fitness value differences between high-quality basins (global optima) and low-quality basins (local optima) (*ela_local.best2mean_contr*.{*orig*, *ratio*}), basin size difference between high, average and low-quality basins (i.e. difference between number of optima in high and low-quality basins) (*ela_local.basin_sizes*.{*avg_best*, *avg_non_best*, *avg_worst*}), statistics of basin centroids Euclidean distances and statistics of number of evaluated functions to find optima from initial samples (i.e. minimum, maximum, lower quantile, median, mean, upper quantile and standard deviation of basin centroids Euclidean distances and number of evaluated functions) (*ela_local*.{*center.dist*, *fun_evals*}_{*min*, *max*, *lq*, *median*, *mean*, *uq*, *sd*}).

Local search metrics are able to provide information on multimodality, as the level of multimodality is highly related to the identified number of optima by

using a gradient algorithm. They are also able to provide information on global to local optima contrast, as fitness landscapes with a high degree of global and local contrast are likely to have significantly different fitness values between global and local basins/optima and vice versa. Additionally, local search metrics are able to provide information on basin size homogeneity, as they can present the size difference between high and low-quality basins, in order to check whether basins have the same quality. On the other hand, the evaluated number of functions can also show the depth of different basins. Finally, they are able to provide information on search space homogeneity, as they are able to show the distribution of centroids on fitness landscapes. The distance between centroids can indicate whether basins are converged to a small region or widely distributed on the whole fitness landscape, referring to whether different regions in search space have the same feature.

C6. Curvature Metrics:

As can be seen in Table 2.3, curvature metrics are able to provide information on a number of fitness landscape features including plateaus and variable scaling. Their calculation requires implementation of the general following steps:

(i) Calculate the gradient information

$$f'(x_i) = df(x_i)/dx \qquad (C7)$$

where $f(x)$ is the fitness function and $f'(x_i)$ is the first order derivative of f at the variable x_i. Based on $f'(x)$ in all directions, the total gradient length L of a sample is calculated as

$$L = \sqrt{\sum_{i=1}^{n} \left(f'(x_i)\right)^2} \qquad (C8)$$

where n is the number of variables (dimensions) of the fitness function. The gradient condition C_G of a sample is calculated as

$$C_G = \max\{f'(x)\}/\min\{f'(x)\} \qquad (C9)$$

(ii) Calculate the Hessian matrix

$$H(X) = \partial^2 f(x)/\partial x_i \partial x_j \qquad (C10)$$

so the eigenvalues $\boldsymbol{\lambda}$ of $H(X)$ can be calculated from

$$H(X) - \boldsymbol{\lambda} \cdot \boldsymbol{I}_n = \mathbf{0} \qquad (C11)$$

where $\boldsymbol{I}_n$ is the identity matrix with size n. The Hessian condition C_H of a sample is calculated as

$$C_H = \max\{\boldsymbol{\lambda}\}/\min\{\boldsymbol{\lambda}\} \qquad (C12)$$

(iii)In total, 24 curvature metrics are considered, which differ in terms of 8 statistics of L, C_G and C_H of all samples, which are the minimum, maximum, lower quantile, median, mean, upper quantile, standard deviation and proportion of samples with no L, C_G and C_H (i.e. *ela_curv*.{*grad_norm*, *grad_scale*, *hessian_cond*}.{*min*, *max*, *lq*, *med*, *mean*, *uq*, *sd*, *nas*})

Curvature metrics are able to provide information on the plateaus of fitness landscapes, as they refer to the gradient information of fitness landscapes. Plateau-like fitness landscapes contain limited gradient information, resulting in small L of all samples in general. They are also able to provide information on variable scaling, as C_G and C_H can show the differences of contribution between variables to the fitness values. Large C_G and C_H values generally

indicate that there are variables which have very small contributions to the fitness values, prviding little guidance to the search algorithm.

C7. ICoFS Metrics:

As can be seen in Table 2.3, ICoFS metrics are able to provide information on a number of fitness landscape features, including global structure, multimodality and plateaus. Their calculation requires implementation of the following general steps:

(i) Firstly sort all samples into a sequence. In this study, two sampling ordering methods are used to generate different ICoFS metrics: (1) nearest neighbouring (nn), as part of which the following sample x_{i+1} of one sample x_i is the closest sample to the corresponding by Euclidean distance; (2) random (rand) order, as part of which the following sample x_{i+1} of one sample x_i is randomly selected from the entire set of samples.

(ii) Build a symbol sequence $\emptyset(\epsilon)$ by using the following rule:

$$\emptyset_i = \begin{cases} -1, if\ y_{i+1} - y_i < -\epsilon \\ 0, if\ |\ y_{i+1} - y_i| <\ \epsilon \\ 1, if\ y_{i+1} - y_i >\ \epsilon \end{cases} \qquad (C13)$$

where $\epsilon \geq 0$ is the accuracy parameter of the symbol sequence and y_i is the fitness value of sample x_i. It can be seen that $\emptyset(\epsilon)$ is controlled by the value of ϵ. If ϵ is small, $\emptyset(\epsilon)$ can be quite sensitive and contain frequent symbol changes in the sequence (for example sequence [-1, 1, 1, -1, 1]). If ϵ is big, on the other hand, $\emptyset(\epsilon)$ can be insensitive and contain many 0 values in the sequence (for example sequence [0, 0, 0, 0, 0]).

(iii)Calculate the information content $H(\epsilon)$ of the sequence based on

the definition:

$$\boldsymbol{H(\epsilon) = -\sum_{a \neq b} P_{ab} log_6 P_{ab}} \quad (\boldsymbol{C14})$$

where $a, b \in \{-1, 0, 1\}$ and P_{ab} is the probability that two neighboured symbols a, b are different.

(iv) Build a new sequence $\emptyset'(\epsilon)$ by removing all 0 values in $\emptyset(\epsilon)$, and calculate the partial information content $M(\epsilon)$, which is defined as:

$$\boldsymbol{M(\epsilon) = |\emptyset'|/(n-1)} \quad (\boldsymbol{C15})$$

where n is the length of sequence $\emptyset(\epsilon)$.

(v) In total, 10 ICoFS metrics are considered and both of two sample orders contain 5 metrics. The typical result curves of $H(\epsilon)$ and $M(\epsilon)$ against $\epsilon$are shown in Figure C.1. Munoz et al., (2015b) provides 5 metrics to summarise the curves, which are

$$H_{max} = \max\{H(\epsilon)\} \quad (C16)$$

$$\epsilon_{max} = log_{10}\epsilon\,, where\ H(\epsilon) = H_{max} \quad (C17)$$

$$\epsilon_s = log_{10} \min(\epsilon)\,, where\ H(\epsilon) < 0.05 \quad (C18)$$

$$M_0 = M(\epsilon = 0) \quad (C19)$$

$$\epsilon_{ratio} = log_{10} \max\{\epsilon\}, where\ M(\epsilon) > 0.5M_0 \quad (C20)$$

The detailed dot plots of the 5 metrics are also shown in Figure C.1. The total 10 metrics are shown as format {*nn, rand*}.*ic*.{*h.max, eps.max, eps.s, m0, eps.ratio*}.

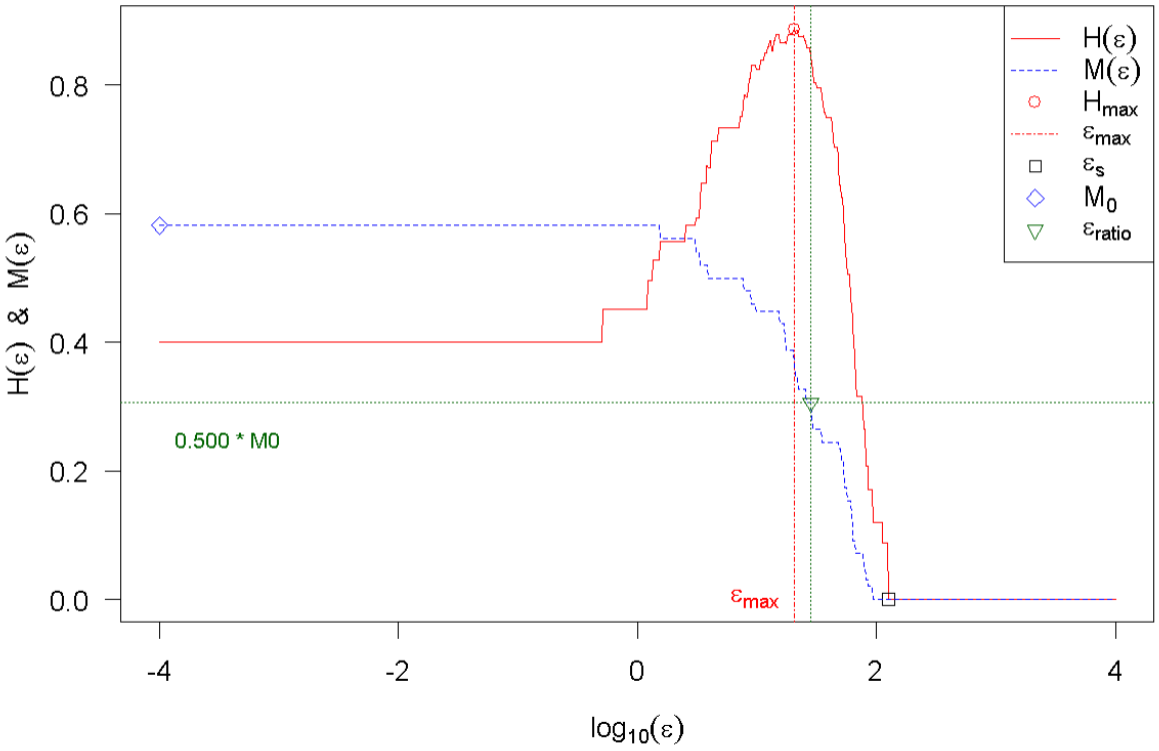


Figure C.1 Typical Results of ICoFS

ICoFS metrics are able to provide information on multimodality, as the symbols in the sequence represent information about the smoothness of fitness landscapes. Rough landscapes are likely to have high values of H_{max} and M_0, and if landscapes are rough, they have the potential to have a high-degree of multimodality. Additionally, they are able to provide information on plateaus and global structure, as plateau-like landscapes should contain many 0s in their symbol sequence. Even when ϵ is very small, this is likely to return a small ϵ_s, ϵ_{max} and ϵ_{ratio} for a plateau-like landscape. In contrast, fitness landscapes with good global structure should have a level of scaling in terms of fitness values, as a result, small ϵ_s, ϵ_{max} and ϵ_{ratio} should be bigger than those for flat fitness landscapes.

Appendix D: Slopes and Coefficients of Determination (r^2) of Benchmark Test for All Tested Metrics

Table D.3 Slope of dimension, sample size and r^2 of plane regression for all metrics in different classes and clusters (unclassified metrics are not included)

Metric Number	Metric	Slope of Dim	Slope of Sample Size	Plane r^2	Cluster	Class
1	ela_conv.conv_prob	0.00	0.02	0.39	1	Convexity
2	ela_conv.lin_prob	0.01	0.00	0.41	1	Convexity
3	ela_conv.lin_dev.orig	0.01	0.01	0.11	1	Convexity
4	ela_conv.lin_dev.abs	0.00	0.01	0.12	1	Convexity
5	ela_distr.skewness	0.02	0.05	0.70	1	Y-Distribution
6	ela_distr.kurtosis	0.05	0.07	0.79	1	Y-Distribution
7	ela_distr.number_of_peaks	0.13	0.08	0.70	2	Y-Distribution
8	ela_level.mmce_lda_10	0.12	0.10	0.66	2	LevelSet
9	ela_level.mmce_qda_10	0.26	0.12	0.85	3	LevelSet
10	ela_level.mmce_mda_10	0.12	0.12	0.70	2	LevelSet
11	ela_level.lda_qda_10	0.27	0.12	0.84	3	LevelSet
12	ela_level.lda_mda_10	0.05	0.08	0.57	1	LevelSet
13	ela_level.qda_mda_10	0.27	0.11	0.82	3	LevelSet
14	ela_level.mmce_lda_25	0.13	0.11	0.76	2	LevelSet
15	ela_level.mmce_qda_25	0.24	0.13	0.84	3	LevelSet
16	ela_level.mmce_mda_25	0.12	0.16	0.86	2	LevelSet
17	ela_level.lda_qda_25	0.25	0.12	0.82	3	LevelSet
18	ela_level.lda_mda_25	0.08	0.13	0.81	2	LevelSet
19	ela_level.qda_mda_25	0.25	0.10	0.73	3	LevelSet
20	ela_level.mmce_lda_50	0.13	0.11	0.80	2	LevelSet

21	ela_level.mmce_qda_50	0.24	0.10	0.80	3	LevelSet
22	ela_level.mmce_mda_50	0.12	0.15	0.86	2	LevelSet
23	ela_level.lda_qda_50	0.28	0.09	0.80	3	LevelSet
24	ela_level.lda_mda_50	0.08	0.12	0.82	2	LevelSet
25	ela_level.qda_mda_50	0.24	0.09	0.82	3	LevelSet
26	ela_meta.lin_simple.adj_r2	0.01	0.04	0.49	1	MetaModel
27	ela_meta.lin_simple.intercept	0.00	0.01	0.16	1	MetaModel
28	ela_meta.lin_simple.coef.min	0.08	0.08	0.79	1	MetaModel
29	ela_meta.lin_simple.coef.max	0.08	0.05	0.74	1	MetaModel
30	ela_meta.lin_simple.coef.max_by_min	0.07	0.07	0.78	1	MetaModel
31	ela_meta.lin_w_interact.adj_r2	0.13	0.10	0.62	2	MetaModel
32	ela_meta.quad_simple.adj_r2	0.02	0.05	0.46	1	MetaModel
33	ela_meta.quad_simple.cond	0.16	0.09	0.83	2	MetaModel
34	ela_meta.quad_w_interact.adj_r2	0.19	0.11	0.79	3	MetaModel
35	nn.ic.h.max	0.01	0.06	0.42	1	ICoFS
36	nn.ic.eps.s	0.07	0.08	0.54	1	ICoFS
37	nn.ic.eps.max	0.03	0.06	0.50	1	ICoFS
38	nn.ic.eps.ratio	0.06	0.06	0.42	1	ICoFS
39	nn.ic.m0	0.03	0.04	0.24	1	ICoFS
40	rand.ic.h.max	0.03	0.08	0.81	1	ICoFS
41	rand.ic.eps.s	0.01	0.02	0.19	1	ICoFS
42	rand.ic.eps.max	0.01	0.01	0.14	1	ICoFS
43	rand.ic.eps.ratio	0.00	0.03	0.22	1	ICoFS
44	rand.ic.m0	0.00	0.01	0.11	1	ICoFS
45	ela_curv.grad_norm.min	0.01	0.12	0.69	4	Curvature
50	ela_curv.grad_norm.max	0.01	0.12	0.64	4	Curvature
52	ela_curv.grad_norm.nas	0.00	0.00	NA	1	Curvature
53	ela_curv.grad_scale.min	0.01	0.12	0.68	4	Curvature

58	ela_curv.grad_scale.max	0.01	0.11	0.64	4	Curvature
60	ela_curv.grad_scale.nas	0.04	0.00	0.70	1	Curvature
61	ela_curv.hessian_cond.min	0.01	0.12	0.68	4	Curvature
66	ela_curv.hessian_cond.max	0.00	0.06	0.41	1	Curvature
68	ela_curv.hessian_cond.nas	0.01	0.01	0.41	1	Curvature
69	ela_local.n_loc_opt.abs	0.17	0.04	0.70	5	LocalSearch
70	ela_local.n_loc_opt.rel	0.06	0.12	0.62	2	LocalSearch
71	ela_local.best2mean_contr.orig	0.14	0.05	0.61	5	LocalSearch
72	ela_local.best2mean_contr.ratio	0.12	0.06	0.30	5	LocalSearch
73	ela_local.center.dist_min	0.15	0.05	0.70	5	LocalSearch
74	ela_local.center.dist_lq	0.07	0.04	0.33	1	LocalSearch
75	ela_local.center.dist_mean	0.08	0.04	0.36	1	LocalSearch
76	ela_local.center.dist_median	0.08	0.04	0.35	1	LocalSearch
77	ela_local.center.dist_uq	0.07	0.04	0.34	1	LocalSearch
78	ela_local.center.dist_max	0.15	0.05	0.73	5	LocalSearch
79	ela_local.center.dist_sd	0.09	0.02	0.30	1	LocalSearch
80	ela_local.basin_sizes.avg_best	0.16	0.03	0.73	5	LocalSearch
81	ela_local.basin_sizes.avg_non_best	0.17	0.03	0.70	5	LocalSearch
82	ela_local.basin_sizes.avg_worst	0.18	0.04	0.76	5	LocalSearch
83	ela_local.fun_evals.min	0.10	0.07	0.69	1	LocalSearch
84	ela_local.fun_evals.lq	0.02	0.03	0.35	1	LocalSearch
85	ela_local.fun_evals.mean	0.01	0.00	0.01	1	LocalSearch
86	ela_local.fun_evals.median	0.00	0.02	0.30	1	LocalSearch
87	ela_local.fun_evals.uq	0.00	0.02	0.18	1	LocalSearch
88	ela_local.fun_evals.max	0.02	0.07	0.33	1	LocalSearch
89	ela_local.fun_evals.sd	0.00	0.02	0.11	1	LocalSearch

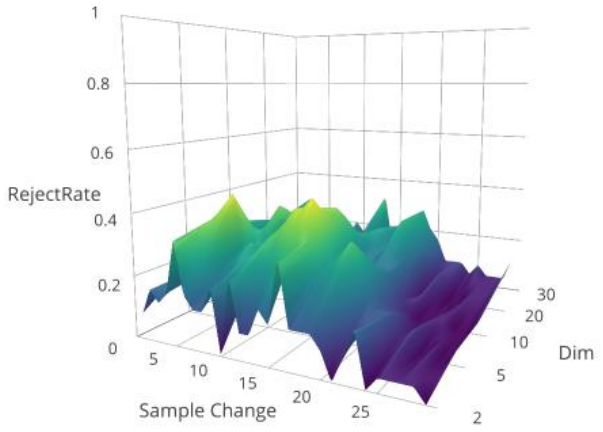


Figure D.4 Reject rate plots of metrics with low r^2

Appendices of Chapter 4 (Paper 3)

Appendix A – Details of CRR Models

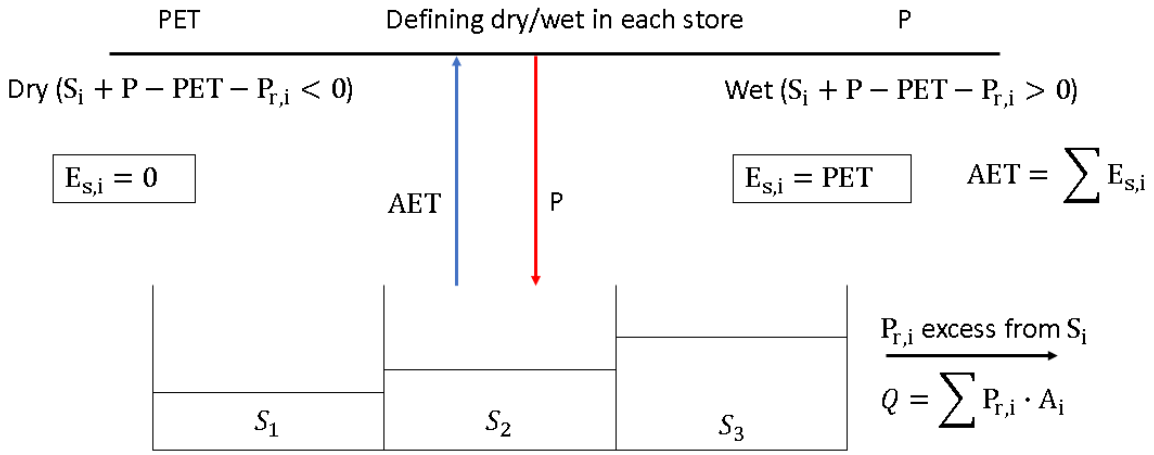


Figure A.1 – Model Structure of AWBM (adapted from Boughton et al., (2004))

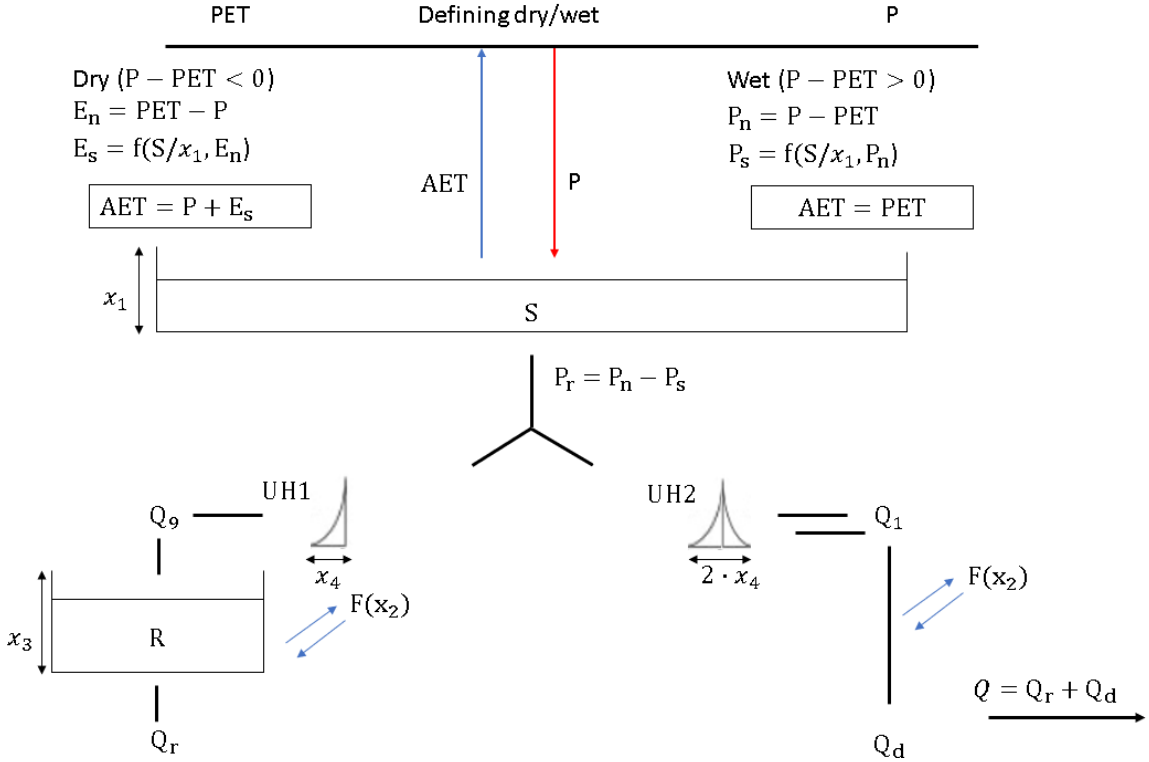


Figure A.2 – Model Structure of GR4J (adapted from Perrin et al., (2003))

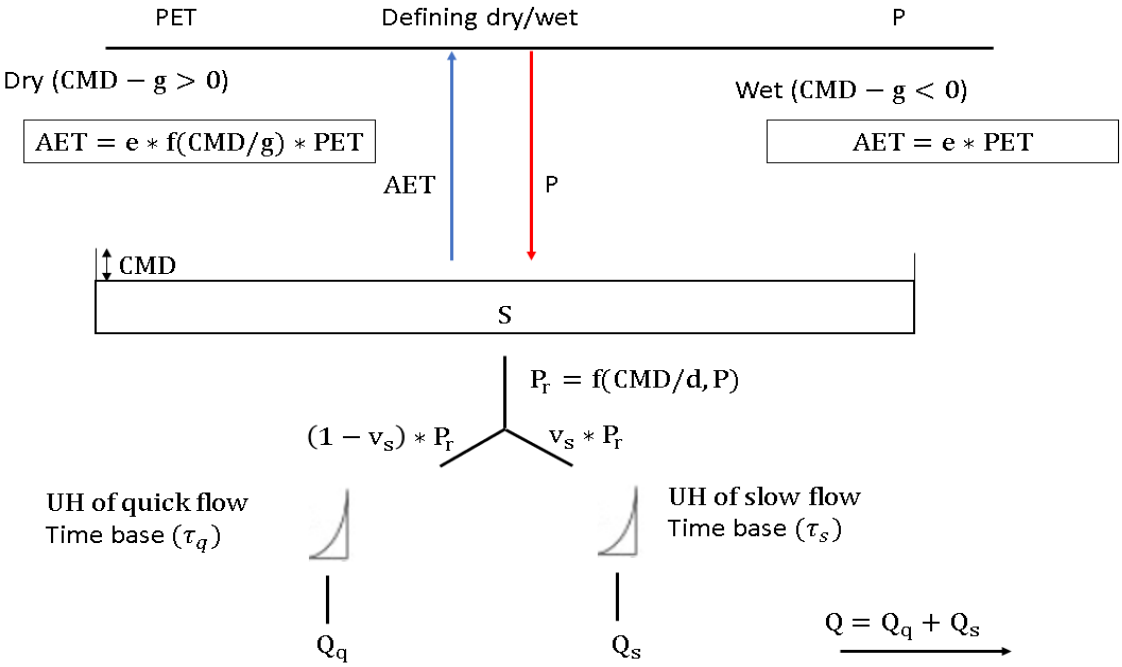


Figure A.3 – Model Structure of IHACRES-CMD (adapted from Croke and Jackman (2004) and Jackman et al., (1990))

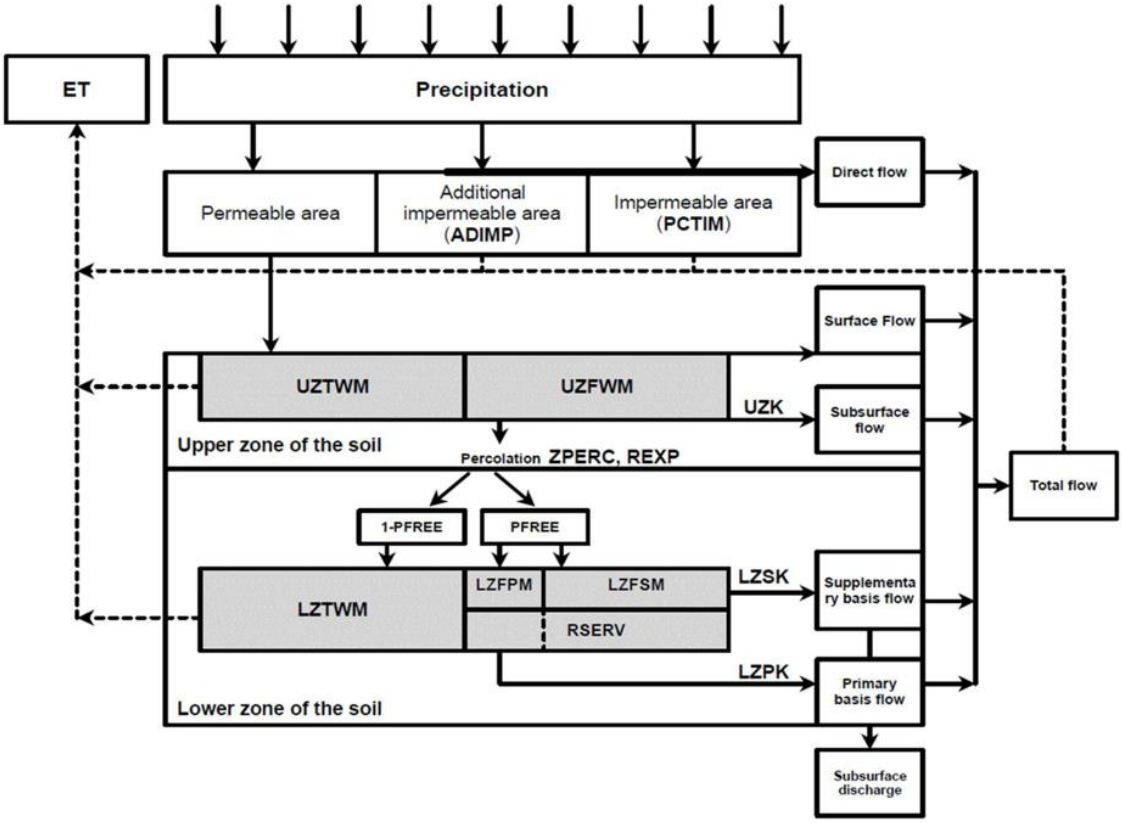


Figure A.4 – Model Structure of Sacramento (adapted Uliana et al., 2019)

Table A.1 CRR Model Parameters and Corresponding Ranges

Model	**Parameter**	**Description**	**Range**
AWBM	*cap.ave*	average soil water storage capacity (mm)	1—1000
	emult	multiplier for the PET	0.01—1
GR4J	x_1	maximum capacity of the production store (mm)	100—1200
	x_2	groundwater exchange coefficient (mm)	-5—3
	x_3	one day ahead maximum capacity of the routing store (mm)	20—300
	x_4	time base of unit hydrograph UH1 (time steps)	1.1—2.9
IHACRES	*f*	CMD stress threshold as a proportion of *d*	0.01—3
	e	temperature to PET conversion factor	0.01—1.5
	d	CMD threshold for producing flow	50—550
	v_s	fractional volumes for the slow flow	0—1
	$\boldsymbol{\tau_s}$	time base of unit hydrograph of slow flow	10—350
	$\boldsymbol{\tau_q}$	time base of unit hydrograph of quick flow	0.5—10
Sacramento	*uztwm*	upper zone tension water maximum capacity (mm)	1—150
	uzfwm	upper zone free water maximum capacity (mm)	1—150
	uzk	lateral drainage rate of upper zone free water expressed as a fraction of contents per day	0.1—0.5
	pctim	the fraction of the catchment which produces impervious runoff during low flow conditions	10^{-6}—0.1

	adimp	the additional fraction of the catchment which exhibits impervious characteristics when the catchment's tension water requirements are met	0—0.4
	zperc	maximum percolation rate coefficient	1—250
	rexp	the rate of change of the percolation rate with changing lower zone water contents	0—5
	lztwm	lower zone tension water maximum capacity (mm)	1—500
	lzfsm	lower zone supplemental free water maximum capacity (mm)	1—1000
	lzfpm	lower zone primary free water maximum capacity (mm)	1—1000
	lzsk	lateral drainage rate of lower zone supplemental free water	0.01—0.25
	lzpk	lateral drainage rate of lower zone primary free water	10^{-4}—0.25
	pfree	direct percolation fraction from upper to lower zone free water	0—0.6

Appendix B – Results of Clustering of ELA Metrics for the 420 Error Surfaces Considered

Table B.1 Metric results of H_{max} (metric related to roughness) and corresponding degree of relative roughness and score (cluster) for all error surfaces (as cases with same data length have very close results, the average of 1-year and 5-year data length results of corresponding cases are presented instead of each single case)

Catchment Location	Model	Data Length	Error Metric	Metric Result	Degree of Relative Roughness	Score (Cluster)
Burnie	AWBM	1_year	NS	0.64	Low	2
Burnie	AWBM	5_year	NS	0.71	Medium	3
Burnie	AWBM	10_year	NS	0.72	Medium	3
Darwin	AWBM	1_year	NS	0.65	Low	2
Darwin	AWBM	5_year	NS	0.67	Low	2
Darwin	AWBM	10_year	NS	0.70	Low	2
Adelaide	AWBM	1_year	NS	0.59	Very Low	1
Adelaide	AWBM	5_year	NS	0.59	Very Low	1
Adelaide	AWBM	10_year	NS	0.59	Very Low	1
Wagga Wagga	AWBM	1_year	NS	0.57	Very Low	1
Wagga Wagga	AWBM	5_year	NS	0.54	Very Low	1
Wagga Wagga	AWBM	10_year	NS	0.54	Very Low	1
Alice Springs	AWBM	1_year	NS	0.59	Very Low	1
Alice Springs	AWBM	5_year	NS	0.58	Very Low	1
Alice Springs	AWBM	10_year	NS	0.56	Very Low	1
Burnie	AWBM	1_year	LOGNS	0.74	Medium	3
Burnie	AWBM	5_year	LOGNS	0.67	Low	2
Burnie	AWBM	10_year	LOGNS	0.65	Low	2
Darwin	AWBM	1_year	LOGNS	0.65	Low	2
Darwin	AWBM	5_year	LOGNS	0.62	Very Low	1
Darwin	AWBM	10_year	LOGNS	0.64	Low	2
Adelaide	AWBM	1_year	LOGNS	0.61	Very Low	1
Adelaide	AWBM	5_year	LOGNS	0.58	Very Low	1
Adelaide	AWBM	10_year	LOGNS	0.56	Very Low	1
Wagga Wagga	AWBM	1_year	LOGNS	0.64	Low	2
Wagga Wagga	AWBM	5_year	LOGNS	0.62	Very Low	1
Wagga Wagga	AWBM	10_year	LOGNS	0.62	Very Low	1
Alice Springs	AWBM	1_year	LOGNS	0.62	Very Low	1
Alice Springs	AWBM	5_year	LOGNS	0.61	Very Low	1
Alice Springs	AWBM	10_year	LOGNS	0.64	Low	2

Burnie	AWBM	1_year	WLS	0.73	Medium	3
Burnie	AWBM	5_year	WLS	0.73	Medium	3
Burnie	AWBM	10_year	WLS	0.74	Medium	3
Darwin	AWBM	1_year	WLS	0.75	Medium	3
Darwin	AWBM	5_year	WLS	0.73	Medium	3
Darwin	AWBM	10_year	WLS	0.71	Medium	3
Adelaide	AWBM	1_year	WLS	0.73	Medium	3
Adelaide	AWBM	5_year	WLS	0.73	Medium	3
Adelaide	AWBM	10_year	WLS	0.74	Medium	3
Wagga Wagga	AWBM	1_year	WLS	0.73	Medium	3
Wagga Wagga	AWBM	5_year	WLS	0.73	Medium	3
Wagga Wagga	AWBM	10_year	WLS	0.75	Medium	3
Alice Springs	AWBM	1_year	WLS	0.70	Low	2
Alice Springs	AWBM	5_year	WLS	0.72	Medium	3
Alice Springs	AWBM	10_year	WLS	0.73	Medium	3
Burnie	GR4J	1_year	NS	0.78	High	4
Burnie	GR4J	5_year	NS	0.79	High	4
Burnie	GR4J	10_year	NS	0.80	High	4
Darwin	GR4J	1_year	NS	0.78	High	4
Darwin	GR4J	5_year	NS	0.78	High	4
Darwin	GR4J	10_year	NS	0.78	High	4
Adelaide	GR4J	1_year	NS	0.74	Medium	3
Adelaide	GR4J	5_year	NS	0.77	High	4
Adelaide	GR4J	10_year	NS	0.75	Medium	3
Wagga Wagga	GR4J	1_year	NS	0.76	Medium	3
Wagga Wagga	GR4J	5_year	NS	0.77	High	4
Wagga Wagga	GR4J	10_year	NS	0.78	High	4
Alice Springs	GR4J	1_year	NS	0.76	Medium	3
Alice Springs	GR4J	5_year	NS	0.72	Medium	3
Alice Springs	GR4J	10_year	NS	0.72	Medium	3
Burnie	GR4J	1_year	LOGNS	0.81	Very High	5
Burnie	GR4J	5_year	LOGNS	0.80	High	4
Burnie	GR4J	10_year	LOGNS	0.78	High	4
Darwin	GR4J	1_year	LOGNS	0.80	High	4
Darwin	GR4J	5_year	LOGNS	0.79	High	4
Darwin	GR4J	10_year	LOGNS	0.81	Very High	5
Adelaide	GR4J	1_year	LOGNS	0.76	High	4
Adelaide	GR4J	5_year	LOGNS	0.78	High	4
Adelaide	GR4J	10_year	LOGNS	0.78	High	4
Wagga Wagga	GR4J	1_year	LOGNS	0.78	High	4
Wagga Wagga	GR4J	5_year	LOGNS	0.79	High	4
Wagga Wagga	GR4J	10_year	LOGNS	0.79	High	4

Alice Springs	GR4J	1_year	LOGNS	0.75	Medium	3
Alice Springs	GR4J	5_year	LOGNS	0.76	Medium	3
Alice Springs	GR4J	10_year	LOGNS	0.77	High	4
Burnie	GR4J	1_year	WLS	0.79	High	4
Burnie	GR4J	5_year	WLS	0.78	High	4
Burnie	GR4J	10_year	WLS	0.77	High	4
Darwin	GR4J	1_year	WLS	0.78	High	4
Darwin	GR4J	5_year	WLS	0.76	High	4
Darwin	GR4J	10_year	WLS	0.80	High	4
Adelaide	GR4J	1_year	WLS	0.77	High	4
Adelaide	GR4J	5_year	WLS	0.78	High	4
Adelaide	GR4J	10_year	WLS	0.79	High	4
Wagga Wagga	GR4J	1_year	WLS	0.76	Medium	3
Wagga Wagga	GR4J	5_year	WLS	0.77	High	4
Wagga Wagga	GR4J	10_year	WLS	0.78	High	4
Alice Springs	GR4J	1_year	WLS	0.74	Medium	3
Alice Springs	GR4J	5_year	WLS	0.76	Medium	3
Alice Springs	GR4J	10_year	WLS	0.75	Medium	3
Burnie	IHACRES	1_year	NS	0.82	Very High	5
Burnie	IHACRES	5_year	NS	0.83	Very High	5
Burnie	IHACRES	10_year	NS	0.83	Very High	5
Darwin	IHACRES	1_year	NS	0.81	Very High	5
Darwin	IHACRES	5_year	NS	0.82	Very High	5
Darwin	IHACRES	10_year	NS	0.84	Very High	5
Adelaide	IHACRES	1_year	NS	0.71	Medium	3
Adelaide	IHACRES	5_year	NS	0.76	Medium	3
Adelaide	IHACRES	10_year	NS	0.76	Medium	3
Wagga Wagga	IHACRES	1_year	NS	0.74	Medium	3
Wagga Wagga	IHACRES	5_year	NS	0.75	Medium	3
Wagga Wagga	IHACRES	10_year	NS	0.74	Medium	3
Alice Springs	IHACRES	1_year	NS	0.70	Low	2
Alice Springs	IHACRES	5_year	NS	0.75	Medium	3
Alice Springs	IHACRES	10_year	NS	0.75	Medium	3
Burnie	IHACRES	1_year	LOGNS	0.83	Very High	5
Burnie	IHACRES	5_year	LOGNS	0.82	Very High	5
Burnie	IHACRES	10_year	LOGNS	0.82	Very High	5
Darwin	IHACRES	1_year	LOGNS	0.81	High	4
Darwin	IHACRES	5_year	LOGNS	0.81	Very High	5
Darwin	IHACRES	10_year	LOGNS	0.82	Very High	5
Adelaide	IHACRES	1_year	LOGNS	0.74	Medium	3
Adelaide	IHACRES	5_year	LOGNS	0.76	High	4
Adelaide	IHACRES	10_year	LOGNS	0.77	High	4

Wagga Wagga	IHACRES	1_year	LOGNS	0.77	High	4
Wagga Wagga	IHACRES	5_year	LOGNS	0.78	High	4
Wagga Wagga	IHACRES	10_year	LOGNS	0.77	High	4
Alice Springs	IHACRES	1_year	LOGNS	0.68	Low	2
Alice Springs	IHACRES	5_year	LOGNS	0.76	Medium	3
Alice Springs	IHACRES	10_year	LOGNS	0.70	Low	2
Burnie	IHACRES	1_year	WLS	0.79	High	4
Burnie	IHACRES	5_year	WLS	0.78	High	4
Burnie	IHACRES	10_year	WLS	0.78	High	4
Darwin	IHACRES	1_year	WLS	0.81	Very High	5
Darwin	IHACRES	5_year	WLS	0.79	High	4
Darwin	IHACRES	10_year	WLS	0.79	High	4
Adelaide	IHACRES	1_year	WLS	0.79	High	4
Adelaide	IHACRES	5_year	WLS	0.78	High	4
Adelaide	IHACRES	10_year	WLS	0.78	High	4
Wagga Wagga	IHACRES	1_year	WLS	0.78	High	4
Wagga Wagga	IHACRES	5_year	WLS	0.78	High	4
Wagga Wagga	IHACRES	10_year	WLS	0.78	High	4
Alice Springs	IHACRES	1_year	WLS	0.73	Medium	3
Alice Springs	IHACRES	5_year	WLS	0.77	High	4
Alice Springs	IHACRES	10_year	WLS	0.74	Medium	3
Burnie	Sacramento	1_year	NS	0.78	High	4
Burnie	Sacramento	5_year	NS	0.80	High	4
Burnie	Sacramento	10_year	NS	0.79	High	4
Darwin	Sacramento	1_year	NS	0.81	High	4
Darwin	Sacramento	5_year	NS	0.82	Very High	5
Darwin	Sacramento	10_year	NS	0.83	Very High	5
Adelaide	Sacramento	1_year	NS	0.79	High	4
Adelaide	Sacramento	5_year	NS	0.80	High	4
Adelaide	Sacramento	10_year	NS	0.80	High	4
Wagga Wagga	Sacramento	1_year	NS	0.80	High	4
Wagga Wagga	Sacramento	5_year	NS	0.78	High	4
Wagga Wagga	Sacramento	10_year	NS	0.78	High	4
Alice Springs	Sacramento	1_year	NS	0.79	High	4
Alice Springs	Sacramento	5_year	NS	0.78	High	4
Alice Springs	Sacramento	10_year	NS	0.79	High	4
Burnie	Sacramento	1_year	LOGNS	0.85	Very High	5
Burnie	Sacramento	5_year	LOGNS	0.85	Very High	5
Burnie	Sacramento	10_year	LOGNS	0.86	Very High	5
Darwin	Sacramento	1_year	LOGNS	0.84	Very High	5
Darwin	Sacramento	5_year	LOGNS	0.84	Very High	5
Darwin	Sacramento	10_year	LOGNS	0.85	Very High	5

Adelaide	Sacramento	1_year	LOGNS	0.85	Very High	5
Adelaide	Sacramento	5_year	LOGNS	0.83	Very High	5
Adelaide	Sacramento	10_year	LOGNS	0.82	Very High	5
Wagga Wagga	Sacramento	1_year	LOGNS	0.83	Very High	5
Wagga Wagga	Sacramento	5_year	LOGNS	0.84	Very High	5
Wagga Wagga	Sacramento	10_year	LOGNS	0.85	Very High	5
Alice Springs	Sacramento	1_year	LOGNS	0.83	Very High	5
Alice Springs	Sacramento	5_year	LOGNS	0.85	Very High	5
Alice Springs	Sacramento	10_year	LOGNS	0.85	Very High	5
Burnie	Sacramento	1_year	WLS	0.83	Very High	5
Burnie	Sacramento	5_year	WLS	0.80	High	4
Burnie	Sacramento	10_year	WLS	0.79	High	4
Darwin	Sacramento	1_year	WLS	0.80	High	4
Darwin	Sacramento	5_year	WLS	0.81	Very High	5
Darwin	Sacramento	10_year	WLS	0.84	Very High	5
Adelaide	Sacramento	1_year	WLS	0.80	High	4
Adelaide	Sacramento	5_year	WLS	0.83	Very High	5
Adelaide	Sacramento	10_year	WLS	0.79	High	4
Wagga Wagga	Sacramento	1_year	WLS	0.81	Very High	5
Wagga Wagga	Sacramento	5_year	WLS	0.80	High	4
Wagga Wagga	Sacramento	10_year	WLS	0.78	High	4
Alice Springs	Sacramento	1_year	WLS	0.79	High	4
Alice Springs	Sacramento	5_year	WLS	0.79	High	4
Alice Springs	Sacramento	10_year	WLS	0.80	High	4

Table B.2 Metric results of the *epsilon ratio of information content* (metric related to flatness) and corresponding degree of relative flatness and score (cluster) for all error surfaces (as cases with same data length have very close results, the average of 1-year and 5 year data length results of corresponding cases are presented instead of each single case)

Catchment Location	Model	Data Length	Error Metric	Metric Results	Degree of Relative Flatness	Score (Cluster)
Burnie	AWBM	1_year	NS	-1.18	Medium	3
Burnie	AWBM	5_year	NS	-0.85	Medium	3
Burnie	AWBM	10_year	NS	-0.90	Medium	3
Darwin	AWBM	1_year	NS	-1.04	Medium	3
Darwin	AWBM	5_year	NS	-0.97	Medium	3
Darwin	AWBM	10_year	NS	-1.04	Medium	3
Adelaide	AWBM	1_year	NS	-0.56	Low	2
Adelaide	AWBM	5_year	NS	-0.32	Low	2
Adelaide	AWBM	10_year	NS	-0.20	Low	2
Wagga Wagga	AWBM	1_year	NS	-0.99	Medium	3
Wagga Wagga	AWBM	5_year	NS	-0.28	Low	2
Wagga Wagga	AWBM	10_year	NS	-0.40	Low	2
Alice Springs	AWBM	1_year	NS	-0.26	Low	2
Alice Springs	AWBM	5_year	NS	-1.32	Medium	3
Alice Springs	AWBM	10_year	NS	-1.20	Medium	3
Burnie	AWBM	1_year	LOGNS	-1.74	High	4
Burnie	AWBM	5_year	LOGNS	-1.58	High	4
Burnie	AWBM	10_year	LOGNS	-1.62	High	4
Darwin	AWBM	1_year	LOGNS	-1.74	High	4
Darwin	AWBM	5_year	LOGNS	-1.72	High	4
Darwin	AWBM	10_year	LOGNS	-1.84	High	4
Adelaide	AWBM	1_year	LOGNS	-1.15	Medium	3
Adelaide	AWBM	5_year	LOGNS	-1.01	Medium	3
Adelaide	AWBM	10_year	LOGNS	-0.84	Medium	3
Wagga Wagga	AWBM	1_year	LOGNS	-1.42	Medium	3
Wagga Wagga	AWBM	5_year	LOGNS	-1.07	Medium	3
Wagga Wagga	AWBM	10_year	LOGNS	-1.00	Medium	3
Alice Springs	AWBM	1_year	LOGNS	-0.43	Low	2
Alice Springs	AWBM	5_year	LOGNS	-1.60	High	4
Alice Springs	AWBM	10_year	LOGNS	-1.60	High	4
Burnie	AWBM	1_year	WLS	-0.72	Low	2
Burnie	AWBM	5_year	WLS	0.02	Very Low	1
Burnie	AWBM	10_year	WLS	0.27	Very Low	1
Darwin	AWBM	1_year	WLS	-0.70	Low	2

Darwin	AWBM	5_year	WLS	-0.02	Very Low	1
Darwin	AWBM	10_year	WLS	0.37	Very Low	1
Adelaide	AWBM	1_year	WLS	-1.23	Medium	3
Adelaide	AWBM	5_year	WLS	-0.43	Low	2
Adelaide	AWBM	10_year	WLS	-0.08	Very Low	1
Wagga Wagga	AWBM	1_year	WLS	-1.00	Medium	3
Wagga Wagga	AWBM	5_year	WLS	-0.26	Low	2
Wagga Wagga	AWBM	10_year	WLS	0.07	Very Low	1
Alice Springs	AWBM	1_year	WLS	-2.01	High	4
Alice Springs	AWBM	5_year	WLS	-0.78	Low	2
Alice Springs	AWBM	10_year	WLS	-0.52	Low	2
Burnie	GR4J	1_year	NS	-1.94	High	4
Burnie	GR4J	5_year	NS	-2.02	High	4
Burnie	GR4J	10_year	NS	-2.06	Very High	5
Darwin	GR4J	1_year	NS	-2.01	High	4
Darwin	GR4J	5_year	NS	-1.99	High	4
Darwin	GR4J	10_year	NS	-1.98	High	4
Adelaide	GR4J	1_year	NS	-1.84	High	4
Adelaide	GR4J	5_year	NS	-1.96	High	4
Adelaide	GR4J	10_year	NS	-1.96	High	4
Wagga Wagga	GR4J	1_year	NS	-1.99	High	4
Wagga Wagga	GR4J	5_year	NS	-1.97	High	4
Wagga Wagga	GR4J	10_year	NS	-2.02	High	4
Alice Springs	GR4J	1_year	NS	-2.25	Very High	5
Alice Springs	GR4J	5_year	NS	-2.48	Very High	5
Alice Springs	GR4J	10_year	NS	-2.38	Very High	5
Burnie	GR4J	1_year	LOGNS	-1.95	High	4
Burnie	GR4J	5_year	LOGNS	-2.08	Very High	5
Burnie	GR4J	10_year	LOGNS	-2.06	Very High	5
Darwin	GR4J	1_year	LOGNS	-2.18	Very High	5
Darwin	GR4J	5_year	LOGNS	-2.27	Very High	5
Darwin	GR4J	10_year	LOGNS	-2.30	Very High	5
Adelaide	GR4J	1_year	LOGNS	-1.71	High	4
Adelaide	GR4J	5_year	LOGNS	-1.82	High	4
Adelaide	GR4J	10_year	LOGNS	-1.80	High	4
Wagga Wagga	GR4J	1_year	LOGNS	-1.70	High	4
Wagga Wagga	GR4J	5_year	LOGNS	-1.70	High	4
Wagga Wagga	GR4J	10_year	LOGNS	-1.76	High	4
Alice Springs	GR4J	1_year	LOGNS	-2.16	Very High	5
Alice Springs	GR4J	5_year	LOGNS	-2.13	Very High	5
Alice Springs	GR4J	10_year	LOGNS	-1.96	High	4
Burnie	GR4J	1_year	WLS	-1.17	Medium	3

Burnie	GR4J	5_year	WLS	-0.50	Low	2
Burnie	GR4J	10_year	WLS	-0.26	Low	2
Darwin	GR4J	1_year	WLS	-1.33	Medium	3
Darwin	GR4J	5_year	WLS	-0.62	Low	2
Darwin	GR4J	10_year	WLS	-0.30	Low	2
Adelaide	GR4J	1_year	WLS	-1.44	Medium	3
Adelaide	GR4J	5_year	WLS	-0.72	Low	2
Adelaide	GR4J	10_year	WLS	-0.36	Low	2
Wagga Wagga	GR4J	1_year	WLS	-1.16	Medium	3
Wagga Wagga	GR4J	5_year	WLS	-0.49	Low	2
Wagga Wagga	GR4J	10_year	WLS	-0.22	Low	2
Alice Springs	GR4J	1_year	WLS	-2.44	Very High	5
Alice Springs	GR4J	5_year	WLS	-0.96	Medium	3
Alice Springs	GR4J	10_year	WLS	-0.66	Low	2
Burnie	IHACRES	1_year	NS	-1.71	High	4
Burnie	IHACRES	5_year	NS	-1.68	High	4
Burnie	IHACRES	10_year	NS	-1.66	High	4
Darwin	IHACRES	1_year	NS	-1.81	High	4
Darwin	IHACRES	5_year	NS	-1.74	High	4
Darwin	IHACRES	10_year	NS	-1.78	High	4
Adelaide	IHACRES	1_year	NS	-0.90	Medium	3
Adelaide	IHACRES	5_year	NS	-1.28	Medium	3
Adelaide	IHACRES	10_year	NS	-1.24	Medium	3
Wagga Wagga	IHACRES	1_year	NS	-1.32	Medium	3
Wagga Wagga	IHACRES	5_year	NS	-1.26	Medium	3
Wagga Wagga	IHACRES	10_year	NS	-1.22	Medium	3
Alice Springs	IHACRES	1_year	NS	-0.58	Low	2
Alice Springs	IHACRES	5_year	NS	-1.80	High	4
Alice Springs	IHACRES	10_year	NS	-2.02	High	4
Burnie	IHACRES	1_year	LOGNS	-1.63	High	4
Burnie	IHACRES	5_year	LOGNS	-1.36	Medium	3
Burnie	IHACRES	10_year	LOGNS	-1.32	Medium	3
Darwin	IHACRES	1_year	LOGNS	-1.55	High	4
Darwin	IHACRES	5_year	LOGNS	-1.41	Medium	3
Darwin	IHACRES	10_year	LOGNS	-1.42	Medium	3
Adelaide	IHACRES	1_year	LOGNS	-0.89	Medium	3
Adelaide	IHACRES	5_year	LOGNS	-1.06	Medium	3
Adelaide	IHACRES	10_year	LOGNS	-1.00	Medium	3
Wagga Wagga	IHACRES	1_year	LOGNS	-1.02	Medium	3
Wagga Wagga	IHACRES	5_year	LOGNS	-0.92	Medium	3
Wagga Wagga	IHACRES	10_year	LOGNS	-0.96	Medium	3
Alice Springs	IHACRES	1_year	LOGNS	-0.47	Low	2

Alice Springs	IHACRES	5_year	LOGNS	-1.39	Medium	3
Alice Springs	IHACRES	10_year	LOGNS	-1.58	High	4
Burnie	IHACRES	1_year	WLS	-1.02	Medium	3
Burnie	IHACRES	5_year	WLS	-0.18	Low	2
Burnie	IHACRES	10_year	WLS	0.11	Very Low	1
Darwin	IHACRES	1_year	WLS	-1.14	Medium	3
Darwin	IHACRES	5_year	WLS	-0.23	Low	2
Darwin	IHACRES	10_year	WLS	0.09	Very Low	1
Adelaide	IHACRES	1_year	WLS	-1.06	Medium	3
Adelaide	IHACRES	5_year	WLS	-0.27	Low	2
Adelaide	IHACRES	10_year	WLS	0.11	Very Low	1
Wagga Wagga	IHACRES	1_year	WLS	-0.94	Medium	3
Wagga Wagga	IHACRES	5_year	WLS	-0.17	Low	2
Wagga Wagga	IHACRES	10_year	WLS	0.15	Very Low	1
Alice Springs	IHACRES	1_year	WLS	-1.61	High	4
Alice Springs	IHACRES	5_year	WLS	-0.59	Low	2
Alice Springs	IHACRES	10_year	WLS	-0.16	Low	2
Burnie	Sacramento	1_year	NS	-0.77	Low	2
Burnie	Sacramento	5_year	NS	-1.62	High	4
Burnie	Sacramento	10_year	NS	-2.00	High	4
Darwin	Sacramento	1_year	NS	-1.58	High	4
Darwin	Sacramento	5_year	NS	-2.11	Very High	5
Darwin	Sacramento	10_year	NS	-2.32	Very High	5
Adelaide	Sacramento	1_year	NS	0.47	Very Low	1
Adelaide	Sacramento	5_year	NS	-0.96	Medium	3
Adelaide	Sacramento	10_year	NS	-1.16	Medium	3
Wagga Wagga	Sacramento	1_year	NS	-0.04	Very Low	1
Wagga Wagga	Sacramento	5_year	NS	-0.65	Low	2
Wagga Wagga	Sacramento	10_year	NS	-1.04	Medium	3
Alice Springs	Sacramento	1_year	NS	1.96	Very Low	1
Alice Springs	Sacramento	5_year	NS	-0.77	Low	2
Alice Springs	Sacramento	10_year	NS	-1.14	Medium	3
Burnie	Sacramento	1_year	LOGNS	-1.81	High	4
Burnie	Sacramento	5_year	LOGNS	-2.43	Very High	5
Burnie	Sacramento	10_year	LOGNS	-2.54	Very High	5
Darwin	Sacramento	1_year	LOGNS	-2.24	Very High	5
Darwin	Sacramento	5_year	LOGNS	-2.38	Very High	5
Darwin	Sacramento	10_year	LOGNS	-2.38	Very High	5
Adelaide	Sacramento	1_year	LOGNS	-1.01	Medium	3
Adelaide	Sacramento	5_year	LOGNS	-1.78	High	4
Adelaide	Sacramento	10_year	LOGNS	-1.88	High	4
Wagga Wagga	Sacramento	1_year	LOGNS	-1.24	Medium	3

Wagga Wagga	Sacramento	5_year	LOGNS	-1.84	High	4
Wagga Wagga	Sacramento	10_year	LOGNS	-2.00	High	4
Alice Springs	Sacramento	1_year	LOGNS	0.35	Very Low	1
Alice Springs	Sacramento	5_year	LOGNS	-1.56	High	4
Alice Springs	Sacramento	10_year	LOGNS	-1.90	High	4
Burnie	Sacramento	1_year	WLS	-1.38	Medium	3
Burnie	Sacramento	5_year	WLS	-0.78	Low	2
Burnie	Sacramento	10_year	WLS	-0.44	Low	2
Darwin	Sacramento	1_year	WLS	-1.38	Medium	3
Darwin	Sacramento	5_year	WLS	-0.85	Medium	3
Darwin	Sacramento	10_year	WLS	-0.54	Low	2
Adelaide	Sacramento	1_year	WLS	-1.42	Medium	3
Adelaide	Sacramento	5_year	WLS	-0.98	Medium	3
Adelaide	Sacramento	10_year	WLS	-0.64	Low	2
Wagga Wagga	Sacramento	1_year	WLS	-1.34	Medium	3
Wagga Wagga	Sacramento	5_year	WLS	-0.72	Low	2
Wagga Wagga	Sacramento	10_year	WLS	-0.44	Low	2
Alice Springs	Sacramento	1_year	WLS	-1.56	High	4
Alice Springs	Sacramento	5_year	WLS	-1.06	Medium	3
Alice Springs	Sacramento	10_year	WLS	-0.88	Medium	3

Table B.3 Metric results of the *median basin centroidal distance* (metric related to optima dispersion) and corresponding degree of relative flatness and score (cluster) for all error surfaces (as cases with same data length have very close results, the average of 1-year and 5-year data length results of corresponding cases are presented instead of each single case)

Catchment Location	Model	Data Length	Error Metric	Metric Result	Degree of Relative Optima Dispersion	Score (Cluster)
Burnie	AWBM	1_year	NS	10.45	Low	2
Burnie	AWBM	5_year	NS	7.73	Low	2
Burnie	AWBM	10_year	NS	2.88	Very Low	1
Darwin	AWBM	1_year	NS	8.49	Low	2
Darwin	AWBM	5_year	NS	6.22	Low	2
Darwin	AWBM	10_year	NS	5.37	Low	2
Adelaide	AWBM	1_year	NS	8.61	Low	2
Adelaide	AWBM	5_year	NS	7.64	Low	2

Adelaide	AWBM	10_year	NS	7.31	Low	2
Wagga Wagga	AWBM	1_year	NS	9.30	Low	2
Wagga Wagga	AWBM	5_year	NS	8.68	Low	2
Wagga Wagga	AWBM	10_year	NS	0.00	Very Low	1
Alice Springs	AWBM	1_year	NS	10.81	Low	2
Alice Springs	AWBM	5_year	NS	14.30	Medium	3
Alice Springs	AWBM	10_year	NS	13.86	Medium	3
Burnie	AWBM	1_year	LOGNS	9.64	Low	2
Burnie	AWBM	5_year	LOGNS	7.53	Low	2
Burnie	AWBM	10_year	LOGNS	8.55	Low	2
Darwin	AWBM	1_year	LOGNS	7.94	Low	2
Darwin	AWBM	5_year	LOGNS	8.73	Low	2
Darwin	AWBM	10_year	LOGNS	4.10	Very Low	1
Adelaide	AWBM	1_year	LOGNS	8.78	Low	2
Adelaide	AWBM	5_year	LOGNS	5.41	Low	2
Adelaide	AWBM	10_year	LOGNS	7.03	Low	2
Wagga Wagga	AWBM	1_year	LOGNS	9.12	Low	2
Wagga Wagga	AWBM	5_year	LOGNS	7.46	Low	2
Wagga Wagga	AWBM	10_year	LOGNS	6.81	Low	2
Alice Springs	AWBM	1_year	LOGNS	12.35	Low	2
Alice Springs	AWBM	5_year	LOGNS	12.68	Low	2
Alice Springs	AWBM	10_year	LOGNS	12.54	Low	2
Burnie	AWBM	1_year	WLS	1.37	Very Low	1
Burnie	AWBM	5_year	WLS	0.00	Very Low	1
Burnie	AWBM	10_year	WLS	0.00	Very Low	1
Darwin	AWBM	1_year	WLS	1.12	Very Low	1
Darwin	AWBM	5_year	WLS	0.00	Very Low	1
Darwin	AWBM	10_year	WLS	0.00	Very Low	1
Adelaide	AWBM	1_year	WLS	12.46	Low	2
Adelaide	AWBM	5_year	WLS	11.17	Low	2
Adelaide	AWBM	10_year	WLS	0.00	Very Low	1
Wagga Wagga	AWBM	1_year	WLS	5.44	Low	2
Wagga Wagga	AWBM	5_year	WLS	0.00	Very Low	1
Wagga Wagga	AWBM	10_year	WLS	0.00	Very Low	1
Alice Springs	AWBM	1_year	WLS	12.03	Low	2
Alice Springs	AWBM	5_year	WLS	10.19	Low	2
Alice Springs	AWBM	10_year	WLS	10.66	Low	2
Burnie	GR4J	1_year	NS	6.32	Low	2
Burnie	GR4J	5_year	NS	7.61	Low	2
Burnie	GR4J	10_year	NS	7.59	Low	2
Darwin	GR4J	1_year	NS	8.39	Low	2
Darwin	GR4J	5_year	NS	16.16	Medium	3

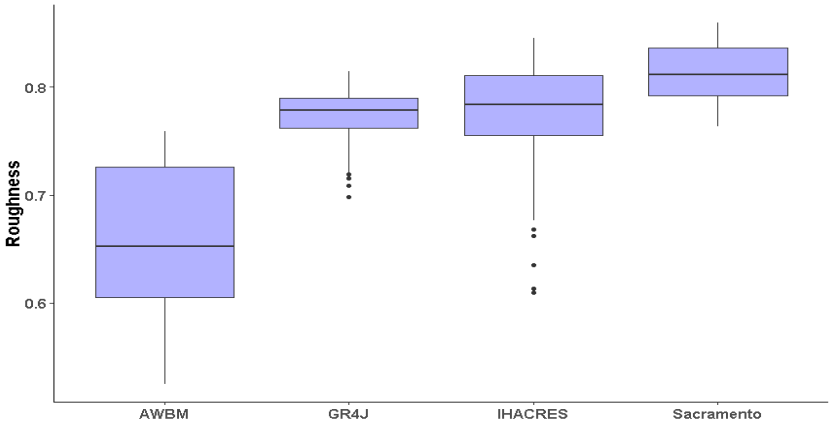
Darwin	GR4J	10_year	NS	15.69	Medium	3
Adelaide	GR4J	1_year	NS	17.54	Medium	3
Adelaide	GR4J	5_year	NS	3.38	Very Low	1
Adelaide	GR4J	10_year	NS	0.00	Very Low	1
Wagga Wagga	GR4J	1_year	NS	3.57	Very Low	1
Wagga Wagga	GR4J	5_year	NS	5.64	Low	2
Wagga Wagga	GR4J	10_year	NS	0.00	Very Low	1
Alice Springs	GR4J	1_year	NS	14.22	Medium	3
Alice Springs	GR4J	5_year	NS	15.51	Medium	3
Alice Springs	GR4J	10_year	NS	0.00	Very Low	1
Burnie	GR4J	1_year	LOGNS	0.72	Very Low	1
Burnie	GR4J	5_year	LOGNS	0.69	Very Low	1
Burnie	GR4J	10_year	LOGNS	2.25	Very Low	1
Darwin	GR4J	1_year	LOGNS	6.84	Low	2
Darwin	GR4J	5_year	LOGNS	0.00	Very Low	1
Darwin	GR4J	10_year	LOGNS	0.00	Very Low	1
Adelaide	GR4J	1_year	LOGNS	17.75	Medium	3
Adelaide	GR4J	5_year	LOGNS	0.00	Very Low	1
Adelaide	GR4J	10_year	LOGNS	0.00	Very Low	1
Wagga Wagga	GR4J	1_year	LOGNS	18.98	Medium	3
Wagga Wagga	GR4J	5_year	LOGNS	15.32	Medium	3
Wagga Wagga	GR4J	10_year	LOGNS	12.50	Low	2
Alice Springs	GR4J	1_year	LOGNS	18.39	Medium	3
Alice Springs	GR4J	5_year	LOGNS	15.17	Medium	3
Alice Springs	GR4J	10_year	LOGNS	20.90	High	4
Burnie	GR4J	1_year	WLS	17.34	Medium	3
Burnie	GR4J	5_year	WLS	8.75	Low	2
Burnie	GR4J	10_year	WLS	3.89	Very Low	1
Darwin	GR4J	1_year	WLS	10.72	Low	2
Darwin	GR4J	5_year	WLS	12.07	Low	2
Darwin	GR4J	10_year	WLS	11.48	Low	2
Adelaide	GR4J	1_year	WLS	16.12	Medium	3
Adelaide	GR4J	5_year	WLS	9.87	Low	2
Adelaide	GR4J	10_year	WLS	8.76	Low	2
Wagga Wagga	GR4J	1_year	WLS	15.83	Medium	3
Wagga Wagga	GR4J	5_year	WLS	10.66	Low	2
Wagga Wagga	GR4J	10_year	WLS	0.00	Very Low	1
Alice Springs	GR4J	1_year	WLS	17.55	Medium	3
Alice Springs	GR4J	5_year	WLS	14.53	Medium	3
Alice Springs	GR4J	10_year	WLS	12.09	Low	2
Burnie	IHACRES	1_year	NS	14.29	Medium	3
Burnie	IHACRES	5_year	NS	10.16	Low	2

Burnie	IHACRES	10_year	NS	7.00	Low	2
Darwin	IHACRES	1_year	NS	15.53	Medium	3
Darwin	IHACRES	5_year	NS	16.46	Medium	3
Darwin	IHACRES	10_year	NS	24.15	High	4
Adelaide	IHACRES	1_year	NS	16.87	Medium	3
Adelaide	IHACRES	5_year	NS	18.54	Medium	3
Adelaide	IHACRES	10_year	NS	15.28	Medium	3
Wagga Wagga	IHACRES	1_year	NS	18.73	Medium	3
Wagga Wagga	IHACRES	5_year	NS	13.02	Low	2
Wagga Wagga	IHACRES	10_year	NS	0.00	Very Low	1
Alice Springs	IHACRES	1_year	NS	20.07	High	4
Alice Springs	IHACRES	5_year	NS	19.53	Medium	3
Alice Springs	IHACRES	10_year	NS	14.07	Medium	3
Burnie	IHACRES	1_year	LOGNS	18.55	Medium	3
Burnie	IHACRES	5_year	LOGNS	18.16	Medium	3
Burnie	IHACRES	10_year	LOGNS	19.85	High	4
Darwin	IHACRES	1_year	LOGNS	23.15	High	4
Darwin	IHACRES	5_year	LOGNS	21.60	High	4
Darwin	IHACRES	10_year	LOGNS	21.72	High	4
Adelaide	IHACRES	1_year	LOGNS	16.90	Medium	3
Adelaide	IHACRES	5_year	LOGNS	21.51	High	4
Adelaide	IHACRES	10_year	LOGNS	22.15	High	4
Wagga Wagga	IHACRES	1_year	LOGNS	21.17	High	4
Wagga Wagga	IHACRES	5_year	LOGNS	10.88	Low	2
Wagga Wagga	IHACRES	10_year	LOGNS	17.14	Medium	3
Alice Springs	IHACRES	1_year	LOGNS	17.19	Medium	3
Alice Springs	IHACRES	5_year	LOGNS	13.76	Medium	3
Alice Springs	IHACRES	10_year	LOGNS	23.57	High	4
Burnie	IHACRES	1_year	WLS	18.99	Medium	3
Burnie	IHACRES	5_year	WLS	21.57	High	4
Burnie	IHACRES	10_year	WLS	5.68	Low	2
Darwin	IHACRES	1_year	WLS	18.62	Medium	3
Darwin	IHACRES	5_year	WLS	17.11	Medium	3
Darwin	IHACRES	10_year	WLS	14.54	Medium	3
Adelaide	IHACRES	1_year	WLS	18.41	Medium	3
Adelaide	IHACRES	5_year	WLS	18.50	Medium	3
Adelaide	IHACRES	10_year	WLS	18.96	Medium	3
Wagga Wagga	IHACRES	1_year	WLS	19.49	Medium	3
Wagga Wagga	IHACRES	5_year	WLS	19.18	Medium	3
Wagga Wagga	IHACRES	10_year	WLS	20.51	High	4
Alice Springs	IHACRES	1_year	WLS	18.49	Medium	3
Alice Springs	IHACRES	5_year	WLS	17.47	Medium	3

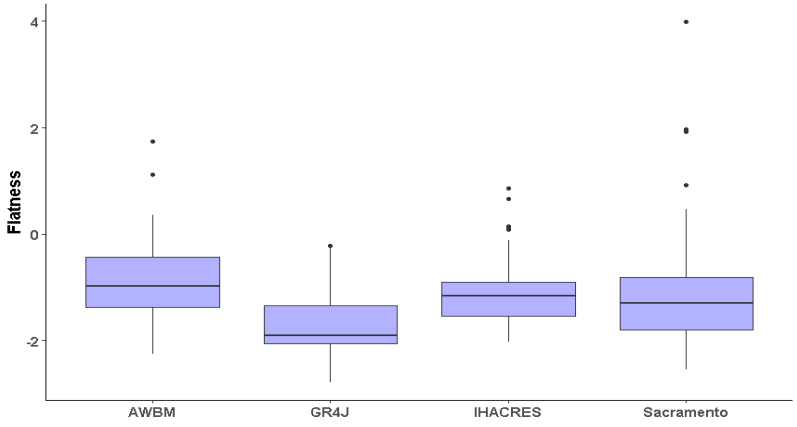
Alice Springs	IHACRES	10_year	WLS	18.41	Medium	3
Burnie	Sacramento	1_year	NS	23.98	High	4
Burnie	Sacramento	5_year	NS	23.97	High	4
Burnie	Sacramento	10_year	NS	23.79	High	4
Darwin	Sacramento	1_year	NS	26.53	Very High	5
Darwin	Sacramento	5_year	NS	NA	NA	NA
Darwin	Sacramento	10_year	NS	NA	NA	NA
Adelaide	Sacramento	1_year	NS	22.46	High	4
Adelaide	Sacramento	5_year	NS	22.23	High	4
Adelaide	Sacramento	10_year	NS	21.26	High	4
Wagga Wagga	Sacramento	1_year	NS	23.72	High	4
Wagga Wagga	Sacramento	5_year	NS	NA	NA	NA
Wagga Wagga	Sacramento	10_year	NS	22.51	High	4
Alice Springs	Sacramento	1_year	NS	19.59	Medium	3
Alice Springs	Sacramento	5_year	NS	24.90	Very High	5
Alice Springs	Sacramento	10_year	NS	26.10	Very High	5
Burnie	Sacramento	1_year	LOGNS	23.05	High	4
Burnie	Sacramento	5_year	LOGNS	NA	NA	NA
Burnie	Sacramento	10_year	LOGNS	22.85	High	4
Darwin	Sacramento	1_year	LOGNS	21.23	High	4
Darwin	Sacramento	5_year	LOGNS	19.52	Medium	3
Darwin	Sacramento	10_year	LOGNS	19.91	High	4
Adelaide	Sacramento	1_year	LOGNS	23.67	High	4
Adelaide	Sacramento	5_year	LOGNS	23.02	High	4
Adelaide	Sacramento	10_year	LOGNS	23.81	High	4
Wagga Wagga	Sacramento	1_year	LOGNS	NA	NA	NA
Wagga Wagga	Sacramento	5_year	LOGNS	NA	NA	NA
Wagga Wagga	Sacramento	10_year	LOGNS	21.55	High	4
Alice Springs	Sacramento	1_year	LOGNS	21.28	High	4
Alice Springs	Sacramento	5_year	LOGNS	NA	NA	NA
Alice Springs	Sacramento	10_year	LOGNS	NA	NA	NA
Burnie	Sacramento	1_year	WLS	23.31	High	4
Burnie	Sacramento	5_year	WLS	23.64	High	4
Burnie	Sacramento	10_year	WLS	24.60	Very High	5
Darwin	Sacramento	1_year	WLS	30.15	Very High	5
Darwin	Sacramento	5_year	WLS	28.50	Very High	5
Darwin	Sacramento	10_year	WLS	28.93	Very High	5
Adelaide	Sacramento	1_year	WLS	24.93	Very High	5
Adelaide	Sacramento	5_year	WLS	24.15	High	4
Adelaide	Sacramento	10_year	WLS	22.66	High	4
Wagga Wagga	Sacramento	1_year	WLS	21.65	High	4
Wagga Wagga	Sacramento	5_year	WLS	19.43	Medium	3

Wagga Wagga	Sacramento	10_year	WLS	21.92	High	4
Alice Springs	Sacramento	1_year	WLS	25.54	Very High	5
Alice Springs	Sacramento	5_year	WLS	22.00	High	4
Alice Springs	Sacramento	10_year	WLS	19.31	Medium	3

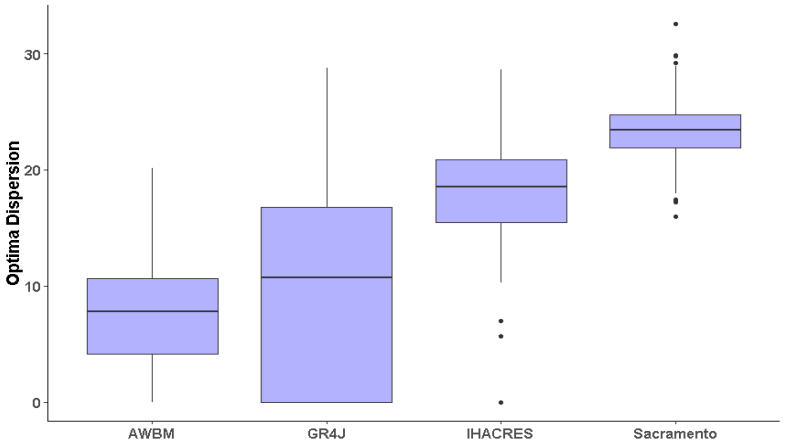
Appendix C – Influence of model structure / complexity on error surface features (Raw Results)



(a)



(b)



(c)

Figure C.1 Influence of model structure / complexity on error surface features (raw results): Roughness (a); Flatness (b); Optima Dispersion (c). The complexity of the CRR models increases from left to right: AWBM has 2 parameters, GR4J has 4 parameters, IHACRES has 6 parameters and Sacramento has 13 parameters.

Supplementary Materials of Chapter 2 (Paper 1)

S1: Example R-code for calculating ELA metric results

The code implementation should follow the following process:

1. install the packages listed in the script;
2. confirm the problem for fitness landscape analysis, this includes:
   a. initial set up of problem dimension, sample size for the test;
   b. problems/functions for the test (here use BBOB function as example);
3. load the function for local search metric if the local search test is needed (i.e. running codes under the text annotation "Function of modified version of local search metric from FLACCO“);
4. running the implementation codes for metric calculation (i.e. running codes under the text annotation “Metric calculation”).

```
# Package Install
library(lhs)
library(flacco)
library(smoof)
library(mlr)
library(class)
library(mda)

# Initial setup
Nsamp = 500 # Sample size
dimNum = 2  # problem dimension
funcNum = 20 # BBOB function ID, can be set up from 1 to 24
funciid = 1 # BBOB function instance, any random positive value
## initial sampling
ini.sample = as.matrix(randomLHS(Nsamp,dimNum))
ini.sample = (ini.sample-0.5)*10
## BBOB function setup
fn = makeBBOBFunction(dimension = dimNum, fid = funcNum, iid = funciid)
## Search space range setup
low = replicate(dimNum, -5)
upp = replicate(dimNum, 5)

# Function of modified version of local search metric from FLACCO
################################################################################
calculateLocalSearchFeatures1 = function(feat.object, low, upp,
nor.clust.cutfun) {
  measureTime(expression({
    f = initializeCounter(feat.object$fun)
    X = extractFeatures(feat.object)
    y = extractObjective(feat.object)
    d = feat.object$dim
    N=nrow(X)
    opt.algo = "L-BFGS-B"
    opt.algo.control = list()
    if (!feat.object$minimize) {
      y = -1 * y
      opt.algo.control$fnscale = -1
    } else {
      opt.algo.control$fnscale = 1
    }
    id.seed = sample(1:1e6, 1)
    clust.method = "single"
    clust.cutfun = nor.clust.cutfun*sqrt(sum((upp-low)^2))

    calcOptim = function(par, ...) {
      res = optim(as.numeric(par), fn, method = opt.algo, control =
opt.algo.control, lower = low, upper = upp, ...)
      return(list(par = res$par, counts = resetCounter(fn)))
    }
    set.seed(id.seed)
    ids = sample(nrow(X), N)

    fn = initializeCounter(f)
    result = lapply(ids, function(i) calcOptim(drop(X[i,])))

    pars = t(vapply(result, function(i) i$par, double(d)))
    fun.evals = vapply(result, function(i) i$counts, integer(1))
```

```
    cl = hclust(dist(pars), clust.method)
    clust = cutree(cl, h = clust.cutfun)

    clust.size = tapply(clust, clust, length)
    clust.size = clust.size / sum(clust.size) ## Normalize!

    centers = t(vapply(seq_along(clust.size),
                       function(i) colMeans(pars[clust == i, , drop = FALSE]),
double(d)))
    pdist.center=dist(centers)
    centers.funvals = apply(centers, 1, f)
    centers.best = which(centers.funvals == min(centers.funvals))
    centers.worst = which(centers.funvals == max(centers.funvals))
    list(ela_local.n_loc_opt.abs = max(clust),
         ela_local.n_loc_opt.rel = max(clust) / N,
         ela_local.best2mean_contr.orig = min(centers.funvals) /
mean(centers.funvals),
         ela_local.best2mean_contr.ratio = (mean(centers.funvals) -
min(centers.funvals)) /
           (max(centers.funvals) - min(centers.funvals)),
         ## Metrics related to distance between local regions of attractions,
which is not included in FLACCO

##############################################################################
         ela_local.center.dist_min=ifelse(length(pdist.center) == 0, 0,
min(pdist.center)),
         ela_local.center.dist_lq=ifelse(length(pdist.center) == 0, 0,
as.numeric(quantile(pdist.center, 0.25))),
         ela_local.center.dist_mean=ifelse(length(pdist.center) == 0, 0,
mean(pdist.center)),
         ela_local.center.dist_median=ifelse(length(pdist.center) == 0,
0,median(pdist.center)),
         ela_local.center.dist_uq=ifelse(length(pdist.center) == 0, 0,
as.numeric(quantile(pdist.center, 0.75))),
         ela_local.center.dist_max=ifelse(length(pdist.center) == 0, 0,
max(pdist.center)),
         ela_local.center.dist_sd=ifelse(length(pdist.center) == 0, 0,
sd(pdist.center)),

##############################################################################
         ela_local.basin_sizes.avg_best = mean(clust.size[centers.best]),
         ela_local.basin_sizes.avg_non_best = ifelse(length(clust.size[-
centers.best]) == 0L,
                                                     0, mean(clust.size[-
centers.best])),
         ela_local.basin_sizes.avg_worst = mean(clust.size[centers.worst]),
         ela_local.fun_evals.min = min(fun.evals),
         ela_local.fun_evals.lq = as.numeric(quantile(fun.evals, 0.25)),
         ela_local.fun_evals.mean = mean(fun.evals),
         ela_local.fun_evals.median = median(fun.evals),
         ela_local.fun_evals.uq = as.numeric(quantile(fun.evals, 0.75)),
         ela_local.fun_evals.max = max(fun.evals),
         ela_local.fun_evals.sd = sd(fun.evals),
         ela_local.costs_fun_evals = showEvals(f)
    )
  }), "ela_local")
}
```

```
initializeCounter = function(fn) {
  force(fn)
  count = 0L
  structure(function(x, ...) {
    count <<- count + if (is.matrix(x))
      ncol(x)
    else 1L
    fn(x, ...)
  })
}

extractFeatures = function(feat.object) {
  as.matrix(subset(feat.object$env$init, select = feat.object$feature.names))
}

extractObjective = function(feat.object) {
  feat.object$env$init[, feat.object$objective.name]
}

showEvals = function(fn) {
  environment(fn)$count
}

resetCounter = function (fn) {
  counts = environment(fn)$count
  environment(fn)$count = 0L
  counts
}
################################################################################
# Metric calculation
feat.object = createFeatureObject(X = ini.sample, fun = fn)
## ELA Convexity Metric
convexityresult =  calculateFeatureSet(feat.object = feat.object, set =
"ela_conv",
                                       control = list(ela_level.parallel.cpus
= 1,
                                                      allow_cellmapping =
FALSE,ela_conv.nsample = Nsamp))
## ELA Y_distribution Metric
y_distributionresult = calculateFeatureSet(feat.object = feat.object, set =
"ela_distr",
                                           control =
list(ela_level.parallel.cpus = 1, allow_cellmapping = FALSE))
## ELA Levelset Metric
levelsetresult = calculateFeatureSet(feat.object = feat.object, set =
"ela_level", control = list(ela_level.parallel.cpus = 1, allow_cellmapping =
FALSE))

## ELA Meta Model Metric
metamodelresult = calculateFeatureSet(feat.object = feat.object, set =
"ela_meta", control = list(ela_level.parallel.cpus = 1, allow_cellmapping =
FALSE))

## Information Content Metric Nearest Neighbouring Sampling Sequence
nn_icofsresult = calculateFeatureSet(feat.object = feat.object, set = "ic",
control = list(ela_level.parallel.cpus = 1, allow_cellmapping = FALSE,
ic.sorting = "nn"))
```

```
## Information Content Metric random Sampling Sequence
rand_icofsresult = calculateFeatureSet(feat.object = feat.object, set = "ic",
control = list(ela_level.parallel.cpus = 1, allow_cellmapping = FALSE,
ic.sorting = "random"))

## ELA Curvature Metric
curvaturresult = calculateFeatureSet(feat.object = feat.object, set =
"ela_curv",
                                     control = list(ela_level.parallel.cpus =
1,
                                                    allow_cellmapping =
FALSE,ela_curv.sample_size = Nsamp))

## ELA Local Search Metric
localsearchresult = calculateLocalSearchFeatures1(feat.object = feat.object,
low = low, upp = upp, nor.clust.cutfun = 0.05)
```

S2: Example R-Code for dependency analysis

The code implementation should follow the following process:

1. load the data for building the regression model. The data should be data frame with 3 columns, which are named as Dim (i.e. dimensionality), SampleSize (i.e. problem sample size), Result (i.e. corresponding ELA metric result). An example data is used in this script;
2. build the regression model and calculate the slope of dimension, slope of sample size and r2 of the regression model for the output by running codes under the text annotation "regression model setup".

```
library("rstudioapi")
# Example data loading (RData file should be in the same directory with the
script)
setwd(dirname(getActiveDocumentContext()$path))
load("exampledata.RData")

# regression model setup
regmodel = lm(Result ~ log(Dim) + log(SampleSize), data = RegresionMatrix)
regsum = summary(regmodel)
Dimslope =abs(regsum$coefficients[2, 1])  # slope of dimension
SSSlope = abs(regsum$coefficients[3, 1])  # slope of sample size
Rsquared = regsum$r.squared               # r-squared value of regression
```