Improved Evolutionary Algorithm Optimisation of Water Distribution Systems Using Domain Knowledge

by
Weiwei Bi

Thesis submitted to School of Civil, Environmental & Mining Engineering
of the University of Adelaide
in fulfillment of the requirements for
the degree of

Doctor of Philosophy

Improved Evolutionary Algorithm Optimisation of Water Distribution Systems Using Domain Knowledge

By:
Weiwei Bi

Supervised by:
Graeme C. Dandy, *B.E.* (*Hons*), *MEngSc*, *Ph.D.*
Professor, School of Civil, Environmental & Mining Engineering,
The University of Adelaide
Holger R. Maier, *B.E.* (*Hons*), *Ph.D.*
Professor, School of Civil, Environmental & Mining Engineering,
The University of Adelaide

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

School of Civil, Environmental & Mining Engineering
Faculty of Engineering, Computer and Mathematical Sciences
The University of Adelaide
North Terrace, Adelaide, SA 5005, Australia
Telephone: +61 8303 6139
Facsimile: +61 8303 4359
Web: http://www.adelaide.edu.au/directory/weiwei.bi
Email: weiwei.bi@adelaide.edu.au

Contents

Abstract

Water distribution systems (WDSs) are becoming increasingly complex and larger in scale due to the rapid growth of population and fast urbanization. Hence, they require high levels of investment for their construction and maintenance. This motivates the need to optimally design these systems, with the aim being to minimize the investment budget while maintaining high service quality. Over the past 25 years, a number of evolutionary algorithms (EAs) have been developed to achieve optimal design solutions for WDSs, representing a focal point of much research in this area.

One issue that hinders EAs' wide application in industry is their significant demand on computational resources when handling real-world WDSs. In recognition of this, there has been a move from aiming to find the globally optimal solutions to identifying the best possible solutions within constrained computational resources. While many studies have been undertaken to attain this goal, there have been limited efforts that use engineering knowledge to reduce the computational effort. The research undertaken in this thesis is such an attempt, as it aims to efficiently identify near-optimal solutions with the aid of WDS design knowledge.

This thesis presents a domain-knowledge based optimization framework that enables the near-optimal solutions (fronts) of WDS problems to be identified within constrained computing time. The knowledge considered includes (i) the relationship between pipe size and distance to the water source(s); (ii) the impact of flow velocities on optimal solutions; and (iii) the relationship between flow velocities and network resilience.

This thesis consists of an Introduction, three chapters that are based around a series of three journal papers and a set of Conclusions and Recommendations for Further Work.

The first paper introduces a new initialization method to assist genetic algorithms (GAs) to identify near-optimal solutions in a computationally efficient manner. This is attained by incorporating domain knowledge into the generation of the initial population of GAs. The results show that the proposed method performs better than the other three initialization methods considered, both in terms of computational efficiency and the ability to find near-optimal solutions.

The second paper investigates the relative impact of different algorithm initializations and searching mechanisms on the speed with which near-optimal solutions can be identified for large WDS design problems. Results indicate that EA parameterizations, that emphasize exploitation relative to exploration, enable near-optimal solutions to be identified earlier in the search, which is due to the "big bowl" shape of the fitness function for all of the WDS problems considered. Using initial solutions that are informed using domain knowledge can further increase the speed with which near-optimal solutions can be identified.

The third publication extends the single-objective method in the first paper to a two-objective problem. The objectives considered are the minimization of cost and maximization of network resilience. The performance of the two-objective initialization approach is compared with that of randomly initializing the population of multi-objective EAs applied to range of WDS design problems. The results indicate that there are considerable benefits in using the proposed initialization method in terms of being able to identify near-optimal fronts more rapidly.

Although all of the results obtained in this research have shown that the proposed method is effective for improving the efficiency of EAs in finding near-optimal solutions, only gravity fed water distribution systems with a single loading case were considered as case studies. One important area for future research is the extension of the proposed method to more complex WDSs which may include tanks, pumps and valves.

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 any university or other tertiary institution. To the best of my knowledge and belief in 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 give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying, subject to the provision of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis (as list below) 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 catalogue, and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signature Date

Acknowledgements

I would take this opportunity to thank my academic supervisors Professor Graeme C. Dandy and Professor Holger R. Maier for their continuous support and guidance. I appreciate the research freedom I have enjoyed under their supervision, and thank them deeply for always being available to provide constructive suggestions and discuss any issues with great patience.

I would like to acknowledge the support of my family: to my husband Feifei Zheng for his love, patience and continuous support; to my parents Mr. Keming Bi and Mrs. Yuqin Guo for their help and encouragements; and to my baby boy Zhuohao Zheng for the happiness he brings to me over the journey of my PhD candidature.

I acknowledge the research scholarship provided by the University of Adelaide, as well as the assistance from the colleagues and staff in the School of Civil, Environmental and Mining Engineering, The University of Adelaide.

List of Figures

List of Tables

List of Acronyms

ABS	Average of the best solutions
BBS	Best of the best solutions
DE	Differential evolution
DMO	Deviation of the mean cost relative to the best known solution
EA	Evolutionary algorithms
GA	Genetic algorithm
KLSM	Kang and Lansey's sampling method
LHS	Latin hypercube sampling
MOEA	Multi-objective evolutionary algorithms
MOPHSM	Multi-objective prescreened heuristic sampling method
NYTP	New York tunnel problem
PCX	Patent-centric crossover
PHSM	Prescreened heuristic sampling method
RS	Random sampling
SBX	Simulated binary crossover
SPD	Standardized average population diversity
SPX	Simplex crossover
UM	Uniform mutation
UNDX	Unimodal normal distribution crossover
WDS	Water distribution systems

Chapter 1. Introduction

Water distribution systems (WDS) are used to deliver water from water sources or treatment plants to end-users, representing one of the basic forms of civil infrastructure within cities. A typical WDS consists of pipes, reservoirs, pumps, valves and other hydraulic elements, which are all cost intensive in construction and management. Furthermore, the maintenance and rehabilitation costs for WDSs are often very large, which can be on the order of millions of dollars (e.g. Simpson et al. 1994; Nicklow et al. 2010). This motivates a number of studies to optimise these systems, aiming to potentially save significant costs while meeting the required demand as well as satisfying supply pressures (Marchi et al. 2014a).

For a given water distribution network layout, the design problem typically involves the selection of the pipe sizes as well as the sizes of other system components (e.g. valves and pumps), such that the system can be constructed or operated with the minimum total life cycle cost while satisfying all of the design constraints (e.g. Dandy et al. 1996; Zheng et al. 2011a,b; Kang and Lansey 2012). However, the complex system structure (e.g. loops), highly nonlinear relationship between pipe head loss and flows and the discrete nature of the availability of pipe sizes that can be used create a highly complex search space for a WDS design problem (Zecchin et al. 2012). This results in the presence of many local optimal solutions, bringing significant challenges for finding a high quality solution.

Traditionally, a trial-and-error approach or deterministic optimisation techniques (e.g. linear programming and nonlinear programming) have been used to find efficient solutions for simple WDSs (Fujiwara and Khang 1990; Bragalli et al. 2012). However, solutions found using these approaches are often unsatisfactory, especially for large, real-world problems (Simpson et al. 1994; Maier et al. 2014). More recently, evolutionary algorithms (EAs) have been employed to optimise the design of WDSs, and have been often demonstrated to be able to find significantly improved solutions compared to traditional methods (Maier et al. 2015). This is because EAs differ from deterministic optimisation techniques in that they navigate through the search space by

means of stochastic evolution rather than using gradient information, thereby leading to a higher likelihood that globally optimal solutions will be reached (di Pierro et al. 2009; Fu et al. 2012). Another important advantage of EAs over traditional optimisation techniques is for multi-objective problems, where they can identify a set of Pareto optimal solutions in a single run, with trade-offs between multiple competing objectives being identified (Ostfeld et al. 2014).

From the literature, it can be seen that the research area of EAs applied to WDS design optimisation has undergone significant development over the past few decades (e.g. Nicklow et al. 2014; Maier et al. 2014). This is supported by the following: (i) a broad range of EA types has been successfully applied to WDS design problems; and (ii) EAs have provided an improved understanding of the WDS optimisation problem for both single objective and multi-objective problems. However, the application of EAs is not without difficulties, with one of the main issues being their significant demand on computational resources, which is especially the case when dealing with real-world problems (Fu et al. 2012; Kang and Lansey 2012). In fact, their computational intensity has been one of the main reasons for practitioners' reluctance to use EAs in practice.

In recognition of this, there has been a move from attempting to find the global optimal solutions, which may require very large computational effort, to identifying near-optimal solutions within limited computational budgets in recent years (Gibbs et al. 2008, 2015; Tolson et al. 2007, 2009; Maier et al. 2014, 2015). This is because, for many water resource problems, finding near-optimal solutions in a reasonable amount of time (rather than attempting to find the global optimum) is often sufficient from a practical perspective. Finding best possible solutions within a limited time framework is challenging for many EA-based optimisation techniques, which are typically developed to find globally optimal solutions without considering the constraint of the available computational resources (Maier et al. 2014).

To address this issue, a number of approaches have been developed for efficiently arriving at near-optimal solutions in recent years. Examples include the hybridisation of EAs with deterministic techniques (Tolson et al. 2009; Zheng et al. 2011a), EA

parametrisations based on improved understanding of their run-time searching behaviour (Zecchin et al. 2012; Zheng et al. 2015a), and partitioning of large problems into a set of smaller and manageable sub-problems (Zheng et al. 2013 a,b). However, so far there have been limited efforts to use domain knowledge to assist EAs to efficiently identify near-optimal solutions (Kang and Lansey 2012; Bi et al. 2015a). Domain knowledge is often derived from a physical understanding of the system, as well as engineering experience. The research undertaken in this thesis is an attempt to develop new techniques that efficiently identify near-optimal solutions with the aid of WDS design knowledge.

1.1 Objectives of research

This research aims to improve EA optimisation of WDSs with the aid of domain knowledge in both single and multiple objective spaces. The specific objectives are given below:

Objective 1: *To incorporate domain knowledge into the initialization of EAs, enabling EAs to commence their search in the areas surrounding promising regions and hence improve their performance in efficiently identifying near-optimal solutions.* The EA used to meet this objective is the genetic algorithm (GA), which is the most frequently used EA for water resources problems.

Objective 2: *To gain an improved understanding of the relative impact of the EA starting position and parameterisations on the speed with which near-optimal solutions can be identified for large optimization problems.* In order to meet this objective, fitness function and run-time behavioural statistics are used to gain such an increased understanding.

Objective 3: *To extend the domain-knowledge based single-objective EA initialization method in Objective 1 to a multi-objective problem.* The objectives considered are the minimization of cost and the maximization of network resilience (one way to represent WDS reliability). Two different types of multiobjective EAs (MOEAs) are considered in order to meet this objective, which are NSGA-II, representing one of

the standard MOEAs for industry application, and Borg, representing a recent state-of-the-art MOEA for water resources.

1.2 Outline of the thesis

This thesis consists of five chapters, with the main body (**Chapters 2-4**) being a collection of published, accepted or submitted papers from internationally recognised Journals (Bi et al., 2015a; Bi et al., 2015b; Bi et al., 2015c). Table 1.1 summarises the main information of each paper and how they link to the stated objectives of this thesis, with details given below.

Table 1.1 Publication information

Publication	Aims	Linking to the objectives	The number of case studies (the range of the number of decision variables)
Publication 1 (Chapter 2)	To develop a new initialization method for EAs with the incorporation of domain knowledge in *single objective (cost) space*	**Objective 1**	Seven (34 to 1274)
Publication 2 (Chapter 3)	To improve understanding of the relative impact of EA starting position and parameterisation on the speed with which near-optimal solutions are found in *single objective (cost) space*	**Objective 2**	Four (164 to 1274)
Publication 3 (Chapter 4)	To extend the initialization method in single-objective (cost) space to a *two-objective (cost and network resilience) space*	**Objective 3**	Five (34 to 1274)

In **Chapter 2 (publication 1)**, a new initialization method is developed to assist EAs to find near-optimal solutions in an efficient manner, in which domain knowledge with regard to the relationship between pipe size and distance to the source(s) of WDS(s), as well as the impact of flow velocities on optimal solutions are considered. Three steps are involved in the proposed approach, including (i) the selection of pipe sizes based on knowledge that pipe diameters generally get smaller the further they are from the source; (ii) dynamic adjustment of the velocity threshold to account for the fact that appropriate velocity thresholds are likely to be network dependent; and (iii) control of initial population diversity by sampling from distributions centred on the solutions determined using the heuristic procedures of (i) and (ii). The performance of the proposed method is

compared with that of another heuristic sampling method and two non-heuristic sampling methods applied to seven WDS design case studies with the number of decision variables ranging from 34 to 1274. *This chapter links to **Objective 1** of this thesis (Table 1.1).*

Chapter 3 (publication 2) aims to investigate the relative impact of different algorithm initializations and searching mechanisms on the speed with which near-optimal solutions can be identified. While the impact of the initialization (initial population, starting position in solution space) of EAs on the *speed* (in terms of computational effort) with which *near*-optimal solutions can be found has been investigated previously, the impact of using different EAs (e.g. genetic algorithm, differential evolution) and EA parameterizations (e.g., mutation rate) on the relative performance of these initialization methods has not been studied previously. This study addresses this issue. In order to obtain a better understanding of the relative performance of different algorithm initialization methods and searching behaviours, a secondary objective of this research is to examine the properties of the fitness functions of the case studies and the run-time behavioural statistics of the different algorithms and their parameterizations, and how they relate to observed algorithm performance. *This chapter links to **Objective 2** of this thesis (Table 1.1).*

Chapter 4 (publication 3) develops and tests a method for identifying high quality initial populations for multi-objective EAs (MOEAs) applied to WDS design problems aimed at minimizing cost and maximizing network resilience. The proposed multiobjective initialization method not only considers the relationship between pipe size and distance to the source(s) of water, as for the method in **Chapter 2**, but also accounts for the relationship between flow velocities and network resilience. The benefit of using the proposed approach compared with randomly generating initial populations in relation to finding near-optimal fronts more efficiently is tested on five WDS optimization case studies of varying complexity with two MOEAs (NSGA-II and Borg). *This chapter links to **Objective 3** of this thesis (Table 1.1).*

While the manuscripts have been reformatted (the sections have also been renumbered) in accordance with University guidelines, the material within this thesis is otherwise presented herein as published or submitted for publication. A copy of the manuscript that has already been published is provided in Appendix.

Conclusions of the research within this thesis are provided in Chapter 5, which summarises research contributions, research limitations and recommendations for further research.

Chapter 2. Journal Paper 1-Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge

Statement of Authorship

Title of Paper	Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Publication Style
Publication Details	Bi, W., Dandy, G. C., and Maier, H. R. (2015). "Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge." *Environmental Modelling & Software*, 69(0), 370-381.

Principal Author

Name of Principal Author (Candidate)	Weiwei Bi		
Contribution to the Paper	Develop the approach, perform the simulation study and prepare the manuscript		
Overall percentage (%)	50%		
Signature		Date	

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	Graeme C. Dandy		
Contribution to the Paper	Research supervision and review of manuscript.		
Signature	25%	Date	

Name of Co-Author	Holger R. Maier		
Contribution to the Paper	Research supervision and review of manuscript.		
Signature	25%	Date	

Abstract

Over the last two decades, evolutionary algorithms (EAs) have become a popular approach for solving water resources optimization problems. However, the issue of low computational efficiency limits their application to large, realistic problems. This paper uses the optimal design of water distribution systems (WDSs) as an example to illustrate how the efficiency of genetic algorithms (GAs) can be improved by using heuristic domain knowledge in the sampling of the initial population. A new heuristic procedure called the Prescreened Heuristic Sampling Method (PHSM) is proposed and tested on seven WDS cases studies of varying size. The EPANet input files for these case studies are provided as supplementary material. The performance of the PHSM is compared with that of another heuristic sampling method and two non-heuristic sampling methods. The results show that PHSM clearly performs best overall, both in terms of computational efficiency and the ability to find near-optimal solutions. In addition, the relative advantage of using the PHSM increases with network size.

Keywords: Optimization; Genetic algorithms, Water distribution systems; domain knowledge; heuristics; computational efficiency.

2.1 Introduction

Evolutionary algorithms (EAs) have been used successfully and extensively for solving water resources optimization problems in a number of areas, such as engineering design, the development of management strategies and model calibration (Nicklow et al. 2010; Zecchin et al. 2012). However, a potential shortcoming of EAs is that they are computationally inefficient, especially when applied to problems of realistic size. Consequently, there is a need to improve the computational efficiency of EAs to make them easier to use for the optimization of realistic water resources problems (Maier et al. 2014a).

One application area where this is the case of is the design of water distribution systems (WDSs) (Marchi et al. 2014a; Stokes et al. 2014). Over the past two decades, a variety of EAs have been applied to this problem, as detailed in Zheng et al. (2013a). Among these,

genetic algorithms (GAs) have been used most extensively (Simpson et al. 1994; Dandy et al. 1996, Gupta et al. 1999; Vairavamoorthy et al. 2005; Krapivka and Ostfeld 2009; Kang and Lansey 2012; Zheng et al. 2013b). However, GAs have been primarily applied to relatively simple benchmark problems, such as the 14-pipe problem (Simpson et al. 1994), the New York Tunnels problem with 21 tunnels (Dandy et al. 1996), and the Hanoi problem with 34 pipes (Zheng et al. 2011a). In recent years, there has been a move towards increasing the complexity and realism of the case studies to which GAs are applied, including the Balerma network with 454 pipes (Reca and Martínez 2006), the Rural network with 476 pipes (Marchi et al. 2014b), the BWN-II network with 433 pipes (Zheng et al. 2013b), and the network used by Kang and Lansey (2012), which has 1274 pipes and will be referred to as the "KL" network for the remainder of this paper.

Increased network size and complexity result in significant challenges in terms of achieving good quality near-optimal solutions given the computational budgets that are typically available in practice (di Pierro et al. 2009; Fu et al. 2012). This is because (i) the time for hydraulic simulation increases appreciably for large WDSs; and (ii) the complexity and size of the search space associated with a large WDS are increased significantly. As a result, computational efficiency has been identified as a key concern for the widespread uptake of GAs for the optimization of large, real-world WDSs (di Pierro et al. 2009).

In order to address this issue, two main approaches have been adopted in the literature. As part of the first approach, it is argued that for large, real problems, the focus should be on finding the best possible solution within a realistic computational budget, rather than on attempting to find the global optimal solution (e.g. Tolson and Shoemaker 2007; Gibbs et al. 2008; Tolson et al. 2009; Gibbs et al. 2010, 2011). This is because for such large problems, the global optimal solution is unlikely to be found within a reasonable computational timeframe.

As part of the second approach, efforts have been made to increase the computational efficiency of the optimization process. This has been done in a number of ways,

including the use of increased computational power, such as parallel and distributed computing (Wu and Zhu 2009; Roshani and Filion 2012; Wu and Behandish 2012), the use of surrogate- and meta-modeling to speed up the simulation process (e.g. Broad et al. 2005; di Pierro et al. 2009; Broad et al. 2010; Razavi et al. 2012), and the seeding of the initial population of EAs with good solutions obtained using a variety of analytical techniques (e.g. Keedwell and Khu 2006; Zheng et al. 2011a; Fu et al. 2012; Zheng et al. 2014(b,c,d)). It should be noted that similar concepts have recently also been used in conjunction with other optimization techniques (e.g. Zhang et al. 2013; Housh et al. 2013) and non-optimization based WDS design approaches (e.g. Sitzenfrei et al. 2013).

Although some of the methods mentioned above utilize engineering knowledge in their development (e.g. Keedwell and Khu 2006; Zheng et al. 2011a), there have been limited attempts to incorporate engineering knowledge and experience directly. Only Kang and Lansey (2012) have combined engineering experience with GAs in order to increase the computational efficiency of the optimization process. This was achieved by seeding half of the initial GA population with solutions that result in flow velocities below a threshold selected from within a pre-defined velocity range. However, the approach has only been applied to a single case study thus far and its relative performance has not yet been assessed in a rigorous and comprehensive manner. In addition, the approach has a number of potential shortcomings. Firstly, selection of an appropriate range for the velocity threshold is subjective, which might make the method difficult to apply and could result in inconsistent results from repeated, independent implementation of the method. Secondly, pipe sizes that result in appropriate velocities are determined using a structured trial-and-error process. However, in practice, pipe sizes generally reduce with distance from the source (Walski 2001). Consequently, there exists an opportunity to incorporate this domain knowledge into the initial pipe sizing process. Finally, there is limited control over population diversity, as this is achieved by seeding the initial population with 50% of randomly generated solutions and 50% of the solutions obtained based on engineering experience.

In order to address these shortcomings, the objectives of this paper are (i) to introduce a new heuristic sampling method for determining the initial population of GAs for the least-cost design of WDSs that is based on engineering experience / domain knowledge and that overcomes the potential shortcomings of the method of Kang and Lansey (2012); and (ii) to provide a rigorous assessment of the performance of this method compared with that of Kang and Lansey's sampling method (KLSM) and two sampling methods that do not consider any domain knowledge (i.e. random sampling (RS) and Latin hypercube sampling (LHS)) on seven WDS design case studies of varying size and complexity.

The remainder of this paper is organized as follows. The proposed heuristic, domain knowledge based sampling method for determining the initial population of GAs for the least-cost design of WDSs is introduced in next section, followed by the methodology for assessing the performance of this method against that of the KLSM and the two non-heuristic sampling methods. Next, the results are presented and discussed, followed by a summary and conclusions.

2.2 Proposed prescreened heuristic sampling method for WDS design

The proposed heuristic sampling method for initializing the population of GAs for the least-cost design of WDSs based on domain knowledge is called the Prescreened Heuristic Sampling Method (PHSM). It uses a three-step procedure that (i) selects pipe sizes based on knowledge that pipe diameters generally get smaller the further they are from the source; (ii) dynamically adjusts the velocity threshold to account for the fact that appropriate velocity thresholds are likely to be network dependent; and (iii) enables the diversity of the initial population to be controlled by sampling from distributions centred on the solutions determined using the heuristic procedures in (i) and (ii). The PHSM has some similarities to the KLSM in that it aims to find initial pipe sizes that restrict flow velocities to lie within certain ranges. However, it

overcomes the potential limitations of the KLSM outlined in the Introduction. Details of the three steps of the PHSM are given below.

Step 1: Assign pipe diameters based on distances between demand nodes and supply sources

As mentioned above, the first step of the PHSM is motivated by the knowledge that, in real WDSs, the diameters of upstream pipes are generally larger than those further downstream (Walski 2001). However, for WDSs, each demand node usually has a number of different paths that connect it to the supply source (reservoir). This indicates that the spatial distance between each demand node and the reservoir may vary according to the paths selected to deliver the required demands. In the proposed method, the shortest delivery path to each demand node is selected and used to represent the spatial distance between that node and the source node. The rationale behind this is that it has been demonstrated that the majority of the demand at a node is supplied by the path with the shortest distance for an optimal design of WDSs (Zheng et al. 2011a). The detailed process of step 1 of the PHSM is as follows:

i: Find the shortest distance to a reservoir in the water network, l_i for each node i (i=1,2…..,n, where n is the total number of demand nodes in the network) using the Dijkstra algorithm (Zheng et al. 2011a). When dealing with a water network with multiple reservoirs, an augmented source node is created to connect all the reservoirs to enable the determination of l_i following Deuerlein (2008) and Zheng et al. (2011a).

ii: Obtain the largest value of the shortest distance L by L=max(l_i).

iii: Divide the network into P specific areas with the shortest distance to the source node interval of L/P, where P is the number of available pipe diameters for the design.

iv: Assign pipes in each area a different diameter, with the largest diameter assigned to the pipes in the area nearest to the source and the smallest diameter to the pipes

in the area furthest from the source (reservoir). All pipes in a single area are assigned the same diameter.

For example, for the WDS introduced by Zheng et al. (2011a), which has 164 pipes (Figure 2.1), the largest shortest distance of all nodes (L) is obtained after steps i and ii. If there are five diameter options for this network (i.e. $P=5$), the network will be divided into five areas in step iii. In order to do this (i) all nodes that have a shortest-distance that is not greater than L/P (i.e. $0<l_i \leq L/5$) form Area 1; (ii) all nodes that have a shortest-distance that is larger than L/P but not greater than $2L/5$ (i.e. $L/5<l_i \leq 2L/5$) form Area 2; (v) all nodes that have a shortest-distance larger than 4L/P (i.e. $4L/5<l_i \leq L$) form Area 5. The resulting division of the network is given in Figure 2.1. Finally, (i) all pipes in Area 1 are assigned the largest diameter; (ii) all pipes in Area 2 are assigned the second largest diameter; and so on until all pipes in Area 5 are assigned the smallest pipe diameter. As such, the diameters of the upstream pipes are generally larger than those of the downstream pipes.

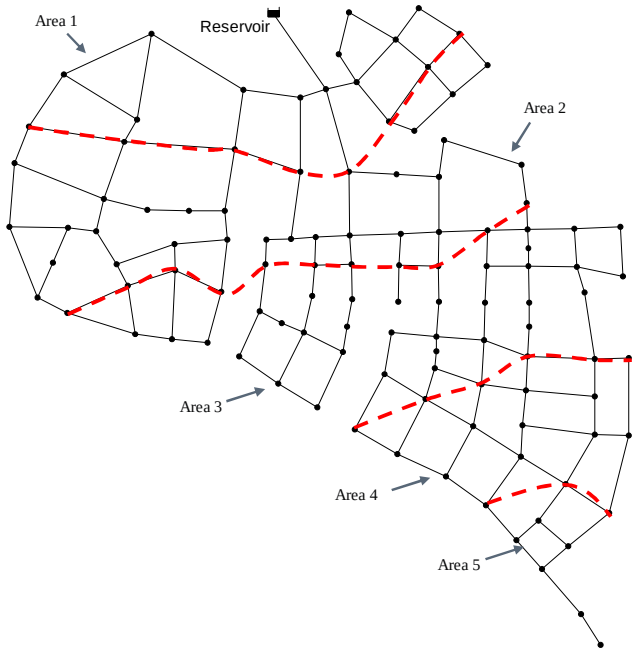

Figure 2.1 WDS used to illustrate the result of network division of the PHSM (The red dot lines represent the distance boundary used to assign diameters)

Step 2: Adjust pipe diameters based on velocities

In this step, the diameters obtained in step 1 are refined to achieve flow velocities in all pipes that are close to a particular threshold. This is based on the domain knowledge that the velocity in each pipe of an optimal solution for a WDS is in a limited range. In addition, in order to ensure that the chosen pipe diameters approach optimal values, the velocity threshold is selected to result in solutions that are on the boundary between feasibility and infeasibility. This is because the optimal solution is often located on the boundary of the feasible and infeasible areas of the search space. The stages in the process for achieving this are shown in Figure 2.2.

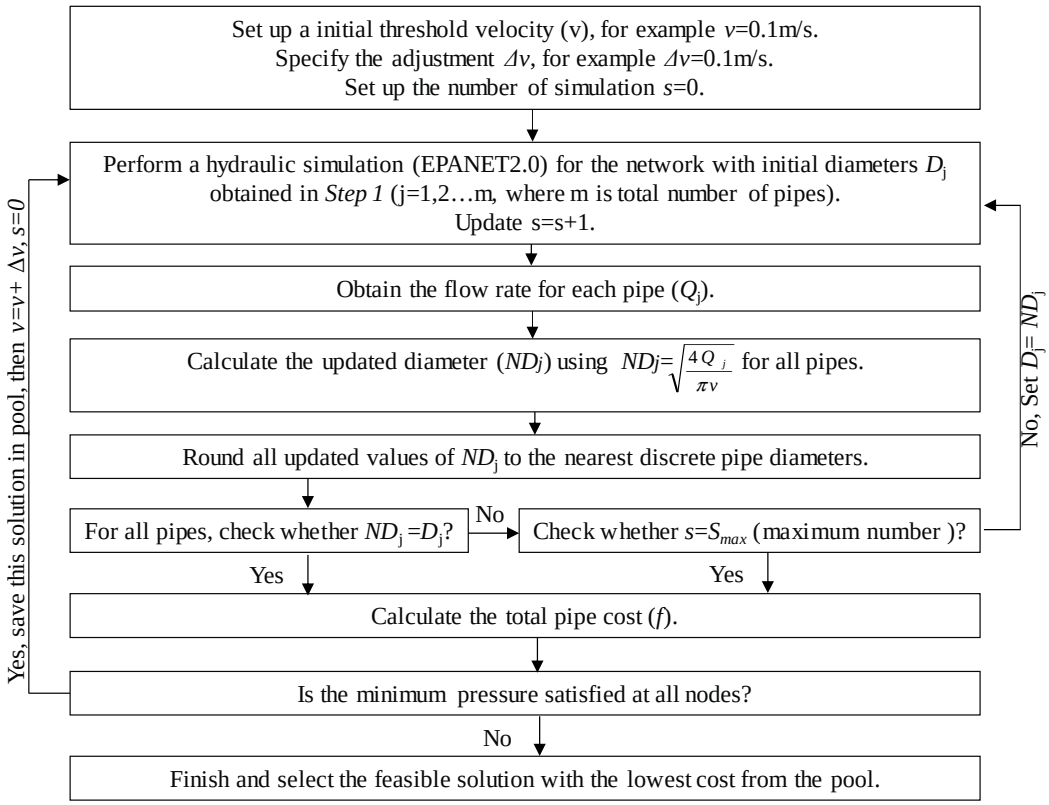

Figure 2.2 Flowchart of the algorithm for adjusting pipe diameters based on flow velocity

As can be seen from Figure 2.2, an inner loop and an outer loop are involved in the algorithm. The inner loop is used to determine the network configuration based on pipe velocities. To do this, a threshold value v for velocity needs to be assigned at

the beginning (e.g. v =0.1 m/s), which represents the expected velocity for each pipe in the network. The network with initial diameters determined in Step 1 is then simulated using a hydraulic solver to obtain the flow rate for each pipe. Based on this flow rate, the new diameter ND_j for each pipe can be calculated using:

$$ND_j = \sqrt{\frac{4Q_j}{\pi v}} \qquad (2.1)$$

where j=1,….,m is the j^{th} pipe in the water network and m is the total number of pipes. As continuous diameter values are generated using Equation (2.1), these values need to be rounded up or down to the nearest discrete diameter based on the available options.

The inner loop continues until there is no further change in diameter in accordance with Equation (2.1) or the number of simulations (s) reaches the specified maximum number of allowable simulations (S_{max}), at which point the cost (f) and the minimum pressure head of this design are determined. If this solution is feasible (i.e., the pressure head constraints are satisfied), the network configuration and its associated network cost are saved to an archive. As part of the outer loop, the inner loop is repeated for successive increases in the velocity threshold (i.e. $v = v + \Delta v$) until no feasible solution can be found. If the solution found at the completion of the inner loop is infeasible, the outer loop is not performed and the process of adjusting diameters is terminated. Finally, the feasible solution with the lowest cost for the different velocity thresholds considered is selected from the archive and denoted as an approximate optimal solution for the WDS being optimized. This solution is then used as the starting point for Step 3, as outlined below.

Step 3: Generate distribution functions based on the approximate optimal solution determined in Step 2.

In order to ensure sufficient diversity in the initial solution, the initial diameter for each pipe is generated from a distribution, such that the pipe diameter obtained in

Step 2 has the highest probability of being selected. The logic behind this is that the approximate diameter for a pipe determined in Step 2 is most likely to be the optimal diameter relative to other diameter options. Hence, a relatively higher density function value is assigned to this diameter (i.e. it is more likely to be selected during sampling).

The density function $f(D_k)$ and the distribution function $F(D_k)$ for selecting each initial diameter are given by the following equations:

$$f(D_k) = \frac{1}{1 + a \mid x \mid} \qquad k = 1,....,P \qquad (2.2)$$

$$F(D_k) = \frac{f(D_k)}{\sum_{k=1}^{P} f(D_k)} \qquad k = 1,....,P \qquad (2.3)$$

where a is a constant factor to control the density of each diameter D_k, details of which are discussed in Section 4; x is the distance between D_k and D_c (the diameter for a pipe in the approximate optimal solution determined in Step 2) in terms of integer coding; and P is the total number of available pipe diameters.

In order to illustrate how the approach outlined above is used to generate the pipe diameters in the initial solution, the following example is used. Table 2.1 presents the assumed total pipe diameter options and their corresponding integer coding values. If D_c=200mm in Step 2 for a particular pipe, its integer code is 1, as shown in Table 2.1. The absolute distance $|x|$ between each D_k and D_c is then calculated and presented in the third column of Table 2.1. The density function and distribution function values for generating each available diameter for this pipe during sampling are calculated based on Equations (2.2) and (2.3), respectively (assuming a=1). The results are given in the fourth and fifth columns of Table 2.1. As can be seen, a diameter of 200mm has the largest probability of being selected during sampling, as this diameter is selected based on the heuristic rules used in Steps 1 and 2. In contrast,

a diameter of 600mm has the smallest probability of being selected, since it has the largest distance to the optimal diameter of 200mm.

Table 2.1 An example to illustrate the application of Step 3 of the PHSM

Pipe diameters D_k (mm)	Integer coding number	Absolute distance to D_c ($\|x\|$)	Density function values $f(D_k)$	Distribution function values $F(D_k)$
100	0	1	0.5	0.19
200	1	0	1	0.39
300	2	1	0.5	0.19
500	3	2	0.33	0.13
600	4	3	0.25	0.10

It should be noted that the assumption made in Step 1 that the upstream diameters are typically larger than those further downstream might not hold for all networks due to the influence of network topology and zoning, However, as the initial diameters obtained in Step 1 are adjusted based on flow velocities in Step 2, the influence of network topology and zoning is accounted for in the overall approach.

2.3 Methodology

As stated in the Introduction, one of the objectives of this paper is to provide a rigorous assessment of the relative performance of the PHSM compared with that of the KLSM and two sampling methods that do not consider domain knowledge. The flowchart of the process for achieving this is shown in Figure 2.3. As can be seen, four different sampling methods, including two heuristic methods (i.e. the PHSM and the KLSM) and two non-heuristic methods (i.e. RS and LHS), are used to obtain initial GA populations. The two non-heuristic sampling methods are considered as they provide a benchmark against which the performance of the two heuristic sampling methods can be assessed. RS is used as this is the conventional method for initializing GA populations and LHS is used as it provides a more structured approach for sampling the solution space. It should be noted that, although there are some other analytical techniques for seeding the initial population of EAs (e.g. Keedwell and Khu 2006; Zheng et al. 2011a; Fu et al. 2012), they do not incorporate engineering knowledge and experience directly and hence are not considered in this paper.

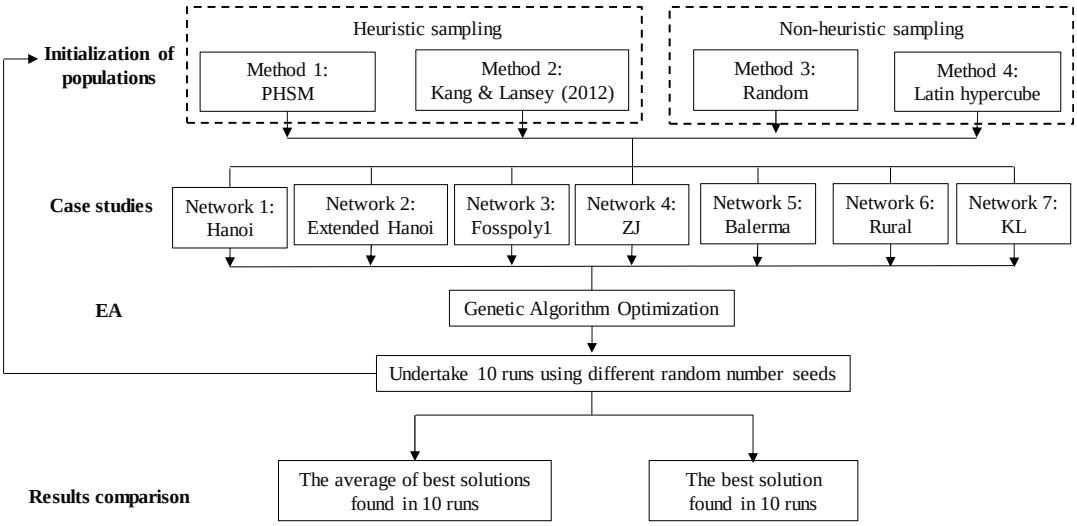

Figure 2.3 Flowchart of the assessment process

Each of the sampling approaches is applied to seven WDSs of varying size and complexity, including the Hanoi, Extended Hanoi, Fosspoly 1, ZJ, Balerma and Rural networks, as well as a modified version of the KL network (KLmod). The networks are optimized for total life cycle costs while satisfying pressure head constraints at each demand node. The hydraulic simulations required to check pressure constraints are performed using EPANET 2.0, as demand-driven modelling is most commonly used in optimization studies, although pressure-driven modelling is likely to be a better alternative under some circumstances (Laucelli et al. 2012). Each of the GA optimization runs is repeated 10 times with different sets of initial solutions and GA operators generated using different random number seeds for each network and sampling method. The results are compared in terms of the best and average solutions found during these ten runs. Details of each of the components of the process are provided in subsequent sections.

2.3.1 Sampling methods

Details of the KLSM (Method 2, Figure 2.3) and the two non-heuristic sampling methods (Methods 3 and 4, Figure 2.3) are given below. Details of the PHSM (Method 1, Figure 2.3) are given in the previous section.

2.3.1.1 The KLSM (Kang and Lansey 2012) (Method 2)

As mentioned previously, in this approach, initial solutions are generated by adjusting pipe diameters to ensure that the velocities in all pipes are less than a pre-set velocity threshold selected from a practical range of velocities for average and peak flows in water supply networks. The heuristic procedure for achieving this is as follows:

(1) All pipes to be optimized are set to the minimum allowable diameter.

(2) A hydraulic simulation is carried out to obtain the flow velocity in each pipe.

(3) The resulting velocity in each pipe is compared with a pre-set velocity threshold selected from within the range of 0.45-1.5 m/s (e.g. 1 m/s). If the velocity is larger than the threshold, this pipe diameter is increased to the next larger commercial size.

Steps (2) and (3) are performed repeatedly until all velocities in all pipes are below the threshold and the resulting pipes sizes are used to form one solution of the initial population. A number of different initial solutions is generated by varying the value of the velocity threshold within the pre-defined velocity range of 0.45 - 1.5m/s. In order to maintain solution diversity, half of the initial solutions are generated using this heuristic method, while the other half are generated randomly. In this study, the velocity thresholds of the KLSM are obtained using the following equation:

$$VT_r = 0.45m/s + r\frac{(1.5m/s - 0.45m/s)}{\frac{1}{2}N} \tag{2.4}$$

where VT_r (m/s) is the r^{th} (r=1,2…,$\frac{1}{2}N$) velocity threshold used for generating the heuristic solutions; N is the total population size.

2.3.1.2 Random Sampling (Method 3)

In random sampling (RS), each diameter option has the same probability of being selected for each pipe within the WDS. When generating a solution, each decision variable (i.e. pipe) is assigned a diameter value that is randomly selected from all available diameter options.

2.3.1.3 Latin Hypercube Sampling (Method 4)

Latin Hypercube Sampling (LHS) is a type of stratified sampling method that ensures that all portions of the sample space of each variable are sampled (McKay et al. 1979). In this study, Simlab2.2 (JRC 2008) is used to generate initial solutions using LHS for each case study. A detailed description of the process of LHS can be found in the manual of Simlab2.2 (JRC 2008).

2.3.2 Case studies

Details of each case study are given in Table 2.2. For each case study, the decision variables are the pipe diameters and the objective is to find the minimum cost solution while satisfying the pressure head constraints. Consequently, the optimization problem to be solved can be represented as follows:

Minimize $$F = \sum_{j=1}^{m} C_j(D_j) \quad (2.5)$$

Subject to:

$$H_i^{\min} \le H_i \le H_i^{\max} \quad i = 1,2.....,n \quad (2.6)$$

$$G(H_i, \mathbf{D})=0 \quad (2.77)$$

$$D_j \in \{A\} \quad (2.8)$$

where F is the network cost that is to be minimized; $C_j(D_j)$ is the cost function for pipe j =1,2…,m with assigned diameter D_j; m and n are the total number of pipes and demand nodes in the network, respectively; $G(H_i, \mathbf{D})$=nodal mass balance and loop (path) energy balance equations for the whole network with pipe combinations of $\mathbf{D}=[D_1, D_2,\ldots D_m]^T$, which is solved using EPANET2.0; H_i=head at node i=1,2….,n; H_i^{min} and H_i^{max} are the minimum and maximum allowable head limits at the nodes; and A = a set of commercially available pipe diameters.

Table 2.2 Details of the seven case studies

Case study	Reference	No. of decision variables[1]	No. of diameter options[2]	Size of total search space	Pressure head constraint	Current best solution	Current best solution found by GAs
Hanoi	Fujiwara and Khang (1990)	34	6	2.86×10^{26}	≥ 30 m	$6.081 million by Reca and Martínez (2006) using GENOME	$6.081 million by Reca and Martínez (2006)
Extended Hanoi	current study	34	10	1×10^{34}	≥ 30 m	-[3]	-[3]
Fosspoly1	Bragalli et al. (2012)	58	22	7.26×10^{77}	≥ 40 m	$0.0291 million by Bragalli et al. (2012) using MINLP	-[3]
ZJ	Zheng et al. (2011a)	164	14	9.23×10^{187}	≥ 22 m	$7.082 million by Zheng et al. (2011a) using NLP-DE	-[3]
Balerma	Reca and Martínez (2006)	454	10	1×10^{454}	≥ 20 m	€1.923 million by Zheng et al. (2011a) using NLP-DE	€2.302 million by Bolognesi et al. (2010)
Rural network	Marchi et al. (2014)	476	15	6.58×10^{559}	≥ 0 m	$ 31.22 million by Marchi et al. (2014) using DE	$ 36.25 million by Marchi et al. (2014)
KLmod network	Adapted from Kang and Lansey (2012)	1274	10	1×10^{1274}	≥ 45 m	-[3]	-[3]

[1]The decision variables are the pipe diameters. [2]The pipe diameter options for the Extended Hanoi and KL network are given in this paper and those for the other case studies are given in the references provided. [3]The current best solution is unknown or the network has not been optimized previously using an Evolutionary Algorithm.

As shown in Table 2.2, the seven case studies vary in size and complexity. Details of each network, including the network layout, the available pipe diameters and the cost of each diameter for the Extended Hanoi and the KLmod network are given in this paper and those of the other case studies are given in the corresponding references in the second column of Table 2.2. The EPANet input files for these seven networks are provided as supplementary material. The current best known solution for each case

study (if available) is presented in the second last column in Table 2.2. The best known solutions (least-cost solutions) for the Hanoi, Balerma and Rural case studies found by GAs are given in the last column, while no GA solutions can be found in the literature for the other case studies.

The Extended Hanoi case study is developed based on the original Hanoi problem (Fujiwara and Khang 1990), and has not been used in previous studies. The only difference between the original and Extended Hanoi case studies is the number of available diameters for each pipe, while the other information is the same. As it is acknowledged that infeasible solutions dominate the search space for the Hanoi case study, a larger number of diameter options is included for this case study in order to test the performance of the various sampling methods when dealing with a search space with a larger feasible proportion. For the Extended Hanoi problem, ten pipe diameters, including 12, 16, 20, 24, 30, 40, 50, 60, 70 and 80 inches are available instead of the six smallest diameters from this list that were available for the original Hanoi case study (Fujiwara and Khang 1990).

The topology of the KLmod network case study is taken from the network used by Kang and Lansey (2012), without consideration of pumps and fire-fighting conditions. For this network, a total of ten diameters, including 150, 200, 300, 400, 500, 600, 700, 800, 900, and 1,000 mm are available for all pipes, with the unit costs given in Kadu et al. (2008).

2.3.3 Genetic algorithm optimization

The description of genetic algorithms (GAs) has been well documented (see e.g. Simpson et al. 1994) and hence, this information is not repeated in this paper. In this study, the GA used integer coding, two-point crossover, bitwise mutation, and tournament selection, as these have been demonstrated to be effective in terms of finding optimal solutions (Deb 2000; Vairavamoorthy and Ali 2005; Zheng et al. 2011b). Although a number of different GA variants have been developed over the past four decades in order to improve search performance (Dandy et al. 1996;

Nicklow et al. 2010), the use of a relatively standard GA formulation was considered adequate, as the focus of this study is on the evaluation of different methods for obtaining initial GA populations. In addition, all of the sampling approaches considered in this paper can be used in conjunction with any GA variant or other type of EA.

2.4 Computational Experiments

The four sampling methods (i.e. the PHSM, the KLSM, RS and LHS) were used to generate the initial solutions for GAs applied to each of the seven WDS case studies (Figure 2.3). The results of GAs seeded using these four sampling methods were compared in terms of objective function value and computational efficiency.

For the PHSM, the value of the initial threshold velocity v used in Step 2 was selected to be 0.1 m/s for all case studies based on the results of preliminary trials with several different values, although variations of this initial value were found to have only a slight impact on the results. It was found that the overall number of simulations required for adjusting pipe diameters in Step 2 was less than 200 for the seven case studies, and hence the maximum number of allowable simulations S_{max} was set to 1000. In Step 3 of the PHSM, a number of different values of a (see Equation 2.2) ranging from 0.1 to 2 were tried and a = 0.5 was ultimately selected, as it produced slightly better results than other a values. However, as was the case for the initial threshold velocity v, slight variations in a did not significantly influence the final results. For the KLSM, velocity thresholds were generated in accordance with Equation (2.4).

The parameter values of the GAs applied to each case study were fine-tuned with the aid of a large-scale sensitivity analysis. For the crossover probability, values ranging from 0.1 to 0.9 were tried. For the mutation probability, 10 different values around the value of *1/ND* (where *ND* is the number of decision variables) were tried for each study, as it has been demonstrated that a value of approximately *1/ND* is an effective value and is normally used for GAs (Simpson et al. 1994). The parameter values that

exhibited the best performance in terms of efficiently finding good quality optimal solutions were selected and are presented in Table 2.3. For each case study, the GAs seeded using the four sampling methods considered used the same parameter values. A penalty cost was added to the objective function value for infeasible solutions, with a penalty multiplier of 10^5/metre of head being used for all case studies (Simpson et al. 1994). The tournament size in the selection operator was two for all GAs. The maximum allowable number of evaluations for each case study is given in the last column of Table 2.3, with the larger networks assigned larger computational budgets.

Table 2.3 Parameters values of GAs for each case study

Case study	Number of decision variables (*ND*)	Network group based on the size of WDSs	Population size (*N*)	Crossover probability	Mutation probability	Total number of evaluations
Hanoi	34	G1 (*ND*<100)	100	0.9	0.02	300,000
Extended Hanoi	34		100	0.9	0.02	300,000
Fosspoly1	58		500	0.8	0.02	500,000
ZJ	164	G2 (100<*ND*<500)	500	0.9	0.006	500,000
Balerma	454		1000	0.9	0.002	1,000,000
Rural network	476		1000	0.8	0.002	1,000,000
KLmod network	1274	G3(*ND*>500)	1,000	0.9	0.0008	2,000,000

In order to facilitate easier discussion of the results, the seven case studies were assigned to three groups based on the number of decision variables (*ND*), as shown in the third column of Table 2.3. The first three case studies (Hanoi, Extended Hanoi and Fosspoly1) were assigned to G1, as their values of *ND*<100, while the ZJ, Balerma and Rural network case studies were allocated to G2 with 100<*ND*<500. The KLmod network was assigned to G3, as its *ND*>500.

The performance of each sampling method was assessed using the method outlined below:

1. For each case study, ten GA runs were performed for each of the four sampling methods using different random number seeds, resulting in a total of 40 final optimal solutions.
2. The best final solution from the 40 solutions was selected for each case study and used as a benchmark against which the performance of each sampling

method was assessed. This benchmark optimal solution was also compared with the current best known solution in the literature obtained using similar GAs, if available (see Table 2.2), in order to ensure that the results obtained in the current study are reasonable.

3. For each sampling method, the average of the best solution at each GA generation was calculated for each case study based on the ten runs with different starting random number seeds (denoted ABS). In addition, among the ten best solutions at each generation, the one with the lowest cost was selected (denoted as BBS).
4. The deviation of ABS and BBS from the corresponding benchmark optimal solution was plotted against the number of evaluations for each sampling method. This resulted in four convergence curves on the same plot, enabling a comparison of the performance of the four sampling methods considered.
5. The performance of each sampling method was also assessed in terms of its computational efficiency in being able to find near-optimal solutions. For this purpose, optimal solutions that had objective function values within 5% of the benchmark optimal solution were defined as being near-optimal.

In order to enable a fair comparison between the methods, the computational overheads associated with implementing the proposed PHSM are also considered (Table 2.4). This was achieved by converting the computational time required for each step of the proposed PHSM (see Section 2.2) to the equivalent number of network simulations using the same computer configuration (Pentium PC (Inter R) at 3.0 GHz). As shown in Table 2.4, the proposed PHSM is very efficient in computing the shortest-distance values for the network (Step 1) and generating distribution functions based on the approximate optimal solutions (Step 3), while it is relatively more time-consuming in adjusting pipe diameters based on the velocities in Step 2. This is expected, as this step involves an iterative process (see Figure 2.2). The number of equivalent network simulations that correspond to the total computational overhead required by the PHSM method is presented in the last column of Table 2.4.

As can be seen, this computational effort is negligible compared with the total computational budgets used in Table 2.3, and hence is not considered in the subsequent discussions in Section 2.5.

Table 2.4 Computational overhead analysis for the proposed sampling method (PHSM)

Case study	Number of decision variables (*ND*)	Equivalent simulations of the computational overhead used in Step 1[1]	Equivalent simulations of the computational overhead used in Step 2[1]	Equivalent simulations of the computational overhead used in Step 3[1]	Total computational overhead[2]
Hanoi	34	10	102	1	113 *(0.38%)*
Extended Hanoi	34	10	126	1	137 *(0.46%)*
Fosspoly1	58	11	153	1	165 *(0.33%)*
ZJ	164	6	147	2	155 *(0.31%)*
Balerma	454	8	165	3	176 *(0.18%)*
Rural network	476	9	171	3	183 *(0.18%)*
KLmod network	1274	14	190	4	208 *(0.10%)*

[1]The computational overhead used for each step has been converted to the equivalent number of network simulations for each case study.

[2]The computational overhead is expressed as the equivalent number of simulations and the fraction of the total computational budget this represents (in brackets)

2.5 Results and discussion

The costs of the best solutions found using the GAs initialized with the four sampling methods considered for each of the seven case studies are given in Table 2.5, with the lowest cost solutions found highlighted in bold. In addition, for the case studies to which GAs had been applied previously in the literature, the percentage deviation of the solutions found in this study compared with the best solution found using GAs reported in the literature are shown in brackets (i.e. negative percentage changes indicate that the solutions found in this study are better and vice versa). It should be noted that the results presented here are compared with those obtained using GAs in previous studies because the purpose of this study is to compare the relative performance of different initial sampling approaches. This requires the impact of the sampling approaches to be isolated from the impact of algorithm searching behavior as much as possible. Consequently, as a GA is used as the EA in this study

for reasons outlined previously, the final results obtained in this study should only be compared with those obtained using other GAs.

Table 2.5 Cost of the best solution found by each sampling method for each case study

Case study	Cost of the best solution found by each sampling method (Million)			
	RS	LHS	KLSM	PHSM
Hanoi	\$6.195 *(1.87%)*	\$6.217 *(2.24%)*	\$6.224 *(2.35%)*	**\$6.109** *(0.46%)*
Extended Hanoi	\$5.365	\$5.366	\$5.360	**\$5.346**
Fosspoly1	\$0.0294	\$0.0294	\$0.0309	**\$0.0290**
ZJ	\$7.562	\$7.560	\$7.655	**\$7.431**
Balerma	€2.125 *(-7.69%)*	€2.146 *(-6.78%)*	€2.130 *(-7.47%)*	**€2.061** *(-10.47%)*
Rural	\$36.108 *(-0.39%)*	\$36.265 *(0.04%)*	\$36.255 *(0.01%)*	**\$35.173** *(-2.97%)*
KLmod	\$8.686	\$8.737	\$8.418	**\$8.307**

Note: The result of each sampling method for each case study was obtained over 10 runs with different random number seeds. The percentage of the cost of each best solution relative to the best solution found by GAs is given in italics in the brackets. The benchmark optimal solution for each case study is indicated in bold.

From Table 2.5, it can be clearly seen that by using the proposed PHSM, better quality solutions could be found for each case study within the given computational budgets than when the other approaches were used. The KLSM produced better solutions for the large KLmod network compared to the other two non-heuristic sampling methods (RS and LHS). This agrees well with the observations made by Kang and Lansey (2012). However, for five of the other six case studies, RS performed better than the KLSM in terms of the quality of the final solutions.

The convergence plots for each of the algorithms for the case studies belonging to the three different groups defined in Table 2.3, as defined in the previous section, are given in Figures 2.4 (G1), 2.5 (G2) and 2.6 (G3) and provide an indication of both solution quality and computational efficiency. A common observation is that the PHSM generally found significantly better initial solutions than the non-heuristic sampling approaches. This is most likely because the initial solutions obtained using the PHSM were feasible and the diameters for these solutions were generally in a reasonable range based on velocities, nodal demands and elevations. For the larger case studies, the PHSM also found significantly better initial solutions than the KLSM. The initial solutions obtained using the other sampling methods were

typically infeasible for the larger case studies or feasible with high costs for the simpler WDSs. This demonstrates that the proposed domain knowledge based sampling method is effective in identifying good quality starting solutions. A detailed discussion of the results for the three groups of case studies is given in the subsequent sections.

2.5.1 Group 1 (G1) case studies

As can be seen from Table 2.5 and Figure 2.4, the performance of the GAs initialized with the four different sampling methods is very similar for the G1 case studies (i.e. Hanoi, Extended Hanoi and Fosspoly1), both in terms of the ability to find optimal solutions and computational efficiency. While the GAs initialized with the PHSM were able to find the best solution for all three case studies, the variation in the cost of the best-found solutions was relatively small (Table 2.5). Similarly, while GAs initialized with the PHSM found better initial solutions and generally converged more quickly than the GAs initialized with the other methods (Figure 2.4), this difference was not very large. Consequently, based on the results obtained, there does not appear to be a significant advantage of using domain knowledge for the initialization of GAs for small problems, such as those considered for the G1 case studies.

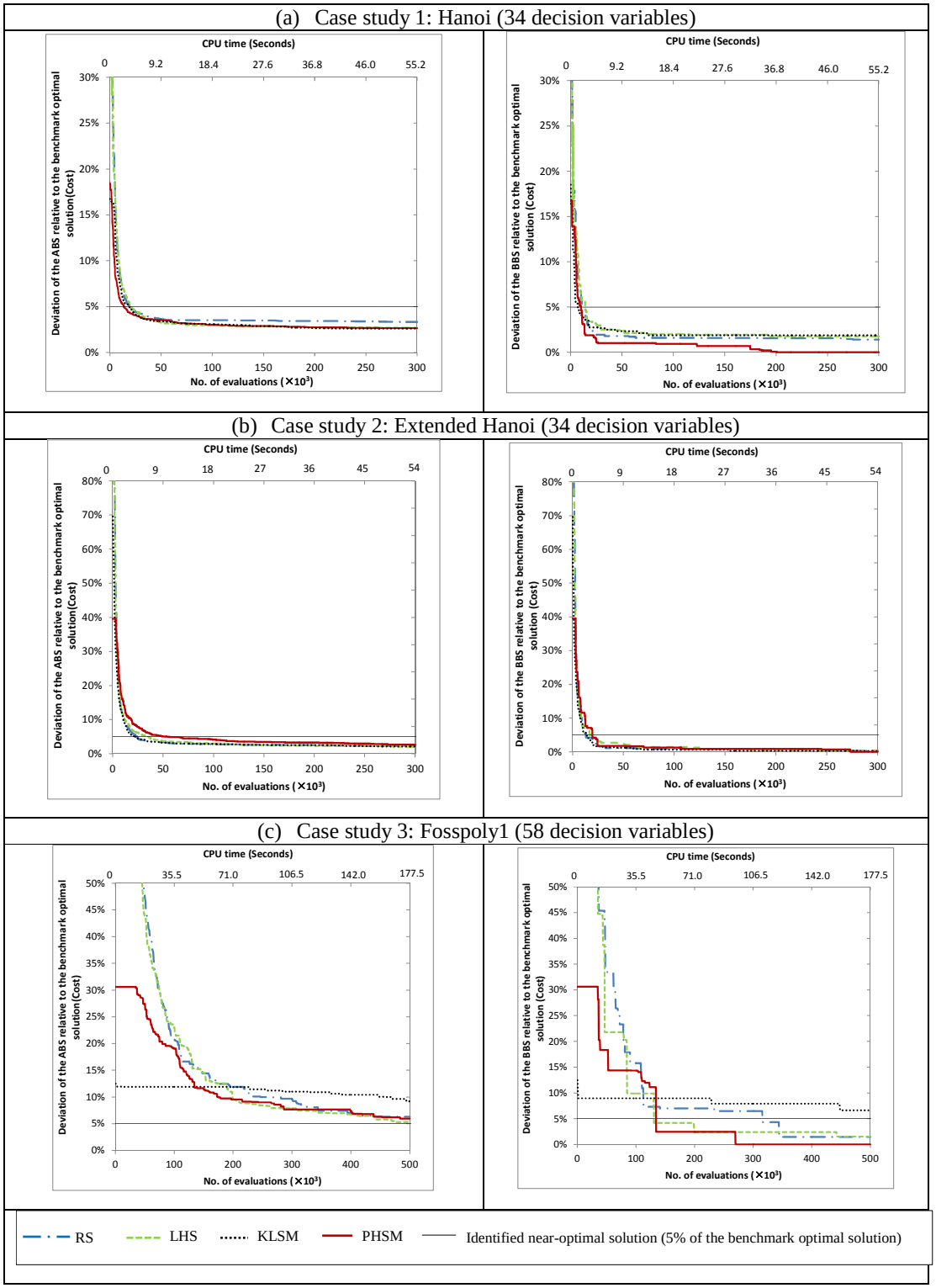

Figure 2.4 Results of the GAs with the four sampling methods applied to case studies in Group 1 (G1 in Table 2.3)

2.5.2 Group 2 (G2) case studies

As can be seen from Table 2.5 and Figure 2.5, the performance of the GA initialized with the PHSM is noticeably better than that of the GAs initialized with the other three methods for the G2 (ZJ, Balerma, Rural) networks, both in terms of the best-found solution and computational efficiency. This suggests that while for the simpler G1 case studies the GAs were able to find good solutions relatively quickly with the aid of their evolutionary operators, irrespective of the starting position in solution space, this is not the case for the more complex G2 case studies. This demonstrates that the better starting positions in solution space identified using the PHSM are able to assist the GA with finding better regions of larger search spaces, as indicated by the better solutions found when the GAs were initialized with the PHSM (Table 2.5 and Figure 2.5). This trend was already noticeable for the Fosspoly1 case study, which is the most complex of the G1 case studies (Figure 2.4).

The results in Table 2.5 and Figure 2.5 also indicate that the solutions found using the PHSM were not only better than those obtained using RS and LHS, but also better than those obtained using the other heuristic sampling method (i.e. the KLSM). This appears to be both as a result of the quality and diversity of the initial solutions. For example, for the ZJ and Rural networks, the PHSM was able to identify significantly better initial solutions than the KLSM, resulting in more rapid convergence and better final solutions (Figure 2.5). In contrast, for the Balerma network, use of the KLSM resulted in better initial solutions than use of the PHSM. However, despite this initial disadvantage, use of the PHSM resulted in more rapid convergence and the identification of better solutions, which is likely due to the additional control over population diversity offered by the PHSM. A similar trend was also observed for the Fosspoly1 network (Figure 2.4), which is the largest of the G1 networks. It should be noted that the better performance of the PHSM was not affected by the presence of multiple source reservoirs, as is the case for the Balerma network. This suggests that the approach of using an augmented source node for networks with more than one source reservoir (as described in Step 1 for the PHSM) is effective.

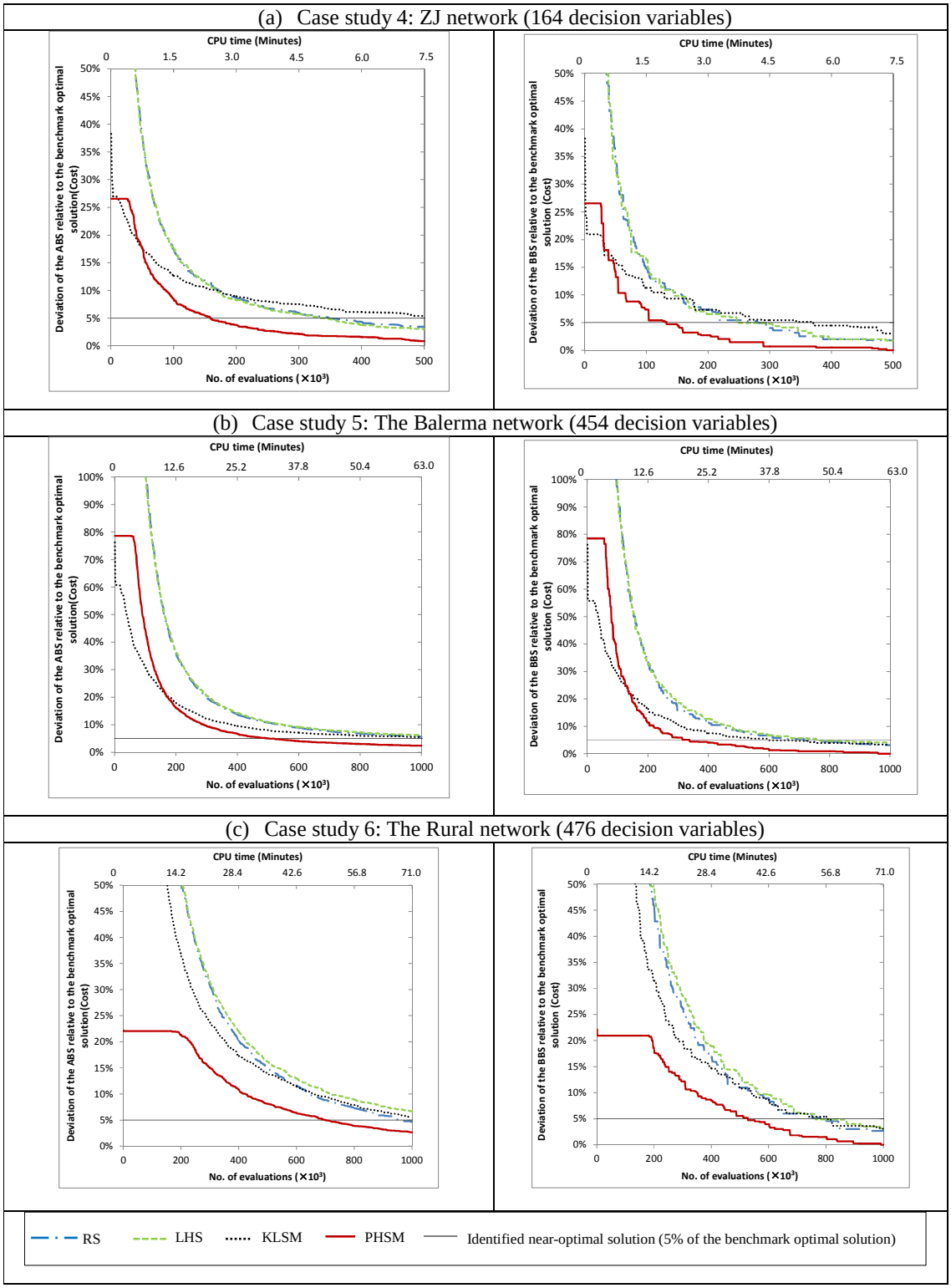

Figure 2.5 Results of the GAs with the four sampling methods applied to case studies in Group 2 (G2 in Table 2.3)

In terms of the quality of the solutions found, use of the PHSM resulted in the best solutions for all three G2 case studies by some margin (Table 2.5, Figure 2.5). In contrast, the quality of the solutions found using the other three initialization methods is quite similar, with no advantage of using the KLSM. It should also be noted that for the two case studies to which similar GAs had been applied in previous studies, the GA initialized with the PHSM found solutions that were 10.47% and 2.97% better than those found in previous studies for the Balerma and Rural networks, respectively (Table 2.5).

As far as convergence speed is concerned, use of the PHSM results in significantly faster convergence to near-optimal solutions (i.e. solutions that are within 5% of the benchmark optimal solution, as defined previously) than use of the other three initialization methods, which all performed similarly (Figure 2.5). This indicates that there is likely to be a significant advantage in using the PHSM when trying to find the best possible solution within reasonable computational budgets for complex networks.

2.5.3 Group 3 (G3) case studies

As can be seen from Table 2.5 and Figure 2.6, for this very large network (i.e. KLmod), the performance of the GAs initialized with both heuristic sampling methods (i.e. PHSM and KLSM) are noticeably better than that of the GAs initialized with the two non-heuristic sampling methods (i.e. RS and LHS), both in terms of the best-found solution and computational efficiency. While the GAs initialized using the two heuristic sampling methods were able to find near-optimal solutions after approximately 800,000 evaluations for the average solutions based on ten runs, which is equivalent to approximately 3 hours in terms of CPU time, the GAs initialized with the non-heuristic sampling methods (i.e. RS and LHS) were not even able to find solutions of this quality at the end of the optimization run (using nearly 2,000,000 evaluations and approximately 7 hours of CPU time). Although Figure 2.6 suggests that the GAs initialized with RS and LHS had not converged yet and might ultimately

find solutions of a similar quality to those found when the heuristic sampling methods were used, the computational effort required to do so is likely to be very large. This clearly highlights the advantage of using heuristic sampling methods for initializing GA populations for larger networks.

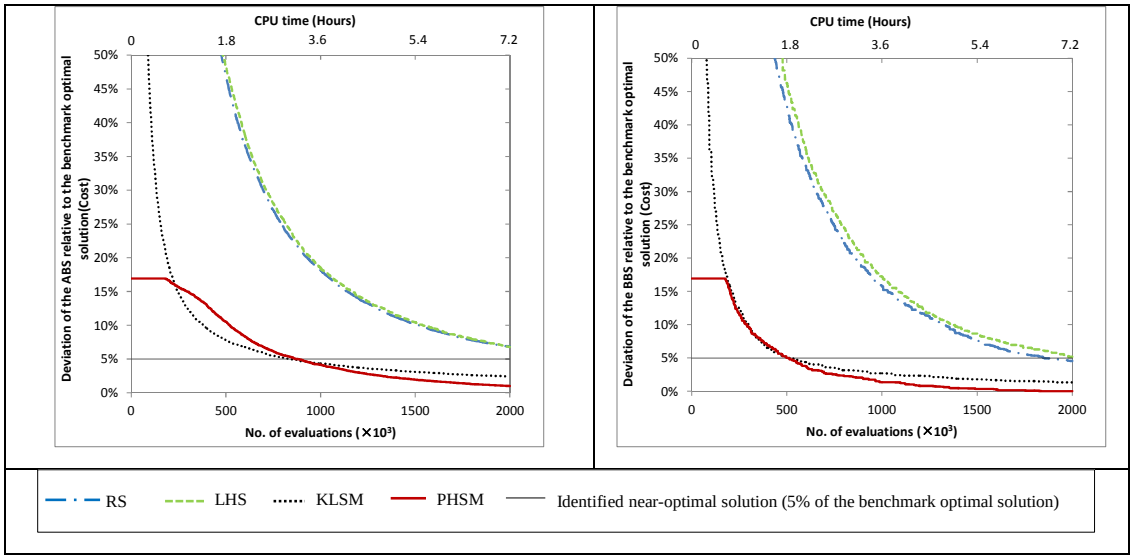

Figure 2.6 Results of the GAs with the four sampling methods applied to case studies in Group 3 (G3 in Table 2.3)

In terms of the relative performance of the two heuristic sampling methods, while both converged to near-optimal solutions after approximately the same number of iterations, use of the PHSM resulted in clearly better best-found solutions. This is likely to be due to a combination of the better initial solutions identified using the PHSM, as well as the additional control over population diversity afforded by the PHSM. However, the relative performance of the KLSM compared with that of the PHSM was much better for the KLmod case study, which is most likely because the KLSM was designed for a modified version of this problem.

2.6 Summary and conclusions

In order to improve the ability of GAs to find optimal or near-optimal solutions in reasonable timeframes for realistic-sized water distribution optimization problems, a new heuristic sampling method (the PHSM) for initialising GA populations was introduced

and evaluated in this paper. The performance of the PHSM was compared with that of an existing heuristic sampling method (the KLSM) and with that of more traditional sampling methods, including RS and LHS, for seven WDSs of varying size and complexity.

The results obtained based on the seven WDS optimization (pipe-sizing) problems considered indicate that overall, the proposed PHSM performed significantly better than the other three sampling methods, both in terms of solution quality and computational efficiency. It was also found that the relative advantage of the PHSM increased with network size and complexity. While for the smaller (G1) networks, the performance of the GAs initialised using the four different methods was very similar, there were clear advantages in using the PHSM for the larger (G2) networks and in using both heuristic sampling methods (i.e. PHSM and KLSM) for the largest network considered (G3). This advantage is likely to be due to the ability to find better initial solutions, enabling more favourable regions of the solution space to be explored more quickly. The results also indicate that PHSM outperforms the KLSM, which is likely due to a combination of the ability to find better initial solutions and the additional population diversity provided by the PHSM.

As the focus of this paper was on the development and evaluation of the PHSM, all analyses were conducted using a reasonably standard GA. However, as the PHSM is independent of the optimization algorithm used, it can be tested in combination with other algorithms. Such investigations would be useful in terms of assessing the generality of the results obtained in this paper. In addition, it would be useful to extend and apply the proposed approach to a larger number of case studies with increased hydraulic complexity, such as the inclusion of tanks, valves and pumps. However, given that pipe sizes generally represent the largest number of decision variables, application of the PHSM to the subset of the decision variables consisting of pipe diameters is still likely to be beneficial for WDSs including tanks, valves and pumps. Finally, it would be interesting to compare the performance of the PHSM with that of

other methods that could be used for initialising EAs, such as the cellular automata network design algorithm of Keedwell and Khu (2006).

Chapter 3. Journal Paper 2- Impact of starting position and searching mechanism on evolutionary algorithm convergence rate

Statement of Authorship

Title of Paper	Impact of starting position and searching mechanism on evolutionary algorithm convergence rate
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Publication Style
Publication Details	Bi, W., Maier, H. R. and Dandy, G. C. (2015). "Impact of starting position and searching mechanism on evolutionary algorithm convergence rate." *Submitted to Journal of Water Resources Planning and Management, July, 2015.*

Principal Author

Name of Principal Author (Candidate)	Weiwei Bi		
Contribution to the Paper	Develop the approach, perform the simulation study and prepare the manuscript		
Overall percentage (%)	50%		
Signature		Date	

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 R. Maier		
Contribution to the Paper	Research supervision and review of manuscript.		
Signature	25%	Date	

Name of Co-Author	Graeme C. Dandy		
Contribution to the Paper	Research supervision and review of manuscript.		
Signature	25%	Date	

Abstract

Traditionally, evolutionary algorithms (EAs) have been used to attempt to find globally optimal solutions for water distribution system (WDS) optimization problems. However, as these algorithms are being applied to increasingly complex systems, computational efficiency is becoming an issue, and hence approaches that enable near-optimal solutions to be identified within reasonable computational budgets have received increasing attention. One of these approaches is the initialization of EAs in a manner that accounts for domain knowledge of WDS design problems. While the effectiveness of these initialization approaches has been studied previously, the impact of algorithm searching behavior on the speed with which near-optimal solutions can be found has not yet been examined. To this end, this study aims to investigate the relative impact of different algorithm initialization methods and searching mechanisms on the speed with which near-optimal solutions can be identified for large WDS optimization problems. Fitness function and run-time behavioral statistics are used to gain an increased understanding of the behaviour. The results show that both the starting population and algorithm searching mechanism have an impact on the speed with which near-optimal solutions are identified. The fitness function and run-time behavioral statistics indicate that EA parameterizations that favor exploitation over exploration enable near-optimal solutions to be identified earlier in the search, which is due to the “big bowl” shape of the fitness function for all of the WDS problems considered. Using initial populations that are informed by domain knowledge further increases the speed with which near-optimal solutions can be identified.

CE Database subject headings: searching mechanism; evolutionary algorithm; optimization; sampling method; water distribution system.

Author Keywords: searching mechanism; evolutionary algorithm; optimization; sampling method; water distribution system.

3.1 Introduction

Evolutionary algorithms (EAs) have been used extensively for various water resources optimization problems over the past few decades (Nicklow et al. 2010; Maier et al. 2014, 2015). Their main advantages compared with traditional deterministic approaches include (i) increased ability in exploring the entire search space, leading to a higher likelihood of arriving at good quality solutions (Nicklow et al. 2010); (ii) the ability to be linked with any simulation models (Zheng et al. 2013a; Beh et al. 2015), and (iii) greater adaptability in handling water resources problems with multiple conflicting objectives (Kapelan et al. 2005; Ostfeld et al. 2014).

However, the application of the EAs is not without difficulties, with one of the typical challenges being their larger demands on computational time (di Pierro et al. 2009; Zheng et al. 2013b). This is especially the case when dealing with realistic water resources problems, such as the design of large-scale water distribution systems (WDSs) (Marchi et al. 2014b), which are investigated in the present study. In fact, as highlighted in Maier et al. (2014), the relatively low computational efficiency of EAs has become a main barrier to their wider up-take in industry.

In order to address this issue, there is general consensus that finding near-optimal solutions as quickly as possible, rather than trying to find the best possible solution, or finding the best possible solution within a given computational budget (e.g. Gibbs et al. 2010; 2015), is of great importance (Maier et al. 2014). This is because, as, from a practical perspective, there is generally insufficient time to run the optimization until no further improvement in objective function values are obtained when dealing with real-world problems.

One way to increase the computational efficiency of EAs so that near-optimal solutions can be found within realistic timeframes is to initialize their searching in promising regions of the solution space based on an understanding of the physics of the problem being solved. In terms of WDS design optimization, Keedwell and Khu (2006) considered intuitive knowledge of the way in which WDSs function to

generate the initial solutions for EAs. As part of their approach, if a demand node in the WDS has a head deficit or surplus, the diameters of the pipes that are connected to this node are increased or decreased (respectively). Subsequently, Zheng et al. (2011a) seeded EAs with solutions from an optimal tree network based on the knowledge that the optimal solution for a WDS subject to a single loading condition consists of a branched topology without any loops.

More recently, Kang and Lansey (2012) incorporated engineering experience into the initialization of EAs through the use of the optimal flow velocity within the pipes. Three steps are involved in their method, which are (1) all pipes to be optimized are given the minimum allowable diameter, (2) a hydraulic simulation model is used to calculate the flow velocity in each pipe, and (3) the diameters are increased to the next available size if the obtained flow velocity is larger than the preset optimal velocity that is determined based on engineering experience, and vice versa. Steps (2) and (3) are repeated until flow velocities in all pipes are below the given optimal velocity, and the resulting pipe sizes form an initial solution that overall has a velocity close to the optimal velocity in each pipe (domain knowledge). A set of different initial solutions is obtained through the use of different optimal velocities according to engineering experience, and the EAs are seeded with these solutions to find near-optimal solutions with increased computational efficiency (Kang and Lansey 2012).

Building on the work of Kang and Lansey (2012), Bi et al. (2015) proposed an initialization (sampling) method that accounted for the fact that pipe sizes generally reduced with distance from the source (Walski, 2001), in addition to considering optimal flow velocities. As part of the approach, initial EA populations are generated by sampling in the vicinity of the solutions identified based on this domain knowledge, in order to avoid premature convergence to local optima in solution space. Bi et al. (2015) found that their method outperformed the approach of Kang and

Lansey (2012) for a set of WDS case studies with different sizes and complexity in terms of the speed with which near-optimal solutions were identified.

While the impact of the initialization (initial population, starting position in solution space) of EAs on the *speed* (in terms of computational effort) with which *near*-optimal solutions can be found has been investigated previously, as outlined above, the impact of using different EAs (e.g. genetic algorithm, differential evolution) and EA parameterizations (e.g., mutation rate) on the relative performance of these initialization methods has not yet been studied. This is despite the fact that algorithm searching behavior, as represented by different EAs and parameterizations, is known to have a significant impact on algorithm convergence (Zheng et al. 2015a). In other words, while there have been many studies that have examined the impact of different EAs and EA parameterizations on convergence rate and the ability to find globally optimal solutions, their impact on the ability to find *near*-optimal solutions for real-life problems with limited computational budgets using EAs that have been initialized by methods using varying degrees of domain knowledge has not yet been investigated.

In order to overcome the above shortcoming, the primary objective of this paper is to investigate the relative impact of different algorithm initializations, different EAs and different EA parameterizations on the speed with which near-optimal solutions can be identified for a number of WDS optimization problems of varying complexity. In order to obtain a better understanding of the relative performance of different algorithm initialization methods and searching behaviours, a secondary objective of this paper is to examine the properties of the fitness functions of the case studies and the run-time behavioral statistics of the different algorithms and their parameterizations, and how they relate to observed algorithm performance, as suggested by Maier et al. (2014). The remainder of this paper is organized as follows. The methodology is given in the next section, followed by details of the computational experiments that have been conducted in order to meet the objectives.

Next, the results are presented and discussed, before the paper is summarized and conclusions are drawn.

3.2 Methodology

Figure 3.1 presents the overall methodology used in the present study. As stated previously, the relative impact of the EA initializations (starting positions) and algorithm behaviours (algorithms and their parameterizations) on the convergence rate to near-optimal solutions are assessed. Two different initialization (sampling) methods and two different types of EAs are considered, as shown in Figure 3.1. For each EA, a suite of different parameterizations is used. The two EAs with different initialization methods and parameterizations are applied to four large WDS design problems, for which the number of decision variables ranges from 164 to 1,274. To gain an improved understanding of the results in terms of the speed with which near-optimal solutions are identified, the problem characteristics, as well as the run-time algorithm searching behavior, are analyzed. Details of each element in Figure 3.1 are discussed in following sub-sections.

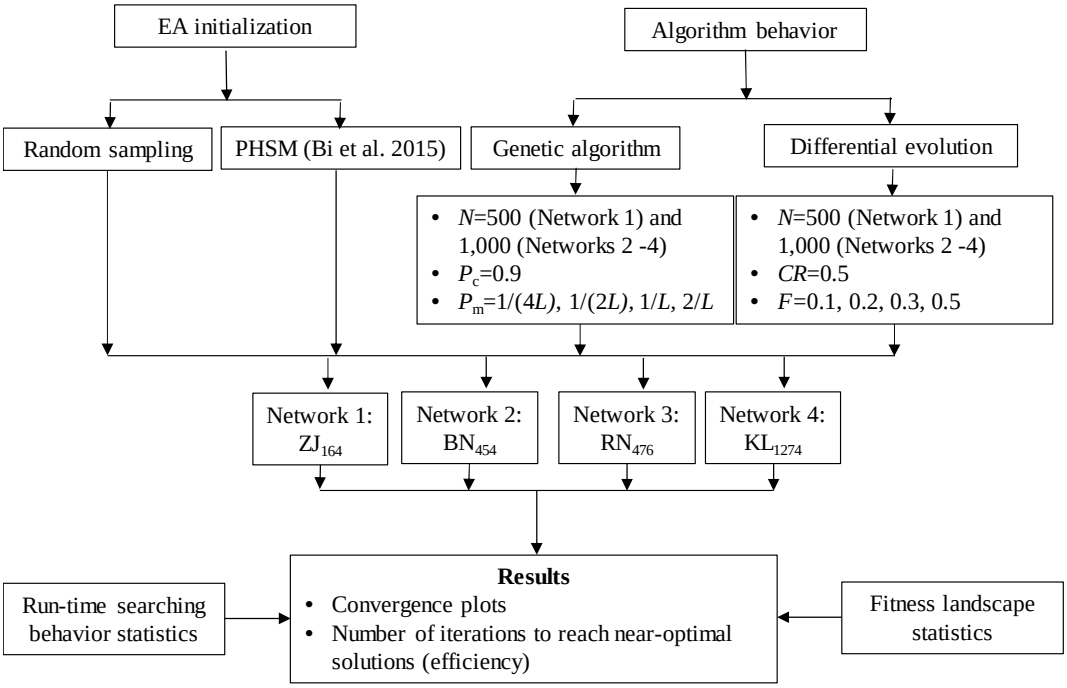

Figure 3.1 Flowchart of the assessment process, where *N* is the population size, *L* is the number of decision variables, P_c and *CR* are the crossover probabilities for the genetic algorithm (GA) and differential evolution (DE), respectively, and P_m and *F* are their mutation probabilities. The subscript of each case study indicates the number of decision variables.

3.2.1 Initialization approaches

As shown in Figure 3.1, the two initialization methods considered are random sampling (RS), which is the most commonly used initialization method, and the Prescreened Heuristic Sampling Method (PHSM) (Bi et al. 2015). The PHSM is used to represent the class of initialization approaches that take into account domain knowledge, as it performed better than the method of Kang and Lansey (2012) when applied to case studies used in this paper (Bi et al. 2015). The three main steps in the PHSM (Bi et al. 2015) include:

- Step 1: Assign pipe diameters to all pipes based on the distance between demand nodes and supply sources. This is motivated by the knowledge that, in real WDSs, the diameters of upstream pipes are generally larger than those of pipes further downstream (Walski 2001).

- Step 2: Adjust pipe diameters based on velocities. In this step, the diameters obtained in Step 1 are refined to achieve flow velocities in all pipes that are close to a particular threshold. This is based on the domain knowledge that the velocity in each pipe of an optimal solution for a WDS is within a limited range.
- Step 3: Generate initial population by sampling from distribution functions centered around the approximate optimal solution determined in Step 2. In order to ensure sufficient diversity in the initial solution, the initial diameter for each pipe is generated from a distribution, such that the pipe diameter obtained in Step 2 has the highest probability of being selected.

3.2.2 Evolutionary algorithms and their parameterization

Genetic algorithms (GAs) and differential evolution algorithms (DEs) are considered in the present study. These EAs are selected because GAs have been widely recognized as an industry standard optimization technique (Wang et al. 2015), while DEs have been shown to outperform GAs in terms of computational efficiency and the ability to find optimal solutions in recent WDS studies (Vasan and Simonovic 2010; Zheng et al. 2013c). Details of the GA algorithm adopted are given in Bi et al. (2015), but an elitism scheme was added in this study to facilitate better comparison with the DE. Details of the DE algorithm used are given in Zheng et al. (2011a). It should be noted that, for each algorithm, constraint tournament selection is used to handle infeasible solutions (Deb et al. 2000).

Many studies have shown that the mutation operator in GAs and DEs can have a more significant impact on searching behaviour, and hence algorithm performance, than other parameters, such as crossover and population size (Reca and Martinez 2006; Zheng et al. 2015a). This is because different mutation rates can substantially alter the balance between exploration (i.e. broadly searching the solution space) and exploitation (i.e. focusing on the local regions) (see Maier et al. 2014), during an algorithm's search. In order to explore the influence of different degrees of

exploration and exploitation on EA convergence rate and the ability to find near-optimal solutions, a number of different mutation probabilities are therefore used for both the GA and DE. Details of the adopted values of the mutation rates and other algorithm parameters are shown in Figure 3.1 for both EAs. As can be seen, in addition to the recommend mutation probability of $P_m=1/L$ (where L is the number of decision variables in the real-value coding scheme) (Wang et al. 2015), mutation rates of $1/(4L)$, $1/(2L)$ and $2/L$ are considered for the GAs applied to each case study.

Zheng et al. (2015a) conducted a comprehensive study to analyze the impact of the DE parameters (mutation factor F and crossover rate CR) on its searching performance. They concluded that DE performance was more influenced by the value of F rather than CR, and F=0.3 and CR=0.5 were recommended as default parameters for relatively large optimization problems. Following their work, CR=0.5 is used for each case study and mutation factors of F=0.1, 0.2, 0.3, and 0.5 are considered, as they represent significantly different searching behavior. For each case study, the population size N outlined in Bi et al. (2015) is also used in the present study for both EAs, which is N=500 for the ZJ_{164} problem and N=1,000 for the other three WDS design case studies as shown in Figure 3.1.

3.2.3 Case studies

As shown in Figure 3.1, four large-scale case studies are considered, the details of which are given in Bi et al. (2015). These large problems are chosen because they are more relevant than simpler case studies for the purposes of this study, which is aimed at assessing the effectiveness of different algorithm initializations and behaviours in terms of finding near-optimal solutions for large WDSs that are representative of those encountered when solving real-world problems as quickly as possible.

The aim of the optimization problem is to identify the n pipe diameters $D=[d_1,d_2,\cdots,d_n]^T$ that correspond to the least cost design solution D^*, subject to the satisfaction of a number of constraints. That is

$$D^* = \text{argmin} \sum_{i=1}^{n} c_i(d_i) \tag{3.1}$$

where

$$P(D^*) \geq P_{\min} \tag{3.2}$$

$$d_i \in \{A\} \tag{3.3}$$

where d_i is the diameter of pipe i=1,2….,n; c_i is the cost function for pipe i associated with the choice of decision variable d_i; $P(D^*)$ is the nodal pressure vector for design solution D^*, which has to be greater than the minimum allowable pressure vector $P_{\min}$ for demand nodes under a set of design demands for the WDS (feasible solution). A hydraulic simulation model (EPANET2.0 in this study) is typically used to determine $P(D^*)$. $\boldsymbol{A}$ is a set of commercially available pipe diameters (discrete) for the given WDS design problem.

3.2.4 Performance assessment

The results of the optimization runs are presented in terms of the convergence plots and the number of iterations required to identify near-optimal solutions, as was done by Bi et al. (2015). Near-optimal solutions are defined as solutions that are within 5% of the best-known solution for each of the case studies, which are also given in Bi et al. (2015). The selection of 5% is based on the fact that it is commonly used in statistical analysis and also, from a practical point of view, it could be considered that being within 5% of the minimum cost solution is sufficiently close. However, it is recognized that, in certain circumstances a smaller value (e.g. 2% or 1%) would be more appropriate.

In order to minimize any influence of the stochastic elements of the two EAs considered, the results presented are averaged over 10 runs with different random number seeds. The metric used for this purpose is the percentage deviation of the mean cost to the best known solution (DMO%), to enable the results from different case studies to be compared in an objective fashion.

3.2.5 Performance explanation

To gain an improved understanding of the optimization results achieved by using different starting positions, EAs and EA parameterizations, the characteristics of the fitness functions and run-time behavior statistics are assessed (see Figure 3.1), as detailed in the sub-sections below.

Fitness function properties

Previous studies have recognized the importance of the characteristics of the problem on the success of a given optimization algorithm (Gibbs et al. 2011, 2015). Therefore, there is a growing interest in quantifying the characteristics of water resources optimization problems to provide guidance for selecting the most effective searching algorithms and algorithm parameterizations (Maier et al. 2014). For instance, Gibbs et al. (2011) proposed a number of statistics to quantify the properties of the fitness function (see Maier et al., 2014) of two operational WDS optimization problems, including the spatial correlation and mutual information between decision variables, which were used to guide the determination of the most appropriate GA parameters. Subsequently, the same authors quantified the characteristics of the fitness function of two optimization problems with regard to water quality within WDSs, including one real-world WDS in Sydney, Australia. This information was used to determine the most appropriate GA population size when the number of available evaluations was limited (Gibbs et al. 2015). In order to better understand and explain the algorithm behaviour observed in this study, the fitness function characteristics of the WDS design problems considered are calculated using two metrics: the spatial correlation (Gibbs et al. 2011) and the dispersion metric (Arsenault et al. 2014).

Spatial correlation

Spatial correlation is often used to identify the macrostructure of the solution space, and is typically represented by the correlation length R_l (Weinberger 1990), the total correlation strength in the fitness function over the correlation length R_s (Gibbs et al.

2011) and the total positive correlation strength R_p (Gibbs et al. 2015). Figure 3.2 illustrates the meaning of the R_l, R_s and R_p metrics using a hypothetical case. As shown in this figure, R_l corresponds to the shortest distance at which the correlation value falls below zero, R_s is given by the area under the plot of correlation for a distance no greater than R_l, and R_p corresponds to the area under the correlation plot with positive values.

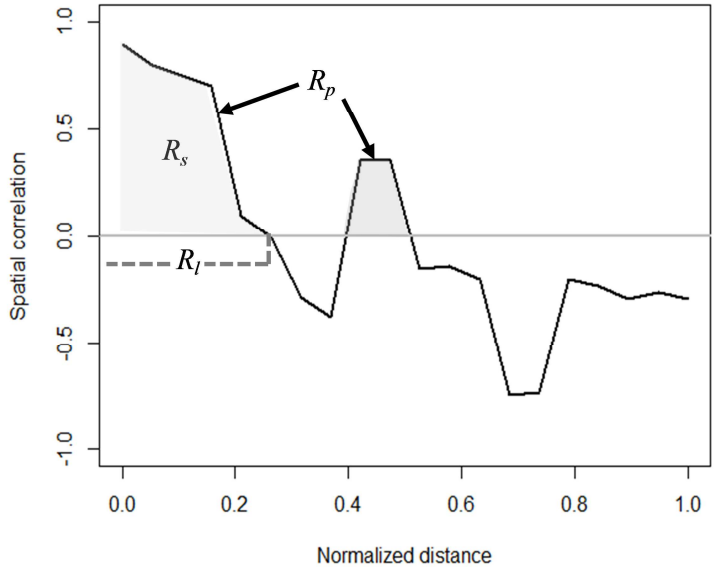

Figure 3.2 Illustration of the spatial correlation statistics using a hypothetical case.

The difference between R_s and R_p provides useful insight into the global structure of the search space. For example, if $R_s < R_p$, points that are far apart in the search space are positively correlated, suggesting a complex global structure with multiple correlated regions, as illustrated in Figure 3.2. In contrast, if $R_s = R_p$, then the plot of correlation versus distance in the search space does not become positive at distances that are greater than the correlation length, which is indicative of a single big bowl shape. More details of the spatial correlation are given in Gibbs et al. (2011, 2015).

Dispersion metric

While the spatial correlation outlined above focuses on the macrostructure of the solution space, the dispersion metric aims to provide greater insight to the

microstructure (promising regions). This metric uses iterative random sampling of the search space to measure the average pairwise distance between the *m*-best solutions (i.e., solutions with lowest cost) from a population of *N* samples, where the value of *m* is fixed (e.g. 100) and *n* is variable (e.g. from 1,000 to 10,000). A fast decrease in the average pairwise distance when *N* is increased means the fitness function has a smooth microstructure. In contrast, the value of the dispersion metric is expected to gradually decrease or even remain constant for a complex and rugged search space. This metric has previously been used to measure landscape properties of hydrological calibration problems (Arsenault et al. 2014), but this is the first time it has been used to investigate the fitness function structure of WDS design optimization problems.

The dispersion metric is illustrated in Figure 3.3 using a hypothetical case. The top and bottom panels indicate relatively smooth (a single global optimum) and rugged (multiple optimal solutions) search spaces, respectively. The mean pairwise distance between the 30 best solutions selected from the random solutions (lowest objective function values) in the top panel is expected to decrease more quickly compared with that in the bottom panel when increasing the sample size *N* from 100 to 200. This is because the pairwise distance between the top solutions in the bottom panel is dominated by the distance between the optimal solutions in different sub-regions.

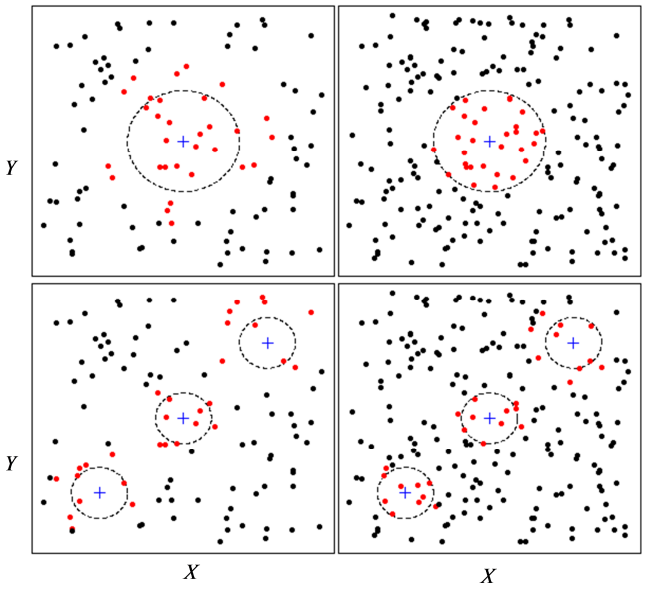

Figure 3.3 Illustration of the dispersion metric using a hypothetical case. Dots represent solutions in the two-dimensional domain and the red dots indicate the 30 best solutions across different sample sizes. The blue "+" indicate the local optima and the dashed circles show the promising regions.

Searching behavior metrics

Two metrics from the literature are selected in order to better understand the run-time searching behavior for the different optimization runs, namely: objective function cost (in objective space) and population diversity (in decision space). The most straightforward metric for assessing search quality during an optimization run is the objective function value of the best solution found at each generation $f_{best}(G)$ (Zecchin et al. 2012). For a single-objective minimization problem, this can be expressed as

$$f_{best}(G) = \min f(\mathbf{X}_G) \tag{3.4}$$

where $\mathbf{X}_G = [X_{1,G}, X_{2,G}, \cdots, X_{N,G}]^T$ is the population with N solution vectors at generation G=1, 2….,G_{max}. This metric can effectively characterize an algorithm's searching behavior, such as how algorithms with different parameterizations and starting points approach the optimal solution, and how an algorithm's searching

performance temporally varies (e.g., which stage of searching is most productive in reducing the cost).

The population diversity in decision space is typically measured by the averaged pairwise Hamming distance of the population (Zecchin et al. 2012; Zheng et al. 2015a). However, the Hamming distance only measures the existence of the difference between each bit of the solution string without considering the magnitude of this difference. Consequently, in the present study, the mean of the magnitude of the pairwise distance, $d_{mean}(G)$, is used to characterize population diversity, where

$$d_{mean}(G) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{k=i+1}^{N} H(X_{i,G}, X_{k,G}) \tag{3.5}$$

where $N(N-1)/2$ is the total number of pairs of candidate solutions (N is the population size); $H(X_{i,G}, X_{k,G})$ measures the degree of the topological distance between solutions $X_{i,G}$ and $X_{k,G}$. For example, by using the proposed metric, if all available options of the decision variables are [200, 400, 500, 600], $X_{i,G}$=200, and $X_{k,G}$=600, then $H(X_{i,G}, X_{k,G})$=3, rather than 1 (Hamming distance). As such, in addition to accounting for the presence of any differences in the solutions, the magnitude of the difference between two solutions is also considered. The metric in Equation (3.5) quantitatively measures the spread of solutions over the decision space. A large value of $d_{mean}(G)$ reveals that the current search is broadly exploring the decision space, while a low value of $d_{mean}(G)$ is indicative of localized exploitation (Maier et al. 2014).

The population diversity metric presented in this paper (SPD%) is standardized as follows

$$SPD(\%) = \frac{d_{mean}(G)}{LC} \times 100\% \tag{3.6}$$

where L is the number of the decision variables and C the number of possible pipe diameters. As such, population diversity can be compared for different case studies.

To better understand the relationship between searching quality (objective function cost, DMO%) and searching diversity (SPD%), dynamic correlations are estimated between these two run-time series. The correlation at generation *G* between DMO% and SPD% is estimated using DMO%[1:*G*] and SPD%[1:*G*].

3.3 Computational Experiments

As mentioned previously, for each EA with a different starting position and parameterization, ten runs with different random number seeds are performed for each case study. The averaged results over the ten runs are presented to minimize the impact of the stochastic nature of the EAs. For the large case studies considered, the typical computational budgets in terms of the maximum number of generations ranged from 1,000 (Kang and Lansey 2012) to 2,500 (Wang et al. 2015). To enable a more comprehensive analysis on the run-time searching behavior during different stages, the maximum number allowable generations for each case study is set as 5,000 in the present work. This results in a total of 2×10^8 simulations for the ZJ_{164} problem, and 4×10^8 simulations for each of the other three case studies (BN_{454}, RN_{476} and KL_{1274}) for all combinations of EA parameterizations. This takes approximately 60 days using a Pentium PC (Inter R) at 3.0 GHz.

To obtain the fitness function statistics, a sample size of 10,000 is used for each case study, guided by Gibbs et al. (2011). The samples are generated using both random and Latin hypercube sampling. As the resulting statistics were very similar, only the results for random sampling are presented. The mean of the magnitude of the pairwise distances given in Equation (3.5) is used to calculate the dispersion metric, with m=100 (the number of the best solutions) and N (sample sizes) ranging from 1,000 to 10,000 following Arsenault et al. (2014). The analysis of spatial correlation and dispersion metric were repeated ten times, using 10,000 samples generated with different random number seeds, and the results obtained were similar.

The dispersion metric was also calculated for the New York Tunnel problem (NYTP), which only has 21 decision variables and is widely acknowledged in the literature as

a rather simple problem with a large proportion of feasible regions (Zheng et al. 2015a). This enables the dispersion metric for the NYTP to be used as a benchmark against which the relative roughness of the microstructure of the four complex case studies considered in this paper can be assessed.

3.4 Results and discussions

Figures 3.4 and 3.5 present the changes in the solution quality and diversity metrics (Equations 3.4 and 3.5) over the different optimization runs for the GA and DE with different initializations and parameterizations, applied to the four case studies. The black and red lines represent results for initializations using the RS and PHS methods, respectively, with different line types indicating different algorithm parameterizations.

3.4.1 Impact of the starting positions and searching mechanisms

As can be seen from the left panels in Figures 3.4 and 3.5, overall, the EAs that are initialized with the PHSM are able to converge more rapidly than those initialized with random sampling for the same parameterizations and case studies. This is independent of case study and algorithm searching behavior (i.e. type of EA and EA parameterization), thereby extending the findings of Bi et al. (2015), who used a single EA with a single parametrization. The greater effectiveness of the PHSM is mainly due to the good starting positions it is able to identify. For example, the deviations from the best-known values of the initial solutions for the ZJ_{164}, RN_{476} and KL_{1274} problems are approximately DMO% = 30%, 40% and 20%, respectively. These values are appreciably lower relative to the corresponding values obtained using the RS method.

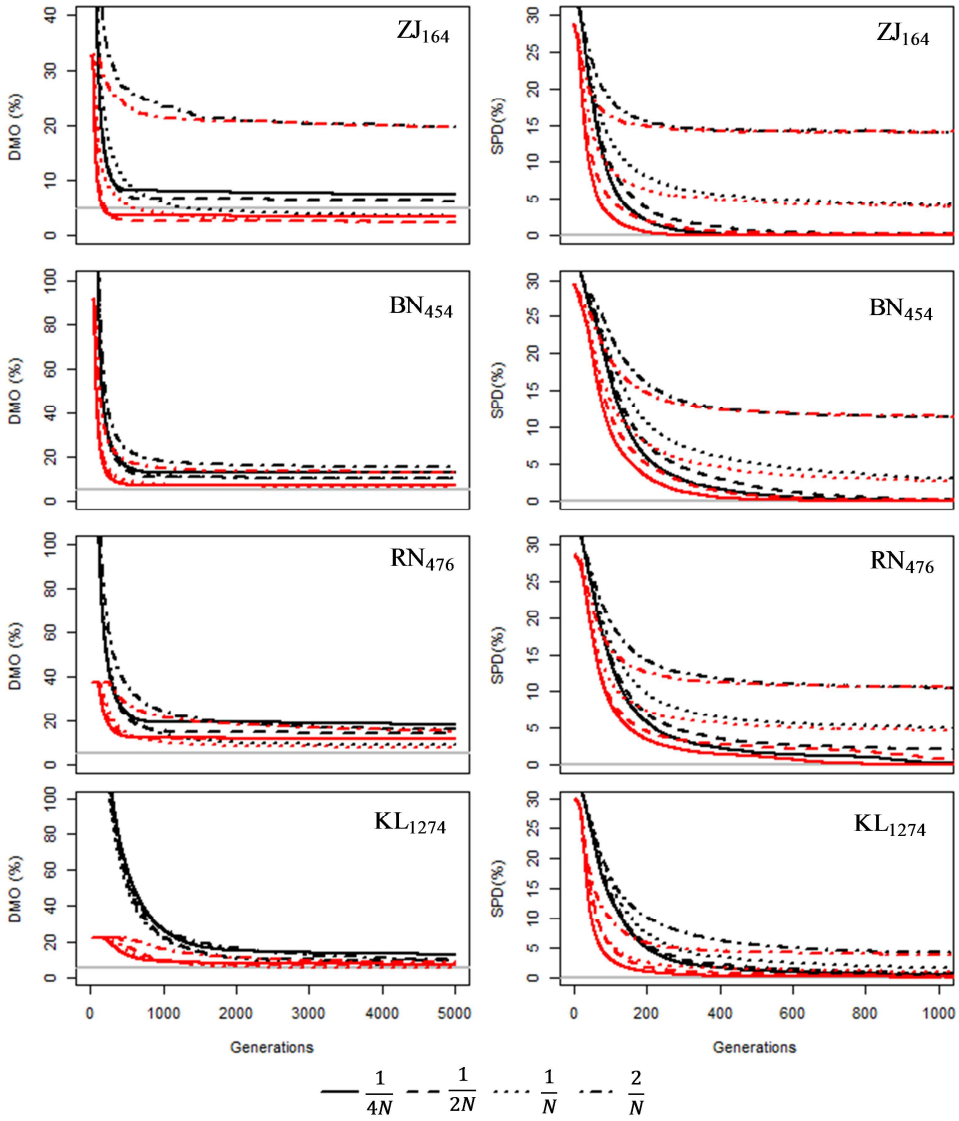

Figure 3.4 Results for GAs for which initial solutions are obtained using the RS method (black lines), and the PHSM (red lines). Different line types represent different mutation probabilities P_m. Left panel: Deviation of the mean cost from the best known solution (DMO%), with the horizontal grey lines showing 5% deviation. Right panel: Standardized average population diversity SPD (%) with the horizontal grey lines indicating complete convergence.

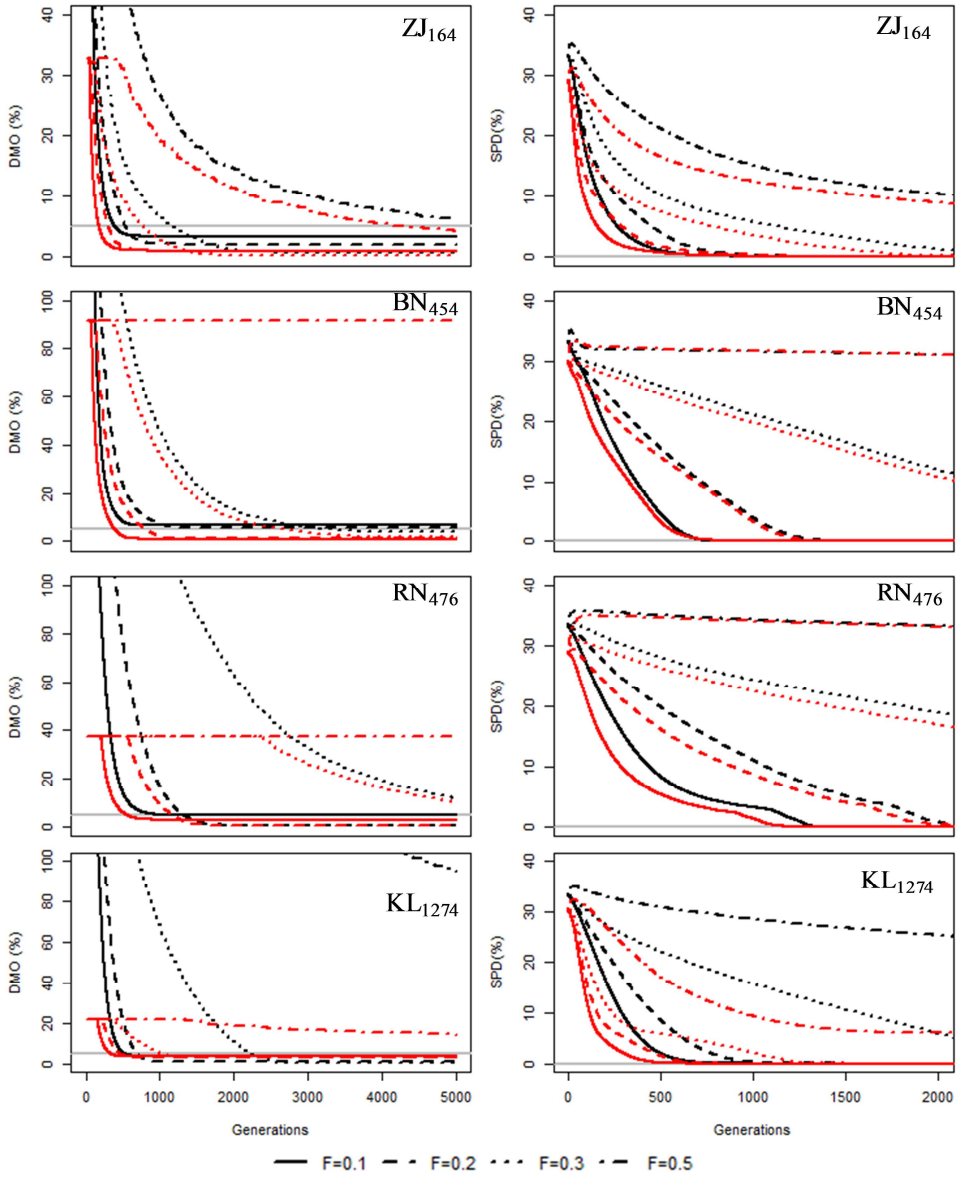

Figure 3.5 Results for DEs for which initial solutions are obtained using the RS method (black lines), and the PHSM (red lines). Different line types represent different mutation weighting factors *F*. Left panel: Deviation of the mean cost to the best known solution (DMO%), with the horizontal grey lines showing 5% deviation to the best-known solution. Right panel: Standardized average population diversity SPD (%) with the horizontal grey lines indicating complete convergence.

However, the faster convergence rate associated with the use of the PHSM does not always result in the ability to find near-optimal solutions more quickly. This is because, for certain combinations of algorithm type and parameterization,

near-optimal solutions cannot be identified at all, regardless of the initialization method used (e.g. none of the GA parameterizations investigated are able to find near-optimal solutions for the BN_{454} and RN_{476} networks, Figure 3.4, left panel). In other words, while the use of the PHSM is able to "shift down" the convergence curve for a particular algorithm and associated parameterization during the early stages of searching, the overall shape and location of this curve is a function of algorithm searching behavior.

In terms of searching behavior, the results suggest that EAs with relatively lower mutation rates are able to find or approach near-optimal solutions more quickly, although this may not necessarily result in the best final solution. For example, for the KL_{1274} problem, the DE algorithm that was initialized using random sampling (RS) found lower-cost solutions with F=0.3 than with F=0.1 after 2,400 generations, but the latter located near-optimal solutions (DMO% < 5%) after only 600 generations, which is appreciably less than the 2,200 generations required when F=0.3 is used.

The difference in observed convergence rates for the different algorithms and their parameterizations can be explained by examining the properties of the fitness functions for the four case studies, as well as the run-time behavioral statistics of the different algorithms and parameterizations as shown below.

3.4.2 Relationship between observed performance and problem statistics

The fact that parameters with reduced explorative and increased exploitative behavior are able to find near-optimal solutions more quickly and consistently is a function of the global structure of the fitness functions for the case studies investigated. As can be seen from Figure 3.6(a), all four case studies have a very similar global structure, with a correlation length $R_l \approx 0.5$ and $R_s = R_p$. This indicates that approximately 50% of the search space for each case study forms a "big bowl" shape that is positively correlated as illustrated in Figure 3.7. Consequently, little exploration is needed to

identify near-optimal regions, providing an explanation for why EAs with smaller mutation probabilities exhibit better performance for all case studies. In contrast, higher levels of exploration (high values of mutation) increase the time and effort taken to find promising regions in the solution space, resulting in a slower convergence to near-optimal solutions (Figures 3.4 and 3.5, left panel).

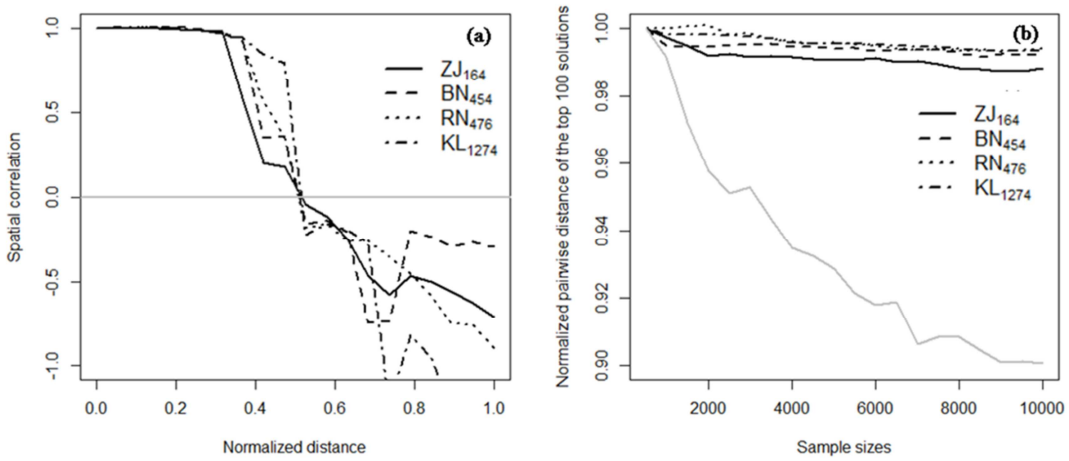

Figure 3.6. (a) Fitness function statistics (spatial correlation) of the four WDS problems considered; (b) Change in normalized pairwise distance of the top *m*=100 solutions with increase in sample size *N*. The grey line represents the values for the benchmark NYTP.

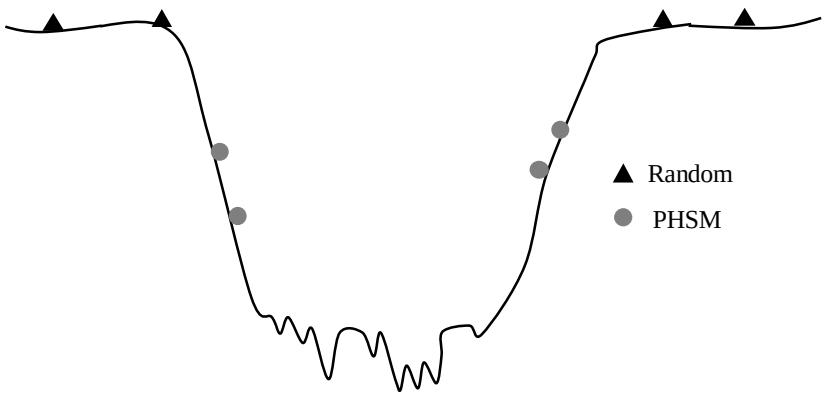

Figure 3.7. Stylized representation of the cross-section of the fitness function of the WDS design problems considered.

Although EAs with lower mutation rates converge more rapidly during the early phases of searching, their objective function values tend to stagnate during the later stages of the search. This can be explained by examining the values of the normalized dispersion metrics for each case study. As can be seen from Figure 3.6(b), for the four WDSs considered, the mean of the pairwise distance of the selected top 100 solutions decreases slowly or remains approximately constant. This is in contrast to the values for the benchmark NYTP, for which there is a rapid reduction in the value of the normalized metric, as was the case for the hydrological model calibration study considered by Arsenault et al. (2014). These results suggest that the microstructure of the promising regions (i.e. at the base of the big bowl shape) is very complex and rugged for the case studies considered (see Figure 3.7), whereas the opposite is the case for the NYTP. For the rugged microstructure of the four larger WDSs, EAs with relatively low mutation rates do not have sufficient exploration power during the later searching stages to enable them to find better solutions, resulting in stagnant performance, as observed from Figures 3.4 and 3.5 (left panel).

The fitness function characteristics discussed above can also be used to explain why the convergence plots of optimization runs initialized with the PHSM are always below those of the corresponding optimization runs initialized randomly, but generally follow the same shape (Figures 3.4 and 3.5, left panel). Due to the "big-bowl" macrostructure of the fitness functions of all case studies, the better initial solutions identified with the aid of the PHSM are likely result in searches that commence "part-way" down the "big bowl" (grey dots in Figure 3.7), shifting the convergence curve "down" during the initial stages of the search. In contrast, the randomly generated solutions (black triangles in Figure 3.7) are generally scatted in regions at the "top" of the "big bowl".

3.4.3 Relationship between observed performance and population diversity

While the relative ability of EAs with different levels of mutation to identify near-optimal solutions can be explained by the properties of the fitness function, as discussed above, the absolute performance of the algorithms with the different parameterizations is somewhat more difficult to explain. The same applies to the relative performance of the GA and DE. However, additional insight into the speed with which different algorithms and parametrizations are able to identify near-optimal solutions can be obtained by examining the relationship between population diversity (Figures 3.4 and 3.5, right panel) and searching quality (Figures 3.4 and 3.5, left panel). As can be seen, better optimization performance generally corresponds to a faster reduction in population diversity, which makes sense, given that the macrostructure of the fitness function of all four case studies has a “big bowl” shape (see Figure 3.7). The strong relationship between solution quality and population diversity is confirmed by the generally high values (between 0.8 and 1.0) of the dynamic correlation between DMO% and SPD% as shown in Figure 3.8, irrespective of EA and parameterization. This highlights that the speed with which near-optimal solutions can be identified is a property of population diversity, which is a property that transcends algorithm type and parameterization. The results suggest that different algorithms and parameterizations could well be merely means by which different population diversities are achieved and that it is likely that similar searching behavior can be achieved by different algorithms and parameterizations. In other words, while the results in the left panel of Figures 3.4 and 3.5 would suggest that the DE outperforms the GA, this is might not be an intrinsic property of these algorithms, provided different algorithm parameterizations can produce the desired population diversities.

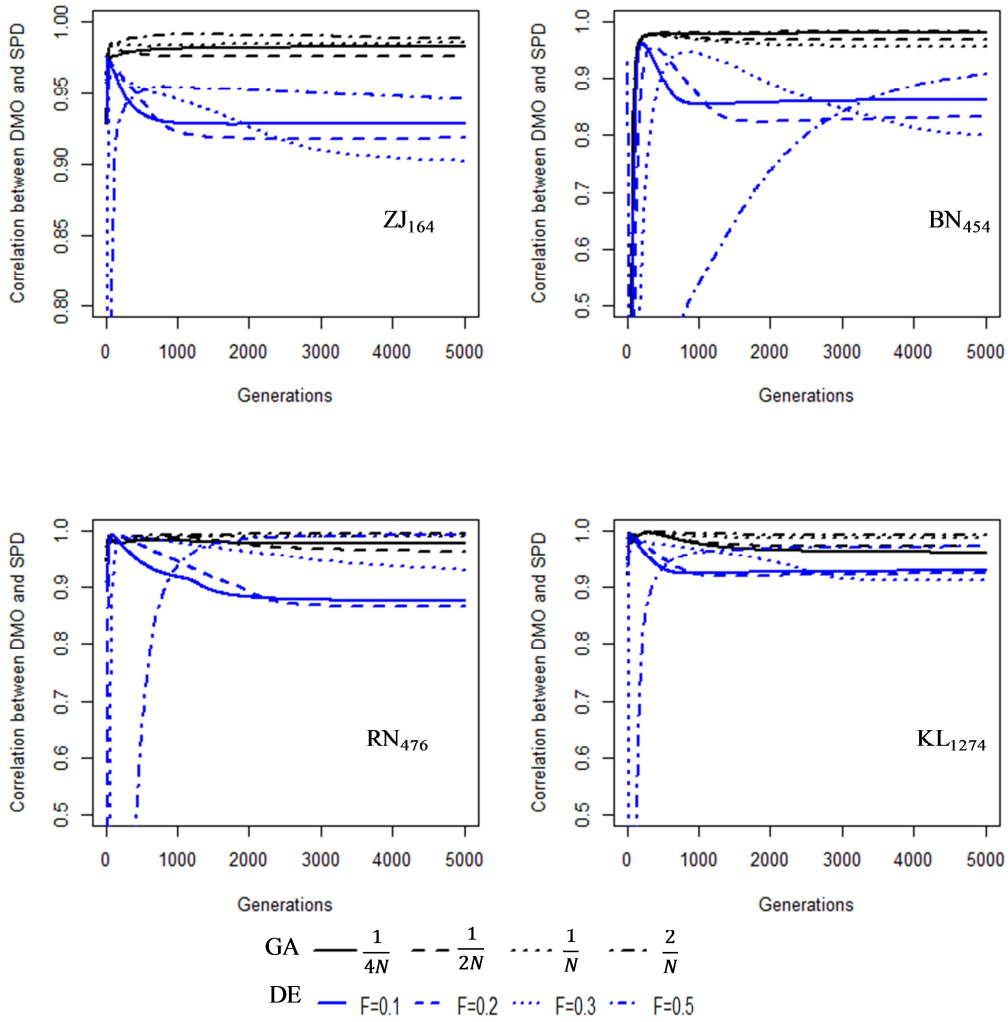

Figure 3.8. Results of the dynamic correlations between DMO% and SPD%. The correlation at generation *G* was estimated using DMO%[1:*G*] and SPD%[1:*G*]. The results shown are for random initialization. Similar results are obtained for initialization with the PHSM.

3.4.4 Summary

Overall, the results suggest that both algorithm initialization and searching mechanisms can significantly affect EA convergence rates, and hence the speed with which they can identify near-optimal solutions for large problems. For the WDS case studies considered, which all have fitness functions with a "big bowl" macrostructure, population diversity, as controlled by EA type and parameterization, has the biggest impact on the shape and location of the convergence plot, with algorithm behavior

favoring exploitation resulting in the ability to identify near-optimal solutions, and to do this more quickly. The use of the PHSM for algorithm initialization enables better starting solutions to be identified, thereby enabling near-optimal solutions to be found more quickly. However, this is conditional on the selection of a combination of algorithm and parameterization that results in a rapid reduction in population diversity.

3.5 Conclusions

Evolutionary algorithms (EAs) have been used widely to optimize the design and operation of water distribution systems (WDSs) over the past four decades. Starting from relatively simple benchmarking problems, there has been a move towards applying EAs to real-world, and large WDS design problems (Maier et al. 2014). One of the challenges associated with the application of EAs to large problems is their relative computational inefficiency, which make them difficult to apply to real-world problems. In recognition of this, there is growing interest in finding near-optimal solutions of large optimization problems within manageable computational budgets, instead of necessarily seeking the global optimum.

One way to enable near-optimal solutions to be determined more quickly is to seed the initial solutions of EAs within promising regions of the solution space. This can be achieved by generating initial solutions with the aid of engineering experience or domain knowledge, as has been done in a number of previous studies, such as Kang and Lansey (2012) and Bi et al. (2015). While these initialization methods have often been reported to exhibit better performance than random sampling, their performance as a function of different EAs and EA parameterizations has not yet been investigated. Furthermore, there is a lack of understanding of the relative impact of different starting positions and searching mechanisms on convergence rate in the context of finding near-optimal solutions for problems with real-life complexity with limited computational budgets.

The present study aims to address the above issues by investigating the impact of starting positions and searching mechanisms on the rate with which EAs converge to near-optimal solutions. Two initialization methods are considered, namely: random sampling (RS) and the Prescreened Heuristic Sampling Method (PHSM, Bi et al. 2015). Different searching mechanisms are represented by two EAs, including genetic algorithms and differential evolution algorithms, with different parameterizations. Four large WDS design problems, for which the number of decision variables ranges from 164 to 1,274, are considered as case studies. To gain a better understanding of the relative performance of different algorithm initialization methods and searching mechanism, the fitness function characteristics of the case studies and the run-time behavioral statistics of the different algorithms and their parameterizations are assessed.

The results of the present study and their implications can be summarized as follows.

(i) Both starting position and searching mechanism significantly affect the capacity of EAs to efficiently identify near-optimal solutions for large WDS design problems, with the latter exhibiting relatively more noticeable impacts.

(ii) Strong correlations are observed between improvements in objective cost function and reduction in population diversity during the run-time behavior analysis, for each type of EA and parameterization. This indicates that the convergence properties in the decision space heavily affect the searching quality in the objective space. Such an observation sheds new light on the causes of performance differences between different algorithms, parameterizations and starting positions, making it possible to modify an algorithm's performance through manipulating its population diversity.

(iii) The performance variation between different initialization methods and searching mechanisms (algorithms and parameterizations) can be related to the properties of the fitness function of the WDS design problems considered. The results show that the fitness functions for the case studies considered are likely to consist of

a single “big bowl” structure with a rugged base that is likely to have many local optima. This can explain the greater utility of the PHSM over the RS method, as the initial solutions obtained using the former enable the search to commence part way down the “side” of the “big bowl”. This is supported by the high quality of the starting positions obtained using the PHSM for the case studies considered. Based on the observed fitness function characteristics of the case studies considered, the use of EAs with reduced explorative capability (lower mutation rates) is expected to be more effective at being able to converge to near-optimal solutions more quickly.

In closing, the results of this study indicate that the use of EA initialization methods that are based on domain knowledge, such as the PHSM, in combination with EAs and their parametrizations that enable population diversity to be reduced rapidly, has the potential to enable near-optimal solutions to WDS optimization problems of real-world complexity to be obtained with significantly reduced computational budgets.

Chapter 4. Journal Paper 3- Use of domain knowledge to increase the convergence rate of evolutionary algorithms for optimizing the cost and resilience of water distribution systems

Statement of Authorship

Title of Paper	Use of domain knowledge to increase the convergence rate of evolutionary algorithms for optimizing the cost and resilience of water distribution systems
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Publication Style
Publication Details	Bi, W., Dandy, G. C., and Maier, H. R. (2015). "Use of domain knowledge to increase the convergence rate of evolutionary algorithms for optimizing the cost and resilience of water distribution systems." *Submitted to Journal of Water Resources Planning and Management, September 2015.*

Principal Author

Name of Principal Author (Candidate)	Weiwei Bi		
Contribution to the Paper	Develop the approach, perform the simulation study and prepare the manuscript		
Overall percentage (%)	50%		
Signature		Date	

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	Graeme C. Dandy		
Contribution to the Paper	Research supervision and review of manuscript.		
Signature	25%	Date	

Name of Co-Author	Holger R. Maier		
Contribution to the Paper	Research supervision and review of manuscript.		
Signature	25%	Date	

Abstract: Evolutionary algorithms (EAs) have been used extensively for the optimization of water distribution systems (WDSs) over the last two decades. However, computational efficiency can be a problem, especially when EAs are applied to complex problems that have multiple competing objectives. In order to address this issue, there has been a move towards developing EAs that identify near-optimal solutions within acceptable computational budgets, rather than necessarily identifying globally optimal solutions. This paper contributes to this work by developing and testing a method for identifying high quality initial populations for multi-objective EAs (MOEAs) for WDS design problems aimed at minimizing cost and maximizing network resilience. This is achieved by considering the relationship between pipe size and distance to the source(s) of water, as well as the relationship between flow velocities and network resilience. The benefit of using the proposed approach compared with randomly generating initial populations in relation to finding near-optimal solutions more efficiently is tested on five WDS optimization case studies of varying complexity with two different MOEAs (NSGA-II and Borg). The results indicate that there are considerable benefits in using the proposed initialization method in terms of being able to identify near-optimal solutions more quickly. These benefits are independent of MOEA type and are more pronounced for larger problems and smaller computational budgets.

CE Database subject headings: Multiobjective evolutionary algorithm; optimization; initialization method; water distribution system; near-optimal fronts

Author Keywords: Multiobjective evolutionary algorithm; optimization; initialization method; water distribution system; near-optimal fronts.

4.1 Introduction

Evolutionary algorithms have been used extensively and successfully for the optimization of water distribution systems (WDSs) over the last 20 years (Nicklow et al., 2010; Maier et al., 2014). However, as demonstrated in the Battle of the Water Networks II (Marchi et al., 2014), it is extremely difficult to find globally optimal

Pareto fronts for large WDS optimization problems with more than one objective. As a result, heuristic information and domain knowledge are commonly used to either reduce the size of the search space or to identify promising regions in the solution space from which to commence the search (Marchi et al., 2014). Both of these approaches are designed to ensure that near-optimal solutions are obtained within reasonable computational budgets, rather than to necessarily attempt to find the globally optimal solution(s) (e.g. Tolson and Shoemaker, 2007; Gibbs et al., 2008, 2011, 2015; Tolson et al., 2009). The use of such approaches is of particular importance when evolutionary algorithms (EAs) are applied to real-world problems, as they often require the use of computationally intensive simulation models for objective function and/or constraint evaluation (Maier et al., 2014). Consequently, there is a need to develop approaches that enable near-optimal solutions to be found for the optimization of WDSs within computational budgets that are acceptable from a practical perspective. This is important for the successful application of EAs in both the research domain and in practice, thereby enabling their full potential to be realized (Maier et al., 2014).

The use of domain knowledge is an important approach to achieving the above goal, as demonstrated in a number of engineering problem domains, including mechanical design (Sapuan, 2011), aircraft wing design (Ong and Keane, 2002) and reservoir system optimization (Li et al., 2014). In the area of the optimization of WDSs, Keedwell and Khu (2006) considered the fact that the diameters of the pipes that are connected to demand nodes with a pressure deficit (or surplus) can be increased (or decreased) to increase pressure (or reduce cost) in the determination of the initial population of EAs. Subsequently, Zheng et al. (2011) incorporated knowledge that the most cost-effective solution for a looped WDS with a single demand case is a tree-branched topology into the initialization of EAs, and Kang and Lansey (2012) developed an initialization method that uses engineering experience about optimal flow velocities in WDSs. More recently, Bi et al. (2015) proposed an initialization approach that not only considers optimal flow velocities in pipes, as in Kang and

Lansey (2012), but also allows for the fact that pipe diameters generally reduce with distance from the source (Walski, 2001). As part of the approach, initial EA populations are obtained by sampling in the vicinity of the solutions generated based on the above principles in order to avoid premature convergence to local optima in solution space (Bi et al., 2015).

The initialization methods outlined above have been reported to significantly outperform the random initialization approach in terms of their ability to identify near-optimal solutions at reduced computational cost. However, they are only applicable to single-objective WDS optimization problems, or at least have not been applied to multi-objective problems to date. In contrast, most real-world problems have more than one competing objective and, in recent years, increasing effort in the optimization of WDSs has been devoted to multi-objective optimization problems (Nicklow et al., 2010), with the minimization of cost and the maximization of various network reliability measures receiving the most attention (Tolson et al., 2004; Prasad and Park, 2014; Raad et al. 2010; Wu et al., 2013; Zheng et al., 2014; Wang et al., 2015). While a number of previous studies have been successful in improving the computational efficiency of such problems (e.g. Zheng et al., 2011; Creaco and Franchini, 2012), there remains a need to develop a formal approach that enables domain knowledge to be used to identify good initial populations for multi-objective EAs (MOEAs) applied to WDS design problems.

In order to address this shortcoming, an approach that uses WDS domain knowledge to identify good initial populations for EAs that minimize cost and maximize network resilience is introduced in this paper. The approach extends the Prescreened Heuristic Sampling Method (PHSM) of Bi et al. (2015), which only considers cost minimization as an objective, to a multi-objective problem. The proposed Multi-Objective Prescreened Heuristic Sampling Method (MOPHSM) is tested on a number of benchmark WDS design problems, ranging in size from 34 to 1274 pipes, and the performance of the proposed MOPHSM is compared with that of randomly

initializing the population of MOEAs, which is most commonly used in literature. In order to assess the utility of the proposed MOPHSM, three run-time performance metrics are used. These are the hypervolume index (Hadka and Reed, 2015), the generational distance (Kollat et al., 2008) and the extent of front. NSGA-II (Deb et al.,2002) and Borg (Hadka and Reed, 2012) are used as MOEAs, as NSGA-II, or algorithms based on it, have been used extensively for the multi-objective optimization of WDSs (e.g. Wu et al., 2010; Stokes et al., 2015a, b; Wang et al., 2015) and Borg is a more recently developed algorithm that is being applied increasingly to a range of problems, including WDS optimization (e.g. Stokes et al., 2015c; Wang et al., 2015).

The remainder of this paper is organized as follows. The optimization problem is given in the next section followed by the proposed MOPHSM. The methodology used and computational experiments performed for testing the approach then follow. The results are presented and discussed in the following section, before a summary and conclusions are provided.

4.2 The optimization problem

The WDS design problem typically involves the selection of pipe diameter sizes for a predefined network topology, in order to meet selected design objectives, and satisfy hydraulic and design constraints. Following Wang et al. (2015), the minimization of pipe cost and the maximization of network resilience (a surrogate measure of network reliability) were taken as the two objectives, which can be described as follows:

Minimize the cost:
$$F_c = a\sum_{i=1}^{n} D_i^b L_i \tag{4.1}$$

Maximize the network resilience:
$$I_n = \frac{\sum_{j=1}^{m} U_j DM_j (H_j - H_j^*)}{\sum_{r=1}^{R} q_r H_r^R - \sum_{j=1}^{m} DM_j (H_j^* + z_j)} \tag{4.2}$$

Subject to:

Indicator of diameter uniformity

$$U_j = \frac{\sum_{p \in M_j} D_p}{|M_j| \times \max_{p \in M_j}\{D_p\}} \tag{4.3}$$

Pressure and velocity constraints

$$H_j^* \le H_j \le H_j^{\#}$$
$$V_i^* \le V_i \le V_i^{\#} \tag{4.4}$$

Hydraulic constraints:

$$\mathbf{H} = f(\mathbf{D}, \mathbf{DM}) \tag{4.5}$$

Diameter choices:

$$D_i \in A \quad i = 1, \ldots, n \tag{4.6}$$

where F_c is the total network cost, including pipe material and construction costs; D=$[D_1, \ldots, D_n]^T$, where D_i is the diameter of pipe i=1,……n; L_i is the length of pipe i; a, b are specified cost function coefficients; n is the total number of pipes in the network; $\mathbf{H} = [H_1 \cdots H_m]^T$ is the vector of pressure heads at network nodes; $\mathbf{DM} = [DM_1 \cdots DM_m]^T$ is the predefined vector of nodal demands; m is the total number of demand nodes in the network; H_j and DM_j are the pressure head and the nodal demands for node j=1,……m, respectively; z_j is the elevation of node j; H_j^* and $H_j^{\#}$ are the design minimum and maximum allowable pressure head at node j, respectively; q_r and H_r^R are total demands and total heads (pressure head plus the elevation head) provided by the supply source (reservoirs or tanks) r=1,……R, respectively.

D_p is the diameter belonging to set M_j, which represents all pipes connected to node j; $|M_j|$ is the cardinality of M_j; V_i^* and $V_i^{\#}$ are the design minimum and maximum allowable flow velocity for pipe i, respectively; and A is the set of commercially available pipe diameters.

It is noted that the network resilience defined in Prasad and Park (2014) included pumps within the WDS. These are not considered in this paper for consistency with

Wang et al. (2015). U_j in Equation (4.3) is an indicator of diameter uniformity for pipes that immediately connect to node j(M_j), with a larger value representing a higher reliability of the network loop, since the diameter variations between these pipes are overall lower (U_j=1 when all pipe diameters are identical) (Prasad and Park, 2014).

4.3 Proposed multi-objective prescreened heuristic sampling method

The proposed MOPHSM to identifying good initial populations of MOEAs used to minimize the cost and maximize the resilience of WDSs is based on a basic understanding of the relationship between some of the characteristics of optimal WDSs and the two objectives considered. The proposed MOPHSM consists of three steps, details of which are given below.

4.3.1 Step 1: Identify initial solutions using domain knowledge related to cost

In actual WDSs, the diameters of upstream pipes are generally equal to or larger than those further downstream (Walski, 2001). In this step of the MOPHSM, this knowledge is used to assign initial pipe diameters in accordance with their distances to the supply sources, as was done by Bi et al. (2015) as part of the single-objective PHSM. A brief summary of the main steps for achieving this is given below, with full details provided in Bi et al. (2015).

1) Find the shortest distance tree-network for the WDS design problem being solved using the Dijkstra algorithm (Zheng et al., 2011).
2) Obtain the largest value of the shortest distance *L*in the tree network, i.e. the largest distance that the supply sources have to deliver the demands.
3) Divide the WDS network into *P* specific areas with the shortest distance to the source node interval of *L/P*, where *P* is the number of available pipe diameters for the design.

4) Assign pipes in each area a different diameter, with the largest diameter assigned to the pipes in the area nearest to the source and the smallest diameter to the pipes in the area furthest from the source (reservoir).All pipes in a single area are assigned the same diameter.

The main benefit of the above approach is that it is able to identify a much greater proportion of lower-cost solutions that are near the boundary of feasibility in terms of being able to satisfy pressure constraints than can be achieved with random initialization. Consequently, EAs that are seeded with these solutions can commence their search in more promising regions of the solution space, leading to improved efficiency in terms of being able to identify near-optimal solutions, as demonstrated in Bi et al. (2015).

4.3.2 Step 2: Identify an initial front by adjusting the solutions obtained in Step 1 based on domain knowledge related to network resilience

While step 1 results in the identification of good starting positions in solution space in relation to the cost objective, additional considerations are required in order to identify good initial solutions in relation to both the cost and network resilience objectives. This requires a good understanding of the factors that have an impact on network resilience. As shown in Equations (4.2) and (4.3), the two factors that affect network resilience are pressure head H_j and diameter uniformity U_j. For a WDS design problem, the nodal demands are typically fixed, and hence a network design solution with a relatively higher pressure head at each node would have overall larger diameters and, accordingly, relatively lower flow velocities V. With the aid of this knowledge, it is possible to adjust the initial solutions from Step 1 to produce an extended front of initial solutions with a range of cost and network resilience values by using the methodology illustrated in Figure 4.1.

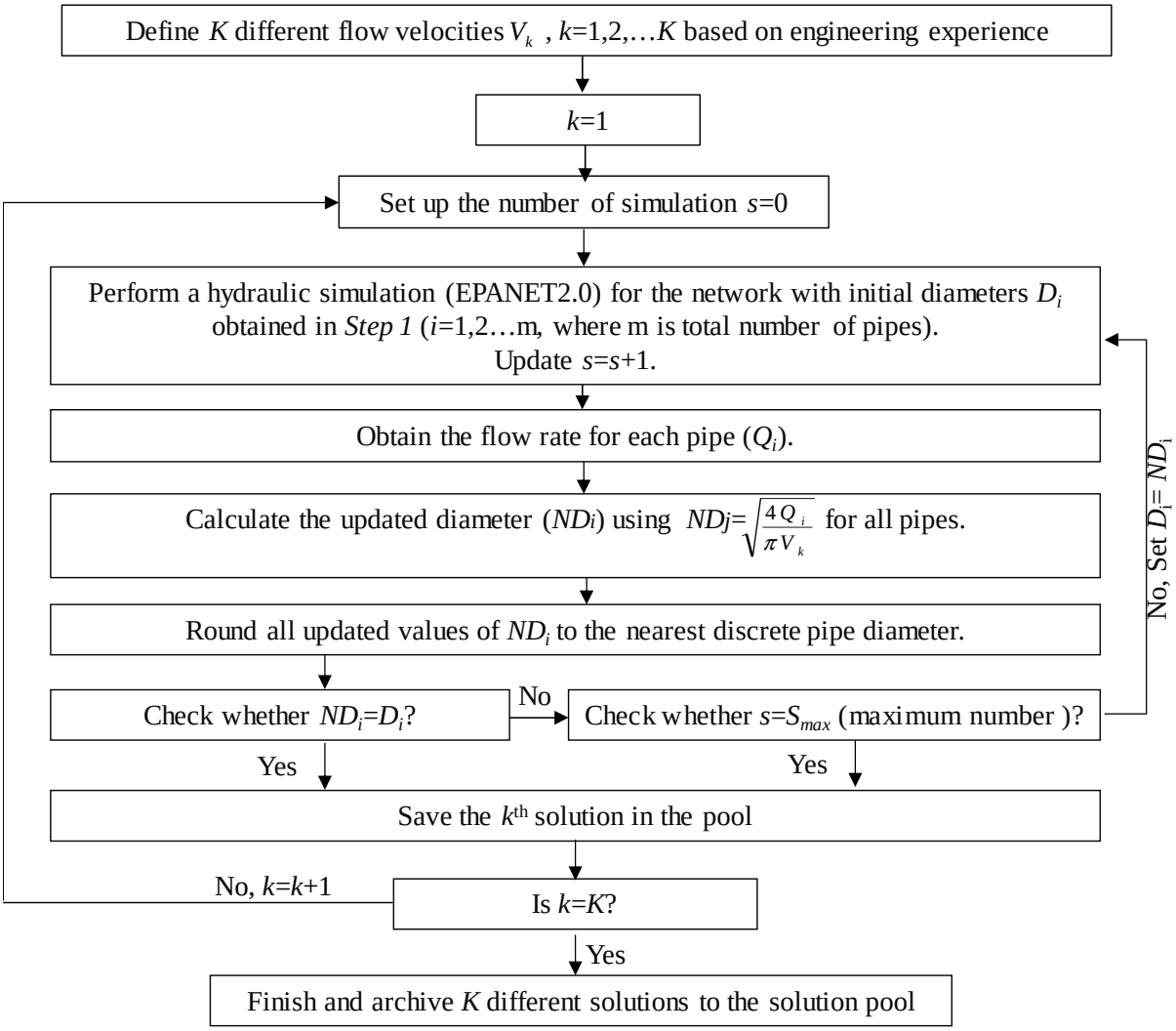

Figure 4.1 Flowchart of the proposed methodology for adjusting the pipe diameters obtained from step 1 based on flow velocity in order to identify good initial solutions in relation to both cost and network resilience

As shown in Figure 4.1, as part of this proposed methodology, the diameters obtained in Step 1 are refined to achieve flow velocities that are close to a particular threshold V_k (k=1,2…,K) in all pipes. The different values of V_k can be determined based on engineering experience, as well as the type of water network being designed (e.g. potable supply network, irrigation network). For relatively smaller values of V_k, the overall diameters of the design solution are larger, resulting in an overall larger set of pressure heads H_j in Equation (4.1),

and vice versa. Since all pipes are assumed to have an identical expected value of V_k, the diameters for the pipes with similar flows are more likely to be of overall similar diameter, leading to a relatively large value of diameter uniformity U_j (Equation 4.2) and hence a greater value of network resilience. As such, an approximate front that accounts for domain knowledge related to both cost (Step 1) and network resilience (Step 2) is formed by the solutions archived in the solution pool, as shown in Figure 4.1.

4.3.3 Step 3: Generate initial MOEA population by sampling in the vicinity of the initial front identified in Step 2

In order to ensure sufficient diversity in the initial MOEA population, density functions are developed around each of the solutions identified in Step 2, from which samples can be drawn to form the initial population, as was carried out by Bi et al. (2015) for the single objective case. The proposed density function takes the following form (Bi et al., 2015)

$$f(D_b) = \frac{1}{1 + a \parallel D_b - D_c \parallel} \qquad b = 1,....,P \tag{4.7}$$

where a is constant; $\parallel D_b - D_c \parallel$ is the distance between D_b and D_c (the diameter for a pipe in the approximate optimal solution determined in Step 2) in terms of integer coding (for details see Bi et al. (2015)); and P is total number of available pipe diameters.

Figure 4.2 illustrates the distributions of the initial samples for different values of a for a pipe with D=400 mm (grey line) obtained in Step 2, with P={100, 200, 300, 400, 500, 600, 700}. As shown in this figure, the diameters closer to the heuristic pipe diameter obtained in Step 2 have a higher probability of being selected, as they are more likely to be the optimal diameter compared with other diameter options for the given flow velocity value. It can also be seen that a larger value of a will produce samples that are closer overall to the initial solution obtained in Step 2, and vice versa.

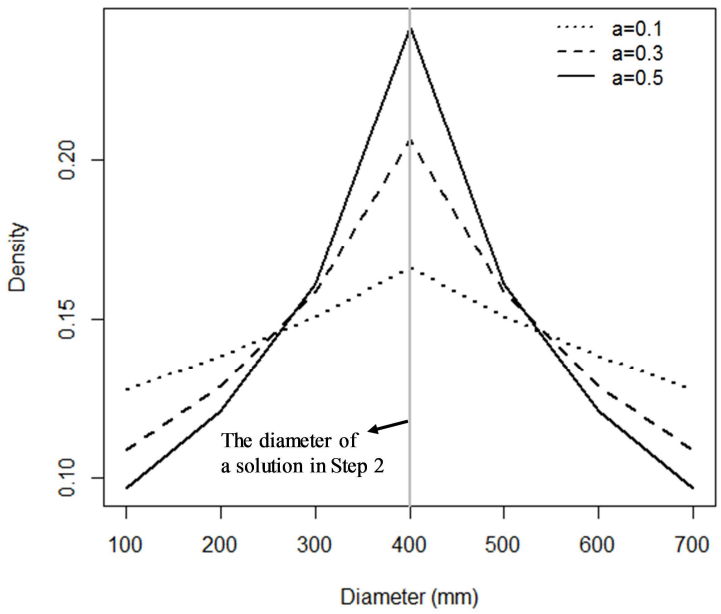

Figure 4.2 Distribution of samples for different values of *a*

For each solution on the initial front obtained in Step 2, a set of $NK=\lceil N/K \rceil$ samples is generated using the density function in Equation (4.7), where N is the population size of the MOEA, K is the total number of solutions in the initial front (Step 2), and $\lceil N/K \rceil$ is the ceiling function with a smallest integer not less than N/K. In this way, the required N initial solutions of the MOEA are generated.

4.4 Methodology

The methodology used to test the utility of the MOPHSM introduced in the previous section is summarized in Figure 4.3. As can be seen, two different MOEA initialization methods are considered, namely the proposed MOPHSM with the most commonly used random initialization as a benchmark. As mentioned previously, two different MOEAs, including NSGA-II and Borg, are applied to both initialization methods in order to ensure that the impact of the different initialization schemes is not algorithm specific. Both initialization methods and MOEAs are applied to five WDN design problems, for which the number of decision variables ranges from 34 to 1,274. To gain an improved understanding of the results in terms of the speed with which near-optimal fronts are identified, three run-time performance metrics are

analyzed. These are the hypervolume, the generational distance and the extent-of-front. Details of each step in Figure 4.3 are discussed in following sub-sections.

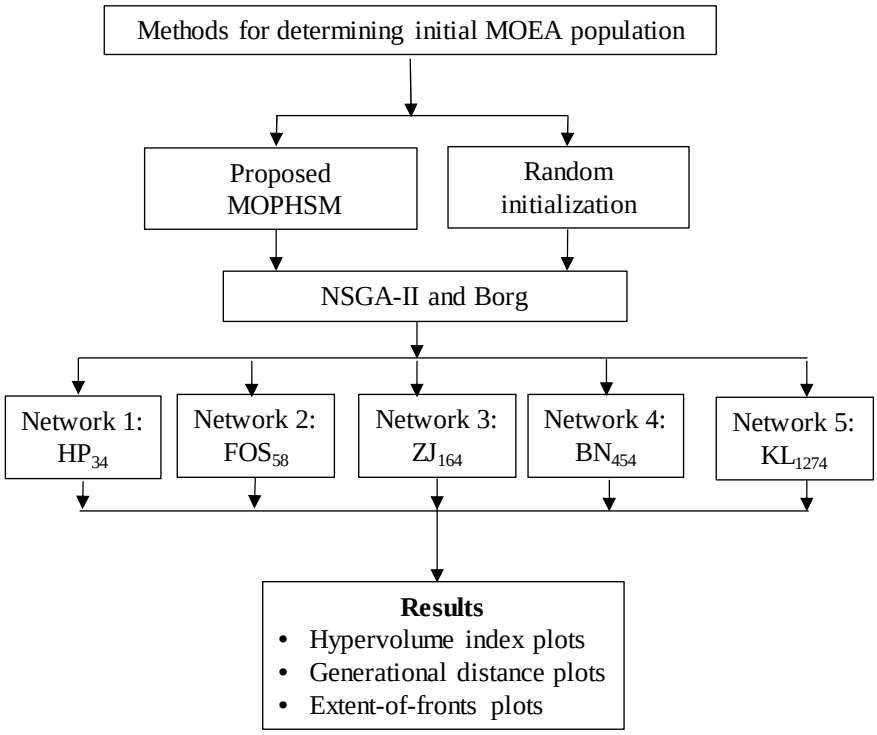

Figure 4.3 Flowchart of the assessment process. The subscript of each case study indicates the number of decision variables.

4.4.1 Methods for determining initial MOEA population

As shown in Figure 4.3, the performance of the proposed MOPHSM is compared with that of the random initialization method in terms of efficiently finding near-optimal fronts. For the MOPHASM, the *K* different flow velocities (see Figure 4.1) are determined within the specified range [0.1 m/s, 4 m/s], with an interval of Δv =0.1 m/s. These upper and lower bounds of velocity, as well as the velocity interval, are selected based on a preliminary analysis of the best known fronts of a number of the WDS benchmarking case studies outlined in Wang et al. (2015).

The most appropriate values of *a* for the density function used to generate the initial population (see Equation (4.7)) are determined using a trial-and-error process. As part of this process, the impact of different values of *a* (0.1, 0.3, 0.5 and 0.7 for each of the case studies) on the resulting optimal solutions is assessed by visual inspection. The results of this analysis show that values of $a \geq 0.5$ are more likely to result in premature convergence during the optimization run, whereas values of $a \leq 0.1$ do not exhibit a significant advantage relative to the random initialization approach. Hence, a value of $a = 0.3$ is used in this study.

4.4.2 Multiobjective evolutionary algorithms

As mentioned previously, NSGA-II (Deb et al., 2002) and Borg (Hadka and Reed, 2012) are selected as MOEAs. NSGA-II uses a fast non-dominated sorting strategy to rank solutions, which is followed by selecting population members of the next generation according to Pareto dominance and crowding distance. As outlined in Deb et al. (2002), a Simulated Binary Crossover (SBX) operator and a polynomial mutation approach are used to carry out crossover and mutation, respectively, and a constraint tournament selection strategy is used to handle infeasible solutions.

Borg is a unified optimization framework combining ε-dominance, ε-progress, randomized restart and auto-adaptive multioperator recombination, with details given in Hadka and Reed (2012). The operators used in Borg include SBX, Differential Evolution (DE), Patent-Centric Crossover (PCX), Unimodal Normal Distribution Crossover (UNDX), Simplex Crossover (SPX), and Uniform Mutation (UM). One of Borg's important features is its auto-adaptive multi-operation selection scheme, where a feedback loop is established in which operators that produce more successful offspring are rewarded by increasing their selection probabilities for generating new solutions for the next generation (Hadka and Reed, 2012). Another feature is the implementation of a restart strategy (adaptive population sizing) in order to avoid premature convergence.

It should be noted that both algorithms use the constraint tournament selection method (Deb, 2000) to handle infeasible solutions.

4.4.3 Case studies

As shown in Figure 4.3, five case studies of varying sizes are included, ranging from 34 to 1274 pipes. All of these case studies were considered as single objective problems in Bi et al. (2015) and the HP_{34}, FOS_{58} and BN_{454} problems were also considered in the multi-objective study by Wang et al. (2015). As stated previously, the minimization of network cost and the maximization of network resilience (a surrogate measure of network reliability) are taken as the two objectives in the present work, as was the case in the benchmarking study by Wang et al. (2015), in which further details about the objectives are provided.

4.4.4 Run-time performance metrics

As mentioned above, three performance metrics are used to assess the quality of the fronts that are generated during the optimization runs. The first metric is the *hypervolume index* (Zitzler and Thiele 1999), which calculates the hypervolume of the multi-dimensional region enclosed by a front and a reference point. This metric is able to represent overall performance, jointly measured by solution quality, solution diversity and the uniformity of the solutions on the front (Hadka and Reed 2012). In this study, the specific hypervolume performance metric used is the hypervolume index at generation G, denoted as $HI(G)$, which is defined as the ratio of the hypervolume at generation G relative to that of the best-known Pareto front PF^*. As such, $HI(G)$ lies within [0,1], with larger values representing a hypervolume that is closer to that of PF^*.

It should be noted that there is no accepted definition of what constitutes near-optimal fronts in multi-objective space. Hadka and Reed (2012) defined near-optimal fronts in terms of achievement of 90% of the hypervolumes of the best-found (Pareto) fronts, as the hypervolume is widely accepted as the best overall performance metric for multi-objective optimization problems. Consequently, a

similar measure is used in this study. However, in this study, near-optimal fronts are defined as those with hypervolumes that lie within 5% of the hypervolume of the best-known solution, i.e. HI(G) 0.95. This is done in order to align this study with that of Bi et al. (2015), in which, in a single-objective optimization context used to assess the performance of the PHSM, near-optimal solutions were defined as those that lie within 5% of the best-known solution.

The second metric is the *generational distance*, which is typically used to represent an MOEA's convergence status in objective space (Kollat and Reed, 2006). It is calculated as the mean of the Euclidean distance between each solution point on the approximate front and its nearest solution point on PF^*. Each dimension of the solution vectors on PF^* is normalized to [0,1] initially, followed by the normalization of each dimension of the approximate front using the data ranges from PF^*. As such, the value of the generational distance is normalized within the range [0, 1]. A lower value of generational distance indicates a better front, as it possesses an overall shorter distance to the Pareto front in objective space.

The third metric is the *extent of the front*, which is another important indicator for assessing an MOEA's searching quality in terms of explorative ability. Although $HI(G)$ can partly represent the extent of the front, solution quality and diversity also affect its value. Consequently, the extent of the front measure is used in this study to obtain a deeper understanding of this particular aspect of front quality. This is achieved by comparing the extent of the front at a particular generation G with that of the best-known front PF^*. More specifically, the extent of front value is equal to the maximum Euclidean distance between two solution points on the front at generation G divided by the equivalent distance for the best-known Pareto front PF^*. The normalization method described for computing the generational distance is also used to calculate the extent of front metric, and hence its value is within the range [0, 1].

It should be noted that the best-known Pareto front PF^* is needed in the calculation of all of the above performance metrics. For the HP_{34}, FOS_{58} and BN_{454} problems,

these fronts are taken from Wang et al. (2015), which were developed by comprehensive runs of multiple MOEAs. The PF^* of the ZJ_{164} and KL_{1274} problems are not available from the literature, and hence are obtained by non-dominated sorting (Deb et al., 2002) of the merged fronts from all of the results generated in this study. For the two objectives considered, a hypothetical solution with the maximum cost (i.e. all pipes are assigned the largest available diameter) and the minimum network resilience (i.e. I_n=0) is considered as the reference point (the worst solution) for the computation of $HI(G)$ for each WDS problem, which is consistent with Wang et al. (2015).

4.5 Computational experiments

For each case study, the default parameter values of NSGA-II and Borg are used (Wang et al. 2015). For both NSGA-II and Borg, these include a crossover probability of 0.9 (SBX for NSGA-II, and all other crossover operators for Borg) and a mutation rate of 1/*LN* (polynomial mutation), where *LN* is the number of decision variables, as shown in Table 1. The population sizes (*N*) for the HP_{34}, FOS_{58} and BN_{454} problems are taken from Wang et al. (2015), and for the remaining case studies they are taken from Bi et al. (2015), with values given in Table 4.1. It is noted that the population sizes given in Table 1 are the initial value for Borg applied to each case study, as its population size is dynamically increased as the search progresses (Hadka and Reed 2012). The maximum allowable number of generations for each case study is 2,500, which is consistent with those used in Wang et al. (2015).

Table 4.1 MOEA parameters used for each case study

Case studies	Crossover probability (SBX)	Mutation probability (polynomial mutation)[1]	Population size (*N*)	Maximum allowable of generations (*MG*)	Equivalent number of generations for identifying the initial fronts
HP_{34}	0.9	0.0294	240	2500	2.6
FOS_{58}	0.9	0.0172	400	2500	2.9
ZJ_{164}	0.9	0.0061	500	2500	4.5
BN_{454}	0.9	0.0022	1000	2500	5.7
KL_{1274}	0.9	0.0008	1000	2500	6.2

[1]Mutaton rate is 1/*LN*, where *LN* is the number of decision variables (the subscript number of the case studies).

It should be noted that additional computational effort is required to determine the initial front (Steps 1-2) in the proposed method (mainly Step 2), which has been converted to the equivalent number of generations for each case study using a Pentium PC (Inter R) at 3.0 GHz, as shown in Table 4.1. It can be seen that the proposed method is very efficient in producing the initial front for each case study, as its computational overhead is negligible compared with the total computational budgets allowed for the NSGA-II and Borg optimization runs. As stated previously, the number of solutions on the initial front is K=40, and hence the number of samples generated based on each of these solutions is $\lceil N/K \rceil$ using the density function in Equation (4.3). As such, the sample size for each solution on the initial front is 6, 10, 13, 25 and 25 for the HP_{34}, FOS_{58} ZJ_{164}, BN_{454} and KL_{1274} problems, respectively.

For each case study, all NSGA-II and Borg optimization runs with each of the two initialization methods are repeated ten times using different starting random number seeds, and the mean value of the run-time measure metrics over the ten runs is presented for discussion (Zecchin et al., 2012; Zheng et al., 2015). In addition, the approximate fronts from the two initialization methods at three different generations (*G*) are shown to enable a direct visual comparison of their performance.

Given that the ability to find near-optimal fronts (rather than the end-of-run front) is the focus of this study, performance metrics are shown up to a maximum of 500 generations, which is 20% of the total computational budget of 2,500 generations.

The approximate fronts at G=10, 100 and 500 are presented for the HP_{34}, FOS_{58} ZJ_{164}, BN_{454} problems. For the KL_{1274} case study, G=100, 300 and 500 are used, as there are many infeasible solutions at small numbers of generations for this large problem. The fronts at these three generation numbers are selected as they provide an indication of (i) the quality of the initial fronts, (ii) performance with very limited computational budgets (e.g. 100 generations), and (iii) the ability to identify near-optimal fronts with reasonable computational overheads (e.g. G=500), respectively.

4.6 Results and discussion

The approximate fronts obtained using the proposed MOPHSM (red '+') and the random initialization approach (black circles) are shown in Figures 4.4 (for NSGA-II) and 4.5 (for Borg), with grey triangles representing the best-known fronts. These results are from a typical run for the two initialization methods considered, and similar performances were observed for the other runs. It is noted that an archive was used in Borg to store non-dominated solutions obtained from the ε-dominance operator, with details given in Hadka and Reed (2012), meaning that the number of solutions in the archive increased over the generations. This is the reason why Borg produced fewer non-dominated solutions at the earlier generations relative to NSGA-II, as shown in Figures 4.4 and 4.5.

As can be seen, the fronts produced using the MOPHSM clearly dominate those generated using the random initialization approach, irrespective of which MOEA is used, with the advantages of the MOPHSM more noticeable for larger problems and smaller numbers of generations. For example, (i) for the BN problem with 454 decision variables, the costs of the solutions identified using the MOPHSM for both NSGA-II and Borg were approximately half of those obtained with the aid of the random method at G=10 for similar values of network resilience (Figures 4.4 and 4.5), and (ii) for the largest problem (KL_{1274}) with similar values of network resilience, the costs of solutions from the NSGA-II and Borg fronts seeded by the MOPHSM were

around 15% and 25% lower than those obtained using the random approach at G=100, respectively.

It is noted that, for NSGA-II, the MOPHSM's superior performance is not evenly distributed across the whole front, but is significantly more prominent in regions with relatively low costs. For solution regions with very high costs, both initialization methods exhibited comparable performance, as shown in Figure 4.4. This finding is consistent with the observations of Zheng et al. (2014), in that good starting positions for NSGA-II are more likely to show advantages in searching regions with relatively low costs. This is because such regions typically have more complex fitness functions, as they are often located at the boundary between feasible and infeasible areas. Interestingly, MOPHSM's better performance was observed across the entire front relative to the random method for Borg, as shown in Figure 4.5. This is most likely due to the differences in searching mechanisms between NSGA-II and Borg. For the Borg algorithm applied to the KL network, MOPHSM gave a front that was more limited in extent than the random method and this limitation applied up to 500 generations.

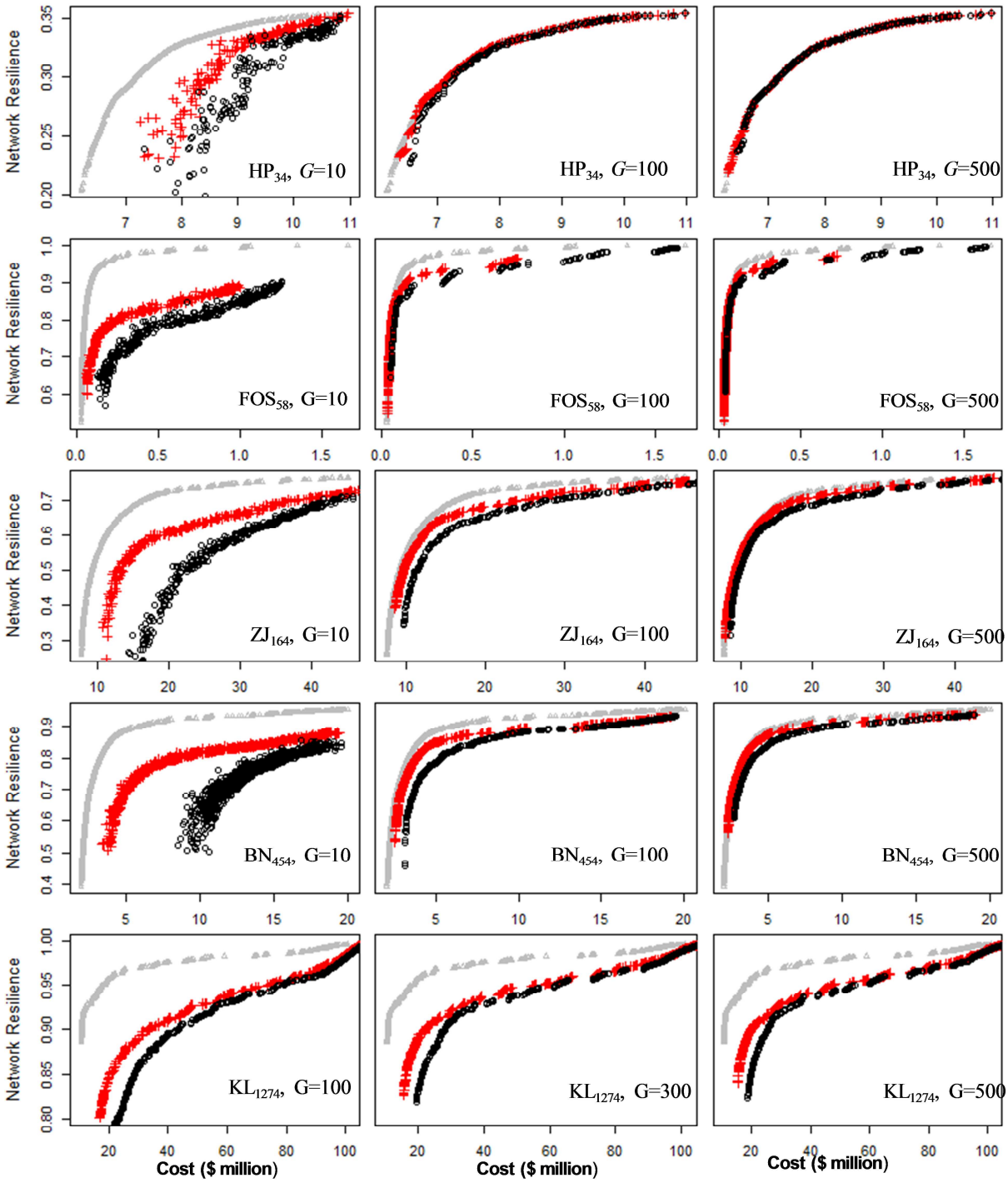

Figure 4.4 Approximate fronts of the proposed MOPHSM (red '+') and the random initialization approach (black circles) obtained using NSGA-II. The grey triangles are the best-known fronts.

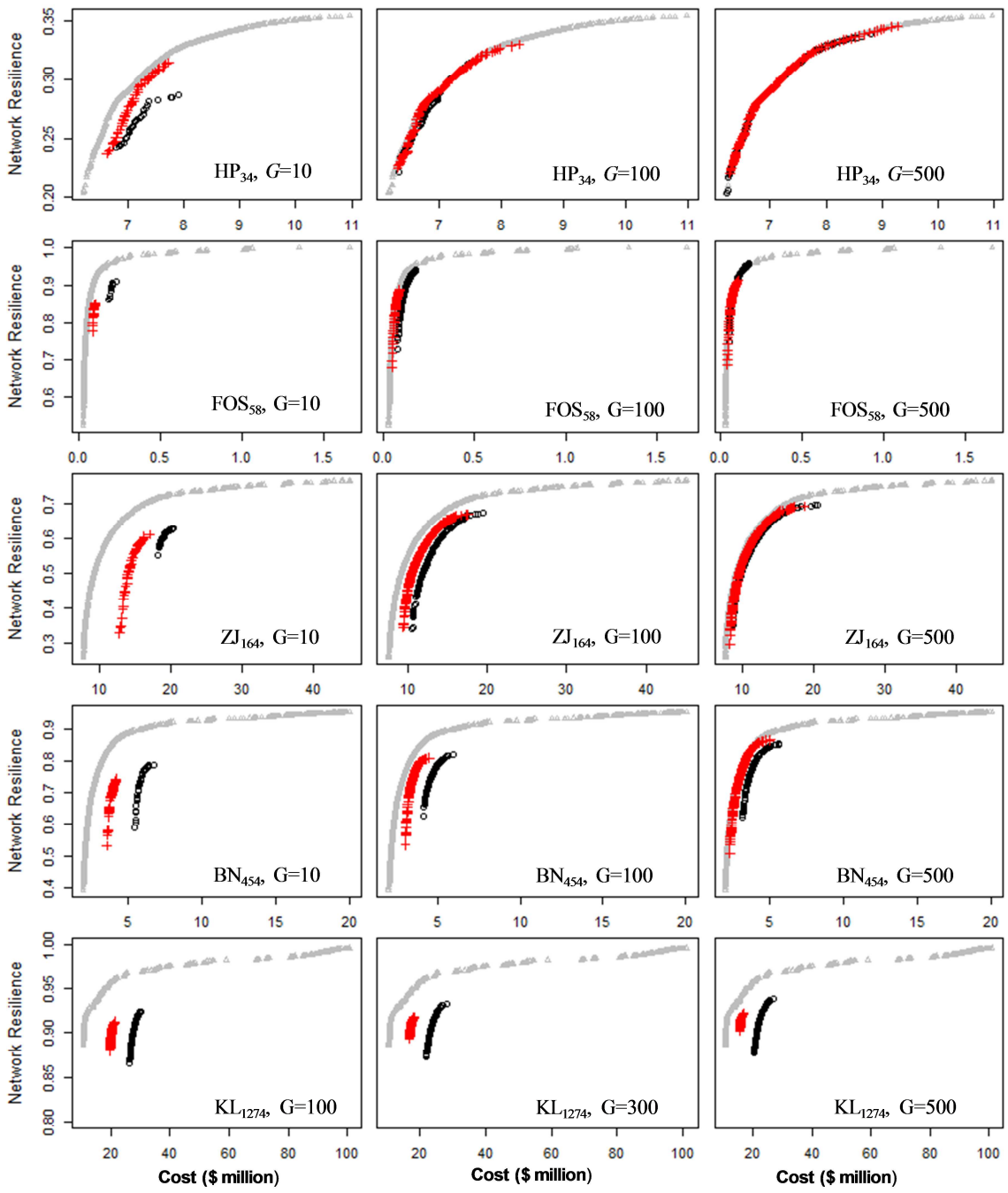

Figure 4.5 Approximate fronts of the proposed MOPHSM (red '+') and the random initialization approach (black circles) obtained using Borg. The grey triangles are the best-known fronts

Figures 4.6 and 4.7 present the mean values of the three run-time performance metrics for both initialization methods for NSGA-II and Borg, respectively. In both

figures, red dashed and black solid lines represent results obtained using the MOPHSM and the random approach, respectively. As can be seen, when the MOPHSM was used instead of the random initialization approach, near-optimal fronts (95% of the hypervolume of the best-known front) were able to be identified much more quickly on a consistent basis, irrespective of which MOEA was used. It can also be seen that the relative advantage of using the MOPHSM is more pronounced for larger problems. For example, with the aid of the MOPHSM (i) NSGA-II only required approximately 130 generations to identify a near-optimal front for the BN_{454} problem, which is about 50% of the generations needed when the random initialization approach was used (Figure 4.6) and (ii) Borg was able to reach the near-optimal front using approximately 420 generations for the BN_{454} case study, with *HI* values consistently higher than those from the random method throughout the run up to 500 generations (Figure 4.7). In addition, for the largest problem (KL problem with 1274 decision variables), both MOEAs (NSGA-II and Borg) produced substantially larger values of *HI* compared with those obtained using the random initialization approach, as shown in Figures 4.6 and 4.7.

The results in the middle panel of Figures 4.6 and 4.7 show that the main advantage of the MOPHSM over the random method is its greater ability to produce fronts with lower generational distance to the best-known fronts. In terms of the extent of the fronts, as shown in the right panels of Figures 4.6 and 4.7, use of both initialization methods exhibited comparable performance although the random method gave a higher value of this measure when Borg is applied to the KL problem.

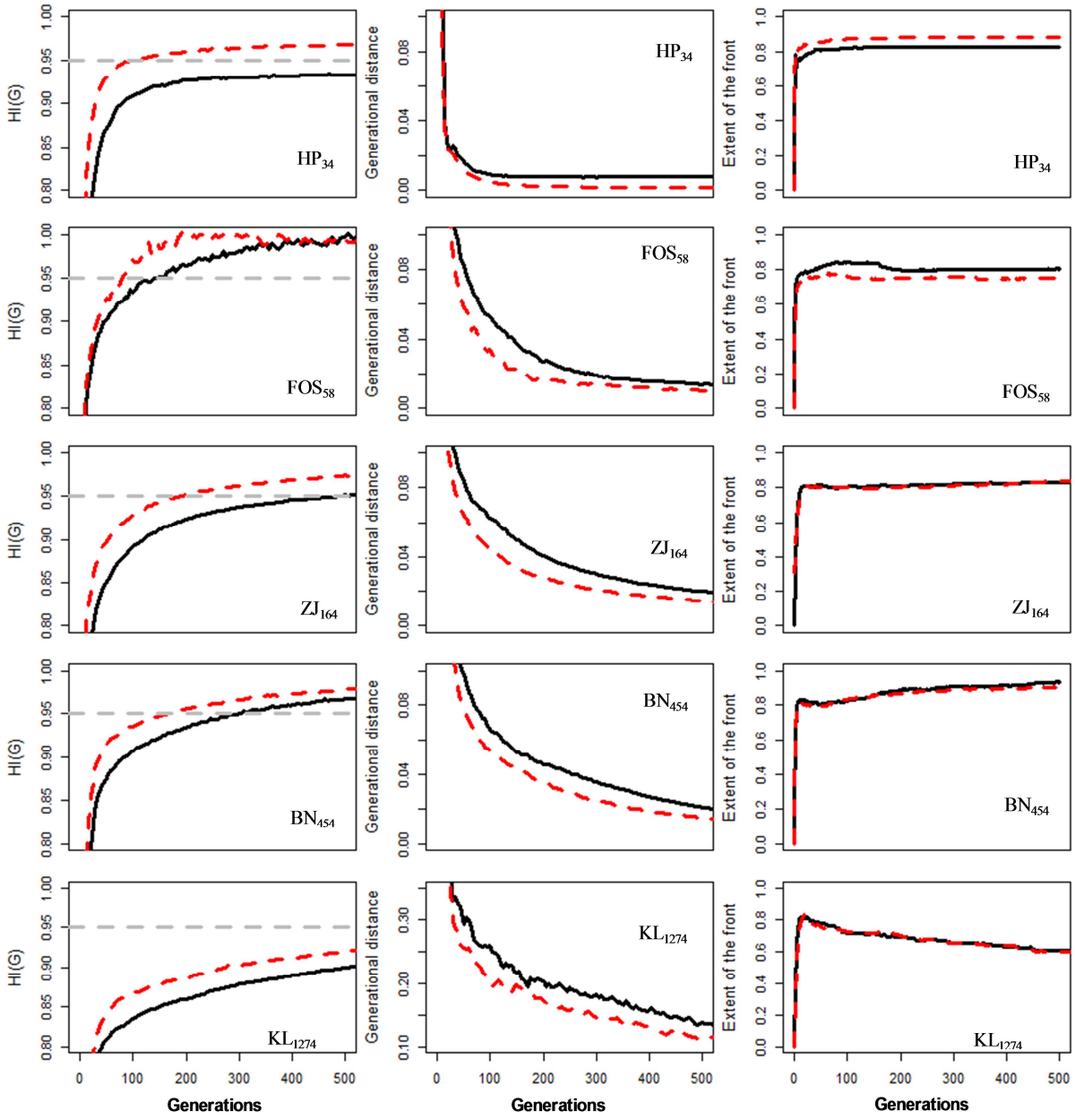

Figure 4.6 Run-time performance metrics of the proposed MOPHSM (red dashed line) and the random initialization approach (black solid lines) for NSGA-II. The horizontal grey line in the left panel indicates 95% of the best-known front hypervolume (near-optimal fronts)

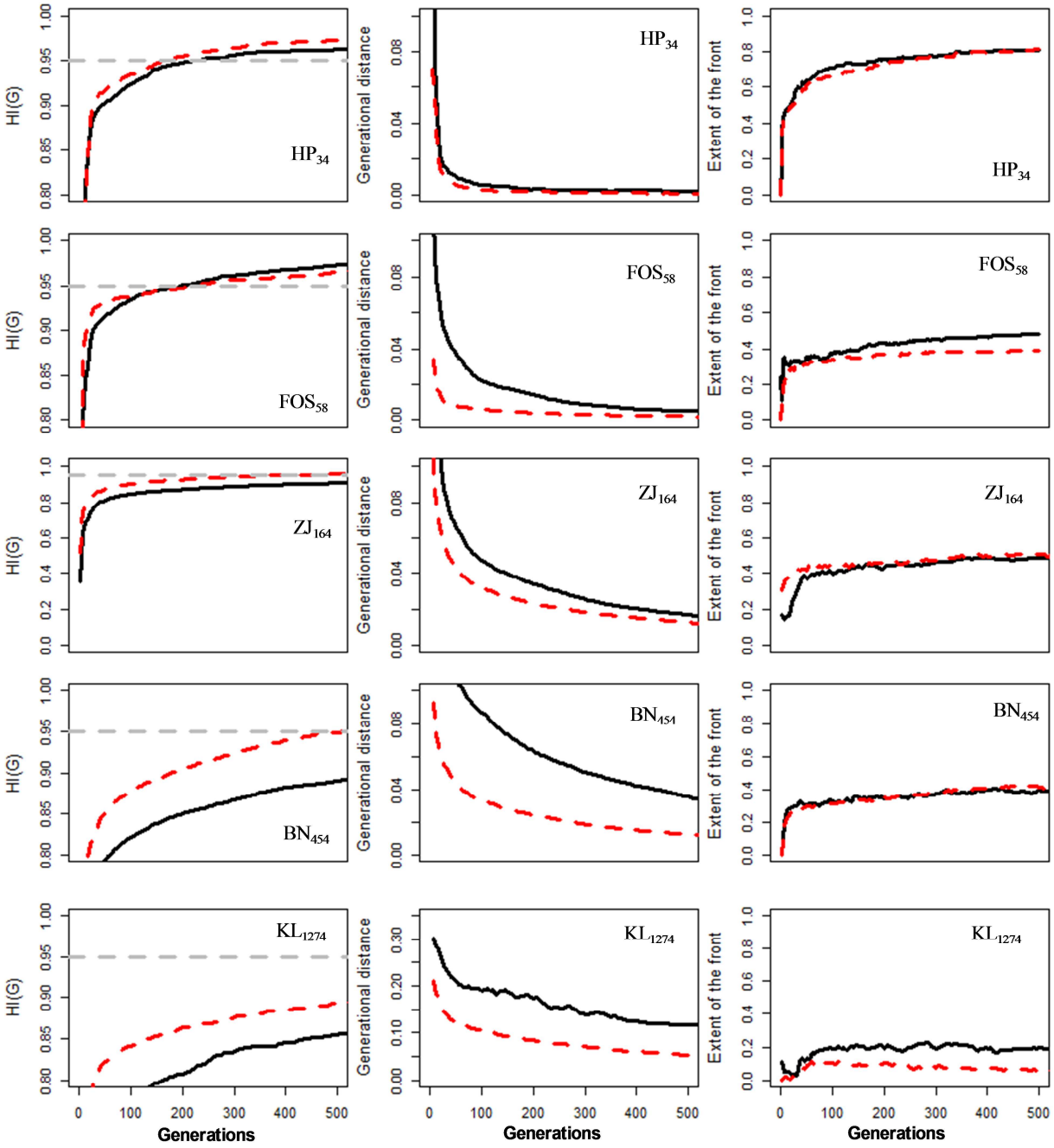

Figure 4.7 Run-time performance metrics of the proposed MOPHSM (red dashed line) and the random initialization approach (black solid lines) for Borg. The horizontal grey line in the left panel indicates 95% of the best-known front hypervolume (near-optimal fronts)

It should be noted that the end-of-run (G=2500) fronts and performance statistics obtained using the two initialization methods considered (not shown) are similar,

suggesting that there is no advantage in using the proposed MOPHSM method if the aim is to identify the globally optimal front. However, as the above results clearly demonstrate, if the aim is to identify near-optimal solutions to complex problems within realistic computational budgets, there are significant advantages in using the proposed approach. It should also be noted that overall, the relative performance of NSGA-II and Borg is in line with that obtained by Wang et al. (2015) and explained in Zheng et al. (2015).

4.7 Summary and conclusions

Over the past few decades, EAs have been used extensively for the optimization of WDSs. In recent years, there has been increased focus on the application of EAs to more complex WDSs and on the inclusion of multiple objectives, resulting in high computational demands and long run-times. In order to address this issue and to enable EAs to be applied more easily in practice, a significant amount of research has focused on the development of methods that enable near-optimal solutions to be identified within reasonable computational budgets, rather than on necessarily finding the globally optimal solution or Pareto front.

This paper makes a contribution in this field of research by developing and testing an approach to identifying good initial solutions for WDS design problems that aim to minimize network cost and maximize network resilience. The method builds on that proposed by Bi et al. (2015) for single objective WDS design problems aimed at minimizing network cost and uses domain knowledge about the attributes of good designs, including the relationship between pipe diameter and the distance to the supply sources and the interaction between flow velocities and network resilience.

The relative advantage of using the proposed MOPHSM compared with using a random initialization method in terms of the computational effort required to identify near-optimal solutions was assessed on five case studies of varying size and complexity using two different MOEAs (NSGA-II and Borg). Performance was

assessed using three run-time metrics, namely the hypervolume, the generational distance to the best-known fronts, and the extent of the fronts.

The results show that while the proposed MOPHSM is unable to improve the quality of solutions at the end of the run, its use enables near-optimal solutions to be identified at much smaller computational expense, irrespective of which of the two MOEAs is used. The advantage of using the MOPHSM is particularly noticeable when dealing with larger problems and smaller computational budgets. This is appealing from a practical perspective, as in many real-life applications, there is insufficient time to run MOEAs until no further improvement in the optimal fronts is obtained. While the use of the MOPHSM has been found to be beneficial, its utility should be assessed further in future studies by comparing its performance with that of other approaches to developing good initial solutions (see Introduction), which are currently generally only applicable to single-objective problems.

Chapter 5. Conclusions and Recommendations for Future Work

Evolutionary algorithms (EAs) have been used extensively to optimise the design of water distribution systems (WDSs) over the last 20 years. However, these applications are still mainly limited to the research domain due to their large computational requirements when applied to real-world problems. These requirements often go beyond the computational budgets that are typically available in practice. In order to address this issue, there is general consensus that identifying near-optimal solutions in a reasonable timeframe, rather than trying to find the globally optimal solution in an unaffordable timeframe, is of great practical importance. While many studies have been undertaken to achieve this goal, to date there have been limited efforts that consider the use of domain knowledge for this purpose.

5.1 Research Contributions

The overall contribution of this thesis is *the development of methods for generating initial solutions with the aid of WDS domain knowledge, thereby enabling EAs to identify near-optimal solutions (fronts) as quickly as possible. A further contribution is the use of run-time convergence statistics to provide an improved understanding of the speeds with which the near-optimal solutions are found.* While these proposed initialization methods do not necessarily improve the final solution quality compared to the random initialization method after very long run times, they are capable of identifying near-optimal solutions with significantly reduced computational effort. This is verified through the results obtained for a number of WDS case studies with increasing complexity. Such a feature is of particular importance when EAs are applied to real-world problems, as they often require the use of computationally intensive simulation models for objective function and/or constraint evaluation. It is anticipated that the initialization methods outlined in this thesis will enable a wider up-take of EAs in practice, thereby enabling their full potential to be realized within the WDS design domain.

The research contributions in each chapter are outlined below to specifically meet the objectives of this research stated in **Chapter 1**.

1. In the first publication given in **Chapter 2**, a new heuristic initialization method for seeding GA populations was introduced and evaluated, in which domain knowledge about the relationship between pipe size and distance to the supply sources, as well as the impact of the flow velocities on optimal solutions, were considered. This initialization method was compared with three other methods (an existing heuristic sampling method and two more traditional sampling methods, including random sampling and Latin Hypercube sampling) on seven WDS optimisation (pipe-sizing) problems with increasing complexity. The results obtained indicated that overall, the proposed initialization method significantly outperformed the other three sampling methods, both in terms of solution quality (single-objective cost) and computational efficiency. It was also found that the relative advantage of the proposed method was greater for larger networks. This demonstrates that the incorporation of domain knowledge into the generation of initial solutions is effective in guiding EAs' searching quickly towards promising regions, thus enabling near-optimal solutions to be reached within very limited timeframes(meeting **Objective 1**).

2. In the second publication (**Chapter 3**), it was found that both EAs' starting positions and searching mechanisms significantly affect their capacity to efficiently identify near-optimal solutions, and the latter exhibited relatively more noticeable impacts. With the aid of run-time behavior analysis, it was observed that improvements in objective cost function and reduction in population diversity were strongly correlated, implying that the convergence properties in the decision space heavily affect the searching quality in objective space. This observation shed new light on the causes of performance differences between different algorithms, parameterizations and starting positions, making it possible to modify an algorithm's performance through manipulating its population diversity. Another important finding in this study in that the fitness functions of WDS design problems are likely to consist of a single "big bowl" structure with a rugged base that is likely to

have many local optima. This can explain the great utility of the proposed initialization method in **Chapter 2** (paper 1), as the initial solutions obtained using this method enabled the search to commence part way down the "side" of the "big bowl". Based on the observed fitness function characteristics of the case studies considered, the use of EAs with reduced explorative capability (lower mutation rates) is expected to be more effective in terms of efficiently identifying near-optimal solutions. This is useful guidance for EA parameterization.

3. In the third publication outlined in **Chapter 4**, a multiobjective initialization method was proposed to identify high quality initial populations for multi-objective EAs (MOEAs) applied to WDS design problems, aimed at minimizing cost and maximizing network resilience (a measure of WDS supply reliability). In addition to engineering experience about the relationship between pipe size and distance to the source(s) of water as considered in the first publication (**Chapter 2**), domain knowledge about the relationship between flow velocities and network resilience was also accounted for. The proposed approach was compared with randomly generating initial populations in relation to finding near-optimal solutions more efficiently based on five WDS case studies of varying complexity with two different MOEAs (NSGA-II and Borg). The results indicate that there are considerable benefits in using the proposed initialization method in terms of being able to identify near-optimal solutions (fronts) more quickly, irrespective of MOEA type, with benefits being more pronounced for larger problems and smaller computational budgets.

5.2 Research Limitations

The limitations of this research are discussed below.

1. As part of the proposed initialization method, only domain knowledge with regard to pipe-sizing is considered. In other words, the WDS case studies

considered are purely pipe-sizing problems, without considering the design of other hydraulic elements, such as valves, tanks and pumps.

2. Domain knowledge is only used in the initialization phase of EAs in this thesis, while it has not been implemented to dynamically guide the searching during the optimization process. For example, the relationship between the pipe size and distance to the source(s) of water can also be considered at each generation after the algorithm operators (e.g. crossover and mutation) have been applied, in addition to the initial population.
3. The effectiveness of the proposed multiobjective initialization method was demonstrated for the objectives of the minimization of cost and the maximization of network resilience. There are many other objectives within WDS design that should be considered, such as minimization of greenhouse gas emissions

5.3 Recommendations for Future Work

This research has developed new initialization methods that have successfully assisted EAs to identify near-optimal solutions (fronts) for WDS design problems in a computationally efficient manner. However, there are still opportunities to address the limitations outlined above as part of future studies along this research line:

1. Incorporating the domain knowledge in relation to other hydraulic components into EA optimization of WDSs, in addition to pipes. It is possible to extend the proposed domain-knowledge based methodology to deal with pipe cleaning and relining within the optimization process through considering the reasonable range of the velocities. However, the inclusion of the domain knowledge for the design of pumps, tanks and valves is not straightforward, requiring further investigation.
2. Applying the proposed method to solve other loading cases, in addition to a single loading case as considered in this thesis. For example, fire loading cases (multiple loading cases) and water quality could be solved by checking whether the velocities lie within acceptable limits.

3. Development of a more advanced EA framework that is able to dynamically implement domain knowledge within the optimization process, in addition to the initialization stage, as considered in this thesis.
4. Modification of the proposed multiobjective initialization method to account for a number of other objectives for WDS design problems. For example, leakage losses as a function of pipe sizes and pressures could be considered as a separate objective during the design optimization. Similarly, water quality could also be treated as a separate objective (where appropriate).
5. Development of advanced EA algorithms to enable dynamic adjustment in parameterization according to the improved understanding between the searching quality in objective space and convergence in decision space obtained in this thesis.

References

Arsenault, R., Poulin, A., Côté, P., and Brissette, F. (2014). "Comparison of Stochastic Optimization Algorithms in Hydrological Model Calibration." *Journal of Hydrologic Engineering*, 19(7), 1374-1384.

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

Bi, W., Dandy, G. C., and Maier, H. R. (2015a). "Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge." *Environmental Modelling & Software*, 69(0), 370-381.

Bi, W., Maier, H. R., and Dandy, G. C. (2015b). " Impact of starting position and searching mechanism on evolutionary algorithm convergence rate " *submitted to Journal of Water Resources Planning and Management*, July, 2015.

Bi, W., Dandy, G. C., and Maier, H. R. (2015c). "Use of domain knowledge to increase the convergence rate of evolutionary algorithms for optimizing the cost and resilience of water distribution systems" *submitted to Journal of Water Resources Planning and Management*, September, 2015.

Bolognesi, A., Bragalli, C., Marchi, A., and Artina, S. (2010). "Genetic Heritage Evolution by Stochastic Transmission in the optimal design of water distribution networks." *Advances in Engineering Software*, 41(5), 792-801.

Bragalli, C., D'Ambrosio, C., Lee, J., Lodi, A., and Toth, P. (2012). "On the optimal design of water distribution networks: a practical MINLP approach." *Optimization and Engineering*, 13(2), 219-246.

Broad, D. R., Dandy, G. C., and Maier, H. R. (2005). "Water distribution system optimization using metamodels." *Journal of Water Resources Planning and Management*, 131(3), 172-180.

Broad, D. R., Maier, H. R., and Dandy, G. C. (2010). "Optimal Operation of Complex Water Distribution Systems Using Metamodels." *Journal of Water Resources Planning and Management*, 136(4), 433-443.

Creaco E., and Franchini M. (2012). "Fast network multi-objective design algorithm combined with an a-posteriori procedure for reliability evaluation under various operational scenarios", *Urban Water Journal*,9(6), 385-399.

Dandy, G. C., Simpson, A. R., and Murphy, L. J. (1996). "An improved genetic algorithm for pipe network optimization." *Water Resources Research*, 32(2), 449-458.

Deb, K. (2000). "An efficient constraint handling method for genetic algorithms." *Computer Methods in Applied Mechanics and Engineering*, 186(2-4), 311-338.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). "A fast and elitist multiobjective genetic algorithm: NSGA-II." *Evolutionary Computation, IEEE Transactions on*, 6(2), 182-197.

Deuerlein, J. W. (2008). "Decomposition model of a general water supply network graph." *Journal of Hydraulic Engineering*, 134(6), 822-832.

di Pierro, F., Khu, S.-T., Savic, D., and Berardi, L. (2009). "Efficient multi-objective optimal design of water distribution networks on a budget of simulations using hybrid algorithms." *Environmental Modelling & Software*, 24(2), 202-213.

Fu, G., Kapelan, Z., and Reed, P. (2012). "Reducing the Complexity of Multiobjective Water Distribution System Optimization through Global Sensitivity Analysis." *Journal of Water Resources Planning and Management*, 138(3), 196-207.

Fujiwara, O., and Khang, D. B. (1990). "A two-phase decomposition method for optimal design of looped water distribution networks." *Water Resources Research*, 26(4), 539-549.

Gibbs, M. S., Dandy, G. C., and Maier, H. R. (2008). "A genetic algorithm calibration method based on convergence due to genetic drift." *Information Sciences*, 178(14), 2857-2869.

Gibbs, M. S., Dandy, G. C., and Maier, H. R. (2010a). "Calibration and Optimization of the Pumping and Disinfection of a Real Water Supply System." *Journal of Water Resources Planning and Management*, 136(4), 493-501.

Gibbs M.S., Maier H.R. and Dandy G.C. (2010b). "Comparison of genetic algorithm parameter setting methods for chlorine injection optimization." *Journal of Water Resources Planning and Management*, 136(2), 288-291, DOI: 10.1061/(ASCE)WR.1943-5452.0000033

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.

Gibbs, M. S., Maier, H. R., and Dandy, G. C. (2015). "Using characteristics of the optimisation problem to determine the Genetic Algorithm population size when the number of evaluations is limited." *Environmental Modelling & Software*, 69(0), 226-239.

Gupta, I., Gupta, A., and Khanna, P. (1999). "Genetic algorithm for optimization of water distribution systems." *Environmental Modelling & Software*, 14(5), 437-446.

Hadka, D., and Reed, P. (2012). "Borg: An Auto-Adaptive Many-Objective Evolutionary Computing Framework." *Evolutionary Computation*, 21(2),231-259.

Hadka, D., and Reed, P. (2012). "Diagnostic Assessment of Search Controls and Failure Modes in Many-Objective Evolutionary Optimization." *Evolutionary Computation*, 20(3), 423-452.

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 & Software*, 69(0), 353-369.

Housh, M., Ostfeld, A., and Shamir, U. (2013). "Limited multi-stage stochastic programming for managing water supply systems." *Environmental Modelling & Software*, 41(0), 53-64.

JRC, 2008. European Commission joint Research Center. http://simlab.jrc.ec.europa.eu/.

Kadu, M. S., Gupta, R., and Bhave, P. R. (2008). "Optimal design of water networks using a modified genetic algorithm with reduction in search space." *Journal of Water Resources Planning and Management*, 134(2), 147-160.

Kang, D., and Lansey, K. (2012). "Revisiting Optimal Water-Distribution System Design: Issues and a Heuristic Hierarchical Approach." *Journal of Water Resources Planning and Management*, 138(3), 208-217.

Kapelan, Z. S., Savic, D. A., and Walters, G. A. (2005). "Multiobjective design of water distribution systems under uncertainty." *Water Resources Research*, 41(11), W11407.

Keedwell, E., and Khu, S.-T. (2006). "Novel Cellular Automata Approach to Optimal Water Distribution Network Design." *Journal of Computing in Civil Engineering*, 20(1), 49-56.

Khedr, A., and Tolson, B. (2015). "Comparing Optimization Techniques with an Engineering Judgment Approach to WDN Design." *Journal of Water Resources Planning and Management*, C4015014. DOI:10.1061/(ASCE)WR.1943-5452.0000611.

Kollat, J. B., and Reed, P. M. (2006). "Comparing state-of-the-art evolutionary multi-objective algorithms for long-term groundwater monitoring design." *Advances in Water Resources*, 29(6), 792-807.

Kollat, J. B., Reed, P. M., and Kasprzyk, J. R. (2008). "A new epsilon-dominance hierarchical Bayesian optimization algorithm for large multiobjective monitoring network design problems." *Advances in Water Resources*, 31(5), 828-845.

Krapivka, A., and Ostfeld, A. (2009). "Coupled Genetic Algorithm---Linear Programming Scheme for Least-Cost Pipe Sizing of Water-Distribution Systems." *Journal of Water Resources Planning and Management*, 135(4), 298-302.

Laucelli, D., Berardi, L., and Giustolisi, O. (2012). "Assessing climate change and asset deterioration impacts on water distribution networks: Demand-driven or pressure-driven network modeling?" *Environmental Modelling & Software*, 37(0), 206-216.

Li, X., Wei, J., Fu, X., Li, T., and Wang, G. (2014). "Knowledge-Based Approach for Reservoir System Optimization." *Journal of Water Resources Planning and Management*, 140(6), 04014001.

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 & Software*, 62(0), 271-299.

Maier, H.R., Kapelan, Z., Kasprzyk, J. and Matott, L.S. (2015). “Thematic issue on evolutionary algorithms in water resources.” *Environmental Modelling & Software*, 69, 222-225, DOI:10.1016/j.envsoft.2015.05.003

Marchi, A., Salomons, E., Ostfeld, A., Kapelan, Z., Simpson, A., Zecchin, A., Maier, H., Wu, Z., Elsayed, S., Song, Y., Walski, T., Stokes, C., Wu, W., Dandy, G., Alvisi, S., Creaco, E., Franchini, M., Saldarriaga, J., Páez, D., Hernández, D., Bohórquez, J., Bent, R., Coffrin, C., Judi, D., McPherson, T., van Hentenryck, P., Matos, J., Monteiro, A., Matias, N., Yoo, D., Lee, H., Kim, J., Iglesias-Rey, P., Martínez-Solano, F., Mora-Meliá, D., Ribelles-Aguilar, J., Guidolin, M., Fu, G., Reed, P., Wang, Q., Liu, H., McClymont, K., Johns, M., Keedwell, E., Kandiah, V., Jasper, M., Drake, K., Shafiee, E., Barandouzi, M., Berglund, A., Brill, D., Mahinthakumar, G., Ranjithan, R., Zechman, E., Morley, M., Tricarico, C., de Marinis, G., Tolson, B., Khedr, A., and Asadzadeh, M. (2014a). "Battle of the Water Networks II." *Journal of Water Resources Planning and Management*, 140(7), 04014009.

Marchi, A., Dandy, G., Wilkins, A., and Rohrlach, H. (2014b). "Methodology for Comparing Evolutionary Algorithms for Optimization of Water Distribution Systems." *Journal of Water Resources Planning and Management*, 140(1), 22-31.

Manache, G., and Melching, C. S. (2004). “Sensitivity Analysis of a Water-Quality Model Using Latin Hypercube Sampling.” *Journal of Water Resources Planning and Management* 130(3), 232-242.

Marques, J., Cunha, M., and Savić, D. A. (2015). "Multi-objective optimization of water distribution systems based on a real options approach." *Environmental Modelling & Software*, 63(0), 1-13.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). “A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code.” *Technometrics* 21(2), 239-245.

Nicklow, J., Reed, P., Savic, D., Dessalegne, T., Harrell, L., Chan-Hilton, A., Karamouz, M., Minsker, B., Ostfeld, A., Singh, A. (2010). "State of the art for genetic algorithms and beyond in water resources planning and management." *Journal of Water Resources Planning and Management*, 136(4), 412-432.

Ong, Y. S., and Keane, A. J. (2002). "A domain knowledge based search advisor for design problem solving environments." *Engineering Applications of Artificial Intelligence*, 15(1), 105–116.

Ostfeld, A., Oliker, N., and Salomons, E. (2014). "Multiobjective Optimization for Least Cost Design and Resiliency of Water Distribution Systems." *Journal of Water Resources Planning and Management*, 140(12), 04014037.

Prasad, T. D., and Park, N.-S. (2014). "Multiobjective genetic algorithms for design of water distribution networks." *Journal of Water Resources Planning and Management*, 130(1), 73-82.

Prasad, T.D., Sung-Hoon, H., and Namsik, P., (2003).“Reliability based design of water distribution networksusing multiobjective genetic algorithms”. *KSCE Journal ofCivil Engineering*, 7 (3), 351 – 361.

Raad, D. N., Sinske, A. N., and van Vuuren, J. H. (2010). "Comparison of four reliability surrogate measures for water distribution systems design." *Water Resources Research*, 46(5), W05524.

Razavi, S., Tolson, B. A., and Burn, D. H. (2012). "Review of surrogate modeling in water resources." *Water Resources Research*, 48(7), W07401.

Reca, J., and Martínez, J. (2006). "Genetic algorithms for the design of looped irrigation water distribution networks." *Water Resources Research*, 42(5), W05416.

Roshani, E., and Filion, Y. (2012). "Using parallel computing to increase the speed of water distribution network optimization." The 14th Water Distribution Systems Analysis Conference, ASCE, Adelaide, South Australia.

Sapuan, S. M. (2001). "A knowledge-based system for materials selection in mechanical engineering design." *Materials & Design*, 22(8), 687-695.

Simpson, A. R., Dandy, G. C., and Murphy, L. J. (1994). "Genetic algorithms compared to other techniques for pipe optimization." *Journal of Water Resources Planning and Management*, 120(4), 423-443.

Sitzenfrei, R., Möderl, M., and Rauch, W. (2013). "Automatic generation of water distribution systems based on GIS data." *Environmental Modelling & Software*, 47(0), 138-147.

Stokes C. S., Simpson A. R., and Maier H. R. (2014). "The Cost - Greenhouse Gas Emission Nexus for Water Distribution Systems including the Consideration of Energy Generating Infrastructure: An Integrated Optimization Framework and Review of Literature." *Earth Perspectives, 1:9 doi:10.1186/2194-6434-1-9.*

Stokes, C. S, Maier, H. R, and Simpson, A. R. (2015a). "Water Distribution System Pumping Operational Greenhouse Gas Emissions Minimization by Considering Time-Dependent Emissions Factors." *Journal of Water Resources Planning and Management*, 141(7), 04014088.

Stokes, C. S., Simpson, A. R., and Maier, H. R. (2015b). "A computational software tool for the minimization of costs and greenhouse gas emissions associated with water distribution systems." *Environmental Modelling & Software*, 69, 452-467.

Stokes C. S., Maier H. R. and Simpson A. R. (2015c). "Effect of storage tank size on the minimization of water distribution system cost and greenhouse gas emissions while considering time-dependent emissions factors." *Journal of Water Resources Planning and Management*, DOI: 10.1061/(ASCE)WR.1943-5452.0000582, 04015052.

Tolson, B. A., Maier, H. R., Simpson, A. R., and Lence, B. J. (2004). "Genetic algorithms for reliability-based optimization of water distribution systems." *Journal of Water Resources Planning and Management*, 130(1), 63-72.

Tolson, B. A., and Shoemaker, C. A. (2007). "Dynamically dimensioned search algorithm for computationally efficient watershed model calibration." *Water Resources Research*, 43(1), W01413.

Tolson, B. A., Asadzadeh, M., Maier, H. R., and Zecchin, A. (2009). "Hybrid discrete dynamically dimensioned search (HD-DDS) algorithm for water distribution system design optimization." *Water Resources Research*, 45(12), W12416.

Vairavamoorthy, K., and Ali, M. (2005). "Pipe index vector: a method to improve genetic-algorithm-based pipe optimization." *Journal of Hydraulic Engineering*, 131(12), 1117-1125.

Vasan, A., and Simonovic, S. P. (2010). "Optimization of water distribution network design using differential evolution." *Journal of Water Resources Planning and Management*, 136(2), 279-287.

Wang, Q., Guidolin, M., Savic, D., and Kapelan, Z. (2015). "Two-Objective Design of Benchmark Problems of a Water Distribution System via MOEAs: Towards the Best-Known Approximation of the True Pareto Front." *Journal of Water Resources Planning and Management*, 141(3), 04014060.

Walski, T. M. (2001). "The Wrong Paradigm—Why Water Distribution Optimization Doesn't Work." *Journal of Water Resources Planning and Management*, 127(4), 203-205.

Weinberger, E. (1990). "Correlated and uncorrelated fitness landscapes and how to tell the difference." *Biological Cybernetics*, 63(5), 325-336.

Wu, Z., and Zhu, Q. (2009). “Scalable Parallel Computing Framework for Pump Scheduling Optimization.” World Environmental and Water Resources Congress 2009, 1-11.

Wu, Z. Y., and Behandish, M. (2012). “Comparing methods of parallel genetic optimization for pump scheduling using hydraulic model and GPU-based ANN meta-model.” The 14th Water Distribution Systems Analysis Conference, ASCE, Adelaide, South Australia.

Wu, W., Maier, H. R., and Simpson, A. R. (2010). "Single-objective versus multiobjective optimization of water distribution systems accounting for greenhouse gas emissions by carbon pricing." *Journal of Water Resources Planning and Management*, 136(5), 555-565.

Wu, W., Maier, H. R., and Simpson, A. R. (2013). "Multiobjective optimization of water distribution systems accounting for economic cost, hydraulic reliability, and greenhouse gas emissions." *Water Resources Research*, 49(3), 1211-1225.

Zecchin, A. C., Simpson, A. R., Maier, H. R., Marchi, A., and Nixon, J. B. (2012). "Improved understanding of the searching behavior of ant colony optimization algorithms applied to the water distribution design problem." *Water Resources Research*, 48(9), W09505.

Zhang, W., Chung, G., Pierre-Louis, P., Bayraksan, G., and Lansey, K. (2013). "Reclaimed water distribution network design under temporal and spatial growth and demand uncertainties." *Environmental Modelling & Software*, 49(0), 103-117.

Zheng, F., Simpson, A. R., and Zecchin, A. C. (2011a). "A combined NLP-differential evolution algorithm approach for the optimization of looped water distribution systems." *Water Resources Research*, 47(8), W08531.

Zheng, F., Simpson, A. R., and Zecchin, A. C. (2011b). "Dynamically expanding choice-table approach to genetic algorithm optimization of water distribution systems." *Journal of Water Resources Planning and Management*, 137(6), 547-551.

Zheng, F., Simpson, A. R., and Zecchin, A. C. (2013a). "A decomposition and multistage optimization approach applied to the optimization of water distribution systems with multiple supply sources." *Water Resources Research*, 49(1), 380-399.

Zheng, F., Simpson, A. R., Zecchin, A. C., and Deuerlein, J. W. (2013b). "A graph decomposition-based approach for water distribution network optimization." *Water Resources Research*, 49(4), 2093-2109.

Zheng, F., Zecchin, A., and Simpson, A. (2013c). "Self-Adaptive Differential Evolution Algorithm Applied to Water Distribution System Optimization." *Journal of Computing in Civil Engineering*, 27(2), 148-158.

Zheng, F., Zecchin, A., Simpson, A., and Lambert, M. (2014a). "Noncrossover Dither Creeping Mutation-Based Genetic Algorithm for Pipe Network Optimization." *Journal of Water Resources Planning and Management*, 140(4), 553-557.

Zheng, F., Simpson, A., and Zecchin, A. (2014b). "Coupled Binary Linear Programming–Differential Evolution Algorithm Approach for Water Distribution System Optimization." *Journal of Water Resources Planning and Management*, 140(5), 585-597.

Zheng, F., and Zecchin, A. (2014c). "An efficient decomposition and dual-stage multi-objective optimization method for water distribution systems with multiple supply sources." *Environmental Modelling & Software*, 55(0), 143-155.

Zheng, F., Simpson, A. R., and Zecchin, A. C. (2014d). "An efficient hybrid approach for multiobjective optimization of water distribution systems." *Water Resources Research*, 50(5), 3650-3671.

Zheng, F., Zecchin, A. C., and Simpson, A. R. (2015a). "Investigating the run-time searching behavior of the differential evolution algorithm applied to water distribution system optimization." *Environmental Modelling & Software*, 69(0), 292-307.

Zheng, F., Simpson, A., and Zecchin, A. (2015b). "Improving the efficiency of multi-objective evolutionary algorithms through decomposition: An application to water distribution network design." *Environmental Modelling & Software*, 69(0), 240-252.

Zheng, F. (2015c). "Comparing the Real-Time Searching Behavior of Four Differential-Evolution Variants Applied to Water-Distribution-Network Design Optimization." *Journal of Water Resources Planning and Management*, 141(10), 04015016.

Zitzler, E., and Thiele, L. (1999). "Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach." *Evolutionary Computation, IEEE Transactions on*, 3(4), 257-271.

Appendix

Environmental Modelling & Software xxx (2014) 1–12

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge☆

W. Bi*, G.C. Dandy, H.R. Maier

School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, South Australia 5005, Australia

ARTICLE INFO

Article history:
Received 24 March 2014
Received in revised form
1 September 2014
Accepted 10 September 2014
Available online xxx

Keywords:
Optimization
Genetic algorithms
Water distribution systems
Domain knowledge
Heuristics
Computational efficiency

ABSTRACT

Over the last two decades, evolutionary algorithms (EAs) have become a popular approach for solving water resources optimization problems. However, the issue of low computational efficiency limits their application to large, realistic problems. This paper uses the optimal design of water distribution systems (WDSs) as an example to illustrate how the efficiency of genetic algorithms (GAs) can be improved by using heuristic domain knowledge in the sampling of the initial population. A new heuristic procedure called the Prescreened Heuristic Sampling Method (PHSM) is proposed and tested on seven WDS cases studies of varying size. The EPANet input files for these case studies are provided as supplementary material. The performance of the PHSM is compared with that of another heuristic sampling method and two non-heuristic sampling methods. The results show that PHSM clearly performs best overall, both in terms of computational efficiency and the ability to find near-optimal solutions. In addition, the relative advantage of using the PHSM increases with network size.

1. Introduction

Evolutionary algorithms (EAs) have been used successfully and extensively for solving water resources optimization problems in a number of areas, such as engineering design, the development of management strategies and model calibration (Nicklow et al., 2010; Zecchin et al., 2012). However, a potential shortcoming of EAs is that they are computationally inefficient, especially when applied to problems of realistic size. Consequently, there is a need to improve the computational efficiency of EAs to make them easier to use for the optimization of realistic water resources problems (Maier et al., 2014).

One application area where this is the case of is the design of water distribution systems (WDSs) (Marchi et al., 2014a,b; Stokes et al., 2014). Over the past two decades, a variety of EAs have been applied to this problem, as detailed in Zheng et al. (2013a). Among these, genetic algorithms (GAs) have been used most extensively (Simpson et al., 1994; Dandy et al., 1996; Gupta et al., 1999; Vairavamoorthy and Ali, 2005; Krapivka and Ostfeld, 2009; Kang and Lansey, 2012; Zheng et al., 2013b). However, GAs have been primarily applied to relatively simple benchmark problems, such as the 14-pipe problem (Simpson et al., 1994), the New York Tunnels problem with 21 tunnels (Dandy et al., 1996), and the Hanoi problem with 34 pipes (Zheng et al., 2011a). In recent years, there has been a move towards increasing the complexity and realism of the case studies to which GAs are applied, including the Balerma network with 454 pipes (Reca and Martínez, 2006), the Rural network with 476 pipes (Marchi et al., 2014a), the BWN-II network with 433 pipes (Zheng et al., 2013c), and the network used by Kang and Lansey (2012), which has 1274 pipes and will be referred to as the "KL" network for the remainder of this paper.

Increased network size and complexity result in significant challenges in terms of achieving good quality near-optimal solutions given the computational budgets that are typically available in practice (di Pierro et al., 2009; Fu et al., 2012). This is because (i) the time for hydraulic simulation increases appreciably for large WDSs; and (ii) the complexity and size of the search space associated with a large WDS are increased significantly. As a result, computational efficiency has been identified as a key concern for the widespread uptake of GAs for the optimization of large, real-world WDSs (di Pierro et al., 2009).

In order to address this issue, two main approaches have been adopted in the literature. As part of the first approach, it is argued

☆ Thematic Issue on Evolutionary Algorithms.

* Corresponding author. Postal address: N223, Engineering North, School of Civil, Environmental and Mining Engineering, North Terrace Campus, University of Adelaide, Adelaide, South Australia, 5005, Australia. Tel.: +61 8 83136139.

E-mail addresses: weiwei.bi@adelaide.edu.au (W. Bi), graeme.dandy@adelaide.edu.au (G.C. Dandy), holger.maier@adelaide.edu.au (H.R. Maier).

http://dx.doi.org/10.1016/j.envsoft.2014.09.010
1364-8152/

that for large, real problems, the focus should be on finding the best possible solution within a realistic computational budget, rather than on attempting to find the global optimal solution (e.g. Tolson and Shoemaker, 2007; Gibbs et al., 2008; Tolson et al., 2009; Gibbs et al., 2010, 2011). This is because for such large problems, the global optimal solution is unlikely to be found within a reasonable computational timeframe.

As part of the second approach, efforts have been made to increase the computational efficiency of the optimization process. This has been done in a number of ways, including the use of increased computational power, such as parallel and distributed computing (Wu and Zhu, 2009; Roshani and Filion, 2012; Wu and Behandish, 2012), the use of surrogate- and meta-modeling to speed up the simulation process (e.g. Broad et al.,. 2005; di Pierro et al., 2009; Broad et al., 2010; Razavi et al., 2012), and the seeding of the initial population of EAs with good solutions obtained using a variety of analytical techniques (e.g. Keedwell and Khu, 2006; Zheng et al., 2011a, 2014a,b; Zheng and Zecchin, 2014; Fu et al., 2012). It should be noted that similar concepts have recently also been used in conjunction with other optimization techniques (e.g. Zhang et al., 2013; Housh et al., 2013) and non-optimization based WDS design approaches (e.g. Sitzenfrei et al., 2013).

Although some of the methods mentioned above utilize engineering knowledge in their development (e.g. Keedwell and Khu, 2006; Zheng et al., 2011a), there have been limited attempts to incorporate engineering knowledge and experience directly. Only Kang and Lansey (2012) have combined engineering experience with GAs in order to increase the computational efficiency of the optimization process. This was achieved by seeding half of the initial GA population with solutions that result in flow velocities below a threshold selected from within a pre-defined velocity range. However, the approach has only been applied to a single case study thus far and its relative performance has not yet been assessed in a rigorous and comprehensive manner. In addition, the approach has a number of potential shortcomings. Firstly, selection of an appropriate range for the velocity threshold is subjective, which might make the method difficult to apply and could result in inconsistent results from repeated, independent implementation of the method. Secondly, pipe sizes that result in appropriate velocities are determined using a structured trial-and-error process. However, in practice, pipe sizes generally reduce with distance from the source (Walski, 2001). Consequently, there exists an opportunity to incorporate this domain knowledge into the initial pipe sizing process. Finally, there is limited control over population diversity, as this is achieved by seeding the initial population with 50% of randomly generated solutions and 50% of the solutions obtained based on engineering experience.

In order to address these shortcomings, the objectives of this paper are (i) to introduce a new heuristic sampling method for determining the initial population of GAs for the least-cost design of WDSs that is based on engineering experience/domain knowledge and that overcomes the potential shortcomings of the method of Kang and Lansey (2012); and (ii) to provide a rigorous assessment of the performance of this method compared with that of Kang and Lansey's sampling method (KLSM) and two sampling methods that do not consider any domain knowledge (i.e. random sampling (RS) and Latin hypercube sampling (LHS)) on seven WDS design case studies of varying size and complexity.

The remainder of this paper is organized as follows. The proposed heuristic, domain knowledge based sampling method for determining the initial population of GAs for the least-cost design of WDSs is introduced in next section, followed by the methodology for assessing the performance of this method against that of the KLSM and the two non-heuristic sampling methods. Next, the results are presented and discussed, followed by summary and conclusions.

2. Proposed prescreened heuristic sampling method for WDS design

The proposed heuristic sampling method for initializing the population of GAs for the least-cost design of WDSs based on domain knowledge is called the Prescreened Heuristic Sampling Method (PHSM). It uses a three-step procedure that (i) selects pipe sizes based on knowledge that pipe diameters generally get smaller the further they are from the source; (ii) dynamically adjusts the velocity threshold to account for the fact that appropriate velocity thresholds are likely to be network dependent; and (iii) enables the diversity of the initial population to be controlled by sampling from distributions centred on the solutions determined using the heuristic procedures in (i) and (ii). The PHSM has some similarities to the KLSM in that it aims to find initial pipe sizes that restrict flow velocities to lie within certain ranges. However, it overcomes the potential limitations of the KLSM outlined in the Introduction. Details of the three steps of the PHSM are given below.

Step 1: Assign pipe diameters based on distances between demand nodes and supply sources

As mentioned above, the first step of the PHSM is motivated by the knowledge that, in real WDSs, the diameters of upstream pipes are generally larger than those further downstream (Walski, 2001). However, for WDSs, each demand node usually has a number of different paths that connect it to the supply source (reservoir). This indicates that the spatial distance between each demand node and the reservoir may vary according to the paths selected to deliver the required demands. In the proposed method, the shortest delivery path to each demand node is selected and used to represent the spatial distance between that node and the source node. The rationale behind this is that it has been demonstrated that the majority of the demand at a node is supplied by the path with the shortest distance for an optimal design of WDSs (Zheng et al., 2011a). The detailed process of step 1 of the PHSM is as follows:

i Find the shortest distance to a reservoir in the water network, l_i for each node i ($i = 1,2 \ldots ..,n$, where n is the total number of demand nodes in the network) using the Dijkstra algorithm (Deo, 1974). When dealing with a water network with multiple reservoirs, an augmented source node is created to connect all the reservoirs to enable the determination of l_i following Deuerlein (2008) and Zheng et al. (2011a).
ii Obtain the largest value of the shortest distance L by $L = \max(l_i)$.
iii Divide the network into P specific areas with the shortest distance to the source node interval of L/P, where P is the number of available pipe diameters for the design.
iv Assign pipes in each area a different diameter, with the largest diameter assigned to the pipes in the area nearest to the source and the smallest diameter to the pipes in the area furthest from the source (reservoir). All pipes in a single area are assigned the same diameter.

For example, for the WDS introduced by Zheng et al. (2011a), which has 164 pipes (Fig. 1), the largest shortest distance of all nodes (L) is obtained after steps i and ii. If there are five diameter options for this network (i.e. $P = 5$), the network will be divided into five areas in step iii. In order to do this (i) all nodes that have a shortest-distance that is not greater than L/P (i.e. $0 < l_i \leq L/5$) form Area 1; (ii) all nodes that have a shortest-distance that is larger than L/P but not greater than $2L/5$ (i.e. $L/5 < l_i \leq 2L/5$) form Area 2; …; (v)

Please cite this article in press as: Bi, W., et al., Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.09.010

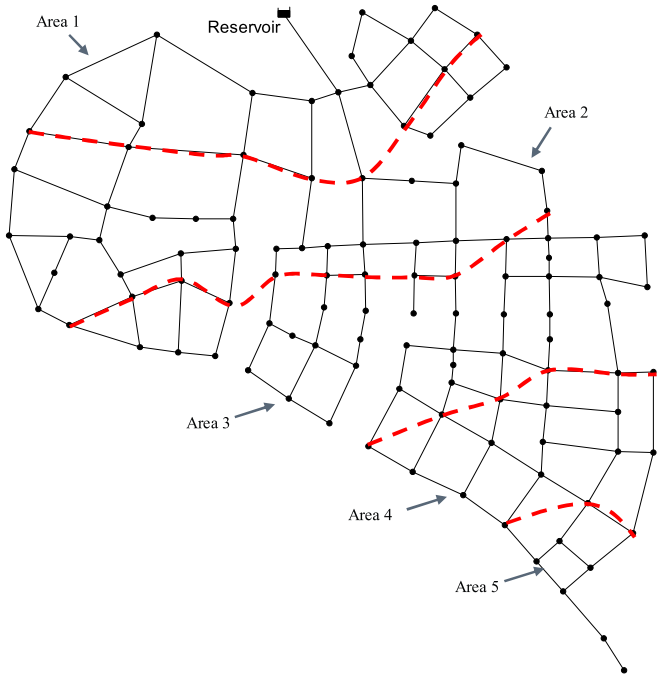

Fig. 1. WDS used to illustrate the result of network division of the PHSM (The red dotted lines represent the distance boundary used to assign diameters). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

all nodes that have a shortest-distance larger than 4L/P (i.e. $4L/5 < l_i \leq L$) form Area 5. The resulting division of the network is given in Fig. 1. Finally, (i) all pipes in Area 1 are assigned the largest diameter; (ii) all pipes in Area 2 are assigned the second largest diameter; and so on until all pipes in Area 5 are assigned the smallest pipe diameter. As such, the diameters of the upstream pipes are generally larger than those of the downstream pipes.

Step 2: Adjust pipe diameters based on velocities

In this step, the diameters obtained in step 1 are refined to achieve flow velocities in all pipes that are close to a particular threshold. This is based on the domain knowledge that the velocity in each pipe of an optimal solution for a WDS is in a limited range. In addition, in order to ensure that the chosen pipe diameters approach optimal values, the velocity threshold is selected to result in solutions that are on the boundary between feasibility and infeasibility. This is because the optimal solution is often located on the boundary of the feasible and infeasible areas of the search space. The stages in the process for achieving this are shown in Fig. 2.

As can be seen from Fig. 2, an inner loop and an outer loop are involved in the algorithm. The inner loop is used to determine the network configuration based on pipe velocities. To do this, a threshold value v for velocity needs to be assigned at the beginning (e.g. $v = 0.1$ m/s), which represents the expected velocity for each pipe in the network. The network with initial diameters determined in Step 1 is then simulated using a hydraulic solver to obtain the flow rate for each pipe. Based on this flow rate, the new diameter ND_j for each pipe can be calculated using:

$$ND_j = \sqrt{\frac{4Q_j}{\pi v}} \tag{1}$$

where $j = 1, \ldots, m$ is the jth pipe in the water network and m is the total number of pipes.

As continuous diameter values are generated using Equation (1), these values need to be rounded up or down to the nearest discrete diameter based on the available options.

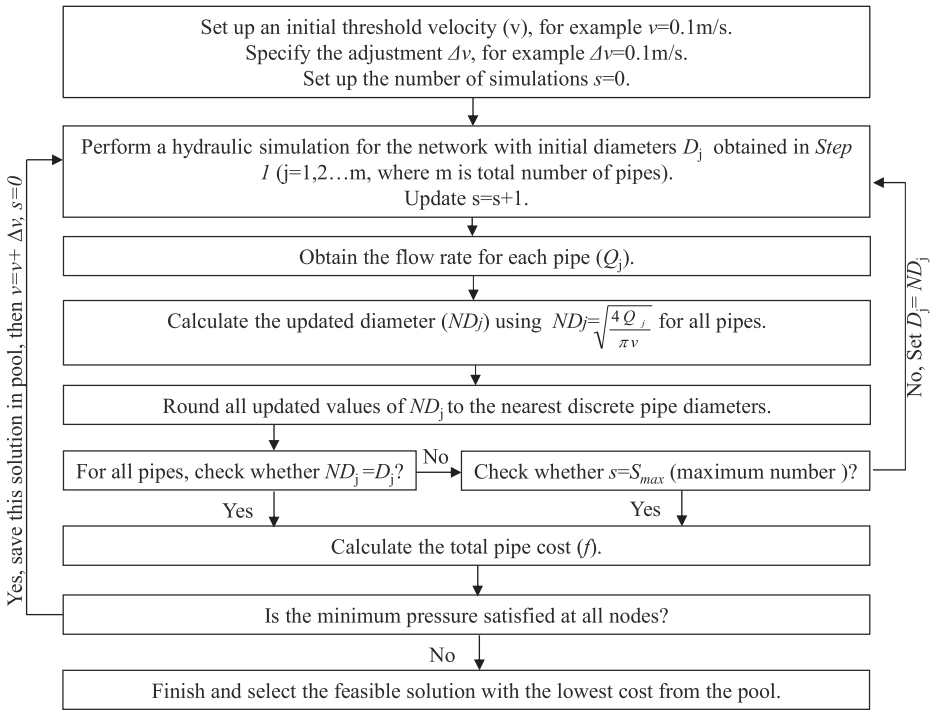

Fig. 2. Flowchart of the algorithm for adjusting pipe diameters based on flow velocity.

Please cite this article in press as: Bi, W., et al., Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.09.010

The inner loop continues until there is no further change in diameter in accordance with Equation (1) or the number of simulations (s) reaches the specified maximum number of allowable simulations (S_{max}), at which point the cost (f) and the minimum pressure head of this design are determined. If this solution is feasible (i.e., the pressure head constraints are satisfied), the network configuration and its associated network cost are saved to an archive. As part of the outer loop, the inner loop is repeated for successive increases in the velocity threshold (i.e. $v = v + \Delta v$) until no feasible solution can be found. If the solution found at the completion of the inner loop is infeasible, the outer loop is not performed and the process of adjusting diameters is terminated. Finally, the feasible solution with the lowest cost for the different velocity thresholds considered is selected from the archive and denoted as an approximate optimal solution for the WDS being optimized. This solution is then used as the starting point for Step 3, as outlined below.

Step 3: Generate distribution functions based on the approximate optimal solution determined in Step 2.

In order to ensure sufficient diversity in the initial solution, the initial diameter for each pipe is generated from a distribution, such that the pipe diameter obtained in Step 2 has the highest probability of being selected. The logic behind this is that the approximate diameter for a pipe determined in Step 2 is most likely to be the optimal diameter relative to other diameter options. Hence, a relatively higher density function value is assigned to this diameter (i.e. it is more likely to be selected during sampling).

The density function $f(D_k)$ and the distribution function $F(D_k)$ for selecting each initial diameter are given by the following equations:

$$f(D_k) = \frac{1}{1 + a|x|} \quad k = 1, ..., P \tag{2}$$

$$F(D_k) = \frac{f(D_k)}{\sum_{k=1}^{P} f(D_k)} \quad k = 1, ..., P \tag{3}$$

where a is a constant factor to control the density of each diameter D_k, details of which are discussed in Section 4; x is the distance between D_k and D_c (the diameter for a pipe in the approximate optimal solution determined in Step 2) in terms of integer coding; and P is the total number of available pipe diameters.

In order to illustrate how the approach outlined above is used to generate the pipe diameters in the initial solution, the following example is used. Table 1 presents the assumed total pipe diameter options and their corresponding integer coding values. If $D_c = 200$ mm in Step 2 for a particular pipe, its integer code is 1, as shown in Table 1. The absolute distance $|x|$ between each D_k and D_c is then calculated and presented in the third column of Table 1. The density function and distribution function values for generating each available diameter for this pipe during sampling are calculated based on Equations (2) and (3), respectively (assuming $a = 1$). The results are given in the fourth and fifth columns of Table 1. As can be seen, a diameter of 200 mm has the largest probability of being selected during sampling, as this diameter is selected based on the heuristic rules used in Steps 1 and 2. In contrast, a diameter of 600 mm has the smallest probability of being selected, since it has the largest distance to the optimal diameter of 200 mm.

Table 1
Example to illustrate the application of Step 3 of the PHSM.

Pipe diameter D_k (mm)	Integer coding number	Absolute distance to D_c ($\|x\|$)	Density function value $f(D_k)$	Distribution function value $F(D_k)$
100	0	1	0.5	0.19
200	1	0	1	0.39
300	2	1	0.5	0.19
500	3	2	0.33	0.13

It should be noted that the assumption made in Step 1 that the upstream diameters are typically larger than those further downstream might not hold for all networks due to the influence of network topology and zoning, However, as the initial diameters obtained in Step 1 are adjusted based on flow velocities in Step 2, the influence of network topology and zoning is accounted for in the overall approach.

3. Methodology

As stated in the Introduction, one of the objectives of this paper is to provide a rigorous assessment of the relative performance of the PHSM compared with that of the KLSM and two sampling methods that do not consider domain knowledge. The flowchart of the process for achieving this is shown in Fig. 3. As can be seen, four different sampling methods, including two heuristic methods (i.e. the PHSM and the KLSM) and two non-heuristic methods (i.e. RS and LHS), are used to obtain initial GA populations. The two non-heuristic sampling methods are considered as they provide a benchmark against which the performance of the two heuristic sampling methods can be assessed. RS is used as this is the conventional method for initializing GA populations and LHS is used as it provides a more structured approach for sampling the solution space. It should be noted that, although there are some other analytical techniques for seeding the initial population of EAs (e.g. Keedwell and Khu, 2006; Zheng et al., 2011a, 2014a,b; Zheng and Zecchin, 2014; Fu et al., 2012), they do not incorporate engineering knowledge and experience directly and hence are not considered in this paper.

Each of the sampling approaches is applied to seven WDSs of varying size and complexity, including the Hanoi, Extended Hanoi, Fosspoly 1, ZJ, Balerma and Rural networks, as well as a modified version of the KL network (KLmod). The networks are optimized for total life cycle costs while satisfying pressure head constraints at each demand node. The hydraulic simulations required to check pressure constraints are performed using EPANET 2.0, as demand-driven modelling is most commonly used in optimization studies, although pressure-driven modelling is likely to be a better alternative under some circumstances (Laucelli et al., 2012). Each of the GA optimization runs is repeated 10 times with different sets of initial solutions and GA operators generated using different random number seeds for each network and sampling method. The results are compared in terms of the best and average solutions found during these ten runs. Details of each of the components of the process are provided in subsequent sections.

3.1. Sampling methods

Details of the KLSM (Method 2, Fig. 3) and the two non-heuristic sampling methods (Methods 3 and 4, Fig. 3) are given below. Details of the PHSM (Method 1, Fig. 3) are given in the previous section.

3.1.1. The KLSM (Kang and Lansey, 2012) (Method 2)

As mentioned previously, in this approach, initial solutions are generated by adjusting pipe diameters to ensure that the velocities in all pipes are less than a pre-set velocity threshold selected from a practical range of velocities for average and peak flows in water supply networks. The heuristic procedure for achieving this is as follows:

Please cite this article in press as: Bi, W., et al., Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.09.010

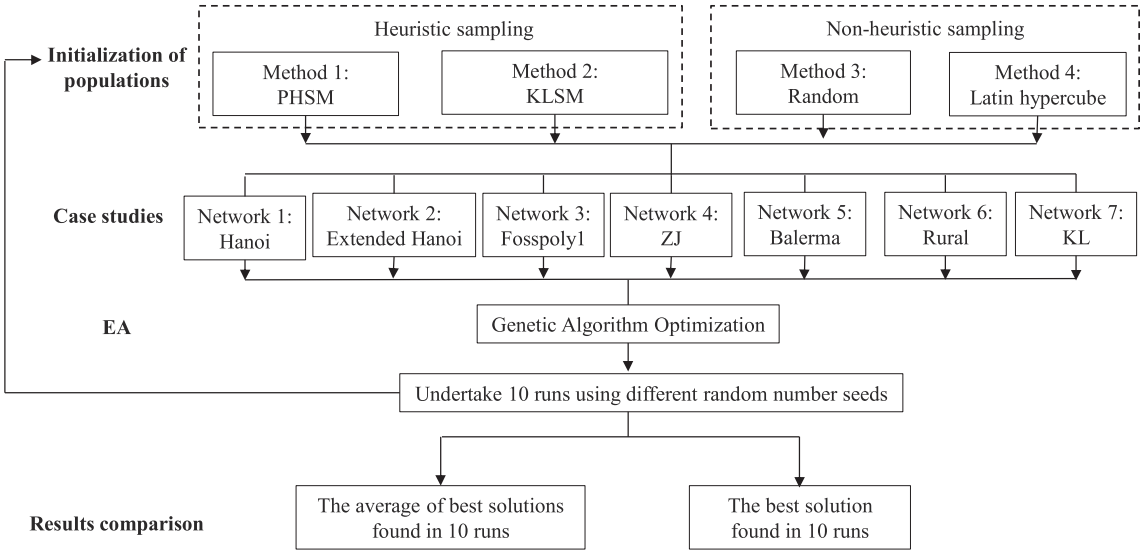

Fig. 3. Flowchart of the assessment process.

(1) All pipes to be optimized are set to the minimum allowable diameter.
(2) A hydraulic simulation is carried out to obtain the flow velocity in each pipe.
(3) The resulting velocity in each pipe is compared with a preset velocity threshold selected from within the range of 0.45–1.5 m/s (e.g. 1 m/s). If the velocity is larger than the threshold, this pipe diameter is increased to the next larger commercial size.

Steps (2) and (3) are performed repeatedly until all velocities in all pipes are below the threshold and the resulting pipes sizes are used to form one solution of the initial population. A number of different initial solutions is generated by varying the value of the velocity threshold within the pre-defined velocity range of 0.45–1.5 m/s. In order to maintain solution diversity, half of the initial solutions are generated using this heuristic method, while the other half are generated randomly. In this study, the velocity thresholds of the KLSM are obtained using the following equation:

$$VT_r = 0.45m/s + r\frac{(1.5m/s - 0.45m/s)}{\frac{1}{2}N} \quad (4)$$

where VT_r (m/s) is the rth ($r = 1,2 \ldots, \frac{1}{2}N$) velocity threshold used for generating the heuristic solutions; N is the total population size.

3.1.2. Random sampling (Method 3)

In random sampling (RS), each diameter option has the same probability of being selected for each pipe within the WDS. When generating a solution, each decision variable (i.e. pipe) is assigned a diameter value that is randomly selected from all available diameter options.

3.1.3. Latin Hypercube Sampling (Method 4)

Latin Hypercube Sampling (LHS) is a type of stratified sampling method that ensures that all portions of the sample space of each variable are sampled (McKay et al., 1979). In this study, Simlab2.2 (JRC, 2008) is used to generate initial solutions using LHS for each case study. A detailed description of the process of LHS can be found in the manual of Simlab2.2 (JRC, 2008).

3.2. Case studies

Details of each case study are given in Table 2. For each case study, the decision variables are the pipe diameters and the objective is to find the minimum cost solution while satisfying the pressure head constraints. Consequently, the optimization problem to be solved can be represented as follows:

Minimize.

$$F = \sum_{j=1}^{m} C_j(D_j) \quad (5)$$

Subject to:

$$H_i^{\min} \le H_i \le H_i^{\max} \quad i = 1, 2\ldots\ldots, n \quad (6)$$

$$G(Hi, \mathrm{D}) = 0 \quad (7)$$

$$D_j \in \{A\} \quad (8)$$

where F is the network cost that is to be minimized; $C_j(D_j)$ is the cost function for pipe $j = 1,2 \ldots, m$ with assigned diameter D_j; m and n are the total number of pipes and demand nodes in the network, respectively; $G(H_i, \mathbf{D}) =$ nodal mass balance and loop (path) energy balance equations for the whole network with pipe combinations of $\mathbf{D} = [D_1, D_2, \ldots .D_m]^T$, which is solved using EPANET2.0; $H_i =$ head at node $i = 1,2 \ldots .,n$; $H_i^{\min}$ and $H_i^{\max}$ are the minimum and maximum allowable head limits at the nodes; and $A =$ a set of commercially available pipe diameters.

As shown in Table 2, the seven case studies vary in size and complexity. Details of each network, including the network layout, the available pipe diameters and the cost of each diameter for the Extended Hanoi and the KLmod network are given in this paper and those of the other case studies are given in the corresponding references in the second column of Table 2. The EPANet input files for these seven networks are provided as supplementary material. The current best known solution for each case study (if available) is presented in the second last column in Table 2. The best known solutions (least-cost solutions) for the Hanoi, Balerma and Rural case studies found by GAs are given in the last column, while no GA solutions can be found in the literature for the other case studies.

Please cite this article in press as: Bi, W., et al., Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.09.010

Table 2
Details of the seven case studies.

Case study	Reference	No. of decision variables[a]	No. of diameter options[b]	Size of total search space	Pressure head constraint	Current best solution	Current best solution found by GAs
Hanoi	Fujiwara and Khang (1990)	34	6	2.86×10^{26}	≥ 30 m	$6.081 million by Reca and Martínez (2006) using GENOME	$6.081 million by Reca and Martínez (2006)
Extended Hanoi	current study	34	10	1×10^{34}	≥ 30 m	-[c]	-[c]
Fosspoly1	Bragalli et al. (2012)	58	22	7.26×10^{77}	≥ 40 m	$0.0291 million by Bragalli et al. (2012) using MINLP	-[c]
ZJ	Zheng et al. (2011a)	164	14	9.23×10^{187}	≥ 22 m	$7.082 million by Zheng et al. (2011a) using NLP-DE	-[c]
Balerma	Reca and Martínez (2006)	454	10	1×10^{454}	≥ 20 m	€1.923 million by Zheng et al. (2011a) using NLP-DE	€2.302 million by Bolognesi et al. (2010)
Rural network	Marchi et al. (2014a,b)	476	15	6.58×10^{559}	≥ 0 m	$ 31.22 million by Marchi et al. (2014a,b) using DE	$ 36.25 million by Marchi et al. (2014a,b)
KLmod network	Adapted from Kang and Lansey (2012)	1274	10	1×10^{1274}	≥ 45 m	–[c]	–[c]

[a] The decision variables are the pipe diameters.
[b] The pipe diameter options for the Extended Hanoi and KL network are given in this paper and those for the other case studies are given in the references provided.
[c] The current best solution is unknown or the network has not been optimized previously using an Evolutionary Algorithm.

The Extended Hanoi case study is developed based on the original Hanoi problem (Fujiwara and Khang, 1990), and has not been used in previous studies. The only difference between the original and Extended Hanoi case studies is the number of available diameters for each pipe, while the other information is the same. As it is acknowledged that infeasible solutions dominate the search space for the Hanoi case study, a larger number of diameter options is included for this case study in order to test the performance of the various sampling methods when dealing with a search space with a larger feasible proportion. For the Extended Hanoi problem, ten pipe diameters, including 12, 16, 20, 24, 30, 40, 50, 60, 70 and 80 inches are available instead of the six smallest diameters from this list that were available for the original Hanoi case study (Fujiwara and Khang, 1990).

The topology of the KLmod network case study is taken from the network used by Kang and Lansey (2012), without consideration of pumps and fire-fighting conditions. For this network, a total of ten diameters, including 150, 200, 300, 400, 500, 600, 700, 800, 900, and 1000 mm are available for all pipes, with the unit costs given in Kadu et al. (2008).

3.3. Genetic algorithm optimization

The description of genetic algorithms (GAs) has been well documented (see e.g. Simpson et al., 1994) and hence, this information is not repeated in this paper. In this study, the GA used integer coding, two-point crossover, bitwise mutation, and tournament selection, as these have been demonstrated to be effective in terms of finding optimal solutions (Deb, 2000; Vairavamoorthy and Ali, 2005; Zheng et al., 2011b). Although a number of different GA variants have been developed over the past four decades in order to improve search performance (Dandy et al., 1996; Nicklow et al., 2010), the use of a relatively standard GA formulation was considered adequate, as the focus of this study is on the evaluation of different methods for obtaining initial GA populations. In addition, all of the sampling approaches considered in this paper can be used in conjunction with any GA variant or other type of EA.

4. Computational experiments

The four sampling methods (i.e. the PHSM, the KLSM, RS and LHS) were used to generate the initial solutions for GAs applied to each of the seven WDS case studies (Fig. 3). The results of GAs seeded using these four sampling methods were compared in terms of objective function value and computational efficiency.

For the PHSM, the value of the initial threshold velocity v used in Step 2 was selected to be 0.1 m/s for all case studies based on the results of preliminary trials with several different values, although variations of this initial value were found to have only a slight impact on the results. It was found that the overall number of simulations required for adjusting pipe diameters in Step 2 was less than 200 for the seven case studies, and hence the maximum number of allowable simulations S_{max} was set to 1000. In Step 3 of the PHSM, a number of different values of a (see Equation (2)) ranging from 0.1 to 2 were tried and $a = 0.5$ was ultimately selected, as it produced slightly better results than other a values. However, as was the case for the initial threshold velocity v, slight variations in a did not significantly influence the final results. For the KLSM, velocity thresholds were generated in accordance with Equation (4).

The parameter values of the GAs applied to each case study were fine-tuned with the aid of a large-scale sensitivity analysis. For the crossover probability, values ranging from 0.1 to 0.9 were tried. For the mutation probability, 10 different values around the value of $1/ND$ (where ND is the number of decision variables) were tried for each study, as it has been demonstrated that a value of approximately $1/ND$ is an effective value and is normally used for GAs (Simpson et al., 1994). The parameter values that exhibited the best performance in terms of efficiently finding good quality optimal solutions were selected and are presented in Table 3. For each case study, the GAs seeded using the four sampling methods considered used the same parameter values. A penalty cost was added to the objective function value for infeasible solutions, with a penalty multiplier of 10^5/metre of head being used for all case studies (Simpson et al., 1994). The tournament size in the selection operator was two for all GAs. The maximum allowable number of evaluations for each case study is given in the last column of Table 3, with the larger networks assigned larger computational budgets.

In order to facilitate easier discussion of the results, the seven case studies were assigned to three groups based on the number of decision variables (ND), as shown in the third column of Table 3. The first three case studies (Hanoi, Extended Hanoi and Fosspoly1) were assigned to G1, as their values of $ND < 100$, while the ZJ, Balerma and Rural network case studies were allocated to G2 with $100 < ND < 500$. The KLmod network was assigned to G3, as its $ND > 500$.

Please cite this article in press as: Bi, W., et al., Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.09.010

Table 3
Parameters values of GAs for each case study.

Case study	Number of decision variables ($_{ND}$)	Network group based on the size of WDSs	Population size ($_{N}$)	Crossover probability	Mutation probability	Total number of evaluations
Hanoi	34	G1 ($ND < 100$)	100	0.9	0.02	300,000
Extended Hanoi	34		100	0.9	0.02	300,000
Fosspoly1	58		500	0.8	0.02	500,000
ZJ	164	G2 ($100 < ND < 500$)	500	0.9	0.006	500,000
Balerma	454		1000	0.9	0.002	1,000,000
Rural network	476		1000	0.8	0.002	1,000,000
KLmod network	1274	G3($ND > 500$)	1000	0.9	0.0008	2,000,000

The performance of each sampling method was assessed using the method outlined below:

1. For each case study, ten GA runs were performed for each of the four sampling methods using different random number seeds, resulting in a total of 40 final optimal solutions.
2. The best final solution from the 40 solutions was selected for each case study and used as a benchmark against which the performance of each sampling method was assessed. This benchmark optimal solution was also compared with the current best known solution in the literature obtained using similar GAs, if available (see Table 2), in order to ensure that the results obtained in the current study are reasonable.
3. For each sampling method, the average of the best solution at each GA generation was calculated for each case study based on the ten runs with different starting random number seeds (denoted ABS). In addition, among the ten best solutions at each generation, the one with the lowest cost was selected (denoted as BBS).
4. The deviation of ABS and BBS from the corresponding benchmark optimal solution was plotted against the number of evaluations for each sampling method. This resulted in four convergence curves on the same plot, enabling a comparison of the performance of the four sampling methods considered.
5. The performance of each sampling method was also assessed in terms of its computational efficiency in being able to find near-optimal solutions. For this purpose, optimal solutions that had objective function values within 5% of the benchmark optimal solution were defined as being near-optimal.

In order to enable a fair comparison between the methods, the computational overheads associated with implementing the proposed PHSM are also considered (Table 4). This was achieved by converting the computational time required for each step of the proposed PHSM (see Section 2) to the equivalent number of network simulations using the same computer configuration (Pentium PC (Inter R) at 3.0 GHz). As shown in Table 4, the proposed PHSM is very efficient in computing the shortest-distance values for the network (Step 1) and generating distribution functions based on the approximate optimal solutions (Step 3), while it is relatively more time-consuming in adjusting pipe diameters based on the velocities in Step 2. This is expected, as this step involves an iterative process (see Fig. 2). The number of equivalent network simulations that correspond to the total computational overhead required by the PHSM method is presented in the last column of Table 4. As can be seen, this computational effort is negligible compared with the total computational budgets used in Table 3, and hence is not considered in the subsequent discussions in Section 5.

5. Results and discussion

The costs of the best solutions found using the GAs initialized with the four sampling methods considered for each of the seven case studies are given in Table 5, with the lowest cost solutions found highlighted in bold. In addition, for the case studies to which GAs had been applied previously in the literature, the percentage deviation of the solutions found in this study compared with the best solution found using GAs reported in the literature are shown in brackets (i.e. negative percentage changes indicate that the solutions found in this study are better and vice versa). It should be noted that the results presented here are compared with those obtained using GAs in previous studies because the purpose of this study is to compare the relative performance of different initial sampling approaches. This requires the impact of the sampling approaches to be isolated from the impact of algorithm searching behavior as much as possible. Consequently, as a GA is used as the EA in this study for reasons outlined previously, the final results obtained in this study should only be compared with those obtained using other GAs.

From Table 5, it can be clearly seen that by using the proposed PHSM, better quality solutions could be found for each case study within the given computational budgets than when the other approaches were used. The KLSM produced better solutions for the

Table 4
Computational overhead analysis for the proposed sampling method (PHSM).

Case study	Number of decision variables ($_{ND}$)	Equivalent simulations of the computational overhead used in step 1[a]	Equivalent simulations of the computational overhead used in step 2[a]	Equivalent simulations of the computational overhead used in step 3[a]	Total computational overhead[b]
Hanoi	34	10	102	1	113 (0.38%)
Extended Hanoi	34	10	126	1	137 (0.46%)
Fosspoly1	58	11	153	1	165 (0.33%)
ZJ	164	6	147	2	155 (0.31%)
Balerma	454	8	165	3	176 (0.18%)
Rural network	476	9	171	3	183 (0.18%)
KLmod network	1274	14	190	4	208 (0.10%)

[a] The computational overhead used for each step has been converted to the equivalent number of network simulations for each case study.
[b] The computational overhead is expressed as the equivalent number of simulations and the fraction of the total computational budget this represents (in brackets).

Please cite this article in press as: Bi, W., et al., Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.09.010

Table 5
Cost of the best solution found by each sampling method for each case study.

Case study	Cost of the best solution found by each sampling method (Million)			
	RS	LHS	KLSM	PHSM
Hanoi	$6.195 *(1.87%)*	$6.217 *(2.24%)*	$6.224 *(2.35%)*	**$6.109** *(0.46%)*
Extended Hanoi	$5.365	$5.366	$5.360	**$5.346**
Fosspoly1	$0.0294	$0.0294	$0.0309	**$0.0290**
ZJ	$7.562	$7.560	$7.655	**$7.431**
Balerma	€2.125 *(-7.69%)*	€2.146 *(-6.78%)*	€2.130 *(-7.47%)*	**€2.061** *(-10.47%)*
Rural	$36.108 *(-0.39%)*	$36.265 *(0.04%)*	$36.255 *(0.01%)*	**$35.173** *(-2.97%)*
KLmod	$8.686	$8.737	$8.418	**$8.307**

Note: The result of each sampling method for each case study was obtained over 10 runs with different random number seeds. The percentage of the cost of each best solution relative to the best solution found by GAs is given in italics in the brackets. The benchmark optimal solution for each case study is indicated in bold.

large KLmod network compared to the other two non-heuristic sampling methods (RS and LHS). This agrees well with the observations made by Kang and Lansey (2012). However, for five of the other six case studies, RS performed better than the KLSM in terms of the quality of the final solutions.

The convergence plots for each of the algorithms for the case studies belonging to the three different groups defined in Table 3, as defined in the previous section, are given in Fig. 4 (G1), 5 (G2) and 6 (G3) and provide an indication of both solution quality and computational efficiency. A common observation is that the PHSM generally found significantly better initial solutions than the non-heuristic sampling approaches. This is most likely because the initial solutions obtained using the PHSM were feasible and the diameters for these solutions were generally in a reasonable range based on velocities, nodal demands and elevations. For the larger case studies, the PHSM also found significantly better initial solutions than the KLSM. The initial solutions obtained using the other sampling methods were typically infeasible for the larger case studies or feasible with high costs for the simpler WDSs. This demonstrates that the proposed domain knowledge based sampling method is effective in identifying good quality starting solutions. A detailed discussion of the results for the three groups of case studies is given in the subsequent sections.

5.1. Group 1 (G1) case studies

As can be seen from Table 5 and Fig. 4, the performance of the GAs initialized with the four different sampling methods is very similar for the G1 case studies (i.e. Hanoi, Extended Hanoi and Fosspoly1), both in terms of the ability to find optimal solutions and computational efficiency. While the GAs initialized with the PHSM were able to find the best solution for all three case studies, the variation in the cost of the best-found solutions was relatively small (Table 5). Similarly, while GAs initialized with the PHSM found better initial solutions and generally converged more quickly than the GAs initialized with the other methods (Fig. 4), this difference was not very large. Consequently, based on the results obtained, there does not appear to be a significant advantage of using domain knowledge for the initialization of GAs for small problems, such as those considered for the G1 case studies.

5.2. Group 2 (G2) case studies

As can be seen from Table 5 and Fig. 5, the performance of the GA initialized with the PHSM is noticeably better than that of the GAs initialized with the other three methods for the G2 (ZJ, Balerma, Rural) networks, both in terms of the best-found solution and computational efficiency. This suggests that while for the simpler G1 case studies the GAs were able to find good solutions relatively quickly with the aid of their evolutionary operators, irrespective of the starting position in solution space, this is not the case for the more complex G2 case studies. This demonstrates that the better starting positions in solution space identified using the PHSM are able to assist the GA with finding better regions of larger search spaces, as indicated by the better solutions found when the GAs were initialized with the PHSM (Table 5 and Fig. 5). This trend was already noticeable for the Fosspoly1 case study, which is the most complex of the G1 case studies (Fig. 4).

The results in Table 5 and Fig. 5 also indicate that the solutions found using the PHSM were not only better than those obtained using RS and LHS, but also better than those obtained using the other heuristic sampling method (i.e. the KLSM). This appears to be both as a result of the quality and diversity of the initial solutions. For example, for the ZJ and Rural networks, the PHSM was able to identify significantly better initial solutions than the KLSM, resulting in more rapid convergence and better final solutions (Fig. 5). In contrast, for the Balerma network, use of the KLSM resulted in better initial solutions than use of the PHSM. However, despite this initial disadvantage, use of the PHSM resulted in more rapid convergence and the identification of better solutions (Fig. 5), which is likely due to the additional control over population diversity offered by the PHSM. A similar trend was also observed for the Fosspoly1 network (Fig. 4), which is the largest of the G1 networks. It should be noted that the better performance of the PHSM was not affected by the presence of multiple source reservoirs, as is the case for the Balerma network (see Reca and Martínez (2006) for network configuration). This suggests that the approach of using an augmented source node for networks with more than one source reservoir (as described in Step 1 for the PHSM) is effective.

In terms of the quality of the solutions found, use of the PHSM resulted in the best solutions for all three G2 case studies by some margin (Table 5, Fig. 5). In contrast, the quality of the solutions found using the other three initialization methods is quite similar, with no advantage of using the KLSM. It should also be noted that for the two case studies to which similar GAs had been applied in previous studies, the GA initialized with the PHSM found solutions that were 10.47% and 2.97% better than those found in previous studies for the Balerma and Rural networks, respectively (Table 5).

As far as convergence speed is concerned, use of the PHSM results in significantly faster convergence to near-optimal solutions (i.e. solutions that are within 5% of the benchmark optimal solution, as defined previously) than use of the other three initialization methods, which all performed similarly (Fig. 5). This indicates that there is likely to be a significant advantage in using the PHSM when trying to find the best possible solution within reasonable computational budgets for complex networks.

5.3. Group 3 (G3) case studies

As can be seen from Table 5 and Fig. 6, for this very large network (i.e. KLmod), the performance of the GAs initialized with both heuristic sampling methods (i.e. PHSM and KLSM) are noticeably better than that of the GAs initialized with the two non-heuristic sampling methods (i.e. RS and LHS), both in terms of the best-found solution and computational efficiency. While the GAs initialized using the two heuristic sampling methods were able to find near-optimal solutions after approximately 800,000 evaluations for the average solutions based on ten runs, which is equivalent to approximately 3 h in terms of CPU time, the GAs initialized with the non-heuristic sampling methods (i.e. RS and LHS) were not even able to find solutions of this quality at the end of the optimization run (using nearly 2,000,000 evaluations and approximately 7 h of CPU time). Although Fig. 6 suggests that the

Please cite this article in press as: Bi, W., et al., Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.09.010

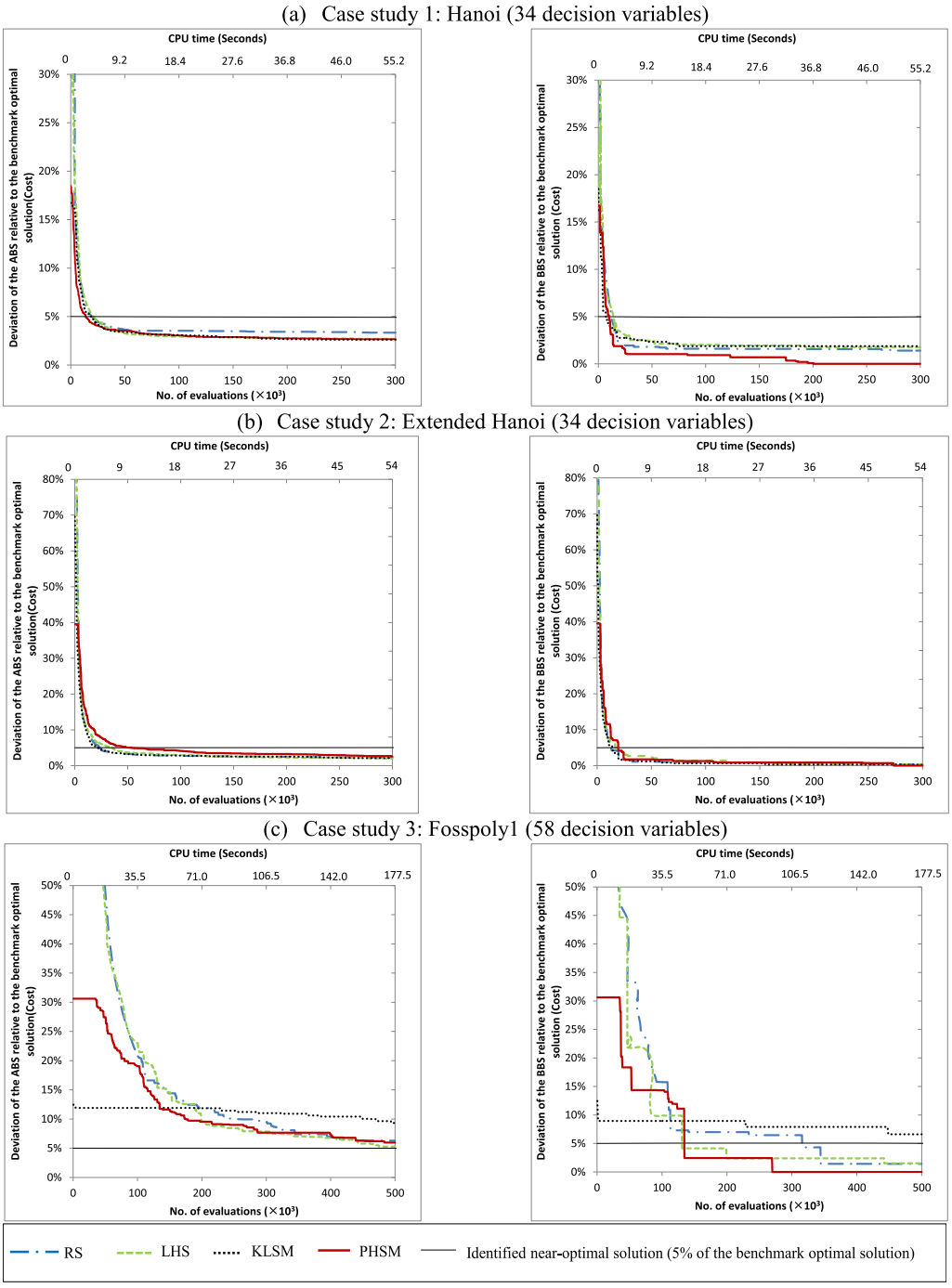

Fig. 4. Results of the GAs with the four sampling methods applied to case studies in Group 1 (G1 in Table 3).

GAs initialized with RS and LHS had not converged yet and might ultimately find solutions of a similar quality to those found when the heuristic sampling methods were used, the computational effort required to do so is likely to be very large. This clearly highlights the advantage of using heuristic sampling methods for initializing GA populations for larger networks.

In terms of the relative performance of the two heuristic sampling methods, while both converged to near-optimal solutions after approximately the same number of iterations, use of the PHSM resulted in clearly better best-found solutions. This is likely to be due to a combination of the better initial solutions identified using the PHSM, as well as the additional control over population diversity afforded by the PHSM. However, the relative performance of the KLSM compared with that of the PHSM was much better for the KLmod case study, which is most likely because the KLSM was designed for a modified version of this problem.

6. Summary and conclusions

In order to improve the ability of GAs to find optimal or near-optimal solutions in reasonable timeframes for realistic-sized water distribution optimization problems, a new heuristic sampling method (the PHSM) for initialising GA populations was introduced and evaluated in this paper. The performance of the PHSM was

Please cite this article in press as: Bi, W., et al., Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.09.010

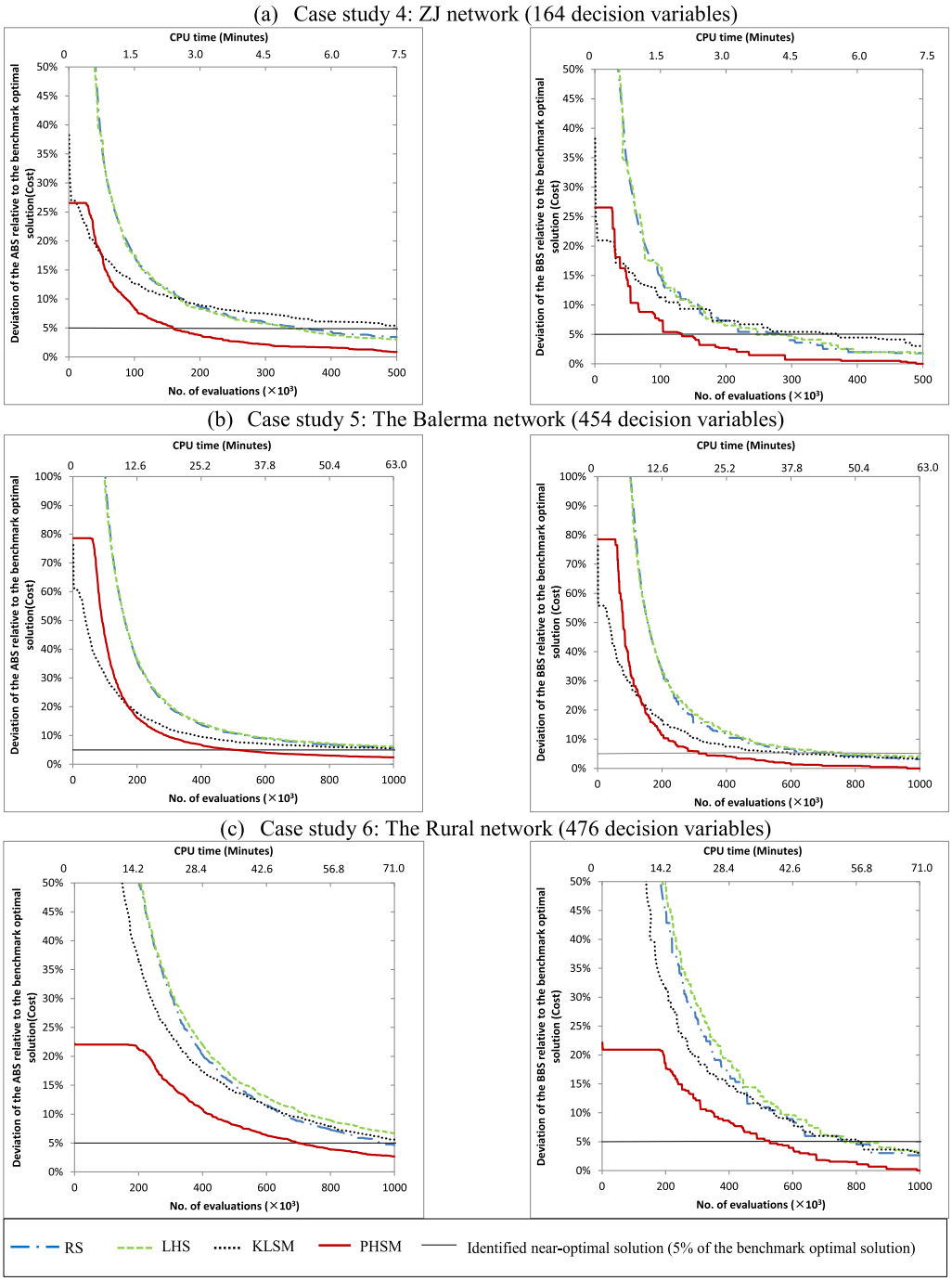

Fig. 5. Results of the GAs with the four sampling methods applied to case studies in Group 2 (G2 in Table 3).

compared with that of an existing heuristic sampling method (the KLSM) and with that of more traditional sampling methods, including RS and LHS, for seven WDSs of varying size and complexity.

The results obtained based on the seven WDS optimization (pipe-sizing) problems considered indicate that overall, the proposed PHSM performed significantly better than the other three sampling methods, both in terms of solution quality and computational efficiency. It was also found that the relative advantage of the PHSM increased with network size and complexity. While for the smaller (G1) networks, the performance of the GAs initialised using the four different methods was very similar, there were clear advantages in using the PHSM for the larger (G2) networks and in using both heuristic sampling methods (i.e. PHSM and KLSM) for the largest network considered (G3). This advantage is likely to be due to the ability to find better initial solutions, enabling more favourable regions of the solution space to be explored more quickly. The results also indicate that PHSM outperforms the KLSM, which is likely due to a combination of the ability to find better initial solutions and the additional population diversity provided by the PHSM.

As the focus of this paper was on the development and evaluation of the PHSM, all analyses were conducted using a reasonably standard GA. However, as the PHSM is independent of the

Please cite this article in press as: Bi, W., et al., Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.09.010

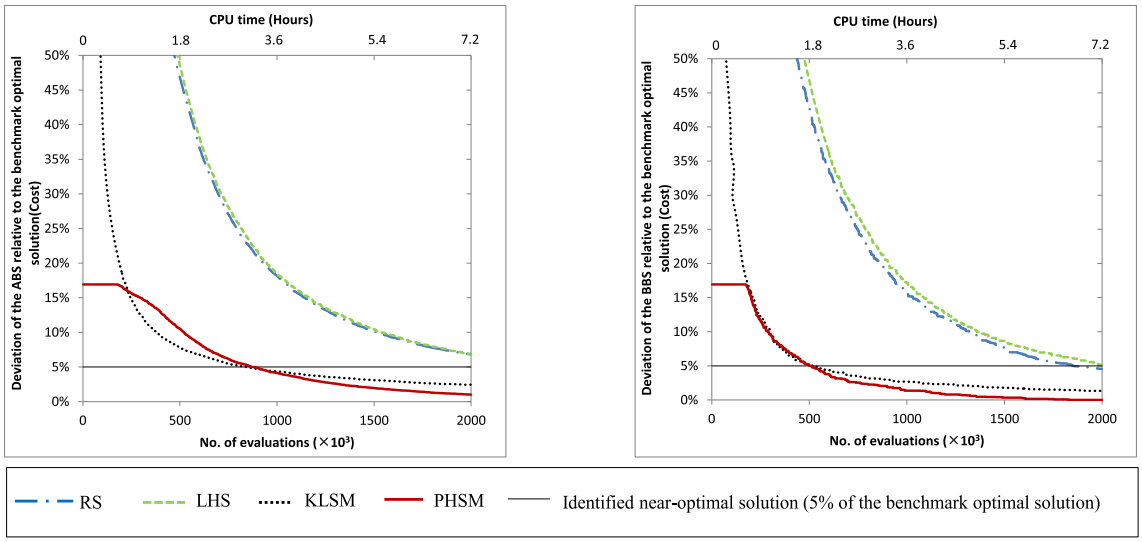

Fig. 6. Results of the GAs with the four sampling methods applied to case studies in Group 3 (G3 in Table 3).

optimization algorithm used, it can be tested in combination with other algorithms. Such investigations would be useful in terms of assessing the generality of the results obtained in this paper. In addition, it would be useful to extend and apply the proposed approach to a larger number of case studies with increased hydraulic complexity, such as the inclusion of tanks, valves and pumps. However, given that pipe sizes generally represent the largest number of decision variables, application of the PHSM to the subset of the decision variables consisting of pipe diameters is still likely to be beneficial for WDSs including tanks, valves and pumps. Finally, it would be interesting to compare the performance of the PHSM with that of other methods that could be used for initialising EAs, such as the cellular automata network design algorithm of Keedwell and Khu (2006).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2014.09.010.

References

Bolognesi, A., Bragalli, C., Marchi, A., Artina, S., 2010. Genetic heritage evolution by stochastic transmission in the optimal design of water distribution networks. Adv. Eng. Softw. 41 (5), 792–801.

Bragalli, C., D'Ambrosio, C., Lee, J., Lodi, A., Toth, P., 2012. On the optimal design of water distribution networks: a practical MINLP approach. Optim. Eng. 13 (2), 219–246.

Broad, D.R., Dandy, G.C., Maier, H.R., 2005. Water distribution system optimization using metamodels. J. Water Resour. Plan. Manag. 131 (3), 172–180.

Broad, D.R., Maier, H.R., Dandy, G.C., 2010. Optimal operations of hydraulically complex water distribution systems using metamodels. J. Water Resour. Plan. Manag. 136 (4), 433–443.

Dandy, G.C., Simpson, A.R., Murphy, L.J., 1996. An improved genetic algorithm for pipe network optimization. Water Resour. Res. 32 (2), 449–458.

Deb, K., 2000. An efficient constraint handling method for genetic algorithms. Comput. Methods Appl. Mech. Eng. 186 (2–4), 311–338.

Deo, N., 1974. Graph Theory with Applications to Engineering and Computer Science. Prentice-Hall, Englewood Cliffs, N. J.

Deuerlein, J.W., 2008. Decomposition model of a general water supply network graph. J. Hyd. Eng. 134 (6), 822–832.

di Pierro, F., Khu, S.-T., Savic, D., Berardi, L., 2009. Efficient multi-objective optimal design of water distribution networks on a budget of simulations using hybrid algorithms. Environ. Model. Softw. 24 (2), 202–213.

Fu, G., Kapelan, Z., Reed, P., 2012. Reducing the complexity of multiobjective water distribution system optimisation through global sensitivity analysis. J. Water Resour. Plan. Manag. 138 (3), 196–207.

Fujiwara, O., Khang, D.B., 1990. A two-phase decomposition method for optimal design of looped water distribution networks. Water Resour. Res. 26 (4), 539–549.

Gibbs, M.S., Dandy, G.C., Maier, H.R., 2008. A genetic algorithm calibration method based on convergence due to genetic drift. Inform. Sci. 178 (14), 2857–2869. http://dx.doi.org/10.1016/j.ins.2008.03.012.

Gibbs, M.S., Maier, H.R., Dandy, G.C., 2010. Comparison of genetic algorithm parameter setting methods for chlorine injection optimization. J. Water Resour. Plan. Manag. 136 (2), 288–291.

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

Gupta, I., Gupta, A., Khana, P., 1999. Genetic algorithm for optimization of water distribution systems. Environ. Model. Softw. 14 (5), 437–446.

Housh, M., Ostfeld, A., Shamir, U., 2013. Limited multi-stage stochastic programming for managing water supply systems. Environ. Model. Softw. 41 (0), 53–64.

JRC, 2008. European Commission Joint Research Center. http://simlab.jrc.ec.europa.eu/.

Kadu, M.S., Gupta, R., Bhave, P.R., 2008. Optimal design of water networks using a modified genetic algorithm with reduction in search space. J. Water Resour. Plan. Manag. 134 (2), 147–160.

Kang, D., Lansey, K., 2012. Revisiting optimal water-distribution system design: issues and a heuristic hierarchical approach. J. Water Resour. Plan. Manag. 138 (3), 208–217.

Keedwell, E., Khu, S.-T., 2006. Novel cellular automata approach to optimal water distribution network design. J. Comput. Civ. Eng. 20 (1), 49–56.

Krapivka, A., Ostfeld, A., 2009. Coupled genetic algorithm–-linear programming scheme for least-cost pipe sizing of water-distribution systems. J. Water Resour. Plan. Manag. 135 (4), 298–302.

Laucelli, D., Berardi, L., Giustolisi, O., 2012. Assessing climate change and asset deterioration impacts on water distribution networks: demand-driven or pressure-driven network modeling? Environ. Model. Softw. 37 (0), 206–216.

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., Reed, P.M., 2014. Evolutionary algorithms and other metaheuristics in water resources: current status, research challenges and future directions. Environ. Model. Softw. http://dx.doi.org/10.1016/j.envsoft.2014.09.013 accepted Sept. 12, 2014.

Marchi, A., Dandy, G., Wilkins, A., Rohrlach, H., 2014a. Methodology for comparing evolutionary algorithms for optimization of water distribution systems. J. Water Resour. Plan. Manag. 140 (1), 22–31.

Marchi, A., Salomons, E., Ostfeld, A., Kapelan, Z., Simpson, A., Zecchin, A., Maier, H., Wu, Z., Elsayed, S., Song, Y., Walski, T., Stokes, C., Wu, W., Dandy, G., Alvisi, S., Creaco, E., Franchini, M., Saldarriaga, J., Páez, D., Hernández, D., Bohórquez, J., Bent, R., Coffrin, C., Judi, D., McPherson, T., van Hentenryck, P., Matos, J., Monteiro, A., Matias, N., Yoo, D., Lee, H., Kim, J., Iglesias-Rey, P., Martínez-Solano, F., Mora-Meliá, D., Ribelles-Aguilar, J., Guidolin, M., Fu, G., Reed, P., Wang, Q., Liu, H., McClymont, K., Johns, M., Keedwell, E., Kandiah, V., Jasper, M., Drake, K., Shafiee, E., Barandouzi, M., Berglund, A., Brill, D., Mahinthakumar, G., Ranjithan, R., Zechman, E., Morley, M., Tricarico, C., de Marinis, G., Tolson, B., Khedr, A., Asadzadeh, M., 2014b. The battle of the water networks II (BWN-II). J. Water Resour. Plan. Manag. http://dx.doi.org/10.1061/(ASCE)WR.1943-5452.0000378.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245.

Nicklow, J., Reed, P., Savic, D., Dessalegne, T., Harrell, L., Chan-Hilton, A., Karamouz, M., Minsker, B., Ostfeld, A., Singh, A., Engineering, E. Z. A. T. C. o. E. C. i. E., and Water, R, 2010. State of the art for genetic algorithms and beyond in water resources planning and management. J. Water Resour. Plan. Manag. 136 (4), 412–432.

Razavi, S., Tolson, B.A., Burn, D.H., 2012. Review of surrogate modeling in water resources. Water Resour. Res. 48, W07401. http://dx.doi.org/10.1029/2011WR011527.

Reca, J., Martínez, J., 2006. Genetic algorithms for the design of looped irrigation water distribution networks. Water Resour. Res. 42 (5), W05416.

Roshani, E., Filion, Y., 2012. Using parallel computing to increase the speed of water distribution network optimization. In: The 14th Water Distribution Systems Analysis Conference, ASCE, Adelaide, South Australia.

Simpson, A.R., Dandy, G.C., Murphy, L.J., 1994. Genetic algorithms compared to other techniques for pipe optimization. J. Water Resour. Plan. Manag. 120 (4), 423–443.

Sitzenfrei, R., Möderl, M., Rauch, W., 2013. Automatic generation of water distribution systems based on GIS data. Environ. Model. Softw. 47 (0), 138–147.

Stokes, C.S., Simpson, A.R., Maier, H.R., 2014. The cost - greenhouse Gas emission nexus for water distribution systems including the consideration of Energy generating infrastructure: an integrated optimization framework and review of literature. Earth Perspect. 1–9. http://dx.doi.org/10.1186/2194-6434-1-9.

Tolson, B.A., Shoemaker, C.A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resour. Res. 43, W01413. http://dx.doi.org/10.1029/2005WR004723.

Tolson, B.A., Asadzadeh, M., Maier, H.R., Zecchin, A., 2009. Hybrid discrete dynamically dimensioned search (HD-DDS) algorithm for water distribution system design optimization. Water Resour. Res. 45 (12), W12416.

Vairavamoorthy, K., Ali, M., 2005. Pipe index vector: a method to improve genetic-algorithm-based pipe optimization. J. Hyd. Eng. 131 (12), 1117–1125.

Walski, T.M., 2001. The wrong paradigm—why water distribution optimization doesn't work. J. Water Resour. Plan. Manag. 127 (4), 203–205.

Wu, Z., Zhu, Q., 2009. Scalable parallel computing framework for pump scheduling optimization. World Environ. Water Resour. Congr. 2009, 1–11.

Wu, Z.Y., Behandish, M., 2012. Comparing methods of parallel genetic optimization for pump scheduling using hydraulic model and GPU-based ANN meta-model. In: The 14th Water Distribution Systems Analysis Conference, ASCE, Adelaide, South Australia.

Zecchin, A.C., Simpson, A.R., Maier, H.R., Marchi, A., Nixon, J.B., 2012. Improved understanding of the searching behavior of ant colony optimization algorithms applied to the water distribution design problem. Water Resour. Res. 48 (9), W09505.

Zhang, W., Chung, G., Pierre-Louis, P., Bayraksan, G., Lansey, K., 2013. Reclaimed water distribution network design under temporal and spatial growth and demand uncertainties. Environ. Model. Softw. 49 (0), 103–117.

Zheng, F., Simpson, A.R., Zecchin, A.C., 2011a. A combined NLP-differential evolution algorithm approach for the optimization of looped water distribution systems. Water Resour. Res. 47 (8), W08531.

Zheng, F., Simpson, A.R., Zecchin, A.C., 2011b. Dynamically expanding choice-table approach to genetic algorithm optimization of water distribution systems. J. Water Resour. Plan. Manag. 137 (6), 547–551.

Zheng, F., Simpson, A.R., Zecchin, A.C., 2013a. A decomposition and multistage optimization approach applied to the optimization of water distribution systems with multiple supply sources. Water Resour. Res. 49 (1), 380–399.

Zheng, F., Zecchin, A., Simpson, A., Lambert, M., 2013b. Non-crossover dither creeping mutation genetic algorithm for pipe network optimization. J. Water Resour. Plan. Manag. http://dx.doi.org/10.1061/(ASCE)WR.1943-5452.0000351.

Zheng, F., Simpson, A.R., Zecchin, A.C., Deuerlein, J.W., 2013c. A graph decomposition-based approach for water distribution network optimization. Water Resour. Res. 49 (4), 2093–2109.

Zheng, F., Simpson, A., Zecchin, A., 2014a. Coupled binary linear programming–differential evolution algorithm approach for water distribution system optimization. J. Water Resour. Plan. Manag. 140 (5), 585–597.

Zheng, F., Zecchin, A., 2014. An efficient decomposition and dual-stage multiobjective optimization method for water distribution systems with multiple supply sources. Environ. Model. Softw. 55 (0), 143–155.

Zheng, F., Simpson, A., Zecchin, A., 2014b. An efficient hybrid approach for multiobjective optimization of water distribution systems. Water Resour. Res. 50 (5), 3650–3671.

Please cite this article in press as: Bi, W., et al., Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.09.010