Diagnostic benchmarking of many-objective evolutionary algorithms for real-world problems

Zatarain Salazar, Jazmin; Hadka, David; Reed, Patrick; Seada, Haitham; Deb, Kalyanmoy

DOI
10.1080/0305215X.2024.2381818

Publication date
2024

Document Version
Final published version

Published in
Engineering Optimization

Citation (APA)
Zatarain Salazar, J., Hadka, D., Reed, P., Seada, H., & Deb, K. (2024). Diagnostic benchmarking of many-objective evolutionary algorithms for real-world problems. *Engineering Optimization*, *57*(1), 287-308. https://doi.org/10.1080/0305215X.2024.2381818

Engineering Optimization

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/geno20

Diagnostic benchmarking of many-objective evolutionary algorithms for real-world problems

Jazmin Zatarain Salazar, David Hadka, Patrick Reed, Haitham Seada & Kalyanmoy Deb

To cite this article: Jazmin Zatarain Salazar, David Hadka, Patrick Reed, Haitham Seada & Kalyanmoy Deb (22 Aug 2024): Diagnostic benchmarking of many-objective evolutionary algorithms for real-world problems, Engineering Optimization, DOI: 10.1080/0305215X.2024.2381818

To link to this article: https://doi.org/10.1080/0305215X.2024.2381818

ENGINEERING OPTIMIZATION
https://doi.org/10.1080/0305215X.2024.2381818

Taylor & Francis
Taylor & Francis Group

Diagnostic benchmarking of many-objective evolutionary algorithms for real-world problems

Jazmin Zatarain Salazar[a,c], David Hadka[b], Patrick Reed[c], Haitham Seada[d] and Kalyanmoy Deb[d]

[a]Faculty of Technology Policy and Management, Delft University of Technology, Delft, The Netherlands; [b]Microsoft, Charlotte, NC, USA; [c]Department of Civil and Environmental Engineering, Cornell University, Ithaca, NY, USA; [d]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

ABSTRACT

Despite progress in multiobjective evolutionary algorithms (MOEAs) research, their efficacy in real-world scenarios remains unclear. This article introduces a diagnostic benchmarking framework to evaluate MOEAs, comprising (1) flexible MOEA construction software, (2) performance evaluation metrics and (3) real-world applications for benchmarking, reflecting diverse mathematical challenges. Utilizing this framework, NSGA-II, NSGA-III, RVEA, MOEA/D and Borg MOEA were evaluated across four applications with three to ten objectives. Collectively, the four applications capture challenges such as stochastic objectives, severe constraints, nonlinearity and complex Pareto frontiers. The study demonstrates how MOEAs that have shown strong performance on standard test problems can struggle on real-world applications. The benchmarking framework and results have value for enhancing the design and use of MOEAs in real-world applications. Further, the results highlight the need to improve the adaptability and ease-of-use of MOEAs given the often ill-defined nature of real-world problem-solving.

ARTICLE HISTORY
Received 4 July 2024
Accepted 7 July 2024

KEYWORDS
Many objective evolutionary algorithm; optimization; benchmarking; diagnostics

1. Introduction

Remarkable progress has been made in building well-defined test problems to assess the efficiency and effectiveness of multiobjective evolutionary algorithms (MOEAs). These test problems, typically comprising a benchmark suite, are designed to reflect challenging problem characteristics, including but not limited to multimodality, deception, isolated optima, non-convexity, discreteness, non-uniformity, non-separability and scalability in the number of decision variables and/or objectives (Deb 1999).

Efforts began with the development of the ZDT test suite (Zitzler, Deb, and Thiele 2000), and were quickly followed by the DTLZ suite (Deb *et al.* 2001; Deb, Thiele, *et al.* 2002) and the WFG suite (Huband *et al.* 2006), each introducing new test problems with challenging characteristics. The DTLZ suite introduced the idea of scalable problems, where the same problem can be instantiated with a different number of objectives, enabling the systematic study of many-objective optimization. The WFG suite introduced the idea of formulating problems with distinct characteristics by applying a series of transformations, which the authors called shape and transition functions, which remain popular benchmarking suites currently used. Since then, the literature has continued to evolve with

CONTACT Jazmin Zatarain Salazar J.ZatarainSalazar@tudelft.nl

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/0305215X.2024.2381818.

Table 1. Multiobjective benchmark suites.

Test suite	Objectives	Properties	Year	Reference
ZDT	2	Unconstrained	2000	Zitzler, Deb, and Thiele
DTLZ	2+	Unconstrained	2002	Deb, Thiele, *et al.*
WFG	2+	Deceptive, non-separable	2002	Deb, Thiele, *et al.*
LZ	2–3	Complicated Pareto sets	2009	H. Li and Zhang
CEC2009	2–5	Unconstrained and constrained	2009	Zhang, Liu, and Li
CDTLZ	2+	Constrained	2014	Deb and Jain
LSMOP	2–10	Large number of decision variables	2017	Cheng, Jin, *et al.*
MAF	2–10	Irregular Pareto fronts	2017	Cheng, Li, *et al.*
Generative benchmarking	2–30	Scalable and customizable Pareto fronts	2020	Meneghini *et al.*; Wang *et al.*
Visualizable test problems	10	Scalable and customizable Pareto fronts	2017	Fieldsend *et al.*; M. Li *et al.*
Visualizable test problems	10	Scalable and customizable Pareto fronts	2017	Fieldsend *et al.*; M. Li *et al.*
Real-world problems	2–9	Pareto front shapes, and types of design variable	2020	Tanabe and Ishibuchi
Box-constrained real-world problems	2–7	Applied engineering applications	2023	Zapotecas-Martínez, García-Nájera, and Menchaca-Méndez

new additions to the benchmarking toolbox. For example, the LZ suite leverages complex Pareto sets to create even more challenging problems (H. Li and Zhang 2009), while the CDTLZ suite offers constrained and scalable test problems (Deb and Jain 2014), the modified-DTLZ (mDTLZ) suite was designed to circumvent the regularly-oriented Pareto front shapes and the single distance functions across the objectives from the original DTLZ suite (Wang, Ong, and Ishibuchi 2018). The CEC2009 suite, which was introduced at a workshop during the CEC 2009 conference, presented both constrained and unconstrained multiobjective test problems (Zhang, Liu, and Li 2009). The large-scale multiobjective and many-objective test suite (LSMOP) focuses on testing multiobjective and many-objective optimization problems that have a high number of decision variables (Cheng, Jin, *et al.* 2017). Meanwhile, the MAF test suite was designed for high-dimensional problems with irregular Pareto fronts (Cheng, Li, *et al.* 2017). Recently, generative benchmarking offers customization by allowing researchers to adjust a set of parameters controlling the problem characteristics. Particularly, multimodality, deceptivity and features can be observed in the search spaces (Wang *et al.* 2018). Other frameworks bring together characteristics from common test suites (Meneghini *et al.* 2020). In other cases, the construction of visualizable test problems introduced a class of multi-objective test problems scalable in the number of objectives called multi-line distance minimization problems (ML-DMP). In this test suite, the Pareto optimal solutions lie in a regular polygon in a 2-D decision space, and these solutions are similar in their distribution of the objective vector set (*e.g.* its uniformity and coverage over the Pareto front) via observing the solution set in the 2-D decision space (Fieldsend *et al.* 2021; M. Li *et al.* 2017). Both Tanabe and Ishibuchi (2020) and Zapotecas-Martínez, García-Nájera, and Menchaca-Méndez (2023), aim to bridge the gap between theoretical performance and practical applicability of MOEAs, but they differ in their approach. Tanabe and Ishibuchi (2020) discuss the challenges MOEAs face with both regular and irregular Pareto fronts, implying a diverse set of applications. Zapotecas-Martínez, García-Nájera, and Menchaca-Méndez (2023) model well-defined optimization problems across various practical engineering domains. These benchmarks are summarized in Table 1. In contrast, the primary motivation of this study is to ensure that the benchmarking framework reflects the complex, stochastic and often ill-defined nature of real-world problems. This complexity includes severe constraints, nonlinearity and complex Pareto frontiers. An MOEA benchmarking framework is presented where MOEAs can easily and rigorously be evaluated for real-world studies.

While the benchmarks in Table 1 capture challenging characteristics, not all of them replicate real-world complexity.

Despite MOEAs' potential in unbiased decision support (Eiben and Smith 2015), a gap exists between benchmark test problems and real-world applications. Studies in water resources, with diverse complexities, reveal MOEAs often struggle (Gupta *et al.* 2020; Reed and Kollat 2013; Zatarain-Salazar *et al.* 2016).

To address this, the proposed benchmarks mirror real-world complexity, offering control, reliability and efficiency for uncertain fronts. These studies should evaluate algorithmic architectures and parametric sensitivities.

The chosen applications encompass domains such as automotive safety, environmental management and aerospace engineering, demonstrating the versatility and robustness of MOEAs across diverse real-world challenges. The article offers a rigorous and comprehensive framework for benchmarking MOEA performance on real-world problems. While the four problems highlighted are not exhaustive, they capture a range of complexities and mathematical characteristics, and they pave the way for emerging MOEA applications in multidisciplinary design optimization (MDO) and environmental systems management (ESM).

To begin, a three-objective car crash problem is considered that is smaller in scale but less common in real-world scenarios compared to two-objective problems. This problem offers a lower level of difficulty in terms of its objective count and mathematical traits, making it a valuable lower-bound benchmark that modern MOEAs should readily address. It also has a strong legacy connection to the MDO engineering literature. The two stochastic ESM problems, the Lake Problem and the Lower Rio Grande Valley (LRGV) problem, are valuable for MOEA benchmarking owing to the rapid growth in applications in challenging environmental and water resources problems. The Lake Problem represents a growing class of social-ecological issues involving societal management of uncertain environmental tipping points. Similarly, the LRGV problem captures the mathematical challenges common to environmental portfolio and investment problems in major infrastructure, which are increasingly relevant for planning under climate change uncertainty. Lastly, the general aviation aircraft (GAA) problem is a well-known and severely challenging MDO benchmark. It provides a valuable test to assess the state of the field for modern MOEAs in solving high-objective count, complex real-world design problems, thus strongly challenging modern MOEAs.

The framework is in line with the increasing recognition that the suitability of MOEAs should be taken outside the scope of test problems and into the sphere of intended users and disciplines (Tanabe and Ishibuchi 2020; Zapotecas-Martínez, García-Nájera, and Menchaca-Méndez 2023). It is essential to have robust and diverse benchmarks that accurately reflect the complexity of real-world problems. Such benchmarks will provide a more comprehensive evaluation of MOEAs' potential for decision support contexts, reducing the risk of algorithmic overfitting and parametric bias.

In MOEA benchmarking, overfitting occurs when evaluation narrows down to specific algorithm architectures and a limited set of test problems. This narrow focus can lead to an algorithm that excels on these selected benchmarks but struggles to adapt to a wider range of real-world problems. In essence, the algorithm becomes too specialized and lacks versatility beyond artificial test scenarios. This approach is not conducive to demonstrating the algorithm's value across diverse real-world challenges.

Giagkiozis and Fleming (2015) point out that a strict emphasis on Pareto ranking in MOEA algorithm design, such as through decomposition or reference points, may overlook broader challenges. For instance, while innovations like self-adaptive or meta-heuristic multi-operator search have been addressed in some notable works (Burke *et al.* 2013; Hadka and Reed 2013; Santiago *et al.* 2019; Vrugt, Robinson, and Hyman 2008; Zhou *et al.* 2019), there's an ongoing opportunity for significant research and development in this area.

Parametric bias in MOEAs is a well-documented concern in the literature (Hadka and Reed 2012; Purshouse and Fleming 2007). However, many practitioners still assume that default parameter recommendations from the literature will suffice for achieving optimal performance. Experimental studies have demonstrated that the ideal parameterization, often referred to as the 'sweet spot' (Goldberg 2002a), for a given MOEA can vary significantly or may not even exist across different artificial

test problems. Moreover, these studies reveal that strong non-separable parametric sensitivities can lead to unpredictable changes in the 'sweet spot' as the number of objectives increases in scalable problems. Relying on recommended parameters derived solely from test problems can result in significant MOEA search failures when applied to real-world applications, substantially reducing their practical utility.

It has long been recognized that no single performance metric is sufficient for fully characterizing the success or failure of MOEAs (Knowles and Corne 2002; Zitzler *et al.* 2003). As noted by Ishibuchi, Pang, and Shang (2022), there are several difficulties for fairly comparing MOEAs, where a careful selection of metrics is required to provide a range of characteristics. For a single parameter MOEA trial run, Hadka and Reed (2012) recommend a mixture of metric diagnostics including generational distance (GD) for a pure proximity measure (Deb and Jain 2002), hypervolume for proximity and diversity, and an ϵ-indicator to identify gaps in approximation sets (*i.e.* consistency with the Pareto frontier) (Zitzler *et al.* 2003). Nonetheless, overfitting and parametric bias artifacts are reinforced by traditional assessments of performance that typically employ a single MOEA parameter set over multiple random seed trials to compute performance metrics. This traditional assessment approach provides a highly local and limited representation of the broader global joint probabilistic performance of an algorithm across proximity, diversity and consistency measures. A broader ensemble-based MOEA diagnostic framework (Hadka and Reed 2012) has demonstrated the importance of globally sampling MOEA parameter spaces and exploiting ensemble-based metric assessments. The assessment seeks for ideal traits for MOEAs: (1) effectiveness; (2) efficiency; (3) reliability; and (4) controllability. In simple terms, a useful MOEA should attain high quality approximation sets ('effectiveness') for a real-world application using the minimum number of function evaluations (NFEs, 'efficiency'), with minimal attainment variability ('reliability') and not be sensitive to algorithmic parameter settings ('controllability'). These traits are particularly important in real-world contexts where user time has real costs and investments in tailoring algorithms compete directly with realized returns from analysis of the actual problems of focus (Woodruff, Reed, and Simpson 2013; Woodruff, Simpson, and Reed 2015).

In light of these observations, it is of paramount importance to assess the performance of MOEAs in real-world applications. While the literature contains numerous case studies where one or more MOEAs are applied to solve real-world problems, aggregating results across these studies proves challenging owing to variations in testing methodologies. Consequently, this study makes a significant contribution by introducing and demonstrating a comprehensive MOEA benchmarking framework that serves the following three key purposes.

(1) Flexible MOEA Architectures. The framework allows for the flexible construction of MOEA architectures.
(2) Global Parameter Sampling. It supports rigorous global sampling of MOEA parameters and employs metric-based evaluations that encompass proximity, diversity and consistency.
(3) Real-World Benchmark Applications. The framework offers an extensive collection of real-world benchmark applications that span a range of mathematical problem difficulties.

This benchmarking framework aims to assist the field in developing best practices, establishing a standardized methodology for running MOEAs, collecting results, computing performance metrics and statistically comparing joint probabilistic performance results.

The study showcases the benchmarking framework through four applications, ranging from a three-objective car crash problem to a ten-objective general aviation aircraft design problem. The framework also employs five popular MOEAs, *i.e.* NSGA-II, NSGA-III, MOEA/D, RVEA and the Borg MOEA, specifically selected for their utilization of various mechanisms for many-objective search, such as Pareto dominance, reference directions, decomposition and ϵ-dominance with adaptive operators. By using this benchmark suite, the study rigorously explores each algorithm's performance across its feasible parameter space. This approach eliminates parametric bias when comparing

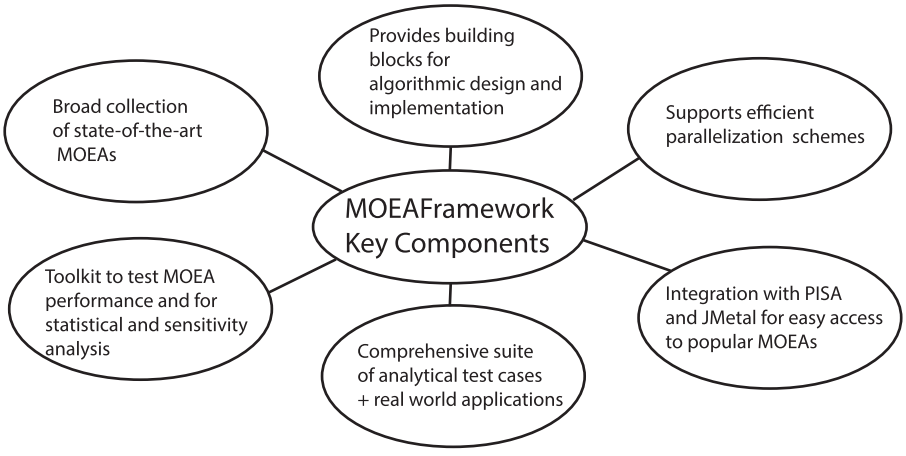

Figure 1. The study introduces and builds upon the MOEA Framework, which provides robust support for a wide range of state-of-the-art MOEAs and serves as a foundation for developing new algorithmic designs. The framework offers an extensive collection of search operators suitable for various decision variable representations. Moreover, it allows for algorithmic parallelization, enabling distribution across multiple computing cores, and seamlessly integrates with the PISA and JMetal platforms for convenient access to popular MOEAs. For detailed information about the repository used in this study, please refer to Hadka (2023).

MOEAs and allows for an in-depth examination of key parametric controls. Moreover, the diverse selection of algorithms captures a wide range of algorithmic architectures.

Section 2 outlines the experimental framework and methodology for evaluating MOEAs. It begins with an overview of the MOEA Framework, highlighting its support for various algorithm architectures and performance diagnostics. The section then details the performance metrics used for assessment, describes the specific MOEAs employed in the study and discusses the real-world benchmarks utilized.

Section 3 presents the results of applying these MOEAs to real-world benchmarks. It emphasizes the contributions of different MOEAs to the best-known reference set and evaluates a range of performance metrics, including convergence, diversity and consistency. Attainment plots are used to measure the reliability of achieving good solutions consistently. The section also examines the algorithms' controllability, analysing the efficiency and sensitivity of MOEAs to parameter variations. Finally, runtime dynamics are assessed to evaluate variability across different seed trials, indicating sensitivity to initial conditions.

Finally, Section 4 presents the conclusions drawn from evaluating the MOEAs in real-world applications, providing insights into effective algorithmic design choices to enhance adaptability across various domains. It also proposes the benchmarking framework as a tool for future algorithmic improvements and encourages collaborative research within the evolutionary computation community.

2. Method

2.1. Experimental framework

Rigorous MOEA benchmarking for real-world applications requires a flexible software framework that can be used to accommodate both common and new algorithm architectures while also supporting ensemble-based performance diagnostics. These features are provided in the open-source MOEA Framework (Hadka 2024). A summary of the key features of the MOEA Framework is shown in Figure 1.

As part of the study, several extensions to the MOEA Framework are made to incorporate real-world benchmarks, allowing distinct MOEAs' capabilities to be assessed across various research

domains, as well as the extension to recent MOEAs. Additionally, the framework offers a comprehensive sensitivity analysis library, facilitating rigorous MOEA diagnostics.

The MOEA Framework is an open-source library developed in Java. It utilizes object-oriented design principles to provide a flexible and extensible framework for creating, testing and distributing MOEAs. At its core, the MOEA Framework defines several key classes and interfaces—including algorithm, problem, variable, solution and variation—that are fundamental for implementing MOEAs. This serves to separate the task of defining specific MOEAs from problem representation and genetic operators.

The MOEA Framework supports a plug-and-play system of adding implementations of both new and already existing algorithms. When adding new algorithm implementations, the MOEA Framework does not require modifications or re-building the framework itself. A newly compiled algorithm implementation simply requires specification of an appropriate path and the MOEA Framework will capture it for full diagnostic analyses automatically. Also, where applicable, the MOEA Framework attempts to determine the appropriate operators to use automatically, based on the decision variable type. For example, when given a real-valued problem, the MOEA Framework will instantiate the MOEA with real-valued variation operators, or report an error if an invalid combination was requested by the user. Again, this provides separation of concerns by providing a clear demarcation between the search algorithm and the problem representation.

The MOEA Framework exploits the service provider interface mechanism for introducing new optimization problems where it will automatically configure MOEAs given the problem properties. It provides default implementations of the most popular MOEAs, including NSGA-II, NSGA-III and MOEA/D. Given the framework's object-oriented design, it is easy to build new MOEAs using existing components or adapt existing MOEA implementations. The MOEA framework has built-in support for global MOEA parametric sampling studies that include many popular performance indicators, including hypervolume (HV), generational distance (GD), inverted generational distance (IGD), additive epsilon-indicator (EI), spacing and the R indicators (Coello Coello, Lamont, and Van Veldhuizen 2007; Zitzler *et al.* 2003). As a result, it contains all the necessary modules to design, test and analyse MOEAs.

The experimental methodology adopted in this study builds on the work of Hadka and Reed (2012). Their methodology is in accordance with the following five steps, and is illustrated in Figure 2.

(1) Generate a global sample of an MOEA's parameters. The algorithmic parameters for each MOEA are statistically sampled via a statistical design of experiments (*e.g.* Latin hypercube sampling) to avoid any parameter bias. Additionally, by testing an MOEA across its feasible parameter space, the 'sweet spots' can be identified, where top performance is expected, observe how these 'sweet spots' change across applications, and compare these 'sweet spots' against the performance of an algorithm's recommended default parameter settings.
(2) Initiate replicate MOEA trial runs for each parameterization. Each parameterization of the MOEA is run for multiple random seed trials on the problem of focus and the resulting approximation sets are recorded. In this study, 50 random seed trial runs for each sampled MOEA parameterization were run for 100,000 NFEs. Search progress was recorded by outputting approximation sets every 10,000 NFEs.
(3) Generate the best-known reference sets. For real-world applications, the true Pareto front is not known a priori. Instead, the best-known approximation to the Pareto front is developed across all runs of all algorithms for a given problem. Likewise, an algorithm-specific best approximation reference set can be attained by ranking the combined results from its individual runs.
(4) Computing performance metrics. In this study, generational distance is used to measure proximity to the reference set, hypervolume to measure both proximity and diversity, and additive epsilon-indicator to measure consistency (no large solution gaps in approximation sets).

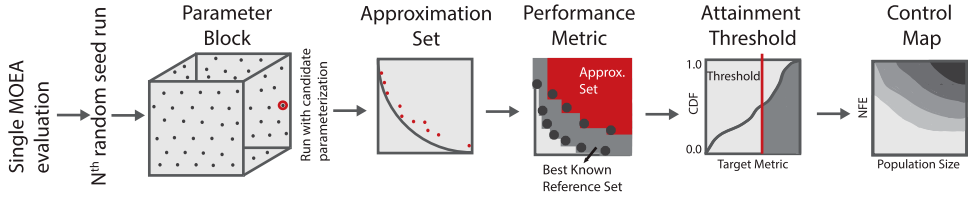

Figure 2. Latin hypercube sampling (LHS) is employed to sample the feasible parameter space for each MOEA. Each point within the parameter block represents a complete set of parameters for generating the corresponding approximation set. This process is repeated for each of the 50 seeds to account for the impact on initial populations. To assess the performance of MOEAs comprehensively, generational distance, epsilon-indicator and hypervolume are computed relative to the best-known reference set (depicted in grey). This reference set embodies the best solutions across all parameterizations, seeds and MOEAs. The probability of attainment measures the percentage of single-seed MOEA runs that achieve good performance, offering insights into algorithmic reliability. Lastly, the control map evaluates the MOEA's sensitivity to its parameterizations, avoiding any specific assumptions about highly-tuned parameters for each algorithm. (Inspiration was taken from Reed and Kollat 2013.)

(5) Generate probabilistic attainment and control figures. Attainment plots show the best overall single MOEA metrics attained as well as the probabilities for attaining higher levels of performance.
(6) Control maps provide a visual means of understanding how sensitive an MOEA's search performance is to its parametric settings (*i.e.* controllability or ease of use).

This process is repeated for each MOEA under study. Parallelization and high-performance computing can be used to execute the type of real-world diagnostics demonstrated in this study efficiently. For example, this study required approximately four days of computing resources with 256 processing cores.

2.2. Multiobjective evolutionary algorithms

In this study, five representative MOEAs were investigated: NSGA-II, NSGA-III, RVEA, MOEA/D and the Borg MOEA. With the exception of NSGA-II, which serves to establish a historical MOEA baseline in performance, these algorithms each have unique mechanisms for handling many objectives. NSGA-III and RVEA utilize reference points/vectors to direct the search towards diversified solutions, MOEA/D uses decomposition, and the Borg MOEA uses its epsilon-dominance archiving to inform auto-adapted changes in selection, population sizing and mixtures of variational operators. Summaries for each algorithm are provided below, and readers are directed to the algorithms' original cited articles for in-depth details. The intent of this diagnostic assessment is to demonstrate how the default or most common versions of these MOEAs perform on the real-world test problem suite.

2.2.1. NSGA-II

In 2002, Deb *et al.* proposed NSGA-II (Deb, Pratap, *et al.* 2002). With a core focus on non-dominated solutions, NSGA-II sorts all population members into a sequence of fronts. Members belonging to each front are non-dominated with respect to each other and are given the same rank. The lower the rank the better. It is worth mentioning that a lower rank solution (A) does not necessarily dominate a higher rank solution (B). If this is the case, there must be at least a single third solution (C) that dominates (B) and whose rank is equal to or higher than (A). After sorting, solutions are incorporated into the next generation, front by front. Typically, the algorithm will reach a front that has more members than the remaining slots in the next population. In such a case, NSGA-II employs a crowding distance operator. Among those solutions awaiting to be incorporated, priority is given to those found in sparse regions of the objective space. NSGA-II is an elitist algorithm. It also uses minimum and maximum values of the current non-dominated set of solutions to normalize objectives.

2.2.2. MOEA/D

Because of its ability to target both convergence and diversity in bi-objective problems, NSGA-II dominated the field for many years. A major drawback of NSGA-II is the performance of its crowding distance operator beyond two objectives. For three and more objectives, NSGA-II struggles to maintain a set of well-distributed solutions over the non-dominated front. In 2007, Zhang and Li sought to address this limitation by introducing their decomposition-based multiobjective optimization algorithm, MOEA/D (Zhang and Li 2007; Zhang, Liu, and Li 2009). One of the most appealing ideas in MOEA/D is the use of reference directions in optimization. MOEA/D breaks down a multiobjective optimization problem into several single-objective optimization subproblems, each using information only from its neighbours. All objective functions are combined to form the fitness functions of these single objective subproblems. This combination (scalarization) is done using Tchebycheff (Miettinen 2012) or boundary intersection (BI) (Das and Dennis 1998) approaches. In contrast to NSGA-II, MOEA/D does not use non-dominated sorting as its main pulling force. It rather tries to minimize the distance between each solution and a reference point (convergence) while minimizing the distance between this solution and its closest reference direction (diversity) at the same time. MOEA/D has been shown to achieve better distributions for many-objective test function suites. The algorithm uses a parameter T to control the size of the neighbourhood, as well as another penalty parameter θ used to penalize infeasible solutions in the optimization process. The original study did not include any constraint-handling strategy. In this study, the MOEA/D-DE algorithm as described by H. Li and Zhang (2009) is used, which incorporates a penalty function based on the sum of constraint violations following the method outlined in Jan and Zhang (2010).

2.2.3. NSGA-III

In 2013, Deb and Jain proposed NSGA-III (Deb and Jain 2014; Jain and Deb 2014) building on top of both NSGA-II and MOEA/D. NSGA-III was developed from a many-objective perspective, and was able to solve problems having up to 20 objectives (Deb and Jain 2014). It employs the same non-dominated sorting as in NSGA-II, but uses a different niching strategy. It uses the idea of reference directions only in niching where reference directions are only used to achieve better diversity since the main pulling force is non-dominated sorting. Unlike MOEA/D, NSGA-III does not impose any neighbourhood restrictions. It also uses a different normalization approach, where the extreme solutions are identified using an achievement scalarization function (ASF) approach. Then the hyperplane containing these extreme points is calculated and used to normalize all other solutions. Selection pressure in NSGA-III is mild compared to NSGA-II. The algorithm only favours feasible over infeasible solutions; however, it flips a coin if two feasible solutions go head to head. This is intended to give the algorithm more time for exploration, preventing the potential rush into local traps. The original study used a set of predefined evenly distributed reference directions, and briefly explored the idea of repositioning these directions during optimization.

2.2.4. RVEA

Cheng *et al.* (2016) proposed RVEA, another reference-direction-based multiobjective optimization algorithm. A distinctive feature of RVEA is its scalarization approach, known as angle-penalized distance (APD), which combines the distance between a solution and the ideal point (as a measure of convergence) with the acute angle the solution makes with its reference direction (as a measure of diversity). RVEA looks at normalization from an opposite perspective. Instead of normalizing solutions, RVEA repositions reference directions to achieve a better distribution when dealing with differently scaled objectives. Two additional parameters are used, namely α, which controls the penalty function, and f_r, which controls the frequency of employing the adaptation strategy (normalization). RVEA shows results competitive with several state-of-the-art algorithms on solving problems with up to 10 objectives.

2.2.5. Borg MOEA

The Borg MOEA utilizes auto-adaptive selection pressure, population sizing and multioperator recombination to search for Pareto optimal solutions efficiently in high-dimensional spaces (Hadka and Reed 2013). This algorithm is characterized by its epsilon-dominance archiving, which helps maintain a diverse set of solutions on the Pareto frontier, while its auto-adaptive search dynamically adjusts the mixture of variational operators to explore the problem domain effectively.

The Borg MOEA's adaptive configuration of simulated binary crossover (SBX), differential evolution (DE), parent-centric recombination (PCX), unimodal normal distribution crossover (UNDX), simplex crossover (SPX), polynomial mutation (PM) and uniform mutation (UM) allows it to adapt to a problem's local characteristics and adjust as required throughout its execution. Moreover, the algorithm's auto-adaptive operator selection can identify which variation operators were successful on a particular problem.

In addition to diversity and selection pressure, the successful and efficient evolution of candidate solutions is crucial for many-objective optimization. The Borg MOEA's auto-adaptive search algorithm directly addresses this issue, leading to dynamically changing mixtures of search operators that yield a broad array of cooperative and evolving exploration strategies. The algorithm also monitors search progress using the epsilon-progress metric and initiates adaptive time continuation (Goldberg 2002a; Srivastava 2002) to inject fresh diversity into a stagnant population, ensuring efficient evolution.

2.3. Test problems

While the proposed real-world benchmark is designed to be extensible, four initial implementations of real-world applications ranging from three to ten objectives are provided. A summary of these problems can be found in Table 2. This table also provides information about the ϵ-values employed by the Borg MOEA for ϵ non-dominated sorting.

2.3.1. Car side impact

The Car side impact problem is a conceptual model for designing cars to withstand car side impacts (Jain and Deb 2014). The aim is to minimize the weight of the car while simultaneously minimizing the pubic force experienced by the passengers and minimizing the average velocity of the V-pillar responsible for withstanding the impact load. The problem represents a mildly constrained three-objective application with seven real decision variables.

2.3.2. Lake problem

The Lake problem is a classical socio-ecological systems decision problem with the goal of managing the eutrophication of lakes subject to irreversible tipping points—*i.e.* nonlinear threshold dynamics (Carpenter, Ludwig, and Brock 1999). In this problem, a hypothetical town emitting phosphorus into a shallow lake system must balance economic needs with the goal of avoiding irreversible pollution of the lake. It entails making annual phosphorus emission control decisions for 100 years, aiming to maximize economic benefits, minimize significant year-to-year emission changes, maximize the reliability of avoiding the pollution threshold and minimize lake phosphorus levels. From an MOEA perspective, this benchmarking application exhibits highly nonlinear threshold behaviour, and its 100-variable decision space is susceptible to genetic drift failures, where less important decisions may not receive sufficient selection pressure (discussed as temporal salience structure by Thierens, Goldberg, and Pereira 1998). For a more detailed formulation and implementation of this benchmark, interested readers can refer to Ward *et al.* (2015) and Hadka *et al.* (2015).

2.3.3. Lower Rio Grande Valley

The Lower Rio Grande Valley (LRGV) problem is a risk-based water supply portfolio planning problem exploring water markets for the Lower Rio Grande Valley in Texas, USA (Kasprzyk *et al.* 2012).

Table 2. Real-world applications.

Name	Decisions (no. of variables)	Objectives	ϵ-Values	Constraints (no.)	Properties	Reference
Car side impact	Design variables (7)	Min: weight, avg. vel. of V-pillar, pubic force experienced.	0.5, 0.25 0.05	Design response constraints (10)	Real-valued, Constrained	Jain and Deb (2014)
Lake Problem	Phosphorous emission controls (100)	Min. average daily P. Max: utility, inertia, day where the critical phosphorous threshold is met.	0.01 0.01, 0.0001 0.0001	Reliability $>$ 85%	Real-valued, Nonlinear, Stochastic objectives	Carpenter, Ludwig, and Brock (1999)
LRGV	Permanent rights, adaptive options contract, leases (8)	Min: cost, surplus water, dropped transfers, no. of leases, dropped transfers' cost. Max. critical reliability.	0.0009, 0.03 0.004, 0.004 0.004 0.002	Reliability $>$ 0.98 Crit. reliability $=$ 1.0 Cost variability $<$ 1.1	Real-valued, Constrained, Stochastic objectives, Disjoint PF	Kasprzyk *et al.* (2012)
GAA	Design variables (27)	Min: noise, empty weight, cost, ride roughness, fuel weight, price. Max: travel range, reuse, lift-to-drag ratio, cruising speed.	0.15, 30.0, 6.0 0.03, 30.0, 3000.0 150.0, 0.3 0.3, 3	Aggregate performance constraint (1)	Real-valued, Constrained Non-separable decisions	Simpson and D'Souza (2004)

In this problem, real-valued decision variables are used to represent the city's selection among three supply sources. These sources include a portion of reservoir inflows (referred to as non-market permanent rights), monthly leases of water from regional irrigation districts with pricing based on volatile demand, and adaptive options contracts that guarantee fixed prices for a specified volume in drought periods.

One critical constraint in the LRGV application ensures that the city always avoids catastrophic water supply shortfalls, defined as situations where supply falls below 40% of demand. To evaluate alternative combinations of market-based and traditional reservoir sources for water supply, Monte Carlo based model assessments are employed to compute the resulting six-objective performance while considering these constraints. The objectives include minimizing expected costs, cost variability, surplus water volume, unused purchased market water volume, and the count of water leases used. Additionally, a sixth objective focuses on maximizing the system's reliability in meeting demands.

Previous MOEA benchmarking efforts for the LRGV problem (Reed *et al.* 2013) involved running a random Latin hypercube sampling of the decision space, consisting of 500,000 members, to assess constraint difficulty. This sampling approach resulted in a very low success rate of 0.06% in generating feasible solutions, and the identified feasible solutions were strongly dominated. Notably, the LRGV problem's best-known Pareto front exhibits a disjointed geometry with two distinct families of solutions, characterized by high surplus water and expensive non-market reservoir-dominated supplies on one hand, and low surplus, cost-effective water market-dominated solutions on the other. Complex disjoint Pareto front geometries often arise in real-world applications where discrete families of solutions, representing different solution types, coexist, such as discrete levels of capital cost investment. For a comprehensive overview of the LRGV problem, readers can refer to Kasprzyk *et al.* (2012).

2.3.4. General aviation aircraft

The general aviation aircraft (GAA) problem has as its origin in the United States National Aeronautics and Space Administration aircraft design challenge from the 1990s (Simpson and D'Souza 2004; Woodruff, Reed, and Simpson 2013). The current version of the GAA problem, as presented in this study, focuses on creating a product family consisting of three general aviation aircraft. In this product family design problem, the objective is to maximize manufacturing economies of scale by maximizing the reuse of existing components (commonality of parts). The problem involves optimizing nine minimization performance objectives: takeoff noise, empty weight, direct operating cost, ride roughness, fuel weight, purchase price, and three maximization performance objectives: travel range, lift-to-drag ratio, and cruising speed. Each candidate product family comprises three aircraft, and each aircraft is characterized by these nine performance objectives, resulting in a challenging problem with 27 performance measures and one commonality objective. To manage the complexity, this study adopts an approach initially introduced by Shah, Reed, and Simpson (2011), where each of the nine performance objectives is transformed into a minimax objective across the three aircraft. The worst-performing aircraft in any of the nine performance measures defines the reported objective measure. While this minimax formulation reduces the objective count, it introduces non-separable interactions among the decision variables across the three aircraft. In this study, three aircraft designs are implemented to accommodate 2, 4 and 6 occupants, each characterized by nine real-valued decision variables, resulting in a total of 27 decision variables. Furthermore, the GAA problem is highly constrained, with strict minimum or maximum performance requirements for takeoff noise, empty weight, direct operating cost, ride roughness, fuel weight and flight range. In a prior benchmarking exercise (Shah, Reed, and Simpson 2011), even with a Latin hypercube sampling of 50 million candidate GAA designs, only three feasible solutions were found, highlighting its significant difficulty. Interested readers can refer to the studies of Shah, Reed, and Simpson (2011) and Woodruff, Reed, and Simpson (2013).

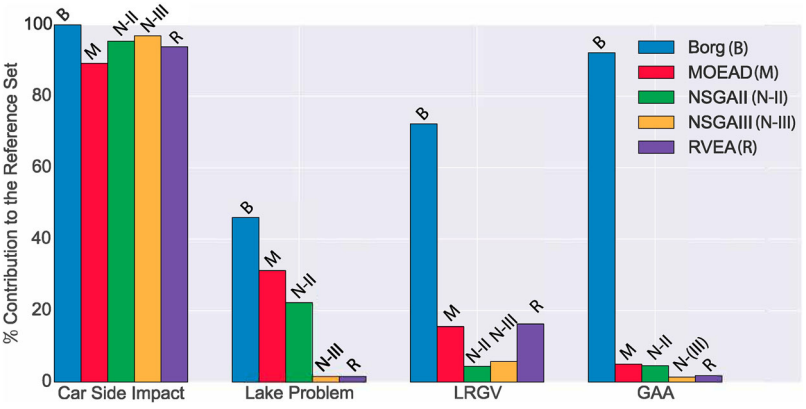

Figure 3. Percentage contribution of each MOEA to the reference set for the four real-world applications' best-known approximate reference sets.

3. Results

3.1. Reference set contributions

Figure 3 illustrates the contributions of the tested MOEAs to the best-known Pareto approximate reference sets in the four benchmark applications. These contributions encompass reference set members identified by multiple algorithms as well as solutions uniquely contributed by a single algorithm. The final reference sets were constructed by aggregating contributions from all runs of all algorithms. To ensure mathematical consistency in ranking definitions, the final reference sets were ϵ-non-dominated sorted, with specific ϵ-values provided in Table 2.

Among the applications, the Car Side Impact problem was the easiest, with the highest proportion of its reference points identified within the same ϵ-box across the MOEAs. The Borg MOEA achieved the best overall performance with a 100% identification rate, while MOEA/D, the lower performer, reached a set contribution as high as 90%. In general, NSGA-II and the reference point MOEAs exhibited very similar performance owing to the large number of runs and the relatively straightforward nature of the application.

For the Lake Problem's reference set, contributions were notably lower. The Borg MOEA (45%), MOEA/D (30%) and NSGA-II (22%) emerged as the top-performing algorithms. Notably, NSGA-III and RVEA made limited contributions, suggesting that the reference point frameworks struggled to achieve optimal convergence. The Lake Problem presents two primary mathematical challenges: (1) a highly nonlinear irreversible threshold affecting control decisions; and (2) the potential diminishing impact of the 100 decision variables on system performance, characterized by temporal salience structure and domino convergence—see Thierens, Goldberg, and Pereira (1998). These challenges, particularly domino convergence and potential selection pressure failures, have been observed in prior diagnostic assessments (Ward *et al.* 2015).

In contrast, the LRGV application, despite having a lower-dimensional decision space, presents significant challenges due to its highly constrained six-objective stochastic formulation and disjoint Pareto front. Previous MOEA diagnostics have revealed that many MOEAs struggle with maintaining diversity and achieving convergence on this problem (Reed *et al.* 2013). Notably, the Borg MOEA contributed approximately 75% of the solutions to the reference set. MOEA/D and RVEA performed similarly, each contributing around 15%. On the other hand, both NSGA-II and NSGA-III faced difficulties in contributing solutions to the reference set. The higher dimensionality of the disjoint Pareto set in the LRGV problem poses deterioration challenges for reference point methods and scaling challenges for MOEA/D's Tchebycheff weighting schemes used in its decomposition strategy.

The ten-objective, highly constrained GAA test yielded the most contrasting results in terms of reference set contributions, as depicted in Figure 3. Despite the suite of algorithms being tailored for many-objective optimization, the GAA problem presented severe challenges for the non-adaptive MOEAs. The Borg MOEA, as a meta-heuristic algorithm, leverages feedback in its search process to adjust population size, selection pressure and mixtures of variational operators dynamically. This adaptivity enabled the Borg MOEA to contribute 95% of the GAA reference set.

While theoretical limits of Pareto ranking have spurred the growth of decomposition and reference point MOEAs (Teytaud 2006, 2007), the results shown in Figure 3 underscore that high-dimensional real-world problems pose substantial challenges for these approaches. As noted in previous assessments (Hadka and Reed 2012; Woodruff *et al.* 2012), the GAA problem's severe constraints, coupled with non-separable aircraft family design decisions, strongly activate the Borg MOEA's auto-adaptivity, particularly in terms of its dynamic utilization of variational operator mixtures. Although the reference set contribution results in Figure 3 offer a broad view of the relative ease or difficulty of the applications, they do not provide specific insights into the modes of search failure, such as convergence, proximity, diversity and consistency of the approximation sets achieved by the evaluated MOEAs.

3.2. Best performance and attainment probabilities

The attainment plots shown in Figure 4 provide a probabilistic metrics-based comparison of the five MOEAs across the four benchmark applications. Figures 4(a)–4(c) show the best single run attainment results for the generational distance (GD, 'convergence'), ϵ-indicator (EI, 'consistency') and hypervolume (HV, 'proximity and diversity') metrics, respectively. In panels (a)–(c), two key aspects of performance for each application are captured: (1) each MOEA's best single run performance for each application (designated with circles); and (2) the probability of attaining between 0 to 100% of the best overall single run's metric value. The 100% of best metric performance level represents the best metric value attained by a single trial run across all the test MOEAs' parameterizations and random seed trials. For each application and metric, the attainment plots provide a comparative assessment of the single most effective runs as well as the algorithms' relative reliability in attaining each level of the metrics' values.

Given that high GD performance only requires a single approximation set point near the reference set, this pure convergence measure is most interesting for diagnosing abject search failures. In Figure 4(a), there is very little difference in the MOEAs' best run and the GD attainments for the Car Side Impact and Lake Problem test cases. The GD results for the LRGV show reduced attainment probabilities across the MOEAs. RVEA and NSGA-II maintain at least 90% attainment probabilities for up to 85% of the best GD value. Beyond that level of performance, both algorithms struggle. Although the Borg MOEA is less reliable than RVEA and NSGA-II at 85% of the best GD value, its attainment probabilities decrease less abruptly at the highest levels of GD performance. Both MOEA/D and NSGA-III show substantially lower GD attainments. The LRGV problem's disjoint Pareto set poses a challenge to the reference point scheme for NSGA-III, as has been shown in other studies (Deb and Jain 2014; Jain and Deb 2014). The relative scaling of the LRGV application's objectives poses a challenge to MOEA/D (Reed *et al.* 2013). The GAA problem shows the most pronounced differences in GD attainments in Figure 4(a). Although all five MOEAs had at least a single best run with high GD performance, the ten-objective constrained GAA problem did yield significant reductions in attainment probabilities for MOEA/D, NSGA-II and RVEA. Overall, the Borg MOEA and NSGA-III maintained high attainment probabilities near peak GD metric values.

The EI attainment plots in Figure 4(b) show that gap-free, highly consistent approximations of the four applications is far more challenging than the GD attainment. Car Side Impact remains the easiest of the application test problems. MOEA/D showed the largest decline in both its best single run as well as its overall EI attainment probabilities. The Borg MOEA is the top performer of the tested MOEAs for the Car Side Impact problem followed by NSGA-III. NSGA-II and RVEA had

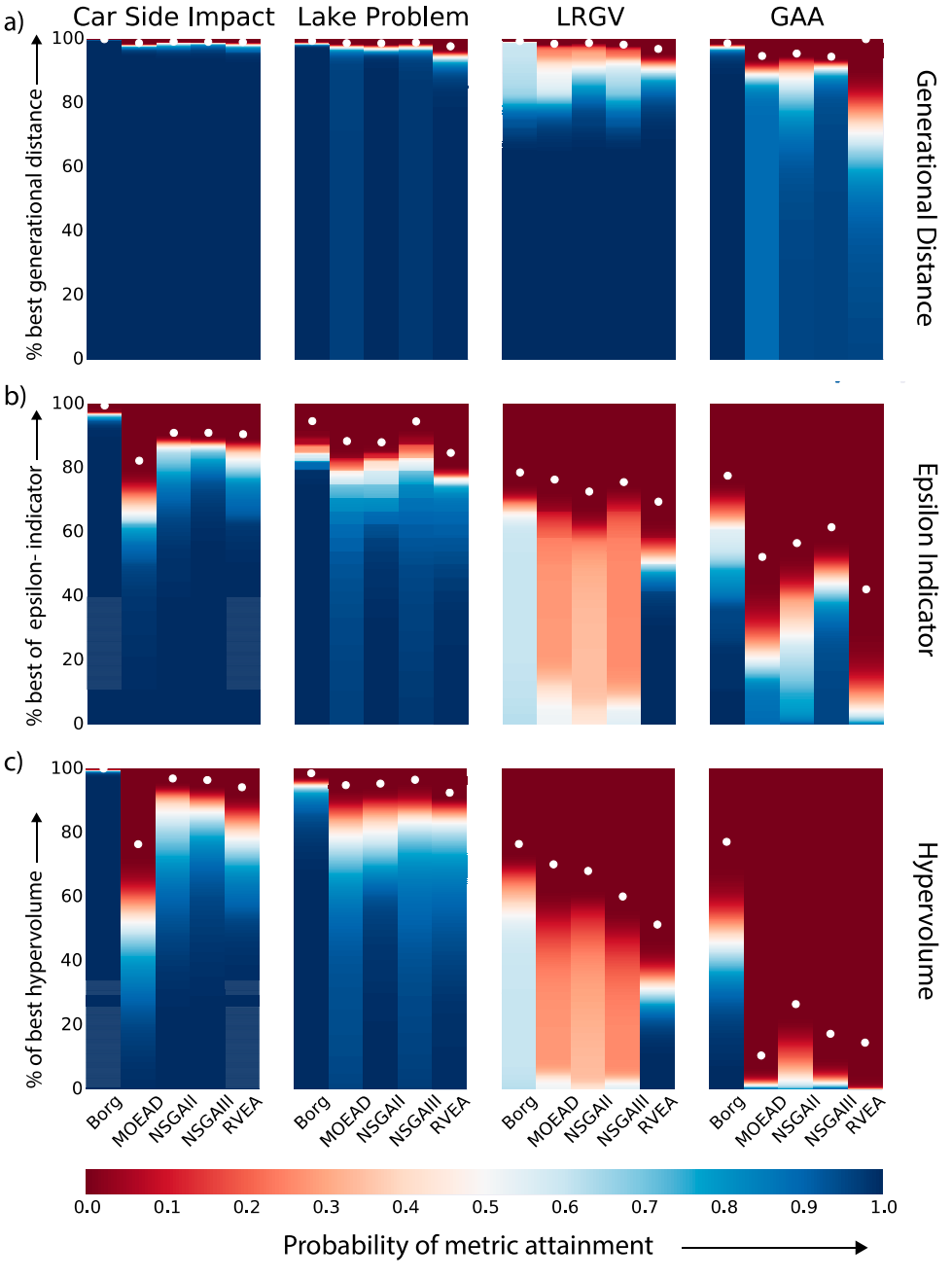

Figure 4. Best performance and attainment probabilities of each MOEA on the four real-world applications.

similar EI attainment performance for the Car Side Impact application. The Borg MOEA again has the top EI attainment performance for the Lake Problem. All the MOEAs struggled to attain results reliably beyond 80% of the best overall reference EI metric value. MOEA/D and NSGA-III showed the worst overall EI attainment performance on the Lake Problem, which is also reflected in the GD attainments in Figure 4(a). For the LRGV EI attainment probabilities and the best single run results in Figure 4(b), all the MOEAs show a strong decline in performance relative to the GD attainment results in Figure 4(a). RVEA shows strong reliability up to approximately 40% of the best EI metric

value and then declines very rapidly. The Borg MOEA maintains a 70% probability of attainment up until approximately 65% of the best EI value.

In the case of the LRGV, given that it is a constrained application with a disjoint Pareto front, it is typically easier and quicker to identify a larger region of attraction compared to isolated groups of solutions involving different non-market water supply options. The EI metric is sensitive to gaps, in this case, to the missing second group of solutions. Notably, none of the algorithms can completely avoid gaps in a single trial run for the LRGV. Specifically, MOEA/D, NSGA-II and NSGA-III face challenges throughout the entire range of EI attainment for the LRGV.

The GAA EI attainments in Figure 4(b) emphasize the complexity of algorithm selection. In this scenario, RVEA's EI attainment for the GAA displays failure across the full spectrum of potential EI values. In contrast, the Borg MOEA maintains a degree of reliable EI attainment across all tested applications. For the GAA, the Borg MOEA achieves the highest EI attainment probabilities for the best overall metric values.

Comparing the EI attainments in Figure 4(b) with the HV attainments in Figure 4(c), consistent performance trends across all applications are observed, except for the GAA test problem. Notably, the HV attainment results for the GAA in Figure 4(c) sharply contrast with the EI attainments in Figure 4(b). In this context, the Borg MOEA stands out by achieving the highest single attained solutions and reliably attaining up to 40% of the best metric values for the GAA application. However, it's important to highlight that most MOEAs struggle to attain high HV values for the GAA problem. This particular instance has the highest dimensionality among those tested in this study, featuring ten objective trade-offs, and some MOEAs may not scale effectively to address it.

Overall, when considering both EI attainments in Figure 4(b) and HV attainments in Figure 4(c) for the GAA, the following is observed: NSGA-III succeeds in creating a relatively consistent and gap-free local approximate set. However, it does not achieve close proximity to the actual reference set, leading to modest HV performance. On the other hand, despite facing challenges presented by the GAA problem's high dimensionality and constraints, the Borg MOEA stands out for its single-run performance, effectively addressing these issues.

3.3. Efficiency and controllability

In this study, a diagnostic framework is used to explore 500 Latin hypercube samples for each MOEA's complete parameter space. Figure 4 presents probabilistic attainments, summarizing the search effectiveness and reliability across parameterizations and random seed trials for each MOEA. While statistically robust, these results may not reflect typical real-world usage, where extensive parameter analysis is often impractical owing to time and resource constraints. Instead, MOEAs that demonstrate reliable performance and quick convergence without extensive tuning are preferable. Real-world applications often involve costly efforts to identify suitable parameterizations, as discussed in prior research (Kasprzyk *et al.* 2012; Walker *et al.* 2003; Zeleny 1989). In such contexts, robust MOEAs are preferable, as they minimize operational costs and maximize decision-making reliability without extensive trial-and-error parameter tuning.

Figure 5 contains control maps assessing MOEAs' controllability, effectiveness and efficiency based on expected HV_{REF} performance relative to reference sets. Each plot displays expected HV_{REF} results from 50 random trials for a specific algorithm's parameterization. These maps highlight performance variations related to NFEs, search population size and reference point divisions. Ideally, control maps would show consistently high HV_{REF} performance across parameterizations with solid dark shading.

In Figure 5, for the Car Side Impact problem, the Borg MOEA achieves nearly ideal performance across its tested parameterizations when NFE values exceed 10,000. While this problem is the easiest among the tested applications, both the HV attainments in Figure 4 and the control maps in Figure 5 emphasize the potential for significant parametric sensitivities and reduced reliability in the four other algorithms. MOEA/D consistently exhibits the lowest expected HV performance regardless of NFE count. Additionally, Figure 5 reveals that NSGA-II is the second most effective algorithm for

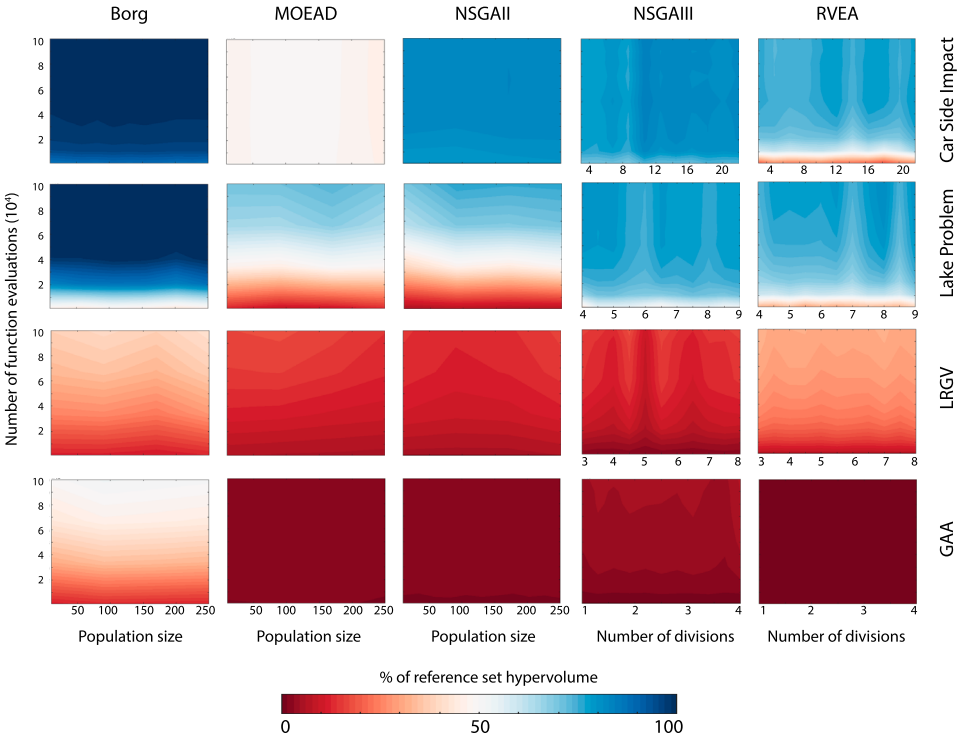

Figure 5. Control map plots showing the efficiency and controllability of each MOEA on the four real-world applications.

the Car Side Impact problem. On the other hand, RVEA and NSGA-III display increased parametric sensitivities (*i.e.* reduced controllability) and a relatively unpredictable dependence on NFE count.

A key distinction between NSGA-II, NSGA-III and RVEA compared to the Borg MOEA is how they manage parametric sensitivities. The Borg MOEA incorporates several ϵ-dominance-based feedback mechanisms, enabling the algorithm to auto-adapt its selective pressure, population size, utilization of variational operators and search population diversity. These feedbacks operate internally within the algorithm, requiring no user intervention. In contrast, achieving high performance with the other MOEAs on the Car Side Impact problem demands external user experimentation to navigate the parametric sensitivities that influence search performance.

In Figure 5, the control maps for the Lake Problem indicate greater complexity compared to the results for the Car Side Impact problem across all tested MOEAs. In particular, the control maps for the Borg MOEA clearly demonstrate a trend of reduced variability and improved HV performance as the NFE count surpasses 20,000. On the other hand, NSGA-III and RVEA display a more complex relationship between expected HV performance and their parameter settings when compared to the Borg MOEA. Notably, the Borg MOEA exhibits enhanced controllability with increasing NFE counts, revealing distinct parametric 'sweet spots' (Goldberg 2002b). Conversely, MOEA/D and NSGA-II present complex control maps for the Lake Problem in Figure 5, indicating higher sensitivity to the algorithmic parameterizations. These algorithms exhibit large NFE requirements before showing improvements in expected HV performance.

All MOEAs showed limited regions of high HV performance for the LRGV problem. Moreover, the LRGV problem involves Monte Carlo based function evaluations, making it the most computationally demanding among the four real-world benchmarking problems. Notably, both the Borg MOEA and RVEA demonstrated the highest degree of controllability when addressing the LRGV problem, displaying a clear advantage in both expected HV performance and controllability as the NFE count

increased for this particular problem. Conversely, MOEA/D, NSGA-II and NSGA-III exhibited subpar controllability when attempting to solve the LRGV problem within the allocated computational time.

Furthermore, as illustrated in Figure 5, it is evident that the ten-objective GAA problem poses the greatest challenges among the four problems considered. It exhibits the most severe constraints, the highest-dimensional objective space and non-separable decision factors. In single trial runs, the expected HV_{REF} performance is notably limited for all tested MOEAs.

Remarkably, at 50,000 NFEs, the Borg MOEA begins to demonstrate notable HV performance and exhibits consistent improvement as the NFE count increases. Prior studies (Hadka and Reed 2012; Shah, Reed, and Simpson 2011; Woodruff *et al.* 2012; Woodruff, Reed, and Simpson 2013) have indicated that auto-adaptive population sizing and time continuation, involving the injection of diverse solutions, significantly enhance MOEA performance when dealing with the GAA problem. Additionally, Hadka and Reed (2012) conducted detailed sensitivity analyses and runtime analysis, demonstrating that the Borg MOEA's cooperative and adaptive utilization of multiple variation operators, including those tailored for non-separable problems, contributes to its success in solving the GAA problem. Surprisingly, the decomposition and reference point strategies, designed to address issues in Pareto ranking, seem to face challenges when applied to the GAA problem.

In general, the controllability results of Figure 5, as well as findings from prior diagnostic studies that carefully explored MOEA parameter space sensitivities (Gupta *et al.* 2020; Hadka and Reed 2012; Purshouse and Fleming 2007; Reed and Kollat 2013) challenge the common notion that stable problem independent MOEA parameter settings ever exist for non-trivial many-objective problems, especially for non-adaptive algorithms. For example, Ward *et al.* (2015) show that even a modest change in the Lake Problem's statistical assumptions caused drastic changes in many non-adaptive MOEAs' control maps, where in many instances successful zones of performance completely disappear (*i.e.* complete algorithmic failure).

3.4. Comparative runtime dynamics for default parameterizations

Figures 4 and 5 provide diagnostic results to understand the overall efficiency, effectiveness, reliability and controllability of the four tested MOEAs. However, these results do not represent the typical operational use of MOEAs. In reality, most MOEA applications and studies use the author-recommended default parameterization and run algorithms for multiple random seed trials.

Figure 6 presents the HV runtime search dynamics of the default parameterizations for the five algorithms addressing the four benchmarking problems. Panels (a)–(d) compare the efficiency (NFE) and reliability (90th interquantile range) of standard MOEA implementations to the best single MOEA runs for each benchmark application. For instance, panel (a) illustrates that the Borg MOEA's default implementation achieves high HV results in under 10,000 NFEs for the Car Side Impact problem. Even the least favourable random seed trial of the Borg MOEA outperforms the other algorithms. NSGA-II and NSGA-III also efficiently and reliably solve the Car Side Impact problem but generally do not reach the same HV performance level as the Borg MOEA. RVEA, on the other hand, requires almost double the NFE (over 20,000 NFEs) to attain Car Side Impact HV results. Concerning the Lake Problem, in panel (b), the default runtime dynamics are displayed, emphasizing substantial differences in the algorithms' runtime behaviours. Both NSGA-III and the Borg MOEA rapidly achieve high HV performance levels (between 15,000 and 20,000 NFEs), with NSGA-III exhibiting greater reliability in the early stages of the search process.

When compared to the Borg MOEA and NSGA-III, RVEA shows lower efficiency, effectiveness and reliability when applied to the Lake Problem. Further, the best random seed trials of RVEA perform less effectively than the worst trials of the top-performing algorithms. Moreover, the default implementation of MOEA/D experiences a significant delay in search progress, requiring more than 50,000 NFEs before achieving substantial HV improvements.

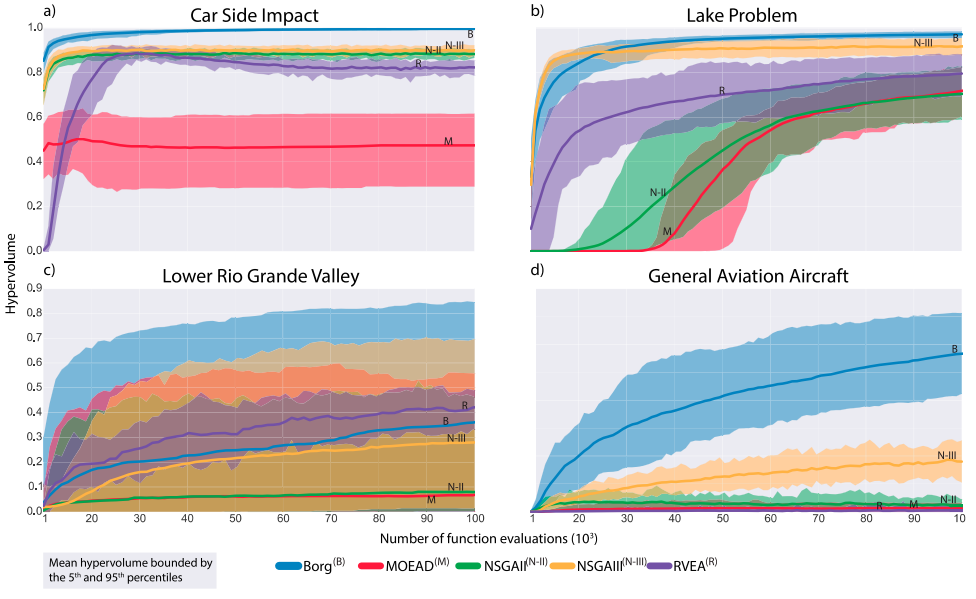

Figure 6. Runtime dynamics for each MOEA using their default parameterizations across the four real-world applications.

In Figure 6(c), the default runtime dynamics for the Lower Rio Grande Valley (LRGV) problem highlight its complexity and the substantial performance uncertainty that real-world practitioners would face when using MOEAs' HV results. The wide variability in the 90th inter-quantile ranges of random seed trials indicates significant unpredictability, making mean runtime traces poor reflections of actual performance. In this context, the Borg MOEA offers a more accurate representation of LRGV application tradeoffs, improving decision support.

In the case of the GAA problem, only the Borg MOEA successfully solved this challenging ten-objective benchmark. Even the worst-performing Borg MOEA trials with default settings outperformed the best trials of all other algorithms. NSGA-III was the only other algorithm that showed some progress with default settings. Notably, the Borg MOEA, being a hyper-heuristic, stands out for its adaptability, adjusting population size, selective pressure, search diversity and variation operators based on ϵ-dominance archiving.

In the discussion by Giagkiozis and Fleming (2015), they argue against solely transitioning away from Pareto ranking schemes to address MOEA limitations. They emphasize that search progress depends on various factors, including population diversity, topological complexities and the avoidance of stagnation due to a lack of selective pressure. Additionally, they highlight the need for MOEAs to be flexible, adaptable and user-friendly in real-world decision contexts.

While decomposition techniques and reference point frameworks have received significant attention in the MOEA research community, this study's results motivate a shift beyond non-adaptive search mechanisms. Overall, the findings in Figures 4–6 underscore the importance of moving beyond fine-tuned parameterizations in MOEA applications, emphasizing the considerable differences in algorithm runtime dynamics. These results emphasize the need for hyper-heuristic approaches capable of adapting to each problem's unique characteristics, providing more accurate tradeoff representations for decision support.

4. Conclusion

In this study, a new open-source real-world benchmarking framework was introduced. Through this framework, a rigorous comparison of five popular MOEAs on real-world problems was conducted.

The results indicate performance disparities among the different MOEAs. These findings offer valuable insights into algorithmic design choices. First, the choice of selection strategy and/or archive mechanism must avoid deterioration, provide sufficient selection pressure and maintain genetic diversity throughout the entire run. ϵ-Dominance archiving and reference points have proven to be successful in this regard. Secondly, it is often overlooked that genetic operators should be stable and flexible across various problem domains in order to drive the search towards the Pareto front reliably. Historically, the choice of genetic operator has been the responsibility of the MOEA practitioner, but recent advances in hyper-heuristics and multi-operator MOEAs provide an automated approach. Lastly, it is essential to recognize that selection and genetic operators are mutually important factors and should not be studied independently in isolation.

These findings suggest that, when aiming to explore the entire Pareto front, ϵ-non-dominated sorting is effective in preserving genetic diversity and maintaining selection pressure on problems with up to ten objectives. Additionally, the results highlight the advantages of using auto-adaptive search operators that can dynamically adapt to the search in diverse real-world applications. While reference point/vector methods have gained popularity, the results indicate a potential deterioration in performance at higher dimensions. Further research is needed to isolate the root cause of deteriorating performance.

The study reveals a notable flaw in existing test benchmarks when it comes to distinguishing the performance of MOEAs (Hadka and Reed 2012). While significant emphasis is placed on gaining marginal improvements on a suite of test problems, the subpar performance of MOEAs in real-world applications persists. The study focused on a selected number of cases spanning design, environmental and water management applications, areas where the authors have expertise. However, a larger pool of test cases is necessary to derive generalizable insights, and the framework would benefit from a broader range of cases representative of real-world environments. The MOEAs were chosen to encompass a wide spectrum of design characteristics and operational challenges, though the selection was not exhaustive. Further, to ensure fairness, the number of function evaluations across all MOEAs was standardized, regardless of test case complexity. This approach aimed to prevent sampling biases due to the high sensitivity of some MOEAs to exploration periods. However, MOEAs with auto-adaptive operator selection might perform better with extended exploration. Additionally, the methodology included 50 random seed trials with 500 samples each, a computationally demanding endeavour that may be impractical in settings with limited resources.

Despite these limitations, the purpose of the benchmarking framework and the results presented in this study is to initiate collaborative efforts within the field, with the aim of expanding a diverse suite of real-world benchmarking problems for a range of domains. The MOEA Framework allows for future studies so as also to flexibly explore new or mixed algorithm architectures as part of benchmarking efforts. The testing framework and real-world benchmarks are freely available to all researchers, and can serve as a foundation for developing the next round of innovations and algorithmic improvements.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The Cornell authors were funded in this work by the US National Science Foundation (NSF)'s Innovations at the Nexus of Food, Energy and Water Systems (INFEWS) programme [Award No. 1639268]. The views expressed in this work represent those of the authors and do not necessarily reflect the views or policies of the NSF.

Data and code availability

The mathematical formulations of the four real-world test problems, the hyperparameters of the algorithms, the best-known non-dominated sets and the analysis settings are available from Hadka (2023). Please also refer to the online supplemental data for further experimental details, operator dynamics and visualizations of the Pareto fronts.

References

Burke, Edmund K., Michel Gendreau, Matthew Hyde, Graham Kendall, Gabriela Ochoa, Ender Özcan, and Rong Qu. 2013. "Hyper-Heuristics: A Survey of the State of the Art." *Journal of the Operational Research Society* 64 (12): 1695–1724.

Carpenter, S. R., D. Ludwig, and W. A. Brock. 1999. "Management of Eutrophication of Lakes Subject to Potentially Irreversible Change." *Ecological Applications* 9 (3): 751–771. https://doi.org/10.1890/1051-0761(1999)009[0751:MOEFLS]2.0.CO;2.

Cheng, Ran, Yaochu Jin, Markus Olhofer, and Bernhard Sendhoff. 2016. "A Reference Vector Guided Evolutionary Algorithm for Many-Objective Optimization." *IEEE Transactions on Evolutionary Computation* 20 (5): 773–791. http://dx.doi.org/10.1109/TEVC.2016.2519378.

Cheng, Ran, Yaochu Jin, Markus Olhofer, and Bernhard Sendhoff. 2017. "Test Problems for Large-Scale Multiobjective and Many-Objective Optimization." *IEEE Transactions on Cybernetics* 47 (12): 4108–4121. https://doi.org/10.1109/TCYB.2016.2600577.

Cheng, Ran, Miqing Li, Ye Tian, Xingyi Zhang, Shengxiang Yang, Yaochu Jin, and Xin Yao. 2017. "A Benchmark Test Suite for Evolutionary Many-Objective Optimization." *Complex & Intelligent Systems* 3 (1): 67–81.

Coello Coello, Carlos A., Gary B. Lamont, and David A. Van Veldhuizen. 2007. *Evolutionary Algorithms for Solving Multi-Objective Problems*. New York: Springer Science+Business Media.

Das, Indraneel, and John E. Dennis. 1998. "Normal-Boundary Intersection: A New Method for Generating the Pareto Surface in Nonlinear Multicriteria Optimization Problems." *SIAM Journal on Optimization* 8 (3): 631–657.

Deb, Kalyanmoy. 1999. "Multi-Objective Genetic Algorithms: Problem Difficulties and Construction of Test Problems." *Evolutionary Computation* 7 (3): 205–230.

Deb, Kalyanmoy, and Sachin Jain. 2002. *Running Performance Metrics for Evolutionary Multi-Objective Optimization*. KanGAL Report No. 2002004. Kanpur, India: Kanpur Genetic Algorithms Laboratory (KanGAL), Indian Institute of Technology.

Deb, Kalyanmoy, and Himanshu Jain. 2014. "An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems with Box Constraints." *IEEE Transactions on Evolutionary Computation* 18 (4): 577–601.

Deb, Kalyanmoy, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. 2002. "A Fast Elitist Multi-Objective Genetic Algorithm: NSGA-II." *IEEE Transactions on Evolutionary Computation* 6 (2): 182–197.

Deb, Kalyanmoy, Lothar Thiele, Marco Laumanns, and Eckart Zitzler. 2001. *Scalable Test Problems for Evolutionary Multi-Objective Optimization*. TIK-Technical Report No. 112. Zürich, Switzerland: Computer Engineering and Networks Laboratory (TIK), Swiss Federal Institute of Technology (ETH).

Deb, Kalyanmoy, Lothar Thiele, Marco Laumanns, and Eckart Zitzler. 2002. "Scalable Multi-Objective Optimization Test Problems." In *Proceedings of the 2002 Congress on Evolutionary Computation (CEC'02)*, 825–830. Piscataway, NJ: IEEE. http://dx.doi.org/10.1109/CEC.2002.1007032.

Eiben, Agoston E., and Jim Smith. 2015. "From Evolutionary Computation to the Evolution of Things." *Nature* 521 (7553): 476–482.

Fieldsend, Jonathan E., Tinkle Chugh, Richard Allmendinger, and Kaisa Miettinen. 2021. "A Visualizable Test Problem Generator for Many-Objective Optimization." *IEEE Transactions on Evolutionary Computation* 26 (1): 1–11.

Giagkiozis, I., and P. J. Fleming. 2015. "Methods for Multi-Objective Optimization: An Analysis." *Information Sciences* 293:338–350.

Goldberg, David E. 2002a. "Design of Competent Genetic Algorithms." In *The Design of Innovation*, 187–216. New York: Springer.

Goldberg, David E. 2002b. *Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Norwell, MA: Kluwer Academic.

Gupta, Rohini S., Andrew L. Hamilton, Patrick M. Reed, and Gregory W. Characklis. 2020. "Can Modern Multi-Objective Evolutionary Algorithms Discover High-Dimensional Financial Risk Portfolio Tradeoffs for Snow-Dominated Water–Energy Systems?." *Advances in Water Resources* 145:103718.

Hadka, David. 2023. "MOEAFramework/RealWorldBenchmarks: Engineering Optimization Study".

Hadka, David. 2024. "MOEA Framework." http://moeaframework.org. Version 3.11.

Hadka, David, Jonathan Herman, Patrick Reed, and Klaus Keller. 2015. "An Open Source Framework for Many-Objective Robust Decision Making." *Environmental Modelling & Software* 74:114–129.

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

Hadka, David, and Patrick Reed. 2013. "Borg: An Auto-Adaptive Many-Objective Evolutionary Computing Framework." *Evolutionary Computation* 21 (2): 231–259.

Huband, Simon, Phil Hingston, Luigi Barone, and Lyndon While. 2006. "A Review of Multiobjective Test Problems and a Scalable Test Problem Toolkit." *IEEE Transactions on Evolutionary Computation* 10 (5): 477–506.

Ishibuchi, Hisao, Lie Meng Pang, and Ke Shang. 2022. "Difficulties in Fair Performance Comparison of Multiobjective Evolutionary Algorithms." In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*

(GECCO '22), 937–957. New York: Association for Computing Machinery (ACM). https://doi.org/10.1145/3520304.3533634.

Jain, Himanshu, and Kalyanmoy Deb. 2014. "An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point Based Nondominated Sorting Approach, Part II: Handling Constraints and Extending to An Adaptive Approach." *IEEE Transactions on Evolutionary Computation* 18 (4): 602–622.

Jan, Muhammad Asif, and Qingfu Zhang. 2010. "MOEA/D for Constrained Multiobjective Optimization: Some Preliminary Experimental Results." In *Proceedings of the 2010 UK Workshop on Computational Intelligence (UKCI)*, 1–6. Piscataway, NJ: IEEE.

Kasprzyk, J. R., P. M. Reed, B. R. Kirsch, and G. W. Characklis. 2012. "Many-Objective De Novo Water Supply Portfolio Planning Under Deep Uncertainty." *Environmental Modelling & Software* 34:87–104.

Knowles, Joshua, and David Corne. 2002. "On Metrics for Comparing Non-Dominated Sets." In *Proceedings of the 2002 Congress on Evolutionary Computation (CEC'02)*, 711–716. Piscataway, NJ: IEEE. http://dx.doi.org/10.1109/CEC.2002.1007013.

Li, Miqing, Crina Grosan, Shengxiang Yang, Xiaohui Liu, and Xin Yao. 2017. "Multiline Distance Minimization: A Visualized Many-Objective Test Problem Suite." *IEEE Transactions on Evolutionary Computation* 22 (1): 61–78.

Li, Hui, and Qingfu Zhang. 2009. "Multiobjective Optimization Problems with Complicated Pareto Sets, MOEA/D and NSGA-II." *IEEE Transactions on Evolutionary Computation* 13 (2): 284–302.

Meneghini, Ivan Reinaldo, Marcos Antonio Alves, António Gaspar-Cunha, and Frederico Gadelha Guimaraes. 2020. "Scalable and Customizable Benchmark Problems for Many-Objective Optimization." *Applied Soft Computing* 90:106139.

Miettinen, Kaisa M. 2012. *Nonlinear Multiobjective Optimization*. Vol. 12 of the book series *International Series in Operations Research & Management Science (ISOR)*. Softcover reprint of the 1998 1st ed. New York: Springer Science & Business Media. https://doi.org/10.1007/978-1-4615-5563-6.

Purshouse, Robin C., and Peter J. Fleming. 2007. "On the Evolutionary Optimization of Many Conflicting Objectives." *IEEE Transactions on Evolutionary Computation* 11 (6): 770–784.

Reed, Patrick M., David M. Hadka, Jonathan D. Herman, Joseph R. Kasprzyk, and Joshua B. Kollat. 2013. "Evolutionary Multiobjective Optimization in Water Resources: The Past, Present, and Future." *Advances in Water Resources* 51:438–456.

Reed, Patrick M., and Joshua B. Kollat. 2013. "Visual Analytics Clarify the Scalability and Effectiveness of Massively Parallel Many-Objective Optimization: A Groundwater Monitoring Design Example." *Advances in Water Resources* 56:1–13.

Santiago, Alejandro, Bernabé Dorronsoro, Antonio J. Nebro, Juan J. Durillo, Oscar Castillo, and Héctor J. Fraire. 2019. "A Novel Multi-Objective Evolutionary Algorithm with Fuzzy Logic Based Adaptive Selection of Operators: FAME." *Information Sciences* 471:233–251.

Shah, R., P. M. Reed, and T. Simpson. 2011. "Many-Objective Evolutionary Optimization and Visual Analytics for Product Family Design." In *Multi-Objective Evolutionary Optimisation for Product Design and Manufacturing*, edited by Lihui Wang, Amos H. C. Ng, and Kalyanmoy Deb, 137–159. London: Springer.

Simpson, T. W., and B. D'Souza. 2004. "Assessing Variable Levels of Platform Commonality Within a Product Family Using a Multiobjective Genetic Algorithm." *Concurrent Engineering: Research and Applications* 12 (2): 119–130.

Srivastava, Ravi P. 2002. *Time Continuation in Genetic Algorithms*. IlliGAL Report No. 2001021. Urbana-Champaign, IL: Illinois Genetic Algorithms Laboratory (IlliGAL), University of Illinois.

Tanabe, Ryoji, and Hisao Ishibuchi. 2020. "An Easy-to-Use Real-World Multi-Objective Optimization Problem Suite." *Applied Soft Computing* 89:106078.

Teytaud, Olivier. 2006. "How Entropy-Theorems Can Show that On-Line Approximating High-Dim Pareto-Fronts is Too Hard." In *Proceedings of the International Conference on Parallel Problem Solving from Nature (PPSN), Bridging the Gap between Theory and Practice (BTP) Workshop*. https://hal.archives-ouvertes.fr/hal-00113362.

Teytaud, Olivier. 2007. "On the Hardness of Offline Multi-Objective Optimization." *Evolutionary Computation* 15 (4): 475–491.

Thierens, Dirk, David E. Goldberg, and Ângela Guimarães Pereira. 1998. "Domino Convergence, Drift, and the Temporal-Salience Structure of Problems." In *Proceedings of the 1998 IEEE International Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence (Cat. No. 98TH8360)*. Piscataway, NJ: IEEE. https://doi.org/10.1109/ICEC.1998.700085.

Vrugt, Jasper A., Bruce A. Robinson, and James M. Hyman. 2008. "Self-Adaptive Multimethod Search for Global Optimization in Real-Parameter Spaces." *IEEE Transactions on Evolutionary Computation* 13 (2): 243–259.

Walker, Warren E., Poul Harremoës, Jan Rotmans, Jeroen P. van der Sluijs, Marjolein B. A. van Asselt, Peter Janssen, and Martin P. Krayer von Krauss. 2003. "Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support." *Integrated Assessment* 4 (1): 5–17.

Wang, Zhenkun, Yew-Soon Ong, and Hisao Ishibuchi. 2018. "On Scalable Multiobjective Test Problems with Hardly Dominated Boundaries." *IEEE Transactions on Evolutionary Computation* 23 (2): 217–231.

Wang, Zhenkun, Yew-Soon Ong, Jianyong Sun, Abhishek Gupta, and Qingfu Zhang. 2018. "A Generator for Multiobjective Test Problems with Difficult-to-Approximate Pareto Front Boundaries." *IEEE Transactions on Evolutionary Computation* 23 (4): 556–571.

Ward, V., R. Singh, P. Reed, and K. Keller. 2015. "Confronting Tipping Points: How Well Can Multi-Objective Evolutionary Algorithms Support the Management of Environmental Thresholds?." *Environmental Modelling & Software* 73:27–43.

Woodruff, Matthew J., David Hadka, Patrick Reed, and Timothy Simpson. 2012. "Auto-Adaptive Search Capabilities of the New Borg MOEA: A Detailed Comparison on Product Family Design Problem Formulations." In *Proceedings of the 12th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference and 14th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, AIAA Paper AIAA-2012-5442. https://doi.org/10.2514/6.2012-5442.

Woodruff, Matthew J., Patrick M. Reed, and Timothy Simpson. 2013. "Many-Objective Visual Analytics: Rethinking the Design of Complex Engineered Systems." *Structural and Multidisciplinary Optimization* 48 (1): 201–219.

Woodruff, Matthew J., Timothy W. Simpson, and Patrick M. Reed. 2015. "Multi-Objective Evolutionary Algorithms' Performance in a Support Role." In *Proceedings of the ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, V02BT03A028. New York: American Society of Mechanical Engineers (ASME).

Zapotecas-Martínez, Saúl, Abel García-Nájera, and Adriana Menchaca-Méndez. 2023. "Engineering Applications of Multi-Objective Evolutionary Algorithms: A Test Suite of Box-Constrained Real-World Problems." *Engineering Applications of Artificial Intelligence* 123:106192.

Zatarain-Salazar, Jazmin, Patrick M. Reed, Jonathan D. Herman, Matteo Giuliani, and Andrea Castelletti. 2016. "A Diagnostic Assessment of Evolutionary Algorithms for Multi-Objective Surface Water Reservoir Control." *Advances in Water Resources* 92:172–185. https://doi.org/10.1016/j.advwatres.2016.04.006.

Zeleny, M. 1989. "Cognitive Equilibrium: A New Paradigm of Decision Making." *Human Systems Management* 8:185–188.

Zhang, Qingfu, and Hui Li. 2007. "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition." *IEEE Transactions on Evolutionary Computation* 11 (6): 712–731.

Zhang, Qingfu, Wudong Liu, and Hui Li. 2009. "The Performance of a New Version of MOEA/D on CEC09 Unconstrained MOP Test Instances." In *Proceedings of the Congress on Evolutionary Computation (CEC'09)*, 203–208. Piscataway, NJ: IEEE.

Zhou, Shengchao, Lining Xing, Xu Zheng, Ni Du, Ling Wang, and Qingfu Zhang. 2019. "A Self-Adaptive Differential Evolution Algorithm for Scheduling a Single Batch-Processing Machine with Arbitrary Job Sizes and Release Times." *IEEE Transactions on Cybernetics* 51 (3): 1430–1442.

Zitzler, E., K. Deb, and L. Thiele. 2000. "Comparison of Multiobjective Evolutionary Algorithms: Empirical Results." *IEEE Transactions on Evolutionary Computation* 8 (2): 173–195. https://doi.org/10.1162/106365600568202.

Zitzler, Eckart, Lothar Thiele, Marco Laumanns, Carlos M. Fonseca, and Viviane Grunert da Fonseca. 2003. "Performance Assessment of Multiobjective Optimizers: An Analysis and Review." *IEEE Transactions on Evolutionary Computation* 7 (2): 117–132.