Real World Diagnostic Benchmarking for Modern Many-Objective Evolutionary Algorithms

David Hadka
davidhadka@gmail.com
Applied Research Laboratory, The Pennsylvania State University, University Park, PA 16802

Jazmin Zatarain Salazar
jz583@cornell.edu
Department of Civil and Environmental Engineering, Cornell University, Ithaca, NY 14853

Patrick Reed
patrick.reed@cornell.edu
Department of Civil and Environmental Engineering, Cornell University, Ithaca, NY 14853

Haitham Seada
seadahai@msu.edu
Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824

Kalyanmoy Deb
kdeb@egr.msu.edu
Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824

Abstract
Despite the significant progress that has been made investigating the performance of multiobjective evolutionary algorithms (MOEAs) on well-defined test problems, it is not well known if these results translate to real world applications. Here, we contribute a real world diagnostic benchmarking framework to rigorously test modern multiobjective evolutionary algorithms. The open source benchmarking framework consists of (1) software for flexibly constructing MOEAs; (2) a suite of metrics for statistical evaluations of MOEA performance; and (3) an extensible means of collecting real-world benchmark applications that capture a range of mathematical problem difficulties. We demonstrate the benchmarking framework by carefully diagnosing the capabilities of the NSGA-II, NSGA-III, RVEA, MOEA/D, and the Borg MOEA on four real world benchmark applications with 3 to 10 objectives. Collectively, the four applications capture a range of mathematically challenging properties including stochastic objectives, severe constraints, strong nonlinearity, and complex Pareto frontiers. We demonstrate how MOEAs that have shown strong performance on standard test problems can struggle on real-world applications. Our benchmarking framework and results presented here have value for enhancing the design and use of MOEAs in real-world applications. Our results highlight the need to enhance the adaptability and ease-of-use of MOEAs given the often ill-defined nature of real-world problem solving.

Keywords
Many objective evolutionary algorithm, optimization, benchmarking, diagnostics.

1 Introduction
Significant 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 et al. (2000), and were quickly followed by the DTLZ suite (Deb et al., 2001, 2002b) 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 one can instantiate the same problem with a different
number of objectives, enabling the systematic study of many-objective optimization. The WFG suite introduced the idea of formulating problems with various characteristics by applying a series of transformations, which the authors call shape and transition functions. These are perhaps the most popular benchmarking suites currently used in the literature. More recent additions include the LZ suite using complex Pareto sets (Li and Zhang, 2009), the CDTLZ suite with constrained, scalable test problems (Deb and Jain, 2014), and the CEC2009 suite introduced at a workshop at the CEC 2009 conference (Zhang et al., 2009). These benchmarks are summarized in Table 1.

While these benchmarks capture many challenging problem characteristics, it still remains an open question if these benchmarks are sufficient to capture the complexity observed in real world applications. One significant concern arises because these numerically-defined test problems are unable to adequately capture the intricate details of complex real-world problems, which often involve running black-box simulations or non-linear computer programs. Furthermore, relying solely on artificial test problems could lead to a myopic understanding of MOEAs translational value for real world application contexts. A broader evaluation of candidate algorithmic architectures as well as their parametric sensitivities is necessary to address problems in the MOEA research community in associated with algorithmic overfitting and parametric bias.

Overfitting is a well-documented phenomenon in machine learning, where a derived model does not fit an independently-sampled validation set as well as it fits the original training set. In the context of MOEA benchmarking, overfitting relates to the breadth of algorithm architectures as well as the breadth of benchmarking test problems. Overfitting can emerge from too narrowly sampling candidate combinations of search traits as well as failing to generalize to problems not contained in common artificial test suites. In other words, focusing on a limited number of highly similar MOEAs that maximize their performance on artificial test suites. This approach is not a strong strategy for generalizing their value to diverse real-world problems. Giagkiozis and Fleming (2015) highlight that MOEA algorithm design with a strict focus on Pareto ranking deficiencies (e.g., decomposition or reference points) may be neglecting broader challenges. For example, innovations in self-adaptive or meta-heuristic multi-operator search (Burke et al., 2013; Hadka and Reed, 2013) have not been well explored in the MOEA literature to date.

Parametric bias in MOEAs is another concern that has been identified in the literature (Purshouse and Fleming (2007); Hadka and Reed (2012)) but practitioners still tend to assume that default parameter recommendations from the literature will provide sufficient level of performance. Experimental studies have shown that the ideal parameterization, or “sweet spot” (Goldberg, 2002a), for a given MOEA can change significantly or may not even exist across artificial test problems. Additionally, these studies show strong non-separable parametric sensitivities can yield difficult to predict variation in the “sweet spot” with increasing objective counts for scalable problems. Relying on recommended parameters derived from test problems could yield significant MOEA search failures on real-world applications and strongly diminish their usefulness.

It has long been recognized that no single performance metric is sufficient for fully characterizing the success or failure of MOEAs (Zitzler et al., 2003; Knowles and Corne, 2002). 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 \( \epsilon \)-indicator to identify gaps in approximation sets and evaluate 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.

Recently, a broader ensemble-based MOEA diagnostic framework (Hadka and Reed, 2012) has demonstrated the importance of globally sampling MOEAs 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 minimum number of function evaluations (NFE, “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 et al., 2013, 2015).

Given these observations, it is of paramount importance to investigate the performance of MOEAs on real-world applications. The literature contains many case studies where one or several MOEAs are applied to solve a real-world application; however, it remains difficult to aggregate results across multiple case studies due to variations in testing methodologies. Therefore, this study contributes and demonstrates an open source MOEA benchmarking framework that (1) facilitates flexibly constructing MOEA architectures; (2) supports rigorous global sampling of MOEA parameters in metrics-based evaluations capturing proximity, diversity, and consistency; and (3) provides an extensible collection of real-world benchmark applications that capture a range of mathematical problem difficulties. The benchmarking framework is meant to aid the field in developing best practices and provide a standard methodology for running MOEAs, collecting results, computing performance metrics, and statistically comparing the joint probabilistic performance results.

The four real-world applications provided for this study, range from a three-objective car crash problem up to a ten-objective general aircraft design problem. We demonstrate the benchmarking framework using five popular MOEAs, including NSGA-II, NSGA-III, MOEA/D, RVEA, and the Borg MOEA. These MOEAs are specifically chosen for their use of different mechanisms for many-objective search: Pareto dominance, reference directions, decomposition, and \( \epsilon \)-dominance with adaptive operators. Using the proposed benchmark suite, we rigorously explore the performance of each algorithm across its feasible parameter space. Doing so allows us to eliminate any parametric bias when comparing MOEAs and also explore the key parametric controls. Moreover, the choice of algorithms is sufficiently diverse as to capture a broad range of algorithmic architectures.

2 Methods

2.1 Experimental Framework

Rigorous MOEA benchmarking for real-world applications requires a flexible software framework that can be used to capture common and new algorithm architectures while also supporting ensemble-based performance diagnostics. These features are provided in the open source MOEA Framework library (Hadka, 2012) developed and managed by the lead author of this study. A brief overview of the MOEA Framework, version 2.11 follows.

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 — includ-
D. Hadka, J. Zatarain Salazar, P. Reed, H. Seada, K. Deb

ing Algorithm, Problem, Variable, Solution, and Variation — that are fundamental for implementing MOEAs. This serves to separate the task of defining specific MOEAs from the 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 MOEA Framework does not require modifications or re-building the framework itself. A new 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 automatically determine the appropriate implementation to use. 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 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 includes many popular performance indicators, including hypervolume (HV), generational distance (GD), inverted generational distance (IGD), additive epsilon-indicator (EI), spacing, and the R indicators (Zitzler et al., 2003; Coello Coello et al., 2007). As a result, it contains all the necessary modules to design, test, and analyze MOEAs.

The experimental methodology adapted in this study builds on the work of Hadka and Reed (2012). Their methodology follows the five steps described below.

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 their feasible parameter space, we can identify “sweet spots” 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 setting.

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 NFE. Search progress was recorded by outputting approximation sets every 10,000 NFE.

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, we utilize generational distance to measure convergence or proximity, hypervolume to measure both proximity and diversity, and additive epsilon-indicator to measure consistency (no large solution gaps in approximation sets).

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. Control maps provide a visual means of understanding how sensitive a 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 efficiently execute the type of real-world diagnostics demonstrated in this
study. For example, this study required approximately four days of computing resources with 256 processing cores.

2.2 Test Algorithms

In this study, we are particularly interested in the ability of modern MOEAs to solve many-objective real-world applications. As such, we have selected five representative MOEAs: 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 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 papers for in-depth details. The intent of this diagnostic assessment is to demonstrate how the default or the 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 the NSGA-II (Deb et al., 2002a). 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 hence 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 et al., 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. This combination (scalarization) is done using Tchebycheff (Miettinen, 2012) or Boundary Intersection (BI) approaches (Das and Dennis, 1998). Counter 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 neighborhood, as well as another penalty parameter $\theta$ in case of using BI approach. The original study did not include any constraints handling strategy.

2.2.3 NSGA-III

In 2013, Deb and Jain proposed the 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-objectives perspective. It employs the same non-dominated sorting used 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 neighborhood restrictions. It also uses a different normalization approach, where the extreme solutions are identified using an Achievement Scalarization Function (ASF) based 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 favors 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 uses a set of predefined evenly distributed set of reference directions, and briefly explored the idea of repositioning these directions during optimization. NSGA-III was able to solve problems having up to 20 objectives (Deb and Jain, 2014).

2.2.4 RVEA

Cheng et al. (2016) propose 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 tries to reposition reference directions to achieve better distribution when dealing with differently scaled objectives. Two additional parameters are used, namely $\alpha$ which controls the penalty function, and $f_r$ which controls the frequency of employing the adaptation strategy (normalization). RVEA shows competitive results to several state-of-the-art algorithm up to 10 objectives.

2.2.5 Borg MOEA

The Borg MOEA is characterized by auto-adapted selection pressure, population sizing, and multioperator recombination in conjunction with its epsilon-dominance archiving (Hadka and Reed, 2013). The Borg MOEA is actually a search framework that instantiates as a specific auto-adaptive search algorithm based on direct closed loop feedback on how well it is generating new dominant solutions (i.e., epsilon-progress). Its auto-adapted search leads to dynamically changing mixtures of variational operators. These emerging mixtures of search operators yield a broad array of cooperative and evolving exploration strategies to enhance search on a wide assortment of problem domains. The 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) permits the Borg MOEA to adapt to a problem’s local characteristics and adjust as required throughout its execution. Epsilon-dominance archiving serves as its method to maintain a diverse set of Pareto optimal solutions in high-dimensional space. Additionally, the Borg MOEA monitors for insufficient search progress, using a metric called epsilon-progress, and triggers adaptive time continuation (Srivastava, 2002; Goldberg, 2002a) to enlarge and inject new diversity into a stagnant population.

In addition to the issues of diversity and selection pressure, Hadka and Reed (2013) also argue that the ability to successfully and efficiently evolve candidate solutions is an important characteristic for many objective optimization. This reasoning is based on the observation that many-objective algorithms are often required to store a diverse population, where solutions spanning the Pareto frontier can exhibit large variation in phenotype and genotype, which may cause difficulty in generating fit offspring capable of surviving in the population if a single operator is selected a priori. The Borg MOEA’s auto-adaptive operator selection can help identify which variation operators were successful on a particular problem, which can also help the analyst to infer problem properties (e.g., separability of decision variables if SBX is dominant by itself).

2.3 Test Problems

While the proposed real-world benchmark is designed to be extensible, we have provided initial implementations of four real-world applications ranging from three to ten objectives. These prob-
Table 2: Real World Applications

<table>
<thead>
<tr>
<th>Name</th>
<th>Decisions (# variables)</th>
<th>Objectives</th>
<th>Epsilons</th>
<th>Constraints (#)</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car Side Impact</td>
<td>Design variables (7)</td>
<td>Min: weight, avg. vel. of V-pillar, pubic force.</td>
<td>0.5</td>
<td>Design response constraints (10)</td>
<td>Real-valued, Constrained</td>
</tr>
<tr>
<td>Lake Problem</td>
<td>Phosphorous emission controls (100)</td>
<td>Min: average daily P, Max: utility, inertia, days P crit is met.</td>
<td>0.01</td>
<td>Rel. &gt; 85%</td>
<td>Real-valued, Non-linear, Stochastic objectives</td>
</tr>
<tr>
<td>LRGV</td>
<td>Permanent rights, adaptive options contract, leases (8)</td>
<td>Min: cost, surplus water, dropped transfers, # of leases, dropped transfers cost, Max: critical reliability.</td>
<td>0.0009</td>
<td>Rel. &gt; 98% Crit. Rel = 1.0 Cost Var. &lt; 1.1</td>
<td>Real-valued, Constrained, Stochastic objectives, Disjoint PF</td>
</tr>
<tr>
<td>GAA</td>
<td>Design variables (27)</td>
<td>Min: noise, empty weight, cost, ride roughness, fuel weight, price. Max: travel range, reuse, lift-to-drag ratio, cruising speed.</td>
<td>0.15</td>
<td>Aggregate performance constraint (1)</td>
<td>Real-valued, Constrained Non-separable decisions</td>
</tr>
</tbody>
</table>

Table 2 also shows the $\epsilon$ values used by the Borg MOEA for $\epsilon$ 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 3-objective application with 7 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 et al. (1999)). In this decision problem, the economic needs of a hypothetical town emitting phosphorous into a shallow lake system must be balanced with the desire to avoid breaching a threshold where the lake becomes irreversibly polluted. The Lake Problem requires the specification of annual phosphorous emission control decisions for the town made over a 100 year period while seeking to maximize economic benefits, minimize drastic inter-annual changes in emission rates, maximize the reliability of not crossing the known pollution tipping point, and minimize the resulting phosphorous levels in the lake. From an MOEA perspective, this real world benchmarking application presents severely nonlinear threshold behavior and its 100 variable decision space is strongly prone to genetic drift failures (i.e., a lack of selection pressure on the least important decisions, see theoretical discussion of temporal salience structure as a heuristic search challenge in...
D. Hadka, J. Zatarain Salazar, P. Reed, H. Seada, K. Deb

Thierens et al. (1998)). Readers interested in a more detailed summary of this benchmarking application’s formulation and implementation can see the following studies: 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). The real-valued decision variable formulation abstracts the city’s choice among three sources of supply: (1) a percent of reservoir inflows (termed non-market permanent rights), (2) monthly leases of water from regional irrigation districts with volatile demand-based pricing, and (3) adaptive options contracts that guarantee fixed prices for a given volume being made available in drought periods. A key constraint for the LRGV application enforces that the city always avoids catastrophic water supply shortfalls where supply is more than 40% below demands (i.e., catastrophic failures). Monte Carlo-based model evaluations are used to compute the resulting constrained 6-objective performance of alternative mixtures of market-based and traditional reservoir sources for the water supply. The LRGV’s minimization objectives include expected costs, cost variability, the volume surplus water, unused volumes of purchased market water, and the discrete count of water leases used. The sixth objective maximizes the system’s reliability in meeting demands. Prior LRGV MOEA benchmarking efforts (Reed et al., 2013a) ran a random 500,000 member Latin Hypercube Sampling of the problem’s decision space to assess the difficulty of its constraints. The Latin Hypercube Sampling generated a success rate of 0.06% in generating feasible solutions and the identified feasible solutions where strongly dominated. Beyond its constraints, the LRGV problem’s best known Pareto front has a strongly disjoined geometry where two very different families of solutions reside (i.e., high surplus water, expensive non-market reservoir dominated supplies versus low surplus, less expensive water market dominated, see Kasprzyk et al. (2016, 2009)). Complex disjoint Pareto front geometries often emerge in real world applications where discrete families of different solution types emerge (e.g., discrete levels of capital cost investment). Detailed formulation and implementation information of the LRGV benchmarking application are presented in Kasprzyk et al. (2012).

2.3.4 General Aviation Aircraft

The General Aviation Aircraft (GAA) problem has as its origin a United States National Aeronautics and Space Administration aircraft design challenge from the 1990s (Simpson and D’Souza, 2004; Woodruff et al., 2013). The present form of the GAA problem presented in this study aims to construct a product family of three general aviation aircraft. As a product family design problem, economies of scale are sought for the manufacturing of the three aircraft by maximizing the re-use of existing components (i.e., “commonality of parts”), while addressing 9 additional performance objectives. The minimization performance objectives include takeoff noise, empty weight, direct operating cost, ride roughness, fuel weight, and purchase price. The three remaining maximization performance objectives measure travel range, lift-to-drag ratio, and cruising speed. Given that each candidate product family is composed of three aircraft each of which are characterized by the 9 performance objectives, the GAA problem presents a severe dimensional challenge with 27 performance measures and 1 commonality objective. Following the approach first presented by Shah et al. (2011), the GAA problem variant explored in this study reduces the objective dimension by making each of the 9 performance objectives a minimax across the aircraft. The worst performing aircraft in any of the 9 performance measures defines the reported objective measure. Although the minimax formulation has the advantage of reducing objective count, it does however introduce non-separable decision variable interactions across the three aircraft. In this study, we are designing three aircraft to seat 2, 4, and 6 occupants each of which are defined by 9 real-valued decision variables yielding a total of 27 decision variables. Moreover, the GAA problem presented is highly constrained because the individual aircraft have 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 et al., 2011), a Latin Hypercube Sampling of 50-million candidate GAA designs only
yielded 3 feasible solutions, demonstrating its severe difficulty. Readers interested in a more detailed description and formulation of the GAA problem as benchmarked in this study can reference the following studies: Shah et al. (2011) and Woodruff et al. (2013).

3 Results

3.1 Reference Set Contributions

Figure 1 provides the tested MOEAs’ percentage contributions to the four applications’ best known Pareto approximate reference sets. The contributions designate both reference set members identified by multiple algorithms as well as solutions uniquely contributed by a single algorithm. The final reference sets were developed across all runs of all algorithms. Each algorithm’s contributions were accumulated across all of its tested runs. The final reference sets were $\epsilon$-non-dominated sorted to maintain a mathematically consistent ranking definition across the point and block dominance approaches. The $\epsilon$ values for generating the final reference sets are shown in Table 2. The Car Side Impact problem was the easiest of the applications tested and had the highest proportion of its reference points that were identified within the same epsilon-box across the MOEAs. The Borg MOEA attained the best overall performance with 100% identified and MOEA/D the worst with 90%. Overall NSGA-II and the reference point MOEAs all attained very similar performance, which is not surprising given the large sample of runs and the relative ease of the application. In Figure 1, contributions for the Lake Problem’s reference set were significantly worse. The Borg MOEA (45%), MOEA/D (30%), and NSGA-II (22%) were the three top performing algorithms. Perhaps surprising here is the limited contributions from NSGA-III and RVEA, suggesting that the reference point frameworks struggled in their best overall convergence results. The Lake Problem as discussed in the prior sections presents two primary mathematical difficulties: (1) the underlying control decisions are subject to a highly nonlinear irreversible threshold where the system is permanently polluted and (2) the 100 decision variables have potentially a diminishing impact on the system performance (i.e., temporal salience structure and domino convergence, see Thierens et al. (1998)). Of these challenges, domino convergence and potential selection pressure failures have been observed in prior diagnostic assessments (Ward et al., 2015).

Figure 1: Percent contribution of each MOEA to the reference set for the four real-world applications best-known approximate reference sets.

Alternatively, the LRGV application has a much lower dimensional decision space but poses challenges due to its highly constrained 6-objective stochastic formulation and disjoint Pareto reference set. Prior MOEA diagnostics exploring older algorithms have shown that MOEAs often strug-
ingle with diversity maintenance and convergence on this problem (Reed et al., 2013b). The Borg MOEA had a pronounced advantage in contributing approximately 75% of the reference set solutions. MOEA/D and RVEA performed similarly contributing approximately 15%. Both NSGA-II and NSGA-III struggled to contribute reference set solutions. The higher dimensional disjoint Pareto set for the LRGV poses deterioration challenges for reference point methods and scaling challenges for MOEA/D’s Chebyshev weighting schemes used in its decomposition strategy. The ten objective highly constrained GAA test yielded the starkest contrast in the reference set contribution results of Figure 1. Although the suite of algorithms tested in this study were designed for many-objective optimization, the GAA problem posed severe challenges for the non-adaptive MOEAs. As a meta-heuristic algorithm, the Borg MOEA uses feedback in its search process to dynamically adjust its population size, selection pressure, and mixtures of variational operators. This endogenous feedback clearly distinguished its contribution of 95% of the GAA reference set. Although theoretical limits of Pareto ranking (Teytaud, 2006, 2007) have motivated a rapid growth of decomposition and reference point MOEAs, the GAA results of Figure 1 highlight that high dimensional real world problem contexts pose severe challenges for these approaches. As highlighted in prior assessments (Hadka and Reed, 2012; Woodruff et al., 2012), the GAA problem is severely constrained with non-separable air craft family design decisions that strongly activate the Borg MOEA’s auto-adaptivity, especially in terms of its dynamic use of mixtures of variational operators. Although the reference set contribution results shown in Figure 1 provide a coarse portrait of the relative ease or difficulty of the applications, they lack specificity in understanding modes of search failure in terms of the convergence, proximity, diversity, and consistency of the approximation sets attained from the evaluated MOEAs.

3.2 Best Performance & Attainment Probabilities

The attainment plots shown in Figure 2 provide a probabilistic metrics-based comparison of the five MOEAs across the four real world test applications. Figures 2a-2c show the best single run attainment results for the generational distance ($GD_{BR}$, “convergence”), $\epsilon$-indicator ($EI_{BR}$, “consistency”), and hypervolume ($HV_{BR}$, “proximity and diversity”) metrics, respectively. In panels a-c of Figure 2, 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 shown in Figure 2 represents the best metric value attained by a single trial run across all of the test MOEAs’ parameterizations and random seed trials. For each application and metric, the attainment plots provide 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 $GD_{BR}$ 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 2a, there is very little difference in the MOEAs’ best run and $GD_{BR}$ attainments for the Car Side Impact and Lake Problem test cases. For the LRGV, the MOEA’s best run results are very similar. The $GD_{BR}$ attainments for the LRGV do show reduced attainment probabilities across the MOEAs. RVEA and NSGA-II maintain at least 90% attainment probabilities up to 85% of the best $GD_{BR}$ 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 $GD_{BR}$ value, its attainment probabilities decrease far less abruptly at the highest levels of $GD_{BR}$ performance. Both MOEA/D and NSGA-III show substantially lower $GD_{BR}$ attainments. The LRGV problem’s disjoint Pareto set poses a challenge to 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., 2013b). The GAA problem shows the most pronounced differences in $GD_{BR}$ attainments in Figure 2a. Although all five MOEAs had at least a single best run with high $GD_{BR}$ 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 $GD_{BR}$ metric values.
Figure 2: Best performance and attainment probabilities of each MOEA on the four real-world applications.
The $\text{EI}_{BR}$ attainment plots in Figure 2b show that gap free, highly consistent approximations of the four applications is far more challenging than the $\text{GD}_{BR}$ attainment. The 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 $\text{EI}$ 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 similar $\text{EI}_{BR}$ attainment performance for the Car Side Impact application. The Borg MOEA again has the top $\text{EI}_{BR}$ attainment performance for the Lake Problem. All of the MOEAs struggled to reliably attain results beyond 80% of the best overall reference $\text{EI}_{BR}$ metric value. MOEA/D and NSGA-III showed the worst overall $\text{EI}_{BR}$ attainment performance on the Lake Problem, which is also reflected in the $\text{GD}_{BR}$ attainments in Figure 2a as well. For the LRGV $\text{EI}_{BR}$ attainment probabilities and the best single run results in Figure 2b, all of the MOEAs show a strong decline in performance relative to the $\text{GD}_{BR}$ attainment results in Figure 2a. RVEA shows strong reliability up to approximately 40% of the best $\text{EI}_{BR}$ metric value and then declines very rapidly. The Borg MOEA maintains a 70% probability of attainment up until approximately 65% of the best $\text{EI}_{BR}$ value.

As was discussed above for the reference set contributions in Figure 1, the LRGV is a constrained application with a disjoint reference set of solutions. In this application, a larger and potentially easier to identify region of attraction is typically found much more quickly than an isolated grouping of solutions that are composed of different non-market water supply options (i.e., two distinct families of solutions). The $\text{EI}_{BR}$ metric is particularly sensitive to the gap generated by missing the second grouping of solutions. None of the algorithms are capable of avoiding gaps in a single trial run for the LRGV. MOEA/D, NSGA-II, and NSGA-III all struggle for the full range of $\text{EI}_{BR}$ attainment for the LRGV. The GAA $\text{EI}_{BR}$ attainments in Figure 2b highlight the challenge that algorithm selection presents. In this case RVEA’s $\text{EI}_{BR}$ attainment for the GAA problem is the worst overall, showing failure over the full range of potential $\text{EI}_{BR}$ values. The Borg MOEA maintains some level of reliable $\text{EI}_{BR}$ attainment across all of the tested applications. For the GAA, the Borg MOEA maintained the highest $\text{EI}_{BR}$ attainment probabilities for the best overall metric values. Comparing the $\text{EI}_{BR}$ attainments in Figure 2b and the $\text{HV}_{BR}$ attainments in Figure 2c. Generally, the same trends in performance are seen for all but the GAA test problem.

The $\text{HV}_{BR}$ attainment results for the GAA in Figure 2c strongly contrast the $\text{EI}_{BR}$ attainments in Figure 2b. The Borg MOEA was capable of producing both proximate and diverse representations of the GAA applications’ ten objective tradeoffs. All of the other algorithms struggle in their $\text{HV}_{BR}$ attainments for the GAA. In combination, the $\text{EI}_{BR}$ attainments in Figure 2b and the $\text{HV}_{BR}$ attainments in Figure 2c for the GAA highlight that while NSGA-III was able to develop a relatively consistent, gap free local approximate set, it was not sufficiently proximate to the actual reference set to register modest $\text{HV}_{BR}$ performance. The Borg MOEA while not without weaknesses when addressing the GAA, with a single run it clearly distinguished itself in its performance to address an extremely constrained, severely difficult many-objective problem.

### 3.3 Efficiency & Controllability

As was discussed in the Methods, the diagnostic framework employed in this paper explores a 500 member Latin Hypercube sampling of each MOEA’s full parameterization space. The probabilistic attainments in Figure 2 provide a summative statistical assessment of search effectiveness and reliability across each MOEA’s sampled parameterizations and random seed trials. Although statistically rigorous, the attainment results are less reflective of the typical operational use of the MOEAs. In practice, one often does not have sufficient time or resources to perform an extensive analysis of an algorithm’s parameterizations. Instead, one desires an MOEA that exhibits reliable performance and fast convergence without the need for fine-tuning settings. In other words, they want an MOEA that is efficient and controllable. The control maps of Figure 3 characterize the controllability, effectiveness and efficiency of the MOEAs using the expected $\text{HV}_{REF}$ performance. $\text{HV}_{REF}$ is defined relative to the best known reference sets. Each control map plot in Figure 3 reports the expected $\text{HV}_{REF}$ results attained for a given parameterization’s 50 random trials for a given algorithm. Although the 500 member Latin Hypercube samples of the MOEAs’ parameterization spaces explore more than...
the NFE and search population size, the control maps in Figure 3 illustrate important performance subspaces. NFE captures the efficiency of the algorithm trials. The search population or reference point divisions capture the algorithms’ sensitivity to diversity maintenance and deterioration, especially as the test application’s objective counts increase. Ideal control maps for each algorithm would be solid dark blue shading for the highest expected HV$_{REF}$ performance (i.e., effective performance regardless of the parameterization employed). In real world applications, MOEA parameter sensitivities are strongly problematic and expensive in terms of actual staff costs of the time spent seeking to find non-generalizable zones of high performance (i.e., seeking to find a parameterization that works). MOEA parameters are not a decision relevant focus for real-world applications. Moreover, real world problems typically require an iterative progression through candidate formulations (Kasprzyk et al., 2012; Walker et al., 2003; Zeleny, 1989), amplifying the real-world operational costs of MOEA parameter sensitivities. Algorithmically, real world decision contexts need MOEAs where it is difficult to make them fail versus being dependent on trial-and-error analyses seeking parameterizations that work.

In Figure 3, the control maps for the Car Side Impact problem show that the Borg MOEA attains nearly ideal performance across its tested parameterizations for NFE values beyond 10,000. Although the Car Side Impact problem is easiest of the tested applications, the HV$_{REF}$ attainments in Figure 2c and the control maps in Figure 3 highlight the potential for strong parametric sensitivities and reduced reliability for the four remaining algorithms. MOEA/D shows lowest expected HV$_{REF}$
performance regardless of NFE count. Figure 3 shows that NSGA-II is the second most effective algorithm for the Car Side Impact problem. RVEA and NSGA-III have increased parametric sensitivities (i.e., reduced controllability) and a relatively difficult to predict or understand dependence on NFE count. A core difference between NSGA-II, NSGA-III, and RVEA from the Borg MOEA is the nature of how parametric sensitivities are managed. The Borg MOEA has several epsilon-dominance-based feedbacks where the algorithm auto-adapts its selective pressure, its population size, its combined use of multiple variational operators, as well as the diversity of its search population. The feedbacks are endogenous to the algorithm and do not require user interaction. Alternatively, high performance for the other MOEAs on the Car Side Impact problem requires exogenous user trial-and-error experimentation to navigate the parametric sensitivities influencing search performance.

In Figure 3, all of the tested MOEAs have far more complex control maps for the Lake Problem versus the Car Side Impact results. All of the control maps for the Borg MOEA show a clear trend of reducing variability and improving HV performance when increasing the NFE count beyond 20,000. NSGA-III and RVEA show a more complex relationship between expected \( \text{HV}_{\text{REF}} \) performance and their parameter settings relative to the Borg MOEA. The Borg MOEA’s controllability improves much more clearly with NFE (i.e., parametric “sweet spots”, Goldberg (2002b)). MOEA/D and NSGA-II show complex control maps for the Lake Problem in Figure 3 indicating that far more search would be necessary. They have a very broad range of variation for NFE requirements before improving their expected HV performance. Users seeking to understand or find these zones would be distracted from the original intent of solving a real world problem and would struggle with the non-separable nature of MOEA parameter sensitivities (e.g., see Hadka and Reed (2012)).

The LRGV test problem has already shown to be quite difficult for all of the tested MOEAs. Consequently, the range of expected \( \text{HV}_{\text{REF}} \) performance plotted in Figure 3 for the LRGV problem is substantially reduced. Additionally, the LRGV includes Monte Carlo-based function evaluations, making it the most computationally demanding of the four real world benchmarking problems. The Borg MOEA and RVEA achieved the highest level of controllability for the LRGV problem, showing clear improvement in both expected \( \text{HV}_{\text{REF}} \) and controllability with increasing NFE for this problem. The LRGV results in Figure 3 show that MOEA/D, NSGA-II, and NSGA-III are not controllable when solving the problem within the allocated computational time.

Figure 3 highlights that the ten-objective GAA problem is the most difficult of the four problems in many respects (i.e., most severely constrained, highest dimension objective space, non-separable decisions, etc.). The single trial run expected \( \text{HV}_{\text{REF}} \) performance is very limited for all four of the tested MOEAs. The Borg MOEA has substantially higher efficiency, effectiveness, and controllability relative to the algorithms. At 50,000 NFE the Borg MOEA begins to contribute \( \text{HV}_{\text{REF}} \) performance and shows a predictable improvement of performance as the NFE count increases. Surprisingly, the decomposition and reference point strategies designed to overcome failures in Pareto ranking appear to struggle on the GAA application. Prior studies (Hadka and Reed, 2012; Shah et al., 2011; Woodruff et al., 2012, 2013) have shown that the auto-adaptive population sizing and time continuation (i.e., injecting diverse solutions) substantially enhances MOEA performance on the GAA. Moreover, Hadka and Reed (2012) also showed through detailed sensitivity analyses and runtime analysis that the Borg MOEA’s cooperative, adaptive use of multiple variation operators, including those that specifically address non-separability, also enhanced the Borg MOEA’s success in solving the GAA problem.

As a general note, the controllability results of Figure 3 as well as findings from prior diagnostic studies that carefully explored MOEA parameter space sensitivities (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 relatively 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

Although the diagnostic results of Figures 1-3 above are useful for understanding the overall efficiency, effectiveness, reliability, and controllability of the four tested MOEAs, they do not represent the typical operational use of MOEAs. A vast majority of MOEA related applications and studies employ the author recommended default parameterization and run algorithms for multiple random seed trials. Figure 4 reports the default parameterizations’ HV$_{BR}$ runtime search dynamics for each of the five algorithms solving the four real world benchmarking problems. Panels (a)-(d) provide a relativistic view of the efficiency (i.e., NFE) and reliability (i.e., the 90th interquantile range) of the standard implementations of the algorithms relative to the best single MOEA runs for each of the applications. Figure 4a shows that the default implementation of the Borg MOEA very rapidly (<10,000 NFE) and reliably attains high HV$_{BR}$ results for the Car Side Impact problem. The worst default parameterization random seed trial of the Borg MOEA for the Car Side Impact exceeds the performance of all the other algorithms. NSGA-II and NSGA-III also efficiently and reliably solve the Car Side Impact problem, although they never attain the HV$_{BR}$ performance of the Borg MOEA. RVEA required nearly double the NFE to attain (>20,000 NFE) Car Side Impact HV$_{BR}$ results that were competitive. Interestingly, RVEA exhibits a moderate decrease in hypervolume as the run progresses. This demonstrates clearly the problem of deterioration, where the selection strategy results in replacing population members with inferior alternatives with respect to a performance indicator. MOEA/D is comparatively the worst performing MOEA on the Car Side Impact problem when using its default parameterization.

Figure 4: Runtime dynamics for each MOEA using their default parameterizations across the four real-world applications.

The default parameterization Lake Problem results in Figure 4b highlight very strong differences in the algorithms’ runtime dynamics. Both NSGA-III and the Borg MOEA very rapidly attain high levels of HV$_{BR}$ results (15,000-20,000 NFE), where NSGA-III is more reliable earlier in the search dynamics. RVEA is less efficient, effective and reliable relative to the Borg MOEA and NSGA-III for the
Lake Problem. It’s best random seed trials are less effective than worst trials of the top performing algorithms. Nonetheless, RVEA outperforms both NSGA-II and MOEA/D. The default implementation of MOEA/D’s search progress is strongly delayed, requiring more than 50,000 NFE before making substantial HVBR progress. As discussed in the reference set contributions of Figure 1, the Lake Problem is strongly prone to selective pressure failures due to the decreasing effects of late period decision variables. The strong separation of the algorithms’ default runtime dynamics in Figure 4b likely reflect their ability to sustain selection pressure. The Borg MOEA monitors the relative size of its archive to its search population size to auto-adaptively trigger increases in both the search population size as well as tournament sizes in its selection operator (Hadka and Reed, 2013).

The default runtime dynamics for the LRGV in Figure 4c reinforce that the problem is challenging and that real world practitioners would be confronting significant uncertainty due to the algorithms’ HVBR performance. The 90th inter-quantile ranges of their random seeds show a tremendous degree of variability where the mean runtime traces are often poor reflections of performance. Prior LRGV MOEA diagnostic studies (Kasprzyk et al., 2016; Reed and Kollat, 2013) have shown that often the best attained reference sets for the problem are the result of a small number of “super seed trials” that capture both zones of the application’s disjoint set. This is also evident when considering the reference set contributions of Figure 1 and the Borg MOEA’s default parameterization runtime dynamics in Figure 4c. The Borg MOEA’s mean HVBR runtime trace is dramatically lower than its top performing trials. That being said, if users were developing a local reference set for the LRGV application using only a single algorithm and the 50 random seed trials illustrated in Figure 4c, the Borg MOEA would yield a far better representation of the LRGV application’s tradeoffs for subsequent decision support. NSGA-III also shows similar behaviors to the Borg MOEA, although its highest performing random seed trials in Figure 4c show reduced relative performance. MOEA/D has the most variability in its performance between its mean HVBR runtime dynamics versus its best random seed trials. This is reflective MOEA/D’s high performing best single trial and extremely low HVBR attainment probabilities for the LRGV illustrated in Figure 2c. RVEA displays the best mean HVBR default parameterization search trace in Figure 4c, but its top performing random seed trial runs are worse than all of the other algorithms’ except for NSGA-II. RVEA’s default parameterization runs converge quickly to the largest zone of solutions in the LRGV’s disjoint Pareto set, but never discovers the problem’s other isolated region of solutions. In the real world decision support context, even a local reference set developed across all of RVEA’s 50 default random seed trials would miss an entire zone of decision relevant solutions (i.e., a poor representation of candidate actions and their tradeoffs).

Lastly, the GAA runtime HVBR dynamics for the algorithms in Figure 4d show only the Borg MOEA as having any capability of solving this ten objective, heavily constrained real world benchmark problem. The Borg MOEA’s worst default parameterization random seed trials outperformed the best trials from all of the other algorithms. Figure 4d shows that NSGA-III is the only other MOEA whose default parameterization made any HVBR progress. As a hyper-heuristic, the Borg MOEA is substantially different in its ability to exploit epsilon-dominance archiving feedback to auto-adapt its population size, its selective pressure, the diversity of its search, and the combination of variation operators used to search the GAA problem’s challenging decision space. Collectively, Figures 4a-4d show that the 50 random seed trials of the default parameterization for the Borg MOEA would provide the best approximation of the benchmark real world problems’ tradeoffs. As noted in the discussion of Giagkiozis and Fleming (2015) the assumption that MOEA limitations are addressed through transitioning away from Pareto ranking schemes fails to acknowledge the strong dependence of search progress on a myriad of other failure modes (i.e., population diversity, topological complexities, drift stall from a lack of selective pressure, etc.). Moreover, as MOEAs are used in actual decision contexts they need to be flexible, adaptable, and easy-to-use. Although decomposition techniques and reference point frameworks have received significant recent focus by the MOEA research community, the results of this study motivate the need to move beyond traditional non-adaptive search mechanisms.
4 Conclusion

In this study, we demonstrated a new open source real-world benchmarking framework. This framework, while intended to be extensible to allow inclusion of additional real-world applications, currently contains four benchmarking problems ranging from three up to ten objectives. This has allowed us to conduct a rigorous comparison of five popular MOEAs on complex, real-world problems.

The results clearly indicate disparity in the performance among the different MOEAs, with the Borg MOEA being the strongest contender across all problems. Still, there is significant diversity in the observed results, which suggests that future work in expanding this benchmarking framework is necessary. These preliminary results give insight, however, into successful algorithmic design choices. First, the choice of selection strategy and/or archive mechanism must avoid deterioration and provide sufficient selection pressure and maintain genetic diversity throughout the entirety of a run. Epsilon-dominance archiving and reference points appear to be successful in this regard. Second, and often overlooked, is the need to choose genetic operators that are stable and flexible across a range of problem domains in order to drive search reliably towards the Pareto front. Historically, the choice of genetic operator has been the responsibility of the MOEA practitioner, but recent advances in hyper-heuristics and multi-operator MOEAs provides an automated approach. Lastly, many studies have investigated selection and genetic operators independently, when they should be considered mutually important.

For situations in which the entire Pareto front is desired, these results indicate that \( \epsilon \)-non-dominated sorting is sufficient to maintain genetic diversity and selection pressure on problems up to ten objectives. They also demonstrate the benefit of auto-adaptive search operators which can adapt to the search landscape across an assortment of real-world applications. Reference point/vector methods have grown in popularity, but these results show signs of deterioration at higher dimensions. Further research is needed to isolate the root cause of deteriorating performance.

It is also apparent that existing test benchmarks are inadequate for distinguishing MOEAs based on performance (Hadka and Reed, 2012). Substantial emphasis is placed on gaining marginal improvements on a suite of test problems, while subpar performance persists in real-world applications. The benchmarking framework and results of this study are meant to initiate broader collaborative efforts in the field, including growing a diverse suite of candidate real world benchmarking problems. Although this study focused on standard implementations of existing MOEAs, the MOEA Framework underlying our work will permit future studies to also flexibly explore new or mixed algorithm architectures as part of benchmarking efforts. Our 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.

Acknowledgment

The Cornell authors were funded in this work by U.S. National Science Foundation (NSF)’s Innovations at the Nexus of Food, Energy and Water Systems (INFEWS) program (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.

References


