Utilizing Innovization to Solve Large-scale Multi-objective Chesapeake Bay Watershed Problem

COIN Report Number 2023001

Gregorio Toscano*, Hoda Razavi†, A. Pouyan Nejadhashemi†, Kalyanmoy Deb*, and Lewis Linker‡

*Electrical and Computer Eng.
Michigan State University
East Lansing, U.S.A.
toscano5@msu.edu, kdeb@egr.msu.edu

† Biosystems and Agricultural Eng.
Michigan State University
East Lansing, U.S.A.
{razavi,pouyan}@msu.edu

‡ Chesapeake Bay Program Office
Environmental Protection Agency
Annapolis, U.S.A.
Linker.Lewis@epa.gov

***Abstract*—*Innovization* is a task for analyzing multiple Pareto-optimal solutions obtained by an evolutionary multi-objective optimization (EMO) algorithm to extract common features in the decision variables, leading to design rules or solution principles. The principles derived from *innovized* principles can provide valuable insights to the users about "how to create an optimal solution?". Manual or automated machine learning-based innovization methods were proposed in the literature to extract innovized principles in a problem. Although different problems may demand different structures of the rules, the innovized rules can also be utilized to improve the performance of the subsequent iterations of the optimization algorithm or help in executing an efficient re-optimization of the same problem. In this paper, we consider a large-scale and multi-objective complex optimization task of minimizing cost and nitrogen loading in certain counties within the Chesapeake Bay Watershed (CBW) and find multiple trade-off solutions using the NSGA-III approach applied to the CBW's real evaluator tool (The Chesapeake Assessment Scenario Tool–CAST). 205 Best Management Practices (BMPs) are considered to be implemented at each land-river segment within a county, leading to as many as 65,260 variables for the resulting multi-objective optimization procedure. First, hundreds of trade-off solutions found by the CAST-NSGA-III procedure are analyzed manually to find the top-most BMPs used in them. After that, a re-optimization of CAST-NSGA-III is run with a few critical BMPs (resulting in a decrease of the variable to a range between 3% and 33%) found to commonly appear in the trade-off solution set of the previous runs. Interestingly, the resulting trade-off front with reduced BMPs is similar to the original run achieved with tens of thousands of variables. The findings are intriguing and demonstrate the efficacy of innovation in addressing intricate, real-world issues at a significant scale.**

***Index Terms*—Innovization, Watershed Management, Large-scale Optimization, Multi-objective Optimization**

I. Introduction

Most practical search and optimization problems involve more than one conflicting objective. By definition, these problems have not one but multiple trade-off Pareto-optimal solutions. While conventional methods scalarize multiple objectives into a single criterion [1], these methods suffer from several shortcomings. First, the scalarization method requires additional preference information of objectives over the entire search space, which is challenging to provide before even a single optimal solution is found. Second, to arrive at a representative set of Pareto-optimal solutions, the scalarized problem must be solved multiple times independently, each with a different parameter setting without any gain in computational effort from multiple runs.

Recent population-based optimization methods, such as evolutionary multi-objective optimization (EMO) algorithms, have been shown to find and store multiple Pareto-optimal solutions in a single application [2], [3]. This is one sole reason why these methods have become increasingly popular. A by-product of finding multiple Pareto-optimal solutions is that, being optimal, these solutions are likely to possess certain common properties among their variables, objective, and constraint values. Finding Pareto-optimal solutions and analyzing them to reveal common properties was termed as the task of *innovization* [4]. The name is so given due to the fact that the common properties often lead to new and innovative properties involving problem parameters [5]–[7]. When the Pareto-optimal solutions are analyzed to unveil such common hidden properties (or rules), the rules are helpful as knowledge, which can then be used to enhance future optimizations of similar problems.

The Chesapeake Bay is the largest estuary in the United States and the third-largest in the world. As a result, the Bay has enormous historical, social, economic, and ecological importance, with natural benefits estimated at more than $100 billion per year [8]. With a drainage area of about 166,000 km^2, the Chesapeake Bay Watershed (CBW) includes parts of six states in the Mid-Atlantic region and is home to more than 18 million people. Since the middle of the twentieth century, human activities such as livestock and crop production, urban development, and stream alteration have resulted in nutrients and sediment excess in waterbodies throughout the watershed, causing water quality impairment, freshwater ecosystems' degradation, and loss of recreational values. To address these issues, the Chesapeake Bay Program (CBP) partnership has coordinated restoration efforts since 1984 [9]. During the last decade, these efforts have been guided by the Chesapeake Bay Total Maximum Daily Load (TMDL), which established limits to nutrients and sediment loadings. The TMDL has been used to formulate comprehensive restoration plans known as Watershed Implementation Plans (WIPs) and outline major

This work is funded by U.S. Environmental Protection Agency under Grant Number 96373801.

goals, timelines, and expected outcomes as established in the CBW Agreement in 2014 [9].

Due to the complexity of evaluating restoration plans' effectiveness, watershed management in the CBW has been extensively supported by modeling tools.The Chesapeake Assessment Scenario Tool (CAST) is the core modeling tool that reports management scenarios' impacts on reducing pollution loads. Those management scenarios (i.e., portfolios of Best Management Practices, BMPs) have been primarily defined based on expert elicitation. Recently, single-objective, gradient-based optimization has been used to identify cost-effective BMPs portfolios to improve water quality in the CBW [10], [11]. Since deciding on the implementation of these portfolios involves many stakeholders (e.g., government agencies, commercial entities, nonprofit organizations, academic institutions) at multiple spatial scales (e.g., county, state), a multi-objective optimization approach exploring trade-offs among conflicting objectives was followed in an earlier study [12].

This study links a specific evolutionary many-objective (EMaO) algorithm – NSGA-III[1] [13] – with the CAST evaluation software directly using RESTful APIs developed here. RESTful APIs allow an easy update and use of instruction codes from different programming languages. In a major application area involving thousands of users, it becomes essential to have coding flexibility and the generality of such linking APIs. Since the CAST evaluation system is managed by different individuals and hosted in a secured location different from the developers and the location of the NSGA-III code, developing the CAST-NSGA-III framework became challenging. In addition, the highlight of this paper is the data analysis study to extract practical innovized rules from the CAST-NSGA-III solutions applied to a few counties in West Virginia, USA. After the rules are extracted, they help reduce the original problem's search space size. A re-optimization based on the learned rules has produced similar results as the original study. The reduced search space and computational time promise to extend the study to the watershed level.

In the remainder of this paper, we briefly describe the CBW optimization problem formulation in Section II. Then, the linking of an EMO algorithm (NSGA-III) with the CAST evaluation tool through API development is discussed in Section III. Next, the innovization procedure and related existing studies are presented in Section IV. Results of CAST-NSGA-III on the CBW optimization problem are discussed in Section V, by discussing the manual innovization procedure adopted for this study and demonstrating that the extracted innovized principles help solve the same CBW problem quickly and accurately. Finally, conclusions are drawn in Section VI.

II. Chesapeake Bay Watershed Optimization Problem for Multiple Objectives

The CBW spans parts of six states in the northeastern United States. This large drainage area encompasses more than 100,000 tributaries. These tributaries bring a large amount of nitrogen, phosphorus, and sediment to the Bay each year. As a result, the water quality conditions of the Bay have been deteriorating to the point where a large area of the Bay is uninhabitable for the aquatic ecosystem's life and endanger human health. To address these issues, it is essential to implement BMPs throughout the watershed to enhance the water quality while minimizing costs. However, due to the size of the watershed and the socioeconomic complexity throughout the region, the optimization problem is difficult to solve as it consists of several objectives (e.g., nitrogen and phosphorus load reductions), different scales (e.g., county, state), and have many constraints (e.g., BMP implementation area, restoration monetary resources).

Regarding the search space, each county involves an average of 10 land-river segments in which 2,000 BMPs can be implemented within a single land-river segment. In addition, each BMP is expected to have a maximum of 50 land use options. Therefore, when considering a single county, the variables can easily reach 20,000, resulting in a huge search space of $50^{20,000}$ different management options. Therefore, a multi-objective optimization method that is efficient and reliable is crucial to address the challenges mentioned above. Furthermore, the results can further narrow the search space and produce knowledge (i.e., innovization) to help with the future BMP implementation plan.

We provide a general formulation of the optimization problem. The number of variables and constraints depends on the CBW's area of interest (e.g., county, state, watershed) [12]:

$$
\begin{aligned}
&\text{Min.} f_1(\mathbf{x}) = \sum_{s\in S}\sum_{h\in H_s}\sum_{u\in U}\sum_{b\in B_u} \tau_b x_{s,h,u,b},\\
&\text{Min.} f_2(\mathbf{x}) = \sum_{s\in S}\sum_{h\in H_s}\sum_{u\in U}\left[\alpha_{s,h,u}\phi_{s,h,u}\prod_{\mathbf{G}^{s,h,u}\in\mathcal{G}}\left(1-\sum_{b\in\mathbf{G}^{s,h,u}}\eta^N_{s,h,b}\frac{x_{s,h,u,b}}{\alpha_{s,h,u}}\right)\right],\\
&s.t. \sum_{b\in\mathbf{G}^{s,h,u}} x_{s,h,u,b} = \alpha_{s,h,u}, \quad \forall s\in S, h\in H_s, u\in U_s, \mathbf{G}^{s,h,u}\in\mathcal{G},\\
&\qquad x_{s,h,u,b}\geq 0, \quad \forall s\in S,\ h\in H_s,\ u\in U_s,\ b\in B_u.
\end{aligned}
\tag{1}
$$

The variable $x_{s,h,u,b}$ indicates the acres used for implementing a BMP b to reduce a load-source (LS) u on an agency h and land-river segment (LRS) s. We describe more about variables in the next paragraph. Objective $f_1(\mathbf{x})$ defines the total cost of implementing BMPs, while objective $f_2(\mathbf{x})$ estimates the nitrogen loading released to the environment. The parameter τ_b indicates the cost per unit acre of implementing BMP b. $\eta^N_{s,h,b}$ refers to the efficiency of BMP b in removing nitrogen when applying on agency h and LRS s. α refers to the available acres, and the group of BMPs $\mathbf{G}^{s,h,u}$ contains all BMPs from $\mathcal{G}$ that can be applied on a given (s,h,u). All fixed parameter values, such as unit cost, available acreage, etc., are chosen from CBW's documentation. For efficient management of CBW, our optimization goal is to develop BMP allocations that minimize the implementation cost and the nitrogen loading at a pre-specified area of interest (counties or a cluster of counties).

For every LRS-Agency-LS (s,h,u) combination (known as a parcel), multiple BMPs are chosen from a group of allowable non-overlapping BMP groups $\mathbf{G}^{s,h,u}$. The variable $x_{s,h,u,b}$ indicates the acres of the BMP b appearing in one of the BMP groups ($\mathbf{G}^{s,h,u}$) allocated to the specific LRS-Agency-

[1]While this study uses two objectives, the CBW problem requires future handling of three and more objectives, thereby motivating us to use NSGA-III.

LS (s, h, u) combination. All BMPs chosen from the all BMP groups associated with (s, h, u) must add up to the parcel's total acres $\alpha_{s,h,u}$. To provide an example, let us consider the Berkeley county in West Virginia. This county contains 248 parcels, each with an average of 56.81 independent BMPs (organized into groups). Each BMP group, in turn, comprises an average of 7.89 BMPs. Thus, for the Berkeley county, there are roughly 248×56.81 (exactly 14,090) real-valued variables. A nonlinear optimization problem with so many variables arising from one county makes the CBW management problem challenging to any optimization algorithm. This detailed representation allows us to systematically explore and optimize the implementation of BMPs in a single or in cluster of counties.

III. Linking EMO Algorithms with CAST evaluator: CAST-NSGA-III

Here, we present the fourth phase of optimization algorithm development for CBW (Figure 1). Phase I encompasses

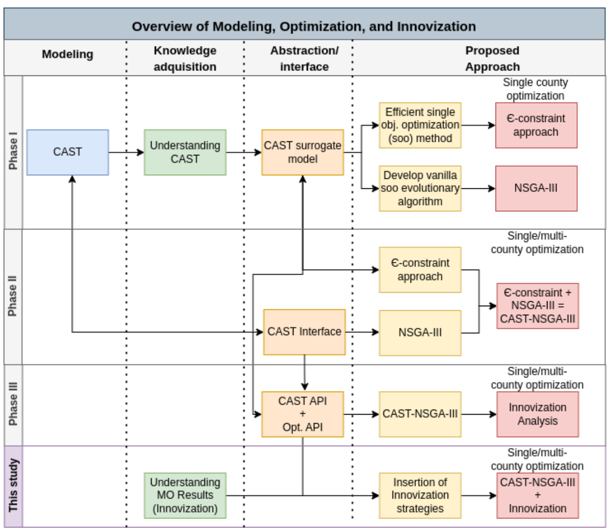


Fig. 1: Overview of the proposed modeling, optimization, and innovization tasks as a solution methodology for the CBW management problem. The final phase is the main crux of this study.

understanding CAST, modeling an accurate surrogate model, and developing single objective optimization methods, which can be solved using two multi-objective optimization methods – the ϵ-constraint approach and the customized version of the NSGA-III approach. Phase II proposes the development of an optimization interface that can interact with CAST. In Phase II, we developed the core of the optimization framework by hybridizing the ϵ-constraint and the NSGA-III approaches together. The resulting approach in which the hybrid method uses the CAST evaluation tool is called CAST-NSGA-III. Finally, phase III is devoted to creating Application Programming Interfaces (APIs) to build a bridge with CAST and an external API to enable developers and practitioners to use our optimization framework.

In addition to the CAST model, we employ a surrogate model to perform the search procedure [10]. The adopted surrogate model captures CAST interactions (with a low accuracy error) and preserves the Pareto-dominance relation between pairs of solutions. However, we also use high-fidelity evaluations on CAST to correct any bias from the surrogate model. Our ϵ-constraint component uses an interior-point-based approach that is highly effective in finding solutions in promissory zones.

Figure 1 shows the modeling, optimization, and innovization framework workflow. Our framework implements an API to deploy our optimization approach. APIs can help the development of new algorithms by providing a platform for developers to access data and resources for development. Additionally, APIs provide other researchers and practitioners to test, debug, and validate new algorithms. Our API is developed on two levels. The first one grant access to CAST, so we can send BMP implementation scenarios to evaluate and retrieve executions. The API isolates all problem communications and error transmissions within a single input. This API, which is meant for internal use, is implemented in C++20.

Our second API level is to expose our optimization approaches to users and decision-makers. Due to this reason, we adopt a *RESTful* approach using Django 4.1.1 [14].

The CAST-NSGA-III procedure starts with n solutions found by the ϵ-constraint approach applied to the surrogate model of the CAST system. Then, the solutions are injected into the initial population of the NSGA-III procedure. Next, the NSGA-III uses the surrogate model to evaluate solutions until we reach a predefined number of generations. After that, the approach uses the expensive CAST evaluation tool for better accuracy and confidence in the final solutions. Finally, CAST-NSGA-III stores the final results used for a post-optimal analysis, which is discussed next.

IV. Innovization Task To Find Innovized Rules

A multi-objective optimization problem gives rise to a number of Pareto-optimal solutions, each of which is an optimal solution corresponding to a specific trade-off among the objectives. Since they are optimal solutions, they may possess certain common features (involving problem variables and objectives/constraints) exhibiting the optimality principles. Such common features are not expected to be present in non-optimal and arbitrary solutions from the search space, as they would not meet the optimality conditions.

Finding common features in Pareto-optimal solutions involves (i) multi-objective optimization, followed by (ii) a feature extraction task. In many practical problems, this dual task has resulted in new and innovative solution principles that were not known before. Hence, the dual task was called *innovization* [4]. It started with manual innovization tasks, in which pairwise plots of variables from the Pareto solution set are created to find if specific pairs of variables possess an interesting pattern of variation across the Pareto set. For example, if a plot of variable pair (x_i-x_j) shows a monotonically increasing or decreasing pattern, a polynomial regression fit will reveal a mathematical rule. If the quadratic fit works well, the rule structure $x_j = c_1 + c_2 x_i + c_3 x_i^2$ (with coefficients (c_i) obtained from the regression process) will become an extracted rule. Such a rule indicates that optimal solutions follow the above relationship between x_i and x_j. Later, a machine learning-based procedure was developed to extract such rules in the form of rules (power laws [5] or decision trees [15], [16]) from an EMO-obtained solution set. However, the type of rules that can exist in a problem depends mainly on the nature of the problem.

If a Pareto-set contains innovized rules, they are essential for the user to gain additional knowledge about ‘what makes a solution optimal?’. If the innovized rules are simple and easily comprehensible, they can act as “thumb-rules” for future applications. On the other hand, if the innovized rules are long and involved, they can be used as a ‘procedure’ for use in the future.

Follow-up innovization studies [6], [17] have unveiled the innovized rules from non-dominated sets during a multi-objective optimization run and used the learned rules back into the optimization process for subsequent iterations as additional constraints. The rules are then used *softly* to modify the new solutions created by the optimization process. This rule-based update procedure, termed online innovization, has enhanced the convergence to Pareto-optimality compared to their original versions.

By manually analyzing the final non-dominated solutions obtained from CAST-NSGA-III runs. We learn new rules using the innovization procedure to re-optimize the problem to make the complex watershed management problem attainable and scalable.

A. Innovization for the CBW Problem

For the CBW problem, applying the CAST-NSGA-III procedure creates several optimized solutions with specific BMP allocations to every LRS. There are 205 BMPs allowed in the optimization process, but the final optimized solutions may not use them equally. Depending on the nature of the LRS, some BMPs may have been used more than others to make the solutions optimal. In this subsection, we describe the manual innovization procedure applied to the CAST-NSGA-III solutions to retrieve the frequency of BMP usage to gather helpful information (knowledge) about optimal solutions (Phase III).

The goal of the innovization task is to learn common or frequently occurring properties from the optimization results so that our optimization approaches can perform a more efficient search. In our context, this process helps us narrow down a set of BMPs that can be further used to accelerate the optimization process by reducing the overall search complexity. To generalize the innovization results, the optimization procedure is performed in four counties within the state of West Virginia. We select two urban-dominated counties, named Berkeley and Mineral, and two agricultural-dominated counties, named Hardy and Jefferson, to examine the results of innovization.

As the evaluation in CAST is computationally expensive, all our runs set the population size to 20, and the ϵ-constraint approach provides five solutions to start the NSGA-III procedure. Finally, we use standard values for the remaining parameters of the NSGA-III [12]. We execute 11 times our proposed approach for each of the previously mentioned configurations to keep a low number of evaluations on CAST. We use three ranking strategies to analyze the optimized solutions to identify the top-ranking BMPs. The first strategy identifies the list of BMPs based on the total implementation acreage. The second strategy ranks the BMPs based on the ratio of the BMP implementation area to the maximum available area. Finally, the third strategy ranks the top BMPs based on cost-effectiveness in pollution reduction. After averaging the individual ranking from the three strategies, the top-ranked BMPs are identified for combined urban and agricultural lands and used to improve our understanding of the BMP implementation strategy based on the innovization technique.

B. Extracted Innovized Principles

The innovization study is performed for agriculture and urban areas separately. Out of over 205 BMPs initially considered for developing the BMP implementation plan, the top seven BMPs are identified and presented in Figure 2. According to the BMP ranking, the ‘Nutrient Management Plan High-Risk Lawn’ is the top-ranked BMP in the three ranking strategies. This BMP has been used in the largest implementation area compared to other BMPs, as it is less costly and can also cause a large load reduction. After that, ‘Off Stream Watering without Fencing’, ‘Nutrient Management N Timing’, ‘Nutrient Management N Rate’, ‘Cover Crop Traditional Rye Early Drilled’, ‘Nutrient Management N Placement’, and ‘Barnyard Runoff Control’ are found to be the next ranked BMPs. Among these BMPs, ‘Nutrient Management N Timing’, ‘Nutrient Management N Rate’, ‘Cover Crop Traditional Rye Early Drilled’, and ‘Nutrient Management N Placement’ are entirely implemented in agricultural lands, while ‘Off Stream Watering without Fencing’ and ‘Barnyard Runoff Control’ are usually used to control pollution generated from animal production.

The ranking score in Figure 2 is calculated based on the overall position of BMPs in urban- and agricultural-dominated counties.

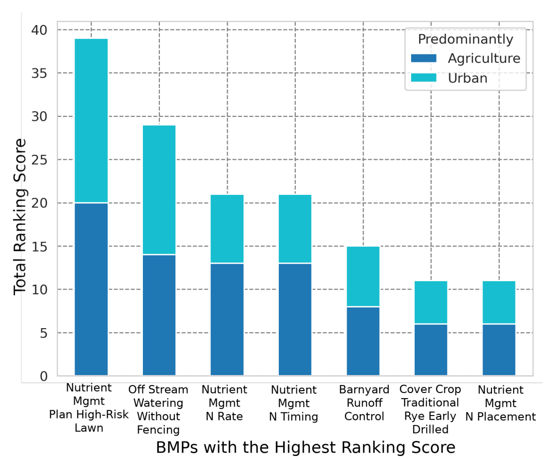


Fig. 2: The overall ranking score versus BMP type and dominated land use classification obtained from the innovization study.

V. Re-optimization Using Innovized Principles

The optimization algorithm is then modified to incorporate the knowledge from the manual innovization task (Phase III) mentioned above. We develop three different re-optimization strategies:

- **Control strategy**: The control strategy runs the optimization approach using all variables, utilizing no information from the innovization study.
- **Static strategy**: Section IV identifies a set of only seven BMPs that provide the best benefit according to the innovization analysis. The Static strategy uses these seven chosen BMPs in the re-optimization, thereby reducing 205 options to only 7 for each parcel.
- **Dynamic strategy**: Besides seven BMPs, the innovization study produces a priority list (a ranked list) of all 205 BMPs. For the Dynamic strategy, we select all necessary BMPs from the priority list until an accumulative percentage of 99.9% of the total implementation area is covered. Due to this large coverage, the number of options for each parcel can be larger than that in the Static case. Table I present the number and reduction in variables.
- **Preferred strategy**: In this strategy, more BMP options than in Static and Dynamic strategies are chosen, aiming to perform a better re-optimization. For every parcel, if any of the seven innovized BMPs (used in the Static strategy) exist in the group of allowable BMPs, then all BMPs other than the innovized BMPs are eliminated as variables in the optimization. Thus, this strategy reduces BMP options only to those parcels where one of the innovized BMPs is used. As can be seen from the table, the number of variables is much larger than the first two strategies but are still about one-third of that in the control problem.

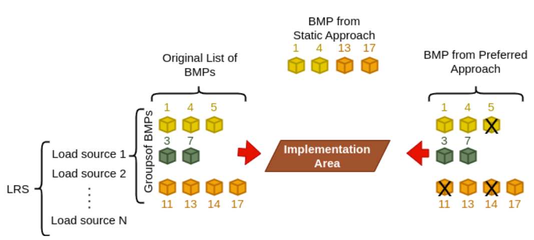


Fig. 3: Groups of BMPs can be applied to load sources with an overlay. Selected BMPs from the innovization process are preferred in their group in the Preferred strategy.

The strategies to incorporate the innovization knowledge into the algorithm eliminate the use of a large number of BMPs; thus, they reduce the search space size. Table I shows the number of variables used in each strategy. It can be seen that the Static and Dynamic strategies use less than 10% of the variables used by the Control strategy, and the Preferred strategy uses only about one-third of the total variables. Using innovization principles substantially reduces the number of variables in the optimization process.

We incorporate these three strategies into the CAST-NSGA-III and execute them 11 times each in the four selected counties. To measure the performance of each strategy, we

TABLE I: Total variables involved by (i) Control: executions considering all variables (no innovization incorporation), (ii) Static: Use the BMPs proposed by the innovization study (seven BMPs, exclusively), (iii) Dynamic: Dynamic selection of BMPs following an accumulative approach until reaching 99.9% of coverage, and (iv) Preferred: Selection of the BMP from the innovization study exclusively in their groups of BMPs. In brackets, we present the percentage of reduced number of variables compared to the original number of variables used in the Control strategy.

	Control	Static	Dynamic	Preferred
Berkeley	14,090	510 (3%)	1,023 (7%)	4,823 (34%)
Hardy	18,607	725 (3%)	751 (4%)	5,307 (28%)
Jefferson	12,303	456 (3%)	456 (3%)	4,079 (33%)
Mineral	20,260	765 (3%)	1,650 (8%)	6,415 (31%)

calculate the hypervolume metric. In this practical problem, the exact location of the nadir point is not known a priori. Moreover, they vary from one county to another. To present results from various counties in the same plot, we use the Hypervolume Ratio (HVR), calculated as follows. First, all non-dominated objective vectors from 11 runs are collected together for all four strategies. Second, dominated solutions are removed, and the ideal and nadir points of the combined set are calculated. Third, each run's objective vectors are normalized using the combined set's maximum and minimum objective values. Fourth, the nadir point $(1.1, 1.1)$ is set to compute HVR of each run. Thus, the maximum HVR for any run is expected to be at most one.

Figure 4 shows the variability of HVRs for different innovization approaches. From the plot, it is clear that the Control strategy, which consists of the CAST-NSGA-III executed with all variables, achieves competitive results. However, in all cases, its performance is overshadowed by the Preferred strategy, which except in the Mineral county, performs the best across the remaining counties. On the other hand, the Static strategy produces the worst results across all the counties, probably due to using a fixed set of seven BMPs. Finally, it is worth noting that the Dynamic strategy also provides competitive results, despite a large reduction in variables.

Although these results show the Preferred strategy as the clear winner, we apply Friedman's χ^2 test to compare our results statistically. The test is based on the rank of the observed values within each group and indicates whether there are significant differences between the results. Furthermore, as the application of Friedman's χ^2 indicates that results differ statistically, we compute pairwise comparisons using the Nemenyi post hoc test [18] to identify which strategy performs statistically differently.

Figure 5 shows the results of applying the Nemenyi post hoc test to the data. The figure gathers the pairwise comparison among all the strategies and counties. The rows indicate the strategy employed. Columns refer to compared configurations. The main diagonal of each county is empty. We include the p-value on each square and colored p-values less than 0.05. To compare a strategy **a** to any other strategy, such as **b**, we select the row of strategy **a** and the column of strategy **b**. The intersection presents the p-value of the compared strategies. For example, if the intersection square is red-colored, there is a statistical difference between these two strategies. In that case, we use a central tendency statistic to show which strategy performs the best.

Following the previous instructions to read the comparison maps, we see that the Preferred strategy outperforms the Static strategy in all cases and the Dynamic strategy in two cases. However, no statistical difference exists regarding the Control strategy and our Preferred strategy. Nonetheless, the median value of the Preferred strategy is better than the other strategies (including the Control strategy) in all counties.

These results show that the innovization study can capture a few critical BMPs (out of 205 BMPs) that help reduce pollutants in the CBW and cause the minimal cost of implementation. However, when we check the performance of the Static strategy (with the top seven BMPs), we see that it is the worst-performing strategy. It is worth noting that the Static strategy requires only 3% of the total variables, indicating that too much reduction of variables to achieve a faster re-optimization may produce valuable solutions. The Dynamic strategy, which chose BMPs based on a good coverage of land, thereby using a slight increase in the number of BMPs, produced almost equivalent trade-off front as that obtained by the Control strategy. Furthermore, the number of variables is not more than 8% of the total variables. These results support the importance and benefits of our innovization study. Figure 6(a) and (b) show the accumulated non-dominated solutions obtained by the four compared methods in Berkeley and Mineral counties, respectively. These plots show that our Dynamic and Preferred strategies are competitive with the control execution, with the Dynamic strategy having a substantial reduction in computational effort. The Static approach shows similar results for Berkeley County, but stays behind the other approaches for Mineral County.

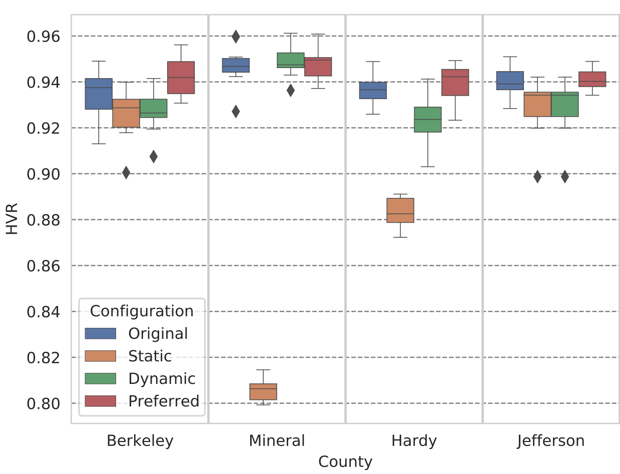


Fig. 4: Boxplot showing the results by the ratio of the hypervolume for four different strategies.

A. Transfer Learning to Other Counties

As a final step of this study, we employ the Preferred strategy to new counties as a transfer learning task in which the innovized BMPs were obtained from different counties. For this purpose, we choose Grant, Morgan, Pendleton, and Preston counties, as new counties and previously considered counties (Berkeley, Hardy, Jefferson, and Mineral) to develop the innovized BMPs. All these counties are in West Virginia and are part of the CBW. Table II shows each strategy's total variables. The Preferred strategy uses around 33% of the original variables.

Like the previous experiment, we performed 11 executions with our Control strategy (CAST-NSGA-III-All with all variables) and the Preferred strategy (CAST-NSGA-III-Pref). The results are presented in Figure 7. At first glance, innovization seems to help CAST-NSGA-III-Pref outperform the CAST-NSGA-III-All in Morgan and Preston counties. The Nemenyi post hoc test shows a p-value of 0.025 for Morgan and Preston, corroborating our observation in Figure 7. However, for Pendleton, CAST-NSGA-III-Pref performs similarly to CAST-NSGA-III-All, but with a minor standard deviation among the runs. For Grant, the median of CAST-NSGA-III-All is slightly better than the Preferred strategy.

TABLE II: Total variables involved by Control and Preferred strategies are presented. The percentage of reduced number of variables to that in the Control strategy is shown in brackets.

	Control	**Preferred (% reduced to)**
Grant	25,228	8,036 (31%)
Morgan	11,880	3,947 (33%)
Pendleton	33,083	10,483 (31%)
Preston	1,470	447 (30%)

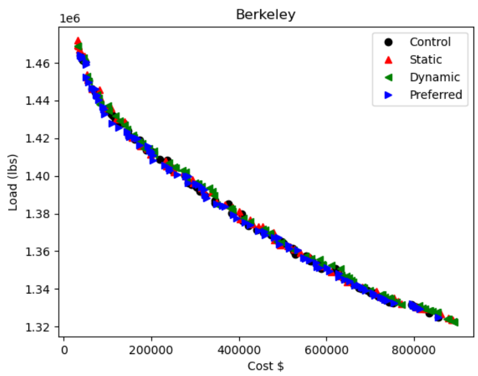
B. Comparative Study

CBW management problem has not been optimized before. In this pilot study, we discuss the challenges of solving the problem, mainly due to a large number of variables, nonlinear objective and constraint functions, and the time-consuming evaluation of solutions using a black-box and EPA-proprietary CAST evaluation tool. However, we also find in this study that using a smaller set of innovized BMPs and re-optimizing the problem with a much-reduced set of variables achieves a similar performing trade-off front and can be extended to address new counties with results from existing studies in a transfer learning mode. Thus, a comparison study to our proposed approach must be another newly developed optimization methodology that has the ability to handle a large number of variables and other complexities mentioned above.

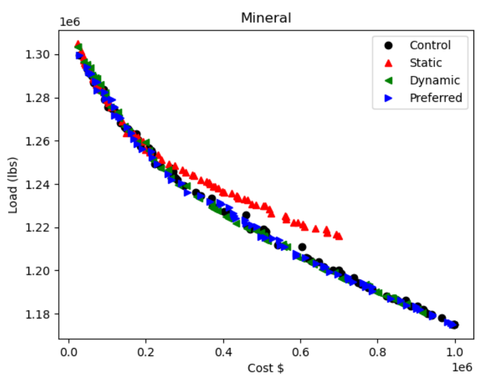
To show that our proposed methodology is efficient and the results of it cannot be produced easily, we compare our results with the Static strategy but using seven randomly-chosen BMPs (without any innovization study). We select a multi-county scenario with four counties simultaneously: Berkeley, Hardy, Jefferson, and Mineral. The Control strategy involves 65,260 variables, while the Preferred strategy reduces the variables to 2,456 (reducing to around 3.7% of the original variables).

The results in Figure 8 show that CAST-NSGA-III-Static has a competitive performance with CAST-NSGA-III-All; however, as expected, the randomly selected BMPs do not perform well at all.

Fig. 5: Pairwise Nemenyi post-hoc test among the different configurations for Berkeley, Hardy, Jefferson, and Mineral counties for the re-optimization study.



(a) Berkeley County



(b) Mineral County

Fig. 6: Comparison of the accumulated Pareto fronts of CAST-NSGA-III-All (Control), CAST-NSGA-III-Static (Static), CAST-NSGA-III-Dynamic (Dynamic), and CAST-NSGA-III-Pref (Preferred).

In addition to examining the quality of the solutions, we now compare the computational time required for each strategy for a similar number of evaluations. NSGA-III is executed on an AMD Ryzen 9 5950X 16-core, 32-Thread, and 128 GB of RAM. CAST is evaluated on the cloud using AWS and five concurrent processors. Each processor executes an instance of the problem in parallel (the times of this experiment can increase if we reduce the number of concurrent processors, and vice versa). Figure 9 shows that utilizing a subset of BMPs led to improvements in computational time, as the execution of our framework with all BMPs demanded more time. The strategy with randomly chosen BMPs is similar to the innovization-based Static approach.

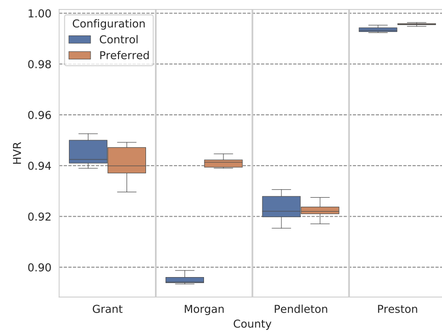


Fig. 7: Boxplot showing the results by the ratio of the hypervolume for the transfer learning cases.

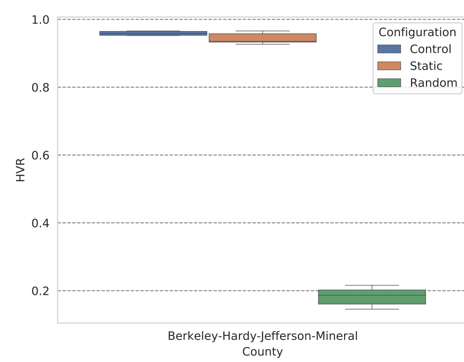


Fig. 8: Boxplot showing the results by the ratio of the hypervolume for the transfer learning case over a multi-county scenario: Berkley, Mineral Hardy, and Jefferson counties involving 65,260 variables (Control). CAST-NSGA-III-Static (Static) uses only 2,456 variables but achieves similar performance. On the other hand, using the random selection of BMPs contained a similar search space to our CAST-NSGA-III-Static strategy, but it could not reach the Pareto front.

VI. Conclusions

This paper has considered a challenging practical optimization problem involving tens of thousands of variables required to be optimized for multiple objectives and involving a computationally expensive evaluation tool. We have proposed a

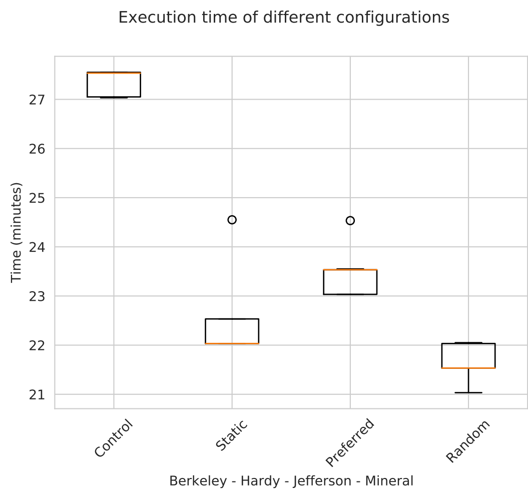


Fig. 9: Computational time needed by CAST-NSGA-III-All (Control), CAST-NSGA-III-Static (Static), CAST-NSGA-III-Pref (Preferred), and the random selection of BMPs (Random). On average, the Static strategy required 20% less time, the Preferred strategy required 14% less time, and the Random strategy required 22% less time than the time needed by the Control strategy.

hybrid multi-objective optimization algorithm that embeds the initial population with optimal solutions found using a point-based scalarization method worked on a mathematical model (surrogate model) of the expensive evaluation tool. To facilitate the use by practitioners, we have used RESTful APIs to link an evolutionary multi-objective optimization procedure (NSGA-III) with the evaluator tool called CAST. The resulting CAST-NSGA-III has been applied to four counties of West Virginia to find the optimal allocation of 200+ land use management, known BMPs .

Although the above refers to a significant study yielding a number of trade-off solutions among the cost of land use implementation and reduction in nitrogen loading to the environment, this paper has brought out another significant finding that can be generically applicable to other practical problems. The multi-objective solutions have been analyzed to find common BMP allocation principles stored in them. In this context, we have identified the most frequent BMPs among optimized non-dominated solutions and used them as a significantly reduced set of variables (to the tune of 3 to 33% of the original variable set) using three different innovization strategies. Remarkably, the re-optimization results with reduced variables are equivalent to (and better on some occasions than) the original application having tens of thousands of variables. Also, innovized strategies have produced better-performing re-optimization results compared to randomly chosen BMPs. The power of a post-optimal innovization task in making a large-scale societal problem solvable stays as the hallmark feature of this study.

Moreover, it has also been shown that the critical BMPs collected from optimization studies on certain counties can solve other new counties or multi-counties in a transfer learning mode. The trade-off solution set is similar to or better than the original BMP optimization of these counties. Inherent problem similarity in problems from different counties enables such transfer learning possible.

The study can be extended to find reduced number of solution evaluations needed by the innovization-based re-optimization strategies to achieve a similar hypervolume to the Control strategy. Its scalability to more counties and to the watershed would also be interesting. Nevertheless, this study has considered 1,470 to 65,260 real-parameter linked variables, making it one of the largest practical problems solved using any evolutionary algorithm. Furthermore, the obtained results showed that incorporating the post-optimal innovization principle to extract critical knowledge can speed up the search process and allow for solving similar problems without performing any new optimization.

References

[1] K. Miettinen, *Nonlinear Multiobjective Optimization*. Boston: Kluwer, 1999.

[2] K. Deb, *Multi-objective optimization using evolutionary algorithms*. Chichester, UK: Wiley, 2001.

[3] C. A. C. Coello, D. A. VanVeldhuizen, and G. Lamont, *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer, 2002.

[4] K. Deb and A. Srinivasan, "Innovization: Innovating design principles through optimization." in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2006)*, New York: ACM, 2006, pp. 1629–1636.

[5] S. Bandaru and K. Deb, "Towards automating the discovery of certain innovative design principles through a clustering based optimization technique," *Engineering Optimization*, vol. 43, no. 9, pp. 911–941, 2011.

[6] R. Datta and K. Deb, "A classical-cum-evolutionary multi-objective optimization for optimal machining parameters," in *World Congress on Nature and Biolog. Inspired Computing (NaBIC-09)*, 2009, pp. 607–612.

[7] K. Deb, S. Bandaru, D. Greiner, A. Gaspar-Cunha, and C. C. Tutum, "An integrated approach to automated innovization for discovering useful design principles: Case studies from engineering," *Applied Soft Computing*, vol. 15, pp. 42 – 56, 2014.

[8] S. Phillips and B. McGee, "The Economic Benefits of Cleaning Up the Chesapeake Bay: A Valuation of the Natural Benefits Gained by Implementing the Chesapeake Clean Water Blueprint," http://www.cbf.org/economicbenefits, 2014, accessed: 2023-01-27.

[9] Chesapeake Bay Program, "Chesapeake Bay Watershed Agreement," https://d18lev1ok5leia.cloudfront.net/chesapeakebay/Chesapeake-Bay-Watershed-Agreement-Amended.pdf, 2014, accessed: 2022-12-21.

[10] D. E. Kaufman, G. W. Shenk, G. Bhatt, K. W. Asplen, O. H. Devereux, J. R. Rigelman, J. H. Ellis, B. F. Hobbs, D. J. Bosch, G. L. Van Houtven, A. E. McGarity, L. C. Linker, and W. P. Ball, "Supporting cost-effective watershed management strategies for chesapeake bay using a modeling and optimization framework," *Environmental Modelling & Software*, vol. 144, p. 105141, 2021.

[11] J. Talberth, M. Selman, S. M. Walker, and E. Gray, "Pay for performance: Optimizing public investments in agricultural best management practices in the chesapeake bay watershed," *Ecological Economics*, vol. 118, pp. 252–261, 2015.

[12] G. Toscano, J. Hernández-Suárez, J. Blank, P. Nejadhashemi, and K. Deb, "Large-scale Multi-objective Optimization for Water Quality in Chesapeake Bay Watershed," in *2022 IEEE Congress on Evol. Computation (CEC'2022)*, A. Sperduti and M. Gori, Ed. Padua, Italy: IEEE Press, 07 2022, pp. 1–10.

[13] H. Jain and K. Deb, "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*, vol. 18, pp. 602–622, 2014.

[14] Django Software Foundation, "Django." [Online]. Available: https://djangoproject.com

[15] C. Dudas, A. H. Ng, L. Pehrsson, and H. Boström, "Integration of data mining and multi-objective optimisation for decision support in production systems development," *International Journal of Computer Integrated Manufacturing*, no. ahead-of-print, pp. 1–16, 2013.

[16] Y. Dhebar and K. Deb, "Interpretable rule discovery through bilevel optimization of split-rules of nonlinear decision trees for classification problems," *IEEE Transactions on Cybernetics*, vol. 51, no. 11, pp. 5573–5584, 2020.
[17] A. Gaur and K. Deb, "Effect of size and order of variables in rules for multi- objective repair-based innovization procedure," in *Proceedings of Congress on Evolutionary Computation (CEC-2017) Conference*. Piscatway, NJ: IEEE Press, 2017.
[18] P. Nemenyi, "Distribution-Free Multiple Comparisons," New York, 1963.