AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2016WR019987

Key Points:
- Observational experiments with an online interactive genetic algorithm and real humans were conducted for conservation planning
- User behavior on the web interface was used to identify the quality of user-provided data guiding the IGA
- Similarities and dissimilarities were found in the objective space of plans generated by participants with varying interests

Correspondence to:
M. Babbar-Sebens,
meghna@oregonstate.edu

Citation:
Piemonti, A. D., M. Babbar-Sebens, S. Mukhopadhyay, and A. Kleinberg (2017), Interactive genetic algorithm for user-centered design of distributed conservation practices in a watershed: An examination of user preferences in objective space and user behavior, *Water Resour. Res.*, *53*, 4303–4326, doi:10.1002/2016WR019987.

Received 24 OCT 2016
Accepted 31 MAR 2017
Published online 31 MAY 2017

Interactive genetic algorithm for user-centered design of distributed conservation practices in a watershed: An examination of user preferences in objective space and user behavior

Adriana Debora Piemonti[1], Meghna Babbar-Sebens[1], Snehasis Mukhopadhyay[2], and Austin Kleinberg[1]

[1]School of Civil and Construction Engineering, Oregon State University, Corvallis, Oregon, USA, [2]Department of Computer and Information Sciences, Indiana University-Purdue University Indianapolis, Indianapolis, Indiana, USA

Abstract Interactive Genetic Algorithms (IGA) are advanced human-in-the-loop optimization methods that enable humans to give feedback, based on their subjective and unquantified preferences and knowledge, during the algorithm's search process. While these methods are gaining popularity in multiple fields, there is a critical lack of data and analyses on (a) the nature of interactions of different humans with interfaces of decision support systems (DSS) that employ IGA in water resources planning problems and on (b) the effect of human feedback on the algorithm's ability to search for design alternatives desirable to end-users. In this paper, we present results and analyses of observational experiments in which different human participants (*surrogates* and *stakeholders*) interacted with an IGA-based, watershed DSS called WRESTORE to identify plans of conservation practices in a watershed. The main goal of this paper is to evaluate how the IGA adapts its search process in the objective space to a user's feedback, and identify whether any similarities exist in the objective space of plans found by different participants. Some participants focused on the entire watershed, while others focused only on specific local subbasins. Additionally, two different hydrology models were used to identify any potential differences in interactive search outcomes that could arise from differences in the numerical values of benefits displayed to participants. Results indicate that *stakeholders*, in comparison to their *surrogates*, were more likely to use multiple features of the DSS interface to collect information before giving feedback, and dissimilarities existed among participants in the objective space of design alternatives.

1. Introduction

Watershed planning and management studies are driven by a need to identify efficient and, at the same time, acceptable decisions related to the spatiotemporal allocation, use, storage, and regulation of resources required by inhabiting humans. To better represent the human conditions and needs, "bottom up" participation of humans (or, stakeholders) in the development of planning and management decisions for their watersheds has been advocated as being vital for success [*Palmer et al.*, 1995; *Lorenzoni et al.*, 2000; *Welp*, 2001; *Van Asselt Marjolein and Rijkens-Klomp*, 2002; *Assaf et al.*, 2008; *Voinov and Bousquet*, 2010; *McIntosh et al.*, 2011; *Hamilton et al.*, 2015].

However, engagement of stakeholders to assist with development of decision alternatives is not an easy task. Multiple conflicting criteria and constraints commonly exist in stakeholder communities, only some of which may be known to or formulated by the entity or agency overseeing the planning process. Additionally, decisions involving spatial features may include a variety of local and/or subjective constraints, which only the local decision maker (e.g., a producer preferring only certain types of decisions on her/his farms) may be aware of. Additionally, many of these stakeholder-specific preferences and knowledge may change with time, as stakeholders continue to learn and evolve. Hence, actions performed by a community of humans in a watershed may not always seem to be rational to a watershed planning entity that assumes a global utility function to represent preferences of watershed inhabitants. Multiple researchers in Cognitive Psychology and Behavioral Economics [e.g., *Kahneman and Tversky*, 1979; *Metcalfe and Shimamura*, 1994;

Schwartz, 1994; *Nelson*, 1996; *Reder*, 1996] have also provided evidence that humans often do not seem to follow principles of rational choice, and limitations in humans' cognition play an important role in determining their choices. For example, an individual driven by local site-scale conditions (e.g., on a farm) may be more likely to accept a watershed Plan X that has local site-scale decisions in agreement with her/his personal preferences for a specific type of solution strategy, goals, and/or constraints. Such an individual is less likely to accept an alternate watershed-scale Plan Y that does not satisfy her/his personal preferences at the local site of interest, even if Plan Y is better than Plan X in meeting the larger watershed goals (or a global utility function). In addition, not all individuals are interested in "optimizing" their decision all the time. Many individuals may be "satisficers" (*satisfy* + *suffice*) for some criteria, and will accept choices based on some degree of satisfaction and until a threshold of acceptability is achieved [*Simon*, 1955, 1977].

One of the biggest limitations in conventional optimization-simulation methods, which are often used to generate decisions for watershed planning problems [e.g., *Randhir et al*., 2000; *Seppelt and Voinov*, 2002; *Perez-Pedini et al*., 2005; *Arabi et al*., 2006; *Artita et al*., 2008; *Lethbridge et al*., 2010; *Babbar-Sebens et al*., 2013], is that many of these methods assume a rational decision maker (DM) who is able to articulate and formulate the problem and is able to provide quantitative information that accurately identifies her/his preference for objectives and/or solutions. While these methods can be immensely effective in searching for efficient alternatives in large and complex decision spaces, they are limited by the assumption that a mathematical solution to the optimization-simulation problem will produce a true Pareto optimal frontier—one that is expected to contain the DM's most-preferred solution. This simplistic view of a decision maker's cognition, attitudes, preferences, and knowledge limits the ability of these optimization-simulation methods to deal with real-world applications where the above-described complexities in human dimensions exist. For example, in a previous study [*Piemonti et al*., 2013] it was found that landowner attitudes toward specific conservation practices and/or criteria at a local-site may motivate them to further alter local-site decisions in a prescribed watershed plan that had been optimized for watershed-scale goals. A global decision maker (e.g., an agency) focused on watershed-scale objective functions (or, objective space) may regard these modified watershed plans as inferior plans in comparison to the original optimized plans, in contrast to local stakeholders who find them more acceptable instead. Other researchers [e.g., *Babbar-Sebens and Minsker*, 2008, 2010, 2012; *Singh et al*., 2008; *Rosenberg and Madani*, 2014; *Read et al*., 2014] have also begun to advocate for advanced search and optimization techniques that enable decision makers to identify solutions that are more acceptable to a community, and perhaps even in the proximity of the most mathematically optimal solutions in objective space.

One way to find the "sweet spot" or desirable region of efficient as well as acceptable watershed plans is by improving methods for engaging with stakeholders, learning from them, and integrating them in the design process. Currently, most participatory planning approaches (e.g., Integrated Water Resources Management [*Schramm*, 1980; *Viessman et al*., 2008], Shared Vision Modeling and Planning [*Hamlet*, 1996a, 1996b; *Palmer*, 1998; *Werick et al*., 1996], Agent-based Land-Use Models [*Millington et al*., 2011]) use traditional stakeholder engagement methods (e.g., focus groups, phone interviews, computer-mediated questionnaires, mail-in surveys, open community forum, etc.) in combination with decision support systems to identify a "shared" vision of goals, constraints, and priorities among all stakeholders, thereby facilitating a better representation of human dimensions in the planning process. However, these engagement methods are limited in terms of the data they can collect on community knowledge, preferences, and risks. They only provide a single "snapshot in time" assessment of communities. This makes a higher-order understanding of stakeholder's iterative learning processes difficult, because such processes tend to unfold over time [*Gunderson and Holling*, 2002; *Prell et al*., 2007; *Reed*, 2008]. Furthermore, commonly used decision support systems [e.g., *Fedra*, 1992; *Loucks and Da Costa*, 1991; *Georgakakos and Martin*, 1996] have rigid model-building and decision-making environments that were developed for a specialized subset of users, and may not be sensitive or attractive to the diverse community of watershed stakeholders whose engagement is necessary for ensuring success.

Over the last few years, a growing community of researchers has begun to explore Web 2.0 technologies as an alternate media for improving stakeholder engagement during the planning process [*Kelly et al*., 2012]. The underlying motivation for this shift is that Web 2.0 provides a promising social networking platform for connecting and assimilating a large number of humans into the planning process. Such technologies, when supported by ubiquitous computing devices, also provide opportunities for conducting continuous multiple

stakeholder interactions, and for improving a community's situational awareness and learning over time. However, the functionality of most of these web-based stakeholder engagement and planning tools (e.g., Web-based, Water-Budget, Interactive, Modeling Program (WebWMPI) [*Matsuura et al*., 2009]; Sierra Nevada Adaptive Management Project (SNAMP) [*Fry et al*., 2015]; Agricultural Conservation Planning Framework (ACPF) [*Tomer et al*., 2015a, 2015b]) has been limited to visualizing, reviewing, and sharing of data and model results. In comparison, *Babbar-Sebens et al*. [2015] recently developed a novel, web-based interactive design tool called WRESTORE (Watershed REstoration using Spatio-Temporal Optimization of REsources; http://wrestore.iupui.edu/), which not only engages with stakeholders via the Web but also uses their feedback ("wisdom of the crowd") to dynamically guide the underlying computational design algorithms in identifying user-preferred watershed plans. The design algorithm in WRESTORE belongs to a family of Interactive Optimization [*Fisher*, 1985; *Klau et al*., 2010; *Meignan et al*., 2015] methods, which unlike conventional optimization-simulation techniques, are useful for design problems containing additional subjective criteria, constraints, and preferences that are not easy to quantify into a global utility function, and/or may not be known a priori. Hence, by engaging a human (decision makers and stakeholders) in the iterative loops of the design process, both the human and the algorithm have the potential to communicate and learn about such subjective preferences from each other via graphical user interfaces on the Web.

As the current science in models and technologies for human-computer collaboration and online social networking [*Nielsen*, 2009] continues to advance, the potential for improved stakeholder engagement via online participatory design assisted by machine agents is enormous. However, there is a serious lack of data and understanding of how users behave in these online design environments and what types of solutions to watershed plans can be generated when multiple humans engage in such participatory design environments. In this paper, for the first time, we present and examine results of observational experiments conducted with different types of human participants who used WRESTORE to generate user-preferred scenarios of spatial allocations of conservation practices in a watershed. To the authors' knowledge, this paper represents the first known study that unifies data on online behaviors of real humans with results of a human-in-the-loop search algorithm employed for designing watershed alternatives. In this paper, we focus on examining the similarities and dissimilarities among generated watershed plans in the objective function space (or, objective space). We specifically examined the following research questions:

1. How effective is the interactive optimization algorithm in assisting users generate highly preferred design alternatives, especially for different types of users with varying interests, preferences, and online interaction behavior?
2. In objective space of global, watershed-scale goals, how similar or dissimilar are the design alternatives found by user-driven interactive optimization algorithm, in comparison to the design alternatives found by a conventional, noninteractive optimization algorithm, and in comparison to the design alternatives found by other users?
3. In the objective space pertaining to local, subbasin-scale goals, how similar or dissimilar are the design alternatives found by different stakeholders, when some of them are focused only in certain local areas (e.g., landowners) and while others are focused on the entire watershed scale (e.g., agency personnel)?

In the following sections, we first describe the overall methodology (section 2) of the WRESTORE decision support system and the interactive genetic algorithm, the watershed study site where the conservation planning problem was solved using interactive genetic algorithm, the experimental setup for user studies, and the metrics proposed for analyzing results in the objective space. In section 3 we present and discuss the results from the user experiments that were used to investigate the research questions 1–3, followed by section 4 that presents the overall conclusions of this study along with directions for future work.

2. Methodology

2.1. Overview of WRESTORE Methodology

The WRESTORE tool was developed by *Babbar-Sebens et al*. [2015] to enable communities to engage in online participatory design of plans on spatial allocation of conservation practices on their landscape. The underlying interactive optimization (or, human-guided search) algorithm in WRESTORE is based on the Interactive Genetic Algorithm with Mixed Initiative Interaction (IGAMII) algorithm, proposed originally by *Babbar-Sebens and Minsker* [2012]. Figure 1 is an overview of workflow that users experience when they

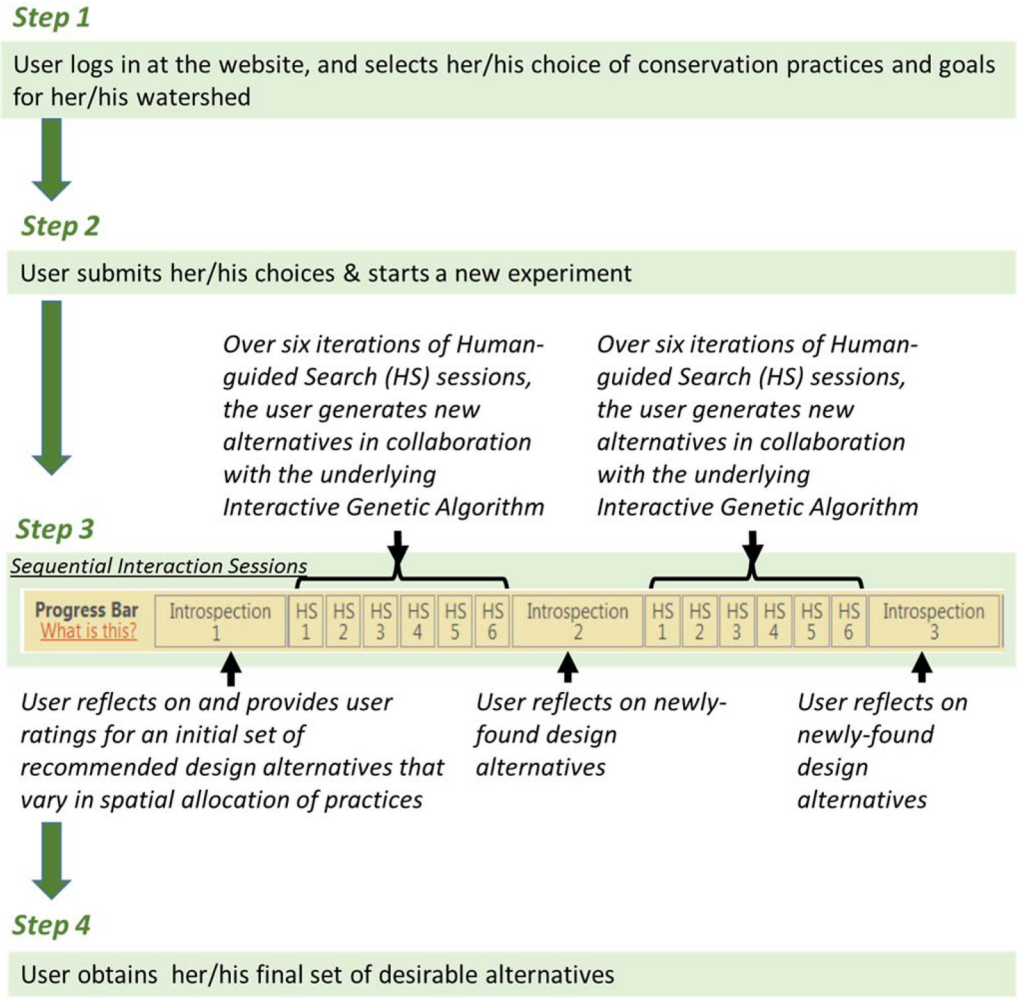


Figure 1. User workflow in WRESTORE (see *Babbar-Sebens et al.* [2015] for details).

engage with WRESTORE. While details can be found in *Babbar-Sebens et al.* [2015], here we discuss only an overview of the methodology.

Step 1. The user logs in, selects her/his choice of conservations practices from a set of seven possible practices (Wetlands, Filter Strips, Grassed Waterways, Strip Cropping, Cover Crops, Crop Rotation, and No-till Tillage practice), and her/his choice of watershed goals (cost, peak flow reduction, sediments reduction, and nitrates reduction).

Step 2. When the user starts a new experiment, each of the chosen practices is mapped into decision variables for each of the subbasins in the watershed. So, if a watershed has 100 subbasins and the user chooses two practices, and if one decision variable exists for each of the practice, then 200 decision variables will be initialized for the interactive optimization experiment. Decisions can be binary (yes/no) or real numbers (e.g., filter strip width). Similarly, the watershed goals will be initialized as objective functions in the experiment. In WRESTORE, a calibrated watershed model of the study site, based on Soil and Water Assessment Tool (SWAT) [*Neitsch et al.*, 2005], is used to simulate the impact of a candidate plan on the watershed, and calculate values for objective functions.

Step 3. The user then goes through a sequence of online interaction sessions (see Figure 1 that shows an example scenario of sessions) to generate efficient as well as desirable watershed plans, with the assistance of IGAMII algorithm. The online interactions sessions are of two types—introspection session and human-guided search (HS) session. The introspection session provides the user to reflect on designs previously

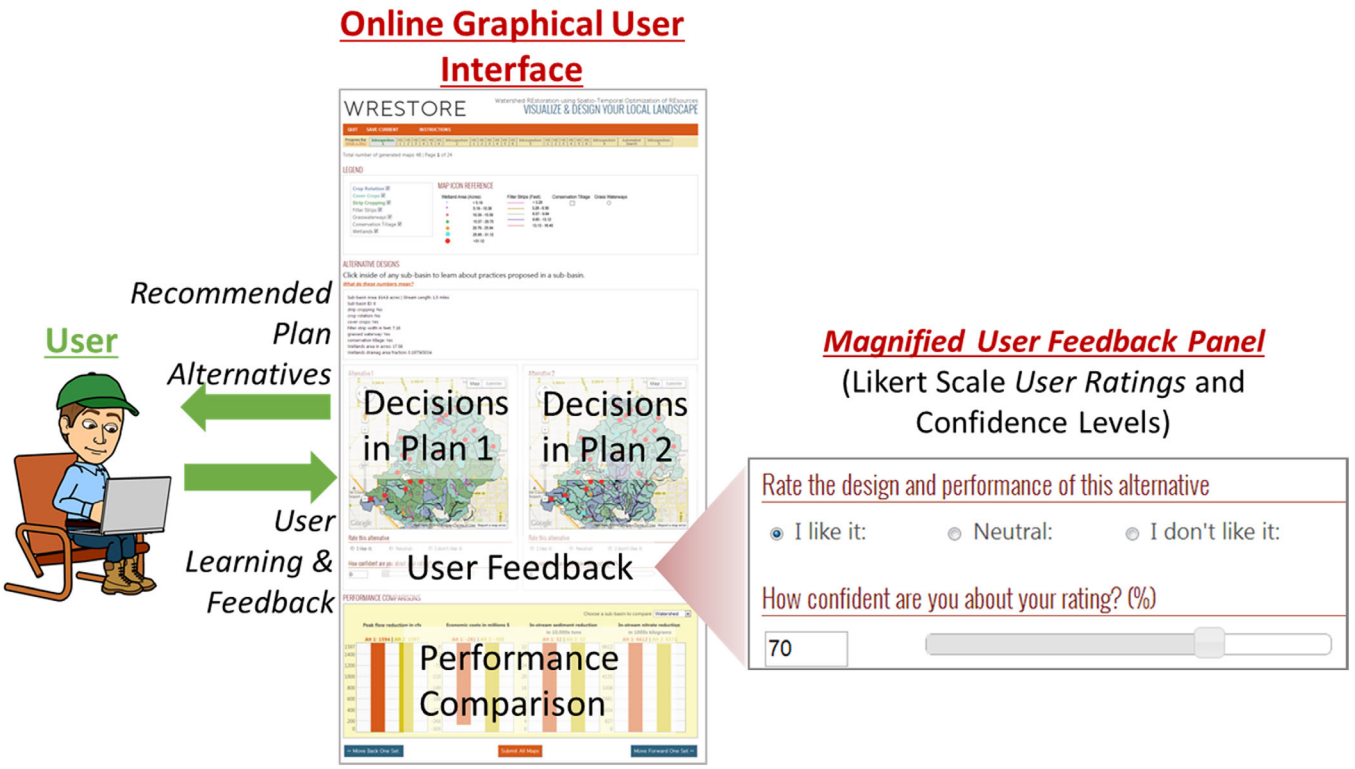


Figure 2. Graphical user interface (GUI) used in the sequential interaction sessions.

found either via a prior noninteractive search (i.e., in Introspection 1), or on the user-preferred alternatives found by the user in the most recently completed human-guided search run (e.g., in Introspections 2 and 3). The human-guided search (HS) runs are conducted using an interactive genetic algorithm (see *Babbar-Sebens et al.* [2015] for details), and, hence, each of the HS sessions displays the design alternatives (or, watershed plans) in the population of the genetic algorithm for user evaluation. For example, in Figure 1 there are six human-guided search (HS) sessions—the initial parent population using in generation 0 of the genetic algorithm are shown in HS 1, while the new child population created in generations 1–5 in the genetic algorithm are shown in HS 2–6. Currently in WRESTORE, we limit human fatigue by limiting the genetic algorithm's population size and the number of design alternatives shown to the user in each session to 20, and also via an inbuilt "reminder-by-email" feature that enables users to return at a later time for evaluating design alternatives in an ongoing session. However, adaptive methods for managing trade-offs between human fatigue and human feedback are being actively investigated in ongoing research, and we expect that future findings will enable significant improvements in how human effort is administered in IGAMII. Finally, the other parameters for the multiobjective genetic algorithm used during HS sessions were set as following: crossover probability = 0.9, mutation probability = 0.05, and selection strategy = $\mu + \lambda$ (i.e., mu + lambda).

In each of the sessions, an online graphical user interface (GUI), as shown in Figure 2, is used to display design alternatives to the user. The GUI has multiple interface features (e.g., maps, charts, drop down menus, and clickable information boxes) that encourage the user to learn about the recommended decisions in candidate plans, compare performances of these plans (i.e., objective functions) in local areas (e.g., one or two specific subbasins) and at the larger watershed-scale, and provide feedback on the quality of the design via a Likert-scale type *user rating* and self-confidence in her/his rating. Note that while the genetic algorithm in WRESTORE uses the watershed-scale performance to generate plans for conservation practices in all the feasible subbasins in a watershed, the user has the option to provide feedback based only on the decisions in her/his local subbasins of interests. Hence, both local and global stakeholders have the ability to participate and interactively generate solutions based on the scale they are most interested in. The *user rating* is used as an additional objective function that guides the operations of the genetic algorithm to generate the next iteration of child population.

Step 4. Once all the sequential interaction sessions are completed, the user's search experiment concludes and all the new desirable alternatives (e.g., those with *user ratings* "I like it" in Figure 2) are stored in a case based memory. The user has access to this memory for extraction and postprocessing of desirable alternatives at a later stage when a final decision has to be identified and/or negotiated.

2.2. Study Site for Participatory Design Experiments

The study site that was used to test WRESTORE in this research is Eagle Creek Watershed (ECW), located 10 miles NW of Indianapolis, IN [*Babbar-Sebens et al*., 2013; *Piemonti et al*., 2013]. This Midwestern watershed is primarily agricultural, with corn and soybean being the major crops grown on the landscape. Growing concerns in downstream water quality due to export of nutrients and sediments from the landscape into the river channel, and finally into the Mississippi river basin and the Gulf of Mexico [*The Conservation Fund*, 2016], have led to region-wide efforts in multiple Midwestern watersheds (including ECW) that are focused on increasing implementation of conservation practices (or, best management practices) on the landscape. In addition to water quality concerns, this region has also been facing increasing frequencies of floods in the last few years, and conservation practices also provide potential for mitigating flooding impacts during storm events. Hence, multiple stakeholders, including state and federal agency personnel, are interested in examining if a suite of spatially distributed conservation practices would provide a range of environmental benefits in ECW and similar watersheds in the Midwest.

To assess the impacts of conservation practices on water quality and peak flows, the hydrology and water quality in ECW were simulated using the Soil and Water Assessment Tool 2005 (SWAT 2005) model [*Neitsch et al*., 2005]. The model was used to simulate not only baseline conditions with no conservation practices, but also scenarios when conservation practices are implemented on the landscape. SWAT uses the topography, land use, soil type, and regional weather information to estimate the water routing and the water quality through the watershed, at a daily time scale. *Babbar-Sebens et al*. [2015] and *Piemonti et al*. [2013] give a detailed description of the model construction, calibration, and how the SWAT model outputs were used to calculate four physically based environmental objective functions (Peak Flow Reduction, Sediment Reduction, Nitrates Reduction, and Costs) estimated at the entire watershed scale. As discussed earlier, besides the four watershed-scale environmental objective functions, a *user rating* objective function is also used to guide the search process of the interactive genetic algorithm in WRESTORE. The values for the *user rating* function are decided by stakeholders who are engaged in the search process. The values are based on a Likert-type scale—"I like it" (R_3), "Neutral" (R_2), and "I do not like it" (R_1)—that users can utilize to indicate their personal subjective preference for an alternative. An overview of these objective functions is shown in Table 1.

ECW was divided into 130 different subbasins (SBs) to simulate the local implementation of a set of conservation practices. In a previous study [*Babbar-Sebens et al*., 2013], it had been identified that 108 of the 130 subbasins were suitable for implementing conservation practices in agricultural areas. Hence, the decisions for spatial allocation of practices were limited to these 108 subbasins in the study site. Currently, WRESTORE is capable of generating design alternatives for seven different BMPs (strip cropping, crop rotation, cover crops, filter strips, grassed waterways, no-till practices, and wetlands) in all the subbasins considered for allocation. However, for this study, the researchers focused only on two practices—Cover Crops (CC) and Filter Strips (FS)—that are represented as binary and real number decision variables, respectively. Table 2 described how the decision variables in WRESTORE's IGAMII algorithm were converted into SWAT model parameters relevant to the practice.

2.3. Recruitment of Participants and Setup of Testing Scenarios

In this study, we evaluated results of 20 participants who volunteered to interact with the WRESTORE tool, and test its capabilities in finding watershed-scale plans that agreed with their individual subjective preferences. These users included 14 *surrogate* users who were volunteers with appropriate science and engineering backgrounds, including students from both Indiana University and Oregon State University, and who helped us with initial evaluation of the tool. After the initial evaluation was done, we successfully tested the tool with six *stakeholder* users (state/federal agency personnel, nongovernmental organization personnel, and watershed individuals) who evaluated the tool during a training workshop. Though the *surrogates* were not directly involved in the watershed, they are useful representatives of potential participants in a community who may be only cursorily interested in the decisions. The *stakeholders* were more closely associated

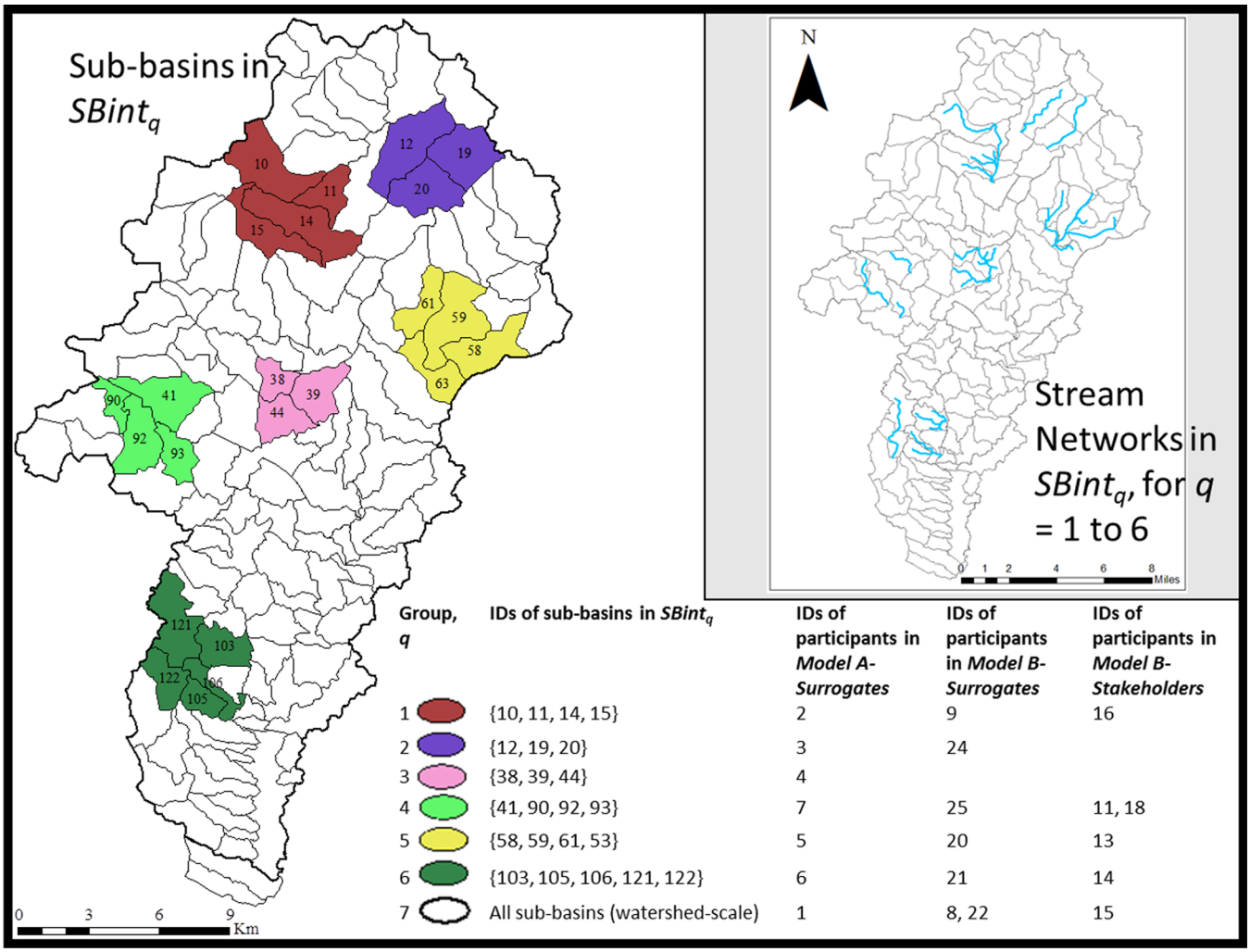
Table 1. Objective Functions Used to Optimize Design Alternatives in WRESTORE[a]

m	Objective	Function
1	Peak flow reduction (***PFR***)	$\boldsymbol{PFR} = Min\left[-Max_{i,t}\left(PF_{i,t,base} - PF_{i,t,alt}\right)\right]$ where $PF_{i,t,case} = \left\{\begin{array}{c} flowout_{i,t,case};\ \ if\ \ flowout_{i,t,case} > fowout_{i,t-1,case} \\ AND \\ flowout_{i,t,case} > flowout_{i,t+1,case} \\ 0;\ \ otherwise \end{array}\right\}$ and, *case* represents the baseline (*base*) or the design alternative (*alt*)
2	Sediments reduction (***SR***)	$\boldsymbol{SR} = Min\left\{-\sum_{i=1}^{N}\left[\sum_{t=1st\ \ day}^{last\ \ day}\ \sum_{day\ \ T1}^{T2}\left(Sout_{i,t,base} - Sout_{i,t,alt}\right)\right]\right\}$
3	Nitrates reduction (***NR***)	$\boldsymbol{NR} = Min\left\{-\sum_{i=1}^{N}\left[\sum_{t=1st\ \ day}^{last\ \ day}\ \sum_{day\ \ T1}^{T2}\left(Nout_{i,t,base} - Nout_{i,t,alt}\right)\right]\right\}$
4	Cost (***C***)	$\boldsymbol{C} = Min\left[\sum_{i=1}^{N} NPV_i\right]$ where $NPV_i = \sum_{c=1}^{BMP}\left[CI_c * A_{i,c}\right] + \sum_{ty=T1}^{T2}\left\{\sum_{c=1}^{BMP}\left[\left(OM_{c,ty} - Rin_{c,ty}\right) * A_{i,c}\right] - PI_{ty} - SP_{ty}\right\} * PWF_{ty}$
5	User rating function (***R***)	*Min* [***R***] where $\boldsymbol{R} = -1$ when user selects R_1 (i.e., "I do not like it") radio button on GUI's feedback panel, $\boldsymbol{R} = -2$ when user selects R_2 (i.e., "Neutral") radio button on GUI's feedback panel, $\boldsymbol{R} = -3$ when user selects R_3 (i.e., "I like it") radio button on GUI's feedback panel. These subjective ratings are determined by a user and are related to user's personal preferences for decisions, goals, and/or interests in $SBint_q$

[a]Definition of all variables in the columns above (see *Piemonti et al.* [2013] for details). *m* = ID of objective function; *i* = subbasin ID; *t* = day; *PF* = peak flow (m^3/s); *base* = baseline SWAT model with no new conservation practices; *alt* = SWAT model with conservation practices indicated in the design alternative; $flowout_{i,t,case}$ = daily flow predicted by SWAT model at the outlet of subbasin *i*, on day *t*, and for a specific scenario *case*; *N* = total number of subbasins (SB) in watershed; *T*1 = initial year of simulation; *T*2 = final year of simulation; *Sout* = daily sediment load (tons); *Nout* = daily nitrate load (Kg); *NPV* = net present value ($/watershed); *c* is the identification number of a practice that varies from 1 to *BMP*; *BMP* is the total number of practices being considered in planning problem; CI_c is the cost of implementation in dollars per acre for *c*th conservation practice; $A_{i,c}$ is the area in acres of *c*th conservation practice in a subbasin *i*; *ty* is the simulated year that varies from *T*1 to *T*2; $Om_{i,ty}$ is the operation and maintenance costs in dollars per acre, for *c*th conservation practice in year *ty*; $Rin_{c,ty}$ is the rent received by the conservation program in dollars per acre for those lands that are taken out of production due to *c*th conservation practice in year *ty*; SP_{ty} is the savings in crop production costs in dollars for all land taken out of production by conservation practices in year *ty*; PI_{ty} represents the net profits, in dollars, obtained from increased productivity in year *ty*; PWF_{ty} is the single payment present worth for year *ty*, based on interest rate *int* and is given by $PWF_{ty} = 1/(1+int)^{ty}$; $SBint_q$ = subbasins of interest of the decision maker belonging to group *q* (see Figure 3 for details).

Table 2. Changes Made on the SWAT Model to Simulate Conservation Practices

Practice	SWAT Variable Modified	File	Notes
Filter strips (can be implemented in each of the 108 subbasins; decision variable is value of filter strip width between 0 and 5 meters for each of the subbasins)	FILTERW	.mgt	The Field Office Technical Guide (FOTG, https://efotg.sc.egov.usda.gov) gives total estimated costs per acre of filter strips. The installation scenario in this work assumes filter strip length of 37 m, and filter strip width of value obtained from the decision variable FILTERW, in a total field area of 19 ha
Cover crops (can be implemented in each of the 108 subbasins; decision variable is value of yes (1) or no (0) for each subbasins)	Operation schedule	.mgt	An example of operations schedule with corn-winter wheat in 1 year is given below. The winter wheat operations (last three lines) can be used at the scale of hydrologic response units, for both corn and soybeans. Year HU* operation kg/ha 1 0.28 harvest and killing 1 0.1 pesticide application 1.12 1 0.12 plant corn 1 0.3 fertilizer application 200.00 1 1.5 harvest and killing 1 0.997 generic fall tillage 1 0.998 plant winter wheat



Group, q	IDs of sub-basins in $SBint_q$	IDs of participants in *Model A-Surrogates*	IDs of participants in *Model B-Surrogates*	IDs of participants in *Model B-Stakeholders*
1	{10, 11, 14, 15}	2	9	16
2	{12, 19, 20}	3	24	
3	{38, 39, 44}	4		
4	{41, 90, 92, 93}	7	25	11, 18
5	{58, 59, 61, 53}	5	20	13
6	{103, 105, 106, 121, 122}	6	21	14
7	All sub-basins (watershed-scale)	1	8, 22	15

Figure 3. Map of spatial distribution of subbasins of interest (*SBint*) in the watershed. The table on the bottom right indicates the assignment of *q* groups of *SBint* to testing participants in *Model A-Surrogates*, *Model B-Surrogates*, and *Model B-Stakeholder* groups. The inset on top right corner only shows stream networks in the local subbasins of interest belonging to groups $q = 1$ to 6.

with ECW via land ownership or professional responsibilities. Note that though eight stakeholders had originally participated in the stakeholder workshop, only six of them completed the experiment correctly; hence, results of only six stakeholders have been analyzed in this paper. This also illuminates the challenges in testing and implementing novel decision support tools in real-world conditions. All participants were first coached on the various features and benefits of the tool and then instructed on how to proceed with the experiments. They were all asked to select the two conservations practices (Cover Crops and Filter Strips) and all four environmental goals, before starting the experiment. However, to explore the effect of varying user interests, perceptions, and preferences of human participants on the search process of the interactive genetic algorithm, we created two types of testing scenarios. Both of these scenarios and their rationale are explained below.

2.3.1. Scenarios Based on Subbasins of Interest

As discussed earlier, stakeholders can comprise of individuals who may be more interested in examining the problem at local scale (e.g., subbasin or farm scale) and/or at the global scale (e.g., at larger watershed or river basin scale). The interests and constraints of these stakeholders may not always converge or overlap. Hence, seven groups of subbasins were identified on the watershed landscape, and each of the participants was assigned to a *q*th ($q = 1$ to 7) group with preselected IDs of "subbasins of interest" ($SBint_q$). Figure 3 gives a map of the subbasins in the various groups, and the IDs of participants in each of the *q*th group. Note that six groups consisted of randomly selected sets of neighboring upstream and downstream

subbasins in different local regions of the watershed, whereas the seventh group consisted of the entire watershed. Once the participants were notified about IDs of their subbasins of interest, they were asked to evaluate candidate watershed plans based on their subjective preference for decisions in only $SBint_q$ and/or based on their perception of performance of the candidate plans in only $SBint_q$. Hence, the participants were asked to provide values for *user ratings* based on only the assigned $SBint_q$. Note that the groups consisted of hydrologically connected as well as hydrologically disconnected subbasins, in order to examine for potential relationships between hydrologic dependencies within subbasins, and preferences expressed by participants.

2.3.2. Scenarios Based on Simulation Models

A stakeholder's preference for a decision may be influenced not only by the actual decision itself (e.g., more preference for one practice versus another), but also by how well the decision performs in goals/criteria valued by the stakeholder. The performances of practices in WRESTORE are estimated by the underlying SWAT watershed model. Hence, two models—*Model A* and *Model B*—were generated and used to explore the potential differences due to users' varying preferences for performances of conservation plans in objective space. *Model A* was the original calibrated SWAT model of the ECW study site (see *Piemonti et al*. [2013] for details on calibration), whereas *Model B* was a SWAT model of the watershed created by artificially enhancing the flow and water quality benefits predicted by the calibrated SWAT model (i.e., *Model A*). Enhancement in benefits was accomplished by activating a few wetlands in the watershed. This change considerably increased the benefits of peak flow reduction, sediment reduction, and nitrate reduction, when the conservation practices under consideration—i.e., Cover Crops and Filter Strips—were allocated in the subbasins by the search optimization algorithm. However, the users were not informed of this enhancement, and hence, they evaluated the design alternatives under the assumption that only Cover Crops and Filter Strips were being implemented in the watershed. The rationale for this artificial enhancement of benefits was to examine if an improvement in flow, nitrate, and sediment benefits would change the participants' preferences for practices, design alternatives, or spatial locations on the interface. In the rest of the paper, all surrogates whose experiments included *Model A* will be identified as *Model A-Surrogates* group, all surrogates whose experiments included *Model B* will be identified as *Model B-Surrogates* group, and all stakeholders whose experiments included *Model B* will be identified as *Model B-Stakeholders*. The *Model A-Surrogates* and *Model B-Surrogates* groups contained seven participants (four females and three males in each of the groups), whereas the *Model B-stakeholders* contained six participants (five males and one female). Please note that because of limited resources and logistical constraints we were not able to conduct a *Model A-Stakeholders* experiment. However, since we already conducted the scenario experiment with *surrogates*, the findings of this study are not expected to be affected significantly by the absence of *Model A-Stakeholders* data from such an experiment.

2.4. Metrics for Evaluation of Research Questions

Once test experiments with all participants were over, the results from all the interaction sessions were analyzed for every user. Multiple metrics were estimated in these analyses, in order to enable investigation of the three research questions that were stated earlier. Below are descriptions of how these metrics were calculated.

2.4.1. Metrics for Assessing User Interaction Behavior

An improved understanding of the nature and amount of interactions an end-user may have with a decision support system (DSS) are critical for facilitating a user's acceptance of the DSS and her/his personal confidence in solutions generated by such a system [*Belton et al*., 2008; *Meignan et al*., 2015]. Hence, evaluating the nature of a user's interaction behavior in DSSs, such as WRESTORE, is important for evaluating whether the DSS provides a user adequate opportunity to examine the multiple "what-if-scenarios" in design alternatives, and, as a result, learn about the decision-making problem in-hand. To assess the nature of user interactions in WRESTORE, we calculated the following two interface usability metrics and one confidence metric.

Mean percentage of time spent in gathering information. We assessed every participant's ability to navigate and gather information on design alternatives by tracking how much time they spent on browsing and clicking on maps where individual decisions were displayed, and time spent on browsing and clicking on bar graphs where performance of decision alternatives were compared on the GUI (Figure 2). In a previous study [*Piemonti et al*., 2017], we had learned that time spent by a user in gathering information on the GUI

can provide cues to her/his underlying motivations, ongoing learning, and interest in using the tool to assist with solving the problem in hand. Task times for each interaction session were recorded in the database. When the participants were taking a break and not using the tool, they were instructed to press the "Save all" button in GUI, so that these off-task time intervals could be considered as outliers and excluded from the analyses. However, there were some occasions when some participants did not click the "Quit" or "Save all" buttons, resulting in excessively long task times. To remove these outliers, we manually excluded these task times that were greater than two standard deviations from the mean task time across all sessions, for each participant. At the end, we calculated the sample estimate of percent of time spent in gathering information via the following equation:

$$PTS_k = \frac{\sum_{h=1}^{H} \sum_{l=1}^{L_{h,k}} \Delta t_{l,k,h}}{\sum_{h=1}^{H} ttotal_{k,h}} * 100, \quad (1)$$

where PTS_k is the percentage of time spent by a participant with ID k in gathering information per experiment, H is total number of sessions, $L_{h,k}$ is the total number of events associated with gathering information in hth session, $\Delta t_{l,k,h}$ is the interval of time spent in lth information gathering event by kth participant in hth session, and $ttotal_{k,h}$ is the total time (including gathering information, making decisions, etc.) spent by kth participant's in hth session.

Mean percentage of mouse clicking events related to information gathering. Besides tracking the time spent in gathering information, we also tracked how many mouse clicks were made in GUI areas that provided clickable information (e.g., drop down menus, pop-up boxes in maps and charts, etc.). For each participant, we used the following general formula for sample estimate of percentage of mouse clicking events related to information gathering,

$$PMC_k = \frac{\sum_{h=1}^{H} NC_{k,h}}{\sum_{h=1}^{H} TC_{k,h}} * 100, \quad (2)$$

where PMC_k is the percentage of mouse clicks made by kth participant in GUI areas where information can be gathered, $NC_{k,h}$ is the number of clicks made in areas that provide information in the hth session by kth participant, and $TC_{k,h}$ is the total number of clicks in hth session by kth participant.

Confidence trends. While giving *user ratings* on the desirability of design alternatives (see magnified feedback panel in Figure 2), the users were also asked to indicate how confident she/he felt about her/his own *user rating*. The confidence levels, along with the user's interaction behavior, can offer potential insights into the quality of a user's evaluation and how much the user trusts in her/his own feedback [*Fischer and Budescu*, 2005]. The users moved the bar on the confidence slider to identify a suitable confidence level between 0 and 100. At the end of the experiment, the average values of confidence levels were estimated for every hth session. In this study, we calculated an overall average value for the entire set of twenty design alternatives in a session irrespective of their *user ratings*, and then we also calculated average values of subsets of designs in a session that had that the same *user rating* (i.e., R_1, R_2, or R_3). Trend analyses were then performed on these average values using the Mann-Kendall hypothesis test [*Helsel and Hirsch*, 2002] to assess whether the average confidence levels monotonically increased or decreased over the span of sequential interaction sessions. A significance alpha level of 0.1 was used to determine the statistical significance of the trend. Since the main focus of this test was to minimize the risk of not detecting an existing trend (i.e., Type II error), a larger alpha value was chosen. At the end, three types of trends (positive, negative, and no trend) were identified for each of the average confidence level types (i.e., average confidence of all designs, average confidence of R_1 designs, average confidence of R_2 designs, and average confidence of R_3 designs) per participant.

2.4.2. Metrics Based on User Ratings for Assessing Efficiency of Interactive Optimization Algorithm

To evaluate how effective the interactive optimization algorithm is in finding desirable alternatives for individual users, we calculated the percent number of design alternatives with a specific *user rating* at the end of search experiments. This metric, $PRate_{i,k}$, was calculated using equation (3) below.

$$PRate_{i,k} = \frac{X_{i,k}}{TD} * 100, \quad (3)$$

where $PRate_{i,k}$ is the percentage of design alternatives with ith *user rating* (i.e., R_i in Table 1) found by the kth user, $X_{i,k}$ is the total number of designs with *user rating* R_i found by kth user, and TD is the total number

of designs presented to a participant. As explained earlier, all the participants were shown at least 260 design alternatives that were included in the initial *Introspection* session and first two cycles of the *Human-guided Search* sessions (i.e., I1, HS1-HS6 after I1, and HS1-HS6 after I2). Therefore, TD had a value of 260.

To examine the overall outcome for users working with one of the three *modelScen* simulation models scenarios (i.e., *modelScen* = *Model A-Surrogates*, *Model B-Surrogates*, or *Model B-Stakeholders*), a summary group metric $GPRate_{i,modelScen}$ in equation (4) was also estimated for all R_i *user ratings*. This metric was based on the average of the $PRate_{ik}$ for all users in each of the model scenario groups, and for a specific *i*th *user rating*.

$$GPRate_{i,modelScen} = \frac{\sum_{k=1}^{MaximumID} (PRate_{i,k} * bel_{k,modelScen})}{N_{modelScen}}, \tag{4}$$

where $N_{modelScen}$ is the total number of users in a particular *modelScen* model scenario group, and $bel_{k,modelScen} = 1$ if *k*th user belongs to *modelScen* group under consideration, else $bel_{k,modelScen} = 0$. Note that the summation term in the numerator sums from lowest value of participant ID (i.e., $k = 1$) to maximum value of participant ID (i.e., $MaximumID = 25$ in Figure 3).

2.4.3. Assessment of Similarities and Dissimilarities in Global Objective Space

The similarities and dissimilarities among design alternatives in the watershed-scale (i.e., global) objective function space (described in Table 1) were evaluated using a metric proposed by *Piemonti et al*. [2013] on the overall distance between Pareto Fronts. This distance metric was first estimated for each of the participants, and then an average of the metric values across the participants was calculated to summarize the results for participants in each of the *modelScen* model scenario groups.

For every participant, in order to assess the effect of interactive optimization on search results, the distance in the objective space between the design alternatives found via the participant's interactive search experiment and the design alternatives found via a noninteractive search was estimated. Note that even though the participant was not involved in the noninteractive search process, he/she had the opportunity to review and rate twenty of the noninteractive search's final nondominated design alternatives in the first introspection session *I1*. All design alternatives found via interactive optimization were first separated into three separate groups based on their *user rating* R_i (i.e., "I don't like it," "Neutral," and "I like it"). Then the distance metric was calculated by comparing the group of alternatives with specific R_i user rating with all the twenty nondominated design alternatives found earlier via noninteractive search. For the purpose of simplifying visualization, two physical objectives at a time (e.g., cost and peak flow reduction or cost and nitrates reduction) were selected to calculate and visualize the distance. This metric is based on an average relative Euclidean distance (equation (5)) and compares the position of every *j*th design alternative from the noninteractive Pareto Front, with each *d*th design alternative that has *i*th *user rating* and is from the set of solutions found via interactive search.

$$DR_{i,k}(A,B) = \frac{\sum_{j=1}^{J} \sum_{d=1}^{ND_{i,k}} \sqrt{(A_j - A_{i,d,k})^2 + (B_j - B_{i,d,k})^2}}{J * ND_{i,k}}. \tag{5}$$

$DR_{i,k}$ (A, B) represents the distance between two objective functions (*A* is the scaled value between 0 and 1 of peak flow reduction, sediment reduction, or nitrate reduction and *B* is the scaled value between 0 and 1 of cost), *J* is the total number of alternatives in the noninteractive Pareto Front shown to the user in first Introspection session *I1* and $ND_{i,k}$ is the total number of design alternatives with *i*th *user rating* found by the *k*th participant at the end of the interactive optimization experiment. The scaled *A* and *B* values were obtained from the original values of objective functions (i.e., peak flow reduction, sediment reduction, nitrate reduction, or cost) by using equation (6). In this equation, *X* is the original value of one of the four objective functions under consideration, *min X* is minimum value of the objective function under consideration and across all users working with the same simulation model (i.e., *Model A* or *Model B*), and *max X* is maximum value of the objective function under consideration and across all users working with the same simulation model (i.e., *Model A* or *Model B*).

$$X_{scaled} = \frac{X - min\,X}{max\,X - min\,X}, \tag{6}$$

where X_{scaled} can be *A* or *B*.

Once the distance metric was calculated for each participant, a representative group distance metric was also calculated in a manner similar to equation (4) for the multiple participants belonging to each of *modelScen* groups. Equation (7) shows how this group metric, $GDR_{i,modelScen}$, was calculated as the average of the distances $DR_{i,k}$ of all participants in a *modelScen* group, and for each of the *i*th *user ratings*. Note that the summation term in the numerator of equation (7) sums from lowest value of participant ID (i.e., $k = 1$) to maximum value of participant ID (i.e., *MaximumID* = 25 in Figure 3).

$$GDR_{i,modelScen}(A,B) = \frac{\sum_{k=1}^{MaximumID} (DR_{i,k}(A,B) * bel_{k,modelScen})}{N_{modelScen}}. \quad (7)$$

2.4.4. Assessment of Similarities and Dissimilarities in Local Objective Space for User's Influenced by Local "Subbasins of Interest"

As mentioned earlier, user's feedback can be based on what spatial scale they are most interested in. For example, stakeholders focused on local subbasin-scale impacts may be more interested in viewing and learning about the performance of proposed practices in their local subbasins of interests. Hence, their user rating may be more influenced by this local-scale performance, even though the physical objective functions in the genetic algorithm are based on the global watershed-scale performance of the recommended distribution of practices. To evaluate evidence of such user behavior in the test experiments, we examined for any potential overlaps or biases in performance functions—i.e., costs, peak flow reduction, sediment reduction, and nitrate reduction—at local scales for individual users, and potential similarities/dissimilarities in these biases among users with interests focused on local versus global scales. This was done by first generating the histograms of performances of desirable (i.e., R_3 *user ratings*) design alternatives at the individual subbasins of interest (Figure 3) that were of interest to participants.

3. Results and Discussion

This section has been divided into three subsections, each of which examines the results relevant to three research questions stated in section 1. In each subsection, we discuss results for the experimental scenarios described earlier for simulation models, participant types, and participant areas of interest. Because of the exploratory nature of this work, we have discussed comparisons among individual participants, including comparisons among groups. A major benefit of these comparisons is that they can help:

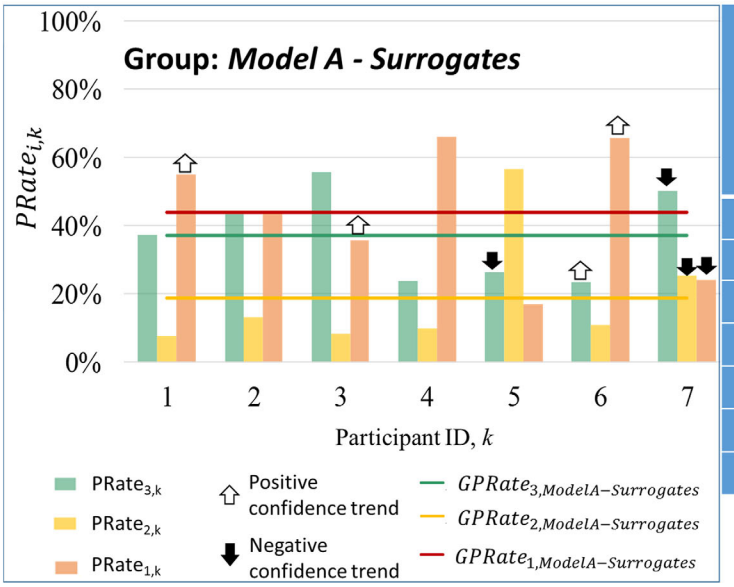
a. Improve our current understanding of how effective interactive optimization can be in generating personalized design alternatives for users with different interests, preferences, and behaviors.
b. Evaluate agreement among participants and variability across participants who belong to a certain group. This can further provide insight into whether, and to what extent, the personalized interactive search can also assist in the generation of desirable alternatives that may satisfy the requirements of most of the community members within a group, even though each individual user interacts independently with the WRESTORE system.

3.1. Assessment of User Behavior and Efficiency of the Interactive Optimization Algorithm

In this subsection, we investigate the first research question related to efficiency of the interactive genetic algorithm in generating user-preferred design alternatives, for different types of users with varying interests, preferences, and online interaction behavior.

3.1.1. Assessment for Individual Participants

In this study, the percentage of *user ratings* (equation (3)) provides insight into how successful the interactive optimization algorithm was in identifying desirable alternatives (i.e., alternatives with user rating R_3), for the different types of participants (*Surrogates* and *Stakeholders*), and for different watershed models (i.e., *Model A* and *Model B*) that were used to evaluate cost-benefits of the design alternatives. The light red, yellow, and green vertical bars in Figure 4 show the percentage of user ratings ($PRate_{i,k}$) at the end of the search experiment, and for each *i*th rating (i.e., R_i) ranked by each *surrogate* participant who was working with watershed *Model A*. Mann-Kendall trends in average confidence levels of design alternatives with the same R_i *user rating* were also estimated for each of the participant with ID "*k*." These temporal trends in average confidence levels are represented by white arrows (positive trend) and black arrows (negative trend) in this figure. Absence of arrow indicates a lack of a statistically significant temporal trend. Note that the trends in confidence levels provide an insight into the participant's learning process and a measure of self-trust of his/her own evaluation of alternatives through time. The red, yellow, and green horizontal lines



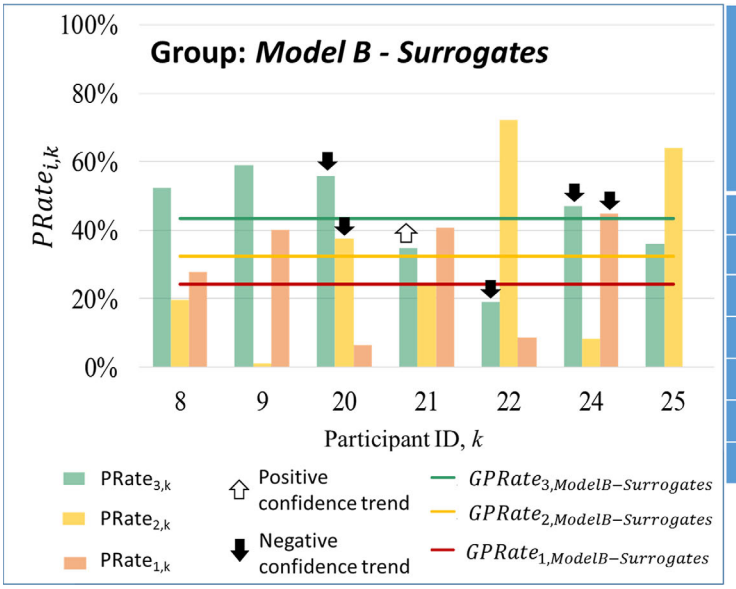
Participant	Average % of mouse Clicks made in Info	Average % of time spent in Info
1	1	3
2	72	62
3	60	56
4	10	14
5	40	45
6	66	67
7	14	34

Figure 4. Percentage of designs with different ratings, found by the participants in the group *Model A-Surrogates*.

in this figure indicate the group average values of $PRate_{1,k}$, $PRate_{2,k}$, $PRate_{3,k}$ across all participants, respectively. These group statistics for different user ratings were calculated using equation (4), and are discussed in details in section 3.1.2 on group assessment. The table in the top right corner of Figure 4 lists values of metrics used to track individual user's behavior (equations (1) and (2)) on the web-based GUI.

By the end of the experiments, as shown in bar graph of Figure 4, four (i.e., 57%) out of seven of these *surrogate* participants had a higher value of $PRate_{1,k}$ (for R_1 or "I don't like it" alternatives) than $PRate_{3,k}$. This indicates that more than half of these participants did not like most of the design alternatives produced by the interactive search algorithms. However, the confidence arrows indicate that only two of these four (i.e., 50%) *surrogate* participants experienced an increase (white arrows) in the average confidence levels of R_1 design alternatives over time (Participant 1 and Participant 6 in Figure 4). This suggests that only half of these four individuals by the end of the experiment were increasingly self-confident about the design alternatives they did not like. Moreover, when behavioral data on the participant's interaction with the GUI was also taken into account (see table in Figure 4), we observed that the Participant 6, unlike Participant 1, spent a considerable amount of time and made a large number of mouse clicks in tasks involving gathering of information on the GUI. This indicates that his/her assignment of R_1 *user rating* to design alternatives, and his/her self-confidence in his/her own assessment of designs, are most likely supported by a cognitive learning process that involved the use of semantic information on the design and the costs and benefits. Also, notice that for "I like it" alternatives, Participants 3 and 7 have the highest $PRate_{3,k}$ than $PRate_{2,k}$ and $PRate_{1,k}$. However, *user rating* provided by Participant 3 may be more reliable than Participant 7 because Participant 3 shows no evidence of decreasing confidence levels through time for her/his $PRate_{3,k}$, shows an increase in confidence levels for her/his $PRate_{1,k}$, and also has high percentage of her/his interaction with the tool in the information gathering areas (67% of time and 66% of mouse clicks in gathering information reported in the table). Participant 7, on the other hand, had significantly lower interaction with the GUI (34% of time and 14% of mouse clicks) in information gathering areas of the GUI, and also experienced a decrease in all of his/her confidence levels through time. This indicates that even though Participant 7 liked most of his/her designs, the user behavior, and self-confidence in his/her feedback do not suggest that the *user ratings* may be a reliable portrayal of her/his assessment of designs alternatives.

Figures 5 and 6 show the range of values of percentage of *user ratings* for participants in the *Model B-Surrogates* and *Model B-Stakeholder* experiments, respectively. Contrary to results of *Model A-Surrogates* in Figure 4, user experiments with artificially enhanced *Model B* (i.e., *Surrogates* and *Stakeholder*) generated a higher percentage of participants that found a high proportion of design alternatives that they liked. Specifically, four out of seven (i.e., 57%) of *Model B-Surrogates* participants and four out of six (66%) of the *Model*

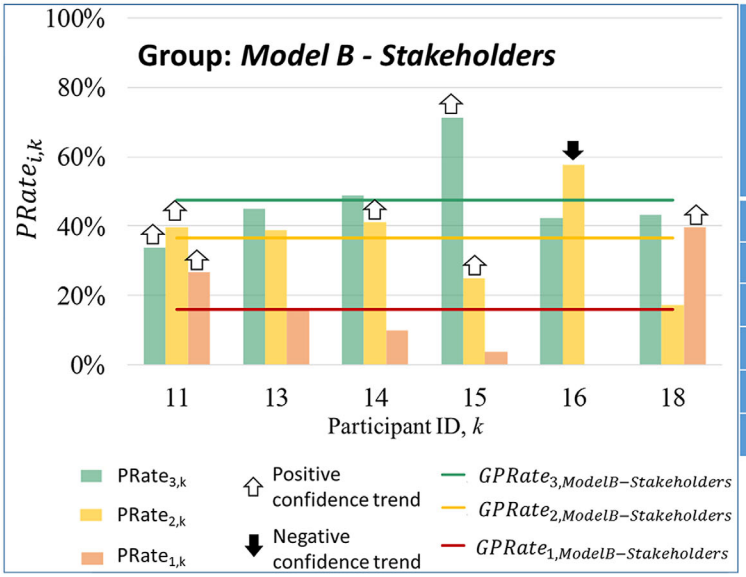


Participant	Average % of mouse Clicks made in Info	Average % of time spent in Info
8	0	0
9	57	47
20	71	36
21	34	47
22	1	0
24	21	15
25	2	1

Figure 5. Percentage of designs with different ratings, found by the participants in the group *Model B-Surrogates*.

B-Stakeholders participants had values for $PRate_{3,k}$ higher than $PRate_{1,k}$ and $PRate_{2,k}$. These results indicate that participants (stakeholders and nonstakeholders) seemed to be a lot more satisfied with their design alternatives when the watershed simulation model overpredicted the performance of the conservation practices, than participants in *Model A-Surrogates*.

In Figure 5 it is also interesting to note that *Model B-Surrogates* had only one participant (Participant 21) whose confidence level increased over time for design alternatives rated R_3, even though his/her $PRate_{3,k}$ was smaller than $PRate_{1,k}$. This participant has a moderate amount of interaction (the average percent of time in information gathering was 34% and the average percent of mouse clicks in information gathering was 47%) with the GUI. On the other hand, Participants 20, 22, and 24, had $PRate_{3,k}$ higher than $PRate_{1,k}$ (Figure 5), but with a decrease in their confidence level trends for R_3 designs over time. While the decrease in confidence level of Participants 22 and 24 is supported by their lack of time spent and mouse clicks made toward gathering of information, Participant 20 demonstrated unexpected interaction behavior and



Participant	Average % of mouse Clicks made in Info	Average % of time spent in Info
11	76	73
13	54	47
14	38	40
15	3	2
16	69	65
18	45	50

Figure 6. Percentage of designs with different ratings, found by the participants in the group *Model B-Stakeholders*.

confidence trends. This participant (i.e., Participant 20) spent moderate amount of time (average percent of time was 36%) and significant amount of browsing effort (average percent of mouse clicks was 71%) in gathering information, but did not yield any increases in trends of her/his self-confidence levels.

For *Model B-Stakeholders* (Figure 6) group, majority of the participants (i.e., four out of six participants or ~66%) were also found to have increasing confidence levels over time, similar to *Model B-Surrogates*. In particular, Participant 15 had $PRate_{3,k}$ close to 70%, suggesting that he/she was satisfied with most of the design alternatives found by him/her. However, unlike other participants, Participant 15 did not spend a lot of time interacting with the GUI to collect information on the attributes of individual design alternatives. This can be attributed to the fact that Participant 15 was interested in the entire watershed scale (Group 7 in Figure 3) and was not focused on a smaller region (local subbasins) of interest. Since the default visualization in the GUI shows information on decisions and cost-benefits at the watershed scale, this user did not need to do additional mouse clicks or use drop down menus in order to procure attribute information at this scale. Also, only one of the participants in this group presented a decrease in the confidence level trends (Participant 16) for design alternatives with *user rating* R_2. In summary, even though $PRate_{3,k}$ was found to be highest for the majority of the participants in both the two groups using *Model B*, most of the participants in *Model B-Stakeholders* were found to be increasingly (white arrows) confident in the accuracy of their user ratings throughout the experiment, contrary to participants in *Model B-Surrogates*.

3.1.2. Overall Group Assessment

The above results on individual participants demonstrate the unique differences in individual user behavior and search outcomes, when humans are included in the loop of interactive optimization experiments. To measure the differences in effectiveness of search at the group level, we compared the horizontal red, green, and yellow lines in Figures 4–6. These lines indicate the average of the percentage of solutions (for each *user rating*) across all participants in a group. Figure 4 shows that $GPRate_{1,modelScen} > GPRate_{3,modelScen-1,modelScen} > GPRate_{3,modelScen} > GPRate_{2,modelScen}$ when $modelScen = Model\ A\text{-}Surrogates$, whereas Figures 5 and 6 show that $GPRate_{3,modelScen} > GPRate_{2,modelScen} > GPRate_{1,modelScen}$ when $modelScen = Model\ B\text{-}Surrogates$ or *Model B-Stakeholders*. This indicates that when participants used the enhanced watershed *Model B*, the search algorithm was able to better identify a larger percent of "I like it" alternatives, thereby, delivering more design alternatives that the participants would be satisfied with. One potential reason that could be attributed toward this behavior is that majority of participants seemed to be more influenced by the performance of the decisions in the objective space. Hence, they indicated satisfaction with a lot more designs when *Model B* was used in contrast to *Model A*, since *Model B* overpredicted the performances. Second, a larger set of "I like it" alternatives also could have potentially made it easier for the underlying genetic algorithm to identify genes that coincided with higher user satisfaction, and hence, leading to better search effectiveness. However, to further examine and validate these potential reasons, a much larger sample size of user experiments, than what was possible in this study, would be required. While such types of experiments were beyond the scope of this initial observational study, the user behaviors observed in these experiments provide insight into potential hypotheses to test for future planned studies with large number of stakeholders.

3.2. Assessment of Similarities and Dissimilarities in Global Objective Space

In this subsection, we investigate the second research question related to similarity and dissimilarity among design alternatives (found via interactive as well as noninteractive search) in objective space of global, watershed-scale goals. The physical objective space reflects the range of physical environmental benefits and costs that different design alternatives would be expected to encompass. As also mentioned earlier, the similarity among results from multiple user experiments in watershed-scale objective space was assessed by calculating the distance (equation (5)) between the set of a nondominated design alternatives found by every participant and the initial set of noninteractively optimized design alternatives. Design alternatives generated via user's participation in interactive optimization were separated into three groups based on the user ratings R_i. Below is a discussion of similarities and dissimilarities in objective space of various design alternatives with *user rating* R_i in each set, and found by the participants belonging to groups *Model A-Surrogates*, *Model B-Surrogates*, and *Model B-Stakeholders*.

3.2.1. Assessment of Individual Participants

Table 3 presents the results of the distance metric calculated for participants in *Model A-Surrogates* group, for each of the user rating R_i, and using two objectives at a time (as described in equation (5)). It can be

Table 3. Distance Between Noninteractive and Interactive Pareto Fronts in Objective Space of Functions A (Peak Flow Reductions (PFR), Sediments Reduction (SR), or Nitrates Reduction (NR)) and B (Cost), for *Model A-Surrogates*

	A: PFR, B: Cost			A: SR, B: Cost			A: NR, B: Cost		
Participant, *k*	$DR_{1,k}$	$DR_{2,k}$	$DR_{3,k}$	$DR_{1,k}$	$DR_{2,k}$	$DR_{3,k}$	$DR_{1,k}$	$DR_{2,k}$	$DR_{3,k}$
1	0.22	0.09	0.12	0.24	0.08	0.11	0.22	0.07	0.09
2	0.25	0.16	0.11	0.27	0.17	0.12	0.23	0.15	0.10
3	0.21	0.14	0.11	0.24	0.15	0.14	0.20	0.14	0.11
4	0.19	1.3E-4	0.11	0.21	1.4E-4	0.11	0.18	1.2E-4	0.10
5	0.24	0.14	0.16	0.27	0.16	0.18	0.22	0.13	0.16
6	0.23	0.08	0.10	0.25	0.10	0.12	0.23	0.08	0.09
7	0.31	0.12	0.12	0.34	0.15	0.15	0.32	0.12	0.12

observed that for any pair of objective functions (e.g., cost versus *PFR*), the distances (e.g., $DR_{1,k}$) for the same *user rating* (e.g., R_1) was similar across all users. Note that this distance metric not only represents how far the centers of mass of two sets of design alternatives are from each other, but it also captures the spread of the alternatives around their center of mass. For all the participants in this group, it can be seen that $DR_{1,k}$ was greater than $DR_{3,k}$, suggesting that designs classified as R_3 (i.e., "I like it") are closer to the design alternatives in the noninteractive Pareto Front. For some participants—i.e., Participants 1, 4, 5, and 6—$DR_{3,k}$ was greater than $DR_{2,k}$. This could be attributed to the fact that for these participants the set of R_2 design alternatives was less spread than the set of R_3 design alternatives, leading to a smaller value of $DR_{2,k}$.

Figure 7, which graphically illustrates the distribution of design alternatives with R_1, R_2, and R_3 *user ratings* in the objective space, further exemplifies the difference in spreads for Participants 1 and 6. Participant 1 was interested in the performance of design alternatives at the scale of the entire watershed (Figure 7a), and Participant 6 was interested in the performance in a small set of local subbasins (SBs) (Figure 7b). From the perspective of quality of user interaction behavior, both of these participants had also demonstrated increasing trends in their self-confidence in evaluation of *user ratings*. Figure 7 also shows the set of design alternatives (labeled as "Noninteractive Pareto") that were used for the initial evaluation in Introspection 1 (I1) sessions, for all participants. These 20 design alternatives had been previously found via an exhaustive noninteractive search based on only the four physical objective functions (Cost, PFR, SR, and NR) and the calibrated *Model A*. This set of design alternatives gave the participant a good starting point for her/his search. The first observation that one can make from these figures is that both participants found multiple design alternatives with R_1 and R_2 *user ratings* in noninteractive Pareto Front that they did not find

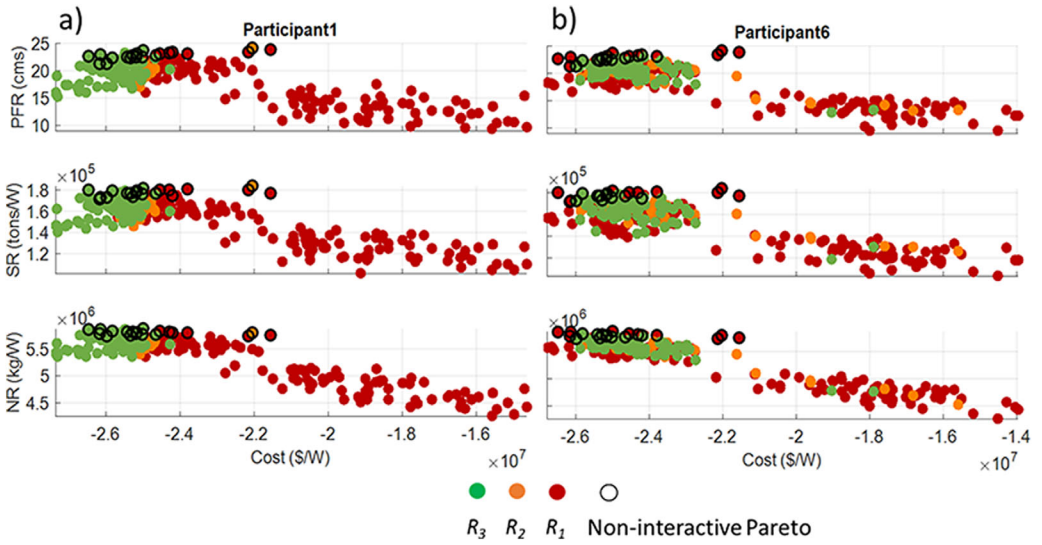


Figure 7. Pareto Front representation of watershed performance for (a) Participant 1 and (b) Participant 6 in *Model A-Surrogates* group. Participant 1 was asked to provide the *user rating* for the design alternative based on the watershed performance, while Participant 6 was asked to provide the user rating for the design alternative based on the group of SBs: 103, 105, 106, 121, and 122. Note that negative costs indicate positive revenue.

Table 4. Distance between Noninteractive and Interactive Pareto Fronts in Objective Space of functions A (Peak Flow Reductions (PFR), Sediments Reduction (SR) or Nitrates Reduction (NR)) and B (Cost), for *Model B-Surrogates* and *Model B-Stakeholders*

		A: PFR, B: Cost			A: SR, B: Cost			A: NR, B: Cost		
Participant, k		$DR_{1,k}$	$DR_{2,k}$	$DR_{3,k}$	$DR_{1,k}$	$DR_{2,k}$	$DR_{3,k}$	$DR_{1,k}$	$DR_{2,k}$	$DR_{3,k}$
Model B-Surrogates	8	0.50	0.41	0.33	0.47	0.41	0.35	0.46	0.41	0.35
	9	0.27	0.01	0.21	0.38	0.01	0.32	0.33	0.01	0.28
	20	0.63	0.47	0.29	0.58	0.44	0.28	0.57	0.45	0.29
	21	0.33	0.37	0.46	0.33	0.37	0.45	0.32	0.36	0.45
	22	0.34	0.39	0.41	0.38	0.38	0.39	0.37	0.38	0.40
	24	0.40	0.37	0.41	0.39	0.37	0.40	0.39	0.38	0.40
	25	0	0.39	0.38	0	0.39	0.36	0	0.38	0.35
Model B-Stakeholders	11	0.24	0.13	0.26	0.37	0.19	0.42	0.32	0.17	0.37
	13	0.29	0.28	0.18	0.41	0.36	0.26	0.36	0.32	0.23
	14	0.24	0.30	0.18	0.36	0.37	0.26	0.31	0.34	0.23
	15	0.30	0.11	0.24	0.47	0.16	0.37	0.41	0.13	0.33
	16	0	0.23	0.28	0	0.30	0.36	0	0.28	0.33
	18	0.45	0.23	0.42	0.44	0.23	0.41	0.45	0.23	0.41

desirable. Additionally, both participants, along with other participants not shown in this figure, found a much greater number of R_3 ("I like it") designs in the space slightly suboptimal to noninteractive Pareto Front. The shape and size of this desirable region of alternatives with R_3 designs were found to be unique to each participant, with potential overlaps in some areas. It can also be seen that for Participant 1 there are distinct clustered regions in the objective space where the user found most of her/his preferred and less-preferred design alternatives. In this example, cost at the watershed scale seems to be the deciding criteria based on which a user decided the R_1, R_2, and R_3 *user ratings*. However, this clear distinction in the regions of desirable and less-desirable alternatives does not seem to exist for Participant 6, who was more concerned about the design at the subbasin scale. Even when most of the preferred design alternatives lie on the low cost (on the left side) region of the objective space for Participant 6, there are multiple design alternatives in the same region that were also rated R_1 by this participant. This also indicates that while the user rating objective function was able to guide the optimization algorithm to identify large number of design alternatives with R_3 rating, the search process was less sensitive to watershed-scale performance for this user. Another conclusion that can be made by this result is that for many humans watershed-scale performance might not be the only criteria for identifying solutions that are acceptable to them.

Table 4 presents the $DR_{i,k}$ for participants in groups *Model B-Surrogates* and *Model B-Stakeholders*. Notice that the values of these distances are significantly higher than for *Model A-Surrogates* in Table 3. This is an effect of the artificially enhanced peak flow reduction, nitrate reduction, and sediment reduction estimated by *Model B* during interactive optimization, as explained in the Methodology section. Results of *Model B-Surrogates* showed that only 43% of the participants seem to have $DR_{1,k} > DR_{3,k}$, while in contrast 83% of the participants for *Model B-Stakeholders* showed distances values of $DR_{1,k} > DR_{3,k}$. Overall, these results suggest that in both the groups not all participants necessarily preferred solutions with enhanced PFR, SR, and NR values closer to the Pareto Front of noninteractive design alternatives.

Figure 8 illustrates examples of the distribution of *user ratings* in the watershed-scale objective space for two participants (IDs 8 and 20) in *Model B-Surrogates* and two participants (IDs 15 and 11) in *Model B-Stakeholders* groups. Note that Participants 8 and 15 rated design alternatives based on their performance at the scale of the entire watershed (Figures 8a and 8c), and Participants 22 and 11 rated designs based on the performance of alternatives in a particular subset of local subbasins (Figures 8b and 8d). Notice that even when Figures 8a and 8c are for participants who were concerned with optimizing the solutions for the entire watershed, the design alternatives with different *user ratings* are considerably scattered in the objective space for Participant 15 in Figure 8c. Participant 8 in Figure 8a, on the other hand, has well-defined clusters of R_1, R_2, and R_3 design alternatives, similar to Participant 1 from *Model A-Surrogates* (shown in Figure 7). Conversely, Figure 5 indicates that Participant 8 did not have any clear trend in confidence levels, whereas Participant 15 had an increase in self-confidence levels over time for R_2 and R_3 ratings (Figure 6). Hence, the results of Participant 15's user experiments could be considered more reliable from the user's perspective, in spite of the lack of clear clusters in objective space. This

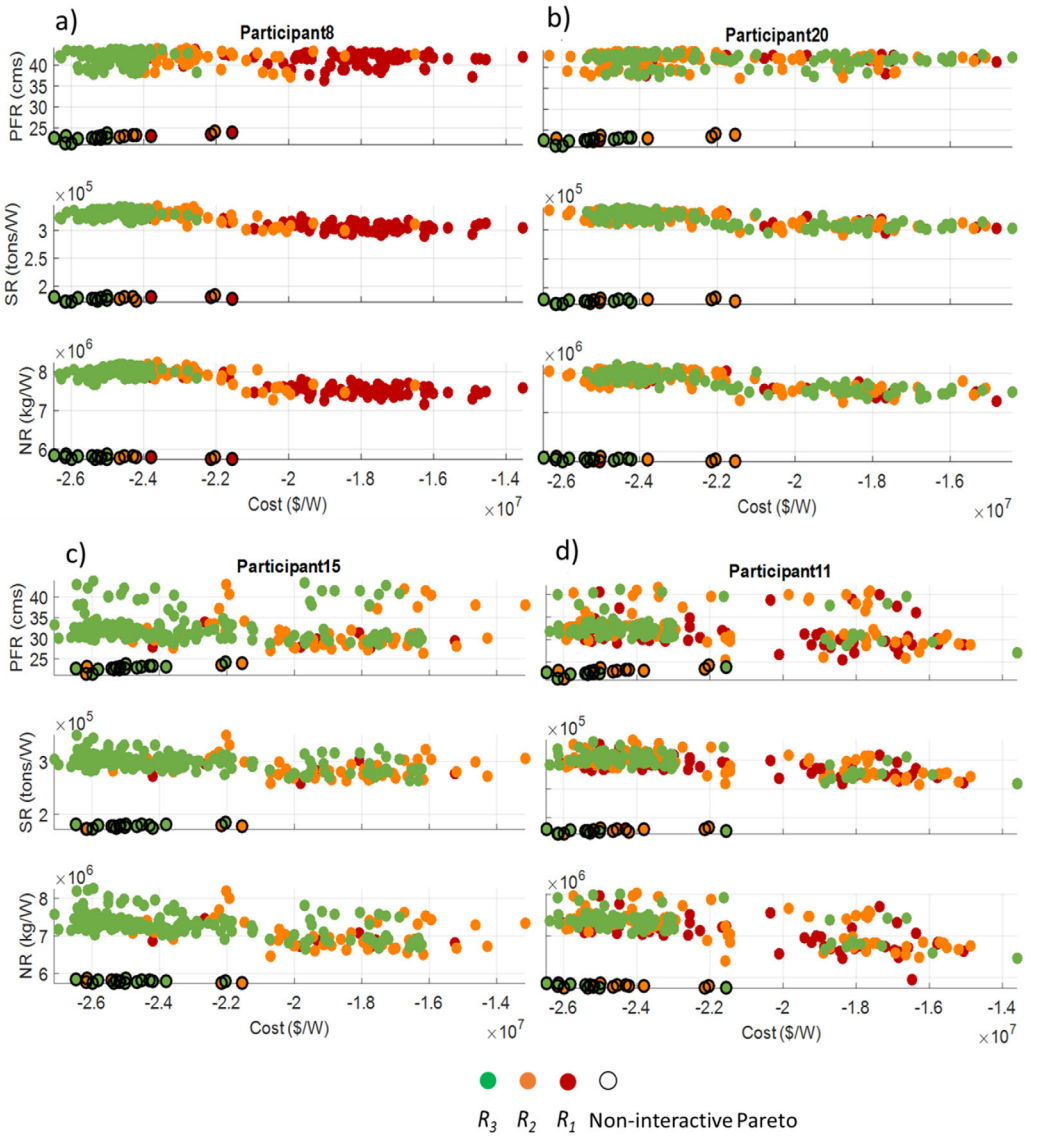


Figure 8. Pareto Front representation of watershed performance for (a) Participant 8 and (b) Participant 20 in *Model B-Surrogates* group. Participant 8 was asked to provide the *user rating* for the design alternative based on the watershed performance, while Participant 20 was asked to provide the user rating for the design alternative based on the group of SBs: 58, 59, 61, and 63. Similarly, the Pareto Fronts for (c, d) Participants 15 and 11, respectively, in *Model B-Stakeholders* group are shown. Participant 15 was asked to provide the *user rating* for the design alternative based on the watershed performance, while Participant 11 was asked to provide the user rating for the design alternative based on the group of SBs: 41, 90, 92, and 93. Note that negative costs indicate positive revenue.

further suggests that Participant 15 is a lot more flexible about her/his criteria in objective space, and not necessarily motivated by the need to mathematically optimize watershed-scale goals to the best possible value. Two other examples in Figure 8 of Participants 11 and 20 who rated the design alternatives based on a particular set of SBs have R_1, R_2, and R_3 designs overlapping each other in watershed-scale objective space. Comparing these with the confidence trends in Figure 6 indicates that Participant 11 has an increase in confidence levels for all the trends, thereby suggesting a higher reliability in the acceptability of these designs. In contrast, Participant 20 has negative trends for alternatives with *user ratings* R_3 and R_2 in Figure 5. Moreover, even when both participants seem to have a good percentage of mouse click events (76 and 71% respectively), the discrepancies in the average percentage time spent (73 and 36% respectively) also suggest that the reliability of the results of Participant 11 may be higher than that of the results found by Participant 20.

Table 5. Group Averages, $GDR_{i,modelScen}$, for Distance Between Noninteractive Pareto Front and Design Alternatives Found by Participants via Interactive Optimization

	A: PFR, B: Cost			A: SR, B: Cost			A: NR, B: Cost		
Group, *modelScen (or mS)*	$GDR_{1,\ mS}$	$GDR_{2,\ mS}$	$GDR_{3,\ mS}$	$GDR_{1,\ mS}$	$GDR_{2,mS}$	$GDR_{3,mS}$	$GDR_{1,mS}$	$GDR_{2,mS}$	$GDR_{3,mS}$
Model A-Surrogates	0.24	0.10	0.12	0.26	0.12	0.13	0.23	0.10	0.11
Model B-Surrogates	0.35	0.34	0.36	0.36	0.34	0.36	0.35	0.30	0.34
Model B-Stakeholders	0.25	0.21	0.26	0.34	0.27	0.35	0.32	0.25	0.32

3.2.2. Overall Group Assessment

Table 5 shows the group averages in distance metrics for the different participant groups and models, calculated using equation (7). Notice that for group working with *Model A*, the values of $GDR_{1,modelScen}$ was found to be greater than $GDR_{3,modelScen}$ for the different combinations of A and B objective functions. This suggests that this group, on an average, preferred design alternatives located closer to the design alternatives on the noninteractive Pareto Front. The groups working with *Model B*, on the other hand, had values of $GDR_{1,modelScen}$ close to values of $GDR_{3,modelScen}$. This finding seems counterintuitive, at first glance, for participants working with *Model B-Surrogates* or *Model B-Stakeholders*. Note that *Model B* overestimated the *PFR*, *SR*, and *NR* benefits in comparison to *Model A*, and *Model A* was used to estimate benefits of alternatives on the noninteractive Pareto Front. Hence, a user assessing the quality of the design alternative based on only the physical objective function values estimated by *Model B* should be expected to prefer designs with higher $GDR_{3,modelSecn}$ than $GDR_{1,modelScen}$. However, this was not always observed, indicating that even *Model B* participants may not have been entirely motivated by the performance of design alternative estimated at the watershed scale in order to decide what design alternatives they liked. Additional factors may have been more important to these participants when they were evaluating design alternatives. For example, some participants may have been more influenced by the value of the design decisions (e.g., certain locations may be more favorable for a BMP from the user's perspective, in spite of the performance). This issue will be examined in a forthcoming article.

3.3. Assessment of Similarities and Dissimilarities in Local Objective Space for User's Influenced by Local "Subbasins of Interest"

In this subsection, we present results for the third research question related to similarity and dissimilarity among design alternatives in objective space of local, subbasin-scale goals. Note that even though the underlying optimization algorithm used four of the cost-benefit objective functions to assess watershed-scale goals (Table 1), the participants who were focused on local subbasin scales (Figure 3) had the option to provide their *user ratings* based on their perception of costs and benefits at the subbasin scale. Hence, the *user rating* gave participants an indirect mechanism to guide the multiobjective search on the basis of performance at subbasin scales, in addition to watershed-scale performance.

To assess for similarities and dissimilarities in subbasin-scale performance across participants, we generated and compared histograms of subbasin-scale objective function values for participants in the six ($q = 1$ to 6) groups of subbasins of interest ($SBint_q$ shown in Figure 3). Figure 9 shows the scaled histograms for R_3 ("I like it") design alternatives, found by only six of the participants whose watershed-scale objective function values were shown earlier in Figures 7 and 8. Figures 9a–9f are histograms of pairs of participants in groups *Model A-Surrogates*, *Model B-Surrogates*, and *Model B-Stakeholders*, respectively. For each pair, the histogram on the right indicates the distribution of objective function values in the $SBint_q$ that were of interest to the participants focused on the subbasin-scale goals (i.e., Participants 6, 20, and 11 in Figures 9b, 9d, and 9f, respectively), whereas the histograms on the left illustrate the subbasin-scale distributions of design alternatives for the same subbasins on the right, but found by participants who were instead focused on the larger watershed-scale goals (i.e., Participants 1, 8, and 15 in Figures 9a, 9c, and 9e, respectively). The curved blue arrows indicate the hydrologic connectivity in the subbasins of interest; the directions of arrows indicate the direction of flow from upstream subbasin to downstream subbasin. For example, in Figures 9a and 9b, Participant 6 was focused on local, subbasins belonging to sixth group in Figure 3. In these subbasins, with IDs 103, 105, 106, 121, and 122, subbasin 105 is directly upstream of 106, and subbasin 121 is directly upstream of 122. Hence, the blue arrows are directed from 105 to 106 and 121 to 122 in Figures 9a and 9b.

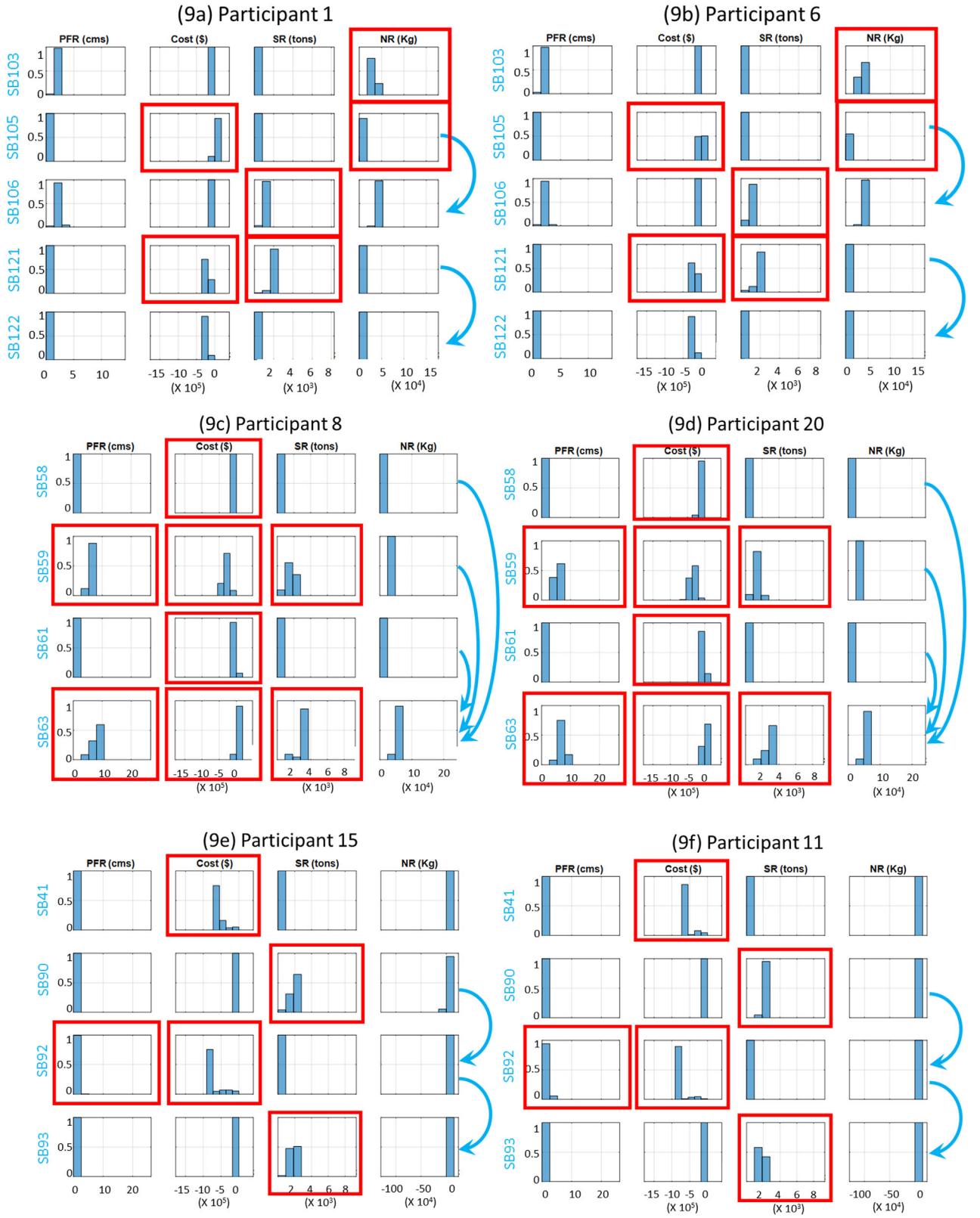


Figure 9. Scaled histograms of subbasin-scale objective function values of "I like it" design alternatives, found by Participants 1 and 6 in *Model A-Surrogates* group, Participants 8 and 20 in *Model B-Surrogates* group, and Participants 15 and 11 in *Model B-Stakeholders* group. The *x* axes gives the magnitude of objective functions (PFR, Cost, SR, and NR) listed on the top, and the *y* axes indicate the frequency of designs with corresponding objective function value on the *x* axis. The light blue curvy arrows indicate the direction of streamflow between subbasins that are hydrologic connected. The red boxes indicate histograms where noticeable dissimilarities between left and right histograms exist. Note that negative costs indicate positive revenue.

In Figure 9, at first glance, the histograms of "I like it" design alternatives found by the participants focused on goals at the entire watershed scale (Figures 9a, 9c, and 9e) look somewhat similar to histograms of participants on the right (Figures 9b, 9d, and 9f). On closer inspection, however, noteworthy differences can be detected among some of the histograms. These dissimilar histograms are indicated by red boxes drawn around them in Figure 9. The first thing to notice in these dissimilar histograms is the difference in distribution shapes between the participant focused on watershed scale on the left and participant focused on subbasin scale on the right. For example, in Figures 9a and 9b, the red-outlined distribution of nitrate reductions of subbasin 103 are peaked at lower value for Participant 1 than for Participant 6. Similar differences in the location of the peak along the *x* axis and the peak height along the *y* axis can also be noticed for other pairs of dissimilar histograms in Figures 9a–9f. While the occurrence of dissimilarities can be considered as potential outcomes of the algorithm's search operations (specifically, crossover, mutation, and selection operations of the underlying genetic algorithm in WRESTORE), the differences in shape of distributions and the intensity of peaks in distributions of user-favored designs also indicate a possible influence of users' unique preferences and known biases for specific subbasins and criteria. However, in order to confirm the extent of individual contributions from user preferences and search operations on the occurrence of dissimilarities, additional research needs to be conducted in the future with larger number of participants and detailed tracking of user's mouse clicks beyond those done for this study.

The second thing to notice in Figure 9 is that while for some participants there is no one criterion that dominates the dissimilar histograms (e.g., Figure 9a versus Figure 9b, and Figure 9e versus Figure 9f), for others one unique criterion clearly dominated where the histograms of left participant were different from those of the right participant. For example, in Figure 9c versus Figure 9d the histogram for cost function between watershed-scale Participant 8 and subbasin-scale Participant 20 were different for all of the subbasins 58, 59, 61, and 63. Hence, Participant 20, who was only interested in subbasins 58, 59, 61, and 63, was clearly driven by values of the costs objective function in determining her/his feedback (on *user ratings*) to the interactive genetic algorithm.

The third thing to notice in Figure 9 is that for most of the hydrologically connected subbasins occurrence of dissimilarities among histograms in upstream subbasins coincided with dissimilarities among histograms in downstream subbasins. These dissimilarities in histograms of downstream subbasins were a lot more noticeable when there were multiple dissimilar histograms in multiple upstream subbasins. For example, in Figures 9c and 9d, SBs 58, 59, and 61 drain into 63, and have a total of five dissimilar histograms in the three upstream SBs that coincide with three noticeably dissimilar cost-benefit histograms in downstream SB 63. In contrast, Figures 9a and 9b and Figures 9e and 9f have only one upstream subbasin for every downstream subbasin and with one or two dissimilar histograms in the upstream subbasins; these coincide with less noticeable dissimilarities in histograms of downstream SBs. While in these experiments it was not possible to know whether the participants were as much concerned about the downstream subbasins as about the upstream subbasins in the $SBint_q$, this result has general implications for subbasin-scale decision makers (e.g., Participants 6, 20, and 11) who may guide the algorithm's search based on subbasin-scale criteria. For such decision makers, if histograms of subbasin-scale objective functions are significantly different than those of the designs found by the watershed-scale decision makers (e.g., Participant 1, 8, and 15) for the same subbasins, then the benefits (and any cost or revenue impacted by the benefits) in the downstream subbasins may also end up being noticeably different because of the hydrologic connectivity that conveys benefits to lower subbasins.

4. Conclusions and Future Work

In this research, we conducted an observational study to improve our understanding of how variability in the nature of users' behavior, preferences, interests, and feedback can affect the results of an interactive genetic algorithm employed to search for user-desired design alternatives of watershed conservation plans. In this paper, we focused on examining the user-machine-generated results in objective function space of the conservation planning problem. While not all design alternatives found via this interactive search process were rated "I like it" (R_3 *user rating*) by participants, 55% of the participants found a higher percentage of design alternatives that they liked than percentages of alternatives rated R_1 ("I don't like it") or R_2 ("Neutral"). This suggests that over half of the population were able to find more options for acceptable design

alternative when they were engaged in the search process in collaboration with the underlying interactive genetic algorithm than with the number of desirable designs generated by the noninteractive genetic algorithm. The effectiveness of IGA in finding user-desired design alternatives has also been reported by other studies that have investigated IGAs for water resources problems [e.g., *Babbar-Sebens and Minsker*, 2008, 2010, 2012; *Singh et al*., 2008]. This demonstrates the effectiveness (Research Question 1) of such interactive search methods for assisting users during participatory design of conservation plans.

We also examined variability (Research Question 1) in user behavior of *surrogates* as well as *stakeholder* participants via usability metrics (e.g., time spent in gathering information, etc.), and via evaluation of users' feedback on self-confidence levels and *user ratings*. The usability and confidence metrics were used to evaluate the quality of individual user's participation and have implications for research on group decision-making approaches that engage multiple types of individuals with varied backgrounds via web-based design platforms. Note that user behavior metrics are important not only for comparing DSS tools and measuring GUI efficiency, but, as illustrated in this study, can be useful for evaluating the nature of a user's interaction with the tool. User's GUI interactions can further help improve our understanding on whether their behavior correlates with their own individual learning process. However, it is also worth noting that user learning can be noisy and multidimensional in nature and can happen in a noncontinuous manner; hence, we recommend that additional methods for observing the learning process of interacting users are needed so that developers of decision support tools can integrate them in their tools in the future. Such an improved understanding on user behavior and learning has the potential to also enable researchers to develop more human-like simulation models of their stakeholders' preferences (i.e., user surrogates or "avatars").

The use of *Model A* versus enhanced *Model B* helped us identify how users would respond to values of costs and benefits that were either closely accurate or artificially inflated. The user-preferred design alternatives identified for both the cases of models were found to be in the region close to the design alternatives generated via noninteractive search, implying that users were not necessarily only driven by the goal to "optimize" the physical cost-benefit objective functions. They may have also been driven by the values of the decision variables themselves; hence, we recommend that this issue should be further investigated. It is also worth pointing out that the information contained in Figure 8 can be computed "on-the-fly" as a user is interacting with the DSS, and can be used to compute a measure of "reliability" of the user (based on the distance from the noninteractive Pareto Front and the spread in the user-desired solutions). A possible consequence of such a reliability computation could, then, be to exclude or give less weight to an "unreliable" user in a group decision-making exercise, if a user prefers solutions with very large distances from the noninteractive Pareto Front. Such unreliable users could be inexperienced or even malicious. However, a detailed study of such dynamics of a collaborative decision-making exercise was beyond the scope of this paper, and is recommended for potential future research.

In this research, similarities and dissimilarities in the objective function space of user-desired alternatives generated by participants in the groups *Model A-surrogates*, *Model B-surrogates*, and *Model B-stakeholders* were also evaluated (Research Questions 2 and 3). While most participants who based their *user ratings* on watershed-scale goals tended to find desirable and undesirable alternatives that were clearly clustered in separate regions (except for Participant 15), participants who were focused on subbasin-scale goals did not exhibit such clustered patterns at the watershed scale. Moreover, for these participants with interests at subbasin scales the histograms of subbasin-scale goals were also found to have distributions different than the histograms of subbasin-scale goals generated and preferred by the watershed-scale participants. While this observational study provides an important insight into the types of patterns one may find in the solutions generated by different humans with different scale biases and interests, we expect these findings to be transferrable to studies involving other watersheds. This will especially be true when the stakeholder group consists of individuals with watershed-scale interests and individuals with local-scale interests, e.g. a group of agency personnel and landowners.

Finally, for most of the hydrologically connected subbasins in this study's watershed, occurrence of dissimilarities among histograms of participants in upstream subbasins coincided with occurrence of dissimilarities among histograms of participants in downstream subbasins. These dissimilarities in histograms of downstream subbasins were a lot more noticeable when the number of dissimilar histograms in multiple

upstream subbasins increased. We recommend that in order to evaluate the generality of this finding, similar analyses should be conducted in future studies on other watersheds with complex subbasin networks.

Acknowledgments
This study has been supported by U.S. National Science Foundation (award IDs 1332385 and 1014693). We would also like to thank Jill Hoffmann of Empower Results LLC for her help with organizing multiple workshops with stakeholders and to enable testing of the tool and Jon Eynon for the website and the initial interface development. We are grateful to the many volunteers and students who have participated in the user tests and/or participated in conducting research and code developments on multiple aspects of this tool. Finally, Institutional Review Board (IRB) permission was obtained for conducting the experiments with humans for this study. Hence, while we cannot make the raw data publicly available, anonymized data are available upon request, as per IRB guidelines, by the corresponding author (meghna@oregonstate.edu).

References

Arabi, M., R. S. Govindaraju, and M. M. Hantush (2006), Cost-effective allocation of watershed management practices using a genetic algorithm, *Water Resour. Res.*, *42*, W10429, doi:10.1029/2006WR004931.

Artita, K., P. Kaini, and J. W. Nicklow (2008), Generating alternative watershed-scale BMP designs with evolutionary algorithms, paper presented at World Environmental Water Resources Congress, pp. 12–16, Honolulu, Hawaii, doi:10.1061/40976(316)127.

Assaf, H., et al. (2008), Chapter thirteen generic simulation models for facilitating stakeholder involvement in water resources planning and management: A comparison, evaluation, and identification of future needs, *Dev. Integr. Environ. Assess.*, *3*, 229–246.

Babbar-Sebens, M., and B. S. Minsker (2008), Standard interactive genetic algorithm—Comprehensive optimization framework for groundwater monitoring design, *J. Water Resour. Plann. Manage.*, *134*, 538–547.

Babbar-Sebens, M., and B. Minsker (2010), A case-based micro interactive genetic algorithm (CBMIGA) for interactive learning and search: Methodology and application to groundwater monitoring design, *Environ. Modell. Software*, *25*, 1176–1187.

Babbar-Sebens, M., and B. S. Minsker (2012), Interactive genetic algorithm with mixed initiative interaction for multi-criteria ground water monitoring design, *Appl. Soft Comput.*, *12*(1), 182–195.

Babbar-Sebens, M., et al. (2013), Spatial identification and optimization of upland wetlands in agricultural watersheds, *Ecol. Eng.*, *52*, 130–142.

Babbar-Sebens, M., et al. (2015), A web-based software tool for participatory optimization of conservation practices in watersheds, *Environ. Modell. Software*, *69*, 111–127.

Belton, V., et al. (2008), Interactive multiobjective optimization from a learning perspective, *Multiobjective Optimization, Lecture Notes in Computer Science*, edited by J. Branke, K. Deb, K. Miettinen and R. Słowiński, vol. 5252, pp. 405–433, Springer, Berlin, Heidelberg.

Fedra, K. (1992), *Advanced Computer Applications, Options*, Int. Inst. for Appl. Syst. Anal., Laxenburg, Austria.

Fischer, I., and D. V. Budescu (2005), When do those who know more also know more about how much they know? The development of confidence and performance in categorical decision tasks, *Organ. Behav. Human Decis. Processes*, *98*(1), 39–53.

Fisher, M. L. (1985), Interactive optimization, *Ann. Oper. Res.*, *5*(3), 539–556.

Fry, D. L., et al. (2015), Appendix A: Fire and forest ecosystem health team final report, in *Learning How to Apply Adaptive Management in Sierra Nevada Forests: An Integrated Assessment. Final report of the Sierra Nevada Adaptive Management Project*, edited by P. Hopkinson and J. J. Battles, 72 p., Center for Forestry, UC Berkeley, Berkeley, Calif. [Available at http://snamp.cnr.berkeley.edu/snampfinal-report/index.html.]

Georgakakos, A. P., and Q. W. Martin (Eds.) (1996), An international review of decision support systems in river basin operation, in *Proceedings of the Fifth Water Resources Operations Management Workshop*, Am. Soc. of Civ. Eng., Arlington, Va.

Gunderson, L., and C. S. Holling (2002), *Panarchy Synopsis: Understanding Transformations in Human and Natural Systems*, Island Press, Washington, D. C.

Hamilton, S. H., et al. (2015), Integrated assessment and modelling: Overview and synthesis of salient dimensions, *Environ. Modell. Software*, *64*, 215–229.

Hamlet, A., et al. (1996a), *Basic STELLA II User's Manual for the ACT-ACF Shared Vision Models*, U.S. Army Corps of Eng., Mobile, Ala.

Hamlet, A., et al. (1996b), *Simulating Basinwide Alternatives Using the ACT-ACF Shared Vision Models*, U.S. Army Corps of Eng., Mobile, Ala.

Helsel, D. R., and R. M. Hirsch (2002), Statistical methods in water resources techniques of water resources investigations, *U.S. Geol. Surv. Tech. Water Resour. Invest., Book 4, Chap. A3*, 522 pp.

Kahneman, D., and A. Tversky (1979), Prospect theory: An analysis of decision under risk, *Econometrica*, *47*(2), 263–292.

Kelly, M., et al. (2012), Expanding the table: The web as a tool for participatory adaptive management in California forests, *J. Environ. Manage.*, *109*, 1–11.

Klau, G., et al. (2010), Human-guided search, *J. Heuristics*, *16*(3), 289–310.

Lethbridge, M. R., et al. (2010), Optimal restoration of altered habitats, *Environ. Modell. Software*, *25*(6), 737–746.

Lorenzoni, I., et al. (2000), A co-evolutionary approach to climate change impact assessment: Part I. Integrating socio-economic and climate change scenarios, *Global Environ. Change*, *10*(1), 57–68.

Loucks, D. P., and J. R. Da Costa (Eds.) (1991), *Decision Support Systems, NATO Ser.*, Springer, Berlin.

Matsuura, K., Willmott, C. and Legates, D. (2009), Web-Based, Water-Budget, Interactive, Modeling Program (WebWIMP). [Available at: http://climate.geog.udel.edu/~wimp/.]

McIntosh, B. S., et al. (2011), Environmental decision support systems (EDSS) development—Challenges and best practices, *Environ. Modell. Software*, *26*(12), 1389–1402.

Meignan, D., et al. (2015), A review and taxonomy of interactive optimization methods in operations research, *ACM Trans. Interact. Intell. Syst.*, *5*(3), 1–43.

Metcalfe, J., and A. P. Shimamura (1994), *Metacognition: Knowing About Knowing*, MIT Press, Cambridge, Mass.

Millington, J. D. A., D. Demeritt, and R. Romero-Calcerrada (2011), Participatory evaluation of agent-based land-use models, *J. Land Use Sci.*, *6*(2–3), 195–210.

Neitsch, S., et al. (2005), Soil and water assessment tool, theoretical documentation, version 2005, Grassland Soil and Water Res. Lab., ARS, Blackland Res. Cent., Tex. Agric. Exp. Stn., Temple, Tex.

Nelson, T. O. (1996), Consciousness and metacognition, *Am. Psychol.*, *51*, 102–116.

Nielsen (2009), Global faces and networked places: A Nielsen report on social networking ' s new global footprint, 16 pp. [Available at http://blog.nielsen.com/nielsenwire/wp-content/uploads/2009/03/nielsen_globalfaces_mar09.pdf.]

Palmer, R. N. (1998), A history of shared vision modeling in the ACT-ACF comprehensive study: A modeler's perspective, in *Coordination: Water Resources and Environment*, pp. 221–226, Am. Soc. of Civ. Eng., Reston, Va.

Palmer, R. N., A. M. Keyes, and S. Fisher (1995), Empowering stakeholders through simulation in water resources planning, in *Water Management in the '90s@sA Time for Innovation*, pp. 451–454, Am. Soc. of Civ. Eng., Seattle, Wash. [Available at http://cedb.asce.org/cgi/wwwdisplay.cgi?9302067.]

Perez-Pedini, C., J. Limbrunner, and R. Vogel (2005), Optimal location of infiltration-based best management practices for storm water management, *J. Water Resour. Plann. Manage.*, *131*, 441–448.

Piemonti, A. D., M. Babbar-Sebens, and E. Jane Luzar (2013), Optimizing conservation practices in watersheds: Do community preferences matter?, *Water Resour. Res.*, *49*, 6425–6449, doi:10.1002/wrcr.20491.
Piemonti, A. D., K. Macuga, and M. Babbar-Sebens (2017), Usability evaluation of an interactive decision support system for user-guided design of scenarios of watershed conservation practices, *J. Hydroinf.*, doi:10.2166/hydro.2017.017, in press.
Prell, C., et al. (2007), If you have a hammer everything looks like a nail: Traditional versus participatory model building, *Interdiscip. Sci. Rev.*, *32*(3), 263–282.
Randhir, T. O., J. G. Lee, and B. Engel (2000), Multiple criteria dynamic spatial optimization to manage water quality on a watershed scale, *Trans. ASAE*, *43*, 291–299.
Read, L., K. Madani, and B. Inanloo (2014), Optimality versus stability in water resource allocation, *J. Environ. Manage.*, *133*, 343–354.
Reder, L. M. (1996), *Implicit Memory and Metacognition*, Lawrence Erlbaum Assoc., Mahwah, N. J.
Reed, M. S. (2008), Stakeholder participation for environmental management: A literature review, *Biol. Conserv.*, *141*(10), 2417–2431.
Rosenberg, D. E., and K. Madani (2014), Water resources systems analysis: A bright past and a challenging but promising future, *J. Water Resour. Plann. Manage.*, *140*(4), 407–409.
Schramm, G. (1980), Integrated river basin planning in a holistic universe, *Nat. Resour. J.*, *20*, 787–806.
Schwartz, B. L. (1994), Sources of information in metamemory: Judgments of learning and feelings of knowing, *Psychon. Bull. Rev.*, *1*, 357–375.
Seppelt, R., and A. Voinov (2002), Optimization methodology for land use patterns using spatially explicit landscape models, *Ecol. Modell.*, *151*, 125–142.
Simon, H. A. (1955), A behavioral model of rational choice, *Quar. J. Econ.*, *69*(1), 99–118.
Simon, H. A. (1977), *The New Science of Management Decision*, Prentice-Hall, Upper Saddle River, N. J.
Singh, A., B. S. Minsker, and A. J. Valocchi (2008), An interactive multi-objective optimization framework for groundwater inverse modeling, *Adv. Water Resour.*, *31*(10), 1269–1283.
The Conservation Fund (2016), Mississippi River Basin/Gulf Hypoxia initiative conservation blueprint v 1.0, final report. [Available at https://lccnetwork.org/sites/default/files/MRB-GHI_Conservation_Blueprint_v1_TCF_Report_Final_wAppendices.pdf.]
Tomer, M. D., et al. (2015a), Agricultural conservation planning framework: 1. Developing multipractice watershed planning scenarios and assessing nutrient reduction potential, *J. Environ. Qual.*, *44*(3), 754–767.
Tomer, M. D., et al. (2015b), Agricultural conservation planning framework: 2. Classification of riparian buffer design types with application to assess and map stream corridors, *J. Environ. Qual.*, *44*, 768–779.
Van Asselt Marjolein, B. A., and N. Rijkens-Klomp (2002), A look in the mirror: Reflection on participation in integrated assessment from a methodological perspective, *Global Environ. Change*, *12*(3), 167–184.
Viessman, W., et al. (2008), Integrated water management, in *Transboundary Water Resources: A Foundation for Regional Stability in Central Asia, NATO Science for Peace and Security Series C: Environmental Security*, edited by J. E. Moerlins et al., pp. 263–301, Springer, Dordrecht.
Voinov, A., and F. Bousquet (2010), Modelling with stakeholders, *Environ. Modell. Software*, *25*(11), 1268–1281.
Welp, M. (2001), The use of decision support tools in participatory river basin management, *Phys. Chem. Earth, Part B*, *26*(7–8), 535–539.
Werick, W. J., W. J. Whipple, and J. Lund (1996), *ACTACF Basinwide Study*, U.S. Army Corps of Eng., Mobile, Ala.