HYDROLOGICAL PROCESSES
Hydrol. Process. (2009)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/hyp.7528

On the use of multi-algorithm, genetically adaptive multi-objective method for multi-site calibration of the SWAT model

Xuesong Zhang,[1]* Raghavan Srinivasan[2] and Michael Van Liew[3]

[1] *Joint Global Change Research Institute, Pacific Northwest National Laboratory, 5825 University Research Court, Suite 3500, College Park, MD 20740, USA*

[2] *Spatial Sciences Laboratory, Department of Ecosystem Sciences and Management, Texas A&M University, College Station, TX 77843, USA*

[3] *Montana Department of Environmental Quality, Helena, MT 59620, USA*

Abstract:

With the availability of spatially distributed data, distributed hydrologic models are increasingly used for simulation of spatially varied hydrologic processes to understand and manage natural and human activities that affect watershed systems. Multi-objective optimization methods have been applied to calibrate distributed hydrologic models using observed data from multiple sites. As the time consumed by running these complex models is increasing substantially, selecting efficient and effective multi-objective optimization algorithms is becoming a nontrivial issue. In this study, we evaluated a multi-algorithm, genetically adaptive multi-objective method (AMALGAM) for multi-site calibration of a distributed hydrologic model—Soil and Water Assessment Tool (SWAT), and compared its performance with two widely used evolutionary multi-objective optimization (EMO) algorithms (i.e. Strength Pareto Evolutionary Algorithm 2 (SPEA2) and Non-dominated Sorted Genetic Algorithm II (NSGA-II)). In order to provide insights into each method's overall performance, these three methods were tested in four watersheds with various characteristics. The test results indicate that the AMALGAM can consistently provide competitive or superior results compared with the other two methods. The multi-method search framework of AMALGAM, which can flexibly and adaptively utilize multiple optimization algorithms, makes it a promising tool for multi-site calibration of the distributed SWAT. For practical use of AMALGAM, it is suggested to implement this method in multiple trials with relatively small number of model runs rather than run it once with long iterations. In addition, incorporating different multi-objective optimization algorithms and multi-mode search operators into AMALGAM deserves further research. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS distributed hydrologic model; multi-method search; multi-objective optimization; multi-site calibration; soil and water assessment tool

Received 22 April 2009; Accepted 12 October 2009

INTRODUCTION

In recent years, hydrological models have been increasingly used by hydrologists and water resources managers to understand and manage natural and human activities that affect watershed systems. These models can contain parameters that cannot be measured directly because of measurement limitations and scaling issues (Beven, 2000). For practical applications in solving water resources problems, model parameters are calibrated to produce model predictions that are as close as possible to observed data. When calibrating a hydrological model, one or more objectives are often used to measure the agreement between observed and simulated values. The objectives to be optimized can be the combination of multiple goodness-of-fit estimators (e.g. relative error, coefficient of determination), multiple variables (e.g. water, energy, sediment, and nutrients), and multiple sites (e.g. Yapo *et al*., 1998; Gupta *et al*., 1998; van Griensven and Bauwens, 2003; Van Liew and Garbrecht, 2003; Vrugt *et al*., 2003; Cao *et al*., 2006; Kim *et al*., 2006; Reddy and Kumar, 2007; Engeland *et al*., 2006; Bekele and Nicklow, 2007; Rouhani *et al*., 2007). With the recent development of distributed hydrological models, which can spatially simulate hydrological variables, the use of multi-site observed data to evaluate model performance is becoming more common.

The Soil and Water Assessment Tool (SWAT) (Arnold *et al*., 1998) is a continuous, long-term, distributed-parameter hydrological model. SWAT has been applied worldwide for distributed hydrological modelling and water resources management. For example, the SWAT model has been incorporated into the U.S. Environmental Protection Agency (USEPA) Better Assessment Science Integrating Point & Nonpoint Sources (BASINS) software package, and is being applied by the U.S. Department of Agriculture (USDA) for the Conservation Effects Assessment Project (CEAP, 2008; Gassman *et al*., 2007). The SWAT model has been widely applied in a wide spectrum of areas related to water resources management, such as assessing land use and climate change (Zhang *et al*., 2007; Wang *et al*., 2008), groundwater withdrawal

* Correspondence to: Xuesong Zhang, Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD 20740, USA. E-mail: Xuesong.Zhang@pnl.gov

Copyright © 2009 John Wiley & Sons, Ltd.

(Lee and Chung, 2007), irrigation (Gosain *et al*., 2005), and agricultural conservation (Arabi *et al*., 2008).

In many SWAT applications, the model was calibrated using objective functions at a single site (e.g. Zhang *et al*., 2009b). The application of observed data from multiple monitoring sites to calibrate parameter values of SWAT was discussed and proved to be related to the appropriate application of SWAT (Santhi *et al*., 2001; Van Liew and Garbrecht, 2003; White and Chaubey, 2005; Cao *et al*., 2006; Bekele and Nicklow, 2007; Migliaccio and Chaubey, 2007; Zhang *et al*., 2008b). For simultaneous multi-site automatic calibration of SWAT, two types of calibration methods are usually implemented. The first calibration method aggregates the different objective function values calculated at each monitoring site into one integrated value, and then use the single objective optimization algorithms for parameter estimation (e.g. van Griensven and Bauwens, 2003). Recently, evolutionary multi-objective algorithms are being used to optimize SWAT using the different objective functions calculated at multiple sites simultaneously, and finds a set of multiple Pareto-optimal solutions (e.g. Bekele and Nicklow, 2007; Migliaccio and Chaubey, 2007; Zhang *et al*., 2008b). Zhang *et al*. (2008b) compared these two types of methods for SWAT, and showed the advantage of using evolutionary multi-objective algorithms. Currently, multi-objective optimization algorithms, including the Non-dominated Sorted Genetic Algorithm II (NSGA-II) (Deb *et al*., 2002) and Strength Pareto Evolutionary Algorithm 2 (SPEA2) (Zitzler *et al*., 2001), have been applied for calibrating parameters of SWAT (Bekele and Nicklow, 2007; Remegio *et al*., 2007; Zhang *et al*., 2008b).

When calibrating a distributed hydrologic model, the complex structures and large number of parameters make it a challenging problem. Often, a single evolutionary multi-objective optimization (EMO) method trial may take several days or even longer (Tang *et al*., 2006). Although the speed and capacity of computers have increased multi-fold in the past several decades, the time consumed by running hydrological models (especially complex, physically based, distributed hydrological models) is still a concern for hydrology practitioners. The SWAT model is computationally intensive. A single trial of parameter optimization of SWAT (with 10 000 runs) can take several days or weeks (Zhang *et al*., 2009a). Given limited computational resources and time, selecting efficient and reliable EMO algorithms is necessary. To the best of the authors' knowledge, the previous studies involving multi-objective optimization of SWAT examined only the performance of single-algorithm multi-objective methods (e.g. SPEA2 or NSGA-II). No comparative studies have been conducted to compare the efficacy of different multi-objective algorithms or explore the applicability of multi-algorithm search methods that combine the strength of multiple multi-objective optimization algorithms for calibrating SWAT. Recently, Vrugt and Robinson (2007) proposed a multi-algorithm, genetically adaptive multi-objective, method (AMALGAM) which blends the attributes of several available individual optimization algorithms, including NSGA-II, particle swarm optimization (PSO) (Kennedy and Eberhart, 2001), adaptive metropolis search (AMS) (Haario *et al*., 2001), and differential evolution (DE) (Storn and Price, 1997). They evaluated AMALGAM using several standard test functions and showed the promise of this multi-algorithm method to efficiently calibrate complex multi-objective problems.

There are numerous multi-objective optimization methods available. Previous studies have compared the efficacy of different EMO algorithms for parameter estimation of computationally intensive hydrological models and show that different EMO algorithms may exhibit preferable properties for optimizing different hydrological models under different hydrological conditions (e.g. Kollat and Reed, 2005; Tang *et al*., 2006). Therefore, the major purpose of this study is to compare and evaluate the efficacy and reliability of different EMO methods (single algorithm vs multi-algorithm) for multi-site calibration of SWAT. Based on previous comparative studies and the current application status of EMO for calibrating SWAT, we selected two single-algorithm EMO methods (i.e. SPEA2 and NSGA-II) and one multi-algorithm EMO method (AMALGAM). In order to generalize the performance of different methods across various situations, these different methods were applied and compared in four watersheds with different characteristics. The results of this study are expected to provide the users of SWAT and other distributed hydrological model practitioners with valuable information for selecting EMO methods.

MATERIALS AND METHODS

Study area description

The efficacy of optimization algorithms is dependent on the characteristics of the objective function response surface of the hydrological model (Duan *et al*., 1992), which is related to the watershed characteristics. In order to evaluate the general performances of different optimization methods, SWAT was applied to four watersheds with different climatic and hydrological characteristics. The four watersheds included the Yellow River Headwaters Watershed (YRHW), Reynolds Creek Experimental Watershed (RCEW), Little River Experimental Watershed (LREW), and Mahantango Creek Experimental Watershed (MCEW). The locations of the four watersheds are shown in Figure 1. The basic characteristics of the four test watersheds are listed in Table I and described below.

YR headwaters watershed. The YRHW is a mountainous river basin, which is located in the northeastern part of Tibetan plateau in China. This area is the primary source of water availability for the Yellow River Basin (Liu, 2004). The area slopes downwards from west to east, ranging from a combined landform of low mountains and wide valleys with lakes to smooth plateaus

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

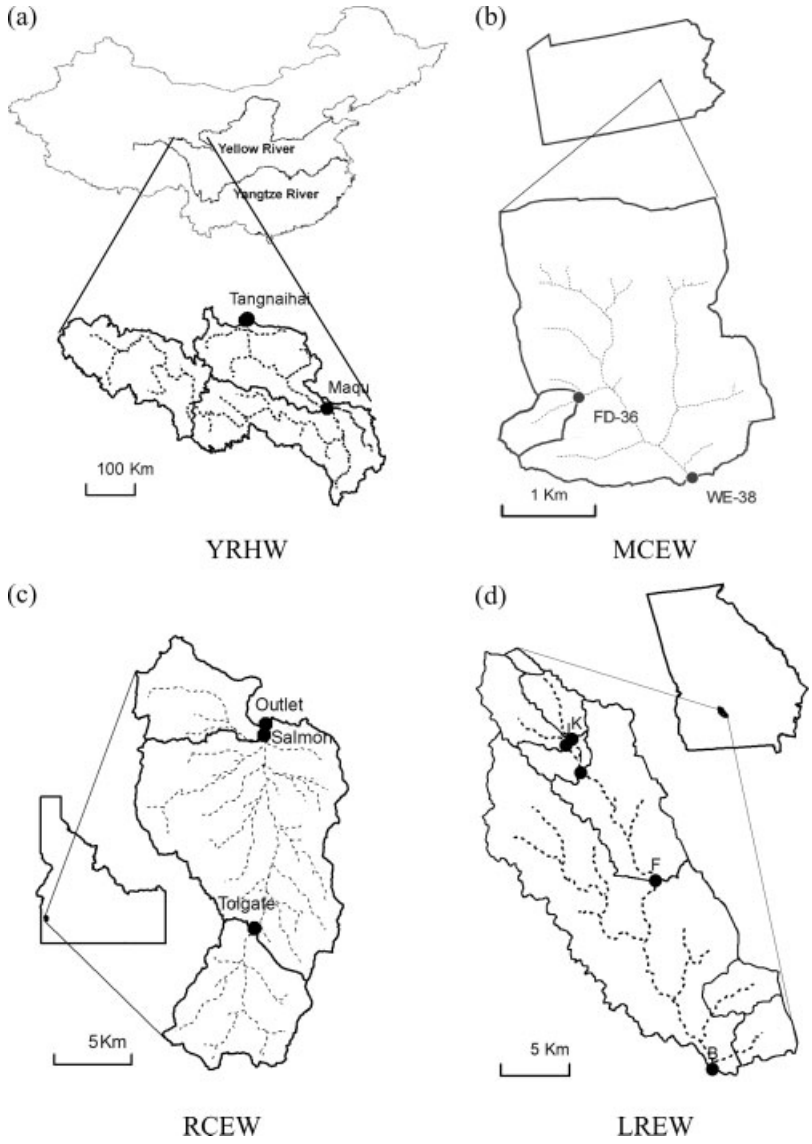

Figure 1. The locations of the four test watersheds and the monitoring stations.

Table I. Major characteristics of the four case study watersheds

Characteristics Watershed	Size (km^2)	Elevation (m)	Climate	Annual precipitation (mm)	Annual temperature (°C)	Major land use types
YRHW	114 345	4217 (2600–6266)	Dry continental alpine	600	−0·7	Grassland
MCEW	7	286 (215–496)	Humid temperate	1100	9·9	Pasture, forest, and cropland
RCEW	239	1526 (1101–2241)	Arid steppe	482	7·3	Rangeland and forest
LREW	334	361 (270–480)	Humid subtropical	1167	14·5	Woodland, pasture and cropland

(Wang *et al*., 2003). The headwaters area has a typical continental alpine cold and dry climate. The annual precipitation amount is around 600 mm and the average annual temperature for the YRHW is near 0 °C. In winter, the average temperature is below 0 °C for most of the weather stations, while in summer the average temperature is above 0 °C. This seasonal temperature variation makes snowmelt a significant process in this area (Zhang *et al*., 2008a). This watershed is characterized by gently sloping upland, river bed, and swamp and wetland. The major types of soils in this area are clay and loam with relatively low infiltration rates. The major land

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

cover in the study area is grassland, which accounts for approximately 90% of the total area. Other land use/land cover (forest land, rangeland, agriculture land, and bare area) account for the remaining 10% of the area. The observed daily streamflow data were collected at two monitoring sites (i.e. Maqu and Tangnaihai in Figure 1a) for multi-site calibration of SWAT.

Mahantango Creek experimental watershed. The MCEW is a tributary of the Susquehanna River in Central Pennsylvania. The MCEW is typical of upland agricultural watersheds within the non-glaciated, folded, and faulted Appalachian Valley and Ridge Physiographic Province (Veith *et al*., 2005). Climate in the region is temperate and humid, with a long-term average annual precipitation of 1100 mm. The watershed is characterized by shallow, fragipan soils in near-stream areas and deep, well-drained soils in the uplands (Van Liew *et al*., 2007). Land use types consist of pasture (38%), forest (34%), mixed croplands (26%), and farmsteads (2%). WE38 and FD-36 are two monitoring sites with daily streamflow data within the MCEW (Figure 1b).

Reynolds Creek experimental watershed. The RCEW is located near the north end of the Owyhee Mountains of southwest Idaho. The topography of the watershed ranges from a broad, flat alluvial valley to steep, rugged mountain slopes, with a range in elevation from 1101 to 2241 m (Seyfried *et al*., 2000). Because of orographic effects, the average annual precipitation ranges from about 250 mm at the outlet to more than 1100 mm at the upper end of the watershed. Perennial streamflow is generated at the highest elevations in the southern part of Reynolds Creek where deep, late-lying snowpacks are the source for most water (Seyfried *et al*., 2000). Land cover on Reynolds Creek consists of rangeland and forest communities of sagebrush, greasewood, aspen, and conifers (94%) and irrigated cropland (6%). Within the RCEW, there are three monitoring sites (i.e. Salman, Tolgate, and Outlet in Figure 1c) available for multi-site observed streamflow data.

Little River experimental watershed. The LREW in southwest Georgia is the upper 334 km^2 of the Little River and is the subject of long-term hydrological and water quality research by USDA-ARS and co-operators (Sheridan, 1997). The region has low topographic relief and is characterized by broad, flat alluvial floodplains, river terraces, and gently sloping uplands (Sheridan, 1997). Climate in this region is characterized as humid subtropical. Precipitation occurs almost exclusively as rainfall, with an annual mean of 1167 mm. Soils on the watershed are predominantly sands and sandy loams with high infiltration rates. Since surface soils are underlain by shallow, relatively impermeable subsurface horizons, deep seepage and recharge to regional ground water systems are impeded (Sheridan, 1997). Land use within the watershed is approximately 50% woodland, 31% row crops (primarily peanuts and cotton), 10% pasture, and 2% water. In the LREW, streamflow data was monitored at five observation stations (i.e. B, F, I, J, and K in Figure 1d).

SWAT model description

SWAT subdivides a watershed into sub-basins connected by a stream network, and further delineates hydrological response units (HRUs) consisting of unique combinations of land cover and soils in each subbasin. It is assumed that there is no interaction between HRUs, i.e. the HRUs are non-spatially distributed. HRU delineation can minimize computational costs of simulations by lumping similar soil and land use areas into a single unit (Neitsch *et al*., 2005a). SWAT allows a number of different physical processes to be simulated in a watershed. The hydrological routines within SWAT account for snow fall and melt, vadose zone processes (i.e. infiltration, evaporation, plant uptake, lateral flows, and percolation), and ground water flows. The hydrologic cycle as simulated by SWAT is based on the water balance equation:

$$\mathrm{SW}_t = \mathrm{SW}_0 + \sum_{i=1}^{t} (R_{\mathrm{day}} - Q_{\mathrm{surf}} - E_a - w_{\mathrm{seep}} - Q_{\mathrm{gw}}) \quad (1)$$

where SW_t is the final soil water content (mm H_2O), SW_0 is the initial soil water content on day i (mm H_2O), t is the time (days), R_{day} is the amount of precipitation on day i (mm H_2O), Q_{surf} is the amount of surface runoff on day i (mm H_2O), E_{a} is the amount of evapotranspiration on day i (mm H_2O), w_{seep} is the amount of water entering the vadose zone from the soil profile on day i (mm H_2O), and Q_{gw} is the amount of return flow on day i (mm H_2O). Surface runoff volume is estimated using a modified version of the Soil Conservation Service (SCS) Curve Number (CN) method (Kannan *et al*., 2008). A kinematic storage model is used to predict lateral flow, whereas return flow is simulated by creating a shallow aquifer (Arnold *et al*., 1998). The Muskingum method is used for channel flood routing. Outflow from a channel is adjusted for transmission losses, evaporation, diversions, and return flow. van Griensven *et al*. (2006) conducted detailed global sensitivity analysis of the parameters in SWAT, which showed that 10 parameters are sensitive to the hydrological simulation of SWAT. Van Liew *et al*. (2007) tested the suitability of SWAT for the CEAP in five USDA Agricultural Research Service watersheds. In the study conducted by Van Liew *et al*. (2007), 16 parameters, which include the 10 parameters identified by van Griensven *et al*. (2006), were adjusted to calibrate the SWAT model for hydrological simulation. The same 16 parameters identified by Van Liew *et al*. (2007) were applied in this study. The general description of the 16 parameters is shown in Table II. The parameters' ranges were determined according to van Griensven *et al*. (2006) and Neitsch *et al*. (2005b).

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

Multi-objective optimization algorithms

In order to describe multi-objective optimization problems, a set of symbols are defined in Appendix. For multi-objective optimization problems, a series of objective functions need to be taken into account simultaneously. The general multi-objective optimization problem can be defined as: find the parameter solution $\boldsymbol{x}^*$ that will optimize the objective function vector $\boldsymbol{F}(\boldsymbol{x}') = [f_1(x), f_2(x), \ldots, f_m(x)]$. As there are multiple objective functions that need to be optimized simultaneously, and different objective functions prefer different parameter solutions, it is difficult to find a single global optimum parameter solution. The Pareto-optimality concept is defined to evaluate whether a parameter set is 'optimal' or not. An objective function vector $\boldsymbol{F}(\boldsymbol{x}') = [f_1(x'), f_2(x'), \ldots, f_m(x')]$ is said to dominate another objective function vector $\boldsymbol{F}(\boldsymbol{x}) = [f_1(x), f_2(x), \ldots, f_m(x)]$ (denoted by $\boldsymbol{F}(\boldsymbol{x}') \succ F(x)]$), if $\forall i \in \{1, 2, \ldots m\}, f_i(x') \geq f_i(x) \wedge \exists i \in \{1, 2, \ldots m\}, f_i(x') > f_i(x)$ (Zitzler and Thiele, 1999). If the objective function vector $\boldsymbol{F}(\boldsymbol{x}^*)$ of a point $x^* \in \Omega$ is not dominated by all the other objective function vectors of the parameter solutions in the feasible parameter space, then $\boldsymbol{x}^*$ is taken as a Pareto-optimal parameter solution. The Pareto-optimal set (P^*) is defined by the set of parameter solutions that is not dominated by other parameter solutions. The objective function vectors corresponding to the Pareto-optimal set comprise the Pareto front (PF^*). The purpose of multi-objective optimization is to search the feasible parameter space and find those parameter solutions which are Pareto-optimal. The three EMO methods applied in this study involve several common procedures: (1) Population size design (i.e. the number of solutions to be evolved); (2) Fitness assignment, which calculates the fitness value of each solution in the population; (3) Environmental selection, which evolves the fitter solutions into next parent population; (4) Offspring reproduction, which creates new promising solutions using different evolutionary algorithms. In the following sections, the procedures of three EMO methods are introduced briefly.

SPEA2. SPEA2, developed by Zitzler *et al.* (2001), is one EMO method that has been successfully applied in hydrological model optimization (e.g. Tang *et al.*, 2006; Zhang *et al.*, 2008b). SPEA2 requires users to initialize a population of parameter solutions (P_t) and an empty external archive ($\overline{P}_t$), which are evolved using fitness assignment, environmental selection, and offspring production operations. In the fitness assignment operator, three basic procedures are implemented. First, each individual in P_t and $\overline{P}_t$ is assigned to a strength value $S(i)$ representing the number of solutions that are dominated by x_i. Second, raw fitness $R(i)$ is calculated as the sum of the strength value of solutions who dominate x_i. Finally, in case many parameter solutions have the same raw fitness when most chromosomes do not dominate each other, the final fitness $F(i)$ is computed through adjusting the raw fitness using a kth nearest neighbour method, which defines the density of a parameter solution as a function of its distance to the kth nearest neighbour in the objective space (Zitzler *et al.*, 2001). The environmental selection operator is used to copy all Pareto-optimal chromosomes in P_t and $\overline{P}_t$ to P_{t+1}. If the size of P_{t+1} exceeds N, then P_{t+1} is reduced by means of truncating the non-dominated chromosomes with less fitness; otherwise, if the size of P_{t+1} is less than N, then P_{t+1} is filled with the best dominated chromosomes in P_t and $\overline{P}_t$. In the offspring reproduction operator, the genetic algorithms (GAs) (Goldberg, 1989) are used to evolve the parameter solutions in P_{t+1} and reproduce promising new candidates for the next generation $\overline{P}_{t+1}$. The fitness assignment, environmental selection, and offspring production are repeated until maximum number of model runs is reached. In this study, the binary tournament selection, simulated binary crossover, and polynomial mutation were used within the GA framework.

NSGA-II. NSGA-II is an elitist multi-objective GA developed by Deb *et al.* (2002). NSGA-II requires users to initialize a population of parameter solutions and an external archive, which are evolved using several operators, including fast non-domination sorting (FNS), crowded distance calculation, elitist selection, and offspring reproduction operations t. The FNS is an efficient operator that assigns ranks to the parameter solutions in P_t and $\overline{P}_t$. The individuals that are not dominated by other individuals are put in the first front FT_1, and are assigned rank 1. The individuals that are not dominated by other individuals except those in FT_1 are put in the second front FT_2, and assigned rank 2. Similarly, all individuals are assigned to a specific front and rank number. After the FNS operation, many individuals are usually located in the same front and have the same rank. NSGA-II uses crowding distance to discriminate the individuals with the same front order. The crowding distance is calculated as the average distance of the two individuals on either side of individual i along each of the objectives as an estimate of the size of the largest cuboid enclosing the point i without including any other point in the population. The individuals with higher crowding distances help preserve the diversity of the population and ensure that the NSGA-II will find solutions along the full extent of the Pareto front. The chromosomes with lower rank and larger crowding distance are selected into P_{t+1}, which will be evolved using GA to populate $\overline{P}_{t+1}$. The settings of GA operators in NSGA-II are similar to those used in SPEA2.

AMALGAM. AMALGAM adaptively and simultaneously employs multiple EMO algorithms to ensure a fast, reliable, and computationally efficient solution to multi-objective optimization problems (Vrugt and Robinson, 2007). AMALGAM starts with a random initial population P_t of size of N. For each individual in P_t, the

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

FNS operator is used to assign a rank. A offspring population $\overline{P}_t$ of size N is subsequently created by implementing each candidate algorithms within AMALGAM to generate a pre-specified number of offspring points, $N = \{N_t^1, N_t^2, \ldots, N_t^k\}$, from P_t. By using the FNS operator, the best N solutions within $R_t = P_t \cup \overline{P}_t$ are selected into P_{t+1}, which is evolved repeatedly by the multi-method search and adaptive offspring creation method until a maximum number of model runs. The two key procedures of AMALGAM are simultaneous multi-method search and self-adaptive offspring creation. In this study, four candidate EMO algorithms, including NSGA-II, PSO, DE, and AMS, were incorporated into AMALGAM following Vrugt and Robinson (2007). As to the self-adaptive offspring creation of AMALGAM, the number of offspring points generated by each candidate algorithms, $\{N_t^1, N_t^2, \ldots, N_t^k\}$, is updated according to $N_t^i = N \times (O_t^i/N_{t-1}^i)/\sum_{n=1}^{k}(O_t^n/N_{t-1}^n)$, where O_t^i/N_{t-1}^i is the ratio of the number of offspring points an algorithm contributes to the new population, O_t^i, and the corresponding number that the algorithm created in the previous generation (N_{t-1}^l) (Vrugt and Robinson, 2007). For the first generation, $N_0^1 = N_0^2 = \ldots = N_0^k = N/k$. And the minimum value of N_t^k is 5.

Optimization test cases design

Optimization objective functions. The Nash–Sutcliffe efficiency (NSE) values for streamflow simulation at the monitoring stations within each test watershed were the objectives to be optimized simultaneously. The formula to calculate NSE is (Nash and Sutcliffe, 1970):

$$\mathrm{NSE} = 1{\cdot}0 - \frac{\sum_{i=1}^{q}(\mathrm{obs}_i - \mathrm{sim}_i)^2}{\sum_{i=1}^{q}(\mathrm{obs}_i - \overline{\mathrm{obs}})^2} \tag{2}$$

where sim_i is the model-simulated value at time step i, obs_i is the observed data at time step i, $\overline{\mathrm{obs}}$ the mean for the entire time period of the evaluation, and q is the total number of pairs of simulated and observed data. NSE indicates how well the plot of the observed versus the simulated values fits the 1 : 1 line, and ranges from $-\infty$ to 1. For the four test watersheds, the objective function vectors that need to be optimized are described as follows.

$$\begin{aligned} F_{\mathrm{YRHW}} &= \{f_1 = \mathrm{NSE}_{\mathrm{Maqu}'} f_2 \\ &= \mathrm{NSE}_{\mathrm{Tangnaihai}}\} \end{aligned} \tag{3}$$

$$\begin{aligned} F_{\mathrm{MCEW}} &= \{f_1 = \mathrm{NSE}_{FD36'} f_2 \\ &= \mathrm{NSE}_{WE38}\} \end{aligned} \tag{4}$$

$$\begin{aligned} F_{\mathrm{RCEW}} &= \{f_1 = \mathrm{NSE}_{\mathrm{Salmom}'} f_2 \\ &= \mathrm{NSE}_{\mathrm{Togate}'} f_2 = \mathrm{NSE}_{\mathrm{Outlet}}\} \end{aligned} \tag{5}$$

$$\begin{aligned} F_{\mathrm{LREW}} &= \{f_1 = \mathrm{NSE}_{B'} f_2 \\ &= \mathrm{NSE}_{F'} f_3 = \mathrm{NSE}_{I'} f_4 \\ &= \mathrm{NSE}_{J'} f_5 = \mathrm{NSE}_{K}\} \end{aligned} \tag{6}$$

Performance evaluation of different EMO algorithms. Generally, it is very difficult to find an analytical expression of the line or surface that contains all the Pareto-optimal parameter solutions. The normal procedure to generate the Pareto front is to compute the feasible solutions x and their corresponding $F(x)$. When there are a sufficient number of feasible points, then it is assumed that the non-dominated points are approximating the Pareto front (Coello Coello *et al.*, 2004), which is referred to as reference set. In this study, the reference set for each test case was obtained by collecting the non-dominated points obtained by all the methods with multiple trials. Compared with this reference set, three performance metrics were calculated to evaluate the quality of the non-dominated parameter solutions (referred to as approximation set) obtained by each method. The three metrics applied in this study are introduced as follows.

Success Identification (SI): The SI metric is calculated as:

$$\mathrm{SI} = \frac{\sum_{i=1}^{n} S_i}{n} \tag{7}$$

where n is the number of parameter solutions in the reference set. $S_i = 1$ if the parameter solution i in the reference set is identified by the approximation set, and $S_i = 0$ otherwise. A value of $\sum_{i=1}^{n} S_i = n$ indicates that all the members in the reference are found. In this study, if the difference between a member in the reference set and any member in the approximation set is less than 0·001, it is assumed that a member in the reference set is successfully counted. This metric is very similar to the Success Counting metric used by Sierra and Coello (2005).

ε-indicator: this metric was proposed by Zitzler *et al.* (2003) to measure how well the algorithms converge to the reference set. The ε-indicator is calculated as the smallest distance that an approximation set needs to be transformed in order to dominate the reference set. The individuals in the transformed approximation set are called reference points. Figure 2a illustrates the computation of the ε-indicator for a two-objective case.

Hypervolume (HP): this metric was proposed by Zitzler and Thiele (1999). The HP metric is represented by the difference between the volume of the objective space dominated by the reference set and the approximation set which measures how well the approximation set performs in identifying solutions along the full extent of the Pareto surface (Tang *et al.*, 2006). In Figure 2b, a two-objective case was used to illustrate the calculation of HP.

When applying the above three metrics to evaluate the performance of different algorithms, larger SI values and smaller ε-indicator and HP values are preferred.

Test cases design. The control parameters of each EMO method were selected on the basis of sets in previous studies (Kollat and Reed, 2005; Tang *et al.*, 2006;

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

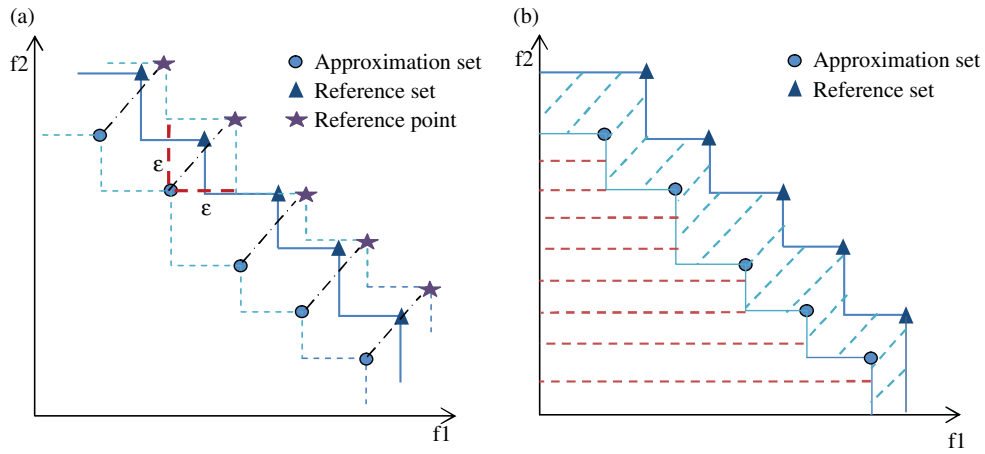

Figure 2. (a) Illustration of the calculation of ε-indicator metric using a two-objective example. (b) Illustration of the calculation of the hypervolume metric using a two-objective example. The shaded area with dashed slash line represents the hypervolume value. Adapted from Tang *et al.* (2006)

Table II. Parameters for calibration in the SWAT model

Code	Parameter	Description	Range
1	CN2	Curve number	±20%
2	ESCO	Soil evaporation compensation factor	0–1
3	SOL_AWC	Available soil water capacity	±20%
4	GW_REVAP	Ground water re-evaporation coefficient	0·02–0·2
5	REVAPMN	Threshold depth of water in the shallow aquifer for re-evaporation to occur (mm)	0–500
6	GWQMN	Threshold depth of water in the shallow aquifer required for return flow to occur (mm)	0–5000
7	GW_DELAY	Groundwater delay (days)	0–50
8	ALPHA_BF	Base flow recession constant	0–1
9	RCHRG_DP	Deep aquifer percolation fraction	0–1
10	CH_K2	Effective hydraulic conductivity in main channel alluvium (mm/h)	0·01–150
11	TIMP	Snow pack temperature lag factor	0–1
12	SURLAG	Surface runoff lag coefficient (day)	0–10
13	SFTMP	Snow melt base temperature (°C)	0–5
14	SMTMP	Snowfall temperature (°C)	0–5
15	SMFMX	Maximum snowmelt factor for June 21 (mm H_2O/°C^{-1} day^{-1})	0–10
16	SMFMN	Minimum snowmelt factor for December 21 (mm H_2O/°C^{-1} day^{-1})	0–10

Table III. Major parameters settings of the three EMO methods

Method Settings	SPEA2	NSGA-II	AMALGAM[a]
Population	100 and 250	100 and 250	100
Termination criterion	10 000 runs	10 000 runs	10 000 runs
Probability of crossover	1·0	1·0	1·0
Crossover distribution Index	15	15	15
Probability of mutation	$1/D$	$1/D$	$1/D$
Mutation distribution index	20	20	20
Variable representation	Real	Real	Real

[a] The settings of other algorithms (i.e. PSO, AMS, and DE) within AMALGAM follow Vrugt and Robinson (2007).

Vrugt and Robinson, 2007). Table III shows the major parameter settings of different methods, among which the population size is an important factor that determines the performance of different algorithms. In this study, the effect of population size on the performance of NSGA-II and SPEA2 was further examined with one relatively small population size (100) and one relatively large population size (250) following Tang *et al.* (2006). For AMALGAM, the population size effect was not further explored. Hence, there were a total of five optimization cases for each test watershed.

The EMO methods applied in this study involve random sampling of the parameter values, so the results obtained by one trial are stochastic and cannot be used to accurately evaluate the algorithm's performance. The average behaviour of multiple trials of each method was used to compare the performance of different methods. This is a popular performance comparison method

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

reported in the literature (Ali *et al*., 2005). Ideally, it is expected that one method can consistently outperform the other methods in terms of all three metrics introduced above. The three EMO methods were run 10 times to obtain SI, ε-indicator, and hypervolume values as indicator of their performance. The multi-objective optimization problems require finding multiple solutions that approximate the Pareto front. With a single trial of an EMO method, it is usually difficult to find a good approximation of the Pareto front. Therefore, we tried to evaluate the performance of each algorithm from two aspects: (1) the performance of the non-dominated set obtained through combining the results of multiple trials of one method, which is referred to as combined performance; and (2) the average performance of the non-dominated sets found by multiple trials of one method, which is represented as average performance.

In this study, the SWAT model was set up for daily flow simulation at the monitoring stations within each watershed. The calibration periods consisted of 10 years (1976–1985) in the YRHW, 4 years (1995–1998) in the MCEW, 3 years (1995–1997) in the LREW, and 4 years (1966–1969) in the RCEW. On a computer with Pentium IV 3 GHZ and 1 GB RAM, the time consumed by one SWAT model run was 30 s for the YRHW, 13 s for the MCEW, 44 s for LREW, and 42 s for RCEW. As time and computer resources are limited, it was not possible to run the SWAT model for a very long simulation period or for an unlimited number of model evaluations. The three algorithms were compared on the basis of the average performance of 10 trials within a limited and affordable number of model evaluations. Considering the time and computer resources available, the maximum number of model evaluations was limited to 10 000 model runs for the four test watersheds. The time consumed by one trial was 84 h in the YRHW, 37 h in the MCEW, 122 h in LREW, and 117 h in RCEW.

RESULTS AND DISCUSSION

Evaluation of different algorithms for the two-objective case in the YRHW

The best known reference set (Figure 3) for the two-objective test case in the YRHW was collected by running the three multi-objective optimization algorithms in multiple trials. This reference set was used to evaluate the performance of different multi-objective optimization methods. It was found that only SPEA2 and AMALGAM contributed to the reference set. SPEA2-100, SPEA2-250, and AMALGAM contributed 3, 1, and 6, respectively, of a total of 10 members in the reference set.

The combined performance of each method was evaluated by collecting the combined approximation sets for each method (Figure 4). Table IV lists the evaluation coefficients for each combined approximation set. AMALGAM consistently performed the best among the three methods. NSGA-II-250 performed the least.

The average performance of each algorithm was evaluated using the average evaluation coefficients obtained

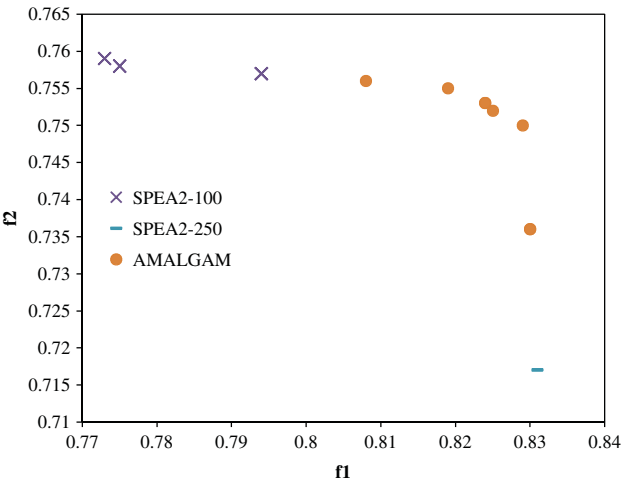

Figure 3. The reference set for the two-objective case in the YRHW

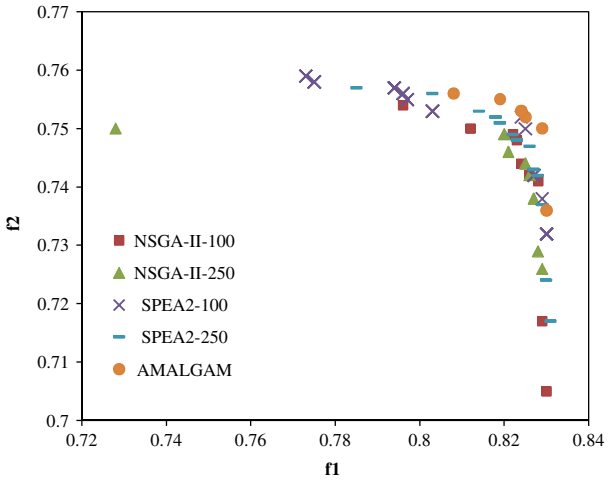

Figure 4. The approximation sets found by different algorithms through multiple trials in the YRHW

across 10 trials (Table IV). It was found that AMALGAM successfully identified more members in the reference set than the other methods, while SPEA2-100 performed best in terms of ε-indicator and HP values. It is worth noting that the average evaluation coefficient values are much less than the combined performance for each algorithm. For example, the HP value obtained by combining multiple trials of AMALGAM is 0·000027, while the average HP value (0·001) was much larger. The SI value obtained by combining multiple trials of AMALGAM is 60%, which is about 10 times its average SI value of 6·62%. The substantial difference between the combined performance and average performance of each method emphasizes the inadequacy of a single trial of EMO methods tested in this case. In order to obtain good approximation for the reference set, multiple trials are required.

Evaluation of different algorithms for the two-objective case in the MCEW

For the two-objective test case in the MCEW, Figure 5 shows the reference set obtained by running different algorithms with multiple trials. Only SPEA2 and

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

Table IV. Average and combined performance of different algorithms for the two-objective test case in theYRHW

Performance metrics / Method	Average performance			Combined performance		
	ε-Indicator	SI	HP	ε-Indicator	SI	HP
NSGA-II-100	0·015 (0·007)	0·00% (0·00%)	0·0009 (0·0005)	0·006	0·00%	0·0003
NSGA-II-250	0·015 (0·005)	0·00% (0·00%)	0·0009 (0·0003)	0·01	0·00%	0·0005
SPEA2-100	0·011 (0·003)	4·09% (12·28%)	0·0006 (0·0003)	0·004	30%	0·0001
SPEA2-250	0·013 (0·005)	1·05% (3·33%)	0·0007 (0·0003)	0·004	10%	0·0001
AMALGAM	0·017 (0·009)	6·62% (12·15%)	0·001 (0·0006)	0·003	60%	2.7E-05

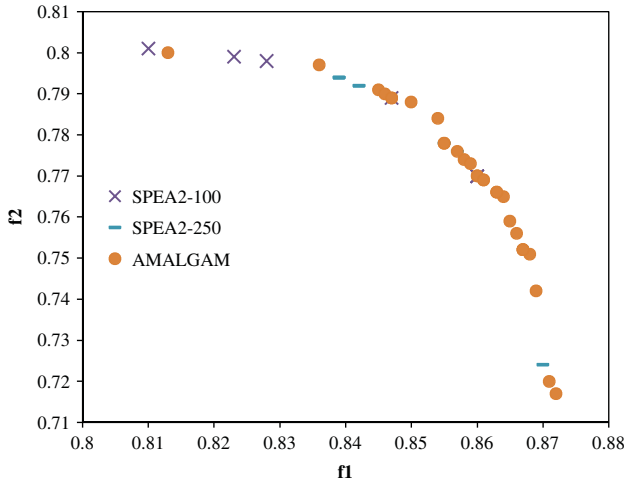

Figure 5. The reference set for the two-objective case in the MCEW

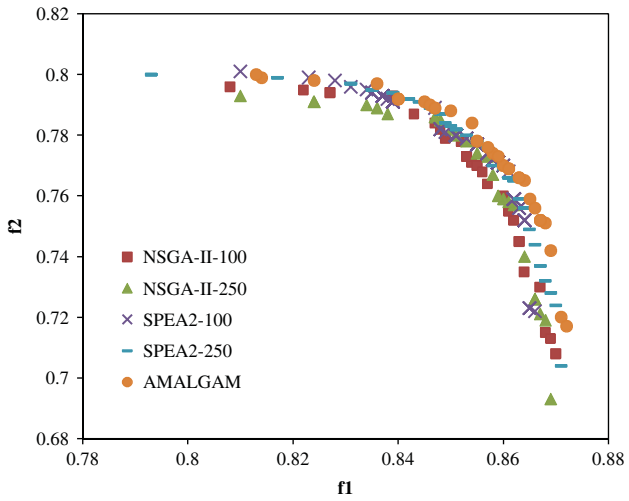

Figure 6. The approximation sets found by different algorithms through multiple trials in the MCEW

AMALGAM contributed to the reference set. SPEA2-100, SPEA2-250, and AMALGAM contributed 12%, 24%, and 64% of the reference set, respectively.

The combined approximation sets found by each of the three methods with multiple trials are shown in Figure 6. Visually, the approximation set found by AMALGAM dominates those found by other methods. The evaluation coefficients listed in Table V confirm the superior performance of AMALGAM, which exhibited smaller ε-indicator and HP values and larger SI values. Further analysis of the evaluation coefficients of average performance of each method shows that AMALGAM still outperforms the other two single-algorithm methods (Table V).

Evaluation of different algorithms for the three-objective case in RCEW

The RCEW test case has three monitoring sites to calibrate SWAT. Three two-dimensional graphics are used to represent the reference set (Figure 7). Among the total of 321 members in the reference set, NSGA-II-100, NSGA-II-250, SPEA2-100, SPEA2-250, and AMALGAM contributed 17, 18, 32, 11, and 243, respectively. The combined approximation set of each method is shown in Figure 8. The evaluation coefficients of the combined performances of each method show that AMALGAM obtained the largest SI value while the lowest ε-indicator and HP values. Similarly, the evaluation coefficients of average performance of the different methods (Table VI) also show the superior performance of AMALGAM over the other methods.

Evaluation of different algorithms for the five-objective case in LREW

In the LREW, there are five objectives that need to be optimized simultaneously. It is difficult to plot the results in five-dimensional space. Therefore, the figures illustrating the reference set and approximation sets found by different algorithms were not plotted. As it is very time consuming to calculate the HP metric for this five-objective case and HP metric is highly correlated to the ε-indicator, only the ε-indicator and SI metrics were computed to evaluate the performance of different algorithms in the LREW. A total of 795 parameter solutions were identified in the reference set. NSGA-II-100, NSGA-II-250, SPEA2-100, SPEA2-250, and AMALGAM found 124, 13, 2, 98, 136, and 422 non-dominated parameter solutions, respectively. Further analysis shows that AMALGAM exhibited better ε-indicator and SI values

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

Table V. Average and combined performance of different algorithms for the two-objective test case in the MCEW

Performance metrics	Average performance			Combined performance		
Method	ε-indicator	SI	HP	ε-indicator	SI	HP
NSGA-II-100	0·0159 (0·0063)	0·00% (0·00%)	0·0012 (0·00057)	0·006	0·00%	0·000513
NSGA-II-250	0·0140 (0·0037)	0·00% (0·00%)	0·0011 (0·00034)	0·008	0·00%	0·000583
SPEA2-100	0·0170 (0·0093)	2·00% (3·27%)	0·0013 (0·00082)	0·006	20·00%	0·000302
SPEA2-250	0·0112 0·005007	3·00% (8·81%)	0·00083 (0·00044)	0·003	30·00%	0·000167
AMALGAM	**0·0072** (0·0015)	**8·60%** (20·00%)	**0·00057** (0·00028)	**0·002**	**80·00%**	**2·6E-05**

The values in bold denote the best performance metric.

than the other methods for both combined performance and average performance assessment (Table VII).

Discussion

The results discussed above, to some extent, agree with the popular 'no free lunch (NFL) theorem' that 'for any algorithm, any elevated performance over one class of problems is offset by performance over another class' (Wolpert and Macready, 1997). For the two single-algorithm methods (i.e. SPEA2 and NSGA-II), neither of them can consistently outperform the other. For example, NSGA-II-250 outperformed SPEA2-100 for the MCEW, while this is reversed in LREW. The advantage of the multi-method search approach (Vrugt and Robinson, 2007; Vrugt *et al*., 2009) employed by AMALGAM is clearly demonstrated. AMALGAM obtained the best average and combined performance in MCEW, RCEW, and LREW. For the YRHW, AMALGAM performed less than SPEA2 in terms of ε-indicator and HP for average performance assessment, but it still performed best with respective to combined performance. The optimum single-objective values are valuable to modellers when they have preference to one specific objective. AMALGAM found most of the optimal single-objective values in the reference sets in the four test watersheds. For example, AMALGAM obtained best values for four out of the five objectives in LREW, two out of the three objectives in RCEW, and one out of the two objectives in MCEW. In YRHW, SPEA2 found the extreme ends of the two objectives. The watershed characteristics of YRHW (Table I) are quite different from those of the other three test watersheds, which means that the YRHW is a special case of applications of SWAT. Overall, AMALGAM proved to be a good choice for multi-objective calibration of SWAT. But, it is worth noting that none of the multi-objective methods can consistently find the extreme end of each single objective. If the modellers emphasize one of the multiple objectives, single-objective optimization methods may provide useful information on the best value of each single objective (Zhang *et al*., 2008b).

AMALGAM provides a flexible framework for simultaneously implementing different EMO algorithms to solve multi-objective optimization problems. The essence of AMALGAM is to combine the strength of different EMO algorithms and dynamically change the contributions of these EMO algorithms to solve the problem on the basis of their performance history. In order to provide insights into how different algorithms alternate in their importance during the optimization process, the numbers of individuals produced by different algorithms are plotted against the iterations for one test trial of AMALGAM in MCEW in Figure 9. The relative contribution by each algorithm within AMALGAM is calculated by dividing the number of individuals it produced (averaged over the 10 trials) by the total number of model runs in one trial. The percentage of individuals contributed by each algorithm within AMALGAM is shown in Table VIII. The relative contribution of each algorithm varied in different watersheds, which indicates the adaptability of AMALGAM to change preference to individual search algorithms for different problems. Overall, NSGA-II contributed about half of the SWAT model runs during the optimization process. The contributions of PSO, DE, and AMS ranked the second, third, and fourth, respectively. Currently, the GA-based EMO algorithm incorporated in AMALGAM is NSGA-II. The test results obtained in this study (Tables IV–VII) indicate the advantage of SPEA2 over NSGA-II for calibrating SWAT. In general, SPEA2 obtained better performance metrics for most cases and find more non-dominated parameter solutions than NSGA-II. Therefore, incorporating SPEA2 into the AMALGAM framework holds the promise to be further exploited.

We also need to pay attention to the difference between the average evaluation coefficient values and those obtained by combining multiple trial results. For all three EMO methods, the average SI values are about 1/10 of those obtained through multiple trials in all four test watersheds. The different ε-indicator and HP values further confirm this difference. Taking AMALGAM as an example, in the four test watersheds, the ε-indicator values obtained through multiple trials are less than one-third of those average ε-indicator values, and HP values are

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

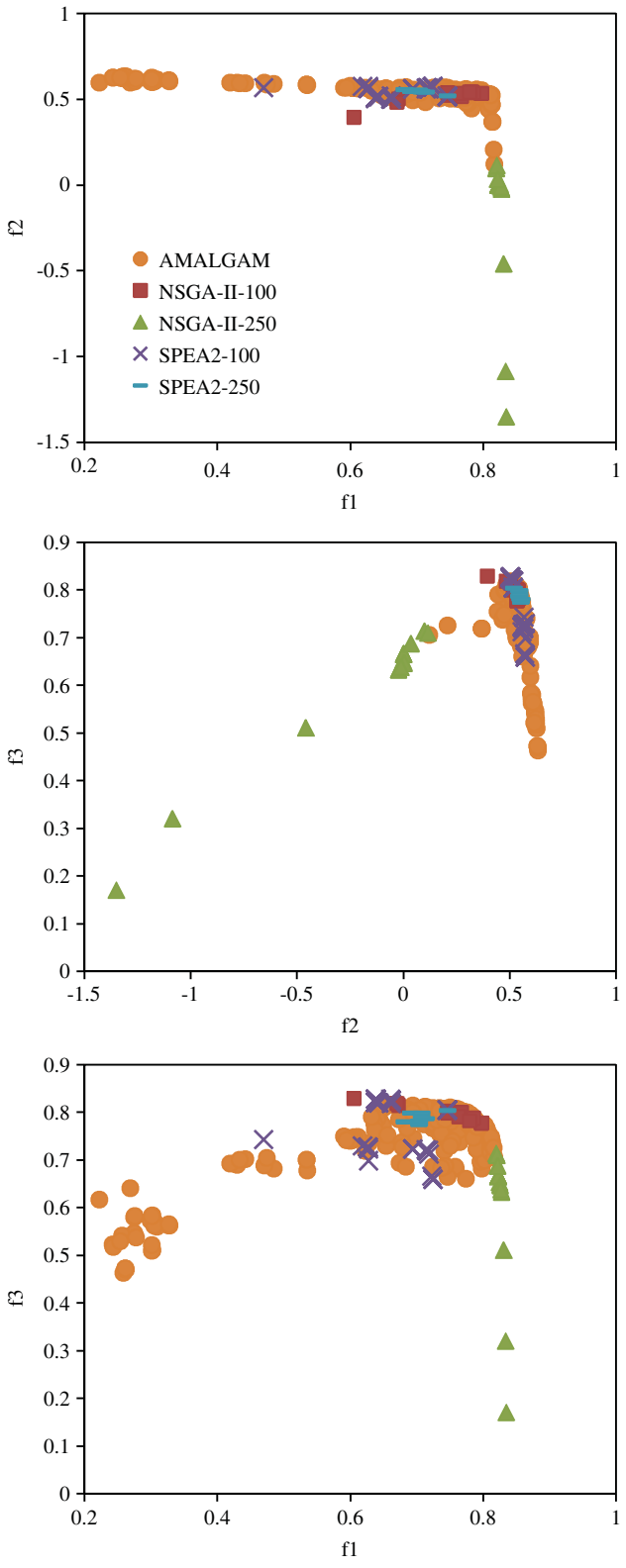

Figure 7. The reference set for the three-objective case in the RCEW

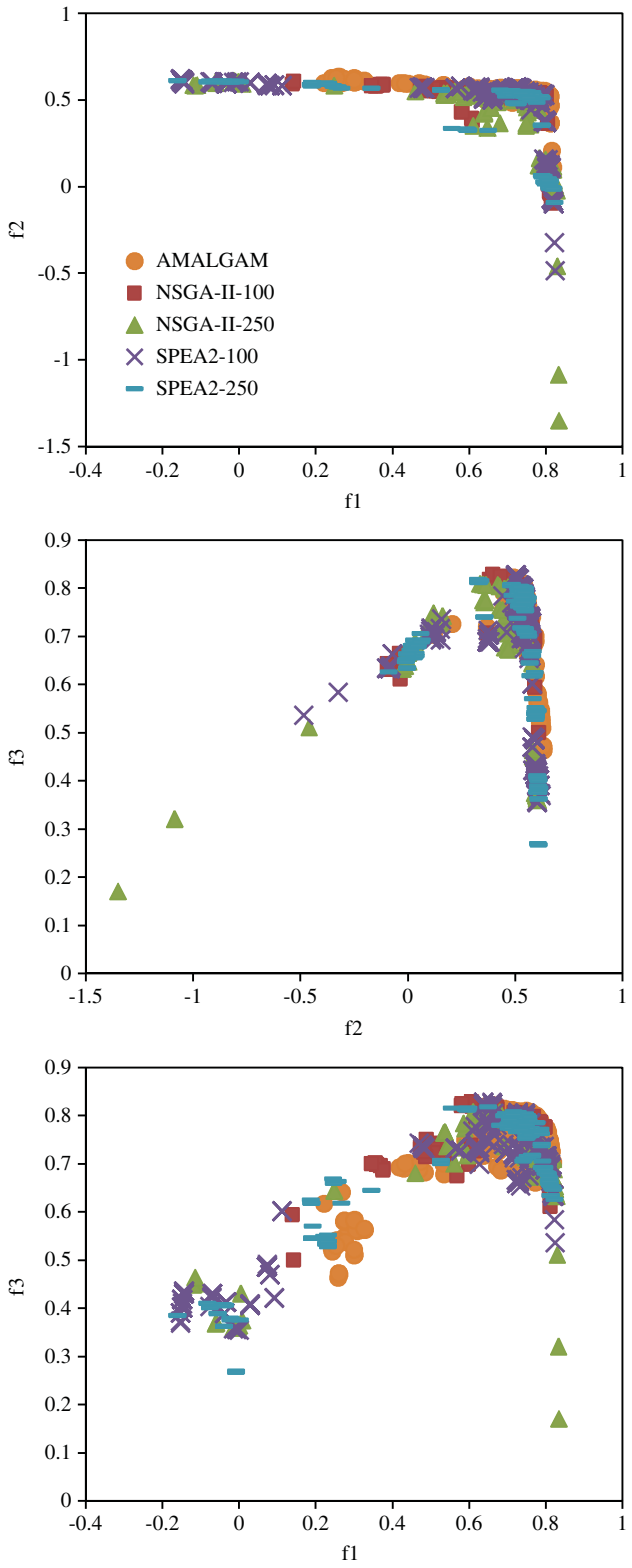

Figure 8. The approximation sets found by different algorithms through multiple trials in the RCEW

less than one-sixth. These results indicate the risk of running one EMO method only once to estimate the trade-off between the multiple objective functions. Multiple trails and more model runs may be required to obtain close approximation of the Pareto front. For practical implementation of AMALGAM, the modellers need to choose between multiple trials with a small number of runs and a single trial with a large number of runs. Although 10 trials with 10 000 runs and 1 trial with 100 000 runs cost similar amount of time, the results obtained by these two schemes may be different. For the four test watersheds, we examined the performance of AMALGAM under the one

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

Table VI. Average and combined performance of different algorithms for the three-objective test case in the RCEW

Performance metrics / Method	Average performance			Combined performance		
	ε-Indicator	SI	HP	ε-Indicator	SI	HP
NSGA-II-100	0·118 (0·04774)	0·34% (0·98%)	0·121 (0·051)	0·054	3·43%	0·036
NSGA-II-250	0·11 (0·019)	0·56% (1·77%)	0·103 (0·017)	0·062	5·61%	0·05
SPEA2-100	0·116 (0·047)	1·00% (2·12%)	0·122 (0·059)	0·057	9·97%	0·02
SPEA2-250	0·102 (0·0339)	0·66% (1·64%)	0·099 (0·038)	0·043	5·30%	0·023
AMALGAM	**0·066** (0·0242)	**7·57%** (8·31%)	**0·06** (0·024)	**0·015**	**75·70%**	**0·011**

Table VII. Average and combined performance of different algorithms for the five-objective test case in the LREW

Performance metrics / Method	Average performance		Combined performance	
	ε-Indicator	SI	ε-Indicator	SI
NSGA-II-100	0·076 (0·014)	0·16% (0·26%)	0·053	1·64%
NSGA-II-2500	(0·085) (0·008)	0·03% (0·06%)	0·058	0·25%
SPEA2-100	0·066 (0·014)	2·24% (1·97%)	0·045	12·34%
SPEA2-250	0·072 (0·015)	2·71% (3·21%)	0·047	17·13%
AMALGAM	**0·055** (0·0047)	**5·90%** (5·42%)	**0·045**	**53·15%**

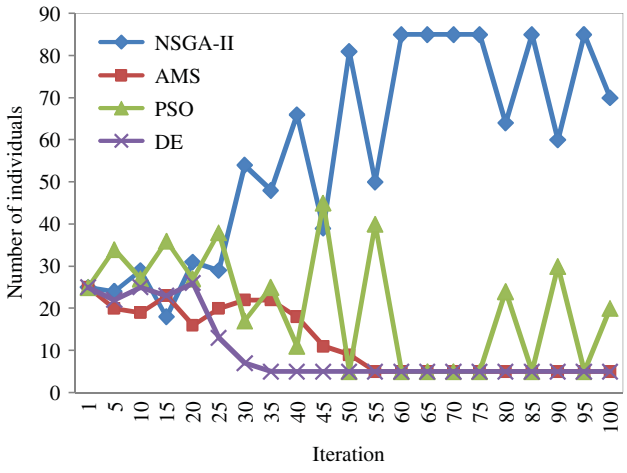

Figure 9. The number of individuals produced by each algorithm in AMALGAM against the iterations for one test trial in MCEW

single trial with 100 000 runs scheme. In order to compare the performance of AMALGAM with these two schemes, we calculated the non-dominated sets found by each of these two schemes. The performance metrics (Table IX) of each non-dominated set were computed with respect to the reference set derived from these two non-dominated sets. The AMALGAM($10 \times 10\,000$) scheme

Table VIII. Percentage of SWAT runs contributed by each algorithm in AMALGAM for the four test watersheds

Method / Watershed	NSGA-II (%)	AMS (%)	PSO (%)	DE (%)
YRHW	52	10	28	9
LREW	45	14	27	13
RCEW	47	5	26	21
MCEW	46	5	38	10

outperformed the AMALGAM($1 \times 100\,000$) scheme in three watersheds (i.e. MCEW, RCEW, and LREW), while AMALGAM($1 \times 100\,000$) scheme performed better in YRHW. In general, the multiple trials with a relatively small number of model runs scheme is preferred for implementing AMALGAM.

Previous applications of SWAT have reported that the time consumed by running SWAT once varies from seconds to hours. Implementing EMO methods for multi-objective parameter estimation of SWAT requires a large number of SWAT model runs, which is very time consuming. Reducing the time consumed by running SWAT by using advanced computational techniques deserves further research. Zhang *et al*. (2009a) compared the artificial neural network (ANN) and support vector machine (SVM) as surrogate of SWAT, and showed the potential of using SVM to save time consumed by calibration and uncertainty of SWAT. Combination of EMO methods and SVM has the potential to reduce computational burden of calibrating SWAT. Parallel computing techniques have also proved to be promising for efficient calibration of computationally intensive models. For example, Vrugt *et al*. (2006) reported that parallel implementation of the Shuffled Complex Evolution Metropolis (SCEM-UA) global optimization algorithm can substantially speed up the computational processes (closely approximates linear speed-up). In Vrugt *et al*. (2006, 2008) showed schemes of extending different optimization algorithms to parallel implementation version using Message Passing Interface (MPI), a mechanism for a specification of passing instructions between

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

Table IX. Performance of AMALGAM between two optimization schemes

Performance metrics	Watershed/optimization scheme	ε-indicator	SI (%)	HP
YRHW	AMALGAM(10 × 10 000)[a]	0·004	11·11	3·5E−05
	AMALGAM(1 × 100 000)[b]	0·001	88·89	1E−06
MCEW	AMALGAM(10 × 10 000)	0·002	88·64	2·2E−05
	AMALGAM(1 × 100 000)	0·003	29·55	0·0001
RCEW	AMALGAM(10 × 10 000)	0·017	81·39	0·0031
	AMALGAM(1 × 100 000)	0·042	18·61	0·0038
LREW	AMALGAM(10 × 10 000)	0·011	99	—
	AMALGAM(1 × 100 000)	0·088	1	—

[a] AMALGAM(10 × 10 000) represents the optimization scheme of 10 trials with 10 000 model runs.
[b] AMALGAM(1 × 100 000) represents the optimization scheme of one trial with 100 000 model runs.

different computational resources. Modification of optimization algorithms source code to implement on a distributed computer system is minor using MPI (Vrugt *et al*., 2006). The major limitation of implementing parallel computation is the availability of distributed computer systems. Given the enormous time consumed by calibration of SWAT, it is important to try different advanced computational techniques to improve the efficiency of EMO methods in the future.

In addition to the improve the efficiency and reliability of multi-objective optimization methods to find the Pareto-optimal solutions that are defined in objective space, their ability to explore the parameter space and provide a wide range of parameter sets for the modellers should be emphasized. Schaefli *et al*. (2004) showed that different parameter values can produce very similar performance metrics, but these different parameter values exhibited very different behaviour in projecting climate change impact on future water resources. In this case, using expert knowledge to select one or several appropriate parameter sets is necessary to reduce the uncertainty associated with parameter. Zhang *et al*. (2009c) also discussed the importance of considering the prior knowledge of parameter uncertainty in hydrological model selection. Finding a wide range of parameter sets with satisfactory performance metrics is the basis for appropriate parameter selection. AMALGAM cannot explicitly consider the differences between parameter values into the optimization process. Incorporating multi-mode search operators into the multi-objective optimization algorithms (e.g. Leyland *et al*., 2001; Schaefli *et al*., 2004) to improve their ability to explore parameter space deserves further research.

CONCLUSIONS

In this study, we compared the efficacy of two single-algorithm EMO methods (i.e. SPEA2 and NSGA-II) and a multi-algorithm, genetically adaptive multi-objective method (AMALGAM), for multi-site calibration of the SWAT model in four test watersheds with different characteristics. The results obtained in this study show that, without tuning population size, AMALGAM produced superior or competitive optimization results compared to other single-algorithm methods for most test cases. More importantly, the multi-method search framework within AMALGAM allows flexible incorporation of different algorithms. Compared with NSGA-II which has been incorporated into AMALGAM, our test results show that SPEA2 can provide better results for calibrating SWAT. It is promising to include SPEA2 into AMALGAM in the future. The difference between the average evaluation coefficient values and those obtained using multiple trials should be noted. Further analysis indicates that AMALGAM exhibited different performance between two implementation schemes (10 trials with 10 000 model runs vs 1 trial with 100 000 model runs). For practical use of AMALGAM, it is suggested to implement this method in multiple trials with a relatively small number of model runs rather than run it once with long iterations. In the future, evaluating and improving the ability of AMALGAM to explore the parameter space to provide a wide range of parameter sets for model selection deserve further research.

ACKNOWLEDGEMENTS

We sincerely appreciate the associate editor's and two reviewers' constructive comments and suggestions for revising this paper. Dr Jasper Vrugt at Los Alamos National Laboratory provided specific comments on revising this paper and guides on the application of AMALGAM. This code of AMALGAM in Visual Basic version used in this study was transferred from the source code in Matlab by Dr Vrugt. Dr Bettina Schaefli at Delft University of Technology provided constructive comments on the discussion of application multi-objective methods in hydrological model optimization. We appreciate Dr Patrick Reed at Pennsylvania State University for providing precious information on multi-objective optimization.

APPENDIX

$\boldsymbol{x}$ is the vector of hydrological parameters in this study;
$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ is the ith parameter solution in the population;

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/hyp

D is the number of optimized parameters;
$\boldsymbol{x}_{iD}$ is the dth dimension of the ith parameter solution;
Ω is the feasible space of parameters;
$\boldsymbol{F}(\boldsymbol{x}) = [f_1(x), f_2(x), \ldots, f_m(x)]$,is an objective function vector that contains multiple individual objective functions that need to be optimized simultaneously;
m is the number of objective functions;
$f_i(x)$, also denoted f_j, is the jth objective function;
P^* is Pareto-optimal set;
PF^* is the Pareto front;
FT_l is the lth front that is defined in NSGA-II;
T is the maximum number of generations;
t is the current generation number;
P_t is the population of parameter solutions at generation number t;
N is the number of parameter solutions in a population;
k is the number of candidate EMO algorithms in AMALGAM;
N_t^k is the number of offspring assigned to the kth model at tth generation;
$\overline{P}_t$ is the external archive at generation t which is used to store the parameter solutions with high fitness values;
$\overline{N}$ is the external archive size;
A is the Pareto optimal set.

REFERENCES

Ali MM, Khompatraporn C, Zabinsky ZB. 2005. A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems. *Journal of Global Optimization* **31**: 635–672.

Arabi M, Frankenberger JR, Engel BA, Arnold JG. 2008. Representation of agricultural conservation practices with SWAT. *Hydrological Processes* **22**(16): 3042–3055.

Arnold JG, Srinivasan R, Muttiah RS, Williams JR. 1998. Large-area hydrologic modeling and assessment: Part I. Model development. *Journal of American Water Resources Association* **34**(1): 73–89.

Bekele GE, Nicklow WJ. 2007. Multi-objective automatic calibration of SWAT using NSGA-II. *Journal of Hydrology* **341**: 165–176.

Beven KJ. 2000. *Rainfall-runoff Modelling : The Primer*. John Wiley & Sons: New York.

Cao W, Bowden BW, Davie T, Fenemor A. 2006. Multi-variable and multi-site calibration and validation of SWAT in a large mountainous catchment with high spatial variability. *Hydrological Processes* **20**: 1057–1073.

CEAP. 2008. *Conservation Effects Assessment Project*. Washington, D.C.: USDA Natural Resources Conservation Service. Available at: www.nrcs.usda.gov/technical/NRI/ceap/. Accessed on14 March 2008.

Coello Coello CA, Pulido GT, Lechuga MS. 2004. Handling multiple objectives with particle wwarm pptimization. *IEEE Transactions of Evolutinary Computation* **8**(3): 256–279.

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

Duan Q, Sorooshian S, Gupta VK. 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research* **28**(4): 1015–1031.

Engeland K, Braud I, Gottschalk L, Leblois E. 2006. Multi-objective regional modelling. *Journal of Hydrology* **327**: 339–351.

Gassman PW, Reyes M, Green CH, Arnold JG. 2007. The Soil and Water Assessment Tool: Historical development, applications, and future directions. *Transactions of the ASABE* **50**(4): 1212–1250.

Goldberg D. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley: Reading.

Gosain K, Rao S, Srinivasan R, Reddy NG. 2005. Return-flow assessment for irrigation command in the Palleru river basin using SWAT model. *Hydrological Processes* **19**(3): 673–682.

van Griensven A, Bauwens W. 2003. Multiobjective autocalibration for semidistributed water quality models. *Water Resources Research* **39**(12): 1348. DOI:10.1029/2003WR002284.

van Griensven A, Meixner T, Grunwald S, Bishop T, Di luzio M, Srinivasan R. 2006. A global sensitivity analysis tool for the parameters of multi-variable catchment models. *Journal of Hydrology* **324**: 10–23.

Gupta HV, Sorooshian S, Yapo PO. 1998. Toward improved calibration of hydrologic models: multiple and noncommensurate measures of information. *Water Resources Research* **34**(4): 751–763.

Haario H, Saksman E, Tamminen J. 2001. An adaptive metropolis algorithm. *Bernoulli* **7**(2): 223–242.

Kannan N, Santhi C, Williams JR, Arnold JG. 2008. Development of a continuous soil moisture accounting procedure for curve number methodology and its behaviour with different evapotranspiration methods. *Hydrological Processes* **22**: 2114–2121.

Kennedy J, Eberhart RC. 2001. *Swarm Intelligence*. Morgan Kaufmann: San Mateo.

Kim T, Heo JH, Jeong CS. 2006. Multireservoir system optimization in the Han River basin using multi-objective genetic algorithms. *Hydrological Processes* **20**: 2057–2075.

Kollat JB, Reed P. 2005. Comparing state-of-the-art evolutionary multi-objective algorithms for long-term groundwater monitoring design. *Advacnces in Water Resources* **29**: 792–807.

Lee KS and Chung ES. 2007. Hydrological effects of climate change, groundwater withdrawal, and land use in a small Korean watershed. *Hydrological Processes* **21**: 3046–3056.

Leyland GB, Molyneaux AK, Favrat D. 2001. *A New, Clustering Evolutionary Multi-objective Optimisation Technique*. Accessed at http://citeseer.ist.psu.edu/cache/papers/cs/22723/http:zSzzSzwww.jeo.orgzSzemozSzmolyneaux01.pdf/a-new-clustering-evolutionary.pdf on Sep. 28, 2009.

Liu C. 2004. Study of some problems in water cycle changes of the Yellow River basin. *Advances in Water Sciences* **15**(5): 608–614 (in Chinese).

Migliaccio KW, Chaubey O. 2007. Comment on Cao W, Bowden BW, Davie T, Fenemor A. 2006. Multi-variable and multi-site calibration and validation of SWAT in a large mountainous catchment with high spatial variability. *Hydrological Processes* **21**(4): 3326–3328.

Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models: Part I. A discussion of principles. *Journal of Hydrology* **10**(3): 282–290.

Neitsch SL, Arnold JG, Kiniry JR, King KW, Williams JR. 2005a. *Soil and Water Assessment Tool (SWAT) theoretical documentation*. Blackland Research Center, Texas Agricultural Experiment Station, Temple, Texas, BRC Report 02–05.

Neitsch SL, Arnold JG, Kiniry JR, Srinivasan R, Williams JR. 2005b. *Soil and Water Assessment Tool (SWAT) users manual*, BRC Report 02–06. Blackland Research Center, Texas Agricultural Experiment Station: Temple, Texas.

Reddy MJ, Kumar DN. 2007. Multi-objective particle swarm optimization for generating optimal trade-offs in reservoir operation. *Hydrological Processes* **21**: 2897–2909.

Remegio B, Confesor J, Whittaker GW. 2007. Automatic calibration of hydrologic models with multi-objective evolutionary algorithm and pareto optimization. *Journal of American Water Resources Association* **43**(4): 981–989. DOI: 10.1111/j.1752-1688.2007.00080.x.

Reyes Sierra M, Coello Coello CA. 2005. Improving PSO-based Multiobjective Optimization using Crowding, Mutation and ε-Dominance. *Lecture Notes in Computer Science* **3410**: 505–519.

Rouhani H, Willems P, Wyseure G, Feyen J. 2007. Parameter estimation in semi-distributed hydrological catchment modelling using a multi-criteria objective function. *Hydrological Processes* **21**(22): 2998–3008.

Santhi C, Arnold JG, Williams JR, Dugas WA, Hauck L. 2001. Validation of the SWAT model on a large river basin with point and nonpoint sources. *Journa of Amereican Water Resources Association* **37**(5): 1169–1188.

Schaefli B, Hingray B, Musy A. 2004. Improved calibration of hydrological models: use of a multi-objective evolutionary algorithm for parameter and model structure uncertainty estimation. In *Hydrology: Science and Practice for the 21st Century*, Webb B (ed.) British Hydrological Society: London; 362–371.

Seyfried MS, Harris RC, Marks D, Jacob B. 2000. *A Geographic Database for Watershed Research, Reynolds Creek Experimental Watershed, Idaho, USA*. USDA ARS Technical Bulletin NWRC-2000-3.

Sheridan JM. 1997. Rainfall-streamflow relations for coastal plain watersheds. *Transactions of the ASAE* **13**(3): 333–344.

Storn R, Price K. 1997. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization* **11**: 341–359.

Copyright © 2009 John Wiley & Sons, Ltd.

Tang Y, Reed P, Wagener T. 2006. How effective and efficient are multiobjective evolutionary algorithms at hydrologic model calibration? *Hydrology and Earth System Sciences* **10**: 289–307.

Van Liew MW, Garbrecht J. 2003. Hydrologic simulation of the Little Washita River Experimental Watershed using SWAT. *Journa of American Water Resources Association* **39**(2): 413–426.

Van Liew MW, Veith TL, Bosch DD, Arnold JG. 2007. Suitability of SWAT for the conservation effects assessment project: a comparison on USDA ARS watersheds. *Journal of the Hydrological Engineering* **12**(2): 173–189.

Veith TL, Sharpley AN, Weld JL, Grurek WJ. 2005. Comparison of measured and simulated phosphorous losses with index site vulnerability. *Transactions of the ASAE* **48**(2): 557–565.

Vrugt JA, Gupta HV, Bastidas LA, Bouten W, Sorooshian S. 2003. Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research* **39**(8): 1214; DOI:10.1029/2002WR001746.

Vrugt JA, Nuallain B, Robinson BA, Bouten W, Dekker SC, Sloot PMA. 2006. Application of parallel computing to stochastic parameter estimation in environmental models. *Computers and Geosciences* **32**(8): 1139–1155.

Vrugt JA, Robinson BA, Hyman JM. 2009. Self-adaptive multimethod search for global optimization in real parameter spaces. *IEEE Transactions on Evolutionary Computation* **13**(2): 243–259. DOI:10.1109/TEVC.2008.924428.

Vrugt JA, Robinson BA. 2007. Improved evolutionary optimization from genetically adaptive multimethod search. *Proceedings of the National Academy of Sciences* **104**: 708–711.

Vrugt JA, Stauffer PH, Wohling T, Robinson BA, Vesselinov VV. 2008. Inverse modeling of subsurface ow and transport properties: a review with new developments. *Vadose Zone Journal* **7**(2): 843–864.

Wang G, Guo X, Shen Y, and Cheng G. 2003. Evolving landscapes in the headwaters area of the Yellow River (China) and their ecological implications. *Landscape Ecology* **18**: 363–375.

Wang S, Kang S, Zhang L, Li F. 2008. Modelling hydrological response to different land-use and climate change scenarios in the Zamu River basin of northwest China. *Hydrological Processes* **22**(14): 2502–2510.

White LK, Chaubey I. 2005. Sensitivity analysis, calibration, and validations for a multisite and multivariable SWAT model. *Journal of the American Water Resources Association* **41**(5): 1077–1089.

Wolpert DH, Macready WG. 1997. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1**(1): 67–82.

Yapo PO, Gupta HV, Sorooshian S. 1998. Multi-objective global optimization for hydrologic models. *Journal of Hydrology* **204**: 83–97.

Zhang X, Srinivasan R, Debele B, Hao F. 2008a. Runoff simulation of the headwaters of the yellow river using the SWAT model with three snowmelt algorithms. *Journal of the American Water Resources Association* **44**(1): 48–61.

Zhang X, Srinivasan R, Hao F. 2007. Predicting hydrologic response to climate change in the Luohe river basin using the SWAT MODEL. *Transactions of the ASABE* **50**(3): 901–910.

Zhang X, Srinivasan R, Van Liew M. 2008b. Multi-site calibration of the SWAT model for hydrologic modeling. *Transactions of the ASABE* **51**(6): 2039–2049.

Zhang X, Srinivasan R, Van Liew M. 2009a. Approximating SWAT model using artificial neural network and support vector machine. *Journa of the American Water Resources Association* **45**(2): 460–474.

Zhang X, Srinivasan R, Zhao K, Van Liew M. 2009b. Evaluation of global optimization algorithms for parameter calibration of a computationally intensive hydrologic model. *Hydrological Processes* **23**: 430–441.

Zhang X, Liang F, Srinivasan R, Van Liew M. 2009c. Estimating uncertainty of streamflow simulation using Bayesian neural networks. *Water Resources Research* **45**: W02403, DOI:10.1029/2008WR007030.

Zitzler E, Laumanns M, Thiele L. 2001. *SPEA2: Improving the Performance of the Strength Pareto Evolutionary Algorithm*. Technical Report 103. Zurich: Computer Engineering and Communication Networks Lab (TIK), Swiss Federal Institute of Technology (ETH).

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

Zitzler E, Thiele L, Laumanns M, Fonseca CM, Grunert da Fonseca V. 2003. Performance assessment of multiobjective optimizers: an analysis and review. *IEEE Transactions on Evolutionary Computation* **7**: 117–132.

Copyright © 2009 John Wiley & Sons, Ltd.