Review and comparison of performance indices for automatic model induction

Jayashree Chadalawada and Vladan Babovic

ABSTRACT

One of the more perplexing challenges for the hydrologic research community is the need for development of coupled systems involving integration of hydrologic, atmospheric and socio-economic relationships. Given the demand for integrated modelling and availability of enormous data with varying degrees of (un)certainty, there exists growing popularity of data-driven, unified theory catchment scale hydrological modelling frameworks. Recent research focuses on representation of distinct hydrological processes using mathematical model components that vary in a controlled manner, thereby deriving relationships between alternative conceptual model constructs and catchments' behaviour. With increasing computational power, an evolutionary approach to auto-configuration of conceptual hydrological models is gaining importance. Its successful implementation depends on the choice of evolutionary algorithm, inventory of model components, numerical implementation, rules of operation and fitness functions. In this study, genetic programming is used as an example of evolutionary algorithm that employs modelling decisions inspired by the Superflex framework to automatically induce optimal model configurations for the given catchment dataset. The main objective of this paper is to identify the effects of entropy, hydrological and statistical measures as optimization objectives on the performance of the proposed approach based on two synthetic case studies of varying complexity.

Key words | automatic model induction, conceptual hydrological modelling, flexible and modular modelling frameworks, genetic programming, performance indices

Jayashree Chadalawada (corresponding author)
Department of Civil and Environmental Engineering,
National University of Singapore,
Block E1-08-24, No. 1 Engineering Drive 2,
Singapore 117578
E-mail: *jayashree@u.nus.edu*

Vladan Babovic
Department of Civil and Environmental Engineering,
National University of Singapore,
Block E1A- 05-09, No. 1 Engineering Drive 2,
Singapore 117576

INTRODUCTION

Increasing demand for integrated environmental modelling and diversity of hydrological modelling approaches motivates the development of unified modelling frameworks (Clark *et al.* 2015b) that integrate, compare and evaluate multiple model representations. Unified hydrological modelling frameworks can be implemented by defining a general set of conservation equations and their numerical implementation, model structure units/reservoirs, flux parameterizations, model complexity that governs granularity of process representations, spatial discretizations and hydrologic connectivities. In this context, flexible multi-model frameworks such as Rainfall-Runoff Modelling Toolbox (RRMT) (Wagener *et al.* 1999), Framework for Understanding Structural Errors (FUSE) (Clark *et al.* 2008) and multi-component frameworks such as FARM (Euser *et al.* 2013), Superflex (Fenicia *et al.* 2011, 2014; Kavetski & Fenicia 2011; van Esse *et al.* 2013) and SUMMA (Clark *et al.* 2015a) have been developed. In earlier applications, selection of the optimal model configuration was carried out based on hypothesizing model representations using prior knowledge of interactions among observed data and field observations, followed by evaluation of all possible combinations of the selected hypotheses using Bayesian framework, which is computationally intensive. On the other side, with advancements in computing power, use of machine learning tools is gaining popularity and data are becoming drivers for automatic optimal model selection, formulating modelling

doi: 10.2166/hydro.2017.078

strategies. One recent study has presented the application of data mining algorithm Automatic Model Configuration Algorithm (AMCA) (Vitolo 2015) to identify the most suitable model configuration using minimum data and FUSE as sample model inventory. The success of an ideal data-driven multi-component modelling framework is not only to perform automatic selection from certain hypotheses, but also to search an entire model space extensively to evolve the most suitable combinations of conceptual model components and generate novel modelling philosophies to fill knowledge gaps. Evolutionary computation (EC) techniques (Goldberg 1994; Babovic *et al.* 1995, 2017; Babovic & Abbott 1997; Babovic & Keijzer 2000) mimic the mechanisms of natural evolution to provide global or near global solutions to real world optimization problems. Literature shows the successful application of EC techniques in modelling hydrological processes of complex and chaotic nature in application areas, namely, rainfall–runoff prediction (Khu *et al.* 2001; Babovic & Keijzer 2002), settling velocity estimation (Babovic *et al.* 2001), water quality (Jeong *et al.* 2003), sediment transport (Kizhisseri *et al.* 2006), model error correction (Zechman & Ranjithan 2007), reservoir operation (Rani & Moreira 2010), water distribution systems (Savic & Walters 1997; Bi *et al.* 2016), etc. Earlier studies involve the use of EC techniques for evolving empirical, grey box models for short-term prediction rather than harnessing the idea of automated knowledge evolution and improving our current understanding of the system. Recent studies have attempted to incorporate hydrological concepts into the framework of an evolutionary data-driven algorithm, say, genetic programming (GP) (Koza 1992) to automatically estimate the input (rainfall) history for the prediction of runoff (Havlíček *et al.* 2013) and to automatically evolve Sugawara tank (Sugawara 1979) model configurations (Chadalawada *et al.* 2017) for the catchments of interest.

In this study, GP is used in conjunction with Superflex titled Evolutionary Superflex framework (http://scholarbank.nus.edu.sg/handle/10635/138678) for automatic evolution of optimal model configuration for the given catchment. As mentioned earlier, Superflex (Fenicia *et al.* 2011; Kavetski & Fenicia 2011) is a multi-component hydrological modelling framework that consists of building blocks that are customized into spatially lumped flow network configurations. The building blocks are reservoirs, flux and lag functions, connection elements. Twelve submodels (M1 to M12) (Figure 1) from Superflex framework (Fenicia *et al.* 2014) of varying complexities along with basic algebraic functions form the function set of GP. The selected submodels consist of five different types of reservoir units (interception (IR), unsaturated (UR), riparian (RR), fast reservoir and slow reservoir (FR and SR)) with serial, linear and parallel connections as illustrated in Figure 1. The constitutive functions (e.g., Monod, Modified logistic, Linear, Power, Reflected Hyperbolic, Tessier) governing storage–discharge relationships and evaporation losses, shape of lag functions (e.g., Triangle lag scheme), connection elements and ranges of associated parameters, which include, correction factors to inputs, splitters, coefficients of storages and outflows, smoothing factors, lag time are defined based on literature (Fenicia *et al.* 2011). The selected submodels represent a subset of possible conceptual model constructs of Superflex framework. Submodels M1 and M2 contain a single reservoir unit varying in terms of functions governing storage–discharge relation. M2 to M6 contain reservoir units with serial connections differing in terms of number and type of reservoir units, lag components, flux functions and associated parameters. M8 has two reservoir units in parallel with a splitter (D) dividing the incoming flux precipitation (P) between them. M9 to M12 represent the complex structures with either one (UR) or two reservoirs (IR and UR) preceding the parallelly connected FR and SR. The unsaturated zone reservoir is initialized to a percentage of its maximum capacity. All other reservoirs are assumed to be empty initially. The inputs to GP framework include observed forcings Precipitation (P) and Evaporation (E), Discharge (Q) and random constants. Other GP settings are listed in the Appendix (Table A1, available with the online version of this paper). For more details on GP framework consult (http://scholarbank.nus.edu.sg/handle/10635/138678). In Figure 1, rectangular boxes represent storage units and variables within parentheses represent parameters of the constitutive functions.

OBJECTIVE FUNCTIONS

The model identification and its performance evaluation should be based on ability to reproduce hydrological behaviour and capture properties of catchments rather than merely matching observed and simulated responses. There

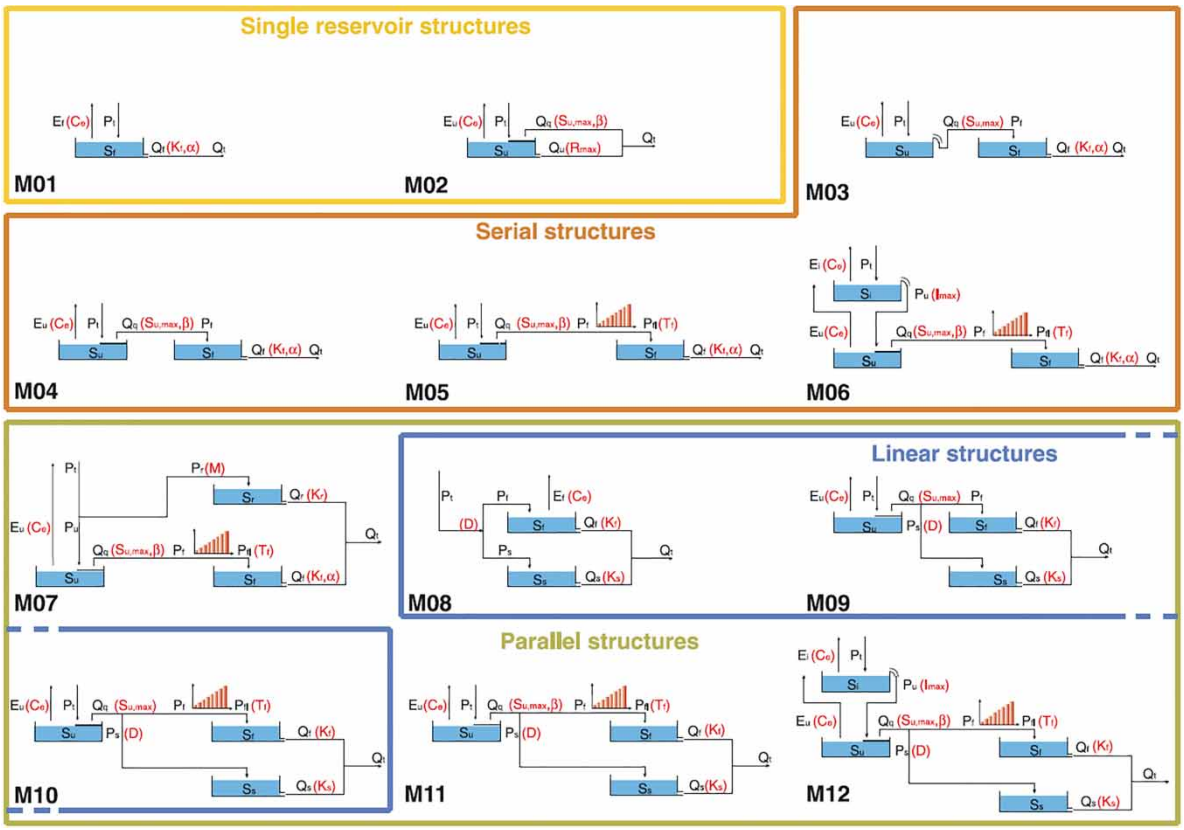

Figure 1 | Twelve Superflex submodels considered in this study (Fenicia *et al.* 2014). *P, E* and *Q* represent precipitation, potential evaporation and streamflow, respectively. S_I, S_U, S_R, S_F and S_S represent storages of IR, UR zone, RR zone, fast reacting and slow reacting reservoirs, respectively.

are a variety of performance indices that measure the ability of a hydrological model to reproduce observed streamflow. It is important for the simulated time series to preserve critical aspects of observed streamflow hydrograph (Vis *et al.* 2015) so that the governing model structure can also be used as a catchment classifier for predictions in ungauged basins. One of the recent studies (Shafii & Tolson 2015) highlights the point that the results are more sensitive to objective function formulation as opposed to the choice of optimization algorithm. Therefore, in this paper, a range of entropy, statistical and hydrological signatures listed in Table 1 are used as optimization objectives of Evolutionary Superflex approach.

In this review, a single objective optimization scheme is adopted to clearly assess each of the 14 objective functions highlighted in Table 1. The objective functions defined consist of one criterion or combined criteria representing different aspects of streamflow. This includes three one criterion (NS0, MD0 and RD0) and the rest as combined criteria objective functions.

The main characteristics of the metrics used in this study can be summarized as follows:

1. Nash–Sutcliffe efficiency (NSE) and mean absolute relative error (MARE) are sensitive to peak and low flows, respectively.
2. Modified NSE and logarithmic NSE emphasize low flows.
3. Mean absolute error (MAE) measures the agreement between the average simulated and observed catchment runoff volume and root mean squared error (RMSE) measures the overall agreement of hydrograph shape.
4. Volumetric efficiency (VE) measures the accuracy in prediction of overall runoff volume over the entire simulation period.
5. Correlation coefficient (r) measures the success in replicating overall timing and magnitude of discharge but disregards differences in absolute values.

Table 1 | Objective functions of evolutionary superflex framework

Objective function	Description	Criteria	Equations
1. Madsen (Madsen 2000)	a. Water balance; b. Hydrograph shape, good agreement of high and low flows	a. F1: Mean absolute error (MAE); b. F2, F3, F4: Overall, high, low flows RMSE	$MAE = \left\vert \frac{\sum_{t=1}^{N}(Qo_t - Qs_t)}{N} \right\vert$ $Overall,\ High\ and\ Low\ RMSE = \sqrt{\frac{\sum_{t=1}^{N_i}(Qo_t - Qs_t)^2}{N_i}},\ i = all,\ h,\ l$ $Madsen = \sqrt{(F_1 + A_1)^2 + \cdots + (F_4 + A_4)^2}$ Range: 0 (best) to ∞ $where,\ A_i = Max\{F_{j,min},\ j = 1\ to\ 4\} - F_{i,min},\ i = 1\ to\ 4$
2. NS0 (Nash & Sutcliffe 1970)	Sensitive to high flows	NA	$NSE = 1 - \frac{\sum_{t=1}^{N}(Qo_t - Qs_t)^2}{\sum_{t=1}^{N}(Qo_t - \overline{Qo_t})^2}$ Range: $-\infty$ to 1 (best) $NS0 = 1 - NSE$
Logarithmic NSE (Krause *et al.* 2005)	Sensitive to low flows	NA	$\log NSE = 1 - \frac{\sum_{t=1}^{N}(lnQo_t - lnQs_t)^2}{\sum_{t=1}^{N}(lnQo_t - ln\overline{Qo_t})^2}$ $logNS0 = 1 - logNSE$
3. Mai0 (Mai *et al.* 2016)	a. Sensitivity to very high flows; b. Weightage to low flows	a. NSE; b. logNSE	$Mai0 = \sqrt{NS0^2 + logNS0^2}$ $Mai = 1 - Mai0$
VE (Criss & Winston 2008)	Volumetric efficiency	NA	$VE = 1 - \frac{\left\vert \sum_{t=1}^{N}(Qo_t - Qs_t) \right\vert}{\sum_{t=1}^{N}(Qo_t)}$ Range: 0 to 1 (best) $VE0 = 1 - VE$
4. Vis_C1 (Vis *et al.* 2015)	a. Sensitivity to very high flows; b, c. Weightage to low flows, runoff volume	a. NSE; b. logNSE; c. VE	$Vis\_C1 = \sqrt{NS0^2 + logNS0^2 + VE0^2}$
MARE (Dawson *et al.* 2007)	Sensitivity to lower magnitudes	NA	$MARE = \frac{1}{N}\sum_{t=1}^{N}\frac{\vert Qo_t - Qs_t \vert}{Qo_t}$ Range: 0 (best) to ∞
5. Vis_C2 (Vis *et al.* 2015)	a. Sensitivity to very high flows; b. Sensitivity to both peaks and lows; c, d. Good agreement of timing, discharge magnitude and runoff volume	a. NSE; b. MARE; c. Pearson correlation coefficient (r); d. VE	$Vis\_C2 = \sqrt{NS0^2 + MARE^2 + (1 - r)^2 + VE0^2}$
6. Vis_C3 (Vis *et al.* 2015)	a, b. Good agreement of timing, discharge magnitude and runoff volume	a. r; b. VE	$Vis\_C3 = \sqrt{(1 - r)^2 + VE0^2}$

(*continued*)

Table 1 | continued

Objective function	Description	Criteria	Equations
7. KG10; 8. KG20 (Gupta *et al.* 2009; Kling *et al.* 2012)	Model efficiency measure based on equal weighting of correlation, bias and flow variability measures	a. r; b. β; c. α or γ; (γ ensures bias and variability ratios are not cross correlated)	KGE1 $= \sqrt{(1-r)^2 + (\alpha-1)^2 + (\beta-1)^2}$ Range: 0 to 1 (best) $\alpha = \frac{\sigma_s}{\sigma_0},\ \beta = \frac{\mu_s}{\mu_0}$ $KG10 = 1 - KGE1$ $KGE2 = \sqrt{(1-r)^2 + (\gamma-1)^2 + (\beta-1)^2}$ Range: 0 to 1 (best) $\gamma = \left(\frac{\sigma_s}{\mu_s} / \frac{\sigma_0}{\mu_0}\right)$ $KG20 = 1 - KGE2$
CED (Pechlivanidis *et al.* 2012)	Importance weighted informational entropy-based metric	NA	$H_x = \sum_{i=1}^{N} p(x_i)\ log_2\ p(x_i)$ $SUSE = \max\left(\left\|H_s^{unscaled} - H_o^{unscaled}\right\|,\ \left\|H_s^{scaled} - H_o^{scaled}\right\|\right)$ $CED = \max\left[SUSE_{(0-sg_1)},\ \ldots..,\ SUSE_{(sg_m-100)}\right]$ Range: 0 (best) to 1
9. CED_new (Pechlivanidis *et al.* 2014)	a. Static information of signal b. Timing information	a. CED; b. KGE1	CED|KG10$= \sqrt{CED^2 + KG10^2}$
10. Borsanyi (Borsányi *et al.* 2016)	Ideal point in 4D space in which correlation, mean squared error (MSE) are weighted more	a. r; b. NSE; c. VE; d. KGE1	Borsanyi $= \sqrt{(1-r)^2 + NS0^2 + VE0^2 + KG10^2}$
Modified Nash Sutcliffe efficiency (MNS) (Price *et al.* 2012)	Modified NSE	NA	$\mathrm{MNS0} = \frac{\sum_{t=1}^{N}\left\|\mathrm{Qo_t} - \mathrm{Qs_t}\right\|}{\sum_{t=1}^{N}\left\|\mathrm{Qo_t} - \overline{\mathrm{Qo_t}}\right\|}$ Range: $-\infty$ to 1 (best) $MNS = 1 - MNS0$
11. Price (Price *et al.* 2012)	a. Emphasize on flood peaks and low flows; b. Prioritize flow variability	a. NSE, MNS; b. α	Price $= \sqrt{NS0^2 + MNS0^2 + (1-\alpha)^2}$
PI (Dawson *et al.* 2007)	Best estimate for future is given by the latest value	NA	$PI0 = \frac{\sum_{t=1}^{N}(Qo_t - Qs_t)^2}{\sum_{t=1}^{N}(Qo_t - Qo_{t-1})^2}$ $PI = 1 - PI0$ Range: $-\infty$ to 1 (best)
12. Dawson (Dawson *et al.* 2012)	a. Emphasize on high flows; b, c. Overall agreement; d. Timing	a. Overall RMSE; b. r^2; c. MAE; d. PI	$Dawson = \sqrt{RMSE^2 + (1-r^2)^2 + PI0^2 + MAE^2}$ r^2: *coefficient of determination*
13. Modified Index of Agreement (Legates & McCabe 1999)	Detects additive and proportional differences in observed and simulated means and variances	NA	$MD0 = \frac{\sum_{t=1}^{N}(Qo_t - Qs_t)^j}{\sum_{t=1}^{N}\left\|Qs_t - \overline{Qo_t}\right\| + \left\|Qo_t - \overline{Qo_t}\right\|^j},$ $\mathrm{j} = 1$ Range: 0 to 1 (best) $MD = 1 - MD0$

(*continued*)

Table 1 | continued

Objective function	Description	Criteria	Equations
14. Relative Index of Agreement (Krause *et al.* 2005)	Sensitive to over- or under-prediction of low flows. (Cannot be computed if any of observed flows are equal to zero)	NA	$RD0 = \frac{\sum_{t=1}^{N}((Qo_t - Qs_t)/Qo_t)^2}{\sum_{t=1}^{N}\left(\left(\left\|Qs_t - \overline{Qo_t}\right\| + \left\|Qo_t - \overline{Qo_t}\right\|\right)/Qo_t\right)^2}$ Range: 0 to 1 (best) $RD0 = 1 - RD$

Notes: N and N_{all} denote entire length of data, N_h, N_l are the number of time steps corresponding to high and low flow values, respectively, Qs_t and Qo_t denote simulated and observed streamflows, respectively. A_i represents transformation constants such that $F_i + A_i$ have the same distance to the ideal point (origin), $F_{i,min}$ denotes minimum values of F_i estimated from initial GP population. $p(x_i)$ denotes probability of outcome x_i such that the probabilities sum to 1, $H^{unscaled}$, H^{scaled} denote unscaled and scaled Shannon entropies and SUSE represents maximum of unscaled and scaled Shannon entropy difference. m is the number of FDC segments and sg is the probability of exceedance for each segment of FDC. σ_0 and σ_s represent standard deviation of observed and simulated flows, respectively, μ_0 and μ_s represent mean of observed and simulated flows, respectively. α denotes relative variability in simulated and observed flows, β (bias) denotes ratio between mean observed and mean simulated flows and γ represents ratio of coefficient of variation of simulated and observed flows.

6. Kling–Gupta efficiency (KGE) combines correlation, bias and variability into one objective metric and is proven to have overcome the problems associated with NSE.
7. Persistence index (PI) aims at comparing the performance of a model against a simple model using the observed value of the previous day as the prediction for the current day as opposed to a constant mean. This is based on the assumption that the process is a Wiener process, i.e., variance increases linearly with time.
8. Index of agreement is based on a two-part squared distance measurement in which the absolute difference between model and simulated value and observed mean is added to the absolute difference between observed record and observed mean. Modified index agreement (MD) involves replacing the squared (power = 2) differences that are sensitive to extreme values by different values of power. On the other hand, relative index of agreement (RD) quantifies the differences between observed and predicted values as relative deviations which reduce the influence of absolute differences during high flows and enhance the influence of absolute low flow differences.
9. Signature measures of variability of runoff generation mechanisms (runoff coefficients and recession curves) are believed to capture significant catchment information that is not conveyed regularly using least squares or conventional variance-based measures of fit. Flow duration curve (FDC) can be considered a hydrologically informative signature of catchment behaviour. Entropy-based descriptors are more sensitive to subtle differences in streamflow (Pechlivanidis *et al.* 2014). Conditioned entropy difference (CED) is a diagnostic FDC signature representing static information of flow frequency distribution. CED is temporally insensitive and is said to result in better performance when combined with measures that characterize temporal dynamics such as KGE.

The main objective of this work is to evaluate the ability of different objective functions in bringing out the best performance of the Evolutionary Superflex approach thereby resulting in simulated response that adequately preserves observed flow characteristics. The outline of the remainder of this paper is as follows. The next section provides a description of the synthetic dataset and Evolutionary Superflex simulations. The results and analysis of the simulations follow and finally the key conclusions are summarized.

DATA AND METHODOLOGY

Data used in this study are collected at headwater catchments Attert Basin, Luxembourg (Fenicia *et al.* 2014), over a period of five years from 1 September 2004 to 31 August 2009, covering a wide range of runoff events. Data consist of daily time steps of precipitation (P) and potential evaporation (E) in mm/h. Two synthetic streamflow time series are generated using the observed forcings (P and E) of two catchments, namely, Huewelerbach and Weierbach.

Huewelerbach and Weierbach catchments have contrasting geology dominated by sandstone-marly and schists, respectively. Huewelerbach is a comparatively larger catchment (2.7 km^2) with landuse consisting of forests and agricultural land, while Weierbach (0.42 km^2) is a fully forested catchment. Runoff behaviour varies between stable, low reactivity of Huewelerbach catchment with runoff coefficient (R_c) of 0.3 and strong seasonality of Weierbach catchment with R_c of 0.98 in wet and 0.095 in dry periods, respectively.

Two Superflex submodels M4 and M12 are used in the generation of synthetic streamflow series for Weierbach (*Q_MIV*) and Huewelerbach (*Q_MXII*) catchments, respectively. Among the Superflex submodels considered in this study, M12 is the most complex while M4 has a comparatively simpler structure (see Figure 1). The idea behind using synthetic data is to assess the ability of the objective function to guide GP towards the optimal known solution. In the case of synthetic data, the existence of the optimal solution can be guaranteed. The entire length of data (five years) is used for every simulation. First, one month is reserved for model spin up. All the reservoirs of the 12 selected Superflex submodels are initially assumed to be empty except for the UR reservoir that is initialized to a percentage of its maximum capacity. Euler implicit stepping scheme with fixed daily time steps is used in the flux computations. In this setting, GP minimizes the objective functions, evolves optimal model configurations using components (Superflex submodels and basic algebraic functions) of the function set, terminal set (P, E, constants) and other settings listed in the Appendix (Table A1). For each of the 14 objective functions listed in Table 1, 10 Evolutionary Superflex trials are carried out resulting in 14 model configurations having the corresponding best training fitnesses. The resultant model configurations are returned as simplified, easily interpretable, and ready to use equations using Yacas (Pinkus *et al.* 2012). To evaluate the performance of the resultant model configurations in representing target flow characteristics, standardized signature index sum criterion (Ley *et al.* 2016) is calculated:

$$Z_{sa} = \frac{|x_{sa}| - \overline{x_a}}{\sigma_a} \tag{1}$$

$$SIS_s = Z_{sFHV} + Z_{sFMV} + Z_{sFMS} + Z_{sFLV} \tag{2}$$

where s indicates model configuration induced by Evolutionary Superflex approach, a indicates type of FDC signature index, x its value, $|x_{sa}|$ indicates absolute values of every signature index, $\overline{x_a}$ and σ_a represent mean and standard deviation of $|x_{sa}|$ for all s. Z_{sa} represents Z score or standardized values. SIS_s represents the sum of standardized indices of each model configuration.

Sum of Standardized Indices (SIS) is a combined measure of signature indices derived from FDC (Yilmaz *et al.* 2008) that distinguishes the merits of the evolved model alternatives. FDC signature indices considered in the formulation of SIS include high flows corresponding to extreme rainfall events (FHV (FDC high-segment volume): probability of flow exceedance <2%), intermediate flows between high and medium runoff (FMV (FDC intermediate-segment volume): probability of flow exceedance between 2% and 20%), medium flows covering the mid slope of FDC corresponding to vertical moisture distribution (FMS (FDC mid-segment slope): flow exceedance probability between 20% and 70%) and low flows corresponding to base flows (FLV (FDC low-segment volume) flow exceedance probability >70%). These indices identify the portion of hydrograph where bias is the highest. SIS compares the deviation of simulated FDCs from observed FDC weighting all parts of FDC equally. Negative SIS values indicate above average performance and the model configuration with the lowest SIS value is the one that performs the best for the given catchment.

RESULTS

The equations of resultant model configurations induced by Evolutionary Superflex simulations using two synthetic case studies are presented in Tables 2 and 3, respectively. It is evident that the respective underlying model structures are induced correctly for all objective functions except for CED|KG10 under *Q_MIV*.

M3 and M4 submodels consist of two serially connected reservoirs (UR and FR) in which precipitation enters UR, whereas its excess storage is routed through FR (Figure 1).

Table 2 | Results of evolutionary superflex simulations based on synthetic data generated using superflex submodel M4

Fitness metric	Equation of the best model configuration (in terms of training fitness)	Best fitness (out of 10 runs)	True positives (out of 10 runs)	False positives (out of 10 runs)	Best models out of 10 runs of 50 generations (fitness range)
Borsanyi	Q_MIV = M4(P, E, 2.10, 0.16, 1.08, 440.10, 8.62, 4.47, 0.86)	0.05	5	M3(1), M9(4)	M3(5), M4(5) (0.05–0.096)
CED|KG10	Q_MIV = M3(P, E, 1.75, 0.79, 1.37, 120.10, 1.13, 0.11, 0.44)	0.05	4	M3(5), M9(1)	M3(5), M4(5) (0.046–0.067)
Dawson	Q_MIV = M4(P, E, 1.86, 0.33, 1.22, 330.10, 5.57, 8.12, 0.81)	0.13	3	M7(1), M9(5), M12(1)	M4(10) (0.132–0.164)
KG10	Q_MIV = M4(P, E, 1.08, 1.37, 1.17, 330.10, 6.67, 7.15, 0.81)	7×10^{-3}	3	M2(3), M9(4)	M4(4), M9(6) (7.4×10^{-3}–0.012)
KG20	Q_MIV = M4(P, E, 1.94, 0.25, 0.98, 230.10, 3.27, 0.15, 0.43)	5×10^{-3}	3	M3(4), M9(3)	M4(4), M9(6) (4.5×10^{-3}–0.016)
Mai0	Q_MIV = M4(P, E, 2.49, 0.09, 1.26, 340.10, 6.56, 7.98, 0.86)	0.02	3	M9(4), M10(2), M12(1)	M4(10) (0.019–0.038)
MD0	Q_MIV = M4(P, E, 1.51, 0.6, 1.14, 250.10, 4, 0.57, 0.73)	0.03	2	M3(1), M9(7)	M3(4), M4(6) (0.03–0.05)
Madsen	Q_MIV = M4(P, E, 2.04, 0.17, 0.98, 250.10, 4.11, 0.19, 0.72)	0.2	6	M9(3), M12(1)	M4(10) (0.193–0.447)
NS0	Q_MIV = M4(P, E, 1.42, 0.83, 1.12, 430.10, 8.54, 10, 0.85)	4×10^{-3}	1	M2(1), M3(1), M9(7)	M3(2), M4(2), M9(6) (4×10^{-3}–0.012)
Price	Q_MIV = M4(P, E, 2.66, 0.04, 1.09, 230.10, 3.42, 0.35, 0.67)	0.05	5	M9(5)	M4(10) (0.05–0.08)
RD0	Q_MIV = M4(P, E, 2.18, 0.18, 1.07, 290.10, 7.24, 1.19, 0.85)	2×10^{-6}	6	M6(1), M9(3)	M4(10) (2.4×10^{-6}–5×10^{-6})
Vis_C1	Q_MIV = M4(P, E, 1.96, 0.21, 1, 230.10, 4.01, 0.2, 0.68)	0.03	4	M9(6)	M4(10) (0.026–0.1)
Vis_C2	Q_MIV = M4(P, E, 2.06, 0.18, 1.1, 250.10, 4.49, 0.43, 0.74)	0.14	4	M6(1), M9(1), M10(4)	M4(10) (0.137–0.317)
Vis_C3	Q_MIV = M4(P, E, 1.83, 0.29, 1.13, 370.10, 6.88, 3.57, 0.82)	0.05	2	M9(7), M10(1)	M4(6), M9(4) (0.05–0.133)
Target	Q_MIV = M4(P, E, α_Qq_FR, K_Qq_FR, Ce, Smax_UR, β_Qq_UR, β_E_UR, SiniFr_UR) Q_MIV = M4(P, E, 2.00, 0.20, 1.00, 200.00, 3.00, 0.10, 0.50)				

The outflow from UR is a power function of storage in M4 rather than a threshold function in M3 and CED|KG10 fails to capture this difference.

Other resultant model configurations show differences in terms of parameter values. Further investigation is required to identify the objective functions and other settings that enable Evolutionary Superflex to efficiently search model parameter space and find the optimum.

In terms of identifying the underlying model structure, the objective functions that are able to find the right structures (termed true positives) in at least five out of ten independent runs are noted. Comparing Tables 2 and 3, the combined objective function Borsanyi has performed consistently well in evolving the correct model structure as the one with minimum fitness in 50% of the total runs considering both synthetic case studies of varying complexities.

The most commonly occurring false positives (wrong model structures established as the ones with minimum fitness in respective runs) in the case of simulations using Q_MIV and Q_MXII are M9, M3 and M9, M10, respectively. The objective functions resulting in less than five true positives and more than two different false positives in the case of Q_MIV and Q_MXII simulations are Dawson, Mai0, NS0, Vis_C2 and CED|KG10, RD0, respectively. The top

Table 3 | Results of evolutionary superflex simulations based on synthetic data generated using superflex submodel M12

Fitness metric	Equation of the best model configuration (in terms of training fitness)	Best fitness (out of 10 runs)	True positives (out of 10 runs)	False positives (out of 10 runs)	Best models out of 10 runs of 50 generations (fitness range)
Borsanyi	Q_MXII = M12(P, E, 3.24, 9.98, 1.05, 0.23, 0.95, 0.58, 220.10, 10.68, 0.16, 0.32, 1.16)	0.06	5	M9(2), M10(3)	M12(10) (0.056–0.094)
CED| KG10	Q_MXII = M12(P, E, 5.65, 0.58, 1.16, 0.07, 0.32, 0.92, 290.10, 7.67, 0.92, 2.23, 1.52)	0.15	3	M5(1), M9(5), M10(1)	M9(1), M10(3), M12(6) (0.146–0.183)
Dawson	Q_MXII = M12(P, E, 6.10, 2.46, 0.99, 0.25, 0.92, 0.74, 370.10, 14.88, 0.2, 0.75, 2.06)	0.09	5	M9(1), M10(4)	M12(10) (0.085–0.1)
KG10	Q_MXII = M12(P, E, 7.03, 0.44, 1.05, 0.33, 0.99, 0.83, 440.10, 20, 0.5, 8.26, 2)	7×10^{-3}	5	M9(1), M10(4)	M12(10) (7.3×10^{-4}–0.01)
KG20	Q_MXII = M12(P, E, 2.89, 0.14, 1.32, 0.43, 0.69, 0.7, 160.10, 17.10, 0.52, 4, 1.08)	0.01	3	M9(5), M10(2)	M12(5), M9(4), M10(1) (0.014–0.023)
Mai0	Q_MXII = M12(P, E, 7.8, 0.33, 0.98, 0.24, 0.98, 0.78, 330.10, 9.61, 0.1, 0.52, 3.2)	0.03	5	M9(4), M10(1)	M12(10) (0.026–0.039)
MD0	Q_MXII = M12(P, E, 7.23, 0.2, 1.04, 0.74, 0.3, 0.8, 390.10, 12.42, 0.26, 3.24, 1.44)	0.03	8	M9(2)	M12(10) (0.027–0.03)
Madsen	Q_MXII = M12(P, E, 3.29, 0.17, 1.04, 0.48, 0.48, 0.54, 200.10, 6.49, 0.02, 0.21, 1.09)	0.12	4	M9(1), M10(4)	M12(10) (0.11–0.258)
NS0	Q_MXII = M12(P, E, 4.66, 0.26, 1.08, 0.27, 0.94, 0.68, 280.10, 9.3, 0.16, 1.63, 2.4)	6×10^{-3}	7	M9(3)	M12(10) (6×10^{-3}–7×10^{-3})
Price	Q_MXII = M12(P, E, 5.3, 8.64, 1.13, 0.22, 0.9, 0.7, 280.10, 7.11, 0.23, 3.56, 1.13)	0.06	4	M9(2), M10(4)	M12(10) (0.057–0.069)
RD0	Q_MXII = M12(P, E, 4.28, 0.12, 0.92, 0.33, 0.9, 0.63, 160.10, 10.10, 0.06, 0.17, 1.14)	10^{-5}	4	M4(1), M6(1), M9(3), M10(1)	M12(10) (10^{-5}–1.6×10^{-5})
Vis_C1	Q_MXII = M12(P, E, 6.13, 0.28, 1.06, 0.33, 0.9, 0.72, 310.10, 11.48, 0.21, 2.26, 1.96)	0.1	5	M4(1), M9(2), M10(2)	M12(10) (0.096–0.106)
Vis_C2	Q_MXII = M12(P, E, 3.43, 0.16, 1.12, 0.46, 0.46, 0.62, 180.10, 7.05, 0.05, 0.34, 1.68)	0.39	6	M9(3), M10(1)	M12(10) (0.39–0.535)
Vis_C3	Q_MXII = M12(P, E, 3.83, 0.17, 1.03, 0.27, 0.94, 0.62, 240.10, 8.99, 0.1, 0.32, 1.94)	0.06	7	M9(2), M10(1)	M12(10) (0.06–0.085)
Target	Q_MXII = M12(P, E, β_Qq_UR, K_Qq_FR, Ce, K_Qq_SR, D_S, SiniFr_UR, Smax_UR, Smax_IR, m_QE_IR, β_E_UR, Tlag) Q_MXII = M12(P, E, 3, 0.2, 1, 1, 0.3, 0.5, 200, 10, 0.1, 0.1, 2)				

ten model configurations (in terms of training fitness) resulting from 50 generations of ten independent runs (500 generations) using each of the 14 objectives are analysed. It is found that the objective functions which identified the right model structure all ten times are Dawson, Mai0, Madsen, Price, RD0, Vis_C1, Vis_C2 (7 out of 14) and Borsanyi, Dawson, KG10, Mai0, MD0, Madsen, NS0, Price, RD0, Vis_C1, Vis_C2, Vis_C3 (12 out of 14) in the case of Q_MIV and Q_MXII simulations, respectively. The objectives other than those mentioned above fail to adequately distinguish between hypotheses (submodels) of the model structure space in the respective case studies. It is evident that greater efficiency in searching model structure space is observed in the case of M12 as compared to M4. This might be due to the fact that M12 is the most unique submodel of GP function set consisting of four reservoirs with both serial and parallel connections, whereas M4 is reflected as not significantly different from M3 (differs from M4 only in terms of UR outflow function) and M9 (differs from M4 in terms of UR outflow function and excess UR storage being routed through two parallel reservoirs FR and SR).

The variability of parameter values for the best model configurations simulating Q_MIV and Q_MXII are analysed using boxplots shown in Figures 2 and 3. For ease

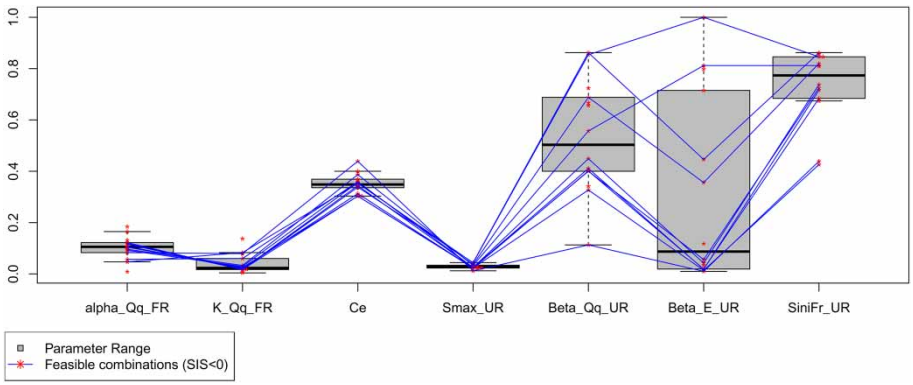

Figure 2 | Parameter uncertainties in the best *Q_MIV* model configurations induced by Evolutionary Superflex simulations using 14 objective functions.

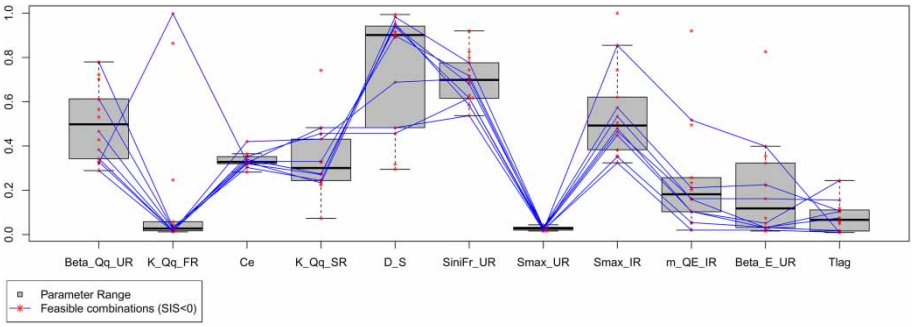

Figure 3 | Parameter uncertainties in the best *Q_MXII* model configurations induced by Evolutionary Superflex simulations using 14 objective functions.

of comparison, all associated parameter values are normalized between 0 and 1. Dots indicate the spread of parameter values and the lines connect the parameters of multiple feasible descriptions (model configurations with corresponding best training fitnesses and negative SIS values) of the target. From Figure 2, it can be observed that the least and the most uncertain parameters of evolved M4 model configurations are Smax_UR (maximum capacity of UR) and β_E_UR (smoothing factor for function governing evaporation loss from UR), respectively. In resultant M4 configurations whose structure consists of serially connected UR and FR, greater variation is observed in the parameters of flux functions of UR, namely, SiniFr_UR (Intial UR storage = SiniFr_UR*Smax_UR), β_E_UR and β_Qq_UR (outflow function parameter).

Figure 3 shows that the least and the most uncertain parameters of the resultant M12 configurations whose structure consists of serial network of IR and UR before parallelly connected FR and SR are K_Qq_FR and D_S, respectively. K_Qq_FR denotes hydraulic conductivity of outflow from FR and D_S is the splitter that diverts excess storage of UR to FR and SR, which implicitly governs the outflows of FR and SR. Greater variation is visually evident in parameters of IR, UR and SR of M12 configurations.

In GP runs simulating Q_MIV and Q_XII, the lowest variation is observed in the values of evaporation multiplication factor (Ce) and FR parameters (α_Qq_FR, K_Qq_FR).

Evolutionary Superflex approach shows greater uncertainty in filtering out combinations of parameters from model parameter space given that inherent model structure is accurately found at most instances. Beven (2006) suggests that this uncertainty should be viewed as a problem of decidability between different probable descriptions of the working of hydrological systems. Sources of uncertainty include model structure (process representations), numerical solution approximations and parameter values. Many different parameter sets within a chosen model structure

may be acceptable in reproducing the observed behaviour of that system and this concept is called 'Equifinality'.

Literature suggests that uncertainty estimation techniques allowing for equifinality can be broadly classified as formal (e.g., Bayesian) limited by requirements on prior knowledge and informal (e.g., generalized likelihood uncertainty estimation, GLUE) limited by subjectivity but simple (Vrugt *et al.* 2009). Here, GLUE, a useful paradigm that avoids overconditioning and estimates prediction quantiles is chosen for uncertainty analysis. All Evolutionary Superflex configurations generated using 14 objective functions that have rightly induced the inherent model structures M4 and M12 for *Q_MIV* and *Q_MXII*, respectively, are listed and the corresponding parameter sets are chosen for uncertainty analysis using GLUE. From the chosen parameter sets belonging to behavioural model space, elimination of those that do not perform adequately is carried out based on SIS criterion (likelihood measure) and the threshold is set as 0. More weight is assigned to parameter sets that result in negative SIS values and the prediction boundaries are constructed as shown in Figures 4 and 5. The percentage of observed values falling outside the prediction limits are observed as 1.49% and 3.58% for *Q_MIV* and *Q_MXII*, respectively. The width of the prediction boundary for *Q_MIV* (0.67) is smaller than for *Q_MXII* (0.99), which can be attributed to higher complexity of inherent model (M12) of *Q_MXII*. It can be seen in Figures 4 and 5 that the observed values closely follow the 50% quantile (median) of the simulated values. It has also been verified that uncertainty bands derived using parameter sets given by Evolutionary Superflex are narrower than using parameter sets randomly drawn from uniform distributions (with user specified parameter ranges) for the same likelihood measure (SIS).

In order to select the most appropriate, the best model configurations (Tables 2 and 3) of both the synthetic case studies are evaluated with respect to biases in FDC signatures. The FDCs (Figures 6 and 7) reflect the differences in the behaviour among the best configurations of *Q_MIV* and *Q_MXII* with negative SIS values.

Figure 6 illustrates that M4 model configuration evolved based on Vis_C1 with the least SIS value (−3.1) matches most of the *Q_MIV* characteristics showing 1% bias in FHV and negligible bias (less than ±0.2%) with respect to other FDC signatures (FMV, FMS and FLV) (see subplot of Figure 6 entitled 'Vis_C1'). KG20 (SIS = −2.69) and Madsen (SIS = −2.16) also resulted in good overall approximation of the target with slightly greater deviations in FMS and FMS, FLV respectively (see subplots of Figure 6 entitled 'KG20' and 'Madsen'), when compared to the best objective

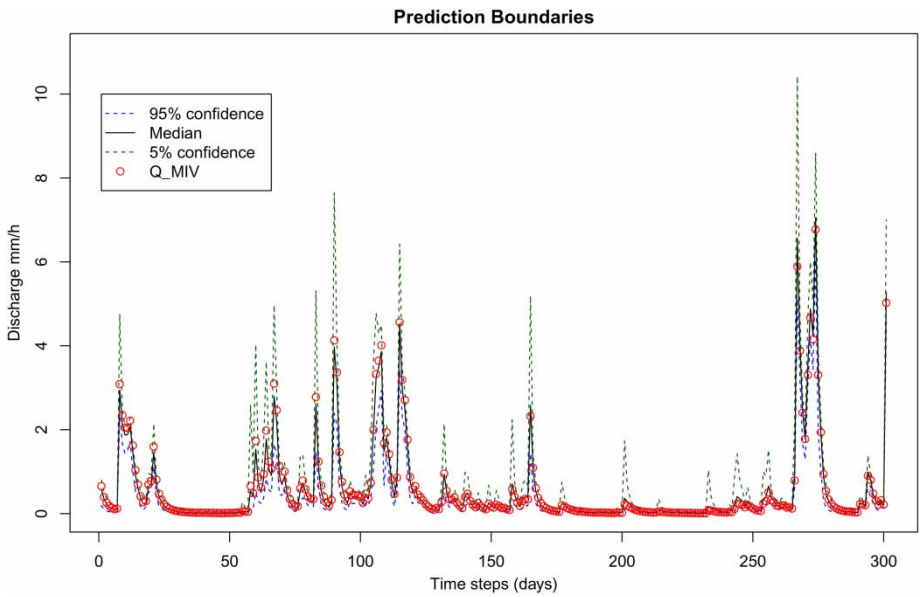

Figure 4 | Uncertainty ranges of predictions of Evolutionary Superflex derived using GLUE for a representative portion of *Q_MIV*.

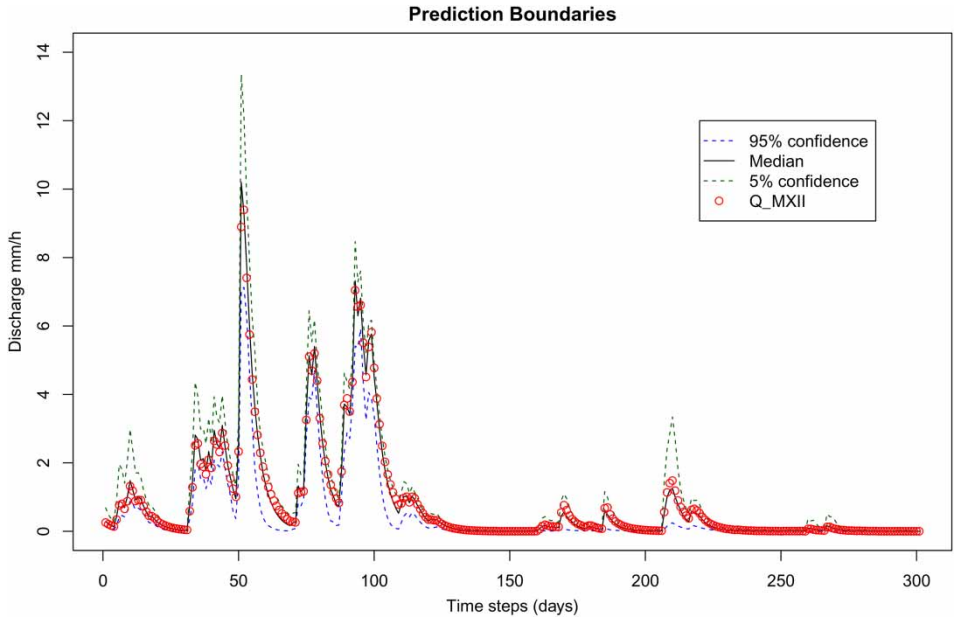

Figure 5 | Uncertainty ranges of predictions of Evolutionary Superflex derived using GLUE for a representative portion of *Q_MXII*.

Vis_C1. Although the wrong model structure M3 (differs from M4 in terms of UR outflow function) identified as the best by CED|KG10 has a negative SIS value, it results in the poorest performance with respect to low flows (see subplot of Figure 6 entitled 'CED_new').

Figure 7 shows that M12 model configurations evolved based on Madsen (SIS = −3.05), Vis_C3 (SIS = −2.83) and Vis_C2 (−2.49) perform well in comparison to others. The FDC signatures of Madsen, Vis_C2 and Vis_C3 show that high and medium flows are better approximated by M12 model configuration evolved based on Vis_C3 in comparison to Madsen and Vis_C2. Intermediate flows are well captured by M12 model configuration evolved based on Madsen. Considering low flows, the lowest value of FLV bias is observed in the case of Vis_C2.

Comparing Figures 6 and 7, it is evident that the optimal model reported by Evolutionary Superflex approach in the case of simulations using *Q_MIV* is superior (close match to the target both in terms of structure and parameters) to those using *Q_MXII*. This can be attributed to higher complexity of M12 submodel both in terms of structure and number of parameters to be optimized. It can be concluded that combined objective functions, say, Vis_C1, KG20 and Madsen for *Q_MIV* simulations and Madsen, Vis_C3 and Vis_C2 for *Q_MXII* simulations resulted in Evolutionary Superflex performing better compared to others. Madsen fitness metric is observed as the common best objective irrespective of the complexity of model search space.

Tables 4 and 5 show the performance evaluation of the best model configurations of Evolutionary Superflex simulations using *Q_MIV* (Table 2) and *Q_MXII* (Table 3), respectively, based on hydrological efficiency criteria KGE1, KGE2, NSE and VE, correlation (r) and numerical accuracy metrics MAE, MD, RMSE and PI. No considerable variation with respect to SUSE, KGE1, KGE2, NSE, MD and r is observed (except in the case of RD0 for *Q_MXII* simulations) making them unfit to decide between different system representations. Very high values of Pearson correlation coefficient (r) (see Tables 4 and 5) associated with all the best configurations indicate the absence of phase errors and good agreement of observed and simulated means and variances.

In Table 4, the low RMSE values of 0.055 and 0.066 correspond to the best M4 model configurations induced based on Vis_C1 and Madsen, respectively, which is in agreement with the decision on the most suitable configurations made using SIS criterion. On the other hand, the highest RMSE and the lowest VE (see Table 4) observed in the case of model configuration induced based on CED|KG10 for *Q_MIV* is not to be considered, although it has a negative SIS value. KG20 is associated with better MAE (1.6×10^{-3})

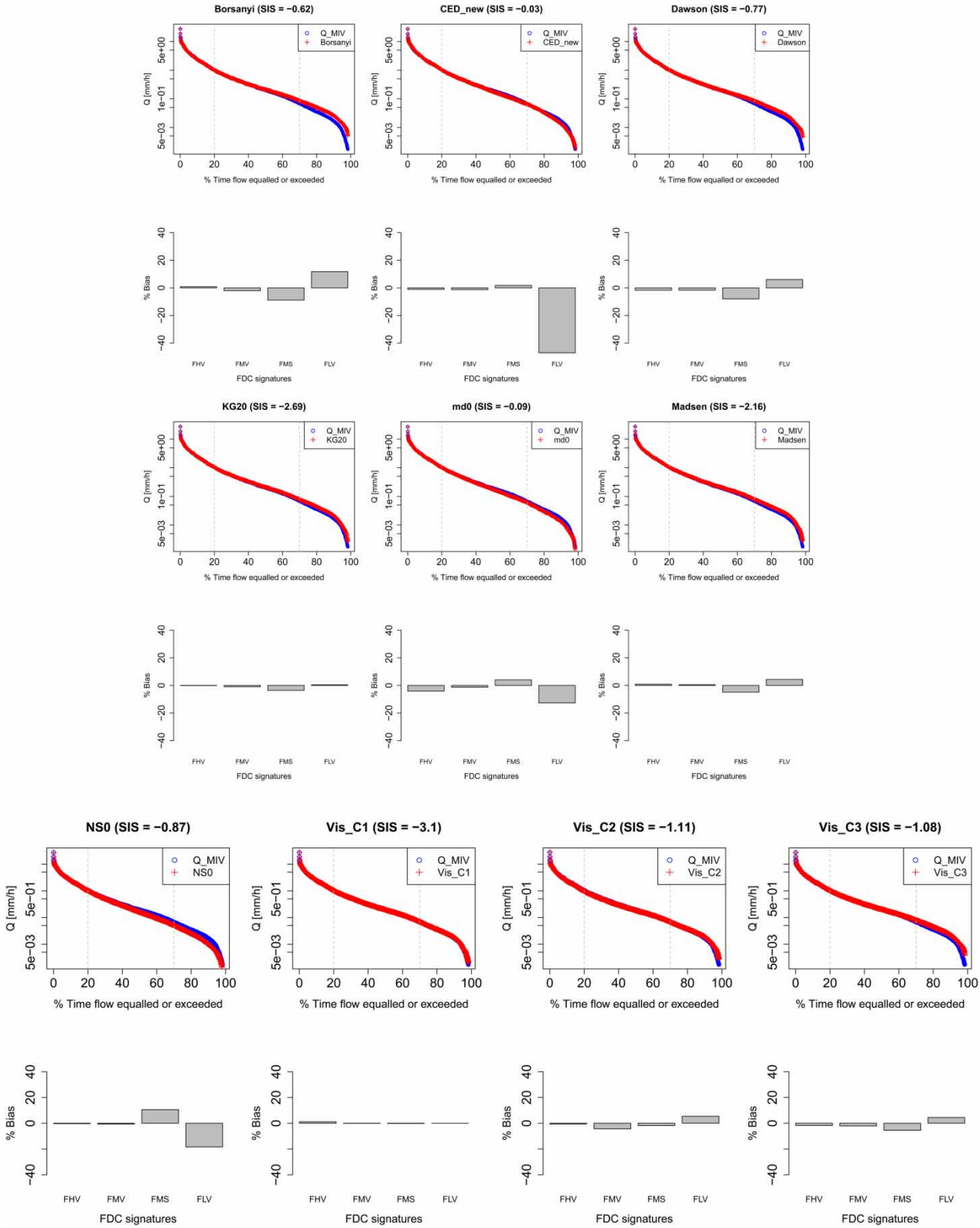

Figure 6 | FDC and % bias of FDC signature indices of the best *Q_MIV* model configurations values evolved by Evolutionary Superflex simulations with negative SIS values.

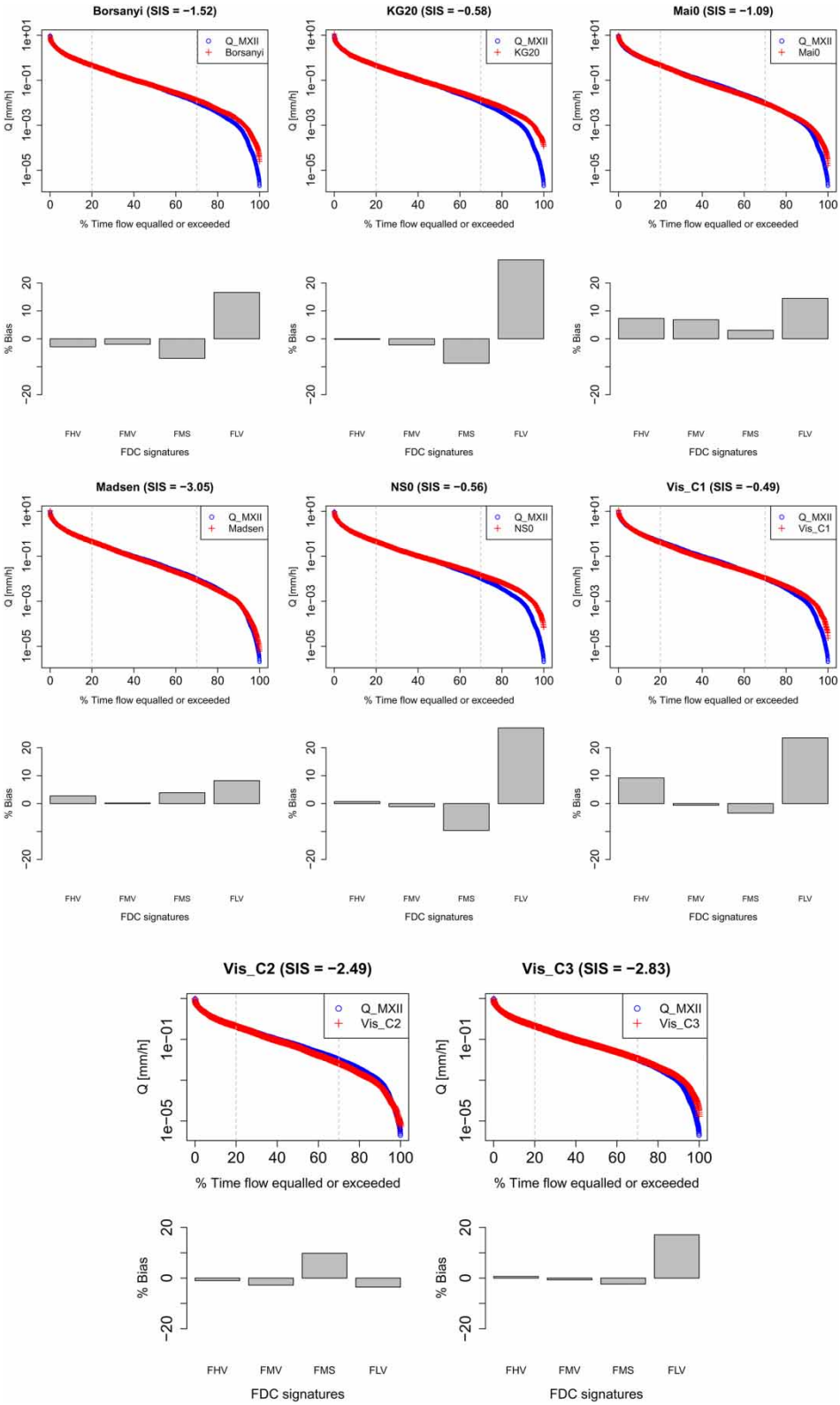

Figure 7 | FDC and % bias of FDC signature indices of the best *Q_MXII* model configurations values evolved by Evolutionary Superflex simulations with negative SIS values.

Table 4 | Performance of the best configurations evolved by evolutionary superflex for *Q_MIV* using 14 objective functions

Objective functions (optimum)	SUSE (0)	KGE1 (1)	KGE2 (1)	MAE (0)	MD (1)	NSE (1)	PI (1)	r (1)	RMSE (0)	VE (1)
Borsanyi	0.0028	0.996	0.994	0.0026	0.979	0.998	0.997	0.999	0.088	0.951
CED|KG10	0.0024	0.962	0.964	0.0025	0.891	0.94	0.914	0.971	0.442	0.741
Dawson	0.005	0.983	0.989	0.0056	0.967	0.996	0.994	0.998	0.116	0.922
KG10	0.013	0.992	0.992	0.0015	0.939	0.985	0.978	0.992	0.222	0.852
KG20	9.7×10^{-4}	0.995	0.995	0.0016	0.969	0.991	0.988	0.996	0.167	0.927
Mai0	0.025	0.922	0.909	0.064	0.958	0.991	0.987	0.996	0.17	0.903
MD0	0.0064	0.957	0.971	0.025	0.97	0.996	0.995	0.999	0.111	0.929
Madsen	0.0064	0.983	0.982	0.013	0.984	0.999	0.998	0.999	0.066	0.963
NS0	0.011	0.973	0.961	0.023	0.962	0.996	0.994	0.998	0.114	0.910
Price	0.0052	0.983	0.987	0.003	0.975	0.998	0.997	0.999	0.082	0.942
RD0	0.027	0.913	0.927	0.012	0.956	0.982	0.974	0.995	0.243	0.893
Vis_C1	0.0036	0.99	0.988	0.0017	0.989	0.999	0.999	0.999	0.055	0.974
Vis_C2	0.011	0.958	0.951	0.034	0.981	0.998	0.997	0.999	0.09	0.954
Vis_C3	0.0031	0.978	0.986	0.012	0.978	0.998	0.997	0.999	0.084	0.949

Table 5 | Performance of the best configurations evolved by evolutionary superflex for *Q_MXII* using 14 objective functions

Objective functions (optimum)	SUSE (0)	KGE1 (1)	KGE2 (1)	MAE (0)	MD (1)	NSE (1)	PI (1)	r (1)	RMSE (0)	VE (1)
Borsanyi	0.002	0.971	0.982	0.006	0.973	0.993	0.951	0.997	0.071	0.932
CED|KG10	0.03	0.953	0.936	0.015	0.905	0.951	0.627	0.975	0.196	0.759
Dawson	0.015	0.985	0.988	0.004	0.971	0.997	0.974	0.998	0.052	0.926
KG10	0.013	0.992	0.992	1.3×10^{-4}	0.94	0.985	0.887	0.993	0.108	0.848
KG20	0.022	0.975	0.978	0.0018	0.922	0.965	0.738	0.983	0.164	0.8
Mai0	0.022	0.904	0.941	0.018	0.954	0.982	0.867	0.995	0.117	0.878
MD0	0.023	0.981	0.974	0.0073	0.973	0.997	0.977	0.999	0.049	0.930
Madsen	0.029	0.966	0.963	0.0012	0.966	0.990	0.921	0.995	0.09	0.911
NS0	0.0029	0.992	0.988	0.0014	0.966	0.994	0.956	0.997	0.067	0.912
Price	0.016	0.991	0.989	0.0017	0.958	0.985	0.881	0.993	0.108	0.894
RD0	0.029	0.583	0.7	0.114	0.879	0.847	−0.16	0.98	0.345	0.657
Vis_C1	0.039	0.917	0.916	1.4×10^{-4}	0.952	0.982	0.863	0.994	0.119	0.875
Vis_C2	0.019	0.965	0.952	0.013	0.96	0.99	0.923	0.995	0.089	0.896
Vis_C3	0.0073	0.987	0.979	0.004	0.973	0.996	0.968	0.998	0.057	0.930

than Madsen (0.013) while RMSE of Madsen (0.066) is superior to KG20 (0.167). Both SIS criterion and performance evaluation based on a wide range of metrics indicate that M4 configurations evolved based on KG20 (QGP_KG20), Madsen (QGP_Madsen) and Vis_C1 (QGP_Vis_C1) can be considered the most suitable for Weierbach catchment synthetic dataset used in this study (see Figure 8).

Table 5 shows that the best M12 model configurations evolved based on CED|KG10 (SIS = 1.25) and RD0 (SIS = 6.18) with low PI values of 0.63 and −0.16, high RMSE values of 0.2 and 0.34, low VE values of 0.76 and 0.66, respectively, can be ignored. The best configurations evolved using the objective functions Borsanyi, Dawson, MD0, Madsen, NS0, Vis_C2 and Vis_C3 have

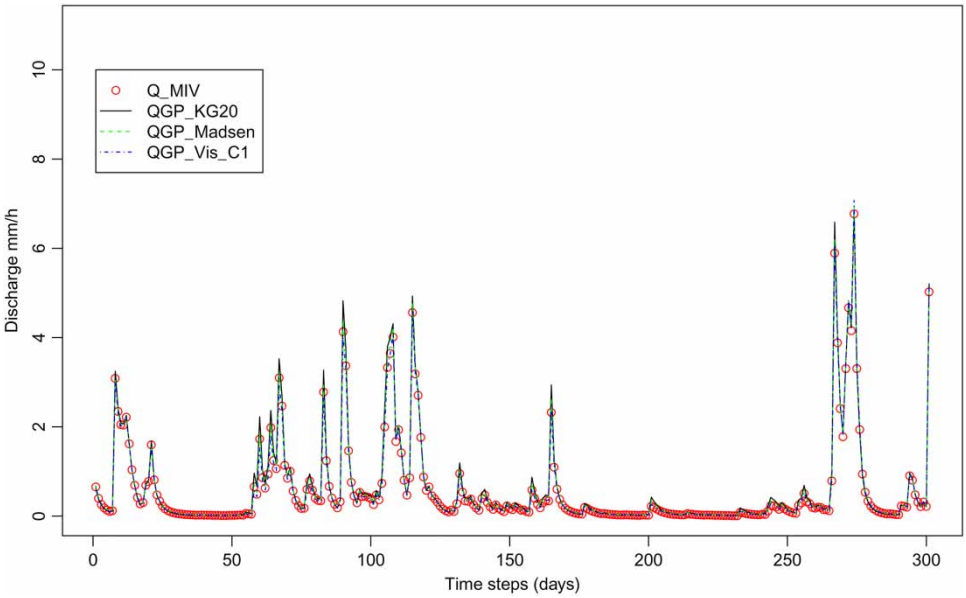

Figure 8 | Observed and simulated hydrographs of the most suitable Evolutionary Superflex models for a portion of synthetic Weierbach catchment data.

better fitness in comparison to the rest (Table 5). Among the above mentioned, all configurations have negative SIS values except for those evolved using Dawson (SIS = 1.03) and MD0 (SIS = 1.45). As the complexity of the system increases, the model selection decisions based on FDC signature indices and other performance evaluation metrics do not match each other completely. In such cases, the choice based on SIS is considered superior as it equally weights different flow aspects. Thus, M12 model configurations evolved using Madsen (QGP_Madsen), Vis_C2 (QGP_Vis_C2) and Vis_C3 (QGP_Vis_C3) are considered the most suitable for Huewelerbach catchment synthetic dataset used in this study (see Figure 9).

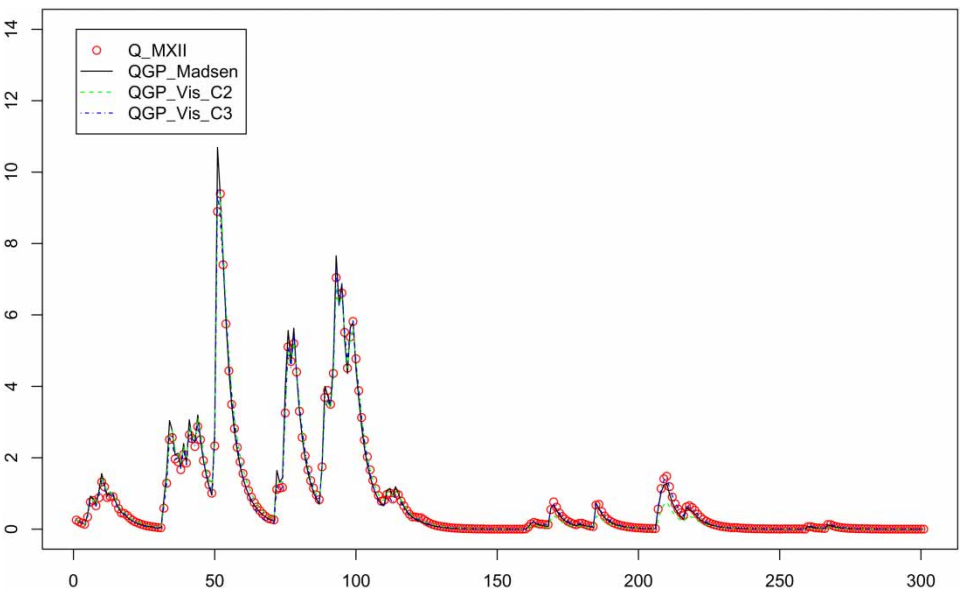

Figure 9 | Observed and simulated hydrographs of the most suitable Evolutionary Superflex models for a portion of synthetic Huewelerbach catchment data.

SUMMARY AND FUTURE WORK

This review provides an assessment of abilities of different objective functions that enable Evolutionary Superflex modelling framework to identify the most suitable model configurations (structure + parameters set) that reproduce several target flow regime signatures simultaneously. As expected, model configurations evolved based on one criterion objective function do not ensure good agreement with respect to others. NS0 has performed better than the other one criterion objective functions used in this study. Combined objective functions have delivered good performance across a range of flow characteristics. The proposed approach performs well with respect to refinement of model structure space using objective functions defined in this study. Greater variation is observed in filtering out feasible parameter combinations from model parameter space. The uncertainty bands of simulated model parameters are quantified using GLUE approach and it is found that observed values closely follow the median of simulated values. The uncertainty in the prediction of parameter values increases as the complexity of the structure increases. Among the five types of reservoirs (IR, RR, UR, FR and SR) that constitute the structures of submodels employed in this study, the parameters of FR show the least variation. The best configurations of Evolutionary Superflex approach evolved using each of the 14 objective functions can be considered as multiple descriptions of the system under study. Deciding on the most suitable configurations among them is done using SIS criterion and other selected performance evaluation metrics. Each performance evaluation metric measures one aspect of the difference between observed and simulated responses. Some of the metrics, say, SUSE, KGE, NSE, MD and r assign similar values to different hydrographs and are not suitable for selection of the most suitable configurations. Difference is observed between selection of suitable configurations based on SIS criterion and other performance evaluation metrics, as complexity of the underlying system increases. The model configuration selected based on SIS criterion overrules in case of disparity as SIS equally weights (different portions of the hydrograph). The combined objective functions, say, Madsen, KG20, Vis_C1 and Madsen, Vis_C2, Vis_C3 are found to be the best for induction of inherent models of Weierbach and Huewelerbach catchments synthetic datasets, respectively. Madsen metric that equally weights agreement of overall volume, overall shape of system response, high and low flows is found as the common best objective for both case studies, irrespective of the complexity of search space. It remains a challenge to pinpoint an objective function that brings out the best performance of Evolutionary Superflex approach in effectively searching the model space to evolve optimal conceptual model constructs and parameters for given datasets. This study aims to run the evolutionary modelling framework using different, strategically aggregated objective functions and to select the most suitable model based on the collective analysis of results obtained.

This research will be extended to capture hydrological processes of larger and complex catchments in order to reinforce the current findings. An attempt can be made to constrain equifinal parameters to ranges associated with observed catchment behaviour using identifiability analysis which can result in narrow uncertainty bands. Also, the efficiency of the proposed modelling approach will be tested in a multiobjective optimization context.

ACKNOWLEDGEMENTS

The authors thank Dr Laurent Pfister, Luxembourg Institute of Science and Technology (LIST) for granting the permission to use Luxembourg catchments dataset for this research study. Additionally, we are grateful to Dr Fabrizio Fenicia for his assistance with Superflex framework.

REFERENCES

Babovic, V. & Abbott, M. B. 1997 The evolution of equations from hydraulic data part I: theory. *Journal of Hydraulic Research* **35**, 397–410.

Babovic, V. & Keijzer, M. 2000 Genetic programming as a model induction engine. *Journal of Hydroinformatics* **2**, 35–60.

Babovic, V. & Keijzer, M. 2002 Rainfall runoff modelling based on genetic programming. *Hydrology Research* **33**, 331–346.

Babovic, V., Zhengyi, W. & Larsen, L. 1995 Calibrating hydrodynamic models by means of simulated evolution. *Oceanographic Literature Review* **11**, 1025.

Babovic, V., Keijzer, M., Aguilera, D. R. & Harrington, J. 2001 An evolutionary approach to knowledge induction: genetic programming in hydraulic engineering. In: *Bridging the Gap: Meeting the World's Water and Environmental Resources Challenges* (D. Phelps & G. Shelke, eds). American Society of Civil Engineers, pp. 1–10.

Babovic, V., Xin, L. & Chadalawada, J. 2017 Rainfall-runoff modeling based on genetic programming. *Encyclopedia of Water: Science, Technology, and Society* (in press).

Beven, K. 2006 A manifesto for the equifinality thesis. *Journal of Hydrology* **320**, 18–36.

Bi, W., Dandy, G. C. & Maier, H. R. 2016 Use of domain knowledge to increase the convergence rate of evolutionary algorithms for optimizing the cost and resilience of water distribution systems. *Journal of Water Resources Planning and Management* **142**, 04016027.

Borsányi, P., Hamududu, B., Wong Kwok, W., Magnusson, J. & Shi, M. 2016 First steps in incorporating data-driven modelling to flood early warning in Norway's flood forecasting service. In: *EGU General Assembly Conference Abstracts, EGU General Assembly*, Vienna, Austria, pp. 7661.

Chadalawada, J., Havlicek, V. & Babovic, V. 2017 A genetic programming approach to system identification of rainfall-runoff models. *Water Resources Management* **31**, 3975–3992.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T. & Hay, L. E. 2008 Framework for understanding structural errors (FUSE): a modular framework to diagnose differences between hydrological models. *Water Resources Research* **44** (12), W00B02.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W. & Brekke, L. D. 2015a *The Structure for Unifying Multiple Modeling Alternatives (SUMMA), Version 1.0: Technical Description.* NCAR Tech. Note NCAR/TN-5141STR.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W. & Brekke, L. D. 2015b A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research* **51**, 2498–2514.

Criss, R. E. & Winston, W. E. 2008 Do Nash values have value? discussion and alternate proposals. *Hydrological Processes* **22**, 2723–2725.

Dawson, C. W., Abrahart, R. J. & See, L. M. 2007 Hydrotest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling and Software* **22**, 1034–1052.

Dawson, C. W., Mount, N. J., Abrahart, R. J. & Shamseldin, A. Y. 2012 Ideal point error for model assessment in data-driven river flow forecasting. *Hydrology and Earth System Sciences* **16**, 3049–3060.

Euser, T., Winsemius, H., Hrachowitz, M., Fenicia, F., Uhlenbrook, S. & Savenije, H. 2013 A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences* **17**, 1893–1912.

Fenicia, F., Kavetski, D. & Savenije, H. H. 2011 Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research* **47** (11), W11510.

Fenicia, F., Kavetski, D., Savenije, H. H., Clark, M. P., Schoups, G., Pfister, L. & Freer, J. 2014 Catchment properties, function, and conceptual model representation: is there a correspondence? *Hydrological Processes* **28** (4), 2451–2467.

Goldberg, D. E. 1994 Genetic and evolutionary algorithms come of age. *Communications of the ACM* **37**, 113–120.

Gupta, H. V., Kling, H., Yilmaz, K. K. & Martinez, G. F. 2009 Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology* **377**, 80–91.

Havlíček, V., Hanel, M., Máca, P., Kuráž, M. & Pech, P. 2013 Incorporating basic hydrological concepts into genetic programming for rainfall-runoff forecasting. *Computing* **95**, 363–380.

Jeong, K.-S., Kim, D.-K., Whigham, P. & Joo, G.-J. 2003 Modelling microcystisaeruginosa bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach. *Ecological Modelling* **161**, 67–78.

Kavetski, D. & Fenicia, F. 2011 Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resources Research* **47**.

Khu, S. T., Liong, S. Y., Babovic, V., Madsen, H. & Muttil, N. 2001 Genetic programming and its application in real-time runoff forecasting. *Journal of the American Water Resources Association* **37**, 439–451.

Kizhisseri, A. S., Simmonds, D., Rafiq, Y. & Borthwick, M. 2006 An evolutionary computation approach to sediment transport modelling. In: *Coastal Dynamics 2005: State of the Practice, Proceedings of the 5th Conference*, Barcelona, Spain, pp. 1–14.

Kling, H., Fuchs, M. & Paulin, M. 2012 Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology* **424**, 264–277.

Koza, J. R. 1992 *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, Cambridge, MA, USA.

Krause, P., Boyle, D. & Bäse, F. 2005 Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences* **5**, 89–97.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35**, 233–241.

Ley, R., Hellebrand, H., Casper, M. C. & Fenicia, F. 2016 Comparing classical performance measures with signature indices derived from flow duration curves to assess model structures as tools for catchment classification. *Hydrology Research* **47**, 1–14.

Madsen, H. 2000 Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. *Journal of Hydrology* **235**, 276–288.

Mai, J., Cuntz, M., Shafii, M., Zink, M., Schäfer, D., Thober, S., Samaniego, L. & Tolson, B. 2016 Multi-objective vs. single-objective calibration of a hydrologic model using single- and multi-objective screening. In: *EGU General Assembly Conference Abstracts, EGU General Assembly*, Vienna, Austria, pp. 8997.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I – A discussion of principles. *Journal of Hydrology* **10**, 282–290.

Pechlivanidis, I., Jackson, B., McMillan, H. & Gupta, H. V. 2012 Using an informational entropy-based metric as a diagnostic of flow duration to drive model parameter identification. *Global Nest Journal* **14**, 325–334.

Pechlivanidis, I., Jackson, B., McMillan, H. & Gupta, H. 2014 Use of an entropy-based metric in multiobjective calibration to improve model performance. *Water Resources Research* **50**, 8066–8083.

Pinkus, A., Winitzki, S. & Niesen, J. 2012 The Yacas computer algebra system.

Price, K., Purucker, S. T., Kraemer, S. R. & Babendreier, J. E. 2012 Tradeoffs among watershed model calibration targets for parameter estimation. *Water Resources Research* **48** (10), W10542.

Rani, D. & Moreira, M. M. 2010 Simulation–optimization modeling: a survey and potential application in reservoir systems operation. *Water Resources Management* **24**, 1107–1138.

Savic, D. A. & Walters, G. A. 1997 Genetic algorithms for least-cost design of water distribution networks. *Journal of Water Resources Planning and Management* **123**, 67–77.

Shafii, M. & Tolson, B. A. 2015 Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resources Research* **51**, 3796–3814.

Sugawara, M. 1979 Automatic calibration of the tank model/ L'étalonnage automatique d'un modèle à cisterne. *Hydrological Sciences Journal* **24**, 375–388.

van Esse, W., Perrin, C., Booij, M., Augustijn, D., Fenicia, F., Kavetski, D. & Lobligeois, F. 2013 The influence of conceptual model structure on model performance: a comparative study for 237 French catchments. *Hydrology and Earth System Sciences* **17**, 4227–4239.

Vis, M., Knight, R., Pool, S., Wolfe, W. & Seibert, J. 2015 Model calibration criteria for estimating ecological flow characteristics. *Water* **7**, 2358–2381.

Vitolo, C. 2015 *Exploring Data Mining for Hydrological Modelling*. PhD thesis.

Vrugt, J. A., Ter Braak, C. J., Gupta, H. V. & Robinson, B. A. 2009 Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment* **23**, 1011–1026.

Wagener, T., Lees, M. & Wheater, H. 1999 *A Rainfall-Runoff Modelling Toolbox (RRMT) for Matlab–User Manual*. Imperial College, London, UK.

Yilmaz, K. K., Gupta, H. V. & Wagener, T. 2008 A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. *Water Resources Research* **44** (9), W09417.

Zechman, E. M. & Ranjithan, S. R. 2007 Evolutionary computation-based approach for model error correction and calibration. *Advances in Water Resources* **30**, 1360–1370.

First received 31 May 2017; accepted in revised form 4 August 2017. Available online 6 December 2017