www.cafetinnova.org

Indexed in
Scopus Compendex and Geobase Elsevier, Chemical Abstract Services-USA, Geo-Ref Information Services-USA, List B of Scientific Journals, Poland, Directory of Research Journals

International Journal of Earth Sciences and Engineering

April 2015, P.P.80-87

ISSN 0974-5904, Volume 08, No. 02

Implications of Inter-Relationships between Evaluation Criteria on Calibration of Hydrologic Models

M RANJIT KUMAR[1], T MEENAMBAL[1] AND V KUMAR[2]
[1]*Department of Civil Engineering, Govt. College of Technology, Coimbatore-641013, Tamil Nadu*
[2]*Department of Agricultural Engineering, Agricultural College & Research Institute, Madurai 625104*
Email: *kumarkncsaga@gmail.com*

Abstract: Evaluation criteria are used for assessing the relative performance of a model and for estimating the parameters of the model. As each criterion is distinctly oriented towards one characteristic of the data, calibration with a single criterion will give distorted values for parameters. This has prompted the modellers to switch to multi-criteria calibration, wherein more than one unrelated criteria will be considered for optimising the parameters. But objectivity is lacking in selection of evaluation criteria to constitute the right mix. Hence an analysis is carried out to find the unrelatedness between different criteria using correlation coefficients. As the study demands model-based data, Watershed Processes Simulation (WAPROS) model is applied to a real watershed to generate simulated data for various values of parameters. For each set of parameters, a new set of simulated values and a new set of criteria values are generated. Then correlations between different criteria are estimated. Based on the strength of correlation coefficients, the criteria are classified into three groups, each group consisting of highly correlated criteria among themselves. It is recommended that one criterion from each group may be selected to constitute the right mix of criteria, for use in multi-criteria calibration.

Keywords: *Evaluation criteria, interrelationship, correlated criteria, multi-criteria calibration, right mix*

1. Introduction

Watershed simulation models or hydrologic models are being developed for over five decades, each with the aim of offering a new face to the catalogue of models. Various simulation models are created by the developers keeping their ideas, logics and algorithms distinct, attributing a signature to their own models. The developer is at liberty to explore the vast domain of hydrology, making a set of rules and building a simulation model of his choice, as a mark of opening up his mind and sowing his ideas.

But the targeted user is fraught with so many models and confused with the utility or versatility of those models and pushed to a state of selection dilemma. The user choses few evaluation criteria that he deems good and fixes screening values for each criterion and select those models that stand out. The problem is in choosing the right criteria that suit the conditions the most.

Many evaluation criteria are used during validation to indicate the performance of the model. These are classified into statistical, error and efficiency criteria. But for calibration, all the criteria cannot be used, and only a selected few criteria are used, as a mix or in combination, to assist in multi-criteria approach. It is recommended to have three criteria for calibration [5]. Similarly aggregated metrics or integrated indices are proposed as combined criteria for use in optimization [7]. While selecting the mix of evaluation criteria for calibration, personal judgement and bias dominate [15] [27]. This often results in selection of like-criteria, having high correlations, in the mix diluting the effectiveness of multi-criteria calibration and parameter identification. Hence there arises a need to know which criteria are related and which are not related.

The analysis demands large sets of simulated channel flow data. Every set of simulated data is compared to the same set of observed data and one set of values for all evaluation criteria is obtained. This produces large sets of values for evaluation criteria which are used for estimating correlation. But the basic data shall have to be modelled hydrologic data. Use of random data or hypothetical data will distort or misrepresent correlations. Hence WAPROS model is applied to Ebbanad watershed for generating simulation data and compared with the observed data from the watershed, for estimating values of evaluation criteria.

In this paper an analysis is carried out to identify the inter-relationships between various criteria, using correlation coefficients [9] [15]. The correlation coefficients are estimated for intra-class criteria and inter-class criteria and presented as correlation matrices. Based on the values of correlation coefficients, the

criteria are placed in different groups, each comprising a set of criteria more strongly correlated among themselves. For multi-criteria calibration, it is recommended to select not more than one criterion from each group to constitute a right mix to have uncorrelated criteria [9]. The importance of formulating uncorrelated criteria is also emphasized [8].

2. Evaluation Criteria

The evaluation criterion is a numerical formulation devised to compare and evaluate the modelled and the observed series of data and assign a value as an index of comparison. It is variously known as calibration criterion, objective function, misfit function and goodness of fit criterion.

Among the different classes of criteria, the statistical criteria give the comparative status, the error functions indicate the deficiency or gap in performance and efficiency indices show how far the simulated values match the observed values [10]. Lower the better for error criteria and higher the better for efficiency criteria.

The values of criteria are again dependent on the time step of the model, the stringent to lenient values respectively for small to large time steps [19]. The value of an efficiency criterion for an hourly time step of a simulation model will be smaller than daily time step, which will be still smaller than monthly time step, due to accumulation of data or averaging. The value of an error criterion follows an inverse interpretation; its value for hourly time step will be larger than daily time step, which will be still larger than monthly time step, for the same reasons. This difference in criterion values due to time step difference does not indicate improvement in efficiency levels. Hence it makes obligatory on the part of every modeller to report the performance value along with the time step used in simulation.

Few researchers have suggested a set of criteria to be used for calibration; but these suggestions have been made based on their experience with calibration [1] [9] [16] and not based on any specific study. There is no consensus on the use of any reliable and consistent evaluation criterion for optimisation of parameters in the calibration runs [6] [19] [24], as well as on the indication value of a criterion to represent and judge the overall performance of the model. Evaluation is considered as under-investigated aspect of modelling research, and that the recommendations based on different criteria are frequently found contradictory [24].

The criteria most often used for optimisation include: the Nash Sutcliffe's Efficiency, the coefficient of determination, the mean square error, the root mean square error, etc. [14]. But these criteria are subjected to microscopic analysis and criticised for being non-representative of an ideal criterion to judge the performance of the model [11] [15] [16] and a series of new and innovative evaluation criteria have been formulated, such as: the Legates McCabe's Efficiency Index [16], the Kling Gupta's Efficiency Index (KGEα) [11], the Willmott's Index of Agreement [22], the Robinson's Coefficient of Agreement [21], the Lei Ji's Coefficient of Agreement [17], etc.

The new criteria of evaluation also suffer from the limitation of being untried, unused and unpopular, lacking published data for reference [19]. In the absence of specific guidance [7] [20] the modeller has no choice, except to trade-off between the criteria, according to his own judgement and experience. However one's obsession with the Nash Sutcliffe's Efficiency remains a pragmatic choice [18] [19], overriding the logic of rationality. Next comes the RMSE criterion for frequent use as a single objective function in automatic calibration [5] [18].

The best value of one criterion does not match with the best values of other criteria [18], causing a dilemma in deciding which criteria is to be chosen as the best [11]. The Pareto optimality may help to resolve this issue [14], but, as more objectives are included in the calibration, the set of Pareto optimal solutions tends to be impractically extended [8].

3. Selection of Criteria

The evaluation criteria are called as calibration criteria, when used for calibration. When automatic calibration is resorted to, the calibration criteria are used as objective functions, such that efficiency functions are maximised and the error functions are minimised in the optimisation procedure. While comparing and evaluating different evaluation criteria, few are labelled as unscientific [26], inappropriate [25], and misleading [23] and even criticised as unfit and not to be used [23]. The selection of evaluation criteria is also influenced by personal bias. Hence all the known and reported criteria are considered [3] [20] for the study. Of them the following are not adopted for the reasons explained against each: (i) the Mean Square Logarithmic Error (MSLE) is not suitable for zero stream discharges prevalent in ephemeral streams, as log (0) is invalid; (ii) the Coefficient of Persistence (CP) [12], though useful, is not bounded and produces results that are difficult to interpret, besides being more influenced by the number of data; and (iii) the Coefficient of Gain from Daily Mean (DG) [1] requires the average measured discharge data from the past years for each hour/day of the period for comparison, which are not available; and it is seldom used in practice.

For the purpose of study, 30 evaluation criteria are chosen at the rate of 10 for each class, and analysed for inter relationships. As the nomenclature and formulation of evaluation criteria are used ambiguously, the formulae for all evaluation criteria considered for analysis are given in Appendix 1.

4. Methodology Adopted for the Analysis

Under this study, WAPROS model is run with a set of values for six parameters for simulating hourly data for 94 days and for generating values for 30 evaluation criteria, against one set of observed data. One set of values of parameters gives one set of values for 30 evaluation criteria, representing one model performance level. This is repeated for 90 times with 90 sets of values for parameters to get 90 sets of values for 30 evaluation criteria against the fixed set of observed data. The 90 sets of values represent 90 levels of model performance. Each criterion will have 90 values for 90 levels of performance. Now the correlation coefficient between two criteria, each having a series of 90 values, is estimated. This is repeated for all combinations of pairs of criteria and correlation coefficients between all the chosen criteria are estimated. As 90 levels of performance are fixed, the values of correlation coefficients between the criteria indicate the direction and strength of association between them for 90 levels of performance in the same order.

Two identical criteria will have a correlation coefficient of +1.0. Two dissimilar criteria will have a correlation coefficient of –1.0, which indicates the two criteria are strongly but negatively correlated: in such cases treating a criterion with deduction from 1.0 (1.0 -x) or multiplication by –1 (–1*x) will make the correlation positive and strong. Hence those criteria whose correlation coefficients are greater than ± 0.90 are considered as more correlated.

For the convenience of interpretation, the values of correlation coefficients are assigned to the concerned criteria under the three classes: statistical, error and efficiency criteria, each class having 10 criteria. The correlations among the statistical criteria are called as intra-class (statistical) correlation coefficients. Similarly for error and efficiency criteria classes, there will be intra-class (error) and intra-class (efficiency) correlation coefficients. These are represented in the form of correlation matrices (10x10). Here the diagonal values (10) represent the self-correlation (one against the same) and the upper diagonal values (45) will be equal to the lower diagonal values (45). Hence for three classes, there will be a total number of 135 (45*3) intra-class correlation coefficients.

The correlation between a criterion from one class and another criterion from another class is called as inter class correlation coefficient. There will be three combinations of inter class correlation coefficients as: statistical-error, error-efficiency and efficiency-statistical inter class correlation coefficients. When these are represented in the form of correlation matrices, there will be three inter class correlation matrices, each having 100 (10x10) correlation coefficients. For three inter class correlation matrices, there will be 300 (3x100) correlation coefficients. Altogether the analysis will have 435 correlation coefficients, which will be sufficient to assess the relatedness between the evaluation criteria.

5. WAPROS Model and Application

WAPROS is specifically developed for simulating hydrologic processes in small and medium sized watersheds. It is a new lumped, deterministic hourly model exclusively for simulating hydrologic processes and water balance in watersheds of size from 100 to 10000 ha to suit Indian conditions. As the space and time scales objectivized for WAPROS are finer, the model shall have to be more sensitive than that for a large basin and monthly simulation.

5.1. WAPROS Model Development

The model is developed to simulate 15 hydrologic processes, with 10 hydrologic storages. The model simulates hydrologic storage positions and process values on hourly basis and integrates hourly data into daily data for use in watersheds where only hourly rainfall and daily channel flow data are available. The model also synthesizes elemental processes into lumped processes usable for water balance. The model generates two water balance equations, closure errors for 10 storages and water balance ratios.

5.2. Application of the Model

The model is applied to a real watershed, called ‘Ebbanad’, which is located in the Nilgiris district of Tamil Nadu State, India. The centroid of the watershed is at 11^0 26’ 15”N. A Silt monitoring station, located in the drain point of the watershed and equipped with continuous hydrologic monitoring with logger recordings, is the data source.

5.3. Watershed Characteristics

Ebbanad is a mountainous watershed in humid agro-climatic region, with a mean elevation of 2084.0 m above MSL. The stream originates from the Doddabetta peak and joins the Moyar River. The total area of the watershed is 3582.0 hectares, with a drainage density of 2.904 km per sq. km. The land use pattern in the watershed is: area under forest: 1722 ha; area under agricultural crops including tea plantations: 1797 ha; and impervious area under rocks, habitations and roads: 63 ha. The average longitudinal slope of the watershed

is 7.01 % and the average cross sectional slope is 32.52 %. The weighted average of the soil constituents are estimated as: sand: 55.01 %; silt: 17.40 %; clay: 27.59 %; organic matter: 1.45 %; and coarse fragments: 1.23 %.

5.4. Model Simulation Results

The simulated channel flows from WAPROS are compared to the observed flows and the values of criteria, with respect to hourly and daily data are: Nash-Sutcliffe's Efficiency (NSE) = 0.8588; 0.9029; Volume Handling Efficiency (VHE) = 0.9409; 0.9409; Mean Square Error (MSE) = 0.4030; 0.2413; Ratio of RMSE to Standard Deviation of Observed flow (RSR) = 0.3758; 0.3117; and Coefficient of determination: (r^2) = 0.8623; 0.9073 (Formulae in Appendix 1).

Simulation results with a coefficient of determination < 0.6 or VDE > ± 10% are generally considered too poor to be acceptable [4]. It is also recommended that both the NSE and correlation coefficient are to be > 0.8 for an acceptable calibration for monthly stream flow [2]. The following values of evaluation criteria are recommended for monthly simulation data: (i) NSE: > 0.75 for very good; 0.75 - 0.65 for good and 0.65 - 0.50 for satisfactory ratings; (ii) ME (Bias) (±): < 0.10 for very good; 0.10 - 0.15 for good and 0.15 - 0.25 for satisfactory ratings; and (iii) RSR: < 0.50 for very good; 0.50 - 0.60 for good and 0.60 - 0.70 for satisfactory ratings [19]. It could be seen that the performance of the model for hourly and daily data surpass the ratings recommended for monthly data, suggesting a 'very good' rating for WAPROS model.

6. Results and Discussion

As discussed before, 435 correlation coefficients are estimated for the study and these are presented in the form of correlation matrices, as: Intra-class correlation coefficients between Statistical criteria (Table 1), Intra-class correlation coefficients between Error criteria (Table 2) and Intra-class correlation coefficients between Efficiency criteria (Table 3); Inter-class correlation coefficients between Error criteria and Statistical criteria (Table 4), Inter-class correlation coefficients between Efficiency criteria and Statistical criteria (Table 5) and Inter-class correlation coefficients between Efficiency criteria and Error criteria (Table 6).

The correlation coefficient values are furnished in the tables as above said, in the popular format of correlation matrix. The objective of the analysis is to identify and group those criteria which are closely related to each other, and to recommend use of only one criterion from one group to assist in effective multi-criteria calibration.

For better presentation and interpretation, the following codes are used for evaluation criteria: S for Statistical criteria, E for Error criteria and EF for Efficiency criteria classes. The prefix H stands for Hourly data. For clarity, the evaluation criteria are given both Evaluation codes (from H1 to H30) and Criteria codes (from S1 to S10, from E1 to E10 and from EF1 to EF10).

From Table: 1, it has been noticed that S1 stands out as distinct and loner; S2 is highly and positively correlated with S3, S6, S8 and S9; and S2 is highly and negatively correlated with S7 and S10; S3 is highly and positively correlated with S6, S8 and S9; and S3 highly and negatively correlated with S7 and S10; S4 and S5 are highly and positively correlated with each other; S6 is highly and positively correlated with S2, S3, S8 and S9; and S6 is highly and negatively correlated with S7 and S10; S7 is highly and positively correlated with S10; and S7 is highly and negatively correlated with S2, S3, S6, S8 and S9; S8 is highly and positively correlated with S2, S3, S6 and S9; and S8 is highly and negatively correlated with S7 and S10; S9 is highly and positively correlated with S2, S3, S6 and S8; and S9 is highly and negatively correlated with S7 and S10; S10 is highly and positively correlated with S7; and S10 is highly and negatively correlated with S2, S3, S6, S8 and S9.

From the preceding discussions, the following results can be summarized: (i) S1 is not correlated to others; (ii) S4 and S5 are related to each other and not related to others; and (iii) S2, S3, S6, S8, S9 and S10 are highly positively and negatively correlated. This summary can be presented in the following groups:

Group A: Ratio of Means;

Group B: Correlation coefficient and Coefficient of determination;

Group C: Ratio of Standard Deviations, Ratio of Coefficient of Variations, Covariance between Sim and Obs Flow, Regression Line Slope: (Sim on Obs), Regression Line Slope: (Obs on Sim), Regression Line Intercept: (Sim on Obs) and Regression Line Intercept: (Obs on Sim).

Applying the procedure explained above, the results in respect of other tables will be restricted to presenting summarized groupings, without repetition of the logical procedures.

From Table: 2, the following groupings are suggested:

Group D: Mean Error and Volume Deviation Error;

Group E: Mean Absolute Error, Mean Absolute Relative Error, Mean Square Error, Root Mean Square Error, Normalised RMSE, Relative RMSE, Fourth Power RMSE and Transformed RMSE;

From Table: 3, the following groupings are suggested:

Group F: Volume Handling Efficiency;

Group G: Volume-Time Matching Efficiency, Nash-Sutcliffe's Efficiency, Relative Nash-Sutcliffe's Efficiency, Legates-McCabe's Efficiency Index and Kling-Gupta's Efficiency Index (γ);

Group H: Willmott's Index of Agreement (d), Willmott's Modified Index (d1), Robinson's Coefficient of Agreement and Lei Ji's Coefficient of Agreement;

From Table: 4, the following groupings from inter correlations are suggested:

Group I: Mean Error, Volume Deviation Error and Ratio of Means;

Group J: Mean Absolute Error; Mean Absolute Relative Error, Transformed RMSE, Ratio of Standard Deviations and Ratio of Coefficient of Variations.

From Table: 5, the following groupings are suggested:

Group K: Volume Handling Efficiency and Ratio of Means;

Group L: Willmott's Modified Index (d1);

Group M: Volume-Time Matching Efficiency, Nash-Sutcliffe's Efficiency, Relative Nash-Sutcliffe's Efficiency, Legates-McCabe's Efficiency Index, Ratio of Standard Deviations and Ratio of Coefficient of Variations;

Group N: Willmott's Index of Agreement (d), Robinson's Coefficient of Agreement and Lei Ji's Coefficient of Agreement, correlation coefficient and Coefficient of determination.

From Table: 6, the following groupings are suggested:

Group O: Volume Handling Efficiency, Mean Error and Volume Deviation Error;

Group P: Volume-Time Matching Efficiency, Nash-Sutcliffe's Efficiency, Relative Nash-Sutcliffe's Efficiency, Legates-McCabe's Efficiency Index and Kling-Gupta's Efficiency Index (γ), Willmott's Modified Index (d1), Mean Absolute Error, Mean Absolute Relative Error, Mean Square Error, Root Mean Square Error, Normalised RMSE, Relative RMSE, Fourth Power RMSE and Transformed RMSE;

Table-1 *Intra-correlation matrix for statistical criteria*

S. No	Evaluation criteria	Stat. Code	H1 S1	H2 S2	H3 S3	H4 S4	H5 S5	H6 S6	H7 S7	H8 S8	H9 S9	H10 S10
1	Ratio of Means	S1	1.00									
2	Ratio of Standard Deviations	S2	-0.18	1.00								
3	Ratio of Coefficient of Variations	S3	-0.27	1.00	1.00							
4	Correlation Coefficient:(r)	S4	0.07	-0.59	-0.59	1.00						
5	Coefficient of Determination:(r^2)	S5	0.07	-0.58	-0.58	1.00	1.00					
6	Covariance bet. Sim and Obs Flow	S6	-0.18	0.99	0.99	-0.49	-0.48	1.00				
7	Reg. Line Slope:(Sim on Obs)	S7	0.11	-0.93	-0.92	0.32	0.31	-0.95	1.00			
8	Reg. Line Slope:(Obs on Sim)	S8	-0.18	0.99	0.99	-0.49	-0.48	1.00	-0.95	1.00		
9	Reg. Line Intercept:(Sim on Obs)	S9	-0.28	0.93	0.93	-0.31	-0.30	0.95	-0.98	0.95	1.00	
10	Reg. Line Intercept:(Obs on Sim)	S10	0.30	-0.98	-0.99	0.48	0.47	-0.99	0.94	-0.99	-0.95	1.00

Table-2 *Intra-correlation matrix for error criteria*

S. No	Evaluation criteria	Err. code	H11 E1	H12 E2	H13 E3	H14 E4	H15 E5	H16 E6	H17 E7	H18 E8	H19 E9	H20 E10
1	Mean Error (ME)	E1	1.00									
2	Mean Absolute Error (MAE)	E2	-0.21	1.00								
3	Mean Absolute Relative Error	E3	-0.22	0.97	1.00							
4	Volume Deviation Error (VDE)	E4	1.00	-0.21	-0.22	1.00						
5	Mean Square Error (MSE)	E5	-0.22	0.97	0.90	-0.22	1.00					
6	Root Mean Square Error:(RMSE)	E6	-0.23	0.98	0.91	-0.23	0.99	1.00				
7	Normalised RMSE:(Max-Min Obs)	E7	-0.23	0.98	0.91	-0.23	0.99	1.00	1.00			
8	Relative RMSE:(SD Obs):(RSR)	E8	-0.23	0.98	0.91	-0.23	0.99	1.00	1.00	1.00		
9	Fourth Power RMSE:(R4MS4E)	E9	-0.25	0.96	0.88	-0.25	0.99	0.99	0.99	0.99	1.00	
10	Transformed RMSE:(Box-Cox)	E10	-0.21	1.00	0.97	-0.21	0.96	0.98	0.98	0.98	0.95	1.00

Table-3 *Intra-correlation matrix for efficiency criteria*

S. No	Evaluation criteria	Eff. Code	H21 EF1	H22 EF2	H23 EF3	H24 EF4	H25 EF5	H26 EF6	H27 EF7	H28 EF8	H29 EF9	H30 EF10
1	Volume-Time Matching Efficiency	EF1	1.00									
2	Volume Handling Efficiency	EF2	0.22	1.00								
3	Nash-Sutcliffe's Efficiency	EF3	0.97	0.22	1.00							
4	Relative Nash-Sutcliffe's Efficiency	EF4	0.99	0.21	0.95	1.00						
5	Legates-McCabe's Efficiency Index	EF5	1.00	0.22	0.97	0.99	1.00					
6	Kling-Gupta's Efficiency Index (γ)	EF6	0.99	0.27	0.96	0.96	0.99	1.00				
7	Willmott's Index of Agreement (d)	EF7	0.88	0.20	0.90	0.85	0.88	0.91	1.00			
8	Willmott's Modified Index (d1)	EF8	0.95	0.19	0.90	0.93	0.95	0.95	0.96	1.00		
9	Robinson's Coeff. of Agreement	EF9	0.88	0.21	0.90	0.85	0.88	0.91	1.00	0.96	1.00	
10	Lei Ji's Coefficient of Agreement	EF10	0.89	0.16	0.91	0.87	0.89	0.91	1.00	0.96	1.00	1.00

Table-4 *Inter-correlation matrix- error criteria vs. statistical criteria*

S. No	Evaluation criteria	Err. Code	H1 S1	H2 S2	H3 S3	H4 S4	H5 S5	H6 S6	H7 S7	H8 S8	H9 S9	H10 S10
1	Mean Error	E1	1.00	-0.18	-0.27	0.07	0.07	-0.18	0.11	-0.18	-0.28	0.30
2	Mean Absolute Error	E2	-0.21	0.93	0.93	-0.79	-0.78	0.90	-0.75	0.90	0.74	-0.90
3	Mean Absolute Relative Error	E3	-0.22	0.92	0.92	-0.68	-0.68	0.90	-0.76	0.90	0.76	-0.90
4	Volume Deviation Error	E4	1.00	-0.18	-0.27	0.07	0.07	-0.18	0.11	-0.18	-0.28	0.30
5	Mean Square Error	E5	-0.22	0.90	0.91	-0.87	-0.87	0.84	-0.71	0.84	0.71	-0.84
6	Root Mean Square Error	E6	-0.23	0.89	0.90	-0.86	-0.86	0.84	-0.68	0.84	0.69	-0.84
7	Normalised RMSE	E7	-0.23	0.89	0.90	-0.86	-0.86	0.84	-0.68	0.84	0.69	-0.84
8	Relative RMSE (RSR)	E8	-0.23	0.89	0.90	-0.86	-0.86	0.84	-0.68	0.84	0.69	-0.84
9	Fourth Power RMSE	E9	-0.25	0.88	0.89	-0.87	-0.87	0.82	-0.67	0.82	0.68	-0.83
10	Transformed RMSE	E10	-0.21	0.92	0.92	-0.78	-0.78	0.89	-0.73	0.89	0.73	-0.89

Table-5 *Inter-correlation matrix- efficiency criteria vs. statistical criteria*

S.No	Evaluation criteria	Eff. Code	H1 S1	H2 S2	H3 S3	H4 S4	H5 S5	H6 S6	H7 S7	H8 S8	H9 S9	H10 S10
1	Volume-Time Matching Efficiency	EF1	0.21	-0.93	-0.93	0.79	0.78	-0.90	0.75	-0.90	-0.74	0.90
2	Volume Handling Efficiency	EF2	0.99	-0.19	-0.28	0.09	0.08	-0.18	0.12	-0.18	-0.29	0.30
3	Nash-Sutcliffe's Efficiency	EF3	0.22	-0.90	-0.91	0.87	0.87	-0.84	0.71	-0.84	-0.71	0.84
4	Relative Nash-Sutcliffe's Efficiency	EF4	0.20	-0.93	-0.93	0.77	0.77	-0.89	0.76	-0.89	-0.76	0.89
5	Legates-McCabe's Efficiency Index	EF5	0.21	-0.93	-0.93	0.79	0.78	-0.90	0.75	-0.90	-0.74	0.90
6	Kling-Gupta's Efficiency Index (γ)	EF6	0.26	-0.90	-0.91	0.80	0.80	-0.86	0.68	-0.86	-0.69	0.86
7	Willmott's Index of Agreement (d)	EF7	0.20	-0.67	-0.68	0.94	0.94	-0.59	0.37	-0.59	-0.37	0.60
8	Willmott's Modified Index (d1)	EF8	0.18	-0.77	-0.78	0.85	0.85	-0.72	0.50	-0.72	-0.50	0.72
9	Robinson's Coeff. of Agreement	EF9	0.20	-0.66	-0.68	0.94	0.94	-0.59	0.36	-0.59	-0.37	0.60
10	Lei Ji's Coefficient of Agreement	EF10	0.16	-0.68	-0.69	0.95	0.95	-0.61	0.39	-0.61	-0.39	0.61

Table-6 *Inter-correlation matrix- efficiency criteria vs. error criteria*

S. No	Evaluation criteria	Eff. Code	H11 E1	H12 E2	H13 E3	H14 E4	H15 E5	H16 E6	H17 E7	H18 E8	H19 E9	H20 E10
1	Volume-Time Matching Efficiency	EF1	0.21	-1.00	-0.97	0.21	-0.97	-0.98	-0.98	-0.98	-0.96	-1.00
2	Volume Handling Efficiency	EF2	0.99	-0.22	-0.23	0.99	-0.22	-0.23	-0.23	-0.23	-0.25	-0.22
3	Nash-Sutcliffe's Efficiency	EF3	0.22	-0.97	-0.90	0.22	-1.00	-0.99	-0.99	-0.99	-0.99	-0.96
4	Relative Nash-Sutcliffe's Efficiency	EF4	0.20	-0.99	-0.99	0.20	-0.95	-0.96	-0.96	-0.96	-0.93	-0.99
5	Legates-McCabe's Efficiency Index	EF5	0.21	-1.00	-0.97	0.21	-0.97	-0.98	-0.98	-0.98	-0.96	-1.00
6	Kling-Gupta's Efficiency Index (γ)	EF6	0.26	-0.99	-0.93	0.26	-0.96	-0.99	-0.99	-0.99	-0.97	-0.99
7	Willmott's Index of Agreement (d)	EF7	0.20	-0.88	-0.80	0.20	-0.90	-0.93	-0.93	-0.93	-0.92	-0.88
8	Willmott's Modified Index (d1)	EF8	0.18	-0.95	-0.92	0.18	-0.90	-0.94	-0.94	-0.94	-0.91	-0.95
9	Robinson's Coeff. Of Agreement	EF9	0.20	-0.88	-0.80	0.20	-0.90	-0.92	-0.92	-0.92	-0.92	-0.88
10	Lei Ji's Coefficient of Agreement	EF10	0.16	-0.89	-0.81	0.16	-0.91	-0.93	-0.93	-0.93	-0.92	-0.89

Group Q: Willmott's Index of Agreement (d), Robinson's Coefficient of Agreement, Lei Ji's Coefficient of Agreement, Root Mean Square Error, Normalised RMSE, Relative RMSE and Fourth Power RMSE.

Now the inferences drawn from tables 1 to 6 are combined to give consolidated results; while doing so the intra-class correlation coefficients are also considered, even though direct correlation coefficients between criteria are not strong enough:

Group I: (Four criteria) Ratio of Means, Mean Error, Volume Deviation Error and Volume Handling Efficiency;

Group II: (Five criteria) Correlation coefficient, Coefficient of determination, Willmott's Index of Agreement (d), Robinson's Coefficient of Agreement and Lei Ji's Coefficient of Agreement;

Group III: (Twenty one criteria) Ratio of Standard Deviations, Ratio of Coefficient of Variations, Covariance between Sim and Obs Flow, Regression Line Slope: (Sim on Obs), Regression Line Slope: (Obs on Sim), Regression Line Intercept: (Sim on Obs) and Regression Line Intercept: (Obs on Sim); Mean Absolute Error, Mean Absolute Relative Error, Mean Square Error, Root Mean Square Error, Normalised RMSE, Relative RMSE, Fourth Power RMSE and Transformed RMSE; Volume-Time Matching Efficiency, Nash-Sutcliffe's Efficiency, Relative Nash-Sutcliffe's Efficiency, Legates-McCabe's Efficiency Index, Kling-Gupta's Efficiency Index (γ) and Willmott's Modified Index (d1).

From the above analysis, it is recommended that only one criterion from a group shall be considered for optimization under multi-criteria calibration; and considering more than one criteria from the same group amounts to repetition of the same. However it is emphasized that the criteria under a group are only correlated, and not equivalent; such criteria can continue to serve the purpose for which those are devised.

It may also be interesting to notice that the Kling-Gupta's Efficiency Index (KGE1), formulated as triple component criterion [11], consists of Ratio of Means (β), Ratio of Standard Deviations (α) and correlation coefficient (r); it is also known as KGEα.

It is later revised as KGE2 (or KGEγ) [13], by substituting Ratio of Coefficient of Variations (γ) in place of Ratio of Standard Deviations (α). The formulations are described below:

$$(I)\ ED_1 = \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

$$r = \text{Correlation Coefficien t};$$

$$\alpha = \left(\frac{\sigma_x}{\sigma_y}\right); \quad \beta = \left(\frac{\overline{X}}{\overline{Y}}\right);$$

$$\text{KGE}_1 = \text{KGE}_\alpha = (1.0 - \text{ED}_1)$$

$$(II)\ ED_2 = \sqrt{(r-1)^2 + (\gamma-1)^2 + (\beta-1)^2}$$

$$\gamma = \left(\frac{CV_x}{CV_y}\right);$$

$$KGE_2 = KGE_\gamma = (1.0 - ED_2)$$

It could be seen that the three components selected for KGE formulations are from different groups, i.e. respectively from Group I, III and II; and substitution of one criterion by another criterion is also drawn from the same Group III. Though KGEs have been devised after decomposition analysis of MSE and NSE, this exemplifies the logical utility of correlation studies now undertaken. In this paper, KGEγ version is used for analysis.

From the published literature and review, it could be seen that the researchers used both right and wrong combinations of evaluation criteria for calibration [8] [9] [20].

When different independent criteria are combined to form an aggregated or integrated metric, sometimes they lose their individual characters and become an entirely new different criterion. Hence while devising a new aggregated metric, its characteristics need to be thoroughly studied.

Based on the values of correlation coefficients between different evaluation criteria, the criteria are brought under three groups. Each group now consists of closely correlated criteria, which are considered to be effective substitutes among them, for the purpose of being used as objective functions in multi-criteria calibration. This implies that, as of now, three criteria, at the rate of one from each group, will be sufficient for calibration. This recommendation will also equally apply while devising aggregated indices for calibration.

7. Conclusion

While selecting the mix of evaluation criteria for multi-criteria calibration, personal judgement and bias dominate, resulting in selection of like-criteria having high correlations. This results in sub-optimal calibration and non-optimal set of parameters. An objective analysis is made to identify the inter-relationships between various criteria, using correlation coefficients. Based on the values of correlation coefficients, all the criteria are brought under three groups, each comprising a set of criteria more strongly correlated among them. The intra-group criteria are considered to be effective substitutes among themselves. For multi-criteria

calibration, it is recommended to select not more than one criterion from each group and selecting more than one criterion from the same group amounts to using inter-correlated criteria. This study and the grouping of criteria will pave way for objective selection of evaluation criteria for calibration. The results of the study will also be useful in devising the aggregated indices for calibration and new indices for evaluation.

References

[1] ASCE. Criteria for evaluation of watershed models. *J. Irrigation Drainage Eng*. 119(3): 429-442. 1993.

[2] Bari, M. A. and K. R. J. Smettem. A conceptual model of daily water balance following partial clearing from forest to pasture. *Hydrol. Earth Syst. Sci*. 10, 321–337. 2006.

[3] Bellocchi, Gianni, Mike Rivington, Marcello Donatelli and Keith Matthews. Validation of biophysical models: issues and methodologies. A review. *Agron. Sustain. Dev*. 1-22. 2009.

[4] Boughton, W. Catchment Water Balance Modelling in Australia 1960-2004. *Agricultural Water Management.* Vol. 71, No 2, 91-116. 2005.

[5] Boyle, Douglas, P., Hoshin V. Gupta and Soroosh Sorooshian. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resour. Res*. Vol. 36, No. 12: 3663-3674. 2000.

[6] Dawson, C.W., R.J. Abrahart, and L.M See. *HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts*. http:// creative commons. org/licenses/by-nc-nd/2.5/. pp 1-68. 2005.

[7] Dawson, C. W., R. J. Abrahart, A. Y. Shamseldin, and N. J. Mount. Ideal point error for model assessment. *Hydrol. Earth Syst. Sci.* Discuss. 9: 1671–1698. 2012.

[8] Efstratiadis, A. and D. Koutsoyiannis. One decade of multi-objective calibration approaches in hydrological modelling: a review. *Hydrol. Sci. J.* 55(1): 58–78. 2010.

[9] Gupta, Hoshin V., Soroosh Sorooshian and Patrice Ogou Yapo. Toward improved calibration of hydrologic models: Multiple and non-commensurable measures of information. *Water Resour. Res*. Vol. 34, No. 4: 751-763. 1998.

[10] Gupta, Hoshin V., Thorsten Wagener and Yuqiong Liu. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrol. Process*. 1-12. 2008.

[11] Gupta, Hoshin V., Harald Kling, Koray, K. Yilmaz and Guillermo F. Martinez. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*. 377: 80–91. 2009.

[12] Kitanidis, P.K. and R.L. Bras. Real-time forecasting with a conceptual hydrologic model: 2. Application and results, *Water Resour. Res*. Vol. 16(6): 1034 – 1044. 1980.

[13] Kling, H., M. Fuchs, and M. Paulin. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology.* Volumes 424-425: 264-277. 2012.

[14] Kim, S. M., B. L. Benham, K. M. Brannan, R. W. Zeckoski, and J. Doherty. Comparison of hydrologic calibration of HSPF using automatic and manual methods. *Water Resour. Res*. 43: 1-12. 2007.

[15] Krause, P., D. P. Boyle, and F. Base. Comparison of different efficiency criteria for hydrological model Assessment. *Advances in Geosciences.* 5: 89–97. 2005.

[16] Legates, David R. and Gregory J. McCabe Jr. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydro-climatic model validation. *Water Resour. Res*. Vol. 35, No. 1: 233–241. 1999.

[17] Lei Ji and K. Gallo. An agreement coefficient for image comparison. *Photogrammetric Engineer ing & Remote Sensing.* Vol. 72, No. 7: 823–833. 2006.

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

[19] Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, T. L. Veith. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*. Vol. 50(3): 885-900. 2007.

[20] Moriasi, D. N., B. N. Wilson, K. R. Douglas-Mankin, J. G. Arnold, P. H. Gowda. Hydrologic and water quality models: use, calibration and validation. *Transactions of the ASABE*. Vol. 55(4): 1241-1247. 2012.

[21] Robinson, W.S. The statistical measurement of agreement. *American Sociological Review*. Vol. 22. No.1: 17-25. 1957.

[22] Willmott, C. J. On the validation of models. *Physical Geography*. 2, 2: 184-194. 1981.

[23] Willmott, C. J. Some comments on the evaluation of model performance. *Bulletin American Meteorological Society*. Vol. 63: No. 11: 1309-1313. 1982.

[24] Willmott, C. J., S.G. Ackleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R. Legates, James O' Donnell, and C.M. Row. Statistics for the evaluation and comparison of models. *Journal of Geophysical Research*. Vol. 90, No. C5. 8995-9005. 1985.

[25] Willmott, C. J. and Kenji Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res*. Vol. 30: 79–82. 2005.