ANALYSING FOR MEASURE CHARACTERISTICS OF NASH-SUTCLIFFE EFFICIENCY TYPE INDICATORS USED FOR MODEL VALIDATION

Kai-Bin Zhao [(a)], Ke Fang[(b)], Ming Yang[(c)]

Control & Simulation Center, School of Astronautics, Harbin Institute of Technology, Harbin, China

[(a)]kaibin.zhao.HIT@hotmail.com, [(b)]hitsim@163.com, [(c)]myang@hit.edu.cn

ABSTRACT

The validation of simulation model is commonly quantified by various goodness-of-fit indicators. The selection and use of specific goodness-of-fit indicators and the interpretation of their results can be a challenge for even the most experienced evaluators since each indicator may put different emphasis on different types of simulated and observed behaviours. In this paper, first of all, two goodness-of-fit indicators calculated on weighted errors are developed, presented and recommened, which are more suitable for evaluating simulation model compared with high and low magnitude within observed data respectively. Then, the specific measure characteristics (that is, which part of the simulated and observed datasets most influences the various indicators) of several Nash-Sutcliffe efficiency (*NSE*) type indicators is analyzed and compared through two hypothetical examples using a simple observed and simulated datasets. Finally, some misunderstanding and ambiguity is corrected and clarified for measure characteristics of *NSE* type indicators.

Keywords: methods of quantitative validation, Nash-Sutcliffe efficiency, comparison between indicators, measure characteristics

1. INTRODUCTION

Krause, Boyle, and Bäse (2005) put forward three main reasons for model validation requirement: (1) to provide a process of quantitative indicator of model's ability to replicating past and predicting future behavior; (2) to provide a means for evaluating improvements to the model, such as through adjustment of model parameter values and model structural modifications, and the important characteristics expression changes of model behavior in space and time; (3) to compare modeling efforts in model development life cycle process.

Model validation deals with the issue of whether or not a model is a sufficient accurate representation of real system for the intended applications of the model (Martens, Put, and Kerre 2006). From the perspective of practicality and operability, a commonly pragmatic approach is comparison of simulated data (modeled data, predicted data, or model output data) with observed data (reference data, measured data, experimental data, test data, expected data, or real-word data) to determine the degree of behavioral similarity between simulation model and real system. This process of model performance evaluation may appear the following situations (Pushpalatha, Perrin, and Le Moine 2012): (1) data set used for model validation vary span several orders of magnitude that may not be equally important for the modeller who may be interested in agreement between the part of observed and simulated graph; (2) the model may be used for diferent applications, which may require specific criteria (e.g. either high or low magnitude simulation studies); (3) the variance of error varies significantly throughout validation period with space and time, instead of a nearly constant, which often leads to emphasize larger errors within a wider range of the observations while smaller errors within a small range tend to be neglected.

For these situations, a large variety of goodness-of-fit indicators have been proposed and used over the years in the field of model validation, as shown for example by the lists of indicators given by Bennett, Croke, and Guariso (2013); Hauduc, Neumann, and Muschalla (2015); and Crochemore, Perrin, and Andreassian (2015). Among these indicators, the Nash and Sutcliffe (1970) efficiency (*NSE*) type indicators has received considerable attention from many evaluators, especially in the area of environmental science, due to its flexibility to be applied to various types of simulation models (Oudin, Andreassian, and Mathevet 2006; Nicolle, Pushpalatha, and Perrin 2014). Modifications of the *NSE* have been proposed by several authors, including those based on transformed variables, others using relative or absolute instead of squared errors, and those adopting benchmark series different from the mean of observations. Until by now, the focus of discussion work with the existing *NSE* type indicators is primarily on their interrelationship and measure characteristics, which have gained some research achievements that can be found in the literature (i.e. Krause, Boyle, and Bäse 2005; Oudin, Andreassian, and Mathevet 2006; Dawson, Abrahart, and See 2007; Pushpalatha, Perrin, and Le Moine 2012; Hauduc, Neumann, and Muschalla 2015).

Although there are a lot of pre-analysis for *NSE* type indicators, but there are still some limitations. The

ISBN 978-88-97999-85-0; Affenzeller, Bruzzone, Jiménez, Longo and Piera Eds.

present work intends to complement previous studies, and this paper has two main objectives: first, in order to effectively select and use different *NSE* type indicator and interpret their results, we wish to discuss further specific measure characteristics of each indicator，that is, which part of the $\{O_i, S_i\}$ most influences the various indicators; Second, we wish to propose two goodness-of-fit indicators suited for the evaluation of high-magnitude and low-magnitude simulations using actually observed data.

The remainder of this paper is arranged as follows: Section 2 describes different *NSE* type indicators and our proposed goodness-of-fit indicators on weighted errors; Section 3 provide a hypothetical example used for analysing measure characteristics of various indicators; Results and discussion are presented in Section 4, while conclusions are drawn in Section 5.

2. GOODNESS-OF-FIT INDICATORS

A number of goodness-of-fit indicators have been proposed in the simulation literature, however, there is no consensus on the best approach. In this section, the goodness-of-fit indicators included in the study are provided. These are the seven indicators: Nash-Sutcliffe efficiency and some of its modified forms: Legates-McCabe efficiency (LME), NSE on squared transformed series, NSE on root-squared transformed series, NSE on logarithmic transformed series, NSE on inverse transformed series, LME on deviations weighted by observations cubed, LME on deviations weighted by observations to the power of minus three.

These indicators above are generally viewed as different normalized forms of different error measures calculated on point-by-point concurrently and are generally used in order to make decisions about the validity of the simulation model.

In the following equations, $\boldsymbol{O}$ represents the sample containing the observations obtained from real system, but could more generally containing the expected values, $\boldsymbol{O} = (O_1, O_2, \ldots, O_n)^{\mathrm{T}} \in \mathrm{R}^n$; $\boldsymbol{S}$ represents the sample containing the model simulated outputs, $\boldsymbol{S} = (S_1, S_2, \ldots, S_n)^{\mathrm{T}} \in \mathrm{R}^n$; O_i and S_i are respectively the ith observed and simulated value; The subscript i is data index (usually referring to time, but could more generally indicate spatial location or some other kind of index); n is is total number of pairwise-matched observed and simulated data used for validation; $\overline{O_i}$ is the mean of the observations.

2.1. Nash-Sutcliffe Efficiency

The *NSE* suggested by Nash and Sutcliffe (1970) is expressed as one minus the mean of the squared errors between the observed and simulated datasets normalized by the variance of the observed values during the validation period, which is the most widely attention goodness-of-fit indicator in the model validation fields due to its flexibility. It is calculated as:

$$NSE = 1 - \frac{\sum_{i=1}^{n} (O_i - S_i)^2}{\sum_{i=1}^{n} (O_i - \overline{O_i})^2}, \quad -\infty < NSE \le 1 \tag{1}$$

The *NSE* ranges from minus infinity to one: an *NSE* value of one implies that observed value are in complete agreement with simulated value for all subscript i, which is generally viewed as model's ability to reproduce and predict system behaviour perfectly; *NSE* values between zero and one implies that the simulation model is superior than the use of the benchmark model (the mean of the observation series) as a accurate representation of real system; an *NSE* value of zero implies that the simulation model, on average, performs as good as the use of the benchmark model; whereas negative *NSE* values implies that the simulation model is poor than the use of the benchmark model, which ususlly indicates unacceptable or unsatisfactory model performance rating.

2.2. Legates-McCabe Efficiency

The *LME* suggested by Legates and McCabe (1999) is expressed as one minus the mean of the absolute errors between the observed and simulated datasets normalized by the mean absolute deviation of the observed values. It is calculated as:

$$LME = 1 - \frac{\sum_{i=1}^{n} |O_i - S_i|}{\sum_{i=1}^{n} |O_i - \overline{O_i}|}, \quad -\infty < LME \le 1 \tag{2}$$

The modified *NSE* is less sensitive to large extreme absolute errors values due to using absolute instead of squared errors within the numerator of the fractional part of *NSE*, but this indicator was not catch much attention and deeply discussed because of its limited application and resulting relative lack of reported information. The range of *LME* lies between minus infinity and one (perfect fit). An *LME* value of lower than zero indicates that the mean value of the observation series series would have been a better predictor than the model.

2.3. NSE Based on Transformed Data Series

Additional modifications of Nash-Sutcliffe efficiency are based on transforming observed outcomes and model outputs, including NSE calculated on a squared transformation of $\boldsymbol{O}$ and $\boldsymbol{S}$ (NSE_{squ}) (Oudin, Andreassian, and Mathevet 2006; Crochemore, Perrin, and Andreassian 2015), NSE calculated on a root-squared transformation of $\boldsymbol{O}$ and $\boldsymbol{S}$ (NSE_{sqrt}) (Perrin, Miche, and Andréassian 2001), NSE calculated on a logarithmic transformation of $\boldsymbol{O}$ and $\boldsymbol{S}$ (NSE_{ln}) (Oudin, Andreassian, and Mathevet 2006), NSE calculated on inverse transformation of $\boldsymbol{O}$ and $\boldsymbol{S}$ (NSE_{inv}) (Le Moine 2008). The NSE_{squ}, NSE_{sqrt}, NSE_{ln} and NSE_{inv} ranges

ISBN 978-88-97999-85-0; Affenzeller, Bruzzone, Jiménez, Longo and Piera Eds.

from minus infinity to one (perfect fit). They is calculated as:

$$NSE_{\text{squ}} = 1 - \frac{\sum_{i=1}^{n}(O_i^2 - S_i^2)^2}{\sum_{i=1}^{n}(O_i^2 - \overline{O_i^2})^2},\ -\infty < NSE_{\text{squ}} \le 1 \quad (3)$$

$$NSE_{\text{sqrt}} = 1 - \frac{\sum_{i=1}^{n}(O_i^{0.5} - S_i^{0.5})^2}{\sum_{i=1}^{n}(O_i^{0.5} - \overline{O_i^{0.5}})^2},\ -\infty < NSE_{\text{sqrt}} \le 1 \quad (4)$$

$$NSE_{\text{ln}} = 1 - \frac{\sum_{i=1}^{n}[\ln(O_i) - \ln(S_i)]^2}{\sum_{i=1}^{n}[\ln(O_i) - \overline{\ln(O_i)}]^2},\ -\infty < NSE_{\text{ln}} \le 1 \quad (5)$$

$$NSE_{\text{inv}} = 1 - \frac{\sum_{i=1}^{n}(O_i^{-1} - S_i^{-1})^2}{\sum_{i=1}^{n}(O_i^{-1} - \overline{O_i^{-1}})^2},\ -\infty < NSE_{\text{inv}} \le 1 \quad (6)$$

The NSE_{squ} indicator is more sensitive to very high magnitudes and shows nearly no reaction on low magnitudes. The NSE_{sqrt} indicator provides more balanced information as the errors are more equally distributed on high- and low- magnitudes parts. The NSE_{ln} indicator is more sensitive to low magnitudes but still reacts to peak magnitudes. The NSE_{inv} indicator is more sensitive to very low magnitudes and shows nearly no reaction on high magnitudes.

2.4. LME on Deviations Weighted by Observations

Nash-Sutcliffe efficiency type indicators based on data transformations use the model error series computed after arithmetic transformation instead of original error series to evaluate model performance, which result in the fact that the implicit weightings to emphasise specific magnitudes of interest usually involve simulated variable. As a result, this fact is very adverse to the interpretation and choice of the efficiency criteria. In this section, we present a kind of Modified Legates-McCabe efficiency indicators calculated on weighted deviations between the observed and simulated datasets as a alternate solution to address the problem described above. These are the two indicators: LME on deviations weighted by observations cubed (LME_{vh}), LME on deviations weighted by observations to the power of minus three (LME_{vl}). They is described respectively as:

$$LME_{\text{vh}} = 1 - \frac{\sum_{i=1}^{n}\left|O_i^3 \cdot (O_i - S_i)\right|}{\sum_{i=1}^{n}\left|O_i^3 \cdot (O_i - \overline{O_i})\right|},\ -\infty < LME_{\text{vh}} \le 1 \quad (7)$$

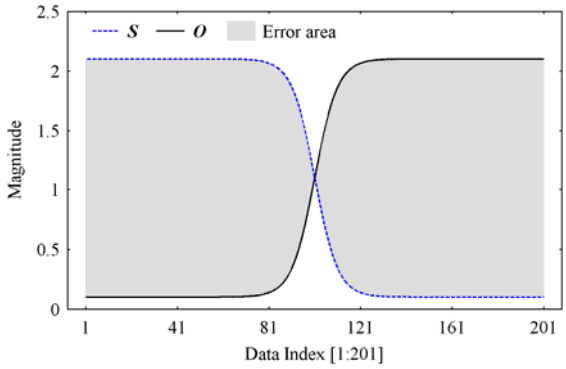
$$LME_{\text{vl}} = 1 - \frac{\sum_{i=1}^{n}\left|O_i^{-3} \cdot (O_i - S_i)\right|}{\sum_{i=1}^{n}\left|O_i^{-3} \cdot (O_i - \overline{O_i})\right|},\ -\infty < LME_{\text{vl}} \le 1 \quad (8)$$

The LME_{vh} is used to emphasize errors that occur in very high mignitudes of absolute $\boldsymbol{O}$. The LME_{vl} is used to emphasize errors that occur in very low mignitudes of absolute $\boldsymbol{O}$. Through this modifications to the *LME*, the differences between the observed and simulated values are quantified as absolute errors weighted by given observation series, which reduce the adverse effects of factors such as interpretation of indicator measures and the choice of efficiency indicators;

3. APPROACH FOR INDICATORS ANALYSIS

$$\begin{cases} \{x_i;\ i = 1,\ 2,\ \ \ , n\} = [-10 : 0.1 : 10] \\ O_i = 1.1 + \tanh(x_i) \\ S_i = 1.1 - \tanh(x_i) \end{cases} \quad (9)$$

In this example, $\{O_i, S_i\}$ generated by equation (9), as shown in Figure 1. The 201 separate synthetic simulated graph were generated by the following processes: for the simulated graph number 1, the first ordinate was the same as the first ordinate of the observed graph and the remaining ordinates of the simulated graph stays the same; for the second simulated graph, the first two ordinate was the same as the first two ordinate of the observed graph and the remaining ordinates of the simulated graph stays the same; for the simulated graph number 3, the first three ordinate was the same as the first three ordinate of the observed graph and the remaining ordinates of the simulated graph stays the same; and in the remaining model simulations (4 to 201), the observed graph values were progressively substituted for the simulated graph until the last model simulation (number 201) was the actual observed graph. Figure 2 shows the observed graph and simulated graph for model simulation number 100 (the first 100 time steps are the same as the observed and the remaining values are the arithmetic mean of entire observed graph).



ISBN 978-88-97999-85-0; Affenzeller, Bruzzone, Jiménez, Longo and Piera Eds.

Figure 1: Original Observed (Black Line) and Original Simulated (Blue Dotted Line) Graph and (Gray Area) Error Area

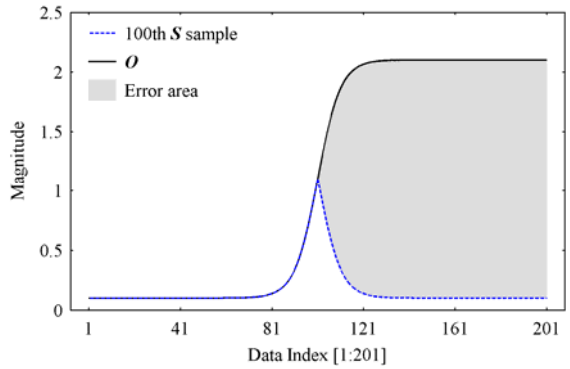


Figure 2: Original Observed (Black Line) and 100th Simulated (Blue Dotted Line) Graph and (Gray Area) Error Area

4. RESULTS AND DISCUSSION

The computed values of the goodness-of-fit indicators described in section 2 for each of 201 separate synthetic simulated graph are shown in Figures 3-4 and Table 1.

The plot in Figures 2 shows the Nash-Sutcliffe efficiency and its modified forms (NSE_{squ}, NSE_{sqrt}, NSE_{ln}, NSE_{inv}), the plot in Figures 4 shows Legates-McCabe efficiency and LME on deviations weighted by observations (LME_{vh}, LME_{vl}). Table 1 shows the Sample Index (colums) 1, 41, 81, 121, 161, 201 and the computed values (rows) of the different indicators at Sample Index.

For simulated sample index 1, all goodness-of-fit indicators have negative values, which implies that the simulation model is poor than the use of the mean of the observation series as a accurate representation of real system. The computed values of NSE, NSE_{squ}, NSE_{sqrt}, NSE_{ln}, NSE_{inv} adopting squared errors is less than LME, LME_{vh}, LME_{vl} adopting absolute errors.

For simulated sample number 1 to 81, there were major increases of of NSE (1.76), NSE_{squ} (1.74), NSE_{sqrt} (1.77), NSE_{ln} (1.78), NSE_{inv} (1.74), LME (0.85), LME_{vl} (1.94), which indicates that these indicators very sensitive to very low magnitudes conditions. LME_{vh} (0.00) exhibited nearly no reaction on or is not sensitive to very low magnitudes conditions.

For simulated sample number 81 to 121, there were major increases of of NSE (0.46), NSE_{squ} (0.45), NSE_{sqrt} (0.41), NSE_{ln} (0.31), NSE_{inv} (0.13), LME (0.29) which indicates that these indicators very sensitive to low magnitudes conditions. LME_{vh} (0.23) showed a immediate reaction on or is sensitive to high magnitudes conditions, and slight increases of LME_{vl} (0.03) showed nearly no reaction on or is not sensitive to high magnitudes conditions.

For simulated sample number 121 to 201, there were major increases of of NSE (1.76), NSE_{squ} (1.74), NSE_{sqrt} (1.77), NSE_{ln} (1.78), NSE_{inv} (1.75), LME (0.85), which indicates that these indicators very sensitive to high magnitudes conditions. LME_{vh} (1.77) exhibited showed a strong reaction on very high magnitudes conditions, and LME_{vl} (0.00) showed nearly no reaction on or is not sensitive to very high magnitudes conditions.

The modified forms LME_{vh} did exhibited nearly no reaction on low magnitudes of observed and simulated series and therefore was mostly sensitive for better model realisation during high magnitudes of observation series, while LME_{vl} did show nearly no reaction on high magnitudes of observed and simulated series and was mostly sensitive for better model realisation during low magnitudes of observation series.

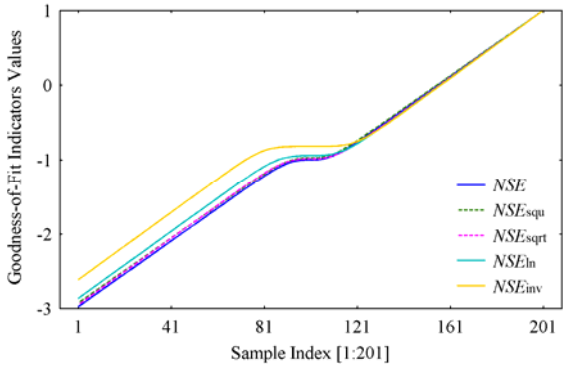


Figure 3: Evolution of NSE, NSE_{squ}, NSE_{sqrt}, NSE_{ln}, NSE_{inv} During Example in Sect. 3

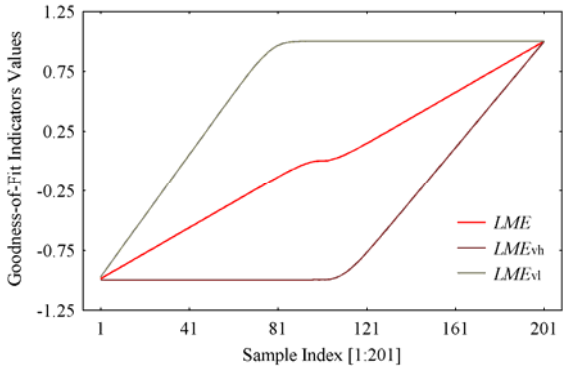


Figure 4: Evolution of LME, LME_{vh}, LME_{vl} during Example in Sect. 3

Table 1: Goodness-of-fit Indicators Values for The Six Sample Index of Example

S.S.I.[a]	1	41	81	121	161	201
NSE	-2.98	-2.09	-1.22	-0.76	0.12	1.00
LME	-0.99	-0.56	-0.14	0.15	0.57	1.00
NSE_{squ}	-2.93	-2.05	-1.19	-0.74	0.13	1.00
NSE_{sqrt}	-2.95	-2.06	-1.18	-0.77	0.11	1.00
NSE_{ln}	-2.87	-1.96	-1.09	-0.78	0.10	1.00
NSE_{inv}	-2.62	-1.71	-0.88	-0.75	0.09	1.00
LME_{vh}	-1.00	-1.00	-1.00	-0.77	0.11	1.00
LME_{vl}	-0.97	0.05	0.97	1.00	1.00	1.00

[a] S.I. = Simulation Sample Index.

5. CONCLUDING REMARKS

Eight different goodness-of-fit measures for the evaluation of model performance or validity were investigated with a simple example. Through the computed results of the different goodness-of-fit indicators obtained for a simple example, their measure

ISBN 978-88-97999-85-0; Affenzeller, Bruzzone, Jiménez, Longo and Piera Eds.

characteristics in terms of emphasis on different types of errors are analyzed and identified. To increase the sensitivity of indicators measures to very high- and very low- magnitudes conditions and reduce adverse effect on the choice of the goodness-of-fit indicators, LME on deviations weighted by observations cubed (LME_{vh}) and LME on deviations weighted by observations to the power of minus three (LME_{vl}) were proposed. Contrary to what was expected, the results from the a simple example revealed that LME_{vh} did show nearly no reaction on low magnitudes of observed and simulated series and therefore was mostly sensitive for better model realisation during high magnitudes of observation series, while LME_{vl} did show nearly no reaction on high magnitudes of observed and simulated series and was mostly sensitive for better model realisation during low magnitudes of observation series.

ACKNOWLEDGMENTS

The paper was also made possible through the financial support of the National Natural Sciences Foundation of China (Grant No. NNSFC 61374164), for which the authors are most grateful.

REFERENCES

Bennett, N.D., Croke, B.F., Guariso, G., et al., 2013. Characterising performance of environmental models. Environmental Modelling & Software, 40: 1-20.

Crochemore, L., Perrin, C., Andreassian, V., et al., 2015. Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrological Sciences Journal, 60(3): 402-423.

Dawson, C.W., Abrahart, R.J. and See, L.M., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environmental Modelling & Software, 22(7): 1034-1052

Hauduc, H., Neumann, M.B., Muschalla, D., et al., 2015. Efficiency criteria for environmental model quality assessment: a review and its application to wastewater treatment. Environmental Modelling & Software, 68: 196-204.

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

Legates, D.R. and McCabe, G.J., 1999. Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35(1): 233-241.

Le Moine, N., 2008. Le bassin versant de surface vu par le souterrain: une voie d'amélioration des performances et du réalisme des modéles pluie–débit? PhD Thesis, Université Pierre et Marie Curie, Pairs, France, pp. 324.

Martens, J., Put, F. and Kerre, E., 2006. A fuzzy set theoretic approach to validate simulation models. ACM Transactions on Modeling and Computer Simulation, 16(4): 375-398.

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models: part I - a discussion of principles. Journal of Hydrology, 10(3): 282-290.

Nicolle, P., Pushpalatha, R., Perrin, C., et al., 2014. Benchmarking hydrological models for low-flow simulation and forecasting on French catchments. Hydrology and Earth System Sciences, 18(8): 2829-2857.

Oudin, L., Andreassian, V., Mathevet, T., Perrin, C. and Michel, C., 2006. Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. Water Resources Research, 42.

Perrin, C., Michel, C. and Andréassian, V., 2001. Does a large number of parameters enhance model performance? comparative assessment of common catchment model structures on 429 catchments. Journal of Hydrology, 242(3): 275-301.

Pushpalatha, R., Perrin, C., Le Moine, N. and Andréassian, V., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. Journal of Hydrology, 420: 171-182.

AUTHORS BIOGRAPHY

Kai-Bin Zhao received the M.S. degree in guidance navigation and control from the Harbin Engineering University, China, in 2013. He is currently working toward the Ph.D. degree in control science and engineering at Harbin Institute of Technology, China. His current research focuses on model validation for complex simulation system. His e-mail address is kaibin.zhao.HIT@hotmail.com.

Ke Fang is an Associate Professor of the Control & Simulation Center, School of Astronautics at Harbin Institute of Technology, China. He holds a Ph.D. in control science and engineering from Harbin Institute of Technology, and was a visiting scholar at Arizona State University from year 2014 to 2015 in AZ, USA. His research interests include complex simulation systems, model validation and VV&A. His e-mail address is hitsim@163.com.

Ming Yang is a Professor of the Control & Simulation Center, School of Astronautics at Harbin Institute of Technology, China. He holds a Ph.D. in control science and engineering from Harbin Institute of Technology. He involved in major simulation conference and committees in China. His e-mail address is myang@hit.edu.cn.