36th European Modeling & Simulation Symposium, 002
21st International Multidisciplinary Modeling & Simulation Multiconference

2724-0029 © 2024 The Authors.
doi: 10.46354/i3m.2024.emss.002

Credibility Metric Weighting for Accelerating Qualitative Model Evaluation

Ke Hu[1], Jiabei Gong[1], Jiayi Zhang[1], Bijun Tang[4], Yuanjun Laili[1, 2, *], Lin Zhang[1, 3] and Lei Ren[1, 2]

[1]School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100191, China
[2]Zhongguancun Laboratory, Beijing 100094, China
[3]Key Laboratory of Intelligent Manufacturing Systems Technology, Beijing, 100854, China
[4]Beijing Institute of Astronautical Systems Engineering, Beijing 100076, China

[*]Corresponding author. Email address: lailiyuanjun@buaa.edu.cn

Abstract

Credibility evaluation of a simulation model is an important premise of the simulation, since incredible model can produce unconvincing or even wrong results. The overall credibility of a simulation model is evaluated by weighting a group of metrics. These metrics are calculated by different kinds of quantitative or qualitative methods. As the weighting criteria of different kinds of models is hard to define uniformly, existing qualitative methods depend highly on expert scoring. However, expert scoring with metric weighting process is subjective and time-consuming. It is challenging to weight the metrics appropriately and evaluate a simulation model efficiently. Therefore, this paper proposes an automated metric weighting method for accelerating the credibility evaluation. It applies the historical scoring records of similar models as the reference. Then, it introduces an evolutionary algorithm to calculate the possible weights of the metrics inversely and calculate the overall credibility value of the simulation model. Experimental results on a typical simulation model verify that the proposed method is able to weight the metrics within seconds and calculate the credibility value with high degree of alignment with expert scoring.

Keywords: credibility evaluation; qualitative expert scoring; weighting rules; evolutionary algorithm

1. Introduction

Credibility evaluation is indispensable for a simulation model before its application (Balci, 1986; Law, 2022). A simulation model without credibility evaluation may produce unconvincing results, provide a wrong view for the model user, and even lead to wrong decisions. The credibility evaluation of a simulation model can be implemented quantitatively or qualitatively. If the reference data of the target object is provided, quantitative methods can be applied to compare the simulation results to the reference data. However, if the reference data is unreachable, or the simulation is expensive, qualitative methods are introduced as an key alternative (Ho and Ma, 2018).

However, existing qualitative methods depends highly on expert scoring. When the number of metrics for credibility evaluation is high, the scoring process becomes time-consuming. Experts are required to designate a set of rules to score the metrics and define weighting criteria for a comprehensive evaluation. For the models that share similar weighting criteria, the scoring process becomes tedious. It is necessary to establish an automatic weighting process to accelerate the credibility evaluation.

Therefore, this paper presented a method of credibility metric weighting to accelerate qualitative model evalu-

ation. The evaluation framework adopted in this paper is derived from NASA-STD-7009 (Min et al., 2010). The basic idea is to utilize historical scoring results as a reference, employing polynomial expansion to approximate the metric formulae. First, the perceptually important points algorithm (PIP) (Tsinaslanidis and Kugiumtzis, 2014) is introduced to extract key points of the simulation results. Subsequently, polynomial expansion is applied to establish the formulae for fitting the evaluation metrics. Then the credibility of the simulation model is constructed by weighting the evaluation metrics. Following this, three typical evolutionary algorithms (Slowik and Kwasnicka, 2020) are introduced to estimate the optimal polynomial coefficients and weights for fitting the above process. Experimental results show that the metric values and model credibility value calculated by the proposed method have an average error within 8.5% compared to the expert scoring results.

The rest of this paper is organized as follows. Section II presents a literature review of qualitative assessment methods for simulation models. Section III outlines the problem scenarios. Section IV proposes a stochastic hyper-heuristic-based differential evolution algorithm for solving the problem. Section V conducts experiment on the *wolf-sheep* predation model. Section VI summarizes the paper.

2. Literature review

Qualitative analysis is a crucial part of credibility evaluation. It mainly refers to the process of evaluating certain metrics based on expert scoring and experience-based weighting. In the early stages, researchers proposed some typical methods for assessing the effectiveness of simulation models, including face validation method (Hermann, 1967), Turing test method (Schruben, 1980), and directed judgment method (Wright, 1972) based on graphical comparisons.

Based on these foundations, Goerger et al. (2005) proposed improvement measures by identifying performance biases and anchoring biases present in experts, thereby enhancing the accuracy of face validation. Gao et al. (2019) presented a petrochemical process simulation model validation framework based on symbolic directed graphs, comprehensively validating the model at multiple levels to improve its correctness and accuracy. Zhang et al. (2013) introduced a model validation and verification method using symbolic directed graphs and qualitative trend analysis. Additionally, Samlaus and Fritzson (2015) utilized semantic constraints to establish role models for verifying and analyzing the interactions, behaviors, and parameters of physical models. Ahn et al. (2014) proposed a Delphi method for assessing the credibility of M&S procedures and validated its effectiveness through case studies, providing a structured and objective approach for M&S credibility assessment.

Analytic hierarchy process (AHP) and Technique for order preference by similarity to ideal solution (TOPSIS) theory also find wide applications in qualitative model evaluation. Zhang et al. (2011) proposed a group-AHP evaluation method that integrates the wisdom of multiple experts and avoids subjective biases, providing new insights for the credibility assessment of complex simulation systems. Lu and Yuan (2018) presented a novel credibility assessment scheme for cloud computing services based on TOPSIS, considering both objective aspects of service quality and user subjective preferences.

With the increasing complexity of simulation models and challenges such as missing simulation data, researchers are gradually employing knowledge-based qualitative assessment methods. Typical approach includes the categorizing complex simulation behaviors into five types and combining expert experience and domain knowledge for analyzing simulation systems Min et al. (2010).

Li et al. (2016) proposed a group assessment method for the credibility of complex simulation systems based on second-order additive fuzzy measures, considering the correlation between evaluation metrics and evaluation experts, making the evaluation method more reasonable and objective. Foures et al. (2016) introduced a qualitative measurement method based on specification descriptions, combining simulation objectives and formal methods to evaluate the simulation models in different scenarios.

From the perspective of credibility assessment methods, most of the classical qualitative assessment methods are highly subjective and time-consuming. Their effectiveness is greatly influenced by external factors and requires significant manpower. Therefore, establishing automated qualitative methods for the credibility evaluation of simulation model is imperative.

3. Problem description

The historical data of the simulation model consists of two parts: first, the historical outputs of the model; second, the historical scoring results by experts. The historical outputs and scoring results of the model were used as reference cases for constructing the metric fitting formulae. The outputs of the reference cases are denoted as $Y = \{Y_1, Y_2, Y_3, \ldots, Y_j\}$, where each $Y_j, j \in [1, J]$ represents a one-dimensional sequence $Y_j = \{y_{j1}, y_{j2}, \ldots, y_{jn}\}$. J represents the number of reference cases, and n represents the length of the output for the reference case. Historical metrics values are expressed as $I = \{I_1, I_2, I_3, \ldots, I_j\}$, where each $I_j, j \in [1, J]$ represents the metrics values of the j_{th} reference case, one-dimensional sequence $I_j = \{i_{j1}, i_{j2}, \ldots, i_{jm}\}$ where $m \in [1, M]$ represents the index of the metrics. Historical credibility scoring results can be represented as $c = \{c_1, c_2, c_3, \ldots, c_j\}$. Here, $c_j, j \in [1, J]$ denotes the credibility score of the j_{th} reference case.

The construction of a formula for fitting one metric is illustrated as an example, and the process of fitting other metrics is the same. The formula fitting process for metric i_{jm} based on the j_{th} reference case is exempli-

fied. The output of the j_{th} reference case after PIP compression (detailed in Section 4.2) to k dimensions is expressed as: : $Y'_j = \{y'_{j1}, y'_{j2}, \ldots, y'_{jk}\}$ and the metric value is $I_j = \{i_{j1}, i_{j2}, \ldots, i_{jm}\}$. This paper uses polynomials to fit the formula for calculating metric values. The formula for forecast metric $\widehat{i}_{jm}$ is expressed as follows.

$$\widehat{i}_{jm} = \sum_{\lambda=1}^{k} \left(a_{\lambda 1} \cdot y'_{j\lambda} + a_{\lambda 2} \cdot {y'_{j\lambda}}^{2} + \ldots + a_{\lambda k} \cdot {y'_{j\lambda}}^{k} + b \right) \quad (1)$$

where $a_{\lambda k}$ represents the coefficient of the k_{th} power. b denotes a constant term. When normalizing the simulation output to fall within the range of [0,1], the values of higher-order terms in Equation 1 will diminish. Therefore, to simplify, this paper retains only the linear term and the constant term in Equation 1. Following this, the predicted metrics are weighted and summed to obtain a predictive credibility score, which is calculated as follows.

$$\widehat{c}_j = \sum_{m=1}^{M} w_{jm} \cdot \widehat{i}_{jm} \quad (2)$$

The objective function is the sum of the differences between the predicted metric values and historical metric values, and the differences between the predicted credibility scores and historical credibility scores.

$$\min f = \delta \sum_{m=1}^{M} \left| i_{jm} - \widehat{i}_{jm} \right| + (1 - \delta) \left| c_j - \widehat{c}_j \right| \quad (3)$$

δ represents the weighting coefficient, which is set to 0.8 in the experiment.

General framework is such that the output of each reference case can construct a set of metric calculation formulae. For example, if there are J reference cases, then J sets of metric calculation formulae can be constructed. When there is a new case, it can be substituted into J sets of fitting formulae for metrics. In this case, J sets of metric values will be obtained, defined as pseudo-metric values. The pseudo-metric values can be expressed as $P = \{P_1, P_2, P_3, \ldots, P_j\}$, where each $P_j, j \in [1, J]$ represents a one-dimensional sequence $P_j = \{p_{j1}, p_{j2}, \ldots, p_{jm}\}$. p_{jm} denotes the value of the m_{th} pseudo-metric for the new case under reference case j.

The pseudo-metric value of the new case to the final metric value still needs a weighting operation. The weight of each pseudo-metric requires the normalised euclidean distance between the output of the new case and the reference case. The output of the new case after normalization is expressed as $Z = \{z_1, z_2, z_3, \ldots, z_n\}$ and the normalized output of the reference case j is denoted as $H_j = \{h_{j1}, h_{j2}, \ldots, h_{jn}\}$. Then the normalised Euclidean distance is defined as follows:

$$d_j = \sqrt{(z_1 - h_{j1})^2 + (z_2 - h_{j2})^2 + \ldots + (z_n - h_{jn})^2} \quad (4)$$

The weight ψ_j of the pseudo-metric value can be defined as follows:

$$\psi_j = \frac{1}{1 + d_j} \quad (5)$$

The distance between the new case and all reference cases is represented as: $D_{distance} = \{d_1, d_2, d_3, \ldots, d_j\}, j \in (1, J)$. The set of weights for the pseudo-metrics is denoted as $\Psi = \{\psi_1, \psi_2, \psi_3, \ldots, \psi_j\}, j \in (1, J)$. The values of the metrics for new case can be expressed as $\hat{I} = \{\hat{i}_1, \hat{i}_2, \hat{i}_3, \ldots, \hat{i}_m\}, m \in (1, M)$. The reference case β furthest from the new case is considered to lack reference value and will be discarded. The value of the m_{th} predicted metric for the new case is calculated as follows:

$$\hat{i}_m = \sum_{j=1}^{j=J, j \neq \beta} \psi_j \cdot p_{jm} \quad (6)$$

Once the predicted metric values for the new case were obtained, the model credibility score can be obtained from the following equation.

$$\hat{c} = \sum_{j=1}^{j=J, j \neq \beta} w_{jm} \cdot \psi_j \cdot p_{jm} \quad (7)$$

The methodology for obtaining the weight w_{jm} and Polynomial fitting coefficients $a_{\lambda k}$ will be covered in Section 4. In summary, the process involves constructing a formula to adjust the calculation of the assessment metrics based on the reference cases. Subsequently, the output of the new case is inputted into this formula to derive the pseudo-metric value. Finally, the weights, derived from the output distances, are utilized to calculate the predicted metric, as depicted in Figure 1.

4. Methodology

4.1. Framework of the proposed methodology

The construction of the metrics fitting formulae is the most critical aspect. After dimensionality reduction of the reference cases' outputs through PIP, a stochastic hyper-heuristic-based differential evolution algorithm was designed to dynamically select the underlying operators. Based on this foundation, we analyze the evolutionary patterns of the simulation model to obtain the functional representations of metrics outlined in NASA standards. Subsequently, the model's credibility is established based on these assessment metrics. Finally, experimental is conducted within *wolf-sheep* predation model.

4.2. PIP

When dealing with high-dimensional data, using the original data would be computationally burdensome and might

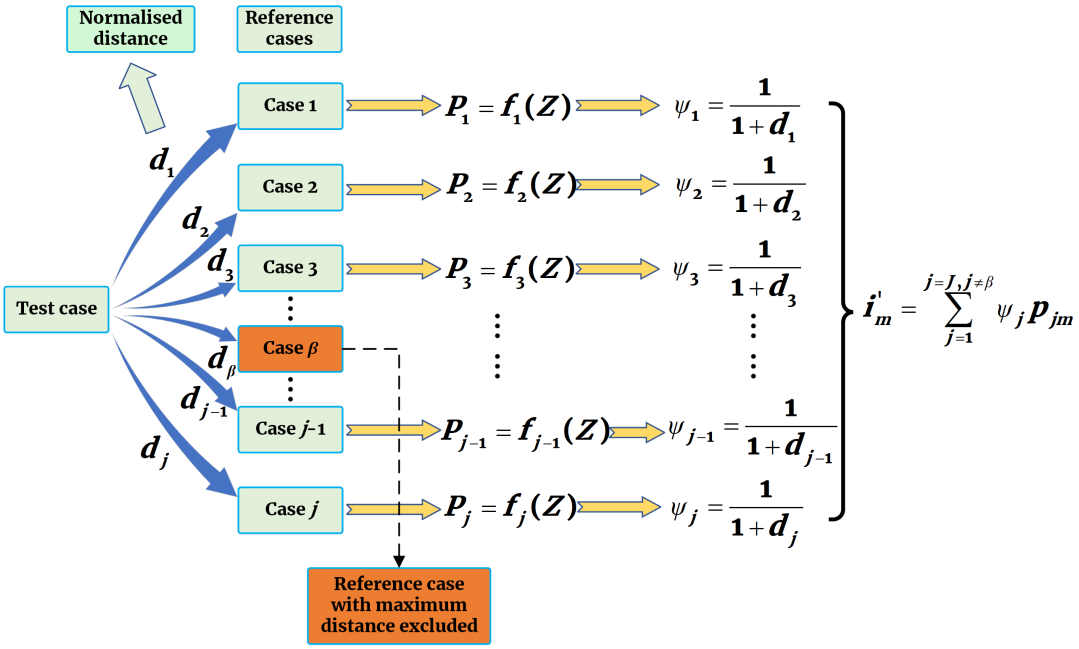

Figure 1. Explanation of the process for calculating metrics using the new output as a test case. The formula $f(Z)$ for calculating the metric established through reference cases, followed by distance weighting, allows for the prediction of the metric value for new cases.

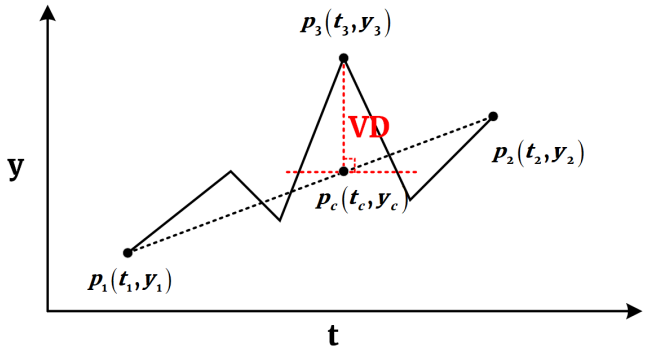

Figure 2. The vertical distance VD in PIP.

overlook important data points. Therefore, compressing the data patterns is necessary. This paper adopts the PIP (Perceptually Important Point) algorithm for dimensionality reduction. The algorithm proceeds as follows:

1. Store the first and last points of the sampled time series in the downsampled data point set.
2. Calculate the distance from each remaining unsampled point to its adjacent two keypoints.
3. Sample the point with the maximum distance and store it in the downsampled data point set.
4. Repeat steps 2 and 3 until the number of sampled keypoints reaches the specified k dimensions.

To compute the distance shown in Figure 2 between adjacent keypoints, the vertical distance can be calculated as follows:

$$VD(t_3, y_3) = |y_c - y_3| = \left| y_1 + (y_2 - y_1) \cdot \frac{t_c - t_1}{t_2 - t_1} - y_3 \right| \quad (8)$$

For multi-dimensional output data of the simulation model, an additional joint distance needs to be introduced. Suppose the output of simulation model consists of D-dimensional data, then the joint distance from $P_3(t_3, y_{31}, y_{32}, \ldots, y_{3D})$ to its adjacent feature points $P_1(t_1, y_{11}, y_{12}, \ldots, y_{1D})$ and $P_2(t_2, y_{21}, y_{22}, \ldots, y_{2D})$ is defined as follows:

$$d_{\text{joint}}(P_3) = \left| k(P_3) - k(P_1) - \frac{k(P_2) - k(P_1)}{t_2 - t_1}(t_3 - t_1) \right|$$
$$k(P_i) = \sqrt{\sum_{j=1}^{D} y_{ij}^2} \quad (9)$$

The distance of multi-dimensional data is defined as the maximum value between the distance of each dimension and the joint distance.

$$d_{mul}(P_3) = \max\{d_1(P_3), d_2(P_3), \ldots, d_D(P_3), d_{joint}(P_3)\} \quad (10)$$

where $d_D(P_3) = VD(t_3, y_{3d}), d \in (1, D)$. This definition of multi-dimensional distance aims to capture the most significant peak (valley) values as much as possible, ensuring that the compressed data closely resembles the original data.

4.3. Stochastic hyper-heuristic-based differential evolution algorithm

Stochastic hyper-heuristic-based differential evolution algorithm is an intelligent optimization algorithm combining stochastic hyper-heuristic and differential evolution (DE) algorithm. In the low-level problem space, problem solutions are encoded as floating-point numbers, ranging from 0 to 1. In the high-level strategy space, strategies for the DE algorithm are encoded using a combination of integers and floating-point numbers, encompassing both operator and parameter selection.

4.3.1. Encode

The problem's solution comprises the coefficients of the linear terms and the constant term in Equation 1, along with the weight values in Equation 2. In the lower-level problem domain, the length of the individual encoding is $N = 2nM + M$, where n represents the length of the output data, and M represents the number of metrics. The range of the code is $[0, 1]$.

4.3.2. Stochastic hyper-heuristic strategy

A strategy based on stochastic hyper-heuristic is employed for selecting operators.The operator decision encoding in the high-level strategy domain adopts a hybrid encoding of integers and floating-point numbers. Each individual in the population corresponds to an operator decision encoding and a problem solution encoding. The operator decision encoding is a sequence S of length 9.

- S[0]: the new individual chooses the historical optimal solution *ibest* or the global optimal solution *gbest* or the random individual historical *rbest* optimal solution in a new iteration.
- S[1]: select different differential evolution operators
- S[2]: differential evolution operator parameter selection
- S[3]: decision making whether to perform the difference operation again or not
- S[4]: differential evolution operator parameter selection
- S[5]:Scale factor of operator 1
- S[6]:Scale factor of operator 2
- S[7]:Probability of crossover
- S[8]:Variation probability

S[0] is a random integer from 0 to (*popsize* + 3), S[1] and is a random integer from 0 to (*popsize* + 1), S[2] and is a random integer from 0 to popsize, S[5] to S[7] is a random floating-point number from 0.2 to 0.7, and S[8] is a random floating-point number from 0.1 to 0.5. The variable *popsize* represents the population size.

Individual i is represented during the G_{th} iteration as $X_{i,G} = \{x^1_{i,G}, x^2_{i,G}, x^3_{i,G}, \ldots, x^N_{i,G}\}$. For each individual $X_{i,G}$, the corresponding mutant vector can be represented as $V_{i,G} = \{v^1_{i,G}, v^2_{i,G}, v^3_{i,G}, \ldots, v^N_{i,G}\}$. The two mutation strategies employed in this paper are listed as follows:

1. $V_{i,G} = X_{i,G} + F_1 \times (X_{rbest1,G} - X_{rbest2,G})$
2. $V_{i,G} = X_{i,G} + F_2 \times (X_{gbest,G} - X_{rbest3,G})$

F_1: The scaling factor of Operator 1, controlled by S[5]. F_2: The scaling factor of Operator 2, controlled by S[6]. $X_{rbest1,G}, X_{rbest2,G}, X_{rbest3,G}$: Historical best solutions of randomly selected individuals during the G_{th} iteration, where individual selection is controlled by S[1], S[2], S[4] respectively. $X_{gbest,G}$: Global best solution in the G_{th} iteration.

If any element in vector $V_{i,G}$ exceeds the encoded upper and lower bounds, it is reset to a random number within the bounds. After mutating, a crossover operation is performed on each individual to obtain the trial vector $U_{i,G} = \{u_{1i,G}, u_{2i,G}, u_{3i,G}, \ldots, u_{Ni,G}\}$, and the crossover operation can be defined as follows:

$$u^e_{i,G} = \begin{cases} x^e_{i,G}, \text{if}(\text{rand}_e[0,1) > \text{CR}) \text{and} (e \neq e_{\text{rand}}) \\ v^e_{i,G}, \text{otherwise} \end{cases} \tag{11}$$

where CR: a fixed differential crossover probability within the range [0.2, 0.7], controlled by S[7]. e_{rand}: a random integer within the range [1, N].

To enhance solution diversity, a single-point mutation is applied to the trial vector, with the mutation probability controlled by S[8].

$$u^e_{i,G} = \begin{cases} \text{rand}_{num}, \text{if}(\text{rand}_p[0,1) < \text{CR}') \\ u^e_{i,G}, \text{otherwise} \end{cases} \tag{12}$$

If a random number $rand_p$ is less than the single-point mutation probability CR', the e_{th} element of the vector is reset to a random number within the coding range, otherwise the value of the element remains unchanged. After mutation, crossover and single point mutation, the objective function values of the experimental vector and the original individual are calculated. If the objective function value of the trial vector is less than that of the original individual, the high-level strategy operator selection coding S is retained; otherwise, the operator selection coding S is reinitialised.

5. Experimental discussion

We set up with 6 groups of reference cases and 2 groups of test cases and score them in advance by experts. The reference cases are used to construct the metrics fitting formulae, and then the outputs of the new cases are processed according to Section 3 and Section 4 to obtain the predicted metrics value and credibility value. The calculated metric values and credibility values of the new cases are compared with the expert scoring values, and if they are within the error allowance, it means that the proposed qualitative assessment method is effective.

Table 1. main parameter setting table

	case1	case2	case3	case4	case5	case6	case7	case8
wolf_reproduce	4	12	4	4	12	12	6	6
sheep_reproduce	4	4	4	8	4	8	5	3
grass_regrowth_time	40	40	100	40	100	40	80	25

Table 2. Assessment metrics values and credibility values for cases

	case1	case2	case3	case4	case5	case6	case7	case8
i_1	0.765	0.537	0.238	0.312	0.663	0.438	0.306	0.654
i_2	0.426	0.738	0.536	0.436	0.384	0.523	0.258	0.637
i_3	0.327	0.683	0.573	0.268	0.568	0.794	0.354	0.695
i_4	0.776	0.367	0.362	0.376	0.612	0.427	0.247	0.633
i_5	0.267	0.462	0.487	0.215	0.247	0.537	0.238	0.524
i_6	0.421	0.767	0.458	0.431	0.375	0.683	0.284	0.675
i_7	0.697	0.548	0.638	0.374	0.534	0.712	0.347	0.739
i_8	0.415	0.735	0.382	0.437	0.763	0.549	0.296	0.697
c	0.713	0.764	0.536	0.347	0.685	0.674	0.323	0.801

5.1. Experimental settings

To validate the establishment of the proposed credibility assessment metric system, the experimental section adopts the classic *wolf-sheep* predation model as the simulation experiment model. The simulation of the wolf-sheep predation model is conducted using *NetLogo version 6.3*. The main model parameters involved in the *wolf-sheep* predation model are *wolf-gain-from-food*, *wolf-reproduce*, *sheep-gain-from-food*, *grass-regrowth-time*, *sheep-reproduce*. Fixing *initial-number-wolves*, *initial-number-sheep*, *wolf-gain-from-food*, and *sheep-gain-from-food*, the variables of the experiment are set to *wolf-reproduce*, *sheep-reproduce*, *grass-regrowth-time*. These three parameters were chosen as variable parameters because they have a significant impact on the output of the *wolf-sheep* predation model. The output of the simulation model varies even with identical parameter settings. Therefore, for each case, we conducted five repetitions of the experiment, generating 500 dimensions of output data in each repetition. The experiment comprises a total of 8 case. Cases 1 to 6 serve as reference cases for establishing the metric system, while Cases 7 to 8 are designated as test cases to evaluate the effectiveness of the metric system. The parameter settings for each case are provided as shown in Table 1:

According to the metric system proposed in Section 3, expert ratings were obtained to derive the reference metrics I and credibility c for the 8 cases, as shown in Table 2.

The experiment is divided into two groups, namely: comparison of optimization algorithms, and credibility assessment experiment for test cases.

5.2. Results

5.2.1. Algorithm comparison experiment

A comparison experiment of different optimization algorithms including DE, GA_GT and MBO is conducted. The convergence graphs of fitness values under different optimization algorithms for various cases are illustrated in Figure 3.

Through comparison, it can be observed that in this experiment, the stochastic hyper-heuristic DE algorithm outperforms other algorithms in terms of fitness convergence speed across different cases.

Table 3. The calculated metric and credibility values for case 7

	case1	case2	case3	case4	case5	case6	sum
$\hat{i}_1$	0.0570	0.0387	0.0325	0.0841	0.0492	0.0322	0.2550
$\hat{i}_2$	0.0312	0.0542	0.0754	0.0992	0.0275	0.0396	0.2729
$\hat{i}_3$	0.0246	0.0500	0.0841	0.0723	0.0419	0.0552	0.2780
$\hat{i}_4$	0.0549	0.0280	0.0508	0.0823	0.0449	0.0317	0.2645
$\hat{i}_5$	0.0203	0.0332	0.0731	0.0671	0.0189	0.0384	0.2178
$\hat{i}_6$	0.0314	0.0555	0.0680	0.0923	0.0277	0.0507	0.2702
$\hat{i}_7$	0.0512	0.0387	0.0914	0.0926	0.0395	0.0521	0.3268
$\hat{i}_8$	0.0304	0.0545	0.0571	0.1007	0.0571	0.0406	0.2859
d	8.4540	8.5160	7.6250	7.7920	8.3440	8.3450	

The data in the table have been approximated to four decimal places.

Table 4. The calculated metric and credibility values for case 8

	case1	case2	case3	case4	case5	case6	sum
$\hat{i}_1$	0.1598	0.1211	0.1389	0.2403	0.1411	0.0865	0.6475
$\hat{i}_2$	0.0997	0.1550	0.2146	0.2662	0.0975	0.1016	0.6685
$\hat{i}_3$	0.0753	0.1448	0.2205	0.2330	0.1187	0.1417	0.7009
$\hat{i}_4$	0.1684	0.0722	0.1779	0.2386	0.1212	0.0836	0.6233
$\hat{i}_5$	0.0547	0.1058	0.1930	0.2189	0.0486	0.1064	0.5085
$\hat{i}_6$	0.0974	0.1680	0.1962	0.2665	0.0846	0.1226	0.6688
$\hat{i}_7$	0.1483	0.1234	0.2254	0.2507	0.1166	0.1296	0.7433
$\hat{i}_8$	0.0925	0.1510	0.1877	0.2716	0.1509	0.1003	0.6825
d	7.0540	5.8760	10.4030	10.8500	7.9340	6.9390	

The data in the table have been approximated to four decimal places.

5.2.2. Test experiment

Use the established metric calculation formulae based on reference cases to assess the credibility of cases 7 to 8 as test cases. Each case consists of 8 metric values. The weight of each pseudo-metric value is calculated according to the equation 4, and then each weighted metric value under the same metric is added together to obtain the final metric value.

After the output of the test case is processed by PIP downscaling, the metrics values and credibility values for each test case are calculated according to Equation 1, 7.The results are summarised in Table 3 and Table 4.

Table 3 shows the calculated metric and distance values for test case 7. The reference case with the furthest distance is marked in red and was discarded. This operation is repeated for case 8.

The metric values and credibility values calculated for the test cases were compared with the expert scoring values, as shown in Figure 4. The horizontal axis consists of 9 columns, with the last column representing the credibility value of the test case and the remaining columns representing metric values.

The bar comparison chart shows that the calculated

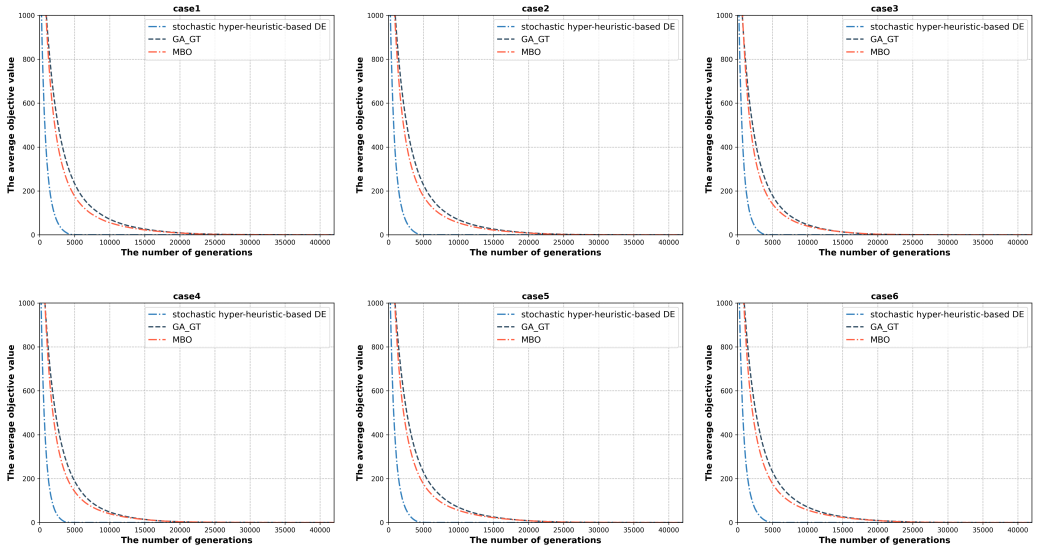

Figure 3. The convergence curve of fitness values during the evolution process of reference cases.

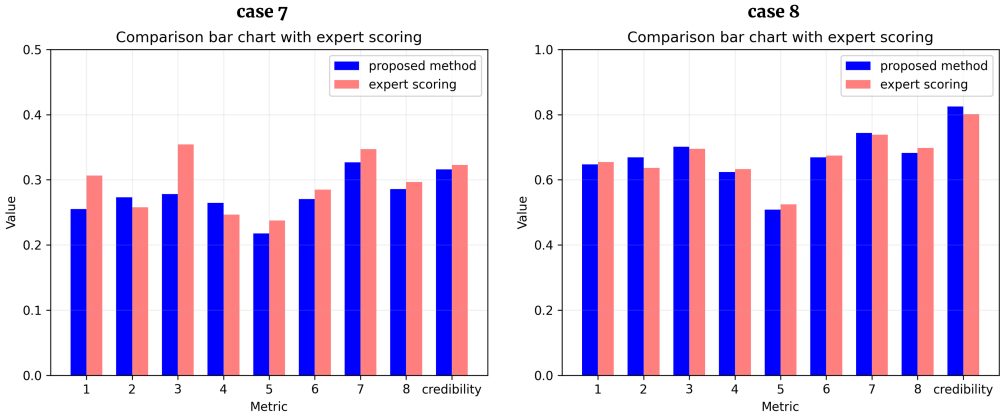

Figure 4. The bar chart comparing the metric values and credibility scores calculated for case 7 and case 8 with those of expert ratings.

metrics and credibility values closely match the expert scoring values in terms of overall trends. The average error for case 7 is 8.48% and for case 8 is 2%. Although there is still a discrepancy with the expert scores, the method proposed in this paper is a valuable qualitative model evaluation method. The weight coefficients $a_{\lambda 1}$ and ω_{jm} in Equation 1 and Equation 2 are derived by the stochastic hyper-heuristic-based differential evolution algorithm presented in Section 4. The weight coefficient ω_{jm} in Equation 2 reflects the degree of influence of the metrics on the credibility of the model, and the weight coefficient in Equation 1 $a_{\lambda 1}$ reflects the degree of influence of the compressed data on the metrics. These weights help determine which model outputs and evaluation metrics are more worthy of attention. This part of the research will be carried out in our subsequent work.

6. Conclusions

To achieve automated qualitative assessment for simulation model, this paper establishes a evolutionary algorithm-based evaluation fitting method using historical scoring cases as the reference. It utilizes the stochastic hyper-heuristic-based differential evolution algorithm to generate the optimal fitting formulae of the metrics and their weights, and then obtains the overall credibility value of a simulation model. New cases can be evaluated automatically using the fitting formulae and the weights. The experiments show that the average error of the test cases is within 8.5%, demonstrating the effectiveness of the method. The proposed method for qualitative evaluation of models effectively overcomes the subjectivity of traditional expert scoring and accelerates the evaluation process. In future work, we will focus on studying the impact of the weights in the metric fitting formula on model evaluation. Additionally, we will investigate the relation-

ship between these weights and the model's key outputs and important evaluation metrics.

7. Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62173017).

References

Ahn, J., de Weck, O. L., and Steele, M. (2014). Credibility assessment of models and simulations based on nasa's models and simulation standard using the delphi method. *Systems Engineering*, 17(2):237–248.

Balci, O. (1986). Credibility assessment of simulation results. In *Proceedings of the 18th conference on Winter simulation*, pages 38–44.

Foures, D., Albert, V., and Nketsa, A. (2016). A new specification-based qualitative metric for simulation model validity. *Simulation Modelling Practice and Theory*, 66:1–15.

Gao, D., Zhang, B., Xu, X., and Xiao, Y. (2019). Signed directed graph based simulation model validation framework for petrochemical process. In *2019 Chinese Control Conference (CCC)*, pages 7132–7136. IEEE.

Goerger, S. R., McGinnis, M. L., and Darken, R. P. (2005). A validation methodology for human behavior representation models. *The Journal of Defense Modeling and Simulation*, 2(1):39–51.

Hermann, C. F. (1967). Validation problems in games and simulations with special reference to models of international politics. *Behavioral science*, 12(3):216–231.

Ho, W. and Ma, X. (2018). The state-of-the-art integrations and applications of the analytic hierarchy process. *European Journal of Operational Research*, 267(2):399–414.

Law, A. M. (2022). How to build valid and credible simulation models. In *2022 Winter Simulation Conference (WSC)*, pages 1283–1295. IEEE.

Li, S., Peng, X., Peng, T., and Yang, C. (2016). A group evaluation method for complex simulation system credibility based on 2-order additive fuzzy measure. In *2016 Chinese Control and Decision Conference (CCDC)*, pages 147–154. IEEE.

Lu, L. and Yuan, Y. (2018). A novel topsis evaluation scheme for cloud service trustworthiness combining objective and subjective aspects. *Journal of Systems and Software*, 143:71–86.

Min, F.-Y., Yang, M., and Wang, Z.-C. (2010). Knowledge-based method for the validation of complex simulation models. *Simulation Modelling Practice and Theory*, 18(5):500–515.

Samlaus, R. and Fritzson, P. (2015). Semantic validation of physical models using role models. *Simulation*, 91(4):383–399.

Schruben, L. W. (1980). Establishing the credibility of simulations. *Simulation*, 34(3):101–105.

Slowik, A. and Kwasnicka, H. (2020). Evolutionary algorithms and their applications to engineering problems. *Neural Computing and Applications*, 32:12363–12379.

Tsinaslanidis, P. E. and Kugiumtzis, D. (2014). A prediction scheme using perceptually important points and dynamic time warping. *Expert Systems with Applications*, 41(15):6848–6860.

Wright, R. D. (1972). Validating dynamic models: An evaluation of tests of predictive power. In *Proceedings of 1972 Summer Computer Simulation Conference*, pages 13–16.

Zhang, B., Xu, X., Gao, D., Ma, X., and Wu, C. (2013). Model verification based on qualitative trend and sdg. *Ciesc Journal*.

Zhang, Z., Fang, K., and Yang, M. (2011). Method for complex simulation credibility evaluation based on group ahp. *Systems Engineering and Electronics*, 33(11):2569–2572.