Journal of Systems Engineering and Electronics
Vol. 37, No. 3, June 2026, pp.867 – 877

GPR-based model validation method for small samples

YANG Fan[1,2], MA Ping[1,2], ZHANG Huan[1,2], LI Wei[1,2,*], and YANG Ming[1,2]

1. Control and Simulation Center, Harbin Institute of Technology, Harbin 150080, China;
2. National Key Laboratory of Complex System Modeling and Simulation, Harbin 150080, China

Abstract: Validation for simulation models often confronts challenges with small samples due to the costs of time and money. To address this issue, this paper presents a validation method for small-sample dynamic outputs based on Gaussian process regression (GPR) models. Firstly, a validation framework based on Bayes statistics is proposed, shifting the focus from merely analyzing validation data to a more comprehensive analysis of posterior distributions. Subsequently, the posterior distributions of both the simulation outputs and the reference data are separately captured through segmented GPR. Then, the consistency of these posterior distributions is evaluated in terms of the central tendency and the distribution range. This consistency serves as a quantitative measure of the simulation model's credibility, expressed as a value ranging from 0 to 1, where a value closer to 1 indicates higher credibility. Finally, the effectiveness of this validation method is demonstrated through a numerical example and an application example, highlighting its capability in uncertainty description and adaptability to small samples.

Keywords: model validation, small samples, Gaussian process regression (GPR), Bayes statistics.

DOI: 10.23919/JSEE.2026.000113

1. Introduction

Nowadays, simulation technology has entered into a new stage focused on complex system simulations [1]. Due to the complexity of the simulation models, issues such as lack of reference data, coupling between sub-models, uncertainty and variety of outputs have raised concerns about model validation [2]. Validating simulation results is a common method for assessing the credibility of the simulation models by measuring the consistency between the simulation outputs and the reference data [3,4].

To validate the simulation results under uncertainty, it is necessary to analyze the consistency of two sets of multi-sample validation data. Various methods have been proposed to solve this problem, including validation methods based on copulas [5], Bayesian networks [6], differences in multivariate probability distributions [7], grey cloud clustering [8], and improved Bayesian factor [9]. However, due to prolonged experimental cycles, substantial costs, and intricate processes, the outputs of both real systems and simulation models are typically small samples, which makes these methods unsuitable. For instance, in experiments such as airplane flights and rocket launches, the sample size of validation data is significantly limited.

Small-sample data is defined as the data with a sample size not exceeding 30 [10]. Various scholars have explored simulation validation techniques designed for small samples, but the broader practical application of these methods remains somewhat limited. The validation method based on bootstrap and Bayes parameter estimation [11], as well as the validation method based on the clustered cloud model [12], are only applicable to static outputs. The Bayesian inference-based method focuses on assessing the credibility at the system level [7]. The method based on transfer learning and discrete sequence generative adversarial networks could be used to validate the small-sample dynamic outputs [13]. However, it is hard to meet the data dependency required by transfer learning in practice. There is still a lack of effective validation methods for small-sample dynamic outputs.

Validation methods for dynamic outputs have evolved from directly measuring the original data to approaches based on data features or even data models. Common methods of measuring original data include grey relational analysis (GRA), and theil inequality coefficient (TIC) [14]. However, directly measuring the consistency of original data often reveals only superficial similarities and fails to capture the intrinsic characteristics of the data. Additionally, this approach can be significantly influenced by the uncertainty in the data. Validation methods based on data features assess the consistency of the features such as mean, standard deviation, kurtosis, period, autocorrelation coefficient, and spectral density

Manuscript received May 29, 2024.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (62273119).

[9], which are designed and selected according to the specific tasks and types of data involved. Nevertheless, it is challenging to propose guidelines for selecting features for specific systems. Model-based validation methods, in contrast, compare the intrinsic structure and patterns of the data, providing a deeper level of consistency information [15]. These methods can effectively learn the intrinsic structure and capture underlying trends, even in small samples. Therefore, this paper aims to explore a model-based validation method for small-sample dynamic outputs.

The representative models used for time-series modeling and validation mainly include the autoregressive (AR) model [16], the vector AR moving average (VARMA) model [17], the Markov chain model [18], the hidden Markov model [19], and the Gaussian process regression (GPR) model [20]. GPR exhibits strong performance in modeling data uncertainty and complex nonlinear relationships [21,22]. By employing various kernel functions, it can effectively analyze different types of time series data [23]. Thus, we model both the simulation outputs and the reference data by GPR and analyze the consistency between the GPR models.

This paper proposes a validation method based on GPR for small-sample dynamic outputs. The remainder of this paper is structured as follows: Section 2 primarily describes and analyzes the problem of validating small-sample dynamic outputs. Section 3 elaborates a small-sample validation method based on GPR. Within this section, Subsection 3.1 establishes a validation framework based on Bayes statistics, outlining a two-stage validation process that includes segmented GPR and consistency measurement. These two stages are thoroughly described in Subsection 3.2 and Subsection 3.3, respectively. Section 4 provides two examples to demonstrate the effectiveness of the proposed method in describing uncertainty and validating small samples. Finally, the conclusions and future work are presented in Section 5.

2. Problem description and analysis

The outputs of real systems and simulation models often exhibit uncertainty due to variations in input parameters and the intrinsic randomness of the system. In such cases, it is inaccurate to validate uncertain simulation results based on the outputs of a single experiment. To address this issue, multiple experiments are typically conducted, allowing for the validation of simulation results that include multi-sample simulation outputs and multi-sample reference data. However, complex simulation models cannot be repeated frequently due to high computational costs and extensive time consumption. Moreover, the number of experiments that can be conducted on many real systems is even more limited, considering the costs in terms of money and time. This leads to a common issue: the validation data are often small samples. Hence, our aim is to validate univariate, dynamic, and small-sample outputs.

Suppose that R denotes the real system while S denotes the simulation model. Let R and S run N_R times and N_S ($N_R, N_S \leqslant 30$) times respectively. $\boldsymbol{Y}_R = \left[\boldsymbol{y}_R^1, \boldsymbol{y}_R^2, \cdots, \boldsymbol{y}_R^{N_R}\right]$ is an $N_R \times M$ matrix which represents N_R samples of reference data across M timestamps. Similarly, $\boldsymbol{Y}_S = \left[\boldsymbol{y}_S^1, \boldsymbol{y}_S^2, \cdots, \boldsymbol{y}_S^{N_S}\right]$ represents N_S samples of simulation outputs across M timestamps. A single sample $\boldsymbol{y}^n$ ($n = 1, 2, \cdots, N_R + N_S$) within $\boldsymbol{Y}_R$ or $\boldsymbol{Y}_S$ can be denoted as $\boldsymbol{y}^n = \left[y_{t_1}^n, y_{t_2}^n, \cdots, y_{t_M}^n\right]$ where $t_1, t_2, \cdots, t_M$ are the time points.

Simulation result validation methods obtain the simulation credibility by measuring the consistency between the simulation outputs and the reference data. Suppose that $C(\boldsymbol{Y}_S, \boldsymbol{Y}_R)$ denotes the simulation credibility, and let $C(\boldsymbol{Y}_S, \boldsymbol{Y}_R) \in [0,1]$. $C(\boldsymbol{Y}_S, \boldsymbol{Y}_R) = 1$ when $\boldsymbol{Y}_S$ and $\boldsymbol{Y}_R$ are completely consistent, and $C(\boldsymbol{Y}_S, \boldsymbol{Y}_R) \to 0$ as the consistency between $\boldsymbol{Y}_S$ and $\boldsymbol{Y}_R$ diminishes. Since model-based validation methods capture the intrinsic characteristics of the validation data and require only small-sample data [24], we intend to explore a model-based validation method for small-sample dynamic outputs.

Subsequently, we will study a model-based validation method for small-sample dynamic outputs from two aspects:

(i) We will select an appropriate regression method to model the small-sample dynamic outputs and to capture both the intrinsic characteristics and the uncertainty in the validation data.

GPR is not only applicable to modeling complex data, such as nonlinear and non-smooth data, but also capable of describing the uncertainty in data [25,26]. Hence, the relationship between the outputs and their influences can be fitted by GPR to capture the intrinsic structure and patterns of the validation data. Assuming that factors other than time do not influence the outputs, and given two sets of validation data and the corresponding timestamps, we employ two distinct GPR models to individually describe the reference data $\boldsymbol{Y}_R$ and the simulation outputs $\boldsymbol{Y}_S$. The elements in $\boldsymbol{Y}_R$ and $\boldsymbol{Y}_S$ can be respectively regarded as $N_R \times M$ samples and $N_S \times M$ samples obtained from their corresponding models.

(ii) We will measure the consistency between the simulation outputs and the reference data based on the characteristics captured by the GPR models.

The posterior distributions obtained by GPR describe both the intrinsic characteristics and the uncertainty in the data. According to Bayesian statistical method [27,28],

we use the posterior distributions to make inferences about the consistency between two sets of validation data.

3. Model validation method based on GPR

3.1 Validation framework based on Bayes statistics

In response to the above two aspects, we propose a small-sample validation framework based on Bayesian statistics, as shown in Fig. 1, which mainly consists of two parts: segmented GPR and consistency measurement.

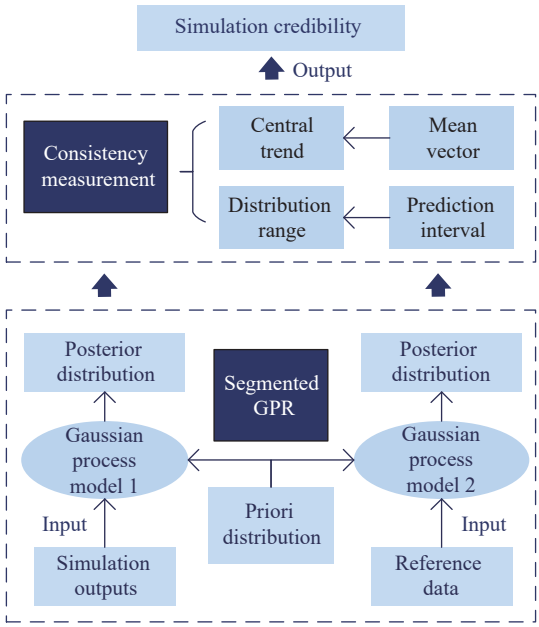


Fig. 1 Validation framework for small samples

We adopt GPR models, suitable for modeling small-sample time series, to model the validation data. Given the same prior distribution, we model the simulation outputs and the reference data separately by segmented GPR and obtain the posterior distributions.

The consistency between the simulation outputs and the reference data is assessed by comparing the consistency of their posterior distributions in two ways. Firstly, we compare the central tendencies of the two sets of data support by measuring the consistency of the posterior mean vectors. Secondly, we compare the distribution ranges of the two sets of data exhibit by measuring the consistency of the prediction intervals.

3.2 Segmented GPR for modeling the validation data

3.2.1 Applicability of GPR for modeling small-sample time series

A Gaussian process (GP) is a stochastic process composed of Gaussian random variables, where any finite number of these variables follow a joint Gaussian distribution, as shown in Fig. 2.

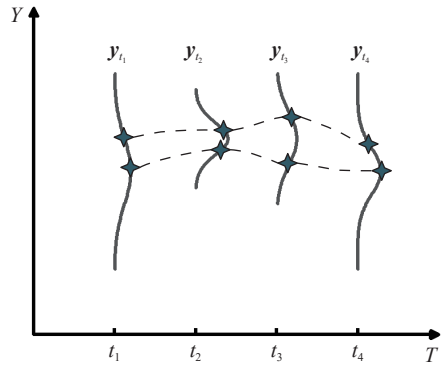


Fig. 2 Overview of GP

According to the central limit theorem, the superposition of a large number of independent random variables exhibits a normal distribution when influenced by multiple small, independent uncertainty factors [29]. Therefore, the output data obtained from a single run of a simulation model or a real system can usually be regarded as a single realization of a GP. When the response of a real system or a simulation model is a single dynamic variable with uncertainty, the outputs obtained by repeatedly running are multi-sample time series with the data at each time point following a Gaussian distribution. GPR is a regression algorithm based on a Bayesian approach, relying on observations to model and predict data through the definition of a GP prior [30]. Hence, we could train two independent GPR models to model the simulation outputs and the reference data.

A GP is specified by a mean function $\mu(t)$ and a covariance function $k(t,t')$, thereby the validation data are modeled as follows:

$$\begin{cases} y_t^n = f(t) + \varepsilon \\ f \sim \mathrm{GP}(\mu(t), k(t,t')) \end{cases} \tag{1}$$

where y_t^n is the element in $\boldsymbol{y}^n$ corresponding to the time point t; t and t' are any two points in $[t_1, t_2, \cdots, t_M]$; f is the unknown function which models the potential relationship between the validation data and time; $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ is the variance, assumed to be independently and identically distributed at different points, indicating the uncertainty in the validation data. The mean function $\mu(t)$ defines the expected value of the function given the input t and is often assumed to be constant or a linear function. The covariance function $k(t,t')$ (also known as the kernel function) defines the covariance between the function values at any two time points t and t', capturing the dependencies and structure in the data.

GPR infers the posterior distribution from the prior distribution and the observations. The mean vector and covariance matrix of the posterior distribution can be considered as the characteristics of the time series. The mean vector provides the optimal estimate for each time point and represents the central tendency of the time series. The covariance matrix describes the correlation between the time points and reflects the uncertainty in the time series.

3.2.2 Modeling process of validation data based on segmented GPR

To model the small-sample validation data based on the one-dimensional GPR model, the data need to be preprocessed firstly. Subsequently, we apply the same prior distribution to the unknown functions, selecting an appropriate kernel function based on the characteristics of the validation data. Next, we train and optimize two distinct GPR models, adjusting the model parameters accordingly. After that, two sets of posterior distributions are obtained, which are then used for further consistency analysis. The process is shown in Fig. 3.

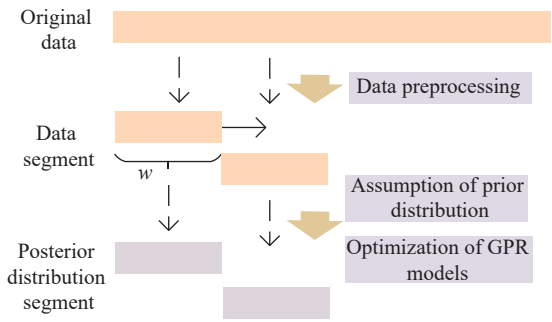


Fig. 3 Process of segmented modeling

Data preprocessing includes three parts: matrix transformation, normalization and segmentation. Since we model the validation data based on the one-dimensional GPR, before being input into the regression model, $\boldsymbol{Y}_R$ needs to be converted from an $N_R \times M$ matrix to a vector of length $N_R M$, denoted as $\boldsymbol{Y}'_R$. Similarly, $\boldsymbol{Y}_S$ needs to be converted from an $N_S \times M$ matrix to a vector of length $N_S M$, denoted as $\boldsymbol{Y}'_S$. Two sets of validation data and their corresponding time data, $\boldsymbol{T}'_R$ and $\boldsymbol{T}'_S$ are as follows:

$$\begin{cases} \boldsymbol{Y}'_R = \left[y^1_{Rt_1}, y^2_{Rt_1}, \cdots, y^{N_R}_{Rt_1}, \cdots, y^1_{Rt_M}, y^2_{Rt_M}, \cdots, y^{N_R}_{Rt_M}\right]_{(1\times N_R M)} \\ \boldsymbol{T}'_R = [t_1, \cdots, t_1, \cdots, t_M, \cdots, t_M]_{(1\times N_R M)} \end{cases}, \tag{2}$$

$$\begin{cases} \boldsymbol{Y}'_S = \left[y^1_{St_1}, \cdots, y^{N_S}_{St_1}, \cdots, y^1_{St_M}, \cdots, y^{N_S}_{St_M}\right]_{(1\times N_S M)} \\ \boldsymbol{T}'_S = [t_1, \cdots, t_1, \cdots, t_M, \cdots, t_M]_{(1\times N_S M)} \end{cases}, \tag{3}$$

where $y^{n_1}_{Rt}(n_1 = 1,2,\cdots,N_R, t = t_1, t_2, \cdots, t_M)$ is the element in the n_1th sample of reference data corresponding to the time point t; and $y^{n_2}_{St}(n_2 = 1,2,\cdots,N_S)$ is the element in the n_2th sample of simulation outputs corresponding to the time point t.

Additionally, in order to improve model performance and stability, normalization is performed on $\boldsymbol{Y}'_R$ and $\boldsymbol{Y}'_S$ as follows:

$$y^i_{R\text{norm}} = \frac{y^i_R - \min\left\{y^i_R, y^j_S\right\}}{\max\left\{y^i_R, y^j_S\right\} - \min\left\{y^i_R, y^j_S\right\}}, \tag{4}$$

$$y^j_{S\text{norm}} = \frac{y^j_S - \min\left\{y^i_R, y^j_S\right\}}{\max\left\{y^i_R, y^j_S\right\} - \min\left\{y^i_R, y^j_S\right\}}, \tag{5}$$

where $y^i_{R\text{norm}}$ and $y^i_{S\text{norm}}$ are the elements in the normalized validation data $\boldsymbol{Y}_{R\text{norm}}$ and $\boldsymbol{Y}_{S\text{norm}}$; $i = 1,2,\cdots,N_R M$; $j = 1,2,\cdots,N_S M$; y^i_R and y^j_S are the elements in the validation data $\boldsymbol{Y}'_R$ and $\boldsymbol{Y}'_S$; $\min\left\{y^i_R, y^j_S\right\}$ denotes the minimum element in $\boldsymbol{Y}'_R$ and $\boldsymbol{Y}'_S$; and $\max\left\{y^i_R, y^j_S\right\}$ denotes the maximum element in $\boldsymbol{Y}'_R$ and $\boldsymbol{Y}'_S$.

Due to the numerous matrix inverse operations involved in the training process of the GPM, the calculations would increase with the size of the dataset. The complexity of GPR is $O(l^3)$ when the length of the sequence is l [31]. The validation data in simulation validation problem are usually long time series which would result in a large amount of computation and long time-consumption. To solve this problem, we perform segmented GPR with a sliding window for data segmentation.

The size of the sliding window mainly depends on the characteristics of the data itself. If the time series is relatively smooth or exhibits no obvious trend changes, the window can be set to a fixed size based on the amount of data. For example, for periodic data, the size of the window can be determined according to the period of the data. If the time series shows an obvious trend change, the data can be divided according to the trend changes at different stages using windows of varying sizes. Positions in the time series where there is a clear change in trend or a jump in value are reasonable segmentation points, and the window size can be adjusted accordingly. Additionally, the number of samples in each segment should not be too small; otherwise, the fitting effect of GPR may be suboptimal. On the other hand, the number of samples in each segment should not be too large, as it may affect computational efficiency. Typically, the number of samples per segment can range between 50 and 200. This ensures both a good fitting effect and manageable computational cost.

Assuming that the length of the window is w timestamps, we slide w timestamps each time so that two sets

of normalized validation data, $\boldsymbol{Y}_{R\text{norm}}$ and $\boldsymbol{Y}_{S\text{norm}}$, are both divided into q segments. The xth segment of the validation data, $\boldsymbol{Y}_{Rx}$ and $\boldsymbol{Y}_{Sx}$, and their corresponding time data, $\boldsymbol{T}_{Rx}$ and $\boldsymbol{T}_{Sx}$, are as follows:

$$\begin{cases}\boldsymbol{Y}_{Rx}=\left[y_{R\text{norm}}^{xw+1},\cdots,y_{R\text{norm}}^{xw+N_Rw}\right]_{(1\times N_Rw)}\\ \boldsymbol{T}_{Rx}=[t_{xw+1},\cdots,t_{xw+1},\cdots,t_{xw+w},\cdots,t_{xw+w}]_{(1\times N_Rw)}\end{cases}, \tag{6}$$

$$\begin{cases}\boldsymbol{Y}_{Sx}=\left[y_{S\text{norm}}^{xw+1},\cdots,y_{S\text{norm}}^{xw+N_Sw}\right]_{(1\times N_Sw)}\\ \boldsymbol{T}_{Sx}=[t_{xw+1},\cdots,t_{xw+1},\cdots,t_{xw+w},\cdots,t_{xw+w}]_{(1\times N_Sw)}\end{cases}, \tag{7}$$

where $y_{R\text{norm}}^{i}(xw+1\leqslant i\leqslant xw+N_Rw,i\in\mathbb{N})$ is the ith element in $\boldsymbol{Y}_{R\text{norm}}$; $x=1,2,\cdots,q$, $q=\dfrac{M}{w}$; and $y_{S\text{norm}}^{j}(xw+1\leqslant j\leqslant xw+N_Sw,\ \ j\in\mathbf{N})$ is the jth element in $\boldsymbol{Y}_{S\text{norm}}$.

Let the unknown functions corresponding to $(\boldsymbol{T}_{Rx},\boldsymbol{Y}_{Rx})$ and $(\boldsymbol{T}_{Sx},\boldsymbol{Y}_{Sx})$ be denoted as f_{Rx} and f_{Sx}, then we aim to learn the posterior distributions of them. Due to the same methods and processes for modeling each segment of the validation data, including the reference data and the simulation outputs, the following is a unified description. The xth segment of the validation data, the time data and the function are respectively denoted as $\boldsymbol{Y}_x$, $\boldsymbol{T}_x$ and f_x, and N is the sample size of the validation data.

We need to define the prior distribution of the function after data preprocessing. Let the prior distribution of the unknown function f_x be a GP with zero mean function and a kernel function, denoted as $f_x\sim\mathrm{GP}(0,k_x(t,t'))$. GPR effectively learns complex trends and nonlinear relationships in data by utilizing appropriate kernel functions, making it adaptable to a wide variety of data forms. Through the selection of suitable kernel functions, GPR can capture intricate relationships in time series data, including periodicity, trends, and self-similarity. Common kernel functions include the Gaussian kernel (also known as the radial basis function, RBF), Matern kernel, linear kernel, polynomial kernel, and periodic kernel. The Gaussian kernel is particularly well-suited for modeling locally smooth and nonlinear time series data:

$$k_{\mathrm{RBF}}(t,t')=\sigma_1^2\exp\left(-\frac{\|t-t'\|^2}{2l_1^2}\right) \tag{8}$$

where σ_1^2 is the variance, controlling the vertical variation of the function; and l_1 is the length scale parameter, controlling the horizontal smoothness of the function.

And it often satisfies the fitting requirements of validation data. However, when the validation data exhibit significant fluctuations or abrupt changes, the Matern kernel is more appropriate:

$$k_{\mathrm{Matern}}(t,t')=\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}|t-t'|}{l_2}\right)^{\nu}K_{\nu}\left(\frac{\sqrt{2\nu}|t-t'|}{l_2}\right) \tag{9}$$

where ν is the smoothness parameter; $\Gamma(\nu)$ is the Gamma function; l_2 is the length scale; and K_ν is the modified Bessel function.

In addition, other kernel functions, such as the linear kernel, polynomial kernel, and periodic kernel, can be chosen based on prior knowledge or observed trends in the data.

The prior distribution is used to initialize the GPR models. Then we can optimize the GPR model with the preprocessed validation data $\boldsymbol{Y}_x$ and the corresponding time data $\boldsymbol{T}_x$ as the training data. To ensure the GPR models accurately fits the validation data, the model's parameters, which include the hyperparameters in the kernel function $k_x(t,t')$ and the variance σ^2, are meticulously optimized by the maximum likelihood method. This optimization process involves adjusting the parameters through a gradient descent algorithm, iteratively refining them until the parameters that maximize the log marginal likelihood (LML) are identified. This method improves the reliability and effectiveness of the GPR models in capturing the essential characteristics of the validation data. LML is calculated as follows:

$$\begin{aligned}\lg p(\boldsymbol{Y}_x|\boldsymbol{T}_x,\theta_x)=-\frac{1}{2}\Big(&\boldsymbol{Y}_x^{\mathrm{T}}\left(\boldsymbol{K}_x+\sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{Y}_x+\\&\lg\left|\boldsymbol{K}_x+\sigma^2\boldsymbol{I}\right|+wN\lg(2\pi)\Big)\end{aligned} \tag{10}$$

where $\boldsymbol{Y}_x$ and $\boldsymbol{T}_x$ are the output and input of the GPR model; θ_x denotes the parameters to be optimized; $\boldsymbol{K}_x$ is the covariance matrix of the training time points $\boldsymbol{T}_x$, consisting of the elements calculated by $k_x(t,t'),t,t'\in\boldsymbol{T}_x$; and σ^2 is the variance.

After that, we can input test time points into the optimized model to obtain the posterior distribution. Due to our desire to capture the essential characteristics of the validation data, we make the test time points $\boldsymbol{T}_x^*$ consist of the non-repetitive timestamps corresponding to the validation data, as follows:

$$\boldsymbol{T}_x^*=[t_{xw+1},t_{xw+2},\cdots,t_{xw+w}]_{(1\times w)}. \tag{11}$$

The posterior distribution of the outputs, denoted as $\boldsymbol{y}_x^*$, follows a Gaussian distribution with posterior mean $\boldsymbol{\mu}_x^*$ and posterior covariance $\boldsymbol{\Sigma}_x^*$:

$$\boldsymbol{y}_x^*|\boldsymbol{T}_x^*,\boldsymbol{T}_x,\boldsymbol{Y}_x\sim\mathrm{N}(\boldsymbol{\mu}_x^*,\boldsymbol{\Sigma}_x^*),$$

$$\boldsymbol{\mu}_x^*=\boldsymbol{K}_x^*\left(\boldsymbol{K}_x+\sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{Y}_x, \tag{12}$$

$$\boldsymbol{\Sigma}_x^*=\boldsymbol{K}_x^{**}-\boldsymbol{K}_x^*\left(\boldsymbol{K}_x+\sigma^2\boldsymbol{I}\right)^{-1}(\boldsymbol{K}_x^*)^{\mathrm{T}}+\sigma^2\boldsymbol{I}, \tag{13}$$

where $\boldsymbol{K}_x^*$ is the covariance between the test time points $\boldsymbol{T}_x^*$ and the training time points $\boldsymbol{T}_x$; and $\boldsymbol{K}_x^{**}$ is the covariance among the test time points $\boldsymbol{T}_x^*$.

3.3 Consistency measurement for model validation

Following Bayes statistics method, we place the same prior distribution $\mathrm{GP}(0,k_x(t,t'))$ on the unknown functions f_{Sx} and f_{Rx}; then, given the xth segment of the validation data and their time data $(\boldsymbol{T}_{Sx},\boldsymbol{Y}_{Sx})$ and $(\boldsymbol{T}_{Rx},\boldsymbol{Y}_{Rx})$, we obtain their posterior distributions:

$$\boldsymbol{y}^*_{Rx}\sim \mathrm{N}(\boldsymbol{\mu}^*_{Rx},\boldsymbol{\Sigma}^*_{Rx}),$$

$$\boldsymbol{y}^*_{Sx}\sim \mathrm{N}(\boldsymbol{\mu}^*_{Sx},\boldsymbol{\Sigma}^*_{Sx}).$$

Next, we make inferences about the consistency between $\boldsymbol{Y}_{Sx}$ and $\boldsymbol{Y}_{Rx}$ by comparing the posterior distributions; finally, the validation results for each segment of data are synthesized to indicate the simulation model's credibility.

The posterior mean vector represents the central tendency of the validation data, while the prediction interval reflects the range of data distribution, taking into account the uncertainty of the validation data. If two posterior distributions are similar in terms of central tendency and distribution range, it can be inferred that both the intrinsic characteristics and the uncertainty in two sets of validation data are consistent. Thus, we simplify the task of comparing the consistency of posterior distributions to quantifying the consistency of mean vectors and prediction intervals.

We compare the central tendencies of the function behavior corresponding to the simulation outputs and the reference data by comparing the posterior mean vectors. The Euclidean distance and cosine distance can respectively measure differences in position and direction between two vectors, respectively. Hence, to assess the consistency of the two posterior mean vectors $\boldsymbol{\mu}^*_{Sx}$ and $\boldsymbol{\mu}^*_{Rx}$, we comprehensively measure the differences between them by Euclidean distance and cosine distance, as follows:

$$D_{\mathrm{eu}-x}=\left\|\boldsymbol{\mu}^*_{Sx}-\boldsymbol{\mu}^*_{Rx}\right\|_2, \tag{14}$$

$$D_{\cos-x}=\frac{\boldsymbol{\mu}^*_{Sx}\cdot\boldsymbol{\mu}^*_{Rx}}{\left\|\boldsymbol{\mu}^*_{Sx}\right\|\left\|\boldsymbol{\mu}^*_{Rx}\right\|}. \tag{15}$$

We aim to map the range of consistency between the two posterior mean vectors to [0,1]. The smaller the $D_{\mathrm{eu}-x}$, the higher the degree of consistency. The larger the $D_{\cos-x}$, the higher the degree of consistency. Considering that $D_{\mathrm{eu}-x}\in[0,+\infty]$ and $D_{\cos-x}\in[-1,1]$, we convert the distance metrics as follows:

$$C_{mx}=\frac{1}{2}\left(\frac{1}{1+D_{\mathrm{eu}-x}}+\frac{1+D_{\cos-x}}{2}\right) \tag{16}$$

where C_{mx} denotes the consistency of the posterior mean vectors.

We compare the coverage range of two prediction intervals to quantify and compare the uncertainty in the simulation outputs and the reference data. For computational simplicity and intuitive understanding, we consider the area of the prediction interval as the sum of the widths of the intervals at each point. Furthermore, we use the ratio of the intersection area and union area to indicate the consistency of the prediction intervals $C_{\mathrm{PI}x}$, as follows:

$$C_{\mathrm{PI}x}=\frac{\mathrm{Area}_{Ix}}{\mathrm{Area}_{Ux}}, \tag{17}$$

$$\mathrm{Area}_{Ix}=\sum\max(\min(U_{Sd},U_{Rd})-\max(L_{Sd},L_{Rd}),0), \tag{18}$$

$$\mathrm{Area}_{Ux}=\sum((U_{Sd}-L_{Sd})+(U_{Rd}-L_{Rd}))-\mathrm{Area}_{Ix}, \tag{19}$$

where Area_{Ix} denotes the intersection area; Area_{Ux} denotes the union area; $[L_{Sd},U_{Sd}]$ and $[L_{Rd},U_{Rd}]$ are the prediction intervals at time point $t_d(d=xw+1,xw+2,\cdots,xw+w)$ corresponding to the simulation outputs and the reference data:

$$\begin{cases}L_{Sd}=\mu_{Sd}-z_{\alpha/2}\cdot\sigma_{Sd}\\ U_{Sd}=\mu_{Sd}+z_{\alpha/2}\cdot\sigma_{Sd}\end{cases}, \tag{20}$$

$$\begin{cases}L_{Rd}=\mu_{Rd}-z_{\alpha/2}\cdot\sigma_{Rd}\\ U_{Rd}=\mu_{Rd}+z_{\alpha/2}\cdot\sigma_{Rd}\end{cases}, \tag{21}$$

where $z_{\alpha/2}$ is the z-score corresponding to the $\alpha/2$ quantile of the standard normal distribution, associated with a confidence level $(1-\alpha)\times100\%$; μ_{Sd} and μ_{Rd} are the posterior means at t_d, which are derived from the posterior mean vectors $\boldsymbol{\mu}^*_{Sx}$ and $\boldsymbol{\mu}^*_{Rx}$; σ_{Sd} and σ_{Rd} are the posterior standard deviations at t_d, which are computed as the square roots of the diagonal elements of the posterior covariance matrices $\boldsymbol{\Sigma}^*_{Sx}$ and $\boldsymbol{\Sigma}^*_{Rx}$, respectively.

Then, the validation result of the xth segment of the validation data, denoted as $C(\boldsymbol{Y}_{Sx},\boldsymbol{Y}_{Rx})$, can be calculated by

$$C(\boldsymbol{Y}_{Sx},\boldsymbol{Y}_{Rx})=\omega_1C_{mx}+\omega_2C_{\mathrm{PI}x} \tag{22}$$

where ω_1 and ω_2 control the contribution of two parts of consistency to the validation result; $\omega_1+\omega_2=1,0<\omega_1,\omega_2<1$. In general, assessors prioritize the central tendency of the validation data over its uncertainty. Therefore, ω_1 is typically assigned a larger value than ω_2. The higher the consistency requirement for uncertainty, the

larger ω_2.

Since we have completed segmented model validation, we synthesize the validation result of each segment $C(\boldsymbol{Y}_{Sx},\boldsymbol{Y}_{Rx})$ $(x=1,2,\cdots,q)$ into the final validation result $C(\boldsymbol{Y}_S,\boldsymbol{Y}_R)$ based on the weighted sum method, as follows:

$$C(\boldsymbol{Y}_S,\boldsymbol{Y}_R)=\sum_{x=1}^{q}\lambda_x C(\boldsymbol{Y}_{Sx},\boldsymbol{Y}_{Rx}) \tag{23}$$

where $\lambda_x(x=1,2,\cdots,q)$ denote the weight of $C(\boldsymbol{Y}_{Sx},\boldsymbol{Y}_{Rx})$, generally set equal to $1/q$. Besides, the weights could be adjusted if we focus more on a certain segment of the outputs.

4. Experimental studies

4.1 Description of uncertainty

In this section, we evaluate the effectiveness of the proposed validation method for uncertainty description by a numerical example.

Given two simulation models:

$$\begin{cases} y_1 = 1 + a_1 \sin(0.2\pi x) + b_1 \\ y_2 = 1 + a_2 \sin(0.2\pi x) + b_2 \end{cases} \tag{24}$$

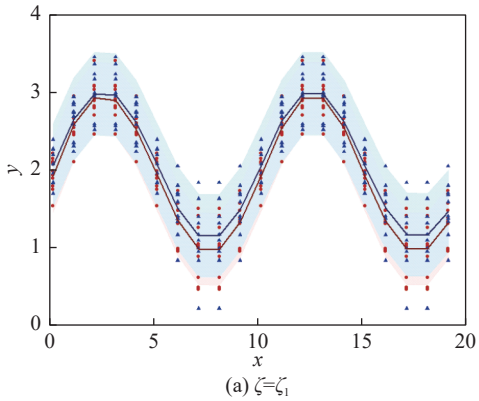
where $0 \leqslant x \leqslant 19, x \in \mathbf{R}$. We set the parameters of the two simulation models $\boldsymbol{\zeta}=[a_1,a_2,b_1,b_2]$ as

$$\boldsymbol{\zeta}_1 :\ a_1,a_2 \sim \mathrm{N}(1,0.25), b_1,b_2 \sim \mathrm{N}(1,0.25),$$

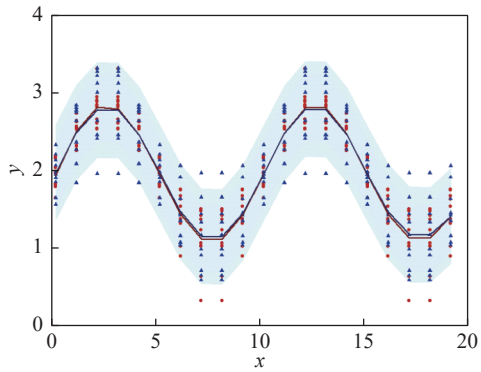
$$\boldsymbol{\zeta}_2 :\ a_1 \sim \mathrm{N}(1,0.25), a_2 \sim \mathrm{N}(1,0.5), b_1,b_2 \sim \mathrm{N}(1,0.25),$$

$$\boldsymbol{\zeta}_3 :\ a_1,a_2 \sim \mathrm{N}(1,0.25), b_1 \sim \mathrm{N}(1,0.5), b_2 \sim \mathrm{N}(1,0.25),$$

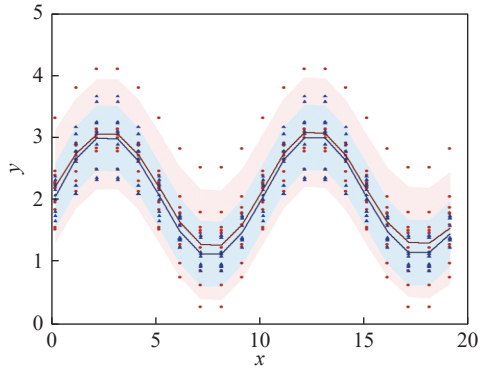
and carry out three sets of experiments. The two models are repeatedly run 10 times to get two sets of observations in each experiment. After training their corresponding GPR models, the posterior distributions are obtained, as shown in Fig. 4.



(a) ζ=ζ_1



(b) ζ=ζ_2



(c) ζ=ζ_3

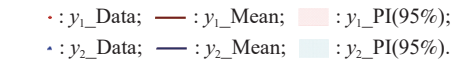


Fig. 4 Prediction intervals of two sets of observations

Each parameter in ζ follows a Gaussian distribution, and the larger the variance the greater the uncertainty in the parameter. As seen in Fig. 4, the prediction interval is larger when the uncertainty in the model parameters is greater. Hence, the GPR model is able to effectively describe the uncertainty in the observations.

4.2 Validation for small samples

In this section, we evaluate the effectiveness of the proposed validation method for small-sample validation by an application example on a force load identification model of a flight vehicle. Considering the strain error, we validate the dynamic outputs under uncertainty, focusing on the axial force F as the target variable.

Let the model be run 20 times with a random strain error of 1% to obtain 20 validation data samples, with 10 of the samples designated as a set of reference data and the other 10 as a set of simulation outputs. Additionally, let the model be run 10 times with a random strain error of 3% to obtain 10 samples as another set of simulation outputs. The reference data are labeled as "Data_R" and two sets of simulation outputs corresponding to 1% and 3% strain errors are labeled as "Data_S1"and "Data_S2", respectively. The information about the validation data is shown in Table 1.

Table 1 Information of validation data

Strain error/%	Sample size	Data type	Data label
1	10	Reference	Data_R
	10	Simulation	Data_S1
3	10	Simulation	Data_S2

We use axial force data from 1.01 s to 2.00 s as the validation data, which are dynamic outputs at intervals of 0.01 s. Each run produces a time series spanning 100 timestamps. Five samples of validation data each with 1% and 3% strain error are randomly selected, and the distribution of the data is shown in Fig. 5.

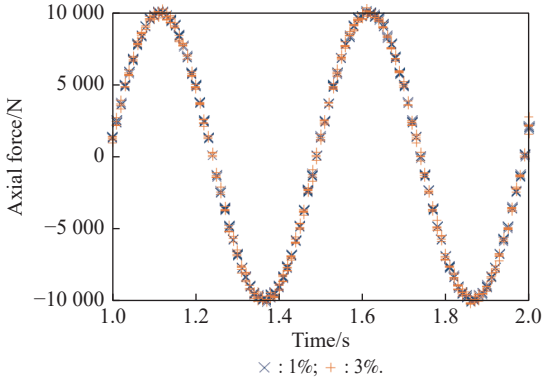
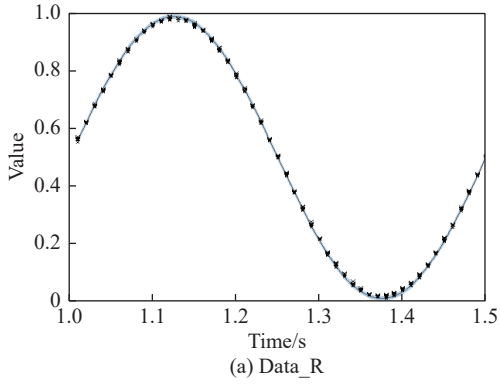


Fig. 5 Original data

As shown in Fig. 5, the axial force data exhibits periodic behavior with a period of 0.50 s. Additionally, the larger the strain error, the greater the uncertainty in the model output. Then, we validate the two sets of simulation outputs separately and show the validation process. For simplicity and efficiency, we set the sliding window size at 50 and adopt a Gaussian kernel function to fit the validation data. After training the GPR models, we obtain the posterior distributions. The posterior distributions of the first segment of reference data and two sets of simulation outputs are shown in Fig. 6.



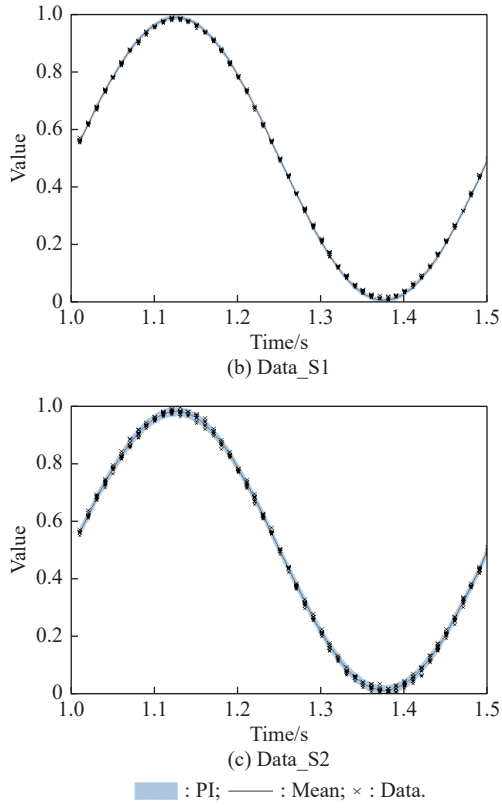


Fig. 6 Posterior distributions (α=0.05)

Fig. 6 shows that the range of prediction intervals of "Data_S2" is significantly larger. This again indicates that the prediction intervals effectively reflect the uncertainty in the observations.

After that, we can quantify the consistency of the posterior mean vectors and prediction intervals by (17)−(24). Next, the validation results of each segment of data are calculated by (25) with $\omega_1 = 0.7$, $\omega_2 = 0.3$. Furthermore, let $\lambda_1 = \lambda_2 = 0.5$ in (26) and obtain the final validation results, as seen in Table 2.

Table 2 Validation results for two sets of simulation outputs

Simulation output	Result	The 1st segment	The 2nd segment	$C(\boldsymbol{Y}_S, \boldsymbol{Y}_R)$
Data_S1	C_{mx}	0.997 9	0.998 0	0.978 2
	$C_{\mathrm{PI}x}$	0.933 4	0.930 7	
	$C(\boldsymbol{Y}_{Sx}, \boldsymbol{Y}_{Rx})$	0.978 6	0.977 8	
Data_S2	C_{mx}	0.994 1	0.994 9	0.832 7
	$C_{\mathrm{PI}x}$	0.439 0	0.471 4	
	$C(\boldsymbol{Y}_{Sx}, \boldsymbol{Y}_{Rx})$	0.827 5	0.837 9	

According to the information in Table 1 and the results shown in Table 2, it is evident that the credibility of the simulation decreases as the divergence in parameter distributions increases. This observation is consistent with theoretical expectations, thereby confirming the effective-

ness of the proposed validation method in achieving reasonable results.

To demonstrate that segmented GPR has minimal or no impact on the overall validation results, validations are conducted using "Data_S2" as the simulation outputs in two cases with window sizes of 20 and 50. The corresponding validation results and times are presented in Table 3.

Table 3 Validation results under different sliding windows

Window size	Result	The 1st segment	The 2nd segment	The 3rd segment	The 4th segment	The 5th segment	$C(\boldsymbol{Y}_S,\boldsymbol{Y}_R)$	Time/s
20	C_{mx}	0.996 4	0.995 5	0.994 4	0.998 3	0.996 3	0.833 5	7.02
	C_{PIx}	0.445 3	0.481 6	0.411 8	0.489 9	0.440 3		
	$C(\boldsymbol{Y}_{Sx},\boldsymbol{Y}_{Rx})$	0.831 1	0.841 3	0.819 6	0.845 8	0.829 5		
Window size	Result	The 1st segment			The 2nd segment		$C(\boldsymbol{Y}_S,\boldsymbol{Y}_R)$	Time/s
50	C_{mx}	0.994 1			0.994 9		0.832 7	8.65
	C_{PIx}	0.439 0			0.471 4			
	$C(\boldsymbol{Y}_{Sx},\boldsymbol{Y}_{Rx})$	0.827 5			0.837 9			

It shows that the overall validation results are very close when setting different window sizes for segmented GPR. Besides, the validation process is less time-consuming with a smaller window. This suggests that segmented GPR has almost no impact on the final model validation results but can significantly enhance the validation efficiency.

Furthermore, we test the performance of the proposed method in validating small samples. Let the force load identification model run 100 times under a strain error of 1%, and the model outputs form the reference data set. Then, let the model run 100 times under a strain error of 3%, and the model outputs form the simulation data set. Following Table 4, in the corresponding validation data set, randomly select a specified number of validation data samples for simulation results validation. The final validation results are also shown in Table 4.

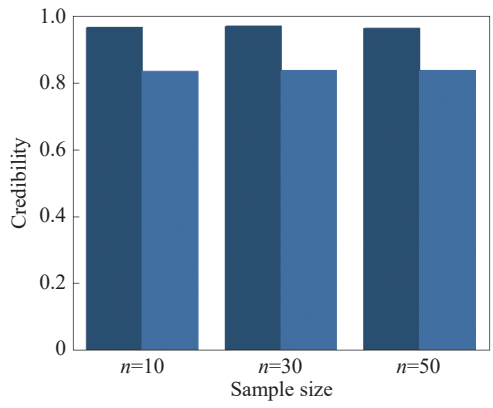
Table 4 Information and results for small-sample validation

Index	Reference data	Simulation output	Validation result
1	5	5	0.837 6
2	10	10	0.833 5
3	20	20	0.836 6
4	30	30	0.838 4
5	5	10	0.838 4
6	5	20	0.840 0
7	5	30	0.840 6

It can be seen that the validation results are very similar, whether only the sample size of simulation outputs is increased or the sample sizes of both simulation outputs and reference data are increased. Therefore, the proposed method can be effectively applied to simulation result validation under small samples.

To verify the superiority of the proposed method, we compare the proposed validation method based on GPR with another simulation dynamic output validation method based on data features proposed in [9]. The validation method based on data features is designed to measure dynamic simulation results under uncertainty based on Bayesian hypothesis testing and it is assumed that the simulation outputs are consistent with the reference data if the validation result exceeds 0.5.

Referring to the validation data information in Table 1, the model outputs under a 1% strain error are used as the reference data. Besides, the model outputs under 1% and 3% strain errors are used as two sets of simulation data, labeled as "Data_S1" and "Data_S2" respectively. The two methods are applied in three cases with sample sizes of 10, 30, and 50 for both reference data and simulation data. The validation results are shown in Fig. 7.



(a) Validation results based on GPR

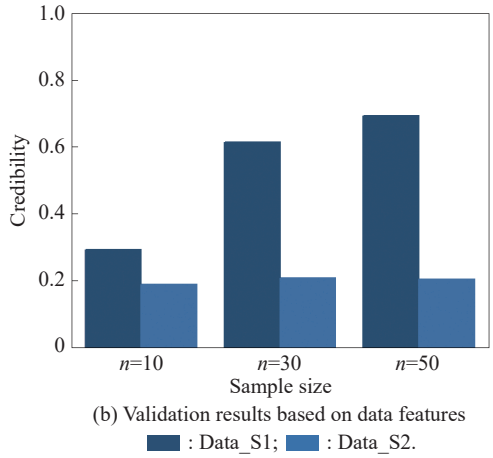


Fig. 7 Validation results by two methods

As can be seen from Fig. 7, the proposed method always yields reasonable validation results even with small sample sizes. However, the validation method based on data features can obtain effective validation results when the samples are large but its performance is poor when validating small samples. In summary, the proposed method demonstrates significant superiority in small-sample validation.

5. Conclusions

To address the challenge of model validation under small samples, this paper proposes a validation method based on GPR to validate small-sample dynamic outputs. We establish a validation framework based on Bayes statistics, shifting the focus from analyzing validation data to analyzing posterior distributions. The validation process is divided into two main stages: segmented GPR and consistency measurement. We respectively model the simulation outputs and the reference data by segmented GPR. Furthermore, the consistency of the validation data is assessed by comparing the posterior distributions of the GPR models, which effectively describe the intrinsic characteristics and the uncertainty in validation data.

We test the effectiveness of the proposed method in terms of both uncertainty description and validation for small samples through a numerical example and an application example. The method has been demonstrated to achieve reasonable results for validating small samples.

However, the proposed method is currently limited to dynamic and univariate outputs. In reality, outputs may be multivariate and correlated. Future work will focus on extending the small-sample validation to accommodate multivariate and correlated outputs.

References

[1] STEVENSON D E. Verification and validation of complex systems. Proc. of the Intelligent Engineering Systems through Artificial Neural Networks, 2002: 159−164.
[2] WANG Y N, LI J Q, SUN H B, et al. A survey on VV&A of large-scale simulations. International Journal of Crowd Science, 2019, 3(1): 63–86.
[3] SARGENT R G. Verification and validation of simulation models. Journal of Simulation, 2013, 7(1): 12–24.
[4] DURST P J, ANDERSON D T, BETHEL C L. A historical review of the development of verification and validation theories for simulation models. International Journal of Modeling, Simulation, and Scientific Computing, 2017, 8(2): 1730001.
[5] XI Z M, YANG R J. Reliability analysis with model uncertainty coupling with parameter and experiment uncertainties: a case study of 2014 V&V challenge problem. Journal of Verification, Validation and Uncertainty Quantification, 2015, 1(1): 011005.
[6] SHANKAR S, SANKARAN M. Integration of model verification, validation, and calibration for uncertainty quantification in engineering systems. Reliability Engineering and System Safety, 2015, 138: 194–209.
[7] LIN S L, LI W, MA P, et al. Structural modelling and bayesian inference for model validation and confidence extrapolation. Journal of Statistical Computation and Simulation, 2020, 90(2): 211–233.
[8] PAN Y L, HE Y Z. Trustworthiness assessment method for guidance simulation system based on DS/AHP and gray cloud clustering. Electronic Measurement Technology, 2017, 40(7): 43–47.
[9] YANG M, QIAN X C, LI W. Simulation dynamic output validation method based on data feature. Systems Engineering and Electronic Technology, 2016, 38(2): 457–463. (in Chinese)
[10] JIA J P, HE X Q, JIN Y J. Statistics. Beijing: Renmin University Press, 2015.(in Chinese)
[11] SONG T. Research on model validation method for small sample and service-oriented tools. Harbin: Harbin Institute of Technology, 2018. (in Chinese)
[12] WANG J M, WU Y J. Credibility assessment of small sample data based on clustered cloud model. Journal of System Simulation, 2019, 31(7): 1263–1271. (in Chinese)
[13] NIE K, LUAN R P. A simulation model validation method based on data enhancement. Command Control and Simulation, 2019, 41(3): 92–96. (in Chinese)
[14] LI W, ZHOU Y C, LIN S L, et al. A review of simulation model validation methods. Journal of System Simulation, 2019, 31(7): 1249–1256. (in Chinese)
[15] ANALLA M. Model validation through the linear regression fit to actual versus predicted values. Agricultural Systems, 1998, 57(1): 115–119.
[16] MAHARAJ E A. Cluster of time series. Journal of Classification, 2000, 17(2): 297–314.
[17] MAHARAJ E A. Comparison and classification of stationary multivariate time series. Pattern Recognition, 1999, 32(7): 1129–1138.
[18] RAMONI M, SEBASTIANI P, COHEN P. Bayesian clustering by dynamics. Machine Learning, 2002, 47(1): 91–121.
[19] BICEGO M, MURINO V, FIGUEIREDO M A T. Similarity-based clustering of sequences using hidden Markov models. Proc. of the International Conference on Machine Learning and Data Mining in Pattern Recognition, 2003: 86−95.
[20] SEEGER M. Gaussian processes for machine learning. International Journal of Neural Systems, 2004, 14(2): 69–106.
[21] ROBERTS S, OSBORNE M, EBDEN M, et al. Gaussian

processes for time-series modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2013, 371(1984): 20110550.

[22] AIGRAIN S, FOREMAN-MACKEY D. Gaussian process regression for astronomical time series. Annual Review of Astronomy and Astrophysics, 2023, 61(1): 329–371.

[23] PALAR P S, PARUSSINI L, BREGANT L, et al. On kernel functions for bi-fidelity Gaussian process regressions. Structural and Multidisciplinary Optimization, 2023, 66(2): 1–22.

[24] LI S. Research on the validation method of missile system simulation model. Changsha: National University of Defense Technology, 2003. (in Chinese)

[25] MANFREDI P. Conservative Gaussian process models for uncertainty quantification and Bayesian optimization in signal integrity applications. IEEE Trans. on Components Packaging and Manufacturing Technology, 2024, 14(7): 1261–1272.

[26] KARVONEN T, WYNNE G, TRONARP F, et al. Maximum likelihood estimation and uncertainty quantification for gaussian process approximation of deterministic functions. SIAM/ASA Journal on Uncertainty Quantification, 2020, 8(3): 926–958.

[27] PARK I, AMARCHINTA H K, GRANDHI R V. A Bayesian approach for quantification of model uncertainty. Reliability Engineering and System Safety, 2010, 95(7): 777–785.

[28] REBBA R, MAHADEVAN S, HUANG S. Validation and error estimation of computational models. Reliability Engineering and System Safety, 2006, 91(10): 1390–1397.

[29] SMITH J. Statistical modeling principles. Journal of Statistical Methods, 2023, 45(2): 234–256.

[30] ANDERSON R, BROWN T. Gaussian process regression for dynamic systems. Computational Bayesian Models, 2023, 29(3): 98–121.

[31] LIU H, ONG Y S, SHEN X, et al. When Gaussian process meets big data: a review of scalable GPS. IEEE Trans. on Neural Networks and Learning Systems, 2020, 31(11): 4405–4423.