2024 32nd Mediterranean Conference on Control and Automation (MED)
June 11-14, 2024 | Chania, Crete, Greece. pp. 909-914.

Uncertainty-Informed Model Selection Method for Nonlinear System Identification and Interpretable Machine Learning*

Yuanlin Gu, Hua-Liang Wei

Abstract—**Modeling uncertainty has been an active and important topic in the fields of data-driven modeling and machine learning. Uncertainty ubiquitously exists in any data modeling process, making it challenging to identify the optimal models among many potential candidates. This article proposes an uncertainty-informed method to address the model selection problem. The performance of the proposed method is evaluated on a dataset generated from a complex system model. The experimental results demonstrate the effectiveness of the proposed method and its superiority over conventional approaches. This method has minimal requirements for the length of training data and model types, making it applicable for various modeling frameworks.**

I. Introduction

The typical process of data-driven modeling involves several stages including data collection, preprocessing, model training and model validation. For most data-driven modeling methods, it is often necessary to define some training parameters prior to model training. For instance, when building a neural network model, parameters such as the number of epochs, the estimation/optimization algorithm, evaluation metrics, and network structure (including layer types, number of layers, and neurons per layer) must be specified first [1], [2]. For some regression-based models, such as the Nonlinear AutoRegressive Moving Average with eXogenous inputs (NARMAX) model [3], [4], the determination of model structure and the generation of candidate linear and nonlinear model terms are essential. The following processes such as the selection of important model terms, the determination of model size (i.e., the number of terms), model validity test, and model performance evaluation plays an important and central role [5]. Due to these factors, for a given data modeling task, there could be a vast number of candidate solutions. In practice it is always challenging to effectively identify the best model or the best set of models among these candidates.

Numerous model selection and model size determination methods have been introduced and widely applied in data modeling, system identification, and machine learning. Among the popular ones are the Akaike Information Criterion (AIC) [6] and Bayesian Information Criterion (BIC) [7]. These criteria aim to find a balance between model performance and complexity by incorporating measures for both. They have found extensive use across various applications [8], [9], [10]. Additionally, methods such as prediction error sum of squares (APRESS) [11] were developed to address model selection challenges in complex nonlinear system identification, which have been proven effective in diverse fields. Cross-validation is another widely utilized technique, particularly suitable for black-box models such as neural networks [12], [13].

In recent years, uncertainty analysis has become an important and hot topic in data modeling fields [14], [15]. When strong uncertainties exist, models can become unreliable, particularly when predicting peak values in some specific applications, e.g., space weather prediction [16]. This can incur substantial financial costs in certain sectors where model risks cannot be accurately assessed. Consequently, various methods have been devised to quantify uncertainties in data-driven modeling. Uncertainties can arise in model structure, parameters, and predictions [16]. The evaluation of prediction uncertainties is particularly crucial for time series prediction tasks. One common approach is to develop methods that generate prediction bands, offering a visual representation of prediction uncertainties. For example, a two-stage support vector machine has been proposed to predict settlement occurrences and evolution while quantifying model uncertainties [17]. A Monte Carlo-based approach has been devised to quantify uncertainties in recurrent neural networks [18]. The Cloud-NARX model has been developed to predict space weather with a prediction interval, aiding in quantifying uncertainties [16]. During the modelling process of the above methods, a bootstrapping method is usually employed, and a huge number of models are generated from sub-datasets. There is a common challenge in quantifying prediction uncertainty: how to effectively determine a set of best and most representative models from a large number of candidate models generated from uncertain data. There is a need to develop a new method that can handle all candidate models simultaneously and identify the best ones. Building on this observation, this article proposes a new uncertainty informed model selection method, for addressing the model selection problem where uncertainty quantification is required during the modeling process.

The main contributions of this article are as follows: 1) The proposed method can be applied to identify the best models in scenarios where multiple sub-datasets and associated models exist. 2) The proposed method leverages uncertainty quantification results from all sub-datasets and incorporates information for model selection. 3) The proposed method can be integrated into various data-driven modeling frameworks, such as neural networks and regression models.

*Research supported by STFC (Ref. ST-Y001524-1), NERC (Ref. NE/V002511/1), and the University of Stirling.

Yuanlin Gu is with the Division of Computer Science and Maths, University of Stirling, Stirling, FK9 4LA, United Kingdom. (e-mail address: yuanlin.gu@stir.ac.uk).

Hua-Liang Wei is with the Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD (*Corresponding author phone: 0044 1142225198; e-mail: w.hualiang@sheffield.ac.uk).

The remainder of the article is below. Section 2 reviews the relevant methods. Section 3 introduces the proposed methods. Experimental results are presented in section 4. Finally, the work is concluded in section 5.

II. Relevant methods

A. Overview of the data modelling process

Given a system, let its input matrix and output vector be $\boldsymbol{X}$ and $\boldsymbol{y}$ respectively as follows:

$$\boldsymbol{X} = \begin{bmatrix} x_1(1) & x_2(1) & & x_M(1) \\ x_1(2) & x_2(2) & \cdots & x_M(2) \\ \vdots & & \ddots & \vdots \\ x_1(N) & x_2(N) & \dots & x_M(N) \end{bmatrix} \quad (1)$$

$$\boldsymbol{y} = \begin{bmatrix} y\,(1) \\ y\,(2) \\ \vdots \\ y\,(N) \end{bmatrix} \quad (2)$$

where the m-th column $(m = 1,2,\dots,M)$ is the m-th input vector, and N is the number of data samples. The modelling task is to establish a mathematical model, denoted by $f(\cdot)$, to describe the relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$, as below:

$$\boldsymbol{y} = f\,[\boldsymbol{x}, \boldsymbol{\theta}] + \boldsymbol{e} \quad (3)$$

where $\boldsymbol{\theta}$ is the vector of estimated parameter and $\boldsymbol{e}$ is the model residual vector. Specifically, when modelling a dynamic system, the model can be written as:

$$y\,(t) = f\,[y(t-d), y(t-d-1), y(t-d-2), \dots, y(t-d-n_y), x_1(t-d-1), x_1(t-d-1), \dots, x_1(t-d-n_x), x_M(t-d-1), x_M(t-d-1), \dots, x_M(t-d-n_x), \boldsymbol{\theta}\,] \quad (4)$$

where t is the time stamp; d is the time delay between the input and output; n_x and n_y are the maximum time lags in inputs and output, respectively.

The approximation accuracy of the actual system model f can be determined by many factors, including the model type and model structure used to represent f, the model training process and the data quality used for model training. Taking the Nonlinear autoregressive (NARX) model as an example, where polynomials are commonly employed as the basis functions to build models, the initial full model can usually be represented as [3], [4]:

$$\boldsymbol{y} = \textstyle\sum_{i=1}^{P} \theta_i\, \boldsymbol{\varphi}_i + \boldsymbol{e} \quad (5)$$

where $\boldsymbol{\varphi}_i$'s $(i = 1,2,\dots,P)$ are the derived model terms, θ_i's $(i = 1,2,\dots,P)$ are the estimated parameters, and $\boldsymbol{e}$ is the model residual. Following the procedure in [19] and [20], the candidate model terms can be generated first, and then a term selection method called Orthogonal Forward Regression (OFR) algorithm can be used to identify the most crucial model terms in a stepwise manner. Prior to this process, the time delay, nonlinear degree, and maximum time lags should be properly defined. The various combination of these hyperparameters can lead to a large number of candidate models, denoted as $f_1, f_2, \dots, f_P$. Sometimes, the number P can be extremely large, making it challenging to identify the best models among these candidates. Model size determination (i.e., the number of model terms) is important, as having too many model terms inevitably increases the model complexity and potentially lead to overfitting. Conversely, having too few model terms can result in poor model performance. Therefore, some model selection criteria are typically applied to identify the optimal number of terms in the final model.

B. Review of model selection criteria

The Akaike Information Criterion (AIC) is a commonly used model selection method [6]. It can be calculated using the formula:

$$AIC = -2\ln(L) + 2p \quad (6)$$

where L is the likelihood based on the given data, and p represents the number of parameters in the model. Another similar approach is the Bayesian Information Criterion (BIC) [7], calculated as:

$$BIC = -2\ln(L) + p \times ln(N) \quad (7)$$

where N is the sample size. More recently, a modified generalized cross validation criteria called adjustable prediction error sum of squares (APRESS) has been developed and used in modeling the NARMAX model [11]. It is calculated as

$$APRESS = \left[\frac{1}{1-\frac{C(p,\alpha)}{N}}\right]^2 \times MSE(p) \quad (8)$$

where the component $[1/(1-\frac{C(p,\alpha)}{N})]^2$ is a penalty function for adding more teams, α is a turning parameter, and $MSE(p)$ is the mean square error. Each of the three criteria consists of two components: one for measuring the error and another for penalizing additional model terms. In the AIC criterion, the component $-2\ln(L)$ measure the model ability to explain the data, while the component $2p$ serves as a penalty for adding more model terms. Therefore, the model with the lowest AIC is regarded as the optimal choice. The distinction between AIC and BIC lies in the penalty component. With a more substantial penalty for model terms, BIC adopts a more conservative approach when determining the number of terms.

Although these criteria are effective in many cases, they cannot be directly applied to modeling problems involving huge number of sub-datasets, especially when a bootstrapping method is employed to quantify model uncertainty. In such scenarios, multiple datasets are typically utilized to derive the distribution of estimated model parameters, generate prediction intervals, and establish a fuzzy representation to describe model uncertainty [16], [17], [18]. Assuming that there are K sub-datasets and the models constructed for each sub-dataset have different model terms, effectively leveraging the information from all sub-datasets and associated models using traditional information criteria becomes challenging. For instance, applying the AIC criteria may result in $K \times P$ values being calculated. However, each of these values just reflects the information of individual models and cannot offer a comprehensive assessment of model performance across all subsets. With this in mind, we aim to develop a new model selection method to identify the optimal number of model terms in such situations.

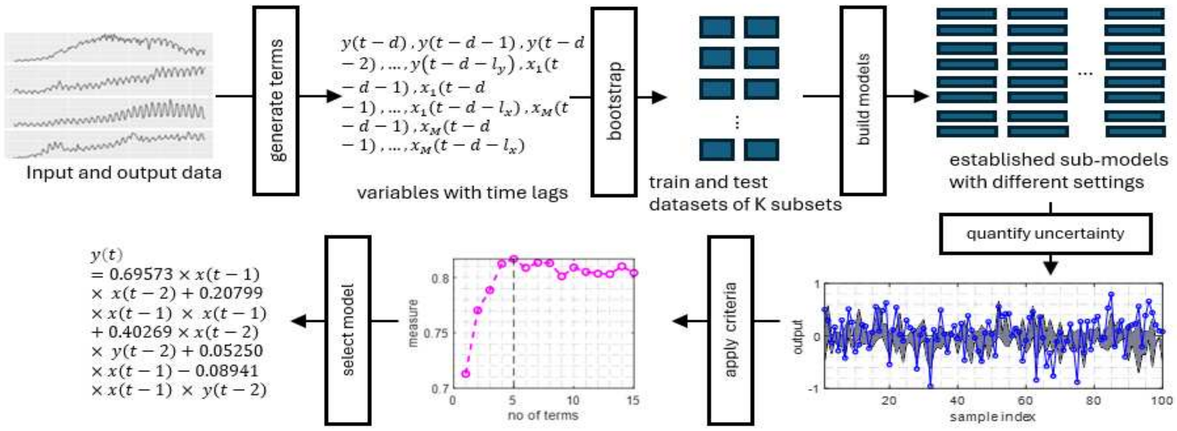
Input and output data
generate terms
$y(t-d), y(t-d-1), y(t-d-2), \dots, y(t-d-l_y), x_1(t-d-1), x_1(t-d-1), \dots, x_1(t-d-l_x), x_M(t-d-1), x_M(t-d-1), \dots, x_M(t-d-l_x)$
variables with time lags
bootstrap
train and test datasets of K subsets
build models
established sub-models with different settings
quantify uncertainty
output
1
0
-1
20
40
60
80
100
sample index
apply criteria
measure
0.8
0.75
0.7
0
5
10
15
no of terms
select model
$y(t) = 0.69573 \times x(t-1) \times x(t-2) + 0.20799 \times x(t-1) \times x(t-1) + 0.40269 \times x(t-2) \times y(t-2) + 0.05250 \times x(t-1) - 0.08941 \times x(t-1) \times y(t-2)$

Figure 1. Flowchart of the proposed method

III. The Proposed Method

The proposed method consists of several stages, as illustrated in Fig. 1. The details are outlined below.

A. Bootstrap

In this stage, a bootstrap procedure is employed to generate several sub-datasets from the original dataset. Assuming that K sub-datasets have been generated, the input and output variables can be represented as follows:

$$\boldsymbol{X} = \begin{bmatrix} x_1^{(k)}(1) & x_2^{(k)}(1) & \cdots & x_M^{(k)}(1) \\ x_1^{(k)}(2) & x_2^{(k)}(2) & & x_M^{(k)}(2) \\ \vdots & & \ddots & \vdots \\ x_1^{(k)}(N') & x_2^{(k)}(N') & \cdots & x_M^{(k)}(N') \end{bmatrix} \tag{9}$$

$$\boldsymbol{y} = \begin{bmatrix} y^{(k)}(1) \\ y^{(k)}(2) \\ \vdots \\ y^{(k)}(N') \end{bmatrix} \tag{10}$$

where $k = 1,2,\dots,K$ is the index of sub-datasets. This step aims to extract uncertainties within the data and build the groundwork for further uncertainty analysis. Typically, the data size of the sub-datasets N' is smaller than that of the original data N, resulting in each sub-dataset having different data samples. Consequently, the model terms and estimated parameters of each model differ, and this information can be analyzed to quantify uncertainty.

B. Individual model identification

In this stage, individual models will be constructed based on the K sub-datasets, resulting in K sets of models denoted as $\boldsymbol{F}^{(1)}, \boldsymbol{F}^{(2)}, \boldsymbol{F}^{(3)}, \dots, \boldsymbol{F}^{(K)}$, with each set containing a variety of candidate models having different numbers of model terms $\boldsymbol{F}^{(k)}\colon \{f_1^{(k)}, f_2^{(k)}, \dots, f_P^{(k)}\}$. Consequently, a total number of $P \times K$ models will be established, and the model selection task is to identify the optimal number of the model terms.

A brief overview of the steps involved in constructing the NARX model is as below. More comprehensive information can be found in [5], [19], [20]. Initially, a set of candidate model terms is established for each sub-dataset. For instance, the candidate terms for the k-th dataset are denoted as:

$$\{\boldsymbol{\varphi}_1^{(k)}, \boldsymbol{\varphi}_2^{(k)} \dots, \boldsymbol{\varphi}_P^{(k)}\} \tag{11}$$

where P is the total number of candidate terms. Next, the OFR algorithm is employed to measure the Error Reduction Ratio (ERR) and identify the most significant terms for each sub-model. Assuming that P' terms are selected and incorporated into the final model, the models for the k-th sub-dataset can be represented as:

$$\boldsymbol{y}^{(k)} = f_{P'}^{(k)}\{\boldsymbol{\varphi}_{l_1}^{(k)}, \boldsymbol{\varphi}_{l_2}^{(k)} \dots, \boldsymbol{\varphi}_{l_{P'}}^{(k)}\} + \boldsymbol{e}^{(k)} \tag{12}$$

where $l_1, l_2, \dots l_{P'}$ are the indices of the selected model terms. Note that there the number of terms P' can range from 1 to P if no model selection is used to identify the optimal value. Thus, the candidate dictionary comprises a total of $P \times K$ models, and the objective of the model selection criteria is to determine the optimal value P'.

C. Model uncertainty quantification

In this stage, we apply certain methods to quantify the uncertainty in model predictions based on the established models. For the models with P' model terms from the k-th sub-dataset, the model predictions can be calculated as follows:

$$\boldsymbol{y}_{pre}{}_{P'}^{(k)} = f_{P'}^{(k)}(\boldsymbol{\varphi}_{l_1}^{(k)}, \boldsymbol{\varphi}_{l_2}^{(k)} \dots, \boldsymbol{\varphi}_{l_{P'}}^{(k)}) \tag{13}$$

Collectively, these predictions constitute a prediction matrix for all the sub-datasets when the number of terms is P'.

$$\boldsymbol{y}_{pre}{}_{P'} = [\boldsymbol{y}_{pre}{}_{P'}^{(1)} \boldsymbol{y}_{pre}{}_{P'}^{(2)}, \dots, \boldsymbol{y}_{pre}{}_{P'}^{(K)}] \tag{14}$$

Consequently, a collection of K models form a model cluster to characterize prediction uncertainty. Unlike traditional predictive models, using such a model cluster for prediction can generate multiple predicted values. These values can be employed to establish a prediction band, offering a representation of prediction uncertainty. To evaluate the performance of such a prediction band, two metrics are proposed. The first metric, named prediction band accuracy, quantifies the proportion of observed values lying within the prediction band. Denote by γ the prediction band, then:

$$\gamma = \frac{N_\gamma}{N'} \tag{15}$$

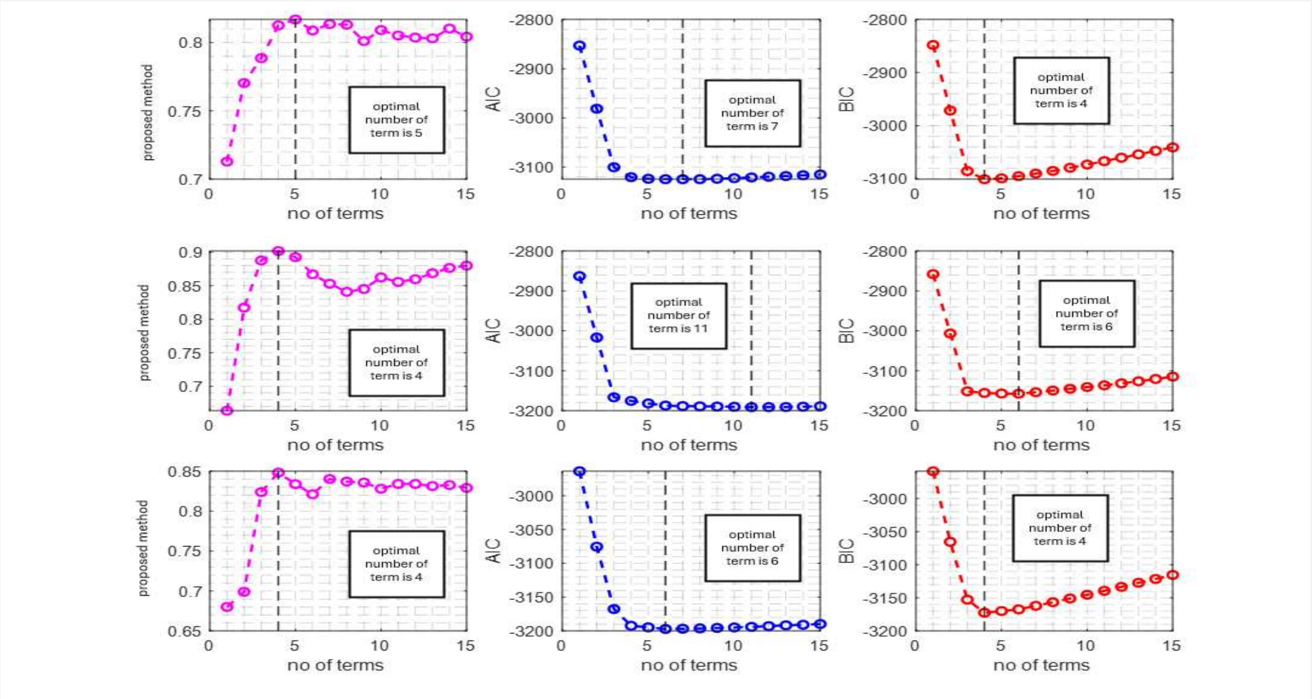

Figure 2. The results of AIC, BIC and the proposed method (upper: 1st simulation, middle: 2nd simulation, lower: 3rd simulation)

where N' is the data size and N_γ is the number of true values within the prediction band. The second metric, prediction band width (denoted as β), measures the width of the prediction band:

$$\beta = \frac{1}{N\prime}\sum_{j=1}^{N\prime}\left|\frac{max[y_{pre}(j)]-min[y_{pre}(j)]}{max[\mathbf{y}]-min\,[\mathbf{y}]}\right| \quad (16)$$

where $max[y_{pre}(j)]$ and $min[y_{pre}(j)]$ represent maximum and minimum values of the generated predicted band for the j-th data sample, respectively. These values define the upper boundary and lower boundary of the prediction band. The component $max[\mathbf{y}] - min\,[\mathbf{y}]$ is the difference of maximum and minimum values of the true observation, which used to normalize β to $[0,1]$.

D. Model selection criteria

The proposed model selection criterion is derived based on the prediction band. Typically, high prediction accuracy is desirable, which can be achieved by widening the prediction band. However, if the prediction band width becomes too large, its effectiveness in quantifying model uncertainty may lose. Therefore, the model selection criteria aim to find a balance between prediction band accuracy and width. Building upon these considerations, the proposed criterion, denoted as Ϲ, is formulated as:

$$\text{Ϲ}= \frac{\gamma}{\beta}\times\left(1-\frac{\alpha\times p}{N\prime-1}\right) \quad (17)$$

where α represents a turning parameter, p stands for the number of terms, N denotes the data size. Here, the component $\frac{\gamma}{\beta}$ assesses the prediction band accuracy while penalizing excessively wide prediction bands. The component $1-\frac{\alpha\times p}{N\prime-1}$ penalizes the addition of model terms. Initially, when a few terms are added, the value of $\frac{\gamma}{\beta}$ increases, while the value of $1-\frac{\alpha\times p}{N\prime-1}$ remains close to 1. However, the penalty becomes more significant as more terms are added, leading to a decrease in the overall value. Consequently, the maximum value of Ϲ indicates the optimal number of terms.

Lastly, additional approaches can be applied to model the uncertainty from the selected models (e.g., the Cloud-NARX model [16]). Since the proposed criterion utilizes post-modeling information, it can be easily integrated into any modeling framework like polynomial models as well as neural networks. The article employs the NARX model as a case study to demonstrate the methodology, as detailed in the following section.

IV. Experiment

To validate the effectiveness of the proposed methods, we performed a case study on simulation data. We generated a dataset from the following designed system:

$$y(t) = 0.1\sqrt{|y(t-1)|}x(t-1) + 0.2x(t-1)^2 + 0.7x(t-1)x(t-2) + 0.4x(t-2)y(t-2) + \varepsilon(t) \quad (18)$$

where the input $x(t)$ is a randomly generated sequence of 100 data samples, ranging from -1 to 1. The noise sequence $\varepsilon(t)$ has zero mean and finite variance. This system has two sources of uncertainty. First, the noise sequence is random and cannot be accurately represented by the model. Second, in this experiment, with a time delay set to 1, maximum time lags of 2 for input and output variables, and a nonlinearity degree of 2, the nonlinear system component $\sqrt{|y(t-1)|}$ cannot be precisely described by any of the specified lagged input and output variables or their interaction product terms, but term $\sqrt{|y(t-1)|}$ may be well approximated by some product terms.

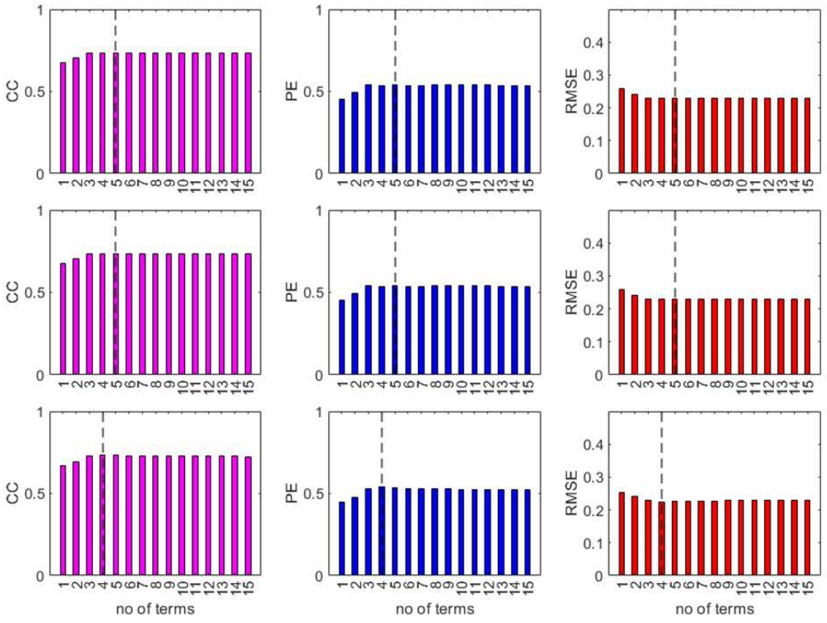

Figure 3. Model performance vs number of model terms (upper: 1st simulation, middle: 2nd simulation, lower: 3rd simulation)

TABLE I. PERFORMANCE COMPARISON OF THE SELECTED MODELS

Method	Measure	Performances of the model picked		
		Simulation 1	Simulation 2	Simulation 3
AIC	CC	0.7340	0.78520	0.7296
	PE	0.5370	0.6145	0.5314
	RMSE	0.2305	0.2225	0.227
BIC	CC	0.7341	0.7867	0.7364
	PE	0.5380	0.6169	0.5414
	RMSE	0.2303	0.2218	0.2247
Proposed criterion	CC	0.7363	0.7873	0.7364
	PE	0.5412	0.6179	0.5414
	RMSE	0.2295	0.2216	0.2247

CC: correlation coefficient, PE: prediction efficiency, RMSE: root mean square error

Following the procedures outlined in Section III, the proposed method was assessed follows. First, a bootstrap process was conducted to generate 10 sub-datasets, each comprising 70% of the original data. For each sub-dataset, NARX models were built with a varying number of model terms, ranging from 1 to 15. Throughout the process, the three criteria including AIC, BIC and the proposed criterion formulated in (17) were applied to identify the optimal number of terms. Three times of simulated were performed independently and the values of AIC, BIC, and the proposed criterion were calculated and recorded, as shown in Fig. 2. Each of the three criteria suggests their own optimal numbers of model terms. Specifically, AIC selected term numbers of 7, 11, and 6, while BIC chose term numbers of 4, 6, and 4. The proposed method picked term numbers of 5, 4, and 4 in the three simulations, respectively. the figure, it is evident that the three criteria suggest different models in most cases, although the proposed criterion suggests the same result as BIC for the third simulation.

The performances of the selected models are detailed in Table I. We employed three metrics to assess model performance: correlation coefficient, prediction efficiency, and root mean square error. It can be noted that the model chosen by the proposed criterion outperforms the models selected by the other two criteria in most cases. The only exception is observed in the 3rd simulation, where BIC selected the same model as that suggested by the proposed criterion. The performances of all the identified models are presented in Fig. 3, with the models selected by the proposed criterion highlighted with a black dashed line. It can be observed that models identified by the proposed criterion show the best performance.

TABLE II. SELECTED MODEL TERMS OF ONE EXAMPLE

	Selected Term	ERR(100%)	Parameters
1	$x(t-1) \times x(t-2)$	41.6924	0.69573
2	$x(t-1) \times x(t-1)$	7.01049	0.20799
3	$x(t-2) \times y(t-2)$	5.20104	0.40269
4	$x(t-1)$	0.59199	0.05250
5	$x(t-1) \times y(t-2)$	0.25551	-0.08941

One of the selected models in detailed Table II, which can be written down as follows:

$$y(t) = 0.69573x(t-1)x(t-2) + 0.20799x(t-1)x(t-1) + 0.40269x(t-2)\,y(t-2) + 0.05250x(t-1) - 0.08941x(t-1)\,y(t-2) \tag{19}$$

A visualization of the prediction band associated with the selected models is presented in Fig. 4. The prediction band covers most of the observations. However, for some peak values, the accuracy of the band decreases. The reason is that the models are built on an imbalanced dataset with strong noise, where most data samples are significantly lower than the peak values. This issue could be partially solved by reducing the sample size of the bootstrap process. Nonetheless, the primary objective of this experiment is to assess the effectiveness of the model selection criteria, and such uncertainty analysis can in turn help better evaluating the effectiveness of these methods. This condition highlights a unique advantage of the proposed method: it can be applied to modeling problems with small-sized data, as the strong uncertainty introduced by the insufficient data is quantified. This offers flexibility in terms of minimal data requirements and improves model robustness under strong uncertainty, which traditional methods cannot provide.

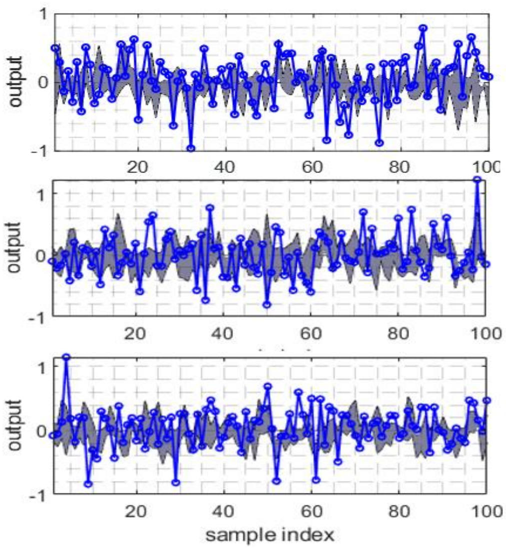

Figure 4. Prediction interval vs observed values (upper: 1st simulation, middle: 2nd simulation, lower: 3rd simulation)

V. Conclusion

This article introduces a new uncertainty-informed model selection method, offering advantages in addressing data modeling complexities associated with substantial uncertainty and multiple datasets. Through three simulation examples, the proposed approach, with the newly introduced model selection criterion, has shown superior performance in identifying better models compared to conventional model selection criteria. Moreover, the method is adaptable to various data modeling and machine learning frameworks. Future work involves evaluating the proposed method on real-world datasets and investigating its adaptability to other modeling techniques, such as recurrent neural networks.

References

[1] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, “A survey of convolutional neural networks: analysis, applications, and prospects,” *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 12, pp. 6999–7019, 2022, doi: 10.1109/TNNLS.2021.3084827.

[2] Y. Zhou, B. Li, J. Wang, E. Rocco, and Q. Meng, “Discovering unknowns: Context-enhanced anomaly detection for curiosity-driven autonomous underwater exploration,” *Pattern Recognit*, vol. 131, p. 108860, 2022, doi: https://doi.org/10.1016/j.patcog.2022.108860.

[3] S. Chen and S. A. Billings, “Representations of non-linear systems: the NARMAX model,” *Int J Control*, vol. 49, no. 3, pp. 1013–1032, 1989, doi: 10.1080/00207178908559683.

[4] S. A. Billings, *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. 2013. doi: 10.1002/9781118535561.

[5] Y. Gu and H. L. Wei, “A robust model structure selection method for small sample size and multiple datasets problems,” *Inf Sci (N Y)*, vol. 451–452, pp. 195–209, 2018, doi: https://doi.org/10.1016/j.ins.2018.04.007.

[6] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans Automat Contr*, vol. 19, no. 6, pp. 716–723, 1974, doi: 10.1109/TAC.1974.1100705.

[7] G. Schwarz, “Estimating the Dimension of a Model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, Mar. 1978, doi: 10.1214/aos/1176344136.

[8] Y. Gu, H. L. Wei, and M. M. Balikhin, “Nonlinear predictive model selection and model averaging using information criteria,” *Systems Science & Control Engineering*, vol. 6, no. 1, pp. 319–328, Jan. 2018, doi: 10.1080/21642583.2018.1496042.

[9] J. J. Dziak, D. L. Coffman, S. T. Lanza, R. Li, and L. S. Jermiin, “Sensitivity and specificity of information criteria,” *Brief Bioinform*, vol. 21, no. 2, pp. 553–565, Mar. 2020, doi: 10.1093/bib/bbz016.

[10] S. Portet, “A primer on model selection using the Akaike Information Criterion,” *Infect Dis Model*, vol. 5, pp. 111–128, 2020, doi: https://doi.org/10.1016/j.idm.2019.12.010.

[11] S. A. Billings and H. L. Wei, “An adaptive orthogonal search algorithm for model subset selection and non-linear system identification,” *Int J Control*, vol. 81, no. 5, pp. 714–724, May 2008, doi: 10.1080/00207170701216311.

[12] D. Gholamiangonabadi, N. Kiselov, and K. Grolinger, “Deep neural networks for human activity recognition with wearable sensors: leave-one-subject-out cross-validation for model selection,” *IEEE Access*, vol. 8, pp. 133982–133994, 2020, doi: 10.1109/ACCESS.2020.3010715.

[13] Y. Zhang and Y. Yang, “Cross-validation for selecting a model selection procedure,” *J Econom*, vol. 187, no. 1, pp. 95–112, 2015, doi: https://doi.org/10.1016/j.jeconom.2015.02.006.

[14] J. Gawlikowski *et al.*, “A survey of uncertainty in deep neural networks,” *Artif Intell Rev*, vol. 56, no. 1, pp. 1513–1589, 2023, doi: 10.1007/s10462-023-10562-9.

[15] E. Hüllermeier and W. Waegeman, “Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods,” *Mach Learn*, vol. 110, no. 3, pp. 457–506, 2021, doi: 10.1007/s10994-021-05946-3.

[16] Y. Gu, H. L. Wei, R. J. Boynton, S. N. Walker, and M. A. Balikhin, “System identification and data-driven forecasting of AE index and prediction uncertainty analysis using a new Cloud-NARX model,” *J Geophys Res Space Phys*, vol. 124, no. 1, pp. 248–263, Jan. 2019, doi: https://doi.org/10.1029/2018JA025957.

[17] Y. Pan, J. Qin, Y. Hou, and J.-J. Chen, “Two-stage support vector machine-enabled deep excavation settlement prediction considering class imbalance and multi-source uncertainties,” *Reliab Eng Syst Saf*, vol. 241, p. 109578, 2024, doi: https://doi.org/10.1016/j.ress.2023.109578.

[18] L. Wang, T. Xiao, S. Liu, W. Zhang, B. Yang, and L. Chen, “Quantification of model uncertainty and variability for landslide displacement prediction based on Monte Carlo simulation,” *Gondwana Research*, vol. 123, pp. 27–40, 2023, doi: https://doi.org/10.1016/j.gr.2023.03.006.

[19] H. L. Wei, S. A. Billings, and J. Liu, “Term and variable selection for non-linear system identification,” *Int J Control*, vol. 77, no. 1, pp. 86–110, Jan. 2004, doi: 10.1080/00207170310001639640.

[20] H. L. Wei and S. A. Billings, “Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information,” *Int J Model Identif Control*, vol. 3, no. 4, pp. 341–356, Jan. 2008, doi: 10.1504/IJMIC.2008.020543.