Revealing Certainties from Uncertainties through Data Mining and Data Modelling

Thesis by

Yuanlin Gu

Submitted for the degree of

Doctor of Philosophy

Department of Automatic Control and Systems Engineering

Faculty of Engineering

University of Sheffield

Nov 2019

ACKNOWLEDGEMENTS

This journey would not have been possible without the support of my family, advisors, mentors and friends.

I am extremely grateful to my supervisor Dr. Hua-Liang Wei for the continuous support and encouragement through my Ph.D. study and research. He has always been a great mentor and helped me by his patience, insightful comments, suggestions, and valuable discussions.

I would like to express my sincere gratitude to my second supervisor Prof. Michael A. Balikhin, my third supervisor Prof. Grant Bigg, for their guidance and suggestions on my research.

I also want to thank Dr. Simon N. Walker and Dr. Richard J. Boynton for their valuable help and suggestions on several papers.

To all my friends, thank you for listening, offering me advice, and supporting me through this entire process.

Finally, I am especially grateful to my parents who support me emotionally and financially. Thank you for encouraging me in all of my pursuits and inspiring me to follow my dreams.

Yuanlin Gu

ABSTRACT

In model identification, the existence of uncertainty normally generates negative impact on the accuracy and performance of the identified model. This thesis focuses on the development of three novel methods to deal with model uncertainty, which are the robust model structure selection (RMSS) method, cloud-NARX model and machine learning enhanced nonlinear autoregressive moving average with exogenous inputs (MLE-NARMAX) model.

First, the RMSS method is developed for model identification problems with small size data and multi-datasets. The proposed method can reduce the model structure uncertainty and therefore improve the model performances. The RMSS method is applied to two real data applications, which are the modelling of Kp index and modelling of cortical response.

Second, the cloud-NARX model is proposed. The cloud-NARX model uses cloud model and cloud transformation to quantify the uncertainty throughout the structure detection, parameter estimation and model prediction. The cloud-NARX model is applied to predict AE index 1 hr ahead. The new predicted band can be generated to forecast system output with confidence interval. The cloud-NARX method provides a new way to evaluate the model based on uncertainty analysis and reveal the reliability of model, and visualize the bias of model prediction.

Third, the MLE-NARMAX model is developed. The MLE-NARMAX model is established based on a NARMAX model structure, which is composed of the most important candidate features (variables). With an extra neural network sub-model, the MLE-NARMAX model is enhanced by the machine learning methods so that the model performance can be improved. The MLE-NARMAX model is applied to predict appliance energy use 10 minutes ahead and predict Dst index 3 hours ahead. The proposed model provides a new way for data modelling problems through machine learning approach with a simple/sparse, interpretable and transparent model structure.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

AIC	Akaike Information Criterion
AICc	Corrected Akaike Information Criterion
APRESS	Adjustable Predicted Residual Sum of Squares
BIC	Bayesian Information Criterion
ERR	Error Reduction Ratio
GBCT	Generic Backward Cloud Transformation
GFCT	Generic Forward Cloud Transformation
MAE	Mean Absolute Error
MI	Mutual Information
MLE-NARMAX	Machine Learning Enhanced Nonlinear AutoRegressive Moving Average with eXogenous inputs
MPO	Model Predicted Output
NARMAX	Nonlinear AutoRegressive Moving Average with eXogenous inputs
NRMSE	Normalised Root Mean Square Error
OFR	Orthogonal Forward Regression
OSA	One Step Ahead
PE	Prediction Efficiency
PRESS	Predicted Residual Sum of Squares

RMSE	Root Mean Squared Error
RMSS	Robust Model Structure Selection
VAF	Variance Accounted For

Chapter 1

INTRODUCTION

1.1 Background and Motivation

Data acquisition is an important aspect of any type of the research because inaccurate data might lead to invalid data learning results. In recent years, the collection of data from a wide range of fields has become more convenient and the quality of data has increased. Benefit from the revolution of computation capacity and data acquisition, data-driven modelling and data analytics approaches have been applied to learn features and behaviours of a wide range of complex systems. As the size and complexity of data increases, the analysis of the uncertainty in the data modelling process becomes ever important for quantifying and improving the reliability of the identified model in many fields (Robinson, Benke & Norng, 2015; Christina, 2016).

The general process of system identification consists of several steps, for example, model type selection, model structure detection, term selection, parameter estimation, model evaluation, etc (Ayala Solares, Wei & Billings, 2017). There are a lot of models which have been developed to deal with the system identification problems, for example, NARMAX model (Billings, 2013; Chen & Billings, 1989), neural network (Chen & Billings, 1992; Chen, Billings & Grant, 1990; Haykin, 1994; Wang, et al., 2017), wavelet models (Billings & Wei, 2005a; Billings & Wei, 2005b; Zhang, 1992), Bayesian network (Guo, Liu & Sun, 2016), fuzzy models (Zadeh, 1965; Bustince, et al., 2016), etc. To establish and optimise the model, a wide range of technologies have been applied, for example, term selection (Chen, Billings & Luo, 1989), model selection (Billings & Wei,

2008), model averaging (Lukacs, Burnham & Anderson, 2010), correlation tests (Billings & Voon, 1983; Billings & Voon, 1986), etc.

Although many data modeling and systems identification methods are capable to describe a wide range of unknown systems, the existence of strong uncertainty may still cause deleterious effect in the modelling process. First, the uncertainty in data collection (e. g. the experimental uncertainty and epistemic uncertainty) might generate incomplete and inaccurate information. If the number of samples are insufficient or some important variables are missing in the dataset, it is extremely difficult to find a suitable model. Second, the model structure uncertainty can directly affect the model performance. It is known that models are usually designed to represent some specific system features and there are no single model type or structure that can perfectly describe all the true systems. Therefore, it is essential to choose a suitable model structure to represent the system. On the contrary, an inappropriate model structure can reduce the model performance. Third, noise/disturbance is another main source of uncertainty. The noise can be brought to the data through many ways, for example, measurement error from physical equipment, external disturbances, etc. The existence of noise could lead to biased parameter estimation, incorrected selected model terms, etc. Based on the above reasons, novel methods are needed to reduce and quantify the uncertainty in the modelling process.

For modelling problems with small size data, there usually exists strong uncertainty in the data because small changes in a few or even a single sample can cause a large effect on the accuracy of parameter estimation. Therefore, the difficulty of finding reliable models is often exacerbated due to the small sample size of data. Consequently, the strong uncertainty of the model structure might generate negative impact on the model performance and accuracy. Finding a robust model structure can reduce the model structure uncertainty. However, sometimes the existence of uncertainty is inevitable, and it is hard to find a robust model structure. Under the effect of uncertainty, the identified model usually cannot perfectly represent the system but only approximately describe the system behaviors. In these situations, a single model may not always work well on future new data, as there might be a risk on trusting and relying on a single model for future system behavior forecasting. In these situations, quantifying the model uncertainty is another way to increase the robustness of the identified model.

In recently years, ‘big’ data becomes a popular topic in engineering and computation fields. As the size and complexity of the data massively increases, the modelling of the complex nonlinear systems requires more efficient and powerful methods. The neural network and its variants are powerful regarding of model prediction performance, but lack capacity to provide an explainable representation. The regression models, for example, the commonly used NARX model, provides a transparent and parsimonious representation. However, sometimes the prediction performance is restricted due to the limitation of the simple model structure. Therefore, it is essential to find a way to deal with data modelling problems through machine learning approach with a simple/sparse, interpretable and transparent model structure.

1.2 Aims and Objectives

The aim of this thesis is to develop novel data modelling and systems identification methods to address the issues brought by model uncertainty. The objectives of the project are given as follows:

- Develop a model structure selection method to establish a robust model structure for modelling problems with strong uncertainty (e. g., small size data problems). The developed method will be applying to some real data case studies, for example, space weather forecast, EEG data, etc.
- Develop a data modelling method to analyze uncertainty. The model will hold the good property of the conventional NARMAX model, but also brings some new abilities to quantify the model uncertainty. It is also desired to develop a new model prediction that provides confidence intervals to describe the model prediction uncertainty.
- Develop a machine learning enhanced NARMAX method. The developed method should be able to provide a transparent and interpretable model structure, which can reveal the most important model terms and systems components. The new model is also expected to achieve better model prediction performance than the conventional NARMAX model.
- Apply the developed methods to a series of case studies, including social science, medical, space weather, environmental data, etc.

1.3 Overview and Contribution

The research in this thesis mainly focus on the modelling, forecasting and uncertainty analysis issues of complex nonlinear systems. The NARMAX methodology, cloud models, neural network and other machine learning techniques are applied and further extended to overcome the negative effects of strong uncertainty in the data modelling problems. The developed methods are illustrated and evaluated via a series of simulation and real data case studies, for example, space weather, appliance energy use, EEG, life satisfaction, etc.

The thesis is organised as follows and the main contributions of the thesis are briefly introduced.

Chapter 3: Robust model structured selection method for small size data modelling problems

This contribution of chapter 3 is the development of a novel RMSS method. Based on a data resampling approach, combined with an orthogonal forward regression algorithm, the RMSS method is designed to reduce model uncertainty and improve model performance. This is especially useful for the following two scenarios of data based modelling problem: (i) modelling from multiple small sample size datasets (e.g. many datasets for a same system but generated under different experimental conditions; (ii) modelling for a non-stationary system where although the key system dynamics can be represented using a single model structure, different model parameters are needed to adaptively reflect the change of system behaviors at different times.

Several simulation examples and case studies are presented to illustrate the advantages of the RMSS method on the modelling of small size data. In addition, a case study on EEG data is presented to show that the RMSS method also works well on data modelling problems with multi-datasets. In the case study, the RMSS method is employed for the modelling and forecasting the cortical responses to mechanical wrist perturbations. The results indicate that the RMSS method can improve the model performance with more than 90% variance accounted for (VAF) when implementing a one-step-ahead prediction and around 50% VAF for a three-step- ahead prediction. The case study significantly

improves modeling of cortical activity in the sensorimotor system in comparison to previous work which uses a truncated Volterra series.

The results of this chapter are published in one journal (Gu & Wei, 2018b) and one conference (Gu & Wei, 2017). Another paper is currently under s review at IEEE Transactions on Biomedical Engineering.

Chapter 4: System identification and uncertainty analysis using a new Cloud-NARX model

In chapter 4, a new cloud-NARX model is developed for: a) describing model structure and parameter uncertainty using a new uncertainty concept 'cloud' model; b) generating a new predicted band, which provides the confidence interval of predicted AE index; c) providing a new way to evaluate the model reliability based on uncertainty analysis. The reliability of the model can be quantified by the proposed uncertainty analysis method, which makes the cloud-NARX model more robust than the conventional NARX model.

The proposed cloud-NARX model is applied to the modelling and forecasting of AE index. The correlation coefficient between averaged prediction and observation is 0.87 and prediction efficiency of 0.81 when benchmarked for the period of 17-21 March 2015 and 22-26 June 2015, which is nearly identical to that produced by the best NARX model. More importantly, the cloud-NARX model is capable to quantify the uncertainty of model structure, model parameter and model prediction and generate new model prediction band with confidence interval. The width of the prediction band indicates the uncertainty of the model prediction and can be used to forecast the arrival of severe geomagnetic activity.

The results of this chapter are published in one journal (Gu, et al., 2018).

Chapter 5: Machine Learning Enhanced NARMAX Model

The contribution of chapter 4 is developing a novel MLE-NARMAX model to improve the model performance and provide a transparent representation. The MLE-NARMAX model consists of two parts, the NARX sub-model and the neural network sub-model. The NARX sub-model reveals the most significant model terms and the neural network

can improve the overall model performance. A simulation example and two real data case studies are presented to illustrate the new MLE-NARMAX model.

One case study presents the MLE-NARMAX model to predict appliance energy use 10 minutes ahead. By taking advantages of neural network and NARMAX model, the proposed interpretable model cannot only provide good forecast result in terms of two prediction skills: correlation coefficient of 0.78 and prediction efficiency of 0.61, but also provide an interpretable NARMAX model structure. In another case study, the MLE-NARMAX model is used to generate 3 hours ahead predictions for Dst index, on three typical test periods of strong storms. The results are compared with those produced by the conventional NARX and neural networks. The main feature of the MLE-NARMAX model are: 1) the resulting models are transparent and easy to interpret, and 2) the model possesses good prediction performance.

The results of this chapter are summarized in two papers. One paper is accepted at the one conference and another paper will be submitted to a journal soon.

1.4 List of Publications

The present research and related works have been published in several journals and conferences, which are listed below:

1.4.1 Journal Papers

- Gu, Y., & Wei, H. L. (2018). A robust model structure selection method for small sample size and multiple datasets problems. Information Sciences, 451–452, 195–209.
- Gu, Y., Wei, H. L., Boynton, R. J., Walker, S. N., & Balikhin, M. A. (2019). System identification and data driven forecasting of AE index and prediction uncertainty analysis using a new cloud-NARX model. Journal of Geophysical Research: Space Physics, 124.
- Gu, Y., Wei, H. L., & Balikhin, M. A. (2017) Nonlinear predictive model selection and model averaging using information criteria. Systems Science & Control Engineering, 6:1, 319-328.

- Gu, Y., & Wei, H. L. (2018). Significant indicators and determinants of happiness: Evidence from a UK survey and revealed by a data-driven system modelling approach. Social Sciences, 7(4).
- Akinola, T. E., Oko, E., Gu, Y., Wei, H. L., Wang, M. (2019) Non-linear system identification of solvent-based post-combustion CO2 capture process. Fuel. 239. pp. 1213-1223.

1.4.2 Conference Papers

- Gu, Y., Wei, H. L., Balikhin, M. A., Boynton, R. J. & Walker, S. N (2019). Machine Learning Enhanced NARMAX Model for Dst Index Forecasting. In ICAC 2019 IEEE International Conference on Automation and Computing. Accepted.
- Gu, Y., Wei, H. L., & Balikhin, M. A. (2017). Nonlinear dynamic predictive model selection and interference using information criteria. In ICAC 2017 - 2017 23rd IEEE International Conference on Automation and Computing: Addressing Global Challenges through Automation and Computing. http://doi.org/10.23919/IConAC.2017.8082005
- Gu, Y., Wei, H. L., Boynton, R. J., Walker, S. N., & Balikhin, M. A. (2017). Prediction of Kp index using NARMAX models with a robust model structure selection method. In Proceedings of the 9th International Conference on Electronics, Computers and Artificial Intelligence, Vol. 2017–January, pp. 1–6.
- Gu, Y., & Wei, H. L. (2016). Analysis of the relationship between lifestyle and life satisfaction using transparent and nonlinear parametric models. In 2016 22nd International Conference on Automation and Computing.

1.4.3 Poster Presentations

- Gu, Y., Wei, H. L., Boynton, R. J., Walker, S. N., & Balikhin, M. A. (2017). Nonlinear predictive model identification for Kp index forecasting. 14th Europeam Space Weather Week.

- Gu, Y., & Wei, H. L. (2016). K-fold voting method with normal cloud transformation - assessment and analysis of model uncertainties. In 2016 ACSE Symposium.

1.4.4 Submitted Papers

- Gu, Y., Yuan, Y., Dewald, J. P. A., van del Helm, F. C. T. & Wei, H. L. Nonlinear Modelling of Cortical Response to Mechanical Wrist Perturbation using the NARMAX Method. Under review at IEEE Transactions on Biomedical Engineering.

Chapter 2

AN OVERVIEW OF SYSTEMS IDENTIFICATION, DATA MODELLING AND UNCERTAINTY ANALYSIS

2.1 Introduction

This chapter provides an in-depth review of the system identification and uncertainty analysis approaches, focusing on the NARMAX model, orthogonal forward regression (OFR) algorithm, cloud model, etc. In addition, a brief discussion of model selection approaches, model validation methods and other commonly used technologies in data modelling process are presented.

2.2 The NARMAX Method

The general procedure of data modelling and systems identification is shown in figure 2.1. Among the many data modelling methods, the NARMAX model is one of the most commonly used model types for many real-world applications including engineering (Zhang, Zhu, & Gu, 2017), ecological (Marshall et al., 2016), environmental (Bigg et al., 2014), geophysical (Balikhin et al., 2011; Boynton, Balikhin, Billings, Wei, & Ganushkina, 2011), medical (Billings, Wei, Thomas, LMLE-NARMAXane, & Hope-Gill, 2013), and neurophysiological sciences (Li, Wei, Billings, & Sarrigiannis, 2016).

This section presents brief reviews of NARMAX and NARX model, along with the associated term selection, parameter estimation, model selection and model validation methods.

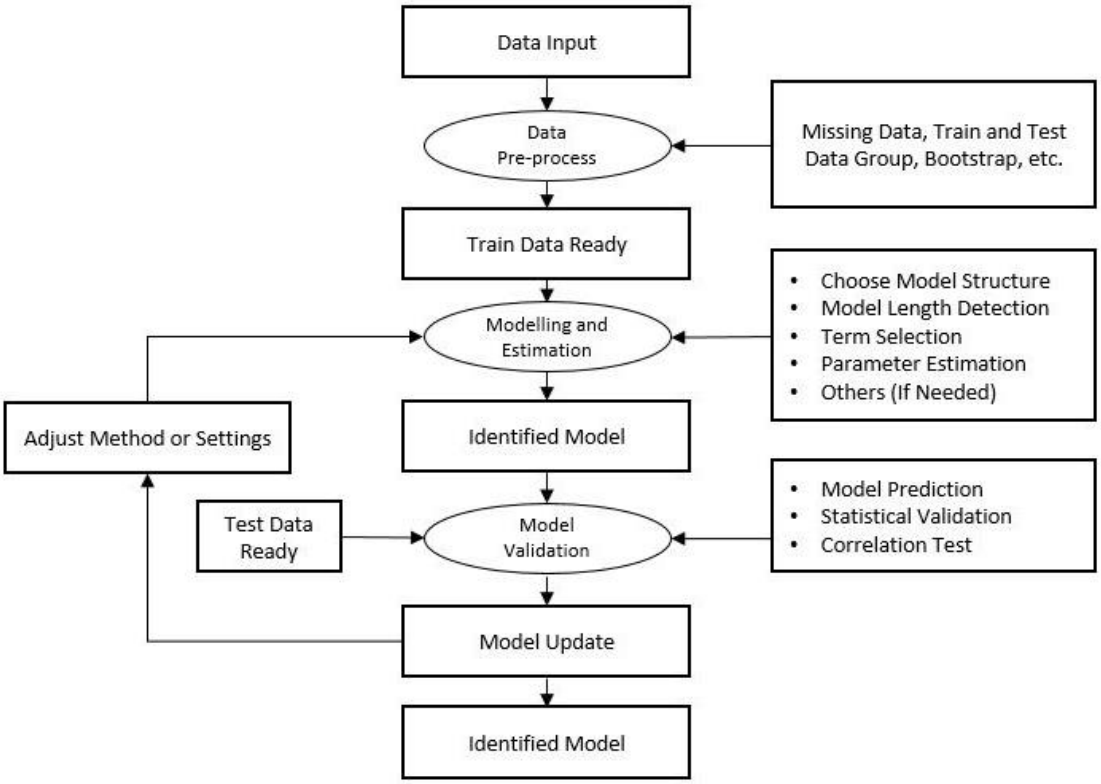

Figure 2.1 The general procedure of system identification

2.2.1 The NARMAX Model and the NARX Model

The nonlinear autoregressive moving average with exogenous inputs (NARMAX) model (Chen & Billings, 1992; Billings, 2013) was developed for black-box system identification where the true model structure is assumed to be unknown. The general NARMAX model structure is:

$$y(t) = F\left[y(t-1), \dots, y\left(t-n_y\right), u(t-1), \dots, u(t-n_u), e(t-1), \dots, e(t-n_e)\right] + e(t) \quad (2.1)$$

where $y(t)$ and $u(t)$ are systems output and input signals; $e(t)$ is a noise sequence with zero-mean and finite variance. n_y, n_u, and n_e are the maximum lags for the system output, input and noise. $F[\cdot]$ is some nonlinear function. Many of the traditional linear

and nonlinear model type, for example, AR, ARM and NARX model can be treated as special cases of NARMAX model. The commonly-used NARX model can be described as:

$$y(t) = F\big[y(t-1), \dots, y\big(t-n_y\big), u(t-1), \dots, u(t-\ n_u)\big] + e(t) \qquad (2.2)$$

There are several advantages of NARX and NARMAX model: first, the model structure can be determined in a stepwise way by selecting the significant model terms by an orthogonal forward regression (OFR) algorithm (Chen, Billings & Luo, 1989); second, the identification procedure is not time consuming and easy to implement; third, the polynomial form of the model provides a transparent and parsimonious representation of the system which is easy to understand and use. These advantages can be realized using an OFR method, which can effectively and efficiently select model terms, from a huge number of candidate model terms.

2.2.2 Term Selection with Orthogonal Forward Regression

Note that a comprehensive expansion or representation of the function $F[\cdot]$ in (2.1) might be very complex. This is because that a candidate full model which includes all the available or possible terms (both linear and cross-product terms) may involve a huge number of unknown parameters. However, in many situations, the fact is that only a small number of model terms are effective in describing the system behaviours and many of the candidate model terms are redundant. Therefore, the model structure of (2.1) is usually overfitting and contains too much unnecessary model components. It is essential to pick out these effective model terms from the full candidate model terms and uses these terms to establish a model that is simpler and more parsimonious. The term selection process can be realized by an orthogonal forward regression (OFR) algorithm (Chen, Billings & Luo, 1989).

The classic OFR algorithm, firstly introduced in (Chen, Billings & Luo, 1989), was originally developed as a subset selection method for nonlinear modelling problems where the nonlinearity is unknown in advance and the desirable model terms cannot be specified. The OFR method was proposed in solving such 'black-box' system identification problems. The basic idea behind this method is to use an error reduction ratio (ERR) index, to measure the significance of candidate model terms and generate a

rank according to the contribution made by each of the model terms to explaining the variation of the response variable. At each step, one model term can be selected from the candidate sets according to their ERR ranking. After each term is selected, it is removed from the bases and the bases are then transformed to new orthogonalized bases for the next terms selection procedure.

The general process of the OFR algorithm is presented as follow. First, the polynomial NARX model can be written as the following linear-in-the-parameters form:

$$y(t) = \sum_{m=1}^{M} \theta_m \varphi_m(t) + e(t) \tag{2.3}$$

where $\varphi_m(t) = \varphi_m(\vartheta(t))$ are the model terms generated from the regressor vector $\vartheta(t) = [y(t-1), \dots, y(t-n_y), u(t-1), \dots, u(t-n_u)]^T$ (T indicates the transpose of the vector), θ_m are the unknown parameters and M is the number of candidate model terms.

The OFR algorithm is briefly introduced as follows (Chen, Billings & Luo, 1989). First, the regression model and prediction error can be written in a compact matrix form:

$$y = \Phi\ \boldsymbol{\theta} + \varepsilon \tag{2.4}$$

where

$$y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} \tag{2.5}$$

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_M \end{bmatrix} \tag{2.6}$$

$$\varepsilon = \begin{bmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \vdots \\ \varepsilon(N) \end{bmatrix} \tag{2.7}$$

$$\Phi = [\varphi_1, \varphi_2, \dots, \varphi_M] \tag{2.8}$$

$$\varphi_i = \begin{bmatrix} \varphi_i(1) \\ \varphi_i(2) \\ \vdots \\ \varphi_i(N) \end{bmatrix} \qquad i = 1, 2, \dots., M \tag{2.9}$$

where N is the number of observations, M is the number of candidate model terms, $\{\theta_1, \theta_2, \dots, \theta_M\}$ are the unknown model parameters, $\{\varphi_1, \varphi_2, \dots, \varphi_M\}$ are the associated candidate basis vectors generated from the candidate model terms $\{u_1, u_2 \dots u_m\}$.

Let $D = \{\varphi_i : 1 \leq i \leq M\}$ be the initial dictionary of all the candidate model terms, the objective of OFR algorithm is to select a number of significant model terms to form a subset, which can be described as $D_n = \{\varphi_{l_1}, \dots, \varphi_{l_n}\}$. The output can then be described with the selected terms as follows:

$$y = \sum_{i=1}^{n} \theta_{l_i} \varphi_{l_i} + e \tag{2.10}$$

At first step of the term selection, the ERR index of each candidate model term of the initial dictionary can be calculated by:

$$ERR^{(1)}[i] = \frac{({r_0}^T \varphi_i)^2}{({r_0}^T r_0)(\varphi_i^T \varphi_i)} \tag{2.11}$$

where $i = 1, 2, \dots, M$. The first selected model term is the candidate model term with highest ERR value, as:

$$l_1 = arg \max_{1 \leq i \leq M} \{ERR^{(1)}[i]\} \tag{2.12}$$

The 1st selected model term is φ_{l_1}, and the its associated orthogonal variable can be defined as $q_1 = \varphi_{l_1}$. The selected term φ_{l_1} is then removed from the initial dictionary and the dictionary D is then reduced to a sub-dictionary D_{M-1} which consists of $M - 1$ model candidates. The residual sum of squares can be calculated as:

$$\|r_1\|^2 = \|y\|^2 - \frac{({r_0}^T q_1)^2}{q_1^T q_1} \tag{2.13}$$

At step s $(s \geq 2)$, the $M - s + 1$ bases are first transformed into new group of orthogonalised base $[q_1^{(s)}, q_2^{(s)}, \dots, q_{M-s+1}^{(s)}]$ with an orthogonal transformation as below:

$$q_j^{(s)} = \delta_j - \sum_{r=1}^{s-1} \frac{\varphi_j^T q_r}{q_r^T q_r} q_r \tag{2.14}$$

where $q_r (r = 1, 2, \dots, s-1)$ are orthogonal vectors, $\varphi_j (j = 1, 2, \dots, M - s + 1)$ are the basis of unselected model terms of subset D_{M-s+1} and $q_j^{(s)} (j = 1, 2, \dots, M - s + 1)$ are the new orthogonalised bases. The rest of the model terms can then be identified step by step using the ERR index of orthogonalised subsets D_{M-s+1}:

$$ERR^{(s)}[j] = \frac{(y^T q_j^{(s)})^2}{(y^T y)(q_j^{(s)^T} q_j^{(s)})} \tag{2.15}$$

$$l_s = arg \max_{1 \le j \le M-s+1} \{ERR^{(1)}[j]\} \tag{2.16}$$

The s-th significant model term of the subset is φ_{l_s}, and its associated orthogonal variable can be defined as $q_s = q^{(s)}{}_{l_s}$. The residual sum of squares can be updated by (Wei & Billings, 2006):

$$\|r_s\|^2 = \|r_{s-1}\|^2 - \frac{(r_{s-1}{}^T q_s)^2}{q_s^T q_s} \tag{2.17}$$

Recursively, the model terms of the subset $\{ \varphi_{l_1}, \dots, \varphi_{l_n} \}$ can be identified step by step, each at one step. By summing (2.17) for s from 1 to n, yields:

$$\|r_n\|^2 = \|y\|^2 - \sum_{s=1}^{n} \frac{(r_{s-1}{}^T q_s)^2}{q_s^T q_s} \tag{2.18}$$

The $\|r_n\|^2$ is called residual sum of squares, or sum squared error of the final model. The mean square error (MSE) of the model can be calculated as $\|r_n\|^2/n$, which can be used to form model selection criteria such as AIC, BIC and APRESS.

Initially, the OFR algorithm is used with the ERR metric. However, the ERR only measures the linear dependencies. Some new metrics have been developed to measure the nonlinear dependencies, for example, the mutual information. The mutual information can be defined as:

$$I(x, y) = \sum_{x \in \chi} \sum_{y \in \gamma} p(x, y) \ln \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{2.19}$$

where x and y are two random discrete variables with alphabet χ and γ, and with a joint probability mass function $p(x, y)$ and marginal probability mass function $p(x)$ and $p(y)$ (Wei & Billings, 2008a). Mutual information measures the amount of information that one variable shares with another and can be incorporated into the OFR algorithm in the same way as the ERR metric.

In recent years, several variants have been introduced to improve the performance of NARX model and OFR algorithm, for example, the wavelet NARX model (Billings & Wei, 2005a; Wei & Billings, 2005b; Wei & Billings, 2004a; Wei & Billings, 2004b), the

iterative search algorithm (Wei & Billings, 2008a), the common/robust model structure selection method (Li et al., 2016; Gu & Wei, 2018a), etc.

2.2.3 Parameter Estimation

Assume that a total of n model terms are selected. Through an orthogonalization procedure, a unity upper triangular matrix A, along with auxiliary parameter vectors $g = [g_1, g_2, \dots, g_n]$, can be calculated as:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & & a_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix} \tag{2.20}$$

$$a_{11} = a_{22} = \cdots = a_{nn}^{(k)} = 1 \tag{2.21}$$

$$a_{rj} = \frac{(q_r)^T \varphi_{l_j}}{(q_r)^T q_r} \quad (r = 1, 2, \dots, j-1 \; and \; j = 2, 3, \dots n) \tag{2.22}$$

$$g_j = \frac{(y)^T q_j}{(q_j)^T q_j} \; (j = 1, 2, \dots, n) \tag{2.23}$$

Then the model parameter vector $\theta = [\theta_{l_1}, \theta_{l_2}, \dots, \theta_{l_n}]$, can be estimated from the triangular equations $A\theta = g$.

2.2.4 Model Selection

Among various model selection methods, Akaike information criterion (AIC) and Bayesian information criterion (BIC) are two most popular measures. Since AIC was firstly proposed in 1974 (Akaike, 1974), many variations of AIC have been developed for model selection. For example, the second-order Akaike information criterion (AICc) was developed for small sample size data modelling problems in 1989 (Hurvich & Tsai, 1989; Brockwell & Davis, 2013); the AIC was designed to approximately estimate the Kullback-Leiber information of models in 1998 (Akaike, 1998); also, the delta AIC and the Akaike weights were introduced to measure how much better the best model is when compared with the other models. In the model selection process, the AIC, delta AIC and AIC weights are calculated for each candidate model. Usually, the 'best' model is chosen to be the model with the smallest AIC; the delta AIC calculates the difference between

the AIC of each model and the smallest AIC of the 'best' model (Symonds & Moussalli, 2011); the AIC weight is ranged from 0 to 1, which is an analogous to the probability that a candidate model is the best choice (Buckland, Burnham & Augustin, 1997). Drawn on these theories, some model averaging approaches were also developed, for example, the natural averaging method (Buckland, Burnham & Augustin, 1997) and full model averaging method (Lukacs, Burnham & Anderson, 2010). Over the past few decades, AIC and its variations have been used to solve a wide range of model selection problems including those in ecology (Johnson & Omland, 2004) and phylogenetics (Posada & Buckley, 2004), among others. Both AIC and BIC have been widely applied on model selection problems. However, there still exists large room for improvement. For example, it lacks evidence that the two criteria can also work well for complex nonlinear system identification problems. AIC and BIC can be calculated as (Akaike, 1974; Gchwarz, 1978):

$$AIC(k) = -2\,ln(L) + 2k \tag{2.24}$$

$$BIC(k) = -2\,ln(L) + k\,ln(N) \tag{2.25}$$

where k is the number of fitted parameters in the model, L is the maximum likelihood estimate for the model and N is the sample size. For least square based regression analysis, AIC and BIC can be directly calculated by using MSE, as (Hurvich & Tsai, 1989):

$$AIC(k) = N\,ln(MSE(k)) + 2k \tag{2.26}$$

$$BIC(k) = N\,ln(MSE(k)) + k\,ln(N) \tag{2.27}$$

Equations (2.26) and (2.27) are and their variants have been applied for nonlinear and generalized linear model identification (see for example Blake and Kapetanios, 2003; Liu et al., 2007; Wei et al., 2007; Egrioglu et al., 2008).

Although AIC and BIC have been widely applied on model selection problems. However, there still exists large room for improvement. For example, it lacks evidence that the two criteria can also work well for complex nonlinear system identification problems. Although AIC and BIC can usually produce good model selection result based on the assumption that the 'true' model is among the candidate models, they may fail to select the best model when the system is very complex and neither of the candidate models can sufficiently represent the data. These situations often occur when the model

structure or some prior information is unknown. To solve the model selection problem of nonlinear system identification, the cross-validation (CV) based criterion (Stone, 1974) and its two variations, the Leave-One-Out (LOO), also called Predicted Residuals Sum of Squares (PRESS) (Allen, 1974; Hong, Sharkey & Warwick, 2003; Chen et al., 2004), and generalised cross-validation (GCV) (Golub, Health & Wahba, 1979), were developed. Most recently, a modified generalised cross-validation criterion, also known as adjustable predicted error sum of squares (APRESS), was also proposed for nonlinear systems identification (Billings & Wei, 2008).

Table 2.1 The advantage and disadvantage of AIC, BIC and APRESS

Criterion	Advantage	Limitation
AIC	• AIC minimizes useful risk function when true model is not a candidate and the model is complex.	• AIC-based model performs not well for out-of-sample data. • AIC-based model is often more complicated
BIC	• BIC is consistent in selecting true model when model is a candidate. • BIC-based model has better out-of-sample performance	• BIC is not consistent when the model is too complex or the uncertainty is too strong
APRESS	• APRESS is easy to implement in the OFR algorithm for nonlinear dynamic modelling. • APRESS have been applied for nonlinear model selection of many applications.	• APRESS has a tuning parameter so that it needs a figure to determine the optimal turning point

The APRESS is given as (Billings & Wei, 2008):

$$APRESS(n) = \left(\frac{N}{N-\lambda n}\right)^2 MSE(n) \tag{2.28}$$

where N is the number of observations, n is the number of selected model terms, λ is a small positive number and $MSE(n)$ is the mean square error. The optimal number of model terms is often chosen as:

$$n_{optimal} = arg \min_{1 \le n \le M} \{APRESS(n)\} \tag{2.29}$$

From the investigation of the literature, a summary of the reported advantages and limitations of the AIC/BIC/APRESS is given in Table 2.1. It can be noted that each of the three criteria contains two components: the first component measures the prediction error, which indicates how well the model fits the data. The second component is the cost function, which is used to penalize the model when more model terms (also called parameters in statistics) are added to the model. Therefore, there is a trade-off between the better fit and the model complexity. In general, the value of the criterion decreases when a first few model terms are included in the model, because of the reduction of prediction error. When an enough number of model terms are included, the penalty component becomes significant, leading to increased value. Thus, the model with a minimum value is then treated as an optimal choice with both good prediction performance as well as parsimonious representation of the system.

2.2.6 Correlation Test

In system identification procedure, model validation is a fundamental part to examine whether the model represent the system correctly. In fact, for many real data modelling problems, the existence of uncertainty might lead to biased estimation. Therefore, model validation is important to evaluate the efficiency and accuracy of the identified model.

A set of statistical correlation tests were developed for nonlinear model testing and validation (Billings & Voon, 1983; Billings & Voon, 1986; Zhang, Zhu & Longden, 2007). The model residual will be unpredictable if and only if these following conditions are satisfied (Billings & Voon, 1986):

$$\begin{cases} \phi_{\xi\xi}(\tau) = \delta(\tau), \ \forall\tau \\ \phi_{u\xi}(\tau) = 0, \ \forall\tau \\ \phi_{\xi(\xi u)}(\tau) = 0, \ \tau \ge 0 \\ \phi_{(u^2)'\xi}(\tau) = 0, \ \forall\tau \\ \phi_{(u^2)'\xi^2}(\tau) = 0, \ \forall\tau \end{cases} \tag{2.30}$$

where $\xi(t)$ is the model residual of the OSA model prediction, $(u^2)'\xi = u^2(t) - \overline{u^2}$ and $(\xi u)(t) = \xi(t+1)u(t+1)$, and the cross correlation function ϕ between two signials is defined as:

$$\phi_{xy}(\tau) = \frac{\sum_{t=1}^{N-\tau}[x(t)-\bar{x}][y(t+\tau)-\bar{y}]}{\sqrt{\sum_{t=1}^{N}[x(t)-\bar{x}]^2 \sum_{t=1}^{N}[y(t)-\bar{y}]^2}} \tag{2.31}$$

Other approaches such as chi-squared test can also be used for model validation. These validations tests are compatible with many other model types including Volterra series, wavelet models, etc (Billings, 2013; Wei & Billings 2004a).

2.2.7 Model Prediction

To evaluate the performance of model prediction, the one step ahead (OSA) and model predicted output (MPO) prediction are generated and compared to the system observation. Consider a simple linear model:

$$y(t) = ay(t-1) + bu(t-1) + cu(t-2) + e(t) \tag{2.32}$$

Assume that a number of observations for the system input $u(t)$ and output $y(t)$ are available, the OSA can be calculated as:

$$\begin{cases} \hat{y}(3) = ay(2) + bu(2) + cu(1) \\ \hat{y}(4) = ay(3) + bu(3) + cu(2) \\ \quad\quad ... \\ \hat{y}(k) = ay(t-1) + bu(t-2) + cu(t-2) \end{cases} \tag{2.33}$$

Different from OSA prediction, the MPO is calculated from the identified model driven only by the given input. The MPO is defined as:

$$\begin{cases} \hat{y}(1) = y(1) \\ \hat{y}(2) = y(2) \\ \hat{y}(3) = a\hat{y}(2) + bu(2) + cu(1) \\ \hat{y}(4) = a\hat{y}(3) + bu(3) + cu(2) \\ \quad\quad ... \\ \hat{y}(k) = a\hat{y}(t-1) + bu(t-2) + cu(t-2) \end{cases} \tag{2.34}$$

The problem with the OSA prediction is that the measured output is used at each step of the calculation, so that the errors are suppressed. As the MPO uses the predicted output at each step, it is usually used to evaluate the long-term prediction of the model. Therefore, a model with good OSA prediction can still be insufficient, biased and unstable.

Sometimes it is essential to use MPO to validate the model, especially when the long-term prediction is desired.

Commonly used statistics include correlation coefficient, prediction efficiency (PE), absolute fraction of variance (R^2), prediction efficiency (PE), normalized root-mean-square error (NRMSE), root mean square error (RMSE), mean square error (MSE) and mean absolute deviation (MAD). Some of the statistics can be calculated as follows:

$$R^2 = 1 - [\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 / \sum_{i=1}^{N}(y_i - \bar{y})^2] \tag{2.35}$$

$$MSE = \sum_{i=1}^{N}(\hat{y}_i - y_i)^2 / N \tag{2.36}$$

$$RMSE = [\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 / N]^{1/2} \tag{2.37}$$

$$MAD = \sum_{i=1}^{N}|\hat{y}_i - y_i| / N \tag{2.38}$$

$$PE = 1 - \frac{\sigma_{error}^2}{\sigma_{observed}^2} \tag{2.39}$$

where $\hat{y}_i$ is the model prediction, y_i is the observation, N is the number of samples; $\sigma_{observed}^2$ is the variance of the observation and σ_{error}^2 is the variance of the error between the model prediction and observation.

2.2.8 Hypothesis Test

The hypothesis testing can be used to detect the spurious model terms and refine the resultant model when the input is poorly designed (Wei & Billings, 2008a). The hypothesis for testing the significance of the regression coefficient, for instance θ_j in the model (2.10), is:

$$H_0: \theta_j = 0 \qquad H_1: \theta_j \neq 0 \tag{2.40}$$

The corresponding regressor x_j can be removed from the model if there is no sufficient reason to reject the null hypothesis $H_0: \theta_j = 0$. The test statistic for the hypothesis is:

$$t_0 = \frac{|\widehat{\theta_j}|}{se(\widehat{\theta_j})} \tag{2.41}$$

where $se(\widehat{\theta_j})$ is the standard error of the regression coefficient θ_j. The details of implementing the hypothesis testing in the system identification procedure is summarised in (Wei & Billings, 2008a).

2.2.9 Extended Least Square Method

Note that the noise signal $e(t)$ in Eq. (2.2) may be a correlated or colored noise sequence, which is generally unobserved and is often replaced by the model residual sequence. Let $\hat{f}(\cdot)$ represent an estimator for the model $f(\cdot)$, the model residuals $\varepsilon(t)$ can then be estimated as

$$\varepsilon(t) = y(t) - \hat{y}(t) = y(t) - f[y(t-1), \dots, y(t-n_y), x_1(t) \dots,$$
$$x_1(t-n_u), x_2(t), \dots, x_2(t-n_u) \ \dots, x_M(t), \dots, x_M(t-n_u)] \quad (2.42)$$

To reduce the effect of the noise, the algorithm includes an ELS-type procedure to compute the prediction errors $\varepsilon(t)$ and use the value of $\varepsilon(\cdot)$ from the previous iteration so that noise model terms are included in model $f(\cdot)$ (Billings, 2013). With the extra moving average components, the NARX model can be further developed to the NARMAX model, which can be described as:

$$y = f[y(t-1), \dots, y(t-n_y), x_1(t) \dots, x_1(t-n_u), x_2(t), \dots, x_2(t-n_u) \ \dots, x_M(t), \dots, x_M(t-n_u) + f^{[pn]}[[y(t-1), \dots, y(t-n_y), x_1(t) \dots, x_1(t-n_u), x_2(t), \dots, x_2(t-n_u) \ \dots, x_M(t), \dots, x_M(t-n_u), \varepsilon(t-1), \dots, \varepsilon(t-n_e)] + f^{n}[\varepsilon(t-1), \dots, \varepsilon(t-n_e)] \quad (2.43)$$

where $f(\cdot)$ is the NARX sub-model identified in first step, $f^{[pn]}(\cdot)$ is the process input-output noise-related sub-model and $f^{[n]}(\cdot)$ is the purely noise process sub-model. In some situations, it may be possible to use just a linear noise model where

$$f^{[n]}(\cdot) = \alpha_1 \varepsilon(t-1) + \cdots + \alpha_{n_e} \varepsilon(t-n_e) \quad (2.44)$$

If this is insufficient, then $\varepsilon(t-p)$ for $p = 1,2, \dots\ n_e$ can be included in model, where the basic regressor vector is defined as $y(t-1), \dots, y(t-n_y), u(t-1), \dots, u(t-n_u), \varepsilon(t-1), \dots, \varepsilon(t-n_e)$.

2.3 Review of Data Modelling and Data Mining Methods

Many techniques have been used for time series forecasting and modelling problems, for example, neural network (Haykin, 1994; Wang et al., 2017; Chen & Billings, 1992; Chen & Billings, 1990), wavelet models (Billings & Wei, 2005; Wei & Billings, 2004; Zhang

& Benvenise, 1992), etc. This section presents brief review on some commonly-used data modelling and data analysis methods.

2.3.1 LASSO Method and Regularization Methods

LASSO is a regression analysis method with variable selection and regularization. It was originally proposed as one of the least square methods. Assume that $y(t)$ is the output variable and $x_j(t)$ $(j = 1,2, \dots, M)$ are the input variables. The objective of lasso is to solve

$$\min_{\beta_0,\beta} \left(\frac{1}{N}\sum_{t=1}^{N}\left(y(t) - \beta_0 - \sum_{j=1}^{M}\beta_j x_j(t)\right)^2\right) \text{ subject to } \sum_{j=1}^{M}|\beta_j| \leq z \qquad (2.45)$$

where z is a prespecified parameter that determines the amount of the regularisation. The estimator of the LASSO method can be considered as (Tibshirani, 1996):

$$\hat{\beta}_j = \hat{\beta}_j^{\,OLS} \max\left(0, 1 - \frac{N\lambda}{\left|\hat{\beta}_j^{\,OLS}\right|}\right) \qquad (2.46)$$

where $\hat{\beta}_j^{\,OLS} = (X^T X)^{-1} X^T y$ is the least square estimate. The formula (2.46) is a sub gradient method that translates values toward zero instead of setting smaller values to zero and leave large values unchanged. This can be compared to ridge regression (Hoerl & Kennard, 1970), where the estimator is:

$$\hat{\beta}_j = (1 + N\lambda)^{-1} \hat{\beta}_j^{\,OLS} \qquad (2.47)$$

The ridge regression shrinks all the values by the factor $(1 + N\lambda)^{-1}$. Another method is the best subset selection, which is defined as:

$$\hat{\beta}_j = \hat{\beta}_j^{\,OLS} I\left(\left|\hat{\beta}_j^{\,OLS}\right| > \sqrt{N\lambda}\right) \qquad (2.48)$$

where the indicator function I is 1 if its argument is true and 0 otherwise. It can be seen that the LASSO method shrink all the values but also set some of the values to 0. In addition, some variants of the LASSO method have been developed, for example, elastic net (Zhou & Hastie, 2005), group LASSO (Yuan & Lin, 2006), fused LASSO (Tibshirani, et al., 2005), etc.

However, LASSO suffers from some weakness. For example, when the number of variables M is much large than the number of samples N, lasso is only capable to select

M features, due to the nature of convex optimization problem. In addition, some authors have found that LASSO is not able to select a group of correlated terms (Hong & Chen, 2012).

2.3.2 Wavelet NARX Model

Wavelet are usually chosen as the function components in additive models, because of its approximation capabilities (Wei, Billings & Balikhin, 2002). The additive models uses an ordinary linear-in-the-parameters form which can be solved by the least square algorithms. From the wavelet theory (Baford, Fazzio & Smith, 1992), any function can be expressed as the wavelet multiresolution expansions:

$$g(x) = \sum_k \alpha_{j0,k}\, \varrho_{j0,k} + \sum_{j \geq j0} \sum_k \beta_{j,k}\, \varpi_{j,k}(x) \tag{2.49}$$

where ϱ is the mother wavelet and ϖ is the associated scale function, $\alpha_{j0,k}$ and $\beta_{j,k}$ are the wavelet decomposition coefficients. Although there are many mother functions that can be used in the decomposition, few of them are suitable for nonlinear system identification. The limitation is that most existing wavelet networks are limited to handling problems in low-dimensional space due to the curse of dimensionality. Later, the wavelet based NARX model and wavelet networks were developed for the identification of nonlinear input-out systems (Wei & Billings, 2004; Billings & Wei, 2005a; Billings & Wei, 2005b). The new models uses sub-models to approximate the nonlinear function, which can be defined as:

$$F[x(k)] = c_0 + F_1[x(k)] + F_2[x(k)] + \cdots + F_n[x(k)] \tag{2.50}$$

where c_0 is constant and the individual wavelet sub-models $F[\cdot]$ are of the form:

$$F_1[x(k)] = \sum_{i=1}^{n} f_i[x_i(k)] \tag{2.51}$$

$$F_2[x(k)] = \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_{ij}\left[x_i(k), x_j(k)\right] \tag{2.52}$$

$$F_M[x(k)] = f_{12\ldots M}[x_1(k), x_2(k), \ldots, x_M(k)] \tag{2.53}$$

The function components f can be wavelet networks or multi-resolution wavelet decomposition (Billings & Wei, 2005b). The new model holds the attractive features possessed by the wavelets and can be used in nonlinear dynamical systems identification. Wavelet is effective to describe data at different scales with a wide range of functions and

can be tuned or refined without interfering with the rest of the model. Thus, a complex nonlinear system can be well represented using only a limited number of basis functions.

2.3.3 State-space Model

Another commonly used model is the state-space model. The advantage of state-space model is that it provides a clear structure of model variables that can represent the physical variables in the real world, in both static and dynamic, linear and nonlinear structure. The general form of discrete-time state-space model for single-input, single-output systems can be expressed as:

$$\begin{cases} x_1(k) = F_1[x_1(k-1), \dots, x_M(k-1), u(k-1)] + e_1(k) \\ x_2(k) = F_2[x_1(k-1), \dots, x_M(k-1), u(k-1)] + e_2(k) \\ \vdots \\ x_M(k) = F_M[x_1(k-1), \dots, x_M(k-1), u(k-1)] + e_M(k) \\ y(k) = G[x_1(k), \dots, x_M(k), u(k)] + \eta(k) \end{cases} \tag{2.54}$$

The disadvantage of the state-space model is that all the variables in the model need to be measured and the relationship of the variables need to be known (Billings, 2013). This is because in order to establish a state-space model, a series of sub-models must be identified, which can only be achieved only if the observations of systems input, output and the state variables are available. It is extremely difficult when dealing with 'black-box' modelling task or the systems is nonlinear and complicated.

2.3.4 Neural Network

Neural network is one of the most commonly-used model type for data-driven modelling task (Zurada, 1992). Over the past few decades, it has been developed and applied in many research areas such as data modelling, signal processing, control, etc. The neural network can be used to represent the data using some learning algorithm (Haykin, 1994; Wang, et al., 2017). Commonly used neural network for SISO system contains one input layer, one hidden layer and one output layer. The nodes of the layers are connected with associated weights, which define the relationship between the system input and output. The identification of neural network contains several steps:

Determine the input layers and output layer.

For the neural network fitting problem, the input variables are entered through the input layers of the neural network and the response variable is defined as the output layer of the neural network. Note that dynamic problems involve lagged variables, which needs to be derived from the input and output signals.

Initialization of the network

For conventional neural network, the number of neurons of hidden layer needs to be determined. For deep neural network, the number of the hidden layers needs to be chosen. Once the number of the hidden layers and neurons are chosen, the structure of neural network can be initialized. The weights of the connections are initialized with random values. The output of that node (also called neuron) is defined by the activation function. This output is then used as input for the next node and so on until a desired representation of the data is found.

There are many activation functions which can be used in neural network, for example, sigmoid tangent function, saturation function, sigmoid function, hyperbolic tangent function, Gaussian function, multi-quadratic function, fractional multi-quadratic function, inverse multi-quadratic function, fractional inverse multi-quadratic function and thin-plate spline function, etc. Some of the activation functions are defined as (Billings, 2013):

- Sigmoid function

$$\varphi(v) = \frac{1}{1+e^{-av}}, \quad a > 0 \tag{2.55}$$

- Gaussian function

$$\varphi(v) = e^{-\frac{v^2}{2\sigma^2}}, \quad \sigma > 0 \tag{2.56}$$

- Multi-quadractic function

$$\varphi(v) = \sqrt{v^2 + \alpha^2}, \quad \alpha > 0 \tag{2.57}$$

- Saturation (threshold) function

$$\varphi(v) = \begin{cases} -a, & v \le -c \\ v, -c \le v \le c, & a > 0, c \ge 0 \\ a, & v \ge c \end{cases} \tag{2.58}$$

- Hyperbolic tangent function

$$\varphi(v) = \frac{e^{av} - e^{-av}}{e^{av} + e^{-av}}, \quad a > 0 \tag{2.59}$$

where c is the vector representing function center and a is a parameter of the function.

Estimation of the weights of each layer.

The network is trained by operating on the prediction error between the actual output and desired output of the network, to change the connections between the nodes. With the Matlab Toolbox, the weights of the neural network can be estimated using several methods, for example, Levenbery-Marquardt method, Bayesian Regularization method or Scaled Conjugate Gradient methods, etc. Levenbery-Marquardt method is the default algorithm when the computation memory is sufficient.

The advantage of neural network is that it can achieve relatively higher performance for complex data in high dimensional space. However, the neural network is too complicated to understand. The structure of neural network is not transparent and it is impossible to know the role of the model terms/components throughout the neural network.

2.4.5 Deep Learning

Comparing to the conventional neural network with single hidden layer, deep learning allows the use multiple-layers network to process the data. In recent years, deep learning methods have been successfully applied to a wide range of research areas, for example, speech recognition (Hinton, et al, 2012), face and pose detection (Garcia & Delakis, 2004), etc. With multiple level of representation, deep-learning methods can learn very complex function by composing simple but non-linear modules at each level. At each level, the raw input is transformed into a representation at a higher, slightly more abstract level (Lecun, Bengio & Hinton, 2015). These layers are not designed by human, but learned from data using a general-purpose learning procedure.

Deep learning is making major advances in solving supervised and unsupervised problems. However, deep neural network is not widely used for systems identification problems. Most of the deep learning models are designed for image recognition, classification, etc. Therefore, it is necessary to design and apply deep learning network

for system identification problems. Another issue is that the deep learning network is even more complex than neural network, which makes it impossible to obtain a transparent and simple representation.

2.4 Review of Uncertainty Analysis Methods

The existence of uncertainty in the modelling process could cause negative effect on the performance of identified model. This section reviews some uncertainty analysis approaches, for example, cloud model and cloud transformation, fuzzy sets, noise modelling, etc. One of the objectives of this research is to incorporate the novel uncertainty analysis methods into the nonlinear systems identification and data modelling problems. Thus, the uncertainty analysis methods are investigated to access the feasibility of their ability in data modelling and systems identification.

2.4.1 Cloud Model and Cloud Transformation

Cloud model is a cognitive model which provides a way of bidirectional transformation between a qualitative concept 'cloud' and the quantitative data 'cloud drops' (Wang, Xu & Li, 2014). The concept cloud is described by three numerical characteristics, namely ex (expectation), en (entropy) and he (hyper entropy). Similar to normal distribution, ex is the expectation of all the elements in the set and en is the variance of the distribution. he depicts the degree of departure from normal distribution of cloud model (Wang, Xu & Li, 2014). Based on the theorem that any distribution can be represented by the sum of several normal distributions, the cloud model can be seen as an extension of normal distribution: when he equals to 0, the cloud model become actually a normal distribution. he is often regarded as an extra variable in practical situation, such as psychological quality of an athlete.

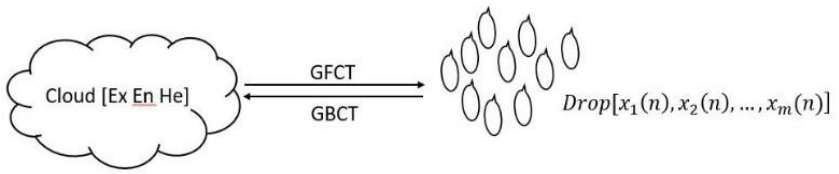

Figure 2.2 Cloud model and generic forward/backward cloud transformation

The bridges between cloud model and cloud drops is cloud transformation. The common used cloud transformation is generic forward and backward cloud transformation (GFCT and GBCT). The forward transformation is used to generate cloud drops from a known cloud model. The backward transformation is used to identify the cloud model from a sequence of cloud drops. In previous research, an ideal cloud backward transformation is also studied (Zhang et al., 2016). However, it is not feasible in real life as the groups of cloud drops could hardly be obtained in advance. The representation of forward and backward cloud transformation can be illustrated as follows:

$$[x_1, x_2, \dots\ x_n] \overset{cloud\ transformation}{\Longleftrightarrow} cloud\ (ex, en, he) \tag{2.60}$$

where $cloud\ \ (ex, en, he)$ is a cloud concept of characteristics modelled from n samples numerical data 'cloud drops' $[x_1, x_2, \dots\ x_n]$.

Algorithm 1: *Generic backward cloud transformation (GBCT)*

Input: *Cloud drops* $\{x_1\ x_2\ \dots\ x_n\}$.

Output: ex, *en and* he.

Step 1: *Calculate the sample mean* ex

$$ex = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{2.61}$$

Step 2: *Make the cloud drops* $\{x_1\ x_2\ \dots\ x_n\}$ *divide into* α *groups randomly, and each group will have* β *samples (note that* $\alpha \times \beta = n$*, so that the number of cloud drops do not change after resampling). The resampled cloud drops can be described as* x_{ij} *,where* $i = 1, 2, \dots, \alpha$ *and* $j = 1,2, \dots, \beta$.

Step 3: *Calculate the sample mean and variance of each group:*

$$u_i = \frac{1}{\beta}\sum_{j=1}^{\beta} x_{ij} \tag{2.62}$$

$$\sigma_i = \frac{1}{\beta-1}\sum_{j=1}^{\beta}(x_{ij} - u_i)^2 \tag{2.63}$$

where $i = 1,2, \dots, \alpha$.

Step 4: *Calculate estimated* en *and* he.

$$en^2 = \frac{1}{2}\sqrt{4EY^2 - 2DY}, \qquad he^2 = EY - En^2 \tag{2.64}$$

where $EY = \frac{1}{\alpha}\sum_{i=1}^{\alpha}\sigma_i$ *and* $DY = \frac{1}{\alpha-1}\sum_{i=1}^{\alpha}(\sigma_i - EY)^2$.

Algorithm 2: *Generic forward cloud transformation (GFCT)*

Input: ex, en *and* he

Output: *Cloud drops* x_{ij} $(i = 1, 2, \dots, \alpha', j = 1, 2, \dots, \beta')$

Step 1: *Generate* α' *normally distributed random numbers* δ_i $(i = 1, 2, \dots, \alpha')$ *with expectation* en *and variance* he^2 *;*

Step 2: *For each* δ_i *in step 1, generate* β' *normally distributed random numbers* x_{ij} $(i = 1, 2, \dots, \alpha', j = 1, 2, \dots, \beta')$ *with expectation* Ex *and variance* ${\delta_i}^2$.

Step 3: *Calculate the certainty degree* $\mu(x_{ij}) = exp\left\{-\frac{(x_{ij}-Ex)^2}{2\delta_i^2}\right\}$ *for each* x_{ij} $(i = 1, 2, \dots, \alpha', j = 1, 2, \dots, \beta')$.

Step 4: x_{ij} $(i = 1, 2, \dots, \alpha', j = 1, 2, \dots, \beta')$ *are the cloud drops. The total number of the generated cloud drops is* $\alpha' \times \beta'$.

The generic cloud transformation achieves the transformation between intension and extension of the cloud concept. The advantage of cloud model is that it provides a way to describe a distribution with only three parameters that cannot be characterized by traditional normal distribution. The cloud transformation is better and more powerful than normal distribution in that: i) it includes normal distribution as a special case; and ii) many data in real life do not follow a normal distribution.

2.4.2 Probabilistic Model

A series of researches have been conducted to deal with uncertainty using techniques such as regression, machine learning and statistical analysis, and so on. Probability theory is one of the effective tools in uncertainty analysis. The central topics of probability theory are random variables and stochastic processes. Thus, the probabilistic model can be applied to describe random uncertainty, with different probabilistic distributions for example Gaussian distribution being used as an approximation to a large number of random phenomena (Wang, Xu & Li, 2014).

Numerous studies on these probabilistic approaches have been conducted over the centuries since then. Bayesian theory and related techniques are among the commonly used methods to calculate the probability density of distributions, providing quantitative descriptions of uncertainty. The probabilistic models often provide confidence intervals with specific distributions of model parameters and predictions, for example, Gaussian

process model (Arendt, Apley & Chen, 2012). However, in some cases, the distributions of variables cannot be known in advance, or need to assume some very specific distributions. Thus, finding a robust and adjustable representation of the uncertainty is still an open question for data modelling problems.

2.4.3 Fuzzy Set

Fuzzy set provides an alternative to represent uncertainty (Zadeh, 1965). As an extension of the classical notion of set, it has developed to be the main tool dealing with fuzzy uncertainty and successfully achieved a lot of applications (Bustince, et al, 2016; Kim, 2015). The main extensions of the fuzzy sets include: Type-n fuzzy set, Interval-Valued Fuzzy Set, Set-valued Fuzzy Set, Bipolar-Valued Fuzzy Set, Hesitant Fuzzy Set, m-Polar-Valued Fuzzy set, etc. (Bustince, et al, 2016).

The use of membership function in fuzzy theory provides a novel gradual assessment, to evaluate how much degree of an element belongs to a fuzzy concept, and can be applied in a wide range of fields where the information in incomplete and imprecise. The biggest challenge of fuzzy set is to find the optimal fuzzy rules to represent the randomly distributed data. The process of identifying the fuzzy rules can be time consuming, due to the variation of data types in real world.

2.4.4 Noise Modelling

Another approach to deal with uncertainty is noise modelling techniques. The uncertainty is often regarded as a noise sequence. Many algorithms have been proposed addressing the noise modelling process.

In some situations where the system is nonlinear, the model residual $e(t)$ is highly unlikely to be Gaussian. In this case, there is still correlation between the noise $e(t)$ and the model inputs. The noise can be modelled in many ways, including the traditional recursive prediction error method (PEM), generalised least squares (GLS), instrumental variables (IV), or an extended least square procedure (Young, 1984; Norton, 1986; Lennart, 1999; Sodestrom & Stoica, 1989; Billings, 2013). As discussed in the section 2.2.9, the noise sequence of the NARX model can be learnt as part of the model fitting using the extended least square (ELS) method (Billings, 2013).

2.4.5 Model Averaging

A single model may not be reliable for some worse-case data scenarios. The collective use of information from many models, however, may help improve the overall model performance. Model averaging is therefore also a widely applied method to reduce or eliminate the negative effect caused by model uncertainty. It was argued that model averaging is a much more reliable method than other techniques such as the statistical tests (Plumper & Neumayer, 2012). Model averaging often involves a resampling process of the original data, through some resampling approach such as bootstrap method (Smith, et al, 2014). The resampling and related methods have been applied to effectively minimise and reduce the variance of estimated parameters for dynamic data modelling (Wei & Billings, 2009). Other resampling methods for example cross validation and jack-knife have also been widely applied to data modelling and analysis (Devijver & Kittle, 1982; Efron, 1983; Efron & Tibshirani, 1993). Model identified from different resampled data may be different. So, how to effectively make good use of a number of models to produce a robust or reliable model that well represents the original data is the core interest of model averaging. Extensive research on model averaging has been done to gather information from several or many sub-models identified from a number of sub-datasets, including Bayesian averaging. A key point is to employ resampling and model averaging process to reduce or eliminate the overfitting and biased estimation in particular when the available data is small.

Model averaging approaches such as AIC and BIC based averaging methods have been used in many applications (Cade, 2015; Asatryan & Feld, 2015; Moral-Benito, 2015; Kontis et al., 2017). The model averaging approach with AIC involves the computation of the delta AIC and the Akaike weights. The delta AIC can be calculated as (Symonds & Moussalli, 2011):

$$\Delta AIC_{c_i} = AIC_{c_i} - AIC_{c_{min}} \tag{2.65}$$

where $\mathrm{AIC_{c_i}}$ is the AIC value for the *i*-th candidate model, $\mathrm{AIC_{c_{min}}}$ is the minimum AIC of all the M candidate models, and $i = 1, 2, \dots, M$. The Akaike weight indicates the probability that an individual candidate model is the best model. The Akaike weight for *i*-th candidate mode is computed as (Buckland, Burnham & Augustin, 1997):

$$\omega_i = \frac{exp\,(-0.5\Delta AIC_{c_i})}{\sum_{j=1}^{M} exp\,(-0.5\Delta AIC_{c_j})} \tag{2.66}$$

where ω_i is the Akaike weight for the *i*-th candidate model and $i = 1, 2, \ldots, M$. Then, the averaged parameter estimate of 'full model averaging' is calculated as follows:

$$\hat{\bar{\beta}} = \sum_{i=1}^{M} \omega_i \hat{\beta}_i \tag{2.67}$$

To produce averaged model based on BIC and APRESS, a simple approach is to replaced AIC by BIC and APRESS, to calculate the BIC and APRESS weights of model parameters of all candidate models. The advantage of the averaged model is that it is in general more robust than the single 'best' model determined by the model selection criterion. This is because a single model only contains a limit number of model terms suggested by model selection criterion. If a model selection criterion fails to detect the correct number of model terms, the model terms of the single model may be insufficient to well represent the system. On the contrary, the averaged model uses the information of all the candidate models and each candidate model gives its contribution according to their weights based on the model selection criterion. Therefore, when the single model selected by the model selection criterion is not the best, the performance of the averaged model is usually better than that of the single model. However, it should also be noted that a model with more terms is not necessarily always better than a model with less terms, because some terms may be redundant and may deteriorate the model prediction performance. Therefore, it is not always true that the averaged model is better than a single model, but the averaged model is often more robust in case where there is large uncertainty in the data collection, model structure and model parameter, etc.

2.5 Summary

This chapter gives an overview of the system identification methods and uncertainty analysis approaches which are used in this thesis. The general process of system identification including the term selection, parameter estimation and model validation are reviewed. The discussion focusses on the implementation of the NARMAX model and OFR algorithm, which is popular and effective for the data-driven modeling problems. Some other common-used modelling approaches are discussed. In addition, a brief review of uncertainty analysis approach is presented, emphasizing on the new concept cloud model. The cloud transformation provides an effective tool to describe variable which is beyond normal distribution.

Chapter 3

ROBUST MODEL STRUCTURE SELECTION METHOD FOR SMALL SIZE DATA MODELLING PROBLEMS

3.1 Introduction

In model identification, the existence of uncertainty normally generates negative impact on the accuracy and performance of the identified models, especially when the size of data used is rather small. With a small data set, least squares estimates are biased, the resulting models may not be reliable for further analysis and future use. This chapter introduces a novel robust model structure selection (RMSS) method for model identification. The proposed method can successfully reduce the model structure uncertainty and therefore improve the model performances. Case studies on simulation data and real data are presented to illustrate how the proposed metric works for robust model identification.

3.2 Small Size Data Modelling Problems

Broadly speaking, data-based modelling approaches can be categorized into two groups: parametric and nonparametric. Nonparametric methods are those that do not make strong assumptions about the form of the mapping functions (that map the model "input" variables to the model "output" variables). Most existing artificial neural networks are

nonparametric approaches. In (Russell & Norvig, 2010) it is stated that "Nonparametric methods are good when you have a lot of data and no prior knowledge, and when you don't want to worry too much about choosing just the right features" (p.757). One of the advantages of neural networks is that in general they can achieve relatively higher performances in dealing with complicated data modelling problems defined in high dimensional space. However, the model structure of most neural networks is very complicated and cannot be simply written down. In addition, neural networks models often involve a large number of variables and take a long time for training. General neural networks models cannot provide a transparent model structure, where the significance of individual variables and the role of their interactions are invisible. Moreover, the implementation of some nonparametric approaches for example Bayesian networks normally would need a huge number of samples. In comparison with neural networks models, parametric NARX models use a nonlinear polynomial structure and often only need a small number of effective model terms to describe the system. It can be achieved by selecting a number of most important model terms by an orthogonal forward regression (OFR) algorithm (Chen, Billings & Luo, 1989; Wei, Billings & Liu, 2004), so that it generally only requires a relatively small number of input and output data points (Wei & Billings, 2008a; Billings & Wei, 2008)). In many applications (e.g. Bigg, et al., 2014; Billings, et al., 2013), where the main objective of the modelling tasks is not only to predict future behaviour, but also reveal and understand which model variables are most important and how the candidate variables interactively affect the system behaviour, parametric models are usually become a first choice.

Under some specific conditions and assumptions, most existing model identification methods work well and can provide sufficiently reliable models for most applications. However, in many cases where there is modelling uncertainty (e.g. in data, model form and structure, parameters, noise level, etc.), the identified models may lack reliability and thus less useful. This is particularly true when the available data set is small. This study focuses on parametric models and aims to answer the following challenging question. Given a small set of experimental data of a system, how to build a model that best represents the underlying system dynamics hidden in the data? Most data modelling approaches can generate good models that best fit the data themselves, but the models may not be able to represent well the inherent dynamics of the original system because of different kinds of uncertainties. For small data modelling problems, the difficulty of

finding reliable models is often exacerbated due to the small sample size of data. It is observed that for a small data modelling problem, small changes in a few or even a single sample can cause a large effect on model estimation. Thus, another question that arises is: how to reduce the model uncertainty (i.e. increase the model reliability) for small size data modelling problems?

It is not straightforward, if not impossible, to induce a robust model from a small sample size data, no matter what kind of system identification or data modelling algorithm are employed. In additional to noise and the size of samples, other types of factors can also lead to model uncertainty. For example, a data based modelling approach may just simply assumes a specific model type to represent the data but the specified model structure is completely different from the true system model; some driven variables may be immeasurable or ignored. All this is embedded in the aphorism "all models are wrong, but some are useful" (Box & Draper, 1987). In fact, for all system identification problems, model type selection and structure detection is usually an instrumentally important task. For the same data based modelling problem, different types of models often have different properties and performance, with different interpretation of the data. Even for the same model type, different algorithms could lead to different final model representations. The reason is simple: when the true model is unknown, all the identified models could be wrong because of uncertainty and the incompleteness of information. Effectively dealing with uncertainty (model structure, parameter, prediction, etc.) has become an important topic in many research fields, for example, soil changes (Robinson, Benke & Norng, 2015), carbon and water fluxes at the tree scale (Christina, 2016). In all scientific research, it nearly always needs to consider uncertainty, from various perspective such as, sources of uncertainty, techniques of quantifying uncertainty, decision making under strong uncertainty conditions, etc.

With the above observations, this study aims to develop a new approach to find a robust model structure to reduce uncertainty in model identification especially when sample size is small. Based on a data resampling approach, combined with an orthogonal forward regression (OFR) algorithm (Chen, Billings & Luo, 1989; Wei, Billings & Liu, 2004), a robust model structure selection (RMSS) method is designed to reduce model uncertainty and improve model performance. This is especially useful for the following two scenarios of data based modelling problem: i) modelling from multiple small sample size datasets (e.g. many datasets for a same system but generated under different experimental

conditions; ii) modelling for a non-stationary system where although the key system dynamics can be represented using a single model structure, different model parameters are needed to adaptively reflect the change of system behaviours at different times. In summary, the main contribution of this chapter lies in the new robust common model structure detection method for solving two challenging problems frequently encountered in practical system identification and data-driven modelling, namely, (a) reliable model identification from small sample data, and (b) robust common model determination from several or many experimental datasets.

3.3 Robust Model Structure Selection Method

Following the discussions in the previous section, the OFR method is used to select a small number of significant terms to establish a best model structure. For many real modelling tasks, there are several commonly seen situations where the OFR algorithm cannot be directly used to generate best models, for example: i). the data are usually recorded from a series of experiments under different experimental conditions, or the system itself is non-stationary and needs to be observed for a long-time scale. In these scenarios, the model structure might be varying with time and/or with the change of external environmental conditions. ii). The true model structure of the system is unknown and cannot be well represented by any of the candidate model terms in the dictionary. Thus, it is impossible to find a perfect model structure and there will always be uncertainty of model structure. iii). the data is corrupted with strong noises which makes the OFR estimation biased. The bias could be extremely obvious when data size is small, since a small change of a single term can bring a huge difference on the estimated model. Under these conditions, the OFR method may fail to find a best model structure that can well represent the system. Therefore, the RMSS method is needed for capturing and reducing the model uncertainty and thus improving the overall model predictive performance.

In the following, the RMSS method is proposed. The basic idea of the new method is first illustrate using a simple example, and the procedure of the method is then presented.

3.3.1 Basic Idea

Consider a scenario where a total number of K datasets are available, all of which are generated from a same system under some different conditions. The primary objective is to find a common model that best fits all the K datasets. The new method uses a concept of overall mean absolute error (OMAE); it is defined as the average of K individual mean absolute errors (MAE) which are calculated when a model (or a new model term is included in an existing model) to fit all the K datasets. Consider two datasets (as shown in Table 3.1) generated from the true system $y = 1.5x_1+0.1x_2+0.02x_3$, the OMAE can be calculated. Note that the first dataset is noise free, while the second dataset is affected by some noises $N(0,0.01)$.

Table 3.1 Variables of two datasets

	x_1	x_2	x_3	y
dataset 1	-0.3	0.1	-0.7	-0.4540
	0.1	-0.1	0.2	0.1440
	-0.9	1.0	-0.5	-1.460
	-0.9	-0.3	0.3	-1.3740
dataset 2	0.1	-0.7	0.4	0.0040
	0.6	0.6	0.5	1.1055
	0.9	-0.4	-0.1	1.2008
	-0.8	0.1	-0.9	-1.1119

Assuming that one and only one variable (among x_1, x_2, and x_3) is needed to fit the two datasets, then which one can give a minimum OMAE value? This can be done by calculating the individual MAE values one by one. For example, the individual mean absolute error $\epsilon_1^{(1)}$ of the variable $\mathrm{x}_1^{(1)}$ for dataset 1 can be calculated as:

$$\epsilon_1^{(1)} = \frac{1}{4}\left\|y^{(1)} - \alpha_1^{(1)} x_1^{(1)}\right\|_1 = \frac{1}{4}\left\|y^{(1)} - \frac{x_1^{(1)^T} y^{(1)}}{x_1^{(1)^T} x_1^{(1)}} x_1^{(1)}\right\|_1 = 0.0290 \qquad (3.1)$$

MAEs for x_2 and x_3 can be calculated in a similar way for datasets 1. Similar calculations can be performed to dataset 2. There is a total number of 6 individual MAEs.

The OMAEs can be calculated, as shown in Table 3.2. As the OMAE value of x_1 is smaller than the other two, x_1 should be the best choice for fitting the two datasets. Note that once the first model term is determined, a second model term can be chosen to join the first one, and then a third one, and on. The detailed descriptions of the general procedure of the RMSS method is given in next section.

Table 3.2 MAE and OMAE values of x_1, x_2, and x_3

Term	MAE (dataset 1)	MAE (dataset 2)	OMAE
x_1	0.0290	0.1313	0.0802
x_2	0.4954	0.8657	0.6805
x_3	0.6926	0.5910	0.6418

3.3.2 Robust model structure selection method

The RMSS method can be summarized into several steps:

a). Resampling process (for small size data)

Assume that the original data can be described by a $N \times M$ matrix $\boldsymbol{d}$ as follows:

$$\boldsymbol{d} = [\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_M] = \begin{bmatrix} \varphi_1(1) & \varphi_2(1) & \cdots & \varphi_M(1) \\ \varphi_1(2) & \varphi_2(2) & & \varphi_M(2) \\ \vdots & & \ddots & \vdots \\ \varphi_1(N) & \varphi_2(N) & \cdots & \varphi_M(N) \end{bmatrix} \tag{3.2}$$

where $\{\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_M\}$ is M candidate basis vectors (generated from M candidate model terms) and N is the number of data points. The original dataset can be regrouped to form K sub-datasets through some resampling methods e.g. random sampling or bootstrap (see (Wei & Billings, 2009a; Wei & Billings, 2009b) and the references therein). The k-th sub-dataset can be described by a $N' \times M$ matrix:

$$\boldsymbol{d}^{(k)} = \left[\boldsymbol{\varphi}_1^{(k)}, \dots, \boldsymbol{\varphi}_M^{(k)}\right] = \begin{bmatrix} \varphi_1^{(k)}(1) & \varphi_2^{(k)}(1) & \cdots & \varphi_M^{(k)}(1) \\ \varphi_1^{(k)}(2) & \varphi_2^{(k)}(2) & & \varphi_M^{(k)}(2) \\ \vdots & & \ddots & \vdots \\ \varphi_1^{(k)}(N') & \varphi_2^{(k)}(N') & \cdots & \varphi_M^{(k)}(N') \end{bmatrix} \tag{3.3}$$

where the associated candidate basis vectors become $\left\{\boldsymbol{\varphi}_1^{(k)}, \ldots, \boldsymbol{\varphi}_M^{(k)}\right\}$ and N' is the number of data points in each sub-dataset.

Remark 1: For small size data, the original dataset is resampled by removing one of the data points each time until all the data points have been picked out once (leaving one sample out), so that $N' = N - 1$ and $K = N$. Thus, the uncertainty brought by removing or adding a single data point can be reduced by finding a single common model for the K sub-datasets. The resampling process is used for the situations when the data size is small and the effect of a single data point can be significant for determining the final model structure and model parameters.

b).The OMAEs of model terms for K sub-datasets

To find a robust model structure that best fits all the K sub-datasets, an MAE matrix is calculated using the data from all the K sub-datasets. In the first step search, the MAE matrix is defined as:

$$\boldsymbol{\Psi}^{(1)} = \begin{bmatrix} \epsilon_1^{(1)} & \epsilon_2^{(1)} & \cdots & \epsilon_M^{(1)} \\ \epsilon_1^{(2)} & \epsilon_2^{(2)} & & \epsilon_M^{(2)} \\ \vdots & & \ddots & \vdots \\ \epsilon_1^{(K)} & \epsilon_2^{(K)} & \cdots & \epsilon_M^{(K)} \end{bmatrix} \tag{3.4}$$

where $\epsilon_m^{(k)}$ $(m = 1,\ 2, \ldots, M \ and \ k = 1,\ 2, \ldots, K)$ is the individual MAE value when the m-th candidate model term is used to approximate output $y^{(k)}$ in the k-th sub-dataset. It is calculated as:

$$\epsilon_m^{(k)} = \frac{1}{N'} \left\| \boldsymbol{y}^{(k)} - \alpha_m^{(k)} \boldsymbol{\varphi}_m^{(k)} \right\|_1 \tag{3.5}$$

where $\alpha_m^{(k)}$ is the parameter. Then, the OMAE associated with the m-th candidate model term which is used to represent all the K sub-datasets is defined as:

$$\bar{\epsilon}_m = \frac{1}{K}(\epsilon_m^{(1)} + \epsilon_m^{(2)} + \ \ldots + \ \epsilon_m^{(K)}) \tag{3.6}$$

Remark 2: In addition to the OMAE, there are several other metrics for measuring the overall predicted error of each model term, for example:

$$\phi_1(y, \hat{y}) = \frac{1}{N} \sum_{t=1}^{N} |y_t - \hat{y}_t| \tag{3.7}$$

$$\phi_2(y,\hat{y}) = \frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2 \tag{3.8}$$

$$\phi_3(y,\hat{y}) = \frac{\sum_{t=1}^{N}|y_t-\hat{y}_t|}{\sum_{t=1}^{N}|y_t|+\sum_{t=1}^{N}|\hat{y}_t|} \tag{3.9}$$

$$\phi_4(y,\hat{y}) = \frac{\sqrt{\frac{1}{N}\sum_{t=1}^{N}|y_t-\hat{y}_t|}}{\sqrt{\frac{1}{N}\sum_{t=1}^{N}|y_t|}+\sqrt{\frac{1}{N}\sum_{t=1}^{N}|\hat{y}_t|}} \tag{3.10}$$

where y and $\hat{y}$ are the observed and predicted system outputs and N is the number of data points. As will be illustrated later that $\phi_1(y,\hat{y})$ (MAE) is a better choice. It was argued in some studies that MAE is a better metric for model evaluation (Chai & Draxler, 2012).

c). OMAE-based term selection and parameter estimation

Define:

$$l_1 = arg \min_{1\le m\le M} \{\bar{\epsilon}_m\} \tag{3.11}$$

Then the 1st significant model terms can be selected as φ_{l_1}. After removal of the basis $\boldsymbol{\delta}_{l_1}^{(k)}$ from the k-th sub-dataset $(k = 1, 2, \dots, K)$, the dictionaries of all the K sub-datasets are then reduced and consists of $M-1$ model candidates. Similar to that in the conventional OFR algorithm, at step s $(s \ge 2)$, the K dictionaries consist of $M-s+1$ candidate model terms. The K bases are all transformed into a new group of K orthogonalized bases. The orthogonal transformation can be implemented using (2.14) for each single sub-dataset. The MAE matrix at step s can be calculated using the new group of K bases, and the MAE matrix is:

$$\boldsymbol{\Psi}^{(s)} = \begin{bmatrix} \epsilon_1^{(1)} & \epsilon_2^{(1)} & \cdots & \epsilon_{M-s+1}^{(1)} \\ \epsilon_1^{(2)} & \epsilon_2^{(2)} & & \epsilon_{M-s+1}^{(2)} \\ \vdots & & \ddots & \vdots \\ \epsilon_1^{(K)} & \epsilon_2^{(K)} & \cdots & \epsilon_{M-s+1}^{(K)} \end{bmatrix} \tag{3.12}$$

The OMAEs of all the candidate terms can then be calculated and the s-th robust model term can be selected to be φ_{l_s}, with:

$$l_s = arg \min_{1\le m\le M-s+1} \{\bar{\epsilon}_m\} \tag{3.13}$$

Repeating the recursive process, a number of model terms can be selected to form a linear-in-parameters robust model structure. Similar to OFR algorithm, the selection procedure can be terminated when specific conditions are met. Assume that a total of n model terms are selected, and for the k-th sub-dataset let the output $y^{(k)}$ be represented by the n selected model terms as:

$$\boldsymbol{y}^{(k)} = {\theta_{l_1}}^{(k)}\, {\boldsymbol{\varphi}_{l_1}}^{(k)} + {\theta_{l_2}}^{(k)}\, {\boldsymbol{\varphi}_{l_2}}^{(k)} + \ldots + {\theta_{l_n}}^{(k)}\, {\boldsymbol{\varphi}_{l_n}}^{(k)} \tag{3.14}$$

Following (Chen, Billings & Luo, 1989; Chen & Billings, 1989), the model parameters ${\theta_{l_1}}^{(k)}, {\theta_{l_2}}^{(k)}, \ldots, {\theta_{l_n}}^{(k)}$ can be calculated through an iterative procedure. According to the orthogonalization procedure (Chen, Billings & Luo, 1989; Chen & Billings, 1989), here we define K unity upper triangular matrices first:

$$\boldsymbol{A}^{(k)} = \begin{bmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & & a_{2n}^{(k)} \\ & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn}^{(k)} \end{bmatrix} \tag{3.15}$$

where $a_{11}^{(k)} = a_{22}^{(k)} = \cdots = a_{nn}^{(k)} = 1$.From the orthogonalization procedure, the elements of $\boldsymbol{A}^{(k)}$ can be calculated as:

$$a_{rj}^{(k)} = \frac{(\boldsymbol{q}_r^{(k)})^T {\boldsymbol{\delta}_{l_j}}^{(k)}}{(\boldsymbol{q}_r^{(k)})^T {\boldsymbol{q}_r}^{(k)}} \quad (r = 1, 2, \ldots, j-1 \; and \; j = 2, 3, \ldots n) \tag{3.16}$$

$$g_j^{(k)} = \frac{(\boldsymbol{y}^{(k)})^T {\boldsymbol{q}_j}^{(k)}}{(\boldsymbol{q}_j^{(k)})^T {\boldsymbol{q}_j}^{(k)}} \; (j = 1, 2, \ldots, n) \tag{3.17}$$

The estimates of K groups of parameter vector $\boldsymbol{\theta}^{(k)} = [{\theta_{l_1}}^{(k)}, {\theta_{l_2}}^{(k)}, \ldots, {\theta_{l_n}}^{(k)}]$ can then be calculated from the triangular equations $\boldsymbol{A}^{(k)}\boldsymbol{\theta}^{(k)} = \boldsymbol{g}^{(k)}$. The final model parameter estimation is chosen to be the average of the K parameter estimates, with:

$$\theta_{\,l_j} = \frac{1}{K}\sum_{i=1}^{K} \theta_{l_j}^{(i)} \; (j = 1, 2, \ldots, n) \tag{3.18}$$

Detailed derivation and explanation for the mechanism of the above calculations can be found in (Chen, Billings & Luo, 1989; Chen & Billings, 1989).

Remark 3: The proposed RMSS method can be summarized into several steps: 1). calculate the OMAE of each candidate model term; 2). select the model term according to the OMAEs; 3). remove the selected terms in the dictionary and transformed the rest

of bases to form new orthogonalized bases; 4) repeat the first 3 steps until a specific model selection criterion is met. 5). parameter estimation. The whole procedure can be described by a diagram as shown in Fig. 3.1.

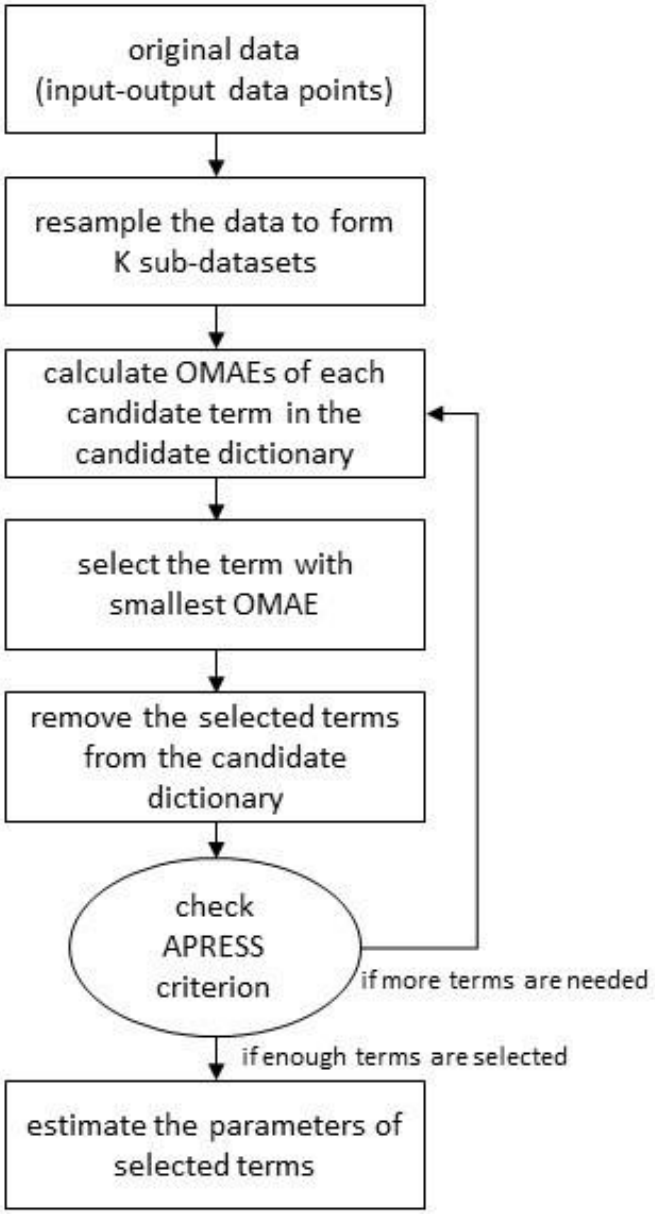

Figure 3.1 Robust model structure selection (RMSS) method

Remark 4: Note that different from traditional L2-norm based algorithms, e.g. the orthogonal projection pursuit (OPP) algorithm (Wei & Billings, 2008b) that can be proven to converge, the proof of the convergence of the proposed RMSS method is not straightforward. In this study, the focus is on choosing a set of most powerful model

terms from a given pool consisting of a large number of candidate model terms, through an iterative manner, one term at each search step, until a model with an appropriate model terms that gives satisfactory fit to the data is obtained. Instead of strictly prove the convergence of the proposed method, we demonstrate the overall performance of the new method through numerical case studies which are presented in the next section.

3.4 Simulation

Two simulation examples are presented to test the efficiency of the RMSS method and to show under which conditions the proposed method can improve the model performance. The first example aims to test if the proposed method can pick out the correct model terms when data are noise free. The second example investigates the performance of the proposed method for modelling problems with different levels of uncertainty (noise). Finally, case studies are carried out to demonstrate the power of the new method solving a real-world problem. For the convenience of comparative analysis, the model identified by OFR method will be referred as 'regular model' and the model identified by RMSS method will be referred as 'robust model'.

3.4.1 Example 1- noise free data modelling

It is known that most existing model structure selection methods can provide sufficiently reliable model, when data are clean (i.e. not corrupted with noise). In the following it will show that both the RMSS method and classic OFR method can generate perfect model structure from noise free data. Consider a nonlinear system:

$$\mathrm{y(t)} = 0.5\mathrm{y(t-1)} + 0.8u(t-2) + \mathrm{u}^2\mathrm{(t-1)} - 0.05y^2\mathrm{(t-2)} + 0.5 \tag{3.19}$$

where the input $\mathrm{u(t)}$ was assumed to be uniformly distributed on $[-1, 1]$. A total number of 100 input-output data points were generated. The first 70 points were used for model estimation and the remaining 30 points were used for performance test. From the results of some pre-experiments, the following candidate variable vector was used for model construction:

$$\boldsymbol{\vartheta}(t) = [\mathrm{y(t-1)}, \mathrm{y(t-2)}, \mathrm{u(t-1)}, \mathrm{u(t-2)}]^{T} \tag{3.20}$$

Table 3.3 Selected terms by classic OFR method

No.	Term	ERR(100%)	Parameter
1	*y(t-1)*	78.7770	0.5000
2	*u(t-2)*	10.6233	0.8000
3	*u(t-1) ×u(t-1)*	8.8996	1.0000
4	*constant*	1.3601	0.5000
5	*y(t-2) ×y(t-2)*	0.3401	-0.0500

Table 3.4 Selected terms by RMSS method

No.	Term	OMAE	Parameter
1	*y(t-1)*	0.5639	0.5000
2	*u(t-2)*	0.3831	0.8000
3	*u(t-1) ×u(t-1)*	0.1610	1.0000
4	*constant*	0.0652	0.5000
5	*y(t-2) ×y(t-2)*	0.0000	-0.0500

The initial full model was chosen to be a polynomial form with nonlinear degree of $l = 3$. Firstly, the OFR method was applied to find the significant model terms according to the ERR ranking. The APRESS values suggest that a model of 5 terms can be a good choice. Not surprisingly, all the model terms are correctly selected and the parameters are estimated correctly. The selected terms and the associated ERR values are shown in Table 3.3. The RMSS method was also applied to the same train data, to select significant terms according to their OMAEs relating to a total number of 70 sub-datasets generated through the resampling process. As a result, the RMSS method selected exactly the same model terms as the OFR method. The associated OMAEs are shown in Table 3.4.

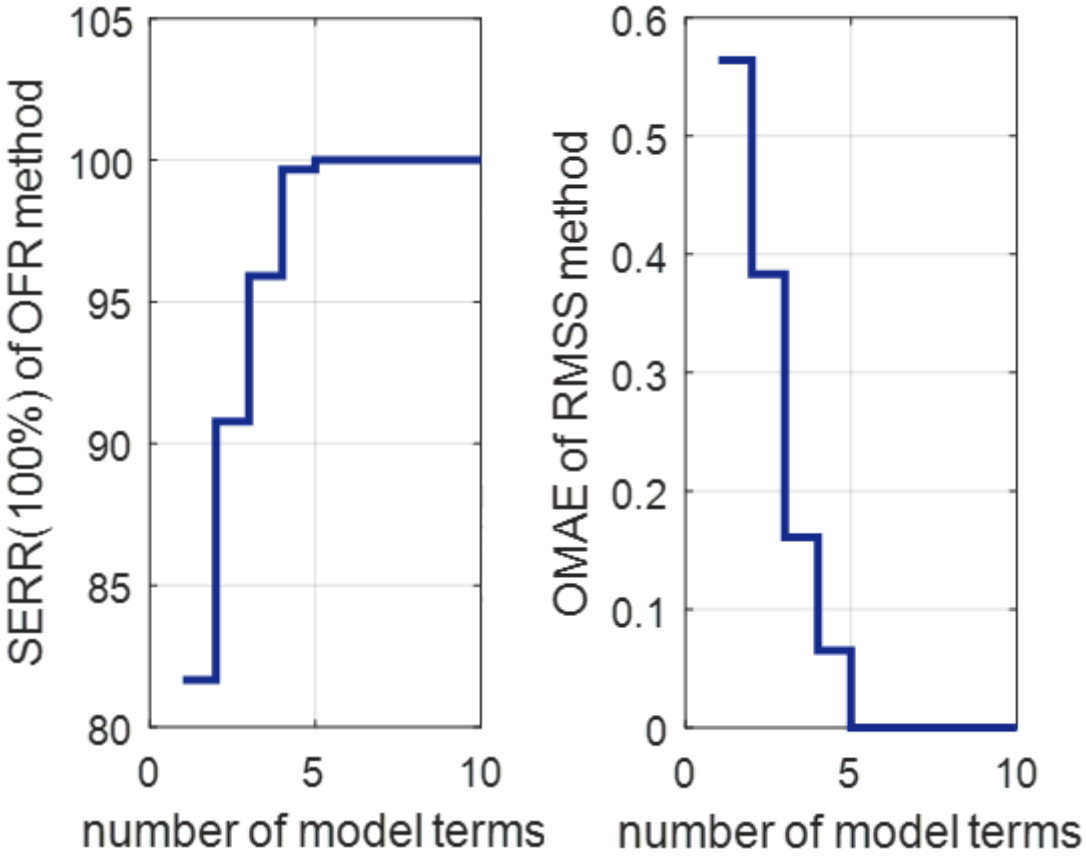

Figure 3.2 SERR and OMAE versus the number of iterations of term selection

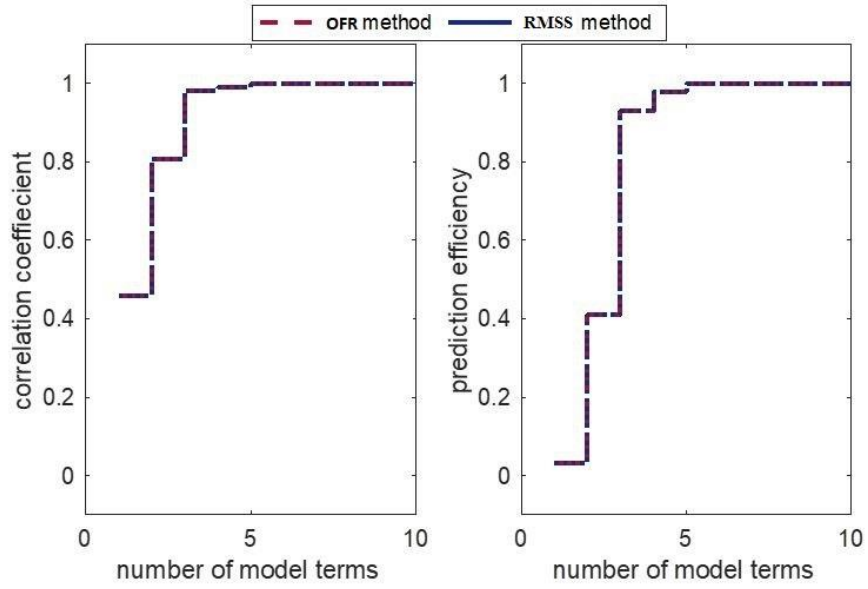

Figure 3.3 Statistics prediction performance of regular model and robust model versus the model complexity

Note that the OFR and RMSS methods employ two different indicators (i.e., the ERR index and OMAEs to measure the contribution of each model term to explaining the variance of response variable. During the process of OFR, the SERR (sum of ERR values) is increasing to the maximum value of 100%, which indicates that 100% of the variance of response variable can be explained by the selected terms. For the RMSS method, the OMAE is decreasing to 0, which means that there is no error in the identified model. The variation of SERR and OMAE of the OFR and RMSS are displayed in Fig. 3.2. It can be easily seen that the model with 5 terms is perfect and can describe 100% of the variance of the response variable. The variation of the correlation coefficient and prediction efficiency, with the inclusion of model terms, one by one, is shown in Fig. 3.3.

3.4.2 Example 2- data with additive white noise

Now consider a nonlinear system:

$$y(t) = -u(t-1)\sqrt{|y(t-1)|} + 0.4u^2(t-1) + 0.8u(t-2)u(t-1) + \xi(t) \quad (3.21)$$

where the input $u(t)$ was assumed to be uniformly distributed on $[-1, 1]$ and $\xi(t)$ is a white noise with zero mean and finite variation. Note that there is no strict definition of 'small' size data. Normally if the number of data points are around 100 or less, the data is a small size data. Also, the SNR is important to determine if there is strong uncertainty in the data. With five different levels of signal to noise ratio, namely, noise-free and SNR = 50, 15, 10, 0 dB, respectively, the system was simulated five times. For each SNR case, a total number of 100 input-output data points were generated. The first 70 points were used for model estimation and the remaining 30 points were used for performance test. From the results of some pre-experiments, the initial full model was chosen to be a polynomial form with maximum time lags of $n_y = n_u = 2$ and nonlinear degree of $l = 3$. Note that the model term $\sqrt{|y(t-1)|}$ was not included in the specific library of candidate model terms. As a consequence, it is impossible to identify a 'true' model structure that perfectly represents every single component of the system. However, it is possible to use both the OFR and the RMSS method to find model that can well represent the simulated data. In what follows, it presents analysis and discussions on whether the RMSS methods can find satisfactory models with good predictive performance, under different level of noise.

Table 3.5 Selected terms by OFR and RMSS method

SNR	No.	OFR method	RMSS method
noise	1	*u(t-2) ×u(t-2)*	*u(t-2) ×u(t-2)*
	2	*u(t-1)*	*u(t-1)*
free	3	*u(t-1) ×u(t-1) ×y(t-2)*	*u(t-1) ×u(t-1) ×y(t-2)*
	4	*u(t-1) ×u(t-2) ×u(t-2)*	*u(t-1) ×u(t-2) ×u(t-2)*
	5	*u(t-1) ×u(t-2)*	*u(t-1) ×u(t-2)*
	6	*u(t-1) ×u(t-2) ×y(t-1)*	*u(t-1) ×u(t-2) ×y(t-1)*
	7	*u(t-2) ×y(t-1)*	*y(t-1) ×y(t-2)*
50db	1	*u(t-2) ×u(t-2)*	*u(t-2) ×u(t-2)*
	2	*u(t-1)*	*u(t-1)*
	3	*u(t-1) ×u(t-1) ×y(t-2)*	*u(t-1) ×u(t-1) ×y(t-2)*
	4	*u(t-1) ×u(t-2) ×u(t-2)*	*u(t-1) ×u(t-2) ×u(t-2)*
	5	*u(t-1) ×u(t-2)*	*u(t-1) ×1(t-2)*
	6	*u(t-1) ×u(t-2) ×y(t-1)*	*u(t-1) ×u(t-2) ×y(t-1)*
	7	*u(t-2) ×y(t-1)*	*y(t-1)*
15db	1	*u(t-2) ×u(t-2)*	*u(t-2) ×u(t-2)*
	2	*u(t-1)*	*u(t-1)*
	3	*u(t-1) ×u(t-1) ×y(t-2)*	*u(t-1) ×u(t-1) ×y(t-2)*
	4	*u(t-1) ×u(t-2) ×u(t-2)*	*u(t-1) ×u(t-2) ×u(t-2)*
	5	*u(t-1) ×u(t-2)*	*u(t-1) ×u(t-2)*
	6	*u(t-1) ×u(t-2) ×y(t-1)*	*u(t-1) ×u(t-2) ×y(t-1)*
	7	*u(t-1) ×u(t-2) ×y(t-2)*	*u(t-1) ×u(t-2) ×y(t-2)*
	8	*u(t-2) ×y(t-1)*	*u(t-1) ×u(t-1)*
10db	1	*u(t-2) ×u(t-2)*	*u(t-2) ×u(t-2)*
	2	*u(t-1)*	*u(t-1)*
	3	*u(t-1) ×u(t-1) ×y(t-2)*	*u(t-1) ×u(t-1) ×y(t-2)*
	4	*u(t-1) ×u(t-2)*	*u(t-1) ×u(t-2)*
	5	*u(t-1) ×u(t-2) ×y(t-1)*	*u(t-1) ×u(t-2) ×y(t-1)*
	6	*y(t-1) ×y(t-2)*	*u(t-2) ×y(t-2)*
	7	*y(t-1) ×y(t-2) ×y(t-2)*	*u(t-1) ×u(t-1) ×u(t-2)*
	8	*u(t-1) ×u(t-2) ×u(t-2)*	*y(t-2) ×y(t-2) ×y(t-2)*
	9	*u(t-1) ×u(t-1) ×u(t-2)*	*u(t-1) ×u(t-2) ×y(t-2)*
0db	1	*u(t-2) ×u(t-2)*	*u(t-2) ×u(t-2)*
	2	*u(t-1)*	*u(t-1)*
	3	*u(t-2) ×u(t-2) ×y(t-2)*	*u(t-2) ×u(t-2) ×y(t-2)*
	4	*u(t-1) ×u(t-1) ×y(t-1)*	*y(t-1) ×y(t-1)*
	5	*u(t-1) ×u(t-2)*	*u(t-2) ×y(t-2)*
	6	*u(t-1) ×u(t-1)*	*u(t-2) ×u(t-2) ×u(t-2)*
	7	*y(t-1) ×y(t-1)*	*u(t-2) ×y(t-2) ×y(t-2)*
	8	*y(t-1) ×y(t-2)*	*y(t-1) ×y(t-1) ×y(t-1)*
	9	*y(t-1) ×y(t-1) ×y(t-1)*	*u(t-1) ×u(t-1) ×y(t-2)*

Both the OFR and RMSS methods were applied to the simulated data with different levels of noises (noise-free, SNR = 50, 15, 10, 0 dB). The model complexity was determined by the APRESS metric (Billings & Wei, 2008). The selected model terms by

the two methods are shown in Tables 3.5. It can be observed that for most cases, the two methods select the same model terms for the first few steps. This is reasonable because these terms are the most significant terms and make major contribution to explaining the variance of system output and leaving one sample out (this scheme is used in RMSS method but not in OFR) does not affect the order of the selected terms. However, the two methods start to select different model terms when the SNR is decreased. For example, for 15db, the 8^{th} terms are different; for 10db, the 6^{th} terms become different. These model terms give smaller contributions to explaining the variance in output signal, and a small change of single sample might affect result of selection of these terms. In other words, the less significant model terms are more sensitive to the effect of noise.

Table 3.6 Performance statistics of the regular model, robust model, lasso algorithm and neutral networks under different noises

SNR Level	performance statistic	regular NARX model	robust NARX model	lasso algorithm	neural network*
noise-free	correlation coefficient	0.9365	0.9497	0.9335	0.9070
	predicted efficiency	0.8534	0.8754	0.8573	/
50 dB	correlation coefficient	0.9374	0.9463	0.9343	0.9273
	predicted efficiency	0.8560	0.8721	0.8587	/
15 dB	correlation coefficient	0.9117	0.9208	0.9114	0.8292
	predicted efficiency	0.7899	0.8135	0.7808	/
10 dB	correlation coefficient	0.8339	0.8758	0.8550	0.7712
	predicted efficiency	0.6219	0.7366	0.7025	/
0 dB	correlation coefficient	0.3780	0.4311	0.4931	0.3740
	predicted efficiency	0.0426	0.1846	0.2221	/

* The training algorithm is Levenberg-Marquardt. The algorithm was run for 10 times and the averaged correlation coefficient is recorded.

As mentioned earlier, the classic OFR method uses ERR index as measure to select model terms; the measure is defined as how much (in percentage) of the variance in the response signal can be explained by a newly included model term. The RMSS method uses OMAE instead, which is a measure of the averaged prediction error in relation to a great number (say K) of models estimated from K sub-datasets generated from the original data through a resampling process. Therefore, the resulting robust model should provide better overall predictive performances than the regular model. The performance statistics of the regular and robust models are given in Table 3.6. The results show that with the decrease in SNR values, the performance of the models identified by both the OFR method and the robust method decreases, due to the increase of uncertainty. It should be stressed that even for the noise-free case, both of the two methods fail to detect the true model structure, because the model component $u(t-1)\sqrt{|y(t-1)|}$ is actually not in the pre-defined library of candidate model terms.

Comparing the performance statistics of the regular and robust NARX models given, the robust models outperform the regular models in all the cases. In addition, the improvement of the robust models is significant when SNR is quite low say at 10 dB and 0 dB. Fig. 3.4-3.6 show the model prediction of the regular and robust models for the three cases: noise-free and SNR=15dB and 0dB, respectively. As can be seen from the figures, the differences of predicted and observed output become more significant with the increase of noise level. It can be noted in Fig. 3.6 that there are some extremely large values in predicted output from the regular model, and the robust model is more conservative in prediction, where the amplitudes of the predicted values are in general smaller than that of the classical model but closer to the true values. The prediction of robust method has smoother curve than that of the regular method.

We also compared the performances of proposed RMSS method with other two nonlinear identification methods: lasso and neural networks. Lasso aims to the degree of the freedom of a given model structure by shrinking the coefficients of unnecessary model terms to zero. The lasso method can be easily adapted to many application scenarios where the desired response signal is assumed to be of a sparse representation of a set of independent signals (predictors). However, lasso could fail to produce stable subset selection results when the predictors are highly correlated. The performances of the two methods are evaluated based on the models with the same number of model terms. From the results in Table 3.6, the robust NARX model outperforms the lasso method in most

of the cases (noise-free, SNR=50, 15, 10 dB). This is because the orthogonal forward regression (OFR) algorithm used in RMSS can effectively solve sever correlation and ill-conditioning problems (Wei & Billings, 2008a; Wei, Billings & Liu, 2004).

The applied neural network has one input layer, one hidden layer and one output layer. The number of neurons is 10 and the activation function is sigmoid function. Note that the neural network was evaluated via correlation. So only correlation statistics is calculated. The estimation algorithm was run for 10 times to obtain robust results, as the training of neural network uses a stochastic process. Regarding all the five cases, the performances of the neural network models are lower than those of the other two methods. This might be because that the size of the data is very small, and that the power of neural networks is cannot be fully exploited for this small size data modelling problem. More importantly, the proposed RMSS method has the following superiorities: i). the procedure is easy to implement and not time-consuming; ii). the identified model clearly indicates the information of the most important model terms; iii). the identified model provides a transparent and parsimonious linear-in-the-parameters representation, which can be easily generalized to new data. It is worth mentioning that in this example, all the robust models were built using only 70 data points, which is quite small. This means the proposed RMSS method may promise an effective data driven modelling approach for nonlinear systems, especially for small size data with strong uncertainty. Overall, these results show the clear advantage of the proposed RMSS method in nonlinear model identification.

Table 3.7 Comparison of the performances of robust models identified based on different measures

Measures	ϕ_1	ϕ_2	ϕ_3	ϕ_4
Correlation Coefficient	0.9208	0.9202	0.8667	0.8667
Predicted Efficiency	0.8135	0.8059	0.7018	0.7018

In addition, for the case of SNR=15dB, three extra robust models are obtained based on the other three different measures defined in (3.8)-(3.10), respectively. The performance statistics of all the four models are given in Table 3.7 and it turns out that the robust model selected by OMAE over performs the other three models.

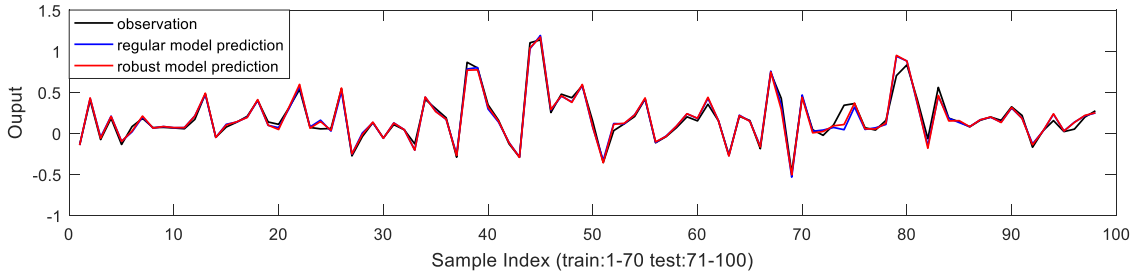

Figure 3.4 One-step-ahead (OSA) predictions of robust model and regular model (noise free)

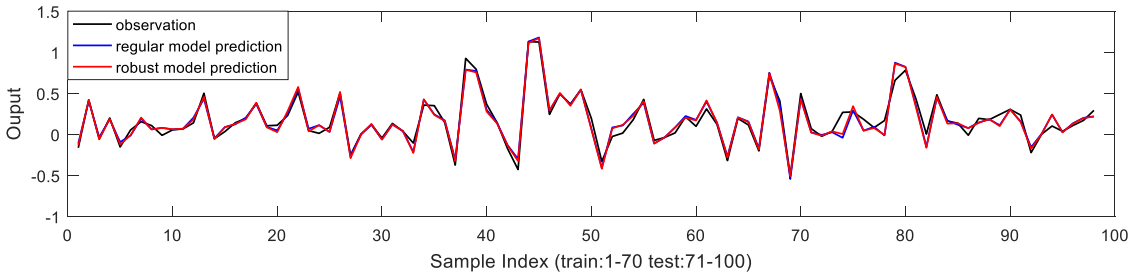

Figure 3.5 One-step-ahead (OSA) predictions of robust model and regular model (SNR is 15dB)

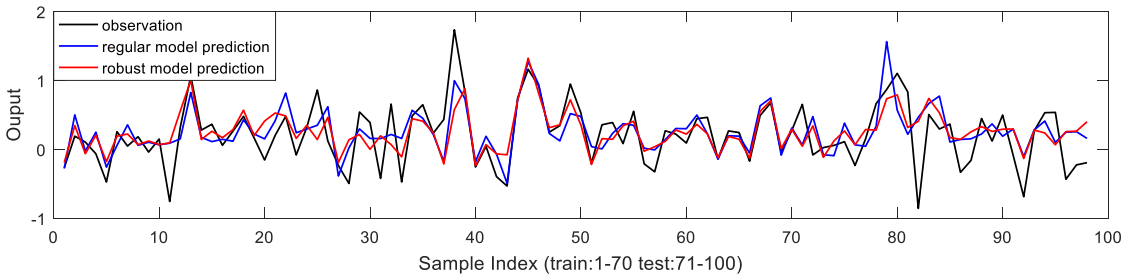

Figure 3.6 One-step-ahead (OSA) predictions of robust model and regular model (SNR is 10dB)

3.5 Real Data Case Studies

This section presents two real data case studies to illustrate the RMSS method. The first case study is modelling and forecasting of Kp index with small size data. the second case study is modelling and forecasting of cortical response with multi-datasets data.

Table 3.8 Kp index and solar wind variables

Name	Model variable	Description
Kp	y	Kp index

V	u_1	solar wind speed/velocity (flow speed) [km/s]
Bs	u_2	interplanetary magnetic field factor [nT]
p	u_3	solar wind pressure (flow pressure) [nPa]
n	u_4	solar wind density (proton density) [n/cc]
VBs	u_5	$V \times Bs/1000$
$\sqrt{p}$	u_6	square root of p

3.5.1 Example 1- Kp index Forecasting

Magnetic disturbance can affect many equipment and systems on or nearby earth, for example, navigation systems, communication systems, satellites, and power grid, etc. They can be paralyzed and unreliable during these severe magnetic situations. In order to understand and forecast the geomagnetic activity, the Kp (planetarische Kennziffer) index was first introduced by Bartels in 1949 (Bartels, 1949). The value of Kp index ranges from 0 (very quiet) to 9 (very disturbed) in 28 discrete steps, resulting values of 0, 0+, 1-, 1, 1+,2-, 2, 2+, …, 9 (Wing, et al., 2005). The Kp index has been recorded and updated since last century and become an important dataset to study space weather. The correlation between Kp index and solar wind parameters has been discovered by many researches. Normally, the solar wind variables are treated as the model inputs and Kp index is treated as the model output. A full description of the solar wind variables and derived variables is summarized in Table 3.8.

Table 3.9 Selected terms by OFR method for Kp model

No	Term	ERR(100%)	Parameter
1	$u_6(t-1)$	79.6551	7.7057e+00
2	$u_2(t-1) \times u_2(t-1)$	5.3507	4.0605e+02
3	$u_1(t-1)$	2.5907	2.3494e+00
4	$u_2(t-2)$	0.3058	7.4787e+00

Table 3.10 Selected terms by RMSS method for Kp model

No	Term	OMAEs	Parameter
1	$u_6(t-1)$	0.85592	6.4929e+00
2	$u_2(t-1)$	0.74081	5.0490e+01
3	$u_1(t-1) \times u_6(t-2)$	0.68803	2.0516e+01
4	$u_5(t-1)$	0.65544	-8.2486e+04

The Kp index was sampled every 3 hours and the solar wind variables were sampled every 1 hour. It should be noted that this study aims to build the models using robust method to predict Kp index 3 hours ahead. Therefore, the unit of time lags of both input and output is 3 hours. For example, $y(t-2)$ is the Kp index recorded 6 hours before $y(t)$ and $u_4(t-1)$ is the solar wind speed recorded 3 hours before $u_4(t)$. A total number of 150 input-output data points of the 2011 are selected for the case study. The maximum time lags are chosen as $n_u = 2$ and the nonlinear degree is 2. The first 100 samples are used for training and the remaining 50 samples are used for testing. The model is selected using only input lag variables, without using autoregressive variables. The first 4 model terms selected by OFR method and RMSS method are shown in the following Table 3.9 and Table 3.10.

The performance statistics of the two models are given in Table 3.11 and Fig. 3.7 presents comparisons between the model outputs and the associated measurements. Clearly, the overall performance of the robust model is better than the regular model and that produced by the lasso algorithm. The performance of the neural network model is slightly better than the robust NARX model. However, it is worth noting that the robust NARX model uses a much less number of model terms to provide a transparent and parsimonious representation, which is easy to interpret and use. Although the correlation between the measurements and the corresponding prediction of the neural network model is higher, the model itself is very complicated and difficult to write down. In contrast, the RMSS method and NARX model provide a transparent and parsimonious representation, which is simple where all the interactive relation among variables is clear. In general, the RMSS method achieves a good trade-off between model complexity and model

performance. Overall, the robust NARX model can be a good choice for Kp index predictions.

Table 3.11 Performance statistics of the regular model and robust model on Kp forecast

Performance Statistics	regular model	robust model	lasso	neural networks*
Correlation Coefficient	0.7132	0.8056	0.6109	0.8368
Predicted Efficiency	0.2927	0.6304	0.3202	/
NRMSE	0.2449	0.1750	0.3506	/

* The training algorithm is Levenberg-Marquardt. The algorithm was run for 10 times and the averaged correlation coefficient is recorded.

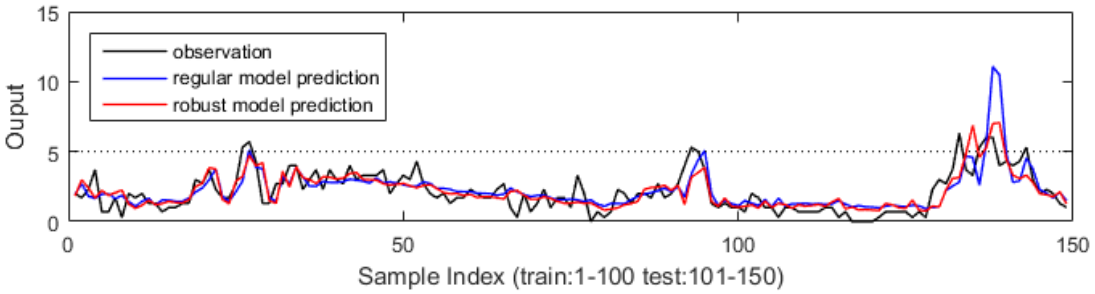

Figure 3.7 One-step-ahead (OSA) predictions of robust model and regular model for Kp index

3.5.2 Example 2- Modelling of Cortical Response

This section presents an example where the RMSS method is applied to an EEG modelling problem to identify a common model structure for 10 the cortical response of 10 different participants.

The data used in this study were recorded from 10 different participants, and each participant took part in 7 experiments with different input signals (mechanical wrist perturbation) (Vlaar, et al., 2017a; Vlaar, et al., 2017b). In total, there are 70 datasets, and each contains 256 sampled input-output data points. Thus, there are 70 datasets in total. Each dataset contains 210 periods (1 s per period) of signals. We average the data over periods to improve its signal-to-noise ratio, leaving 1 s (256 sampled input-output data points) per dataset as shown in Figure 3.8. The first six experiments of each participant were used for model identification and the remaining one was used for model evaluation.

Note that there is a large difference between the amplitudes of the input (i.e., the mechanical perturbation signal) and output signals (i.e., the IC component of EEG signal) in the original experimental datasets. In order to avoid or alleviate ill-conditioning in the relevant procedures (e.g. calculation of designed matrices and associated model parameters), the input signals are scaled up as $u = u' \times 100$, where u is the amplified input signal and u' is the original input signal, so that the amplitude of the input signals used for model identification is at a similar scale as that of output signals.

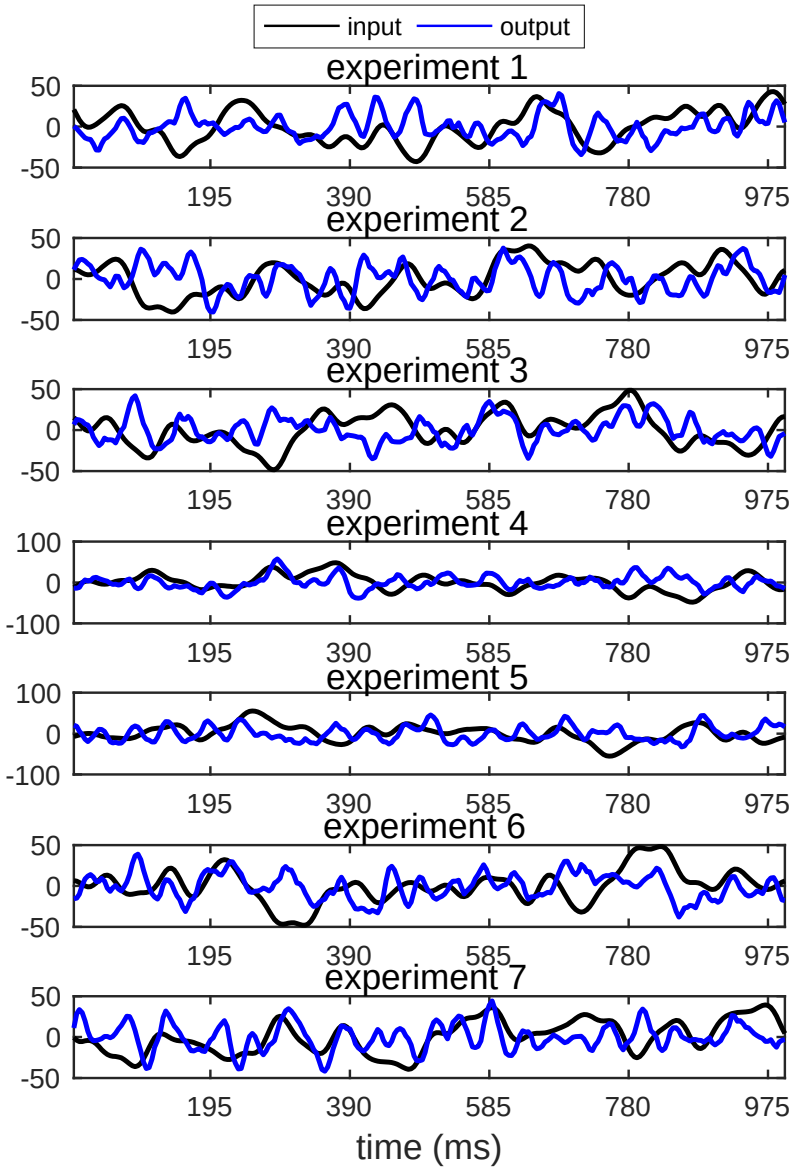

Figure 3.8 Input-output data pairs of the seven realizations of one representative participant (the input signals were amplified 100 times to make the input and the output in the same scale).

Subject-specific Structure Models for Cortical Responses to Mechanical Wrist Perturbations

Subject-specific NARX and Volterra (a nonlinear model without autoregressive terms) models were firstly identified for each participant using the OFR algorithm (Chen, Billings & Luo, 1989). The OFR algorithm uses an error reduction ratio (ERR) index (Chen, Billings & Luo, 1989) to measure the significance of each candidate of model term, and then selects significant model terms based on a stepwise strategy. In each search step, it calculates the associated ERR value for each candidate to create a ranking order. Based on this ranking order, the OFR selects the most significant term for building a model structure.

In this study, the maximum lag of input was set as $n_u = 20$ for both the NARX and Volterra models, and the maximum lags of output was set as $n_y = 5$ for the NARX model. Since previous studies have demonstrated the dominance of second order nonlinearity in this dataset (Vlaar, et al., 2017b), the nonlinear degree is chosen to be 2 for both models.

(i) one-step-ahead model predicted output:

$$\hat{y}(t) = f\left(y(t-1), \dots, y\left(t-n_y\right), u(t-1), \dots, u(t-n_u)\right) \tag{3.22}$$

(ii) two-step-ahead model predicted output:

$$\hat{y}(t+1) = f(\hat{y}(t), y(t-1), \dots, y\left(t-n_y+1\right), u(t), u(t-1), \dots, u(t-n_u+1)) \tag{3.23}$$

(iii) three-step-ahead model predicted output:

$$\hat{y}(t+2) = f(\hat{y}(t+1), \hat{y}(t), y(t-1), \dots, y\left(t-n_y+2\right), u(t+1), u(t) \dots, u(t-n_u+2)) \tag{3.24}$$

where $\hat{y}(t)$ represents the model predicted output, while y(t) is the corresponding measured output at the time instant t. The same evaluation was also performed on the output estimated by the second-order Volterra method (Vlaar, et al., 2017b).

The mean correlation coefficient, VAF and NRMSE of OSA predictions generated by subject-specific NARX models are 0.9710, 94.27% and 0.0458, respectively. The mean

Table 3.12 Ten NARX models with common model structure (Pa: Estimated Parameter; Ts: T Statistics With 95% Confidence).

Common Model Terms	Value	10 Sets of Model Parameters with Standard Errors for 10 Participants									
		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
y(t-1)	PA	1.9136	2.1283	1.9620	1.9062	1.7919	1.9537	1.8207	1.8540	1.7685	2.0378
	TS	39.1586	44.0510	38.6501	37.4519	35.8379	40.7721	35.4509	37.3112	34.4946	40.7741
y(t-2)	PA	-1.6389	-2.3801	-2.0750	-1.9747	-1.6039	-1.8897	-1.6366	-1.9239	-1.5882	-1.9979
	TS	16.1492	22.3020	19.4596	19.2289	16.4262	18.9754	16.4895	19.5723	15.8643	18.2081
y(t-3)	PA	1.1496	1.8781	1.5953	1.5802	1.2751	1.4604	1.2766	1.4740	1.1811	1.4295
	TS	9.6775	14.3512	12.5753	13.2003	11.5258	12.4170	11.3040	12.5733	10.3050	10.8256
y(t-4)	PA	-0.8802	-1.0940	-0.8163	-0.8601	-0.8167	-1.0204	-0.8348	-0.8606	-0.6631	-0.8276
	TS	8.6350	10.3107	7.7028	8.4583	8.3737	10.3214	8.4195	8.8606	6.6553	7.5462
y(t-5)	PA	0.3605	0.3753	0.2594	0.2450	0.2953	0.4125	0.2350	0.2961	0.2067	0.2998
	TS	7.3846	7.9710	5.1839	4.9668	5.9365	8.9470	4.7321	6.3907	4.0945	6.0141
u(t-7)u(t-14)	PA	-0.1579	-0.9576	2.7795	-4.4608	-0.5413	1.9474	0.4822	-4.7443	-0.8010	0.0526
	TS	0.2824	1.2252	1.9677	2.9549	0.6473	1.7759	0.3595	3.1248	0.5467	0.1008
u(t-1)u(t-1)	PA	0.1563	-0.1202	1.3766	-1.2776	-0.1300	0.9525	-1.1649	0.2051	-0.3703	0.1146
	TS	0.7821	0.4234	2.6169	2.2957	0.4287	2.4534	2.4056	0.3871	0.6949	0.5931
u(t-1)u(t-18)	PA	-0.0040	-0.1286	-0.0059	0.0936	0.2731	-0.1931	0.0772	0.3802	-0.0445	0.0494
	TS	0.0598	1.3528	0.0342	0.4992	2.6692	1.5074	0.4736	1.9592	0.2442	0.7698
u(t-20)u(t-20)	PA	-0.0164	-0.0546	0.0148	-0.2992	0.0527	-0.0058	0.2762	-0.5263	-0.2104	0.0240
	TS	0.2980	0.7076	0.1024	2.0001	0.6435	0.0532	2.1118	3.6043	1.4651	0.4670
y(t-1)y(t-1)	PA	-0.0004	-0.0004	0.0001	0.0002	0.0001	-0.0000	0.0003	-0.0006	0.0003	0.0008
	TS	0.6481	1.1333	0.5856	0.6525	0.2848	0.1825	1.1217	1.9511	0.7684	1.4559
u(t-15)u(t-18)	PA	-0.0842	0.1514	-0.4369	0.4028	0.0370	-0.4325	-0.3078	0.8469	0.1073	-0.0686
	TS	1.1332	1.4362	2.2765	1.9458	0.3318	3.0430	1.7342	4.0962	0.5481	0.9779
u(t-6)u(t-12)	PA	0.2432	3.7802	-7.9935	15.7187	2.8887	-4.1672	0.2478	8.9275	2.3809	0.3125
	TS	0.1498	1.6439	1.9138	3.5803	1.1758	1.3221	0.0635	2.0608	0.5526	0.2016
u(t-1)u(t-8)	PA	1.7644	9.3520	-11.8257	16.3080	-3.4110	3.2285	-11.4160	18.8067	9.4023	-1.7272
	TS	1.0645	3.8427	2.7152	3.5468	1.3480	0.9892	2.6911	4.1768	2.1317	1.0826
u(t-4)u(t-10)	PA	-1.0187	-23.0848	40.0287	-77.1972	-7.2711	10.2672	0.5519	-36.9031	-23.7419	-1.1842
	TS	0.1521	2.4104	2.2962	4.2000	0.7122	0.7988	0.0344	2.0595	1.3247	0.1825
u(t-2)u(t-8)	PA	-4.2048	-40.0234	71.4439	-104.4858	3.9345	7.8501	10.7568	-72.9761	-50.9539	3.2769
	TS	0.4693	3.1010	3.0365	4.1776	0.2882	0.4569	0.4999	3.0394	2.1237	0.3766
u(t-4)u(t-5)	PA	0.6918	1.8722	-0.4837	1.2057	-1.1374	2.3516	-5.1799	4.5207	1.1439	-0.5797
	TS	1.5510	2.8550	0.4214	0.9969	1.6396	2.5092	4.1501	3.6059	0.9603	1.3500
u(t-3)u(t-9)	PA	2.6126	45.6948	-81.8456	138.8617	5.2147	-16.2129	-1.8655	75.9933	54.9498	-0.0197
	TS	0.2212	2.7010	2.6538	4.2476	0.2897	0.7171	0.0661	2.4033	1.7373	0.0017
constant	PA	-0.1711	-0.8208	1.0130	-0.2438	-0.0188	0.0589	1.7409	-2.2673	-0.2763	-0.0014
	TS	0.4167	1.3881	0.9611	0.2204	0.0307	0.0745	1.6422	1.9482	0.2595	0.0036
u(t-9)u(t-20)	PA	0.1232	0.1400	-0.0637	0.4682	-0.3325	0.2862	-0.3279	0.4030	0.3258	-0.0239
	TS	1.3271	1.0731	0.2666	1.8478	2.4127	1.5864	1.5145	1.6409	1.3480	0.2735
u(t-1)u(t-6)	PA	-0.1401	3.3975	-13.2583	14.6835	0.5296	-6.0680	7.8915	5.2528	7.6663	-0.2059
	TS	0.0761	1.2961	2.7581	2.8655	0.1893	1.7153	1.7818	1.0726	1.5571	0.1152

correlation coefficient, VAF, and NRMSE of MSA predictions generated by subject-specific NARX method are 0.7431, 54.84% and 0.1281, respectively. The mean correlation coefficient, VAF, and NRMSE of subject-specific Volterra method are 0.6625, 42.84% and 0.1450, respectively. The results from the NARX model is significantly better than those from the Volterra model. Thus, the second-order NARX model provides a simpler model representation with better prediction performances than the second-order truncated Volterra model in all tested datasets.

Common Structure Models for Cortical Responses to Mechanical Wrist Perturbations

A common model structure, with 10 different model parameters, was built to characterize the cortical response behavior of the 10 participants. The first 6 datasets of each participant (recorded from the first 6 experiments) were used for model identification, and the remaining one dataset is used for model evaluation. In total, there were 60 datasets for model identification and 10 datasets for model evaluation. The time lags of input and output were still chosen to be $n_u = 20$ and $n_y = 5$ and the nonlinear degree is chosen to be 2. The common model structure was identified using the proposed CMSD method based on all 60 datasets. According to the results of model length criterion, the optimal number of model terms was determined as 20. The common model structure includes the most important 20 model terms (regressors) selected from a great number of candidates (i.e. 351 candidates). Although the same model structure was obtained for all participants, subject-specific parameters were estimated to indicate the individual differences (see Table 3.12).

Then, the model parameters of each individual NARX model were estimated from the associated 6 datasets of each participant. As shown in Table 3.12, the common model structure comprises of 20 terms selected by the CMSD method, and there are 10 sets of model parameters for different participants. For example, the model for the first participant is $y(t) = 1.9136y(t-1) - 1.6389y(t-2) + \cdots - 0.1401u(t-1)u(t-6)$; while the model for last participant is$y(t) = 2.0378y(t-1) - 1.9979y(t-2) + \cdots - 0.2059u(t-1)u(t-6)$. All participants have the same model structure but different model parameters.

Table 3.13 OMAE values and error reductions (ER) of the selected 20 common model terms (ER= oMAE value of previous term - oMAE value of current term)

Model Terms	oMAE	ER	Model Terms	oMAE	ER
y(t-1)	9.45	-	u(t-15)u(t-18)	5.50	0.0291
y(t-2)	7.16	2.3419	u(t-6)u(t-12)	5.46	0.0375
y(t-3)	6.37	0.7899	u(t-1)u(t-8)	5.43	0.0366
y(t-4)	6.02	0.3456	u(t-4)u(t-10)	5.38	0.0411
y(t-5)	5.70	0.3291	u(t-2)u(t-8)	5.35	0.0323
u(t-7)u(t-14)	5.65	0.0412	u(t-4)u(t-5)	5.30	0.0423
u(t-1)u(t-1)	5.62	0.0311	u(t-3)u(t-9)	5.26	0.0455

u(t-1)u(t-18)	5.59	0.0325	constant	5.23	0.0364
u(t-20u(t-20)	5.56	0.0312	u(t-9)u(t-20)	5.20	0.0317
y(t-1)y(t-1)	5.53	0.0285	u(t-1)u(t-6)	5.17	0.0291

The significance of each model terms can be measured by the proposed oMAE. The oMAE values of all selected model terms in the common structure are presented in Table 3.13. We can see that the inclusion of each model term progressively reduced the overall prediction error, step by step. According to our results, the first five autoregressive terms are important in reducing the prediction error. This result indicates that it is necessary to use the NARMAX method for modeling, since the Volterra model does not have autoregressive terms. Additionally, the t-statistics (with 95% confidence) of each selected model terms are presented in Table 3.12. The t-statistics indicate that the selected model terms are significant for most of the participants. As shown in Table 3.13, the first 5 autoregressive terms are important in reducing the prediction error. However, this does not indicate a linear AR model is sufficient to describe the system. The VAF of the linear AR model with only the 5 AR terms y(t-1) ... y(t-5) is only 36.83% in the 3-step ahead prediction.

Similar to the individual models, we compared the OSA prediction as well as MSA (3-step) model predicted outputs with the measured output using correlation coefficient, VAF and NRMSE to evaluate the models (See Table 3.14). For OSA, the mean correlation coefficient, VAF, and NRMSE for are 0.9691, 93.91% and 0.0472, respectively. For MSA, the mean correlation coefficient, VAF, and NRMSE are 0.6866, 47.09% and 0.1387, respectively.

Table 3.14 Performance statistics of NARX models with the common structure

No. of	Correlation	Correlation	VAF	VAF	NRMSE	NRMSE
P1	0.9773	0.7556	95.52	57.08	0.0397	0.1224
P2	0.9735	0.6366	94.74	39.53	0.0435	0.1459
P3	0.9642	0.5750	92.95	31.17	0.0467	0.1437
P4	0.9591	0.5891	91.94	32.26	0.0543	0.1563
P5	0.9698	0.7848	94.04	61.57	0.0468	0.1191

P6	0.9681	0.7028	93.72	49.18	0.0464	0.1323
P7	0.9784	0.8084	95.73	65.35	0.0487	0.1398
P8	0.9587	0.5952	91.90	32.57	0.0584	0.1689
P9	0.9607	0.6164	92.24	37.98	0.0515	0.1461
P10	0.9813	0.8024	96.28	64.21	0.0362	0.1126
Mean	0.9691	0.6866	93.91	47.09	0.0472	0.1387
Std.	0.0079	0.0898	1.54	13.28	0.0062	0.0165

We compared the OSA prediction as well as k-step ahead (k = 3) model predicted outputs with the measured output using correlation coefficient, VAF and NRMSE to evaluate the models (see Table 3.14). Comparisons of the NARX model predicted output (obtained from the k-step ahead prediction) and the corresponding measured cortical responses are shown in Fig. 3.9 for the ten participants. As shown in Fig.3.9, waveforms of predicted outputs and measured cortical responses look very similar across participants. The autocorrelations of model residuals are shown in Fig. 3.10. Since the common model estimation requires that the model fits different data realizations, the model residual may not be a perfect white noise. For most participants, the statistically significant non-zero auto-correlation values rarely occur with very small magnitudes, indicating that the estimated NARX models describe the inherent dynamics of the cortical response well.

For comparison purpose, common structure Volterra models with 20 model terms are also built. The mean correlation coefficient, VAF, and NRMSE are 0.4893, 23.27% and 0.1690, respectively, which are worse than the NARX model. These results indicate that the inclusion of autoregressive terms, as with a NARX model, improves the model prediction performance substantially

The results indicated that the cortical response can be better explained by the NARMAX method in comparison to previous studies using a linear system identification approach and Volterra kernels (Vlaar, et al., 2017b). The Volterra model can be considered as a special case of the NARMAX model, i.e. a NARMAX model without autoregressive (AR) terms. Our results indicate that the AR terms are essential to reduce the model error and decrease the model complexity.

In modeling, the performance of a common structure model (and using individualized model parameter values) is slightly lower than subject-specific structure models. However, a subject-specific model structure could not summarize common characteristics across subjects. A common model structure attempts to capture the common characteristics shared by and buried in all datasets, by sacrificing local properties hidden in individual datasets. A key advantage of the common model structure for the cortical response is that the model structure reveals the most important inherent features that can explain all data from different participants. Nevertheless, the parameter values may differ from subject to subject when the common model structure is used (see Table 3.12). The common model structure approach may be highly useful for future pathophysiological research to detect abnormalities after neurological dysfunction.

The OSA yielded much better performances than the k-step ahead for both subject-specific models as well as the common model. The k-step ahead prediction for brain activity is still a recognized challenge in the specific field of brain signal modeling due to the complexity of brain dynamics, as well as the poor signal to noise ratio and the non-stationary properties of EEG signals. In this study, the sampling rate of EEG signal is 256 Hz, then each sample time lag is approximately 4 milliseconds (ms). Thus, k-step ahead prediction actually estimates brain activity based on the measured brain “state”, i.e. the output, around 12 ms ago (in case k is 3 steps).

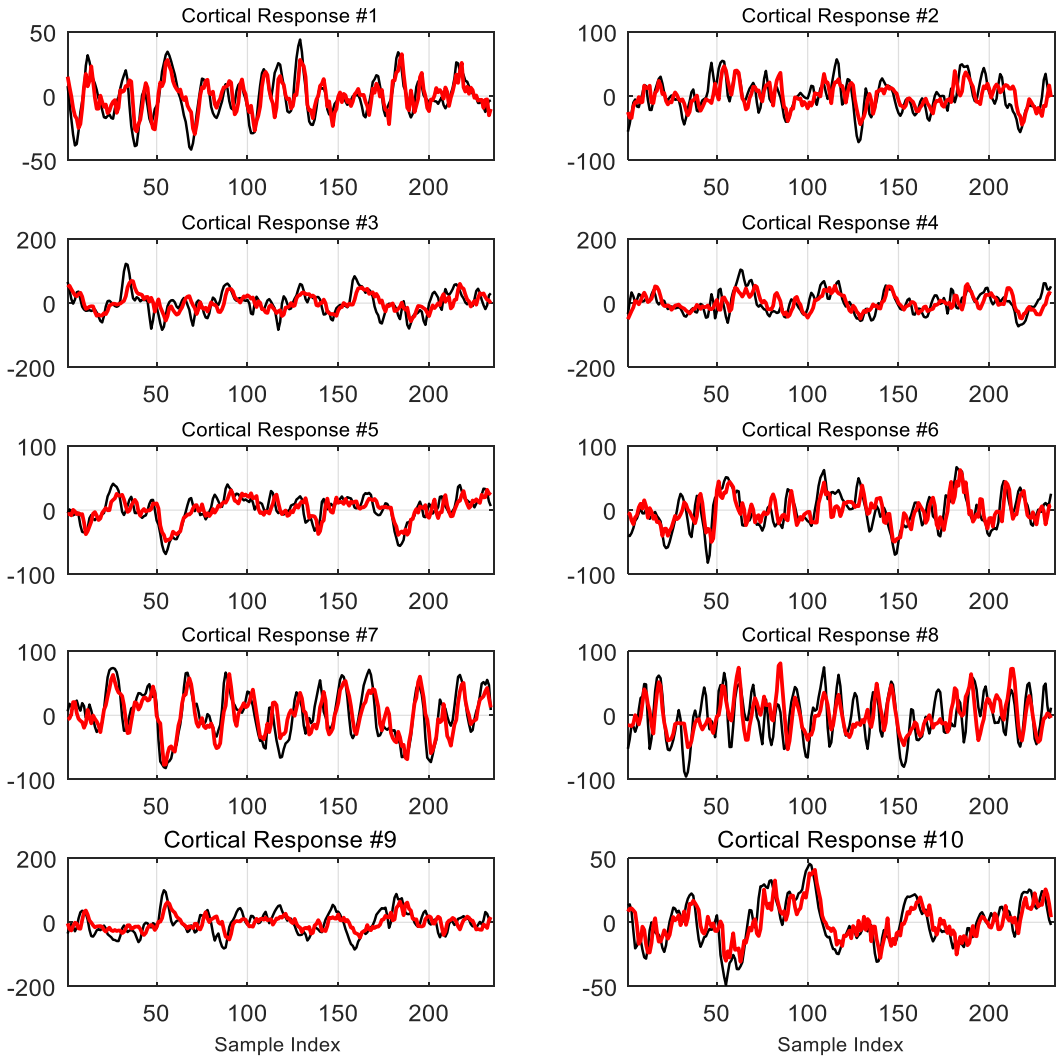

Figure 3.9 Comparisons of model predicted outputs (3-step ahead prediction) and the corresponding measurements of cortical responses for the ten participants (red line: model prediction outputs, black line: measurements of cortical responses).

As shown in Table 3.12, all model terms (except the constant term) are dynamic components with specific time lags. Multiple nonlinear terms and time lags in the common model structure revealed that the processing of somatosensory information in the human nervous system involves multiple neuronal circuitries with different neural transmission delays. These results provide new evidence to support the previous theoretical explanations on neurophysiological mechanisms underlying nonlinear processing of somatosensory information in the human nervous system (Yang, et al., 2018).

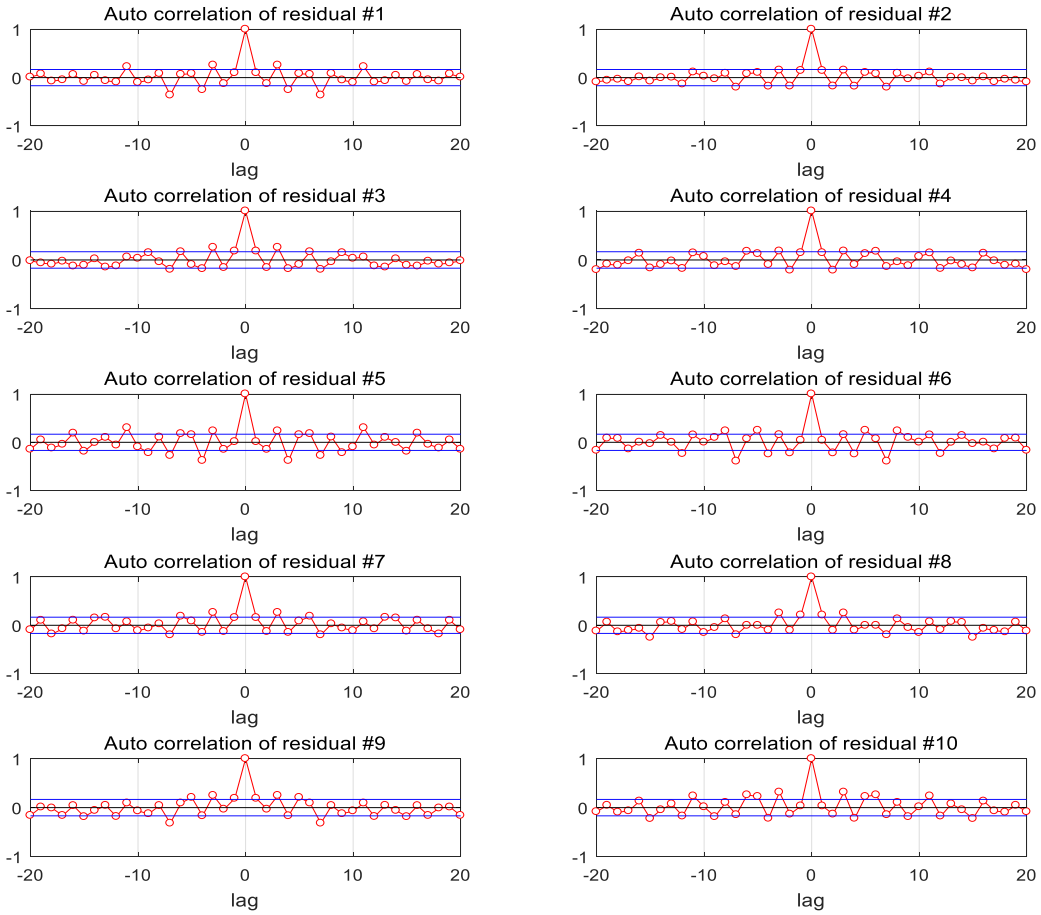

Figure 3.10 Auto-correlations of the model residuals for the ten study participants (blue lines indicate 99% confidence bounds)

The human nervous system receives the mechanical perturbation to the wrist via mechanoreceptors including muscle spindles, Golgi tendon organs, and cutaneous afferents. There are two kinds of sensory fibers in muscle spindles: type Ia primarily sensing muscle stretch velocity and type II primarily sensing muscle stretch. Golgi tendon organ (Ib fibers) detects the tendon strain and as such the force in the muscle-tendon complex. The transmission delays for type Ia fibers are much shorter than those for type II and Ib fibers. Finally, cutaneous afferents (Aβ fibers) conduct the activity of skin sensors resulting from the mechanical perturbation. When the participants are subjected to the mechanical perturbations, all these sensory fibers are active and sense different modalities with different transmission delays. Nonlinear terms with input signal u are likely associated with nonlinear encoding and processing of external inputs in the nervous system. Different time lags may be related to different transmission delays in the sensory input pathways from the mechanoreceptors to the brain.

In the model, we also found (AR) terms with output signal y, both linear (e.g., y(t-5)) and nonlinear (e.g. y(t-1)y(t-1)). These output related terms indicate that both linear and nonlinear neuronal interactions occur at the cortex, presumably caused by cortical neural networks or the inherent dynamics of the cortical processes. Nevertheless, the linear terms have much large weights than the nonlinear terms (see Table 3.12), indicating the dominance of the linear terms in the AR part of the model.

3.6 Conclusion

This chapter focuses on improving model identification methods from small size data. When the size of data is small or data is corrupted with noises, there is large uncertainty of model structure and parameter. These conditions can bring a negative effect on the model structure selection process of the classic OFR method. In this study, the RMSS method is proposed to enhance the classic OFR algorithm by selecting the robust significant model terms according to the OMAEs of resampled sub-datasets. The new method is tested on two simulation examples and real data applications. The results suggest that the new method can improve the prediction performance of modelling problems, especially when the data size is small and there are strong noises and unknown system components. The advantage of this robust model is that it can better capture the inherent dynamics of the whole dataset and thus can be well generalized to new data. Thus, the new method can be applied for small sample size and multiple datasets problems.

As this method does not analyse model uncertainty (e.g. the uncertainty existing in both model structure and model parameters) and its effect on model generalization performance. Inspired by the concepts and ideas proposed for fuzzy and cloud model, one of the future research directions would be focusing on quantitative analysis of model uncertainty, which is presented in the next chapter. In addition, the idea behind Generative Adversarial Network (GAN) (Liu, et al., 2019; Wang, Fan, Zhu & Tang, 2018) would be potentially useful for dealing with small size data modelling problem. A GAN based approach will be considered in future work.

Chapter 4

SYSTEM IDENTIFICATION AND UNCERTAINTY ANALYSIS USING A NEW CLOUD-NARX MODEL

4.1 Introduction

In model identification, the existence of uncertainty normally generates negative impact on the accuracy and performance of the identified models. This chapter introduces a novel cloud NARX model for model identification and uncertainty analysis. It is the first time that a cloud representation is introduced and incorporated with NARX model to provide a nonlinear representation of both the systems and uncertainty.

The presented model uses uncertainty 'cloud' model and cloud transformation to quantify the uncertainty throughout the structure detection, parameter estimation and model prediction. The new predicted band can be generated to forecast AE index with confidence interval. The proposed method provides a new way to evaluate the model based on uncertainty analysis, revealing the reliability of model and visualize the bias of model prediction. Cloud model is a cognitive model which provides a way of bidirectional transformation between a qualitative concept 'cloud' and the quantitative data 'cloud drops' (Wang, Xu & Li, 2014). The cloud concept is described by three numerical characteristics, namely ex (expectation), en (entropy) and he (hyper entropy). ex is the expectation of all elements in the set and en is the variance of the distribution. he depicts

the degree of departure from normal distribution of the cloud model. Thus, the cloud model is an extension of normal distribution. In figure 4.1, several cloud models with the same ex, en and different he are presented as a simple example. When $he = 0$, the cloud model is a normal distribution. As he increases, the cloud model departs from normal distribution and follows a new 'cloud' distribution.

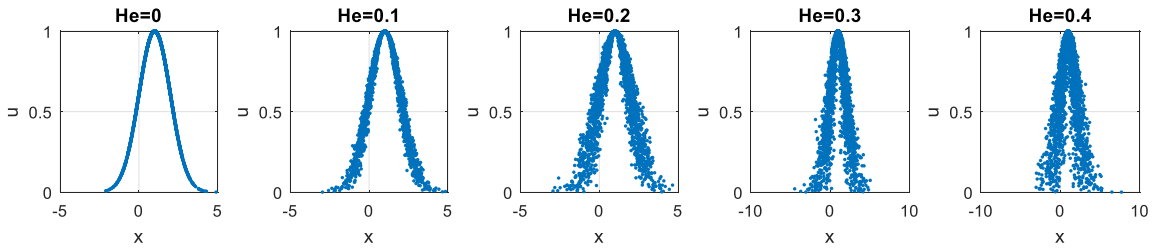

Figure 4.1 The cloud membership functions of cloud models with different values of he $(ex = 1, en = 1)$

Remarks: The advantage of cloud model is that it provides a way to describe a distribution with only three parameters that cannot be characterized by traditional normal distribution. The cloud transformation is better and more powerful than the normal distribution in that: i) it includes normal distribution as a special case; and ii) many data in real life do not follow a normal distribution. In figure 1, each element is a cloud drop and all the cloud drops together form the cloud concept. The bridges between cloud model and cloud drops is cloud transformation. The forward transformation is used to generate cloud drops from a known cloud model. The backward transformation is used to identify the cloud model from a sequence of cloud drops.

4.2 Cloud-NARX Model

Under the assumption that these model structures can perfectly describe the true system components, most of the models are capable to provide accurate representations of the system. However, in many practical scenarios, this assumption is usually invalid due to the existence of uncertainty. The existence of noise could lead to biased parameter estimation, incorrected selected model terms, etc. Therefore, quantifying uncertainty is

essential for establishing the significance of findings and making predictions with known confidence.

From the literature, it is known that estimating the true uncertainty remains an elusive goal. This study aims to quantify uncertainty using a new concept called cloud model (Wang, Xu & Li, 2014). Based on the cloud model and cloud transformation, a novel cloud-NARX model is proposed to quantify the uncertainty in the system modelling process. The new method can quantify the uncertainty efficiently. In addition, the new model is also capable to generate model prediction with known confidence and provide the information of how much uncertainty exists in the model prediction.

4.2.1 The cloud-NARX model structure

Based on cloud model and cloud transformation, a cloud-NARX model is proposed. The idea behind the metrics is to use a new uncertainty 'concept' (cloud model) to replace the traditional model parameters. A series of predicted points can be calculated by performing a transformation (generic cloud transformation) to generate a series of model parameters (cloud drops) from the concept. These predicted points form a predicted distribution (surface/band) with confidence intervals, describing the uncertainty and risk brought by model uncertainty. The cloud-NARX model can be described:

$$y = \sum_{i=1}^{n} Cloud_{l_i}(ex, en, he)\ \varphi_{l_i} \tag{4.1}$$

where $Cloud_{l_i}(ex, en, he)$ $(i = 1, 2, \dots, n)$ are the cloud models, which represent the estimated parameters and the uncertainty of these parameters. The parameters ex, en, he are the characteristics of each cloud model.

4.2.2 Estimation of the cloud-NARX model

The estimation of cloud-NARX model consists of three steps, which are data resampling, sub-model identification and cloud parameter estimation. The general process of estimating the cloud-NARX model is shown in Figure 4.2.

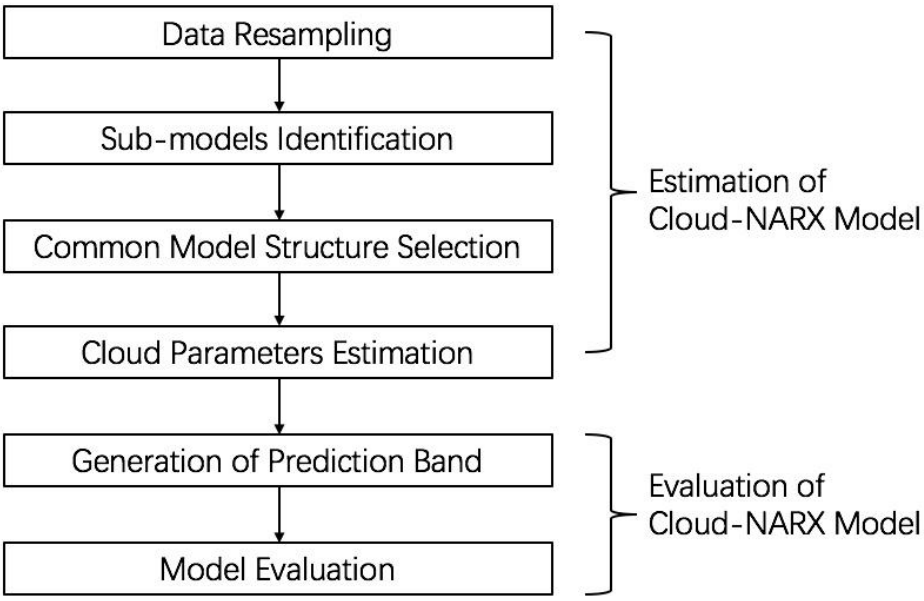

Figure 4.2 The process of estimation and evaluation of the cloud-NARX model

First, the original dataset can be regrouped to form *K* sub-datasets through some resampling methods e.g. random sampling or bootstrap (Wei & Billings, 2009). Assume that the input and output sequence for the *k*-th dataset is $\{u^{(k)}(t)\}_{t=1}^{N_k}$ and $\{y^{(k)}(t)\}_{t=1}^{N_k}$, respectively, for $k = 1, 2, \dots, K$. The model terms $[\varphi_1^{(\mathrm{k})}(\mathrm{t}), \dots, \varphi_{\mathrm{M}}^{(\mathrm{k})}(\mathrm{t})]$ of the *k*-th dataset can be generated from the associated regressor vector relating to the *k*-th dataset $[y^{(k)}(t-1), \dots, y^{(k)}(t-n_y), u^{(k)}(t-1), \dots, u^{(k)}(t-n_u)]^{T}$. The sub-model for the *k*-th sub-dataset can be written in the compact matrix form:

$$y^{(k)} = \sum_{\mathrm{i}=1}^{n} \theta_{l_i}{}^{(k)} \, \varphi_{l_i}{}^{(k)} \tag{4.2}$$

Second, for each sub-dataset, the OFR algorithm can be applied to select a number of significant model terms to establish an individual NARX model. A common model structure can be formed by model terms selected in all the sub-datasets. In addition, a robust model structure selection (RMSS) method is developed as an alternative method, for small size data modelling problem (Gu & Wei, 2018a). With OFR or RMSS method, a common model structure $\{\varphi_{l_1}, \dots, \varphi_{l_n}\}$ can be identified and the associated model parameters for each sub-dataset can be estimated as $[\theta_{l_i}{}^{(1)}, \theta_{l_i}{}^{(2)}, \dots \theta_{l_i}{}^{(K)}]$.

Finally, the cloud model for each model term can be identified from the K groups of model parameters using generic backward cloud transformation (Wang, Xu & Li, 2014).

$$\left[\theta_{l_i}{}^{(1)}, \theta_{l_i}{}^{(2)}, \dots\ \theta_{l_i}{}^{(K)}\right] \xrightarrow{GBCT} Cloud\ {}_{l_i}(ex, en, he) \tag{4.3}$$

where $i = 1, 2, \dots, n$. So that the cloud-NARX model can be identified.

It is known that when the model structure is perfect, and the data is not corrupted with noises, any of the sub-datasets will lead to exact the same model. However, the model structures of the sub-models might be different when there is model uncertainty brought by the noises/disturbances/insufficient information. In these situations, any single model might be unreliable. By removing or adding some data points in the K sub-datasets, the uncertainty can be quantified by the sub-models with different structures and parameters.

4.2.3 Model Predicted Band and Averaged Prediction

With identified cloud model of each parameter, K' groups of cloud drops are generated using cloud forward transformation, as follows:

$$Cloud\ {}_{l_i}(ex, en, he) \xrightarrow{GFCT} [\hat{a}_{l_i}^{(1)}, \hat{a}_{l_i}^{(2)}, \dots, \hat{a}_{l_i}^{(K)}] \tag{4.4}$$

where $\hat{a}_{l_i}^{(\mathrm{k}\prime)}$ is the generated parameters for the model term θ_{l_i} with $k' = 1, 2, \dots, K'$. A number of K' predicted time series of output y can then be calculated as:

$$\hat{y}^{(k\prime)} = \hat{a}_{l_1}^{(k\prime)}\, \varphi_{l_1} + \hat{a}_{l_2}^{k\prime}\, \varphi_{l_2} + \cdots + \hat{a}_{l_n}^{(k\prime)}\, \varphi_{l_n} \tag{4.5}$$

where k' is the index of predicted time series ($k' = 1, 2, \dots, K'$). The K' model prediction can then form a predicted band with density. The upper and lower boundaries of the predicted band can be determined as:

$$\hat{y}_{lower} = min\ (\hat{y}^{(1)}, \hat{y}^{(2)}, \dots, \hat{y}^{(K\prime)}) \tag{4.6}$$

$$\hat{y}_{upper} = max\ (\hat{y}^{(1)}, \hat{y}^{(2)}, \dots, \hat{y}^{(K\prime)}) \tag{4.7}$$

The averaged model prediction can also be calculated as:

$$\hat{y}_{averaged} = \frac{1}{K\prime}\sum_{i=1}^{K\prime} \hat{y}^{(i)} \tag{4.8}$$

4.2.4 Model Performance Evaluation

To evaluate the averaged prediction of the model, the correlation coefficient (ρ), prediction efficiency (PE) and normalized root-mean-square error (NRMSE) are calculated. The PE is defined as:

$$PE = 1 - \frac{\sigma_{error}^2}{\sigma_{observed}^2} \tag{4.9}$$

where $\sigma_{observed}^2$ is the variance of the observed AE values and σ_{error}^2 is the variance of the error between the predicted AE values and observed AE values. The accuracy of the predicted band can be defined as:

$$\gamma = \frac{N_t{}'}{N_t} \tag{4.10}$$

where N_t is the total number of observed data points in test dataset and $N_t{}'$ is Number of the observed data points within the predicted band.

4.3 Simulation

In this section, two simulation examples are presented.

4.3.1 A Simple Linear System

Consider a simple linear system:

$$y = u_1 + 0.7u_2 + 0.4u_3 + 0.2u_4 + \xi(t) \tag{4.11}$$

where the inputs $u_1 \ldots u_4$ was assumed to be uniformly distributed on $[-1, 1]$, and the noise $\xi(t)$ is a Gaussian white noise. Input-output data points were generated and grouped for model estimation and performance test.

A comparison of the NARX and cloud-NARX model is presented in Table 4.1. From the table, the true parameters cannot be estimated due to noise. The conventional NARX model uses single parameters which are biased. The cloud-NARX model uses two extra parameters en and he to quantify the uncertainty of the parameter estimation.

Table 4.1 Estimated parameters of NARX model and Cloud-NARX model

Terms	True Parameter	NARX Parameter	ex	en	he
u_1	1	0.9907	0.9907	0.0433	0.0112
u_2	0.7	0.6975	0.6975	0.0434	0.0125
u_3	0.4	0.4020	0.4020	0.0424	0.0122
u_4	0.2	0.2034	0.2034	0.0454	0.0113

The cloud membership functions of the model terms are presented in figure 4.3. From the figure, the bias of parameter estimation can be well described by the cloud models. According to the sample prediction of the cloud-NARX model in figure 4.4, the cloud-NARX model provide a predicted band/surface to quantify the prediction uncertainty. The prediction error of the conventional NARX prediction can be better described by the Cloud-NARX model.

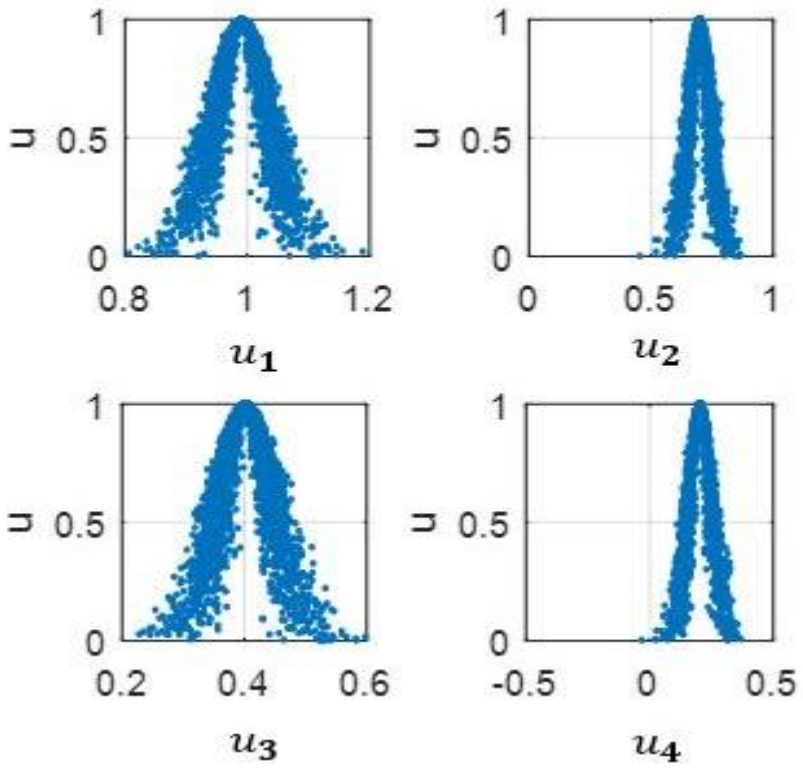

Figure 4.3 The cloud membership functions of the selected model terms u_1, u_2, u_3, u_4

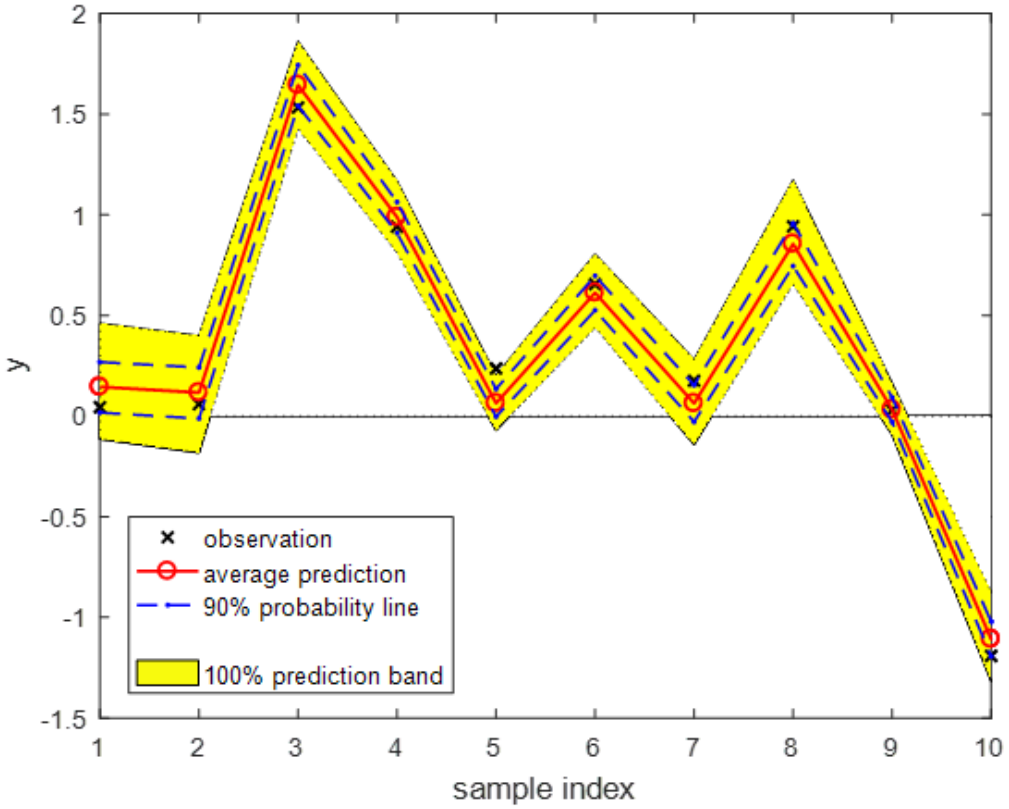

Figure 4.4 Model prediction of the cloud-NARX model

4.3.2 A Nonlinear Dynamic System

Consider a nonlinear dynamic system:

$$y(t) = -0.5y(t-2) + 0.7y(t-1)u(t-1) + 0.6u^2(t-2) + 0.2y^3(t-1) - 0.7y(t-2)u^2(t-1) \quad (4.12)$$

where the input $u(t)$ was assumed to be uniformly distributed on $[-1, 1]$, and the noise $\xi(t)$ is a Gaussian white noise. The SNR of the data is 30 dB. A total number of 1000 input-output data points were generated. The first 500 points were used for model estimation and selection and the remaining 500 points were used for performance test. A regression vector can be defined as:

$$\varphi(t) = [y(t-1), y(t-2), u(t-1), u(t-2)]^T \quad (4.13)$$

with the maximum time lags of $n_y = n_u = 2$. The initial full model was chosen to be a polynomial form with nonlinear degree of $l = 3$. It can be noted that all the components of the system can be well represented by the candidate model terms.

Table 4.2 Cloud-NARX model with cloud parameters

Model Term	ex	en	he
u(t-2) *u(t-2)	0.5937	0.0268	0.0007
y(t-2)	-0.4924	0.0473	0.0026
u(t-1) *y(t-1)	0.6923	0.0673	0.0040
u(t-1) *u(t-1) *y(t-2)	-0.7024	0.1262	0.0110

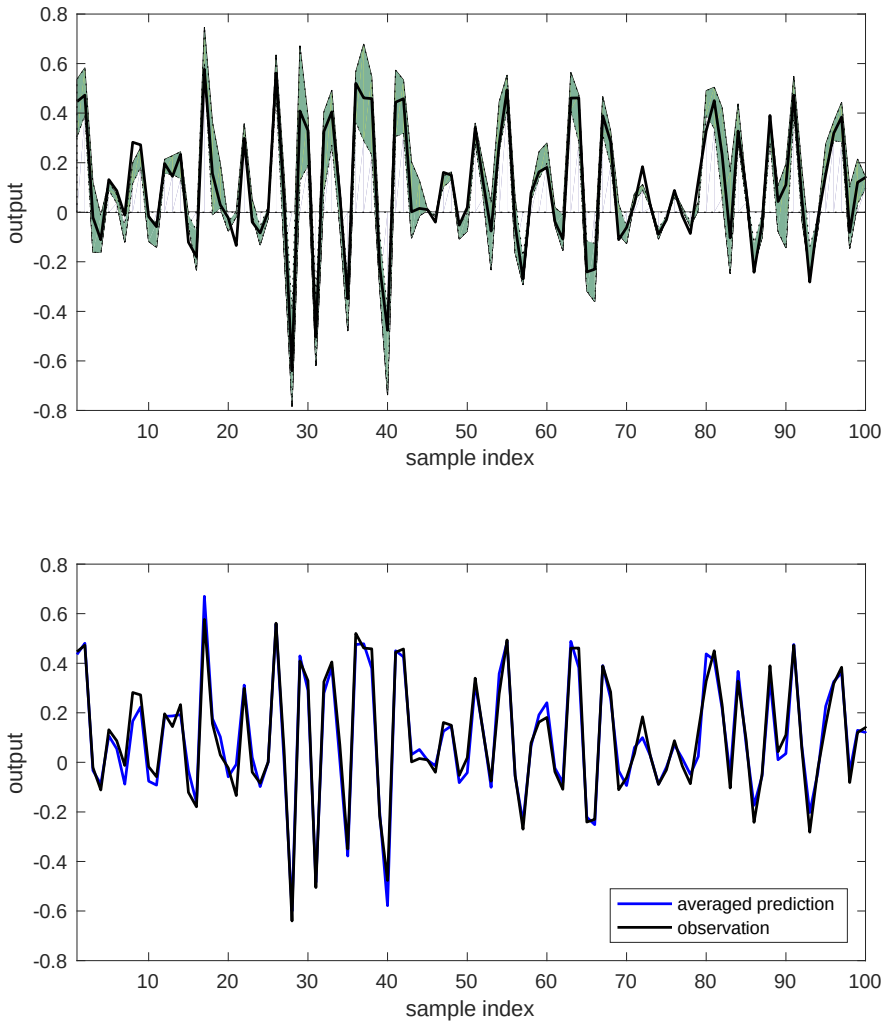

Figure 4.5 Comparison of predicted band, averaged predicted and observation of test dataset

The correlation coefficient, prediction efficiency and NRMSE of the Cloud-NARX model with 5 terms on test dataset are 0.99, 0.99 and 0.01 (averaged prediction), while the conventional NARX model has nearly the same statistics.

The reason that both the two models achieve high performances is that the system in this example has a structure which can be well represented by the selected model terms, so that there is no model structure uncertainty. Also, noise of the data is not very strong so that the parameters are estimated without much disturbances. Thus, it is not surprising that in this example, there are no significant differences of the performances of the two models.

The cloud-NARX model can provide a predicted band to visualize the confidence interval of the model prediction. A comparison of the predicted band/observations and averaged predicted line/observation is shown in Fig. 4.5. The accuracy of the predicted band is 90%, meaning that 90% of the observed points is within the band.

From the figure, the predicted band is narrow for most of the data points. When some points of the averaged line are far from the real observed points, the predicted band becomes quite wide and covers most of the observed points. In this way, it is possible to know when the model uncertainty is large and the averaged prediction become unreliable. For risk analysis and prediction, it is extremely useful to avoid losses caused by the model uncertainty.

Fig. 4.6 shows an example of the model prediction for some selected data points. From the figure, the frequency distribution of the predictions is displayed and the maximum, minimum and average value of the prediction is compared to the observation. There are two clear boundaries, indicating the uncertainty in the prediction. The average prediction can be used as the conventional model prediction. Comparing to the conventional model prediction, the new predicted band provides extra information on how much the uncertainty of model prediction is. Therefore, the cloud-NARX model is more robust and the uncertainty in model prediction can be visualized.

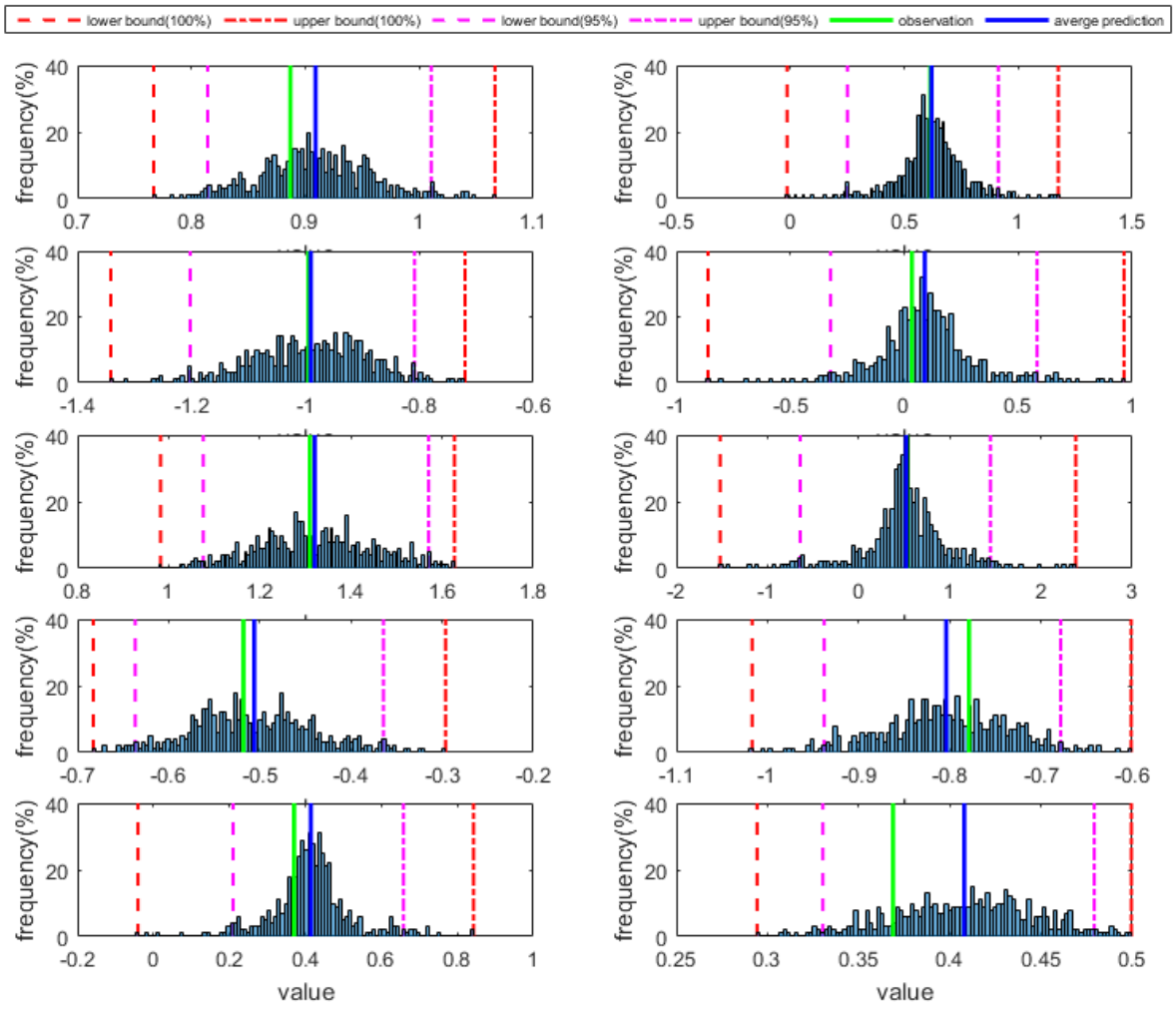

Figure 4.6 Prediction of Cloud-NARX model on 10 randomly selected test data points

4.4 Real Data Case Study: AE Index Modelling

This section presents a real data modelling case study, where the proposed cloud-NARX model is applied to the modelling and forecasting of AE index.

4.4.1 Backgrounds

Many modern technological systems are sensitive to space weather disturbances, such as geomagnetic storms and sub storms and ionosphere variability (Ayala Solares et al., 2016). The severe situations of space weather can have harmful effect on power grid, navigation systems, and satellite system. Thus, it is extremely important to forecast the space weather

disturbances to avoid damages and losses. In addition to the traditional first principle and statistical approaches for understanding the interactions between the solar wind and magnetosphere (e.g., Ala-Lahti et al., 2018 and the references therein), application of data based methods and in particular techniques based on machine learning to the prediction of various geomagnetic indices resulted in many MLE-NARMAXovative forecasting models (e.g. Wintoft & Cander, 2000; Chandorkar, Camporeale, & Wing, 2017; Camporeale et al., 2018).

The AE index, along with the Al and AU indices, was introduced by Davis and Sugiura (1966) as a measurement of global auroral electrojet activity (Mayaud 1980). Changes in AE are driven by variations in the solar wind convection electric field produced by fluctuations in the solar wind velocity and IMF. These two factors govern the efficiency of the coupling between the solar wind and terrestrial magnetosphere with the dominant role being played by a southward directed IMF. In this coupling process, the energy associated with the solar wind flow is converted into magnetic energy which is transferred into the magnetosphere via reconnection processes on the day side and is stored in the magnetotail. This energy is eventually released, energising the plasmasheet, ring current, and ionosphere.

Three classes of interactions have been identified, depending upon the southward turnings of the IMF (see e.g. Gonzales et al., 1994). Short lived southward turnings of the IMF with modest (Bz~-3nT) give rise to minor intensifications of the ring current, yielding substorm events. Repeated southward turnings, referred to as HILDCAA (high intensity, long duration, continuous AE activity) events arise due to the occurrence of interplanetary Alfven wave train embedded within the solar wind flow (Tsurutani and Gonzalez, 1987). These events result in a continued period of AE activity. Finally, Coronal Mass Ejections (CMEs) or magnetic clouds exhibit extended periods in which a strong Bz is observed. This coupling, between the CME and terrestrial magnetosphere, results in a major intensification of the ring current, and large deviations in both AE and Dst and is referred to as an intense magnetic storm.

Previous studies of substorm using AE index have provided accumulated evidence that the magnetosphere behaves as a nonlinear dynamic system, and it can be described by a small number of variables (Kamide et al., 1998). There are plenty of studies aiming to forecast AE index from solar wind measurements. Among the many approaches of

modelling and forecast, neural networks (NN) is a commonly used method. Early in 1997, neural network models were constructed to study prediction of the AE index (Takalo & Timonen, 1997). Later, an ANN algorithm based at interplanetary magnetic field measured on Lagrangian point L1 and plasma measurements was introduced in 2008 to predict AE index from 5 to 60 minutes ahead (Pallocchia et al., 2008). The ANN models were further improved to achieve a correlation coefficient of 0.83 for 1-hour-ahead forecast and 0.80 for 3-hour-ahead forecast, respectively (Bala & Reiff, 2012). In addition, some other approaches are also applied for the analysis, for example, a correlation analysis with a technique of wavelet decomposition and selective reconstruction was applied to analyse the relationship between AE index and solar wind variables (Guarnieri, et al., 2018). The advantage of neural networks and its variants is that it can provide an efficient nonlinear representation to generate good model predictions. However, the identification process of neural networks often involves a large number of variables, so that the model structure of neural networks can be very complicated. From such model structure, it is quite difficult to further understand the nonlinear dynamic of the system, for example, which model term/variables are superior for describing the index and which model terms/variables are redundant. Nevertheless, it is obvious that such a model cannot provide a model structure that is simple and easy for understanding.

Another widely used approach for the modelling and forecast of magnetosphere is the nonlinear autoregressive with exogenous input (NARX) model. The NARX model is developed for the nonlinear system identification and can detect an appropriate model structure by selecting the most important model terms from a dictionary consisting of a huge number of candidate model terms (Billing 2013). Thus, it is very efficient method for the space weather forecast due to the fact that the magnetosphere is a nonlinear process (Kamide et al., 1998). The NARX model have successfully solved the modelling and prediction of many magnetic indices, for example, the Dst index (Balikhin et al, 2011; Boynton et al., 2011; Wei, et al., 2004), the Kp index (Ayala Solares et al., 2016), etc.

Comparing to the neural networks, the NARX method only uses a small number of effective model terms to describe the system, so that the system can be represented a linear-in-the-parameter form which is parsimonious and transparent.

4.4.2 Data Description

A full description of the solar wind variables and the magnetic indices is given in Table 4.3. The AE index is one of the most widely used indices for researchers in geomagnetism, aeronomy and solar-terrestrial physics, to understand the geomagnetic activity. The AE index is the maximum deviation of the horizontal components of geomagnetic field variations from a set of globally distributed ground-based magnetometers located in and near the auroral zone in the Northern Hemisphere (Guarnieri, et al., 2018). It increases when a sub storm event is happening and represents the overall disturbance in both eastward and westward ionospheric electrojets located at around 100 km altitude (Davis & Sugiura, 1966).

Table 4.3 Descriptions of the solar wind variables and AE index

Variable	Description
y	AE index
V	solar wind speed/velocity (flow speed) [km/s]
Bst	Interplanetary magnetic field factor [nT]
n	solar wind density (proton density) [cm^{-3}]
p	solar wind pressure (flow pressure) [nPa]

Note: *Bst* = $B_T \sin^6(\theta/2)$ [nT] (Boynton et al., 2011)

The AE index and solar wind variables used in this study were all sampled hourly. The AE index and solar wind variables are used as the output and input of the systems modelling, respectively. The amplitude of the solar wind velocity is around 250-1000 km/s, which is much larger than those of the other input signals. To avoid producing extreme parameter estimations, the solar wind speed/velocity variable is firstly normalized by $V \to V'/1000$, where V' is the original signal and V is the normalized signal. Two derived variables, $\sqrt{p}$ and $VBst = V \times B_T \sin^6(\theta/2)$ (Boynton et al., 2011), which are effective in describing the magnetic indices, are also used as input variables for the system model.

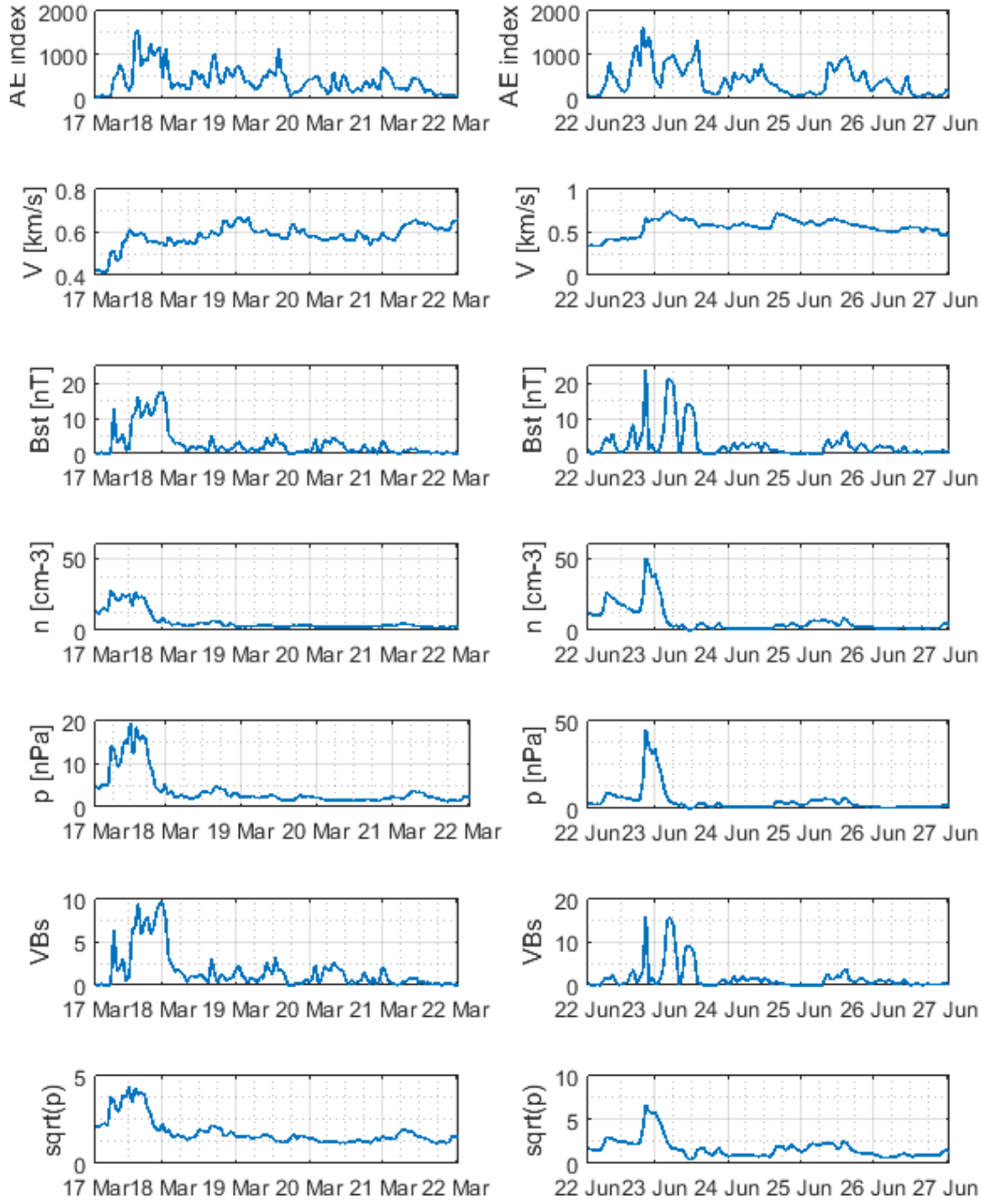

Figure 4.7 Observation of hourly sampled AE index and solar wind variables of two interested periods of 2015

4.4.3 Construction of the cloud-NARX Model

The AE and solar wind data from 2011 to 2013 (around 26000 sampled data points) were used for training the model and the data of 2015 (around 9000 sampled data points) were

used for model validation. In the test dataset, two time periods of strong magnetic storm, 17-21 Mar 2015 and 22-26 Jun 2015 (120 sampled data points for each period) were used as special cases to evaluate the model. The time series of the AE index and solar wind variables of the two interested periods are shown in Figure 4.7. The figure shows that there were two significant storms on 17th Mar 2015 and 22th Jun 2015. Both periods match ICMEs. The first period 17-21 March corresponds to St Patrick storm caused by the CME on the 15th of March [see: https://www.swpc.noaa.gov/sites/default/files/images/u33/StPatrick%27sDay_Geomagnetic_Storm.pdf] whereas the second period 22-26 June 2015 corresponds to the ICME registered by Wind [https://wind.nasa.gov/cycle24.php].

Table 4.4 Cloud-NARX model with cloud parameters

No.	Model Terms	ex	en	he
1	$Bst(t-02)$	-9.7009	8.9120	0.0615
2	$VBs(t-1)$	6.0214	24.8546	16.3577
3	$y(t-01)$	0.6252	0.0108	0.0037
4	$V(t-01) \times Bst(t-01)$	143.6189	33.5311	0.0614
5	$V(t-01) \times Bst(t-02)$	-19.7937	22.3581	0.6368
6	$V(t-01) \times \sqrt{p}(t-01)$	14.3895	15.9883	0.1581
7	$V(t-02) \times p(t-01)$	7.8305	9.0816	0.1808
8	$V(t-02) \times \sqrt{p}\,(t-01)$	2.7969	7.8950	2.9935
9	$Bst(t-2) \times \sqrt{p}(t-02)$	-0.0807	0.5708	0.9887
10	$p(t-2) \times VBst(t-1)$	-0.0795	0.2808	0.4866
11	$VBs(t-01) \times VBst\,(t-01)$	-5.6495	0.4665	0.7133
12	$VBst\,(t-02) \times y(t-01)$	-0.0195	0.0029	0.0008

Note: $Bst = B_T \sin^6(\theta/2)$ [nT], $VBst = V \times Bst$ (Boynton et al., 2011)

In order to determine the maximum time lags for both the input and output variables, we have carried out pre-modelling experiments and simulations, the results suggest that the maximum time lags of the input and output were chosen to be $nu = 2$ and $ny = 2$. The initial full model was chosen to be a polynomial form with nonlinear degree of 2. The input-output data points of training dataset were firstly resampled 100 times with replacement, to form 100 sub-datasets. For each sub-dataset, a NARX model with 6 model terms is identified. For convenience of description, these single NARX models are referred to as 'individual NARX models'. Thus, there are a total number of 100 different individual NARX models and each has its own parameters. A total of 12 different model terms are selected during the 100 runs, and these terms are used for cloud-NARX model construction. The cloud parameters of each of these selected model terms are shown in Table 4.4.

It is noteworthy that the cloud-NARX model consists of 12 model terms, rather than 6 terms, this is because that each individual NARX model has its own structure. There are some common terms which are included in nearly all the individual NARX models, for example, $VBst(t-02)$ and $y(t-01)$. Also, some terms for example, $VBst(t-02) \times y(t-01)$, is selected and included in a relatively small number of times out of the 100 runs. These rarely selected model terms are usually ignored in conventional NARX model because they make small contributions to the whole dataset. However, in some of the sub-datasets, they might be effective in some rare situations, for example, the peak times of the AE index.

Figure 4.8 shows the normal cloud membership functions of the 12 selected model terms. The estimated parameters of the some model terms are normally distributed, for example, $Bst(t-02)$, $V(t-01) * Bst(t-01)$, $V(t-01) \times Bst(t-02)$, $V(t-01) \times \sqrt{p}(t-01)$ and $V(t-02) \times p(t-01)$. The distributions of the parameters of some other model terms (for example $VBs(t-1), y(t-1), VBst(t-1) \times y(t-1)$) are beyond normal distributions. Note that the normal distributions are not always sufficient to describe the distribution of the estimated parameters of these model terms due to the existence of uncertainty which do not necessarily follow a normal distribution law. The three characteristics ex, en and he are used to analyze the uncertainty of each model term. As discussed earlier, ex is the mean of estimated parameters of each model term, which is consistent with the conventional model parameter; en is the variance of

the parameter estimation; he is the hype-parameter to describe the degree of departure of the distribution to normal distribution. The values of en of some model terms (for example $y(t-1)$) are quite small, which indicates that the parameters of these model terms in the individual models are very close. In other words, the contributions of these model terms are consistent in each individual model. On the contrary, the values of en of some model terms (for example $VBs(t-1)$) are quite large, which means that uncertainty of the estimated parameters of these model terms are strong. In the disturbed periods, the contributions of these model terms are different in each individual model and the width of the predicted band increases due to the prediction uncertainty. The cloud parameter he describes how much the distribution is beyond normal distribution. If the value of he is much smaller than that of en, it means the estimated parameters of the model term are normally distributed. With the hyper cloud parameter he, the cloud model can better describe the estimated model parameters which are not normally distributed.

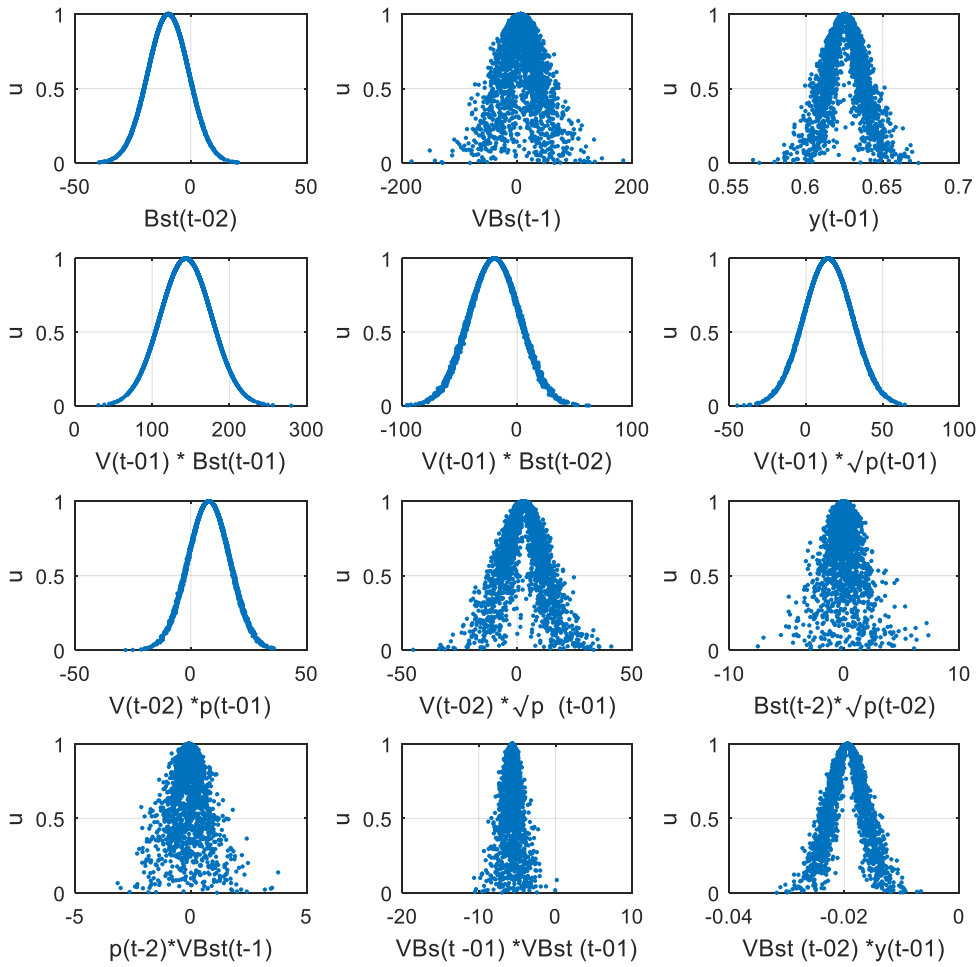

Figure 4.8 The normal cloud membership functions of the 12 selected model terms

4.4.4 One-hour-ahead Prediction of AE cloud-NARX Model

As mentioned earlier, the cloud-NARX model is built on hourly sampled data, so the model can be directly used to generate one-hour ahead (that is one-step-ahead) predictions of AE index. With the cloud parameters, the generic cloud forward transformation was applied to generate 100 sets of model parameters (that is called 'cloud drops' in the transformation) for all the selected terms. A total number of 100 time series of the AE index prediction was calculated. The average prediction and predicted band are presented in Figure 4.9. The predicted band is the quantification of uncertainty throughout the structure detection, parameter estimation and model prediction. If the model structure is perfect and the parameters are estimated unbiased, the predicted band will be narrow. Otherwise, if there are strong uncertainty in the data itself or the model structure and parameter, the uncertainty will be propagated to model prediction and the width of predicted band will be increased.

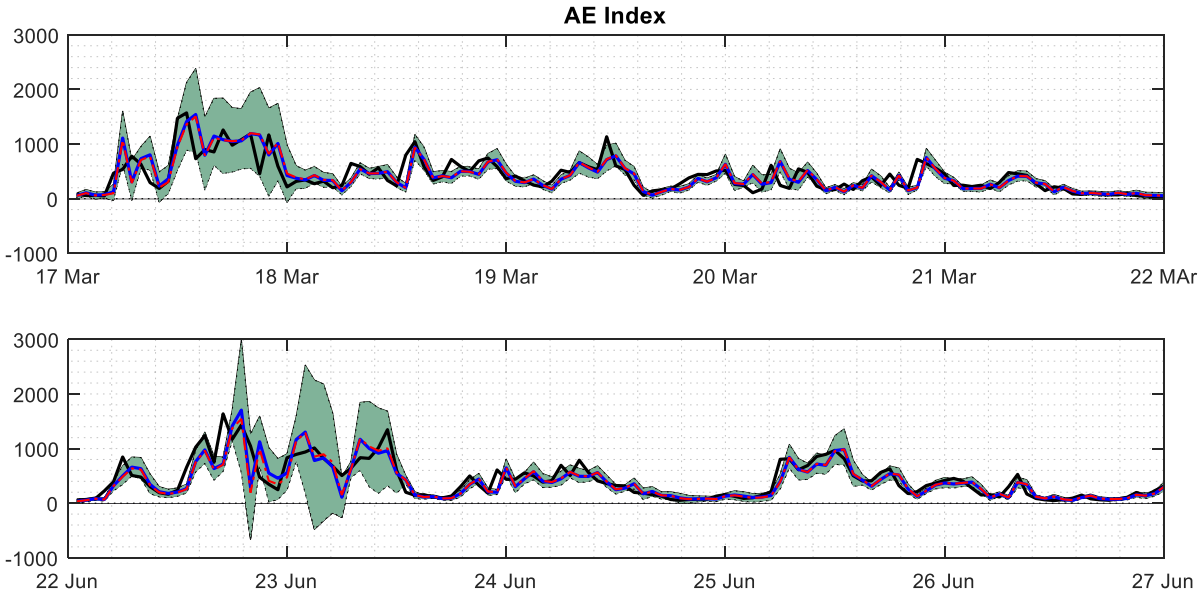

Figure 4.9 One-hour-ahead predicted band (consists of 80% of generated model predictions) and averaged prediction of AE index over 17-21 Mar and 22-26 Jun of 2015 (black line: observation; blue line: averaged prediction; green shadow: predicted band; red line: prediction of conventional NARX model)

From the figure, the predicted band is very wide over 17 Mar 2015 and 22 Jun 2015. This can be explained or understood as follows. First, from the input signals shown in

Figure 4.7, we know that there were interplanetary disturbances over the two periods. It is known that in general most storms last quite a short period in the long-term evolution of the process. As a consequence, most of the training data were sampled at 'quiet' times and only a very small fraction of the training data is for the storm period. This results in that the training data are severely 'imbalanced' (Ayala Solares et al., 2016). Therefore, while a single model may well capture the features and dynamics of the system at 'quiet' times, it may not sufficiently capture the system dynamics at the severe situation times. That is why the prediction band is so wide for these stormy periods. Second, the wide prediction band over the period of 17 Mar 2015 and 22 Jun 2015 implies that no single model would produce reliable prediction of AE over stormy periods, no matter what/which method is used to build the model. That is why we propose to carry out uncertainty analysis in this study.

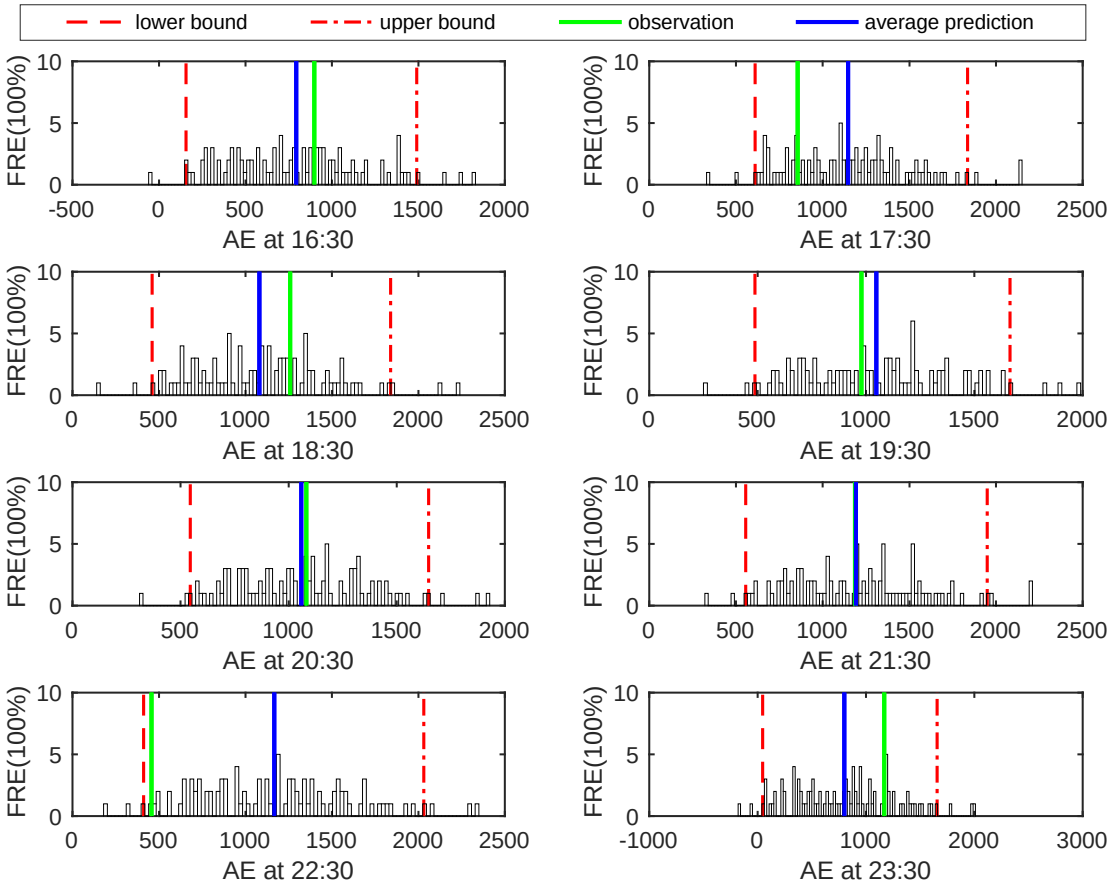

Figure 4.10 Predicted band with density over an 8-hours period on 17 Mar 2015 (FRE: the frequency of predicted AE occurrences in each divided bin of the predicted band)

Note that the predicted band in Figure 4.9 provides only rough quantification of the uncertainty. In many situations, the detailed information of the predicted AE index at a specific time point is often needed. Figure 4.10 and Figure 4.11 are the predicted bands with density over an 8-hours period on 17 Mar 2015 and 23th Jun 2015, respectively. Note that 90% of the prediction vectors are used to form the predicted band. These figures show the probability of the predicted AE index being in each interval. As shown in the figure, the interval of the predicted band for each time point is divided into 100 bins. The histogram shows the probability (frequency) of a single predicted AE value occurs in each bin. The boundaries of the predicted band are also displayed with the histogram, to visualize the prediction uncertainty and make it easier to understand. In addition, it is straightforward to compare the observation (green line) and averaged prediction (blue line) of AE index in the figure. The overall accuracy of the predicted band on the test dataset is 65%. The accuracy of high AE period ($AE > 1000$) is 70%.

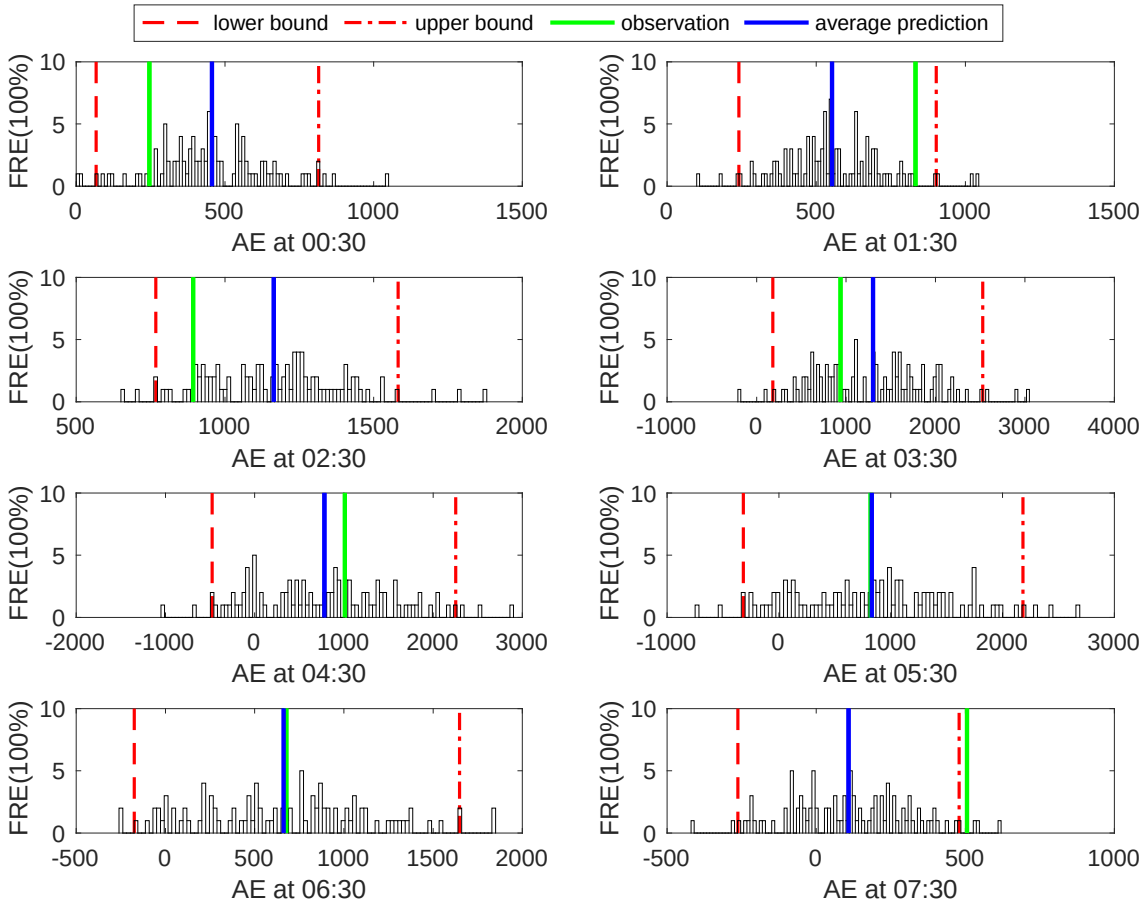

Figure 4.11 Predicted band with density over an 8-hours period on 23 Jun 2015 (FRE: the frequency of predicted AE occurrences in each divided bin of the predicted band)

The only way to reduce the width of the predicted band is to find a model structure which can better describe the true system. However, it is very hard, if not impossible, to obtain a perfect model structure for real-world system identification data modelling problem in the presence of strong uncertainty. Nevertheless, it should be noted that the performance of the model given by Table 4.4 outperforms previous models, for example, the NN model (Bala and Reiff, 2012) (as shown in Table 4.5). Therefore, a wide predicted band might indicate that a severe situation (interplanetary disturbances) is likely to happen. The property of the predicted band could potentially be used to forecast the arrival of the interplanetary disturbances.

4.4.5 Performance and advantage of the cloud-NARX Model

The performance of the averaged prediction of cloud-NARX model is comparable to the best NARX model with very similar structure but fixed model parameters, as shown in Table 4.5. Figure 4.12 presents the scatter plot of the averaged prediction and observation. The correlation coefficient, PE and NRMSE of the averaged prediction is 0.872, 75.97% and 0.0589 (for data of year 2015), which are consistent with the best NARX model. The NARX model outperform the NN model for 1 hour ahead prediction, as the previous NN model achieves the correlation of 0.83 (Bala and Reiff, 2012). More importantly, the cloud-NARX provides a transparent and parsimonious representation. As shown in Table 4.4, the NARX model reveals which of the variables/model terms are significant and which are not, for example, the model terms $V(t-02) \times Bst(t-01)$ indicates that the dayside reconnection 20-40 minutes prior (Balikhin et al., 2010) is an important component of the system, and the model terms $y(t-1)$ suggests that the autoregressive term has a significant effect on the AE index. On the contrary, the NN models are usually very complicated and the training process involves a huge number of model terms and takes a lot of time.

The cloud-NARX model holds all the good properties of conventional NARX model and possess an extra advantage. It provides a tool for understanding and analyzing uncertainty in the model structure and forecasting. For example, the uncertainty band in Figure 4.9 indicates that the model performs well for the period of 18-21 Mar 2015 and 24-26 Jun 2015, but the model is insufficient to characterize the dynamics of the process for the period 17 Mar 2015 and 22-23 Jun 2015 (i.e. when a sharp change occurs in solar

wind variables, for example, $Bst / VBst$). As discussed earlier, this property could potentially be used to forecast the arrival of a solar wind storm.

Table 4.5 Comparison of the Performances of the best NARX model and cloud-NARX model on test data of year 2015

Model	Correlation	PE	NRMSE
Best NARX model	0.8728	0.7611	0.0588
cloud-NARX model	0.8723	0.7597	0.0589
NN model	0.83	/	/

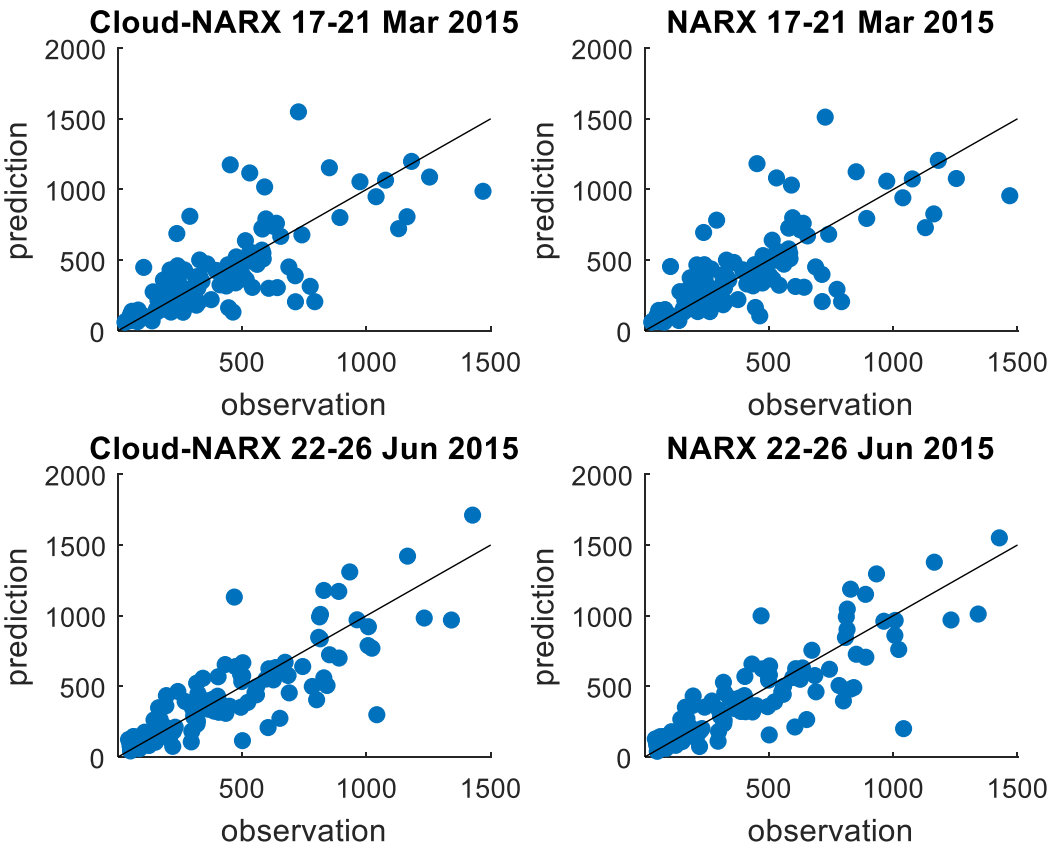

Figure 4.12 Scatter plot of the averaged prediction and observation of the cloud-NARX model and the best NARX model on two test datasets

Note that the model also works well and even better in non-disturbed periods. This is because that the model was trained on the dataset where most of the data were sampled at non-disturbed period. Therefore, the system behaviors in non-disturbed periods are well captured by the identified model. A comparison between the observed and predicted AE index in two selected non-disturbed periods (23 Apr ~ 5 May & 19 Oct ~ 1 Nov) is given in Figure 4.13. According to the figure, the predicted band is narrow, which means that the uncertainty of the model is not strong. From these results, the cloud-NARX model also generates good prediction results in the non-disturbed times.

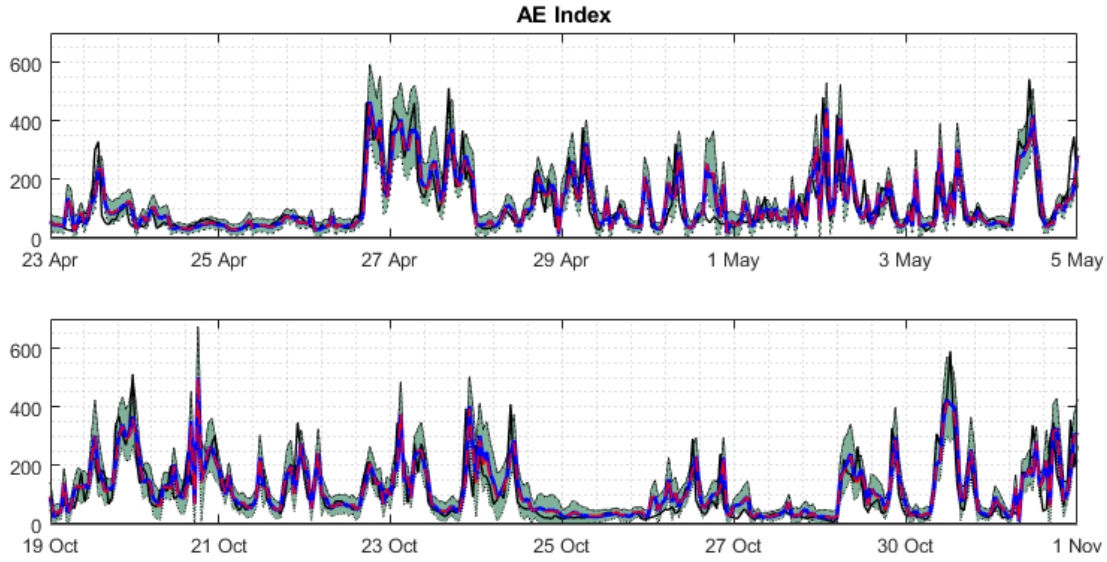

Figure 4.13 One-hour-ahead predicted band and averaged prediction of AE index over 23 Apr ~ 5 May & 19 Oct ~ 1 Nov of 2015 (black line: observation; blue line: averaged prediction; green shadow: predicted band; red line: prediction of convention NARX model)

The model prediction of the cloud-NARX model and the conventional NARX model are consistent in non-disturbed periods. In some disturbed periods, the prediction performance of the cloud-NARX model is better than that of the NARX model. The correlation coefficient and NRMSE of cloud-NARX model in disturbed periods (AE> 1000) are 0.3422 and 0.4454, while the conventional NARX model achieves correlation coefficient and NRMSE of 0.3226 and 0.4518 in the same periods. As discussed earlier, the inclusion of some extra selected model terms in the cloud-NARX model can help improve the model robustness in some severe situations. Therefore, compared to the

conventional NARX model, the cloud-NARX model can better describe the nonlinear effect in the disturbed periods.

In addition, it is easy to generate long-term prediction using the cloud-NARX model. For example, the 3 hours ahead AE index forecast can be achieved by generating 3-step-ahead model predicted output (MPO) with the cloud-NARX model. The correlation coefficient, prediction efficiency and NRMSE of the 3-step-ahead MPO of the cloud-NARX model are 0.8167, 0.6667 and 0.0694, respectively. It is reasonable that the performance of 3 hours head prediction is lower than that of the 1 hour ahead prediction. It is because that at each step of the multiple-step-ahead prediction, the predicted AE index at previous step is used as the model input (as autoregressive variable). Thus, the prediction uncertainty of long-term prediction is increased due to the propagation of the error.

4.5. Conclusion

In this chapter, a new cloud-NARX model was applied to the modelling and forecasting of AE index. Good forecasting results were obtained for 1 hour ahead AE index prediction. The correlation coefficient between averaged prediction and observation is 0.87 and prediction efficiency of 0.81 when benchmarked for the period of 17-21 March 2015 and 22-26 June 2015, which is nearly identical to that produced by the best NARX model. More importantly, the cloud-NARX model is capable to quantify the uncertainty of model structure, model parameter and model prediction.

The advantages of this new model can be summarized as follows. First, the model structure of cloud-NARX model is more robust than that of the conventional NARX model, as the model terms of cloud-NARX model are selected from resampled sub-datasets. Second, the estimated parameters (ex, en and he) of cloud-NARX model can provide more comprehensive information on the model parameter uncertainty. Third, based on cloud forward transformation, the cloud-NARX model can generate the predicted band, which clearly indicates the confidence interval of each predicted AE index. It is extremely important for space weather forecast, because when model becomes unreliable under some severe situations, the biased prediction could cause damages and

losses. With the predicted band, the bias of model prediction can be identified, and the reliability of model can be evaluated.

Chapter 5

MACHINE LEARNING ENHANCED NARMAX MODEL

5.1 Introduction

As the size and complexity of the data largely increase in recent years, the modelling and forecasting of complex nonlinear systems requires more efficient and powerful techniques. In model identification, one of the main objectives is to generate model predicted output that can be relied upon for decision-making, forecasting, etc. Also, it becomes ever more important to develop explainable model structure, to reveal the detailed information of system behaviours. This study introduces a novel machine learning enhanced NARMAX (MLE-NARMAX) model for nonlinear systems identification, which can improve the model prediction performance and provide a transparent and interpretable representation. Case studies on the modelling and forecasting of the appliance energy use and Dst index are presented. The results indicate that the new model generates excellent model prediction and reveals the significant the factors for appliance energy use, for example, humidity in living room and parent room, temperature in laundry room, number of seconds from midnight, etc.

5.2 Limitations of NARMAX model and Neural Network

Neural network was firstly introduced to simulate the way the brain works (Zurada, 1992), and such a model soon became one of the most commonly-used model type for data-

driven modelling task (Zurada, 1992). Note that most traditional neural networks only contains a few (e.g. less than three) hidden layers, whose performance is not always evident among the available data-driven modelling tools.

Compared to conventional neural network, deep neural network is a much more powerful and complicated network (Ciregan, Meier & Schmidhuber, 2012). DNN uses multi-layers network, as shown in Figure 5.1, which can be used to represent a large amount of various systems, linear and nonlinear, static and dynamic. Neural network and its variants, for example, the group method of data handling (GMDH), convolutional neural network (CNN), long short-term memory (LSTM) and DNN have been applied to many modelling problems (Hinton, et al., 2012; Ciregan, Meier & Schmidhuber, 2012).

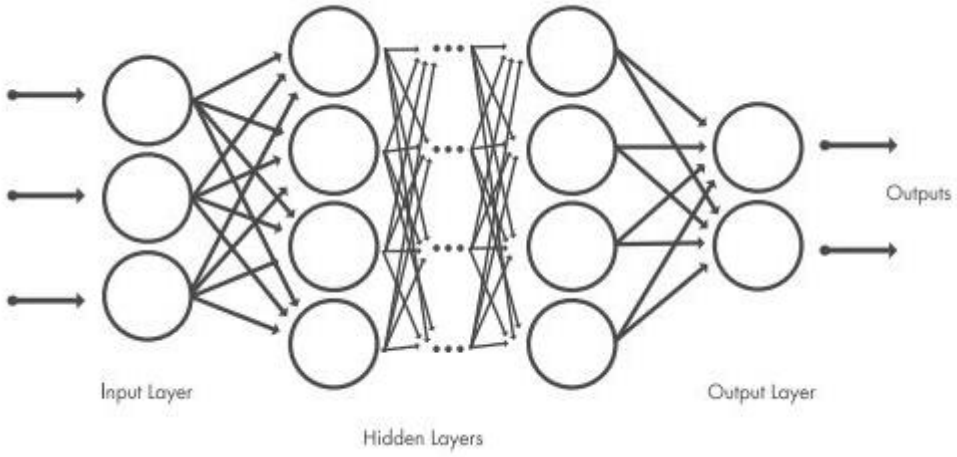

Figure 5.1. Deep neural network

For many modelling and forecasting problems, neural network is capable to describe the behaviours and features of complex systems and can usually achieve excellent model prediction performance. The performance of neural network could become even more powerful when increasing the number of the hidden layers. However, two common issues can arise with neural network. The first issue is the interpretability of the model and the potential overfitting problem. Another issue is that it is known that complicated neural networks usually require a huge amount of computational time. The model structure contains a large number of the connections between each single layer. Improving the training efficiency and convergence capability has always been an important research task for neural network.

Table 5.1 Comparison of the NARMAX model and neural network

	NARMAX	Neural Network
Model Complexity	Only a small number of significant model terms are included in the model	All the available model terms are included in the model
Model Transparency	Transparent linear-in-the-parameters form	Complex network structure
Model Interpretability	Easy to understand; important terms are revealed;	Not interpretable; Cannot know which terms are important or not.
Simulatability	The NARMAX model can be written down, and simulations for a NARMAX model is possible for not only the modeler but also model users.	The NN model is difficult to describe.
Training data size requirement	No special requirement	The number of data points should be much larger than the number of variables.
Prediction Capability	Good prediction performance	Very strong model prediction capability

Compared to neural networks, the nonlinear autoregressive moving average with exogenous (NARMAX) model provides a much simple representation of nonlinear systems (Chen & Billings, 1989; Billings, 2013). It employs an orthogonal forward regression (OFR) algorithm to measure and rank the significance of each candidate model terms, so that the most significant model terms can be selected accordingly (Aguirre & Billings, 1995; Chen, Billings & Luo, 1989; Wei & Billings, 2008; Wei, Billings & Liu,

2004). More importantly, the NARMAX model provides a transparent and parsimonious model structure, which is very useful for understanding and interpreting the system behaviour. The NARMAX model and the OFR algorithm have been successfully applied to solve a wide range of real-world problems in various fields including engineering (Zhang, Zhu, & Gu, 2017), ecological (Marshall et al., 2016), environmental (Bigg et al., 2014), geophysical (Amisigo, et al., 2008; Balikhin et al., 2011; Boynton, Balikhin, Billings, Wei, & Ganushkina, 2011), medical (Billings, Wei, Thomas, LMLE-NARMAXane, & Hope-Gill, 2013), control technology (Tsai, et al., 2010), and neurophysiological (Li, Wei, Billings, & Sarrigiannis, 2016) sciences. A summary of the advantages and limitations of the NARMAX method and the neural network is given in Table 5.1.

In NARMAX model estimation procedure, the moving average model part (i.e. the noise model) is implemented as follows. In each search step, a candidate NARX model is established first, based on which the model error (residual) $\xi(t)$ is calculated and used to estimate an associated candidate NARMAX model. The procedure repeats many times until a NARMAX model with good performance is established. In this study, the estimation of the moving average model part is omitted and replaced by a neural network model.

This study presents a new type of model, which consists of two sub-models, namely, the NARX model component and the neural network model component. The NARX sub-model is established to capture and represent the most important system dynamics in a transparent way, while the neural network sub-model is established to accommodate the error relating to the NARX model. In this way, both the advantages of the NARX model (e.g. transparent, interpretable, simple) and the neural networks (e.g. general strong learning ability) can be well exploited and combined. This is important for many real applications where it requires that the resulting model should be transparent and easy to use to interpret the system behaviour, but in the same time the model should possess excellent prediction performance. The proposed machine learning enhanced NARMAX model is referred to as MLE-NARMAX model.

5.2 MLE-NARMAX Model

This section introduces the novel MLE-NARMAX model and the identification method of MLE-NARMAX model.

5.2.1 Basic Idea

As mentioned earlier, the noise vector $e(t)$ in the NARX model is usually assumed to be independent of any input and output variables. However, in many real data modelling problems, the noise vector $e(t)$ might be correlated with input signal. Consider a nonlinear dynamic single input and single output system as follows:

$$y(t) = -u(t-1)\sqrt{u\,(t-1)} + 0.4u^2(t-1) + 0.8y(t-2)u(t-1) + e(t) \quad (5.1)$$

Assume that the maximum time lags are chosen to be $n_u = n_y = 2$ and the nonlinear degree of the initial full model is 2, the full dictionary of all the candidate model terms is $y(t-1), y(t-2), x(t-1), u(t-2), y(t-1) \times y(t-1), y(t-1) \times y(t-2), y(t-1) \times u(t-1), y(t-1) \times u(t-2), y(t-2) \times y(t-2), y(t-2) \times u(t-1), y(t-2) \times u(t-2), u(t-1) \times u(t-1), u(t-1) \times u(t-2), u(t-2) \times u(t-2)$. Clearly, the true system components $u^2(t-1)$ and $y(t-2)u(t-1)$ are included in the candidate model terms dictiornary but $\sqrt{u\,(t-1)}$ cannot be perfectly represented by any the candidate model terms.

In this case, the polynomial NARX structure cannot perfectly describe the system behaviours. The traditional way to deal with the model residual in to apply the noise modelling process. However, the extra MA components are not useful when generating model prediction, which means that the long term prediction performance of the model can become unreliable due to the unexplained information in the model residual. To overcome this issue, the role of model residual needs to be considered. Based on these considerations, the relationship between the model residual and input signals is considered as a sub-system. This study proposes to use a neural network to characterize the model error relating to the NARX model. Therefore, the final MLE-NARMAX model consists of two sub-models, the NARX sub-model and neural network sub-model.

5.2.2 Identification of the MLE-NARMAX Model

Based on the above considerations, a two-stage identification method is developed, to establish the MLE-NARMAX model. The first stage is to identify the NARX sub-model and the second stage is to use an extra neural network sub-model to fit the model residual of NARX sub-model.

(i). First-stage NARX sub-model

Using the OFR algorithm the first-stage NARX sub-model can be established as:

$$y_{NARX}(t) = \theta_{l_1}\varphi_{l_1}(t) + \cdots + \theta_{l_n}\varphi_{l_n}(t) + e(t) \tag{5.2}$$

where $\varphi_{l_1}(t), \varphi_{l_2}(t), \dots, \varphi_{l_n}(t)$ are the selected model terms and $\theta_{l_1}, \theta_{l_2}, \dots, \theta_{l_n}$ are the estimated parameters. Note that the NARX sub-model is a linear-in-the-parameters representation, where individual model terms are fully transparent, and their contributions are measurable and interpretable. The significant model terms are selected from a pre-specified dictionary and then ranked based on the values of the ERR index. While in most situations, NARX model can provide a good representation of the underlying system dynamics of interest, NARX model might not sufficiently capture all the details of the system. This motivates the use of a neural network sub-model in the second stage to improve the prediction performance.

(ii). Second-stage neural network sub-model

In the second stage, a neural network is used to approximate the model residual of the NARX model. Note that the output (desired signal) of the neural network is the model error, while the inputs of the neural network include not only the original input variables of the NARX model, but also the lagged versions of the model residual variable. The motivation of introducing a neural network model to approximate the model error is to take advantage of neural network approximation capability to accommodate the dependent and correlated relations between the model error and all the candidate explanatory variables that are sufficiently explained by the NARX model.

To avoid any confusion, in this study we use *e*(t) and $\varepsilon(t)$ to represent noise (of a general sense) and model error (residual). The model error of the NARX model (5.2) is:

$$\varepsilon(t) = y(t) - \hat{y}_{NARX}(t) \tag{5.3}$$

The signal ε (t) is used as the desired output signal to train the neural network sub-model of the form:

$$\varepsilon(t) = g[\omega_1(t), \omega_2(t), \dots, \omega_{M'}(t)] \tag{5.4}$$

where $g[\cdot]$ represents the constructed neural network sub-model, and the input vectors $\omega_k(t)$, with $k = 1,2, \dots, M'$, are defined as $y(t-\mathrm{d}), \dots, y(t-n_y), x_1(t-d) \dots, x_1(t-n_u), x_2(t-d), \dots, x_2(t-n_u) \dots, x_r(t-d), \dots, x_r(t-n_u), \varepsilon(t-\mathrm{d})), \dots, \varepsilon(t-n_z)$, where n_ρ is the time lag for the error signal. Note that the neural network sub-model uses all the model terms $\omega_1(t), \omega_2(t), \dots, \omega_{M'}(t)$, meaning that the model structure can be extremely complicated and the modelling process can therefore time-consuming. The applied neural network has one input layer, one hidden layer and one output layer. The number of neurons is 10 and the activation function is sigmoid function.

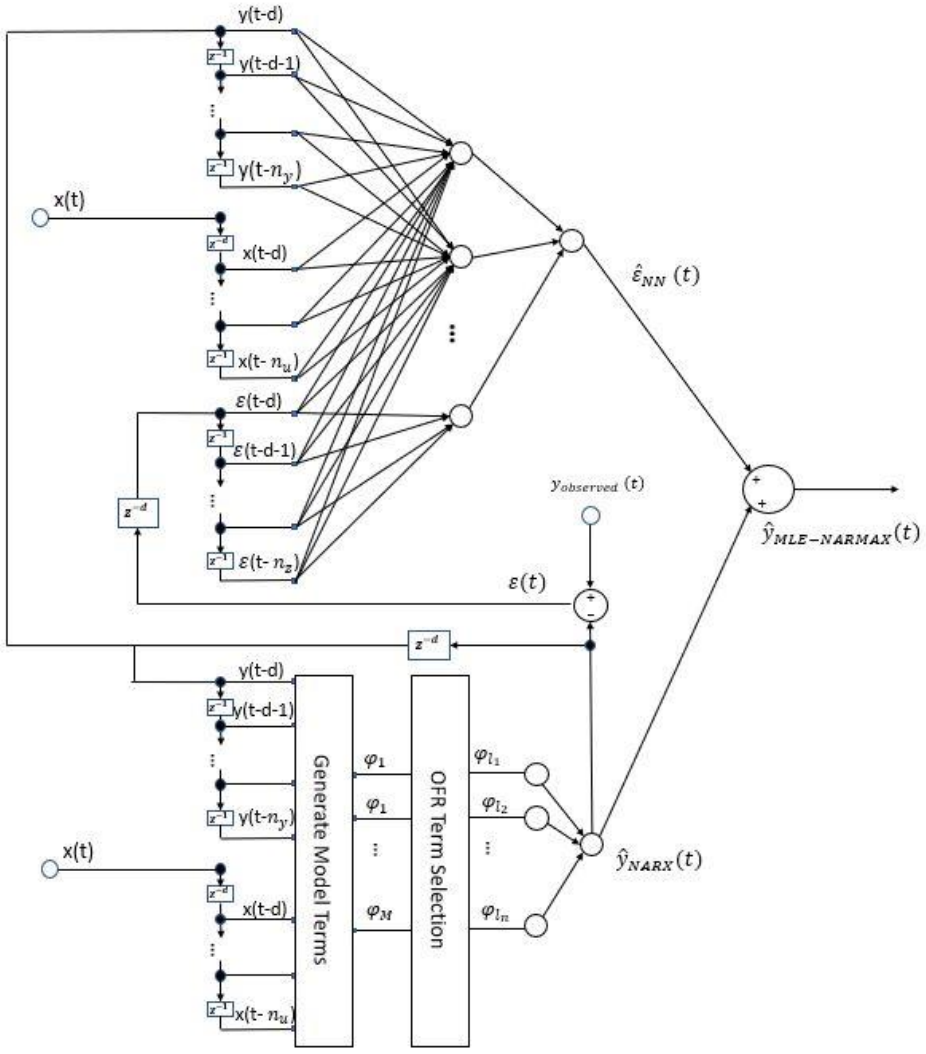

Figure 5.2. The MLE-NARMAX model structure

The general structure of the MLE-NARMAX model is presented in figure 5.2, where the MLE-NARMAX model can be explicitly expressed as:

$$y(t) = \theta_{l_1}\varphi_{l_1}(t) + \theta_{l_2}\varphi_{l_2}(t) + \cdots + \theta_{l_n}\varphi_{l_n}(t) +$$

$$g[\omega_1(t), \omega_2(t), \dots, \omega_M{}'(t)] \quad (5.5)$$

where the $\theta_{l_1}\varphi_{l_1}(t) + \theta_{l_2}\varphi_{l_2}(t) + \cdots + \theta_{l_n}\varphi_{l_n}(t)$ is the NARX sub-model and $g[\omega_1(t), \omega_2(t), \dots, \omega_M{}'(t)]$ is the neural network sub-model. The model prediction of the MLE-NARMAX model can be calculated as:

$$\hat{y}_{MLE-NARMAX}(t) = \hat{y}_{NARX}(t) + \widehat{\varepsilon_{NN}}(t) \quad (5.6)$$

where $\hat{y}_{NARX}(t)$ is the model prediction of NARX sub-model and $\widehat{\varepsilon_{NN}}(t)$ is the model prediction of neural network sub-model.

5.3 Simulation Example

Consider a nonlinear dynamic single input and single output system:

$$y(t) = -u(t-1)\sqrt{u\,(t-1)} + 0.4u^2(t-1) + 0.8y(t-2)u(t-1) + e(t) \quad (5.7)$$

Assume that the maximum time lags are chosen to be $n_u = n_y = 2$ and the nonlinear degree of the initial full model is 2, a number of 14 candidate model terms can be generated. The system component $u(t-1)\sqrt{u\,(t-1)}$ cannot be perfectly described by the model term selected by the OFR algorithm (as shown in Table 5.2).

Table 5.2 Selected model terms by OFR algorithm with associated ERR values and estimated parameters

No.	Model Term	ERR (100%)	Parameter
1	u(t-01) *u(t-01)	36.1871	0.4005
2	u(t-01)	28.1675	-0.2587
3	u(t-01) *y(t-01)	9.6927	-0.5915
4	u(t-01) *y(t-02)	6.1369	0.4808

In this case, the model residual of the NARX model can be further fitted by the neural network sub-model. The second-stage neural network model can be identified. The performance of first-stage NARX model and the new MLE-NARMAX model are shown in Table 5.3.

Table 5.3 Comparison of performances of NARX and MLE-NARMAX models on test dataset

Model Type	Corr	PE	NRMSE
Conventional NARX model	0.8534	0.7281	0.0768
MLE-NARMAX Model	0.8953	0.8015	0.0656

* The algorithm was run for 10 times and the averaged statistics are recorded

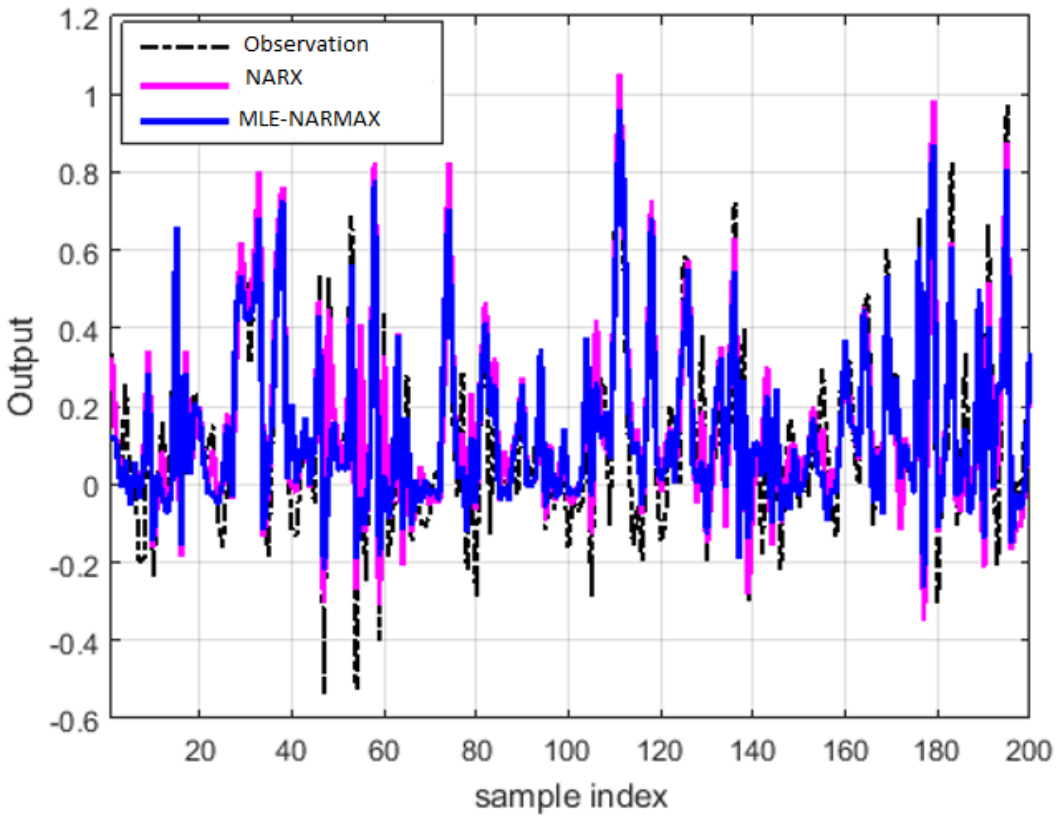

Figure 5.3 Comparison of the observation and prediction of first-stage NARX model and MLE-NARMAX model

A comparison of the model prediction of the NARX, neural network and MLE-NARMAX model is shown in figure 5.3. Note that estimation algorithm was run for 10 times to obtain robust results, as the training of neural network uses a stochastic process. Clearly, the extra second-stage neural network sub-model can improve the model prediction performance. Although the neural network sub-model is not transparent and it is impossible to know what the system is like from the model, the NARX sub-model is interpretable and able to provide the detailed system information. For example, the system components $y(t-2)u(t-1)$ and $u(t-1)u(t-1)$ are revealed and selected in the NARX sub-model. Due to the system noise and uncertainty brought by the 'unknown' component $u(t-1)\sqrt{u\,(t-1)}$, there is bias in the parameter estimation. It is normal as most of the read data comes with strong noise. However, as discussed earlier, the term selection process of OFR algorithm is not affected by noise. Thus, the selected terms in NARX model sub-model is reliable.

5.4 Case Study: Dst Index Forecast

The magnetosphere is a very complex system. To understand the magnetosphere system, the Dst index was developed to measure the magnetic disturbances and it is known to be correlated with a number of solar wind variables (Wei, Billings & Balikhin, 2004; Wei, et al., 2007; Kamide, et al., 1998). In (Wu & Lundstedt, 1996; Wu & Lundstedt, 1997), recurrent neural networks were first proposed for Dst index prediction. Since then, many other neural network models have been introduced for Dst index prediction (Gonazalez, et al., 2004); Amata, et al., 2008; Temerin & Li, 2002; Temerin & Li, 2006). The NARMAX method has also been applied to Dst index forecasting (Boynton, et al., 2011a; Boynton, et al., 2011b). Other methods, for example, wavelets models were also used to forecast Dst index (Wei, Billings & Balikhin, 2004; Wei, et al., 2007). A comparison study of the Dst index forecast models suggests that the neural network by Temerin and Li produces the best predictions when all the events are considered (Ji, et al., 2012). The process of Dst is treated to be a dynamic nonlinear system, where the system inputs are solar wind variables and the system output is the Dst index. The description of the solar wind variables and Dst index is given in Table 5.4. All the variables were sampled every 1 hour. It should be noted that $VBst = V \times B_T \sin^6\left(\frac{\theta}{2}\right)/1000$ is a multiplied input which was suggested to be included in the model inputs (Gonzalez, et al., 1994).

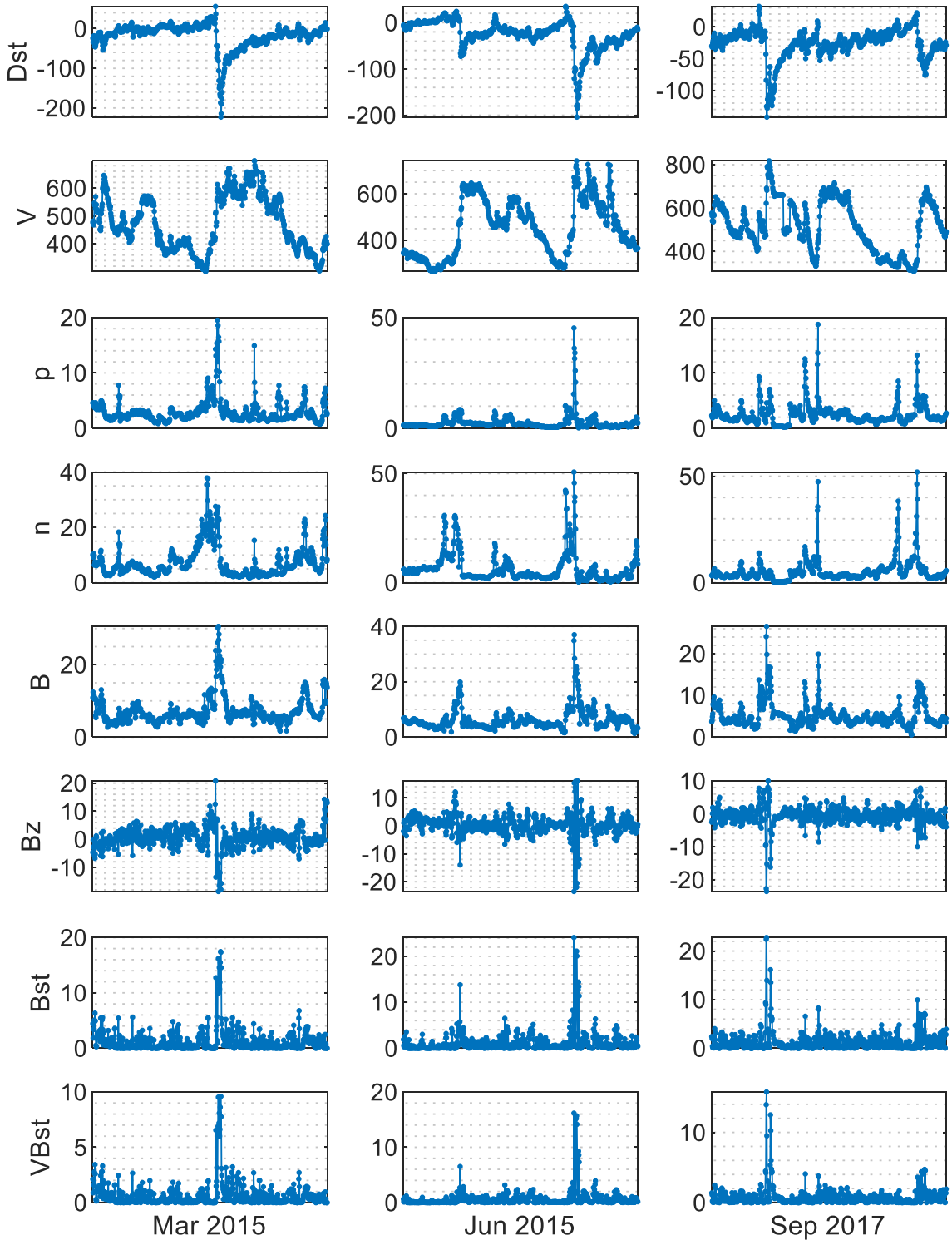

Figure 5.4 Observations of sampled Dst index and solar wind variables of the three test periods.

The Dst and solar wind data of year 2014 were used for training the model. Three time periods of intense storms (Dst<-100nT), Mar 2015, Jun 2015 and Sep 2017 were used to

evaluate the model. The time series of the Dst index and solar wind variables of the three interested periods are shown in figure 5.4. For Dst index forecast, negative peak values are important. From the figure, there are strong storms in these periods. In total, there are around 8700 data points in the training dataset and around 2200 data points in the three test datasets.

Table 5.4 Dst index and solar wind variables

Name	Description
Dst	Dst index [nT]
V	solar wind speed/velocity (flow speed) [km/s]
p	solar wind pressure (flow pressure) [nPa]
n	solar wind density (proton density) [cm^{-3}]
B	interplanetary magnetic field (IMF) [nT]
Bz	the north-south IMF [nT]
Bst	$Bst = B_T \sin^6(\theta/2)$ [nT] [8]

5.4.1 Predict Dst index 3 hours ahead

The 3 hours ahead prediction of Dst can be defined as:

$$Dst(t) = F_{MLE-NARMAX}[Dst(t-3) \dots Dst(t-n_y), V(t-3) \dots V(t-n_u), p(t-3) \dots p(t-n_u), n(t-3) \dots n(t-n_u), B(t-3) \dots B(t-n_u), Bz(t-3) \dots Bz(t-n_u), Bst(t-3) \dots Bst(t-n_u), VBst(t-3) \dots VBst(t-n_u)] \quad (5.8)$$

where $F_{MLE-NARMAX}$ is the MLE-NARMAX framework. To evaluate the prediction of the model, the correlation coefficient, prediction efficiency (PE), and normalized root-mean square error (NRMSE) are calculated.

Table 5.5 Selected model terms of the NARX sub-model

No	Model Term	ERR (100%)	Parameter	t-statistics
1	Dst(t-03)	78.1229	0.8462	103.3129
2	B(t-04) *VBst(t-03)	3.3731	-0.1680	4.6679
3	B(t-04) *VBst(t-04)	0.4205	0.1528	6.1903
4	p(t-03) *p(t-04)	0.3519	-0.2623	16.1041
5	Bz(t-03) *Bst(t-03)	0.2090	0.2002	13.3520
6	p(t-04) *n(t-03)	0.1400	0.0728	11.8165
7	n(t-04) *Dst(t-03)	0.1077	-0.0064	7.0959
8	Bst(t-03)	0.0645	-0.6475	6.9029
9	Bz(t-03) *Bz(t-03)	0.0590	0.0346	8.5650
10	V(t-04) *Bz(t-04)	0.0844	-0.0006	6.5766

5.4.2 The identified MLE-NARMAX model

In order to determine the maximum time lags for both the input and output variables, following the approach described in (Wei, Billings & Liu, 2004) we have carried out pre-modelling experiments and simulations (Wei, Billings & Liu, 2004), the results suggest that the maximum time lags of the input and output were chosen to be $nu = 4$ and $ny = 4$. The initial full model was chosen to be a polynomial form with nonlinear degree of 2.

In the first step, a 10-term bilinear NARX sub-model was identified. The 10 model terms, together with their corresponding ERR values and t-statistics, are shown in Table 5.5. The t-statistics given in the table indicates that all the selected model terms are significant. Note the first-stage bilinear NARX model reported in Table 5.5 can be written as:

$$Dst(t) = 0.8462\, Dst(t-3) - 0.1680\, B(t-4)\, VBst(t-3) + \cdots \quad (10)$$

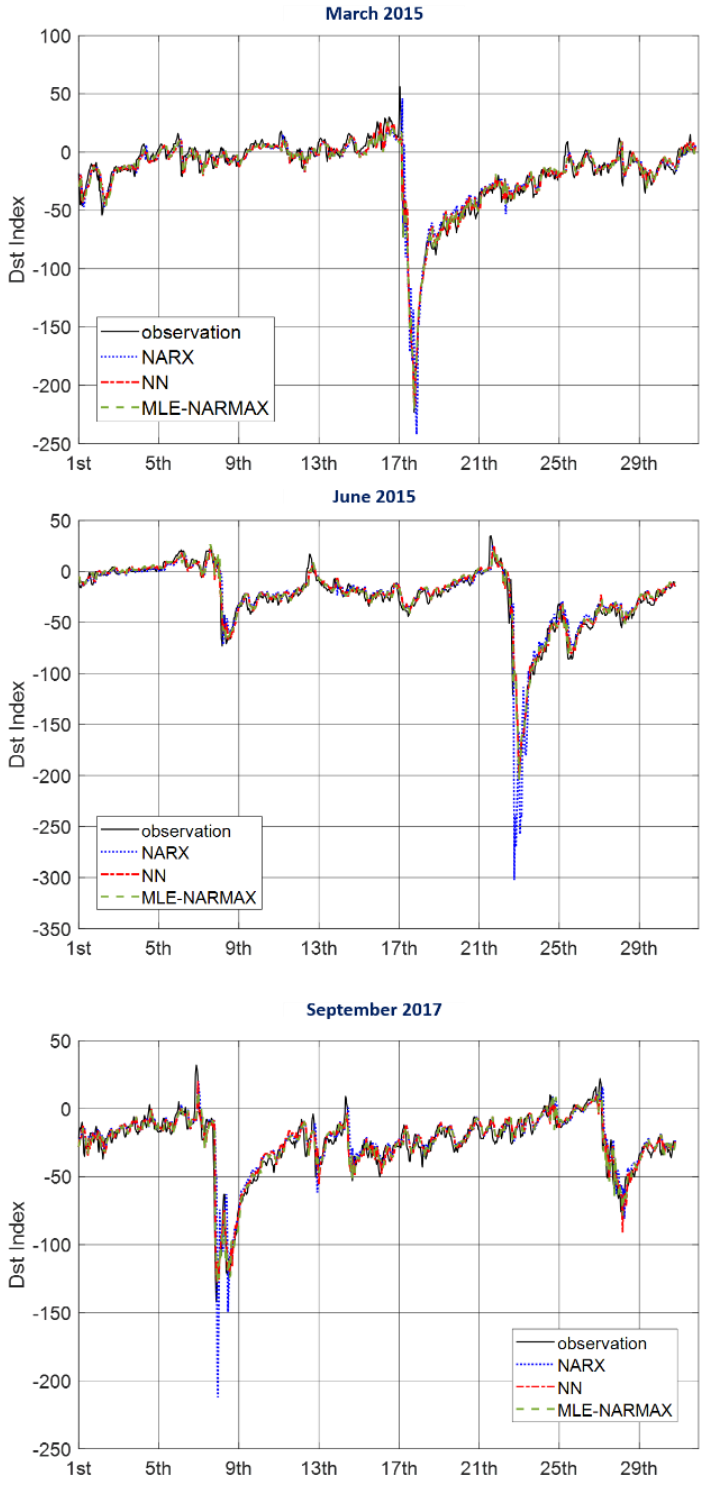

Figure 5.5 Comparison of the predictions of the NARX model, neural network model and MLE-NARMAX model of the three test datasets.

In the second step, the neural network sub-model was estimated to fit the error of NARX sub-model. The estimation algorithm was run for 10 times and the averaged performances were recorded. Then, the final MLE-NARMAX model is obtained with the

NARX sub-model and the neural network sub-model. As the neural network contains too many nodes and connections, the details of the model are not presented here.

Table 5.6 Comparison of the performances of NARX model, neural network and MLE-NARMAX model of the three test periods

Period	Model	Correlation	Prediction	NRMSE
Mar 2015	NARX	0.9502	0.9029	0.0353
	Neural Network*	0.9716	0.9439	0.0269
	MLE-NARMAX*	0.9734	0.9474	0.0260
Jun 2015	NARX	0.8907	0.7368	0.0678
	Neural Network*	0.9599	0.9212	0.0364
	MLE-NARMAX*	0.9598	0.9173	0.0368
Sep 2017	NARX	0.8828	0.7735	0.0642
	Neural Network*	0.9295	0.8487	0.0500
	MLE-NARMAX*	0.9206	0.8333	0.0529

* The algorithm was run for ten times and the averaged statistics are recorded

5.4.3 Performance and advantage of the MLE-NARMAX model

The MLE-NARMAX model was used to generate 3 hours ahead Dst predictions for the three test periods: Mar 2015, Jun 2015 and Sep 2017. Figure 5.5 presents graphical comparisons of the observed and predicted Dst index of the three test periods. The statistical performances of the NARX model, neural network and the MLE-NARMAX model on the three test periods are presented in Table 5.6. From the statistics, the performances of the MLE-NARMAX model are similar to those of the neural networks and better than those of the NARX models for all the three test periods. It can be seen that for the test period of June 2015, the improvement is obviously more significant than those for the other two periods. From Table 5.6 and Figure 5.5, it can be noticed that while the bilinear NARX structure can well capture the features and dynamics of the Dst process at most times of the test periods of June 2015 and September 2017, the model does not sufficiently capture the system dynamics at the severe situation times. The neural network sub-model, however, can help improve the model performance.

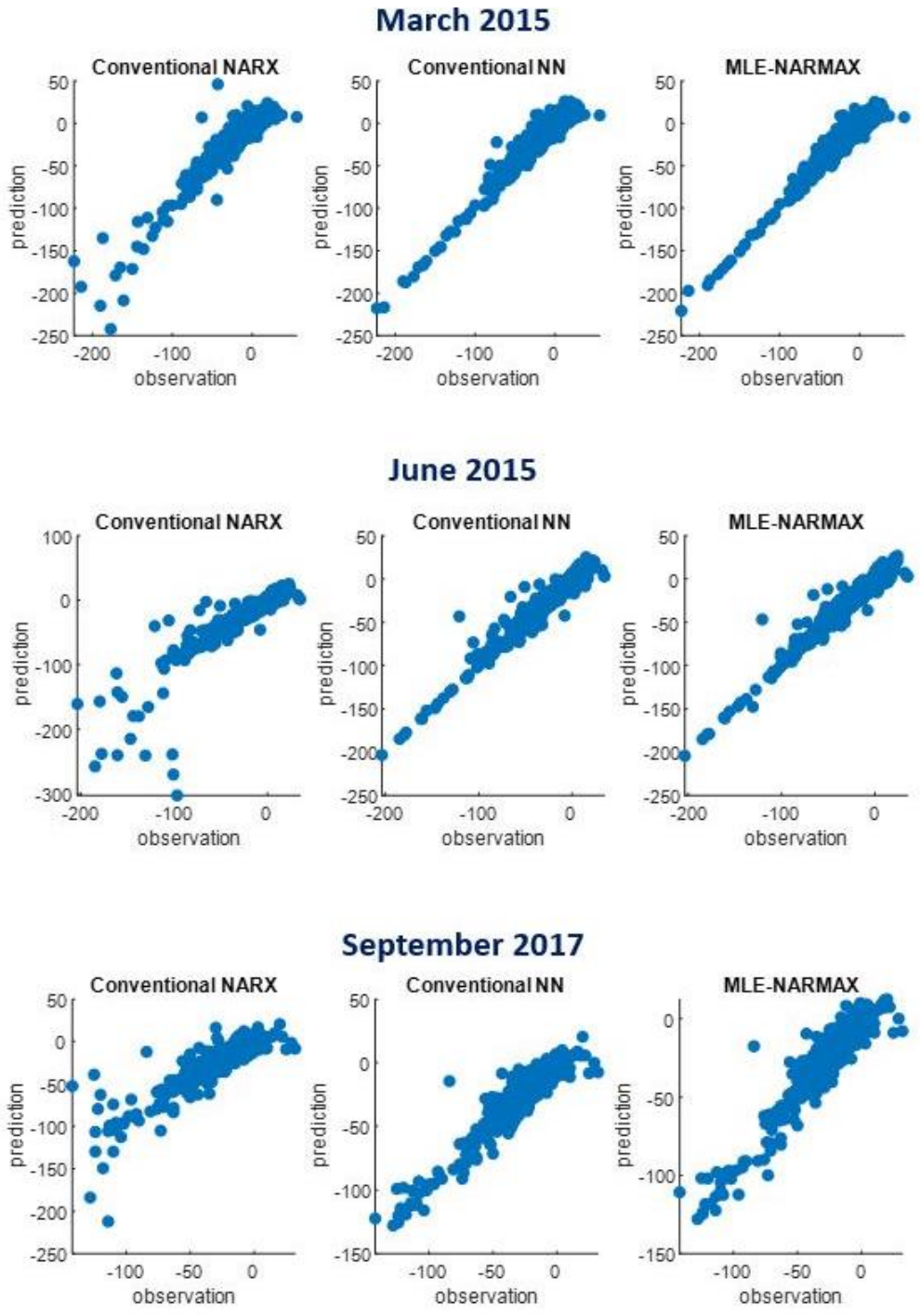

Figure 5.6 Scatter plots of the 3 hours ahead NARX model, neural network model and MLE-NARMAX model of the three test datasets.

Figure 5.6 shows the scatter plots of the NARX model, neural network and MLE-NARMAX model. From these plots, it can be seen that the MLE-NARMAX model produces better predictions for strong storms (Dst< -100nT) than the bilinear NARX

model alone. In other words, the results show that the combination of the NARX sub-model and neural network sub-model can better predict change of the Dst index during strong storm periods.

From the results shown in Table 5.5, the NARX sub-models only consists of 10 significant terms. Obviously, the model provides a parsimonious and transparent representation, where contributions of the selected model terms are clear. However, such a simple bilinear NARX model may not always be sufficient to capture the underlying dynamics of the process, and the model prediction performance may be improved by introducing a sub-model to characterize some dynamics of the system hidden in the model residuals that is not captured by the sub-NARX model. This explains why the two-stage MLE-NARMAX performs better than the NARX model.

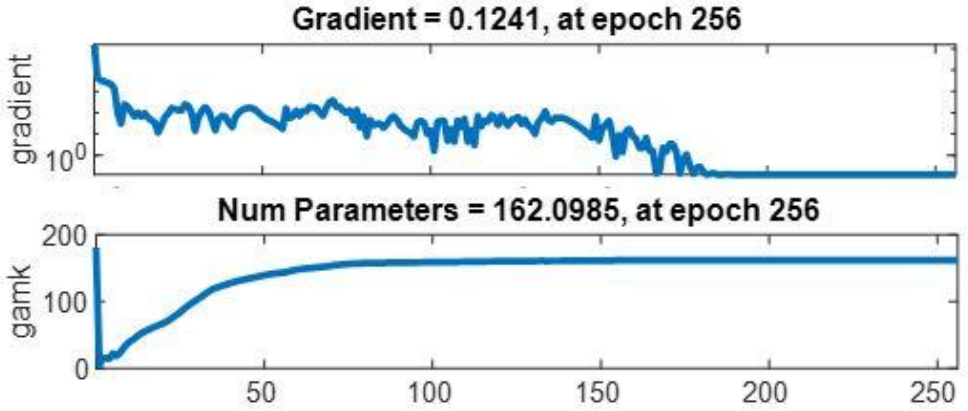

Figure 5.7 Training state of the neural network sub-model of the MLE-NARMAX model.

Note that although the neural network sub-model improves the model performance, the model structure itself cannot be written down. Figure 5.7 shows the training state of one of the neural network sub-models. The two variables 'gradient' and 'gamk' indicate the values of the associated gradient and the effective number of parameters at each iteration, respectively. The figure shows that during the training process of the MLE-NARMAX models, the neural network sub-model contains over 150 parameters. The model complexity of the neural network sub-model is much higher than that of the NARX sub-model. In addition, the neural network sub-model takes many steps to train. For 'big' data modeling problems where the data size is much larger, the training of neural network can take quite long time. On the contrary, the training of the NARX sub-model only takes a

few steps and use relatively much less time. Therefore, the MLE-NARMAX model is developed so that it can take the advantages of both NARX model and neural network model. For example, it provides a transparent representation of complex nonlinear systems, which helps to understand the systems behaviors. Meanwhile, it provides good model prediction performance.

5.5 Case Study: Modelling and Forecasting of Energy Use

The energy use of appliances has received a lot of attention in recent years (Candanedo, Feldheim & Deramaix, 2017). Understanding the relationship between energy use and different potential factors (variables) is very important for many applications, for example, load control of the energy management system (Zhao, Suryanarayanan & Simões, 2013; Barbato, et al., 2011), building performance simulation (Muratori, et al., 2013; Crawley, et al., 2008), control of the energy consumption (Perez-Lombard, Ortiz & Pout, 2008). Different methods have been applied to the analysis of the energy use, including regression models (Candanedo, Feldheim & Deramaix, 2017; Nicoleta, et al., 2012; Candanedo, Dehkordi & Stylianou, 2013), neural networks (Ekici & Aksoy, 2009; Gonzalez & Zamarreno, 2005), machine learning (Li, Bowers & Schnier, 2010; Dong, Cao & Lee, 2005), ensemble modelling (Fan, Xiao & Wang, 2014). As reported in the literature, the energy use of appliances can be explained by many factors, such as humidity and temperature of different areas in the building, weather condition outside the house, number of the seconds from midnight (Candanedo, Feldheim & Deramaix, 2017). Some non-temperature features such as solar radiation were also found to affect the energy use (Fikru & Gautier, 2015). The usage of some energy efficient appliances, programmable thermostats and insulation were correlated with slight increase in energy consumption (Kavousian, Rajagopal & Fischer, 2013). The occupants' behaviour has also been proved to be effective on the energy use (Masoso & Grobler, 2010; Yan, et al., 2015). There are some other factors which were found to be effective, such as socio-enconomic and dwelling factors (Jones, Fuertes & Lomas, 2015).

5.5.1 Data and Variable Description

The appliance energy use data used in this study is at 10 min for about 4.5 months. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. Each wireless node transmitted the temperature and humidity conditions around 3.3 min and the wireless data was averaged for 10 minutes periods. The energy data was logged every 10 minutes with m-bus energy meters (Candanedo, Feldheim & Deramax, 2017). Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru), and merged together with the experimental data sets using the date and time column. Two random variables have been included in the data set for testing the regression models and to filter out non predictive attributes (parameters). The descriptions of all the variables are given in Table 5.7 (Candanedo, Feldheim & Deramax, 2017). The appliances energy use is considered as the output variable and the other features are considered as the input variables. The number of seconds from midnight is a derived variable.

Table 5.7 Descriptions of variables

Variables	No.	Description
Appliances	y	energy use in Wh
lights	u1	energy use of light fixtures in the house in Wh
T1	u2	Temperature in kitchen area, in °C
RH_1	u3	Humidity in kitchen area, in %
T2	u4	Temperature in living room area, in °C
RH_2	u5	Humidity in living room area, in %
T3	u6	Temperature in laundry room area in °C
RH_3	u7	Humidity in laundry room area, in %
T4	u8	Temperature in office room, in °C
RH_4	u9	Humidity in office room, in %

T5	u10	Temperature in bathroom, in °C
RH_5	u11	Humidity in bathroom, in %
T6	u12	Temperature outside the building (north side), in °C
RH_6	u13	Humidity outside the building (north side), in %
T7	u14	Temperature in ironing room, in °C
RH_7	u15	Humidity in ironing room, in %
T8	u16	Temperature in teenager room 2, in °C
RH_8	u17	Humidity in teenager room 2, in %
T9	u18	Temperature in parents room, in °C
RH_9	u19	Humidity in parents room, in %
Tout	u20	Temperature outside (from Chievres weather station), in °C
Pressure	u21	Pressure (from Chievres weather station), in mm Hg
RH_out	u22	Humidity outside (from Chievres weather station), in %
Wind Speed	u23	Wind speed (from Chievres weather station), in m/s
Visibility	u24	Visibility (from Chievres weather station), in km
Tdewpoint	u25	Tdewpoint (from Chievres weather station), in °C
rv1	u26	Random variable 1, nondimensional
rv2	u27	Random variable 2, nondimensional
NSM	u28	number of seconds from midnight, in s

The appliances energy use and the variables in Table 1 are sampled every 10 minutes, from 17:00 of 11 Jan to 18:00 of 27 May of 2016. There are a total number of 19736 observed data points. The first 75% of the data is used for model training and the remaining 25% of the data is used for model testing. Figure 5.8 shows the full time series of the appliances energy use and Figure 5.9 gives an example of the appliances energy use of a representative week. It can be seen that the midday and the evening are two

periods when the appliances energy use increases significantly. The daily appliances energy use of the weekends is lower than that of the weekdays.

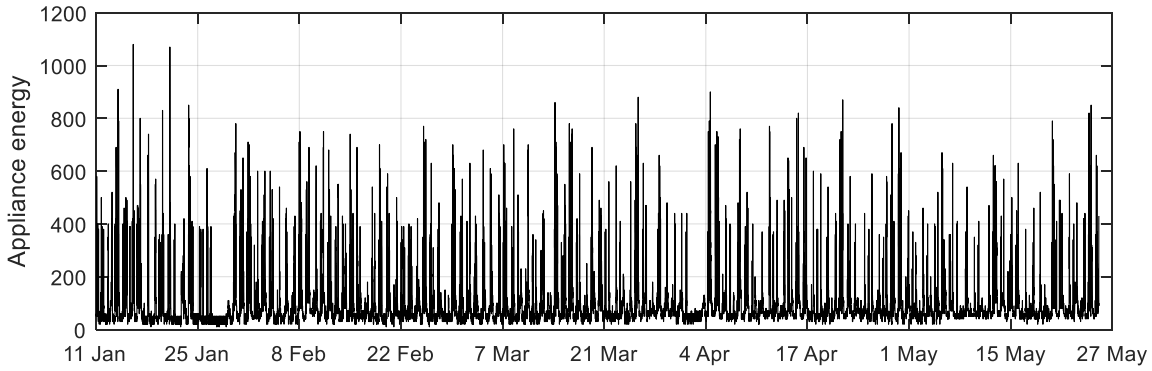

Figure 5.8 Observed appliances energy use from 1 Jan to 27 May

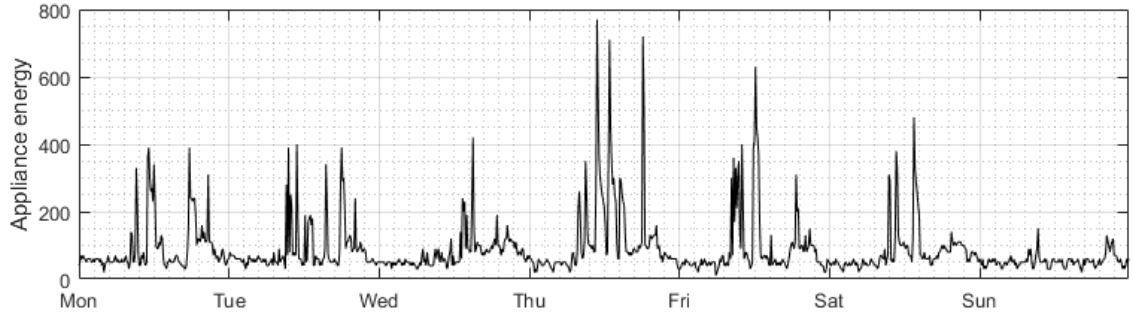

Figure 5.9 Observed appliances energy use in a representative week (From Monday 22 Feb to Sunday 28 Feb)

5.5.2 Model Construction

In this case study, the MLE-NARMAX model was employed to predict energy use 10 minutes ahead. As discussed earlier, the construction of the MLE-NARMAX model contains two steps. At the first step, the NARX sub-model is identified by the OFR algorithm; at the second step, the neural network sub-model is identified using the model residual (noise sequence) of the first-stage NARX sub-model.

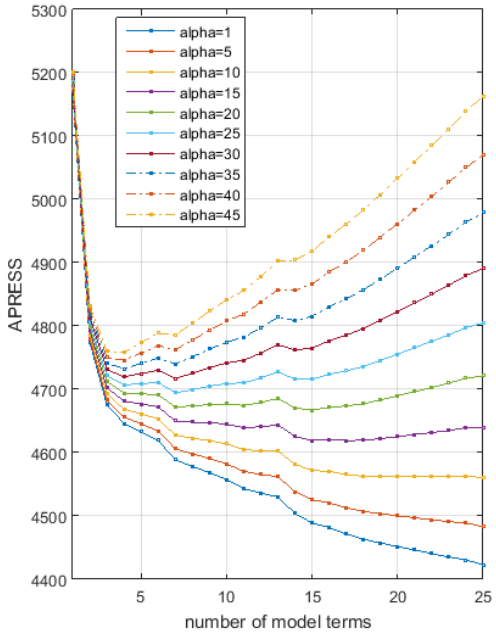

Figure 5.10 Number of selected model terms versus APRESS value (alpha is a tuning parameter)

i) First-stage NARX sub-model

For the NARX sub-model, the maximum time lag of the input and output variables are chosen to be $nu = 2$ and $ny = 2$. The candidate variable vector for model construction is:

$$\vartheta(t) = [y(t-1), y(t-2), u_m(t), u_m(t-1), u_m(t-2)]^T$$

where $u_m(t), \dots, u_m(t-2)$ $(m = 1, 2, \dots, 28)$ are the input variables (as listed in Table 1) and $y(t-1), y(t-2)$ are the autoregressive terms. The initial full model was chosen to be a polynomial form with nonlinear degree of $l = 2$. There are a total number of 1770 of candidate model terms, including the first-order terms, the second-order terms and the constant term. The number of terms that should be included in the model is determined by the APRESS criterion (Billings & Wei, 2008). As shown in Figure 5.10, there is turning points at 4, 7 and 14, which indicates that the optimal number of model terms can be 4, 7 or 14. According to the results of pre-modelling experiments and simulations, a number of 7 model terms are selected to construct the NARX model.

The 7 selected model terms, their associated ERR values and estimated parameters are given in Table 5.8. The importance of these selected model terms are quantified and ranked by the ERR index. The NARX sub-model in Table 2 should read as:

$$y(t) = 0.9553y(t-1) - 0.0004y(t-1)y(t-2) + \cdots \quad (5.9)$$

Table 5.8 Selected model terms by OFR algorithm with associated ERR values and estimated parameters of NARX sub-model

No.	Model Term	ERR (100%)	Parameter	t-statistics
1	y(t-1)	75.5984	9.5526e-01	6.1719e+01
2	y(t-1) *y(t-2)	1.8692	-3.8010e-04	1.7518e+01
3	y(t-1) *y(t-1)	0.4692	-1.4239e-04	5.3335e+00
4	u1(t) *u19(t-2)	0.1441	1.2398e-01	6.9934e+00
5	u1(t) *u5(t)	0.0656	-1.1020e-01	6.0755e+00
6	u6(t) *u28(t)	0.0660	3.3825e-05	1.1715e+01
7	u28(t) *u28(t-2)	0.1480	-7.8836e-09	1.0058e+01

ii) Second-stage neural network sub-model

The model residual of the first-stage NARX sub-model is calculated as:

$$z(t) = y(t) - \hat{y}(t) = y(t) - [0.9553\hat{y}(t-1) - 0.0004\hat{y}(t-1)\hat{y}(t-2) + \cdots] \quad (5.9)$$

The model residual $z(t)$ is considered as the output layer of neural network. The variables $u1 \ldots . u28$ are used as the input layers. The neural network is established by running the algorithm for 10 times and the averaged performances are recorded.

Because that the neural network structure is not transparent, the information of the significance of the input variables can't be known and. Therefore, the neural network model is only useful for generating model prediction, but not for understanding the system

behaviours. Assume that the predicted output signal of the neural network can be described as $\hat{z}(t)$, the model prediction of the final MLE-NARMAX model is:

$$\hat{y}_{INN}(t) = \hat{y}(t) + \hat{z}(t) = 0.9553\hat{y}(t-1) + \cdots + \hat{z}(t) \tag{5.10}$$

5.5.3 Model Performance

The statistics of prediction performances of the first-stage NARX sub-model, the conventional neural network model and the MLE-NARMAX model are shown in Table 5.9. Figure 5.11 shows the scatter plot of the three models and Figure 5.12 shows the comparison of the observed the predicted energy use of MLE-NARMAX model. The correlation coefficient, prediction efficiency and NRMSE of the NARX sub-model on test dataset are 0.7494, 0.5606 and 0.707, respectively. The overall correlation coefficient of the MLE-NARMAX model is 0.7804. The prediction efficiency is about 0.6078 and the NRMSE is about 0.0665. It can be seen that the model prediction is improved by the extra neural network sub-model. From the results, the MLE-NARMAX model and conventional neural network model outperform the conventional NARX model. The performance of MLE-NARMAX model is slightly better than that of the conventional neural network model.

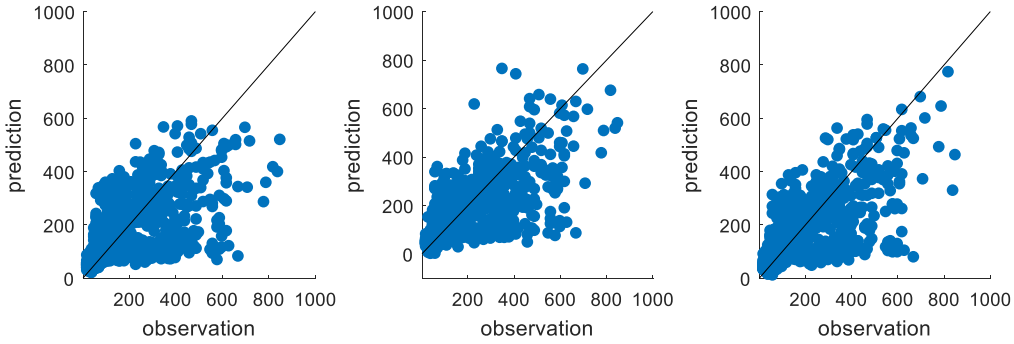

Figure 5.11 Scatter plot of observed and predicted appliance energy use (left: First-stage NARX model, middle: Neural Network, right: MLE-NARMAX Model)

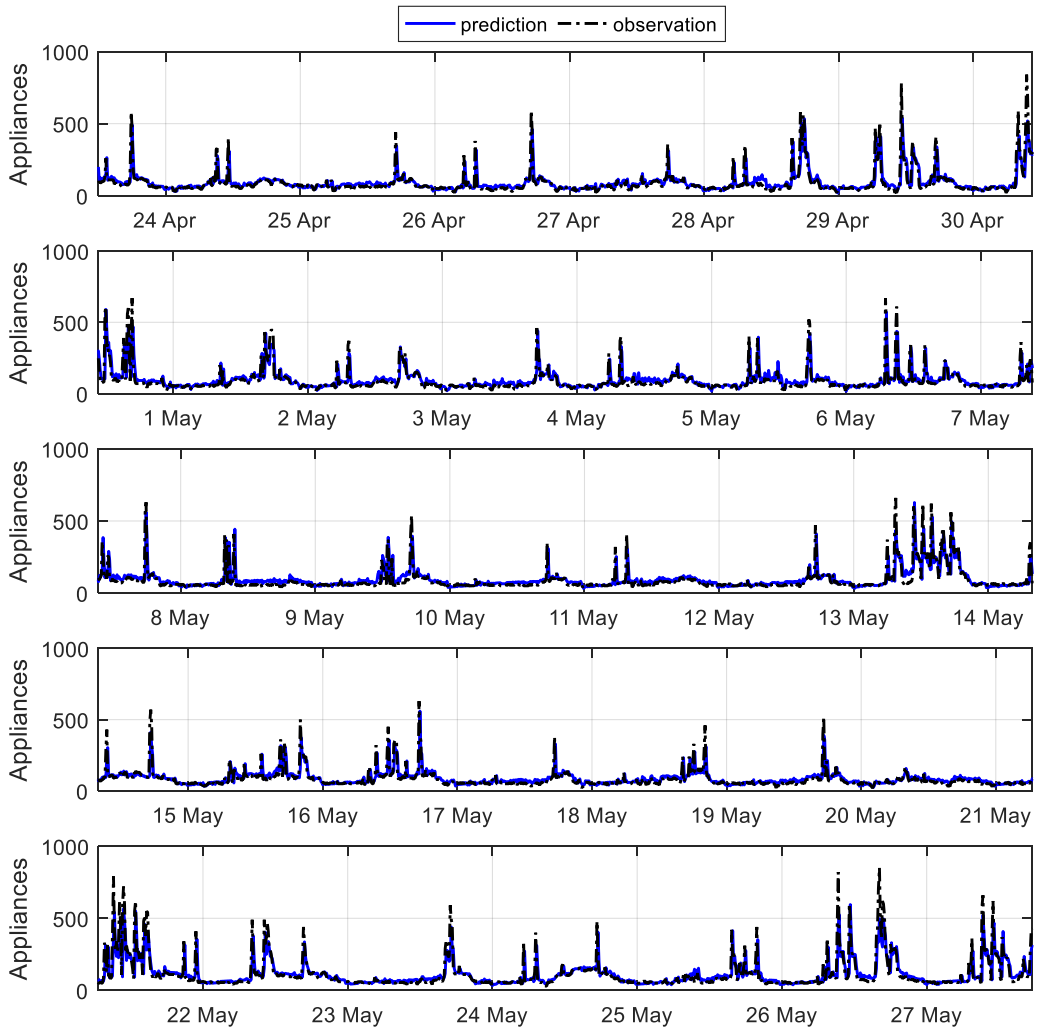

Figure 5.12 Comparison of observed and predicated appliances energy use of MLE-NARMAX model

Table 5.9 Comparison of performances of three models on test dataset

Model Type	Correlation	PE	NRMSE
First-stage NARX model	0.7494	0.5606	0.0707
Neural Network*	0.7790	0.6054	0.0667
MLE-NARMAX Model*	0.7804	0.6078	0.0665

* The algorithm was run for 10 times and the averaged statistics are recorded

5.5.4 Discussion

Our results indicated that the appliances energy use can be better explained by the new MLE-NARMAX method in comparison to conventional NARX and neural network model. The model involving the autoregressive terms and dynamic components reflects the close-loop and dynamic features of the systems. As shown in Table 5.8, all the model terms of first-stage NARX sub-model are dynamic with specific time lags and the model structure of is fully transparent. The significant model terms are picked out from a huge number of candidate terms, which largely reduces the time and cost for data collection and investigation. In general, the first-stage NARX sub-model provides a parsimonious and interpretable representation, which is able to describe the majority of the variance of system output (PE=56%). The second-stage neural network sub-model is developed to further improve the model performance. According to the results shown in Table 5.9, the prediction performance of MLE-NARMAX model is improved by the extra neural network sub-model for around 5%. It is because that the unexplained information in the model residual of the first-stage NARX sub-model can be further described by the neural network sub-model structure.

The selected model terms indicate that the appliances energy use is highly correlated with the house and weather conditions at current time and 10/20 minutes earlier. The results from our study show that, the following factors appear to have significantly impact on the appliance energy use. The previous appliance energy use is extremely significant in the identified model as the first three selected model terms y(t-1), y(t-1) *y(t-2), y(t-1) *y(t-1) are all autoregressive terms, which indicate that the appliance energy use is highly correlated to its history value. The appearance of u1 in the model terms u1(t) *u19(t-2) and u1(t) *u5(t) indicates that large amount of appliance energy use comes from the lights fixtures. The humidity in the living room is also found to be an important factor due to the model term u1(t) *u5(t). The selected model term u6(t) *u28(t) and u1(t) *u19(t-2) also show the significance of the temperature in the laundry room and the humidity in parent's room. The last 3 model terms u6(t) *u28(t), u28(t) *u28(t-2) all consists of the variable u28, which is the number of seconds from the midnights. Clearly, the appliance energy use is highly related to the time period of the day. For the role of energy use of light fixture, temperature in laundry room and number of seconds from midnights, our finding re-confirms the conclusion of previous studies by Candanedo, Feldheim & Deramaix (2008). Our method do not select any weather-related model terms. The reason

might be that for different types of residential building and different seasons, the appliance energy use is not always sensitive to weather changes, which are supported by Fikru & Gautier (2015).

It should be noted that many previous study on the modelling of appliance energy use mainly focus on achieving high performance of the predictor. The previous neural network models cannot provide information on which of the factors are significant and which are not (Gonzalez & Zamarreno, 2005; Ekici & Aksoy, 2009). Although some interpretable model reveals how the appliance energy use replies on the input variables, the prediction performance is lower than that of the neural network (Candanedo, Feldheim & Deramaix, 2008). Rather following the literature, this study advocates to use a new data-driven modelling approach, to identify the most important variables from a huge number of candidate variables by using NARMAX method and improve the model prediction by using an extra neural network sub-model. It is known that the size and complexity of the data is increasing rapidly, there is an increasing demand for quantitative methods for automatic identification of important variables. In this sense, the proposed method provides an effective automatic tool which can save data analysis cost and time and meanwhile produces high prediction performance.

5.6 Conclusion

In this paper, a new MLE-NARMAX model method is proposed. Benefitted from the two-stage modelling process, the MLE-NARMAX model uses an hybrid model structure of NARMAX model and neural network model, to provide interpretable system information and strong model prediction ability. The new MLE-NARMAX model was applied to the modelling and forecasting of appliance energy use. The correlation coefficient between 10 minutes ahead prediction and observation is 0.78 and the prediction efficiency is 0.60, which is nearly identical to that produced by the best neural network model. The MLE-NARMAX method is used for 3 hours ahead prediction of the Dst index. Three periods with typical strong storms were used to test the model performance. The MLE-NARMAX model outperforms the conventional NARX model in terms of correlation coefficient and prediction efficiency. More importantly, the MLE-NARMAX model is capable to provide an interpretable representation of the system, which can reveal the most significant model terms and, in the meantime, show good

generalization properties. For many real data modelling problems, where the central modelling task and objective is not is not only for prediction but also for understanding and explaining the input-output behaviour or cause-effect relationships of the systems, the proposed MLE-NARMAX model is a good choice

For future work, we intend to further develop the neural network sub-model by employing deep learning methods, to improve the prediction performance of the MLE-NARMAX model. The MLE-NARMAX model uses neural network to enhance the NARMAX model. Similarly, other machine learning techniques can be cooperated with the NARMAX method in the same way. Gradient boosting method (GBM), lasso method, and support vector machine (SVM) can be used to model the residual. Further research can be conducted to investigate if these machine learning methods can be combined with the NARMAX method.

Chapter 6

CONCLUSIONS

6.1 Summary and Conclusions

This thesis focuses on developing new approaches for nonlinear dynamic system identification and data modelling, to overcome the negative effect caused by the uncertainty. Three new approaches, namely RMSS method, cloud-NARX model and MLE-NARMAX model have been proposed for data modelling problems with different objectives. The developed methods have been evaluated via simulations and applied to several real data modelling problems, for example, EEG, space weather, energy, etc.

First, the RMSS method is developed to deal with the model structure detection problems with small size data and multi datasets. The RMSS method uses a resampling process and a new oMAE metric to select the important model terms from a series of sub-datasets, to overcome the issue that the change of a single data pair in small size data might bring strong uncertainty to the model structure. In this way, a robust model structure that is robust to all the data points can be identified. In addition, the RMSS method can be directly applied to multi-datasets modelling problems.

Several simulation case studies and two real data case studies are carried out to illustrate the advantages of the RMSS method. In one of the real data case studies, the RMSS method is applied on the modelling and forecasting of Kp index. From this small size data modelling problems, the RMSS method produces more robust model than the conventional NARX and neural network model. In another real data case study, the RMSS

method is applied to a multi-dataset modelling problem, which is the modelling and forecasting of cortical response to mechanical wrist perturbation. There are 10 participants for the collection of EEG data so there are a total number of 10 sub-datasets. The RMSS method establish a common model structure which is robust to all the sub-datasets.

Second, the cloud-NARX model is proposed for uncertainty analysis of the nonlinear dynamic system identification. The cloud-NARX model uses an uncertainty concept, cloud model, to describe and quantify the uncertainty during the modelling process. Benefitted from generic forward and backward cloud transformation, the cloud-NARX model can store the information of the model uncertainty with only three parameters when the model is established and provide visualized information of the model uncertainty with a confidence interval when generating model predictions. The model reliability can be revealed and described using the new model predicted band/surface. This property is useful for detecting strong disturbances in some unstable systems, for example, the space weather.

The cloud-NARX model is firstly evaluated by some simulation examples. Then, the cloud-NARX model is applied to the modelling and forecasting of AE index. The results show that the strong uncertainty caused by the magnetic storm can be detected by the cloud-NARX model. In addition, the cloud-NARX generated excellent 1 hour ahead prediction for AE index.

Third, a novel MLE-NARMAX model for system identification and data modelling is developed. By taking advantages of neural network and NARMAX model, the proposed interpretable model cannot only provide good forecast result. More importantly, the resulting model is established based on an interpretable NARMAX model structure, which is composed of the most important candidate features (variables), it can clearly indicates how the system output depends on these variables. The proposed model provides a new way for data modelling problems through machine learning approach with a simple/sparse, interpretable and transparent model structure.

The proposed method is evaluated via a simulation example and two case studies. In the first case study, the presented a novel MLE-NARMAX is used to predict appliance energy use 10 minutes ahead and achieve good forecasting results in terms of two prediction skills: correlation coefficient of 0.78 and prediction efficiency of 0.61. In

second case study, the MLE-NARMAX is used to predict Dst index 3 hours ahead. The new model outperforms the conventional NARX and neural network model, and also reduces the time cost for the identification process.

In conclusion, the proposed methods provide some solutions for some challenging questions of data modelling and systems identification. The negative effect of the model uncertainty can be reduced or quantified by the proposed novel methods. The applications to data driven modeling and analysis of space weather, energy, social science shows the abilities to establish robust model structure, quantify uncertainty and improve model performance with interpretable model structure.

6.2 Future Work

The proposed novel methods perform the systems identification and data modelling, combined with uncertainty analysis and machine learning. The thesis has laid a framework for such data driven modelling and analysis questions, but further extensions and new directions of research can take this further, which are outlined below.

- The data resampling process is very important for the RMSS method, given that the resample method defines the differences of the sub-datasets and the efficiency of the modelling process. Nevertheless, there is still no systematic approach to determine which resampling method is optimal. Several resampling methods have been proposed but further research is required.

- According to the results of AE index modelling study, the cloud-NARX describe the change of the system (magnetic disturbances) by the predicted band/surface. However, there is no metric to measure the uncertainty brought by these changes. Thus, further research is needed to develop a measure to solve this issue.

- For imbalanced data (for example AE data), the system dynamics is time-variant. In these situations, a single model might be insufficient to describe the system behaviors in different conditions. A hybrid model is needed for the systems which has several different statuses.

- The MLE-NARMAX model can be further improved by employing deep learning to replace the conventional neural network. One challenge is that the training process of deep neural network needs a lot of time and computation resources. NARMAX method and deep learning on the same programing platform.

- The MLE-NARMAX provides a promising framework for combining machine learning techniques and NARMAX method. Thus, it is essential to investigate how the other machine learning methods such as classification, clustering, etc can be combined with the conventional NARMAX method.

BIBLIOGRAPHY

Aguirre, L. A., & Billings, S. A. (1995). Improved structure selection for nonlinear models based on term clustering. International Journal of Control, 62(3), 569-587.

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. Ecology, 95(3), 631-636.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactionson Automatic Control, 19(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle BT - Selected Papers of Hirotugu Akaike. Second International Symposium on Information Theory, 199–213.

Akinola, T. E., Oko, E., Gu, Y., Wei, H. L. & Wang, M. (2019) Non-linear system identification of solvent-based post-combustion CO2 capture process. Fuel. 239, 1213-1223.

Ala-Lahti, M. M., Kilpua, E. K. J., DImmock, A. P., Osmane, A., Pulkkinen, T., & Souček, J. (2018). Statistical analysis of mirror mode waves in sheath regions driven by interplanetary coronal mass ejection. Annales Geophysicae, 36(3), 793–808. https://doi.org/10.5194/angeo-36-793-2018

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. Technometrics, 16(1), 125-127.

Amata, E., Pallocchia, G., Consolini, G., Marcucci, M. F., & Bertello, I. (2008). Comparison between three algorithms for Dst predictions over the 2003–2005 period. Journal of Atmospheric and Solar-Terrestrial Physics, 70(2-4), 496-502.

Arendt, P. D., Apley, D. W., & Chen, W. (2012). Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability. Journal of Mechanical Design, 134(10), 100908. https://doi.org/10.1115/1.4007390

Arghira, N., Hawarah, L., Ploix, S., & Jacomino, M. (2012). Prediction of appliances energy use in smart homes. Energy, 48(1), 128-134.

Asatryan, Z., & Feld, L. P. (2015). Revisiting the link between growth and federalism: A Bayesian model averaging approach. Journal of Comparative Economics, 43(3), 772–781. https://doi.org/10.1016/j.jce.2014.04.005

Ayala Solares, J. R., Wei, H. L., & Billings, S. A. (2019). A novel logistic-NARX model as a classifier for dynamic binary classification. Neural Computing and Applications, 31(1), 11-25.

Ayala Solares, J. R., Wei, H. L., Boynton, R. J., Walker, S. N., & Billings, S. A. (2016). Modeling and prediction of global magnetic disturbance in near-Earth space: A case study for Kp index using NARX models. Space Weather, 14(10), 899–916. https://doi.org/10.1002/2016SW001463

Bala, R., & Reiff, P. (2012). Improvements in short-term forecasting of geomagnetic activity. Space Weather, 10(6). https://doi.org/10.1029/2012SW000779

Balikhin, M. A., Boynton, R. J., Billings, S. A., Gedalin, M., Ganushkina, N., Coca, D., & Wei, H. (2010). Data based quest for solar wind-magnetosphere coupling function. Geophysical Research Letters, 37(24). https://doi.org/10.1029/2010GL045733

Balikhin, M. A., Boynton, R. J., Walker, S. N., Borovsky, J. E., Billings, S. A., & Wei, H. L. (2011). Using the NARMAX approach to model the evolution of energetic electrons fluxes at geostationary orbit. Geophysical Research Letters, 38(18), 1–5. https://doi.org/10.1029/2011GL048980

Ball, R., & Chernova, K. (2008). Absolute income, relative income, and happiness. Social Indicators Research, 88(3), 497–529. https://doi.org/10.1007/s11205-007-9217-0

Barbato, A., Capone, A., Rodolfi, M., & Tagliaferri, D. (2011, October). Forecasting the usage of household appliances through power meter sensors for demand management in the smart grid. In IEEE International Conference on Smart Grid Communications, 404-409.

Barford, L. A., Fazzio, R. S., & Smith, D. R. (1992). An introduction to wavelets. Hewlett-Packard Laboratories, Technical Publications Department.

Bartels, J. (1949). The standardized index, Ks, and the planetary index, Kp. IATME Bull. 12b (97).

Berg, A. I., Hassing, L. B., McClearn, G. E., & Johansson, B. (2006). What matters for life satisfaction in the oldest-old? Aging and Mental Health, 10(3), 257–264. https://doi.org/10.1080/13607860500409435

Berg, A. I., Hassing, L. B., Nilsson, S. E., & Johansson, B. (2008). "As long as I'm in good health". The relationship between medical diagnoses and life satisfaction in the oldest-old. Aging Clinical and Experimental Research, 21(4–5), 307–313. https://doi.org/6688 [pii]

Berg, A. I., Hassing, L. B., Thorvaldsson, V., & Johansson, B. (2011). Personality and personal control make a difference for life satisfaction in the oldest-old: Findings in a longitudinal population-based study of individuals 80 and older. European Journal of Ageing, 8(1), 13–20. https://doi.org/10.1007/s10433-011-0181-9

Berg, A. I., Hoffman, L., Hassing, L. B., McClearn, G. E., & Johansson, B. (2009). What matters, and what matters most, for change in life satisfaction in the oldest-old? A study over 6 years among individuals 80+. Aging and Mental Health, 13(2), 191–201. https://doi.org/10.1080/13607860802342227

Bigg, G. R., Wei, H. L., Wilton, D. J., Zhao, Y., Billings, S. A., Hanna, E., & Kadirkamanathan, V. (2014). A century of variation in the dependence of Greenland iceberg calving on ice sheet surface mass balance and regional climate change. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 470(2166), 20130662. https://doi.org/10.1098/rspa.2013.0662

Billings, C. G., Wei, H. L., Thomas, P., Linnane, S. J., & Hope-Gill, B. D. (2013). The prediction of in-flight hypoxaemia using non-linear equations. Respiratory medicine, 107(6), 841-847.

Billings, S. A., & Wei, H. L. (2005a). The wavelet-NARMAX representation: A hybrid model structure combining polynomial models with multiresolution wavelet

decompositions. International Journal of Systems Science, 36(3), 137–152. https://doi.org/10.1080/00207720512331338120

Billings, S. A., & Wei, H. L. (2005b). A new class of wavelet networks for nonlinear system identification. IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council, 16(4), 862–874. https://doi.org/10.1109/TNN.2005.849842

Billings, S. A., & Wei, H. L. (2008). An adaptive orthogonal search algorithm for model subset selection and non-linear system identification. International Journal of Control, 81(5), 714–724. https://doi.org/10.1080/00207170701216311

Billings, S. A., & Voon, W. S. F. (1983). Structure detection and model validity tests in the identification of nonlinear systems. IEE Proceedings D Control Theory and Applications, 130(4), 193. https://doi.org/10.1049/ip-d.1983.0034

Billings, S. A., & Voon, W. S. F. (1986). Correlation based model validity tests for non-linear models. International Journal of Control, 44(1), 235–244. https://doi.org/10.1080/00207178608933593

Billings, S. A. (2013). Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains. John Wiley & Sons.

Blake, A. P., & Kapetanios, G. (2000). A radial basis function artificial neural network test for ARCH. Economics Letters, 69(1), 15–23. https://doi.org/10.1016/S0165-1765(00)00267-6

Booth, A. L., & Van Ours, J. C. (2008). Job satisfaction and family happiness: the part-time work puzzle. The Economic Journal, 118(526), F77-F99.

Box, G. E., & Draper, N. R. (1987). Empirical model-building and response surfaces. John Wiley & Sons.

Boynton, R. J., Balikhin, M. A., Billings, S. A., Wei, H. L., & Ganushkina, N. (2011). Using the NARMAX OLS-ERR algorithm to obtain the most influential coupling functions that affect the evolution of the magnetosphere. Journal of Geophysical Research: Space Physics, 116(5), 1–8. https://doi.org/10.1029/2010JA015505

Boynton, R. J., Balikhin, M. A., Billings, S. A., Sharma, A. S., & Amariutei, O. A. (2011). Data derived NARMAX Dst model. Annales Geophysicae 29(6), 965-971.

Brockwell, P. J., & Davis, R. A. (1991). Time Series: Theory and Methods. Technometrics, Springer.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model Selection: An Integral Part of Inference. Biometrics, 53(2), 603. https://doi.org/10.2307/2533961

Bustince, H., Barrenechea, E., Pagola, M., Fernandez, J., Xu, Z., Bedregal, B., ... & De Baets, B. (2016). A historical account of types of fuzzy sets and their relationships. IEEE Transactions on Fuzzy Systems, 24(1), 179-194.

Camporeale, E., Wing, S., Johnson, J., Jackman, C. M., & McGranaghan, R. (2018). Space Weather in the Machine Learning Era: A Multidisciplinary Approach. Space Weather, 16(1), 2–4. https://doi.org/10.1002/2017SW001775

Candanedo, J. A., Dehkordi, V. R., & Stylianou, M. (2013). Model-based predictive control of an ice storage device in a building cooling system. Applied Energy, 111, 1032-1045.

Candanedo, L. M., Feldheim, V., & Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. Energy and Buildings, 140, 81–97. https://doi.org/10.1016/j.enbuild.2017.01.083

Carr, D., Freedman, V. A., Cornman, J. C., & Schwarz, N. (2014). Happy marriage, happy life? Marital quality and subjective well-being in later life. Journal of Marriage and Family, 76(5), 930–948. https://doi.org/10.1111/jomf.12133

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chandorkar, M., Camporeale, E., & Wing, S. (2017). Probabilistic forecasting of the disturbance storm time index: An autoregressive Gaussian process approach. Space Weather, 15(8), 1004–1019. https://doi.org/10.1002/2017SW001627

Chen, S., & Billings, S. A. (1989). Representations of non-linear systems: the NARMAX model. International Journal of Control, 49(3), 1013–1032. https://doi.org/10.1080/00207178908559683

Chen, S., & Billings, S. A. (1992). Neural networks for nonlinear dynamic system modelling and identification. International Journal of Control, 56(2), 319–346. https://doi.org/10.1080/00207179208934317

Chen, S., Billings, S. A., & Grant, P. M. M. (1990). Non-linear system identification using neural networks. International Journal of Control, 51(6), 1191–1214. https://doi.org/10.1080/00207179008934126

Chen, S., Billings, S. A., & Luo, W. (1989). Orthogonal least squares methods and their application to non-linear system identification. International Journal of Control, 50(5), 1873–1896. https://doi.org/10.1080/00207178908953472

Chen, S., Hong, X., Harris, C. J., & Sharkey, P. M. (2004). Sparse Modeling Using Orthogonal Forward Regression with PRESS Statistic and Regularization. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 34(2), 898–911. https://doi.org/10.1109/TSMCB.2003.817107

Ciregan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2012.6248110

Christina, M., Nouvellon, Y., Laclau, J. P., Stape, J. L., Campoe, O. C., & Le Maire, G. (2016). Sensitivity and uncertainty analysis of the carbon and water fluxes at the tree scale in Eucalyptus plantations using a metamodeling approach. Canadian Journal of Forest Research, 46(3), 297-309.

Crawley, D. B., Hand, J. W., Kummert, M., & Griffith, B. T. (2008). Contrasting the capabilities of building energy performance simulation programs. Building and environment, 43(4), 661-673.

Davis, T. N., & Sugiura, M. (1966). Auroral electrojet activity index AE and its universal time variations. Journal of Geophysical Research, 71(3), 785-801.

Devijver, P. A., & Kittler, J. (1982). Pattern recognition: A statistical approach. Prentice hall.

Dong, B., Cao, C., & Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical region. Energy and Buildings, 37(5), 545-553.

Easterlin, R. A. (1995). Will raising the incomes of all increase the happiness of all? Journal of Economic Behavior and Organization, 27(1), 35–47. https://doi.org/10.1016/0167-2681(95)00003-B

Efron, B., & Tibshirani, R. J. (1998). An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability). Boca Raton, Fla.: Chapmann and Hall, 57.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. Journal of the American Statistical Association, 78(382), 316–331. https://doi.org/10.1080/01621459.1983.10477973

Eğrioğlu, E., Aladağ, Ç. H., & Günay, S. (2008). A new model selection strategy in artificial neural networks. Applied Mathematics and Computation, 195(2), 591–597. https://doi.org/10.1016/j.amc.2007.05.005

Ekici, B. B., & Aksoy, U. T. (2009). Prediction of building energy consumption by using artificial neural networks. Advances in Engineering Software, 40(5), 356-362.

Ekici, T., & Koydemir, S. (2016). Income Expectations and Happiness: Evidence from British Panel Data. Applied Research in Quality of Life, 11(2), 539–552. https://doi.org/10.1007/s11482-014-9380-9

Elmslie, B. T., & Tebaldi, E. (2014). The determinants of marital happiness. Applied Economics, 46(28), 3452–3462. https://doi.org/10.1080/00036846.2014.932047

Enkvist, Å., Ekström, H., & Elmståhl, S. (2012). What factors affect life satisfaction (LS) among the oldest-old? Archives of Gerontology and Geriatrics, 54(1), 140–145. https://doi.org/10.1016/j.archger.2011.03.013

Fagerström, C., Lindwall, M., Berg, A. I., & Rennemark, M. (2012). Factorial validity and invariance of the Life Satisfaction Index in older people across groups and time: Addressing the heterogeneity of age, functional ability, and depression. Archives of Gerontology and Geriatrics, 55(2), 349–356. https://doi.org/10.1016/j.archger.2011.10.007

Fan, C., Xiao, F., & Wang, S. (2014). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Applied Energy, 127, 1-10.

Fikru, M. G., & Gautier, L. (2015). The impact of weather variation on energy consumption in residential houses. Applied energy, 144, 19-30.

Fletcher, G. J., Fitness, J., & Blampied, N. M. (1990). The link between attributions and happiness in close relationships: The roles of depression and explanatory style. Journal of Social and Clinical Psychology, 9(2), 243–255. https://doi.org/http://dx.doi.org.ezp.lib.unimelb.edu.au/10.1521/jscp.1990.9.2.243

Frey, B. S., & Stutzer, A. (2002). What Can Economists Learn from Happiness Research? Journal of Economic Literature, 40(2), 402–435. https://doi.org/10.1257/002205102320161320

Fujita, F., & Diener, E. (2005). Life Satisfaction Set Point: Stability and Change. Journal of Personality and Social Psychology, 88(1), 158–164. https://doi.org/10.1037/0022-3514.88.1.158

Garcia, C., & Delakis, M. (2002). A neural architecture for fast and robust face detection. In International Conference on Pattern Recognition-ICPR, 26, 44–47. https://doi.org/10.1109/ICPR.2002.1048232

Gerdtham, U. G., & Johannesson, M. (2001). The relationship between happiness, health, and socio-economic factors: Results based on Swedish microdata. Journal of Socio-Economics, 30(6), 553–557. https://doi.org/10.1016/S1053-5357(01)00118-4

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, 21(2), 215–223. https://doi.org/10.1080/00401706.1979.10489751

Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. T., & Vasyliunas, V. M. (1994). What is a geomagnetic storm?. Journal of Geophysical Research: Space Physics, 99(A4), 5771-5792.

Gonzalez, P. A., & Zamarreno, J. M. (2005). Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. Energy and buildings, 37(6), 595-601.

Gonzalez, W. D., Dal Lago, A., De Gonzalez, A. C., Vieira, L. E. A., & Tsurutani, B. T. (2004). Prediction of peak-Dst from halo CME/magnetic cloud-speed observations. Journal of atmospheric and solar-terrestrial physics, 66(2), 161-165.

Gorry, A., Gorry, D., & Slavov, S. N. (2018). Does retirement improve health and life satisfaction?. Health economics, 27(12), 2067-2086.

Gschwandtner, A., Jewell, S. L., & Kambhampati, U. (2016). On the Relationship between Lifestyle and Happiness in the UK: Discussion paper 16/13.

Gu, Y., Wei, H. L., Boynton, R. J., Walker, S. N., & Balikhin, M. A. (2019). System Identification and Data-Driven Forecasting of AE Index and Prediction Uncertainty Analysis Using a New Cloud-NARX Model. Journal of Geophysical Research: Space Physics, 124(1), 248-263.

Gu, Y., & Wei, H.-L. (2016). Analysis of the relationship between lifestyle and life satisfaction using transparent and nonlinear parametric models. In 22nd International Conference on Automation and Computing.

Gu, Y., & Wei, H. L. (2018a). Significant Indicators and Determinants of Happiness: Evidence from a UK Survey and Revealed by a Data-Driven Systems Modelling Approach. Social Sciences, 7(4), 53.

Gu, Y., & Wei, H. L. (2018b). A robust model structure selection method for small sample size and multiple datasets problems. Information Sciences, 451, 195-209.

Gu, Y., Wei, H. L., & Balikhin, M. A. (2017). Nonlinear dynamic predictive model selection and interference using information criteria. In 23rd International Conference on Automation and Computing (ICAC), 1-6.

Gu, Y., Wei, H. L., & Balikhin, M. M. (2018). Nonlinear predictive model selection and model averaging using information criteria. Systems Science & Control Engineering, 6(1), 319-328.

Gu, Y., Wei, H. L., Boynton, R. J., Walker, S. N., & Balikhin, M. A. (2017, June). Prediction of Kp index using NARMAX models with a robust model structure selection method. In 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 1-6.

Guarnieri, F. L., Tsurutani, B. T., Vieira, L. E. A., Hajra, R., Echer, E., Mannucci, A. J., & Gonzalez, W. D. (2018). A correlation study regarding the AE index and ACE solar wind data for Alfvénic intervals using wavelet decomposition and reconstruction.

Nonlinear Processes in Geophysics, 25(1), 67–76. https://doi.org/10.5194/npg-25-67-2018

Guo, H., Liu, X., & Sun, Z. (2016). Multivariate time series prediction using a hybridization of VARMA models and Bayesian networks. Journal of Applied Statistics, 43(16), 2897–2909. https://doi.org/10.1080/02664763.2016.1155111

Hansson, I., Buratti, S., Thorvaldsson, V., Johansson, B., & Berg, A. I. (2017). Changes in life satisfaction in the retirement transition: Interaction effects of transition type and individual resources. Work, Aging and Retirement, 4(4), 352-366.

Hartog, J., & Oosterbeek, H. (1998). Health, wealth and happiness: why pursue a higher education? Economics of Education Review, 17(3), 245–256. https://doi.org/10.1016/S0272-7757(97)00064-2

Haykin, S. (1994). Neural networks: a comprehensive foundation. Prentice Hall PTR.

Henning, G., Hansson, I., Berg, A. I., Lindwall, M., & Johansson, B. (2017). The role of personality for subjective well-being in the retirement transition – Comparing variable- and person-oriented models. Personality and Individual Differences, 116, 385–392. https://doi.org/10.1016/j.paid.2017.05.017

Hills, P., & Argyle, M. (1998). Positive moods derived from leisure and their relationship to happiness and personality. Personality and Individual Differences, 25(3), 523–535. https://doi.org/10.1016/S0191-8869(98)00082-8

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N. & Sainath, T. (2012). Deep neural networks for acoustic modeling in speech recognition. IEEE Signal processing magazine, 29.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hong, X., & Chen, S. (2012). An elastic net orthogonal forward regression algorithm. IFAC Proceedings Volumes, 45(16), 1814-1819.

Hong, X., Sharkey, P. M., & Warwick, K. (2003). A robust nonlinear identification algorithm using PRESS statistic and forward regression. IEEE Transactions on Neural Networks, 14(2), 454-458.

Hooten, M. B., & Hobbs, N. T. (2015). A Guide to Bayesian Model Selection. Ecological Monographs, 85(1), 3–28. https://doi.org/10.1890/14-0661.1

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. Biometrika, 76(2), 297–307. https://doi.org/10.2307/2336663

Jewell, S., & Kambhampati, U. S. (2015). Are happy youth also satisfied adults? An analysis of the impact of childhood factors on adult life satisfaction. Social Indicators Research, 121(2), 543-567.

Ji, E. Y., Moon, Y. J., Gopalswamy, N., & Lee, D. H. (2012). Comparison of Dst forecast models for intense geomagnetic storms. Journal of Geophysical Research: Space Physics, 117(A3).

Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. Trends in ecology & evolution, 19(2), 101-108.

Jones, R. V., Fuertes, A., & Lomas, K. J. (2015). The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings. Renewable and Sustainable Energy Reviews, 43, 901-917.

Kamide, Y., Baumjohann, W., Daglis, I. A., Gonzalez, W. D., Grande, M., Joselyn, J. A. & Sharma, A. S. (1998). Current understanding of magnetic storms: Storm-substorm relationships. Journal of Geophysical Research: Space Physics, 103(A8), 17705-17728.

Kass, R., & Raftery, A. (1995). Bayes Factors. Journal of the American Statistical Association, 90, 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kavousian, A., Rajagopal, R., & Fischer, M. (2013). Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. Energy, 55, 184-194.

Kim, S. G. (2015). Fuzzy multidimensional poverty measurement: An analysis of statistical behaviors. Social Indicators Research, 120(3), 635-667.

Köksal, O., Uçak, H., & Şahin, F. (2017). Happiness and domain satisfaction in Turkey. International Journal of Happiness and Development, 3(4), 323-341.

Kontis, V., Bennett, J. E., Mathers, C. D., Li, G., Foreman, K., & Ezzati, M. (2017). Future life expectancy in 35 industrialised countries: projections with a Bayesian model

ensemble. The Lancet, 389(10076), 1323–1335. https://doi.org/10.1016/S0140-6736(16)32381-9

Kvintova, J., Kudlacek, M., & Sigmundova, D. (2016). Active Lifestyle as a Determinant of Life Satisfaction among university students. Anthropologist, 24(1), 179–185.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436.

Lennart, L. (1999). System identification: theory for the user. PTR Prentice Hall, Upper Saddle River, NJ, 1-14.

Li, X., Bowers, C. P., & Schnier, T. (2010). Classification of energy consumption in buildings with outlier detection. IEEE Transactions on Industrial Electronics, 57(11), 3639-3644.

Li, Y., Wei, H.-L., Billings, S. A., & Sarrigiannis, P. G. (2016). Identification of nonlinear time-varying systems using an online sliding-window and common model structure selection (CMSS) approach with applications to EEG. International Journal of Systems Science, 47(11), 2671–2681. https://doi.org/10.1080/00207721.2015.1014448

Lim, H. E., Shaw, D., & Liao, P. (2017). Revisiting the income-happiness paradox: The case of Taiwan and Malaysia. Institutions and Economies, 9(4), 53–69.

Liu, D., Lin, X., & Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics, 63(4), 1079–1088.

Liu, Z. H., Lu, B. L., Wei, H. L., Chen, L., Li, X. H., & Rätsch, M. (2019). Deep Adversarial Domain Adaptation Model for Bearing Fault Diagnosis. IEEE Transactions on Systems, Man, and Cybernetics: Systems, Early Access

Lukacs, P. M., Burnham, K. P., & Anderson, D. R. (2010). Model selection bias and Freedman's paradox. Annals of the Institute of Statistical Mathematics, 62(1), 117–125. https://doi.org/10.1007/s10463-009-0234-4

Marshall, A. M., Bigg, G. R., Van Leeuwen, S. M., Pinnegar, J. K., Wei, H. L., Webb, T. J., & Blanchard, J. L. (2016). Quantifying heterogeneous responses of fish community size structure using novel combined statistical techniques. Global change biology, 22(5), 1755-1768.

Masoso, O. T., & Grobler, L. J. (2010). The dark side of occupants' behaviour on building energy use. Energy and buildings, 42(2), 173-177.

Mayaud, P. N. (1980). Derivation, meaning, and use of geomagnetic indices. Washington DC American Geophysical Union Geophysical Monograph Series, 22.

Medel, C. A., & Salgado, S. C. (2013). Does the bic estimate and forecast better than the aic? Revista de Analisis Economico, 28(1), 47–64. https://doi.org/10.4067/S0718-88702013000100003

Moral-Benito, E. (2015). Model averaging in economics: An overview. Journal of Economic Surveys, 29(1), 46–75. https://doi.org/10.1111/joes.12044

Mujcic, R., & Oswald, A. J. (2016). Evolution of well-being and happiness after increases in consumption of fruit and vegetables. American Journal of Public Health, 106(8), 1504–1510. https://doi.org/10.2105/AJPH.2016.303260

Muratori, M., Roberts, M. C., Sioshansi, R., Marano, V., & Rizzoni, G. (2013). A highly resolved modeling technique to simulate residential power demand. Applied Energy, 107, 465–473. https://doi.org/10.1016/j.apenergy.2013.02.057

Norton, J. P. (1986). An introduction to identification. New York: Academic Press.

Pallocchia, G., Amata, E., Consolini, G., Marcucci, M. F., & Bertello, I. (2008). AE index forecast at different time scales through an ANN algorithm based on L1 IMF and plasma measurements. Journal of Atmospheric and Solar-Terrestrial Physics, 70(2–4), 663–668. https://doi.org/10.1016/j.jastp.2007.08.038

Pérez-Lombard, L., Ortiz, J., & Pout, C. (2008). A review on buildings energy consumption information. Energy and buildings, 40(3), 394-398.

Plümper, T., & Neumayer, E. (2012). Model uncertainty and robustness tests: Towards a new logic of statistical inference. Available at SSRN 2129113.

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Systematic biology, 53(5), 793-808.

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. Psychological Methods, 17(1), 1–14. https://doi.org/10.1037/a0026804

Puvill, T., Lindenberg, J., de Craen, A. J. M., Slaets, J. P. J., & Westendorp, R. G. J. (2016). Impact of physical and mental health on life satisfaction in old age: a population based observational study. BMC Geriatrics, 16(1), 194. https://doi.org/10.1186/s12877-016-0365-4

Robinson, N. J., Benke, K. K., & Norng, S. (2015). Identification and interpretation of sources of uncertainty in soils change in a global systems-based modelling process. Soil Research, 53(6), 592-604.

Rusell, S., & Norvig, P. (2003). Artificial intelligent: A modern approach.

Sabatini, F. (2014). The relationship between happiness and health: Evidence from Italy. Social Science and Medicine, 114, 178–187. https://doi.org/10.1016/j.socscimed.2014.05.024

Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, 6(2), 461–464. https://doi.org/10.1214/aos/1176344136

Smith, G. C. S., Seaman, S. R., Wood, A. M., Royston, P., & White, I. R. (2014). Correcting for optimistic prediction in small data sets. American Journal of Epidemiology, 180(3), 318–324. https://doi.org/10.1093/aje/kwu140

Sodestrom, T., & Stoica, P. (1989). System Identification. Englewood Cliffs, NJ: Prentice-Hall.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society, 36(2), 111–147. https://doi.org/10.2307/2984809

Suits, D. B. (1957). Use of Dummy Variables in Regression Equations. Journal of the American Statistical Association, 52(280), 548–551. https://doi.org/10.1080/01621459.1957.10501412

Symonds, M. R., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behavioral Ecology and Sociobiology, 65(1), 13-21.

Takalo, J., & Timonen, J. (1997). Neural network prediction of AE data. Geophysical Research Letters, 24(19), 2403–2406. https://doi.org/10.1029/97GL02457

Temerin, M., & Li, X. (2002). A new model for the prediction of Dst on the basis of the solar wind. Journal of Geophysical Research: Space Physics, 107(A12).

Temerin, M., & Li, X. (2006). Dst model for 1995–2002. Journal of Geophysical Research: Space Physics, 111(A4).

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288. http://www.jstor.org/stable/2346178

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 67(1), 91–108. https://doi.org/10.1111/j.1467-9868.2005.00490.x

Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy and Buildings, 49, 560–567. https://doi.org/10.1016/j.enbuild.2012.03.003

Tsurutani, B. T., & Gonzalez, W. D. (1987). The cause of high-intensity long-duration continuous AE activity (HILDCAAs): Interplanetary Alfvén wave trains. Planetary and Space Science, 35(4), 405-412.

Veenhoven, R. (1996). Developments in satisfaction-research. Social Indicators Research, 37(1), 1–46. https://doi.org/10.1007/BF00300268

Vlaar, M. P., Birpoutsoukis, G., Lataire, J., Schoukens, M., Schouten, A. C., Schoukens, J., & Van Der Helm, F. C. T. (2018). Modeling the Nonlinear Cortical Response in EEG Evoked by Wrist Joint Manipulation. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26(1), 205–215. https://doi.org/10.1109/TNSRE.2017.2751650

Vlaar, M. P., Solis-Escalante, T., Vardy, A. N., Van Der Helm, F. C. T., & Schouten, A. C. (2017). Quantifying nonlinear contributions to cortical responses evoked by continuous wrist manipulation. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25(5), 481–491. https://doi.org/10.1109/TNSRE.2016.2579118

Vodopivec, M., & Dolenc, P. (2008). Live longer, work longer: making it happen in the labor market. Financial theory and practice, 32(1), 65-81.

Von Humboldt, S., Leal, I., & Pimenta, F. (2014). Living Well in Later Life: The Influence of Sense of Coherence, and Socio-Demographic, Lifestyle and Health-Related Factors on Older Adults' Satisfaction with Life. Applied Research in Quality of Life, 9(3), 631–642. https://doi.org/10.1007/s11482-013-9262-6

Wang, G., Xu, C., & Li, D. (2014). Generic normal cloud model. Information Sciences, 280, 1-15.

Wang, H., Karimi, H. R., Liu, P. X., & Yang, H. (2017). Adaptive neural control of nonlinear systems with unknown control directions and input dead-zone. IEEE Transactions on Systems, Man, and Cybernetics: Systems, (99), 1-11.

Wang, Q., Fan, H., Zhu, L., & Tang, Y. (2018). Deeply Supervised Face Completion With Multi-Context Generative Adversarial Network. IEEE Signal Processing Letters, 26(3), 400-404.

Ward, S. J., & King, L. A. (2016). Poor but happy? Income, happiness, and experienced and expected meaning in life. Social Psychological and Personality Science, 7(5), 463–470. https://doi.org/10.1177/1948550615627865

Wei, H. L., & Billings, S. A. (2004a). A unified wavelet-based modelling framework for non-linear system identification: The WANARX model structure. International Journal of Control, 77(4), 351–366. https://doi.org/10.1080/00207170420001976 22

Wei, H. L., Billings, S. A., & Balikhin, M. (2004b). Prediction of the Dst index using multiresolution wavelet models. Journal of Geophysical Research: Space Physics, 109(A7). https://doi.org/10.1029/2003JA010332

Wei, H. L., & Billings, S. A. (2006). An efficient nonlinear cardinal B-spline model for high tide forecasts at the Venice Lagoon. Nonlinear Processes in Geophysics, 13(5), 577–584. https://doi.org/10.5194/npg-13-577-2006

Wei, H. L., & Billings, S. A. (2008a). Model Structure Selection Using an Integrated Foward Orthogonal Search Algorithm Assisted by Square Correlation and Mutual Information. International Journal of Modelling, Identification and Control, 3(4), 341–356. https://doi.org/10.1504/IJMIC.2008.020543

Wei, H. L., & Billings, S. A. (2008). Generalized cellular neural networks (GCNNs) constructed using particle swarm optimization for spatio-temporal evolutionary pattern identification. International journal of Bifurcation and Chaos, 18(12), 3611-3624.

Wei, H. L., & Billings, S. A. (2009). Improved model identification for non-linear systems using a random subsampling and multifold modelling (RSMM) approach. International Journal of Control, 82(1), 27–42. https://doi.org/10.1080/00207170801955420

Wei, H. L., Billings, S. A., & Balikhin, M. A. (2006). Wavelet based non-parametric NARX models for nonlinear input-output system identification. International Journal of Systems Science, 37(15), 1089–1096. https://doi.org/10.1080/00207720600903011

Wei, H. L., Billings, S. A., & Liu, J. (2004). Term and variable selection for non-linear system identification. International Journal of Control, 77(1), 86–110. https://doi.org/10.1080/00207170310001639640

Wei, H. L., Billings, S. A., Surjalal Sharma, A., Wing, S., Boynton, R. J., & Walker, S. N. (2011). Forecasting relativistic electron flux using dynamic multiple regression models. Annales Geophysicae, 29(2), 415–420. https://doi.org/10.5194/angeo-29-415-2011

Wei, H. L., Zhu, D. Q., Billings, S. A., & Balikhin, M. A. (2007). Forecasting the geomagnetic activity of the Dst index using multiscale radial basis function networks. Advances in Space Research, 40(12), 1863–1870. https://doi.org/10.1016/j.asr.2007.02.080

Wei, H. L., & Bigg, G. R. (2017). The Dominance of Food Supply in Changing Demographic Factors across Africa: A Model Using a Systems Identification Approach. Social Sciences, 6(4), 122. https://doi.org/10.3390/socsci6040122

Wing, S., Johnson, J. R., Jen, J., Meng, C. I., Sibeck, D. G., Bechtold, K., … Takahashi, K. (2005). Kp forecast models. Journal of Geophysical Research: Space Physics, 110(A4). https://doi.org/10.1029/2004JA010500

Wintoft, P., & Cander, L. R. (2000). Ionospheric foF2 storm forecasting using neural networks. Physics and Chemistry of the Earth, Part C: Solar, Terrestrial and Planetary Science, 25(4), 267–273. https://doi.org/10.1016/S1464-1917(00)00015-5

Wu, J. G., & Lundstedt, H. (1996). Prediction of geomagnetic storms from solar wind data using Elman recurrent neural networks. Geophysical Research Letters, 23(4), 319-322. http://doi.org/10.1029/96GL00259

Wu, J. G., & Lundstedt, H. (1997). Geomagnetic storm predictions from solar wind data with the use of dynamic neural networks. Journal of Geophysical Research A: Space Physics, 102(47), 14255-14268. http://doi.org/10.1029/97JA00975

Y. Yang, J. Dewald, F. C. van der Helm, & A. C. Schouten (2018). Unveiling neural coupling within the sensorimotor system: directionality and nonlinearity. European Journal of Neuroscience, 48(7), 2407-2415.

Yan, D., O'Brien, W., Hong, T., Feng, X., Gunay, H. B., Tahmasebi, F., & Mahdavi, A. (2015). Occupant behavior modeling for building performance simulation: Current state and future challenges. Energy and Buildings, 107, 264-278. http://doi.org/10.1016/j.enbuild.2015.08.032

Young, P. C. (1984). Recursive estimation and time-series analysis: An introduction for the student and practitioner. Berlin: Springer-Verlag.

Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zhang, H. Y, Ji, P., Wang, J. Q, & Chen, X. H. (2016). A Neutrosophic Normal Cloud and Its Application in Decision-Making. Cognitive Computation, 8(4), 649–669. https://doi.org/10.1007/s12559-016-9394-8

Zhang, Q., & Benveniste, A. (1992). Wavelet Networks. IEEE Transactions on Neural Networks, 3(6), 889–898. https://doi.org/10.1109/72.165591

Zhang, W., Zhu, J., & Gu, D. (2017). Identification of robotic systems with hysteresis using Nonlinear AutoRegressive eXogenous input models. International Journal of Advanced Robotic Systems, 14(3). https://doi.org/10.1177/1729881417705845

Zhao, P., Suryanarayanan, S., & Simões, M. G. (2013). An energy management system for building structures using a multi-agent decision-making control methodology. IEEE Transactions on Industry Applications, 49(1), 322-330. http://doi.org/10.1109/TIA.2012.2229682

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 67(2), 301–320. http://doi.org/DOI 10.1111/j.1467-9868.2005.00527.x

Zurada, J. M. (1992). Introduction to artificial neural systems. St. Paul: West publishing company.