Complex dynamic system identification and probabilistic prediction using NARMAX and machine learning methods

Yiming Sun

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

The University of Sheffield
Faculty of Engineering
Department/School of Automatic Control and Systems Engineering

12/10/2023

Declaration

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not been previously presented for an award at this, or any other, university.

All sentences or passages quoted in this document from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure.

Acknowledgement

Embarking on the journey towards earning my PhD has been an enlightening and transformative chapter in my life, illuminated and enriched by the wisdom, support, and encouragement of numerous invaluable individuals.

Foremost, my deepest and most heartfelt gratitude is extended to my esteemed supervisor, Dr Hua-Liang Wei. Your unwavering faith, boundless patience, and insightful advice have been the linchpin in the successful completion of this thesis. Your meticulous critiques, profound wisdom, and the generosity with which you shared your expertise have not only sculpted this academic work but have also significantly influenced my growth as a researcher and as an individual. Your encouragement during times of doubt and celebration of milestones, big or small, have been a constant source of motivation and reassurance.

I would also like to express my appreciation to my second supervisor, Professor Michael Balikhin, for his expertise and guidance throughout this research journey. Your constructive feedback and supportive demeanour have been instrumental in navigating through the complexities of this research.

To my beloved wife, Dr Yiping Meng, your unwavering support and love have been my stronghold. Your sacrifices, understanding, and steadfast belief in me have been a constant source of strength and motivation. Your intellectual contributions and emotional support have been indispensable in this journey, and I am profoundly grateful for your presence in my life.

My heartfelt thanks go to my parents and parents-in-law for their boundless love, prayers, and constant encouragement. Your sacrifices and unyielding belief in my pursuits have been the bedrock upon which my academic and personal achievements have been built.

I am also immensely grateful to my friends and colleagues, especially Dr Yuanli Gu, Dr Yihui Tao, Mr Bo Sun, Mr Guoliang Wang, and Miss Nerfita Nikentari for their camaraderie, insightful discussions, and the countless hours spent pondering over research challenges. Your friendship has been a source of joy and comfort throughout this journey.

A special acknowledgment is extended to my project partners, Professor Edward Hanna, Professor Adam Scaife, and Dr Ian Simpson whose expertise and collaborative spirit have significantly enriched this research. Similarly, I am thankful to Professor Len Shaffrey, and Dr Laura Baker for their valuable insights and constructive feedback which have enhanced the quality and depth of this work. I would like to thank NERC (grant NE/V001787/1) for funding this research.

To all staff in the Department of ACSE, your supportive environment and dedication to excellence have been pivotal in facilitating my research endeavours. Your collective wisdom, support, and encouragement have significantly enriched my academic journey.

In conclusion, I dedicate this thesis to all those mentioned and to the many more who have contributed to this journey in various capacities. Your collective contributions have been instrumental in shaping this work and my academic journey. I am profoundly grateful for the privilege of learning from and being supported by such an inspiring group of individuals.

Abstract

System identification serves as a cornerstone in the formation of mathematical models for dynamical systems from the interpretation of observed data. Not only does this play a pivotal role in expounding complex relationships, but it also optimizes system performance, essential for architecting reliably robust controllers and fostering accurate predictions across multiple scientific and engineering realms. Nevertheless, model uncertainty in system identification poses consequential effects on the dependability of these models and the decision-making mechanisms within intricate systems.

Addressing model uncertainty is fundamental for bolstering the resilience of the identified models. However, contemporary research often falls short, demonstrating a general lack of systematic approaches towards the efficacious evaluation and quantification of uncertainties. Moreover, there lies an inadequacy in research efforts to delve into sophisticated methodologies and emergent tools with the capability of efficiently navigating model uncertainty. This consequently overlooks potential avenues for the optimization of system identification and prediction in the face of inherent ambiguities and complexities inherent within dynamic systems.

This thesis undertakes an examination of model uncertainty within NARMAX, an essential but commonly neglected component within system identification and modelling. The study accentuates the tangible influence model uncertainty wields upon decision-making within complex dynamic systems. The research introduces a polynomial based NARMAX model, leveraging the FROLS algorithm in conjunction with set theory for the quintessential quantification of uncertainty. Additionally, this thesis gives rise to several innovative methodologies, such as the DeepNARMAX network. This network maintains the interpretability and accuracy inherent to NARMAX while proficiently managing high dynamic scenarios. The network's efficacy is put to the test in real-world applications, such as weather and power forecasting models. Further innovations presented within this thesis include the SW-NARMAX model that pertains to seasonal weather forecasting and the MAB-NARMAX model capable of the efficient detection of model structures using a mask matrix.

Table of Contents

List of Tables

List of Figures

List of abbreviations

AI	Artificial intelligence
AIC	Akaike information criterion
AMO	Atlantic multidecadal oscillation
APRESS	The adjustable prediction error sum of squares
ARIMA	Autoregressive integrated moving average
ASOS	The automated surface observing systems
AUV	Autonomous underwater vehicle
AWOS	Automated weather observing system
BC	Beaufort/Chukchi seas
BER	Bering sea
BERT	Bidirectional encoder representations from transformers
BIC	Bayesian information criterion
BK	Barents-kara seas
CI	Confidence interval
CMSS	Common model structure selection
CNN	Convolutional neural network
DIP	North Atlantic dipole
DJF	December January February
DL	Deep learning
DNN	Deep neural networks
DRNN	Diagonal recurrent neural network
EA	East Atlantic
EEG	Electroencephalogram
EISST	E. Indian Ocean
ENSO	El-Niño southern oscillation
EOF	Empirical orthogonal function
EPD	External parameter-dependent
EPR	E. Pacific rainfall
EPSST	E. Pacific
ERA5	ECMWF reanalysis v5
ERR	Error reduction ratio
ESD	External signal-dependent
ESL	E. Siberian/Laptev seas
ETT	Electricity transformers
EUR	Eurasian

FOS	Forward orthogonal search
FROLS	The forward regression with OLS
GAN	Generative adversarial networks
GIN	Greenland/Iceland Norwegian seas
GPT	Generative pre-trained transformer
GPU	Graphics processing unit
GRE	Sub-polar gyre
HUD	Hudson Bay
ID 3	Iterative dichotomise 3
IFOLS	Iterative forward orthogonal least squares regression
IFOS	Integrated forward orthogonal search
IFS	Integrated forecasting system
IPD	Internal parameter-dependent
ISD	Integrated surface data
JA	June August
JAEA	June August east Atlantic
JASCA	July and August Scandinavian
JJA	June July August
LAB	Labrador Sea
LASSO	Least absolute shrinkage and selection operator
LCD	Local climatological data
LMS	Least mean squares
LSTF	Long short-term forecasting
LSTM	Long short-term memory
MAB	Mask attention-based
MAE	Mean absolute error
MI	Mutual information
MIMO	Multi-input, multi-output
MISO	Multi-input, single output
MJO	Madden-Julian oscillation
MOD	Maximizes the overall dependency
MPO	Model predicted output
MSE	Mean squared error
MSMA	Magnetic shape memory alloy
NAH	North Atlantic horseshoe
NAO	North Atlantic oscillation

NARMAX	Nonlinear Autoregressive Moving Average with Exogenous inputs
NARX	Nonlinear Autoregressive Exogenous Model
NN	Neural Network
OFR	Orthogonal forward regression
OLS	Orthogonal least square
OOD	Out-of-distribution
OSA	One-step-ahead
PCA	Principal component analysis
PID	Proportional – integral – derivative
PMSM	Permanent magnet synchronous machines
PSO	Particle swarm optimization
QBO	Quasi-biennial oscillation
RBF	Radial basis function
RELS	Recursive extended least squares
RMSE	Root mean squared error
RNN	Recurrent neural network
SARSA	State–action–reward–state–action
SCA	Scandinavian
SEAS	Seasonal forecasting system
SERR	Sum of error reduction ratios
SISO	Single-input, single-output
SIT	Sparse, interpretable, and transparent
SLP	Sea level pressure
SMB	Surface mass balance
SNAO	Station north Atlantic oscillation
SNR	Signal-noise-ratio
SPG	Sub-polar gyre
SST	Sea surface temperatures
SVM	Support vector machine
SW	Sliding window
TAR	Tropical Atlantic rainfall
TASST	Tropical Atlantic
TRI	North Atlantic tripole
TV	Time-varying
UOFR	Ultra-orthogonal forward regression
WISST	W. Indian ocean

WPR	W. Pacific rainfall
WPSST	W. Pacific
XAI	Explainable artificial intelligence

Chapter 1

Introduction

1.1 Background and motivation

1.1.1 System identification and prediction

In an ideal scenario, most systems in the real world can be approximated by a mathematical model which can be used to evaluate the influence of components, predict the system response for further study and investigate the internal and external interaction of the system [1]. Two typical approaches are usually considered to develop the mathematical models: (i) the theoretical one that utilizes the fundamental laws of matter and energy, and (ii) the empirical one that exploits the system inputs and outputs recorded from experimental or operating data [2]. The theoretical approach might sometimes present challenges in application or may not be completely suitable, as the inherent dynamics of many real-world systems, such as the space weather systems, climate and weather systems, financial systems and other systems, which are different from the physical-based systems, are not fully known or can be ambiguous. In such systems, only a limited number of the fundamental laws of matter and energy may be discernible. On the contrary, by studying the wealth of information from the observation data rather than describe the system by physical laws, the empirical method, also be called system identification is widely applied and more practical in model development. System identification is an empirical method frequently used in developing mathematical models of dynamic and complex systems from observed inputs and outputs of systems using data-driven modelling methods.

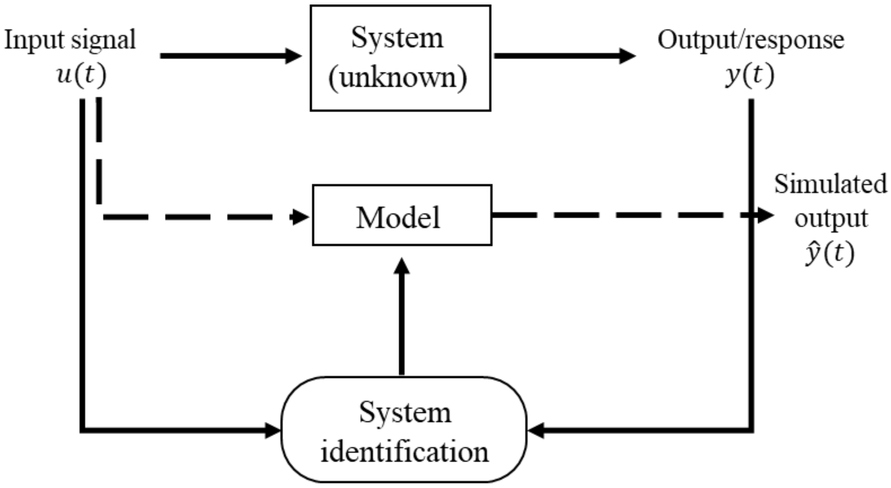


Figure 1 System Identification process

A typical process of system identification is shown in Figure 1. By using data-driven modelling methods, like Linear regression [3], Nonlinear regression [4], Bayesian modelling [5], Lasso regression [6], Long short-term memory (LSTM) [7], Nonlinear autoregressive moving average with exogenous inputs modelling (NARMAX modelling) [1], and methods, an optimal model to describe the unknown system and generate reliable and accurate simulated output is built from the inputs and outputs of the system, which is the primary objective of the system identification [8]. The ideal result is to find the exact model of the unknown system, which is a rare situation as most applications are the black-box systems [1]. Therefore, an optimal model through system identification is more reasonable and desirable for analysing the system and produce accurate prediction [9]. Also, the system identification is applicable to all systems if there are observed data of inputs and outputs of the system [10]. Methods of system identification are often quick and straightforward to be deployed in the study and application [1]. Thus, system identification has been widely applied in many areas of the real world, like space weather [11], industrial design and control [12], economy [13], weather [14], power grid control [15] and etc.

However, several questions in the process of system identification greatly affect the performance of the identified models and understanding of the system, like the accuracy of the observed data, the form of the identified models and the values of parameters in the models. These problems will certainly lead to various identification results, which cause vagueness and uncertainty for studying complex systems. This situation is also defined as model uncertainty [16].

1.1.2 Model uncertainty

Model uncertainty is a pervasive and inherent problem in system identification and prediction of real applications [17], which is hardly to be eliminated. If the system is well understood, where this system can be defined by a theoretical model, then there is no model uncertainty as the model is unique and correct. But as discussed above, in most realistic situations, systems are not fully comprehended. Thus, model uncertainty will exist with the existence of system identification.

The intuitive way to reflect the uncertainties of the identified models is the differences between model output $\hat{y}$ and system response y, which arise from many aspects, such as unpredictable variability and system error in the input data, big noisy signal coupled with the original signal, model parameters, model structures, modelling methods, and the criteria to evaluate the differences [18]. Model uncertainty has increasingly attracted attention from system identification and modelling area, and research results on model uncertainty have put forward higher requirements and more extensive goals for system identification [19]. Often, model uncertainty to be reduced is the main objective as more accurate model definitely has more understanding of the system and more precise forecast of the system in the future. Given the numerous sources of model uncertainty, as illustrated in Figure 2, extensive research has been conducted to enhance model performance and minimize model uncertainty [20] [21] [22].

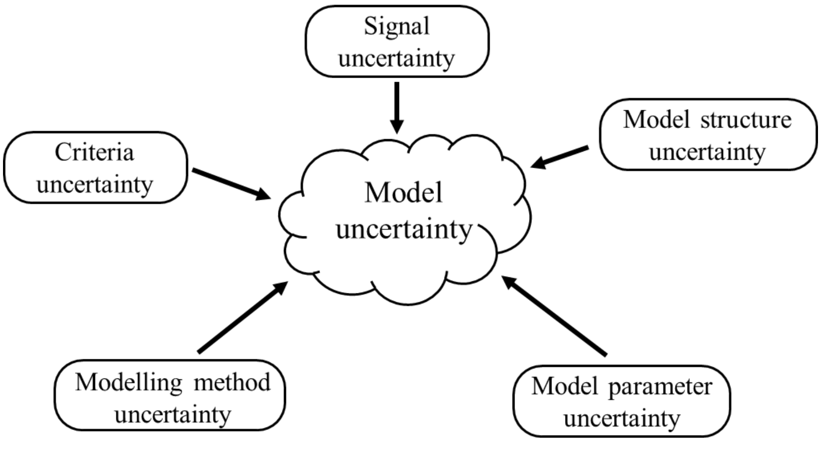

Signal uncertainty
Criteria uncertainty
Model structure uncertainty
Model uncertainty
Modelling method uncertainty
Model parameter uncertainty

Figure 2 Sources of model uncertainty

Model uncertainty has gradually become one of the critical information in analysing the system and understanding the practical cases, not only the problem to be reduced during the process [23, 24]. The model uncertainty information can greatly affect the decisions in human life directly or indirectly [25, 26]. With the uncertainty information, not the optimal model but a set of identified models are constructed to analyse the system and generate either combination predictions or probabilistic forecast

[27]. Therefore, more modelling results are present to the end users and researchers to discover and invest the unknown systems. Thus, there is a trade-off between utilizing and reducing model uncertainty, which is shown in Figure 3.

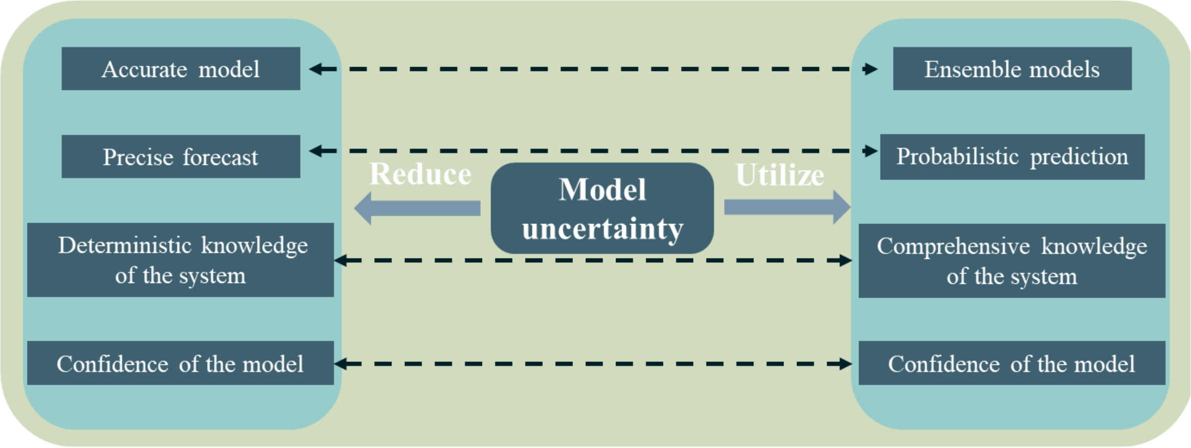


Figure 3 Trade-off between reducing and utilizing model uncertainty

Model uncertainty will bring many identified models with similar performance, which increase the information entropy and provide more options for end users. However, more models represent more riskiness to trust the identified results, while sometimes one best model and most reliable prediction are more considered. Like in the weather prediction, travellers desire a certain predicting result for the preparation and decision, while researchers prefer comprehensive models to discover more possible interaction in the weather system. Therefore, it is important to measure the trade-off of dealing with model uncertainty and provide satisfied identification results.

1.1.3 NARMAX and model uncertainty

NARMAX is a widely applied and efficient data-driven modelling method, which can provide sparse, interpretable and transparent (SIT) parametric models [28]. The SIT NARMAX models are usually desirable and useful for understanding the inherent dynamics and interactions of the system states. Moreover, the NARMAX model is compact and clear to explain and reveal the model structure in many applications where the primary modelling objective and task are to exploit and obtain an insightful description of how the system output explicitly depends on the system inputs, and to provide precise prediction of the system from the system inputs.

For convenience of description, consider the case of systems with single input and single output (SISO), denoted by $u(t)$ and $y(t)$, respectively, where $t=1,2,...,T$ is the sampling index (time instant). The lagged input and output variables are defined as:

$$\begin{aligned} u(t) &\rightarrow u(t-d),u(t-d-1),...,u(t-n_u) \\ y(t) &\rightarrow y(t-1),y(t-2),...,y(t-n_y) \end{aligned} \tag{1.1}$$

where d is a time lag between the system input and output (usually $d=1$ but can be set to zero if the system input $u(t)$ instantly affects the system behaviour), n_u and n_y are the maximum time lags. These lagged variables can be used to create a model term dictionary, which can be used to build models. Therefore, the NARMAX model is defined as [1]:

$$y(t)=F[y(t-1),...,y(t-n_y),u(t-d),...,u(t-d-n_u),e(t-1),...,e(t-n_e)]+e(t) \tag{1.2}$$

where, $e(t)$ is the noise sequence, which is not measurable but can be estimated using model prediction error in practical modelling , n_e is the maximum lag for the system noise; $F[\cdot]$ is some nonlinear function, where in practice, there are many types of model structures that can be used to approximate the unknown mapping $F[\cdot]$, including power-form polynomial models [29], rational models [30], neural networks [31], fuzzy logic-based models [32], and wavelet expansions [33]. The most commonly used model is the power-form polynomial representation [1]. Normally, the NARMAX model can be considered as a generic linear-in-the-parameters representation as:

$$y(t)=\sum_{m=1}^{M}\theta_m p_m(t)+e(t) \tag{1.3}$$

where $p_m(t)=p_m(x(t))$ with $m=1,2,...,M$ are the regressors formed by some combinations of model variables chosen from the variable vector $\boldsymbol{x}(t)=[x_1(t),...,x_n(t)]^T$, where the variables in the vector are the lagged inputs and outputs like: $x(t)=[u(t-1),...,u(t-n_u),y(t-1),...,y(t-n_y),e(t-1),...,e(t-n_e)]^T$. θ_m are the model parameters. In this representation, *M* represents the length of the identified model.

Ideally, the goal of the system identification is to find the most appropriate model to describe the system and reveal the interactions of the inputs and outputs of the system, also to generate the accurate and reliable predictions of the system in the future. Thus, the most appropriate NARMAX model should also be able to predict future system behaviour accurately and reliably.

Based on the source of model uncertainty shown in Figure 2, there are several aspects bringing model uncertainty in NARMAX modelling, including

- Noisy data. The observed inputs and outputs of systems might be noisy and uncertainty due to system observation error, low SNR, ambiguity, data lineage etc.
- Model parameter uncertainty. Like in Eq (1.3) the regressors $p_m(t)$ in the model are determined, while the corresponding parameters θ_m can be diverse from different training subsets. Therefore, many models with the same regressors and different parameters can be generated, and various predictions are produced.
- Model structure uncertainty. Lots of factors lead to model structure uncertainty in NARMAX models. The nonlinear form $F[\cdot]$, the model length *M*, the maximum time lags n_u and n_y are all greatly affect the modelling results.
- Criteria uncertainty. Usually, the sum of error reduction ratios (SERR) is the main criteria to measure the performance of the NARMAX model [34]. When dealing with different real applications, several other criteria, like mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), the R-squared value R^2 or the adjusted R^2, correlation coefficient, etc.

The necessity for discerning and examining model uncertainty in the NARMAX modelling approach is multifaceted, rooted in several pivotal factors:

- Precision and Predictability Enhancement: The reduction of model uncertainty allows for formulating and implementing a more accurate and optimal NARMAX model. Such models are capable of revealing the complex and dynamic interactions prevalent within the system, thereby improving the model's predictability and minimizing previous inaccuracies.
- Deeper Understanding: The inclusion of model uncertainty within NARMAX modelling promotes a broader understanding of dynamical systems. This includes grasping the system's peculiarities, its base functionality, as well as enabling prediction of future behaviours.

- Balancing Trade-Offs: An additional incentive is the delicate balancing required in the complex and intricate NARMAX networks with model uncertainty. Addressing this complexity and uncertainty encourages the development of more accurate, flexible, and practical models that can closely represent dynamic and complex systems.
- Innovation and Progress: Incorporating model uncertainty into established methodologies provides a pathway for creating novel NARMAX frameworks that encourage further progress in this academic field. This aims to stimulate both theoretical advancements and practical implementations, demonstrating the utility and efficiency of these innovative frameworks.
- Real World Applications: Ultimately, deploying these modelling approaches to realistic applications is a compelling factor. By ensuring that improved predictive capabilities and deep understanding offered by managing model uncertainty are applicable to various practical fields - from engineering to technology and economics.

By analysing and studying model uncertainty in NARMAX modelling, this research aims to augment knowledge in this discipline, thereby fostering a sturdy understanding and accurate prediction of complex systems across numerous professional fields. In short, superior management of model uncertainty in NARMAX is essential to fully exploit this modelling paradigm's potential.

1.2 Research hypotheses

This research study seeks to present and defend two hypotheses which find their foundation in the realm of NARMAX. These hypotheses nominally revolve around the strategic integration of model uncertainty as an adjustable hyperparameter, conjecturing that this incorporation could significantly bolster the performance and applicability of NARMAX models.

The first hypothesis maintains that by integrating model uncertainty as a critical hyperparameter within the NARMAX framework, a marked enhancement in predictive accuracy and reliability can be achieved for complex dynamical systems, in contrast to conventional NARMAX models. Existing NARMAX models have typically discounted or minimized the relevance of model uncertainty in their formulation, while this study argues that such an uncertainty holds a key role in computational modelling. Taking into account the inherent challenges and intricate complexities linked with the behaviour of dynamic systems, the need for greater model accuracy and robustness has never been more apparent. Hence, the interlacing of model uncertainty becomes imperative to bolster the performance and predictive capacity of models deployed in scenarios that demand a high degree of precursor understanding and prospective acuity.

The second hypothesis contends that NARMAX models, which have been meticulously engineered with explicit consideration of model uncertainty, exhibit a higher degree of adaptability and overall effectiveness when confronted with real-world applications. Such models, therefore, present a broader comprehension and yield more insightful understanding of complex systems across various domains. In unpredictable, field-based applications, it is common to encounter a wide range of uncontrolled variables and conditions. Here, the rigidity of conventional models may pose an impediment. As a result, there are compelling arguments favouring an alternative approach, one that embraces flexibility over rigidity. Thus, the NARMAX model, designed with a clear accommodation to model uncertainty, emerges as an ideal candidate. Such a model, equipped with the power to adapt to fluid and unpredictable scenarios, can offer a broader spectrum of utility and penetrate deeper into the understanding concerning the behaviour and essential properties of complex systems across a diverse array of domains.

In summation, this research focuses on harnessing the capabilities of NARMAX models and further enhancing their robustness and predictive efficiency by consciously integrating model uncertainty. The examination and defence of these hypotheses forms the core of this research study. Once validated, these hypotheses have the potential to considerably transform and improve the practical applications of dynamic system modelling, injecting an advanced degree of reliability and breadth into the modelling of complex, multifaceted systems. The practical implication of this research, therefore, becomes the demonstration of the sheer power of model uncertainty and how its thoughtful integration into the NARMAX models can augment their overall utility and predictive potency.

1.3 Aims and objectives

In this thesis, the primary focus is centred around the exploration and examination of model uncertainty in NARMAX modelling. Model uncertainty is a pervasive issue in system identification and modelling, extending beyond just NARMAX modelling. This issue is especially significant when dealing with complex dynamical systems. Furthermore, uncertainty can directly influence decision-making by introducing variability into possible outcomes linked to specific actions. Additionally, uncertainty indirectly plays a role by introducing private information. In strategic interactions, the interconnection between uncertainty and private information is inherent: uncertainty creates variations in private information when different agents possess varying access to information regarding uncertain phenomena. Conversely, the presence of private information can also lead to uncertainty if agents are considering its potential implications, resulting in concerns related to moral hazard and adverse.

Numerous studies on the uncertainty of NARMAX models have been conducted, however, a systematic definition, analysis, or quantification of this uncertainty has not been the focus of these investigations. Instead, they tend to concentrate on examining specific prominent or independent uncertainties. As a consequence, in NARMAX modelling, model uncertainty has received less attention compared to hyperparameters, such as model structure, due to the absence of systematic analysis and investigation. The emphasis is often on exploring uncertainties that emerge after the modelling process, resulting in redundancies and complexities in the modelling procedure, and impeding the comprehensive identification and research of complex dynamic systems.

Additionally, with the increased penetration of applications, model uncertainty has transitioned from being merely a singular negative impact to a catalyst for considerable interest and demand for probabilistic modelling and prediction emerging from such uncertainty. Although probabilistic models offer uncertain representations of systems, it is worth noting that, in most cases, as the systems themselves are largely unknown, all system models derived through any method can be viewed as "pseudo-models". Thus, in comparison to deterministic models, probabilistic models can depict systems more comprehensively, while offering highly credible probabilistic predictions for system responses. The research aims are summarized as follows:

- To scrutinize and critically evaluate the nature and implications of model uncertainty within the context of NARMAX modelling.
- To investigate and ascertain the balance between trade-offs and advantages associated with model uncertainty in complex and profound NARMAX networks.
- To design innovative NARMAX models or methods, introducing model uncertainty into the modelling process, seamlessly integrating model uncertainty to enhance the effectiveness of the model and understanding of real systems.
- To implement the constructed NARMAX models in real-world settings, thereby demonstrating their applicability and efficiency across varied practical domains.

The principal objective of this research is to extensively probe and address the relatively uncharted territory of model uncertainty within NARMAX modelling. Propelled by the apparent systematic gaps in the present understanding of model uncertainty and its associated implications, this study aims to delve into the intricate nature of model uncertainty, proposing innovative methods or models that can effectively encompass and manage this uncertainty.

The study is driven by an ambition to devise NARMAX models that offer a comprehensive grasp of real systems and yield precise, efficient predictions, even when applied to multifaceted, complex networks. By achieving this, the research aims to dismantle oft-observed redundant complexities in established modelling procedures, laying a foundation for improved, more reliable applications of NARMAX models across varied real-world scenarios. The detailed objectives are as follows:

- To conduct an exhaustive review of the existing literature concerning model uncertainty in NARMAX modelling to procure a detailed understanding of its current status and identify any discrepancies.
- To execute a theoretical examination of model uncertainty as applied to the NARMAX model, with a focus on its impact on the precision, predictability, and understanding of dynamical systems.
- To formulate an original approach for handling the balance between trade-offs in model uncertainty while preserving the intricacy of NARMAX networks.
- To design and establish sophisticated NARMAX frameworks that incorporate model uncertainty, which in turn enhances their predictive capabilities and overall understanding of dynamics.
- To validate the evolved NARMAX models by employing them in practical scenarios - evaluate their performance and adaptability across assorted fields such as technology, engineering, and economics.

This research study is anchored on in-depth exploration and elucidation of the relatively uncharted domain of model uncertainty within NARMAX. This study postulates that model uncertainty, being an inherent attribute often viewed negatively in problem-solving paradigms, could become a valuable resource in the canvas of advanced NARMAX modelling.

The primary aim of this research is to dismantle the knotted complexities and redundancies in the existing modelling procedures by adeptly incorporating model uncertainty in formulating NARMAX models. It plans to leverage the latent potential of model uncertainty to develop models that ensure a broader understanding of complex systems. Such models are expected to provide precise, insightful, and reliable predictions when implemented to multi-faceted and complex networks across a spectrum of domains including technology, engineering, and economics.

In fulfilling this aim, the research lays out a detailed plan including systematic analysis of the existing literature on model uncertainty in NARMAX modelling to encapsulate a detailed understanding of its current status and underline the existing knowledge gaps. It involves theoretical examination of model uncertainty as applied to NARMAX models prudently outlining its impacts on precision, predictability, and understanding of dynamical systems. The plan also includes crafting novel approaches to balance trade-offs in model uncertainty while maintaining the intricacy of NARMAX networks and the creation of sophisticated NARMAX frameworks that effectively weave model uncertainty enhancing their predictive prowess and dynamic system understanding. Finally, it intends to showcase the practicality and adaptability of the evolved NARMAX models across varied real-world scenarios.

1.4 Outline of the thesis

This thesis comprises seven chapters.

Chapter 1 introduces and explains the background of the study, its aims and objectives.

Chapter 2 provides a comprehensive overview of the evolution of NARMAX, its fundamental models, and prominent feature extraction algorithms. It underscores the importance of implementing NARMAX models in the identification of complex dynamic non-linear systems. Furthermore, an expansive study on model uncertainty is undertaken, including insights into its background, classifications, in addition to research on uncertainty within the context of NARMAX models and machine learning. Through this analysis, the review highlights the significance of studying uncertainty in NARMAX models and pinpoints existing challenges in this domain.

Chapter 3 provides centres on the methodology of this thesis. Acknowledging the complexities of system identification with the NARMAX method due to system inaccessibility, the study explores the identification of key system factors and their impacts through modelling mechanisms. This includes handling model uncertainty originating from diverse NARMAX hyperparameters configurations, as these are known to affect optimal model selection and predictive value generation. Tackling this necessitates a detailed approach to define, examine, and quantify such uncertainties to efficiently reduce them during the modelling procedure. The chapter further looks into the rising prominence of probabilistic models in system research for their comprehensive information provision. However, it also acknowledges their capacity to increase inherent model uncertainties. A qualitative and quantitative scrutiny of uncertainties in polynomial NARMAX models has been done, establishing a link between the uncertainties set's size and corresponding hyperparameters values. It lays the groundwork for defining matching objective functions and optimization processes needed to yield optimal model versions.

Chapter 4 provides a novel approach called the DeepNARMAX network, which enhances the efficiency of the classic polynomial NARMAX model. The new model's construct includes several layers for processing input signals, dimension expansion, feature selection and reduction, non-linear operations, and model creation. The chapter also proposes a PSO-based algorithm for optimizing the calculation of a new hyperparameter - gate weight, which ensures optimal feature selection. The DeepNARMAX model was validated and tested through comprehensive experiments and comparisons with established neural network models. The DeepNARMAX successfully addresses the dimensional explosion issue and improves the capability for dealing with complex dynamic scenarios while maintaining the interpretability and accuracy of the classic NARMAX model.

Chapter 5 provides a novel sliding window NARMAX method to handle the local details of the system. This SW-NARMAX model can divide the dataset through a sliding window, then perform NARMAX modelling on the system within this window, thereby obtaining an ensemble NARMAX model for system description. At the same time, in order to verify the effectiveness of the sliding window NARMAX model, this chapter applies the SW-NARMAX model to seasonal weather forecasting, using data from past 43 years for seasonal weather modelling and prediction. The research results also prove that SW-NARMAX models are effective in describing and predicting weather systems.

Chapter 6 provides a new mask attention-based NARMRAX (MAB-NARMAX) modelling approach for nonlinear dynamic system identification. The mask attention mechanism is borrowed from the extensively utilized neural network Transformer, aiming to diminish the dependencies among the features and neurons. The efficacy of the proposed method is evaluated through three simulation datasets. Results indicate that the advanced MAB-NARMAX modelling scheme displays impressive multi-step-ahead prediction capabilities for nonlinear system identification, as it can generate a valuable model structure. Despite high noise pollution resulting in low Signal-to-Noise Ratio (SNR), the proposed method consistently delivers reliable system models, outperforming leading machine learning methods such as LASSO and LSTM

Chapter 7 summarizes the principal findings and contributions of this research. This study explores model uncertainty within NARMAX framework, filling a significant gap in systematic research on model uncertainty. To tackle this, we innovatively treat model uncertainty as a hyperparameter and introduce three methods: DeepNARMAX, Sliding Window NARMAX, and Mask Attention-based NARMAX. These methods demonstrate reliability and efficiency across high dynamic situations, seasonal weather forecasting, and model structure detection. However, future work needs to address quantification of model uncertainties, inclusion of more non-linear functions for complex system modelling, and validation of these methods in real-world systems. Despite the limitations, the potential of these new methods in significantly enhancing NARMAX-based model utility and efficiency is high. Furthermore, we highlight the relevance of integrating NARMAX into Explainable AI (XAI) and Green AI initiatives. For XAI, NARMAX can effectively handle interpretability challenges while for Green AI, it can model intricate systems without extensive computational requirement, ensure transparency, and display adaptability across diverse environmental contexts. Therefore, NARMAX-based solutions could perfectly align with sustainability goals while maintaining transparency and interpretability.

1.5 Novelty and contributions of the thesis work

1.5.1 Novelty and contributions

The overall novelty of this thesis lies in the development and application of enhanced NARMAX models for improving system identification and handling the model uncertainty.

- A novel comprehensive qualitative and quantitative examination of model uncertainty inherent in NARMAX modelling was carried out, culminating in an aggregated analysis on the concept of quantification associated with NARMAX model uncertainty. Moreover, the study proposes both optimization and probabilistic modelling techniques, predicated on empirically quantified model uncertainty.
- A novel approach, the DeepNARMAX network, that enhances the efficiency of the classic polynomial NARMAX model, was proposed. This new model includes multiple layers for processing input signals, dimension expansion, feature selection and reduction, non-linear operations, and model creation.
- A new hyperparameter-gate weight was present to control the optimization and iteration of the DeepNARMAX network, while a novel PSO-based algorithm for optimizing the calculation of gate weight was proposed. This ensures optimal feature selection, addressing the dimensional explosion issue, improving the capability for dealing with complex dynamic scenarios, and maintaining the interpretability and accuracy of the classic NARMAX model.
- A sliding window NARMAX approach was introduced in the application for seasonal weather forecasting, demonstrating its effectiveness in modelling and prediction of seasonal weather. The sliding window method allowed for an ensemble NARMAX model to describe the system’s local details and generate probabilistic predictions,

- A groundbreaking mask attention-based NARMAX (MAB-NARMAX) approach designed to tackle the daunting task of model structure detection, a prominent challenge in numerous real-world applications. This innovative methodology leverages the advantageous features of the NARMAX model, synergistically integrating the computational efficiency and information processing prowess of the mask matrix commonplace in Transformer neural network models.

1.5.2 Research publications:

1. Sun, Y. and Wei, H.-L. (2022). “Efficient mask attention-based NARMAX (MAB-NARMAX) model identification,” in Proc. 2022 27th International Conference on Automation & Computing (ICAC).
2. Sun, Y., and Wei, H.-L. (2022). “How weather conditions affect the spread of Covid-19: Findings from a study using interpretable machine learning and NARMAX models,” in Jiang, R. (Editor), Crookes, D. (Editor), Wei, H. L.(Editor), Zhang, L. (Editor), Chazot, P. (Editor): Recent Advances in AI-enabled Automated Medical Diagnosis, , pp.238-252.
3. Yiming Sun, Ian Simpson, Hua-Liang Wei, and E. Hanna, "Probabilistic seasonal forecasts of North Atlantic atmospheric circulation using complex systems modelling and comparison with dynamical models," Meteorological Applications, 2023 (Published).
4. Ian Simpson, Edward Hanna, Laura Baker, Yiming Sun, and Hua-Liang Wei, "North Atlantic circulation indices: links with summer and winter temperature and precipitation in north-west Europe, including persistence and variability," International Journal of Climatology, 2023 (Published)

Chapter 2

Literature Review

The literature review presented in this chapter aims to provide a comprehensive background and understanding of model uncertainty in complex dynamic system modelling, with a specific focus on NARMAX and machine learning methods. The field of complex dynamic systems encompasses a wide range of real-world phenomena, such as power grid [35], traffic network [36, 37], robotics [38], biological networks [39], social systems [40, 41], and economic processes [42], which are characterized by intricate interactions between their components and non-trivial system behaviours. As these systems often exhibit high levels of uncertainty, developing accurate models is of paramount importance to facilitate informed decision-making, optimization, and control. NARMAX, as an efficient system identification method, can capture nonlinear relationships and intricate interactions among system components and produce accurate prediction about the system [1]. However, the conventional NARMAX modelling approach suffers from various limitations, particularly with regard to model uncertainty, which encompasses both parameter uncertainty and structure uncertainty. This chapter seeks to explore the existing body of knowledge on NARMAX modelling and the incorporation of machine learning and deep learning techniques to address these uncertainties, with the objective of improving model accuracy and enabling probabilistic prediction.

2.1 Nonlinear autoregressive moving average with exogenous inputs (NARMAX) model

In this section, we will introduce NARMAX modelling, a powerful approach for complex dynamic system identification. We will discuss its history, advantages, limitations, and applications, as well as review existing methodologies and techniques. This foundation will facilitate our exploration of model uncertainty and the integration of machine learning and probabilistic prediction methods in NARMAX modelling.

2.1.1 Brief introduction

The NARMAX model was first introduced and developed in early 1980s as a generalization of linear AMARX models to address inherent nonlinearity and complexity in real-world dynamic systems by Stephen Billings. The concept of early NARMAX modelling was to combine linear dynamic and static nonlinear elements selected by correlation analysis for identification of nonlinear systems [43]. Later, early developments were proposed to explore different aspects of NARMAX modelling and parameter estimation. In [44], a nonlinear difference equation of NARMAX model was introduced and two modified parameter estimation algorithms were proposed. In [45], three modified extended least squares algorithms were considered for the parameter estimation and structure detection of NARMAX models. In [46, 47], deterministic and stochastic nonlinear systems are delved, laying the groundwork for NARMAX modelling by discussing input-output parametric representations, identification methods, and model structures. Collectively, these explorations significantly contributed to the development and formalization of NARMAX methodology.

After the early developments and formalization of the NARMAX methodology, researchers continued to refine and expand NARMAX to improve its performance, robustness, and applicability in various fields. NARMAX was formally introduced versatile and powerful tool for representing and identifying nonlinear systems in [29], combining nonlinear autoregressive, moving average, and exogenous input components to capture complex system dynamics. Furthermore, the orthogonal least squares (OLS) algorithm was proposed for efficient model identification, addressing the challenges of term selection and overfitting in [48]. Since the formal introduction of NARMAX, it gained increased attention due to the capabilities to capture the nonlinear dynamics of various real-world systems, like automotive diesel engine design [49], solar-powered thermoacoustic refrigerator identification [50], modelling of space weather like sunspot [51], electroencephalogram (EEG) signal processing and modelling [52]

As NARMAX continued to gain traction in various fields, additional advancements were made to improve its overall effectiveness and practicality. Researchers focused on improving the parameter estimation techniques of NARMAX models, with a particular emphasis on enhancing computational efficiency and reducing the need for manual intervention. In [53], a recursive extended least squares (RELS) algorithm was proposed, allowing for real-time identification of NARMAX models. This development was especially beneficial for applications that required fast identification, such as adaptive control systems. In [54], A novel forward orthogonal search (FOS) technique, which maximizes the overall dependency (MOD), was introduced to identify significant variables and select a subset from a collection comprising all original variables. The FOS-MOD algorithm operates through a straightforward and easily implementable mechanism, generating efficient subsets while preserving the interpretability of each feature in relation to the original data. In [55], an integrated forward orthogonal search algorithm, assisted by squared correlation and mutual information measures, to effectively identify the model structure of complex nonlinear systems, was proposed. The proposed approach shows potential in identifying accurate and parsimonious models, but a more comprehensive comparison with alternative methods and an investigation into its robustness would further solidify its value and applicability. In [56], a novel approach for sparse model identification in nonlinear system identification by combining the forward orthogonal regression algorithm with mutual information was introduced. The proposed method demonstrates potential in constructing accurate and parsimonious models, but further investigation and comparison with alternative approaches would provide a more comprehensive understanding of its true benefits and limitations.

Another significant milestone in the history of NARMAX was the introduction of different nonlinear form of NARMAX models in [57-61]. These models incorporated wavelet transformations, rational functions and radial basis function network, some powerful mathematical tools for analysing nonlinear system, into the NARMAX framework. The resulting NARMAX models were better suited for handling nonstationary and multiscale phenomena, further extending the applicability of NARMAX to a broader range of systems. The detailed discussion of these NARMAX models is in next section.

In recent years, the research of NARMAX have focused on improving parameter estimation and variable selection algorithm efficiency, incorporating advanced machine learning techniques, developing hybrid approaches, ensuring model interpretability, and expanding applicability to various domains [62-66]. In [67], a novel iterative forward orthogonal least squares regression (IFOLS) algorithm for identifying nonlinear systems with non-persistent excitation was present. The proposed IFOLS method demonstrates improved performance compared to traditional techniques, offering a promising alternative for nonlinear system identification in challenging excitation scenarios. In [68], a novel Maximum Relevance-Minimum Multicollinearity (MRmMC) method for feature selection and ranking was proposed, which addresses the challenges posed by multicollinearity in predictive modelling. However, the MRmMC method involves an iterative process for feature selection and ranking, which could lead to increased computational complexity, especially for high-dimensional datasets or large-scale problems. In [28], a novel mask attention-based NARMAX (MAB-NARMAX) modelling method was proposed to reduce the model structure uncertainty in nonlinear system identification, where the mask attention mechanism came from the attention-based deep neural network (Transformer).

As the research on NARMAX continues to evolve, new applications of NARMAX have also been explored. In [69], a novel non-destructive testing method based NARMAX system identification method for mechanical structural health assessment was proposed. By obtaining frequency response of NARMAX model from time domain, the structural health situation was analysed effectively with the factors affecting the situation of mechanical structures. In [70], a diagonal recurrent neural network (DRNN) based NARMAX method was applied to accurately predict the behaviour of magnetic shape memory alloy (MSMA) actuators and describe the hysteresis nonlinearity degree of the device. In [34], to understand the influenza and influenza-like illness and provide accurate and timely forecasts of seasonal influenza, NAMRAX was introduced as a sparse, interpretable and transparent (SIT) model in the medical and healthcare area. By comparing with state-of-the-art methods, NARMAX can not only provide accurate forecasts, but also present explainable relationship between system inputs and outputs and the interaction of the ‘black box’ system. In [71], the NARMAX method was used to build a nonlinear dynamic model of cortical responses to wrist perturbations. By applying NARMAX, a common model structure across all participants, which could provide a reference for future clinical studies investigating abnormal cortical responses associated with sensorimotor impairments was identified. The study found that the measured cortical response is a mixed outcome of the nonlinear transformation of external inputs and local neuronal interactions or inherent neuronal dynamics at the cortex.

In conclusion, recent studies have demonstrated the advantages of NARMAX models in addressing complex nonlinear problems. The key benefits of NARMAX models include their ability to provide accurate predictions, reveal explainable relationships between system inputs and outputs, and shed light on the interactions within "black box" systems. Researchers are continuously exploring new applications and refining existing methodologies to improve the performance, reliability, and robustness of NARMAX models. The ongoing development and research in NARMAX models hold significant implications for various domains. By harnessing the power of NARMAX modelling techniques, researchers and practitioners can gain deeper insights into complex systems, improve decision-making processes, and develop innovative solutions to pressing challenges. As the field continues to progress, it is expected that the versatility and adaptability of NARMAX models will further establish them as a valuable tool for addressing diverse nonlinear problems across a myriad of disciplines.

2.1.2 Typical version of NARMAX models

Based on the definition of NARMAX, the general NARMAX model is normally defined as [46]:

$$\begin{aligned} y(k) = F[&y(k-1), y(k-2), ..., y(k-n_y), \\ &u(k-d), u(k-d-1), ..., u(k-d-n_u) \\ &e(k-1), e(k-2), ..., e(k-n_e)] + e(k) \end{aligned} \tag{2.1}$$

where $y(k)$, $u(k)$, and $e(k)$ are the system output, input and noise signals, respectively; n_y, n_u, and n_e are the maximum lags for the system output, input, and noise signals, respectively; $F[\cdot]$ is some typical nonlinear function, and *d* is a time delay for the input signals, which is typically set to $d = 1$.

Throughout the years, to enhance the performance of NARMAX, researchers focused on improving the model structure selection and parameter estimation processes in NARMAX models since the successful application in various real-world problems.

Power-form polynomial NARMAX model

Initially, the power-form polynomial model was proposed as the NARMAX model form [29], which means that the unknown nonlinear mapping $F[\cdot]$ is the power-form polynomial function. Therefore, the power-form polynomial representation of NARMAX model is defined as [1]:

$$\begin{aligned} y(k) = \theta_0 + \sum_{i_i=1}^{n} f_{i_1}\left(x_{i_1}(k)\right) + \sum_{i_1=1}^{n}\sum_{i_2=i_1}^{n} f_{i_1 i_2}\left(x_{i_1}(k), x_{i_2}(k)\right) + \cdots \\ + \sum_{i_1=1}^{n} \cdots \sum_{i_t=i_{t-1}}^{n} f_{i_1 i_2 \cdots i_l}\left(x_{i_1}(k), x_{i_2}(k), \ldots, x_{i_\ell}(k)\right) + e(k) \end{aligned} \tag{2.2}$$

$$f_{i_1 i_2 \cdots i_m}\left(x_{i_1}(k), x_{i_2}(k), \ldots, x_{i_m}(k)\right) = \theta_{i_1 i_2 \cdots i_m} \prod_{k=1}^{m} x_{i_k}(k), 1 \le m \le \ell \tag{2.3}$$

$$x_m(k) = \begin{cases} y(k-m) & 1 \le m \le n_y \\ u\left(k-\left(m-n_y\right)\right) & n_y + 1 \le m \le n_y + n_u \\ e\left(k-\left(m-n_y-n_u\right)\right) & n_y + n_u + 1 \le m \le n_y + n_u + n_e \end{cases} \tag{2.4}$$

where $y(k)$, $u(k)$, and $e(k)$, n_y, n_u , and n_e are defined as before; ℓ is the degree of polynomial nonlinearity, $\theta_{i_1 i_2 \cdots i_m}$ are parameters, $n = n_y + n_u + n_e$.

The power-form polynomial NARMAX model has shown great performance in terms of flexibility, adaptability, and interpretability in various applications. Their versatility and adaptability allow for accurate modelling and prediction of system behaviours, providing valuable insights and improved control strategies. Moreover, the power-form polynomial NARMAX models offer a balance between model complexity and interpretability, making them a powerful tool for researchers and practitioners seeking to understand and optimize real-world systems with inherent nonlinearities. These models have shown exceptional performance in fields such as engineering [72], environmental sciences [73], economics [74], and biomedical engineering [75].

Polynomial-based models can exhibit certain limitations, particularly when attempting to capture highly nonlinear behaviours. This is because their fixed structure may not be flexible enough to accurately represent the complex dynamics inherent in severely nonlinear systems. Therefore, alternative nonlinear function forms were introduced to model structure selection. Detailed introduction of polynomial NARMAX is discussed in Section 3.

Rational NARMAX model

Rational NARMAX models are a type of NARMAX models that use rational functions to nonlinear relationships within a dynamic system [30, 59], which is defined as a ratio of two polynomial functions, making it capable of capturing more complex or certain types of singular or near singular behaviours. The rational NARMAX model is defined as [59]:

$$y(k) = \frac{B\left[y(k-1), \ldots, y\left(k-n_y\right), u(k-1), \ldots, u\left(k-n_u\right), e(k-1), \ldots, e\left(k-n_e\right)\right] + e(k)}{A\left[y(k-1), \ldots, y\left(k-n_y\right), u(k-1), \ldots, u\left(k-n_u\right), e(k-1), \ldots, e\left(k-n_e\right)\right] + e(k)} \tag{2.5}$$

where $y(k)$, $u(k)$, and $e(k)$, n_y, n_u , and n_e are defined as before; $A[\cdot]$ and $B[\cdot]$ are functions in polynomial forms. It is assumed that $A[\cdot] \neq 0$. The rational model offers a broader representation compared to the standard polynomial NARMAX model. In fact, setting the denominator polynomial

$B[\cdot]$ to be equal to 1 makes the rational model equivalent to the polynomial NARMAX model [76]. Furthermore, various other rational models, such as integral, recursive, and output-affine models, can be considered as special cases of the rational model [1].

Rational NARMAX models offer increased flexibility and accuracy by approximating a wider variety of nonlinear behaviours with compact, parsimonious representations. However, they come with certain drawbacks related to computational complexity, stability, and interpretability. The choice of using rational NARMAX models depends on the specific problem and the trade-offs between accuracy, complexity, and interpretability.

Wavelet-based NARMAX models

Wavelet-based NARMAX models employ wavelet functions as nonlinear elements within the NARMAX structure to capture the inherent nonlinearities in dynamic systems [77]. Wavelets are mathematical functions with time and frequency localization properties [78], which enable these models to represent non-stationary or transient behaviours more effectively [79]. Additionally, wavelet functions allow for multiresolution analysis, providing a framework for analysing complex system dynamics at multiple levels of detail [80].

In the wavelet-based NARMAX model, the input signal $u(t)$, output signal $y(t)$, and noise signal $e(t)$ at different time lags are represented using a set of wavelet basis functions [81]. The Wavelet-based NARMAX model can be converted into a linear-in-the-parameter form [81]:

$$y(t) = \sum_{m=1}^{M} \theta_m p_m(t) + e(t) \tag{2.6}$$

where $p_m(t)\ (m = 1, 2, \ldots, M)$ are predictors/model terms generated by the scaled and shifted variants of certain mother wavelets; θ_m is the related parameters. The identification and estimation process involves optimizing these coefficients to minimize the prediction error. In most instances, representing a nonlinear dynamical system only necessitates a limited number of significant basis functions, chosen from an extensive library of available functions [58]. Wavelet-based NARMAX models offer a powerful approach for modelling nonlinear dynamic systems, with advantages in time-frequency localization, multiresolution analysis, adaptability to discontinuities, sparse representation, and improved model performance. These characteristics make them particularly suitable for analysing complex systems with non-stationary, time-varying, or transient behaviours [82]. Due to their unique properties, wavelet-based NARMAX models have been successfully applied in various fields, including solar wind modelling and prediction [83], ecological environmental modelling [84], and high tide forecast and modelling [85], among others.

Radial basis function (RBF)-based NARMAX models

Radial basis function (RBF)-based NARMAX models incorporate radial basis functions as nonlinear elements within the NARMAX structure to represent complex nonlinear relationships in dynamic systems [86, 87]. RBFs are a versatile class of functions that can capture a wide range of nonlinear behaviours, making them suitable for modelling various systems [88, 89]. These models are particularly useful in handling systems with localized behaviours, as RBFs are characterized by their limited regions of influence, allowing the model to adapt to changes in the system dynamics more effectively [90]. Similarly, the RBF-based NARMAX can be defined as:

$$y(t) = \hat{f}(\mathbf{x}(t)) + e(t) = \sum_{m=1}^{M} \theta_m \phi_m(\mathbf{x}(t)) + e(t) \tag{2.7}$$

where M is the total number of candidate regressors, $\phi_m(\mathbf{x}(t)), (m = 1, 2, \ldots, M)$ are the model regressors and θ_m are the model parameters. Take an RBF network with Gaussian kernels for the nonlinear form $\hat{f}$ as an example:

$$\hat{f}(\mathbf{x}(t)) = \sum_{i=1}^{M} \theta_i \varphi_i\left(\mathbf{x}(t); \boldsymbol{\sigma}_i, \mathbf{c}_i\right) \tag{2.8}$$

$$\varphi_i\left(\mathbf{x}(t); \boldsymbol{\sigma}_i, \mathbf{c}_i\right) = \exp\left\{-\frac{1}{2}\left[\mathbf{x}(t) - \mathbf{c}_i\right]^{\mathrm{T}} \Lambda_i^{-1} \left[\mathbf{x}(t) - \mathbf{c}_i\right]\right\} \tag{2.9}$$

where φ_i is the general standard Gaussian kernel; $\sigma_i = \left[\sigma_{i,1}, \ldots, \sigma_{i,d}\right]^{\mathrm{T}}$ are the dilation parameters to define the Gaussian kernel width; $\mathbf{c}_i = \left[c_{i,1}, \ldots, c_{i,d}\right]^{\mathrm{T}}$ are the location parameters that determine the kernel positions; Λ_i are positive diagonal matrices [86]. Certainly, there are more types of kernel functions to comprise the RBF-based NARMAX models.

The RBF-based NARMAX models involve optimizing the RBF parameters, such as centers and widths of the kernel function in the RBF network, to minimize the prediction error [91]. While RBF-based NARMAX models offer a flexible and powerful approach for modelling nonlinear dynamic systems, they also come with certain drawbacks, such as increased computational complexity, challenges in selecting appropriate RBF parameters, and potential overfitting risk especially when using a large number of basis functions.

2.1.3 Typical variable selection algorithm

The Orthogonal Least Squares Estimator and the Error Reduction Ratio

The Orthogonal Least Squares (OLS) estimator, initially developed in the late 1980s, was designed for parameter estimation in NARMAX models, which have a linear-in-the-parameters representation [48, 72, 92, 93]. The fundamental concept behind the OLS estimator involves defining and incorporating an auxiliary model with terms orthogonal over the estimation dataset. This approach allows for the sequential and independent estimation of individual model coefficients, even in the presence of correlated measurement noise. Afterward, the auxiliary model is mapped back to the original system model as a final step. Applying this straightforward algorithm repeatedly not only yields unbiased estimates for each coefficient in succession but also reveals the contribution that each model term makes to the system output variance.

Consider a general linear-in-the-parameters NARMAX model defined as:

$$y(k) = \sum_{i=1}^{M} \theta_i p_i(k) + e(k) \tag{2.10}$$

where $y(k), k = 1, 2, ..., N$ is the system output, $p_i(k) = p_i(\mathbf{x}(k))$, $i = 1, 2, ..., M$ are the regressors transformed by the combination of complex functions of selected lagged model variables from the vector $\mathbf{x}(k) = \left[x_1(k), x_2(k), \ldots, x_n(k)\right]^{\mathrm{T}}$, where $x_l(k)$, $l = 1, 2, ...n$ are the lagged system input signal $[u(k-1), ..., u(k-n_u)]$, lagged system output signal $[y(k-1), ..., y(k-n_y)]$ and the lagged noise signal $[e(k-1), ..., e(k-n_e)]$, $n = n_u + n_y + n_e$, θ_i are the model parameters, $e(k)$ is the prediction error. The assumption is that each of the regressors, denoted as $p_i(k)$, is independent of any model parameters, illustrated as θ_i. This independence is represented mathematically as $\partial p_i(k)/\partial\theta_j = 0$, $i = 1, 2, ..., M$ and $j = 1, 2..., M$.

Ideally, the goal of OLS algorithm is to convert the initial non-orthogonal regression model (2.10) into an alternative regression model featuring orthogonal regressors or terms. In other words, the aim is to establish an auxiliary model with orthogonal terms to one another.

$$y(k) = \sum_{i=1}^{M} g_i w_i(k) + e(k) \tag{2.11}$$

where g_i are the orthogonal regression model parameters; $w_i(k)$, $i = 1, 2, ..., M$ are orthogonal vectors:

$$\sum_{k=1}^{N} w_i(k) w_j(k) = \begin{cases} d_i, & i = j \\ 0, & i \neq j \end{cases} \tag{2.12}$$

with $d_i = \sum_{k=1}^{N} w_i^2(k) \neq 0$. As the regressors in model (2.10) are assumed to be independent, therefore, the orthogonalization mechanism is defined as:

$$\begin{cases} w_1(k) = p_1(k) \\ w_2(k) = p_2(k) - a_{1,2} w_1(k) \\ w_3(k) = p_3(k) - a_{1,3} w_1(k) - a_{2,3} w_2(k) \\ \quad\vdots \qquad\quad \vdots \qquad\qquad \vdots \\ w_m(k) = p_m(k) - \sum_{r=1}^{m-1} a_{r,m} w_r, m = 2, 3, \ldots, M \end{cases} \tag{2.13}$$

where

$$a_{r,m} = \frac{\sum_{k=1}^{N} p_m(k) w_r(k)}{\sum_{k=1}^{N} w_r^2(k)}, 1 \leq r \leq m-1 \tag{2.14}$$

$p_1(k)$ is the first model term in model (2.10), which is determined by the error reduction ratio (ERR).

Accordingly, it can be derived that:

$$g_i = \frac{\sum_{k=1}^{N} y(k) w_i(k)}{\sum_{k=1}^{N} w_i^2(k)}, i = 1, 2, \ldots, M \tag{2.15}$$

and the parameters θ_i in model (2.10) can be derived from the orthogonal regression model parameters:

$$\begin{cases} \theta_M = g_M \\ \theta_{M-1} = g_{M-1} - a_{M-1,M} \theta_M \\ \theta_{M-2} = g_{M-2} - a_{M-2,M-1} \theta_{M-1} - a_{M-2,M} \theta_M \\ \quad\vdots \qquad\quad \vdots \qquad\qquad \vdots \\ \theta_m = g_m - \sum_{j=m+1}^{M} a_{m,j} \theta_j, m = M-1, M-2, \ldots, 1 \end{cases} \tag{2.16}$$

During the orthogonalization mechanism, one key area is to select the orthogonal terms in model (2.11). A straightforward yet efficient method for identifying a subset of important regressors, the error reduction ratio (ERR), was proposed to select the most significant terms based on their ERR values. For each orthogonal terms, their ERR values ERR_i is defined as [48]:

$$ERR_i = \frac{g_i^2 < \mathbf{w}_i, \mathbf{w}_i >}{< \mathbf{y}, \mathbf{y} >} \times 100\% = \frac{< \mathbf{y}, \mathbf{w}_i >^2}{< \mathbf{y}, \mathbf{y} >< \mathbf{w}_i, \mathbf{w}_i >} \times 100\%, i = 1, 2, ..., M \tag{2.17}$$

where $\mathbf{w}_i$ is the orthogonalized vector and y is the system output vector.

The OLS estimator boasts higher efficiency compared to the classical least squares algorithm. Its power stems from its capacity to rank and select the most significant or relevant model terms from a potentially large initial set of candidate terms. This capability is often crucial for subset model selection, and in nonlinear dynamic model identification, it assists with model interpretation and frequently improves the generalization properties of the fitted models. However, there are some certain challenges for the basic OLS algorithm, like the location of the first orthogonalized term in model (2.13), the selection of incorrect or redundant model term. Thus, various variable selection algorithms based on the basic OLS algorithm were present.

The Forward Regression OLS Algorithm

The Forward Regression with OLS (FROLS) algorithm, also referred to as the Orthogonal Forward Regression (OFR) algorithm, is widely recognized for its ability to select and rank significant terms [92]. Consequently, it has emerged as a standard algorithm frequently employed in nonlinear model structure detection [94, 95].

Taking the polynomial NARX model of nonlinear degree ℓ as an example, this model can also be considered as the linear-in-the-parameters representation:

$$y(k) = \sum_{m=1}^{M} \theta_m p_m(k) + e(k) \tag{2.18}$$

where θ_m $(m = 1, 2, ..., M)$ are model parameters and $p_m(k)$ are model terms defined as

$$p_m(k) = y(k - m_{y,1}) \cdots y(k - m_{y,my}) u(k - m_{u,1}) \cdots u(k - m_{u,mu}) \tag{2.19}$$

$$\begin{gathered} 1 \le m_{y,1} \le m_{y,2} \le ... \le m_{y,my} \le n_y \\ 1 \le m_{u,1} \le m_{u,2} \le ... \le m_{u,mu} \le n_u \end{gathered} \tag{2.20}$$

With $m = 1, 2, ..., M$, $my, mu \ge 0$; $p_m(\cdot) \equiv 1$ is related to a constant term; $my = 0$ indicates that $p_m(k)$ contains no $y(\cdot)$ variables; $mu = 0$ indicates that $p_m(k)$ contains no $u(\cdot)$ variables. Let $\mathbf{p}_m = [p_m(1), p_m(2), ..., p_m(N)]^T$ be a *m*-th candidate model term vector and $D = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M\}$ be a candidate dictionary, which is normally redundant. The objective of the FROLS algorithm is to find a subset $D_{M_0} = \left\{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{M_0}\right\} = \left\{\mathbf{p}_{i_1}, \mathbf{p}_{i_2}, \ldots, \mathbf{p}_{i_{M_0}}\right\}$ from the dictionary *D*, where $\boldsymbol{\alpha}_m = \mathbf{p}_{i_m}, i_m \in \{1, 2, \ldots, M\}, k = 1, 2, ..., M_0$. Therefore, based on the selected subset, the NARX model in equation (2.18) can be transformed as:

$$\mathbf{y} = \theta_1 \boldsymbol{\alpha}_1 + \cdots + \theta_{M_0} \boldsymbol{\alpha}_{M_0} + \mathbf{e} \tag{2.21}$$

The detailed procedure of FROLS is discussed in Chapter 3.

The Forward Orthogonal Search Algorithm by maximising the overall dependency (FOS-MOD) Algorithm

The FOS-MOD algorithm aims to maximize the overall dependency (MOD) in order to identify significant variables and select a subset from a library containing all original variables [54]. The FOS-MOD algorithm begins by converting a general feature subset selection problem into a multivariate regression problem using the concepts of pseudo-response and pseudo-regression. The core idea behind FOS-MOD is to ensure that the selected subset adequately represents the overall features in the original measurement space. This is achieved by explaining the variation in the overall features using the selected subset with an acceptable accuracy, surpassing a predefined threshold. FOS-MOD generates a ranked list of selected features, ordered based on their percentage contribution to representing the overall features.

Denote $S=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ as the dataset formed by *N* observations and *n* attributes in the measurement space, where $[\mathbf{x}_1(k),\mathbf{x}_2(k),\ldots,\mathbf{x}_n(k)]$ is the *k-th* instance signal vector; $\mathbf{x}_j=[x_j(1),x_j(2),\ldots,x_j(N)]^{\mathrm{T}}$ is the *j-th* attribute of the observation vector. The goal of FOS-MOD algorithm is to find one subset $S_d=\{\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_d\}=\{\mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_d}\}$, to represent the features, where $\mathbf{z}_m=\mathbf{x}_{i_m}, i_m\in\{1,2,\ldots,n\}, m=1,2,...,d$ and $d\leq n$. The fundamental requirement is to ensure that the overall features in the measurement space are adequately represented using S_d. This is achieved by confirming that the variation in the overall features can be explained by the elements of S_d with a satisfactory level of accuracy, like:

$$\mathbf{x}_i=f_i(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_d)+\mathbf{e}_i \tag{2.22}$$

where f_i represents an unknown function that describes the relationship between the *i-th* variable and the selected variables, and $\mathbf{e}_i$ is an unobservable error accounting for the discrepancy in the approximation. The effectiveness of the selected subset S_d can be assessed by examining its ability to approximate individual features $\mathbf{x}_i(i=1,2,\ldots,n)$ in the measurement space. For instance, one can evaluate the percentage of variation in $\mathbf{x}_i$ that can be explained by the elements in S_d. Assuming the percentage of variation in $\mathbf{x}_i$ accounted for by the elements in S_d is $p_i(d)$, the average percentage of variation in the overall features $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ that can be explained by S_d can be defined as $p(d)=(1/n)\sum_{i=1}^{n}p_i(d)$. If the percentage $p(d)$ exceeds a given threshold, S_d can be considered as the final subset. Otherwise, additional significant variables need to be included in S_d. The detailed procedure of FOS-MOD is discussed in [1, 54].

Recently, the OLS algorithm has attracted more attention due to its efficiency and accuracy. In [96], the recursive OLS (ROLS) algorithm is utilized to optimize the radial basis probabilistic neural networks. The ROLS algorithm is introduced to recursively train the weights between the hidden layer and the output layer of the RBNN as the algorithm holds the ability to process a large number of training samples. In [97], a multiple OLS algorithm (m2OLS) was proposed as an extension of the typical OLS algorithm. By using the proposed m2OLS algorithm, the k-sparse signal vector can be recovered from the noiseless measurement. In [98], a greedy OLS algorithm based on the Householder transformation is developed to deal with the multivariable Wiener system identification. The proposed greedy OLS algorithm utilizes the Householder transformation strategy to select the most active information in the potential variables to lower the computation complexity and the system sensitivity. In [99], an accelerated OLS (AOLS) algorithm is introduced, which is a greedy scheme designed for accurate reconstruction of sparse vectors from random linear combinations of their components. The AOLS enables the quantification of relationships between objectives in multi-objective reinforcement learning, which addresses the problem of inferring a sparse vector from random linear combinations of its components, emphasizing efficient and accurate reconstruction in large-scale scenarios.

Collectively, these developments underscore the versatility and evolving nature of the OLS algorithm, highlighting its continued relevance and adaptability to various domains, from neural network optimization and signal processing to system identification, by incorporating strategies like recursive training, sparsity exploitation, efficient variable selection, and computational acceleration.

2.1.4 Summary

The evolution of NARMAX models has undergone a continuous series of refinements and innovations, which have been aimed at enhancing their performance, robustness, and applicability across diverse domains. Originally, NARMAX models were developed as a generalization of linear AMARX models. However, over time these models have evolved considerably, and have incorporated advanced mathematical tools, machine learning techniques, and hybrid approaches specifically to address the challenges associated with nonlinear systems. Additionally, researchers have been exploring novel ways to improve parameter estimation, variable selection, and model interpretability, which are all essential components in the development of a reliable and accurate NARMAX model. Given these ongoing efforts, it is anticipated that NARMAX models will maintain their position as a highly valuable tool for nonlinear system identification and prediction, ultimately expanding their impact and applicability across various fields.

2.2 Model uncertainty

In any statistical model, uncertainty is an inevitable factor that must be considered. One of the most challenging aspects of building a model is to understand and quantify the level of uncertainty in the model. In the context of this literature review, we are concerned with the concept of model uncertainty, which arises when multiple models can be used to explain the data. The problem of model uncertainty is particularly significant in complex models, where there are many parameters and interactions. The aim of this section is to critically review the literature on model uncertainty, with a focus on exploring different methods for quantifying and managing model uncertainty. We will also discuss the implications of model uncertainty for decision-making using statistical models and highlight areas where further research is needed.

2.2.1 Brief introduction of model uncertainty

Model uncertainty pertains to the degree of ambiguity linked with the selection of a specific model that represents a real-world occurrence [100]. This primarily arises when several competing models can equally explain the observed data, or in cases of uncertain model assumptions, a common challenge in research [101-104]. In system identification, the choice of the most suitable model plays a vital role in generating valid predictions and deductions [105, 106]. Nonetheless, model selection is an arduous task, primarily when multiple competitive models exist, each having its unique strengths and limitations [107-109]. The challenge of model uncertainty is especially prominent when handling complex dynamic datasets, where numerous variables and interactions come into play [110]. Therefore, recognizing model uncertainty is critical to the success of data-driven models. The ability to handle and mitigate model uncertainty ensures credible inference and predictions, fostering widespread applicability of the model. Ignoring model uncertainty can lead to challenges such as overfitting, biased results and incorrect conclusions, reducing the usefulness and reliability of the model.

2.2.2 The category of model uncertainty

Model uncertainties result from various factors, such as the scale of data analysis, complexity of the phenomenon, accuracy of measurements, and validity of assumptions underlying the models [111-113]. This section presents a detailed account of the diverse sources of uncertainties involved in modelling and the standard methods utilized for their mitigation.

Data-Related Uncertainty:

Data-related uncertainties refer to the uncertainty, ambiguity, or lack of clarity in the data that could affect the accuracy and reliability of the analysis [114]. Incomplete or noisy data, measurement errors, and outliers are common limitations that could create uncertainties in the data [115, 116]. For instance, if the dataset is incomplete, the results of the analysis may not accurately represent the population, as there could be some critical observations missing [117]. Similarly, if the data contains measurement errors, it could result in incorrect conclusions or interpretations [118]. Finally, outliers or anomalies in the data could also create uncertainties, as they could be due to data-entry errors, or they could represent legitimate data points that are informative and should not be excluded from the analysis [119].

Model parameter uncertainty

Model parameter uncertainty is a fundamental challenge in system identification and modelling. The uncertainty arises from the fact that the values of the parameters in models are not known exactly, and this uncertainty can have a significant impact on the accuracy and reliability of the models [120]. Model parameters are estimates that describe the relationships between variables in the model, influencing the conclusions drawn from analysing data [121]. Estimating model parameters necessarily involves a degree of error due to imperfect modelling assumptions and limitations in data collection and analysis. This intrinsic error causes model parameter uncertainty, which can affect the accuracy and reliability of the conclusions drawn from the data [122]. There are different types of elements to leading to model parameter uncertainty, including non-stationary dataset, model misspecification, heteroscedasticity, etc [123-126].

- Non-stationary data: Non-stationary time series may exhibit dynamic changes in statistical properties, such as trends, seasonal effects, or abrupt shifts in the underlying data generating process, which can pose challenges to accurately estimating model parameters [127]. These changes may lead to the evolution of the relationships between the variables in the model over time, and therefore, require specialized techniques to effectively capture and model the dynamic nature of the data generating process. [128].
- Limited data: Insufficient data may yield imprecise and unreliable parameter estimates, as the limited sample size may not provide adequate information to accurately infer the true values of the model parameters [129]. In such situations, careful data pre-processing and modelling are crucial to minimize bias and variance in the parameter estimates, and to select an appropriate model that can effectively capture the underlying data generating process [130].
- Model misspecification: Misspecified models can also result in biased or inefficient parameter estimates, which can have a significant impact on the accuracy of the model predictions [131]. Such biases can arise due to omitted variables, errors in variable measurement, or nonlinearity in the data generating process [132].
- Heteroscedasticity: Heteroscedasticity denotes the attribute of a dataset wherein the dispersion of residuals is nonuniform across the domain of the predictor variable [133]. With the fluctuations in the predictor variable, the residual magnitudes exhibit variation, yielding an

unequal distribution of error terms. This irregularity complicates the estimation and interpretation of associations amongst variables in statistical modelling [134]. Consequently, heteroscedasticity engenders parameter uncertainty, potentially affecting the dependability of the model's prognostications and deductions [135].

- Outliers: Outliers are characterized as distinct data points that diverge markedly from the prevailing observations in a dataset [136]. The presence of these atypical values can engender considerable uncertainty in the model parameters [137]. This uncertainty arises from the disproportionate impact outliers impart on the estimation of model coefficients, thereby potentially compromising the accuracy and robustness of the model [138]. As such, in the realm of statistical analysis, it is imperative to diligently scrutinize and, if necessary, address these anomalous observations to ensure the reliability of the model's inferences and predictions.
- Correlated predictors/ Multicollinearity: Multicollinearity emerges in statistical modelling when predictor variables manifest significant intercorrelations, complicating the accurate estimation of each predictor's individual impact [139]. The entwined nature of correlated predictors leads to unstable coefficient estimates, hampering the reliability and interpretability of model inferences [140]. Thus, addressing multicollinearity is vital to ensuring the dependability of parameter estimates and model-based conclusions in the realm of statistical analysis.

System identification and modelling require careful consideration of model parameter uncertainty to ensure the validity and reliability of statistical models and the conclusions drawn from them. Effective accounting for sources of model parameter uncertainty is essential to improve data collection and analysis strategies, resulting in more accurate and reliable inferences about the underlying data-generating process. The development of models that capture the complexity of real-world systems can ultimately facilitate more robust decision-making, making model parameter uncertainty a critical factor in system identification and modelling.

Model structural uncertainty

Model structural uncertainty, a crucial aspect of system identification and modelling, encompasses the ambiguity or lack of clarity when selecting an appropriate model structure to represent a given system or phenomenon [141]. Model structural uncertainty emerges as a consequence of the simplifications and approximations employed in model equations, alongside the incomplete or imprecise portrayal of real-world processes and interactions [142, 143]. These uncertainties can stem from limitations in the model structure or incorrect assumptions that fail to capture the full complexity of the investigated system.

Addressing model structural uncertainty is essential for ensuring accurate and reliable predictions, simulations, or control strategies derived from the model. The category of model structure uncertainty is diverse. The typical model structure uncertainty is shown as:

- Functional form: Model structure uncertainty originates from the selection of a simulated function in the context of statistical modelling, potentially engendering biased or inconsistent predictions [144]. This issue emerges when the authentic underlying association between variables is inaccurately captured by the employed function. As a result, it is imperative to meticulously select an appropriate functional form to accurately represent the intrinsic relationships among variables, thereby mitigating the detrimental consequences of biased or inconsistent prognostications [145].

- Complexity: Complex systems often necessitate the development of intricate models that encompass multiple variables and their interactions [146]. The elaborate nature of such systems presents considerable challenges in accurately specifying the suitable model structure. Consequently, researchers must exercise diligence in model selection and validation to ensure a robust representation of the complex system's underlying dynamics and relationships [147, 148].
- Lack of knowledge: Model structure uncertainty can arise due to a lack of knowledge or understanding of the underlying system being modelled, which can make it difficult to choose an appropriate modelling approach or to specify the appropriate variables to include in the model [149].

A comprehensive comprehension of the origins and varieties of model structural uncertainty empowers scholars to make judicious decisions during model development and selection, ultimately augmenting the model's generalization capability, robustness, and reliability. Furthermore, the scrutiny of model structural uncertainty enables researchers to identify potential biases, discrepancies, or constraints inherent in the model, which can be rectified through model refinements, ensemble modelling, or alternative uncertainty quantification methodologies. In summary, the investigation and contemplation of model structural uncertainty serve a pivotal function in the formulation and implementation of dependable and accurate models across diverse domains, thereby fostering enhanced decision-making and system control.

Criteria uncertainty

Criterion uncertainty, within the realm of system identification, encompasses the equivocality or variability associated with the choice and application of germane metrics or assessment benchmarks for appraising the aptness and effectiveness of a particular model [150]. This form of uncertainty arises owing to the multitude of available evaluative indices, the discretionary nature of selecting a suitable criterion, or the inherent compromises between antithetical objectives such as model complexity and generalizability [151]. The recognition and management of criterion uncertainty are thus essential to ensure that the model evaluation process accurately reflects the model's performance, ultimately contributing to the development and selection of robust and reliable models in diverse fields of study.

Acknowledging and tackling criterion uncertainty is vital for ensuring the robustness and dependability of model evaluation and selection procedures. By discerning and alleviating criterion uncertainty, scholars can attain a more nuanced comprehension of model performance across diverse aspects, thereby facilitating well-informed decisions throughout model development, enhancement, and implementation [152]. Moreover, addressing criterion uncertainty can unveil potential limitations or shortcomings in a model's performance, which can be amended through additional model improvements or by examining alternative model structures or parameterizations [153]. In summation, the contemplation and examination of criterion uncertainty are of paramount importance for the formulation of accurate and reliable models, ultimately fostering enhanced decision-making and system control.

2.2.3 Model uncertainty in NARMAX

Uncertainty in NARMAX models can stem from various sources. Based on the analysis of category of model uncertainty, the uncertainty in NARMAX modelling is from measurement noise, model structure inaccuracies, parameter estimation errors and the blur of model validation. Addressing uncertainty in NARMAX models is crucial for ensuring reliable predictions, model validation, and control design.

Data-related uncertainty in NARMAX modelling

Data related uncertainty is normally introduced by the noisy or incomplete observation, inherent variability in data-generating processes, and limitations in system assumptions, which is more related to the quality of the dataset. The most direct approach is to clean the dataset. Before NARMAX modelling, data cleaning methods such as outlier detection, data imputation, and data smoothing can be used to reduce noise and missing values in the data. This helps to mitigate the impact of data uncertainty on modelling results.

In [1], several data uncertainty in NARMAX modelling has been discussed. Firstly, checking and eliminating the input-output data for outliers prior to system identification and to assess any trends is imperative, as minor trends in the data can pose challenges when employing correlation, spectral analysis, and parameter estimation techniques, and thus should be removed. Methods such as least squares trend fitting and removal can be utilized, or data differencing can be applied. However, it is important to note that differencing may amplify the noise level in the signals, as it essentially involves differentiation.

Secondly, Attempting to filter the data for noise removal or enhancement of poorly designed or executed experimental data is typically ineffective and should be circumvented. Determining the distinction between noise and valuable system-generated information prior to identification is unattainable. Consequently, designing filters to attenuate noise might inadvertently eliminate dynamic information about the underlying system. This could result in an identified model that conforms to the dataset, despite the dataset lacking comprehensive system information. Therefore, the identified model will consistently be inaccurate and insufficient.

Finally, the adage "garbage in, garbage out" holds significant relevance in the context of system identification. If the data utilized for identification is substandard due to a poorly designed experiment, inappropriate data sampling or pre-treatment, or substantial noise that cannot be sufficiently addressed through a noise model, the resulting identification is likely to be unsatisfactory. The outcome remains poor irrespective of the algorithm employed, the type of models fitted, or the computing power harnessed for data processing.

Therefore, applying NARMAX to real applications requires exploring and experiments as the quality of dataset in the real problems would differ. Also, the lack of understanding of the real problems would introduce the misjudgement of the noise and signal. Thus, to deal with the data-related uncertainty, exploration and exploitation would be the appropriate method.

Parameter uncertainty in NARMAX modelling

NARMAX models have various advantages in capturing complex relationships between system inputs and outputs and generating reliable and robust system models to present accurate forecast. Still, the parameter uncertainty can adversely impact the accuracy and the reliability of the NARMAX models. The estimated parameters in NARMAX models can be affected by various factors, like the dataset of systems, the complex property of the time-vary dynamic systems and the undiscovered understanding of the complex system. Research on parameter uncertainty in NARMAX has been a crucial step in recent years.

In [64], a time-varying(TV)-NARX model with multiwavelet basis functions and ultra-orthogonal forward regression (UOFR) algorithm was proposed to effectively tracks the changing TV parameters in simulated systems and EEG signal modelling. By expanding the TV coefficients of TV-NARX models with multiwavelet basis functions, the model was transformed into a time-invariant regression problem. Meanwhile, the novel UOFR algorithm aided by mutual information (MI) could identify the most parsimonious model structures and estimate the model parameters. Thus, the parameter uncertainty in TV system could be weakened by converting the TV system into a time-invariant system. In [154], to provide a robust path following control following of an Autonomous Underwater Vehicle (AUV), A NARMAX-based self-tuning PID controller was designed for controlling the Autonomous Underwater Vehicle (AUV) dynamics in real-time with an online recursive extended least squares (RELS) algorithm, which achieved path following for the AUV. The proposed NARMAX based controller could present real-time parameters in the model and handle uncertainties with PID control and a Lyapunov-based backstepping control. In[11], the cloud model and cloud transformation was applied to the NARX modelling, to estimate the cloud parameters and generate the prediction band of NARX models. The proposed cloud-NARX model could not only identify the space weather system, but also quantify the uncertainty of model parameters using the prediction band.

Parameter uncertainty in NARMAX modelling remains as the important aspect as it has a significant impact on the accuracy and reliability of the model. The parameter uncertainty can be handled by using several feedback methods to reduce the uncertainties as much as possible or can be quantified with probabilistic results to illustrate more information from NARMAX models. Both methods can measure the uncertainty in the processing parameters, while the choice of method depends on the needs of the actual application.

Model structure uncertainty in NARMAX modelling

Model structure uncertainty arises from the simplifications and assumptions made during the model formulation process. In NARMAX modelling, the choice of functional form, the number of input/output lags, and the degree of nonlinearity, which are the fundamental issues since the introduction of NARMAX method, are all critical factors that contribute to model structure uncertainty. Several authors have discussed the importance of addressing model structure uncertainty in NARMAX modelling to ensure accurate representation of the underlying system dynamics.

In [155], a novel Integrated Forward Orthogonal Search (IFOS) algorithm, aided by the squared correlation and mutual information was proposed to introduced to improve the reliability of NARMAX model structure and reduce the uncertainty. By incorporating linear and nonlinear dependency measures and employing an incremental construction process, the algorithm facilitates the identification of accurate, interpretable, and robust models. In [156], the application of information criteria, like the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the Adjustable Prediction Error Sum of Squares (APRESS) , for model selection and model averaging of NARMAX has been extensively investigated. Comparing the performance of three criteria, the findings indicate that APRESS yields superior models in terms of generalization performance and model complexity. Moreover, a model averaging technique is introduced, which has the potential to produce an averaged model exhibiting enhanced generalization robustness compared to individual models.

One problem in quantify and measure the model structure uncertainty in NARMAX and other machine learning method is to balance the complexity of models and the performance during the testing period. During the training period, the model structure determination and selection are greatly affected with the introduction of model structure uncertainty. To handle such uncertainty, an effective method is to utilize as many types as possible of model structures to identify the most appropriate NARMAX model. However, for most real applications, the complex and dynamic nature prefers large nonlinearity and high dimensional models, while these models might lead to overfitting problems in the testing period.

Thus, the balance of reducing model structure uncertainty and accurate identification of system should be valued.

Addressing model structure uncertainty is crucial for developing accurate and reliable NARMAX models. Various approaches, including model selection, validation, and averaging techniques, have been proposed and investigated in the literature. As the field continues to evolve, more advanced techniques and hybrid modelling approaches are expected to emerge, offering further improvements in the performance of NARMAX models.

2.2.3 Summary

In summation, model uncertainty constitutes a vital element in the formulation, assessment, and implementation of mathematical and statistical models. A comprehensive comprehension of the distinct categories of model uncertainty—data, parameter, structural, and criterion uncertainty—empowers scholars to make judicious decisions throughout the modelling process. By identifying and addressing these uncertainties, researchers can develop more accurate and reliable models, ultimately fostering enhanced decision-making and system control.

It is essential to acknowledge that model uncertainty can materialize in diverse forms and influence various facets of model performance. As a result, a holistic approach to addressing model uncertainty should encompass the detection, analysis, and mitigation of uncertainties across different categories. This methodology can unveil potential biases, inconsistencies, or constraints within a model, enabling researchers to refine and enhance their models through targeted improvements, alternative model structures, or parameterizations.

Furthermore, addressing model uncertainty can contribute to a more nuanced understanding of model performance across different aspects, thereby informing model development, refinement, and application. By contemplating model uncertainty during model evaluation and selection procedures, researchers can ascertain the robustness and reliability of their models, ultimately facilitating improved decision-making and system control.

Within the realm of system identification and modelling, addressing model uncertainty remains a critical and ongoing research endeavour. Prospective work in this domain should persist in exploring innovative methodologies for quantifying and mitigating model uncertainty, as well as establishing best practices for incorporating uncertainty analysis into the modelling process. By advancing our knowledge and management of model uncertainty, researchers can persist in formulating more precise and dependable models, ultimately promoting improved decision-making and system control across an array of domains.

2.3 Model uncertainty in Machine Learning

As machine learning algorithms are increasingly employed in various critical applications, addressing uncertainty in predictions and decision-making becomes imperative. Uncertainty in machine learning may arise from multiple sources, including noisy or incomplete data, inherent variability in data-generating processes, or limitations in modelling assumptions. Addressing uncertainty enables models to yield reliable and robust predictions, which is particularly crucial in safety-critical applications.

Two primary types of uncertainty exist: aleatoric uncertainty, originating from inherent variability in data and remaining irreducible, and epistemic uncertainty, stemming from a lack of knowledge about the data-generating process and being mitigable through additional data collection or model refinement.

Methods to tackle uncertainty in machine learning encompass Bayesian approaches, ensemble techniques, and confidence intervals. These strategies aim to quantify and propagate uncertainty, facilitating informed decision-making and risk assessment. The handling of uncertainty constitutes a critical area of research, carrying substantial implications for the secure and responsible implementation of machine learning technologies.

2.3.1 Brief introduction of machine learning

Machine learning, an essential subfield of artificial intelligence, has experienced considerable growth and development since its emergence in the mid-20th century [157]. The discipline comprises a wide array of techniques and algorithms that enable computational systems to learn from data, identify patterns, and autonomously make decisions [158]. Applications of machine learning span across various sectors, including healthcare [159], finance [160], natural language processing [161], robotics [162], and computer vision [163]. This introductory section presents an in-depth overview and historical context of machine learning, tracing its progression from early theoretical foundations to the ground-breaking innovations that have significantly reshaped the field in recent years.

The foundations of machine learning can be traced back to the 1940s and 1950s, with seminal works on cybernetics, information theory, and the development of the first artificial neural networks [164]. Key figures such as Alan Turing and Claude Shannon contributed to the groundwork for artificial intelligence and the notion that machines could learn from data. In [165], the famous Turing Test was proposed by Alan Turing, which evaluates a machine’s ability to exhibit human-like intelligence. During this period, other significant works also emerged. For instance, in [166], the first mathematical model of an artificial neuron was present. Moreover, in [167], a fundamental concept, Hebbian learning, in the development of neural networks was introduced.

The 1960s marked the beginning of machine learning as a distinct research area, with the development of early algorithms, such as the perceptron [168] and the Least Mean Squares (LMS) algorithm [169]. In this era, alternative approaches to machine learning garnered increased attention. Investigations concentrated on symbolic methods [170], rule-based systems [171], and expert systems [172]. A prominent instance is the development of the ID3 algorithm, which established the basis for decision tree learning [173]. Other notable methodologies encompassed the progression of the k-nearest neighbours’ algorithm, Bayesian networks, and genetic algorithms.

The k-nearest neighbours (k-NN) algorithm embodies an instance-based learning technique that classifies data points according to the majority class of its *k* nearest neighbours [174]. This non-parametric approach demonstrates efficacy in classification and regression tasks, particularly when working with smaller datasets and fewer dimensions. Bayesian neural networks [175], or belief neural networks, represent directed acyclic graphs capturing probabilistic relationships between variables, facilitating efficient reasoning and learning under uncertain conditions. These networks provide a robust framework for modelling intricate systems and cater to various applications, encompassing diagnostics, decision-making, and anomaly detection. Genetic algorithms [176], drawing inspiration from natural evolutionary processes, employ search heuristics for optimizing solutions within expansive problem spaces. Genetic algorithms find widespread utility in optimization, machine learning, and search tasks across diverse fields.

The 1980s saw a resurgence of interest in artificial neural networks, sparked by the introduction of the backpropagation algorithm [177]. This methodology facilitated the learning of intricate representations in multi-layer neural networks, thereby unlocking novel avenues for machine learning investigation. In the 1990s, the field witnessed further advancements with the development of Support Vector Machines (SVM) [178]. This approach, grounded in statistical learning theory, facilitated efficient learning from high-dimensional data, becoming extensively adopted across diverse applications. The employment of kernel methods in SVMs also permitted learning of non-linear decision boundaries, thereby augmenting their applicability.

The early 2000s witnessed a surge in digital data availability and the requisite computational power for processing it. This period also marked the emergence of ensemble methods, such as AdaBoost [179] and Random Forests [180]. These approaches enhanced the performance of individual learning algorithms through output amalgamation, frequently yielding more accurate and robust models. Concurrently, researchers delved into unsupervised learning techniques, including clustering and dimensionality reduction [181]. Algorithms like k-means clustering [182] and Principal Component Analysis (PCA) [183] gained prominence due to their capacity to discern patterns and structures within unlabelled data. Simultaneously, the development of reinforcement learning algorithms, like Q-Learning [184] and SARSA [185], opened up new possibilities in robotics and control systems. The 2000s also marked the advent of novel machine learning paradigms, such as semi-supervised learning, transfer learning, and active learning [186-188].

The 2010s heralded the deep learning revolution, characterized by deep neural networks' success in various applications, such as image and speech recognition [189]. Notable breakthroughs like CNNs showcased deep learning's potential in computer vision tasks [190]. Jointly, the generative adversarial networks (GANs) ushered in a new era of generative models, capable of generating realistic data types [191]. In the natural language processing realm, advancements included RNNs [192], LSTMs [7], and the Transformer architecture [193], laying the groundwork for modern language models like BERT [194], GPT [195], and RoBERTa [196]. Recent progress in reinforcement learning, exemplified by DeepMind's AlphaGo [197] and AlphaZero [198], has achieved superhuman p0erformance in games like Go and Chess.

Addressing uncertainty, a current challenge, is crucial for reliable decision-making in safety-critical applications [199]. Techniques such as Bayesian deep learning [200] and ensemble methods [201] are employed to quantify and propagate uncertainty. A further challenge in the field involves developing interpretable machine learning models. Explainable AI (XAI) seeks to develop methods that make complex machine learning models understandable and accountable [202].

In summary, machine learning has experienced considerable advancements, and the field's rapid evolution is driven by data availability, computational power, and new research directions. Tackling challenges associated with uncertainty, explanability, and interpretability is essential for the secure application of machine learning technologies. This interdisciplinary area will persistently bring transformative effects on society, fostering innovation and sculpting the future of numerous sectors.

2.3.2 Model uncertainty in machine learning

Uncertainty is a critical aspect of machine learning, particularly in domains where dependable models are essential [203]. The scholarly community has acknowledged the importance of uncertainty in machine learning, identifying novel problems and challenges. At least two types of uncertainty exist in machine learning: aleatoric and epistemic [204]. Aleatoric uncertainty refers to the inherent randomness in the data, while epistemic uncertainty refers to the uncertainty in the model's predictions due to a lack of knowledge or understanding [205].

Researchers have devised diverse techniques to manage uncertainty, including uncertainty estimation and differentiation between the two uncertainty types. The prominence of uncertainty in machine learning has garnered substantial attention, and addressing it is vital for creating reliable models. Quantifying uncertainty in machine learning has been an active area of research, with various approaches proposed to address aleatoric and epistemic uncertainty. Some notable research works include:

Bayesian Methods

Bayesian methods represent a significant approach to tackling uncertainty in machine learning, as they integrate prior knowledge regarding model parameters and subsequently update this information with observed data to produce posterior distributions [206]. Bayesian methods have been applied to various machine learning models, such as Bayesian Neural Networks [207] and Gaussian Processes [208]. These models can provide estimates of both aleatoric and epistemic uncertainty through predictive distributions.

Bayesian methods offer a systematic framework for addressing uncertainty by quantifying and managing it effectively [209-211]. Through the integration of prior knowledge and its subsequent update based on observed data, these methods facilitate the computation of comprehensive posterior distributions over model parameters. This, in turn, allows for the quantification of uncertainty and supports informed decision-making. In [212], "Dropout", a widely-used regularization technique, was proposed and interpreted as an approximation to Bayesian inference in deep learning models. This approach is computationally efficient and can be easily implemented in existing deep learning architectures, making it an attractive option for quantifying model uncertainty.

Bayesian methods, despite their merits, present certain drawbacks. Primarily, the selection of prior distributions substantially influences the model's performance, necessitating meticulous contemplation of appropriate priors. Furthermore, the impact of selected priors may persist even after updating with observed data, potentially yielding biased estimates if the priors do not accurately represent the underlying data-generating process. Secondly, calculating comprehensive posterior distributions can impose significant computational challenges, especially for intricate models such as deep neural networks.

Ensemble Techniques

Ensemble methods, long-established in machine learning, are introduced to enhance model performance and robustness [213]. These techniques amalgamate multiple models' outputs to generate more accurate and stable predictions [214]. By capitalizing on individual learning algorithms' strengths and counterbalancing their weaknesses, ensemble methods bolster overall performance [215]. The primary idea behind ensemble methods is that a group of diverse and complementary models can better capture complex patterns in the data than a single model [216].

Bagging (Bootstrap Aggregating) [217], boosting [218], and stacking (stacked generalization) [219] exemplify such techniques. These ensemble techniques provide a powerful means to improve model performance and stability by leveraging the strengths of multiple base models and compensating for their weaknesses. By training several machine learning models with varying initializations and integrating their predictions, these ensemble methods illustrated the feasibility of obtaining reliable uncertainty estimates. This methodology facilitates the differentiation between epistemic and aleatoric uncertainty by scrutinizing the ensemble's prediction variability.

However, ensemble techniques are not without limitations. Firstly, they can demand significant computational resources, as multiple models must be trained and combined, which may be impractical for large-scale or real-time applications. Secondly, the choice of individual models within the ensemble, as well as the combination strategy, can significantly impact the ensemble's performance. Thus, selecting an optimal ensemble configuration can be challenging and requires extensive experimentation. Finally, while ensemble techniques can improve predictive performance and provide a measure of uncertainty, they may not offer a principled framework for quantifying and managing uncertainty as Bayesian methods do. Despite these drawbacks, ensemble techniques remain valuable tools in machine learning and can provide substantial performance improvements when applied judiciously.

Confidence Intervals

Confidence intervals serve as a valuable tool for quantifying uncertainty in machine learning predictions by establishing a range of plausible values for a particular prediction [220]. By offering insights into the uncertainty surrounding a given prediction, confidence intervals can facilitate more informed decision-making and risk assessment processes [221].

However, it is crucial to recognize the limitations and assumptions underlying confidence intervals. One such assumption is that the underlying data distribution is known or can be accurately estimated. In practice, this assumption may not hold true, potentially leading to inaccurate or misleading intervals. Additionally, the interpretation of confidence intervals can be prone to misinterpretation. A common misconception is that the confidence interval captures the probability of a true parameter lying within the interval, whereas it actually represents the probability that the interval, when computed from multiple samples, will contain the true parameter.

Despite these limitations, confidence intervals remain a valuable tool for assessing uncertainty in machine learning models, provided that practitioners are aware of the assumptions and potential pitfalls. Incorporating confidence intervals into model evaluation can offer a more comprehensive understanding of the model's performance and help inform decision-making processes under uncertainty.

Out-of-distribution detection

One challenge in addressing uncertainty is the detection of out-of-distribution (OOD) inputs, which are examples that differ significantly from the training data. Identifying OOD inputs is vital to avert overconfident predictions in regions where the model's uncertainty is high. Methods for OOD detection encompass Mahalanobis distance-based techniques [222] and Bayesian uncertainty estimates [223]. These approaches strive to pinpoint inputs deviating from the learned distribution and offer uncertainty estimates indicative of the model's limited knowledge regarding these instances. However, discerning OOD inputs can be complex, as the boundary between in-distribution and out-of-distribution data may not be apparent, and the employed techniques may not always provide accurate uncertainty estimates. Consequently, further research and development of robust methods are essential for effective OOD analysis and improved uncertainty quantification.

Adversarial training

Adversarial training constitutes a method for enhancing model robustness through the incorporation of adversarial examples—perturbed inputs intentionally crafted to provoke incorrect predictions. In [224], it has been demonstrated that adversarial training can yield models with increased resilience to adversarial attacks and reduced epistemic uncertainty. By prompting the model to generate accurate predictions despite adversarial perturbations, adversarial training can produce more dependable uncertainty estimates. However, this approach may not always yield models that generalize well to unseen data or entirely new adversarial perturbations. Additionally, adversarial training can be computationally demanding, as it involves generating adversarial examples and updating model weights iteratively. Consequently, while adversarial training offers a viable strategy for augmenting model robustness and refining uncertainty estimates, its limitations necessitate further exploration of alternative methods to address these concerns effectively.

2.3.3 Summary

In summary, the significance of addressing uncertainty in machine learning cannot be overstated, particularly as these models become integral to safety-critical applications. Numerous approaches have emerged to estimate and manage uncertainty, encompassing Bayesian techniques, ensemble methods, active learning, and adversarial training. Although substantial advancements have been achieved in recent years, further investigation is required to develop models proficient in quantifying and propagating uncertainty. Such developments will ensure dependable and robust decision-making across a wide range of domains.

2.4 Challenges

In the realm of engineering and scientific exploration, the challenge of modelling complex, dynamic systems—ranging from engineered structures to natural phenomena like climate and space weather—presents a critical frontier. The precision with which these models can predict system behaviours has profound implications for both the development of technologies and our understanding of the world around us. This thesis delves into the realm of model uncertainty: a pervasive issue in system identification that has garnered extensive study. Yet, when extending the application of system identification methodologies into arenas such as climate science and space weather prediction, where exact models are elusive, we encounter a renewed and intensified significance of model uncertainty [55].

Model Uncertainty in Engineering and System Identification

Historically, the field of data-driven modelling has grappled with model uncertainty through the development and refinement of system identification techniques. These methodologies, including advanced strategies like NARMAX, aim to construct mathematical models that mirror the dynamics of physical systems based on empirical data. While these models are invaluable for predicting system responses and designing control mechanisms, they inherently suffer from uncertainties due to simplifications, assumptions, and the inherent unpredictability of system behaviors. The engineering community has made significant strides in quantifying and mitigating these uncertainties, leveraging both theoretical advancements and computational tools to enhance model reliability and robustness [29] [225] [226].

Expanding Horizons: Complex and Dynamic Systems

The exploration of complex and dynamic systems, such as climate dynamics and space weather, pushes the boundaries of traditional system identification methods. Unlike engineered systems, where physical laws and empirical data guide model development, these natural systems are governed by processes that are not fully understood, making precise modelling an aspirational goal rather than a practical reality [227, 228]. The intrinsic complexity of these systems, coupled with their susceptibility to chaotic behaviours and the vast scale of interactions, exacerbates model uncertainty. The limited availability of observational data, the computational demands of simulating such systems, and the need for interdisciplinary collaboration further compound the challenge.

Bottleneck Issues and the Path Forward

The exploration into model uncertainty within these complex systems has been historically hindered by several bottleneck issues:

Computational Limitations: The complexity and scale of systems like climate and space weather require immense computational resources for modelling and simulation. Until recent advancements in computational power and techniques, these requirements were prohibitive.

Data Availability: Effective modelling demands extensive, high-quality observational data. For many complex systems, such data are sparse or have only recently become available through advancements in sensing and monitoring technologies.

Interdisciplinary Integration: The accurate modelling of complex systems necessitates the integration of knowledge across diverse scientific disciplines. This interdisciplinary approach poses both methodological and logistical challenges, slowing the pace of progress.

Inherent Nonlinearity and Complexity: The predictive modelling of systems characterized by complex nonlinear dynamics (e.g. chaos, randomness and nonstationary) is inherently challenging. Traditional modelling approaches often fall short in capturing the complexity of these dynamics.

Justification for This Research

Given these challenges, this research is motivated by the urgent need to advance our understanding and methodologies for dealing with model uncertainty in the context of complex, dynamic systems. This thesis represents a timely and necessary inquiry into the nature of model uncertainty as it manifests in these less understood domains. By leveraging recent advancements in computational techniques, data availability, and interdisciplinary approaches, this work aims to develop new frameworks and methodologies capable of addressing the nuanced challenges presented by complex systems modelling. This research not only contributes to the theoretical and methodological advancements in the field of system identification but also has practical implications for predicting and mitigating the impacts of climate change, enhancing the reliability of space weather forecasts, and beyond. Through this work, we seek to illuminate the pathways through which model uncertainty can be understood, quantified, and mitigated in the pursuit of more reliable and accurate models for both engineered and natural systems.

2.5 Summary

In summary, addressing uncertainty in NARMAX models is essential for ensuring their reliability and robustness, particularly in safety-critical applications. This literature review has provided an overview of the key concepts, methods, and challenges related to uncertainty in NARMAX models, focusing on Bayesian and ensemble approaches, model selection, and scalability issues. Although significant progress has been made in recent years, further research is needed to develop more efficient and robust methods for quantifying and managing uncertainty in NARMAX models. This will ultimately enable more reliable predictions, model validation, and control design across diverse domains

Chapter 3

Quantification of Model Uncertainty and Probabilistic Decision in NARMAX

Model uncertainty is pervasive in system identification and prediction, which makes it be faced and considered as an essential factor to produce robust and reliable results. Just like other system identification methods, NARMAX serves as an interpretable and effective method for identifying complex nonlinear dynamic systems. During the implementation process, it also considers model uncertainty as a crucial indicator. By measuring the degree of uncertainty in known NARMAX models, we can characterize their reliability and robustness, thereby validating the effectiveness of both the model and method. At the same time, the uncertainty of the model can also serve as a feature for training and optimizing NARMAX models, thereby obtaining more stable system models and further enhancing our understanding of the system. In addition, by accepting and utilizing model uncertainty, we can identify and describe system diversity, such as probabilistic system prediction outputs, composite system model groups etc., thus compensating for deficiencies in generating probabilistic predictions with NARMAX models and enhancing our comprehensive understanding of systems.

While the uncertainty of NARMAX models has been extensively studied and utilized, there is still no unified theory to explain and summarize it. Instead, analysis and discussion are based on modelling tasks for different systems. However, there are commonalities between the ambiguity in different modelling processes and the uncertainty of NARMAX models. Therefore, this chapter first analyzes, summarizes, and categorizes the types of uncertainties in NARMAX models; uses set theory principles to mathematically describe and quantify the uncertainty of NARMAX models; at the same time discusses how to measure and use the uncertainty of NARMAX models based on information theory and optimization theory.

3.1 Introduction

Model uncertainty is an inherent problem for all system identification and modelling, not only for NARMAX modelling, as the system in real world, especially the complex dynamical systems, such as environmental systems (climate change, natural hazards), economic systems (economic and financial market), healthcare systems (epidemiological analysis research, cancer etiologic analysis research), engineering systems (power systems, robotic systems). Also, uncertainty has a direct impact on decision making by introducing variability into the potential outcomes associated with a particular course of action, such as harvesting and weather conditions. Additionally, uncertainty plays an indirect role by giving rise to the presence of private information. In strategic interactions, uncertainty and private information are inherently interconnected. Uncertainty creates variations in private information when different agents possess access to varying information regarding the uncertain phenomena. Similarly, the existence of private information can also give rise to uncertainty if agents are considering its implications, leading to concerns of moral hazard and adverse selection.

In practical settings, uncertainty serves as a primary driver of competitive advantage, thereby creating business opportunities that can be advantageous for entrepreneurial endeavours. In the realm of theory, uncertainty paves the way for the exploration of agent decision-making and strategic interactions, embodying a captivating and intellectually intricate pursuit that diverges significantly from the study of actions and interactions among physical particles. Across both practical and theoretical domains, uncertainty assumes a prominent role.

In the vast majority of system analyses, due to insufficient understanding of the system, lack of prior knowledge about it and inadequate research on its interaction with the environment, there is a tendency to use data-driven system modelling methods. Data-driven system modelling can obtain models and fitting results for the system by analysing its inputs and outputs without relying on prior knowledge of the system. However, the issue of model uncertainty still exists in data-driven system modelling methods. Since these methods rely on the input and output of the system and use parameterized computational models to simulate system behaviour, the structure and parameters of these models are often not unique. At the same time, due to the existence of differences in observational data, such as signal-to-noise ratio and unknown factors affecting the system, coupled with a lack of unified standards for defining model quality, there is significant uncertainty in data-driven system modelling. This greatly affects trust and reliance on these types of models, especially when the system required one best model to describe and forecast.

NARMAX, as an efficient, reliable and accurate nonlinear data-driven method, is widely used in the modelling of complex dynamic systems. The NARAMX model is specifically designed for the purpose of identifying nonlinear systems. Its main function is to determine the appropriate model structure by carefully selecting the most important model items from a large dictionary of numerous candidate models. This selection process is crucial for constructing an accurate representation of the underlying system. Assuming that the identified individual model structural elements can perfectly describe the real system components, most models have the ability to accurately represent the system. But as discussed above, in many practical applications, this assumption is invalid because of the uncertainty.

Thus, conducting an in-depth analysis of the factors attributing to the uncertainty of the NARMAX model, investigating the impact of diverse factors on the model's uncertainty, and studying the role of NARMAX model uncertainty on outcomes is of utmost importance. This endeavour seeks to establish a comprehensive and quantified theory for analysing uncertainties in NARMAX models, which significantly influences the research and utilization of NARMAX. To accomplish this, the present chapter begins by providing a thorough literature review, with a particular focus on polynomial constructed NARMAX models. Subsequently, an analysis and summary of the causes for uncertainties in these models are presented. Leveraging these insights, the chapter then offers mathematical descriptions and quantifications for each influencing factor, thereby enabling the transformation of qualitative challenges into specific quantitative aspects during the analysis of uncertainties within NARMAX models. Building upon this foundation, a quantitative exploration of the contributions made by each influencing factor towards the uncertainties in NARMAX models is carried out. Consequently, an optimized modelling methodology is established, aimed at either eliminating or harnessing these uncertainties.

3.2 NARMAX

3.2.1 NARMAX models

The general NARMAX model, a representation for a broad range of nonlinear systems, is denoted as [1]:

$$\begin{aligned} y(t) = F[&y(t-d_y-1), y(t-d_y-2), ..., y(t-d_y-n_y), \\ &u(t-d_u), u(t-d_u-1), ..., u(t-d_u-n_u), \\ &e(t-d_e-1), e(t-d_e-2), ..., e(t-d_e-n_3)] + e(t) \end{aligned} \tag{3.1}$$

where $y(t)$, $u(t)$, and $e(t)$ are the system output, input and noise sequences, respectively; n_y, n_u, and n_e are the maximum lags for the system output, input, and noise (error), which are one of the key element to determine the model structure and model parameters; $F[\cdot]$ is one of the typical nonlinear functions to describe the complex system, and d_y, d_u, and d_e are time delays that can be set by the prior knowledge. Typically, $d_y = d_e = 0$ and $d_u = 1$. The incorporation of noise terms is essential to account for various factors such as measurement noise, modelling errors, and unmeasured disturbances within the NARMAX framework. Accurately modelling noise is of paramount importance, as practically every real-world dataset is influenced by noise. Neglecting noise in the modelling process inevitably leads to incorrect results. In the context of nonlinear systems, solely relying on the descriptor white noise - characterized by second-order moments - is inadequate. Nonlinear systems require the consideration of higher-order moments since they assume significance and impact system behaviour.

The general process of the NARMAX model is shown in Figure 4. The objective is to find a transparent NARMAX model to explain the unknown system and produce the accurate predictions/simulated output $\hat{y}(t)$.

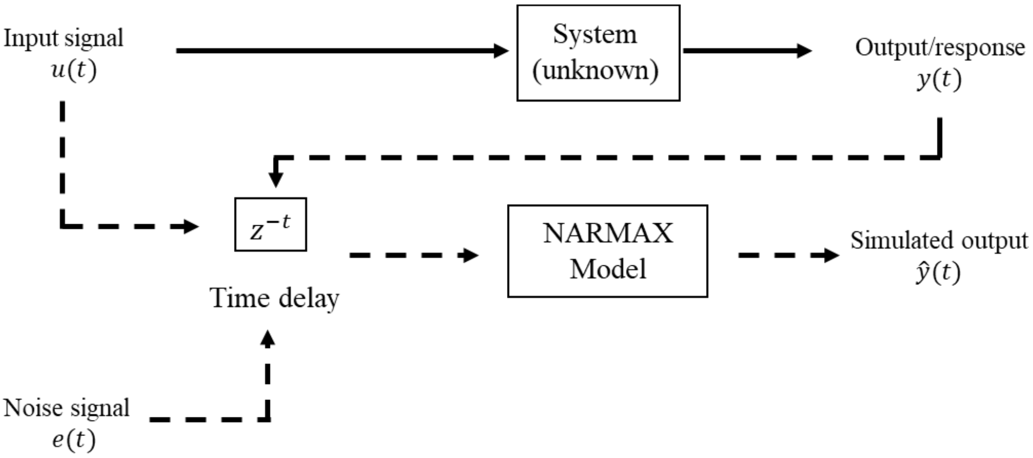


Figure 4 The process of the NARMAX modelling

Originally conceived as a model, NARMAX has evolved into a broader philosophy encompassing nonlinear system identification. The NARMAX approach involves a series of steps:

- Structure detection to identify the model terms.
- Parameter estimation to determine the model coefficients.
- Model validation to assess its unbiasedness and correctness.
- Prediction to determine the future output.
- Analysis to explore the system’s dynamical properties.

Structure detection serves as a fundamental component of NARMAX, particularly in linear parameter estimation where model order determination is relatively straightforward. Typically, the quick and efficient approach involves estimating models of increasing order (e.g., one, two, three, etc.), which are subsequently validated and compared to identify the simplest yet adequate representation of the system. This process is effective due to the assumption of a pulse transfer function representation, wherein each increase in model order augments the number of unknown parameters by two (one additional coefficient for the numerator and denominator). Overfitting is easily discerned through methods such as pole zero cancellations.

However, the straightforward approach employed in linear systems identification does not readily extend to the nonlinear domain. Consider the NARMAX model with a cubic polynomial expansion that includes three lagged input term $u(t-1)$, $u(t-2)$, and $u(t-3)$, one lagged output term $y(t-1)$, and two lagged noise terms $e(t-1)$ and $e(t-2)$. In this case, the total number of potential candidate terms amounts to 84, as all possible combinations within the cubic expansion are encompassed. Naïvely attempting to estimate a model by including all these terms and subsequently pruning them can give rise to significant numerical and computational challenges, which should be diligently avoided. Nonetheless, it is often observed that only a subset of terms holds substantial importance in the model representation. Hence, the process of structure detection, which involves iteratively selecting terms one at a time based on their significance, assumes critical importance. This intuitive approach follows a logical progression of incorporating the most significant terms first, followed by subsequent terms in order of their importance. The termination point is reached when the model achieves adequacy, numerical efficiency, and soundness. Most notably, this pragmatic strategy enables the development of parsimonious models that establish meaningful relationships with the underlying system.

In practical applications, numerous model structures are accessible to approximate the unknown mapping function $F[\cdot]$ in the given equation, like power-form polynomial models, rational models, neural networks, fuzzy logic-based models, wavelet expansions, radial basis function (RBF) networks, and many more. Due to the attractive features of the power-form polynomial functions, like smoothness of polynomials, approximation of functions, transparency, ability of handling noisy data and interpretability to describe the nonlinear systems, the polynomial NARMAX models have been the most popular nonlinear function in the NARMAX method. Therefore, in this thesis, the polynomial NARMAX model is set to be the default NARMAX model.

The polynomial NARMAX model can be represented by the power-form polynomials:

$$\begin{aligned} y(t) = \theta_0 + \sum_{i_i=1}^{n} f_{i_1}\left(x_{i_1}(t)\right) + \sum_{i_1=1}^{n}\sum_{i_2=i_1}^{n} f_{i_1 i_2}\left(x_{i_1}(t), x_{i_2}(t)\right) + \cdots \\ + \sum_{i_1=1}^{n}\cdots\sum_{i_\ell=i_{\ell-1}}^{n} f_{i_1 i_2 \cdots i_l}\left(x_{i_1}(t), x_{i_2}(t), \ldots, x_{i_\ell}(t)\right) + e(t) \end{aligned} \tag{3.2}$$

where ℓ is the degree of polynomial nonlinearity, $\theta_{i_1 i_2 \cdots i_m}$ are model parameters, $n = n_y + n_u + n_e$, and

$$f_{i_1 i_2 \cdots i_m}\left(x_{i_1}(t), x_{i_2}(t), \ldots, x_{i_m}(t)\right) = \theta_{i_1 i_2 \cdots i_m} \prod_{k=1}^{m} x_{i_k}(t), 1 \le m \le \ell \tag{3.3}$$

$$x_m(t) = \begin{cases} y(t-m) & 1 \le m \le n_y \\ u\left(t-\left(m-n_y\right)\right) & n_y + 1 \le m \le n_y + n_u \\ e\left(t-\left(m-n_y-n_u\right)\right) & n_y + n_u + 1 \le m \le n_y + n_u + n_e \end{cases} \tag{3.4}$$

Thus, the equation (3.2) can be explicitly expressed as

$$\begin{aligned} y(t) = \theta_0 + \sum_{i_i=1}^{n} \theta_{i_1} x_{i_1}(t) + \sum_{i_1=1}^{n}\sum_{i_2=i_1}^{n} \theta_{i_1 i_2} x_{i_1}(t) x_{i_2}(t) + \cdots \\ + \sum_{i_1=1}^{n}\cdots\sum_{i_\ell=i_{l-1}}^{n} \theta_{i_1 i_2 \cdots i_l} x_{i_1}(t) x_{i_2}(t) \cdots x_{i_\ell}(t) + e(t) \end{aligned} \tag{3.5}$$

Normally, the degree of a multivariate polynomial is determined by the highest order present among its terms. Taking a polynomial function $f(x_1, x_2, x_3, x_4) = a_0 + a_1 x_1^3 x_2 + a_2 x_2^2 x_3 + a_3 x_4 + a_4 x_1 x_4$, where a_0, a_1, a_2, a_3, and a_4 are the parameters in the polynomial function, as an example, the nonlinearity degree of this function is $\ell = 4$. In this case, one polynomial NARMAX model with nonlinear degree ℓ means that the order of the model terms in this polynomial NARMAX model can be no higher than ℓ.

From the model (3.3), there are several types of NARMAX models based on the inputs of the model. One of the special and widely used models is the NARX model, which does not rely on the noise-related terms, shown in Figure 5.

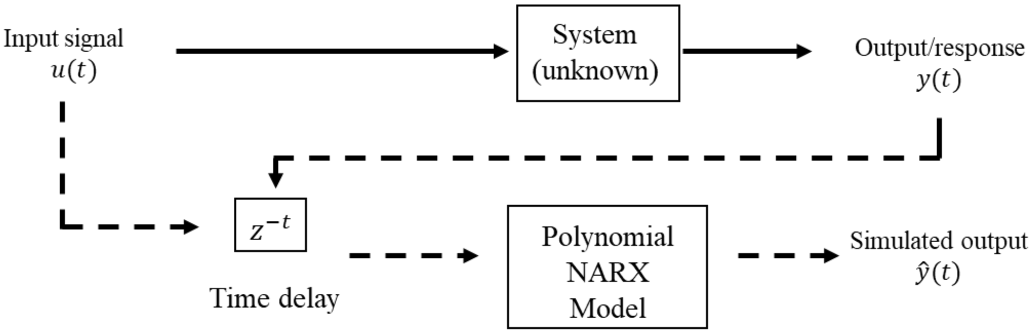


Figure 5 The process of the polynomial NARX modelling

The description of the NARX model is denoted as:

$$\begin{aligned} y(t) = F[y(t-d_y-1), y(t-d_y-2),...., y(t-d_y-n_y), \\ u(t-d_u), u(t-d_u-1),..., u(t-d_u-n_u)] + e(t) \end{aligned} \tag{3.6}$$

In NARX models, the output of the system depends on the inputs of the system and the previous outputs of the system. Similarly, the NARX model can be expressed as a linear-in-the-parameter function like model (3.2), where the model terms in model (3.2) are defined as:

$$x_m(t) = \begin{cases} y(t-m), & 1 \le m \le n_y \\ u\left(t-m+n_y\right), & n_y+1 \le m \le n = n_y + n_u \end{cases} \tag{3.7}$$

Based on the philosophy of the NARMAX, once the model structure of NARMAX model is detected, the next step is to find the key model terms. Due to the time delay information, there are more generated model terms compared to the original inputs and outputs signals, like $u(t-1)$, $y(t-2)$, etc, while the total number of candidate model terms in the polynomial NARMAX model is denoted as:

$$M = (n+\ell)!/[n!\ell!] \tag{3.8}$$

where $n = n_y + n_u + n_e - (d_y + d_u + d_e) + 1$ for the polynomial NARMAX model , d_y , d_u , and d_e are the starters of time delay, typically $d_y = d_e = 1$ and $d_u = 1$, and $n = n_y + n_u - (d_y + d_u) + 1$ for the polynomial NARX model; ℓ is the nonlinearity degree of the model.

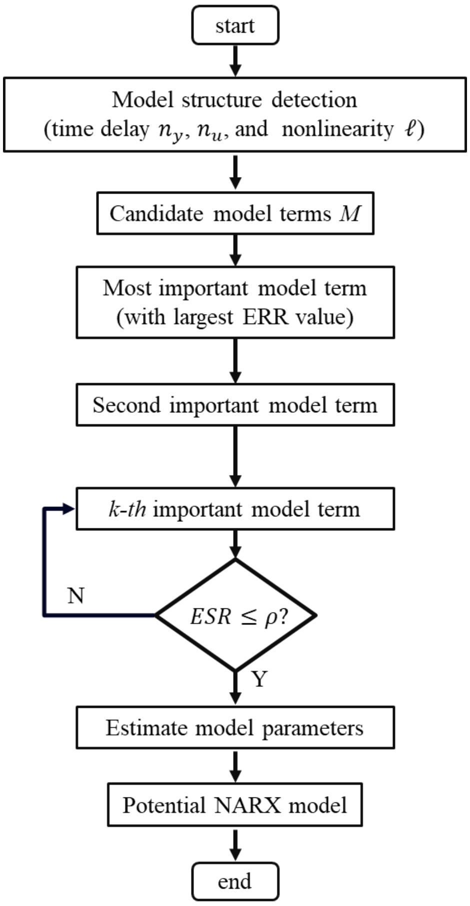

start
Model structure detection
(time delay n_y, n_u, and nonlinearity ℓ)
Candidate model terms M
Most important model term
(with largest ERR value)
Second important model term
k-th important model term
N
$ESR \leq \rho$?
Y
Estimate model parameters
Potential NARX model
end

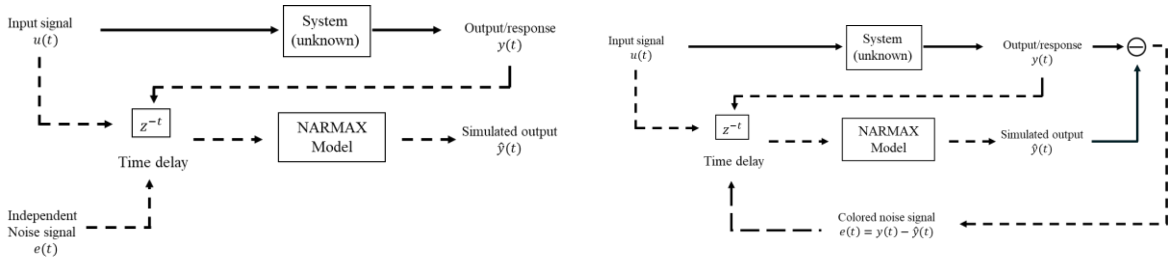
Figure 6 Model feature selection of polynomial NARX model

As shown in Figure 6, the mechanism of determination of the model terms in NARX model is developed based on the Orthogonal Least Squares (OLS) and associated error reduction ratio (ERR). The aforementioned algorithm discerns the requisite dynamic and nonlinear terms in the model by calculating the contribution each possible model term makes to the system's output. This facilitates the construction of the model one term at a time, overtly highlighting the significance of each newly added term. Furthermore, this algorithm presents its results in an intuitive and easily comprehensible manner by demonstrating the percentage contribution each selected model term produces for the output variance. Consequently, this innovative algorithm is user-friendly for both experts and non-experts.

Similar to analytical modelling methods, the algorithm introduces the important model terms first, refining the model subsequently by incorporating less significant effects. The solitary differentiation resides in the fact that, in the NARMAX method, model terms can be identified directly from the data set. Once the model structure has been defined, the unknown parameters within the model can be estimated.

The OLS category of algorithms is compatible with linear-in-the-parameter models for single-input, single-output (SISO) and multi-input, multi-output (MIMO) systems. The input isn't necessitated to take a particular form; even in the presence of unknown nonlinear noise, the algorithms provide unbiased estimates. Moreover, the procedure is scalable to highly intricate nonlinear systems. If an analytical model is wholly or partially identified, the terms derived from this model can be utilized to initiate the model structure selection in orthogonal least squares and help discover any absent model terms.

For extensive n_y , n_u and/or n_e , the initial candidate model terms encompassed in the complete NARMAX or NARX model can be substantial. However, in the majority of practical situations, it is common for only a handful of candidate model terms to be critical for elucidating the intrinsic dynamic relationship. Thus, not every candidate model term necessitates inclusion in the model. Polynomial expansions can exhibit ill-conditioned attributes due to the burgeoning complexity of the involved terms. Nevertheless, these issues can be predominantly mitigated if only the meaningful model terms are elected and incorporated into the model. This aspect further establishes the parsimonious nature of NARMAX.



a) Independent noise model terms

b) coloured noise model terms

Figure 7 Two types of noise models in NARMAX models

For noise terms $\{e(t)\}$ in NARMAX models, they are assumed to be independent of any input and output variables, and therefore, the system does not entail any lagged noise terms or noise-dependent terms, shown in Figure 7 a. Nevertheless, in numerous instances, the noise signal $\{e(t)\}$ might exhibit as a potentially nonlinear, correlated, or coloured noise sequence, which is probable for the majority of genuine data sets. As shown in Figure 7 b, the noise term $e(t)$ is normally defined as:

$$e(t) = y(t) - \hat{y}(t \mid t-1) \tag{3.9}$$

where $\hat{y}(t \mid t-1) = F(y^{[t-1]}, u^{[t-1]}, e^{[t-1]})$, which is the one-step-ahead prediction by the NARMAX model.

As shown in Figure 8, in the case of the NARMAX model, the structure selection method commences with the identification of a process NARX model. It then proceeds by estimating the noise sequence and continues iteratively in accordance with an ELS-type procedure.

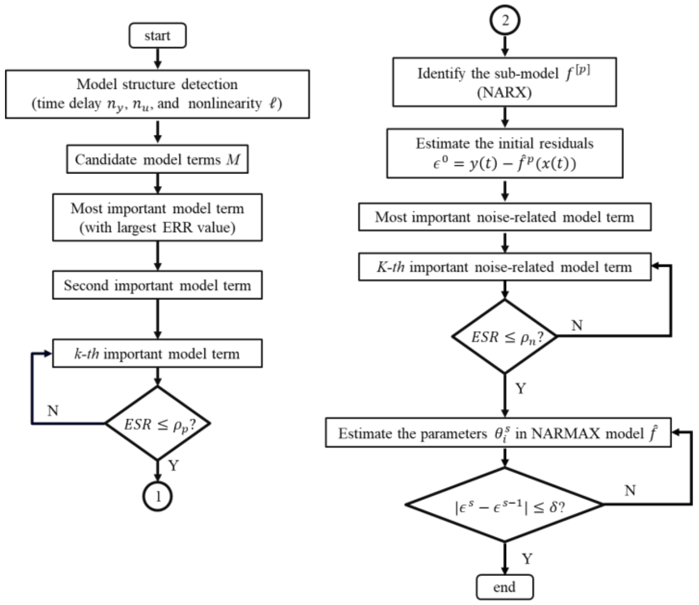

start
Model structure detection
(time delay n_y, n_u, and nonlinearity ℓ)
Candidate model terms M
Most important model term
(with largest ERR value)
Second important model term
k-th important model term
$ESR \leq \rho_p$?
N
Y
1
2
Identify the sub-model $f^{[p]}$
(NARX)
Estimate the initial residuals
$\epsilon^0 = y(t) - \hat{f}^p(x(t))$
Most important noise-related model term
K-th important noise-related model term
$ESR \leq \rho_n$?
N
Y
Estimate the parameters θ_i^s in NARMAX model $\hat{f}$
$|\epsilon^s - \epsilon^{s-1}| \leq \delta$?
N
Y
end

Figure 8 Model feature selection of the polynomial NARMAX model

The NARMAX modelling process and OLS class algorithms can ensure the NARMAX model closely aligns with the complex dynamic system under study. The NARMAX model terms selected by the OLS algorithm optimally interpret and reconstruct the internal interactions of this system, thereby inferring factors that significantly impact it. Not only does the NARAMX model provide a clear structure and parameters, but it also ranks the influence of each term in the system.

3.2.2 The Forward Regression OLS Algorithm

Consider a polynomial NARMAX model of nonlinear degree ℓ as:

$$\begin{aligned} y(t) = c_0 + \sum_{i_i=1}^{n} c_{i_1} x_{i_1}(t) + \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} c_{i_1 i_2} x_{i_1}(t) x_{i_2}(t) + \cdots \\ + \sum_{i_1=1}^{n} \cdots \sum_{i_\ell = i_{t-1}}^{n} c_{i_1 i_2 \cdots i_l} x_{i_1}(t) x_{i_2}(t) \cdots x_{i_\ell}(t) + e(t) \end{aligned} \tag{3.10}$$

where $x_k(t)\ (k = i_1, i_2, ..., i_\ell)$ are defined as (26). Thus, model (3.10) can be represented in the linear-in-the-parameters form:

$$y(t) = \sum_{m=1}^{M} \theta_m p_m(t) + e(t) \tag{3.11}$$

where $\theta_m (m = 1, 2, ..., M)$ are model parameters and $p_m(t)$ are model terms defined as

$$p_m(t) = y\left(t - m_{y,1}\right) \cdots y\left(t - m_{y,my}\right) u\left(t - m_{u,1}\right) \cdots u\left(t - m_{u,mu}\right) e\left(t - m_{e,1}\right) e\left(t - m_{e,me}\right) \tag{3.12}$$

where

$$\begin{gathered} 1 \le m_{y,1} \le m_{y,2} \le \cdots \le m_{y,ny} \le n_y \\ 1 \le m_{u,1} \le m_{u,2} \le \cdots \le m_{u,mu} \le n_u \\ 1 \le m_{e,1} \le m_{e,2} \le \cdots \le m_{e,me} \le n_e \end{gathered} \tag{3.13}$$

It is easy to indicate that if $m_e = 0$, then we have a NARX model, which is considered as the demonstration of the FROLS in the following.

Denote $\mathbf{y} = [y(1), ..., y(N)]^T$ as the output vector with *N* points and $\mathbf{p_m} = [p_m(1), ..., p_m(N)]^T$ $m = 1, 2, ..., M$ as the input vectors by the *m*-th candidate model term. Thus, we have $D = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_M\}$ as the candidate dictionary of the *M* model terms. The goal of the modelling can be defined as finding a subset $D_{M_0} = \{\alpha_1, \alpha_2, ..., \alpha_{M_0}\}$ that:

$$\mathbf{y} = \theta_1 \boldsymbol{\alpha}_1 + \cdots + \theta_{M_0} \boldsymbol{\alpha}_{M_0} + \mathbf{e} \tag{3.14}$$

where $\theta_i\ (i = 1, 2, ..., M_0)$ are the related model parameters.

From the definition we can have

$$D_{M_0} \subset D \tag{3.15}$$

So, the subset D_{M_0} can also be defined as $D_{M_0} = \{\mathbf{p}_{i1}, ..., \mathbf{p}_{i_{M_0}}\}$, where $\alpha_i = \mathbf{p}_i$. The model (3.14) can also be represented by a matrix form as

$$y = \mathbf{A}\boldsymbol{\theta} + \mathbf{e} \tag{3.16}$$

where $A = \left[\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{M_0}\right]$ is a full column rank, $\boldsymbol{\theta} = \left[\theta_1, \ldots, \theta_{M_0}\right]^{\mathrm{T}}$ is the parameter vector, and $\mathbf{e}$ is the error.

The steps of the FROLS algorithm are defined as:

Step 1: Define $\sigma = \mathbf{y}^T \mathbf{y}$, calculate

$$g_m^{(1)} = \frac{\mathbf{y}^{\mathrm{T}} \mathbf{q}_m}{\mathbf{q}_m^{\mathrm{T}} \mathbf{q}_m} \tag{3.17}$$

$$ERR^{(1)}[m] = \left(g_m^{(1)}\right)^2 \left(\mathbf{q}_m^{\mathrm{T}} \mathbf{q}_m\right) / \sigma \tag{3.18}$$

$$v_1 = \arg\max_{1 \le m \le M} \left\{ERR^{(1)}[m]\right\} \tag{3.19}$$

Then denote

$$\begin{gathered} a_{11} = 1 \\ \mathbf{q}_1 = \mathrm{p}_{\ell_1} \\ g_1 = g_{v_1}^{(1)} \\ err[1] = ERR^{(1)}\left[v_1\right] \end{gathered} \tag{3.20}$$

The first selected model term is $\alpha_1 = p_{v1}$, while the first orthogonal vector is $\mathbf{q}_1$. With the selection of first model term α_1, there are $(M_0 - 1)$ model terms needed to be identified. Assuming after the $(n-1)$ *-th* step, there are $(n-1)$ model terms selected forming a subset $D_{n-1} = \{\alpha_1, ..., \alpha_{n-1}\}$ and an orthogonalized set $Q_{n-1} = \{\mathbf{q}_1, ..., \mathbf{q}_{n-1}\}$.

Step n ($n \ge 2$): Let $m \ne \ell_1, m \ne \ell_2, \ldots, m \ne \ell_{s-1}$.

$$\mathbf{q}_m^{(n)} = \mathbf{p}_m - \sum_{r=1}^{n-1} \frac{\mathbf{p}_m^{\mathrm{T}} \mathbf{q}_r}{\mathbf{q}_r^{\mathrm{T}} \mathbf{q}_r} \mathbf{q}_r, \mathbf{p}_j \in D - D_{m-1} \tag{3.21}$$

$$g_m^{(n)} = \frac{\mathbf{y}^{\mathrm{T}} \mathbf{q}_m^{(n)}}{\left(\mathbf{q}_m^{(n)}\right)^{\mathrm{T}} \mathbf{q}_m^{(n)}} \tag{3.22}$$

$$ERR^{(n)}[m] = \left(g_m^{(n)}\right)^2 \left[\left(\mathbf{q}_m^{(n)}\right)^{\mathrm{T}} \left(\mathbf{q}_m^{(n)}\right)\right] / \sigma \tag{3.23}$$

$$v_n = \arg\max_{1 \le m \le M} \left\{ERR^{(n)}[m]\right\} \tag{3.24}$$

Let

$$\begin{gathered} \mathbf{q}_n = \mathbf{q}_{v_n}^{(n)} \\ g_n = g_{v_n}^{(n)} \\ a_{r,n} = \left(\mathbf{q}_r^{\mathrm{T}} \mathbf{p}_{v_n}\right) / \left(\mathbf{q}_r^{\mathrm{T}} \mathbf{q}_r\right), r = 1, 2, \ldots, n-1 \\ a_{nn} = 1 \\ err[n] = ERR^{(n)}\left[v_n\right] \end{gathered} \tag{3.25}$$

The algorithm will end at the M_0*-th* step to form the defined dataset D_{M_0} or satisfying the condition that

$$ESR = 1 - \sum_{n=1}^{L} err(n) \le \rho \tag{3.26}$$

where ρ is a predefined value.

Suppose the algorithm stops at M_0-*th* step, then the model is represented by

$$y(t) = \sum_{i=1}^{M_0} g_i \mathbf{q}_i(t) + e(t) \tag{3.27}$$

which is the same as

$$y(t) = \sum_{m=1}^{M_0} \beta_{\ell_m} \mathbf{p}_{v_m}(t) + e(t) \tag{3.28}$$

where $\boldsymbol{\beta} = \left[\beta_{\ell_1}, \beta_{\ell_2}, \ldots, \beta_{\ell_{M_0}}\right]^{\mathrm{T}}$ is the parameter vector derived from $A\boldsymbol{\beta} = \mathbf{g}$ where $\mathbf{g} = \left[g_1, g_2, \ldots, g_{M_0}\right]^{\mathrm{T}}$ and

$$A = \begin{bmatrix} 1 & a_{12} & \cdots & a_{1M_0} \\ 0 & 1 & \cdots & a_{2M_0} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & a_{M_0-1,M_0} \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{3.29}$$

in which the element a_{ij} $(1 \le i \le j \le M_0)$ is calculated in the above algorithm.

3.3 Qualitative and Quantitative of model uncertainty in NARMAX

3.3.1 Qualitative Analysis of uncertainty in NARMAX models and polynomial NARMAX models

From the literature review, it is discussed that there are four categories of model uncertainty in NARMAX modelling.

- Model uncertainty in data.
- Model structure uncertainty.
- Model parameters uncertainty.
- Model uncertainty in criteria.

As shown in Figure 9, the uncertainties of the aforementioned four NARMAX models correspond to the four stages of NARMAX model training optimization, namely: data pre-process, model structure determination, model parameter estimation and model validation. For each model uncertainty, there are various factors contributing to the uncertainty of each model, driving the emergence of model uncertainty. Therefore, this section qualitatively analyses and categorizes the uncertainties in NARMAX models and polynomial NARMAX models, laying a solid foundation for quantitative analysis of model uncertainty.

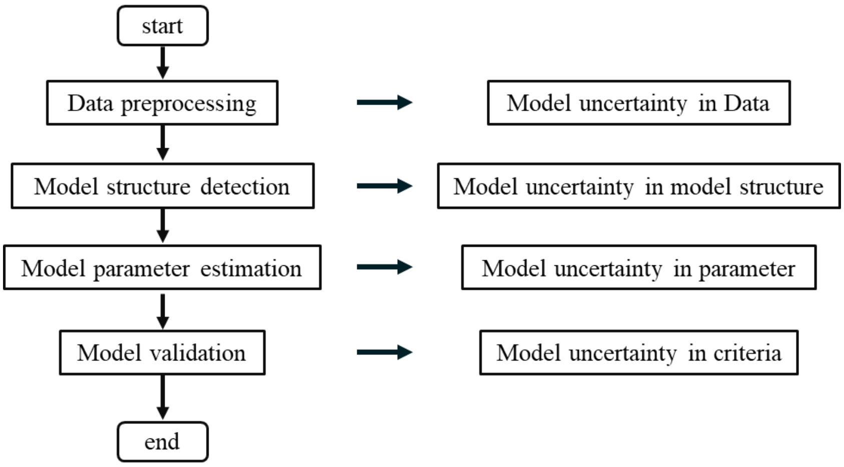

start
Data preprocessing
Model structure detection
Model parameter estimation
Model validation
end
Model uncertainty in Data
Model uncertainty in model structure
Model uncertainty in parameter
Model uncertainty in criteria

Figure 9 The identification steps for NARMAX model and related model uncertainties

3.3.1.1 Model uncertainty in data

Model uncertainty in data typically stems from noisy or incomplete observations, inherent fluctuations in data-generation processes, and constraints in system assumptions, all factors directly influencing the quality of the dataset. And it is this quality that directly influences the accuracy and robustness of the data-driven modelling. Therefore, all factors that affect the quality of a dataset can be collectively referred to as noise.

Noise, particularly that which is found in a dataset, is typically unrelated to the modelling process. Instead, it is associated with the collection, conversion, or storage of the dataset itself [229]. Consequently, the existence of this noise is not obviated by varying modelling methodologies. Furthermore, it is crucial to note that this type of noise is dissimilar to the noise present in the NARMAX modelling process. Specifically, in the context of noise, system responses may be linked to, or as per the coloured-noise definition, correlated with errors. Then again, during the creation of models, specifically, NARX models, the absence of a noise term as a system driver means that the quality, or signal-to-noise ratio (SNR), of the initial datasets heavily influences the ultimate accuracy of the model.

A noise process can be further categorized into one of several classes based on its frequency spectrum or time attributes, as delineated below [230]:

- White noise - purely arbitrary noise characterized by a balanced power spectrum.
- Band-limited white noise - a noise characterized by a balanced spectrum and constrained bandwidth typically encompasses the limited spectrum of the respective device or signal.
- Narrowband noise - a noise process exhibiting a limited bandwidth, for instance, a ‘hum’ of 50-60 Hz typical of an electricity supply.
- Coloured noise - non-white noise, or any broad-band noise with a non-flat spectral shape, exemplified by pink noise, brown noise, and autoregressive noise.
- Impulsive noise - comprising short-duration pulses with variable amplitude and duration.
- Transient noise pules - comprising noise pulses of relatively long duration.

In the vast majority of cases, noise directly impacts the quality of a dataset, thereby preventing data-driven modelling methods from accurately detecting model structures. This leads to more ambiguous parameter estimates and results in less precise system models, causing uncertainties in modelling. Additionally, the presence of noise severely affects model stability. For instance, models optimized through training can precisely reflect system responses due to high signal-to-noise ratios in training sets; however, because test sets have lower signal-to-noise ratios and larger disturbances in model inputs, there is a mismatch between model outputs and system responses on testing machines or real-world applications. This also contributes to model uncertainty.

Therefore, for the NARMAX model, noise affects the NARMAX modelling in two ways, and these two methods bring different types of model uncertainty：

1. For systems driven without noise, that is, systems explainable by the NARX model, the noise in the dataset is either irrelevant or detrimental to the NARX modelling process. Since the modelling process resamples the dataset into training sets, validation sets and test sets, it can be assumed that noise uniformly exists throughout the entire dataset. This prevents scenarios where signal-to-noise ratios are high in training sets but low in test sets or vice versa. In certain applications such as signal processing, even though denoising treatments are applied to original signals, complete elimination of noise isn't achievable; however, it's reasonable to assume that this residual noise is evenly distributed across all data points within a set. Therefore, under these circumstances we can hypothesize that uncertainties caused by noise within a NARMAX model remain constant and do not vary with changes made to other hyperparameters of said NARMAX model.
2. For systems driven by noise, the noise not only affects the quality of the data set, thereby introducing uncertainty into the NARMAX model, but it can also serve as a driver for the system and become an input to the NARMAX model. According to the logic of NARMAX modelling, noise interacts with system inputs and outputs to form new input variables for the system. The inclusion of a general noise model in the NARMAX method is frequently considered judicious since it encapsulates all conceivable noise model components that may surface during the process of nonlinear system identification.
   Consequently, both white and coloured noise could potentially serve as the noise term in NARMAX models, particularly as, in a majority of applications, coloured noise drives the system. Importantly, there are no specific constraints with regard to the noise's probability distribution. In instances where the NARMAX model includes noise, it is generally partitioned into three subsidiary models. To exemplify, consider a polynomial NARMAX model:

$$\begin{aligned}
y(t) &= F[y(t-1),...,y(t-n_y),u(t-1),...,u(t-n_u),e(t-1),...,e(t-n_e)]+e(t)\\
&= f^{[p]}(y(t-1),...,y(t-n_y),u(t-1),...,u(t-n_u))\\
&\quad +f^{[pn]}(y(t-1),...,y(t-n_y),u(t-1),...,u(t-n_u),e(t-1),...,e(t-n_e))\\
&\quad +f^{[n]}(e(t-1),...,e(t-n_e))+e(t)\\
&= f^{[p]}(\mathbf{x}(t))+f^{[pn]}(\mathbf{x}(t),\mathbf{e}(t))+f^{[n]}(\mathbf{e}(t))+e(t)
\end{aligned} \tag{3.30}$$

   where $f^{[p]}(\cdot)$ refers to a NARX model, $f^{[pn]}(\cdot)$ refers to the input-output signal and noise-related sub-model, and $f^{[n]}(\cdot)$ refers to the purely noise sub-model.

   From three sub-models, the noise model terms are denoted as the normal input variables for the NARMAX system. The noisy candidate model terms are generated based on the hyperparameter n_e and nonlinearity degree ℓ, which are key elements to bring the model structure uncertainty and model parameter uncertainty in NARMAX models, while the type of the noise (white or coloured) are the noise uncertainty itself.

In conclusion, the uncertainty brought about by the inherent noise in the dataset is constant and unaffected by the NARMAX modelling process and experimental setup. Therefore, despite its existence, this type of uncertainty can be defined as an independent parameter when analysing and utilizing model uncertainties. On the other hand, the uncertainty caused by noise terms, especially the type of noise, directly affects model performance and system description; thus, requiring quantitative analysis.

3.3.1.2 Model structure uncertainty

The model structure uncertainty is mainly derived from the selection of the fitting function used in the modelling method and its structural form. In the context of NARMAX models, the structural determination is based on critical elements identified in the prior analysis of NARMAX modelling. Specifically, this includes the nonlinear mapping function *F*, the maximum time delay n_y , n_u , and n_e, and the initial delay d_y , d_u , and d_e for system output signal, input signal and noise signals, respectively, These hyperparameters are key in defining the structure of NARMAX models. The uncertainty in a model's structure can thus be regarded as a direct consequence of the variety of choices available for these hyperparameters.

Non-linear mapping functions play a pivotal role in approximating and fitting complex dynamic systems that are unknown. Specifically, for NARMAX models, the frequently utilized functions encompass polynomial functions, rational models, neural networks, fuzzy logic-based models, wavelet functions, Radial Basis Function (RBF) networks, and others.

Each non-linear function carries unique merits and limitations. Polynomial functions operate smoothly and their NARMAX model counterparts demonstrate superior accuracy. Additionally, they are adept at analysing and leveraging noise terms. Despite their transparency and interpretability, polynomial NARMAX models fall short when dealing with severe non-linear behaviours. Compared to its polynomial counterpart, the rational NARMAX model capably emulates severe non-linear behaviours, which represent more general expressions derivatives of the polynomial NARMAX model. This implies that the polynomial NARMAX model is a specific form of the rational NARMAX model. However, its interpretability is not as potent relative to the polynomial NARMAX model. The Neural Network-based NARMAX Model amalgamates the robust non-linear fitting advantages, particularly with a range of activation function choices and neuron weight training algorithms. This ensures such NARMAX models efficiently mimic systems' non-linear behaviours. Nonetheless, due to the complexity in neuron connections and activation function properties, these models lack interpretability. The fundamental concept of employing wavelet representations for dynamical systems hinges on the prospect of realizing the NARMAX model using wavelet decompositions. This involves expanding the unidentified function $F[\cdot]$ using combinations of multi-resolution wavelet basis functions. In most scenarios, only a minimal number of significant basis functions, derived from a redundant library of functions, are requisite for representing a non-linear dynamical system.

Therefore, for the NARMAX model, the selection probability of nonlinear mapping functions is equal. Only when there is prior knowledge or specific requirements from a particular case will there be a single requirement for the nonlinear mapping function. The diversity of choices inevitably brings about uncertainty in the model structure. For the NARMAX model that has already selected a nonlinear mapping function, such as the polynomial NARMAX model, the uncertainty of the model structure caused by the selection of nonlinear functions no longer exists. This is because there is no need to consider other nonlinear mappings in the NARMAX modelling process, and only needs to use the selected nonlinear mapping for fitting, modelling and prediction. Therefore, for a polynomial NARMAX model, its basic structure is determined - it’s based on a polynomial function.

For the NARMAX model that has already selected a nonlinear mapping function, such as the polynomial NARMAX model, the uncertainty of the model structure caused by the selection of nonlinear functions no longer exists. This is because there is no need to consider other nonlinear mappings in the NARMAX modelling process, and only needs to use the selected nonlinear mapping for fitting, modelling and prediction. Therefore, for a polynomial NARMAX model, its basic structure is determined - it’s based on a polynomial function.

Another factor that brings uncertainty to the structure of the NARMAX model is delay information, including the maximum time delay n_y, n_u, and n_e, and the initial delay d_y, d_u, and d_e for system output signal, input signal and noise signals, respectively. As shown in Figure 4, the system output, input and noise signal will be transformed into new variables by adding the time delay information for the NARMAX modelling. During this process, the value of the delay hyperparameter directly affects the candidate model term and its quantity. To more clearly express the uncertainty of model structure brought about by delay information, we first define a set *S*, whose elements are all candidate model terms generated based on delay information. Thus, we have

$$S = \{s_i, i = 1, 2, ..., f(n_y, n_u, n_e, d_y, d_u, d_e, F[\cdot])\} \tag{3.31}$$

where s_i is the candidate model terms, $f(\cdot)$ is the function to determine the number of candidate model terms based on the time delay information and the nonlinear mapping function $F[\cdot]$. For the polynomial NARMAX model, we have

$$f(\cdot) = \frac{(n+\ell)!}{n!\ell!} \tag{3.32}$$

where $n = n_y + n_u + n_e - (d_y + d_u + d_e) + 1$, typically $d_y = d_e = 1$ and $d_u = 1$, and $n = n_y + n_u - (d_y + d_u) + 1$ for the polynomial NARX model; ℓ is the nonlinearity degree of the model. The candidate model terms in polynomial NARMAX model are described as model (26).

For a certain nonlinear mapping function based NARMAX model, the number of elements in *S*, that is, the number of candidate model terms, has a direct relationship with delay information. However, since modelling is a study of unknown systems, the dynamic characteristics of the system cannot be accurately grasped. Therefore, in many cases, due to the inaccuracy of time delay values, it can’t be guaranteed that all valid information is included in the set *S* of candidate model terms, which is

$$S_{true} \not\subset S \tag{3.33}$$

where S_{true} is the set containing all model terms of the real systems.

Usually, d_y, d_u, and d_e are set to 1 or 0, and n_y, n_u, and n_e can increase indefinitely. Ideally, as the time delay information increases, it gets closer to the dynamic characteristics of the real system, and the probability that the actual model item is an element of set *S* becomes higher. However, an infinite maximum time delay still cannot guarantee that all real system model items are included in the generated candidate model items. Moreover, as the number of candidate model items increases, so does modelling redundancy. Redundant model items are considered noise for NARMAX modelling objectives. Therefore, more redundant model items equate to a lower signal-to-noise ratio, leading to greater uncertainty caused by noise. From equation (3.32), for polynomial NARMAX model, ℓ is also a key factor that determines the number of the candidate model terms. With larger ℓ, more candidate model terms will be generated for the polynomial NARMAX models, which leads to the same model structure uncertainty as the time delay.

The model structure uncertainty in NARMAX models is primarily determined by the choice of fitting function used and its structural form. Nonlinear mapping functions, particularly in NARMAX models, play a crucial role in fitting and approximating unknown, dynamic systems. Function types can include polynomial functions, rational models, neural networks, fuzzy logic-based models, wavelet functions, Radial Basis Function networks, etc. Each comes with unique strengths and weaknesses. Hyperparameters such as nonlinear mapping function, maximum time delay, and initial delay for system output, input and noise signals, and form keys in defining the structure of NARMAX models. The uncertainty in a model's structure can, therefore, be seen as a direct result of the varying choices available for these hyperparameters. Such uncertainty is a challenge when working with unknown systems where the system's dynamic characteristics cannot be accurately determined. However, the model structure's uncertainty driven by the selection of nonlinear functions ceases to exist once a specific nonlinear mapping function, like a polynomial, is chosen.

3.3.1.3 Model parameter uncertainty

Model parameter uncertainty in NARMAX refers to the uncertainty caused by parameters of the model terms selected in the NARMAX models. However, numerous cases arise where a system's internal parameters are influenced by other external variables or parameters. External influences that impact a system's dynamic behaviour because what is termed as the External Parameter-Dependent (EPD) or External Signal-Dependent (ESD) problem. This contrasts with the typical Internal Parameter-Dependent (IPD) problem in which the parameters are explicitly present and factored into the system model. In EPD or ESD problems, although the external parameters do not explicitly appear in the system model, they indirectly influence the system behaviour either via internal parameters or another interface.

First, we analyse the model parameter uncertainty caused by the system internal parameters.

Denote a NARMAX model in the linear-in-the-parameters representation:

$$y(t)=\sum_{i=1}^{M}\theta_i\varphi_i(\mathbf{x}(t))+e(t) \tag{3.34}$$

where

$$\mathbf{x}(t)=\left[y(t-1),\ldots,y\left(t-n_y\right),u(t-1),\ldots,u\left(t-n_u\right),e(t-1),\ldots,e\left(t-n_e\right)\right]^{\mathrm{T}} \tag{3.35}$$

and $\varphi_i(\mathbf{x}(t))$, $i=1,2,...,M$ refers to the regressors produced by nonlinear functions and predetermined lagged system variables, M is the number of the selected model terms, θ_i is the parameter related to the model terms $\varphi_i(\mathbf{x}(t))$. Based on the OLS class of algorithms, the estimation of the parameters θ_i does not directly apply the least squares (LS) method, but integrates feature selection, that is, the choice of model terms and corresponding parameter estimation.

Unlike the LS optimization, the objective of NARMAX modelling is to get an auxiliary model

$$y(t)=\sum_{i=1}^{M}g_i w_i(t)+e(t) \tag{3.36}$$

where $w_i(t)\ (i=1,2,...,M)$ are orthogonal terms across the estimation dataset consisting of N samples as

$$\sum_{t=1}^{T} w_i(t)w_j(t) = \begin{cases} d_i, i = j \\ 0, i \neq j \end{cases} \tag{3.37}$$

with $d_i = \sum_{t=1}^{T} w_i^2(t) \neq 0$.

Assuming that the model terms $\varphi_i(\mathbf{x}(t))$ are linear independent across the estimation dataset, the orthogonalization procedure can be summarized as the follows, originating from the first model term

$$\begin{cases} w_1(t) = p_1(t) \\ w_2(t) = p_2(t) - a_{1,2}w_1(t) \\ \dots \\ w_m(k) = p_m(k) - \sum_{r=1}^{m-1} a_{r,m}w_r, m = 2,3,\dots,M \end{cases} \tag{3.38}$$

where

$$a_{r,m} = \frac{\sum_{t=1}^{T} p_m(t)w_r(t)}{\sum_{t=1}^{T} w_r^2(t)}, 1 \leq r \leq m-1 \tag{3.39}$$

Thus, we have

$$g_i = \frac{\sum_{t=1}^{T} y(t)w_i(t)}{\sum_{t=1}^{T} w_i^2(t)}, i = 1,2,\dots,M \tag{3.40}$$

The final NARMAX model is the linear combination of M model terms shown in model (3.36), which is equal to model (3.14). Thus, the parameters θ_i are related to the orthogonal parameters g_i as

$$\begin{cases} \theta_M = g_M \\ \theta_{M-1} = g_{M-1} - a_{M-1,M}\theta_M \\ \theta_{M-2} = g_{M-2} - a_{M-2,M-1}\theta_{M-1} - a_{M-2,M}\theta_M \\ \quad\vdots \qquad\quad \vdots \qquad\qquad \vdots \\ \theta_m = g_m - \sum_{j=m+1}^{M} a_{m,j}\theta_j, m = M-1, M-2,\dots,1 \end{cases} \tag{3.41}$$

Therefore, in the NARMAX model, there is a direct relationship between the NARMAX model structure and its associated parameter θ_i. That is to say, once the specific structure of the NARMAX model and its terms are determined, the related parameter values θ_i can be directly obtained according to equation (3.40) and equation (3.41). Consequently, there's a positive correlation between the model parameter uncertainty by the internal parameters and structural uncertainty.

In [225], the externalparameter-dependent (EPD) problem of NARMAX modelling has been proposed. Define an EPD model as:

$$y(t) = F\left[y(t-1),\dots,y\left(t-n_y\right),u(t-1),\dots,u\left(t-n_u\right),\boldsymbol{\theta}(\boldsymbol{v})\right] + e(t) \tag{3.42}$$

where $F[\cdot]$ is the nonlinear function; $\boldsymbol{\theta}(\boldsymbol{v}) \in \Theta$ is the internal parameter vector, a function dependent on the external parameter set $v \in \Omega$. The sets Θ and Ω correspond respectively to the internal and external parameter sets.

In the EPD problem, the most important factor is assumption that v differs from the different cases. Thus, we have

$$y(t)=\begin{cases} f^{(1)}\left(y(t-1),\ldots,y\left(t-n_y\right),u(t-1),\ldots,u\left(t-n_u\right),\boldsymbol{\theta}_1\left(\boldsymbol{v}_1\right)\right),\text{s.t.}\boldsymbol{v}_1 \\ f^{(2)}\left(y(t-1),\ldots,y\left(t-n_y\right),u(t-1),\ldots,u\left(t-n_u\right),\boldsymbol{\theta}_2\left(\boldsymbol{v}_2\right)\right),\text{s.t.}\boldsymbol{v}_2 \\ \vdots \\ \vdots \\ f^{(L)}\left(y(t-1),\ldots,y\left(t-n_y\right),u(t-1),\ldots,ut\left(k-n_u\right),\boldsymbol{\theta}_L\left(\boldsymbol{v}_L\right)\right),\text{s.t.}\boldsymbol{v}_L \end{cases} \tag{3.43}$$

where L denotes L experiments/cases, $f^i(\cdot)(i=1,2,...,L)$ are different functions including linear and nonlinear sharing the same mapping type; $s.t.v_i$ implies that the individual model is contingent upon the exogenous parameter v_i.The goal of the EPD problem is to find a commonly structured NARMAX model within the L experiments. Suppose a common NARMAX model has been identified through L subsets, the estimation of the relevant model parameters is carried out in each individual subsets of L experiments. The identical NARMAX model serves in estimating parameters across disparate datasets. Consequently, the uncertainty inherent in the model parameters disassociates from the model structure's uncertainty, emerging as an independent variable influencing the model's uncertainty.

3.3.1.4 Model uncertainty in criteria

The uncertainty in a model, which may arise due to varying evaluation standards, can be understood as the ambiguity introduced by different validation methods. Central to all system identification efforts, model validation, inclusive of NARMAX, serves to confirm that the model realistically represents the actual system, as opposed to merely fitting it to system response values. Thus, if an applied validation standard focuses solely on fitting system response values rather than accurately depicting the actual system, it may result in a model that shows excellent performance on existing datasets but lacks the capacity to generalize to real-world applications, particularly in predicting future system responses. Consequently, model validation criteria remain both non-negotiable and essential components of system identification.

Optimally, the evaluation criteria for models should be independent of the modelling methods. This means that all model types, be they NARMAX or otherwise, need to employ the same standards for performance evaluation. As highlighted in the literature review in Chapter 2, there are typically two evaluation methods for non-linear system identification: statistical-based verification and qualitative verification. Either evaluation method proves effective in comparing and measuring models, including NARMAX, facilitating the selection of the most suitable model from a pool of potential options.

The evaluation standard based on the minimum MSE can effectively measure the accuracy of a model. Generally speaking, MSE is used to measure the error between the predicted values and actual values of a model in a test set. There are two methods for obtaining predicted values in the NAMRAX model's test set: one-step-ahead (OSA) and model predicted output (MPO).

OSA refers to one-step forecasting, let's first define this model:

$$y(t)=a_0+a_1y(t-1)+a_2(u(t-2))^2+e(t) \tag{3.44}$$

The OSA predictions of model (3.44) starting from step 3 is defined as:

$$\begin{cases} \hat{y}(3) = a_0 + a_1 y(2) + a_2 (u(1))^2 \\ \hat{y}(4) = a_0 + a_1 y(3) + a_2 (u(2))^2 \\ \ldots \\ \hat{y}(t) = a_0 + a_1 y(t-1) + a_2 (u(t-2))^2 \end{cases} \tag{3.45}$$

It can be seen that the prediction of current time will utilize the observation at the previous time, while the MPO predictions of model (3.45) from step 3 is denoted as:

$$\begin{cases} \hat{y}(3) = a_0 + a_1 \hat{y}(2) + a_2 (u(1))^2 \\ \hat{y}(4) = a_0 + a_1 \hat{y}(3) + a_2 (u(2))^2 \\ \ldots \\ \hat{y}(t) = a_0 + a_1 \hat{y}(t-1) + a_2 (u(t-2))^2 \end{cases} \tag{3.46}$$

In OSA predictions, each calculation is virtually reset at every step given that measured outputs are used on the right-hand side. This inhibits the accumulation of errors. Alternatively, in equation (3.46), after initialization, only the predicted output values are used on the right-hand side. Consequently, any errors in the MPO values accumulate expediently, becoming clearly evident. This phenomenon explains why a deficient model (e.g., insufficient, biased, unstable) may yield satisfactory OSA predictions under certain circumstances. However, MPOs, which essentially serve as long-term predictions, typically illustrate when a model is inadequate.

Hence, when the NARMAX model accurately depicts the system's response and internal structure, the two prediction calculation methods tend to converge in evaluating NARMAX model performance. However, should the NARMAX model fail to comprehensively represent the system, these distinct prediction calculation methods are likely to provide differing assessment results regarding the performance of the NARMAX model.

But MSE is not the only evaluation criterion in NARMAX modelling. The objective in system identification is to pinpoint the simplest unbiased model that can adequately represent the system. An incremental increase in model terms will typically lead to a decrease in MSE. But the increased NARMAX model with smaller MSE may not the exact model to describe the system. Therefore, in order to fully evaluate and select the NARMAX model, other evaluation indicators such as APRESS, AIC and BIC are also used to determine the length of the NARMAX model.

While the uncertainty brought by model standards is similar to that of model structure, the biggest difference lies in that the uncertainty of model evaluation criteria is caused by independent assessment standards. Specifically, it involves measuring a NARMAX model with determined parameters for nonlinear functions, model terms and responses, choosing the optimal one from many possible models without involving the design of the model itself. Therefore, just like in NARMAX modelling, the uncertainty brought about by evaluation criteria forms an integral part of NARMAX modeming's inherent uncertainties.

However, when assessing the uncertainty of the NARMAX model, it is requisite to first establish one or more evaluation criteria to optimize the NARMAX model. Granted, the chosen evaluation criteria may not wholly evaluate the NARMAX model's performance. In such an instance, during the modelling process, should the evaluation criterion transition from uncertain to certain, the uncertainty introduced by these standards is no longer a concern.

Through qualitative analysis, it is discerned that while the uncertainty of the NARMAX model comprises four components, notably model uncertainty in data, model structure uncertainty, model parameters uncertainty, and model uncertainty in criteria. During the NARMAX modelling process, the uncertainty induced by data can be viewed as a constant. Similarly, uncertainty originating from evaluation criteria is perceived as definitive evaluation standards during modelling. Consequently, the uncertainty of the NARMAX model predominantly constitutes uncertainties in both model structure and parameters. In the following section, we intend to undertake a quantitative analysis of these uncertainties to establish a theoretical foundation for leveraging them to derive more suitable NARMAX models.

3.3.2 Quantitative analysis of the model uncertainty in NARMAX models and polynomial NARMAX models

Firstly, this section quantitatively analyses the uncertainty of the NARMAX model. To address the uncertainty mathematically, we introduce set theory to quantify the uncertainties. In order to quantify the uncertainty of the NARMAX model in terms of a set, we first need to define the elements within this set. As can be inferred from the previous section, there are various ways to express the uncertainty of a NARMAX model. In this thesis, however, we choose to represent this uncertainty through the NARMAX model itself. That is, any NARMAX model *F* that has been derived using NARMAX modelling methods on a given dataset *S* and meets system identification requirements while accurately describing and predicting real systems is considered an element within our set representing uncertainties in the NARMAX model.

Definition: Define a set *U*, which is the model uncertainty in NARMAX, as:

$$U = \{M_1(y,u,e,f,(d,n)),...,M_P(y,u,e,f,(d,n))\} \tag{3.47}$$

where $M_i(y,u,e,f,(d,n))\ (i=1,...,p)$, as the element of the uncertainty set *U*, is the NARMAX models that fulfils the objectives of the system identification, p is the number of the NARMAX models in the set; y, u, and e are the system output, input, and noise signals; f is the nonlinear function of the NARMAX model; (d,n) is the time delay information. It is easy to prove that the definition of the set *U* is reasonable.

Proof. In order to validate the rationality of the NARMAX uncertainty set *U*, it is imperative to establish that all elements within this set simultaneously comply with three conditions: disorder, diversity and determinacy [231].

Disorder: Disorder refers to the equal status of elements in a set, without any necessary sequential relationship. The element $M_i(y,u,e,f,(d,n))$ in the set *U* is a NARMAX model obtained from the same dataset *S*. These NARMAX models have met the objectives of system identification and received subjective or objective recognition. There is no issue of hierarchy among these models, thus satisfying the unordered nature of elements in a set.

Uniqueness: The definition of uniqueness is that any two elements in a set are considered to be different, meaning each element can only appear once. It is easy to prove that two NARMAX models M_1 and M_2 cannot be the same model in the set. Let M_1 and M_2 are two polynomial NARMAX models as:

$$M_1 = {}^1a_0 + {}^1a_1\,{}^1x_1 + {}^1a_2\,{}^1x_2 \tag{3.48}$$

$$M_2 = {}^2a_0 + {}^2a_1\,{}^2x_1 + {}^2a_2\,{}^2x_2 \tag{3.49}$$

where ${}^1a_i\ (i=0,1,2)$ and ${}^2a_i\ (i=0,1,2)$ are the model parameters, ${}^1x_i$ and ${}^2x_i$ are the model terms in two NARMAX models. If M_1 and M_2 are two same elements, it means that model terms and parameters are completely consistent

$${}^{1}a_i = {}^{2}a_i \tag{3.50}$$

$${}^{1}x_i = {}^{2}x_i \tag{3.51}$$

Then we have

$$\begin{aligned} M_1 &= {}^{1}a_0 + {}^{1}a_1\,{}^{1}x_1 + {}^{1}a_2\,{}^{1}x_2 \\ &= {}^{2}a_0 + {}^{2}a_1\,{}^{2}x_1 + {}^{2}a_2\,{}^{2}x_2 \\ &= M_2 \end{aligned} \tag{3.52}$$

Thus, M_1 and M_2 are actually the same model.

Therefore, each element of *U* represents a unique NARMAX model, even if the structure and terms of the NARMAX model are exactly the same, there will still be differences in model parameters, thus meeting the requirement for uniqueness.

Determinacy: Determinacy refers to the concept that given a set, any element can only have two states: either belonging or not belonging to this set, and there is no ambiguous situation. In terms of the NARMAX model, according to the definition of elements, as long as it meets the goal of system identification, then this NARMAX model is an element in set *U*. Therefore, the definition of set *U* is complete and reasonable.

Based on the qualitative analysis of NARMAX model uncertainty in the previous section, it is known that the uncertainty of the NARMAX model consists of four aspects: data, structure, parameters, and evaluation criteria. Therefore, set *U* can be defined as the union of uncertainties caused by these four factors.

Based on the qualitative analysis of the model uncertainty in NARMAX, the set *U* can be represented as the union of the sets of four types of model uncertainty, that is

$$U_{NARMAX} = U_{data} \cup U_{para} \cup U_{struc} \cup U_{crit} \tag{3.53}$$

where U_{data}, U_{para}, U_{struc}, and U_{crit} correspond to the sets of uncertainty arising from data, parameters, model structure, and criteria, respectively, as shown in Figure 10. It's crucial to underline that *U* is inherently defined by its respective system, implying that during the NARMAX modelling execution, comparing the uncertainties of distinct systems is not feasible. Every system identification encompasses its exclusive *U*.

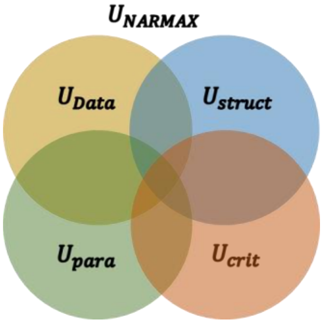


Figure 10 The Venn diagram of the model uncertainty in NARMAX

As defined, the components within *U* are NARMAX models M_i that fulfil the requirements of system identification. Meaning, due to variances in data, structure, parameters, and criteria, numerous NARMAX models exist that meet system identification objectives. The components within U_{data}, U_{para}, U_{struc}, and U_{crit} represent a collection of NARMAX models, M_i, influenced by these four factors to achieve the system identification goal. The definitions for these four subsets, analysed qualitatively in the previous section, are as follows:

$$\begin{aligned}
U_{data} &= \{M_j(y,u,e,f,(d,n)),\ j=1,2,\ldots,p_1 \mid e=e_w,e_c,\ldots\} \\
U_{struc} &= \{M_k(y,u,e,f,(d,n)),\ k=1,2,\ldots,p_2 \mid f=(poly, wavelet,\ldots), d,n \in \mathbb{Z}_0^+\} \\
U_{para} &= \{M_l(y,u,e,f,(d,n)),\ l=1,2,\ldots,p_3 \mid \theta \in \mathbb{R}\} \\
U_{crit} &= \{M_q(y,u,e,f,(d,n)),\ q=1,2,\ldots,p_4 \mid \hat{y}^{OSA}(M_q), \hat{y}^{MPO}(M_q),\ldots\}
\end{aligned} \tag{3.54}$$

where M_j, M_k, M_p, and M_q are identified NARMAX models affected by different factors. p_1, p_2, p_3, and p_4 are the number of NARMAX models in each subset. In set U_{data}, the elements M_j are determined by the noise model terms $e=e_w,e_c,\ldots$, where e_w and e_c represent the white and coloured noise model terms; and other noise signals. In U_{struc}, M_k are primarily influenced by the nonlinear mapping function *f*, and the delay hyperparameters *d* and *n*. According to the literature review analysis, the nonlinear mapping functions in NARMAX modelling mainly include polynomial functions, wavelet-like functions, neural networks etc., hence it's defined that *f* belongs to one of these series of nonlinear mapping functions. And both *d* and *n* take non-negative integer values. In U_{para}, θ is the model parameter belonging to $\mathbb{R}$, while in U_{crit}, M_q are determined independently due to the criteria.

Corollary For an unknown system, compared to the U_{crit}, the other three sets can be defined as infinite sets. While due to the existence of the prior knowledge, all subsets U_{data}, U_{para}, U_{struc}, and U_{crit} can be considered as finite sets. Therefore, set U_{NARMAX} is also a finite set. Consequently, the size of uncertainty in the NARMAX model can be determined by measuring the size of this set. From the definition of *U*, we have

$$L_{uncertainty}^{NARMAX} \propto p \tag{3.55}$$

where $L_{uncertainty}^{NARMAX}$ indicates the level of model uncertainty in NARMAX, *p* refers to the number of NARMAX models in *U*.

Similarly, we have

$$\begin{aligned}
L_{uncertainty}^{data} &\propto p_1 \\
L_{uncertainty}^{struc} &\propto p_2 \\
L_{uncertainty}^{para} &\propto p_3 \\
L_{uncertainty}^{crit} &\propto p_4
\end{aligned} \tag{3.56}$$

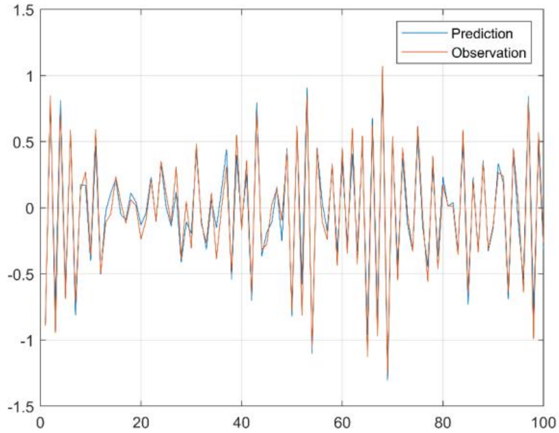
where $L_{uncertainty}^{data}$, $L_{uncertainty}^{struc}$, $L_{uncertainty}^{para}$, and $L_{uncertainty}^{crit}$ are the level of model uncertainty in data, structure, parameter and criteria, and p_1, p_2, p_3, and p_4 are the number of NARMAX models in U_{data}, U_{para}, U_{struc}, and U_{crit}, respectively.

However, for each subset, the number of NARMAX models, that is, the size of the subset does not have a proportional relationship with the impact factor of each subset.

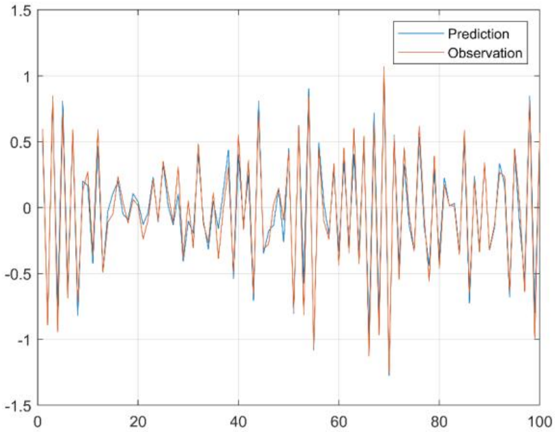
Consider a nonlinear system [232]

$$y(t) = -0.605y(t-1) - 0.163y^2(t-2) + 0.588u(t-1) - 0.240u(t-2) - 0.15\sin(u(t-1)) + \xi(t) \tag{3.57}$$

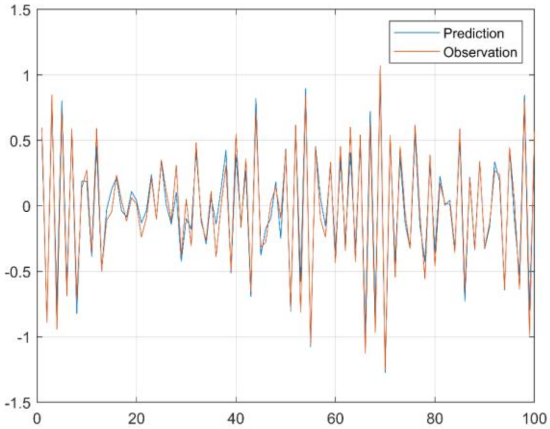
where the system noise $\xi(t) \sim N(0,0.1)$ and $u(t) \sim U[-1,1]$, and the total data points are set to be 500. The aim is to discern both the structure of the model and the unidentified parameters within the model, relying on the documented samples. In the modelling, we set $f = poly$, and $d=1$. Predictions in the testing set (the last 100 points) by NARMAX models with different structures and parameters are shown in Figure 11, and the models by different hyperparameters are listed in Table 1.



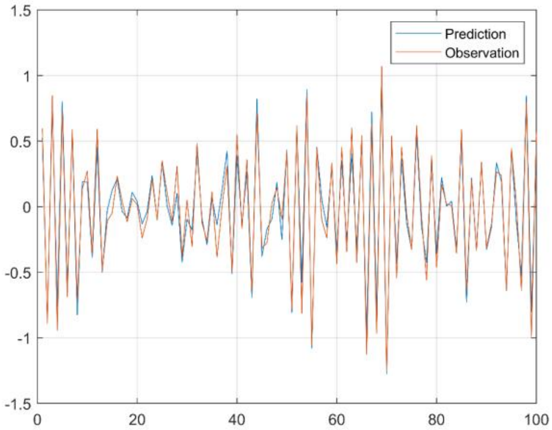
a) $\ell = 2, n_y = n_u = 2$ (MSE=0.0084)



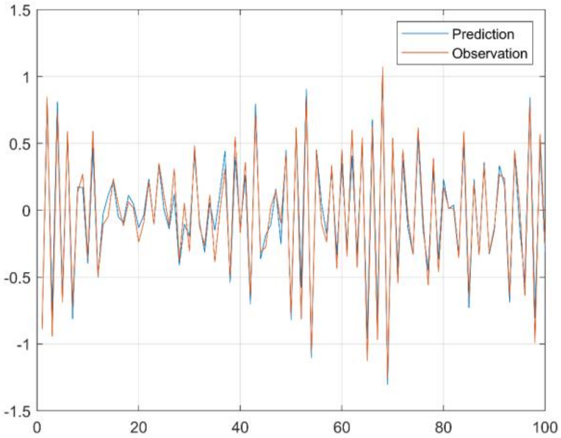
b) $\ell = 2, n_y = n_u = 3$ (MSE=0.0091)



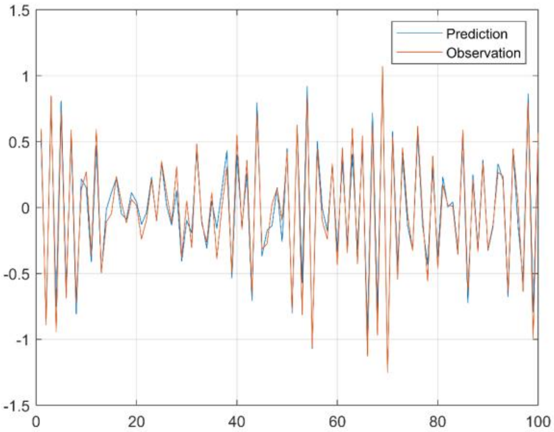
c) $\ell = 2, n_y = n_u = 4$ (MSE=0.0087)



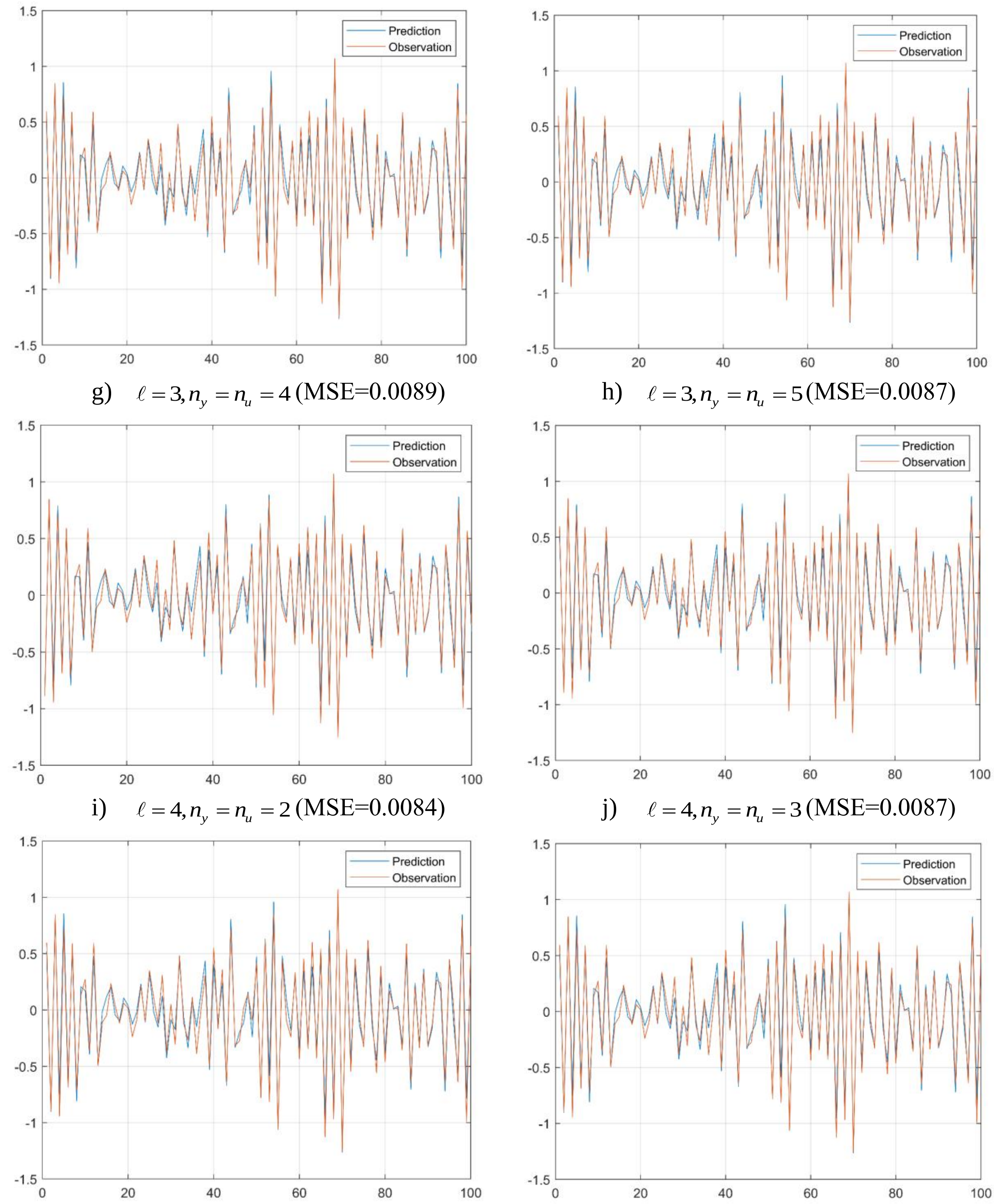
d) $\ell = 2, n_y = n_u = 5$ (MSE=0.0087)



e) $\ell = 3, n_y = n_u = 2$ (MSE=0.0084)



f) $\ell = 3, n_y = n_u = 3$ (MSE=0.0091)



g) $\ell = 3, n_y = n_u = 4$ (MSE=0.0089)

h) $\ell = 3, n_y = n_u = 5$ (MSE=0.0087)

i) $\ell = 4, n_y = n_u = 2$ (MSE=0.0084)

j) $\ell = 4, n_y = n_u = 3$ (MSE=0.0087)

k) $\ell = 4, n_y = n_u = 4$ (MSE=0.0090)

l) $\ell = 4, n_y = n_u = 5$ (MSE=0.0087)

Figure 11 Comparison between observations and predictions in the testing set (100 data points) of different NARMAX models (model length = 5)

By analysing the model (3.57), when $f = poly$, the polynomial NARMAX cannot find the 'true' model as there is a non-polynomial model term $\sin(u(t-1))$, which is a common situation for most applications. Noticed that no system identification method can develop the 'true' model when the prior knowledge about $\sin(u(t-1))$ is missing. Still, we can have several best models and transparent model structures to describe the system.

As shown in Figure 11, the 12 'best' NARMAX models have the similar performance over the testing period indicated by MSEs, which are around 0.008. Moreover, the interpretable model structures are listed in Table 1, where there are 8 unique NARMAX models with 12 different hyperparameters determining the model structure. This indicates that although the number of NARMAX models that meet system identification purposes does not increase proportionally with the setting of hyperparameters, i.e., $N(model) = N(hyperparameter)$, it also suggests a positive correlation between the quantity of NARMAX models and the value of hyperparameters for unknown systems. That is to say, there's a direct relationship between model uncertainty and hyperparameters.

Table 1 The NARMAX models by different hyperparameters (model length = 5)

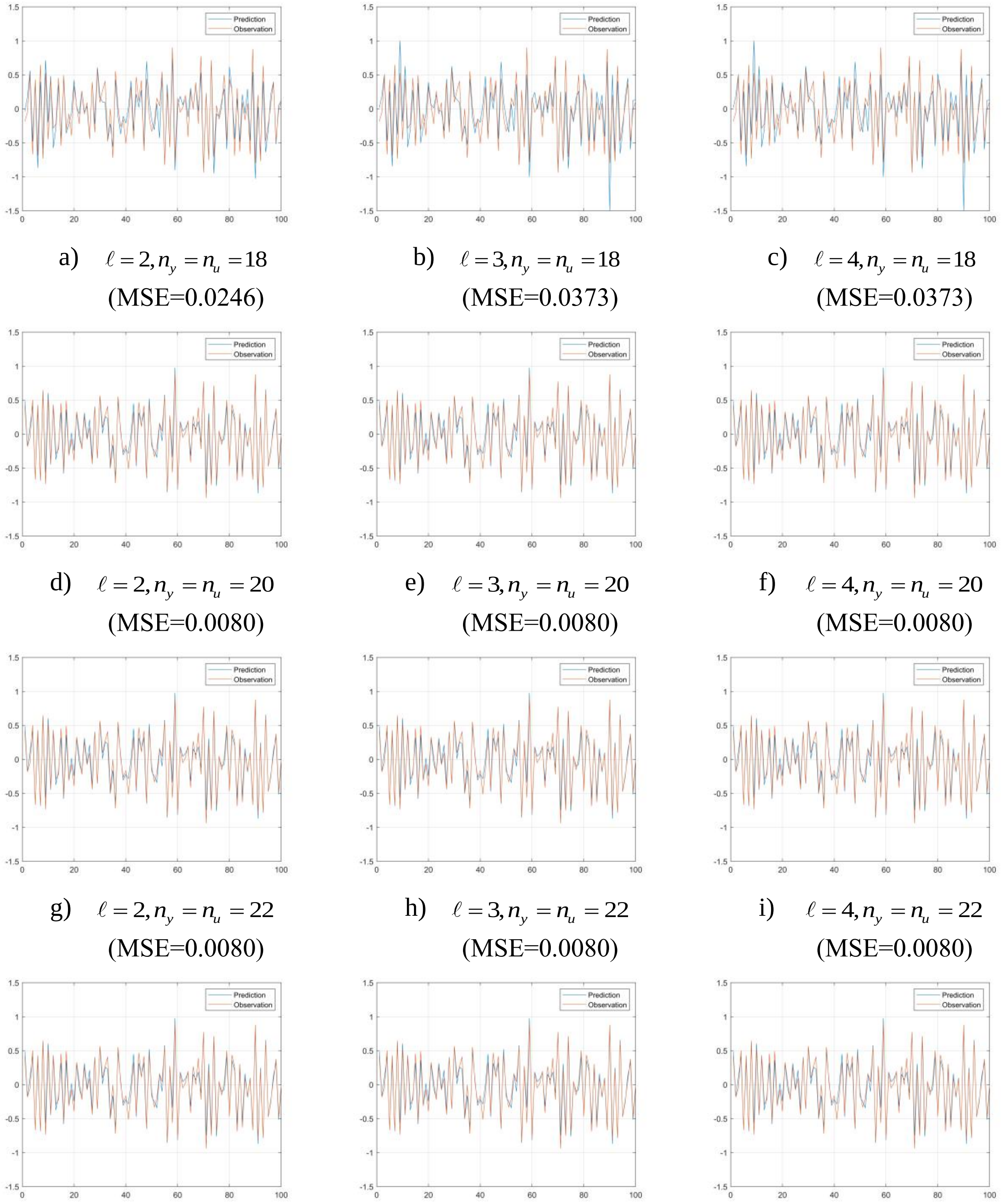
ℓ	Model term index	$n_y, n_u = 2$	$n_y, n_u = 3$	$n_y, n_u = 4$	$n_y, n_u = 5$
2	1	$y(t-1)$	$y(t-1)$	$y(t-1)$	$y(t-1)$
	2	$u(t-1)$	$u(t-1)$	$u(t-1)$	$u(t-1)$
	3	$u(t-2)$	$u(t-2)$	$u(t-2)$	$u(t-2)$
	4	$y(t-2)^2$	$y(t-2)^2$	$y(t-2)^2$	$y(t-2)^2$
	5	$u(t-2)^2$	$u(t-3)y(t-3)$	$u(t-2)u(t-4)$	$u(t-2)u(t-4)$
3	1	$y(t-1)$	$y(t-1)$	$y(t-1)$	$y(t-1)$
	2	$u(t-1)$	$u(t-1)$	$u(t-1)$	$u(t-1)$
	3	$u(t-2)$	$u(t-2)$	$u(t-2)$	$u(t-2)$
	4	$y(t-2)^2$	$y(t-2)^2$	$y(t-2)^2$	$y(t-2)^2$
	5	$u(t-2)^2$	$u(t-3)y(t-3)$	$u(t-1)u(t-4)^2$	$u(t-2)u(t-4)$
4	1	$y(t-1)$	$y(t-1)$	$y(t-1)$	$y(t-1)$
	2	$u(t-1)$	$u(t-1)$	$u(t-1)$	$u(t-1)$
	3	$u(t-2)$	$u(t-2)$	$u(t-2)$	$u(t-2)$
	4	$y(t-2)^2$	$y(t-2)^2$	$y(t-2)^2$	$y(t-2)^2$
	5	$u(t-2)^4$	$u(t-3)u(t-2)^2 y(t-3)$	$u(t-4)^2 y(t-3)u(t-3)$	$u(t-5)^2 u(t-4)u(t-2)$

Consider another nonlinear system:

$$y(t) = -0.605y(t-1) - 0.163y^2(t-2) + 0.588u(t-1) - 0.240u(t-20) + \xi(t) \tag{3.58}$$

where the noise signal $\xi(t)$ and the input signal $u(t)$ have the same definition with model (3.57). Similarly, we define $f = poly$ and $d = 1$. Compared to model (3.57), this system is straightforward and easy to identify. However, model (3.58) is still unknown to users. Predictions in the testing set (the last 100 points) by NARMAX models with different structures and parameters are shown in Figure 12, and the models by different hyperparameters are listed in Table 2.

As shown in Figure 12 and Table 2, when n_y and n_u is not sufficient to the 'true' system, the identified NARMAX models have similar model structure with the 'true' model and relative accurate predictions. However, if the hyperparameters are enough to model, NARMAX models always find the 'best' model, which shall be the only identified NARMAX model in the set *U*. Thus, there is no uncertainty in this situation. However, for most applications in the real world, this situation is rare. Therefore, for most NARMAX models, we define the model uncertainty in NARMAX has a positive correlation with the hyperparameters in equation (3.34).



a) $\ell=2, n_y=n_u=18$ (MSE=0.0246)

b) $\ell=3, n_y=n_u=18$ (MSE=0.0373)

c) $\ell=4, n_y=n_u=18$ (MSE=0.0373)

d) $\ell=2, n_y=n_u=20$ (MSE=0.0080)

e) $\ell=3, n_y=n_u=20$ (MSE=0.0080)

f) $\ell=4, n_y=n_u=20$ (MSE=0.0080)

g) $\ell=2, n_y=n_u=22$ (MSE=0.0080)

h) $\ell=3, n_y=n_u=22$ (MSE=0.0080)

i) $\ell=4, n_y=n_u=22$ (MSE=0.0080)

j) $\ell=2, n_y=n_u=24$ (MSE=0.0080)

k) $\ell=3, n_y=n_u=24$ (MSE=0.0080)

l) $\ell=4, n_y=n_u=24$ (MSE=0.0080)

Figure 12 Comparison between observations and predictions in the testing set (100 data points) of different NARMAX models (model length = 4)

Table 2 The NARMAX models by different hyperparameters (model length = 4)

ℓ	index	$n_y, n_u = 2$	$n_y, n_u = 3$	$n_y, n_u = 4$	$n_y, n_u = 5$
2	1	$u(t-1)$	$u(t-1)$	$u(t-1)$	$u(t-1)$
	2	$y(t-1)$	$y(t-1)$	$y(t-1)$	$y(t-1)$
	3	$y(t-2)^2$	$u(t-20)$	$u(t-20)$	$u(t-20)$

	4	$y(t-18)$	$y(t-2)^2$	$y(t-2)^2$	$y(t-2)^2$
3	1	$y(t-1)$	$u(t-1)$	$u(t-1)$	$u(t-1)$
	2	$y(t-1)$	$y(t-1)$	$y(t-1)$	$y(t-1)$
	3	$y(t-2)^2$	$u(t-20)$	$u(t-20)$	$u(t-20)$
	4	$y(t-1)^2\,y(t-18)$	$y(t-2)^2$	$y(t-2)^2$	$y(t-2)^2$
4	1	$y(t-1)$	$u(t-1)$	$u(t-1)$	$u(t-1)$
	2	$u(t-1)$	$y(t-1)$	$y(t-1)$	$y(t-1)$
	3	$u(t-2)$	$u(t-20)$	$u(t-20)$	$u(t-20)$
	4	$y(t-1)^2\,y(t-18)$	$y(t-2)^2$	$y(t-2)^2$	$y(t-2)^2$

3.3.3 Optimizing and probabilistic NARMAX modelling based on the model uncertainty

The purpose of system identification is to construct a model that can accurately describe the system and precisely predict its future responses. Therefore, NARMAX modelling involves adjusting hyperparameter settings, studying different structures and parameters of NARMAX models, and using one or more evaluation criteria to obtain the 'optimal' NARMAX model. Based on the quantitative analysis in the previous section, an 'optimal' NARMAX model equates to eliminating uncertainty in NARMAX modelling.

Define the modelling objective function as

$$J(y,u,e,f,[d,y],\theta,M_0)=\varepsilon(y,\hat{y}) \tag{3.59}$$

where y, u , and e are the system output, input, and noise signals; f is the nonlinear function of the NARMAX model; (d,n) is the time delay information; θ is the NARMAX model parameters, and M_0 is the length of the NARMAX model. ε denotes the difference between the system y and the NARMAX model $\hat{y}$, which is defined as

$$\begin{aligned} y(t)=\theta_0+\sum_{i_i=1}^{n} f_{i_1}\left(x_{i_1}(t)\right)+\sum_{i_1=1}^{n}\sum_{i_2=i_1}^{n} f_{i_1 i_2}\left(x_{i_1}(t),x_{i_2}(t)\right)+\cdots \\ +\sum_{i_1=1}^{n}\cdots\sum_{i_\ell=i_{\ell-1}}^{n} f_{i_1 i_2\cdots i_l}\left(x_{i_1}(t),x_{i_2}(t),\ldots,x_{i_\ell}(t)\right)+e(t) \end{aligned} \tag{3.60}$$

where

$$x_m(t)=\begin{cases} y(t-m) & d_y+1\le m\le n_y-d_y \\ u\left(t-\left(m-n_y\right)\right) & n_y+d_u+1\le m\le n_y+n_u-d_u \\ e\left(t-\left(m-n_y-n_u\right)\right) & n_y+n_u+d_e+1\le m\le n_y+n_u+n_e-d_e \end{cases} \tag{3.61}$$

Thus, the system identification can be seen as an optimization problem as

$$\begin{aligned}
&\min\ J(y,u,e,f,[d,n],\theta,M_0)=\varepsilon(y,\hat{y})\\
&s.t.\quad e\in\{\textit{white noise, coloured noise,...}\}\\
&\qquad\ f\in\{\textit{polynomial, wavelet, NN,...}\}\\
&\qquad\ d,n\in\mathbb{Z}_0^+\\
&\qquad\ \theta\in\mathbb{R}\\
&\qquad M_0\in\mathbb{Z}^+\\
&\qquad\ \varepsilon\in\{MSE,OSA,MPO,...\}
\end{aligned}\tag{3.62}$$

where the definition of the conditions has been discussed in the previous section.

Since the system is unknown, it can be assumed that all constructed models are 'pseudo-models'. Similarly, for NARMAX, there isn't a 'true' NARMAX model $\hat{y}$ capable of fully describing nonlinear dynamic systems. Despite this, it is desirable to obtain a 'better' model and more accurate system predictions by continuously optimizing the hyperparameters of NARMAX modelling. The process of obtaining a 'better' NARMAX model is essentially the continuous reduction of uncertainty in the NARMAX model as it increasingly approximates the real system. Thus, the optimization problem defined by (3.62) generally resorts to local optimization techniques. The complexities and high dimensionality of the problem pose challenges to finding a global optimum - a solution that yields the minimum possible error across the entire solution space. Depending particularly on the nature of the error surface - which is largely unknown, the optimization procedure could end up being trapped in one of potentially many local minima, rather than finding a global minimum. Implementing local optimization in this context could also be a matter of computational efficiency. The defined optimization problem, while it may not guarantee the best possible model, is computationally less expensive and quicker than more exhaustive techniques such as global optimization. Given the high-dimensional nature of the search space (i.e., the array of hyperparameters in the NARMAX model), using local optimization can expedite the process of finding a suitable model that can then be further refined.

However, in recent years, as opposed to single deterministic predictions, probabilistic forecasts of system responses have gradually gained attention and are widely used. Although there is a certain level of uncertainty in probabilistic forecasts, they provide a vague description rather than a definitive one for the system. This covers more possibilities for practical applications, ensuring a more comprehensive analysis and study of the system.

For NARMAX models, the prediction value under the same system input and output is definite for a specific NARMAX model. To generate probabilistic predictions though, multiple NARMAX models with similar performance are needed. According to the quantification definition from the previous section, the more NARMAX models there are, the greater uncertainty it brings about. Therefore, unlike the 'optimal' singular NARMAX model, probability-based forecast values from NARMAX models utilize this inherent uncertainty.

Optimal NARMAX models based on the model uncertainty The optimal NARMAX model is to minimize the objective function (3.62) to construct the 'best' NARMAX model and generate most accurate predictions of the system. Assume there is no prior knowledge about the system, where the constraint conditions, especially the range of values for $[d,n]$ is the set of non-negative integers, which is an infinite range. During the modelling process, in order to include as much system information as possible, the values of hyperparameters are set to be as large as possible. For example, $[d,n]$ is set to a sufficiently large positive integer. As for the polynomial NARMAX model, ℓ is also defined as a sufficiently large positive integer.

Taking a polynomial NARMAX modelling as an example, the general polynomial NARMAX model is defined as model (27). In the first iteration, let $d_y = d_e = 0$, $d_u = 1$, $n_y = n_u = n_e = n_1^{max}$, and $\ell = \ell_1^{max}$. Based on the equation (30), the number of the candidate model terms is calculated

$$M_1^{max} = \frac{(n + \ell_1^{max})!}{n!\ell_1^{max}!} \tag{3.63}$$

where $n = n_y + n_u + n_e = 3n_1^{max}$, while M_1^{max} candidate model terms form the candidate dictionary D_c^1. By applying the FROLS algorithm, the best desired model M_1^{poly} within the constrains conditions can be achieved. However, the set D_c^1 is redundant, meaning most candidate model terms are equivalent as the noise terms to the model. To mitigate the impact of these interference model terms, hyperparameters can be redefined and a new NARMAX model is trained. In the *s-th* ($s \geq 2$) iteration, let $d_y = d_e = 0$, $d_u = 1$, $n_y = n_u = n_e = n_s^{max}$, and $\ell = \ell_s^{max}$. The number of the candidate model terms in the *s-th* iteration is

$$M_s^{max} = \frac{(n + \ell_s^{max})!}{n!\ell_s^{max}!} \tag{3.64}$$

And the trained NARMAX model is y_s^{poly}. The iterative process will terminate when the newly trained NARMAX model is identical to the previous one, or when it reaches a predetermined number of iterations.

Then the set *U*, describing the model uncertainty and consisting of *q* 'best' NARMAX models, is defined as $U = \{M_1^{poly}, ..., M_q^{poly}\}$. Among all the NARMAX models in set *U*, an optimal model is then selected.

Probabilistic NARMAX modelling based on the model uncertainty Compared to the optimal NARMAX modelling, the probabilistic NARMAX modelling are several 'best' NARMAX models in the set *U* to provide a comprehensive understanding of the system and the probabilistic NARMAX predictions for the system. Consider q_l NARMAX models from the set *U* forming a subset U_{prob} as

$$U_{prob} = \{M^{poly}{}_1, ..., M^{poly}{}_{q_l}\} \tag{3.65}$$

Then the predictions from the q_l NARMAX models is

$$\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_{q_l}]^T \tag{3.66}$$

where $\mathbf{y}_i = [\hat{y}_i(T+1), ..., \hat{y}_i(T+h)]^T$ $(i = 1, 2, ..., q_l)$ is the predictions of the *i-th* NARMAX model over the time period from $T+1$ to $T+h$; *T* is the total data points in the dataset, *h* is the desirable time in the future.

Also, to describe the system, besides the selected NARMAX models in the set U_{prob}, a statistical analysis of the model terms with the related importance is obtained. The simultaneous existence of specific models and statistical analysis of model terms contributes to a more comprehensive understanding of the system.

3.4 Summary

This chapter delves into the methodology adopted for this thesis, with particular focus on the Nonlinear Autoregressive Moving Average model with eXogenous inputs (NARMAX), a recognized and widely referenced technique for the modelling of dynamic nonlinear systems. NARMAX distinctly excels at accurately identifying systems, extracting a clear and interpretable model, and generating precise system predictions. Despite its advantages, actual system identification remains a challenge with NARMAX modelling due to the lack of awareness of the system. Therefore, a decisive task is to identify the critical factors and their internal structural impacts on the system through modelling techniques.

This inevitably necessitates dealing with model uncertainty emerging from differing configurations of NARMAX hyperparameters. The presence of this uncertainty influences the selection of optimal models and the production of predictive values for system responses. Thus, a systematic approach to defining, analysing, and quantifying such uncertainties in NARMAX models is of utmost importance to effectively mitigate or minimize them during the modelling process.

In recent years, probabilistic models have begun to gain traction in system research due to their ability to provide a more extensive set of information and more valuable predictive values. Owing to the structure of NARMAX models, which is established post-training, the need arises for multiple similarly performing NARMAX models to provide probability information. This, however, increases the uncertainty inherent in these models.

Commencing with a brief introduction of NARMAX modelling in conjunction with the Forward Regression Orthogonal Least Squares (FROLS) algorithm, the chapter emphasizes its main attention on polynomial-based models due to their superior interpretability and stability. The chapter also offers both qualitative and quantitative assessments of uncertainties in these models, defining them as ‘NARMAX Model Uncertainty Sets’ using set theory. The conclusion drawn points to a direct link between the size or extent of these sets and the values assigned to their respective hyperparameters. Defining the objective function related to these uncertainties and the optimization processes involved in generating either singularly optimal or probabilistically optimal versions, the entire exercise appears similar to acquiring the best possible version(s) within an accepted range for those parameters.

Chapter 4

Deep Polynomial NARMAX Model

The polynomial NARMAX model is extensively utilized in dynamic non-linear system modelling due to its robust interpretability and simplicity. The Weierstrass theorem posits that any continuously differentiable function within a defined closed space can be expressed as a series of polynomial functions. Therefore, theoretically, the polynomial NARMAX model can describe a broad spectrum of non-linear systems. However, in actual application, most systems are black-box systems with undefined coupling relations between features and variables. This leads to a requirement to define a number of hyperparameters in polynomial NARMAX modelling, including the degree of non-linearity in polynomial functions and the delay information of variables. These hyperparameters have a significant influence on the structure and parameters of the polynomial NARMAX model. As a result, the ambiguity of these hyperparameters can introduce possible uncertainties in the modelling process.

To mitigate the effects of hyperparameter ambiguity in modelling, a large set of hyperparameters that encompasses maximal system information is defined initially. Following this, we create optimal polynomial NARMAX models from potential model elements discerned by these hyperparameters, which satisfies system identification requirements. However, this method poses two distinct problems. Firstly, the majority of potential model elements become redundant in the modelling process. This redundancy increases with the escalating values of these hyperparameters, necessitating more time for modelling. Additionally, it results in noise interference from less effective elements, thereby reducing the signal-to-noise ratio and affecting the quality of modelling. Secondly, the increase in these parameter numbers can cause a significant rise in the dimensionality of potential models. Despite the absence of theoretical limits on dimensions, it becomes infeasible to manage a vast quantity of candidates because of hardware limitations, particularly when using the FROLS algorithm to construct their respective models.

To address these challenges while still utilizing an extensive set of candidate models, efficiently managing redundancy, and improving the signal-to-noise ratio, we propose a novel deep Polynomial Network training optimization algorithm. This algorithm is designed to address potential issues associated with an increase in dimensionality and a reduced signal-to-noise ratio that may result from heightened parameter values. In response to these issues, we have introduced a network structure that includes both the generation layer and the driller layer, explaining their structures and functions separately. This study introduces training optimization algorithms founded on Particle Swarm Optimization (PSO) for the active selection of elements to construct the final model. The performance of the proposed networks is verified using multiple different cases, similar to actual applications. Our DeepNARMAX demonstrations suggest that it has the capacity to effectively identify useful information, construct optimal descriptions, and provide accurate, reliable predictions.

4.1 Introduction

NARMAX models prove to be extensively effective in identifying and predicting nonlinear systems. Particularly, the Polynomial NARMAX models offer a more precise and clear representation of the nonlinear system structure and the pivotal variables that influence the system. Inherently, these models possess an exceptional capability to grasp and simulate both linear and nonlinear dynamics and interactions, displaying their broad applicability. Polynomial NARMAX models extend linear models by accommodating the nonlinearities and interactions that fall beyond the reach of linear approaches.

These models are integral in circumventing hypothesis errors when a statistician mistakenly imposes linearity on a nonlinear system, thus preventing errors leading to inaccurate predictions and system identification.

The structure of the Polynomial NARMAX model comprises the order of the autoregressive part (historical values of the output), the order of the moving average part (past error values), and the maximum lag of the exogenous inputs (past or current values of external or exogenous variables). These parameters dictate the temporal depth at which previous outputs and inputs affect the current output. Primarily, this temporal depth sets the groundwork for the model's exceptional predictive ability. The model encapsulates historical values of the output, past error values, and past or current values of external or exogenous variables. Consequently, its strengths dwell in the ability to embody the stochastic elements, deterministic factors, and any additional influences into one singular equation.

Polynomial NARMAX models function by transposing the primary input data into a higher-dimensional space using a group of non-linear basis functions. This nonlinear transformation permits the formation of a model that remains linear in the parameters, facilitating the employment of efficient linear system identification methods for the estimation of these parameters. Further, Polynomial NARMAX models offer a clear comprehension of the significant variables that impact the system. Utilizing a polynomial form, one can discern the impacts of the interactions between various variables on the output, and the extent of these interactions. These models can symbolize a vast class of nonlinear dynamics and operate as a universal prediction modelling tool, providing them an edge over numerous other model structures that struggle to clearly represent variable interactions. It's pivotal to note that the selection of the model order, coupled with the choice of the non-linear basis functions, requires an expertise. It's a quandary that frequently demands a compromise between complexity and comprehension; a simplistic model might disregard vital details, whereas an overly complex model may vex interpretation.

Define the nonlinearity degree of the polynomial NARMAX model as ℓ, and the maximum time lags for output, input and noise signals as n_y , n_u , and n_e . The candidate model terms are constructed from the system's output y, input u and noise signals e based on the nonlinearity degree ℓ and maximum time delay parameter n_y , n_u , and n_e. Considering a single input single output system, where the output y , input u and noise signals e are linearly expanded with the time delay information as

$$\begin{aligned} &y(t) \to y(t-1), y(t-2), ..., y(t-n_y) \\ &u(t) \to u(t-1), u(t-2), ..., u(t-n_u) \\ &e(t) \to e(t-1), e(t-2), ..., e(t-n_e) \end{aligned} \tag{4.1}$$

Then, based on ℓ, transform the expanded linear model terms into nonlinear exponential terms:

$$\begin{aligned} &y : y(t-1)^{\ell}, y(t-2)^{\ell}, ..., y(t-n_y)^{\ell} \\ &u : u(t-1)^{\ell}, u(t-2)^{\ell}, ..., u(t-n_u)^{\ell} \\ &e : e(t-1)^{\ell}, e(t-2)^{\ell}, ..., e(t-n_e)^{\ell} \end{aligned} \tag{4.2}$$

And nonlinear cross model terms:

$$\begin{aligned}
&y: y(t-1)^{\ell-1}y(t-2),\dots,y(t-n_y+1)y(t-n_y)^{\ell-1}\\
&u: u(t-1)^{\ell-1}u(t-2),\dots,u(t-n_u+1)u(t-n_u)^{\ell-1}\\
&e: e(t-1)^{\ell-1}e(t-2),\dots,e(t-n_e+1)e(t-n_e)^{\ell-1}\\
&y\times u: y(t-1)^{\ell-1}u(t-1),\dots,y(t-n_y)u(t-n_u)^{\ell-1}\\
&y\times e: y(t-1)^{\ell-1}e(t-1),\dots,y(t-n_y)e(t-n_e)^{\ell-1}\\
&u\times e: u(t-1)^{\ell-1}e(t-1),\dots,u(t-n_u)e(t-n_e)^{\ell-1}
\end{aligned} \tag{4.3}$$

The number of the candidate model terms M is calculated as:

$$M = \frac{(n+\ell)!}{n!\ell!} \tag{4.4}$$

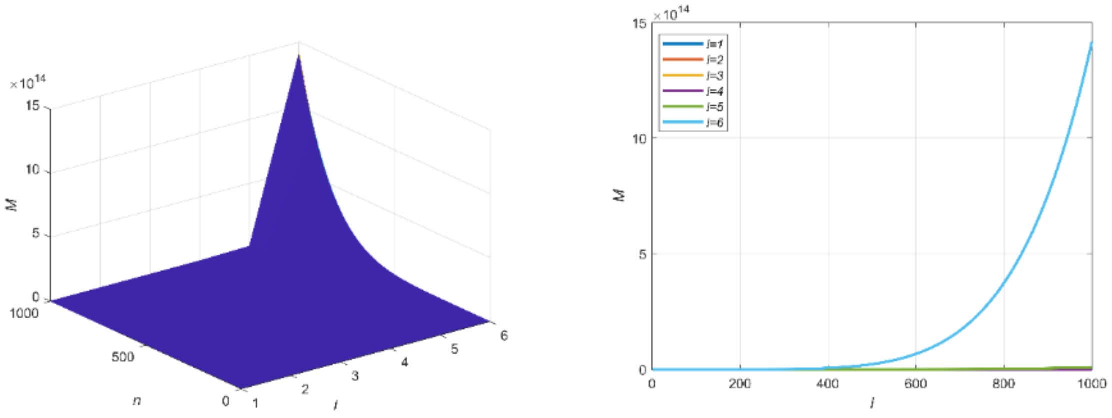
where $n = n_y + n_u + n_e$. Define a set D, in which the elements are the candidate model terms in equation (4.1), (4.2), and (4.3), as follows:

$$D = \{x_i, i = 1,2,\dots,M \mid x_i = y(t-1),u(t-1),\dots,u(t-n_u)e(t-n_e)^{\ell-1}\} \tag{4.5}$$

According to the FROLS algorithm, constructing a polynomial NARMAX model with L terms from set D requires L steps. For the first step, the algorithm involves M ERR calculation operation, while starting from the $s(s \geq 2)$ step, there are $(M-s+1)$ orthogonal operations and $(M-s+1)$ ERR calculation operations. Let the algorithm stops at L step, then the number of operations N_{op} is

$$\begin{aligned}
N_{op} &= \sum_{s=2}^{L} 2(M-s+1)+M\\
&= (2L-1)M - L^2 + 1
\end{aligned} \tag{4.6}$$

From equation (4.6), the number of operations including calculation of ERR values and orthogonalization of model terms rely on the dimension M and model length L. The larger N_{op}, the higher the complexity. Given the system is unknown, a higher degree of nonlinearity ℓ and larger time delay information n are defined during the modelling process to construct more model terms. This ensures a comprehensive description of the system without neglecting important information. However, according to equation (4.4), as ℓ and n increase, there's a risk of dimension explosion in the number or dimensions M of candidate model terms.



a) 3-D graph of dimension M varying with ℓ and n

b) 2-D graph of dimension M varying with ℓ and n

Figure 13 The relationship between the candidate model terms set D and the nonlinearity ℓ and time delay n

Assuming the range of values for ℓ is $[1,6]$ and the range of values for time delay n is $[1,1000]$, the dimension of set D is shown in Figure 13. Overall, as n and ℓ increase, M also increases. Moreover, with the continuous growth of n and ℓ, the value of M shows an explosive rise. When n and ℓ reach their maximum values within the range, the value of M is achieved to 1.4×10^{15}, which is a significant high value, demonstrating the dimension explosion with the growing of ℓ and n.

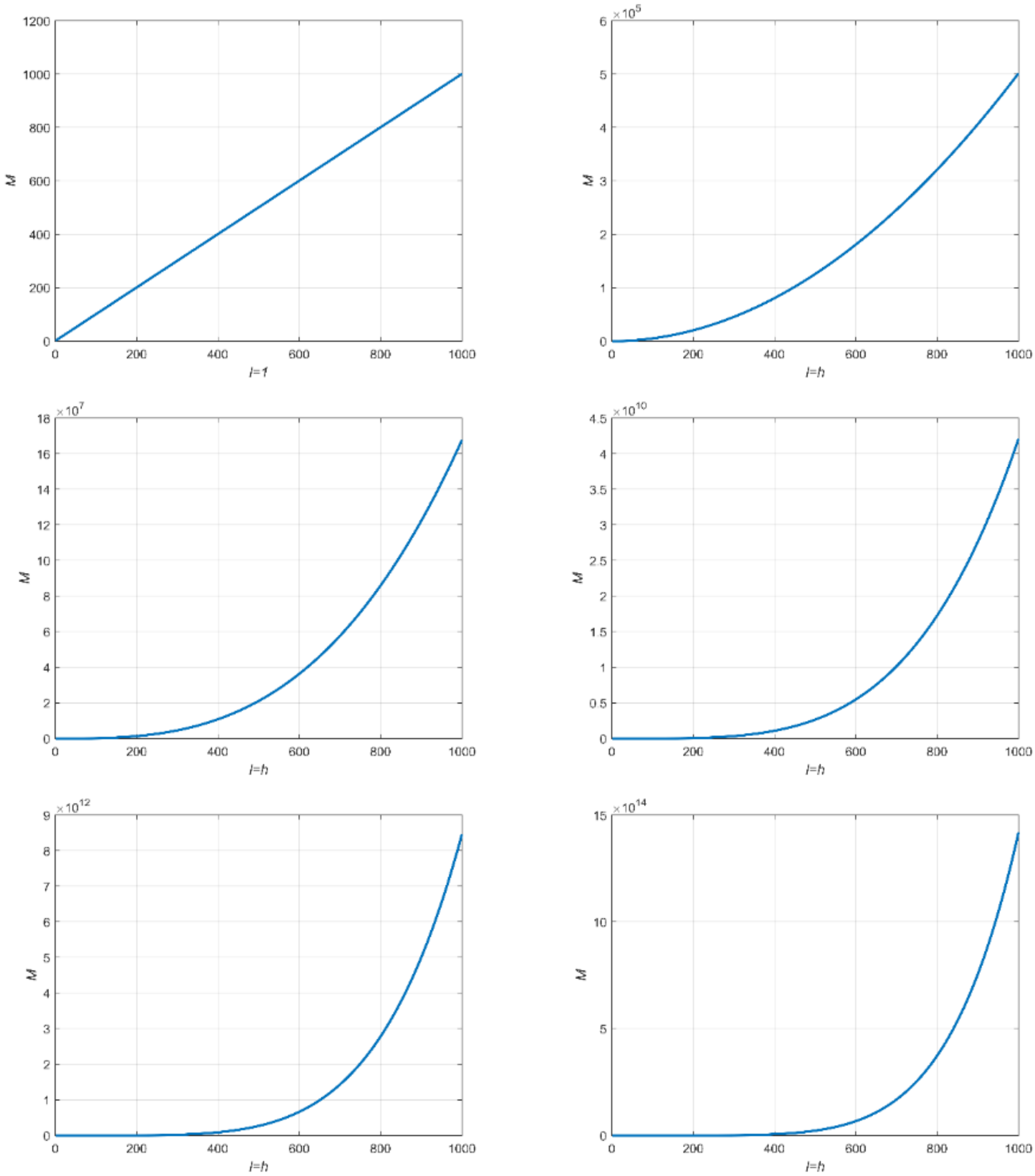


Figure 14 The relationship between M and the n with separate nonlinearity degree ℓ

As shown in Figure 14 , for the linear polynomial NARMAX model, M and n are directly proportional, as $M_{linear}=1000$, when $n=1000$. However, as the degree of nonlinearity ℓ increases, M experiences a dimension explosion. For instance, when $\ell=2$, $M_{\ell=2}=5\times10^5$, which is 500 times of the linear model terms. Similarly, $M_{\ell=3}=1.68\times10^7$, which is 1.68×10^4 times the linear model terms' dimension M_{linear} and 33 times that of when $\ell=2$. While when $\ell=6$, $M_{\ell=6}=1.42\times10^{15}$, which is 1.42×10^{12} the linear dimension M_{linear}, 2.83×10^{9} times $M_{\ell=2}$, 8.46×10^{6} times $M_{\ell=3}$, 33701 times $M_{\ell=4}$, and 167.67 times $M_{\ell=5}$.

The phenomenon of dimensional explosion gives rise to three significant issues. The first pertains to the inflation in computational complexity induced by an excessive number of model terms. According to equation (4.6) the ERR and orthogonal operations in FROLS correspond directly to the dimensionality of candidate model terms. Consequently, as the dimensionality of these candidate model terms escalates, therefore do the operations for polynomial NARMAX models of equivalent lengths, thereby spiralling the computational complexity. This not only extends the training duration but also necessitates massive hardware resources during the modelling process. For instance, ChatGPT-4, which houses approximately 1.7 trillion parameters, required an extensive input of resources by OpenAI – approximately three years of training using around 10,000 high-performance GPUs and a substantial financial investment.

The second issue deals with the redundancy and interference despite having ample candidate model terms that secure exhaustive detail about systems. For instance, if a polynomial NARMAX model with a nonlinearity degree $\ell = 6$ has 10 terms, then the signal-to-noise ratio (SNR) during modelling would be quite low. This scenario exacerbates the task for FROLS in selecting suitable model terms due to the interfering noise, thus undermining the reliability of the identified models.

The third issue is that a surplus of model terms inherently escalates model uncertainty, which becomes especially pronounced when dealing with real-world systems that are predominantly complex, dynamic, and nonlinear. Since a comprehensive model that fully explicates these systems is largely non-existent, it follows that all models serve as approximations. The principal aim of modelling is to depict the real system with minimal error while offering insights into the system. Nevertheless, the rise in candidate model terms introduces an increase in models fulfilling system identification goals, subsequently fostering heightened model uncertainty.

Therefore, the importance of developing new methods to address these issues cannot be overstated. The novel method should not only maintain the interpretability of NARMAX models, but also solve the high-dimensionality and vast model uncertainty. A typical method is the neural network, especially the deep neural networks (DNNs) or the deep learning (DL). Firstly, DNNs are adept at handling high dimensionality, thereby potentially mitigating the computational complexity associated with dimensional explosion. They can process vast amounts of data simultaneously, efficiently reducing training time and hardware resources requirements. Secondly, DNNs' ability to learn and extract relevant features from redundant data can help to reduce the interference from the plethora of model terms. They can deal with non-linear relations, adapting to external noise and disturbances. Finally, DNNs' capacity for accommodating complexities and uncertainties ingrained in real-world systems offers an agile solution to capturing system characteristics with minimal error. While all models are approximations, DNNs, due to their flexibility and learning capability, could potentially provide a closer approximation to the actual system. By so doing, they reduce model uncertainty and enhance the fidelity of system identification.

DNNs, while offering significant advantages for system identification, also present notable challenges. Overfitting, for instance, is a major drawback, where the network learns peculiar patterns within the training data that may not apply universally to new data sets. This scenario can cause erroneous system characterization and incorrect predictions when faced with new data. Further, these networks are often considered 'black boxes' due to their lack of interpretability. Even though they are proficient at mapping inputs to outputs, understanding the particular roles different layers play within the model is complex.

This lack of transparency can inhibit identification of model terms and hamper comprehensive system understanding, which is crucial for system identification. Also, DNNs necessitate large volumes of training data, which may not always be accessible in system identification contexts. Moreover, their significant demand for computational resources and time may restrict their use in real-time identification scenarios. Lastly, these networks might show sensitivity towards parameter initialization and hyperparameter settings. Improper configuration might lead to non-convergence or inferior solutions, thereby compromising performance. Therefore, in the face of these challenges, DNNs' potential for addressing dimensional explosion in system identification needs to be realized only after adequately addressing their limitations.

Thus, by bringing together DNNs and polynomial NARMAX, a novel deep polynomial NARMAX (DeepNARMAX) modelling seeks to leverage the advantages of both. The DL capabilities of DNNs can be used to alleviate the dimensional explosion issues faced by NARMAX, while the transparency and interpretability of NARMAX models can help solve the 'black box' mystery of DNNs, yielding a model that is powerful, efficient, and interpretable. This amalgamation seeks to push the boundaries of system identification and modelling, allowing for more robust, reliable, and meaningful data analysis and system insights.

4.2 The scheme of the DeepNARMAX model

The complex dynamical system identification and prediction problem is to identify the most appropriate and interpretable model(s) and generate the most accurate prediction(s) in the future given the observed data. As mentioned above, we have addressed the difficulties of the complex dynamical system identification: balancing the trade-off between model structure uncertainty and computation efficiency and generating transparent and interpretable models. To overcome the issues, we introduce the 'driller-generator' architecture in D-NARMAX-NN. Moreover, to maximize the utilization of information, we propose the 'gate-weight' concept and 'swarm-based' algorithm for training and optimizing the gate-weight.

4.2.1 The structure of the DeepNARMAX network

The general structure of the proposed network is shown in Figure 15. There are four different novel layers in the hidden layers, which are linear transformation layer, driller layer, generator layer, and model detection layer, along with the input layer and the output layer, combining the proposed DeepNARMAX network.

Initially, the system's input and output signals are defined as the network's input feature parameters. These parameters, upon entering the network, first undergo a linear transformation within the hidden layer. Given specific time delay hyperparameters, these input feature parameters are expanded into high-dimensional linear feature vectors. These vectors subsequently traverse through a driller layer, where a subset is chosen as inputs into the subsequent network layers. After this driller layer, the feature vector dimensionality is reduced before it serves as an input for a generator layer. In this layer, cross-multiplication operations are performed on these feature vectors, resulting in outputs encompassing both original and cross-multiplied, newly generated features.

It is noteworthy that the driller and generator layers are interdependent. The total count of drillers and generators is determined by ℓ, denoting the nonlinearity degree in polynomial NARMAX models, implying there will be l pairs or 2ℓ individual driller-generator layers. Upon navigating through all 2ℓ layers, both drillers and generators, the linear features are inputted into a model detection layer. Here, FROLS algorithm selects optimal features from these inputs, constructing the ideal polynomial NARMAX model, the output for this specific network layer.

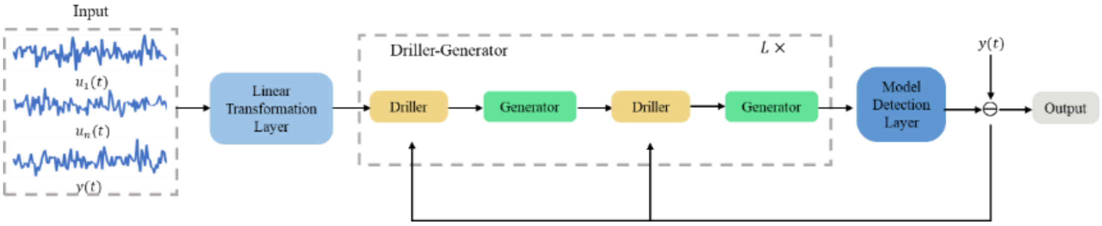


Figure 15 The scheme of DeepNARMAX.

The detailed architecture and neurons are in the left part. Each individual-coloured square stands for the candidate model term. The light blue and dark blue squares in the left part are the description of drilling layer. The light blue squares mean blocking, while the rest allows the model term pass. The brief structure of D-NARMAX-NN is shown in the right part.

Consider a multi input single output (MISO) system, where $\mathbf{u}=[\mathbf{u}_1,\mathbf{u}_2,...,\mathbf{u}_n]^T$ are the *n*-dimensional inputs signals, $\mathbf{u}_i=[u_i(1),u_i(2),...,u_i(T)]^T, i=1,2,...,n$ is the *i-th* input signal vector with *T* observations for each vector, $\mathbf{y}=[y(1),y(2),...,y(T)]^T$ is the system output signal/vector. The objective of deep polynomial NARAMX is to find one or a group of appropriate models to reveal the system's interaction and describe the system. In the following sections, the mathematical definition of the hidden layers will be defined.

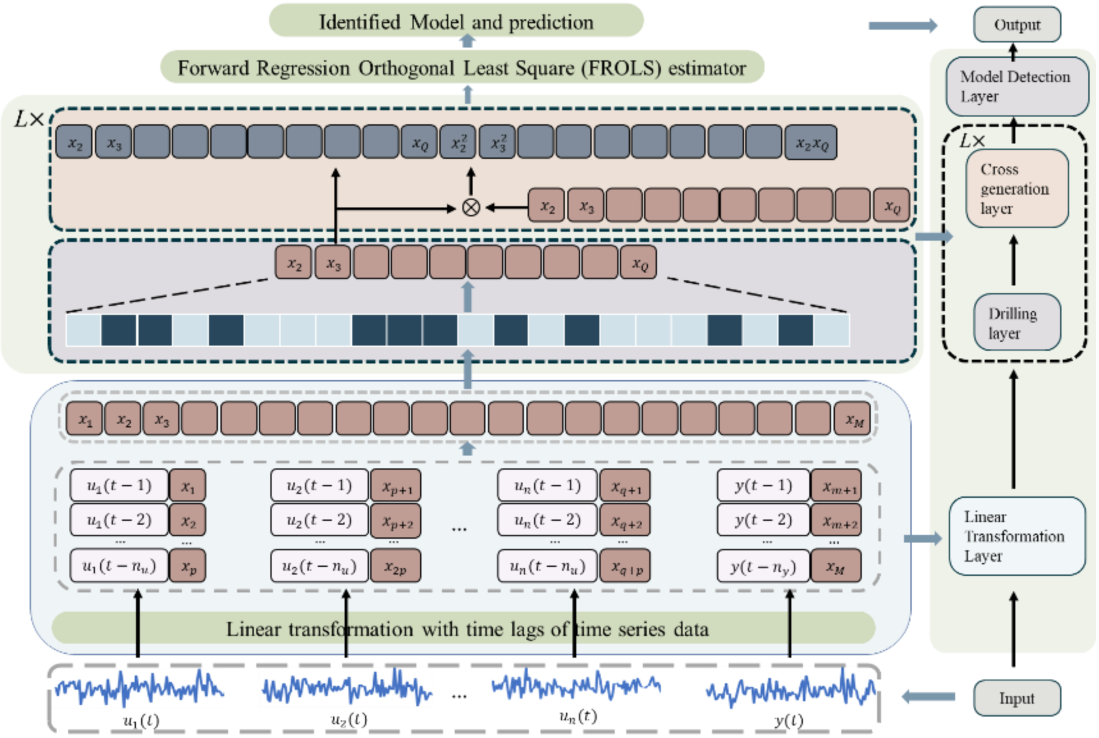


Figure 16 The detailed illustration of the proposed DeepNARMAX network framework

4.2.1.1 Linear Transformation Layer

To map the dynamical interaction between the system output and the observed information, we design the linear transformation layer, which can produce dynamical linear features based on the prior knowledge or end-users' preferences. Supposing the system is a multi-input and single-output (MISO) system, there are several inputs: $u=[u_1,u_2,...,u_n]^T$, and single output: y, where $u_i\in\mathbb{R}^{m\times1}, i=1,2,...,n$ and $y\in\mathbb{R}^{m\times1}$. In D-NARMAX-NN, the system inputs and output would be redefined as the input sample of the network. Concretely, we adjust the input sample with time lags and expand the original features into higher dimensional linear features:

$$
\begin{aligned}
x = [x_1, x_2, ..., x_p]^T &= [u_1(t), u_1(t-d_1), ..., u_1(t-n_u), ..., \\
& u_n(t), u_n(t-d_n), ..., u_n(t-n_u), \\
& y(t-d_y), ..., y(t-n_y)]^T
\end{aligned} \tag{4.7}
$$

where $x = [x_1, ..., x_p]^T$ is the transformed linear features, and $x_i \in \mathbb{R}^{q \times 1}, i = 1, 2, ..., p$; $d_1, ..., d_n, d_y$ are the starting time lags for the system inputs and output. Denote $d_{max} = \max(d_1, ..., d_n, d_y)$ is the maximum starting time lags, while the length of the transformed linear features $x_i, i = 1, 2, ..., p$ is $l = m - d_{max}$.

4.2.1.2 Driller-generator Layer

Driller To reduce the dimensionality of the transformed linear features, we take the idea of mask layer in the Transformer [193] and the dropout in the deep learning [233] and design the driller layer in our D-NARMAX-NN, which reduce the dimensionality of features and the computational complexity. As shown in Figure 16, there are L driller layers in the network. The inputs of each driller layer are the features of dimensional expansion, while the output of each driller layer is the set of features with dimensionality reduction.

Let $P = [p_1, ..., p_s]_{l \times s}$ is the input of the driller layer, where $p_i \in \mathbb{R}^{l \times 1}, i = 1, 2, ..., s$. Without loss of generality, p_i can be linear or nonlinear features from the system inputs. Define the driller layer as:

$$
Dr^v = [dr_1^v(\theta_1^v), dr_2^v(\theta_2^v), ..., dr_s^v(\theta_s^v)] \tag{4.8}
$$

$$
dr_i^v(\theta_i^v) = \begin{cases} [0, 0, ..., 0]^T, & \theta_i^v \le \theta_{th} \\ [1, 1, ..., 1]^T, & \theta_i^v > \theta_{th} \end{cases} \tag{4.9}
$$

where v is the *v-th* layer in the network; $dr_i^v, i = 1, 2, ..., s$ is the *i-th* driller in the *v-th* driller layer; θ_i^v is denoted as the gate weight of the i-th driller, which controls the values of dr_i^v (zero vector or all-ones vector); θ_{th} is the threshold value of the gate weight. If the θ_i^v is smaller than the threshold θ_{th} , then the driller is zero vector; otherwise, the driller is all-ones vector.

Denote $P_l = [p_{l_1}, p_{l_2}, ..., p_{l_r}]^T$ as the output of the driller layer, where $p_{l_j} \in \mathbb{R}^{l \times 1}, l_j = 1, 2, ..., l_r$ and l_r equals the number of all-ones vector in the driller layer. As shown in Figure 17, the output of the driller layer is defined as:

$$
P_l = P \odot Dr^v \tag{4.10}
$$

where $\odot$ denotes the Hadamard product. Clearly, there are less dimensionality of P_l compared with P, leading to reduction of the dimension of the feature vector and computation complexity. The values of drillers in the drilling layer are controlled by the proposed gate weight θ_i^v , which will be optimized based on the modelling performance of each iteration, as shown in Figure 17. The detailed optimization algorithm is present in the next section.

Generator The driller layer can significantly reduce the complexity of the features and select the relative important features. However, we need to increase the nonlinearity information of the features for modelling and prediction. Thus, we introduce the 'generator' layer directly behind the driller layer to generate features with higher nonlinear degree ℓ .

The input of the generator layer is the output of the driller layer, which is P_l as defined above. Define $P_G = [p_{G_1}, p_{G_2}, ..., p_{G_r}, p_{G_{r+1}}, ..., p_{G_R}]^T$ as the output of the generator layer, where $p_{G_i} \in \mathbb{R}^{l \times 1}, i = 1, 2, ..., R$ is the generated complex feature. In the generator layer, there are two parts of the complex features. For the first part, we keep all inputs of the generator layer as the output as shown in Figure 17 which is:

$$P_G = [\underbrace{p_{G_1}, p_{G_2}, \ldots, p_{G_r}}_{l_r}, \ldots]^T = [P_l, \ldots\ldots]^T \tag{4.11}$$

The second part of the output of generator layer is the nonlinear mapping of the input features:

$$P_G = [P_l, p_{G_{r+1}}, \ldots, p_{G_R}]^T = [P_l, F_1(p_{G_{r+1}}), \ldots, F_{R-r}(p_{G_R})]^T \tag{4.12}$$

where $F_i(\cdot), i = 1, \ldots R-r$ are some nonlinear mapping forms. There are plenty of nonlinear mapping functions as discussed above. Here we implement the polynomial form in the generator layer, meaning $F_i(\cdot)$ are cross product form or square product form such as $p_{l_1}^2$ or $p_{l_1} p_{l_3}$ shown in Figure 17. Thus, the output of the generator layer in the DeepNARMAX network is

$$P_G = [P_l, p_{G_{r+1}}, \ldots, p_{G_R}]^T = [P_l, p_{l_1}^2, \ldots, p_{l_r}^2, p_{l_1} p_{l_2}, \ldots, p_{l_1} p_{l_r}, \ldots, p_{l_{r-1}} p_{l_r}]^T \tag{4.13}$$

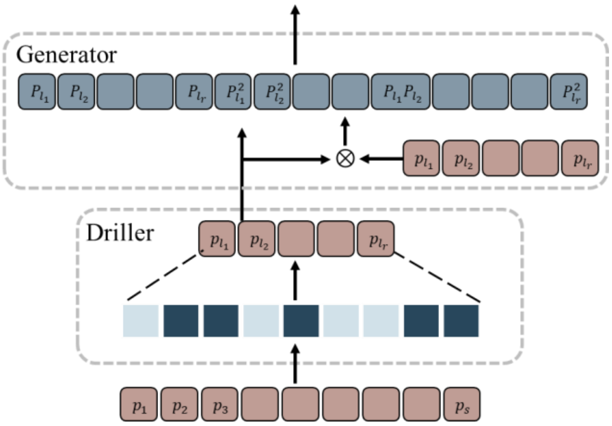


Figure 17 Driller-Generator Layer.

The bottom row represents the inputs of the driller-generator layer. The lower row in the middle box denotes the driller, where the light blue cubes stand for the blocking operation and the navy cubes suggest the passing operation. With the passing operation, less features (third last row) are selected from the input of the layer. The top box indicates the generator, which combine the selected features from driller and the generated features by cross multiplication.

4.2.1.3 Model detection layer

In the model detection layer, the FROLS algorithm is utilized to select the most significant feature vectors of the output of the generator layer and construct the most appropriate NARMAX model with the selected features. The output of the generator layer is defined by equation (4.13). Considering the FROLS algorithm, p_G is the candidate set containing all candidate features to describe the system. Then as discussed in section 3.2, the FROLS algorithm will select M_0 features among the set p_G and calculate the related parameters.

4.2.1.4 Hyperparameter optimization

In the comprehensive domain of deep learning, the hyperparameter optimization is essential as it determines the learning structure and speed of the model. Some typical hyperparameter, like the number of layers and neurons, the activation function, the batch size, epochs, dropout rate, etc. The optimal number of layers and neurons is a critical yet complex task. The selection significantly influences the model's capacity to effectively learn from the data and generate precise predictions. To deal with the problem, there are several methods to exploring the number of layers and neurons in deep learning, like the grid search [234], random search [235], evolutionary algorithm [236], and Bayesian optimization [237].

In the proposed DeepNARMAX architecture, hyperparameter optimization emerges as a critical factor for comprehensive system identification and probabilistic prediction. The model's activation function is defined as a set of polynomial functions, rendering the determination of the number of layers and neurons per layer crucial for overall model performance. Conforming to the fundamental architecture of input-hidden-output, it is conveniently straightforward to determine the number of neurons in both the input and output layers. These are inherently dictated by the system. However, ascertaining the optimal number of hidden layers and neurons within each poses a more complex problem. To address this, we implement a random search methodology for its effective optimization within the scope of this thesis.

Suppose the unrevealed system is a MISO system, with m inputs.Also assume that the actual model of the system is unknown. To obtain satisfactory model approximations of the system, we normally an appropriate nonlinear degree d_{max} and the maximum time lags n_{max} for the input and output signal. According to Eq (3.1), the time lags for input signal and output signal are usually different in the NARMAX model. However, in the proposed DeepNARMAX, we define the initial maximum time lags for input, output and error signals as n_{max}.

For a DeepNARMAX model structure, with the highest nonlinear degree d_{max}, the maximum number of hidden layers L_{max} can be defined as:

$$L_{max} = 2 + 2 \times d_{max} \tag{4.14}$$

Note that there are always a linear transformation layer and a model detection layer. The number of driller-generator layers is determined by the highest nonlinear degree . Similarly, we can derive the maximum number of neurons in the hidden layers based on the initial parameters n_{max} and m, with the driller rate p. Once the maximum number of hidden layers and neurons determined from the initial parameters, the random search can then be applied to optimize the hyperparameters.

4.2.2 Swarm-based Gate weight optimizing mechanism

To obtain the most appropriate model through the proposed D-NARMAX-NN, it is desirable to let the related features more obvious and the irrelevant features less remarkable. Therefore, in the proposed D-NARMAX-NN, the optimal model will be trained through iterations. Consider the model from D-NARMAX-NN in κ-th iteration as:

$$\hat{y}^{\kappa}(\theta, x) = \sum_{i=1}^{I} \delta(\theta_i^{\kappa} \mid x_i) \cdot c_i x_i + e \tag{4.15}$$

$$\delta(\theta_i^{\kappa} \mid x_i) = \begin{cases} 1, & \theta_i^{\kappa} > \theta_{th} \\ 0, & \theta_i^{\kappa} \le \theta_{th} \end{cases} \tag{4.16}$$

where x_i is the selected features in the identified model; $\theta_i^{\kappa} \mid x_i$ is the gate weight of x_i in the κ th iteration; c_i is the parameters related to the selected features; *I* is the total number of terms in the model; $e = |\hat{y} - y|$ is the error between modelling results and observed output. Denote the loss function of identified model in κ-th iteration $\overset{\kappa}{y}(\theta^{\kappa}, x)$ as:

$$L^{\kappa}(\theta \mid x, x) = \| \hat{y}^{\kappa}(\theta \mid x, x) - y \|_2^2 = \| \sum \delta(\theta_i^{\kappa} \mid x_i) \cdot c_i x_i - y \|_2^2 \tag{4.17}$$

The purpose of the training process is to minimize the loss function:

$$\min_{x \in X_{DL}} L^{\kappa}(\theta \mid x, x) \tag{4.18}$$

Which is equal to find the most related model terms which controlled by the gate weight:

$$\underset{\theta_i \in [0,1]}{\arg\max}\ L^{\kappa}(\theta_i \mid x_i, x_i) \qquad (4.19)$$

In this thesis, we propose swarm-based gate weight optimizing algorithm to update the gate weight, optimize the model structure and identify the most fitted model for the system. Suppose $L^{(\kappa+1)}$ as the loss function of the identified model $\hat{y}^{\kappa+1}(\theta^{\kappa+1}, x) = \sum \delta(\theta_i^{\kappa+1} \mid x_i) c_i x_i^{\kappa+1}$ in $\kappa+1$*-th* iteration, while L^{κ} in the κ*-th* iteration.

4.2.2.1 Convergence after iteration: $L^{\kappa+1} < L^{\kappa}$

In this situation, the $\kappa+1$-th identified model $\hat{y}^{\kappa+1}(\theta^{\kappa+1}, x)$ performs better than the κ-th model $\hat{y}^{\kappa}(\theta^{\kappa}, x)$, meaning the identified model terms in model $\hat{y}^{\kappa+1}(\theta^{\kappa+1}, x)$ are more effective than those in model $\hat{y}^{\kappa}(\theta^{\kappa}, x)$. The model terms in $\hat{y}^{\kappa+1}(\theta^{\kappa+1}, x)$ will be kept in the dictionary, while the related gate weights can be updated as follows:

$$\theta_i^{\kappa+1} \mid x_i \leftarrow \alpha \cdot \theta_i^{\kappa+1} \mid x_i, \alpha \in [1, 2] \qquad (4.20)$$

4.2.2.2 Divergence after iteration: $L^{\kappa+1} > L^{\kappa}$

In this situation, the identified model $\hat{y}^{\kappa+1}(\theta^{\kappa+1}, x)$ performs not good as $\hat{y}^{\kappa}(\theta^{\kappa}, x)$. However, some model terms in model $\hat{y}^{\kappa+1}(\theta^{\kappa+1}, x)$ might have critical influence in the prediction and system modelling. Thus, the related gate weights are updated as follows:

$$\theta_i^{\kappa+1} \mid x_i \leftarrow \beta \cdot \theta_i^{\kappa+1} \mid x_i, \beta \in (0,1) \qquad (4.21)$$

4.2.2.3 Equality after iteration: $L^{\kappa+1} = L^{\kappa}$

Though model $\hat{y}^{\kappa+1}(\theta^{\kappa+1}, x)$ has the same performance compared with model $\hat{y}^{\kappa}(\theta^{\kappa}, x)$, model structures of two models can be either totally same or partly same. Despite the difference of two models, all model terms will be recorded in the candidate dictionary. The updating strategy of the related gate weights differ due to the similarity of model structure.

The model structures of two identified models in two iterations are totally same.

If $\hat{y}^{\kappa+1}(\theta^{\kappa+1}, x)$ and $\hat{y}^{\kappa}(\theta^{\kappa}, x)$ share the same model structure, then the model terms identified in $\hat{y}^{\kappa+1}(\theta^{\kappa+1}, x)$ or $\hat{y}^{\kappa}(\theta^{\kappa}, x)$ are more important compared with other options. So, the importance of model terms can be reflected by the related gate weights. In this thesis, we propose a swarm-based gate weight algorithm to optimize the values of the gate weights.

The swarm-based gate weight updating algorithm The gate weights are the key parameters to control the selection of the features shown in equation (4.14). With larger gate weights, the related features will have more opportunity to pass the driller layer. To fully investigate the effectiveness of the features, the swarm-based gate weight updating algorithm is proposed to optimize the values.

Define $P_{\theta_i \mid x_i}$ is as the largest gate weight of model term x_i in the training period, $\varepsilon^{\kappa+1}$ is the prediction error in $\kappa+1$*-th* iteration. For the gate weight of each model terms shown in the identified model, the updating strategy is as follows:

$$\theta_i^{\kappa+1} \mid x_i \leftarrow \eta \theta_i^{\kappa+1} \mid x_i + v^{\kappa+1} \qquad (4.22)$$

where $\eta \in (0,1)$; v is the updating velocity of the gate weight, which is defined as:

$$v^{\kappa+1} \leftarrow \gamma \cdot v^{\kappa} + 0.5 \cdot \tau_1 \cdot (P_{\theta \mid x_i} - \theta_i^{\kappa+1} \mid x_i) + 0.5 \cdot \tau_2 \cdot (\varepsilon^{\kappa+1} - \theta_i^{\kappa+1} \mid x_i) \qquad (4.23)$$

where $\gamma \in (0,1]$, $\tau_1, \tau_2 \in (0.1]$.

The model structures of two identified models in two iterations are partly same.

Model terms in $\hat{y}^{\kappa+1}(\theta^{\kappa+1},x)$ can be divided into two group: model terms within $\hat{y}^{\kappa}(\theta^{\kappa},x)$ and model terms outside of $\hat{y}^{\kappa}(\theta^{\kappa},x)$. For model terms outside of $\hat{y}^{\kappa}(\theta^{\kappa},x)$, the related gate weights can be updated by equation (4.22), while for the other group the related gate weights can be updated by the proposed swarm-based gate weight updating algorithm.

4.3 Case study

In this section, experimental results are presented to demonstrate the performance of the proposed DeepNARMAX network. Firstly, we have designed three simulation models, including various types such as SISO and MISO systems, to verify the feasibility and stability of the proposed network. This is done to demonstrate that the network has the capability to identify real systems and provide accurate interpretive models. Secondly, this thesis selects two actual systems – ETT and weather - to validate the performance of our network in handling real unknown dynamic complex systems. Simultaneously, it compares with state-of-*th*e-art methods like traditional NARAMX, LSTM, Informer, Reformer etc., illustrating its effectiveness and superiority.

4.3.1 Simulating case study

In the following simulating cases, 6 simulated models are used to measure the performance of the proposed DeepNARMAX network. The systems vary from SISO system to MISO system with different inputs, and noises. The data points for each system are all set to be 500, while 80% of the data points is set to be the training set, while the rest 20% is set to be the testing set. To avoid overfitting, in the training set, the k-folder cross validation is utilized. Moreover, the classic polynomial NARMAX model is used to compare the performance of the proposed DeepNARMAX network. For DeepNARMAX, each experiment will run 1000 iterations to gain the best model.

4.3.1.1 Simulating system 1

Consider a single input single output model, with a while noise single as

$$z(t) = -0.6z(t-1) - 0.2u(t-2)\times z(t-3) + 0.5u(t-2) - 0.25u(t-2)u(t-3) + 0.01e(t) \tag{4.24}$$

$$y = z + 0.1\zeta \tag{4.25}$$

where $u(t) \sim U[-1,1]$, $e(t) \sim N(0,0.1)$, and $\zeta \sim N(0,1)$. Define the $\ell = 3$ and the starting of the time delay $d_y = d_u = 1$ and $d_e = 0$, the maximum time delay $n_y = n_u = n_e = 5$ for DeepNARMAX network and classic NARMAX model.

Two transparent identified models by DeepNARMAX and classic NARMAX are listed in Table 3, where the model terms are the exact same with those in the system (4.24), while the parameters of two models are slightly different with those in the system (4.24). Meanwhile, two models by DeepNARMAX and classic NARMAX are identical including the model terms and the related parameters. The comparison between the observation and one-step-ahead predictions of two models is shown in Figure 18. Due to two models are exactly identical, thus, two curves representing the predictions by two models completely overlap. The statistical results mean square error (MSE) and mean absolute error (MAE) are listed in Table 4. Similarly, the MSEs and MAEs of two models over the testing set are same.

Table 3 Comparison of the identified models by DeepNARMAX network and classic NARMAX for case 1

Index	DeepNARMAX	Classic polynomial NARMAX	System

	Model Term	Model Parameter	Model Term	Model Parameter	Model Term	Model Parameter
1	$u(t-2)$	0.528	$u(t-2)$	0.528	$u(t-2)$	0.5
2	$y(t-1)$	-0.534	$y(t-1)$	-0.534	$y(t-1)$	-0.6
3	$u(t-2)u(t-3)$	-0.221	$u(t-2)u(t-3)$	-0.221	$u(t-2)u(t-3)$	-0.25
4	$u(t-2)y(t-3)$	-0.205	$u(t-2)y(t-3)$	-0.205	$u(t-2)y(t-3)$	-0.2

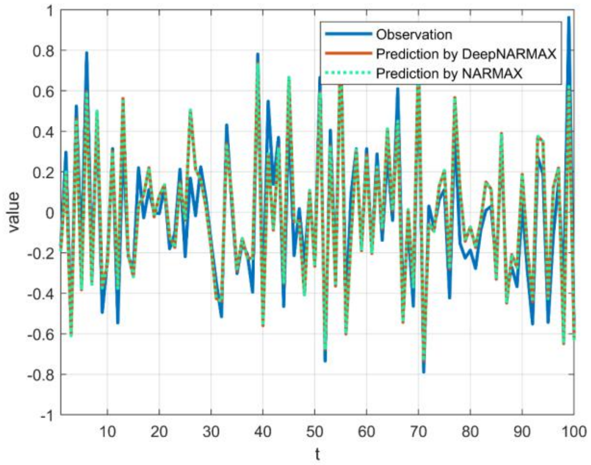


Figure 18 Comparison in the testing set of two models by DeepNARMAX and classic NARMAX and the observation for case 1

Table 4 Statistical performance over the testing period of two models for case 1

	MSE	MAE
DeepNARMAX	0.014818	0.096537
Classic NARMAX	0.014818	0.096537

The experiment 1 highlighted that for SISO systems, DeepNARMAX equates NARMAX in accurately pinpointing factors that influence the system and constructing precise models for their description. While the presence of noise induces some degree of discrepancy between the constructed model parameters and the actual parameters, leading to certain deviation between both models and the test system, results from Table 4 illustrate that these discrepancies do not compromise the models' accuracy and stability.

4.3.1.2 Simulating system 2

Consider a MISO model, with a while noise single as

$$\begin{aligned} y(t) = &-0.6y(t-1)+0.5u_1(t-1)u_4(t-2)-0.2u_2(t-3)y(t-2) \\ &-0.25u_3(t-1)u_4(t-2)+0.4u_1(t-1)y(t-2)-0.15u_4(t-2)u_3(t-4)+0.01e(t) \end{aligned} \tag{4.26}$$

where $u_1(t) \sim U[-1,1]$, $u_2(t) \sim N(0,1)$, $u_3(t) \sim U[0.1]$, and $u_4(t) \sim N(0.5,1)$, $e(t) \sim N(0,1)$. Define $\ell = 4$ and the starting of the time delay $d_y = d_u = 1$ and $d_e = 0$, the maximum time delay $n_y = n_u = n_e = 5$ for DeepNARMAX network and classic NARMAX model.

Two transparent identified models by DeepNARMAX and classic NARMAX are listed in Table 5. From the Table 5, the model terms selected by DeepNARMAX and classic NARMAX modelling are the same as the simulated system (4.25), while the parameters of two models are quite similar to those of system (4.25). Since the two models are completely consistent with the real system, therefore, the predicted values of the model on the test set are almost identical to the observed values, as shown in the

Figure 19.

Table 5 Comparison of the identified models by DeepNARMAX network and classic NARMAX for case 2

	DeepNARMAX		Classic polynomial NARMAX		System	
	Model Term	Model Parameter	Model Term	Model Parameter	Model Term	Model Parameter
1	$u_1(t-1)u_4(t-2)$	0.500	$u_1(t-1)u_4(t-2)$	0.500	$u_1(t-1)u_4(t-2)$	0.5
2	$y(t-1)$	-0.599	$y(t-1)$	-0.599	$y(t-1)$	-0.6
3	$u_3(t-1)u_4(t-2)$	-0.250	$u_3(t-1)u_4(t-2)$	-0.250	$u_3(t-1)u_4(t-2)$	-0.25
4	$u_1(t-1)y(t-2)$	0.400	$u_1(t-1)y(t-2)$	0.400	$u_1(t-1)y(t-2)$	0.4
5	$u_2(t-3)y(t-2)$	-0.201	$u_2(t-3)y(t-2)$	-0.201	$u_2(t-3)y(t-2)$	-0.2
6	$u_4(t-2)u_3(t-4)$	-0.150	$u_4(t-2)u_3(t-4)$	-0.150	$u_4(t-2)u_3(t-4)$	-0.15

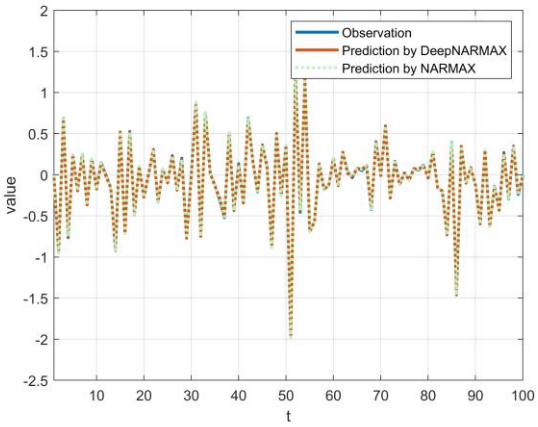


Figure 19 Comparison in the testing set of two models by DeepNARMAX and classic NARMAX and the observation for case 2

Experiment 2 underscored that, for MISO systems, DeepNARMAX aligns with NARMAX in terms of accurately identifying influential factors in the system and creating precise models for those phenomena. Despite the noise-induced discrepancies between the actual parameters and those derived in the models—which lead to certain deviations between both models and the test system—the results presented in Table 6 suggest that these divergences do not undermine the accuracy and stability of the models.

Table 6 Statistical performance over the testing period of two models for case 2

	MSE	MAE
DeepNARMAX	7.5×10^{-5}	0.0069687
Classic NARMAX	7.5×10^{-5}	0.0069687

4.3.1.3 Simulating system 3

Consider a MISO model, with a while noise single as

$$\begin{aligned} y(t) = {} & -0.6y(t-1)+0.5u_1(t-20)u_4(t-2)^2-0.12u_2(t-3)^2y(t-2)^2-0.25u_3(t-1)u_4(t-2) \\ & +0.4u_1(t-1)y(t-2)-0.5u_4(t-2)u_3(t-30)+0.01e(t) \end{aligned} \tag{4.27}$$

where $u_1(t)\sim U[-1,1]$, $u_2(t)\sim N(0,1)$, $u_3(t)\sim U[0.1]$, and $u_4(t)\sim N(0.5,1)$, $e(t)\sim N(0,1)$. Define $\ell=4$ and the starting of the time delay $d_y=d_u=1$ and $d_e=0$, the maximum time delay $n_y=n_u=n_e=5$ for DeepNARMAX network and classic NARMAX model.

Table 7 Comparison of the identified models by DeepNARMAX network and classic NARMAX for case 3

	DeepNARMAX		Classic polynomial NARMAX		System	
	Model Term	Model Parameter	Model Term	Model Parameter	Model Term	Model Paramete r
1	$y(t-1)$	-0.599	X	X	$y(t-1)$	0.5
2	$u_1(t-20)u_4(t-2)^2$	0.498	X	X	$u_1(t-20)u_4(t-2)^2$	-0.6
3	$u_2(t-3)^2 y(t-2)^2$	-0.119	X	X	$u_2(t-3)^2 y(t-2)^2$	-0.25
4	$u_3(t-1)u_4(t-2)$	-0.25	X	X	$u_3(t-1)u_4(t-2)$	0.4
5	$u_1(t-1)y(t-2)$	0.402	X	X	$u_1(t-1)y(t-2)$	-0.2
6	$u_4(t-2)u_3(t-30)$	-0.503	X	X	$u_4(t-2)u_3(t-30)$	-0.15

4.3.2 ETT system

4.3.2.1 Introduction

The Electricity Transformer Temperature (ETT) serves as a pivotal indicator in long-term deployment of electric power. Power distribution, defined as the allocation of electricity to diverse areas based on sequential usage, incorporates a key challenge, namely, predicting subsequent area-specific demands. These demands fluctuate depending on varying factors such as weekdays, holidays, seasons, weather conditions, and temperatures, among others. Nevertheless, to date, no existing method enables long-term forecasting relying on extensive real-world data with elevated precision. Any inaccurate prediction risks damaging the electrical transformer. Consequently, given the absence of a proficient forecasting method, managers currently base decisions on empirical values that substantially exceed actual demands, leading to unnecessary power wastage and equipment depreciation. On the contrary, oil temperatures serve as a reflection of the electricity transformer’s condition. An efficacious strategy involves predicting the safety of the oil temperature in electrical transformers to prevent wastage.

To investigate the granularity in the Long Short-Term Forecasting (LSTF) problem, discrete datasets, designated as {ETTh1, ETTh2}, are utilized at a one-hour level. Each data point comprises seven features: the predictive value, ‘oil temperature’ and six distinct types of power load features. The data is partitioned as follows: 12 months for training, 4 months for validation, and 4 months for testing.

4.3.2.2 Experiments settings

A grid search is conducted to optimize the hyperparameters of the DeepNARMAX. For ℓ, it is chosen from $\{3,4,5,6\}$, meaning the number of driller-generator layers is chosen from $\{6,8,10,12\}$, while the starting time delay of the output d_y is chosen from $\{24,48,168,336,720\}$, while the starting time delay of the input d_u is set from 1. Also, the maximum time delay for the output n_y and the input n_u is set to be 1000. For the other hyperparameter settings, they are tunned following the [238]. Each dataset is standardized to have a zero mean. In the context of the Long Short-Term Forecasting (LSTF) settings, the size of the prediction window, denoted as L_y, is incrementally expanded to include {24, 48,168, 336, 720} for the datasets {ETTh1, ETTh2}. The method of Prophet employs a series-to-point prediction setting, whilst the RNN-based approaches conduct dynamic decoding by left shifting on the prediction windows.

It is noticed that for the DeepNARMAX models, the starting time delay equals to the prediction windows in the LSTF or RNN-based algorithm. Suppose the starting time delay d_y of DeepNARMAX model is 24, thus the identified model should be represented as

$$\begin{aligned}\hat{y}(t) = F(&y(t-24), y(t-25), ..., y(t-n_y),\\ &u(t-1), ..., u(t-n_u), e(t-1), ..., e(t-n_e)) + e(t)\end{aligned} \tag{4.28}$$

Thus, we have

$$\begin{cases} \hat{y}(t+1) = F(y(t-23), ..., y(t-n_y+1), u(t), ..., u(t-n_u+1), ...) + e(t) \\ \hat{y}(t+2) = F(y(t-22), ..., y(t-n_y+2), u(t+1), ..., u(t-n_u+2), ...) + e(t+2) \\ \qquad ... \\ \hat{y}(t+24) = F(y(t), ..., y(t-n_y+24), u(t+23), ..., u(t-n_u+24), ...) + e(t+24) \end{cases} \tag{4.29}$$

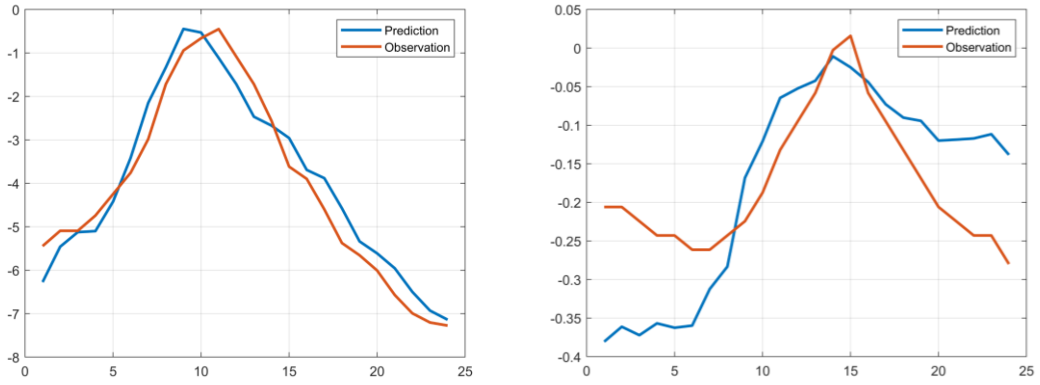
For each model, the choice of the model length is defined by the MSE in the validating set. To verify the performance of the methods, MSE and MAE of the prediction window are calculated.

4.3.2.3 Results of predictions

The statistical results are listed in Table 5. From the table, the proposed model, DeepNARMAX, showcases a considerably enhanced inference performance across all datasets. Moreover, the error rate associated with its predictions increases at a gradual and moderate rate as the predictive horizon expands. This underscores the success of the DeepNARMAX model in effectively broadening the prediction capacity. Besides, the DeepNARMAX model shows improvement of the forecast compared to the Informer, a transformer based deep neural network with self-attention mechanism, by a MSE reduction of 55.2% (at 24), 15.2% (at 48), 9.7% (at 168), 2.1% (at 336), and 0.6% (at 720). From the reduction of MSE, the proposed DeepNARMAX has great prediction ability in the short-time prediction, like $L_y = 24$, while with the increasing of the prediction window, the performance of the DeepNARMAX model gradually aligns with that of Informer. That is, in terms of long-term prediction performance, DeepNARMAX does not show a significant improvement compared to self-attention based deep neural networks like Informer, but its prediction accuracy remains excellent.

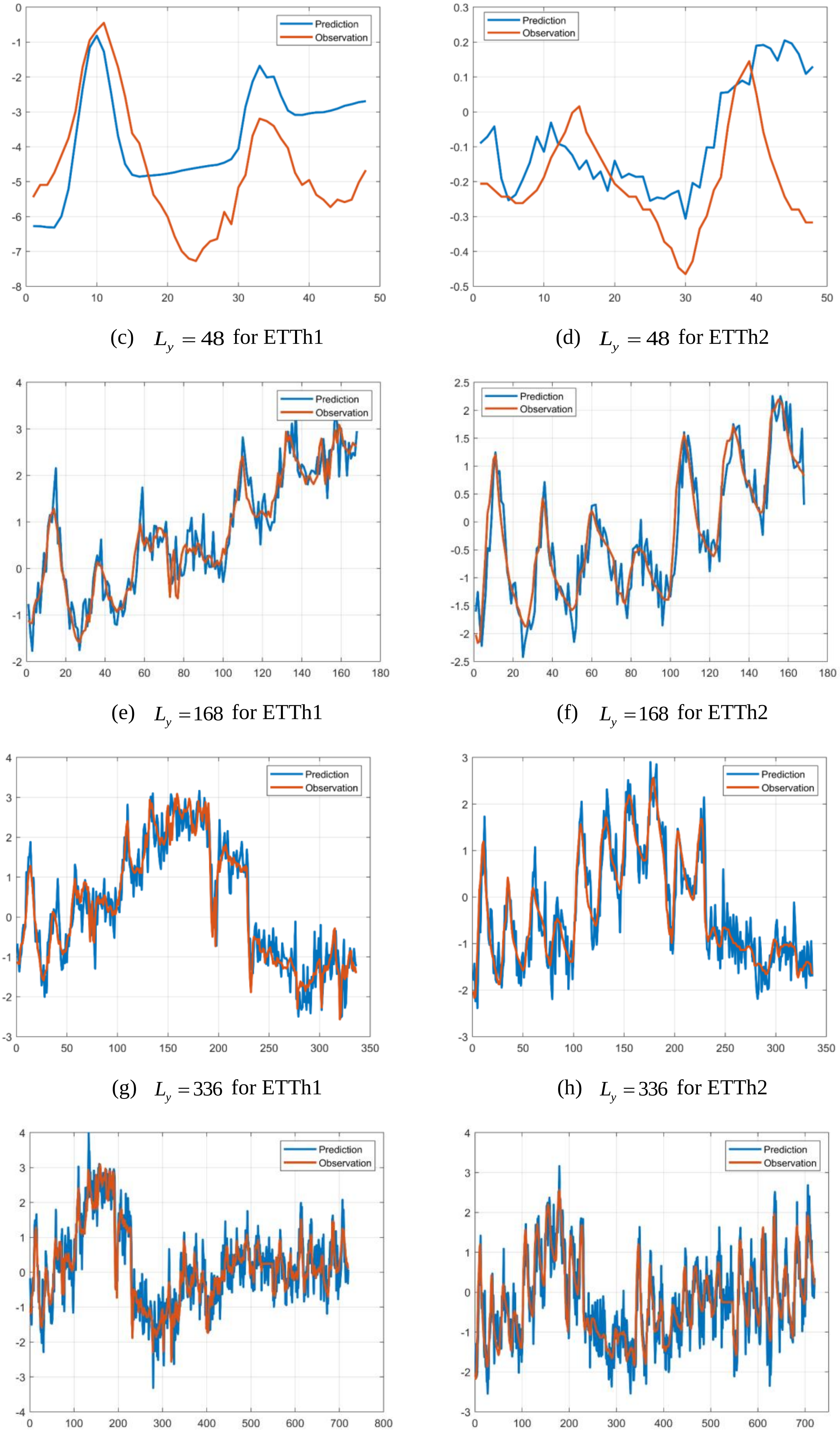
Also, the DeepNARMAX model demonstrates remarkably improved results compared to the LSTM variant of recurrent neural networks (RNNs). With a MSE reduction of 72.4 % (at 24), 47.5% (at 48), 46.2% (at 168), 62.1% (at 336), and 62.5% (at 720), our method suggests that the DeepNARMAX modelling has relative better forecast ability than the RNN-based models. Whether it’s short-term forecasting or long-term forecasting, DeepNARMAX can achieve more accurate predictions.

Meanwhile, on average, the proposed DeepNARMAX method exhibits better performance compared to DeepAR, ARIMA, and Prophet in terms of MSE, achieving reductions 74.6% (at 24), 49.1% (at 48), 65.6% (at 168), 71.9% (at 336), and 73.6% (at 720).



(a) $L_y = 24$ for ETTh1

(b) $L_y = 24$ for ETTh2



(c) $L_y = 48$ for ETTh1

(d) $L_y = 48$ for ETTh2

(e) $L_y = 168$ for ETTh1

(f) $L_y = 168$ for ETTh2

(g) $L_y = 336$ for ETTh1

(h) $L_y = 336$ for ETTh2

(i) $L_y = 720$ for ETTh1 (j) $L_y = 720$ for ETTh2

Figure 20 Comparison between observation and prediction by DeepNARMAX model in the prediction windows for ETT

The comparison between observation and prediction over each prediction window of proposed DeepNARMAX model over two datasets is shown in Figure 20, where the blue lines represent the prediction, and the red lines indicate the observation. The comparison results not only indicate that the DeepNARMAX model can accurately obtain system predictions, but also track changes in the system's trend. That is to say, on a local level, the DeepNARMAX model can keep up with rapid shifts in the system, thereby validating its reliability.

The performance comparison between the DeepNARMAX model and other deep network models is intuitively illustrated in Figure 21 using the ETTh1 dataset. We selected a prediction window of 168 for the comparison. Within this window, the solid blue line depicts observed values, and the solid red line illustrates the values predicted by DeepNARMAX. The series of dotted lines correspond to predictions from other models including Informer, LogTrans, LSTMa, DeepAR, ARIMA, Prophet, and Reformer within the same predictive window. From the figure, it's evident that the predictions from DeepNARMAX (red solid line) exhibit a strong alignment with observed trends (blue solid line) while maintaining relatively small errors. A comparably high level of accuracy is also visible in the predictions from Informer, denoted by light red dotted lines, which track observed trends closely. However, significant discrepancies are noted between the predictions from Prophet and Reformer, and actual observations. As the forecast steps increase, the difference between predicted and actual values widens, indicating a weaker long-term forecasting ability for these models in contrast to their short-term predictions. Contrarily, the DeepNARMAX model, along with models like Informer, showcase remarkable proficiency in both short and long-term forecasting.

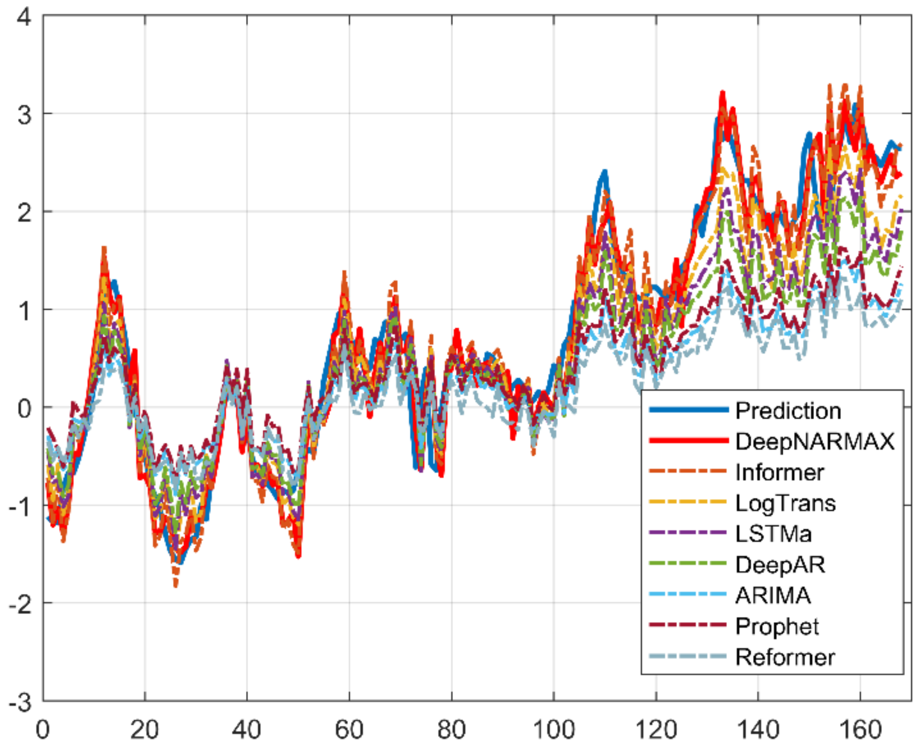


Figure 21 The comparison between observation and predictions of all models over the prediction window with 168 for ETT

4.3.2.3 Results of DeepNARMAX models

Compared to other deep network models, the unique advantage of DeepNARMAX is its interpretability.

That is, it relies on the efficient interpretability of NARMAX modelling, and the structure and parameters of the model are very clear. This provides extremely effective information for system analysis and research. Table 9 and Table 9 list the top 5 effective model items and their parameters in the DeepNARMAX model identified based on different prediction windows for ETTh1 and ETTh2 datasets.

Table 9 highlights that DeepNARMAX, while inheriting the interpretability of the NARMAX model, provides a lucid model structure with corresponding parameters for every model term. These terms unambiguously present the relationship and consequential influence between the system's input and output signals. For instance, in a model based on the ETTh1 dataset, a 24-point prediction window can be reformatted into a 24-step-ahead predictive problem within DeepNARMAX, implying the current system response could be impacted by the response from 24 steps earlier. Table 9 enumerates five crucial terms of our model. A conclusion drawn from this underline the considerable influence of $y(t-28)$ on the predictions made 24 steps ahead, along with meaningful impacts of all six input signals on the ETT's output – exceptionally notable being the second feature – 'high useless load'.

DeepNARMAX's clear structure and robust feature analysis make it adept not only at precise short and long-term sequence forecasting, but also at revealing esoteric information about systems' interactions, especially the effects of input/output signals on them. Even if discrepancies exist between forecast accuracy and actual application, interpretable models offer remarkably insightful information.

4.3.3 weather system

4.3.3.1 Introduction

Local Climatological Data (LCD) summaries furnish an overview of climatic variables for a single weather station over a distinct month in the US. These summaries, derived from surface observations conducted manually and via automated systems such as AWOS and ASOS, utilize source data from the Integrated Surface Data (ISD) dataset curated by the National Centers for Environmental Information. These summaries encompass numerous locations globally, offering measurements of multiple climatic variables. These include hourly, daily, and monthly recordings of temperature, dew point, humidity, winds, sky condition, weather type, atmospheric pressure, and more .

4.3.3.2 Experiment settings

The dataset encompasses localized climatological data from approximately 1,600 sites within the U.S., covering a span of four years from 2010 to 2013. Data points are logged every hour, each comprising 12 features including the predictive feature 'wet bulb' and an additional 11 climate/weather-related features. The breakdown of the data into training, validation, and testing sets is conducted over periods of 28, 10, and 10 months respectively. Similarly, to verify the performance of the methods, MSE and MAE of the prediction window are calculated.

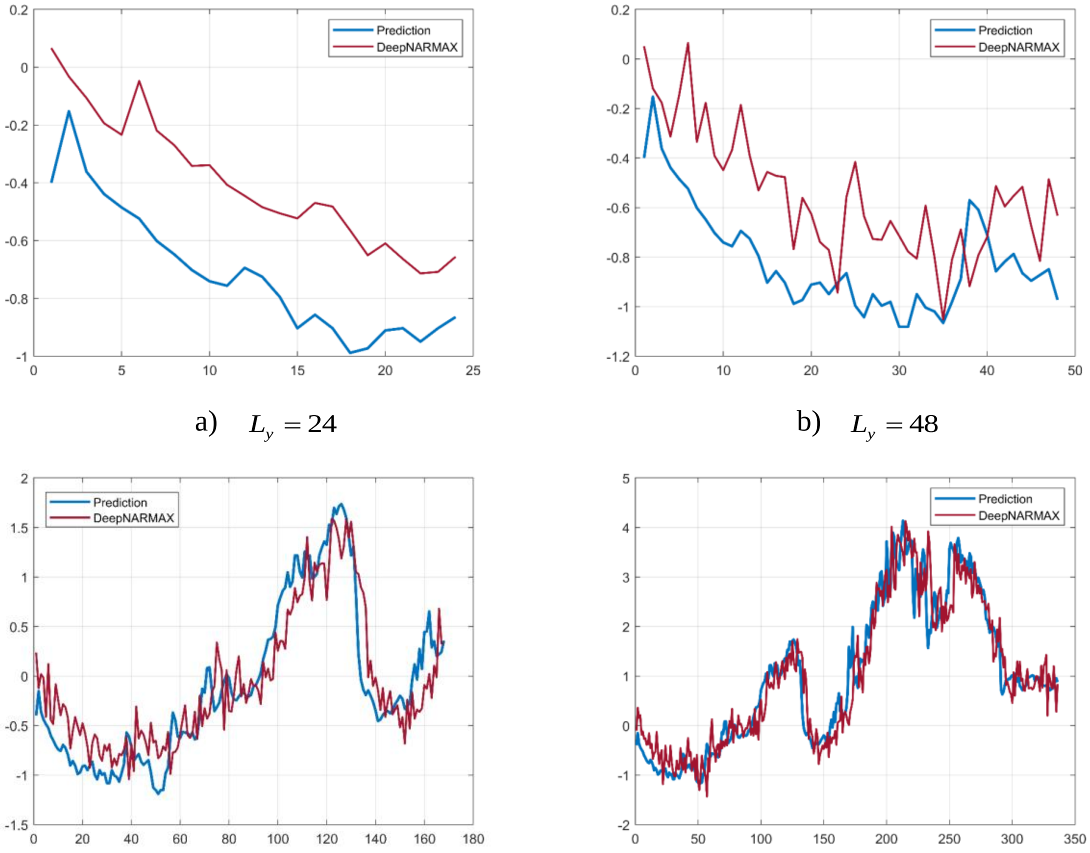
A grid search is executed to optimize the hyperparameters of the DeepNARMAX. The nonlinearity, denoted as ℓ, is chosen from the range $\{3,4,5,6\}$, implying the selection of driller-generator layers from the set $\{6,8,10,12\}$. The initial time delay of the output, denoted as d_y is chosen from the set $\{24,48,168,336,720\}$, while the initial time delay for the input d_u is set at 1. Both the maximum time delay for the output n_y and the input n_u is set at 500 steps. The tuning of other hyperparameters is guided by [238]. Each dataset is standardized to have a zero mean. Each dataset was standardized to have a zero mean. Within the scope of LSTF settings, the prediction window size, denoted as L_y, is incrementally expanded to include {24, 48,168, 336, 720} for the dataset. The method of Prophet employs a series-to-point prediction setting, whilst the RNN-based approaches conduct dynamic decoding by left shifting on the prediction windows.

4.3.3.3 Results of the predictions

The statistical results are listed in Table 11. The table illustrates that the proposed DeepNARMAX model exhibits significantly improved inference performance across all datasets. Additionally, its associated prediction error rate presents a gradual and moderate uptrend as the prediction horizon extends, underscoring DeepNARMAX's successful expansion of its predictive capacity.

DeepNARMAX significantly surpasses the Informer model by exhibiting a MSE reduction of 13% (at 24), 13.5% (at 48), 16.4% (at 168), 19.6% (at 336), and 20% (at 720). This considerable MSE reduction emphasizes DeepNARMAX's exceptional ability to make accurate short-term predictions. As the prediction window expands, the performance of DeepNARMAX gradually converges with that of the Informer model. Thus, while DeepNARMAX may not demonstrate significant improvements in long-term prediction compared to self-attention-based deep neural networks, its predictive accuracy remains commendable.

Further, DeepNARMAX outperforms LSTM. It presents substantial MSE reductions of 17.6% (at 24), 18.9% (at 48), 37.5% (at 168), 45.6% (at 336), and 71% (at 720), implying that DeepNARMAX modelling provides relatively better forecasting abilities than RNN-based models. This holds true for both short and long-term forecasts where DeepNARMAX consistently achieves more accurate predictions. Additionally, DeepNARMAX consistently outperforms DeepAR, ARIMA, and Prophet in terms of MSE, achieving reductions of 43.5% (at 24), 44.4% (at 48), 58.7% (at 168), 70.4% (at 336), and 73.2% (at 720) on average. The value of MSE indicates that the DeepNARMAX model, compared to the DeepAR, ARIMA and Prophet models, possesses both short-term and long-term forecasting capabilities. The aforementioned three models are reliable in short-term forecasting but perform poorly in long-term predictions.



a) $L_y = 24$

b) $L_y = 48$

c) $L_y = 168$

$L_y = 336$

Figure 22 Comparison between observation and prediction by DeepNARMAX model in the prediction windows for weather system

Figure 22 portrays a comparison of observations and predictions for the proposed DeepNARMAX model across the weather dataset, where blue lines represent predictions and red lines denote observations. The comparison substantiates that the DeepNARMAX model can not only generate accurate system predictions but also effectively track changes in system trends. In other words, on a microscale, the DeepNARMAX model efficiently manages rapid system shifts, thereby demonstrating its reliability.

4.3.3.4 Results of the DeepNARMAX models

Table 9 emphasizes that DeepNARMAX, inheriting the interpretability of the NARMAX model, offers a transparent model structure paired with corresponding parameters for each model term. Each term unequivocally delineates the relationship and resulting influence between the system's input and output signals. For instance, when considering a model based on the weather dataset, a 48-point prediction window can be translated into a 48-step-ahead predictive challenge within DeepNARMAX. This suggests that the current system response could be affected by the response registered 48 steps prior, where $y(t-25)$ is significant to the current system response, with u_4 and u_5, which are wind speed and wind direction.

Table 8 Top 5 model terms in the models for ETTh1 with different prediction window

	$y(t)=F(y(t-24),\ldots)$		$y(t)=F(y(t-48),\ldots)$		$y(t)=F(y(t-168),\ldots)$		$y(t)=F(y(t-336),\ldots)$		$y(t)=F(y(t-24),\ldots)$	
	Model term	theta	Model term	theta	Model term	theta	Model term	theta	Model term	theta
1	$y(t-28)^2u_2(t-42)y(t-26)$	0.055	$y(t-49)u_2(t-2)u_2(t-7)$	0.180	$y(t-172)^2$	0.323	$y(t-340)$	0.679298	$y(t-735)$	0.248
2	$y(t-28)^2u_3(t-53)^2$	0.077	$y(t-52)u_1(t-23)^2$	0.197	$u_2(t-4)u_4(t-71)^2$	0.095	$u_3(t-13)y(t-350)$	0.127598	$u_1(t-26)u_5(t-65)$	0.071
3	$y(t-28)^2u_2(t-1)u_6(t-41)$	0.083	$u_2(t-31)u_2(t-80)y(t-49)$	0.136	$u_4(t-76)u_2(t-26)y(t-179)$	0.103	$u_1(t-25)^2$	0.009378	$u_3(t-11)y(t-733)$	0.220
4	$y(t-28)u_5(t-28)u_2(t-42)y(t-26)$	-0.067	$y(t-49)u_5(t-37)u_1(t-92)$	0.134	$u_4(t-76)y(t-172)u_3(t-43)^2$	-0.111	$u_4(t-9)u_2(t-89)$	-0.06196	$u_4(t-34)$	0.180
5	$u_4(t-88)u_1(t-97)$	0.127	$u_2(t-31)y(t-49)u_5(t-47)u_1(t-86)$	-0.077	$u_4(t-90)u_6(t-97)^2$	0.024	$u_6(t-61)y(t-340)$	0.135951	$u_3(t-5)u_3(t-77)$	0.074

Table 9 Top 5 model terms in the models for ETTh2 with different prediction window

	$y(t)=F(y(t-24),\ldots)$		$y(t)=F(y(t-48),\ldots)$		$y(t)=F(y(t-168),\ldots)$		$y(t)=F(y(t-336),\ldots)$		$y(t)=F(y(t-24),\ldots)$	
	Model term	theta	Model term	theta	Model term	theta	Model term	theta	Model term	theta
1	$y(t-42)$	0.879	$y(t-63)$	0.505	$u_5(t-77)u_2(t-162)y(t-189)$	0.109	$u_4(t-22)$	0.109	$y(t-724)$	0.699
2	$y(t-27)^2y(t-29)$	0.210	$y(t-48)u_5(t-21)^2$	0.159	$y(t-174)u_5(t-36)u_5(t-60)$	0.321	$u_2(t-21)$	0.321	$u_6(t-27)$	-0.300
3	$y(t-40)y(t-27)y(t-29)$	-0.593	$y(t-48)y(t-59)y(t-70)$	0.127	$u_6(t-45)y(t-174)u_5(t-36)u_5(t-60)$	0.034	$u_5(t-7)^2$	0.034	$u_6(t-27)u_3(t-45)$	0.089
4	$y(t-37)y(t-25)^2$	0.099	$u_6(t-1)u_5(t-24)y(t-48)$	0.069	$u_3(t-98)$	0.206	$u_4(t-35)u_4(t-45)$	0.206	$u_4(t-92)$	0.225
5	$u_2(t-14)y(t-27)$	0.103	$y(t-63)y(t-48)u_5(t-21)y(t-61)$	-0.096	$u_5(t-77)u_2(t-97)u_5(t-150)y(t-189)$	0.120	$u_1(t-37)u_5(t-46)$	0.120	$u_6(t-27)^2$	-0.016

Table 10 Top 5 model terms in the models for weather with different prediction window

$y(t)=F(y(t-24),\ldots)$		$y(t)=F(y(t-48),\ldots)$		$y(t)=F(y(t-168),\ldots)$		$y(t)=F(y(t-336),\ldots)$		$y(t)=F(y(t-24),\ldots)$	
Model term	theta	Model term	theta	Model term	theta	Model term	theta	Model term	theta

1	$u_3(t-48)$	0.571	$u_4(t-16)u_4(t-5)u_3(t-47)$	0.239	$y(t-169)$	0.479	$u_6(t-1)$	0.420	$u_3(t-21)$	0.782
2	$u_{10}(t-13)u_6(t-1)u_2(t-13)$	0.134	$y(t-25)u_9(t-49)^2$	0.128	$u_{11}(t-20)u_6(t-4)u_2(t-44)$	0.113	$y(t-340)u_2(t-6)u_9(t-33)$	0.157	$y(t-721)u_{11}(t-14)u_{10}(t-23)$	0.160
3	$y(t-25)u_6(t-3)u_4(t-8)$	0.025	$u_{10}(t-6)^2 y(t-25)$	0.027	$u_5(t-35)u_9(t-28)^2$	0.073	$u_3(t-20)^2 y(t-380)$	-0.012	$y(t-730)u_6(t-10)u_6(t-28)$	0.069
4	$u_{10}(t-13)^2 u_6(t-48)$	0.044	$u_5(t-41)u_{10}(t-48)$	0.146	$y(t-169)u_{11}(t-11)^2 u11(t-15)$	0.079	$u_8(t-24)u_3(t-20)y(t-343)$	0.061	$u_3(t-21)u_{11}(t-14)u_{10}(t-23)$	-0.158
5	$y(t-25)u_7(t-4)^2$	0.091	$u_{10}(t-6)^2 u_5(t-5)$	0.084	$u_2(t-12)u_3(t-21)u_2(t-48)$	0.134	$u_8(t-48)u_9(t-48)u_4(t-50)$	0.091	$u_{11}(t-50)u_3(t-21)^2$	-0.244

Table 11 Statistical performance over the testing period on three datasets

Method		DeepNARMAX		Informer		LogTrans		Reformer		LSTMa		DeepAR		ARIMA		Prophet	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	24	**0.032**	**0.151**	0.072	0.206	0.083	0.229	0.222	0.389	0.114	0.272	0.107	0.280	0.108	0.230	0.115	0.275
	48	**0.103**	**0.252**	0.122	0.273	0.131	0.293	0.284	0.445	0.193	0.358	0.142	0.327	0.155	0.274	0.168	0.330
	168	**0.168**	**0.297**	0.172	0.330	0.187	0.355	1.522	1.191	0.236	0.392	0.239	0.422	0.396	0.504	1.224	0.763
	336	**0.234**	**0.374**	0.242	0.417	0.230	0.428	1.860	1.124	0.590	0.698	0.445	0.552	0.468	0.593	1.549	1.820
	720	**0.253**	**0.401**	0.269	0.435	0.273	0.463	2.112	1.436	0.683	0.768	0.658	0.707	0.659	0.766	2.735	3.253
ETTh2	24	**0.042**	**0.175**	0.093	0.240	0.102	0.255	0.263	0.437	0.155	0.307	0.098	0.263	3.554	0.445	0.199	0.381
	48	**0.098**	**0.206**	0.115	0.270	0.129	0.298	0.458	0.545	0.190	0.348	0.141	0.321	3.190	0.474	0.304	0.462
	168	**0.140**	**0.296**	0.169	0.334	0.206	0.392	1.029	0.879	0.385	0.514	0.205	0.394	2.800	0.595	2.145	1.068
	336	**0.203**	**0.357**	0.205	0.375	0.217	0.397	1.668	1.228	0.558	0.606	0.604	0.607	2.753	0.738	2.096	2.543
	720	0.243	0.390	**0.232**	**0.395**	0.263	0.413	2.030	1.721	0.640	0.681	0.429	0.580	2.878	1.044	3.355	4.664
Weather	24	**0.108**	**0.315**	0.117	0.251	0.136	0.279	0.231	0.401	0.131	0.254	0.128	0.274	0.219	0.355	0.302	0.433
	48	**0.154**	**0.318**	0.178	0.318	0.206	0.356	0.328	0.423	0.190	0.334	0.203	0.353	0.273	0.409	0.445	0.536
	168	**0.213**	**0.376**	0.246	0.388	0.309	0.439	0.654	0.634	0.341	0.448	0.293	0.451	0.503	0.599	2.441	1.142
	336	**0.248**	**0.355**	0.297	0.416	0.359	0.484	1.792	1.093	0.456	0.554	0.585	0.644	0.728	0.730	1.987	2.468
	720	**0.251**	**0.389**	0.359	0.466	0.388	0.499	2.087	1.534	0.866	0.809	0.499	0.596	1.062	0.943	3.859	1.144

4.4 Summary

4.4.1 Analysis of model uncertainty in DeepNARMAX

DeepNARMAX as a deepened version of the classical polynomial NARMAX model, has its own deficiency, which lies in the aspect of model uncertainty.

Firstly, the model's sophistication does not necessarily translate into an absolute reflection of the real-world system, despite its capacity to handle complex dynamical scenarios. This limitation often arises due to the difficulty in obtaining an accurate hyperparameter placement during the optimization process. Although the use of PSO-based gate weight optimization algorithm proves to be effective to optimize the selection process, it may still result in a suboptimal selection. This issue could possibly lead to faulty or inaccurate predictions which in turn amplifies the model uncertainty.

Secondly, the decision-making process in the driller layer, which is supposed to prioritize more impactful features, can subject the model to misjudgements. As groundbreaking as this feature might be, inaccurate judgments could arise due to the inherently difficult task of perfectly analyzing the influence degree of different features on the system. This could lead to potentially major features being underrated or minor ones being overestimated, thereby negatively affecting the model's overall accuracy and increasing model uncertainty.

Furthermore, model uncertainty becomes more profound with the increased complexity of the system. This was highlighted with the use of two actual power and weather forecasting models. Whilst they serve as good examples of complex dynamic systems, they also illustrate potential problem areas since such systems are known for their inherent unpredictability. Thus, the same "deepness" that makes DeepNARMAX desirable for complex scenarios might also increase the range of possibilities for errors to creep into the modelling process.

Lastly, the DeepNARMAX relies heavily on iterative training before model refinements are implemented. Although this process is meant to perfectionize the model, each iterative process introduces a level of variation. Hence, it's plausible to consider that subsequent variations could accumulate over time, thereby escalating the model's uncertainty.

In conclusion, while the introduction of a deep learning perspective to the NARMAX modelling process is generally aspirational and beneficial, it brings to light the classical dilemma between model complexity and model uncertainty. Going forward, it will be crucial to continually fine-tune the trade-off between incorporating advanced features and managing the uncertainties that arise thereof. Investigations into more efficient hyperparameter optimization algorithms and feature selection processes may prove beneficial in circumventing some of the highlighted drawbacks.

4.4.2 Summary

The classic polynomial NARMAX model is extensively applied in complex dynamic nonlinear modelling due to its interpretability, reliability, and stability. Its interpretability, worth noting, allows a systematic elucidation of the system's internal logic and the interaction between input/output signals and the system itself, subsequently facilitating the analysis of influential factors in complex dynamic systems. However, practical applications often encounter issues like a low signal-to-noise ratio, escalated model uncertainty, and high demand for hardware computing resources when establishing substantial hyperparameters for the classic polynomial NARMAX model. These issues compromise or even inhibit the effective development of polynomial NARMAX models. To mitigate these problems, this thesis introduces an innovative deep polynomial NARMAX network modelling method.

Initially, we construct and propose the DeepNARMAX network mainly consisting of an input layer, linear transformation layer, driller layer, generator layer, model detection layer, and output layer from basic logical structure of networks. The input layers represent the original inputs *u* and outputs *y* of the system while linear transformation layers linearly expand dimensions based on ℓ and delay information *n* applied to original input signals 'u' and output signals *y* to extract linear dynamic features. The driller layer selects features from previous-layer outputs by reducing their dimensionality with defined driller ratio controlling reduction proportions. The generator performs nonlinear operations on reduced-dimension features, including squaring uniform features or cross-multiplying different ones. Typically, the driller layers pair up with generator layers. The number of these pairs equals ℓ. Finally, the output from the last generator is fed into model detection where the FROLS algorithm is utilized for feature selection, model construction, and parameter calculation, resulting in the final DeepNARMAX model.

To ensure optimal feature selection within the driller's layer and division according to the impact degree on systems, this thesis introduces a novel gate weight hyperparameter describing times each feature vector gets selected during network training or influence degree upon systems. The proposed PSO-based gate weight optimization algorithm updates values during the training process and considers the trade-off between exploration and exploitation thereby optimizing the calculation of characteristic's gate weights. Finally, a check is performed on the optimized DeepNARMAX model against real-world systems over test sets, allowing us to ensure model validity.

A series of comprehensive experiments was performed to affirm the efficiency of the proposed DeepNARMAX networks and algorithms. Three simulated systems were selected, including a single-input-single-output (SISO) system and two multi-input-single-output (MISO) systems, to validate whether the newly proposed DeepNARMAX retains the same interpretability and accuracy as the classic NARMAX. Then, two actual power and weather forecasting models were utilized to validate the performance of the DeepNARMAX under actual, dynamically complex scenarios. After several iterations over the training set, the network obtains one or multiple models then adjusts them using k-fold cross-validation avoiding overfitting problem.

The effectiveness of the proposed DeepNARMAX is further demonstrated by comparison with state-of-the-art deep sequential neural networks such as Informer, LSTM etc. Notably, the interpretability of the DeepNARMAX model is a significant advantage that enables better understanding and utilization of complicated scenarios. In conclusion, the proposed DeepNARMAX addresses the dimension explosion issue encountered in the NARMAX modelling process while enhancing capabilities in complex dynamical situations, particularly analytical abilities towards highly complicated, high-dynamic scenarios, all while retaining the conventional NARMAX models' foundations and effectively resolving uncertainties that arise throughout the modelling process.

Chapter 5

Ensemble Sliding Window NARMAX Models for Seasonal Weather Forecast

Dynamical seasonal forecast models are improving with time but tend to underestimate the amplitude of atmospheric circulation variability and to have lower skill in predicting summer variability than in winter. Here we construct NARMAX to develop the analysis of drivers of North Atlantic atmospheric circulation and jet stream variability, focusing on the East Atlantic (EA) and Scandinavian (SCA) patterns as well as the North Atlantic Oscillation (NAO) index. New time series of these indices are developed from Empirical Orthogonal Function (EOF) analysis. Geopotential height data from the ERA5 reanalysis are used to generate the EOFs. Sets of predictors with known associations with these drivers are developed and used to formulate a sliding window NARMAX model.

5.1 Introduction

5.1.1 Seasonal weather modelling and forecasting

The North Atlantic jet stream strongly influences the weather in Northwest Europe and has a significant role in determining the strength and sign of North Atlantic atmospheric circulation indices such as the North Atlantic Oscillation (NAO), East Atlantic Pattern (EA) and Scandinavian Pattern (SCA); the anomalous weather patterns of a particular season can be described by the interplay of these modes of variability [239]. Recent extreme seasons have been characterised by distinctive jet stream configurations, and jet strength and location are intimately linked with extreme weather conditions (e.g., in temperature and precipitation) experienced across Northwest Europe [239]. Extreme seasonal weather has important socio-economic implications, in terms of risk avoidance, with costs to the insurance industry (e.g., ~£1.5 billion across the UK in winter 2013/14) [240], and impacts on agriculture, food security, energy supply, public health/wellbeing, and severe weather planning.

Until relatively recently, North Atlantic atmospheric variability was thought to be largely due to unpredictable fluctuations [241]. However, dynamical seasonal forecasting systems have been used to develop skilful seasonal forecasts for UK winter weather from a few months ahead [242]. Many factors (drivers) appear to influence the NAO and jet-stream changes, and these potential drivers can be broadly grouped into cryosphere effects from variations in sea-ice extent and snow cover, oceanic effects from NA sea-surface temperatures, tropical influences such as the El-Niño Southern Oscillation (ENSO), and stratospheric effects due to stratospheric circulation variability, solar variability, volcanic eruptions, and the Quasi-Biennial Oscillation [243]. These drivers of jet stream variability can oppose or reinforce one another, and there are indications of interactions between them [14]. Drivers of jet-stream variability show seasonal variation, and distinctive drivers of jet-stream variability operate in different seasons. In addition to these identifiable drivers, a significant part of North Atlantic jet changes is driven by internal unforced variability due to chaotic internal dynamical processes [244, 245]. While a consensus has now been reached that some observed drivers can be reproduced in climate models, improved understanding of more recently identified drivers of the North Atlantic extratropical jet stream is crucial for making progress in UK seasonal climate predictions [243].

The focus of government-funded research is on dynamical forecast systems; however, such forecasts are not always accurate, such as in winter 2004-5 [246] and more recently in 2013-14, when dynamical model forecasts did not well predict the positive winter NAO, and furthermore did not consider the accompanying positive EA pattern and hence the exceptionally heavy rain and flooding in southern England [247]. While dynamical seasonal forecast models are sensitive in winter to tropical forcing such as El Niño events, some evidence suggests that they may be relatively insensitive to Arctic variability [248]. Compared with winter, dynamical model forecasts show relatively little skill in summer, when there is less forcing from the tropics [243]. Recent work on seasonal prediction with dynamical models has also revealed an intriguing conundrum called the Signal to Noise issue: this is where such models reasonably well predict the year-to-year variability of the winter NAO but under-predict its amplitude, due to a systematic underestimation of the mechanisms influencing mid-latitude atmospheric circulation [242, 249, 250]. Supplementing dynamical seasonal forecasting systems, statistical methods identify slowly varying boundary conditions such as sea-ice variability, ocean temperatures and influences from the stratosphere, which are capable of "nudging" the jet stream and providing elements of predictability (e.g., [246, 251, 252]). In the mid latitudes, statistical forecasting has been relatively neglected compared with the tropics; however, recent developments in statistical techniques, under the umbrella of "machine learning" (e.g., [1, 14]) have taken place mainly outside the climate-science community and are relatively quick and cheap to implement.

The novel application of these advanced statistical techniques and systems science methods has significant potential to improve forecast skill and help inform development of the next generation of dynamical seasonal forecasting systems. Here we use a Nonlinear AutoRegressive Moving Average with eXogenous inputs (NARMAX) systems identification approach [1, 14], which is an interpretable machine learning method, to identify and model linear and non-linear dynamic relationships between a range of meteorological and related variables. In addition to its ability to delineate non-linear relations, NARMAX is able to identify non-stationary associations that arise from changes in forcings over time, building on studies where dynamical models have suggested changes in NAO forecast skill over periods of several decades [253], and so NARMAX therefore has significant potential to help improve Northwest Europe seasonal weather forecasts. In a pioneering application of NARMAX in this area, [14] found significant skill in NAO winter forecasting and identified key sources of predictability. Here we extend skilful seasonal forecasting to the summer season, where dynamical seasonal prediction models currently remain most problematic, and identify factors that contribute skill to the forecast. Our results form a firm basis for improving Northwest European regional seasonal weather prediction and should therefore be of interest to potential end-users as well as to model developers and the broader seasonal prediction scientific community.

5.1.2 Description of predictors for indices of seasonal weather forecast

Updated versions of the three principal empirical orthogonal functions (EOFs) of European and North Atlantic atmospheric circulation variability (NAO, EA and SCA) were generated using 500 hPa geopotential height data from the European Centre for Medium-Range Forecasts (ECMWF) ERA5 reanalysis [254], combined with the Python eofs package [255]. ERA5 is based on the Integrated Forecasting System (IFS) and replaces ECMWF's earlier ERA-40 and ERA-Interim reanalysis products. We obtained ERA5 data, specifically of sea surface temperatures (SSTs), sea ice cover, sea level pressure and 500 hPa geopotential heights, from the Copernicus Climate Data Store.

The summer EOFs are based only on high summer (July and August), as using the full June/July/August data generates a poorly defined pattern for EOF1. This is consistent with previous analysis, e.g., [256] which suggest there is a strong summer NAO signal that characterises July and August, while the summer NAO behaves differently during June. The three winter EOFs are based on seasonal data for the full winter quarter (December, January, and February). Maps for the winter and summer EOFs are shown in Figure 23. These are largely similar to the EOFs obtained by [239] but also show some notable differences. For example, the winter Scandinavian pattern has the high-pressure anomaly further west than the [239] version, centred to the north and north-east of the British Isles, while the winter NAO has the low-pressure anomaly centred to the south of Greenland, rather than over Iceland. The differences are mostly down to the different methodology used to calculate the EOFs, rather than different time periods.

Monthly 500hPa geopotential height forecasts from SEAS5 (provided by ECMWF) and GloSea5 (provided by the UK Met Office) were analysed and projected onto the EOFs using the project Field functionality of the elf's package. Seasonal forecast runs were provided by the Copernicus Climate Change Service (C35) via the Climate Data Store website. For both models, hindcasts are available from 1993 to 2016 inclusive. For SEAS5, a complete set of seasonal forecast runs are also available from 2017 onwards, but at the time of analysis for GloSea5, C35 only provided an incomplete set of seasonal forecast runs, covering the winters of 2017/18 to 2019/20 inclusive and the summers of 2018 and 2019.

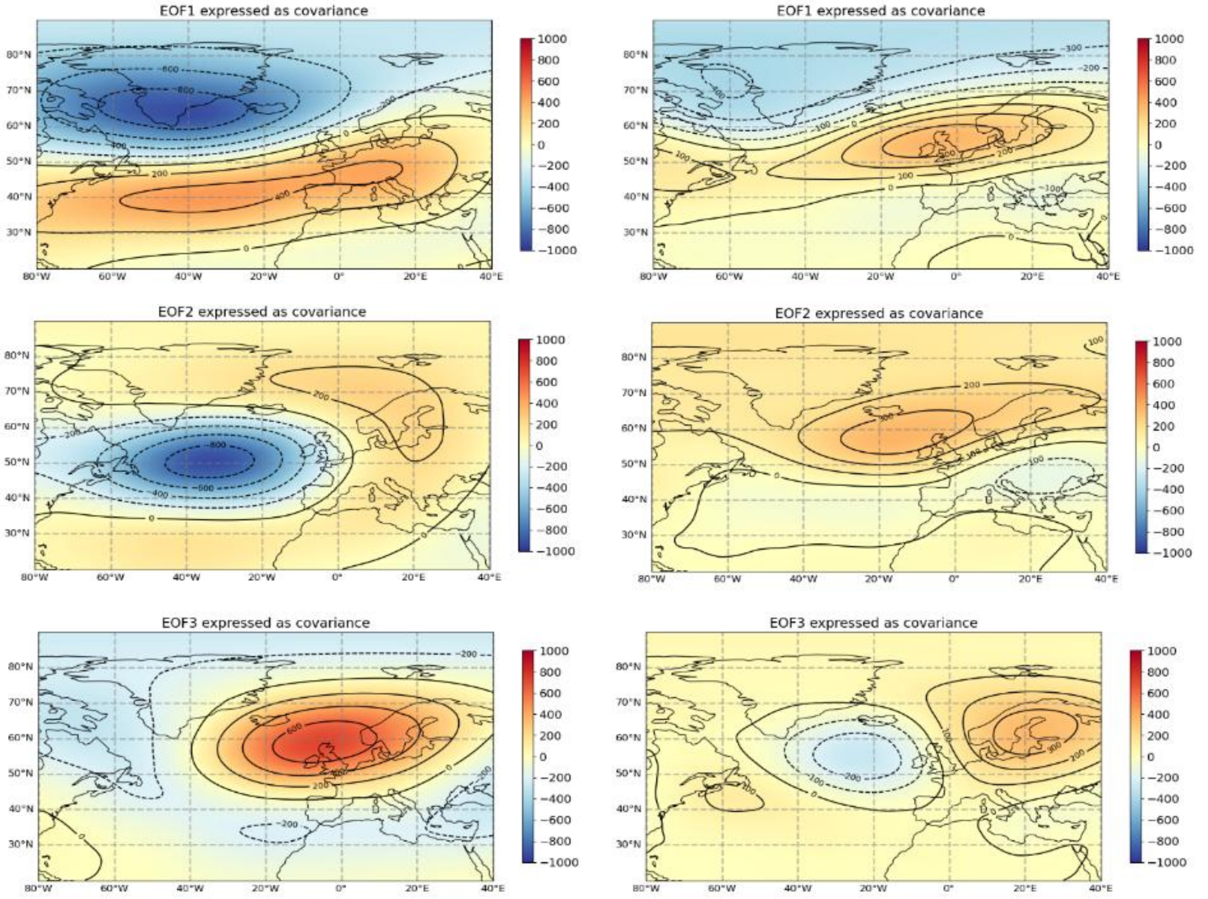

EOF1 expressed as covariance
EOF1 expressed as covariance
EOF2 expressed as covariance
EOF2 expressed as covariance
EOF3 expressed as covariance
EOF3 expressed as covariance
80°N
70°N
60°N
50°N
40°N
30°N
80°W
60°W
40°W
20°W
0°
20°E
40°E
1000
800
600
400
200
0
−200
−400
−600
−800
−1000

Figure 23 The three primary EOFs of atmospheric circulation variability (at the 500 hPa geopotential height level) from ERA5 reanalysis based on 1950 to 2021 for winter and high summer

A number of variables that may be used to predict the North Atlantic jet stream and atmospheric circulation variability, and by extension temperature and precipitation over north-west Europe, have been collected for both winter (DJF) and summer (JJA), building from the drivers identified by [257], [246] and [239]. A wide range of potential drivers has been assembled, so as to be able to select from a wide range of variables for inclusion in NARMAX. SST anomaly patterns are used, including the El Niño Southern Oscillation (ENSO) and the Atlantic Multidecadal Oscillation (AMO), plus sea-ice anomalies, snow cover anomalies and tropical precipitation anomalies. The stratospheric polar vortex and Quasi-Biennial Oscillation (QBO) are used as predictors for winter atmospheric circulation, but not summer, due to a lack of evidence for them having a strong influence on summer atmospheric circulation.

Table 12 Potential drivers of winter and summer atmospheric circulation variability that are used as predictors for NARMAX models

Dataset	Variable used and their abbreviations as used in this study	Region selected	Dates
Atlantic Multidecadal Oscillation (AMO)	ERA5 SST	7-75W, 25-60N, regional SST anomaly minus global SST anomaly	1956-2021
Sea surface temperature	Nino 3.4	170-120W, 5S-5N	
	Tropical Atlantic (TASST)	50W-0E,5S-5N	1956-2021
	W.Indian Ocean (WISST)	50-85E, 5S-5N	1956-2021
	E. Indian Ocean (EISST)	85-120E,5S-5N	1956-2021
	W. Pacific (WPSST)	120-170E,5S-5N	1956-2021
	E. Pacific (EPSST)	140-90W, 5S-5N	1956-2021
	North Atlantic Horseshoe (NAH)	40-15W, 15-30N minus 60-40W,30-45N	1956-2021
	North Atlantic dipole (DIP)	52-40W, 42-52N minus 35-20W, 35-42N	1956-2021
	North Atlantic tripole (TRI)	60-40W, 40-55N minus 80-60W, 25-35N	1956-2021
	Sub-Polar Gyre (GRE)	60-10W, 50-65N	1956-2021
	Barents Sea SST (Bar_SST)	25-70E, 75-80N	1956-2021
	Greenland/Iceland Norwegian Seas (GIN)	20W-20E, 65-80N	1956-2021
	North Atlantic SST gradient (SST_grad)	60-30W, 20-40N minus 60-10W, 50-65N	1956-2021
	Sub-Polar Gyre (SPG_SST)	60-10W, 50-65N	
Sea ice concentration	Barents-Kara Seas (BK)	10-100E,65-85N	1956-2021
	E. Siberian/Laptev Seas (ESL)	100-180E, 68-85N	1956-2021
	Beaufort/Chukchi Seas (BC)	180-120W, 68-85N	1956-2021
	Canadian Archipelago/Baffin Bay (ArB)	120-45W, 63-80N	1956-2021
	Greenland Sea (GRE)	45-0W, 63-85N	1956-2021
	Bering Sea (BER)	195-155W, 55-68N	1956-2021
	Hudson Bay (HUD)	100-70W, 50:63N	1956-2021
	Labrador Sea (LAB)	70-45W, 40-63N	1956-2021
Tropical precipitation	Tropical Atlantic Rainfall (TAR)	50W-0E,5S-5N	1979-2021
	W. Indian Ocean Rainfall (WIR)	50-85E, 5S-5N	1979-2021
	E. Indian Ocean Rainfall (EIR)	85-120E,5S-5N	1979-2021
	W.Pacific Rainfall (WPR)	120-170E,5S-5N	1979-2021
	E. Pacific Rainfal (EPR)	140-90W, 5S-5N	1979-2021

Stratospheric polar vortex	Temperature 100hPa	65-90N	1956-2021
Sea level pressure	Barents SLP	60-120E, 67.5-90N	1956-2021
Carbon dioxide	Annual CO_2 level	NA	1959-2021
QBO	Mean zonal wind, 30hPa	NA	1956-2021
sunspots	Sunspot no.	NA	1956-2021
Snow cover extent	Eurasian snow	55-150E, 45-80N	1979-2021
HadCRUT5	2m Temperature anomaly	90W-90E, 20-80N	1955-2021
MJO Indices	200hPa velocity potential anomalies		1979-2021

Sea surface temperature anomalies, sea ice coverage anomalies, the Atlantic Multidecadal Oscillation (AMO), tropical precipitation anomalies and the strength of the stratospheric polar vortex were calculated based on ERA5 reanalysis data. This version of the AMO is based on the region from 7-75ºW, 25-60ºN, subtracting the global SST anomaly from the regional SST anomaly to remove biases that would result from the upward trend in global SSTs. For summer, an SST based North Atlantic dipole index is used, based on [258] , who provided evidence for a link between this and a high-pressure anomaly to the west of the British Isles, resulting in relatively anticyclonic weather over Britain. The North Atlantic Horseshoe SST pattern [259] is linked with the winter NAO. The North Atlantic tripole is based on the methodology of [260], who provided evidence for this being linked especially with the NAO in winter. Snow cover data are based on. Monthly sunspot numbers were obtained from the Solar Influences Data Analysis Center [261]. The QBO data are obtained from the Free University of Berlin [262]. A full list of the drivers is provided in Table 12.

5.2 The ensemble sliding window NARMAX models

5.2.1 The sliding window NARMAX models

Usually, the NARMAX model can generate reliable predictions and reveal convincing transparent relationships between the system input and output over the entire period of the dataset. However, it is more desirable to pay attention to and make use of local information in the dataset, especially for a complex dynamic time-varying system or process like climate change. Therefore, in the following we introduce the sliding window NARMAX model for seasonal forecasts.

The sliding window is a popular approach used for analysing and processing time-series data, signal processing and machine learning [263-265]. The main idea of this method is to take a subset of data, referred to as a ‘window’, and slide that window across the entire dataset, performing some operation on the data within the window before moving it. Due to the outstanding performance of understanding the time-varying system, the sliding window is introduced into the NARMAX modelling framework to discover the local information and evolution of the system dynamics over time. In [266], a 30-year sliding window is utilized in NARMAX modelling to investigate the intricate nonlinear relationship between iceberg discharge. The outcome demonstrates a significant level of fitness and emphasizes that there’s been a century-long shift in dominant causes from primarily glaciological (SMB) to climatic (ocean temperature). In [267], the sliding window NARMAX was introduced to capture the time-varying relationships and contributions of these factors over the century because a static model is not sufficient for analysing the discharge of icebergs from West Greenland during the 20th century, as the factors contributing to this discharge (including climatic and glaciological conditions) are non-linear and varied greatly over the century.

The sliding window NARMAX is crucial for seasonal weather forecasting due to its ability to model complex and dynamically changing systems. It can adapt over time to reflect changes in influential factors and their interactions, enhancing the accuracy of weather predictions. It’s particularly effective in handling complex and nonlinear relationships within meteorological variables. Despite not wholly eliminating uncertainties, it significantly improves forecast accuracy.

The proposed framework of the sliding window NARMAX model is shown in Figure 24. Taking a single-input, single-output system as an example, where the measured input signal of N samples is denoted by $U=[u_1,u_2,...,u_N]^T$ and the corresponding output is denoted by $Y=[y_1,y_2,...,y_N]^T$. In a matrix format, the dataset of the system can be represented as $D=[U,Y]=[(u_1,...,u_N)^T,(y_1,...,y_N)^T]$.

Let $W=[1,1,...,1]_{w\times 1}$ be a window of length w. With the one-step forward sliding window (which is shown in the dotted rectangles in Figure 24, the original dataset D can be resampled as follows:

$$\begin{cases} \hat{D}_1=[(u_1,...,u_w)^T,(y_1,...,y_w)^T] \\ \hat{D}_2=[(u_2,...,u_{w+1})^T,(y_2,...,y_{w+1})^T] \\ ... \\ \hat{D}_M=[(u_{N-w+1},...,u_N)^T,(y_{N-w+1},...,y_N)^T] \end{cases} \tag{5.1}$$

For each windowed dataset $\hat{D}_m$, a NARMAX model $\hat{Y}_m$ can be generated using the method discussed in Chapter 2. Thus, based on (5.1), there will be M NARMAX models $\hat{Y}=(\hat{Y}_1,...,\hat{Y}_M)$, which are represented by the green rectangles in Figure 24, and M predictions of the system output $\hat{y}=[\hat{y}_1,...,\hat{y}_M]$ will be calculated accordingly. Therefore, the sliding window NARMAX method produces a set of system predictions rather than a single forecast value, and this set forms the basis for a NARMAX-derived probabilistic prediction of the system output.

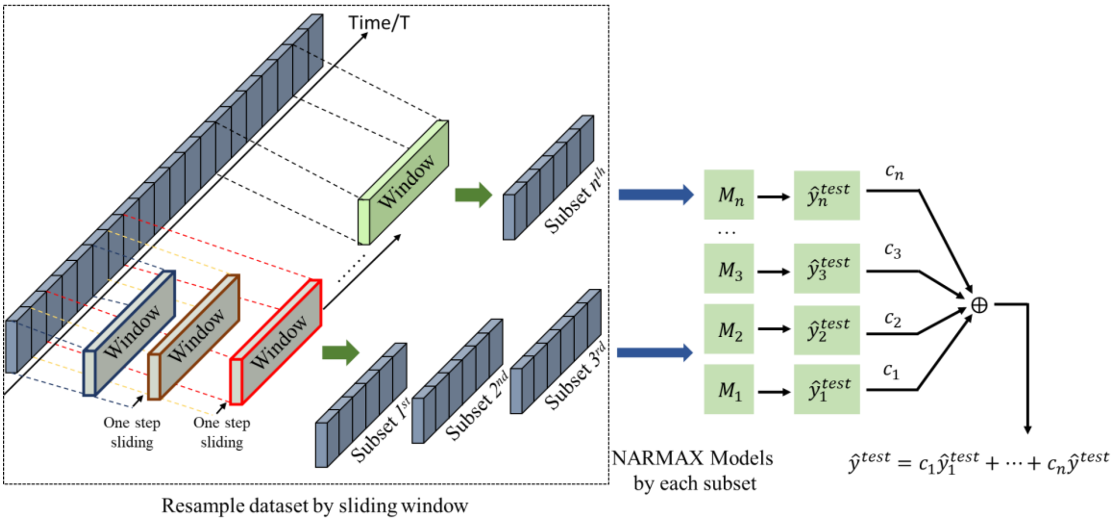


Figure 24 The framework of the sliding window NARMAX models.

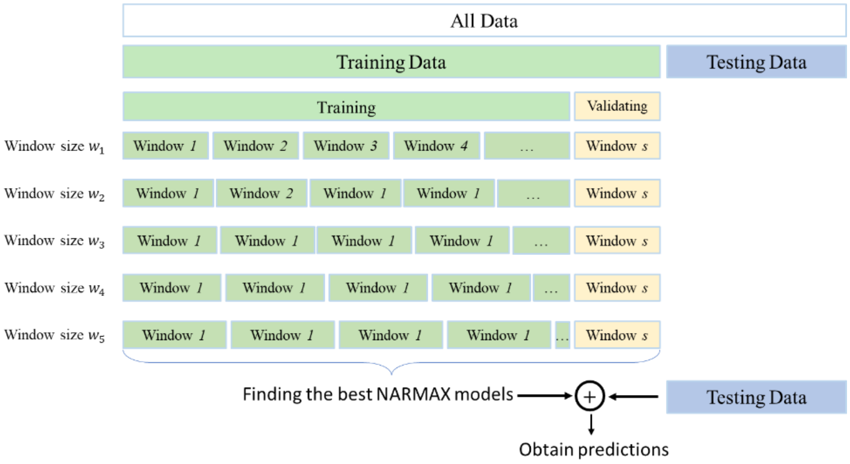


Figure 25 The schematic illustration of the resample by sliding window

Note that the whole available observations are split into two parts: 1) around 80% of the data are used for model training; 2) the remaining 20% are used for model testing. Each windowed dataset is a subset of the training dataset. To find the most appropriate model for each window, the associated windowed dataset is further partitioned into training and validation sub-datasets. For each window, the role of the validation data is threefold: 1) to test and validate the model performance using "unseen" data during the training process; 2) to optimise and adjust model parameters and hyper-parameters, such as the window size and the model structures, where necessary; and 3) to avoid overfitting.

5.2.2 Model selection and verification

NARMAX modelling encompasses both linear and nonlinear models. In this case, as discussed above the polynomial form of model structure, including both linear and nonlinear forms, is considered. Note that different nonlinear degrees can lead to quite different models, which will affect the forecast and the explanation of the system. A model with a higher nonlinear degree may produce a more accurate representation of the system and hence produce better prediction. However, such a model can become very complex, and it needs a large number of samples for model training and estimation. Therefore, the present modelling challenge is a typical small-size problem, where the number of observations is smaller than the number of regressors. Taking a system with *n* input variables as an example, if the nonlinear degree of the model is ℓ, then the number of generated model terms would be $(n+\ell)!/n!\ell!$, where the symbol "!" denotes the factorial function. Thus, it is unlikely to have a high nonlinear degree ℓ, such as $\ell = 4$, otherwise, the number of generated model terms, $(n+\ell)!/n!\ell!$, would be tremendous, leading to an extremely sensitive identified model.

In order to avoid overfitting and reduce the sensitivity of the model but meanwhile guarantee a reliable model structure, the highest nonlinear degree is set to be three in this study. Then for each windowed dataset, there are 3 candidate model forms: M_{mli}, M_{mq}, M_{mc}, representing the linear, quadratic, and cubic models for the *m-th* windowed dataset, respectively. Therefore, with the window of length *w*, there are a total of 3M NARMAX models including linear and nonlinear model forms. To select the best models from the huge number of potential model candidates, the validation set is applied as discussed in above. For each model M_{mi} (identified from the *m-th* windowed dataset), where $i \in \{li, q, c\}$, the prediction for the validation set is denoted by $\hat{y}_{mi}^{v}$. The values of the mean squared error (MSE) of each of the 3M identified models are denoted by, mse_{mli}^{v}, mse_{mq}^{v}, and mse_{mc}^{v}, respectively. The best model is selected by comparing the MSE in each data group. Thus, for each window of length *l*, there are M best NARMAX models.

To reduce the risk that arises from using a single model and in order to constrain model uncertainty, NARMAX predictions are based on a weighted average of multiple models [14]. In this study, the weighted average scheme is also considered to deliver the predicted value. But unlike the method in [14], the weights are calculated based on the MSE of the M models over the validation period, rather than over the training period. NARMAX methods are generally robust for system analysis and prediction but using a single "best" model may be risky in some applications. Therefore, it is reasonable to apply a model-averaging algorithm to reduce the risk associated with depending solely on a single model, especially when dealing with small sample size applications; this can also mitigate the sensitivity of the model to noise or uncertainties [14]. In this study, the weighted mean scheme is also considered to deliver the predicted value. But unlike the method in [14], the weights are calculated based on the MSE of the M models over the validation period, rather than over the training period.

Assume the values of mean squared errors (MSE) of n MARMAX models over their respective training periods are known as $mse_1, ..., mse_s$, respectively. Let

$$l_1 = 1/mse_1, ..., l_n = 1/mse_s \tag{5.2}$$

$$l = l_1 + ... + l_n \tag{5.3}$$

$$c_1 = l_1 / l, ..., c_s = l_s / l \tag{5.4}$$

Then, the averaged model prediction can be defined as:

$$\hat{y}^{test} = c_1 \hat{y}_1^{test} + ... + c_n \hat{y}_s^{test} \tag{5.5}$$

Some typical model validation criteria, including correlation coefficients, mean absolute error (MAE) and root mean square error (RMSE), are used to evaluate model performance. The number of forecasts needs to be sufficiently large to make the statistical conclusions about the skill of the forecast robust and convincing. In this study, we use the out-of-sample set, comprising the validation set and the testing set, to calculate the values of the statistical metrics. Models are trained over the 80% of the original dataset, while model performance is evaluated over the remaining 20% of the dataset. Also, the Brier score [268] and the reliability diagram [269] are used in this study to evaluate the quality of the seasonal forecast models. The Brier score estimates the difference between the observed and expected outcomes. Brier introduced it in 1950 to address the verification of weather forecasts [270], and it is herein defined as:

$$BrierScore = \frac{1}{n}\sum_{i=1}^{n}(d_i - x_i)^2 \tag{5.6}$$

Normally, x takes the value 1 or 0 according to whether or not the event occurred in the predefined class, especially a binary classification forecast, while p_i is the forecast probability for such occasion i. To clearly calculate the Brier scores of the SW-NARMAX models, we define the two class as follows:

$$class\,1: x = 1,\ p_i \in [-0.5, 0.5] \tag{5.7}$$

$$class\,2: x = 0,\ p_i \in [-3, -0.5) \cup (0.5, 3] \tag{5.8}$$

To calculate the Brier score of the SW-NARMAX models, the forecast probability p_i should be obtained first. By applying the sliding window methods in the training period, there will be *s* SW-NARMAX models identified as discussed above. Thus, for each year in the testing set, there are *s* predictions as defined above. We can calculate the forecast probability by

$$p_i = \frac{count(\hat{y}_i \in [-0.5, 0.5])}{s} \tag{5.9}$$

where $count$ means function to calculate the quantity that meets the condition, $\hat{y}_i, i = 1, 2, ..., s$ indicates the predictions of *i-th* SW-NARMAX model.

Based on the definition of the classes, the Brier score verifies the accuracy of the predictions of SW-NARMAX models matching the class of the observation. The smaller the Brier score, the more consistent the category of SW-NARMAX predictions is with the category of observed values; otherwise, they are inconsistent.

5.2.3 Dynamical models

To help assess the accuracy and utility of NARMAX-generated seasonal forecasts, comparisons are made with the seasonal forecasts from two commonly used dynamical models. Monthly forecast runs are obtained from the C35 Copernicus Climate Change Service. For this analysis, hindcast data for 1993 to 2016 are used from the ECMWF SEAS5 model [271] and the Met Office GloSea5 model [272]. The GloSea5 outputs are based on seven ensemble members, while SEAS5 has 25 ensemble members over this period. The monthly runs are aggregated to produce a seasonal forecast that corresponds to the seasonal prediction from one month out (e.g., the winter forecasts are based on December with one month lead time, January with two months lead time, and February with three months lead time, corresponding to a seasonal forecast issued in November).

Predictions of 500hPa heights from the dynamical models are assessed against the three EOFs discussed in Section 2, for both winter and summer. As with NARMAX, in the case of summer, June is considered separately from high summer (July and August), while winter is defined simply as December, January and February and is labelled by the year of the January. Correlation, RMSE and the Brier score are used together to provide an indication of forecast skill. Reliability diagrams are used to give an additional probabilistic indication of the reliability of these model predictions. These compare the observed frequency of an event with the forecast probability of it happening [273], grouped together into probability bins. Reliability is high if the probability of the event happening is close to the forecast probability.

5.3 Experiments and Results of seasonal weather forecast based on sliding window NARMAX models

In this section, sliding window NARMAX models are defined by the indices they use (station-based NAO, EA, and SCA), by the start year of the predictor dataset (1950 or 1979), and by season (summer or winter). For example, the junSNAO79 (JASNAO79) summer models are the sliding window NARMAX models for the summer SNAO in June (July and August average), using the 1979-2022 predictor dataset. In the main part of this thesis, we focus on the NARMAX prediction results based on the 1979 (start date) datasets, while the rest of the prediction results are presented in the Supplementary Information. For a better evaluation of the performance of different models against observations, model results are based on weighted averages of the model ensemble members over validation and test periods.

5.3.1 Summer seasonal prediction results

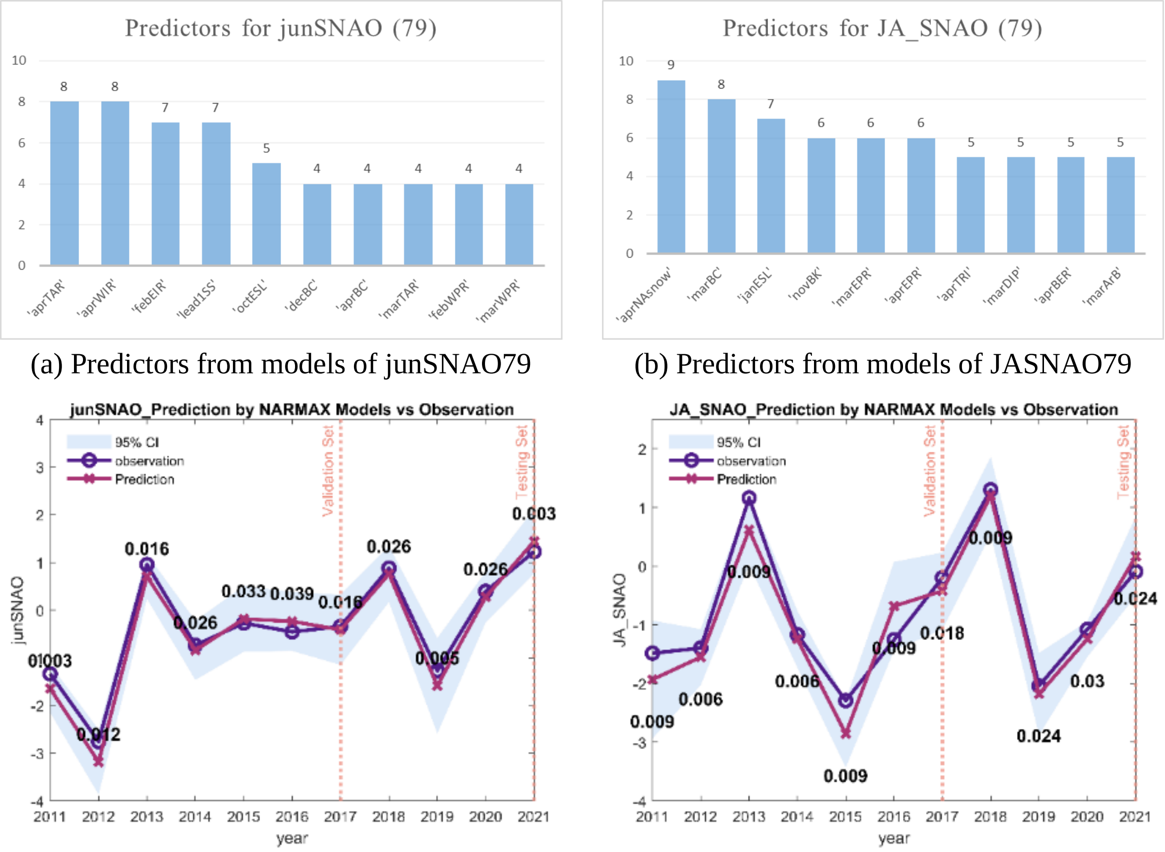
5.3.1.1 SNAO summer results

For each index, the ten most frequent predictors in the sliding window NARMAX models (16

JunSNAO79 models and 15 JASNAO models) are listed in Figure 26 (a-b). Although the frequency of the predictors between the various models is different, there are some common patterns among the predictors. ‘BC’ and ‘ESL’ are both selected in both the JunSNAO79 and JASNAO79 models, 16 and 12 times respectively. For JunSNAO79 models, ‘TAR’, ‘WIR’, ‘EIR’, and ‘WPR’ are more important than other predictors, while ‘NAsnow’, ‘BK’, ‘EPR’, ‘TRI’, ‘DIP’, ‘BER’ and ‘ArB’ have more influence for JASNAO79.

The predictions made by the sliding window NARMAX models are presented alongside the observed SNAO time series in Figure 26 (c-d). In the figures, the red lines with ‘o’ markers are the observation of the indices (i.e., the EOF time series defined in Section 2), while the blue lines with ‘x’ markers are the weighted average predictions by the sliding window NARMAX models. The light blue area is the 95% confidence interval (CI) generated by the identified NARMAX models. As shown in Figure 26 (c-d), the averaged predictions by sliding window NARMAX models closely match the observed NAO changes, with the observed NAO values mainly lying within the 95% CI range. The 4-digit decimal values in Figure 3 (c-d) are the Brier scores based on the NARMAX ensemble weighted mean prediction for each year. These Brier scores are close to zero, with the highest value 0.036 in 2015 in Figure 26 (c), indicating the robustness of the NARMAX models. Thus, the results overall indicate that the sliding window NARMAX models can produce generally accurate and reliable predictions.

NARMAX verification statistics against observed data are summarised in Table 13. Model-observation correlations coefficients of the junSNAO79 models and JASNAO79 models for the entire out-of-sample period (2011-2021) are 0.94 and 0.91 respectively and are highly significant (p<=0.01). To more fully assess model performance, we also present RMSE and MAE values for the validation and testing sets. For JunSNAO79 models, RMSEs (MAEs) of the validation set and testing set are 0.10 (0.07) and 0.33 (0.27). For JASNAO79 models, RMSEs (MAEs) are 0.166 (0.11) and 0.49 (0.39).



(a) Predictors from models of junSNAO79

(b) Predictors from models of JASNAO79

(c) Comparison between observation and prediction band of junSNAO79

(d) Comparison between observation and prediction band of JASNAO79

Figure 26 Results (predictors (a, b) and predictions (c, d) by sliding window NARMAX) of junSNAO79 and JASNAO79.

Predictors are shown according to the month in which that value occurred (e.g., AprTAR = tropical Atlantic rainfall for the month of April). Refer to Table 12 for a full list of predictor names.

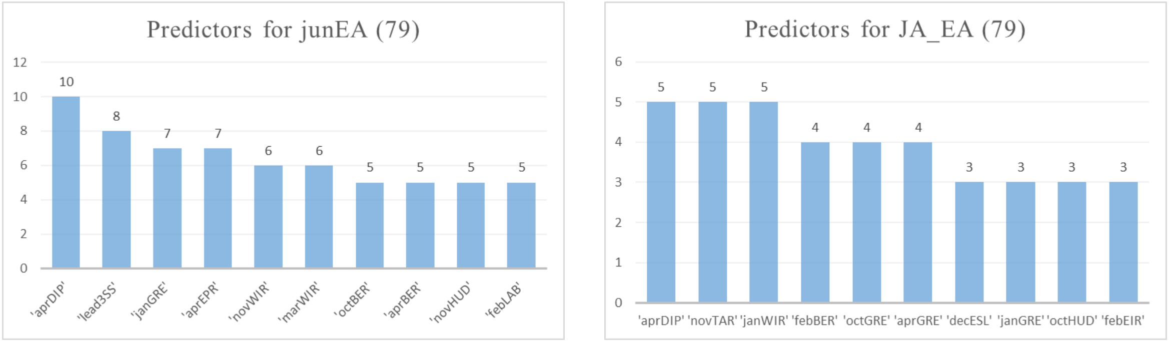
5.3.1.2 EA summer results

The 10 most frequent predictors in the sliding window NARMAX models (16 JunEA79 models and 11 JAEA79 models) are shown in Figure 27 (a-b). There is some consistency in the analysis of the predictors: ‘DIP’, ‘HUD’, ‘GRE’, ‘WIR’, and ‘BER’ are selected in models for two indices with high frequencies. Among them, ‘aprDIP’ is the most frequency selected predictor for the two indices: 10 times for junEA79 and 5 times for JAEA79. Meanwhile, there are several unique predictors in NARMAX models, like ‘lead3SS’ for JunEA79 and ‘TAR’ for JAEA79. ‘lead3SS’ shows high influence on the JunEA79 and less effect on the JAEA79, while ‘TAR’ has the opposite effect on the two indices.

The comparison between the prediction of sliding window NARMAX models for SNAO and the observations is shown in Figure 27 (c-d), while the verification statistics for the validation and testing periods for the average predictions by the model ensembles are shown in Table 13。

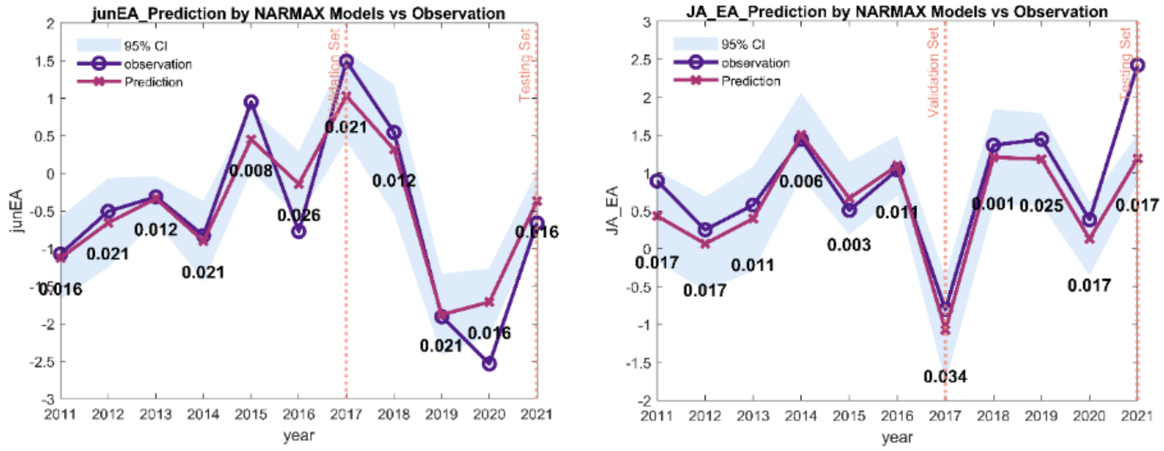
Similarly, Figure 27 (c-d) show that the weighted average predictions by NARMAX models usually perform well in following the observed yearly values from 2011 to 2021 (out-of-sample period) for JunEA79 and JAEA79. Most observations are in the prediction band formed by the 95% CI of the predictions. However, in four years, 2019 and 2020 for JunEA79 and 2019 and 2021 for JAEA79, the NARMAX models did not produce accurate predictions. These four more poorly predicted years are in the testing set, which suggests that the selected NARMAX models are unable to follow year-to-year EA changes without further prior information. Despite this limitation, overall, the Brier scores in Figure 27 support the robustness of the sliding window NARMAX models.

Model-observed correlation coefficients of the JunEA79 and JAEA79 models for the out-of-sample period (2011-2021) are 0.90 and 0.85 respectively (Table 13), which are highly significant (p <= 0.01). In the validation period, models generate accurate predictions for these two indices, with low RMSE (MAE) values of 0.12 (0.09) for JunEA79, and 0.22 (0.18) for JAEA79. The errors between the predictions and observations are marginally greater for the testing period, but the overall performance of the NARMAX models remains robust.



(a) Predictors from models of junEA79

(b) Predictors from models of JAEA79



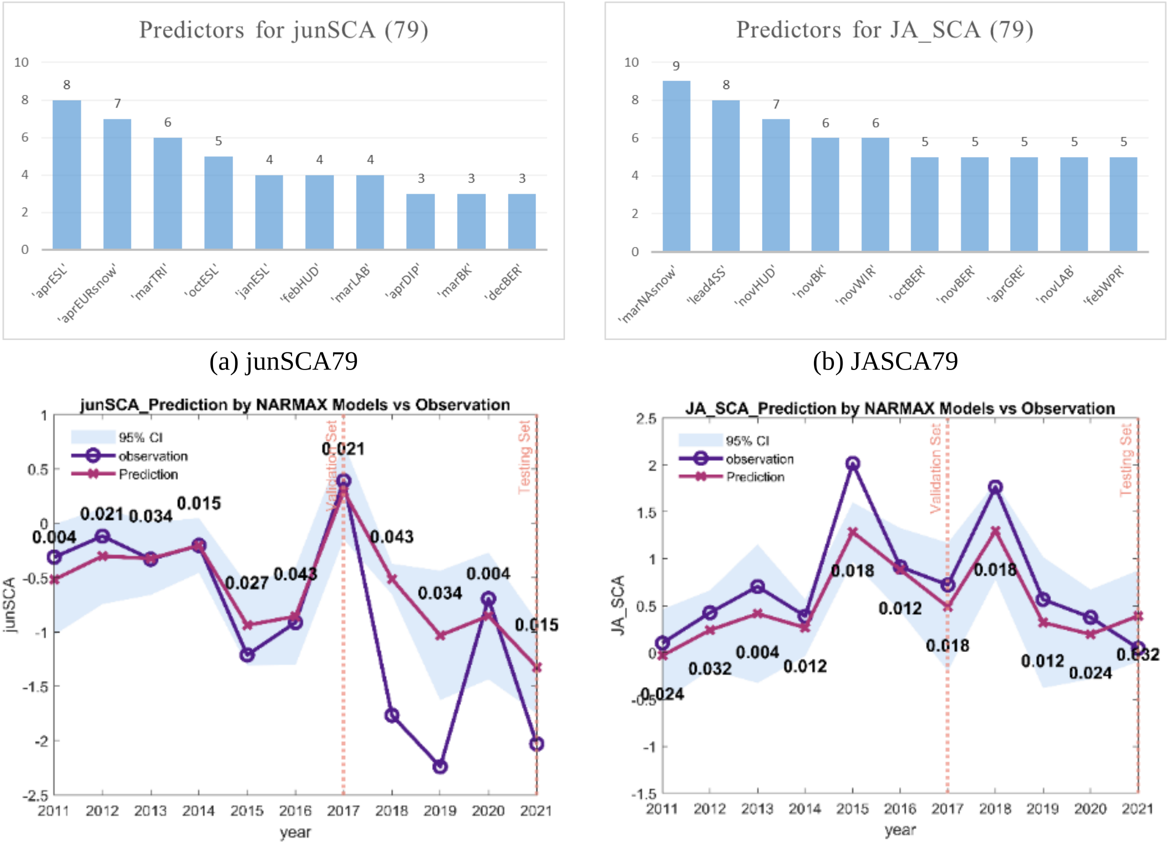
(c) Comparison between observation and prediction band of junEA79

(d) Comparison between observation and prediction band of JAEA79

Figure 27 Results (predictors (a, b) and predictions (c, d) by sliding window NARMAX) of junEA79 and JAEA79.

5.3.1.3 SCA summer results

The statistical analysis of predictors in 14 JunSCA79 NARMAX models and 13 JASCA79 models are shown in Figure 28 (a-b). The models for two SCA indices have some common predictors, including ‘LAB’, ‘HUD’, ‘BK’ and ‘BER’. Similarly, there are some unique predictors in the models of each index: i.e., ‘DIP’, ‘EURsnow’ and ‘TRI’ in the JunSCA79 models, and ‘NAsnow’, ‘lead4SS’ and ‘GRE’ in the JASCA79 models. Ranked by frequency of the predictors in the models for JunSCA79, ‘ESL’ has the most influence on the index, since it occurs 17 times (out of 47) in the analysis.



(a) junSCA79

(b) JASCA79

(c) Comparison between observation and prediction band of junSCA79

(d) Comparison between observation and prediction band of JASCA79

Figure 28 Results (predictors (a, b) and predictions (c, d) by sliding window NARMAX) of junSCA79 and JASCA79.

Similarly, Figure 28 (c-d) shows the comparison between the weighted averaged predictions and observations over the out-of-sample period for SCA. Most observations are, once again, in the prediction band generated by the sliding window models. Brier scores marked in the figures reflect the robustness of the trained models with a highest value of 0.034 (for the JunSCA79 in 2019) However, as before, in a very few years, i.e., 2018 for JunSCA79 and 2015 and 2019 for JASCA79 (i.e., three out of 22 years), the probabilistic prediction bands do not encompass the observations. The statistical verification metrics for these models are once again summarised in Table 13.

Model-observation correlation coefficients of the junSCA79 models and JASCA79 models for the whole out-of-sample period (2011-2021) are respectively 0.77 and 0.75 (Table 13), which again are significant at the $p < 0.01$ level. The values of RMSE and MAE in the validation and testing sets show that the two NARMAX SCA models produce generally good predictions in the out-of-sample period. Although RMSE and MAE in the testing set are slightly higher than those in the validation set, the testing period values remain small, supporting considerable predictive capability of the NARMAX models.

Table 13 Verification statistics for averaged sliding window NARMAX models for summer seasonal prediction (validation period = 2011-2016; test period = 2017-2021).

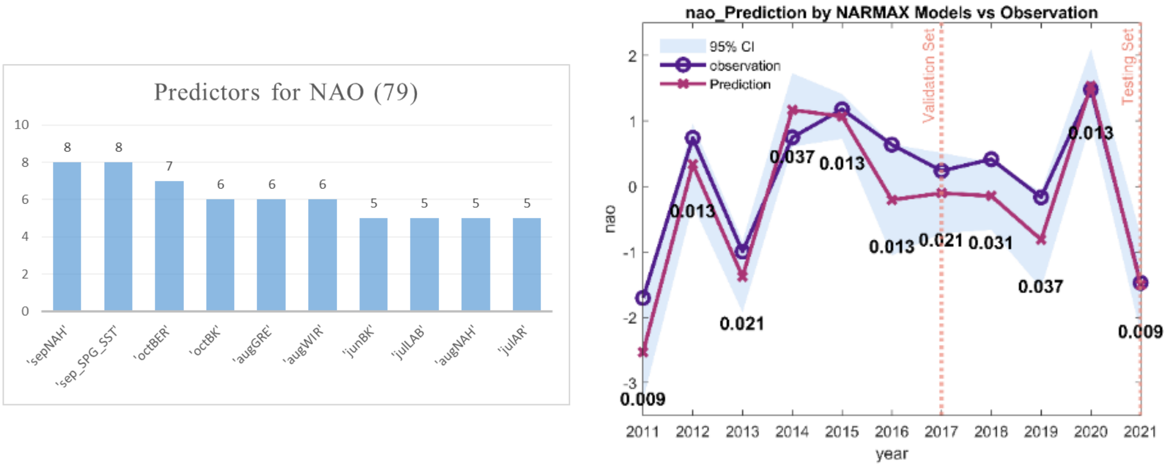
	RMSE		MAE		Correlation (2011-2021)
	Validation	Testing	Validation	Testing	
junSNAO79	0.10	0.33	0.07	0.27	0.94
JASNAO79	0.16	0.49	0.11	0.39	0.91
junEA79	0.12	0.48	0.09	0.46	0.90
JAEA79	0.22	0.55	0.18	0.43	0.85
junSCA79	0.11	0.39	0.08	0.38	0.77
JASCA79	0.36	0.57	0.25	0.54	0.75

5.3.2 Winter seasonal prediction results

5.3.2.1 NAO winter results

The frequency analysis of predictors in the winter NAO models are shown in Figure 29 (a), showing the ten most chosen predictors among all models. As listed in Figure 29 (a), the joint most selected predictors are ‘SepNAH’ and Sep_SPG_SST, with ‘OctBER’ as the next most frequently identified predictor. The performance comparison is shown in Figure 29 (b). The comparison shows that sliding window NARMAX models have accurate prediction capability, with low Brier scores shown in the figures.

The verification statistics of weighted average of models is shown in Table *14*. For the winter NAO79 model, the model-observation correlation coefficients over the out-of-sample (2011-2021) period are 0.92. For this model, the RMSE (0.17) and MAE (0.11) values for the validation set are lower than those for the testing set (RMSE of 0.62 and MAE of 0.52) but both these sets of values indicate high model accuracy.



(a) Predictors analysis of models for winter NAO79

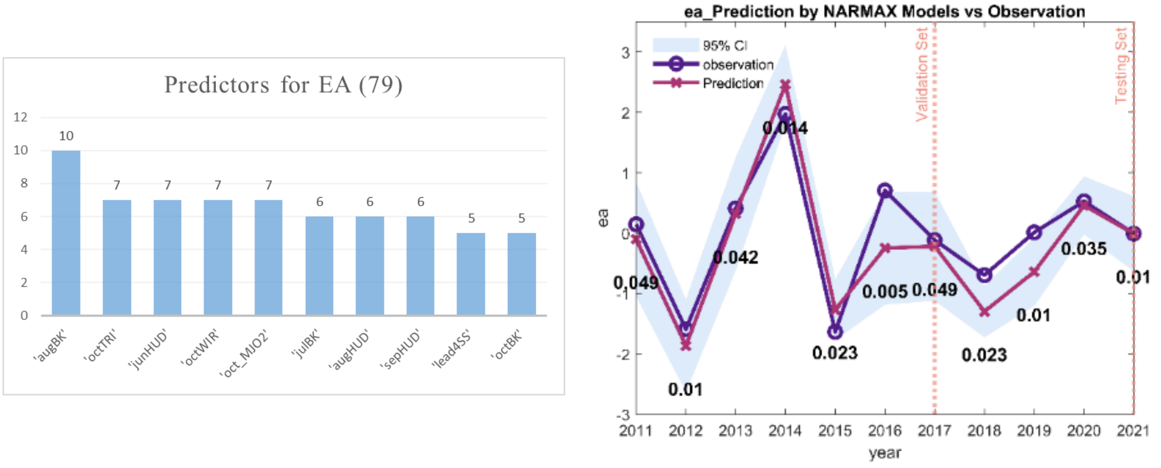
(b) Comparison between observation and averaged prediction for winter NAO79

Figure 29 Results (predictors (a) and predictions (b) by sliding window NARMAX) of NAO79 in winter.

5.3.2.2 EA winter results

Figure 30 (a) shows the predictor analysis in models of the winter EA79. 'augBK' is the most commonly identified predictor in winter EA79 models. Based on the frequency analysis of predictors in EA79 predictor models, the most commonly selected predictors are 'HUD' and 'BK', indicating they have more influence on the winter EA.

Performance comparison is shown in Figure 30 (b). The weighted average prediction by winter EA79 models (blue line) closely follows the observations in both the validation and training sets, and the model prediction band consistently encompasses the observations. This comparison once again highlights the skilful performance of the NARMAX models, which is confirmed by the statistical verification results shown in Table 14.

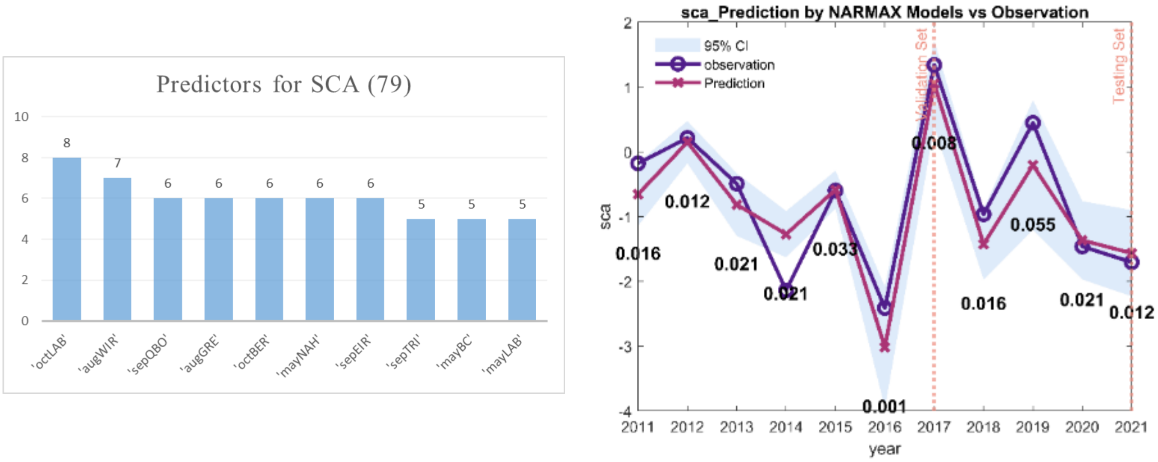


(a) Predictors analysis of models for winter EA79

(b) Comparison between observation and averaged prediction for winter EA79

Figure 30 Results (predictors (a) and predictions (b) by sliding window NARMAX) of EA79 in winter.

5.3.2.3 SCA winter results



(a) Predictors analysis of models for winter SCA79

(b) Comparison between observation and averaged prediction for winter SCA79

Figure 31 Results (predictors (a) and predictions (b) by sliding window NARMAX) of SCA79 in winter.

The 10 most frequent predictors in winter SCA79 models are shown in Figure 31 (a), where sea ice dominates the predictors and 'LAB' is the only predictor to occur more than once. Figure 31 (b) shows the comparison between observations and weighted average prediction of identified models. Most observations are in the prediction band, although the winter SCA79 models did not produce a skilful prediction in 2021. However, the correlation coefficients (0.82 for SCA79) listed in Table 14indicate overall high predictive skill. The RMSE and MAE of SCA79 in the testing set are larger than those of other indices, although the prediction band (average value with 95% CI) usually encompasses the observations.

Table 14 Verification statistics for averaged sliding window NARMAX models for winter seasonal weather (validation period = 2011-2016; test period = 2017-2021).

	RMSE		MAE		Correlation (2011-2021)
	Validation	Testing	Validation	Testing	
winterNAO79	0.17	0.62	0.11	0.52	0.92
winterEA79	0.20	0.36	0.13	0.31	0.83
winterSCA79	0.37	0.79	0.24	0.65	0.82

5.3.3 Seasonal weather forecast by dynamical models

Figure 32 illustrate how SEAS5 fared when predicting the winter and high summer NAO, EA and SCA, using hindcast data for 1993 to 2016. There is a consistent tendency to under-predict the amplitude of the extreme seasons, to a greater extent than is observed for the NARMAX predictions. GloSea5 has higher skill than SEAS5 at predicting the East Atlantic pattern, and for example correctly predicted that the winter of 2004/05 would probably have a negative EA, although the observed outcome was more extreme than predicted by any of the ensemble members. Conversely, SEAS5 performs better with the NAO, although its ensemble did not catch the intensity of the negative NAO in the winter of 2009/10. SEAS5 particularly tended to underpredict the variability in the East Atlantic pattern in winter. For the Scandinavian pattern, again SEAS5 underpredicted the variability but performed better overall than GloSea5, which showed a slight negative correlation with the observed Scandinavian pattern.

In the case of high summer, covering July and August, skill is generally lower for the summer NAO than for the winter NAO, although both SEAS5 and GloSea5 showed a small positive correlation with the observed outcomes. GloSea5 performs better than SEAS5 at predicting the summer EA and SCA but again the overall correlations are low. It appears that high summer is the area where NARMAX may prove especially useful as a means of improving the reliability of our seasonal forecasts.

When June is considered, the dynamical models perform much better, reflecting their greater skill at one month out. SEAS5 shows particularly high skill with predicting the June East Atlantic pattern, with a correlation of 0.81, but both models have correlations of around or above 0.6 for the June NAO and EA. For the Scandinavian pattern, correlations are lower, with evidence of a lag developing for the more recent years, with the dynamical models' predictions of extreme Scandinavian pattern Junes being a month in arrears.

Reliability diagrams and Brier scores have been generated for winter and high summer. Generally, the results from the dynamical models show considerable divergence between the predicted probability of the EOFs being positive, negative or near neutral, and the observed probability. The Brier scores for the dynamical models are mostly around 0.2, compared with Brier scores of less than 0.1 for the predictions from NARMAX. The results for June (not shown) have closer correspondence between the predicted probability and the actual probability, but still with Brier scores consistently above 0.1.

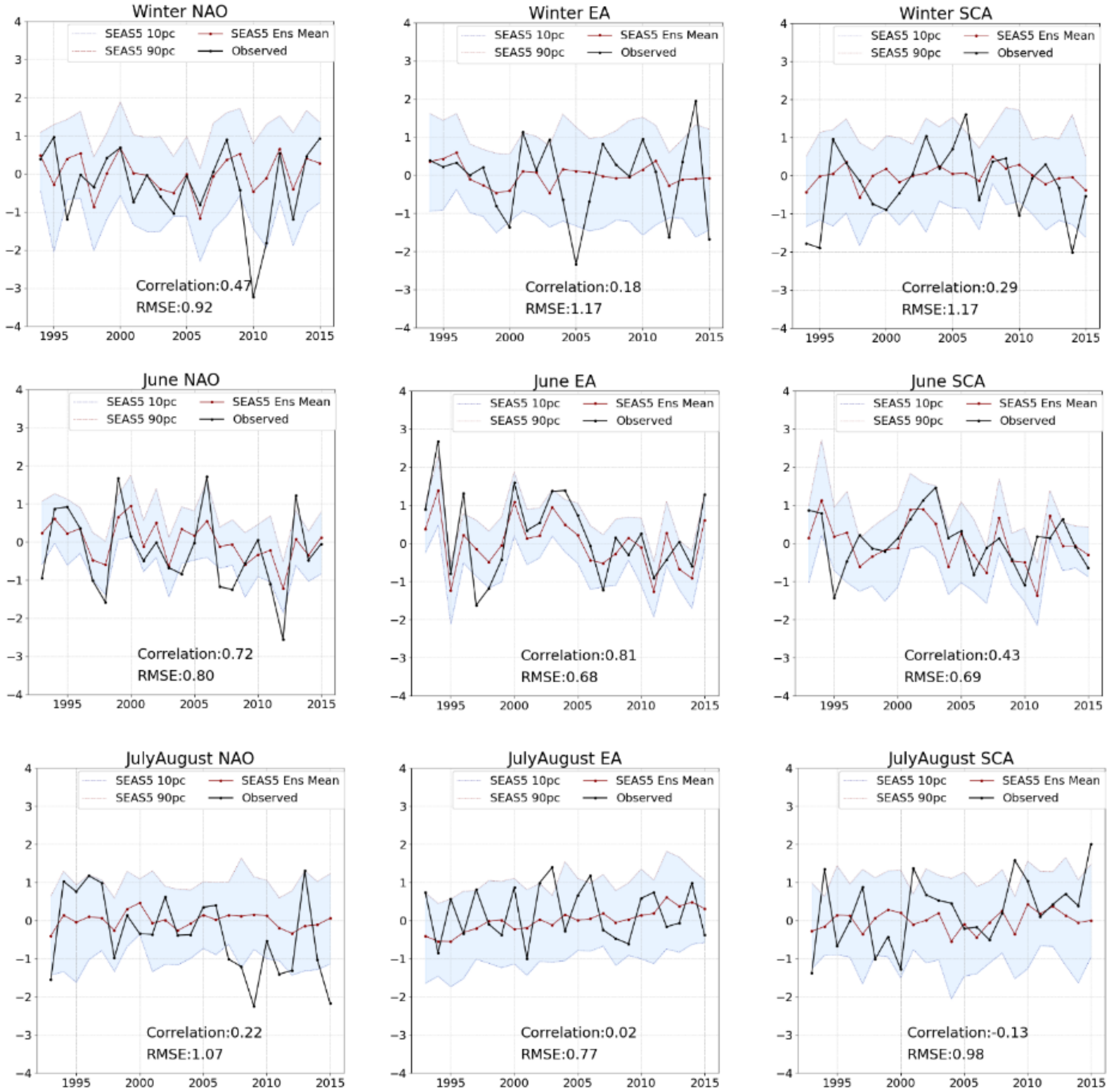


Figure 32 SEAS5 ensemble mean and 10th and 90th percentile predictions of NAO (left), EA (middle) and SCA (right), compared with observed values.

5.3.4 Comparison between dynamic models and sliding window NARMAX models

To more fully measure the performance of NARMAX models, this section presents a comparison of predictions by the SEAS5 and GloSea5 dynamical models with the sliding window NARMAX models discussed earlier. GloSea5 model output were not readily available for after 2017. Therefore, the following two training periods are used to develop NARMAX models: 1979 to 2000 (22 years) and 1979 to 2005 (27 years), where the out-of-sample periods are respectively 2001 to 2016 (16 years) and from 2006 to 2021 (16 years).

Table 15 Verification statistics comparing the dynamical model hindcasts with NARMAX forecasts for the period 2001-2016 (with NARMAX using the training period:1979-2000).

Index	Correlation with observed			RMSE		
	SEAS5	GloSea5	NARMAX	SEAS5	GloSea5	NARMAX
DJF_NAO	**0.59**	0.23	0.13	0.93	1.10	1.10
DJF_EA	-0.15	0.31	**0.55**	1.23	1.07	0.95

DJF_SCA	0.44	0.12	**0.84**	1.21	1.03	0.64
Jun_SNAO	**0.66**	**0.66**	0.46	0.79	0.81	0.90
Jun_EA	**0.78**	**0.71**	**0.81**	0.52	0.68	0.45
Jun_SCA	**0.55**	**0.63**	**0.50**	0.63	0.62	0.73
JA_SNAO	-0.08	0.09	0.25	1.16	1.15	1.22
JA_EA	0.00	0.41	**0.67**	0.73	0.67	0.58
JA_SCA	0.04	0.30	0.27	0.92	0.87	0.84

In Table 15 , the predictions by NARMAX are compared with the SEAS5 and GloSea5 models for winter and separately for June and for high summer (July and August). In most cases, NARMAX models show comparable skill to the dynamical models. For JunSNAO and JASNAO, NARMAX models produce accurate predictions in the early years from 2001 to 2010, while from 2011 onwards, NARMAX predictions are less accurate, mainly because the training period for NARMAX is too short with insufficient information to capture changes in seasonal North Atlantic atmospheric circulation. For JunEA and JAEA, NARMAX models show more skilful performance compared with dynamic models. For JunSCA and JASCA, NARMAX models show similar forecasting capability to dynamic models.

Another comparison between the SEAS5 dynamical model and NARMAX models with different training periods is shown in Table 16. The statistical results for these NARMAX models show higher correlation coefficients and smaller RMSE for all indexes, indicating a more skilful performance than the SEAS5 model. A comparison of the results presented in Table 15 and Table 16 indicates that having a relatively modest amount of extra information in the training period enables NARMAX to capture more details of North Atlantic atmospheric circulation and generate considerably better models.

Table 16 Verification statistics comparing the SEAS5 model hindcasts and forecasts with NARMAX, for the period 2006 to 2021.

Index	Correlation with observed		RMSE	
	SEAS5	NARMAX	SEAS5	NARMAX
DJF_NAO	**0.62**	**0.77**	0.99	0.79
DJF_EA	0.03	**0.77**	0.92	0.57
DJF_SCA	0.26	**0.74**	1.09	0.78
Jun_SNAO	**0.83**	**0.87**	1.36	0.58
Jun_EA	**0.79**	**0.88**	0.93	0.53
Jun_SCA	0.37	**0.83**	0.90	0.52
JA_SNAO	-0.08	**0.70**	0.80	0.82
JA_EA	0.43	**0.90**	0.64	0.40
JA_SCA	0.04	**0.72**	0.95	0.54

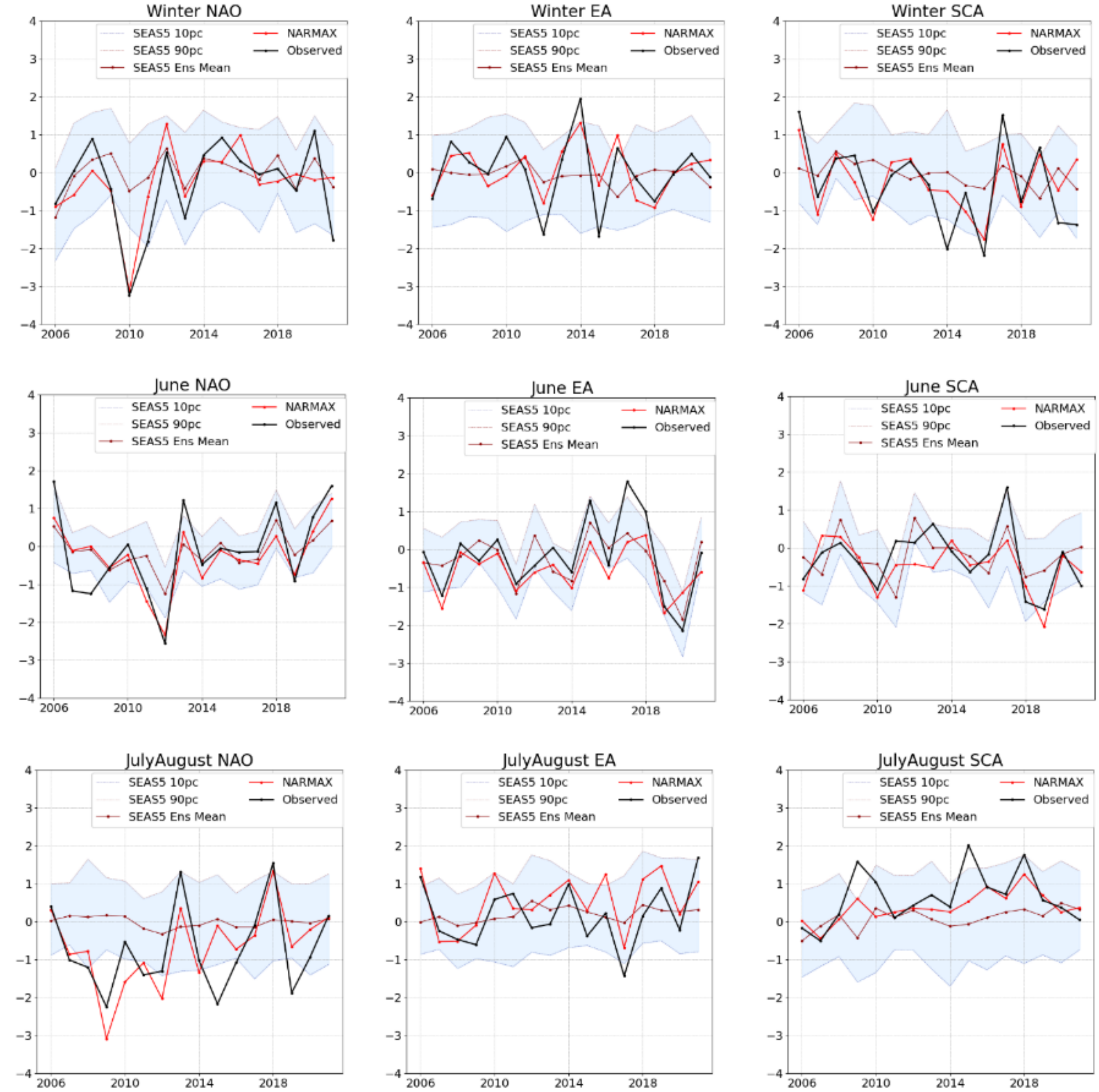


Figure 33 Comparisons between NARMAX predictions and SEAS5 hindcasts and forecasts for the period 2006-2021.

5.4 Summary

5.4.1 Analysis of the seasonal weather system based on sliding window NARMAX models

The results presented here highlight the potential for NARMAX to add considerable value to current dynamical model predictions of NAO, EA and SCA, especially in the case of summer, where dynamical models tend to struggle to a greater extent than for winter. The NAO alone only accounts for some of the variability in temperature and precipitation over north-west Europe, making it useful to predict other important modes of atmospheric circulation variability. For example, [239] attributed the exceptionally high rainfall of winter 2013/14 over much of the UK primarily to a strongly positive East Atlantic pattern. It is therefore encouraging that the NARMAX results are strongly correlated with the observations in the case of EA and SCA as well as NAO.

Splitting the summer into June and July/August shows that the dynamical models perform better at forecasting June NAO, EA and SCA from one month ahead than at forecasting July/August, as would be expected. The highest correlation with the observed data is 0.81, in the case of SEAS5 predictions of June EA (compared with correlations of 0.94 and 0.90 for the 1956 and 1979 NARMAX runs for June EA respectively). Surprisingly, the NARMAX verification rate is similar for June and for July/August, indicating that there is less of an advantage over the dynamical models at relatively short time range, but that the accuracy of NARMAX does not decline as much for two and three months ahead, at least in the case of summer.

Compared with the dynamical models, NARMAX predictions show a reduced "signal to noise" problem, i.e., the year-to-year variability of the NAO, EA and SCA is captured accurately, and while the amplitude of extreme events is at times underpredicted, it is generally underpredicted to a much smaller extent than is observed with the dynamical models. When analysing summer predictions for June and for July/August, it is clear that the "signal to noise" problem with the dynamical models increases markedly when predicting two and three months ahead as opposed to just one month ahead, at least in the case of summer. This is far less apparent with the NARMAX predictions, suggesting that NARMAX-assisted forecasts may be especially useful at reducing the "signal to noise" problem when forecasting further ahead. It is particularly encouraging that this appears to be true for summer, as the dynamical models tend to struggle more with seasonal predictions of summer than of winter.

The lists of predictors that the sliding window NARMAX chooses for summer are mixed, but some consistent results stand out. The 1956 model (see supplementary material) selects the Canadian Archipelago/Baffin Bay sea ice concentrations (ArB) as one of the top two predictors of the June and July/August NAO. Solar activity with a 2-year lead time is most often chosen for the July/August NAO. The 1979 model favours tropical precipitation for predicting the June NAO, and April North American snow cover is most often selected for predicting July/August NAO. For the July/August NAO, the 1956 model most often selects tropical rainfall, the East Siberian/Laptev Seas sea ice concentration is one of the two most selected predictors. ESL is also the most often selected predictor for June SCA, and the ArB is most often selected for July/August SCA.

The 1979 model has less of a strong tendency to select sea ice concentrations. The April North Atlantic Dipole pattern of sea surface temperature anomalies is most selected for predicting the summer EA. For June NAO, April tropical precipitation patterns make up the two most often selected predictors, while for July NAO, April North American snow cover is most often selected.

The recurrence of sea ice concentrations in the top ten predictors must be taken with some caution, for as [246] discussed, the recent sharp decline in sea ice concentrations could contribute to models overestimating the influence of sea ice and potentially issuing poorer forecasts for recent years. However, despite this issue, the NARMAX predictions consistently outperformed the dynamical models, especially in the case of high summer (July and August).

The 1979 model most often selects the September North Atlantic Horseshoe (NAH) to predict the winter NAO, and the 1956 model has the May NAH as the second most often selected predictor. This is a reassuring result, as it ties in well with the findings of [274], which identified plausible physical explanations for observed links between the NAH and the subsequent winter NAO, particularly in relation to the NAH during the preceding summer and autumn. The 1956 model most often selects October Barents-Kara Sea ice concentrations for the winter NAO.

The 1956 model also most often selects the May NAH in predicting the winter EA, but in the case of the 1979 model, the most commonly selected predictors are August Barents-Kara Sea sea ice concentrations for the winter EA, and October Labrador sea ice concentrations for the winter SCA.

[246] discussed links between solar activity (with a lead time of 6 months to 2 years) and the June tripole and the winter NAO. Neither of those were in the top ten predictors of the 1979 NARMAX model, although in the model from 1956, the April and July tripole both featured in the ten most frequently selected predictors. The October stratospheric polar vortex only appears in the top ten predictors of the 1956 model, which selects it second most in predicting winter EA and SCA.

Lagged teleconnection links between sea ice concentrations, sea surface temperature anomalies, tropical precipitation and subsequent atmospheric circulation patterns have already been found. For example, there may be links between Barents-Kara Sea ice concentrations and extratropical atmospheric circulation via complex teleconnections with the Aleutian low and tropical sea surface temperature and rainfall variations [275]. This also ties in well with the 1956 NARMAX model frequently choosing October Barents-Kara sea ice concentrations as a predictor of the winter NAO. There is also evidence for a link between tropical precipitation anomalies and wintertime European precipitation events [276] and, correspondingly, the East Atlantic Pattern [247].

5.4.2 Summary

These results demonstrate that NARMAX models have considerable potential to improve upon purely dynamical model based seasonal weather predictions, especially in the case of high summer (July and August), and therefore significantly extend the pilot study which focused on winter. It is important to provide NARMAX with a reasonably long training period (ideally upwards of 25 years) to ensure a higher level of accuracy. Links with individual indices that are frequently chosen by NARMAX are a basis for future work, both with the aim of evaluating the physical plausibility of the links identified by NARMAX, and for using NARMAX to assist the identification of new teleconnection links that have not previously been identified and explored. Follow-up work will downscale the three principal EOFs to determine the links between the EOF time series (both observed and predicted by NARMAX) and Northwest European temperatures and precipitation, including links with persistence and variability indices as well as maximum, minimum and mean values, that are relevant for end-users such as the agri-food, energy and tourism industries.

Chapter 6

Efficient Mask Attention-Based NARMAX (MAB-NARMAX) Model Identification

6.1 Introduction

Data-driven modelling and complex system identification has consistently been a powerful technique in analysing and investigating systems and their behaviours [1, 28]. Sparse, interpretable, and transparent (SIT) parametric models are usually desirable and useful for understanding the inherent dynamics and interactions of the system states. An effective representation of the SIT modelling method is the famous NARMAX (nonlinear autoregressive moving average with exogenous inputs) model [1]. Moreover, a NARMAX model is usually compact and clear to explain and reveal the model structure in many applications where the primary modelling objective and task are to exploit and obtain an insightful description of how the system output explicitly depends on the system inputs [2]-[5].

In NARMAX modelling, model structure detection and selection are critical procedures for efficient and effective model identification [1, 277][10]. The detection and decision of the model structure can be influenced by lots of factors of the signals from the system, like signal noise, the sampling rate of the signal, and the richness of the input signal. Other factors, which come from the modelling process, like the maximum lags in the system signals and the nonlinear degree for the polynomial model, are also crucially important for reliable system identification [277].

For convenience of description, consider the case of systems with single input and single output (SISO), denoted by $u(t)$ and, respectively, where *t* is the sampling index (time instant). The lagged input and output variables are defined as:

$$\begin{aligned} &u(t) \to u(t-d), u(t-d-1), ..., u(t-n_u) \\ &y(t) \to y(t-1), y(t-2), ..., y(t-n_y) \end{aligned} \tag{6.1}$$

where *d* is a time lag between the system input and output (usually $d=1$ but can be set to zero if the system input $u(t)$ instantly affects the system behaviour), n_u and n_y are the maximum time lags. These lagged variables can be used to create a model term dictionary, which can be used to build models. In doing so, a sparse learning algorithm, e.g., the orthogonal least squares with error reduction ratio (OLS-ERR) algorithm [14], the term clustering based algorithm [278], and random search approaches [279, 280], can be used to determine a set of best models.

It is important to define the model settings appropriately. Taking the choice of n_u and n_y as an example, if these two hyper-parameters are much smaller than that of the 'true' system model, then many important lagged variables will not get involved in model construction, implying that the resulting model term dictionary will only contain partial information of the underlying system. Consequently, the finally identified 'best' model(s) may not be able sufficiently to characterize the system input-output behaviour. This is a fundamental issue in all system identification and data-driven modelling practices and applications, no matter what kind of models (e.g., NARX/NARMAX, traditional neural networks, deep neural networks) are used. Therefore, in practice, n_u and n_y are usually set to be large enough to ensure that the dictionary is large enough for representing the input-output behaviour of the system of study. However, large n_u and n_y means that larger numbers of candidate model terms are included in the dictionary, many of which are irrelevant and not useful for characterizing the system dynamics but can only increase the difficulty in finding the best and most reliable models. So, it is always a challenge to properly define the model settings for transparent, interpretable, and parsimonious model identification of complex systems whose structure is completely unknown. Many methods and algorithms have been proposed for model structure determination in the literature. In [155], a novel mutual information-based integrated forward orthogonal search algorithm was proposed. This algorithm can comprehensively measure the contributions of the model terms for NARMAX model identification. In [281], an adjustable prediction error sum of squares (APRESS) was proposed for model structure detection. In [282], the performance of APRESS was compared with the Akaike information criterion (AIC) and Bayesian information criterion (BIC), and it was shown that APRESS outperforms AIC and BIC for robust and parsimonious model structure selection. In [283], an efficient alternative model structure selection algorithm using an exhaustive-like mechanism was proposed. These methods, as well as many other methods, were designed to prevent spurious model terms from being included in the final models; this is important for obtaining a set of best models that are not only transparent and parsimonious, but also have good generalization properties.

Recently, a novel attention-based neural network named Transformer has been widely used [284]. A particular layer, ‘mask’, in the multi-head self-attention of Transformer is utilized for specializing (rare information), syntactic (dependency syntax and significant relations) and window roles (size of the signals and relative position) [285, 286]. The structure and good properties of mask layer motivate us to introduce the concept and scheme of ‘mask’ into system identification for model structure selection.

This study aims to exploit the potential of the idea of mask layer or mask matrix used in Transformer and make use of it for better model identification of nonlinear dynamic systems. The work focuses on NARMAX model identification and the use of mask operations for model input variable and term selection. The performance of the proposed mask attention-based NARMAX (MAB-NARMAX) method is evaluated via three simulation case studies, and the experimental results confirm the good properties of the new method.

6.2 The Mask attention-based NARMAX

6.2.1 Mask layer

A mask matrix is such a square matrix whose entries are either zero or one, that is, $M \in \mathbb{R}^{T\times T}, M_{i,j} \in [0,1]$ [287]. A mask matrix can improve computational efficiency in complex neural networks [288]. Generally, mask matrices used in Transformer are defined in 2D space and use binary values to indicate which features (variables) or neurons are important and in favour of further processing in the next stage.

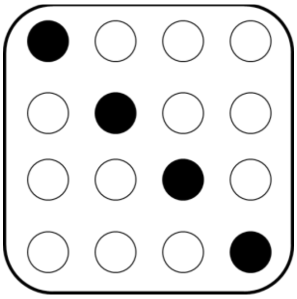


Figure 34 The diagonal two-dimensional mask matrix

The black dots in the diagonal mask matrix mean the value is 1, where $M_{j,j} = 1$, while the white circles (or disks) indicate that the value is 0, where $M_{i,j} = 0, i \neq j$. With the definition of the mask matrix [289], the self-attention in Transformer can be redefined as the mask-attention function:

$$\mathcal{A}_M(Q,K,V) = \mathcal{S}_M(Q,K)V \tag{6.2}$$

$$\mathcal{S}_M(Q,K) = \left[\frac{M_{i,j}\exp(Q_i K_j^T / \sqrt{d_k})}{\sum_k M_{i,k} exp(Q_i K_k^T / \sqrt{d_k})} \right] \tag{6.3}$$

where the queries Q , keys K and values $V \in \mathbb{R}^{T\times d_k}$ are three separate parts transformed from the input $X^{n\times d}$; $M \in \mathbb{R}^{T\times T}$, $M_{i,j} \in [0,1]$ is the mask matrix; $S(\cdot)$ is the softmax function; d_k is the dimension of K . For details about mask matrix and mask-attention function, interested readers are referred to [290, 291], where several mask generation methods are presented.

Note that variable and model term selection for NARMAX can be carried out in 1D space, where the model term index can be represented as a mask vector, as shown in (6.4).

Specifically, the mask vector is defined as:

$$M_{l\times 1} = \begin{cases} m_l = 0, if\ l = s \\ m_l = 1, if\ l = p \end{cases} \tag{6.4}$$

where $s, p = 1, 2, ..., L$, L is the total length of the mask vector; s and p are random indexes of the mask vector, representing the location of '0' in the mask while p is the position of '1'.In this study, a random sampling and random generation approach is used to generate mask matrices (vectors).

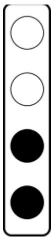


Figure 35 The proposed 1D mask matrix

6.2.2 NARMAX and polynomial NARX

The nonlinear autoregressive moving average with exogenous input (NARMAX) model is defined as [1]:

$$\begin{aligned} y(k) = F[&y(k-1),...,y(k-n_y), \\ &u(k-d),...,u(k-d-n_u), \\ &e(k-1),...,e(k-n_e)] + e(k) \end{aligned} \tag{6.5}$$

where, $y(k)$, $u(k)$ and $e(k)$ are the system output, input and noise sequences, respectively; n_y , n_u , and n_e are the maximum lags for the system output, input, and noise; $F[\cdot]$ is some nonlinear function, and d is a time delay, typically set to $d = 1$. The noise terms $e(k)$ are normally defined as the prediction errors. In practice, there are many types of model structures that can be used to approximate the unknown mapping $F[\cdot]$, including power-form polynomial models [29], rational models [30] , neural networks [31], fuzzy logic-based models [292], and wavelet expansions [79, 293, 294]. The most commonly used model is the power-form polynomial representation [1].

The polynomial NARX model is the special case of the polynomial NARMAX model, which does not include the noise-dependent or noise model terms [1]. NARX model is of the form:

$$y(k) = F[y(k-1),...,y(k-n_y), u(k-d),...,u(k-d-n_u)] + e(k) \tag{6.6}$$

Equation (6.6)**Error!** can usually be rearranged as:

$$y(k)=\theta_0+\sum_{i_i=1}^{n} f_{i_1}(x_i(k))+\sum_{i_1=1}^{n}\dots\sum_{i_\ell=i_{\ell-1}}^{n} f_{i_1 i_2\dots i_\ell}(x_{i_1}(k),x_{i_2}(k),\dots,x_{i_\ell}(k))+e(k) \tag{6.7}$$

where ℓ is the degree of polynomial nonlinearity, $\theta_{i_1 i_2\dots i_m}$ are model parameters, $n=n_y+n_u+n_e$, and

$$f_{i_1 i_2\dots i_m}(x_{i_1}(k),x_{i_2}(k),\dots,x_{i_m}(k))=\theta_{i_1 i_2\dots i_m}\prod_{p=1}^{m} x_{i_p}(k), 1\le m\le \ell \tag{6.8}$$

$$x_m(k)=\begin{cases} y(k-m), & 1\le m\le n_y \\ u(k-m+n_y), & n_y+1\le m\le n_y+n_u \end{cases} \tag{6.9}$$

More specifically, (3) can be explicitly written as:

$$y(k)=\theta_0+\sum_{i_1=1}^{n}\theta_{i_1}x_{i_1}(k)+\sum_{i_1=1}^{n}\dots\sum_{i_\ell=i_{\ell-1}}^{n}\theta_{i_1 i_2\dots i_\ell}x_{i_1}(k)x_{i_2}(k)\dots x_{i_\ell}(k)+e(k) \tag{6.10}$$

The degree of a multivariate polynomial is defined as the highest nonlinear order among all the model terms. Like the model $y(k)=0.5y(k-1)+u(k-1)+0.25y(k-1)u(k-2)$, the degree of nonlinearity is 2, that is, $\ell=2$.

6.2.3 MAB-NARMAX model

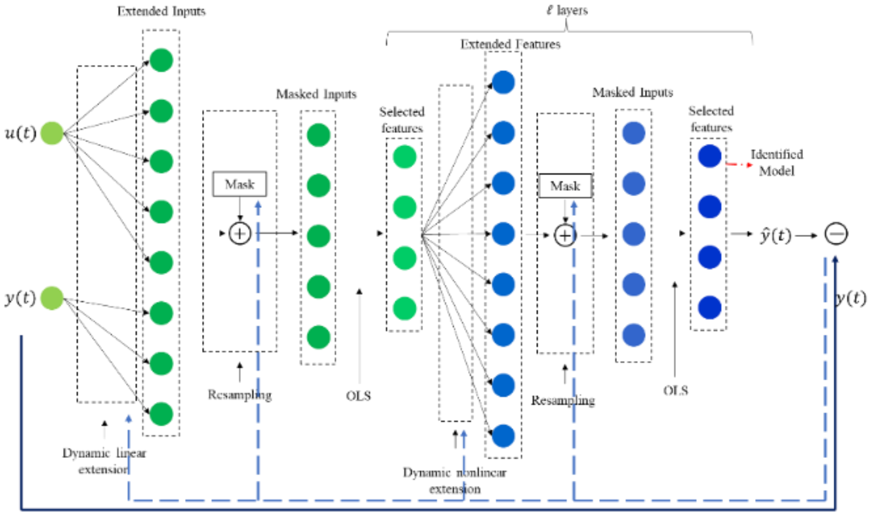


Figure 36 The structure and process of the proposed MAB-NARMAX method

The general process of the proposed MAB-NAMRAX is shown in Figure 36. On the basis of the mask vector and the polynomial NARX model defined by (3) and (9), respectively, we define a mask attention based NARX model as:

$$\hat{y}(k)=\theta_0+\sum_{i_1=1}^{n} m_{i_1}\cdot\theta_{i_1}x_{i_1}(k)+\sum_{i_1=1}^{n}\dots\sum_{i_\ell=i_{\ell-1}}^{n} m_{i_1 i_2\dots i_\ell}\cdot\theta_{i_1 i_2\dots i_\ell}x_{i_1}(k)\dots x_{i_\ell}(k)+e(k) \tag{6.11}$$

$m_p\in\{0,1\}^{n\times 1}, p\in\{i_1,i_1 i_2,\dots,i_1 i_2\dots i_n\}$ is a mask matrix with binary values in each mask layer. Equation (6.11) is the description of the output of the final layer in Figure 36. For each layer in the masked NARMAX model, the NARMAX maps the discrete terms to an output as:

$$S_{M_l}(X_l)=M_l\odot A(X_l,K_l) \tag{6.12}$$

$$A(X_l,K_l)=X_l\times K_l \tag{6.13}$$

$$X_{l+1} = F(S_{M_l}(X_l)) \tag{6.14}$$

where M_l is the mask matrix in the *l-th* layer, $S_{M_l}(X_l)$ is the regressors selected by the mask matrix in the *l-th* layer, $\odot$ means doc product between mask matrix and the model terms from former layer; X_l is the input signal of the *l-th* layer, K_l is the time lagged operation, $A(X_l,K_l)$ are the lagged variables of X_l; X_{l+1} are the selected model terms.

Then, the current best model is compared with the target signal to evaluate the performance of the identified NARMAX model. A backward-optimization algorithm is applied to adapt the values of the mask matrix for a better NARMAX model. Once the error of the identified model is below a specified threshold (tolerance), the MAB-NAMRAX modelling procedure will terminate.

6.3 Case study

To demonstrate the performance of the proposed method, three nonlinear systems with different noise are considered. For comparison purposes, LASSO [295] and LSTM [296] are also utilized to solve the same problems. To better evaluate the these methods, multi-step-ahead (long-term) prediction [297], rather than one-step-ahead prediction is considered.

The each of the three systems, a total number of 500 input-output data points were recorded, 70% of which (consisting of the first 350 points) was for training, 20% (consisting of the next 100 points) was for validating and the remaining 10% (consisting of the final 50 data points) was for test.

Following [155], the maximum time lags for the system signals in the proposed MAB-NARMAX were chosen to be 5, and the nonlinear degree was set to be 3. For comparison purposes, the input vector for the LASSO and LSTM models had the same predictors as those for the MAB-NARMAX. In this study, the configurations of the LSTM network were as follows:

- The number of maximum echoes M = 250.
- The number of neurons in the hidden layer N = 200.
- The initial learning rate $\lambda = 0.005$.
- The number of layers L = 4.

6.3.1 Example 1: A System Contaminated by White Noise

Consider a nonlinear system contaminated by white noise as below:

$$\begin{aligned} y(t) = &-0.6y(t-1) + 0.5u(t-2) - 0.2u(t-2)y(t-3) \\ &-0.25u(t-2)u(t-3) + e(t) \end{aligned} \tag{6.15}$$

where the input $u(t)$ is uniformly distributed on $[-1,1]$, while the noise $e(t)$ is Gaussian with zero mean and standard deviation of 0.025. It was assumed that the ‘true’ system model structure was not known. With these model settings, the proposed method was applied to these 500 data points.

The identified model structures by the two methods are listed in Table I, from which it can be seen that the proposed method correctly identified all the four ‘true’ model terms, whereas Lasso selected two more spurious linear and one more nonlinear model terms and LSTM could not generate a transparent model structure. The root means square error (RMSE) and mean absolute error (MAE) of the three models produced by MAB-NARMAX, LASSO and LSTM are shown in Table II.

Table 17 Identified model structure for Example 1

Index	Term	True model	Model by MAB-NARMAX	Model by LASSO
1	$u(t-1)$	×	×	-0.0064
2	$u(t-2)$	0.5	0.4989	0.4933
3	$u(t-3)$	×	×	-0.0054
4	$y(t-1)$	-0.6	-0.6022	-0.5873
5	$u(t-2)u(t-3)$	-0.25	-0.2476	-0.2373
6	$u(t-2)u(t-5)$	×	×	-0.0115
7	$u(t-2)y(t-3)$	-0.2	-0.2060	-0.1798

Table 18 Statistical criteria performance for Example 1

Method	Testing set	
	RMSE	MAE
MAB-NARMAX	**0.0209**	**0.0173**
LASSO	0.0224	0.0188
LSTM	0.1091	0.0544

Note that in Table 17 here (and in Table 20 and Table 21 in the following sections), all model terms selected by either MAB-NARMAX or Lasso are isted in the 2nd coclumn. The symbol 'X' indicates that the model term is either not in the true system model or not selected by MAB-NARMAX.

To graphically show the performance of all models on the testing dataset, a comparison between the model prediction and the measurement is shown in Figure 37. From which it can be observed that all models show excellent performance for the simulation example here, but a closer inspection can reveal that MAB-NARMAX outperforms LASSO and LSTM.

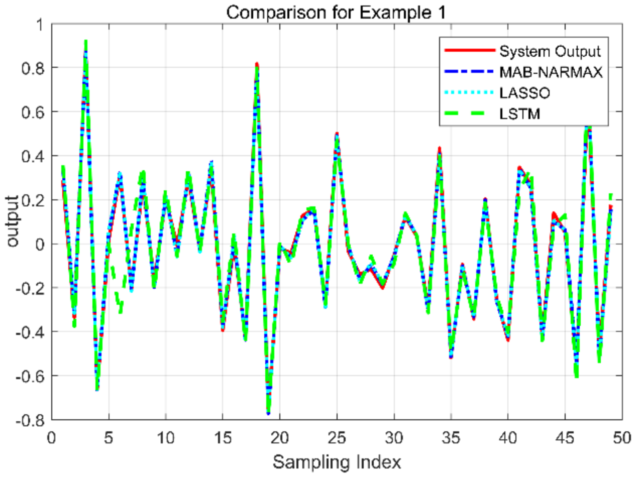


Figure 37 Comparisons between MAB-NARMAX, LASSO and LSTM for Example 1 on the training, validating, and testing sets

6.3.2 Example 2: A System with both Internal and Additive Noise

Consider the following system:

$$z(t) = -0.6z(t-1) - 0.2u(t-2)u(t-3) + 0.5u(t-2) - 0.25u(t-2)u(t-3) + e(t) \tag{6.16}$$

$$y(t) = z(t) + \zeta(t) \tag{6.17}$$

where the input is following the uniformly distribution on [-1,1], the internal noise $e(t) \sim N(0, 0.01^2)$ and the additive noise $\zeta(t) \sim N(0, 0.1^2)$.

Again, the ‘real’ model structure was set to be a black box. The identified model structures by MAB-NARMAX and LASSO are listed in Table 19, where the proposed method correctly identify all four ‘true’ model terms with almost the same parameters, however, Lasso selected a great number of spurious model terms. This results in that the parameters of the selected model terms by Lasso are quite different from the ‘true’ model structure. The values of RMSE and MAE of all models by are listed in Table 20. The predictions from all models on the testing dataset, are shown in (6.16). Again, a closer inspection can reveal that MAB-NARMAX shows better prediction performance.

Table 19 Identified model structure for Example 2

Index	**Term**	**True model**	**Model by MAB-NARMAX**	**LASSO Model**
1	$u(t-2)$	0.5	0.5258	0.5138
…	…	×	×	…
4	$y(t-1)$	-0.6	-0.5407	-0.1549
…	…	×	×	…
7	$u(t-2)u(t-3)$	-0.25	-0.2277	-0.2241
	…	×	×	…
9	$u(t-2)y(t-3)$	-0.2	-0.2157	-0.0388
	…	×	×	…
15	$y(t-3)y(t-5)$	×	×	-0.0110

Table 20 Statistical criteria performance for Example 2

Method	**Testing set**	
	RMSE	***MAE***
MAB-NARMAX	**0.1074**	**0.0840**
LASSO	0.1090	0.0861
LSTM	0.1857	0.1472

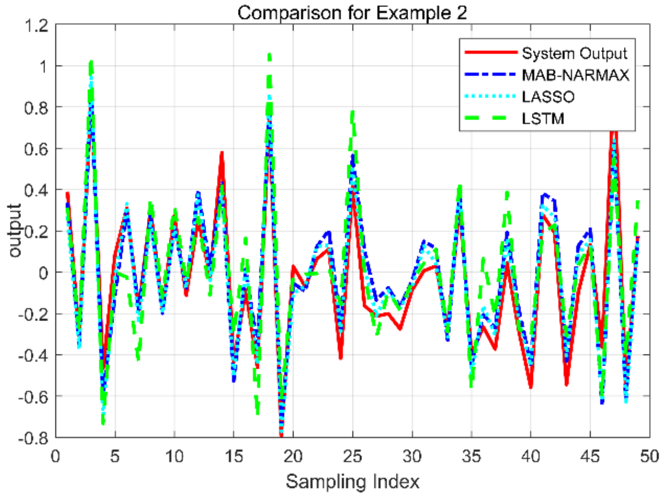


Figure 38 Comparisons between MAB-NARMAX, LASSO and LSTM for Example 2 on the training, validating, and testing sets

6.3.3 Example 3: the input is driven by a non-white noise

Consider the following nonlinear system:

$$w(t) = e(t) - 0.5e(t-1) \quad (6.18)$$

$$y(t) = -0.6y(t-1) - 0.2u(t-2)y(t-3) + 0.5u(t-2) - 0.25u(t-2)u(t-3) + w(t) \quad (6.19)$$

where the input u(t) is uniformly distributed on $[-1,1]$ and the internal noise $e(t) \sim N(0, 0.05^2)$ .Model structures identified by MAB-NARMAX and LASSO methods are listed in Table 21, in which the proposed method can precisely find all the four 'true' model terms, whereas Lasso selected many spurious model terms. The values of RMSE and MAE of all models are listed in

Table *22*.

The predictions from the two models on the whole dataset are shown in Figure 39. While all models give excellent predictions, a closer inspection can show that MAB-NARMAX performs better.

Table 21 Identified model structure for Example 3

Index	**Term**	**True model**	**MAB-NARMAX**	**LASSO Model**
1	$u(t-1)$	×	×	-0.0123
2	$u(t-2)$	0.5	0.5045	0.50119
…	…	×	×	…
Er5	$y(t-1)$	-0.6	-0.6150	-0.7023
…	…	×	×	…
9	$y(t-5)$	×	×	-0.0119
…	…	×	×	…
14	$u(t-2)u(t-3)$	-0.25	-0.2455	-0.2361
…	…	×	×	…
17	$u(t-2)y(t-3)$	-0.2	-0.2065	-0.1630
…	…	×	×	…
33	$y(t-4)^2$	×	×	0.0148

Table 22 Statistical criteria performance for Example 2

Method	**Testing set**	
	RMSE	***MAE***
MAB-NARMAX	**0.0489**	**0.0385**
LASSO	0.0489	0.0402
LSTM	0.2307	0.1493

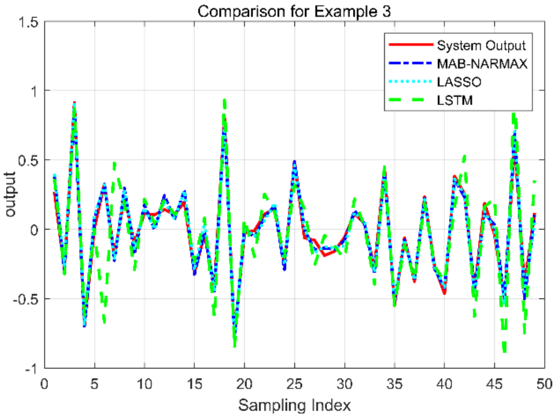


Figure 39 Comparisons between MAB-NARMAX, LASSO and LSTM for Example 3 on the training, validating, and testing sets

6.4 Summary

This study proposes a novel mask attention-based NARMAX (MAB-NARMAX) method for solving the model structure detection problem which is a fundamentally important and challenging task in most real applications. MAB-NARMAX makes use of the power and good properties of NARMAX, and the efficient computational and information processing capability of the mask matrix used in the Transformer neural network model. The mask matrix used in MAB-NARMAX is actually a mask vector; it has a simple structure but can help produce more precise, efficient and parsimonious models than the most commonly used state-of-*th*e-art methods – LASSO and LSTM for the multi-step-ahead (long-term) prediction, as shown by the three numerical examples. The application of MAB-NARMAX model is not limited to the polynomial NARX and NARMAX models. It can also be applied to any linear-in-*th*e-parameters modelling. Still, a further comprehensive study and investigation on the mask matrix and the backward optimization in MAB-NARMAX will be carried out in the future.

The limitations of this study lie in two aspects. First, we have only used mask matrix as feature selection. In the future, more investigations on the backward optimization scheme will be carried out. Second, our experimental studies focus on three numerical examples, but in the future, more analyses on real-world application will be investigated.

Chapter 7

Conclusions and Future Work

This section encapsulates the key findings and contributions of this study. The research commenced with an exploration of model uncertainty, both qualitatively and quantitatively, to afford a comprehensive understanding of model uncertainty in the context of NARMAX. Subsequently, three innovative methods were introduced: DeepNARMAX, Sliding Window NARMAX, and Mask Attention-based NARMAX, where model uncertainty was treated as a hyperparameter of the model. Lastly, the thesis outlined the study's limitations, future work directions, and concluding remarks.

7.1 Summary of the thesis work

This thesis explores model uncertainty in NARMAX modelling, a pervasive issue in system identification and modelling. The study addresses the absence of systematic research on model uncertainty in current literature, highlighting its impact on decision-making, private information, and complex dynamic systems. The research is also driven by a recent shift that treats model uncertainty not merely as a negative impact, but as a catalyst for probabilistic modelling and prediction.

In chapter 3, a systematic study about the model uncertainty in NARMAX modelling is present. NARMAX is lauded for its ability to accurately identify systems, construct clear and interpretable models, and provide precise system predictions. However, real system identification remains a challenge due to an insufficient understanding of the system. Therefore, identifying the key factors and their internal impacts on the system through modelling techniques becomes crucial. A central issue is dealing with model uncertainty, which arises from differing NARMAX hyperparameters configurations. Such uncertainty impacts the selection of optimal models and the production of predictive values for system responses. As a result, defining, analysing, and quantifying NARMAX model uncertainties are vital to mitigate or minimize their effects on the modelling process. Probabilistic models, which provide a broader set of information and valuable predictive values, have gained popularity in recent years. However, due to the NARMAX model's post-training structure, multiple NARMAX models are required to offer probability information, which in turn increases these models' inherent uncertainty.

The chapter starts with a brief introduction to NARMAX modelling and the Forward Regression Orthogonal Least Squares (FROLS) algorithm, focusing primarily on polynomial-based models due to their greater interpretability and stability. It also provides both qualitative and quantitative assessments of uncertainties relating to these models, defining them as 'NARMAX Model Uncertainty Sets' employing set theory. The observations made establish a direct correlation between the size of these uncertainty sets and the values assigned to their corresponding hyperparameters. The chapter concludes by defining the objective function related to these uncertainties and outlining the optimization processes for generating either singularly optimal or probabilistically optimal versions. This process is akin to obtaining the best possible versions within an acceptable range for those parameters. This comprehensive approach to understanding and handling model uncertainties augments the robustness of NARMAX modelling and offers a pathway to more effective and reliable system identification and prediction.

In chapter 4, an innovative method, the DeepNARMAX network modelling is proposed. DeepNARMAX network primarily consists of several layers including the input, linear transformation, driller, generator, model detection, and output layers, each catering to a specialized function towards generating a reduced, linearised, and optimal model. Yet, for optimal feature selection and maintaining a proportional balance in representation from differing systems, the thesis introduces a unique gate weight hyperparameter. This new hyperparameter, optimizable through the PSO-based method, describes the number of times each feature vector is selected during network training or the level of influence it exerts on the system. A series of comprehensive experiments affirm the efficiency of DeepNARMAX, showing that it retains similar interpretability and accuracy as compared to classic NARMAX. The DeepNARMAX network was extensively tested and validated using three simulated systems and two real-world power and weather forecasting models, proving its efficacy in dynamically complex situations. Its effectiveness is further demonstrated through comparison with other state-of-*th*e-art deep sequential neural networks, such as Informer and LSTM.

The main contribution of this work is the introduction of the deep polynomial NARMAX network and its gate weight hyperparameter. These innovations address dimension explosion issues encountered in traditional NARMAX modelling. The DeepNARMAX approach enhances analytical capabilities for high-dynamic scenarios while maintaining the basic pillars of traditional NARMAX model and effectively handling uncertainties during the modelling process. This new approach not only significantly enhances the applicability and efficiency of NARMAX models but also provides a more comprehensive understanding of complex systems.

In Chapter 5, a Sliding Window NARMAX (SW-NARMAX) method is introduced, promising to significantly enhance seasonal weather modelling and forecasting. The implementation of the SW-NARMAX model is shown to efficiently engage with system’s localized characteristics—the model segments the dataset using a concept of a sliding window and subsequently applies NARMAX modelling within this specific window to generate a cohesive NARMAX model ensemble for accurately representing the system. Validation of the SW-NARMAX method involves its application to seasonal weather forecasting, utilizing historical data spanning 43 years to perform modelling and prediction operations. The research results effectively substantiate the high performance and reliability of SW-NARMAX models in the accurate characterization and prediction of weather systems. Delving into the analytical processes, the results from the SW-NARMAX model reveal its potential to significantly augment current dynamical predictions, particularly for summer weather forecasts where traditional models face challenges. The NARMAX-derived models exhibit a reduced ‘signal to noise’ issue, demonstrating consistent accuracy in capturing year-to-year weather variability, even during extreme events.

A notable feature of SW-NARMAX models is their ability to manage data scalability while effectively focusing on specific localized system dynamics. The dynamic segmentation of the data using sliding windows allows the model to create localized NARMAX representations which, when combined, offer an overarching system perspective. This enhances the model’s adaptability in complex and changing environments. The predictors chosen by the SW-NARMAX models for weather forecasting show a promising correlation with known influential factors like sea ice concentrations and atmospheric circulation patterns. However, it’s also noted by researchers that the recent sharp decline in sea ice concentrations could lead to potential overestimations in sea ice influence, yet despite this issue, SW-NARMAX predictions consistently outperformed the dynamical models.

In chapter 6. a novel Mask Attention-based NARMAX (MAB-NARMAX) method to address the fundamentally important and challenging task of model structure detection, which is often an issue in real-world applications. MAB-NARMAX leverages the robust features of NARMAX as well as the computational efficiency and information processing abilities of the mask matrix employed in the

Transformer neural network model. Despite possessing a simple structure, the mask vector used in MAB-NARMAX ensures the generation of more precise, efficient, and parsimonious models compared to commonly used state-of-*th*e-art methods like LASSO and LSTM, particularly for the multi-step-ahead prediction. This superiority is evidenced through three numerical examples. While MAB-NARMAX's application is not confined to polynomial NARX and NARMAX models and could extend to any linear-in-*th*e-parameters modelling, future work will exhibit a comprehensive analysis of the mask matrix and the backward optimization in MAB-NARMAX.

7.2 Limitations and future work

This thesis primarily focuses on the exploration of uncertainty within NARMAX models. For the inaugural time, a systematic qualitative and quantitative analysis of uncertainty in NARMAX models was executed, embedding uncertainty as a fundamental factor within NARMAX modelling. We also introduced three novel modelling techniques rooted in NARMAX and applied them to several intricate dynamic nonlinear system models for validation analysis. The results compellingly showcased their feasibility and dependability.

However, our work has certain shortcomings and deficiencies that necessitate further research and enhancement in future studies.

Firstly, pertaining to the quantification of uncertainty in NARMAX models, we solely analyzed polynomial NARMAX model uncertainties quantitatively, bypassing other variants of NARMAX models such as wavelet variations or rational ones. Hence, additional research about the global model uncertainty is imperative.

Regarding our proposed DeepNARMAX model, which is established on polynomial based NARMX models; since these cannot accurately delineate certain non-linearities, it necessitates the incorporation of more non-linear functions to aptly illustrate systems. Therefore, it is necessary to develop a DeepNARMAX network that includes different non-linear functions for more complex system modelling.

In the context of the Sliding window NARMX model, this study exclusively investigated ensemble-NARMX modelling with varying structures but did not delve into ensemble-NARMX modelling with altering parameters. Practical avenues for future work involve combining structural uncertainties with parameter uncertainties for a comprehensive depiction and examination of systems.

With respect to the Mask-attention based NARMX model, its efficacy was confirmed using simulation models, but validating the model using real-world systems is suggested for future research. This will necessitate comparisons with other modelling methods to accomplish comprehensive verification.

7.3 Future development

7.3.1 Explainable AI for complex and dynamic systems

Explainable Artificial Intelligence (XAI) has gained considerable attention, reflecting a growing interest in understanding and interpreting the predictions of intricate computational models, particularly deep learning models. The complexity inherent in these models often results in a "black box" scenario, where the interpretability of the model becomes challenging. This lack of transparency becomes particularly acute in fields where there is an absence of 'ground truth', such as climate change and space weather forecasting. The deficiency of accurate ground truth data coupled with incomplete physical understanding presents obstacles in interpreting the predictions of the AI systems. Various contingent factors compound this issue, including the lack of comprehensible models or benchmarks for comparison which ultimately impede the evaluation of the AI system's explanatory capabilities.

NARMAX provides a robust approach to these interpretability challenges. These models, which may feature polynomial or other functional forms, enable straightforward interpretation, delivering insights into the dynamics of the system and the interrelation between variables. Furthermore, the NARMAX methodology includes systematic procedures for identifying the most significant terms within the model. This process assists in discerning which variables and interactions influence the system's behaviour most profoundly thereby helping in understanding complex system dynamics. Functioning beyond simple predictive capabilities, NARMAX models can describe the dynamics of a system in a manner that uncovers underlying mechanisms, even in situations where the physical principles are not entirely comprehensible.

Despite the promising potential of NARMAX models as an XAI tool, there remain several challenges related to their application to complex and dynamic systems that are without a well-understood physical basis. These challenges include:

- Model Complexity: The complexity of nonlinear systems can often result in a large number of terms in NARMAX models, which can render the interpretation of these models difficult.
- Data Requirements: The construction of accurate NARMAX models often demands extensive and high-quality data. The availability of such data can be a major challenge, especially in fields like space weather forecasting and climate change.
- Computational Demand: The tasks of structure detection, parameter estimation, and model validation can require considerable computational resources. This computational demand can be particularly problematic for large-scale problems.

The proposed NARMAX-based approaches show remarkable potential for explicitly modelling nonlinear dynamics, especially in contexts where traditional physical modelling might be arduous. As an XAI method, NARMAX-based models pose solutions to these challenges, turning them into valuable tools for extracting insights from complex and dynamic systems. By further developing and refining these techniques, it is predicted that they will increasingly contribute to a more transparent and comprehensible AI landscape.

7.3.2 Green AI

Artificial Intelligent Energy Efficiency, or "Green AI," represents an innovative direction in AI research and development. The primary focus is to meld methods for devising AI models that are environmentally amicable and exhibit high energy efficiency. "Green AI" underscores the importance of mitigating the excessive carbon footprint linked to the deployment and training of AI models, a challenge stemming from the considerable computational requirements of massive-scale deep learning models. The objective is to attain a balance between environmental impact and AI advances and operationality by optimising algorithms, more effective utilization of hardware, and considering the energy source in computational environments.

NARMAX based model functions as a valuable instrument for anticipating and modelling intricate nonlinear systems. The model is influential in scenarios where external factors and the system's past values are integral to its behaviour. Incorporating NARMAX based methods into Green AI entails harnessing the model's efficiency and interpretability to foster sustainable AI solutions. Here are ways NARMAX can contribute to fuelling Green AI:

- Efficiency: NARMAX based models can encapsulate the dynamics of complex systems without the extensive computational requirements inherent in deep learning models. Such efficiency is conducive to the Green AI's objective of reduced energy consumption.
- Interpretability: NARMAX based models, owing to their structured nature, grant a clear insight into how outputs are influenced by inputs. Such transparency lends credibility and accessibility to AI systems, feeding into XAI's element of Green AI by offering models that are environmentally friendly and interpretable.
- Adaptability: NARMAX based models can be adjusted to various domains, evidenced by their use in climate modelling and renewable energy forecasting - domains essential to understanding and mitigating environmental impacts. This adaptability ensures that Green AI initiatives can freely exploit the benefits of accurate and efficient modelling across diverse environmental contexts.
- Optimization: The parameters of NARMAX models can be further optimized for increased accuracy and efficiency, leading to sustainable AI solutions. Advanced optimization techniques can quell the computational load and energy consumption of model training and inference, thereby aligning with Green AI fundamentals.

By blending NARMAX based methods into Green AI's endeavour, researchers can create AI solutions that are efficient in modelling intricate systems, and at the same time, resonates with sustainability goals. Such amalgamation underscoring the potential for AI to wield a positive influence on environmental challenges whilst maintaining its commitment to transparency and interpretability.

Bibliography

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

[2] A. K. Tangirala, "Principles of system identification: theory and practice". Crc Press, 2018.

[3] D. N. Gujarati, "Linear regression : a mathematical introduction". Los Angeles, CA: Los Angeles, CA : SAGE Publications, Inc, 2019, 2019.

[4] D. M. Bates and D. G. Watts, "Nonlinear regression analysis and its applications". Chichester: Chichester : Wiley, c1988, 1988.

[5] P. Congdon, "Bayesian statistical modelling". Chichester: Chichester : Wiley, c2001, 2001.

[6] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1996. [Online]. Available: http://www.jstor.org/stable/2346178.

[7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[8] L. Ljung, "System identification," in Signal analysis and prediction: Springer, 1998, pp. 163-173.

[9] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," Automatica, vol. 50, no. 3, pp. 657-682, 2014/03/01/ 2014, doi: https://doi.org/10.1016/j.automatica.2014.01.001.

[10] T. Söderström and P. Stoica, "Instrumental variable methods for system identification," Circuits, Systems and Signal Processing, vol. 21, no. 1, pp. 1-9, 2002/01/01 2002, doi: 10.1007/BF01211647.

[11] Y. Gu, H.-L. Wei, R. J. Boynton, S. N. Walker, and M. A. Balikhin, "System Identification and Data-Driven Forecasting of AE Index and Prediction Uncertainty Analysis Using a New Cloud-NARX Model," Journal of Geophysical Research: Space Physics, vol. 124, no. 1, pp. 248-263, 2019, doi: https://doi.org/10.1029/2018JA025957.

[12] K.-Y. Lin, "User experience-based product design for smart production to empower industry 4.0 in the glass recycling circular economy," Computers & Industrial Engineering, vol. 125, pp. 729-738, 2018/11/01/ 2018, doi: https://doi.org/10.1016/j.cie.2018.06.023.

[13] C. McHugh, S. Coleman, and D. Kerr, "Hourly electricity price forecasting with NARMAX," Machine Learning with Applications, vol. 9, p. 100383, 2022/09/15/ 2022, doi: https://doi.org/10.1016/j.mlwa.2022.100383.

[14] R. J. Hall, H.-L. Wei, and E. Hanna, "Complex systems modelling for statistical forecasting of winter North Atlantic atmospheric variability: A new approach to North Atlantic seasonal forecasting," Quarterly Journal of the Royal Meteorological Society, vol. 145, no. 723, pp. 2568-2585, 2019, doi: https://doi.org/10.1002/qj.3579.

[15] M. Vaezi and A. Izadian, "Piecewise Affine System Identification of a Hydraulic Wind Power Transfer System," IEEE Transactions on Control Systems Technology, vol. 23, no. 6, pp. 2077-2086, 2015, doi: 10.1109/TCST.2015.2398311.

[16] M. Clyde and E. I. George, "Model Uncertainty," Statistical Science, vol. 19, no. 1, pp. 81-94, 14, 2004. [Online]. Available: https://doi.org/10.1214/088342304000000035.

[17] C. Young and K. Holsteen, "Model Uncertainty and Robustness:A Computational Framework for Multimodel Analysis," Sociological Methods & Research, vol. 46, no. 1, pp. 3-40, 2017, doi: 10.1177/0049124115610347.

[18] S. K. Jain and V. P. Singh, "Water resources systems planning and management [electronic resource]". Boston London: Boston London : Elsevier, 2003, 2003.

[19] J. C. Sadeghi, "Uncertainty Modelling for Scarce and Imprecise Data in Engineering Applications," Ph.D., The University of Liverpool (United Kingdom), England, 28179388, 2020.

[20] H. Wang, "Chapter 14 - Uncertainty quantification and minimization," in Computer Aided Chemical Engineering, vol. 45, T. Faravelli, F. Manenti, and E. Ranzi Eds.: Elsevier, 2019, pp. 723-762.

[21] M. G. Uddin, S. Nash, A. Rahman, and A. I. Olbert, "A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches," Water Research, vol. 229, p. 119422, 2023/02/01/ 2023, doi: https://doi.org/10.1016/j.watres.2022.119422.

[22] Q. Ren, H. Zhang, D. Zhang, X. Zhao, and X. Yu, "Enhancing Seismic Facies Classification Using Interpretable Feature Selection and Time Series Ensemble Learning Model with Uncertainty Assessment," IEEE Transactions on Geoscience and Remote Sensing, pp. 1-1, 2023, doi: 10.1109/TGRS.2023.3317983.

[23] C. Y. D. Yang, K. Ozbay, and X. Ban, "Developments in connected and automated vehicles," Journal of Intelligent Transportation Systems, vol. 21, no. 4, pp. 251-254, 2017/07/04 2017, doi: 10.1080/15472450.2017.1337974.

[24] Y. Gal, "Uncertainty in Deep Learning," Phd, University of Cambridge, 2016.

[25] D. Trafimow and M. Marks, "Editorial," Basic and Applied Social Psychology, vol. 37, no. 1, pp. 1-2, 2015/01/02 2015, doi: 10.1080/01973533.2015.1012991.

[26] R. Nuzzo, "Scientific method: Statistical errors," Nature, vol. 506, no. 7487, pp. 150-152, 2014/02/01 2014, doi: 10.1038/506150a.

[27] N. K. Ajami, Q. Duan, and S. Sorooshian, "An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction," Water Resources Research, vol. 43, no. 1, 2007, doi: https://doi.org/10.1029/2005WR004745.

[28] Y. Sun and H. L. Wei, "Efficient Mask Attention-Based NARMAX (MAB-NARMAX) Model Identification," in 2022 27th International Conference on Automation and Computing (ICAC), 1-3 Sept. 2022 2022, pp. 1-6, doi: 10.1109/ICAC55051.2022.9911110.

[29] S. Chen and S. A. Billings, "Representations of non-linear systems: the NARMAX model," International Journal of Control, vol. 49, no. 3, pp. 1013-1032, 1989/03/01 1989, doi: 10.1080/00207178908559683.

[30] Q. M. Zhu and S. A. Billings, "Identification of Polynomial & Rational Narmax Models," IFAC Proceedings Volumes, vol. 27, no. 8, pp. 259-264, 1994/07/01/ 1994, doi: https://doi.org/10.1016/S1474-6670(17)47725-1.

[31] Y. Gao and M. J. Er, "NARMAX time series model prediction: feedforward and recurrent fuzzy neural network approaches," Fuzzy Sets and Systems, vol. 150, no. 2, pp. 331-350, 2005/03/01/ 2005, doi: https://doi.org/10.1016/j.fss.2004.09.015.

[32] G.-X. Wen and Y.-J. Liu, "Adaptive fuzzy-neural tracking control for uncertain nonlinear discrete-time systems in the NARMAX form," Nonlinear Dynamics, vol. 66, no. 4, pp. 745-753, 2011/12/01 2011, doi: 10.1007/s11071-011-9947-z.

[33] S. A. Billings and W. Hua-Liang, "A new class of wavelet networks for nonlinear system identification," IEEE Transactions on Neural Networks, vol. 16, no. 4, pp. 862-874, 2005, doi: 10.1109/TNN.2005.849842.

[34] S. Billings and H. L. Wei, "NARMAX Model as a Sparse, Interpretable and Transparent Machine Learning Approach for Big Medical and Healthcare Data Analysis," in 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 10-12 Aug. 2019 2019, pp. 2743-2750, doi: 10.1109/HPCC/SmartCity/DSS.2019.00385.

[35] B. Zeng, J. Zhang, X. Yang, J. Wang, J. Dong, and Y. Zhang, "Integrated Planning for Transition to Low-Carbon Distribution System With Renewable Energy Generation and Demand Response," IEEE Transactions on Power Systems, vol. 29, no. 3, pp. 1153-1165, 2014, doi: 10.1109/TPWRS.2013.2291553.

[36] A. M. Nagy and V. Simon, "Survey on traffic prediction in smart cities," Pervasive and Mobile Computing, vol. 50, pp. 148-163, 2018/10/01/ 2018, doi: https://doi.org/10.1016/j.pmcj.2018.07.004.

[37] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, "Traffic Flow Prediction With Big Data: A Deep Learning Approach," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 865-873, 2015, doi: 10.1109/TITS.2014.2345663.

[38] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications," IEEE Transactions on Cybernetics, vol. 50, no. 9, pp. 3826-3839, 2020, doi: 10.1109/TCYB.2020.2977374.

[39] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," Nature Reviews Genetics, vol. 5, no. 2, pp. 101-113, 2004/02/01 2004, doi: 10.1038/nrg1272.

[40] R. Albert and A. L. Barabasi, "Statistical mechanics of complex networks," ArXiv, vol. cond-mat/0106096, 2001.

[41] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," Reviews of Modern Physics, vol. 81, no. 2, pp. 591-646, 05/11/ 2009, doi: 10.1103/RevModPhys.81.591.

[42] J. Foster, "From simplistic to complex systems in economics," Cambridge Journal of Economics, vol. 29, no. 6, pp. 873-892, 2005, doi: 10.1093/cje/bei083.

[43] S. A. Billings and S. Y. Fakhouri, "Identification of systems containing linear dynamic and static nonlinear elements," Automatica, vol. 18, no. 1, pp. 15-26, 1982/01/01/ 1982, doi: https://doi.org/10.1016/0005-1098(82)90022-X.

[44] S. A. Billings and I. J. Leontaritis, "Parameter Estimation Techniques for Nonlinear Systems," IFAC Proceedings Volumes, vol. 15, no. 4, pp. 505-510, 1982/06/01/ 1982, doi: https://doi.org/10.1016/S1474-6670(17)63039-8.

[45] S. A. Billings and W. S. F. Voon, "Least squares parameter estimation algorithms for non-linear systems," International Journal of Systems Science, vol. 15, no. 6, pp. 601-615, 1984/06/01 1984, doi: 10.1080/00207728408547198.

[46] I. J. Leontaritis and S. A. Billings, "Input-output parametric models for non-linear systems Part I: deterministic non-linear systems," International Journal of Control, vol. 41, no. 2, pp. 303-328, 1985/02/01 1985, doi: 10.1080/0020718508961129.

[47] I. J. Leontaritis and S. A. Billings, "Input-output parametric models for non-linear systems Part II: stochastic non-linear systems," International Journal of Control, vol. 41, no. 2, pp. 329-344, 1985/02/01 1985, doi: 10.1080/0020718508961130.

[48] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," International Journal of Control, vol. 50, no. 5, pp. 1873-1896, 1989/11/01 1989, doi: 10.1080/00207178908953472.

[49] S. A. Billings, S. Chen, and R. Backhouse, "The identification of linear and non-linear models of a turbocharged automotive diesel engine," Mechanical Systems and Signal Processing, vol. 3, pp. 123-142, 1989.

[50] S. Chen, S. A. Billings, C. F. N. Cowan, and P. M. Grant, "Practical identification of NARMAX models using radial basis functions," International Journal of Control, vol. 52, no. 6, pp. 1327-1350, 1990/12/01 1990, doi: 10.1080/00207179008953599.

[51] S. Chen, S. A. Billings, and P. M. Grant, "Non-linear system identification using neural networks," International Journal of Control, vol. 51, no. 6, pp. 1191-1214, 1990/01/01 1990, doi: 10.1080/00207179008934126.

[52] Y. Zhao, S. A. Billings, H. Wei, F. He, and P. G. Sarrigiannis, "A new NARX-based Granger linear and nonlinear casual influence detection method with applications to EEG data," Journal of Neuroscience Methods, vol. 212, no. 1, pp. 79-86, 2013/01/15/ 2013, doi: https://doi.org/10.1016/j.jneumeth.2012.09.019.

[53] E. H. K. Fung, Y. K. Wong, H. F. Ho, and M. P. Mignolet, "Modelling and prediction of machining errors using ARMAX and NARMAX structures," Applied Mathematical Modelling, vol. 27, no. 8, pp. 611-627, 2003/08/01/ 2003, doi: https://doi.org/10.1016/S0307-904X(03)00071-4.

[54] H. l. Wei and S. A. Billings, "Feature Subset Selection and Ranking for Data Dimensionality Reduction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 162-166, 2007, doi: 10.1109/TPAMI.2007.250607.

[55] "Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information," International Journal of Modelling, Identification and Control, vol. 3, no. 4, pp. 341-356, 2008, doi: 10.1504/ijmic.2008.020543.

[56] S. A. Billings and H. L. Wei, "Sparse Model Identification Using a Forward Orthogonal Regression Algorithm Aided by Mutual Information," IEEE Transactions on Neural Networks, vol. 18, no. 1, pp. 306-310, 2007, doi: 10.1109/TNN.2006.886356.

[57] H. L. Wei, S. A. Billings, and J. Liu, "Term and variable selection for non-linear system identification," International Journal of Control, vol. 77, pp. 110 - 86, 2004.

[58] S. A. Billings * and H. L. Wei, "The wavelet-NARMAX representation: A hybrid model structure combining polynomial models with multiresolution wavelet decompositions," International Journal of Systems Science, vol. 36, no. 3, pp. 137-152, 2005/02/20 2005, doi: 10.1080/00207720512331338120.

[59] Q. M. Zhu and A. Billings, "Parameter estimation for stochastic nonlinear rational models," International Journal of Control, vol. 57, no. 2, pp. 309-333, 1993/02/01 1993, doi: 10.1080/00207179308934390.

[60] Q. Zhu and S. Billings, "Identification of polynomial & rational NARMAX models," IFAC Proceedings Volumes, vol. 27, no. 8, pp. 259-264, 1994.

[61] S. Chen, X. Hong, and C. J. Harris, "Sparse multioutput radial basis function network construction using combined locally regularised orthogonal least square and D-optimality experimental design," IEE Proceedings - Control Theory and Applications, vol. 150, no. 2, pp. 139-146. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/ip-cta_20030253

[62] Z. H. Liu, H. L. Wei, Q. C. Zhong, K. Liu, X. S. Xiao, and L. H. Wu, "Parameter Estimation for VSI-Fed PMSM Based on a Dynamic PSO With Learning Strategies," IEEE Transactions on Power Electronics, vol. 32, no. 4, pp. 3154-3165, 2017, doi: 10.1109/TPEL.2016.2572186.

[63] Z. H. Liu, B. L. Lu, H. L. Wei, X. H. Li, and L. Chen, "Fault Diagnosis for Electromechanical Drivetrains Using a Joint Distribution Optimal Deep Domain Adaptation Approach," IEEE Sensors Journal, vol. 19, no. 24, pp. 12261-12270, 2019, doi: 10.1109/JSEN.2019.2939360.

[64] Y. Li, W. G. Cui, Y. Z. Guo, T. Huang, X. F. Yang, and H. L. Wei, "Time-Varying System Identification Using an Ultra-Orthogonal Forward Regression and Multiwavelet Basis Functions With Applications to EEG," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 7, pp. 2960-2972, 2018, doi: 10.1109/TNNLS.2017.2709910.

[65] Z. H. Liu, H. L. Wei, X. H. Li, K. Liu, and Q. C. Zhong, "Global Identification of Electrical and Mechanical Parameters in PMSM Drive Based on Dynamic Self-Learning PSO," IEEE Transactions on Power Electronics, vol. 33, no. 12, pp. 10858-10871, 2018, doi: 10.1109/TPEL.2018.2801331.

[66] Y. Sun and H. Wei, "How weather conditions affect the spread of Covid-19 : findings from a study using contrastive learning and NARMAX models," R. Jiang, D. Crookes, H. L. Wei, L. Zhang, and P. Chazot, Eds., ed: Taylor & Francis, 2021.

[67] "Identification of nonlinear systems with non-persistent excitation using an iterative forward orthogonal least squares regression algorithm," International Journal of Modelling, Identification and Control, vol. 23, no. 1, pp. 1-7, 2015, doi: 10.1504/ijmic.2015.067496.

[68] A. Senawi, H.-L. Wei, and S. A. Billings, "A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking," Pattern Recognition, vol. 67, pp. 47-61, 2017/07/01/ 2017, doi: https://doi.org/10.1016/j.patcog.2017.01.026.

[69] H. Chen, L. Huang, L. Yang, Y. Chen, and J. Huang, "Model-based method with nonlinear ultrasonic system identification for mechanical structural health assessment," Transactions on Emerging Telecommunications Technologies, vol. 31, no. 12, p. e3955, 2020, doi: https://doi.org/10.1002/ett.3955.

[70] Y. Yu, C. Zhang, and M. Zhou, "NARMAX Model-Based Hysteresis Modeling of Magnetic Shape Memory Alloy Actuators," IEEE Transactions on Nanotechnology, vol. 19, pp. 1-4, 2020, doi: 10.1109/TNANO.2019.2953933.

[71] Y. Gu, Y. Yang, J. P. A. Dewald, F. C. T. v. d. Helm, A. C. Schouten, and H. L. Wei, "Nonlinear Modeling of Cortical Responses to Mechanical Wrist Perturbations Using the NARMAX Method," IEEE Transactions on Biomedical Engineering, vol. 68, no. 3, pp. 948-958, 2021, doi: 10.1109/TBME.2020.3013545.

[72] S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO non-linear systems using a forward-regression orthogonal estimator," International Journal of Control, vol. 49, no. 6, pp. 2157-2189, 1989/06/01 1989, doi: 10.1080/00207178908559767.

[73] K. Ahmed, A. A. Shah, L. Wang, and S. Wang, "Modeling and identification of power forecasting scheme for real PV system using Grey box neural network based NARMAX model," in 2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 16-20 Aug. 2022 2022, pp. 562-567, doi: 10.1109/IBCAST54850.2022.9990342.

[74] G. P. Zhang and M. Qi, "Neural network forecasting for seasonal and trend time series," European Journal of Operational Research, vol. 160, no. 2, pp. 501-514, 2005/01/16/ 2005, doi: https://doi.org/10.1016/j.ejor.2003.08.037.

[75] S. L. Kukreja, H. L. Galiana, and R. E. Kearney, "NARMAX representation and identification of ankle dynamics," IEEE Transactions on Biomedical Engineering, vol. 50, no. 1, pp. 70-81, 2003, doi: 10.1109/TBME.2002.803507.

[76] S. A. Billings and Q. M. Zhu, "A structure detection algorithm for nonlinear dynamic rational models," International Journal of Control, vol. 59, no. 6, pp. 1439-1463, 1994/06/01 1994, doi: 10.1080/00207179408923140.

[77] H. Wei, "A wavelet-based approach for nonlinear system identification and non-stationary signal processing," 2004, 2004.

[78] C. K. Chui, "An introduction to wavelets". Boston, London: Boston, London : Academic Press, 1992, 1992.

[79] S. A. Billings* and H.-L. Wei, "The wavelet-NARMAX representation: A hybrid model structure combining polynomial models with multiresolution wavelet decompositions," International Journal of Systems Science, vol. 36, no. 3, pp. 137-152, 2005.

[80] H. L. Wei and S. A. Billings, "A unified wavelet-based modelling framework for non-linear system identification: the WANARX model structure," International Journal of Control, vol. 77, pp. 351 - 366, 2004.

[81] S. A. Billings and W. Hua-Liang, "A new class of wavelet networks for nonlinear system identification," IEEE Trans Neural Netw, vol. 16, no. 4, pp. 862-874, 2005, doi: 10.1109/TNN.2005.849842.

[82] H. L. Wei and S. A. Billings, "Long term prediction of non-linear time series using multiresolution wavelet models," International journal of control, vol. 79, no. 6, pp. 569-580, 2006, doi: 10.1080/00207170600621447.

[83] H. L. Wei, S. A. Billings, and M. A. Balikhin, "Wavelet based non-parametric NARX models for nonlinear input–output system identification," International Journal of Systems Science, vol. 37, pp. 1089 - 1096, 2006.

[84] H.-L. Wei and S. A. Billings, "Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information," Int. J. Model. Identif. Control., vol. 3, pp. 341-356, 2008.

[85] H. L. Wei and S. A. Billings, "An efficient nonlinear cardinal B-spline model for high tide forecasts at the Venice Lagoon," Nonlin. Processes Geophys., vol. 13, no. 5, pp. 577-584, 2006, doi: 10.5194/npg-13-577-2006.

[86] S. A. Billings, H.-L. Wei, and M. A. Balikhin, "Generalized multiscale radial basis function networks," Neural Netw, vol. 20, no. 10, pp. 1081-1094, 2007, doi: 10.1016/j.neunet.2007.09.017.

[87] J. X. Peng, K. Li, and G. W. Irwin, "A Novel Continuous Forward Algorithm for RBF Neural Modelling," IEEE Transactions on Automatic Control, vol. 52, pp. 117-122, 2007.

[88] H. Guang-Bin, P. Saratchandran, and N. Sundararajan, "A generalized growing and pruning RBF (GGAP-RBF) neural network for function approximation," IEEE Trans Neural Netw, vol. 16, no. 1, pp. 57-67, 2005, doi: 10.1109/TNN.2004.836241.

[89] C. Sheng, H. Xia, C. J. Harris, and P. M. Sharkey, "Sparse modeling using orthogonal forward regression with PRESS statistic and regularization," IEEE Trans Syst Man Cybern B Cybern, vol. 34, no. 2, pp. 898-911, 2004, doi: 10.1109/TSMCB.2003.817107.

[90] L. Zhang, K. Li, H. He, and G. W. Irwin, "A New Discrete-Continuous Algorithm for Radial Basis Function Networks Construction," IEEE Trans Neural Netw Learn Syst, vol. 24, no. 11, pp. 1785-1798, 2013, doi: 10.1109/TNNLS.2013.2264292.

[91] H. Asato, K. Yamashita, and H. Miyagi, "A method for parameter estimation in the NARMAX model with ARCH errors by RBF networks," in IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028), 12-15 Oct. 1999 1999, vol. 5, pp. 425-428 vol.5, doi: 10.1109/ICSMC.1999.815588.

[92] S. A. Billings, M. J. Korenberg, and S. Chen, "Identification of non-linear output-affine systems using an orthogonal least-squares algorithm," International Journal of Systems Science, vol. 19, no. 8, pp. 1559-1568, 1988/01/01 1988, doi: 10.1080/00207728808964057.

[93] M. Korenberg, S. A. Billings, Y. P. Liu, and P. J. McIlroy, "Orthogonal parameter estimation algorithm for non-linear stochastic systems," International Journal of Control, vol. 48, no. 1, pp. 193-210, 1988/07/01 1988, doi: 10.1080/00207178808906169.

[94] M. Kubat, "Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7," The Knowledge Engineering Review, vol. 13, no. 4, pp. 409-412, 1999, doi: 10.1017/S0269888998214044.

[95] X. Hong, M. Brown, S. Chen, and C. J. Harris, "Sparse model identification using orthogonal forward regression with basis pursuit and D-optimality," IEE Proceedings - Control Theory and Applications, vol. 151, no. 4, pp. 491-498. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/ip-cta_20040693

[96] D.-S. Huang and W.-B. Zhao, "Determining the centers of radial basis probabilistic neural networks by recursive orthogonal least square algorithms," Applied Mathematics and Computation, vol. 162, no. 1, pp. 461-473, 2005/03/04/ 2005, doi: https://doi.org/10.1016/j.amc.2003.12.105.

[97] S. Mukhopadhyay, S. Satpathi, and M. Chakraborty, "A Low Complexity Orthogonal Least Squares Algorithm for Sparse Signal Recovery," in 2018 International Conference on Signal Processing and Communications (SPCOM), 16-19 July 2018 2018, pp. 75-79, doi: 10.1109/SPCOM.2018.8724462.

[98] X. Liu and Y. Liu, "A Greedy Orthogonal Least Squares Algorithm for Nonlinear Systems with Time-Delays," in 2023 42nd Chinese Control Conference (CCC), 24-26 July 2023 2023, pp. 1436-1441, doi: 10.23919/CCC58697.2023.10240249.

[99] A. Hashemi and H. Vikalo, "Accelerated orthogonal least-squares for large-scale sparse reconstruction," Digital Signal Processing, vol. 82, pp. 91-105, 2018/11/01/ 2018, doi: https://doi.org/10.1016/j.dsp.2018.07.010.

[100] K. B. Laskey, "Model uncertainty: theory and practical implications," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 26, no. 3, pp. 340-348, 1996, doi: 10.1109/3468.487959.

[101] K. P. Burnham and D. R. Anderson, "Model selection and multi-model inference [electronic resource] : a practical information-theoretic approach", 2nd ed. ed. 2002.

[102] Y.-J. Chang, J. Brodziak, J. O'Malley, H.-H. Lee, G. DiNardo, and C.-L. Sun, "Model selection and multi-model inference for Bayesian surplus production models: A case study for Pacific blue and striped marlin," Fisheries research, vol. 166, pp. 129-139, 2015, doi: 10.1016/j.fishres.2014.08.023.

[103] I. Osband, J. Aslanides, and A. Cassirer, "Randomized Prior Functions for Deep Reinforcement Learning," ed. Ithaca: Ithaca: Cornell University Library, arXiv.org, 2018.

[104] M. C. Kennedy, "Chapter 42 - Exposure assessment: modeling approaches including probabilistic methods, uncertainty analysis, and aggregate exposure from multiple sources," in Present Knowledge in Food Safety, M. E. Knowles, L. E. Anelich, A. R. Boobis, and B. Popping Eds.: Academic Press, 2023, pp. 614-632.

[105] A. K. Hajdasinski, P. Eykhoff, A. A. H. Damen, and A. J. W. van den Boom, "The Choice and Use of Different Model Sets for System Identification," IFAC Proceedings Volumes, vol. 15, no. 4, pp. 47-55, 1982/06/01/ 1982, doi: https://doi.org/10.1016/S1474-6670(17)62963-X.

[106] Z. Zhang and E. Sejdić, "Radiological images and machine learning: Trends, perspectives, and prospects," (in eng), Comput Biol Med, vol. 108, pp. 354-370, May 2019, doi: 10.1016/j.compbiomed.2019.02.017.

[107] L. Loewe and W. G. Hill, "The population genetics of mutations: good, bad and indifferent," (in eng), Philos Trans R Soc Lond B Biol Sci, vol. 365, no. 1544, pp. 1153-67, Apr 27 2010, doi: 10.1098/rstb.2009.0317.

[108] J. Ding, V. Tarokh, and Y. Yang, "Model Selection Techniques: An Overview," IEEE Signal Processing Magazine, vol. 35, no. 6, pp. 16-34, 2018, doi: 10.1109/MSP.2018.2867638.

[109] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," Journal of Econometrics, vol. 187, no. 1, pp. 95-112, 2015/07/01/ 2015, doi: https://doi.org/10.1016/j.jeconom.2015.02.006.

[110] H. Heesterbeek *et al.*, "Modeling infectious disease dynamics in the complex landscape of global health," (in eng), Science, vol. 347, no. 6227, p. aaa4339, Mar 13 2015, doi: 10.1126/science.aaa4339.

[111] Y. Ovadia *et al.*, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in Proceedings of the 33rd International Conference on Neural Information Processing Systems: Curran Associates Inc., 2019, p. Article 1254.

[112] A. Malinin and M. J. F. Gales, "Predictive Uncertainty Estimation via Prior Networks," in Neural Information Processing Systems, 2018.

[113] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight Uncertainty in Neural Network," presented at the Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research, 2015. [Online]. Available: https://proceedings.mlr.press/v37/blundell15.html.

[114] T. Ekström, S. Burke, M. Wiktorsson, S. Hassanie, L.-E. Harderup, and J. Arfvidsson, "Evaluating the impact of data quality on the accuracy of the predicted energy performance for a fixed building design using probabilistic energy performance simulations and uncertainty analysis," Energy and Buildings, vol. 249, p. 111205, 2021/10/15/ 2021, doi: https://doi.org/10.1016/j.enbuild.2021.111205.

[115] H.-J. von Martens, "Evaluation of uncertainty in measurements—problems and tools," Optics and Lasers in Engineering, vol. 38, no. 3, pp. 185-206, 2002/09/01/ 2002, doi: https://doi.org/10.1016/S0143-8166(02)00010-6.

[116] D. K. Robinson, "Data reduction and error analysis for the physical sciences", 2nd ed. ed. New York: New York : McGraw-Hill, 1992, 1992.

[117] D. S. Moore and G. P. McCabe, "Introduction to the practice of statistics" (Introduction to the practice of statistics.). New York, NY, US: W H Freeman/Times Books/ Henry Holt & Co, 1989, pp. xix, 790-xix, 790.

[118] D. C. Montgomery and G. C. Runger, "Applied statistics and probability for engineers", Seventh edition.; EMEA edition. ed. Hoboken, NJ: Hoboken, NJ : Wiley, 2018, 2018.

[119] P. Newbold, W. L. Carlson, and B. Thorne, "Statistics for business and economics", Ninth, global edition. ed. Harlow, England: Harlow, England : Pearson, 2020, 2020.

[120] R. C. Spear, T. M. Grieb, and N. Shang, "Parameter uncertainty and interaction in complex environmental models," Water Resources Research, vol. 30, no. 11, pp. 3159-3169, 1994, doi: https://doi.org/10.1029/94WR01732.

[121] G. Nolet, "Inverse problem theory, methods for data fitting and model parameter estimation," Physics of the earth and planetary interiors, vol. 57, no. 3-4, pp. 350-351, 1989, doi: 10.1016/0031-9201(89)90124-6.

[122] Z. Xi, "Model-Based Reliability Analysis With Both Model Uncertainty and Parameter Uncertainty," Journal of Mechanical Design, vol. 141, no. 5, 2019, doi: 10.1115/1.4041946.

[123] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, "Bayesian data analysis", Third edition. ed. Boca Raton: Boca Raton : CRC Press, Taylor & Francis Group, an informa business, 2014, 2014.

[124] F. E. Harrell, "Regression modeling strategies : with applications to linear models, logistic and ordinal regression and survival analysis", Second edition. ed. Cham: Cham : Springer, 2015, 2015.

[125] D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye, "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors," Statistical Science, vol. 32, no. 1, pp. 1-28, 28, 2017. [Online]. Available: https://doi.org/10.1214/16-STS576.

[126] D. C. Montgomery, C. L. Jennings, and M. Kulahci, "Introduction to time series analysis and forecasting", Second edition. ed. Hoboken, New Jersey: Hoboken, New Jersey : Wiley, 2015, 2015.

[127] F. Kwasniok, "Predicting critical transitions in dynamical systems from time series using nonstationary probability density modeling," Phys Rev E Stat Nonlin Soft Matter Phys, vol. 88, no. 5, pp. 052917-052917, 2013, doi: 10.1103/PhysRevE.88.052917.

[128] C. Cheng *et al.*, "Time series forecasting for nonlinear and non-stationary processes: a review and comparative study," IIE Transactions, vol. 47, no. 10, pp. 1053-1071, 2015/10/03 2015, doi: 10.1080/0740817X.2014.999180.

[129] M. Saifuddin, R. Z. Abramoff, E. A. Davidson, M. C. Dietze, and A. C. Finzi, "Identifying Data Needed to Reduce Parameter Uncertainty in a Coupled Microbial Soil C and N Decomposition Model," Journal of Geophysical Research: Biogeosciences, vol. 126, no. 12, p. e2021JG006593, 2021, doi: https://doi.org/10.1029/2021JG006593.

[130] F. Khorashadi Zadeh, J. Nossent, B. T. Woldegiorgis, W. Bauwens, and A. van Griensven, "Impact of measurement error and limited data frequency on parameter estimation and uncertainty quantification," Environmental Modelling & Software, vol. 118, pp. 35-47, 2019/08/01/ 2019, doi: https://doi.org/10.1016/j.envsoft.2019.03.022.

[131] S. Lahlou *et al.*, "DEUP: Direct Epistemic Uncertainty Prediction," 2021, doi: 10.48550/arxiv.2102.08501.

[132] Y. Kato, D. M. J. Tax, and M. Loog, "A view on model misspecification in uncertainty quantification," 2022, doi: 10.48550/arxiv.2210.16938.

[133] R. F. Engle, "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," Econometrica, vol. 50, no. 4, pp. 987-1007, 1982, doi: 10.2307/1912773.

[134] H. White, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," Econometrica, vol. 48, no. 4, pp. 817-838, 1980, doi: 10.2307/1912934.

[135] S. T. Mustafa, J. Nossent, G. Ghysels, and M. Huysmans, "Estimation and Impact Assessment of Input and Parameter Uncertainty in Predicting Groundwater Flow With a Fully Distributed Model," Water resources research, vol. 54, no. 9, pp. 6585-6608, 2018, doi: 10.1029/2017WR021857.

[136] R. C. Hill, "Messy data [electronic resource] : missing observations, outliers, and mixed-frequency data". Bingley, U.K.: Place of publication not identified Emerald Group Publishing Limited, 1999, 1999.

[137] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A Review on Outlier/Anomaly Detection in Time Series Data," ACM Comput. Surv., vol. 54, no. 3, p. Article 56, 2021, doi: 10.1145/3444690.

[138] S. Wang and Z. S. Ye, "Distributionally Robust State Estimation for Linear Systems Subject to Uncertainty and Outlier," IEEE Transactions on Signal Processing, vol. 70, pp. 452-467, 2022, doi: 10.1109/TSP.2021.3136804.

[139] K. P. Vatcheva, M. Lee, J. B. McCormick, and M. H. Rahbar, "Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies," (in eng), Epidemiology (Sunnyvale), vol. 6, no. 2, Apr 2016, doi: 10.4172/2161-1165.1000227.

[140] J. I. Daoud, "Multicollinearity and Regression Analysis," J. Phys.: Conf. Ser, vol. 949, no. 1, p. 12009, 2017, doi: 10.1088/1742-6596/949/1/012009.

[141] W. E. Walker *et al.*, "Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support," Integrated assessment, vol. 4, no. 1, pp. 5-17, 2003, doi: 10.1076/iaij.4.1.5.16466.

[142] H. V. Gupta, M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye, "Towards a comprehensive assessment of model structural adequacy," Water Resour. Res, vol. 48, no. 8, pp. np-n/a, 2012, doi: 10.1029/2011WR011044.

[143] H. R. Maier, J. H. A. Guillaume, H. van Delden, G. A. Riddell, M. Haasnoot, and J. H. Kwakkel, "An uncertain future, deep uncertainty, scenarios, robustness and adaptation: How do they fit together?," Environmental Modelling & Software, vol. 81, pp. 154-164, 2016/07/01/ 2016, doi: https://doi.org/10.1016/j.envsoft.2016.03.014.

[144] E. Zio, "Some Challenges and Opportunities in Reliability Engineering," IEEE Transactions on Reliability, vol. 65, no. 4, pp. 1769-1782, 2016, doi: 10.1109/TR.2016.2591504.

[145] J. C. Helton, "Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty," Journal of Statistical Computation and Simulation, vol. 57, no. 1-4, pp. 3-76, 1997/04/01 1997, doi: 10.1080/00949659708811803.

[146] B. J. Minsley, N. L. Foks, and P. A. Bedrosian, "Quantifying model structural uncertainty using airborne electromagnetic data," Geophysical Journal International, vol. 224, no. 1, pp. 590-607, 2020, doi: 10.1093/gji/ggaa393.

[147] V.-P. Parkkinen and M. Baumgartner, "Robustness and Model Selection in Configurational Causal Modeling," Sociological Methods & Research, vol. 52, no. 1, pp. 176-208, 2023, doi: 10.1177/0049124120986200.

[148] F. Cribari-Neto, V. T. Scher, and F. M. Bayer, "Beta autoregressive moving average model selection with application to modeling and forecasting stored hydroelectric energy," International Journal of Forecasting, vol. 39, no. 1, pp. 98-109, 2023/01/01/ 2023, doi: https://doi.org/10.1016/j.ijforecast.2021.09.004.

[149] H. He *et al.*, "Learning to Select External Knowledge with Multi-Scale Negative Sampling," IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 1-7, 2023, doi: 10.1109/TASLP.2023.3301222.

[150] T. Kollmann and A. Kuckertz, "Evaluation uncertainty of venture capitalists' investment criteria," Journal of Business Research, vol. 63, no. 7, pp. 741-747, 2010/07/01/ 2010, doi: https://doi.org/10.1016/j.jbusres.2009.06.004.

[151] I. N. Durbach and T. J. Stewart, "Modeling uncertainty in multi-criteria decision analysis," European Journal of Operational Research, vol. 223, no. 1, pp. 1-14, 2012/11/16/ 2012, doi: https://doi.org/10.1016/j.ejor.2012.04.038.

[152] F. Sitorus and P. R. Brito-Parada, "A multiple criteria decision making method to weight the sustainability criteria of renewable energy technologies under uncertainty," Renewable and Sustainable Energy Reviews, vol. 127, p. 109891, 2020/07/01/ 2020, doi: https://doi.org/10.1016/j.rser.2020.109891.

[153] R. d'Amore-Domenech, Ó. Santiago, and T. J. Leo, "Multicriteria analysis of seawater electrolysis technologies for green hydrogen production at sea," Renewable and Sustainable Energy Reviews, vol. 133, p. 110166, 2020/11/01/ 2020, doi: https://doi.org/10.1016/j.rser.2020.110166.

[154] R. Rout and B. Subudhi, "Inverse optimal self-tuning PID control design for an autonomous underwater vehicle," International Journal of Systems Science, vol. 48, no. 2, pp. 367-375, 2017/01/25 2017, doi: 10.1080/00207721.2016.1186238.

[155] H.-L. Wei and S. A. Billings, "Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information," International Journal of Modelling, Identification and Control, vol. 3, no. 4, pp. 341-356, 2008.

[156] Y. Gu, H.-L. Wei, and M. M. Balikhin, "Nonlinear predictive model selection and model averaging using information criteria," Systems Science & Control Engineering, vol. 6, no. 1, pp. 319-328, 2018.

[157] O. S. Tătaru *et al.*, "Artificial intelligence and machine learning in prostate cancer patient management—current trends and future perspectives," Diagnostics (Basel), vol. 11, no. 2, p. 354, 2021, doi: 10.3390/diagnostics11020354.

[158] D. Kreuzberger, N. Kühl, and S. Hirschl, "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," IEEE Access, vol. 11, pp. 31866-31879, 2023, doi: 10.1109/ACCESS.2023.3262138.

[159] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," Medical Image Analysis, p. 102802, 2023/04/05/ 2023, doi: https://doi.org/10.1016/j.media.2023.102802.

[160] J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik, "Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis," Journal of Behavioral and Experimental Finance, vol. 32, p. 100577, 2021/12/01/ 2021, doi: https://doi.org/10.1016/j.jbef.2021.100577.

[161] C. Zhou *et al.*, "A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT," 2023, doi: 10.48550/arxiv.2302.09419.

[162] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," Cognitive Robotics, vol. 3, pp. 54-70, 2023, doi: 10.1016/j.cogr.2023.04.001.

[163] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in, 2021: IEEE, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986.

[164] M. Mohri, A. Rostamizadeh, and A. Talwalkar, "Foundations of Machine Learning," in Adaptive computation and machine learning, 2012.

[165] A. M. Turing, "Computing Machinery and Intelligence," Mind, vol. 59, no. 236, pp. 433-460, 1950.

[166] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bulletin of mathematical biology, vol. 52, no. 1, pp. 99-115, 1990, doi: 10.1016/S0092-8240(05)80006-0.

[167] D. O. Hebb, "The organization of behavior : a neuropsychological theory". New York: New York : Wiley, 1949, 1949.

[168] J. Orbach, "Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms," Archives of general psychiatry, vol. 7, no. 3, pp. 218-219, 1962, doi: 10.1001/archpsyc.1962.01720030064010.

[169] B. Widrow and M. E. Hoff, "Adaptive switching circuits," Stanford Univ Ca Stanford Electronics Labs, 1960.

[170] H. Chen, "Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms," Journal of the American Society for Information Science, vol. 46, no. 3, pp. 194-216, 1995, doi: https://doi.org/10.1002/(SICI)1097-4571(199504)46:3<194::AID-ASI4>3.0.CO;2-S.

[171] J. A. Bernard, "Use of a rule-based system for process control," IEEE Control Systems Magazine, vol. 8, no. 5, pp. 3-13, 1988, doi: 10.1109/37.7735.

[172] "A guide to expert systems". Reading, Mass.: Reading, Mass. : Addison-Wesley, c1986 [i.e.1985, 1985.

[173] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81-106, 1986/03/01 1986, doi: 10.1007/BF00116251.

[174] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967, doi: 10.1109/TIT.1967.1053964.

[175] J. Pearl, "Probabilistic reasoning in intelligent systems : networks of plausible inference". San Francisco, Calif.: San Francisco, Calif. : Morgan Kaufmann Publishers, 1988, 1988.

[176] M. Mitchell, "An introduction to genetic algorithms". Cambridge, Massachusetts; London, England: Cambridge, Massachusetts; London, England : The MIT Press, 1998, 1998.

[177] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533-536, 1986/10/01 1986, doi: 10.1038/323533a0.

[178] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995/09/01 1995, doi: 10.1007/BF00994018.

[179] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," presented at the Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Bari, Italy, 1996.

[180] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001/10/01 2001, doi: 10.1023/A:1010933404324.

[181] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," presented at the Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996.

[182] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, pp. 100-108, 1979, doi: 10.2307/2346830.

[183] H. Abdi and L. J. Williams, "Principal component analysis," WIREs Computational Statistics, vol. 2, no. 4, pp. 433-459, 2010, doi: https://doi.org/10.1002/wics.101.

[184] C. J. C. H. Watkins and P. Dayan, "Q-learning," Machine Learning, vol. 8, no. 3, pp. 279-292, 1992/05/01 1992, doi: 10.1007/BF00992698.

[185] G. Rummery and M. Niranjan, "On-Line Q-Learning Using Connectionist Systems," Technical Report CUED/F-INFENG/TR 166, 11/04 1994.

[186] Y. Chen, G. Wang, and S. Dong, "Learning with progressive transductive support vector machine," Pattern Recognition Letters, vol. 24, no. 12, pp. 1845-1855, 2003/08/01/ 2003, doi: https://doi.org/10.1016/S0167-8655(03)00008-4.

[187] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345-1359, 2010, doi: 10.1109/TKDE.2009.191.

[188] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," The Journal of artificial intelligence research, vol. 4, pp. 129-145, 1996, doi: 10.1613/jair.295.

[189] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015/05/01 2015, doi: 10.1038/nature14539.

[190] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[191] I. Goodfellow *et al.*, "Generative adversarial networks," Commun. ACM, vol. 63, no. 11, pp. 139–144, 2020, doi: 10.1145/3422622.

[192] Y. Kim, "Convolutional Neural Networks for Sentence Classification," 2014, doi: 10.48550/arxiv.1408.5882.

[193] A. Vaswani *et al.*, "Attention Is All You Need," 2017, doi: 10.48550/arxiv.1706.03762.

[194] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, doi: 10.48550/arxiv.1810.04805.

[195] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[196] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[197] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," Nature, vol. 529, no. 7587, pp. 484-489, 2016/01/01 2016, doi: 10.1038/nature16961.

[198] D. Silver *et al.*, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," Science, vol. 362, no. 6419, pp. 1140-1144, 2018, doi: doi:10.1126/science.aar6404.

[199] D. Betancourt and R. Muhanna, "Interval Deep Learning for Uncertainty Quantification in Safety Applications," 2021, doi: 10.48550/arxiv.2105.06438.

[200] H. Wang and D.-Y. Yeung, "A Survey on Bayesian Deep Learning," ACM Comput. Surv., vol. 53, no. 5, p. Article 108, 2020, doi: 10.1145/3409383.

[201] L. Deng and J. Platt, "Ensemble deep learning for speech recognition," in Proc. interspeech, 2014.

[202] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020/06/01/ 2020, doi: https://doi.org/10.1016/j.inffus.2019.12.012.

[203] E. Hüllermeier, "Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning: Are Conditional Entropy and Mutual Information Appropriate Measures?," 2022, doi: 10.48550/arxiv.2209.03302.

[204] S. Munikoti, D. Agarwal, L. Das, and B. Natarajan, "A general framework for quantifying aleatoric and epistemic uncertainty in graph neural networks," Neurocomputing (Amsterdam), vol. 521, pp. 1-10, 2023, doi: 10.1016/j.neucom.2022.11.049.

[205] E. Hüllermeier and W. Waegeman, "Aleatoric and Epistemic Uncertainty in Machine Learning: A Tutorial Introduction," ArXiv, vol. abs/1910.09457, 2019.

[206] M. Sun, T. Zhang, Y. Wang, G. Strbac, and C. Kang, "Using Bayesian Deep Learning to Capture Uncertainty for Residential Net Load Forecasting," IEEE transactions on power systems, vol. 35, no. 1, pp. 188-201, 2020, doi: 10.1109/TPWRS.2019.2924294.

[207] H. Wang and D.-Y. Yeung, "Towards Bayesian Deep Learning: A Framework and Some Existing Methods," IEEE transactions on knowledge and data engineering, vol. 28, no. 12, pp. 3395-3408, 2016, doi: 10.1109/TKDE.2016.2606428.

[208] C. E. Rasmussen and C. K. I. Williams, "Gaussian processes for machine learning". Cambridge, Massachusetts; Piscataqay, New Jersey: Cambridge, Massachusetts : MIT Press, 2005, 2005.

[209] J. L. Beck and L. S. Katafygiotis, "Updating Models and Their Uncertainties. I: Bayesian Statistical Framework," Journal of Engineering Mechanics, vol. 124, no. 4, pp. 455-461, 1998, doi: doi:10.1061/(ASCE)0733-9399(1998)124:4(455).

[210] M. Schmitt, S. T. Radev, and P.-C. Bürkner, "Meta-Uncertainty in Bayesian Model Comparison," presented at the Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, 2023. [Online]. Available: https://proceedings.mlr.press/v206/schmitt23a.html.

[211] I. Park, H. K. Amarchinta, and R. V. Grandhi, "A Bayesian approach for quantification of model uncertainty," Reliability Engineering & System Safety, vol. 95, no. 7, pp. 777-785, 2010/07/01/ 2010, doi: https://doi.org/10.1016/j.ress.2010.02.015.

[212] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," presented at the Proceedings of The 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research, 2016. [Online]. Available: https://proceedings.mlr.press/v48/gal16.html.

[213] J. Kittler, F. Roli, and J. van Leeuwen, "Ensemble Methods in Machine Learning," vol. 1857. Germany: Germany: Springer Berlin / Heidelberg, 2000, pp. 1-15.

[214] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," Engineering Applications of Artificial Intelligence, vol. 115, p. 105151, 2022/10/01/ 2022, doi: https://doi.org/10.1016/j.engappai.2022.105151.

[215] P. Kazienko, E. Lughofer, and B. Trawiński, "Hybrid and ensemble methods in machine learning J.UCS special issue," Journal of Universal Computer Science, vol. 19, no. 4, pp. 457-461, 2013.

[216] O. Sagi and L. Rokach, "Ensemble learning: A survey," WIREs Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1249, 2018, doi: https://doi.org/10.1002/widm.1249.

[217] T.-H. Lee, A. Ullah, and R. Wang, "Bootstrap Aggregating and Random Forest," in Macroeconomic Forecasting in the Era of Big Data: Theory and Practice, P. Fuleky Ed. Cham: Springer International Publishing, 2020, pp. 389-429.

[218] H. Drucker, C. Cortes, L. D. Jackel, Y. LeCun, and V. Vapnik, "Boosting and Other Ensemble Methods," Neural Computation, vol. 6, no. 6, pp. 1289-1301, 1994, doi: 10.1162/neco.1994.6.6.1289.

[219] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241-259, 1992/01/01/ 1992, doi: https://doi.org/10.1016/S0893-6080(05)80023-1.

[220] B. Efron and R. Tibshirani, "An introduction to the bootstrap". New York London: New York London : Chapman & Hall, 1993, 1993.

[221] C. M. Bishop, "Pattern recognition and machine learning". New York: New York : Springer, 2006, 2006.

[222] H. Anthony and K. Kamnitsas, "On the Use of Mahalanobis Distance for Out-of-distribution Detection with Neural Networks for Medical Imaging," in Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, Cham, C. H. Sudre, C. F. Baumgartner, A. Dalca, R. Mehta, C. Qin, and W. M. Wells, Eds., 2023// 2023: Springer Nature Switzerland, pp. 136-146.

[223] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting Adversarial Samples from Artifacts," 2017, doi: 10.48550/arxiv.1703.00410.

[224] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," 2017, doi: 10.48550/arxiv.1706.06083.

[225] H.-L. Wei, Z.-Q. Lang, and S. A. Billings, "Constructing an overall dynamical model for a system with changing design parameter properties," International Journal of Modelling, Identification and Control, vol. 5, no. 2, pp. 93-104, 2008, doi: 10.1504/ijmic.2008.022014.

[226] T. Baldacchino, S. R. Anderson, and V. Kadirkamanathan, "Computational system identification for Bayesian NARMAX modelling," Automatica, vol. 49, no. 9, pp. 2641-2651, 2013/09/01/ 2013, doi: https://doi.org/10.1016/j.automatica.2013.05.023.

[227] A. Zadra *et al.*, "Systematic Errors in Weather and Climate Models: Nature, Origins, and Ways Forward," (in English), Bulletin of the American Meteorological Society, vol. 99, no. 4, pp. ES67-ES70, 01 Apr. 2018 2018, doi: https://doi.org/10.1175/BAMS-D-17-0287.1.

[228] A. Frassoni *et al.*, "Systematic Errors in Weather and Climate Models: Challenges and Opportunities in Complex Coupled Modeling Systems," (in English), Bulletin of the American

Meteorological Society, vol. 104, no. 9, pp. E1687-E1693, 01 Sep. 2023 2023, doi: https://doi.org/10.1175/BAMS-D-23-0102.1.

[229] "Noise and Distortion," in Advanced Digital Signal Processing and Noise Reduction, 2008, pp. 35-50.

[230] S. V. Vaseghi, "Advanced digital signal processing and noise reduction", 2nd ed. ed. Chichester: Chichester : Wiley, c2000, 2000.

[231] C. law Bylinski, "Some basic properties of sets," Formalized Mathematics, vol. 1, no. 1, pp. 47-53, 1990.

[232] H.-L. Wei and S. A. Billings, "Improved model identification for non-linear systems using a random subsampling and multifold modelling (RSMM) approach," International Journal of Control, vol. 82, no. 1, pp. 27-42, 2009, doi: https://doi.org/10.1080/00207170801955420.

[233] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," 2015, doi: 10.48550/arxiv.1506.02142.

[234] H. Alibrahim and S. A. Ludwig, "Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization," in 2021 IEEE Congress on Evolutionary Computation (CEC), 28 June-1 July 2021 2021, pp. 1551-1559, doi: 10.1109/CEC45853.2021.9504761.

[235] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," Journal of Machine Learning Research, vol. 18, no. 185, pp. 1-52, 2018.

[236] X. Xiao, M. Yan, S. Basodi, C. Ji, and Y. Pan, "Efficient hyperparameter optimization in deep learning using a variable length genetic algorithm," arXiv preprint arXiv:2006.12703, 2020.

[237] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb," Journal of Electronic Science and Technology, vol. 17, no. 1, pp. 26-40, 2019/03/01/ 2019, doi: https://doi.org/10.11989/JEST.1674-862X.80904120.

[238] H. Zhou, J. Li, S. Zhang, S. Zhang, M. Yan, and H. Xiong, "Expanding the prediction capacity in long sequence time-series forecasting," Artificial Intelligence, vol. 318, p. 103886, 2023/05/01/ 2023, doi: https://doi.org/10.1016/j.artint.2023.103886.

[239] R. J. Hall and E. Hanna, "North Atlantic circulation indices: links with summer and winter UK temperature and precipitation and implications for seasonal forecasting," International Journal of Climatology, vol. 38, pp. e660-e677, 2018, doi: https://doi.org/10.1002/joc.5398.

[240] R. Davies. "The Cost of the UK Floods." The floodlist. https://floodlist.com/insurance/uk/cost-of-2013-2014-floods (accessed 02.03.2016, 2016).

[241] Stephenson, David B and Pavan, and R. Valentina and Bojariu, "Is the North Atlantic Oscillation a random walk?," International Journal of Climatology: A Journal of the Royal Meteorological Society, vol. 20, no. 1, pp. 1-18, 2000, doi: https://doi.org/10.1002/(SICI)1097-0088(200001)20:1<1::AID-JOC456>3.0.CO;2-P.

[242] A. A. Scaife *et al.*, "Skillful long-range prediction of European and North American winters," Geophysical Research Letters, vol. 41, no. 7, pp. 2514-2519, 2014, doi: https://doi.org/10.1002/2014GL059637.

[243] R. Hall, R. Erdélyi, E. Hanna, J. M. Jones, and A. A. Scaife, "Drivers of North Atlantic polar front jet stream variability," International Journal of Climatology, vol. 35, no. 8, pp. 1697-1720, 2015, doi: https://doi.org/10.1002/joc.4121.

[244] E. N. Lorenz, "Deterministic nonperiodic flow," Journal of atmospheric sciences, vol. 20, no. 2, pp. 130-141, 1963, doi: https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

[245] Y. Kushnir, W. A. Robinson, P. Chang, and A. W. Robertson, "The physical basis for predicting Atlantic sector seasonal-to-interannual climate variability," Journal of Climate, vol. 19, no. 23, pp. 5949-5970, 2006, doi: https://doi.org/10.1175/JCLI3943.1.

[246] R. J. Hall, A. A. Scaife, E. Hanna, J. M. Jones, and R. Erdélyi, "Simple Statistical Probabilistic Forecasts of the Winter NAO," (in English), Weather and Forecasting, vol. 32, no. 4, pp. 1585-1601, 01 Aug. 2017 2017, doi: 10.1175/waf-d-16-0124.1.

[247] A. Maidens, J. R. Knight, and A. A. Scaife, "Tropical and stratospheric influences on winter atmospheric circulation patterns in the North Atlantic sector," Environmental Research Letters, vol. 16, no. 2, p. 024035, 2021, doi: https://doi.org/10.1088/1748-9326/abd8aa.

[248] J. Cohen *et al.*, "S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts," Wiley Interdisciplinary Reviews: Climate Change, vol. 10, no. 2, p. e00567, 2019, doi: https://doi.org/10.1002/wcc.567.

[249] R. Eade *et al.*, "Do seasonal-to-decadal climate predictions underestimate the predictability of the real world?," Geophysical research letters, vol. 41, no. 15, pp. 5620-5628, 2014, doi: https://doi.org/10.1002/2014GL061146.

[250] T. N. Stockdale, F. Molteni, and L. Ferranti, "Atmospheric initial conditions and the predictability of the Arctic Oscillation," Geophysical Research Letters, vol. 42, no. 4, pp. 1173-1179, 2015, doi: https://doi.org/10.1002/2014GL062681.

[251] R. J. Hall, J. M. Jones, E. Hanna, A. A. Scaife, and R. Erdélyi, "Drivers and potential predictability of summer time North Atlantic polar front jet variability," Climate Dynamics, vol. 48, no. 11, pp. 3869-3887, 2017, doi: https://doi.org/10.1007/s00382-016-3307-0.

[252] M. Baker, D. Bergstresser, G. Serafeim, and J. Wurgler, "Financing the response to climate change: The pricing and ownership of US green bonds," National Bureau of Economic Research, 2018.

[253] A. Weisheimer *et al.*, "How confident are predictability estimates of the winter North Atlantic Oscillation?," Quarterly Journal of the Royal Meteorological Society, vol. 145, pp. 140-159, 2019, doi: https://doi.org/10.1002/qj.3446.

[254] H. Hersbach *et al.*, "The ERA5 global reanalysis," Quarterly Journal of the Royal Meteorological Society, vol. 146, no. 730, pp. 1999-2049, 2020, doi: https://doi.org/10.1002/qj.3803.

[255] A. Dawson, "eofs: A Library for EOF Analysis of Meteorological, Oceanographic, and Climate Data," Journal of Open Research Software, vol. 4, no. 1, p. e14, 2016, doi: http://doi.org/10.5334/jors.122.

[256] C. K. Folland, J. Knight, H. W. Linderholm, D. Fereday, S. Ineson, and J. W. Hurrell, "The summer North Atlantic Oscillation: past, present, and future," Journal of Climate, vol. 22, no. 5, pp. 1082-1103, 2009, doi: https://doi.org/10.1175/2008JCLI2459.1.

[257] R. J. Hall, "The North Atlantic polar front jet stream: variability and predictability, 1871-1914," University of Sheffield, 2016.

[258] A. Ossó, R. Sutton, L. Shaffrey, and B. Dong, "Observational evidence of European summer weather patterns predictable from spring," Proceedings of the National Academy of Sciences, vol. 115, no. 1, pp. 59-63, 2018, doi: doi:10.1073/pnas.1713146114.

[259] C. Cassou, L. Terray, J. W. Hurrell, and C. Deser, "North Atlantic winter climate regimes: Spatial asymmetry, stationarity with time, and oceanic forcing," Journal of Climate, vol. 17, no. 5, pp. 1055-1068, 2004, doi: https://doi.org/10.1175/1520-0442(2004)017<1055:NAWCRS>2.0.CO;2.

[260] J. Marshall *et al.*, "North Atlantic climate variability: phenomena, impacts and mechanisms," International Journal of Climatology: A Journal of the Royal Meteorological Society, vol. 21, no. 15, pp. 1863-1898, 2001.

[261] S. W. D. Center, "The International Sunspot Number," International Sunspot Number Monthly Bulletin and online catalogue, 1956-2021. [Online]. Available: http://www.sidc.be/silso/.

[262] B. Naujokat, "An update of the observed quasi-biennial oscillation of the stratospheric winds over the tropics," Journal of Atmospheric Sciences, vol. 43, no. 17, pp. 1873-1877, 1986, doi: https://doi.org/10.1175/1520-0469(1986)043<1873:AUOTOQ>2.0.CO;2.

[263] C.-S. J. Chu, "Time series segmentation: A sliding window approach," Information Sciences, vol. 85, no. 1, pp. 147-173, 1995/07/01/ 1995, doi: https://doi.org/10.1016/0020-0255(95)00021-G.

[264] E. Dias *et al.*, "Sliding Window Network Coding Enables NeXt Generation URLLC Millimeter-Wave Networks," IEEE Networking Letters, vol. 5, no. 3, pp. 159-163, 2023, doi: 10.1109/LNET.2023.3269387.

[265] G. Wei, Z. Duan, S. Li, X. Yu, and G. Yang, "LFEformer: Local Feature Enhancement Using Sliding Window With Deformability for Automatic Speech Recognition," IEEE Signal Processing Letters, vol. 30, pp. 180-184, 2023, doi: 10.1109/LSP.2023.3241558.

[266] G. R. Bigg *et al.*, "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, vol. 470, no. 2166, p. 20130662, 2014, doi: doi:10.1098/rspa.2013.0662.

[267] Y. Zhao *et al.*, "Inferring the variation of climatic and glaciological contributions to West Greenland iceberg discharge in the twentieth century," Cold Regions Science and Technology, vol. 121, pp. 167-178, 2016/01/01/ 2016, doi: https://doi.org/10.1016/j.coldregions.2015.08.006.

[268] A. A. Bradley, S. S. Schwartz, and T. Hashino, "Sampling uncertainty and confidence intervals for the Brier score and Brier skill score," Weather and Forecasting, vol. 23, no. 5, pp. 992-1006, 2008, doi: https://doi.org/10.1175/2007WAF2007049.1.

[269] J. Bröcker and L. A. Smith, "Increasing the reliability of reliability diagrams," Weather and forecasting, vol. 22, no. 3, pp. 651-661, 2007, doi: https://doi.org/10.1175/WAF993.1.

[270] M. Assel, D. D. Sjoberg, and A. J. Vickers, "The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models," Diagnostic and prognostic research, vol. 1, no. 1, pp. 1-7, 2017, doi: https://doi.org/10.1186/s41512-017-0020-3.

[271] S. J. Johnson *et al.*, "SEAS5: the new ECMWF seasonal forecast system," Geoscientific Model Development, vol. 12, no. 3, pp. 1087-1117, 2019, doi: https://doi.org/10.5194/gmd-12-1087-2019.

[272] C. MacLachlan *et al.*, "Global Seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system," Quarterly Journal of the Royal Meteorological Society, vol. 141, no. 689, pp. 1072-1084, 2015, doi: https://doi.org/10.1002/qj.2396.

[273] A. Weisheimer and T. Palmer, "On the reliability of seasonal climate forecasts," Journal of the Royal Society Interface, vol. 11, no. 96, p. 20131162, 2014, doi: https://doi.org/10.1098/rsif.2013.1162.

[274] C. Cassou, C. Deser, L. Terray, J. W. Hurrell, and M. Drévillon, "Summer sea surface temperature conditions in the North Atlantic and their impact upon the atmospheric circulation in early winter," Journal of Climate, vol. 17, no. 17, pp. 3349-3363, 2004, doi: https://doi.org/10.1175/1520-0442(2004)017<3349:SSSTCI>2.0.CO;2.

[275] J. L. Warner, J. A. Screen, and A. A. Scaife, "Links between Barents-Kara sea ice and the extratropical atmospheric circulation explained by internal variability and tropical forcing," Geophysical Research Letters, vol. 47, no. 1, p. e2019GL085679, 2020, doi: https://doi.org/10.1029/2019GL085679.

[276] R. K. K. Li, T. Woollings, C. O'Reilly, and A. A. Scaife, "Tropical atmospheric drivers of wintertime European precipitation events," Quarterly Journal of the Royal Meteorological Society, vol. 146, no. 727, pp. 780-794, 2020, doi: https://doi.org/10.1002/qj.3708.

[277] H. L. Wei, S. A. Billings, and J. Liu, "Term and variable selection for non-linear system identification," International Journal of Control, vol. 77, no. 1, pp. 86-110, 2004/01/10 2004, doi: 10.1080/00207170310001639640.

[278] L. A. Aguirre and S. Billings, "Improved structure selection for nonlinear models based on term clustering," International journal of control, vol. 62, no. 3, pp. 569-587, 1995.

[279] A. Falsone, L. Piroddi, and M. Prandini, "A randomized algorithm for nonlinear model structure selection," Automatica, vol. 60, pp. 227-238, 2015.

[280] P. F. L. Retes and L. A. Aguirre, "NARMAX model identification using a randomised approach," International Journal of Modelling, Identification and Control, vol. 31, no. 3, pp. 205-216, 2019.

[281] S. Billings and H.-L. Wei, "NARMAX model as a sparse, interpretable and transparent machine learning approach for big medical and healthcare data analysis," in 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2019: IEEE, pp. 2743-2750.

[282] S. A. Billings and H.-L. Wei, "An adaptive orthogonal search algorithm for model subset selection and non-linear system identification," International Journal of Control, vol. 81, no. 5, pp. 714-724, 2008.

[283] E. Mendez and S. A. Billings, "An alternative solution to the model structure selection problem," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 31, no. 6, pp. 597-608, 2001.

[284] A. Vaswani *et al.*, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[285] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," arXiv preprint arXiv:1905.09418, 2019.

[286] D. Wang *et al.*, "Multi-Head Self-Attention with Role-Guided Masks," in European Conference on Information Retrieval, 2021: Springer, pp. 432-439.

[287] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7262-7272.

[288] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," arXiv preprint arXiv:2112.01527, 2021.

[289] Z. Fan *et al.*, "Mask attention networks: Rethinking and strengthen transformer," arXiv preprint arXiv:2103.13597, 2021.

[290] K. Kim *et al.*, "Rethinking the self-attention in vision transformers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3071-3075.

[291] A. Nicolson and K. K. Paliwal, "Masked multi-head self-attention for causal speech enhancement," Speech Communication, vol. 125, pp. 80-96, 2020.

[292] G.-X. Wen and Y.-J. Liu, "Adaptive fuzzy-neural tracking control for uncertain nonlinear discrete-time systems in the NARMAX form," Nonlinear Dynamics, vol. 66, no. 4, pp. 745-753, 2011.

[293] S. A. Billings and H.-L. Wei, "A new class of wavelet networks for nonlinear system identification," IEEE Transactions on neural networks, vol. 16, no. 4, pp. 862-874, 2005.

[294] H.-L. Wei, S. A. Billings, and M. Balikhin, "Prediction of the Dst index using multiresolution wavelet models," Journal of Geophysical Research: Space Physics, vol. 109, no. A7, 2004.

[295] H. Zou, "The adaptive lasso and its oracle properties," Journal of the American statistical association, vol. 101, no. 476, pp. 1418-1429, 2006.

[296] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.

[297] A. G. Parlos, O. T. Rais, and A. F. Atiya, "Multi-step-ahead prediction using dynamic recurrent neural networks," Neural Networks, vol. 13, no. 7, pp. 765-786, 2000/09/01/ 2000, doi: https://doi.org/10.1016/S0893-6080(00)00048-4.