2025 32nd International Conference on Systems, Signals and Image Processing, Skopje, North Macedonia, 24 - 26 June 2025. Final Accepted Manuscript.

System Identification and Interpretable Modelling of Dynamical Systems with Small Data Using Sparse Bayesian Learning

Hua-Liang Wei[1,2]

[1] School of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, UK
[2] Centre of Machine Intelligence, The University of Sheffield, Sheffield, UK

w.hualiang@sheffield.ac.uk

Abstract— **Learning from data plays a major role in understanding complex natural and engineered systems. System identification (SysID), as a data-driven modelling technique, provides a powerful tool for building dynamical system models. Building models from noisy small data is a challenging research question in many practical problems. This paper is concerned with parsimonious and transparent modelling of dynamical systems which are of high interest in many real applications. Sparse Bayesian learning (SBL), due to its ability to use prior information to generate sparse predictive models, is employed in this study to estimate models from small data. The performance of the proposed sparse Bayesian learning approach is tested using real-life data. Experimental results show that the SBL approach shows strong performance for solving small data modelling problems.**

Index Terms—dynamical system, interpretable modelling, nonlinear system, small sample size, small data, sparse Bayesian learning, system identification.

I. Introduction

System identification (SysID) is a methodology that is widely used for building mathematical models of dynamical systems from measured input and output data [1]. SysID and machine learning (ML) share many common aspects in their implementation procedures including data acquisition and processing; model structure selection; model training, parameter estimation; and model verification, validation and refinement [2],[3]. SysID was initially invented in the field of control engineering [1], but gradually it has been widely used in many other fields. Traditionally, the most important primary roles or tasks of SysID is to construct models that enable physical interpretability or mechanistic understanding of black-box systems of interest, so as to facilitate system design [4]-[6], system analysis and operation [7], forecasting and predictive analysis and so on [8]-[11]. To meet these requirements, many linear and nonlinear parametric models have been developed, including the commonly used ARX (AutoRegressive with eXogenous inputs) and ARMAX (AutoRegressive Moving Average with eXogenous inputs) [2], and NARX (Nonlinear ARX) and NARMAX (Nonlinear ARMAX) [3]. These models have several attractive advantages, for example, they are transparent, interpretable, and simulatable; allow analysis in both the time and frequency domains; and can well represent and approximate a wide range of linear and nonlinear dynamical systems in real world problems. Other types of models, e.g., neural networks [12] and deep learning [13],[14] have also been introduced and adapted for system identification where the primary objective is usually to make predictions, with less or little attention being paid to model explanation and interpretation.

In many real applications, SysID is concerned with constructing transparent and interpretable models. For many real problems, especially those relating to complex nonlinear black-box systems, the initially specified full models can be very complex and highly redundant due to the inclusion of many irrelevant model terms (or regresors). This is particularly true for problems involving a great number of input and output variables. Model selection plays a crucial role for achieving best parsimonious models [3]. Many efficient sparse learning methods have been developed over the past years or decades, including greedy search methods such as orthogonal least squares (OLS) [15],[16] and orthogonal matching pursuit (OMP) [17],[18], least absolute shrinkage and selection operator (LASSO) [19],[20], L_0-regularisation [21],[22], sparse Bayesian learning (SBL) [23]-[26], and randomised methods [27]-[29]. Each of these methods has its own strengths and limitations [30]-[31].

Building models with small data is an important and challenging issue in many practical applications, e.g. in seasonal weather and crop yield forecasting [32]. When the data are obtained under non-persistent excitation conditions (e.g., system inputs are not persistently exciting) [33], or the system to be modelled is highly nonlinear and has many inputs [8],[9], it is even more difficult and challenging to develop reliable models.

This paper endeavours to build transparent, interpretable, parsimonious and simulatable (TIPS) models [11] from small data using sparse Bayesian learning techniques. The originality and contributions of the work are as follows: 1) it endeavours to tackle the challenge of small data modelling problem by introducing an SBL approach; 2) it incorporates two model selection schemes, i.e., the Bayesian Information Criterion (BIC) and a penalized error-to-signal ratio (PESR) metric, into the associated modelling procedure to reduce the model redundancy so as to refine the model and control the model complexity; and 3) the ability of the SBL model performance is benchmarked through solving a small data modelling problem in climate relating to annual iceberg prediction in the Northwest Atlantic.

The remaining of the paper is organised as follows. Section 2 introduces SysID and NARX models. Section 3 presents the theory of the SBL framework for NARX model identification. Section 4 provides a case study to demonstrate the SBL efficacy for sparse model identification of a system driven by multiple inputs. A summary and suggestions for future work are provided in Section 5.

II. Identification of Nonlinear Dynamical Systems Using Parametric Models

For a black-box system whose internal structure is completely unknown but whose inputs and outputs can be measured, we would usually have a variety of choices of models and modelling approaches to approximate the system input-output behaviour to some extent. In practice, the ultimate goal of SysID is to find a model or a set of models that can characterise the system input-output relationship as accurate as possible. Moreover, in many cases, model interpretability may be highly desirable or crucial [34]. Keeping these in mind and emphasizing the Occam's razor from a means-ends perspective [35],[36], the focus of the paper is on transparent modelling to be achieved through sparse Bayesian learning.

NARX models, which include many linear and nonlinear models as special cases, are among the most commonly used representations for nonlinear dynamical identification [3]. For a general multiple-input, single-output (MISO) system, the NARX model can be represented as:

$$\left.\begin{aligned} y(t) = f[&y(t-1), y(t-2)..., y(t-n_y), \\ &u_1(t-\tau), u_1(t-\tau-1), ..., u_1(t-n_u), ..., \\ &u_r(t-\tau), u_r(t-\tau+1), ..., u_r(t-n_u)] + e(t) \end{aligned}\right\} \quad (1)$$

where $u_1(t)$, $u_2(t)$, …, $u_r(t)$ are r input signals, $y(t)$ is the output signal, and $e(k)$ is noise signal; n_y and n_u are the associated maximum time lags; τ is the time delay between the response and the inputs, and usually $\tau = 0$ or $\tau = 1$; $f[\bullet]$ is an unknown function to be built from data. Note that $e(t)$ is unobservable and can only be estimated based on an identified model; it is usually estimated using model prediction as $e(t) = y(t) - \hat{y}(t)$ where $\hat{y}(t)$ represents a model predicted value. Eq. (1) can be easily extended to multi-input, multi-output (MIMO) cases in a straightforward way (see e.g. [3],[37],[38]).

Define

$$\begin{aligned} \mathrm{x}(t) &= [x_1(t), ..., x_n(t)] \\ &= [y(t-1), ..., y(t-n_y), u_1(t-1), ..., u_1(t-n_u), ..., \\ &\quad u_r(t-1), ..., u_r(t-n_u)] \end{aligned} \quad (2)$$

Eq. (1) can then be rearranged to a linear-in-the-parameters (LIP) and nonlinear-in-the-variables (NIV) form as follows:

$$y(t) = \theta_0 + \sum_{m=1}^{M} \theta_m \varphi_m(t) + e(t) \quad (3)$$

where $n = n_y + vn_u$, meaning that the values of the regressors φ_m (m =1,2, …, M) are completely determined by the lagged variables defined in (2). Model (3) is usually referred to as an initial candidate full polynomial NARAX model, which is often represented in matrix format as:

$$\mathbf{y} = \mathbf{\Phi\theta} + \mathbf{e} \quad (4)$$

where $\mathbf{\Phi} = [\boldsymbol{\varphi}_0, \boldsymbol{\varphi}_1, ..., \boldsymbol{\varphi}_M]$, with $\boldsymbol{\varphi}_0 = [1,1,...,1]^T$; $\boldsymbol{\varphi}_m = [\varphi_m(1), ..., \varphi_m(t)]^T$ (m = 1, …, N), $\boldsymbol{\theta} = [\theta_0, ..., \theta_M]^T$ is a vector of unknown parameters. Eqs. (3) and (4) are referred to as Nonlinear Lagged Inputs and Outputs (NLIO) model [38], which is represented in a LIP-NIV form.

The main objective of SysID is to find a sparse solution, amounting to finding the best subset model consisting of s elements (model terms/regressors) selected from the M candidates, such that the s elements can well characterise the response y. In many cases, the problems to be solved in nonlinear SysID are ill-conditioned as the number of regressors is far larger than the number of samples (s << M). Therefore, efficient methods for sparse model identification are very important and useful. The following section presents such a method.

III. Sparse Bayesian Learning for NARX Identification

The central task of sparse model identification to significantly reduce the initial full models (3) and (4) through estimating and optimising the model parameters θ_m (m=0, 1, …, M). To emphasise this, the LIP-NIV model (4) can also be written as $\mathbf{y}=f(\mathbf{x};\ \boldsymbol{\theta})= \Phi\boldsymbol{\theta} + \mathbf{e}$. In recent years, SBL methods have been employed to solve the model as well as multiple regression [39]-[42]. In the following, the theory of the SBL method proposed in [23]-[26] is briefly introduced.

A. Multiple Regression from a Bayesian Perspective

Assume the values of the noise sequence $e(k)$ in (4) independent and follow a Gaussian distribution with mean zero and limited variance σ^2, i.e., $e(t)\sim N(0,\sigma^2)$. Thus, the distribution of the noise over the complete data is $p(\mathbf{e}) = \prod_{t=1}^{N} N(e(t) \mid 0, \sigma^2)$ [24], [26]. In practice, the true value of the variance σ^2 is not known; it can and needs to be estimated from data. With this error model, the likelihood of the complete data of interest can be formulated as:

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2) &= \prod_{t=1}^{N} p(y(t) \mid \boldsymbol{\theta}, \sigma^2) \\ &= \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|\, y(t) - f(\mathrm{x}(t);\boldsymbol{\theta}) \,|^2}{2\sigma^2} \right) \\ &= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{\| \mathbf{y} - \mathbf{\Phi\theta} \|^2}{2\sigma^2} \right) \end{aligned} \quad (5)$$

To effectively estimate the M+1 model parameters, a prior function involving M+1 independent hyperparameters is proposed in [24] which takes the form:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = (2\pi)^{-(M+1)/2} \prod_{m=0}^{M} \alpha_m^{1/2} \exp\left(-\frac{1}{2} \alpha_m \theta_m^2 \right) \quad (6)$$

where $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, ..., \alpha_M]^T$, each element individually controls the strength of the prior over its associated model parameter. This function plays a major role in achieving sparse models by eliminating unimportant regressors.

Following the Bayes' rule, the posterior distribution over all the unknowns can be desired as follows:

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \sigma^2 \mid \mathbf{y}) = p(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\alpha}, \sigma^2) p(\boldsymbol{\alpha}, \sigma^2 \mid \mathbf{y}) \tag{7}$$

where $p(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\alpha}, \sigma^2)$ is the posterior distribution of the regression coefficients. Given $\boldsymbol{\alpha}$, this posterior distribution is tractable as follows [23]:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\alpha}, \sigma^2) &= \frac{p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})}{p(\mathbf{y} \mid \boldsymbol{\alpha}, \sigma^2)} \\ &= (2\pi)^{-(M+1)/2} \mid \boldsymbol{\Sigma} \mid^{-1/2} \exp\left[-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}) \right] \end{aligned} \tag{8}$$

where

$$\boldsymbol{\Sigma} = (\sigma^{-2} \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \mathbf{A})^{-1}, \ \boldsymbol{\mu} = \sigma^{-2} \Sigma \boldsymbol{\Phi}^T \mathbf{y} \tag{9}$$

with $\mathbf{A} = \mathrm{diag}(\alpha_0, \alpha_1, ..., \alpha_M)$.

To facilitate Bayesian inference for sparsity determination and model parameter estimation, a most probable point (MPP) approach was proposed in [24],[25] using a type-II maximum likelihood procedure, through which sparse Bayesian learning is achieved by maximising (locally), with respect to $\boldsymbol{\alpha}$, the following logarithm-form marginal likelihood:

$$\begin{aligned} L(\boldsymbol{\alpha}) = \log p(\mathbf{y} \mid \boldsymbol{\alpha}, \sigma^2) &= \int_{-\infty}^{\infty} p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) d\boldsymbol{\theta} \\ &= -\frac{1}{2} [N \log(2\pi) + \log \mid \mathbf{C} \mid + \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y}] \end{aligned} \tag{10}$$

where $\mathbf{C} = \sigma^2 \mathbf{I} + \boldsymbol{\Phi} \mathbf{A}^{-1} \boldsymbol{\Phi}^T$. Denote by $\boldsymbol{\alpha}^{\mathrm{MPP}}$ a solution obtained via the most probable point method. A point estimate $\boldsymbol{\mu}^{\mathbf{MPP}} = \sigma^{-2} \Sigma^{\mathbf{MPP}} \boldsymbol{\Phi}^T \mathbf{y}$ can then be obtained as, leading to a final approximator $f(x; \boldsymbol{\mu}^{\mathbf{MPP}}) = \boldsymbol{\Phi} \boldsymbol{\mu}^{\mathbf{MPP}}$ of $\mathbf{y}$.

The values of the M +1 hyperparameters can be updated using gradient-based approach. The update formulas can be obtained by differentiating $\log p(\mathbf{y} \mid \boldsymbol{\alpha}, \sigma^2)$ with respect to α_m (m=0,1,…,M) and σ^2. By setting the derivatives to zero and rearranging, we obtain the following updated as [26]:

$$\alpha_m^{\mathrm{new}} = \frac{\gamma_m}{\mu_m^2} \tag{11}$$

$$(\sigma^2)^{\mathrm{new}} = \frac{\| \mathbf{y} - \boldsymbol{\Phi} \boldsymbol{\mu} \|^2}{N - \sum_{m=0}^{M} \gamma_m} \tag{12}$$

where $\gamma_m = 1 - \alpha_m \Sigma_{mm}$, with $\gamma_m \in [0,1]$ being a measure of defining the characteristic of the automatic relevance determination (ARD) prior and evaluating how well the parameter θ_m is determined for the data.

B. The Computational Procedure of SBL

The procedure of the sparse Bayesian inference algorithm for NARX model identification is summarised below:

1. Determine the regression matrix $\boldsymbol{\Phi}$ defined in (4).
2. Estimate the variance σ^2 of the target signal $\mathbf{y}$.
3. Initialise the parameter vector α_m (m=0,1, …, M).
4. Compute the posterior statistics $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ using (9).
5. Compute $\gamma_m = 1 - \alpha_m \boldsymbol{\Sigma}_{mm}$ (m=0,1, …, M).
6. Update α_m and σ^2 using (11).
7. Go to step 4 and repeat steps 4-6 until convergence.
8. Remove the corresponding regressiors from the model described by (4) if the computed optimal $\alpha_m = \infty$ as this implies $\mu_m = 0$ from (9).

C. Model Identification Procedure with SBL

In SBL, the importance or relevance of an individual regressors of the model is determined by the weight assigned to them: regressors with higher weights which are not penalised to zero are considered more important than those with relatively smaller weights. SBL determines the importance by maximising the marginal likelihood function with respect to all the hyperparameters α_m (m=0,1, …, M).

Our empirical experience shows that more than occasional SBL may include redundant regressors in the final determined models. To reduce the irrelevant regressors, we introduce two model selection schemes, i.e., the Bayesian Information Criterion (BIC) and a penalized error-to-signal ratio (PESR) metrics [43], into the associate modelling procedure to reduce the redundancy of the model produced by SBL. The procedure is as follows:

1. Perform SBL.
2. Rank the regressors based on their importance.
3. Use BIC and a penalized ESR (PESR) [] to control the model complexity (model length):

$$\mathrm{BIC}_m = N \log(\mathrm{MSE}_m) + m \log(N) \tag{13}$$

$$\mathrm{PESR}_m = \left(\frac{N}{N - \lambda m} \right)^2 \frac{\| \boldsymbol{\varepsilon}_m \|^2}{\| \mathbf{y} \|^2} \tag{14}$$

where $\mathbf{y}$ is an observation vector on a dataset of interest, $\boldsymbol{\varepsilon}_{\mathrm{m}}$ and MSE$_m$ are the model residual vector and mean square error related to the m-term model produced by the proposed sparse modelling approach. λ is an adjustable parameter; it is chosen to be 1 in this study. The metric PESR in (14) is a variant of the metrics defined in [44], [45].

IV. Numerical Experiments

This section provides an example to illustrate how SBL approach works for model identification of nonlinear dynamical with small data of the annual icebergs in the Northwest Atlantic. The dataset used in this work contains 40 annual measurements of the period 1982 – 2021. The descriptions of the response (Iceberg counts) and the nine potential drivers are shown in Table I, and more details can be found in [46]. This is a very typical small sample size data modelling problem, for which most complicated machine learning methods may not work (see e.g. [32]).

A. Experimental Settings

The initial full NARX model includes 220 terms, composed of all the lagged variables, e.g. Winter_NAO(t-1), Air_Temp(t-1), and all their quadratic and cubic cross-product terms e.g. [Winter_NAO(t-1)]×[Sea_Ice(t-1)]×[Sea_Ice(t-1)].

The 40 samples were separated to two parts: the first 30 (1982-2011) were used for model training and validate, and the remaining 10 (2012-2021) were used for model testing. A leave-one-out (LOO) cross-validation was performed for model validation.

TABLE I.

DESCRIPTIONS OF THE RESPONSE AND POTENTIAL DRIVERS [46].

Icebergs and potential drivers	**Descriptions**
Iceberg count	Normalized anomalies of the number of icebergs crossing 48degN on the Grand Banks.
Winter NAO	Average North Atlantic Oscillation over the months of December to March.
Air Temp	Mean normalized anomalies of annual air temperature.
Sea Ice	Mean normalized anomalies of Sea ice maximum area and season duration for Northern Labrador, Southern Labrador and Newfoundland shelves.
SST	Mean normalized anomalies of Sea Surface Temperature over NAFO divisions 2HJ3KLNOP.
S27 Temp	Normalized anomalies of the vertically-averaged temperature at Station 27.
S27 Sal	Normalized anomalies of the vertically-averaged salinity at Station 27.
S27 CIL	Normalized anomalies of the summer (June-August) cold intermediate layer core temperature at Station 27.
CIL area	Mean normalized anomalies of the summer cold intermediate layer area over hydrographic sections Seal Island, Bonavista and Flemish Cap on the Newfoundland and Labrador shelf.
Bottom Temp	Mean normalized anomalies of the bottom temperature during spring (NAFO divisions 3LNOPs) and fall (NAFO divisions 2HJ3KLNO).

B. Results

The results produced by SBL are shown in Table II. For comparison purposes, the elastic-net LASSO method [20] was also performed to the same data, with the same model experimental settings. The results given by elastic-net are shown in Table III. The values of MSE and R^2 (coefficient of determination) of the models clearly show that SBL significantly outperforms elastic-net for the problem here. For graphical illustration purposes, a comparison between model predictions (based on Table I) and the corresponding measurements, on the training period (1982-2011) and the testing period (2012-2021), are shown in Fig. 1.

It is worth mentioning that traditional Bayesian learning (TBL) does work for the data modelling problem here. For example, while a TBL algorithm performs perfect on the training data (1982-2011) with $R^2 = 0.9876$, it does not show any prediction skill on the testing data (2012-2021) where $R^2 = -0.1520$.

V. CONCLUSION

In this study an SBL approach was presented for model identification of nonlinear dynamical systems with small data; the efficacy of the approach was tested on a real-life problem predicting annual icebergs in the Northwest Atlantic. The performance of SBL was analysed and compared with that of elastic-net LASSO and traditional non-sparse Bayesian learning. SBL showed promising performance for solving the small data modelling problem. However, SBL has some drawbacks, for example, our previous experience showed that it was less effective when data are contaminated by non-Gaussian noise; the computational load of SBL could be very high when dealing with data modelling problems in high-dimensional space, e.g. when the number of variables and model training samples are large.

This study has its limitation, e.g., the experiments are far from comprehensive; more work needs to be done to further test and exploit the potentials of SBL.

TABLE II.

THE MODEL PRODUCED WITH SBL.

	Model Term	**Coefficient**
1	Air_Temp(t-1)	0.9266
2	Winter_NAO(t-1) × Sea_Ice(t-1) × Sea_Ice(t-1)	0.3139
3	S27_CIL(t-1)	-0.4385
4	S27_Temp(t-1) × S27_Temp(t-1)	0.4318
5	Winter_NAO(t-1) × Sea_Ice(t-1)	0.0045
6	Winter_NAO(t-1)× Winter_NAO (t-1)× Winter_NAO	-0.2196
7	Winter_NAO(t-1)× S27_CIL(t-1)	-0.3274

MSE = 0.3505 on training data; MSE = 0.5343 on testing data;
R^2 = 0.7098 on training data; R^2 = 0.4033 on testing data.

TABLE III.

THE MODEL PRODUCED WITH LASSO.

	Model Term	**Coefficient**
1	Air_Temp(t-1)	0.3910
2	Winter_NAO(t-1)× Sea_Ice(t-1)	-0.2246
3	Winter_NAO(t-1) × Sea_Ice(t-1) × Sea_Ice(t-1)	0.2164
4	Sea_Ice(t-1) × S27_Sal × S27_Sal	-0.0274
5	S27_Sal(t-1) × CIL area(t-1) × Bottom Temp(t-1)	0.1779

MSE = 0.5169 on training data; MSE = 0.7477 on testing data;
R^2 = 0.6350 on training data; R^2 = 0.0549 on testing data.

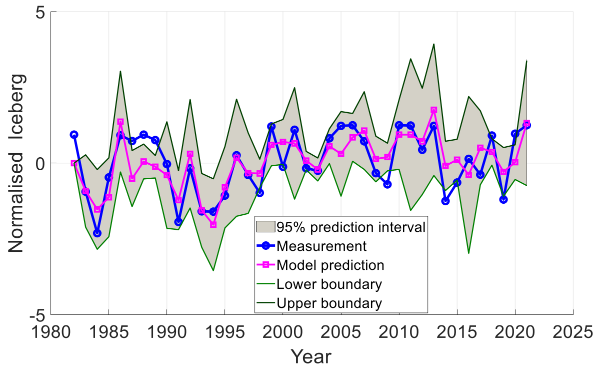

Fig 1. A comparison between model predictions (based on Table II) and measurements.

ACKNOWLEDGMENT

The authors gratefully acknowledge that this work was supported in part by STFC (Ref. ST/Y001524/1), NERC (Ref. NE/W005875/1, Ref. NE/V001787/1, Ref. NE/Y503290/1 and Ref. NE/V002511/1).

References

[1] L. A. Zadeh, "From circuit theory to system theory," *Proc. I.R.E.*, vol. 50, pp. 856-865, May 1962.

[2] L. Ljung, System Identification Toolbox User's Guide. The MathWorks, Inc., 2022.

[3] S. A. Billings, Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains. Hoboken, NJ, USA: John Wiley & Sons, 2013.

[4] C. Deiler, "Aerodynamic modeling, system identification, and analysis of iced aircraft configurations," *J. Aircr.*, vol. 55, pp.145–161. July 2017.

[5] P. Lichota, F. Dul, A. Karbowski, "System identification and LQR controller design with incomplete state observation for aircraft trajectory tracking," *Energies*, vol. 13, 5354, Oct. 2020.

[6] A. Simorgh, A. Razminia, and V. I. Shiryaev, "System identification and control design of a nonlinear continuously stirred tank reactor," *Math. Comput. Simul.*, vol. 173, pp. 16–31, Feb. 2020.

[7] E. Reynders, "System identification methods for (operational) modal analysis: Review and comparison," *Arch. Comput. Methods Eng.*, vol. 19, no. 1, pp. 51–124, Feb. 2012.

[8] 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," *Q. J. R. Meteorol Soc.*, vol. 145, pp. 2568-2585, June 2019.

[9] Y. Sun, I. Simpson, H. L. Wei, and E. Hanna, "Probabilistic seasonal forecasts of North Atlantic atmospheric circulation using complex systems modelling and comparison with dynamical models," *Meteorol. Appl.*, vol. 31, no. 1, e2178, Feb. 2024.

[10] Y. Gu et al., "System identification and data-driven forecasting of AE index and prediction uncertainty analysis using a new cloud-NARX model," *J. Geophys. Res.*, vol. 124, pp. 248-263, Dec. 2018.

[11] H.-L. Wei and S. A. Billinsg, "Modelling COVID-19 pandemic dynamics using transparent, interpretable, parsimonious and simulatable (TIPS) machine learning models: A case study from systems thinking and system identification perspectives," in *Recent Advances in AI-enabled Automated Medical Diagnosis*, R. Jiang et. al, eds., New York: CRC Press, 2022, pp. 13-28.

[12] O. Nelles, Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models. Berlin: Springer-Verlag, 2020.

[13] H. Zhou, C. Ibrahim, W. X. Zheng, and W. Pan, "Sparse Bayesian deep learning for dynamic system identification," *Automatica*, vol. 144, p. 110489, Oct. 2022.

[14] R. Sasaki et al., "A deep neural network with module architecture for model reduction and its application to nonlinear system identification," *IFAC PapersOnLine*, vol. 56, no.2, pp.10650–10655, Nov. 2023.

[15] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, no. 5, pp. 1873–1896, Nov. 1989.

[16] S. A. Billings and H.-L. Wei, "An adaptive orthogonal search algorithm for model subset selection and non-linear system identification," *Int. J. Control*, vol. 81, no. 5, pp. 714-724, April 2008.

[17] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Syst. Comput.*, vol. 1. Pacific Grove, CA, USA, 1993, pp. 40–44.

[18] J. A. Tropp and A. C. Gilbert, ''Signal recovery from random measurements via orthogonal matching pursuit,'' *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[19] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Stat. Soc., Ser. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[20] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York: Springer, 2009.

[21] D. Bertsimas et al., "Best subset selection via a modern optimization lens," *Ann. Statist.*, vol. 44, no. 2, pp. 813-852, April 2016.

[22] J. Huang, Y. Jiao, Y. Liu, X. Lu, "A constructive approach to l0 penalized regression," *J. Mach. Learn. Res.* vol. 19, no. 10, pp. 1–37, 2018.

[23] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Jun. 2001.

[24] M. E. Tipping, "Bayesian inference: an introduction to principles and practice in machine learning," in *Proc. 9th International Workshop on Artificial Intelligence and Statistics*, C. Bishop et al., eds., Key West, FL, USA, 3–6 Jan. 2003, pp. 276–283.

[25] A. C. Faul and M. E. Tipping, "Analysis of sparse Bayesian learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Cambridge, MA, USA, MIT Press, Jan. 2001, pp. 383–389.

[26] M. E. Tipping, "Bayesian inference: an introduction to principles and practice in machine learning," in *Advanced Lectures on Machine Learning*, O. Bousquet et al. eds., Berlin, Heidelberg, Germany: Springer, 2004, pp. 41–62.

[27] A. Falsone et al., "A randomized algorithm for nonlinear model structure selection," *Automatica*, vol. 60, pp. 227–238, Oct. 2015.

[28] P. E. L. Retes and L.A. Aguirre, "NARMAX model identification using a randomized approach," *Int. J. Model. Identif. Control*, vol. 31, no. 3, pp. 205–216, March 2019.

[29] F. Hafiz, A. Swain, and E. Mendes, "Multi-objective evolutionary framework for non-linear system identification: A comprehensive investigation," Neurocomputing, vol. 386, pp. 257–280, April 2020.

[30] X. Liu and X. Yang, "Exploiting spike-and-slab prior for variational estimation of nonlinear systems," *IEEE Trans. Ind. Informat.*, vol. 19, no. 11, pp. 11275-11285, Feb. 2023.

[31] X. Liu et al., "Joint parameter and time-delay estimation for a class of Wiener models based on a new orthogonal least squares algorithm," *Nonlinear Dyn.*, vol. 112, pp. 12159–12170, June 2024.

[32] R.J. Hall et al. "Complex systems modelling of UK winter wheat yield," Comput. Electr. Agric., 209, 107855, June 2023.

[33] Y. Guo et al. "An iterative orthogonal forward regression algorithm," *Int. J. Syst. Sci.*, vol. 46, no. 5, pp. 776-789, April 2015.

[34] D. Materassi, et al., "Explaining complex systems: a tutorial on transparency and interpretability in machine learning models (part II)," *IFAC-PapersOnLine*, vol. 58, no. 15, pp. 497-501, 2024.

[35] F. Petropoulos et al., "Wielding Occam's razor: Fast and frugal retail forecasting," *J. Oper. Res. Soc.*, Nov. 2024.

[36] T. F. Sterkenburg, "Statistical learning theory and Occam's razor: The core argument," *Minds and Machines*, vol. 35, no. 3, pp. 1–28, 2025.

[37] S. A. Billings and Q. M. Zhu, "Model validation tests for multivariable nonlinear models including neural networks," *Int. J. Control*, vol. 62, no. 4, pp. 749–766, Oct. 1995.

[38] H.-L. Wei, "System identification-informed transparent and explainable machine learning with application to power consumption forecasting," in *Proc. 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, Tenerife, Spain, 19-21 July 2023, pp. 1-6.

[39] C. Lu et al., ''Bagging linear sparse Bayesian learning models for variable selection in cancer diagnosis,'' *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 3, pp. 338–347, May 2007.

[40] W. R. Jacobs, T. Baldacchino, and S. R. Anderson, "Sparse Bayesian identification of polynomial NARX models," *IFAC-PapersOnLine*, vol. 48, no. 28, pp. 172–177, Dec. 2015.

[41] S. E. Ament and C. P. Gomes, "Sparse Bayesian learning via stepwise regression," In *Proc. 38th International Conference on Machine Learning*, PMLR, Virtual, July 2021, pp. 264–274.

[42] N. Zheng, Y. Li, W. Shi, and Q. Xie, "Sparse Bayesian based NARX modeling of cortical response: Introducing information entropy for enhancing the stability," *Neurocomputing*, vol. 626, 129569, April 2025.

[43] H.-L. Wei, "Boosting wavelet neural networks using evolutionary algorithms for short-term wind speed time series forecasting," in *Adv. Comput. Intell. IWANN* 2019. *Lecture Notes in Computer Science*, vol. 11506. Springer, Cham.

[44] H.-L. Wei et al., "An adaptive wavelet neural network for spatio-temporal system identification," *Neural Netw*., vol. 23, no. 10, pp. 1286–1299, Dec. 2010.

[45] Y. Zhao et al., "Tracking time-varying causality and directionality of information flow using an error reduction ratio test with applications to electroencephalography data," *Phys. Rev. E, Stat. Phys.*, vol. 86, no. 5, Nov. 2012.

[46] F. Cyr and P. S. Galbraith, "A climate index for the Newfoundland and Labrador shelf," Earth Syst. Sci. Data, 13. Pp. 1807–1828, May 2021.