
Article:
Gu, Y. and Wei, H.-L. orcid.org/0000-0002-4704-7346 (2023) Modelling short-term appliance energy use with interpretable machine learning: a system identification approach. Arabian Journal for Science and Engineering, 48 (11). pp. 15667-15678. ISSN 1319-8025

https://doi.org/10.1007/s13369-023-08084-1

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-023-08084-1

RESEARCH ARTICLE-SYSTEMS ENGINEERING

Modelling Short-Term Appliance Energy Use with Interpretable Machine Learning: A System Identification Approach

Yuanlin Gu[1] · Hua-Liang Wei[2,3]

Received: 1 June 2022 / Accepted: 18 June 2023

Abstract
The modelling and analysis of appliance energy use (AEU) of residential buildings are important for energy consumption control, energy management and maintenance, building performance evaluation, and so on. Although some traditional machine learning methods have been applied to produce good prediction results, these models are usually not interpretable, in that they fail to explain how appliance factors make contributions to the variation of AEU individually and interactively. Explicitly knowing the role played by each of the appliance factors in explaining AEU, however, is very important for energy saving. Motivated by this observation, this study introduces an interpretable machine learning approach which is built upon the nonlinear autoregressive moving average with eXogenous inputs model. The advantage of the proposed model is that in comparison with other state-of-the-art machine learning methods, for example, feedforward neural network, recurrent neural network (e.g., gated recurrent unit), and long short-term memory network, the established model is not only able to produce more accurate energy use prediction, but more importantly, also fully transparent and physically interpretable, clearly and explicitly indicating which factors significantly affect the variation of AEU. The findings of this study provide meaningful insights for improving the AEU efficiency.

Keywords Appliance energy use · Residual building · Modelling · Forecasting · Interpretable machine learning · NARMAX model

1 Introduction

Extensive attention has been paid to the analysis and modelling of appliance energy use (AEU) in the literature [1–3]. Revealing and establishing the inherent dependency relationship of AEU on potential drivers is very useful for energy control and management [4, 5], building performance analysis through simulations [6, 7], and energy consumption control [8]. Many methods have been proposed for AEU modelling and analysis, such as multiple linear regression [3], artificial neural networks [9, 10], outlier detection [11], support vector machines [12], and model ensembles [13]. AEU is determined by many factors, e.g., local temperature in and outside the house, humidity in the building, time of the day, just mention a few [2]. Some space climate factors such as solar radiation also make contributions to AEU [14]. Studies show that the occupants' behaviour is also an import factor that affects AEU [15, 16].

Many machine learning methods including artificial neural networks (ANNs) have been extensively applied to AEU modelling and analysis. Traditional neural networks usually comprise three or less hidden layers. Deep neural networks, which usually have many hidden layers, are more powerful for learning and representing nonlinear features [17, 18]. For most complex systems, well trained neural networks (with a proper number of hidden layers and trained with sufficiently large number of samples) can well capture the inherent nonlinear dynamics and provide good predictions of the future behaviour of complex systems. However, neural networks, as black-box models, have an important shortcoming: the lack of interpretability and explainability. In a neural network, all the input variables are usually collectively coupled into a huge number of hidden neurons (nodes). Although well-trained neural network models can usually show good

✉ Hua-Liang Wei
w.hualiang@sheffield.ac.uk

1 Faculty of Natural Sciences, University of Stirling, Stirling, UK

2 Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield, UK

3 Energy Institute, The University of Sheffield, Sheffield, UK

Published online: 12 July 2023

prediction performance, the resulting structure of such neural networks are internally informative to nobody (even if the model builders themselves). It is difficult to know which inputs or drivers are significantly important and which are not to the system output because the internal structure is opaque. Moreover, a significant amount of time may be needed to build a complex neural network model due to the inclusion of redundant input variables. Usually, it is impossible to know which of the individual input variables play an important role in determining the behaviour of the system, and how the interactions of the input variables affect the system behaviour. In addition, the inclusion of irrelevant input variables may lead to overfitting. In recent years, there have been extensive studies focusing on partially implementing explainable neural networks. For instance, a convolutional neural network (CNN) was developed to predict oil prices, production, consumption, and inventory based on online news [19], where the model was designed to be able to provide the impact values of input features, thereby offering insight into its functioning. Subsequently, an interpretable prediction system named VMD–ADE–TFT was developed for predicting wind speed, where the significance of variables can be evaluated. These systems can greatly enhance the interpretability of neural networks [20]. Nevertheless, for energy appliance prediction, which requires an even more interpretable approach, it is highly desirable to have fully transparent and interpretable models.

In comparison with complex neural networks, NARMAX model is much simple [21]. It uses an orthogonal forward regression (OFR) algorithms [22–25] or other methods, such as random search or input variable-informed approaches [26, 27] to detect the model structure. NARMAX model is transparent and parsimonious, making it easy and straightforward to understand and interpret the model response behaviour. NARMAX methods have been widely applied to system identification and data modelling in a wide range of multidisciplinary areas, including engineering [28–30], energy, ecological and environmental [31–34], space and geophysical [35–38], medical [39–41], biological and neurophysiological studies [42, 43]. A comparison of the features of NARMAX and neural networks is summarized in Table 1.

This study introduces an interpretable NARMAX modelling framework for predicting the energy use of appliances. The model is constructed with following considerations: to capture the inherent system dynamics in an explicitly transparent way, as well as generating accurate predictions. For comparison purposes, the performance of the obtained NARMAX model is compared with three state-of-the-art neural networks, namely, feedforward neural network (FNNs), gated recurrent unit (GRU, a special class of recurrent neural networks), and long short-term memory (LSTM) network.

The main contributions and novelty of the paper are as follows:

- It proposes a transparent and explainable machine learning model for appliance energy use pattern analysis.
- Unlike other machine learning methods which work in a black-box manner, the proposed model is completely explainable due to its transparent, interpretable, reproducible and parsimonious (TRIP) properties. It explicitly tells which factors significantly affect the appliance energy use, and reveals the relationship between appliance energy use and external factors.
- The performance of the proposed model is comparable to that of the state-of-the-art machine learning methods with regards to prediction accuracy when measured by either the normalized root mean square error (NRMSE) or weighted mean absolute percentage error (WMAPE).

The above features possessed by NARMAX model are highly attractive and crucially useful in cases where the primary modelling task is to establish an explicit quantitative representation showing which input variables are important and how the response variable depends on these important predictors. In such cases, it usually requires that models should be transparent and interpretable, but meanwhile the models should have good prediction ability.

The remainder of this paper is organized as follows. Section 2 briefly reviews the NARMAX model. A brief description of the data used is presented in Sect. 3. The experimental results are presented in Sect. 4. Discussions are given in Sect. 5, and finally the work is concluded in Sect. 6.

2 The NARMAX Model

Consider a process with one output (response) and r inputs (independent predictors), for which the NARMAX presentation can be written as:

$$\begin{aligned} & y(t) \qquad (1)\\ & = f\left[y(t-1), \ldots, y\left(t-n_y\right), u_1(t-d), u_1(t-d-1),\right.\\ & \ldots, u_1\left(t-n_u\right), u_2(t-d), u_2(t-d-1), \ldots,\\ & u_2\left(t-n_u\right), \ldots, u_r(t-d), u_r(t-d-1),\\ & \left.\ldots, u_r\left(t-n_u\right), e(t-1), \ldots e\left(t-n_e\right)\right]+e(t) \end{aligned}$$

where $u_k(t)(k = 1, 2, \ldots, r)$, $y(t)$ and $e(t)$ are the system inputs, output and noise, respectively; n_u, n_y, and n_e are the associated maximum time lags; d is the time delay, and for many processes the time delay can be set as $d = 0$ or $d = 1$ (in this study, d is set to be zero); $f[\cdot]$ is an unknown function that needs to be built from available training data. The NARX model, which does not include the noise moving-averaging

Table 1 Basic properties of NARMAX and other neural network models

Feature	NARMAX	Neural networks
Model structure	Clear, specific mathematical structure	Data-driven with artificial neurons
Model complexity	Transparent linear-in-the-parameters with a small number of parameters	Complex network structure with a large number of parameters
Model transparency	Fully transparent	Not transparent or partially transparent
Model interpretability	High, produces interpretable models	Low, can be less interpretable due to complex internal structure
Simulatability	Model results are repeatable	Simulations are not straightforward
Accuracy	Good, but limited by assumed mathematical structure	High, but highly depending the data size
Training data	Can work with either small or large datasets	Training data size should be large enough
System flexibility	Limited, works best for specific types of systems	High, can model a wide range of systems
Training time	Relatively fast, due to limited complexity	Can be slow, due to complex internal structure
Strength	Interpretation, explanation, and prediction	Prediction

model elements $e(t-1), \ldots, e(t-n_e)$, is a special case of the NARMAX model, and can be written in a linear-in-the-parameters form[21]:

$$y(t) = \sum_{m=1}^{M} \theta_m \varphi_m(t) + e(t) \tag{2}$$

where $\varphi_1(t) \ldots \varphi_r(t)$ are the model terms generated from the regressor vector $X(t) = [y(t-1), \ldots, y\big(t-n_y\big), u_1(t-d), \ldots, u_1(t-n_u), u_2(t-d), \ldots, u_2(t-n_u), \ldots, u_r(t-d), \ldots, u_r(t-n_u)]^T$, θ_m are model parameters and M is the number of model elements (that is, candidate model terms). The identification of NARMAX model consists of several key steps including model structure determination, model parameter estimation, model validation, model explanation, and prediction. Detailed description of these steps may be found in [21].

2.1 Identification of the NARX Model

In this study, the OFR algorithm is used to build compact models. The detailed implementation of the OFR algorithm can be found in [21, 24] or [25]. For ease of reading and facilitating the understanding of the OFR algorithm, the associated pseudocode is provided in the Appendix at the end of the paper. The basic idea of the OFR algorithm is to use a simple and effective error reduction ratio (ERR) index, to measure the contribution of each model term to explaining the variation in the system output. Let $D = \{\varphi_1(t), \ldots, \varphi_M(t)\}$ be the dictionary of all the candidate model terms and $D_n = \{\varphi_{l_1}(t), \ldots, \varphi_{l_n}(t)\}$ be the selected significant model terms, the final NARX model can be identified by the OFR algorithm, as:

$$y(t) = \sum_{i=1}^{n} \theta_{l_i} \varphi_{l_i}(t) + e(t) \tag{3}$$

where $l_1, \ldots, l_n$ is the index of the selected model terms and $\theta_{l_i} (i = 1, 2, \ldots n)$ is the estimated parameters. Note that the number of selected model terms, n, is usually much smaller than the number of the candidate model terms, M, so that the final identified NARX model is much simpler and easier to use. With all the significant model terms selected and ranked by ERR index, the importance of model terms can be measured and revealed.

As mentioned early, an attractive advantage of NARX model is that it is fully transparent, that is, it can be explicitly known how predictors are coupled or interacted in the model, and how important of each of the model elements is for explaining the change of response variable.

2.2 Noise Modelling

With the time delay $d = 0$ (between the model inputs and output), the model residual signal $\varepsilon(t)$ can be estimated as

$$\varepsilon(t) = y(t) - \hat{y}(t) \tag{4}$$

where $\hat{y}(t)$ represents the value of output at the time instant t. To refine the model by reducing the impact of the noise, the NARMAX method uses an extended least squares (ELS) scheme to estimate the prediction errors $\varepsilon(t)$ and uses the estimates to update the model structure (e.g. adding noise model terms in the model and update the model structure) [21]. A NARMAX model can be developed based on the associated NARX model by including these moving average

terms as follows:

$$y(t) = f^{[p]}(X_1(t)) + f^{[pn]}(X_2(t)) + f^{[n]}(X_3(t)) + \varepsilon(t) \quad (5)$$

with

$$X_1(t) = [y(t-1), ..., y(t-n_y), u_1(t), u_1(t-1), ..., u_1(t-n_u), ..., u_r(t), u_r(t-1), ..., u_r(t-n_u)]^T \quad (5a)$$

$$X_2(t) = [y(t-1), ..., y(t-n_y), u_1(t), ..., u_1(t-n_u), ..., u_r(t), ..., u_r(t-n_u), \varepsilon(t-1), ..., \varepsilon(t-n_e)]^T \quad (5b)$$

$$X_3(t) = [\varepsilon(t-1), \varepsilon(t-2), ..., \varepsilon(t-n_e)]^T \quad (5c)$$

where $f^{[p]}(\cdot)$ represents the identified NARX model, $f^{[pn]}(\cdot)$ represents the coupled process-noise sub-model, and $f^{[n]}(\cdot)$ represents the noise process sub-model, which are built based on their own regressor vectors, respectively. A simple and fast but less effective way is to use a linear moving average model below:

$$f^{[n]}(\cdot) = \alpha_1 \varepsilon(t-1) + \ldots + \alpha_{n_e} \varepsilon(t-n_e) \quad (6)$$

If a noise model is insufficient, then lagged noise variables $\varepsilon(t-p)$ for $p = 1, 2, \ldots n_e$ should be included in model (2) and (3) in a nonlinear manner. In summary, the basic regressor vectors in (2) and (3) includes all the lagged outputs, inputs and noise variables:$y(t-1), \ldots, y\big(t-n_y\big), u_1(t), u_1(t-1), \ldots, u_1(t-n_u), \ldots, u_r(t), u_r(t-1), \ldots, u_r(t-n_u), \varepsilon(t-1), \ldots, \varepsilon(t-n_e)$.

3 Data Description

The AEU dataset used in this study is obtained from [2]. It involves one response variable, AEU, and a total of 28 predictors (independent variables) such as *energy use of light fixtures in the house*, *Humidity in living room area*, *temperature in laundry room area*, *Humidity in parent room*, *and number of seconds from midnight*. Detailed descriptions of these 28 predictors can be found in [2].

The sampling period for AEU and all the 28 predictors is 10 min. In this study, the sampling period of 10 min was selected based on the typical usage patterns of the appliances being studied and the resolution required for the analysis [2]). The measurement time window is 11 January (17.00) to 27 May (18.00), 2016, making the total number of samples be 19,736. The entire dataset is split to two parts: 75% for model estimation (training) and another 25% for model performance test. The entire 19,736 samples of the AEU are shown in Fig. 1, and the values of AEU in a typical week (Monday – Sunday, 22—28 Feb, 2016) are shown in Fig. 2. From the two figures, it can be observed that there are a few of daily peak periods for AEU, whose values seem obviously lower at weekends (especially on Sunday) than other days.

4 Experimental Results

4.1 The Identified NARMAX Model

Based on pre-modelling experiments and simulations, the settings for building NARMAX models are as follows: (1) the maximum time lag for input variables $n_u = 2$; (2) the time lag for the response variable (the AEU) $n_y = 2$; and (3) polynomials are used as the elementary building blocks to build models, and the nonlinear degree of polynomials is set to be 2. Larger time lags lead to more extensive search of the most important and appropriate model terms from a relatively large candidate dictionary, while smaller time lags allow for faster training process. These parameters were chosen based on a balance between efficiency and accuracy. In this study, following the method proposed in [24], we conducted a series of pre-experiments with various initial hyperparameters and found that the selected values yielded the best performance on the validation set.

The candidate model input vector is:

$$\vartheta(t) = [y(t-1), y(t-2), u_1(t), u_1(t-1), u_1(t-2), \ldots, u_{28}(t), u_{28}(t-1), u_{28}(t-2)]^T \quad (7)$$

where $u_m(t)$, $u_m(t-1)$, $u_m(t-2)(m = 1, \ldots, 28)$ represent the 28 predictors and their time lagged versions, and $y(t-1)$, $y(t-2)$ represent the previous values (10 and 20 min before) of AEU. Initially, the full dictionary that used to build models consists of a total of 1770 elements including all the 28 predictors and all the cross-product terms.

The model complexity (i.e., the number of terms to be included in the final model) is determined by the APRESS criterion [44]. The APRESS value is determined by two components. The first measures the prediction error, and the second one penalizes the model when more model terms are added. Therefore, APRESS decreases at first when model terms are added, and gradually increases with the increase of the model complexity due to the penalty. Thus, the numbers of model terms at these turning points provide a good suggestion on the determination of how many model terms should be included in the final model. As shown in Fig. 3, APRESS has three relatively obvious turning points at 4, 7 and 14, which suggest that the optimal number of model terms can be 4, 7 or 14, where APRESS starts to increase

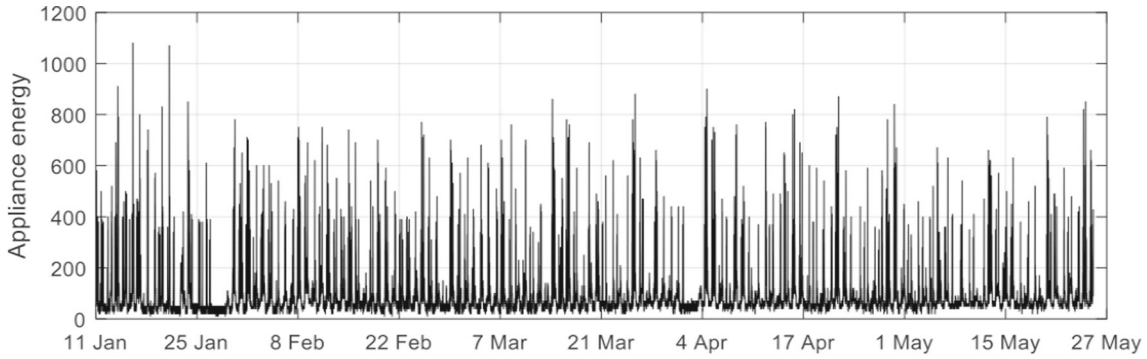

Fig. 1 Graphical illustration of the AEU between 1 January and 27 May, 2016

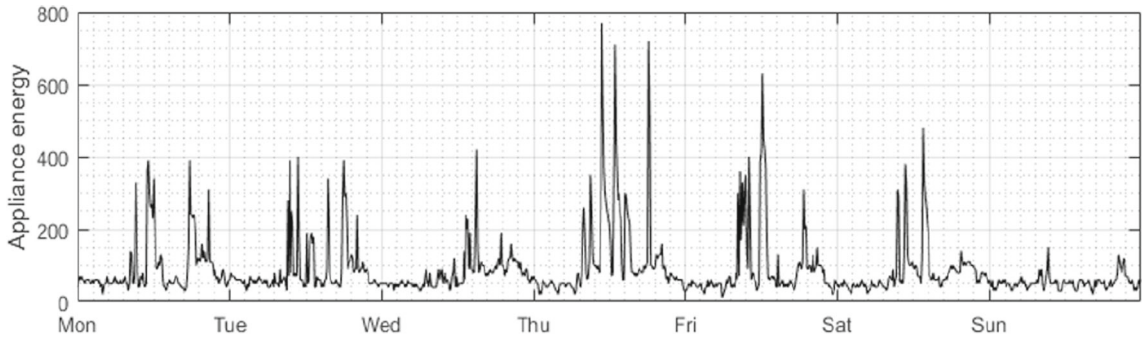

Fig. 2 An illustration of the AEU in a period of a typical week

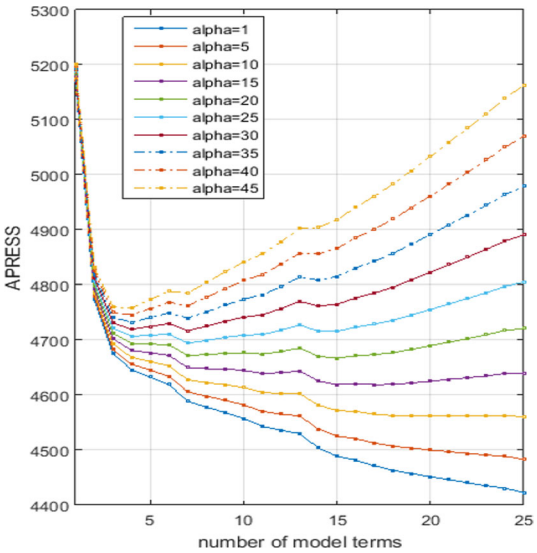

Fig. 3 Number of model terms versus APRESS value (alpha: tuning parameter)

after decreasing for several iterations. Further analysis (e.g. pre-modelling experiments and simulations) suggests that a total number of 7 model terms is a good choice, therefore the identified 7 terms are used to construct the NARX model. The identified model structure is shown in Table 2.

In Table 2, the contributions made by the individual model terms to explaining the change of the AEU, measured by the ERR index, are also given in column 3. The model reported in Table 2 should be written as:

$$AEU(\mathrm{t}) = 0.95526AEU(\mathrm{t}-1) - 0.0003801AEU(\mathrm{t}-1) \times AEU(\mathrm{t}-2) + \ldots \quad (8)$$

Note that these identified model terms provide information of which appliance factors or variables are involved in the model building. The time dependencies between the prediction and explanatory variables are indicated by time lags. For example, $Lights\,(\mathrm{t}) \times H_{pr}(\mathrm{t}-2)$ indicates that the interaction variable of *light fixtures* measured at current time and *humidity in parents room* measured 20 min ago plays an important role in explaining the variation of the appliance energy use.

4.2 Model Performance and Comparisons

To evaluate the performance of the identified NARMAX model, we compare its predictive ability with that of feedforward neural network, long short-term memory (LSTM) network, and gated recurrent unit (GRU) on the same dataset. The feedforward neural network contains one input layer, three fully connected layers and one output layer. The LSTM contains one input layer, three LSTM layers and regression layer. The GRU contains one input layer, three GRU layers and regression layeræ.

Table 2 The best model structure and model terms

No	Model term	ERR (100%)	Parameter	t-statistics
1	$AEU(\mathrm{t}-1)$	75.5866	9.5491×10^{-1}	61.680
2	$AEU(\mathrm{t}-1) \times AEU(\mathrm{t}-2)$	1.8701	-3.8022×10^{-4}	17.520
3	$AEU(\mathrm{t}-1) \times AEU(\mathrm{t}-1)$	0.4688	-1.4187×10^{-4}	5.3132
4	$Lights(\mathrm{t}) \times H_{pr}(\mathrm{t}-2)$	0.1446	1.2402×10^{-1}	6.9959
5	$Lights(\mathrm{t}) \times H_{lr}(\mathrm{t})$	0.0657	-1.1023×10^{-3}	6.0768
6	$T_{lr}(\mathrm{t}) \times N_s(\mathrm{t})$	0.0664	3.3829×10^{-5}	11.717
7	$N_s(\mathrm{t}) \times N_s(\mathrm{t}-2)$	0.1479	-7.8808×10^{-9}	10.054

$Lights =$ energy use of light fixtures in the house (unit: Wh)
$H_{pr} =$ Humidity in parent room (unit: %)
$H_{lr} =$ Humidity in living room (%)
$T_{lr} =$ Temperature in laundry room area (unit: °C)
$N_s =$ Number of seconds from midnight (unit: second)

Table 3 Model performance on test data

Model	CC	PE	NRMSE	WMAPE
NARMAX model	0.7502	0.5619	0.0707	0.3872
Feedforward neural network	0.7383	–	0.1706	0.4241
LSTM	0.7312	0.5345	0.0729	0.4328
GRU	0.7278	0.5283	0.0737	0.7419

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

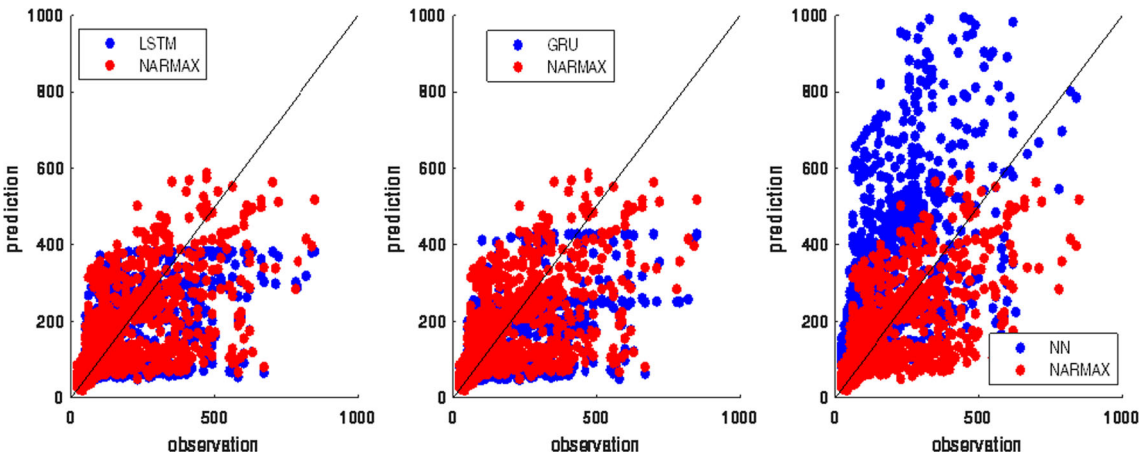

Fig. 4 The scatter plot between the model predicted AEU and the actual observations

The three metrics, correlation coefficient (CC), prediction efficiency (PE), and normalised root mean square error (NRMS), for the NARMAX model, feedforward neural network, LSTM, GRU are presented in Table 3, and the scatter plots between the measured AEU and the predictions from NARMAX, feedforward neural network, LSTM and GRU are shown in Fig. 4. The correlation coefficient, prediction efficiency and NRMSE of the NARMAX model on test dataset are 0.7502, 0.5619 and 0.707, respectively. From the results, the NARMAX model outperforms the other three models with regard to prediction accuracy. More importantly, the NARMAX model is advantageous over the neural networks in that the former is transparent, parsimonious and interpretable.

From Fig. 4, it can be observed that the model prediction errors are significantly large if the actual AEU values are large; this is especially true for the feedforward neural network model. This is a typical observation occurring with most machine learning methods when modelling non-stationary dynamic processes, e.g., a case where (1) the maximum amplitude of the signal is much larger than the minimum amplitude, and (2) most of the time series values are small ('normal' periods) and only a small number of values are very

large ('peak' period'). For such a case, many models may show very good performance for 'normal' periods but may not work well for capturing key behaviors in 'peak' periods. However, ideally, a good model should be able to show satisfactory performance over the entire prediction period rather than only show well fellow the 'normal' values. In fact, for many real applications an accurate prediction of the 'peak' values may be of more interest than that of the prediction of 'normal' values.

To fairly evaluate a model performance for predicting the energy use of not only during the normal periods, but also the peak periods, the weighted mean absolute percentage error (WMAPE), along with the commonly used NRMSE, is considered in this study. The WMAPE is calculated as:

$$WMAPE = \frac{\sum_{t=1}^{N} w_t |e(t)|}{\sum_{t=1}^{N} w_t |y(t)|} = \frac{\sum_{t=1}^{N} w_t \left|y(t) - \hat{y}(t)\right|}{\sum_{t=1}^{N} w_t |y(t)|} \quad (9)$$

where $y(t)$ is the observed appliance energy consumption, $\hat{y}(t)$ is the predicted appliance energy consumption, and w_t is the validation weight for the prediction at time point t. In this study, we define $w_t = 0.7$ for samples where the values of appliance energy consumption are no smaller than 400 and $w_t = 0.3$ for samples where the values of appliance energy consumption are lower than 400. The WMAPE values calculated from NARMAX, feedforward neural network model, LSTM and GRU models are presented in Table 3, where it can be noted that whilst the values of the three metrics, three metrics, correlation coefficient, prediction efficiency and normalised root mean square error are close and comparable, the WMAPE value of the NARMAX model is much smaller than those of the other three models, indicating that the performance of NARMAX model for peak periods is significantly better than those of the other models. A comparison between the predicted AEU values from the NARMAX model and the actual measurements is shown in Fig. 5, from which it can be seen that the NARMAX model well captures the system dynamics. For a closer inspection of the prediction performance, the prediction errors for each of the five periods illustrated in Fig. 5 are shown in Fig. 6. The discrepancy between the predicted and observed values looks large, this is understandable as accurately predicting AEU at peak times is always challenging and difficult, no matter which method is used. However, it should be noted that the NARMAX model demonstrates a degree of improvement in comparison to other models, as evidenced by the scatter plots shown in Fig. 4. Overall, The NARMAX model outperforms the other three methods over the entire prediction period of the test data.

5 Model Interpretability

Our results indicate that the dynamic change of AEU can be well captured by the NARMAX model, which shows better performance than the three compared machine learning methods (e.g., feed neural network, LSTM, GRU). As shown in Table 2, one of the advantages of the NARMAX model is that it is fully transparent, making it easy to explain and interpret. The detection of significant variables and terms is important because this can potentially significantly reduce the time and cost of data for data collection and investigation. The interpretability of NARMAX models can be understood from two perspectives as follows. Firstly, the model terms are intelligible, providing clear information about the variables appearing in the model, e.g., the time lags give insight into the temporal dependencies between the prediction variables and inputs. For instance, the term "*Hpr(t − 2)*" represents the humidity in the parent room measured 20 min prior, while "*Lights(t)* × *Hpr(t − 2)*" signifies the interaction between the current energy usage of light fixtures in the house and the humidity in the parent room 20 min ago. Secondly, the models are constructed using linear-in-the-parameter terms, where the interactions between individual variables and the connections between model terms are completely clear. All this makes the model comprehensible and adaptable to new data and new applications.

The model reported in Table 2 suggests that AEU is closely related to some weather and house conditions at the present time and 10 and 20 min earlier as well. Specifically, the result given in this study shows that the AEU value at present time is highly correlated with that of its history values (e.g., 10 and 20 min earlier as shown by the first three model terms given in column 2 in Table 2). This finding is consistent with that reported in other studies (e.g. [45].) that energy prediction is correlated with historical data. The inclusion of *energy use of light fixtures in the house* in the 4th and 5th model terms probably can be understood that a large fraction of AEU is attributed to the lightning fixtures. Humidity in the house (in the living room and parent room), coupled with the lightning fixtures, might also makes a contribution to explaining AEU. This result is consistent with that reported in [2] that lightning is one of the most significant source of energy consumption.

Our finding that *humidity* plays an important role in AEU is also strongly supported by other studies, e.g., [46], where it is suggested that humidity control material has a great impact on the energy performance of buildings. The last two model terms are related to the *number of seconds from midnight*. It is reasonable that AEU is highly associated with the specific time of the day. This finding is consistent with that reported in some most recent studies (see, e.g., [47, 48]), showing that occupant behaviour has significant impact on building energy consumption. The NARMAX method does not include any predictors that are directly related to weather. This probably

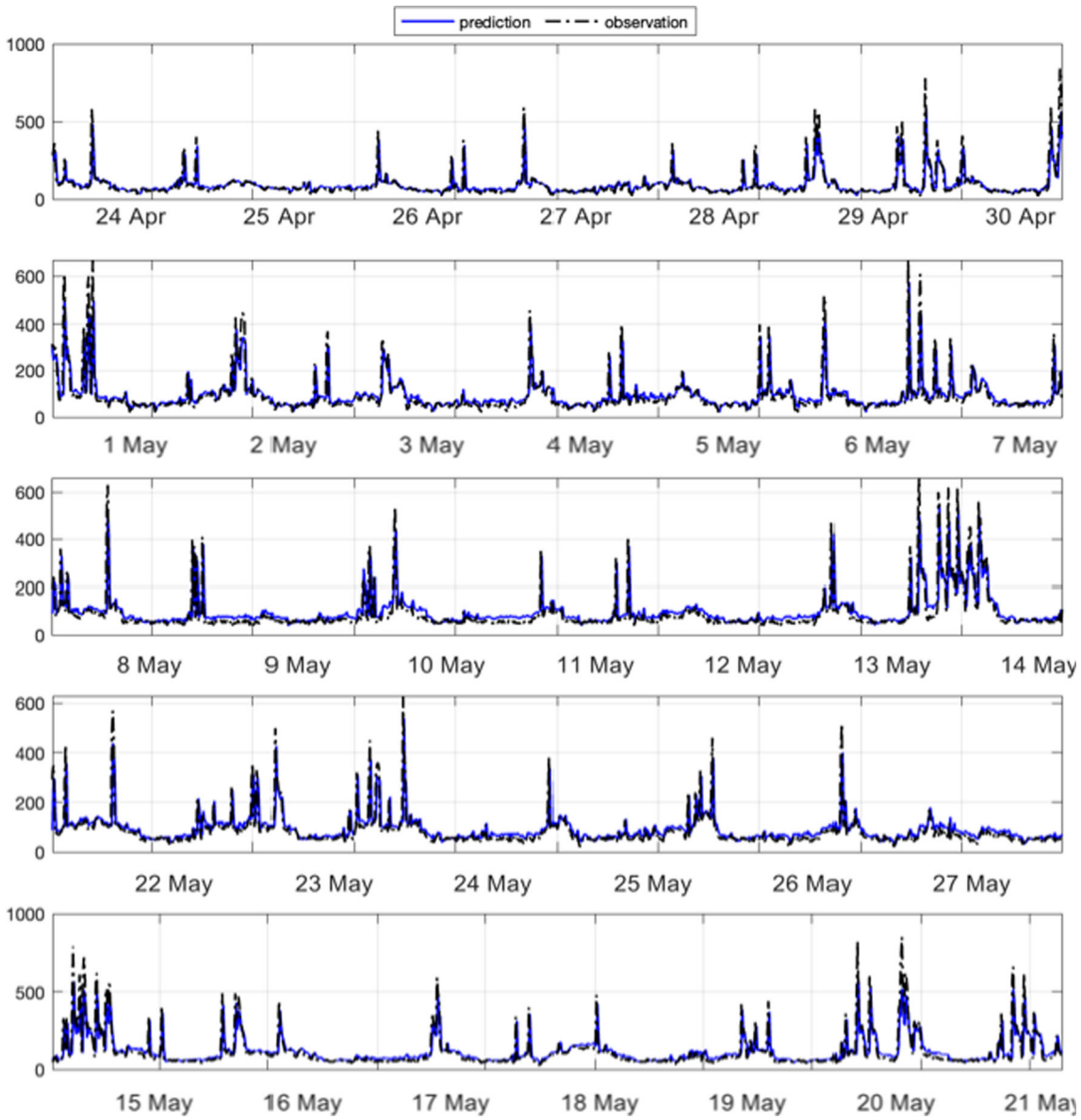

Fig. 5 A comparison between the predicted AEU from the NARMAX model and the actual measurements over the test data

because that at different seasons the AEU of different types of buildings may not always be equally sensitive to the change of weather as explained in [14].

The resulting NARMAX model in this study shows that the following factors play a clear role in explaining the total AEU: *light fixtures*, *humidity in living room*, *temperature in laundry room*, *humidity in parent room*, and *the number of seconds from midnight*. Moreover, the results in this study add important contribution to knowledge by explicitly revealing the interaction variables (cross-product terms), which were ignored in most previous studies with multiple linear regression where only single individual predictors were included in the models as determinants of the appliance energy use.

It is worth mentioning that although neural network models can produce good prediction results, they cannot give clear indications of which inputs are important and which are not for explaining the change of the response variable [9, 10, 49]. Neural network models usually use all available model input variables together regardless of their significance. A consequence of allowing redundant variables to participate the model training process is that the resulting models may be overfitting and lack generalization ability. There are some interpretable machine learning models that have been applied to AEU modelling and analysis, however,

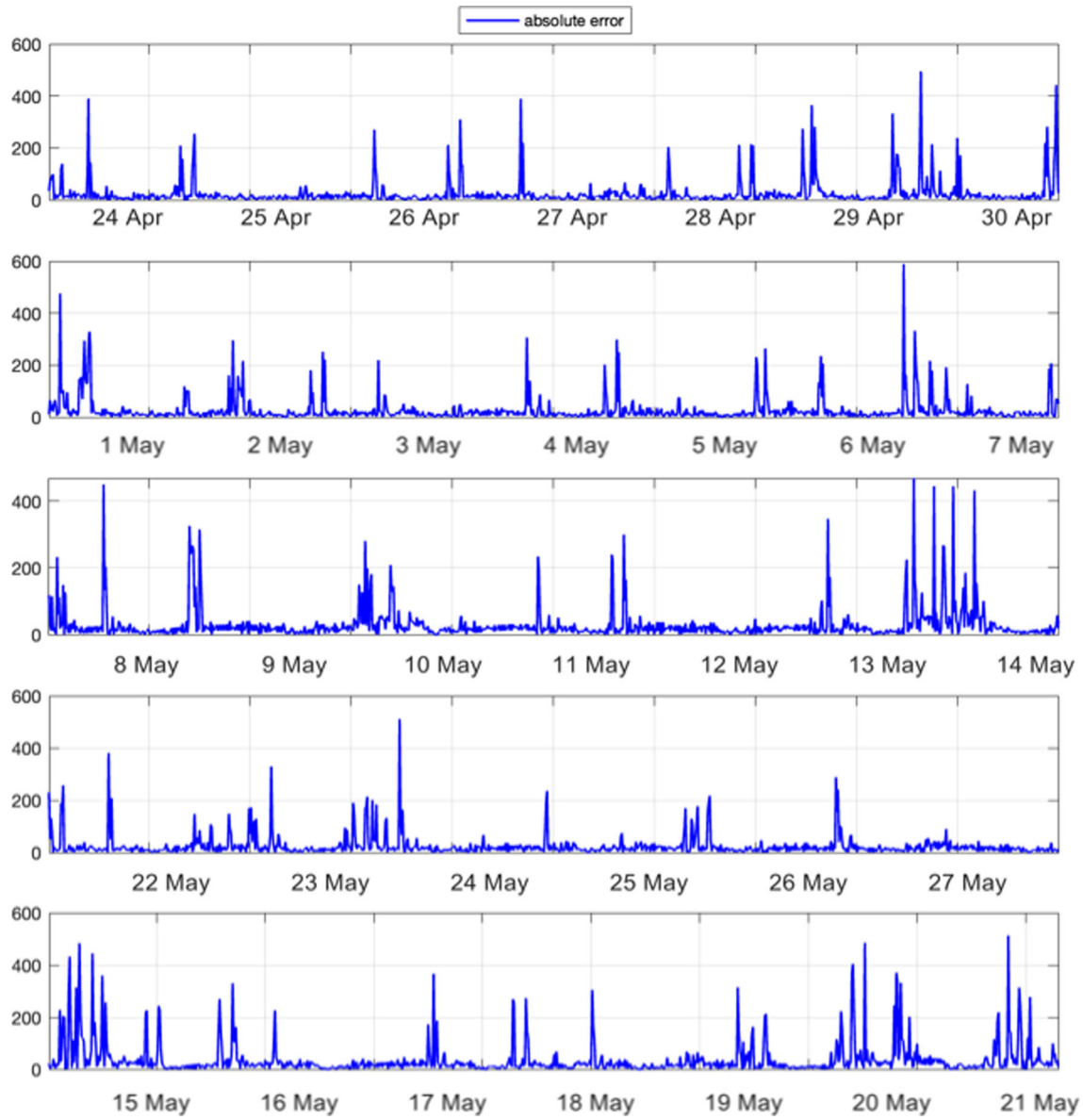

Fig. 6 Errors between the predicted AEU values from the NARMAX model and the actual measurements over the test data

the prediction performance of these models is lower than that of neural networks [2]. The NARMAX method used in this study is unique in comparing with other AEU modelling methods: the resulting models are transparent, parsimonious and interpretable, and meanwhile maintain excellent predictive capability.

6 Conclusions

In the field of building energy consumption, interpretability and accountability of machine learning models are highly desirable and demanded. While traditional machine learning techniques have been widely applied to achieve accurate prediction outcomes, the models used often lack interpretability, failing to explain how the individual appliance contributes to the overall energy use, and interactively and collectively determine the overall AEU patterns. Motivated by these observations, this study developed NARMAX models for explaining and predicting the energy consumption. The proposed method outperforms other state-of-the-art machine learning models, including feedforward neural networks, LSTM, and GRU, in that resulting NARMAX model exhibits a clear dependent relationship between AEU and the associated factors. More importantly, the model developed can be

used to analyse and isolate the sources of energy consumption, such as specific rooms and appliances, and the time lags in historical consumption measurements. Such information is particularly highly valuable for identifying the most important factors from a large number of potential variables. The findings of this study can be used by public organizations and governments for developing more effective energy policies for new buildings.

A limitation of this paper is that it only investigates the impact of 28 predictors. Other factors, such as lockdowns and climate changes can also be considered to improve the model predictive and explanatory ability. Therefore, one of the future research directions will be to collect more data and investigate the effectiveness of other potential predictors, and hence further improve the model predictive ability and interpretability.

Acknowledgements We would like to thank Dr. Luis M. Candaned for sharing the AEU data.

Funding The research work was supported in part by EPSRC (Ref. EP/H00453X/1 and Ref. EP/I011056/1).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article.

Appendix

The pseudo-code of the orthogonal forward regression (OFR) algorithm [21, 25] is given below.

Input vector $\boldsymbol{y}$, candidate terms $\boldsymbol{\varphi}_i$ with $i = 1, \ldots, M$
Initialize the adjustable parameter α
Set $\boldsymbol{r}_o \leftarrow \boldsymbol{y}$
Set $APRESS[0] \leftarrow 0$
Set $\boldsymbol{q}_i^{(1)} \leftarrow \boldsymbol{\varphi}_i$
Initialize $s \leftarrow 1$
while $APRESS[s] < APRESS[s-1]$ do
    Compute $ERR_i^{(s)} \leftarrow \frac{[\boldsymbol{y}^T \boldsymbol{q}_i^{(s)}]^2}{(\boldsymbol{y}^T \boldsymbol{y})(\boldsymbol{q}_i^{(s)T} \boldsymbol{q}_i^{(s)})}$ for $i = 1,2, \ldots, M$
    Find $l_s \leftarrow arg \max_{1 \le j \le M, j \notin l} \left\{ ERR_j^{(s)} \right\}$
    Assign $\boldsymbol{q}_s \leftarrow \boldsymbol{q}_{l_s}^{(s)}$
    Compute $\boldsymbol{q}_i^{(s+1)} \leftarrow \boldsymbol{q}_i^{(s)} - \frac{\boldsymbol{q}_i^{(s)T} \boldsymbol{q}_s}{(\boldsymbol{q}_i^{(s)T} \boldsymbol{q}_s)} \boldsymbol{q}_s$ for $i = 1,2, \ldots, M$
    Compute $\|\boldsymbol{r}_s\| \leftarrow \|\boldsymbol{r}_{s-1}\| - \frac{[\boldsymbol{r}_{s-1}^T \boldsymbol{q}_s]^2}{(\boldsymbol{q}_s^T \boldsymbol{q}_s)}$
    Compute $APRESS[s] \leftarrow \frac{1}{1 - \frac{s \times \alpha}{N}} \times \frac{\|\boldsymbol{r}_s\|}{s}$
    Update $s \leftarrow s + 1$
end while
Update selected terms as $\boldsymbol{\varphi}_l \leftarrow [\boldsymbol{\varphi}_{l_1}, \boldsymbol{\varphi}_{l_2}, \ldots, \boldsymbol{\varphi}_{l_n}]$
Compute weights $\boldsymbol{\theta}_l \leftarrow [\boldsymbol{\varphi}_l^T \boldsymbol{\varphi}_l]^{-1} \boldsymbol{\varphi}_l^T \boldsymbol{y}$
Output selected model terms $\boldsymbol{\varphi}_l$ and estimated parameters $\boldsymbol{\theta}_l$

References

1. Aksanli, B.; Akyurek, A.S.; Rosing, T.S.: User behavior modelling for estimating residential energy consumption. In: Leon-Garcia, A., Lenort, R., Holman, D., Staš, D., Krutilova, V., Wicher, P., Cagáňová, D., Špirková, D., Golej, J., Nguyen, K. (eds.) Smart City 360. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST), vol. 166, pp. 348–361. Springer, Cham (2016)
2. Candanedo, L.M.; Feldheim, V.; Deramaix, D.: Data driven prediction models of energy use of appliances in a low-energy house. Energy and Buildings **140**, 81–97 (2017). https://doi.org/10.1016/j.enbuild.2017.01.083
3. Vérez, D.; Borri, E.; Cabeza, L.F.: Trends in research on energy efficiency in appliances and correlations with energy policies. Energies **15**(9), 3047 (2022). https://doi.org/10.3390/en15093047
4. Zhao, P.; Suryanarayanan, S.; Simões, M.G.: An energy management system for building structures using a multi-agent decision-making control methodology. IEEE Trans. Ind. Appl. **49**(1), 322–330 (2013). https://doi.org/10.1109/TIA.2012.2229682
5. Barbato, A.; Capone, A.; Rodolfi, M.; Tagliaferri, D.: Forecasting the usage of household appliances through power meter sensors for demand management in the smart grid. In: 2011 IEEE International Conference on Smart Grid Communications, pp. 404–409 (2011). https://doi.org/10.1109/SmartGridComm.2011.6102356
6. Muratori, M.; Roberts, M.C.; Sioshansi, R.; Marano, V.; Rizzoni, G.: A highly resolved modeling technique to simulate residential power demand. Appl. Energy **107**, 465–473 (2013). https://doi.org/10.1016/j.apenergy.2013.02.057
7. Crawley, D.B.; Hand, J.W.; Kummert, M.; Griffith, B.T.: Contrasting the capabilities of building energy performance simulation programs. Build. Environ. **43**(4), 661–673 (2008). https://doi.org/10.1016/j.buildenv.2006.10.027
8. Pérez-Lombard, L.; Ortiz, J.; Pout, C.: A review on buildings energy consumption information. Energy Build. **40**(3), 394–398 (2008). https://doi.org/10.1016/j.enbuild.2007.03.007
9. Ekici, B.B.; Aksoy, U.T.: Prediction of building energy consumption by using artificial neural networks. Adv. Eng. Softw. **40**(5), 356–362 (2009). https://doi.org/10.1016/j.advengsoft.2008.05.003
10. Gonzalez, P.A.; Zamarreno, J.M.: Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. Energy Build. **37**(6), 595–601 (2005). https://doi.org/10.1016/j.enbuild.2004.09.006
11. Li, X.; Bowers, C.P.; Schnier, T.: Classification of energy consumption in buildings with outlier detection. IEEE Trans. Ind. Electron. **57**(11), 3639–3644 (2010). https://doi.org/10.1109/TIE.2009.2027926
12. Dong, B.; Cao, C.; Lee, S.E.: Applying support vector machines to predict building energy consumption in tropical region. Energy Build. **37**(5), 545–553 (2005). https://doi.org/10.1016/j.enbuild.2004.09.009
13. Fan, C.; Xiao, F.; Wang, S.: Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl. Energy **127**, 1–10 (2014). https://doi.org/10.1016/j.apenergy.2014.04.016
14. Fikru, M.G.; Gautier, L.: The impact of weather variation on energy consumption in residential houses. Appl. Energy **144**, 19–30 (2015). https://doi.org/10.1016/j.apenergy.2015.01.040
15. Masoso, O.T.; Grobler, L.J.: The dark side of occupants' behaviour on building energy use. Energy Build. **42**(2), 173–177 (2010). https://doi.org/10.1016/j.enbuild.2009.08.009
16. Yan, D.; O'brien, W.; Hong, T.; Feng, X.; Burak Gunay, H.; Tahmasebi, F.; Mahdavi, A.: Occupant behavior modeling for building performance simulation: current state and future challenges. Energy Build. **107**, 264–278 (2015). https://doi.org/10.1016/j.enbuild.2015.08.032
17. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Mohamed, A.R.; Jaitly, N.; Sainath, T.: Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. **29**, 82–97 (2012). https://doi.org/10.1109/MSP.2012.2205597
18. LeCun, Y.; Bengio, Y.; Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015). https://doi.org/10.1038/nature14539
19. Wu, B.; Wang, L.; Wang, S.; Zeng, Y.R.: Forecasting the U.S. oil markets based on social media information during the COVID-19 pandemic. Energy **226**, 120403 (2021). https://doi.org/10.1016/j.energy.2021.120403
20. Wu, B.; Wang, L.; Zeng, Y.R.: Interpretable wind speed prediction with multivariate time series and temporal fusion transformers. Energy **252**, 123990 (2022). https://doi.org/10.1016/j.energy.2022.123990
21. Billings, S.A.: Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains. Wiley, Chichester (2013)
22. Chen, S.; Billings, S.A.; Luo, W.: Orthogonal least squares methods and their application to non-linear system identification. Int. J. Control **50**(5), 1873–1896 (1989). https://doi.org/10.1080/00207178908953472
23. Aguirre, L.A.; Billings, S.A.: Improved structure selection for nonlinear models based on term clustering. Int. J. Control **62**(3), 569–587 (1995). https://doi.org/10.1080/00207179508921557
24. Wei, H.-L.; Billings, S.A.; Liu, J.: Term and variable selection for non-linear system identification. Int. J. Control **77**(1), 86–110 (2004). https://doi.org/10.1080/00207170310001639640
25. Wei, H.-L.; Billings, S.A.: Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information. Int. J. Model. Identif. Control **3**(4), 341–356 (2008). https://doi.org/10.1504/IJMIC.2008.020543
26. Retes, P.F.L.; Aguirre, L.A.: NARMAX model identification using a randomized approach. Int. J. Model. Identif. Control **31**(3), 205–216 (2019). https://doi.org/10.1504/IJMIC.2019.098779
27. Tavares, L.A.; Abreu, P.E.; Aguirre, L.A.: Nonlinearity compensation based on identified NARX polynomials models. Nonlinear Dyn. **107**(1), 709–725 (2022). https://doi.org/10.1007/s11071-021-06797-2
28. Tsai, J.S.H.; Wang, C.T.; Kuang, C.C.; Guo, S.M.; Shieh, L.S.; Chen, C.W.: A NARMAX model-based state-space self-tuning control for nonlinear stochastic hybrid systems. Appl. Math. Model. **34**(10), 3030–3054 (2010). https://doi.org/10.1016/j.apm.2010.01.011
29. Barbosa, B.H.; Aguirre, L.A.; Martinez, C.B.; Braga, A.P.: Black and gray-box identification of a hydraulic pumping system. IEEE Trans. Control Syst. Technol. **19**(2), 398–406 (2011). https://doi.org/10.1109/TCST.2010.2042600
30. Zhang, W.; Zhu, J.; Gu, D.: Identification of robotic systems with hysteresis using nonlinear AutoRegressive eXogenous input models. Int. J. Adv. Robot. Syst. **14**(3), 1–10 (2013). https://doi.org/10.1177/1729881417705845
31. Bigg, G.R.; Wei, H.-L.; Wilton, D.J.; Zhao, Y.; Billings, S.A.; Hanna, E.; Kadirkamanathan, V.: A century of variation in the dependence of Greenland iceberg calving on ice sheet surface mass balance and regional climate change. Proc. R. Soc. A Math. Phys. Eng. Sci. **470**(2166), 20130662 (2014). https://doi.org/10.1098/rspa.2013.0662

32. Marshall, A.M.; Bigg, G.R.; Van Leeuwen, S.M.; Pinnegar, J.K.; Wei, H.-L.; Webb, T.J.; Blanchard, J.L.: Quantifying heterogeneous responses of fish community size structure using novel combined statistical techniques. Glob. Chang. Biol. **22**(5), 1755–1768 (2016). https://doi.org/10.1111/gcb.13190
33. Ayala-Solares, J.R.; Wei, H.-L.; Bigg, G.R.: The variability of the Atlantic meridional circulation since 1980, as hindcast by a data-driven nonlinear systems model. Acta Geophys. **66**(4), 683–695 (2018). https://doi.org/10.1007/s11600-018-0165-734
34. Akinola, T.E.; Oko, E.; Gu, Y.; Wei, H.-L.; Wang, M.: Non-linear system identification of solvent-based post-combustion CO_2 capture process. Fuel **239**, 1213–1223 (2019). https://doi.org/10.1016/j.fuel.2018.11.097
35. Amisigo, B.A.; Van de Giesen, N.; Rogers, C.; Andah, W.E.I.; Friesen, J.: Monthly streamflow prediction in the Volta Basin of West Africa: a SISO NARMAX polynomial modelling. Phys. Chem. Earth Parts A/B/C **33**(1–2), 141–150 (2008). https://doi.org/10.1016/j.pce.2007.04.019
36. Balikhin, M.A.; Boynton, R.J.; Walker, S.N.; Borovsky, J.E.; Billings, S.A.; Wei, H.-L.: Using the NARMAX approach to model the evolution of energetic electrons fluxes at geostationary orbit. Geophys. Res. Lett. **38**(18), L18105 (2011). https://doi.org/10.1029/2011GL048980
37. Ayala-Solares, J.R.; Wei, H.-L.; Boynton, R.J.; Walker, S.N.; Billings, S.A.: Modeling and prediction of global magnetic disturbance in near-Earth space: a case study for Kp index using NARX models. Space Weather **14**(10), 899–916 (2016). https://doi.org/10.1002/2016SW001463
38. Gu, Y.; Wei, H.-L.; Boynton, R.J.; Walker, S.N.; Balikhin, M.A.: System identification and data-driven forecasting of AE index and prediction uncertainty analysis using a new cloud-NARX model. J. Geophys. Res. Space Phys. **124**(1), 248–263 (2019). https://doi.org/10.1029/2018JA025957
39. Billings, C.G.; Wei, H.-L.; Thomas, P.; Linnane, S.J.; Hope-Gill, B.D.: The prediction of in-flight hypoxaemia using non-linear equations. Respir. Med. **107**(6), 841–847 (2013). https://doi.org/10.1016/j.rmed.2013.02.016
40. Wei, H.-L.; Billings, S. A.: 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 (in press) (2022). https://doi.org/10.48550/arXiv.2111.01763
41. Beltran-Perez, C.; Serrano, A.A.A.; Solís-Rosas, G., et al.: A general use QSAR-ARX model to predict the corrosion inhibition efficiency of drugs in terms of quantum mechanical descriptors and experimental comparison for lidocaine. Int. J. Mol. Sci. **23**(9), 5086 (2022). https://doi.org/10.3390/ijms23095086
42. Li, Y.; Cui, W.G.; Guo, Y.Z.; Huang, T.; Yang, X.F.; Wei, H.-L.: Time-varying system identification using an ultra-orthogonal forward regression and multiwavelet basis functions with applications to EEG. IEEE Trans. Neural Netw. Learn. Syst. **29**(7), 2960–2972 (2018). https://doi.org/10.1109/TNNLS.2017.2709910
43. Gu, Y.; Yang, Y.; Dewald, J.P.; Van der Helm, F.C.; Schouten, A.C.; Wei, H.-L.: Nonlinear modeling of cortical responses to mechanical wrist perturbations using the NARMAX method. IEEE Trans. Biomed. Eng. **68**(3), 948–958 (2020). https://doi.org/10.1109/TBME.2020.3013545
44. Billings, S.A.; Wei, H.-L.: An adaptive orthogonal search algorithm for model subset selection and non-linear system identification. Int. J. Control **81**(5), 714–724 (2008). https://doi.org/10.1080/00207170701216311
45. Sun, Y.; Haghighat, F.; Fung, B.C.: A review of the-state-of-the-art in data-driven approaches for building energy prediction. Energy Build. **221**, 1100022 (2020). https://doi.org/10.1016/j.enbuild.2020.110022
46. Wu, Z.; Qin, M.; Zhang, M.: Phase change humidity control material and its impact on building energy consumption. Energy Build. **174**, 254–261 (2018). https://doi.org/10.1016/j.enbuild.2018.06.036
47. Chen, S.; Zhang, G.; Xia, X.; Chen, Y.; Setunge, S.; Shi, L.: The impacts of occupant behavior on building energy consumption: a review. Sustain. Energy Technol. Assess. **45**, 101212 (2021). https://doi.org/10.1016/j.seta.2021.101212
48. Amasyali, K.; El-Gohary, N.: Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings. Renew. Sustain. Energy Rev. **142**, 110714 (2021). https://doi.org/10.1016/j.rser.2021.110714
49. Somu, N.; Gauthama Raman, M.R.; Ramamritham, K.: A deep learning framework for building energy consumption forecast. Renew. Sustain. Energy Rev. **137**, 110591 (2021). https://doi.org/10.1016/j.rser.2020.110591