Input Vector Selection in NARX Models using Statistical Techniques to Improve the Generated Power Forecasting in PV Systems

Eduardo Rangel-Heras, Nun Pitalúa-Díaz, Pavel Zuniga, *Member, IEEE*, E. A. Hernandez-Vargas, *Member, IEEE* and Alma Y. Alanis*, *Senior Member, IEEE*

Abstract— **This paper uses collinearity and causality tests to choose variables for an input vector to forecast the electrical power generated by a photovoltaic system. The collinearity test determines redundant variables, and the causality test determines which variables cause the electric power. The chosen input vector is used to train nonlinear autoregressive models with external inputs neural networks (NARX-NN). We develop an algorithm to generate NARX models with an all variable combinations algorithm (AVCA) to validate the results. Finally, we compare the results of the proposed methodology against the best results obtained by the AVCA; the algorithm tests 502 input vectors with the NARX model to forecast 26 steps (a day ahead) of the electrical power. The best model chosen using the collinearity and causality techniques has an RMSE of 308 W for the electric power using four variables in the input vector; the best model using the AVCA has an RMSE of 305 W using five variables in the input vector. Results show that the collinearity and causality techniques are a direct way to select the input vector without affecting the model's performance and results in a reduction of the input vector length.**

Index Terms—**Neural Network, Electrical Power, Input Vector, Photovoltaic System, Collinearity and Granger Tests**

I. Introduction

Forecasting the solar irradiance and the power generated by (PV) systems is essential for control and management in smart grids. For this reason, many mathematical, numerical, statistical, and machine learning techniques have been implemented [1].

Generally, the best approaches are those models that implement several variables as input vectors. One of the problems of multivariable models is finding the best combination of variables to build the input vector. This problem increases when there are many variables. In this work, we propose using statistical techniques to find the optimal inputs in a NARX model to forecast the electric power of a PV system, focusing on the impact of the input vector on the network's performance; for this, we use nine variables. The approach starts by identifying redundant variables in the models using the collinearity test. The causality test is then used to find the variables containing helpful information to forecast the generated electric power in a photovoltaic system. Finally, we validate the results using NARX models developed with all combinations of the available variables and measure the performance using the MSE, RMSE, and R^2 indexes.

For best comprehension, we divide this work into five sections. Section II presents research works about input vector selection and NARX models applications. It is also pointed out where other authors do not study the selection of the input vector, leaving a gap that can be filled by the proposal in this work. Section III shows the implementation of mathematical tools and the development of the NARX models. Section IV presents the results and discussions. Finally, conclusions are presented in Section V.

II. Related works

Much research has been done to forecast solar irradiance and the power generated by PV systems. The reason is apparent, nowadays it is an essential field and much work is ahead. Different approaches are used for this purpose, such as statistical methods, artificial neural network (ANN) techniques, and a combination of several techniques (hybrid models) [2].

A few works have focused on the variables for input vectors in ANN models, for instance, Hocaoglu *et al.* successfully implemented Granger causality to determine which meteorological variables contain information that causes solar irradiance, but no forecasting model was implemented [3]. Eom *et al.* used a regression model and a significance level of 95% to remove irrelevant variables; they tested the input vector in VAR and long short-term memory (LSTM) models but did not try all combinations to validate their results [4]. In addition, the linear regression models do not consider the lag effect, which is essential in the time series analysis. Rangel *et al.* forecasted 1 hour ahead of solar irradiance and 24 hours ahead of electrical power for a PV module in [5], [6], respectively; they estimated the energy in the PV system using a single diode model. These authors performed an exhaustive analysis of the variables to build the input vector but did not apply the proposed techniques to a PV system; also, not all variable combinations were tested to validate the results.

E.3 Rangel-Heras, Universidad de Guadalajara, Blvd. Gral. Marcelino García Barragán 1421, Olímpica, C.P. 44430 Guadalajara, México, (e-mail: phdeduardorangelheras@gmail.com)

N. Pitalúa-Díaz, Departamento de Ingeniería Industrial, Universidad de Sonora, Blvd. Luis Encinas y Rosales S/N, Col. Centro, C.P. 83000, Hermosillo Sonora, México, (email: nun.pitalua@unison.mx)

P. Zuniga, Universidad de Guadalajara, Blvd. Gral. Marcelino García Barragán 1421, Olímpica, C.P. 44430 Guadalajara, México, (e-mail: pavel.zuniga@academicos.udg.mx)

E. A. Hernandez-Vargas, Department of Mathematics and Statistical Science, University of Idaho, Moscow, Idaho, USA, (e-email: esteban@uidaho.edu)

*Alma Y. Alanis, Universidad de Guadalajara, Blvd. Gral. Marcelino García Barragán 1421, Olímpica, C.P. 44430 Guadalajara, México, (e-mail: alma.alanis@academicos.udg.mx)

An example of a multivariable statistical model is presented by Yanting *et al.* that used an ARMAX model for daily electrical power prediction of a grid-connected PV system. The selection of inputs is based on the Bayesian information criterion (BIC), however, the information on other variable combinations is not reported [7].

Among the ANN paradigms, there are approaches to forecast solar irradiance based on neural networks, such as the feed-forward neural network (FFNN) proposed by Sharma *et al.* in [8], where they performed one hour and fifteen minutes ahead of predictions. Their contribution is the usage of Morlet and Mexican hat wavelets as activation functions in the hidden layer, employing values of the clear sky index as inputs of the FFNN model. However, this work does not consider the impact of other variables on solar irradiance behavior. Liping Liu *et al.* investigated the effects of PV electrical power variability and proposed a data-driven ensemble modeling technique to improve the prediction accuracy of PV power generation one day ahead. Disposing of six meteorological variables, they used SVM, multivariate adaptive regression spline (MARS), and MLP. The authors compared the performance of the mentioned models and concluded that the best results are obtained with a recursive arithmetic average ensemble model. The scope of the work does not include an analysis of the combination of variables for the input vectors [9]. Alkandari & Ahmad implemented deep learning (DL) and statistical methods to predict a day ahead of the electric power generated in a PV system; the DL model included LSTM, gated recurrent unit (GRU), and AutoEncoder LMST (Auto-LSMT), however, it did not included the analysis of the variables for the input vector [10]. Chuluunsaikhan *et al.* implemented machine-learning techniques to predict the power output of solar panels based on fifteen weather and air pollution variables; however, it lacks a deeper study of the input variables to determine the best combination of input vector space [11]. Dawan *et al.* compared electrical power output forecasting on the PV systems using adaptive neural-fussy inference system (ANFIS) and particle swarm optimization – artificial neural network, with temperature and solar radiation as inputs [12]. The authors only used two variables, ignoring the possible effects of other inputs; the best model reported was the ANFIS.

Finally, Ahmad *et al.* compared NARX models against MLP and ARMA; in the case of the multivariable models, twelve input vectors were tested [13]; again, many potential input vectors were not considered for the analysis. In a similar work, Hussain & Al-Alili compared a NARX model against an artificial network fuzzy inference system (ANFIS) using four meteorological variables to build the input vector [14]; in both cases the NARX models performed better than the ANFIS. NARX neural network models are powerful and popular to solve several problems in nonlinear control applications. NARX models are chosen for rapid training, convergence speed, and strong representativeness, and have been implemented successfully in [5], [6], [13], [15], and [16]. Louzazni *et al.* proposed two approaches using NARX models to estimate the electrical power output in a PV system using temperature and solar irradiance; the prediction was not considered in these works [15], [16].

Among hybrid models, Jimenez and Mora present a hybrid model that implements DT, ANNs, and SVMs to forecast hourly global solar radiation (GSR). However, the selection of variables for the input space is not considered as only a pair of input vectors are tested [2].

Azimi *et al.* integrated techniques to predict 1, 24, and 48 hours ahead of solar radiation, such as a modified k-means clustering algorithm and an MLP [17]. Boland *et al*. proposed three methods to forecast one-hour ahead solar radiation, combining statistical models, Fourier series, plus ARMA models [18]. Monjoly *et al.* also proposed hybrid models with multiscale decomposition, neural networks, and ARIMA techniques to predict hourly global solar radiation [19]. Chen & Kartini developed a k-NN-MLP model to predict 60 minutes before global solar irradiance employing past meteorological data [20]. The works mentioned before are univariate models, therefore, the impact of other weather variables is not considered.

Using DL models named FFNN, GRU-RNN, and LSTM-RNN, du Plessis *et al.* carried out short-term power forecasting and investigated the ability of DL models to represent the PV systems' behavior; the models forecasted 1-6 h for a 75 MW rated PV system. The best results were obtained with FFNN and GRU-RNN [21]. Input vectors are formed by power, six weather variables, and three geographical variables. They focused on the parameters of the models, but the effect of input vector variables is not considered.

In the literature, we found many multivariable models [2], [5]–[7], [9]–[16], [20]–[22]. Nevertheless, these authors focused their efforts on selecting the best model, finding the best model combination, and implementing a powerful DL model; other authors implemented univariate models like in [8], [18], [23], and [24]. The model selected is based on the following factors: the available data, the length of the data, the data resolution, and the inclination of the authors in choosing between techniques and methods.

As a result, there is still room to propose selecting the variables for the input vector. Therefore, this work proposes to implement two statistical techniques to identify the appropriate variables for the input vectors of a model. To validate the results of the proposed method, we developed an algorithm to test all variable combinations of input vectors.

III. Implementation of Mathematical Tools and Development of the NARX Models

This section presents the results of the data preprocessing, the collinearity and causality tests, and the development of the NARX models. First, possible missing data is completed by linear interpolation; refer to (1). Next, outlier values are identified by the z-score method using a threshold of three; refer to (2) [25]–[27].

$$\tilde{x}(t) = \frac{x_{k+1} - x_k}{t_{k+1} - t_k}(t - t_k) + x_k \tag{1}$$

where $\tilde{x}$ is the interpolated datum, x_k, x_{k+1} are the data points corresponding to t_k, t_{k+1}, and t is the time for the interpolated datum $\tilde{x}$.

$$Z_t = \frac{Y_t - \bar{Y}}{\sigma} \qquad (2)$$

where Z_t is the i^{th} transformed datum, Y_t is the i^{th} datum of the time-series, $\bar{Y}$ is the time-series average, and σ is the time-series standard deviation [28], [29].

Then, we apply the collinearity test, followed by the causality test [3], [30]–[32], to obtain the input vectors used to develop the NARX models.

A. Missing values and outliers (Data cleaning)

As previously mentioned, we apply a linear interpolation technique for missing data. The data sample used in this work is from August 22 to December 31, 2018. We configure the dataset by completing 26 values for each day from 6:00 to 18:30 hours to forecast 26 steps; the sample resolution is of 30 minutes completing a total data of 3,430.

There are nine variables, electrical power (EP), solar irradiance (SI), wind direction (WD), humidity (H), heat index (HI), pressure (P), dew point (DP), temperature (T), and wind speed (WS).

TABLE I shows the variation in percentage between the missing data with and without linear interpolation in descriptive statistics: mean, median, and standard deviation. Higher variation is observed for 30% and 34% of missing data, however, the difference is generally slight and the original time series patterns are used for the NARX models to learn the appropriate stochastic behavior.

TABLE I.
METRICS DIFFERENCE BETWEEN ORIGINAL AND LINEAR INTERPOLATION DATA

Month	Metric (%)	Variable								
		EP	SI	WD	H	HI	P	DP	T	WS
Aug	Missing data	20.0 %								
	Difference in μ	-6.9	-5.0	0.4	-0.6	0.0	0.0	-0.4	0.1	-0.6
	Difference in M	-6.6	-3.8	16.7	0.0	0.2	0.0	-0.4	0.1	0.0
	Difference in σ	1.6	-0.4	-3.2	-2.4	1.5	2.0	1.8	0.3	-10
Sep	Missing data	30.0 %								
	Difference in μ	-8.1	-6.1	-0.7	1.7	-0.4	0.0	0.4	-0.7	4.0
	Difference in M	-13.1	-12.4	0	1.9	-0.3	0.0	0.6	-0.8	0.0
	Difference in σ	1.5	1.5	-5.1	-1.9	-1.0	-1.2	-2.6	-1.8	5.9
Oct	Missing data	34.0 %								
	Difference in μ	-13.7	-8.5	1.2	3.0	-1.9	0.0	-1.8	-2.8	-0.2
	Difference in M	-20.0	-11.5	0	0.94	-2.8	0.0	-2.4	-3.9	0.0
	Difference in σ	0.71	-0.82	-4.4	3.6	2.9	11.6	-3.3	4.0	4.3
Nov	Missing data	27.0 %								
	Difference in μ	-6.1	-3.3	3.2	3.2	0.3	0.0	5.5	-0.7	-9.7
	Difference in M	-9.7	-2.1	0	0	-0.1	0.0	2.2	-0.3	-50.0
	Difference in σ	-0.6	2.2	-4.8	3.4	-0.4	-3.3	-6.7	0.5	-6.7
Dec	Missing data	6.0 %								
	Difference in μ	-4.2	-3.8	-0.6	1.2	-0.4	0.0	0.0	-0.8	-3.4
	Difference in M	-5.1	-5.4	0.0	0	-0.4	0.0	-0.6	-0.8	-50
	Difference in σ	2.2	2.0	-0.4	1.33	1.5	-1.0	-0.6	1.8	1.0

Once we completed the missing data step, we use the z-score to identify outlier values. TABLE II shows a summary of the results and the descriptive statistics with and without outliers. We only detect outlier values in five variables, HI, P, DP, T, and WS, where WS is the variable with the more significant number of outliers. It is clear that the variation between statistical metrics is slight, suggesting that outliers do not have a significant effect when the data length is large.

TABLE II.
STATISTICAL DESCRIPTIVE RESULTS WITH AND WITHOUT OUTLIERS

Var	Outliers	Without exclude the outliers				Excluding the outliers			
		Mean	Std	Skewness	Kurtosis	Mean	Std	Skewness	Kurtosis
EP (W)	0	1101.9	804.8	0.05	1.53	1101.9	804.8	0.05	1.53
SI (W/m^2)	0	500.1	346.4	0.06	1.55	500.1	346.4	0.06	1.55
WD (°)	0	164.8	73.6	0.06	1.68	164.8	73.6	0.06	1.68
H (%)	0	45.5	17.2	0.35	2.72	45.45	17.2	0.35	2.72
HI (°C)	6	27.9	9.3	0.65	5.68	27.8	8.9	0.14	2.15
P (inHg)	9	29.2	0.1	0.21	3.23	29.2	0.1	0.30	3.05
DP (°C)	3	12.9	8.4	-0.43	2.37	12.9	8.4	-0.42	2.34
T (°C)	10	26.8	6.8	-0.20	2.49	26.9	6.7	-0.15	2.37
WS (m/s)	85	2.9	1.6	1.35	7.37	2.7	1.3	0.30	2.96

Fig. 1 presents the iterative way the z-score detects outliers equal to or greater than the chosen value of three, once the data is standardized with the z-score (the mean and standard deviation are recalculated when an outlier is found and deleted).

The outliers are replaced with the average of the corresponding variable.

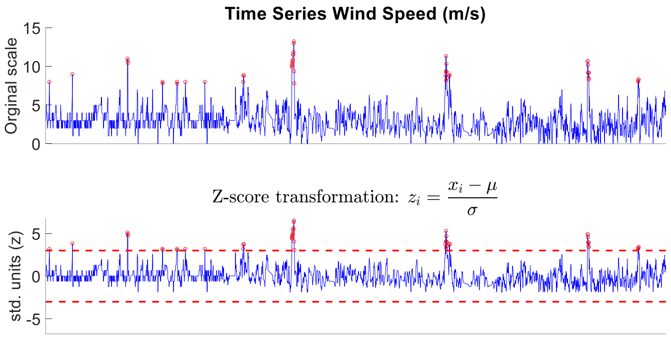


Fig. 1. Iterative way to detect outliers for wind speed.

B. Generating of the Input Vectors First Step

In this section, the collinearity test is applied to the time series. This procedure aims to establish a first approach for building the input vectors in the NARX models, finding redundant variables in the input space. After that, the Granger causality test will be applied to these results to obtain the input vectors used in the NARX models.

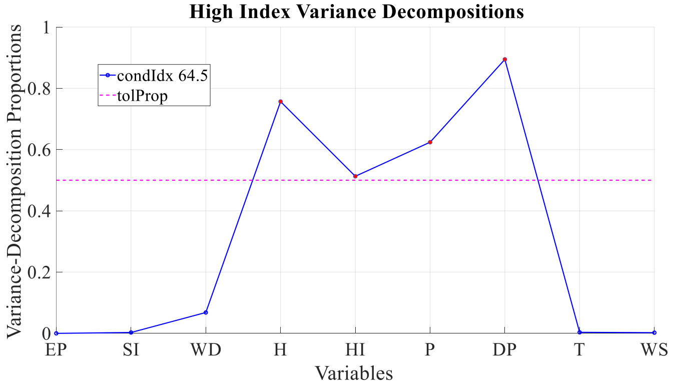


Fig. 2. Collinearity test results.

For the Belsley collinearity diagnostics, the condition index tolerance (conIdx) is set to 30. For the Variance-decomposition proportion, tolerances (tolPorp) are set to 0.5 as suggested by [30]. This is shown in Fig. 2, where the continuous line represents the results when the tolPorp is higher than 0.5; the conIdx is higher than 30. In this case, the redundant variables are **H**, **HI**, **P**, and **DP** (from now on, we represent the collinear variables in bold for easy reference). The condition indices for a scaled matrix X identify the number and strength of any near dependencies in X, and Variance-decomposition proportions determine groups of variables involved in near dependencies and the extent to which these dependencies degrade the regression [30].

TABLE III.
INPUT VECTORS FROM THE COLLINEARITY TEST

Group 1	EP, SI, WD, **H**, T and WS
Group 2	EP, SI, WD, **HI**, T and WS
Group 3	EP, SI, WD, **P**, T and WS
Group 4	EP, SI, WD, **DP**, T and WS

After applying this test, we find four collinear variables, and place them in different groups; TABLE III shows the four groups that result from the collinearity test. Then, we obtain the final input vector by applying the causality test to each group.

C. Generating the Input Vectors Second Step

After the collinearity test, we use the causality test (refer to (3) and (4)) to build the input vectors for the NARX models. The causality test aims to determine which variables contain information to predict the generated power of the PV system.

We can extend the causality test to more than two variables by applying the vector autoregressive (VAR) technique to several variables. In this case, we have nine equations, one for each variable. However, for practical purposes we only show the equation for two variables:

$$Y_t = \sum_{i=1}^{n} \alpha_i X_{t-i} + \sum_{j=1}^{n} \beta_j Y_{t-j} + u_{1t} \quad (3)$$

$$X_t = \sum_{i=1}^{n} \lambda_i Y_{t-i} + \sum_{j=1}^{n} \delta_j X_{t-j} + u_{2t} \quad (4)$$

where X_t represents any variable, for example, the electric power; Y_t can represent any meteorological variable; u_{1t} is the uncorrelated white noise; α_i, β_j, λ_i, and δ_j are parameters to be determined using Ordinary Least Squares (OLS); and n is the number of lags [31].

Before carrying out the causality test, first we apply the Augmented Dickey-Fuller (ADF) test to ensure the stationarity (the statistical properties of a process generating a time series do not change over time) of the time series. TABLE IV displays a summary of the ADF test; the lag length is set using the Schwarz information criterion [33]. The exogenous column in TABLE IV depends on the time series characteristics like the deterministic trends and other patterns present in the time series. The null of the ADF test is that the time series is not stationary. Since the p-value is smaller than 0.05, we reject the null; therefore, the time series are stationary. The Durbin-Watson statistical test is in the range of 1.85 – 2.15, so there is no evidence of autocorrelation [34].

TABLE IV.
SUMMARY UNIT ROOT TESTS

Time series	Exogenous	Lag Length	p-value	Durbin-Watson
EP	Constant	28	0.00	1.99
SI	Constant	27	0.00	2.00
WD	Constant	27	0.00	2.00
H	Constant	27	0.00	1.99
HI	Constant Linear Trend	27	0.00	1.99
P	Constant	26	0.00	1.99
DP	Constant Linear Trend	3	0.00	1.97
T	Constant Linear Trend	29	0.00	1.95
WS	Constant	0	0.00	2.02

Once we guarantee the time series are stationary, we apply the Granger causality test to the four variable groups obtained from the collinearity test. TABLE V summarizes the causality test for the four groups of variables.

TABLE V. CAUSALITY TEST RESULTS

Group 1 (Dependent variable: Power) Lags: 28		Group 2 (Dependent variable: Power) Lags: 28	
Independent variables	p-value	Independent variables	p-value
SI	0.0000	SI	0.0000
WD	0.0449	WD	0.0396
H	0.0000	**HI**	0.0000
T	**0.7763**	T	**0.3165**
WS	**0.3847**	WS	**0.4706**
Group 3 (Dependent variable: Power) Lags: 28		Group 4 (Dependent variable: Power) Lags: 30	
Independent variables	p-value	Independent variables	p-value
SI	0.0000	SI	0.0000
WD	0.0045	WD	0.0232
P	0.0008	**DP**	**0.2409**
T	0.0054	T	0.0002
WS	**0.4802**	WS	**0.5397**

The null indicates that the independent variables have information to predict the dependent variable. Then, for a p-value greater than 0.05 we reject the null, and therefore, we remove the variables with a p-value greater than 0.05. Once we finish with the causality test, we can build the input vector for the NARX models. From Group 1 we obtain Input vector 1: SI, WD, **H**, removing T and WS. From Group 2, we obtain Input vector 2: SI, WD, **HI**, extracting T and WS with p-values in bold (TABLE V) and so on for Input vectors 3 and 4. Finally, TABLE VI shows the input vectors. Notice that we kept the variables that, according to the causality test, have helpful information to forecast the electric power.

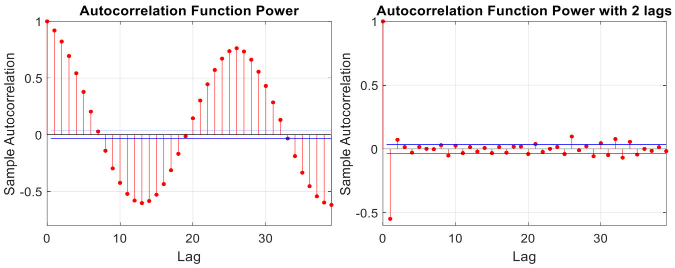
D. Choosing the Lag Number for NARX Models

We implement the autocorrelation function (ACF) test [5], [6] for the lag number. This test is only applied to the dependent variable (EP) and we choose the lag number depending on this result; for the ACF refer to (5).

$$r_k = \sum_{t=k+1}^{n} \frac{(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \tag{5}$$

where r_k is the autocorrelation value in lag k, Y_t is the value in time t, and $\bar{Y}$ is the mean of the time-series.

In Fig. 3 we show the results for the ACF test with and without differentiating the time series; Fig. 3 a) and b), respectively. The ACF test displays a sinusoidal form, indicating the presence of seasonality, as the peak repeats itself every 13 lags. We also see that with the first difference the seasonal behavior disappears.



a) b)

Fig. 3. a) ACF without differentiation and b) ACF first difference.

E. NARX models using inputs from collinearity and causality

We develop the NARX models using Matlab®. From the 3,430 available data, we use a first set of 3,250 data points to establish the NARX models split into three groups: 70% for training, 15% for validation, and 15% for testing. We use the remaining data set to evaluate the accuracy of the NARX models to forecast 156 steps ahead of the electric power of the PV system, and we implement the RMSE and R^2 metrics (refer to 7 and 8) to measure the performance of the forecasted data. To obtain the NARX models for the training stage, we use the MSE as a performance function (refer to 6).

The NARX models are configured with ten neurons in the hidden layer, hyperbolic tangent sigmoid activation function in the hidden layer, and linear transfer function in the output layer. We use the Levenberg-Marquardt backpropagation (LMB) and the Bayesian regularization backpropagation (BRB) algorithms as training functions. The NARX models are trained with the input vector obtained by using the collinearity and causality tests. Hereinafter, the NARX models that implement the collinearity and causality tests are denoted with the prefix CC (collinearity and causality) and the suffix 1, 2, 3 or 4 (number of input vector), e.g., the CC_NARX_1 model corresponds to the NARX model developed with the Input vector 1, obtained with the collinearity and causality tests.

TABLE VI
. NARX MODELS BUILD AFTER APPLIED COLLINEARITY AND CAUSALITY TESTS

Model	Lags Number	Input vectors	Output	retraining
CC-NARX_1	13	EP, SI, WD, **H**	EP	20
CC-NARX_2	13	EP, SI, WD, **HI**	EP	20
CC-NARX_3	13	EP, SI, WD, **P**, T	EP	20
CC-NARX_4	13	EP, SI, WD, T	EP	20

TABLE VI shows the NARX models, number of lags, input vector, output, and the number of times these are retrained with input vectors obtained from the collinearity and causality tests.

TABLE VII shows the metrics for the fit curve with the four NARX models for total performance, training, validation, and testing. As has been established, according to the MSE, the best models are CC-NARX_1 and CC-NARX_2. The function that measures the NARX models' performance during training, validation, and test stages is the MSE.

TABLE VII.
METRICS FOR FIT CURVE NARX MODELS

Model	Total MSE	Train MSE	Val MSE	Test MSE
CC-NARX_1	5.53×10^4	4.73×10^4	7.94×10^4	6.88×10^4
CC-NARX_2	5.59×10^4	5.10×10^4	7.29×10^4	6.21×10^4
CC-NARX_3	6.03×10^4	5.60×10^4	7.12×10^4	6.97×10^4
CC-NARX_4	5.96×10^4	5.65×10^4	7.20×10^4	6.17×10^4

F. NARX Models Development With the all Variable Combinations Algorithm (AVCA)

The collinearity and causality test are techniques that focus on statistical models, and the application along with ANN networks is a new topic. To validate the results, we use all variable combinations for input vectors in NARX models, and the best results are reported in this work; we develop a code that uses to the k −combination formula (6) to generate 502 input vectors.

$$N = \frac{n!}{k!\,(n-k)!} \tag{6}$$

where n is the total number of variables (nine in our case study) and k is the variable number that will be part of the input vector, 2, 3, …, 9, and N is the number of all possible combinations.

TABLE VIII shows the best ten NARX models from all possible combinations of the variables. The results are ordered first by Total MSE (whole sample) value and then by test MSE (only data for test) value; refer to (7). Then, we use the best ten models to compare the results with the NARX models trained with the input vectors obtained from the collinearity and causality test.

We see that the best model is the NARX_406, followed by the NARX_322; the worst model is the NARX_117 at the end of the table.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - F_t)^2 \tag{7}$$

$$RMSE = \sqrt{MSE} \tag{8}$$

$$R^2 = 1 - \sum_{t=1}^{n}(Y_t - F_t)^2 \bigg/ \sum_{t=1}^{n}(Y_t - \bar{Y})^2 \tag{9}$$

TABLE VIII. NARX MODEL RESULTS OF TEN BEST VARIABLES POSSIBLE COMBINATIONS

Index	Input	Total MSE	Train MSE	val MSE	test MSE
NARX_406	EP, P, T	5.10×10^4	4.09×10^4	8.37×10^4	6.52×10^4
NARX_322	SI, WD, P, DP	5.47×10^4	4.65×10^4	6.92×10^4	7.80×10^4
NARX_80	EP, SI, HI, DP, T, WS	5.48×10^4	4.04×10^4	7.79×10^4	9.87×10^4
NARX_223	SI, H, HI, P, WS	5.54×10^4	4.70×10^4	7.66×10^4	7.33×10^4
NARX_199	EP, HI, DP, T, WS	5.63×10^4	4.69×10^4	7.09×10^4	8.55×10^4
NARX_283	EP, WD, HE, P	5.64×10^4	4.48×10^4	5.97×10^4	1.07×10^5
NARX_158	EP, SI, HI, P, WS	5.64×10^4	4.30×10^4	9.28×10^4	8.25×10^4
NARX_328	SI, H, HI, P	5.69×10^4	5.31×10^4	7.05×10^4	6.13×10^4
NARX_295	EP, H, HI, T	5.71×10^4	4.69×10^4	7.34×10^4	8.80×10^4
NARX_117	SI, WD, P, DP, T, WS	5.74×10^4	4.99×10^4	6.81×10^4	8.20×10^4

IV. RESULTS AND DISCUSSIONS

In this section, using the RMSE and R^2 metrics we compare the forecasting performance of the NARX models trained with the vectors obtained from the collinearity and causality tests, to the best 10 NARX models obtained from the algorithm for all variable combinations.

A. Best NARX Models from the Proposed Methodology Versus the Best NARX Models Obtained from all combinations

Fig. 4 and Fig. 5 show the RMSE performance for NARX models. In Fig. 4, we offer the RMSE of the four best NARX models that use input vectors obtained from the collinearity and causality tests. In Fig. 5, we show the RMSE of the ten best NARX models that use input vectors obtained from the AVCA algorithm; for the RMSE calculation refer to (8).

The RMSE performance is calculated using the forecasted data points obtained from each NARX model and the data set (156 data) not used for training.

We see that the best NARX model from the proposed scheme is the CC_NARX_2 with an RMSE of 308 W for the electric power of the PV system, trained with the LM training function. On the other hand, the best NARX model from the AVCA algorithm is the NARX_158 with an RMSE of 305 W, using the LM training function.

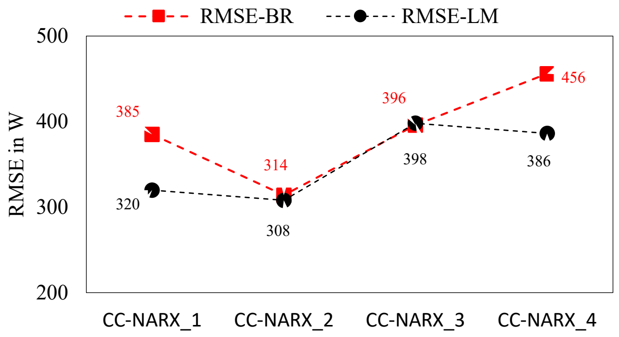


Fig. 4. RMSE for 156 steps ahead implemented CC-NARX Models.

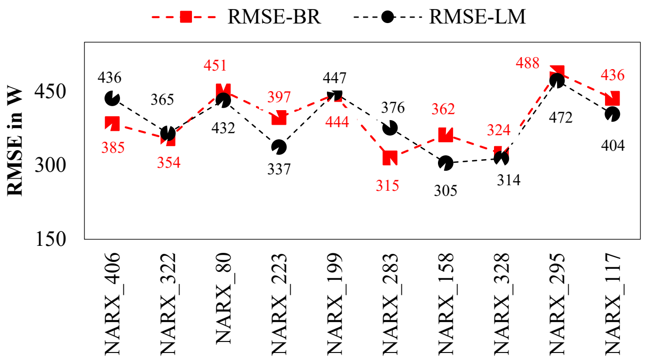
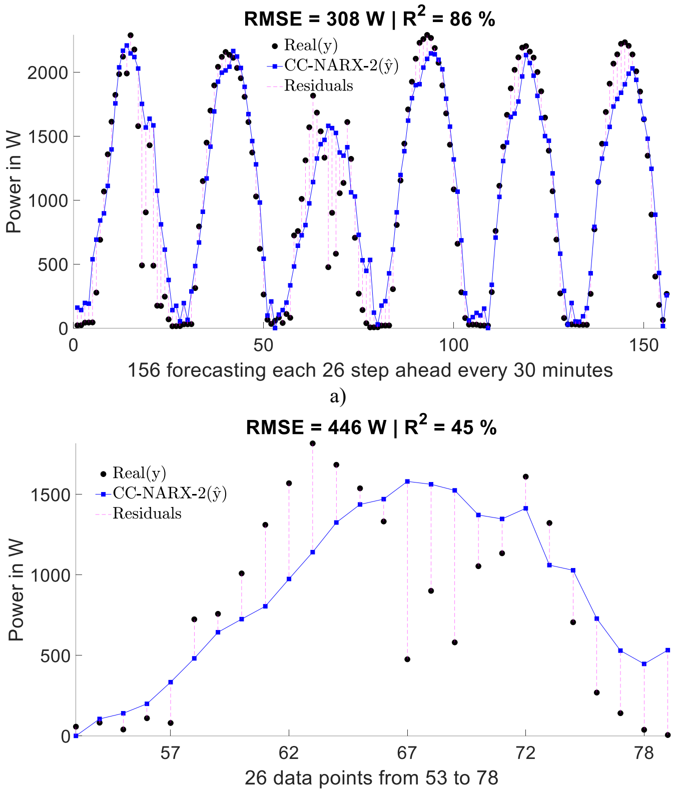


Fig. 5. RMSE for 156 steps ahead using the best ten models obtained by all variable combinations



a)

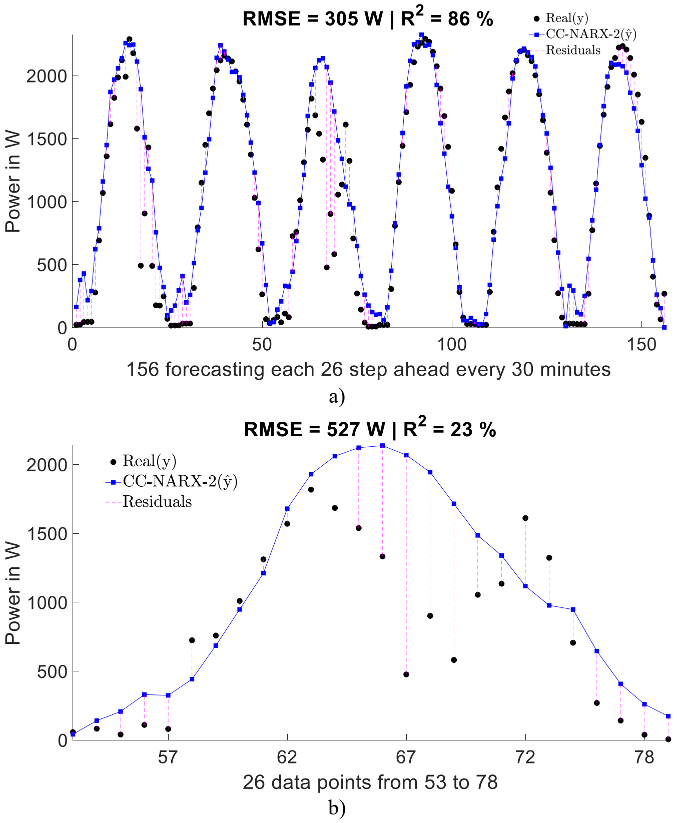
b)

Fig. 6. CC-NARX-2 implementation. a) 156 of real data points versus the forecasting model. b) 26 data points.

Fig. 6 a) displays the actual time series of the power versus the forecasted data obtained with the CC_NARX_2. As can be seen, the model fails mainly on days 1 and 3. Fig. 6 b) shows the third day, where the solid points represent the actual data, whereas the line with square markers results from the forecasted data obtained by the CC_NARX model. The vertical dashed line represents the relative error between the actual and predicted data.

According to the RMSE values, the NARX_158 model is slightly superior to CC-NARX_2; the R^2 metric (refer to (9)) is 86% for both models, as can be seen in Fig. 6 a) and Fig. 7 a).

These results indicate that the NARX models built with the input from the collinearity and causality tests show trusty and quick ways of finding the input vector.



a)

b)

Fig. 7. NARX_158 implementation. a) 156 of real data points versus the forecasting model. b) 26 data points

V. Conclusion

This work aims to fill the gap in selecting the most relevant variables in the input vectors for multivariable models based on machine and deep learning paradigms. We achieve this through collinearity and causality tests, and validate the results by developing the AVCA.

The AVCA tests all possible variable combinations to be used as input vectors in NARX models. We compared the results of the NARX models developed with input vectors obtained with the collinearity and causality tests to the NARX models created with the input vectors obtained from the AVCA.

The results show that the proposed methodology is effective and can be implemented in multivariable models to improve their performance, reducing the number of variables with minimal loss of accuracy and achieving a parsimonious state in models, saving time, money, and computational resources.

Furthermore, the NARX models trained with the LMB algorithm perform better for the NARX models that use the proposed methods than the best NARX models obtained from the AVCA. The models fail on days with substantial solar radiation followed by a day with slight solar radiation; the models try to mimic the previous time series pattern because of lags. The external input in the models tries to correct the forecasting result. The proposed methodology provides a way to choose the variables to build the input vector in multivariable models. Even more, we can use this methodology for selecting the most convenient sensors to measure the meteorological variables containing helpful information to forecast the electric power in smart grids, reducing costs in sensor number and computational time, translating into a saving of time and money. The dataset of the PV system used in this work has a time frame of five months; however, the authors intend to extend its duration, for example, to one year, in future studies. This could help to answer questions such as how stable the proposed methodology is or how the results change as a function of the length of the dataset. This work helped to determine the variables that contain critical information needed to forecast the electric power in a PV system, thus allowing for a reduction in the size of the needed input vector. Although the authors believe this outcome has more to do with the physical nature of the variables and that the results will not considerably change with larger datasets, this also has to be investigated.

References

[1] N. S. Maimouna Diagne, Mathieu David, Philippe Lauret, John Boland, “Review of solar irradiance forecasting methods and proposition for small-scale insular grids,” *Renewable and Sustainable Energy Reviews*, vol. 27, pp. 65–76, 2013.

[2] P. F. Jiménez-Pérez and L. Mora-López, “Modeling and forecasting hourly global solar radiation using clustering and classification techniques,” *Solar Energy*, vol. 135, pp. 682–691, 2016, doi: 10.1016/j.solener.2016.06.039.

[3] F. O. Hocaoglu and F. Karanfil, “A time series-based approach for renewable energy modeling,” *Renewable and Sustainable Energy Reviews*, vol. 28, pp. 204–214, 2013, doi: 10.1016/j.rser.2013.07.054.

[4] H. Eom, Y. Son, and S. Choi, “Feature-selective ensemble learning-based long-term regional PV generation forecasting,” *IEEE Access*, vol. 8, pp. 54620–54630, 2020, doi: 10.1109/ACCESS.2020.2981819.

[5] E. Rangel, E. Cadenas, R. Campos-Amezcua, and J. L. Tena, “Enhanced prediction of solar radiation using NARX models with corrected input vectors,” *Energies (Basel)*, vol. 13, no. 10, pp. 1–22, 2020, doi: 10.3390/en13102576.

[6] E. Rangel-Heras, C. Angeles-Camacho, E. Cadenas-Calderón, and R. Campos-Amezcua, “Short-Term Forecasting of Energy Production for a Photovoltaic System Using a NARX-CVM Hybrid Model,” *Energies (Basel)*, vol. 15, no. 8, Apr. 2022, doi: 10.3390/en15082842.

[7] L. Yanting, S. Yan, and S. Lianjie, “An ARMAX model for forecasting the power output of a grid connected photovoltaic system,” *Renew Energy*, vol. 66, pp. 78–89, 2014.

[8] V. Sharma, D. Yang, W. Walsh, and T. Reindl, “Short term solar irradiance forecasting using a mixed wavelet neural network,” *Renew Energy*, vol. 90, pp. 481–492, 2016, doi: 10.1016/j.renene.2016.01.020.

[9] L. Liu, M. Zhan, and Y. Bai, “A recursive ensemble model for forecasting the power output of photovoltaic systems,” *Solar Energy*, vol. 189, pp. 291–298, Sep. 2019, doi: 10.1016/j.solener.2019.07.061.

[10] M. Alkandari and I. Ahmad, “Solar power generation forecasting using ensemble approach based on deep learning and statistical methods,” *Solar power generation forecasting*, 2019, doi: 10.1016/j.aci.

[11] T. Chuluunsaikhan, A. Nasridinov, W. Seok-Choi, D. bin Choi, S. Hyun Choi, and Y. Myoung Kim, “Predicting the Power Output of Solar Panels based on Weather and Air Pollution Features using Machine Learning,” *Journal of Korea Mutilmedia Society*, vol. 24, pp. 222–232, 2021.

[12] P. Dawan *et al.*, “Comparison of power output forecasting on the photovoltaic system using adaptive neuro-fuzzy inference systems and particle swarm optimization-artificial neural network model,” *Energies (Basel)*, vol. 13, no. 2, 2020, doi: 10.3390/en13020351.

[13] A. Ahmad, T. N. Anderson, and T. T. Lie, “Hourly global solar irradiation forecasting for New Zealand,” *Solar Energy*, vol. 122, pp. 1398–1408, 2015, doi: 10.1016/j.solener.2015.10.055.

[14] S. Hussain and A. Al-Alili, “A new approach for model validation in solar radiation using wavelet, phase and frequency coherence analysis,” *Appl Energy*, vol. 164, pp. 639–649, 2016, doi: 10.1016/j.apenergy.2015.12.038.

[15] M. Louzazni, H. Mosalam, and A. Khouya, “A non-linear auto-regressive exogenous method to forecast the photovoltaic power output,” *Sustainable Energy Technologies and Assessments*, vol. 38, Apr. 2020, doi: 10.1016/j.seta.2020.100670.

[16] M. Louzazni, H. Mosalam, and D. T. Cotfas, “Forecasting of photovoltaic power by means of non-linear auto-regressive exogenous artificial neural network and time series analysis,”

Electronics (Switzerland), vol. 10, no. 16, Aug. 2021, doi: 10.3390/electronics10161953.

[17] R. Azimi, M. Ghayekhloo, and M. Ghofrani, "A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting," *Energy Convers Manag*, vol. 118, pp. 331–344, 2016, doi: 10.1016/j.enconman.2016.04.009.

[18] J. Boland, M. David, and P. Lauret, "Short term solar radiation forecasting: Island versus continental sites," *Energy*, vol. 113, pp. 186–192, 2016, doi: 10.1016/j.energy.2016.06.139.

[19] S. Monjoly, M. André, R. Calif, and T. Soubdhan, "Hourly forecasting of global solar radiation based on multiscale decomposition methods: A hybrid approach," *Energy*, vol. 119, pp. 288–298, 2017, doi: 10.1016/j.energy.2016.11.061.

[20] C.-R. Chen and U. Kartini, "k-Nearest Neighbor Neural Network Models for Very Short-Term Global Solar Irradiance Forecasting Based on Meteorological Data," *Energies (Basel)*, vol. 10, no. 2, p. 186, 2017, doi: 10.3390/en10020186.

[21] A. A. du Plessis, J. M. Strauss, and A. J. Rix, "Short-term solar power forecasting: Investigating the ability of deep learning models to capture low-level utility-scale Photovoltaic system behaviour," *Appl Energy*, vol. 285, Mar. 2021, doi: 10.1016/j.apenergy.2020.116395.

[22] J. Huang and R. J. Davy, "Predicting intra-hour variability of solar irradiance using hourly local weather forecasts," *Solar Energy*, vol. 139, pp. 633–639, 2016, doi: 10.1016/j.solener.2016.10.036.

[23] E. Akarslan and F. O. Hocaoglu, "A novel adaptive approach for hourly solar radiation forecasting," *Renew Energy*, vol. 87, pp. 628–633, 2016, doi: 10.1016/j.renene.2015.10.063.

[24] K. Benmouiza and A. Cheknane, "Small-scale solar radiation forecasting using ARMA and nonlinear autoregressive neural network models," *Theor Appl Climatol*, vol. 124, no. 3–4, pp. 945–958, 2016, doi: 10.1007/s00704-015-1469-z.

[25] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*, Second. New Jersey: Wiley, 2015. doi: 10.1007/978-3-319-28725-6.

[26] N. M. Noor, M. M. al Bakri Abdullah, A. S. Yahaya, and N. A. Ramli, "Comparison of linear interpolation method and mean method to replace the missing values in environmental data set," in *Materials Science Forum*, 2015, vol. 803, pp. 278–281. doi: 10.4028/www.scientific.net/MSF.803.278.

[27] J. Honaker *et al.*, "What to Do about Missing Values in Time-Series Cross-Section Data," *Am J Pol Sci*, vol. 54, no. 2, pp. 561–581, 2010, [Online]. Available: http://gking.harvard.edu

[28] R. E. Shiffler, "Maximum z scores and outliers," *American Statistician*, vol. 42, no. 1, pp. 79–80, 1988, doi: 10.1080/00031305.1988.10475530.

[29] D. Cousineau and S. Chartier, "Outliers detection and treatment: a review," *Int J Psychol Res (Medellin)*, vol. 3, no. 1, pp. 58–67, 2010, [Online]. Available: http://www.redalyc.org/articulo.oa?id=299023509004

[30] D. Belsley, E. Kuh, and R. Welsch, *Regression Diagnostics — Identifying Influential Data and Sources of Collinearity*, vol. 32, no. 2. New Jersey: John Wiley & Sons, Inc., 1980.

[31] D. N. Gujarati and D. C. Porter, *Basic Econometrics*, 5th ed. Douglas Reiner, 2009.

[32] S. G. Makridakis, S. C. Wheelwright, and R. J. Hyndman, *Forecasting Methods and Applications*, 3rd ed. Wiley, 1997.

[33] M. Yang, "Lag length and mean break in stationary VAR models," *Econom J*, vol. 5, no. 2, pp. 374–386, 2002, [Online]. Available: http://www.jstor.orgURL:http://www.jstor.org/stable/23114900http://www.jstor.org/page/info/about/policies/terms.jsp

[34] J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression: I," *Springer*, vol. 37, no. 3, pp. 237–259, 1992, [Online]. Available: https://www.jstor.org/stable/2332391

Eduardo Rangel-Heras was born in Morelia, Michoacán México in 1983. He received his master's degree in 2013 and his Ph.D. degree in 2018 in the field of Sciences in Mechanical Engineering in the Universidad Michoacana de San Nicolas de Hidalgo, Morelia, Michoacán, México. Degree in Mechanical Engineering from Universidad Michoacana de San Nicolas de Hidalgo. During his Ph.D. stay, he developed projects implementing statistics and artificial intelligence techniques, applying neural networks and ARIMA models in forecasting solar irradiance. After he obtained his Ph.D. degree, he works in the private industries designing storage tanks, stripping columns and pressure vessels for the oil & gas industry, direct reduction rectors for iron ore reductions. At the present time, he is doing a post doctorate stay at the Universidad de Guadalajara, México.

Nun Pitalúa-Díaz received his Doctor Science Degree in Electric Engineering from the Research and Advanced Studies Center of the National Polytechnic Institute (CINVESTAV), Guadalajara, México in 2005. He is member of the Mexican National System of Researchers (SNI-CONACYT) since 2011. He is a Professor in Mechatronics and Renewable Energy Areas of the Sonora University (UNISON), México. His current research centers on control design and stability for intelligent systems and energy process.

Pavel Zuniga obtained his M.Sc. and Ph. D. in Electrical Engineering from the Research and Advanced Studies Center (CINVESTAV) Guadalajara Campus, Guadalajara, México in 2001 and 2006, respectively.

Since 2006 he has been with the Graduate Program in Electrical Engineering at the University Center of Exact Sciences and Engineering of the University of Guadalajara, Guadalajara, Jalisco, México. He is the author and co-author of several articles and conference proceedings. His research interests include harmonic analysis, controllers, modeling, and the application of power converters to active filtering and system balancing, renewable generation, and microgrids.

Dr. Zuniga was a recipient of the Institute of Electrical Studies Best Ph. D. Electric Networks National Thesis Award, and the Jalisco Science and Technology Council Science and Technology Award, both in 2006.

Esteban A. Hernandez-Vargas Esteban was born in Mexico. He did his Ph.D. in Mathematics at the Hamilton Institute, NUI, Ireland. After completing his doctoral studies in 2011, he continued his research as a postdoctoral fellow (2011-2014) at the Helmholtz Centre for Infection Research, Braunschweig, Germany. In the summer of 2014, he got a Junior Research Leader position and founded the lab of Systems Medicine of Infectious Diseases at the Helmholtz Centre for Infection Research. In March 2017, he got an independent Research Leader position at the Frankfurt Institute for Advanced Studies in Frankfurt am Main, Germany. In January 2020, just before the COVID-19 pandemic, he got a professor position at the National Autonomous University of Mexico (UNAM). Since August 2022, he is an Assistant Professor in the Department of Mathematics and Statistical Science at the University of Idaho, USA.

Alma Y. Alanis is Professor-Researcher of the Department of Computational Sciences of the University Center of Exact Sciences and Engineering of the University of Guadalajara. She is a member of the National System of Researchers at Level 2 and a member of the Mexican Academy of Sciences since 2017. She has been recognized as a desirable PRODEP profile since 2010. She has the “Senior Member” distinction from the IEEE. In 2013 she received the scholarship for women in science from L'oreal-UNESCO-AMC-CONACYTCONALMEX and in 2015 she received the "Marcos Moshinsky Chair" award from the UNAM Institute of Physics, the Marcos Moshinsky Foundation. and CONACYT. She is currently associate editor of Elsevier's “Journal of Franklin Institute” and Taylor and Francis's “Intelligent Automation & Soft Computing”, both journals indexed in the JCR. Her research interests are: neural modeling and control (“backstepping”, block control, inverse optimal control) among others, as well as its application to automatic control systems and robotics.