IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 5, 2019, accepted February 24, 2019, date of publication March 12, 2019, date of current version April 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2904432

A Hybrid Regression Model for Day-Ahead Energy Price Forecasting

DANIEL BISSING, MICHAEL T. KLEIN, RADHAKRISHNAN ANGAMUTHU CHINNATHAMBI, DAISY FLORA SELVARAJ, AND PRAKASH RANGANATHAN

School of Electrical Engineering and Computer Science, University of North Dakota, Grand Forks, ND 58202, USA

Corresponding author: Prakash Ranganathan (prakash.ranganathan@und.edu)

ABSTRACT Accurate forecast of the hourly spot price of electricity plays a vital role in energy trading decisions. However, due to the complex nature of the power system, coupled with the involvement of multi-variable, the spot prices are volatile and often difficult to forecast. Traditional statistical models have limitations in improving forecasting accuracies and reliably quantifying the spot electricity price under uncertain market conditions. This paper presents a hybrid model that combines the results from multiple linear regression (MLR) model with an auto-regressive integrated moving average (ARIMA) and Holt–Winters models for better forecasts. The proposed method is tested for the Iberian electricity market data set by forecasting the hourly day-ahead spot price with dataset duration of 7, 14, 30, 90, and 180 days. The results indicate that the hybrid model outperforms the benchmark models and offers promising results under most of the testing scenarios.

INDEX TERMS ARIMA, energy price, forecasting, Holt-Winters, hybrid model and regression.

I. INTRODUCTION

The energy trading has seen a rapid growth as result of deregulation and competitive energy markets in the recent years. The electricity price changes hour by hour and these changes typically reflect the variations in the availability of generation resources, fuel costs and demand. This volatility increases as the integration of intermittent sources of electric power generation (e.g., wind and solar) continues to rise. Furthermore, participants in the electricity spot market must submit price bids the day before buying or selling electricity (day-ahead), which means that buyers and sellers must make significant decisions regarding prices well in advance. Therefore, the producers of electric power require reliable forecasting methods to offer competitive bids to the buyers of electric power, while the consumers require reliable forecasting tools to acquire lowest possible price of electricity. Thus, the accurate forecasting methods are crucial for economic decision making and implementation of incentive based ''time-of-use pricing'' scheme for consumers. Moreover, the power grid is a highly complex system, governed by many variables, such as generation and transmission constraints, environmental variables, and seasonal demand variations. Thus, there is a need for an accurate forecasting model that accommodates these influential variables.

This paper presents a method for predicting ''day-ahead'' spot electricity prices for the Iberian electricity market, resulting in forecasted prices for each hour of the following day. The dataset available to the authors contains the hourly spot price for 3 and 6 months period ranging from approximately February to July of 2015. In addition, several other variables such as lagged values of price and demand, power production rates from coal and hydroelectric plants, and environmental variables such as temperature, wind speed, and irradiance are available in this dataset. The spot price for the last day available in the dataset (i.e., July 31, 2015) is forecasted using a novel hybrid method that combines typical forecasting methods such as Auto-Regressive Integrated Moving Average (ARIMA) and the Holt-Winters method with a multiple linear regression (MLR) model. In this model, the ''predictor variables'' such as hourly energy production and environmental variables are combined with forecasted variables using a weighted average.

II. LITERATURE REVIEW

The accurate forecast of short-term price is challenging for the electricity markets due to the complex nature of power system. Furthermore, the data series of electricity prices is typically non-stationary and highly volatile [1],

The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Rathinam.

making fundamental forecasting methods such as the linear regression (LR), moving averages (MA), or exponential smoothing (ES) presented in [2] unsuitable for robust price forecasting. On the other hand, Auto-Regressive Integrated Moving Average (ARIMA) is a more sophisticated and widely used forecasting method that combines predictions based on past values of the target variable (Auto-Regressive) with a moving average of the target value [2], and it is often used as a benchmark for comparing newer candidate models [1].

Chinnathambi *et al.* [3] discuss several forecasting methods for Iberian electricity markets using multiple variables. Here, they use a two-stage approach, where ARIMA method is first deployed in stage 1, and the resulting residuals are used as inputs to stage 2. In stage 2, authors use Locally Weighted Scatterplot Smoothing (LOWESS), Support Vector Machines (SVM), Random Forest (RF), Generalized Linear Model (GLM) for further improving forecasts. de Marcos *et al.* [1] claim that hybrid models combined with conventional forecasting models produce better forecasts. Here, the authors use an optimization model that outputs an estimated price based on different parameters of the power system such as supply, demand, and transmission constraints. This estimated price is then fed to a Function Fitting Neural Network (FFNN) that forecasts hourly prices based on the estimated price and predictor variables such as lagged prices and expected wind and solar generation. The results indicate that the hybrid model produces lower forecast error in terms of mean absolute percentage error (MAPE) and Mean Squared Error (MSE) as compared to ARIMA and FFNN model without the estimated fundamental price from the Cost-Production optimization model. Nowakowska and Lis [4] use a similar hybrid optimization model based on generation and demand in which they first seek to predict the hour-ahead demand using an optimization model that forecasts using the demand for the previous hour and historical data from the previous year. Previous demand forecasts are then compared with the real demand values and used along with the current price and ''price elasticity'' coefficient to forecast the next price. Alshejari and Kodogiannis [5] present the Asymmetric Gaussian Fuzzy Inference Neural Network (AGFINN) that combines neural networks, fuzzy logic, and clustering schemes in a hybrid model to produce improved price forecasts. Saini *et al.* [6] propose a hybrid method that combines linear regression with Support Vector Machine (SVM). Several linear equations are generated using linear regression with different combinations of the predictor variables, and a set of predictor variables resulting in a linear equation with the smallest MAPE is selected as the optimal set of predictor variables. The results of the linear forecasts are then used as the inputs to SVM.

There are also studies [7]–[10] focusing on different methods of using predictor variables as inputs to the forecasting model. Portela *et al.* [7] present an Autoregressive Moving Average Hilbertian (ARMAHX) forecasting method that offers improved capabilities of ARIMA to model seasonal effects and accounts for the effect of predictor (''explanatory'') variables to produce lower MAPE values than that of other benchmark methods. A Singular Spectrum Analysis along with an Artificial Neural Network (ANN) is used to develop a non-linear relationship between the electricity price and the predictor (''exogenous'') variables, such as temperature (t), and grid conditions [8]. In [9], the method of SVR (Support Vector Regression) for developing nonlinear regression relationships between predictor variables and the electricity price is proposed. Mohamed and El-Hawary [10] stress the importance of selecting the right predictor variables (''input features'') for electricity price forecasting, and they propose different methods of predictor variable selection including Attribute Evaluator methods such as ''CFsSubsetEval'' and ''WrapperSubsetEval'' and search methods such as the ''Best-First'', ''Greedy-Step Wise'', and other Exhaustive methods.

The contribution of this paper is to expand the work of Chinnathambi *et al.* [3] by strategically selecting the significant predictor variables for the Iberian electricity market data and using them as inputs to hybrid models that combine multiple linear regression with ARIMA and the Holt-Winters method. Chinnathambi *et al.* [3] do not perform a detailed analysis of the predictor variables to determine which predictor variables are truly significant to the model, and instead they have used all 17 predictor variables. This paper expands the work of Chinnathambi *et al.* [3] with the following contributions:

- A multiple regression method in MATLAB is used to perform an exhaustive search of multiple regression models of all possible combinations of predictor variables and select the best model based on various measures of a predictor variable. Each predictor variable is represented by one digit in a binary number which is toggled to either include or remove the variable from the model between iterations. This method is described in more detail in the following section.
- A weighted averaging technique is developed for combining ARIMA, and regression methods for better forecasts for data duration of 7, 14, 30, 90, and 180 days.

III. STATIONARITY CHECK

Stationarizing a time series data is an essential step to obtain the statistical parameters such as mean, variance and correlation along with other variables, if the original series is non-stationary. If the data series is steadily growing over time, then the mean and variance will also increase with the size of the sample. This may result in underestimated values of mean and variance for the future periods. Hence, the stationarity check was performed for all data series using 'R' software. The stationary test such as Augmented-Dickey-Fuller Unit Root Test was performed for different dataset durations such as 7, 14, 30, 90, and 180 days. Table 1 shows the p-values for the unit root test (URT). Generally, a p-value of less than 0.05 indicates that the data is stationary, and greater than 0.05 requires differencing operations on the data. Hence, this

TABLE 1. Stationarity test results for different datasets.

Data set duration	p-value (Non-differenced)	p-value (First difference)
7 days	0.6145	< 4.229e-15
14 days	0.5293	< 2.2e-16
30 days	0.4164	< 2.2e-16
90 days	0.1643	< 2.2e-16
180 days	0.009699	-

unit root test was performed using a function named "ur.df ()" under the library "urca".

Stationarity check involves a two-step process. In step 1, the original data series is transformed into a time series object using 't_s' function in R software. Step 2 requires performing a unit root test (URT) for the time series data or the non-differenced data to check the stationarity. The URT results show that the original data set is non-stationary, as p-values for various durations (except 180 days) are greater than 0.05. Therefore, the first differencing for the non-stationary data series is performed to make it stationary and the resultant data is tested for stationarity. The results indicate that the data series becomes stationary after the first differencing and this is evident from lower p-values ($p<0.05$). Therefore, the stationarized first-differenced data set is used for the remainder of the paper except 180 days of data, where the non-differenced dataset is used.

IV. SELECTION OF PREDICTOR VARIABLES

The 3-month and 6-month datasets from the Iberian electricity market are used for this study and these datasets contain the hourly electricity price along with the hourly value of 17 other "predictor" variables that may or may not be significantly related to the price. In order to construct a model that can forecast hourly prices of electricity, it is important to identify the significant variables and discard the insignificant variables. The methods used to select the significant predictor variables are described in the following sections.

A. MULTIPLE LINEAR LEAST SQUARES REGRESSION

Multiple linear regression is a widely used method that fits a data set to a model in which the forecasted variable y_i depends linearly on a number of predictor variables $x_{1,i}, x_{2,i} \ldots x_{k,i}$. This multiple linear regression model can be expressed as,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots \beta_k x_{k,i} + e_i \tag{1}$$

Here k is the number of predictor variables, $\beta_1, \beta_2 \ldots \beta_k$ are the regression coefficients and e_i is an error term which represents the difference between the forecasted value ($\hat{y}_i$) and the measured value (y_i) [11]. Therefore, the values of β_j (and the overall model) can be optimized by minimizing the sum of the square of the error (SSE) term $e_{i.}$

Equation (1) can be expressed in matrix form as:

$$Y = X\beta + E \tag{2}$$

Here Y is an N x 1 matrix of the past N measured values of y_i, β is a $(k+1)$ x 1 matrix of the β values, E is an N x 1 matrix of the e_i values, and X is given by equation (3).

$$X = \begin{bmatrix} 1 & x_{1,1} \ldots & x_{1,k} \\ \ldots & \ldots\ldots & \ldots \\ 1 & x_{N,1} \ldots & x_{N,k} \end{bmatrix} \tag{3}$$

The SSE can be minimized and the optimum values of β_j can be selected by the equation:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{4}$$

Here $\hat{\beta}$ contains the optimized values of the linear coefficients [11].

If there is little or no relationship between the dependent and a predictor variable x_j, the value of β_j should be very close to 0. Therefore, a hypothesis test is performed with the null hypothesis that $\beta_j = 0$. Therefore, with a 95% confidence interval, if a value of β_j has a p-value greater than 0.05, the corresponding predictor variable x_j does not significantly contribute to the model. Thus, a predictor variable x_j can be removed from the data and a new model can be generated with updated values of $\hat{\beta}$ and the error term (SSE'). If the value of SSE' determined using hypothesis tests and p-values are not significantly larger than the SSE of the original model, then x_j does not add significantly to the model and it can be eliminated [11]. This process is repeated for all predictor variables, removing them one by one as seen unfit and testing if the SSE increases significantly. This procedure is implemented using the matrix and statistical functions of Microsoft Excel using the 3-month data, and it is observed that the predictor variables 5, 7, 8, 9, 10, and 16 are insignificant.

B. MEASURES OF PREDICTIVE VALUE IN MATLAB

Some statisticians warn against the use of selecting variables based on hypothesis tests and p-values, as these methods solely prove statistical significance, which is not necessarily an accurate measure of predictive value [2]. Rather, it is recommended to test all possible models and compare them based on measures of predictive value, such as the adjusted R^2, corrected Akaike's Information Criterion (AICc) or Bayesian Information Criterion (BIC).

The coefficient of determination (R^2) is a common statistic used to measure the correlation between variables and it is given by:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \tag{5}$$

Here $\hat{y}_i$ is the forecasted target variable, y_i is the measured target variable, and $\bar{y}$ is the mean target variable [2]. However, R^2 is not necessarily a useful tool for measuring the *accuracy* of a model (only correlation). Therefore, the value of R^2 can be adjusted as shown below:

$$R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-k-1} \tag{6}$$

The adjusted R^2 overcomes the limitations of the conventional R^2.

A second measure of predictive value is Akaike's Information Criterion, corrected (AICc), which is given by:

$$AICc = Nlog\left(\frac{SSE}{N}\right) + 2\,(k+2) + \frac{2(k+2)(k+3)}{N-k-3} \tag{7}$$

A third measure of predictive value is the Schwarz Bayesian Information Criterion (BIC), which is given by:

$$BIC = Nlog\left(\frac{SSE}{N}\right) + (k+2)\log(N) \tag{8}$$

Thus, the best model for a given data set is typically the one with the largest value of R^2, lowest values of AICc and BIC [2].

To select the best model for a given data, the values of adjusted R^2, AICc, and BIC must be calculated for *all* possible models [2] as shown in Table 2.

TABLE 2. Predictive values for all possible regression models.

X_1	X_2	X_k	Adj R^2	AICc	BIC
0	0	1	.0217	2625.4	2638
0	1	0	.0801	2594.6	2607.2
0	1	1	.0877	2591.5	2608.3
…	…	…	…	…	…
1	1	1	.4658	2325.9	2351.1

A data set with k predictor variables has 2^k-1 possible models, but it can become unwieldy for data set with a large number of predictor variables [2]. Therefore, a novel MATLAB program is developed to compute the Adjusted R^2, AICc, and BIC for the $2^{17} - 1 = 131,071$ possible models.

The program performs multiple regression on the data set in order to compute the predictive values which are then stored after each regression. This cycle is looped, and the loop index variable (w) is converted to a binary number during each iteration. For example, in the first iteration, $w = 1$, and this index is converted to the binary number 00000000000000001. Each digit of the binary number is then used to "turn on" or "turn off" a predictor variable. Therefore, for the first iteration where $w = 1$, the only predictor variable included in the data set is the 17th variable that corresponds to the first binary digit on the extreme right of the binary number and it is represented by "1" and all other variables are marked "0", resulting in 00000000000000001. This regression process is looped 131,071 times to obtain the regression models from every possible combination of predictor variables.

The results of the MATLAB program are shown in Fig. 1, which plots the values of adjusted R^2, AICc, and BIC versus the iteration number. The maximum adjusted R^2 and minimum AICc and BIC occur at 129,021th iteration and the corresponding values are 0.6159, 18028, and 18055 respectively. This optimized model contains all the variables except variables 6 and 16.

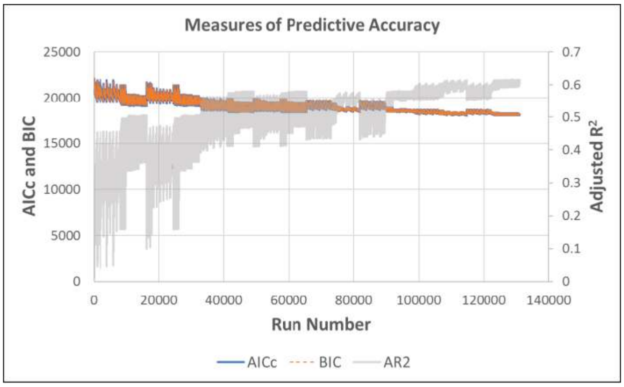

FIGURE 1. Measures of predictive value returned by MATLAB function. The optimal model occurs at run when the adjusted R^2 (gray) is maximized and the AICc and BIC (blue, orange) are minimized at iteration 129,021 when variables 6 and 16 are removed.

The residuals of this optimized model appear to be normally distributed with a mean of zero, indicating a good fit. However, the autocorrelation function of the residuals shows large spikes at the first several lags, indicating that an autocorrelation forecasting model such as ARIMA may be well-suited to this data.

C. STEPWISE AND SUBSET REGRESSION IN R

The Stepwise and Subset regression can be employed in situations where it is undesirable or impossible to test every possible model for determining the significant predictor variables. The Subset regression allows only a certain number of variables to be evaluated as significant (a subset of the original set of variables), while the Stepwise regression removes one predictor variable at a time and keeps the new model if the predictive measures are improved [2]. The process is repeated until the model cannot be further improved.

Both Subset and Stepwise regression were performed on the 6-month data set using R software. The largest subset available in R allows 8 variables, and the best model (judged by adjusted R^2) consisting of a subset of 8 predictor variables including the variables 1, 2, 3, 4, 9, 11, 15, and 17. The Stepwise regression performed in R returned an optimized model which is identical to the MATLAB based model that includes all variables except 6 and 16.

The results of four different predictor selection tests are presented in Table 3. If a test indicates that a given predictor variable should be included in the model, then that particular predictor variable is assigned a value of "1" in the column corresponding to the test or else "-1" is assigned for removal of a predictor variable. If the test did not address the predictor, the space is left blank. The last column adds the corresponding row values and gives the total for each predictor variable, with the largest and smallest sums corresponding to the most and least significant predictor variables, respectively. Based on these results, it is observed that hourly price demand, wind generation (w_g), temperature (t), and wind speed (w_s) (variables 1-4, 11, 15, and 17) are the most

TABLE 3. Results of predictor variable selection tests.

#	Variable	p-value	MATLAB	Stepwise	Subset	Total
1	Hourly Price D		1	1	1	3
2	Hourly Price D-6*		1	1	1	3
3	Hourly Power Demand D-1		1	1	1	3
4	Hourly Power Demand D-6		1	1	1	3
5	Hourly Hydro Generation D-1	-1	1		1	2
6	Hourly Hydro Generation D-6		-1		-1	-2
7	Hourly Solar power D-1	-1	1		1	1
8	Hourly Solar power D-6	-1	1		1	1
9	Hourly Coal Generation D-1	-1	1	1	1	2
10	Hourly Coal Generation D-6	-1	1		1	1
11	Hourly Wind Generation D-1		1	1	1	3
12	Hourly Wind Generation D-6		1		1	2
13	Hourly Combined Cycle Power Generation D-1		1		1	2
14	Hourly Combined Cycle Power Generation D-6		1		1	2
15	Temperature		1	1	1	3
16	Irradiance	-1	-1		-1	-3
17	Wind Speed		1	1	1	3

*D-6 or D-1 indicates hourly variables lagged by one day.

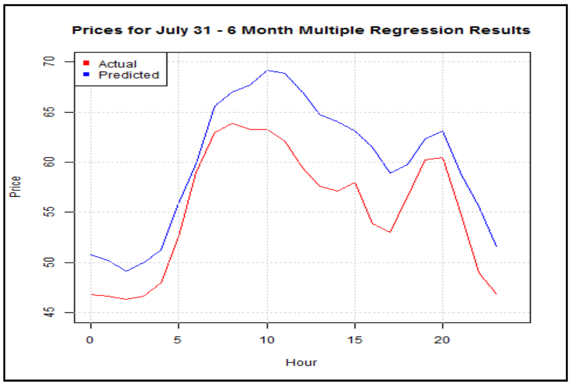

FIGURE 2. Forecast for July 31, based solely on multiple regression using all variables except variables 6 and 16.

significant predictor variables, while the lagged hydroelectric generation and irradiance (variables 6 and 16) are the least significant predictor variables.

Thus, a multiple regression equation is generated by using all predictor variables except the lagged hydroelectric generation and solar irradiance in equation (1). The resulting values of regression coefficients (β_{i}) from equation (1) are shown in Table 4. The regression equation therefore gives the hourly electricity price and it is the sum of product of each predictor variable x_{i} and its corresponding regression coefficient β_i from Table 4.

The results indicate that the price has a negative co-relation with variables 3, 5, 7, 10, 11, 13, and 17 and it is clear that the increase in these predictor variables tends to decrease the price and vice versa. Furthermore, the magnitude of each coefficient may not be indicative of the significance of the predictor variable because the variables have different units and scales. Equation (1) without the insignificant predictor variables was used to forecast the electricity price for July 31 and compared with actual data points for that day. The resulting forecast is shown in Fig. 2 and the Mean Average Percent Error (MAPE) is 8.17%.

TABLE 4. Regression coefficients.

Variable	Description	Coefficient β_i Value
0	Intercept	5.75656380
1	Hourly Price D	0.36240290
2	Hourly Price D-6*	0.34910440
3	Hourly Power Demand D-1	-0.00160260
4	Hourly Power Demand D-6	0.00391450
5	Hourly Hydro Generation D-1	-0.00891350
7	Hourly Solar power D-1	-0.00697340
8	Hourly Solar power D-6	0.00360060
9	Hourly Coal Generation D-1	0.00259170
10	Hourly Coal Generation D-6	-0.00169360
11	Hourly Wind Generation D-1	-0.00095150
12	Hourly Wind Generation D-6	0.00067980
13	Hourly Combined Cycle Power Generation D-1	-0.00660180
14	Hourly Combined Cycle Power Generation D-6	0.00497130
15	Temperature	0.15053270
17	Wind Speed	-1.12510640

V. TIME SERIES DECOMPOSITION

Time series data, such as historical prices of electricity, can typically be decomposed into three components: seasonal, cyclic/trend, and remainder (error). The seasonal component of a time series is the component that fluctuates regularly with known duration and amplitude, while the cyclic/trend component may cause the overall series to rise or fall without any definite period or amplitude. The remainder component is the error that remains when the seasonal and cyclic/trend components have been removed from the data. The time series decomposition for the first 48 days of the data is shown in Fig. 3, where the x-axis represents the day number (Jan 1 = 1, Jan 2 = 2, etc.).

The decomposition shown in Fig. 3 clearly shows a seasonal (daily) component having a period of 24 hours. This is expected, since energy usage has a similar trend from day to day. The data is made stationary using unit root test for forecasting as outlined in section III.

VI. FORECASTING METHODS

In this study, hybrid models of ARIMA with multiple regression and Holt-Winters with regression are used to predict the day-ahead electricity price. These methods are implemented using R software.

A. ARIMA METHOD

ARIMA is a commonly used forecasting method that uses three significant time-series components. These components include AR (Auto-Regressive), I (Integrated), and MA (Moving Average) which are denoted as p, d, and q respectively.

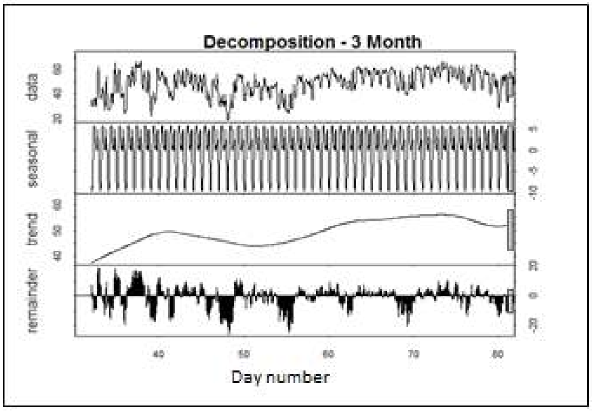

FIGURE 3. Times series decomposition of 48 days of the 3-month data exhibiting trends and seasonality.

Each of these components is optimized to find the best-fit forecast determined by the smallest residual values. ARIMA is a valuable forecasting tool for data that incorporates trend and seasonality [12].

The first step in ARIMA is the computation of the Integrated (I) component in which the data is integrated. This is accomplished by subtracting each data point from the previous data point. The goal of this step is to create a trendless/stationary data set, which can be accomplished through a single difference or multiple differences depending on the characteristics of the dataset. Once the differenced data is trendless, or as close to trendless as possible, the method proceeds to the next step.

The second step involves the computation of Auto-Regressive (AR) component which predicts future values of the trendless dataset based on a weighted sum of past values as shown in equation (9).

$$Y_t = c + \o_1 Y_{t-1} + \o_2 Y_{t-2} + ldots . \o_p Y_{t-p} + e_t \quad (9)$$

Here, Y_t is the price at time t, $\o_t$ denotes the regression coefficient, and e_t denotes the error term.

The final step of ARIMA is the computation of Moving Average (MA). A moving average is similar to autoregression, but instead of using previous target values, it uses previous error values to determine the current value as shown in equation (10). The R software package uses an auto-ARIMA function that optimizes the values to produce the best fit forecast.

$$Y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots . \theta_p e_{t-p} \quad (10)$$

B. HOLT-WINTERS METHOD

The Holt-Winters Method, also known as triple exponential smoothing, uses the principle of exponential smoothing to forecast the data points. Exponential smoothing is a technique used for smoothing time series data that can also be used for forecasting future values of the data. By assigning an exponentially decreasing weight to previous values of data, the future values are predicted with higher deference given to the most recent values. A smoothing factor dictates the amount of weight given to the previous values [13]. The equation for basic exponential smoothing is given in equation (11).

$$S_t = \alpha^* x_t + (1 - \alpha)^* S_{t-1} \quad (11)$$

Here S_t is the predicted value, α is the smoothing factor, and x_t is the value at time t. The smoothing factor ranges from 0 to 1 with smaller values giving more weight to previous data.

The Holt-Winters method uses triple exponential smoothing, allowing it to account for trend and seasonality. Triple exponential smoothing is achieved through equation (12).

$$y = S_t + B_t + U_t \quad (12)$$

Here:

$$\begin{aligned} U_t &= \gamma(x_t - -S_{t-s}) + (1 - \alpha)(U_{t-1} + B_{t-1}) \\ B_t &= \beta(U_t - -U_{t-1}) + (1 - -\beta)B_{t-1} \\ S_t &= \alpha(y_t - -U_t) + (1 - \alpha)S_{t-s} \end{aligned}$$

Here γ, β, and α are the smoothing factors for their respective levels.

The R software automatically optimizes the value of γ, β, and α using this method and Sum of Squared Error (SSE) metric is used to understand the residual errors.

VII. HYBRID FORECASTING METHOD

To provide a forecasting model with higher accuracy, hybrid approaches are explored that include combinations of ARIMA, Holts-Winters and regression methods. These hybrid methods are then tested using dataset of varying durations. The dataset includes the hourly electricity price for 7, 14, 30, 90, and 180 days.

The flowchart for the proposed hybrid model is shown in the figure 4. Step-1 involves data collection phase that collects the information on price, load, generation and temperature for the Iberian electricity market. In step-2, the important predictor variables are selected using different variable selection methods as explained in section-IV. Step-3 involves forecasting the initial price using ARIMA for 7, 14, 30, 90 and 180 days. Finally, two hybrid models are developed by combining (i) ARIMA with multiple linear regression and (ii) ARIMA with Holt-Winters and these models are tested on Iberian electricity market dataset to forecast the day-ahead electricity price. These hybrid models are discussed in detailed in the following subsections.

A. ARIMA COMBINED WITH MULTIPLE REGRESSION

The ARIMA forecast is based on the previous price data (e.g. D-1, D-6), so the addition of forecasting information based on the multiple regression model should increase the accuracy of the forecast.

The following steps were used to combine ARIMA with multiple linear regression model:

Step 1: The dataset of previous price variables (e.g., D-1 and D-6) is fed into the auto-ARIMA function

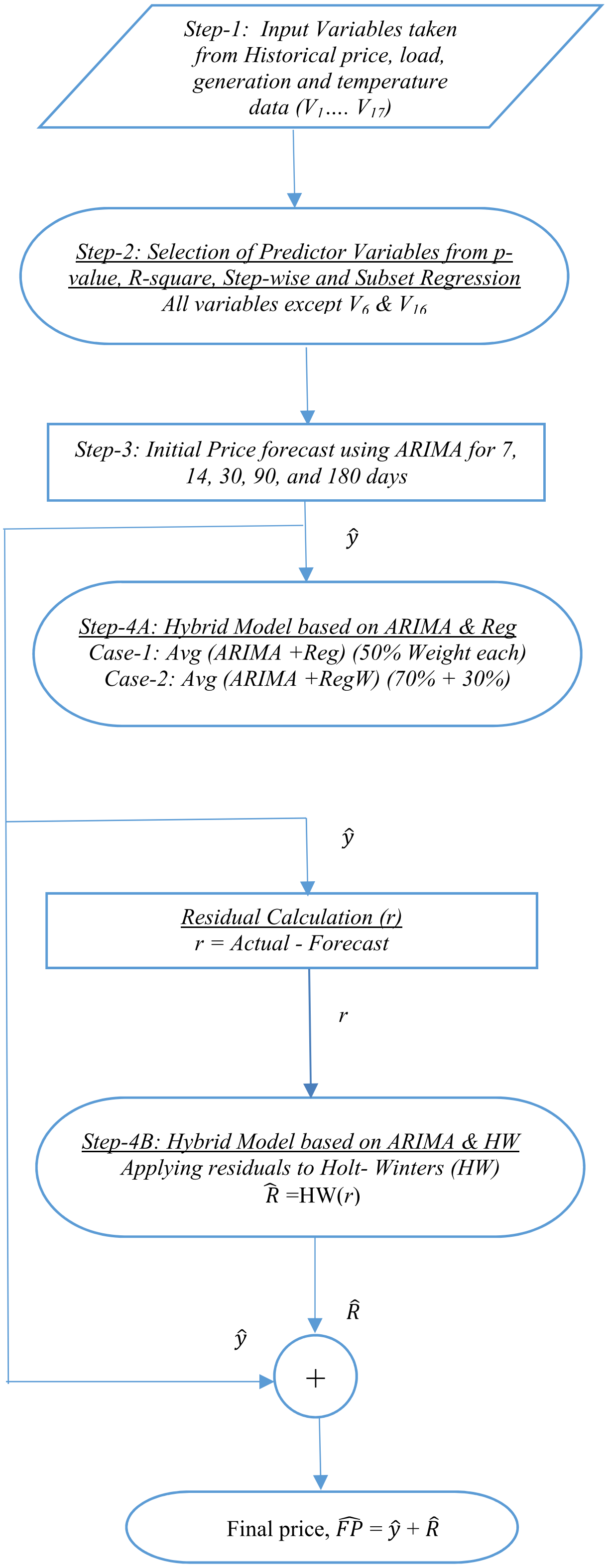
Step-1: Input Variables taken from Historical price, load, generation and temperature data (V_1.... V_{17})
Step-2: Selection of Predictor Variables from p-value, R-square, Step-wise and Subset Regression
All variables except V_6 & V_{16}
Step-3: Initial Price forecast using ARIMA for 7, 14, 30, 90, and 180 days
$\hat{y}$
Step-4A: Hybrid Model based on ARIMA & Reg
Case-1: Avg (ARIMA +Reg) (50% Weight each)
Case-2: Avg (ARIMA +RegW) (70% + 30%)
$\hat{y}$
Residual Calculation (r)
r = Actual - Forecast
r
Step-4B: Hybrid Model based on ARIMA & HW
Applying residuals to Holt- Winters (HW)
$\hat{R}$ =HW(r)
$\hat{R}$
$\hat{y}$
+
Final price, $\widehat{FP} = \hat{y} + \hat{R}$

FIGURE 4. Flowchart for the proposed hybrid Model.

of R software and a forecast for the following day (7/31) is generated.

step 2: A multiple regression model is used to forecast for the same day.

step 3: A new forecast is estimated by averaging the hourly forecasts from steps 1 and 2. This equal allocation of weights (50 % each) is termed as "ARIMA + Reg" method. In "ARIMA + RegW" method, the weights for ARIMA and Regression are adjusted to 70 % and 30 % respectively.

The forecasters may notice that the error in one of the models is larger in magnitude than that of the other model. In these cases, the forecast having a larger magnitude of error can be given less weight. This strategy will result in a more accurate forecast while still incorporating the predictions from both forecasting methods.

B. ARIMA COMBINED WITH HOLT-WINTERS

Generally, the residuals from the ARIMA forecast are trendless. However, auto-ARIMA finds the best-fit model, which is not necessarily perfect. Therefore, the residuals having noticeable trends can be detected and forecasted by the Holt-Winters model. The incorporation of trends from the residuals into the ARIMA forecast will result in higher accuracy. The following steps were used to combine the ARIMA forecast with Holt-Winters method:

step 1: The dataset of previous price is fed into the auto-ARIMA function of R software and a forecast for the following day (7/31) is produced.

step 2: The residuals produced by the auto-ARIMA function are extracted and converted to time series data.

step 3. The residuals are then fed into the Holt-Winters function in the R software and a forecast for the following day (7/31) is generated.

step 4: The forecasted residual values are added to the ARIMA forecast to generate an optimized forecast.

VIII. RESULTS AND DISCUSSION

The day ahead electricity price was predicted using two hybrid models discussed in the previous section and each model was trained using dataset durations of 7, 14, 30, 90, and 180 days. A 24-hour forecast for July 31, 2015 was generated including one data point for each hour of the day.

Mean Average Percentage Error (MAPE) was used as the metric for determining accuracy of the forecast, which is a common technique used in the forecasting field [1], [3]. MAPE is calculated using equation (13).

$$MAPE = \left(\frac{100}{n}\right) * \sum_{t=1}^{n} \left(\frac{A_t - F_t}{A_t}\right) \tag{13}$$

Here A_t = actual price at time t, F_t = forecasted price at time t, and n = number of data points being considered. In this case, the MAPE is determined using all 24 hours of the forecasted day. The MAPE values of different forecast methods for dataset durations of 7, 14, 30, 90, and 180 days are shown in Table 5.

The day-ahead electricity price forecasted for 7/31 (July 31, 2015) and the MAPE values of different forecasting methods for 7, 14, 30, 90 and 180 days are shown in Fig. 5 to Fig. 14. It is clear from Fig. 5 and Fig. 6 that ARIMA

TABLE 5. MAPE values for different forecast methods.

No. of Days	MAPE Values for different forecast methods				
	ARIMA	ARIMA +Holt	Reg	ARIMA +Reg	ARIMA +RegW
7	3.94	3.47	4.94	3.19	3.14
14	4.57	3.96	12.15	7.66	5.87
30	4.56	4.25	10.74	7.07	5.61
90	4.80	5.07	9.83	3.28	2.67
180	2.38	2.39	8.17	3.36	2.35

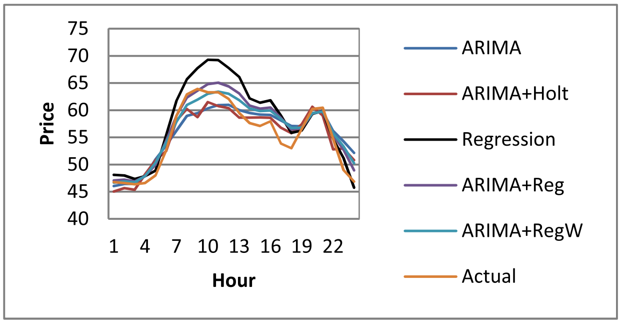

FIGURE 5. 24 hour forecast for 7/31 (July 31, 2015) using different forecast methods trained by the previous 7 days of data.

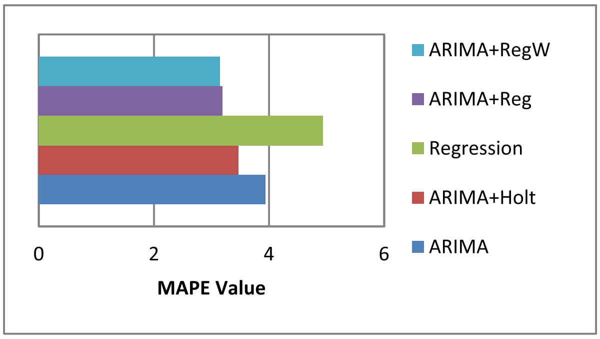

FIGURE 6. MAPE values for different forecasting models trained by the previous 7 days of data.

combined with multiple regression forecasting is the most accurate forecast when using seven days of data to train the model. The weighted model that gives the regression forecast a 30 % weight and the ARIMA forecast a 70 % weight has the least error at 3.14 %. It is also worth noting that the ARIMA combined with Holt-Winters forecast outperforms the ARIMA forecast.

It is noted from Fig. 7 and Fig. 8 that the error increases when 14 days of data are used to train the model. It can be seen that the most accurate model is ARIMA combined with Holt-Winters, and the least accurate is the regression model. Thus, the ARIMA models combined with multiple regression forecast are less accurate than ARIMA alone and this may be due to the large error caused by regression model with a 14-day dataset. Fig. 9 and Fig. 10 show results similar to that of forecast models for 14 days (Fig. 7 and Fig. 8), but with slightly less error. Again, the increased error magnitude of ARIMA + regression models is attributed to larger error of stand-alone regression model.

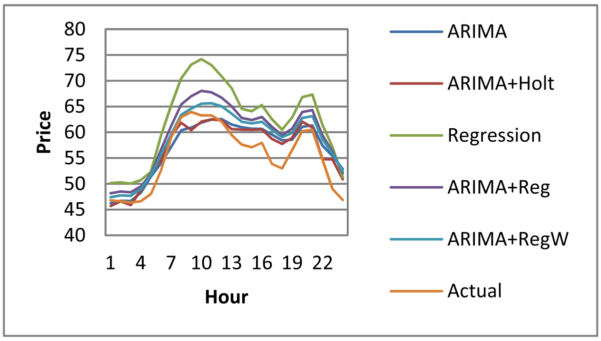

FIGURE 7. 24 hour forecast for 7/31 using different forecast methods trained by the previous 14 days of data.

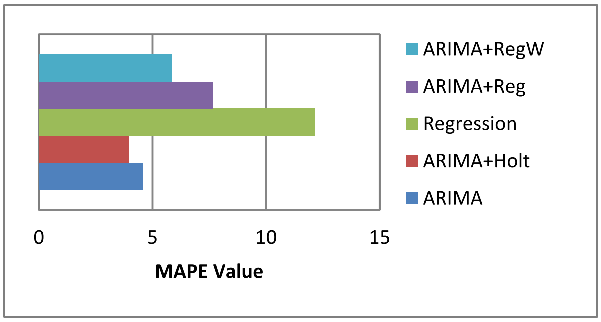

FIGURE 8. MAPE values for different forecasting models trained by the previous 14 days of data.

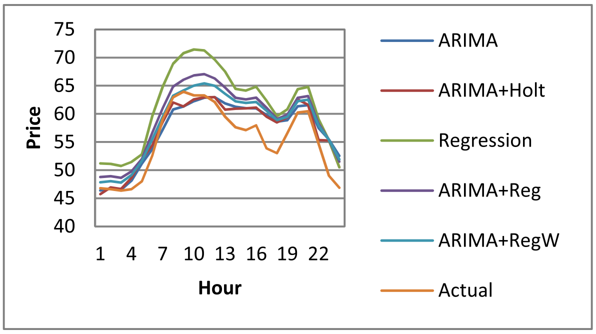

FIGURE 9. 24 hour forecast for 7/31 using different forecast methods trained by the previous 30 days of data.

Fig. 11 and Fig. 12 show that the error decreases when the training data is increased to 90 days. Though the error in regression method remains high, the ARIMA + regression models have the lowest error values. This indicates that the magnitude of error in the individual regression and ARIMA models generally has contradictory signs as seen from Fig. 11.

It can be seen from Fig. 13 and Fig. 14 that there is an overall decrease in error for 180 days of data and it is evident that the hybrid models have performed better than the stand-alone method regression model. Thus, it is clear from the results that the accuracy of the ARIMA forecast decreases

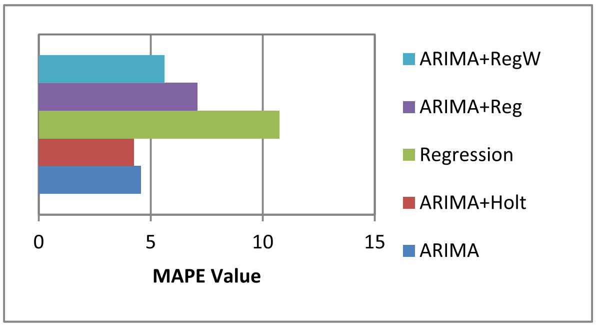

FIGURE 10. MAPE values for different forecasting models trained by the previous 30 days of data.

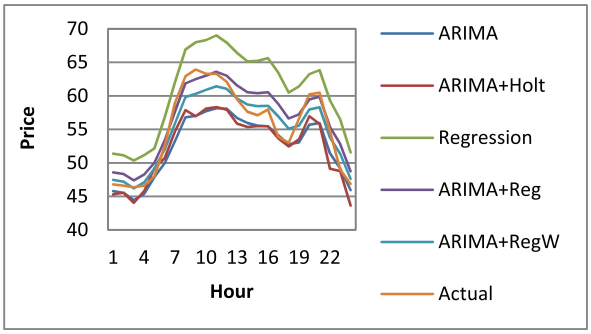

FIGURE 11. 24 hour forecast for 7/31 using different forecast methods trained by the previous 90 days of data.

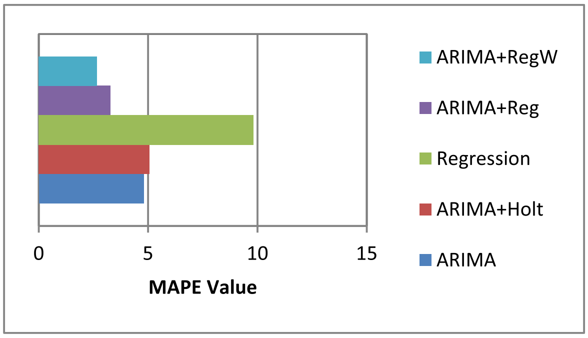

FIGURE 12. MAPE values for different forecasting models trained by the previous 90 days of data.

with an increase in the size of the training data. However, the hybrid models such as ARIMA + Holt, ARIMA + RegW models are performing better for 180 days as compared to the previous training datasets. This gives a clue to the forecaster that the models need to be trained with sufficient amount of data in order to produce good results while the most relevant data for the next day forecast is the most recent data. Future forecasting models can further build on this method by adding data from the same 7-day period of the previous years. This may be useful in maintaining the consistency of accuracy by including more "training" data.

The hybrid models presented in this study have yielded lower values of MAPE in comparison to other hybrid models (ARIMA-GLM, ARIMA-SVM and ARIMA-RF) proposed in [3] for the same Iberian electricity market for the duration of 7, 14, and 30 days. The results also confirm that the hybrid combination of ARIMA with Holt-winters proposed in this study outperforms other hybrid models discussed in [3].

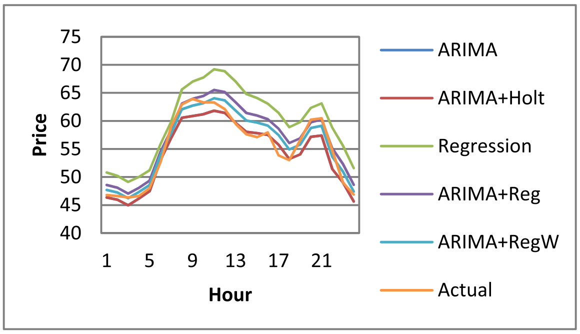

FIGURE 13. 24 hour forecast for 7/31 using different forecast methods trained by the previous 180 days of data.

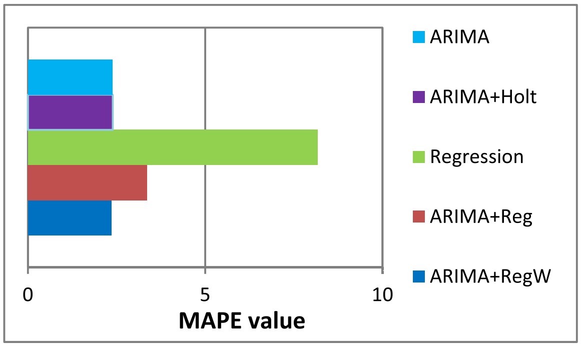

FIGURE 14. MAPE values for different forecasting models trained by the previous 180 days of data.

IX. CONCLUSION

A hybrid forecasting method that investigates the possibilities of combining the regression with Holt-Winters and ARIMA models is explored. Several variable selection methods are deployed to identify the predictor variables (e.g., hourly price, demand, wind generation, temperature, and wind speed) that significantly affect the hourly spot price of electricity.

The multiple regression model of predictor variables appears to be accurately determining the shape of the actual day-ahead electricity price, but it overfits the magnitude of the price. The ARIMA method is good at maintaining the magnitude within range; however, it could not capture the shape very well. A combination of these two methods provide a forecast having a proper magnitude and similar shape on a price versus hour plot, which resulted in a more accurate forecast. The varying weighted approach with regression and ARIMA model also yielded lower MAPE values, with 70 % weight assigned to ARIMA, and 30 % to regression based model. The combination of ARIMA with Holt-Winters outperformed other methods in most scenarios as well as other hybrid methods presented in the literature. It also strengthens the fact that the proposed hybrid model is a promising model to improve the accuracies of the short-term price forecasting model.

REFERENCES

[1] R. A. de Marcos, A. Bello, and J. Reneses, "Short-term forecasting of electricity prices with a computationally efficient hybrid approach," in *Proc. 14th Int. Conf. Eur. Energy Market (EEM)*, Dresden, Germany, 2017, pp. 1–6.

[2] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2013.

[3] R. A. Chinnathambi *et al.*, "A multi-stage price forecasting model for day-ahead electricity markets," *Forecasting*, vol. 1, no. 1, pp. 26–46, 2019.

[4] L. Nowakowska and K. Lis, "Electricity price forecasting as optimization problem using discrete dynamic model," in *Proc. 6th Int. Youth Conf. Energy (IYCE)*, Budapest, Hungary, 2017, pp. 1–6.

[5] A. Alshejari and V. S. Kodogiannis, "Electricity price forecasting using asymmetric fuzzy neural network systems," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Naples, Italy, Jul. 2017, pp. 1–6.

[6] D. Saini, A. Saxena, and R. C. Bansal, "Electricity price forecasting by linear regression and SVM," in *Proc. Int. Conf. Recent Adv. Innov. Eng. (ICRAIE)*, Jaipur, India, 2016, pp. 1–7.

[7] J. P. González, A. M. S. Roque, and E. A. Pérez, "Forecasting functional time series with a new Hilbertian ARMAX model: Application to electricity price forecasting," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 545–556, Jan. 2017.

[8] H. Varshney, A. Sharma, and R. Kumar, "A hybrid approach to price forecasting incorporating exogenous variables for a day ahead electricity market," in *Proc. IEEE 1st Int. Conf. Power Electron., Intell. Control Energy Syst. (ICPEICES)*, New Delhi, India, Jul. 2016, pp. 1–6.

[9] A. Mohamed and M. E. El-Hawary, "On optimization of SVMs kernels and parameters for electricity price forecasting," in *Proc. IEEE Elect. Power Energy Conf. (EPEC)*, Ottawa, ON, Canada, Oct. 2016, pp. 1–6.

[10] A. Mohamed and M. E. El-Hawary, "Effective input features selection for electricity price forecasting," in *Proc. IEEE Can. Conf. Elect. Comput. Eng. (CCECE)*, Vancouver, BC, Canada, May 2016, pp. 1–5.

[11] C. Zaiontz. (2017). *Multiple Regression Analysis, Real Statistics Using Excel*. Accessed: Dec. 8, 2017. [Online]. Available: http://www.real-statistics.com/multiple-regression/multiple-regression-analysis/

[12] R. Upadhyay. (2015). *Step-by-Step Graphic Guide to Forecasting Through ARIMA Modeling Using R—Manufacturing Case Study Example (Part 4)*. Accessed: Dec. 8, 2017. [Online]. Available: http://ucanalytics.com/blogs/step-by-step-graphic-guide-to-forecasting-through-arima-modeling-in-r-manufacturing-case-study-example/

[13] T. Jonsson, P. Pinson, H. A. Nielsen, H. Madsen, and T. S. Nielsen, "Forecasting electricity spot prices accounting for wind power predictions," *IEEE Trans. Sustain. Energy*, vol. 4, no. 1, pp. 210–218, Jan. 2013.

DANIEL BISSING is currently pursuing the degree with the School of Electrical Engineering and Computer Science, University of North Dakota, Grand Forks, USA.

MICHAEL T. KLEIN received the B.S. and M.S. degrees in electrical engineering from the University of North Dakota and the B.S. degree in mechanical engineering from the Benedictine College, in 2018. He was an Engineering Intern of nuclear power plants, from 2015 to 2016. He is currently with Burns & McDonnell, Minneapolis, MN, USA, where he focuses on transmission-level substation design. His interests include process automation, optimization, power system stability and protection, and low-cost rural microgrid design and deployment.

RADHAKRISHNAN ANGAMUTHU CHINNATHAMBI received the bachelor's degree in electrical and electronics engineering from Anna University, Chennai, India, and the master's degree in electrical engineering from the University of North Dakota, Grand forks, USA, in 2018. He was an Electrical Engineer with Doosan Power Systems, Chennai, India, for four years, where he was involved in designing electrical systems for the thermal power plant. He is currently a Graduate Researcher with the Data, Energy, Cyber and Systems (DECS) Laboratory, UND. His research interests include electricity price forecasting for day-ahead electricity market, smart grid, power system markets, energy forecasting, and demand response.

DAISY FLORA SELVARAJ received the B.E. degree in electrical and electronics engineering from Bharathidasan University, India, in 1999, the M.E. degree in high-voltage engineering from Anna University, India, in 2008, and the Ph.D. degree in electrical engineering from Visvesvaraya Technological University (VTU), Belgaum, India, in 2018. From 2013 to 2017, she was a Senior Research Fellow with the Research and Development Management Division, Central Power Research Institute, Bengaluru. She is currently a Postdoctoral Research Fellow with the University of North Dakota (UND), Grand Forks, North Dakota, USA. Her research interests include smart grid, condition monitoring of power apparatus, dielectric studies, and machine learning algorithms.

PRAKASH RANGANATHAN received the Ph.D. degree in software engineering from North Dakota State University (NDSU). He is currently an Assistant Professor of electrical engineering and the Research Director of the Data Energy Cyber and Systems (DECS) Laboratory, University of North Dakota (UND). He plays a leadership role on cyber educational initiatives with North Dakota University System (NDUS). He also plays a mentorship role for several native American students and tribal college faculty across the tribal reservations in ND, USA. He is a member of the Research Institute for Autonomous Systems (RIAS), a center for advancing autonomous systems, data and cyber security research. His research interests include operations research, smart grid, data mining, and cyber security. He was a recipient of the College of Engineering and Mines (CEM)'s Dean's Outstanding Faculty Award. He received the North Dakota Spirit Faculty Achievement Award from the UND Alumni Foundation, recognizing his significant contribution in teaching, research, and service, and the Public Scholar Award from the Center for Community Engagement.