Data-Analytic-Based Adaptive Solar Energy Forecasting Framework

Yashar Sahraei Manjili, Graduate Student Member, IEEE, Rolando Vega, and Mo M. Jamshidi, Fellow, IEEE

Abstract—An adaptive framework for day-ahead forecasting of available solar energy is proposed based on a combination of data-analytic approaches consisting of artificial intelligence and statistical techniques. Models are developed and validated utilizing a large dataset from the National Renewable Energy Laboratory (NREL) archive, the Automated Surface Observing System, and the solar position and intensity calculator (i.e., NREL-SOLPOS) sampled at 1-min intervals during eight years (2005–2012) for a site in Golden, CO, USA. The methodology is now ready for testing and validation in San Antonio, TX, USA, with data collected in the largest solar photovoltaic plant in TX, Alamo 1, which is the first solar plant in TX connected to the transmission grid allowing solar energy bidding into the market. A uniqueness of the methodology developed is that an integrated serial time-domain analysis coupled with multivariate analysis was used for preprocessing. The resulting enhanced dataset is used for adaptive training of the neural-network-based forecast engine. Standard performance measures are obtained. The forecast results are compared to those of the state of the art on day-ahead solar energy forecasting methodologies used in Austria and other European Union members in order to provide a clear understanding of capabilities of the proposed solar energy forecasting framework.

Index Terms—Adaptive systems, data analytic, day-ahead solar forecasting, neural networks, weather synoptic events detection.

I. INTRODUCTION

SOLAR forecasting is an area of research that is gaining more and more visibility due to the increasing penetration of solar energy into conventional electricity grids. Once the amount of solar penetration reaches a certain threshold, the variability of the solar energy production may become a problem for both grid stability and reliable bidding of electricity by electric utilities. Solar forecasting solutions provide electric utilities with predictions of power output from large-scale solar installations or from distributed solar generation with a temporal scale ranging from the next few minutes up to several days ahead forecast [1].

Databases offer a wealth of knowledge to users/operators but are often so restrictively large that information and/or knowledge cannot be extracted, analyzed, and synthesized quickly and efficiently enough for daily, hourly, and intra-hourly decision making. This is where data-analytic and data-mining techniques can be leveraged. Data analytics seek to discover relevant data features and patterns through data mining processes and communicate that information in the most effective way to the end user: either a human or a computerized process [2]–[4].

Various parties, such as system operators, utilities, project developers, and solar farm owners, can benefit from solar forecasting. For system operators, solar forecasts allow them to predict and manage the variability in solar energy to balance supply and demand on the regional or national grid system. Moreover, knowing, in advance, when expected surges in cheap and clean solar energy production will occur could allow for grid operators to reduce costs through the power-down of more expensive natural-gas-fired plants. In addition, there are costs associated with having excess units online, as well as from reduced unit efficiency and increased operations and maintenance. Improved solar energy forecasting can reduce these costs too.

Energy providers and utilities can benefit from solar energy forecasts. Imbalance charges and fines imposed on energy providers that result from deviations in scheduled output will increase energy providers’ operating costs. Solar energy forecasts can help minimize these penalties. Solar energy forecasts can also reduce the significant opportunity costs of being too conservative in bidding output into a forward market, due to uncertainty of resource availability [5]. Furthermore, the energy market operations (EMOs) of utility companies can use the cloud formation predictions, which give credible estimations of available solar energy ahead of time, to elevate their profit even more by leveraging storage units and smart energy management in microgrids [6].

The changing patterns that weather and climate variables follow will affect the cloud formation in the sky, which requires specific conditions for pressure, temperature, and relative humidity with respect to the dew point and air particles present in the sky. The cloud formation consequently influences the amount of sun rays that reach the ground negatively affecting the electrical energy generated by different approaches such as photovoltaic solar cells, concentrated solar power, etc. There are a number of approaches that focus on direct estimation of patterns of cloud formation, their movement, and spread in the sky such as methods based on sky imaging that are high-fidelity approaches for intrahourly solar energy forecast. Other methods such as numerical weather prediction (NWP) and satellite-based
approaches are used for prediction of cloud formation patterns in larger temporal and spatial resolution.

NWP discretizes space and time at the mesoscale/regional and national levels and solves the equations of motion, thermodynamics, and mass transfer at predefined time and space intervals [7]. The problem with NWP is that it cannot be considered site specific to any plant but based on regional commingled datasets and interpolation techniques. Such commingled datasets propagate error from node to node and from time step to time step, and its uncertainties are still too large in the order of mean absolute error (MAE) of 20–40% and root mean square error (RMSE) of 30–50% for day-ahead predictions [7]–[11].

More recently, interest has increased toward smart modeling of weather and climate events through training of systems based on artificial intelligence techniques [1], [3]–[5], [12], [13]. The results of such modeling approaches can be extensively used in solar forecasting.

Our solution, the process of which is illustrated in Fig. 1, is not dependent on theories defined for physical processes that constrain natural processes to a set of equations. If we compare NWP to the solution of a textbook, our solution using adaptive data-analytic-based artificial neural networks (ANNs) can be compared to the experience of a veteran in weather who is able to predict climate conditions accurately because he/she knows the local micrometeorological features and has a robust memory to recall historical anatomies of weather. Besides, the proposed method automatically adapts itself to abrupt and long-term changes in climate and environmental conditions and modifies its forecasts accordingly [14].

Our approach has initially been demonstrated for the prediction of global horizontal irradiance (GHI) of the next day using the irradiance information and weather sensory data of current day. Predicting irradiance on a point in space is more difficult than energy over larger areas, since the local features cannot be averaged out [11], [15]. Therefore, by developing the site-specific irradiance solar forecast tool, the methodology addresses the forecast of energy over larger areas more accurately when aggregating the predictions from the microclimate of each location.

II. PROBLEM STATEMENT AND CURRENT STATE OF THE ART IN SOLAR FORECASTING

A report performed for the Regents of the University of California titled “Current State of the Art in Solar Forecasting” [16] best summarizes current status of this field of research. In the report, Glassley et al. [16] state that satellite and NWP are the preferred methods for longer duration solar forecasting (one hour to a few days) [11], [16]. As for intrahour forecasts, sky-imagery-based forecasting methods are most common. However, several basic assumptions regarding cloud shape and linear cloud movement vectors reduce the potential accuracy of this type of forecasting [17], [18].

Making intrahourly or site-specific forecasts using satellite data is also not very common due to the infrequent sampling interval (30 min) and the low image resolution [11]. This problem is only increased in the NWP method due to a larger sampling interval and lower cloud imaging precision. The limitations of NWP models do not allow for shorter time scale and smaller spatial sampling to be accounted for [7], [11], [17]. Some of the error associated with the aforementioned problems can be corrected through a method known as modeled output statistics (MOS), which determines statistical correlations in observed weather data and related imagery (satellite, sky imagery, etc.) [7], [8], [11], [19].

Practical methods for day-ahead up to week-ahead solar forecasting methods are mostly based on NWP and statistical approaches as mentioned previously [8], [11], [16], [20]. A number of global models are available for this purpose, among them the Global Forecasting Service and the European Center for Medium Range Weather Forecasting (ECMWF) [21], both of which are considered the state of the art of global forecast models. However, in order to increase spatial and temporal resolution of these models, other models have been developed, which are generally called national or mesoscale models among which the North American Model, the High-Resolution Limited-Area Model, and the Weather Research and Forecasting (WRF) are widely used by different communities. A broad range of in-depth expertise is needed in order to obtain accurate results when running these models, due to the wide variety of parameters that can be configured. Also, sophisticated techniques such as data assimilation or statistical postprocessing have been needed in order to calibrate the estimations or obtain a probabilistic point of view of the accuracy of the output. Usually, techniques that mix outputs of different models are used for this, which finally provide a better estimate of those variables along with a degree of uncertainty associated with the forecast results [18].

Recently, artificial intelligence has entered the solar forecasting realm, and a number of approaches are developed based on neural networks, support vector machines, etc. [10], [12], [13], [22], which still need to improve the accuracy of their forecast engine, better stabilize, and further fine-tune the forecast results in order to be comparable with already well-established forecasting methods that have been in the market for the past decades such as NWP-based approaches mentioned above.

NWP and satellite imagery-based forecast techniques are the two most versatile approaches for longer forecast horizons, i.e., have the largest temporal capacity. The mentioned approaches also have the largest coverage, i.e., biggest spatial capacity, in solar energy forecast.

Regarding the MOS for a solar forecasting approach, regardless of the time scale, a database of instrument characteristics is necessary [19]. Unfortunately, the instrument characteristics database by itself does not translate to improved accuracies because the MOS solar forecasting model would need...
The adaptive forecasting framework is implemented based on correlation analysis and feature extraction of weather data and solar irradiance in combination with a synoptic event detection algorithm, which is used to determine the appropriate training sets for the ANN in a continuous and real-time manner. This creates a knowledge-based library that is ever-growing in size and increasing in intelligence. The contents of the mentioned knowledge-based library are information regarding the best sets of data chosen for training of the ANN and the forecast performance for each and every target day in the past. This information enables the forecast framework to modify its decision rules to choose the best training datasets for coming target days in such a way that the accuracy of the future forecasts is increased.

Fig. 2 represents the basic input/output (I/O) setup of the proposed approach used to develop the forecasting engine. The proposed forecast framework, as represented in Fig. 2, is capable of giving estimations of solar irradiance of the next “k” sample times based on current time data and the extracted feature matrix of the past data.

Fig. 3 represents the block diagram of the proposed solar forecast framework. This block diagram can also be regarded as a flowchart of the forecast process.

As can be seen in the block diagram of Fig. 3, the data coming from the sensors include the weather, timing data, and the solar plant output, i.e., the amount of dc power generated by solar array. The sensory data are continuously recorded added to the existing database.

A number of training subsets with different lengths and coverage of training windows are formed, considering the results of synoptic events detection algorithm, and a consecutive training and test process will take place for the target forecast horizon using each and every one of the training subsets. The process goes on by passing each training subset through a correlation analysis, where the correlation function between temperature, pressure, relative humidity, and GHI input with the plant output, i.e., the generated electrical power or the on-site measured GHI, is computed after using low-pass filters of lengths from 1 min to 3 h on the mentioned input variables. After the best filter length and shift values, i.e., the values that maximize the correlation between the mentioned variables and the plant output, are determined for each and every one of the input variables mentioned above, the rates of change (ROC) of the above-mentioned variables are also computed as represented in Fig. 3 in the ROC blocks. These ROCs represent gradients in the atmospheric variables and constitute precursors of changes in other variables, which indicate development or dissipation of clouds.

The best-fit variables obtained by correlation analysis are fed together with the ROCs to the feature extraction block, which transforms the data into a new domain using ICA in order to find the underlying independent features of the variables. These independent components or features will then be used for training of the ANN. After the ANN training is converged, control commands switch the system from training phase to the test phase. In the test phase, as seen (see Fig. 3) in the lower part of the block diagram, results of dynamic performance evaluation will be considered for final aggregation of forecasts. As the first step in this phase, the real-time temperature, pressure, relative humidity, and GHI obtained from the plant, \( D(t) \), are shifted and filtered using the best set of filter length and shift values, obtained from correlation analysis, for each training subset. Then, the ROCs for those weather variables are computed. This will follow by feeding the shifted and filtered variables to the feature computation block together with the ROCs of the aforementioned weather data. The data input to the feature computation unit is transformed into the independent components’ domain using the separation matrix \( W(t) \) obtained for each training subset in the training phase. The computed features will then be fed to the ANN for the prediction of solar energy to take place. The output of the ANN will go through the aggregation process. The forecasts obtained by different training subsets will be aggregated based on a weighted averaging regime, and then, the percentiles will be determined based on the Monte Carlo approach for uncertainty analysis for each and every target day. The 50th percentile, \( p_{50} \), is regarded as the stabilized raw forecast result, which will be then fine-tuned according to the set of relationships as shown in Section III-H. The fine-tuned signal is ready to be used by the electric grid operators and will be fed into a user interface to be used for electric energy bid decisions by the EMOs of utility companies.

A. Data Description

The National Renewable Energy Laboratory (NREL) archive, the nation’s largest combined climate and solar radiation dataset, is the reference big dataset used for implementation of the solar energy forecast framework along with the Automated Surface Observing System [23]–[25]. Data from more than 250 instruments and derived variables were available, including solar inputs of GHI, diffuse horizontal irradiance, direct normal irradiance, solar zenith and azimuth angles, plus atmospheric...
variables such as pressure, temperature, relative humidity, wind speed, wind direction, cloud cover percentages including total cloud cover (TCC) and opaque cloud cover (OCC), which are the variables that the researchers used for the solar forecasting library [1]. The variables used for training and performance analysis of the forecast engine are introduced in Section III-C, which span the whole duration of year 2012 sampled at 1-min intervals. Techniques used for preprocessing include elimination of outliers of measured variables and filtering of the data. The outliers are either removed or replaced by the estimates using extrapolation of the previous samples or a random deviation introduced to the previous sample value. Due to the fact that the sampling interval was only 1 min, both of the approaches involving extrapolation or random deviation introduced to the last sample value were sufficiently accurate. Filtering of some variables with an appropriate low-pass or moving average filter helps modify the amount of deviations in data samples which, in turn, for the desired forecasting time horizons will result in higher performance.

**B. Weather Synoptic Event Detection**

Synoptic weather events occur in a periodic fashion, and since the atmosphere is a fluid, much like in turbulence of fluid dynamics, synoptic events are regarded as stochastic processes. These events may seem random when analyzed by a historical statistical approach. However, when studied in a time-dependent manner, the time-varying coherence of those synoptic systems and the effects of individual variables and their features on atmospheric cloud formation and eventually their influence on the solar irradiance can be established.

The procedure for detection of weather synoptic events is based on locating the low-pressure weather systems, which happen in between two consecutive points of high pressures that meet certain conditions. Two basic criteria are used to distinguish a fully developed synoptic system from a weak one as established in [26]. As represented in Fig. 4, the first two derivatives of the filtered signal go through a set of relationships in order for the minimum and maximum points to be found. Finally, the day types are determined considering the portion of the reference variable each day represents. The day types are directly used for the purpose of adaptive training of the ANN, which is explained later in Section III-E. The outcomes of this algorithm, i.e., the types of days and the extreme points, are also used in order to determine the relationship base of fine-tuning.

The mathematical representation of the filtering procedure for synoptic events detection is as follows [5]:

\[
    f(t_i) = \frac{\sum_{j=\max(1, i-[L/2])}^{\min(N, i+[\frac{L}{2}])} v(t_j)}{\min(N, i+[\frac{L}{2}]) - \max(1, i-[\frac{L}{2}])} \quad (1)
\]

where \(i = 1, 2, \ldots, N\) is the sample number, \(N\) and \(L\), respectively, represent the number of samples for the input variable, \(v(t)\), and the low-pass filter length that works like an averaging window. The relationships used to determine the
It must be noted that the pressure signal gradients have been scaled up by making the delta pressures ten times larger for visualization purposes only, as shown in Fig. 5. Four relationships of climate anomalies; these, in turn, define the criteria.

1) IF \( f'(t_i) = 0 \) AND \( f''(t_i) < 0 \) THEN \( t_i \) is a MAX

2) IF \( f'(t_i) = 0 \) AND \( f''(t_i) > 0 \) THEN \( t_i \) is a MIN

where \( f' \) and \( f'' \) represent the first and the second derivatives of the filtered function \( f(t_i) \), respectively. After the MAX and MIN points are found, the high, i.e., Hi, and low, i.e., Lo, points of the input variable are determined based on the following criteria.

1) If the difference in value of a MAX with its adjacent MIN is less than one unit, the MAX point is neglected.

2) If any two consecutive MAX points are less than 36 h away, the largest is defined as Hi, and the other local MAX is neglected.

3) The smallest MIN point between each two consecutive Hi's is considered as a Lo.

Subsequently, two different signals are used for extracting relationships of climate anomalies; these, in turn, define the training window later. The first signal is the atmospheric pressure, which is used for synoptic events detection. The second signal is calculated based on the ratio of the three variables of atmospheric pressure, temperature, and relative humidity, as

\[
\frac{P \times T}{RH}
\]

The reason to choose this second signal as a reference is the cloud formation pattern itself. Generally, the clouds form when an air parcel rises and attains a relative humidity of 100% because of cooling to its dew point temperature, which causes the water vapor to condense into tiny water droplets or ice crystals [27], [28]. Hence, it is more likely for clouds to form at low pressure, low temperature, and high RH conditions, which translate to the local minimum points of the introduced reference signal. On the other hand, when at high pressure, high temperature, and lower RH conditions, i.e., at the local maxima of the reference signal, the cloud formation is less likely. When both of the signals are rising toward their Hi points, the atmospheric weather mostly behaves in a stable manner, and little to no cloud formation is generally expected.

Figs. 5 and 6, respectively, represent the result of synoptic event detection algorithm based on pressure signal, \( P \), and the reference signal, \( \frac{P \times T}{RH} \), for the same portion of the year 2012. It must be noted that the pressure signal gradients have been scaled up by making the delta pressures ten times larger for visualization purposes only, as shown in Fig. 5. Four relationships used to define types of days are as follows.

1) IF day \( d \) has a \( t_i = \text{MAX} \) AND \( t_i \) meets criteria (a)–(c), THEN \( d \) is a Hi.

2) IF day \( d \) has a \( t_i = \text{MIN} \) AND \( t_i \) meets criteria (a)–(c), THEN \( d \) is a Lo.

3) IF day \( d \) is after a Lo AND \( d \) is NOT a Hi, THEN \( d \) is a Rise.

4) IF day \( d \) is after a Hi AND \( d \) is NOT a Lo, THEN \( d \) is a Fall.

C. Correlation Function Analysis

Correlation analysis and an algorithm for weather synoptic systems detection are the two first steps toward adding intelligence to the proposed solar forecasting solution and making the framework adaptive. This analysis is implemented in time domain and, as represented in Fig. 3, consists of a filtering and shifting block, in loop with the correlation analysis unit. This process helps with adaptive training.

In the first stage of serial time-domain analysis, input data to the system are passed through filters of different lengths, and various shifted versions of these filtered data are generated and the correlation function for each and every one of those input variables with the GHI of the training subsets are obtained. The characteristics of filter and shift for which peak cross correlation between each input variable and the GHI occurs are used for modifying the input variables accordingly both in the training and test phases so that the dataset gets richer with regard to the information content.

Fig. 7 represents the correlation functions for eight input variables with the GHI. The correlation coefficient values between each and every variable with the GHI are represented based on different amounts of shift and using different filter sizes for the year 2012. Weather variables influence the cloud formation significantly, which in turn affects the amount of solar irradiance that reaches the ground. For instance, atmospheric pressure works as the driver for the weathersynoptic system so that a high-pressure system is regarded as the sign of a stable climate and a low-pressure system can be followed by an unstable stormy weather or overcast days. Hence, by computing the correlation functions between each of those variables with GHI, over the training window, in a continuous manner for each consecutive target day, we can not only obtain the refined training...
Fig. 7. Cross-Correlation Functions of Nine Input Variables With GHI Using Different Filters. (a) Time. (b) Zenith. (c) Azimuth. (d) Temperature. (e) Pressure. (f) Relative Humidity. (g) GHI. (h) OCC. (i) TCC.

438 dataset with respect to optimally correlated variables, but also use the information obtained by the climatic variables in order to detect the weather synoptic events’ beginnings and ends.

439 Using the weather synoptic event detection algorithm [26], the forecasting framework is able to adaptively determine the size and the coverage of training windows by eliminating portions of the data that are less information rich.

440 It must be noted that the minimum shift between the above-mentioned variables with respect to the GHI is equal to the forecast horizon. For instance, in order to do the day-ahead forecast, the correlation function of each of the mentioned variables, \( f(t) \), with the measured GHI, \( I(t) \), is computed neglecting the \( f(t) \) values recorded during the past 24 hours. This process is mathematically shown in (2)–(4), at the bottom of this page [5], where \( T_{sh} \) and \( L \), respectively, represent the shift value and the length of the low-pass filter, which is applied to the signal for correlation analysis and \( T_{fh} \) stands for the forecast horizon, which is the default shift value between the variable \( v(t) \) and the irradiance \( I(t) \). Parameter \( N \) represents the total number of samples that exist in the vector \( v(t) \), while \( N_{d}^j \) is a positive integer showing the number of samples in the training subset \( \tau \) for target day \( d \), i.e., \( S_{d}^j \), which is equal to the number of samples in \( I(t) \) itself. Term \( T_{sh} \) ranges from 0 to \( N_{d}^j \), while the filter length \( L \) is between 1 and 1440 samples, i.e., minutes.

\[
\tau_{T_{sh}} = \text{corr}(f(t - T_{fh} - T_{sh}), I(t))
\]

\[
= \frac{\sum_{i=N-N_{d}^j}^{N} (f(t_i - T_{fh} - T_{sh}) - \bar{f}) (I(t_{i-N+N_{d}^j+1}) - \bar{I})}{\sqrt{\sum_{i=N-N_{d}^j}^{N} (f(t_i - T_{fh} - T_{sh}) - \bar{f})^2 (I(t_{i-N+N_{d}^j+1}) - \bar{I})}}
\]

\[
\bar{f} = \frac{\sum_{i=N-N_{d}^j}^{N} f(t_i - T_{fh} - T_{sh})}{N_{d}^j}
\]

\[
\bar{I} = \frac{\sum_{i=N-N_{d}^j}^{N} I(t_{i-N+N_{d}^j+1})}{N_{d}^j}
\]

\[
f(t_i) = \frac{\sum_{j=\max(1,i-\frac{1}{2})}^{\min(N,i+\frac{1}{2})} v(t_j)}{\min(N,i+\frac{1}{2}) - \max(1,i-\frac{1}{2})}
\]

D. Feature Extraction

A fundamental problem in neural network research, as well as in many other disciplines, is finding a suitable representation of multivariate data, i.e., random vectors. For reasons of computational and conceptual simplicity, the representation is often sought as a linear transformation of the original data. In other words, each component of the representation is a linear combination of the original variables. Well-known linear transformation methods based on statistical concepts that use mathematical tools to obtain information in different forms such as feature extraction, blind source separation, etc., include principal components analysis (PCA), factor analysis, projection pursuit, and the ICA [29]–[31].

In the proposed framework, the underlying features of input data are extracted using ICA before being fed to the ANN for training and further pattern recognition.

The preprocessing done on the data prior to extraction of the independent components include centering, whitening, and projection using PCA [31].

E. ANN, Adaptive Training, and Test

The ANN mentioned in the proposed framework uses the upcoming entries of the dataset in order to retrain while it is used in real time for forecasting. This means that instead of being trained using a specific constant window of data and keep running afterwards based on the initial settings and weight values, the ANN continuously retrains itself using the data inside a moving window of varying length. This way, the ANN will keep itself as updated as possible so that it can match with intentional or unintentional changes that may occur in a system’s general structure or in the environment.

461
Taking advantage of an ANN structure as the core of the forecast engine, the proposed framework optimally analyzes and recognizes patterns in meteorological and solar data inputs and understands how different variables are related to each other and what would be the climate-related outcome of certain changing patterns in some variables with respect to others.

The neural network training procedure can be illustrated by the fact that the output of the next day is the target for inputs and output of the current day. This process continues until we get to the reference, or in some cases current, day. Then, the test phase begins. This means that the inputs and output of the reference day are fed into the already trained ANN, and the output of the ANN is kept recorded as the forecast result in order to be compared to the actual target signal, which will happen in future. The target signal may be the output power of the solar plant or the solar irradiance that reaches the ground.

The predefined lengths for the adaptive training windows are stored in a vector called training specifications vector, \( \varphi = [5, 8, 10, 12, 15, 20] \). Those predefined lengths, however, do not mean that every single days before current day will be included in the training until the number of days is reached. The days that will be put in the training subset will be determined based on training relationships. In order to form the training window, the following relationships consider the day types obtained by the weather systems detection algorithm on the pressure signal:

1) IF current day is “Lo” or “Rise” of “P”, THEN include days of type “Lo”, “Rise”, and “Hi”.
2) IF current day is “Hi” or “Fall” of “P”, THEN include days of type “Hi”, “Fall,” and “Lo.”

where \( X_i \) refers to the \( i \)th dataset and \( X \) is the original recorded dataset. \( \text{Rand}(\mu, \sigma) \) is a matrix of random variables with Gaussian distribution of mean \( \mu \) and standard deviation \( \sigma \). For the purposes of this study, we needed \( \mu = 0 \) and \( \sigma = X_{(v_{tol})} \), where \( v_{tol} \) is the tolerance vector that includes the tolerances, or COVs, of sensors designed for each and every one of the variables considered in Monte Carlo analysis. For variables such as time, zenith angle, and azimuth angle, the tolerance was considered to be 0; therefore, the values of those variables are the same in all of the 20 datasets. The COVs are, respectively, chosen to be 0.03, 0.15, 0.01, 0.02, 0.04, and 0.03 for atmospheric pressure, temperature, relative humidity, GHI, TCC percentage, and OCC percentage [32], [35].

After the forecast results are obtained for each of the \( N \) samples from the Monte Carlo approach, i.e., \( N_i \), \( i = 1, 2, \ldots, 20 \), the output percentiles are determined and the 50th percentile, i.e., \( \mu_y \), the median is considered as the stabilized forecast result. The following relationships represent how to compute the percentiles:

\[
\begin{align*}
    p_{50} \ (y) & = \bar{Y} - z_{50} \times \sigma_y \\
    p_{75} \ (y) & = \bar{Y} - z_{75} \times \sigma_y \\
    p_{90} \ (y) & = \bar{Y} - z_{90} \times \sigma_y
\end{align*}
\]

where \( \bar{Y} \) denotes the mean of all 20 outcomes of the 20 different datasets used for forecast and \( \sigma_y \) represents the standard deviation of those forecast results from Monte Carlo simulation. Finally, \( z \) stands for the \( z \)-score of each percentile. The \( z \)-scores for the aforementioned percentiles are as follows:

\[
z_{50} = 0 \quad z_{75} = 0.675 \quad z_{90} = 1.282.
\]

**F. Monte Carlo Uncertainty Analysis**

In order to stabilize the forecast results and to obtain reliable predictions of the available solar energy in the days ahead, it is required to consider the uncertainty associated with the recorded data. This uncertainty can be due to measurement noise, calibration offset, hardware failure, data-acquisition errors, or data-reduction issues. In the Monte Carlo approach for uncertainty analysis, a number of datasets similar to the originally recorded database are generated, the forecast process takes place for each dataset and, the statistical percentiles of the results will be considered as stabilized reliable outputs.

In our forecasts, the Monte Carlo analysis is done on 20 different datasets. Hence, 19 other datasets similar to the originally recorded database are generated considering the sensor tolerances, i.e., the coefficient of variation (COV), obtained from current industry state-of-the-art producers. Variables for which the COV is considered for Monte Carlo approach include atmospheric pressure, temperature, relative humidity, TCC and OCC, and GHI. The COV for each variable was the basis for generation of the other 19 seeds. The recorded value is assumed to be the actual value for each variable, and 19 other values are generated around this actual value as follows:

\[
X_i = \begin{cases} 
    X, & i = 1 \\
    X(1 + \text{Rand}(\mu, \sigma)), & i = 2, 3, \ldots, 20
\end{cases}
\]

where \( \mu \) and \( \sigma \) are the mean and standard deviation of the recorded value, respectively. The COV for each variable was the basis for considering the uncertainty in the forecast results. The COVs are, respectively, chosen to be 0.03, 0.15, 0.01, 0.02, 0.04, and 0.03 for atmospheric pressure, temperature, relative humidity, GHI, TCC percentage, and OCC percentage [32], [35].

**G. Dynamic Performance Evaluation**

Dynamic evaluation of performance helps with the adaptive training and test, and the outcome of this process is used for the aggregation procedure, which improves the accuracy of forecast results. The outcome of dynamic performance evaluation, i.e., the performance matrix, tends to be more reliable as time passes and more new data come into the database. In this section, the performances of each and every training subset used for training of the neural network for the last forecast period, i.e., previous target day are computed. Let us denote those subsets by \( S_d^r \), \( \tau = 1, 2, \ldots, \tau \), where \( \tau \) and \( \tau \) represent the training set number and the total number of sets considered for training, respectively. The beginning of the time until current time, denoted by \( S_d^t \), \( \tau = 1, 2, \ldots, \tau \), \( d = 1, 2, \ldots, d - 2 \), will also be considered in order to determine the coefficients of aggregation.

For each target day \( d \), after the training subsets are determined based on the predefined training window lengths considering the day types determined by the synoptic event detection algorithm, the training phase will take place on the ANN. After the ANN converges during the training process, the test phase begins. In this phase, main input consists of the same variables introduced previously.
before, in Fig. 2, for current day with the relevant filter and
shift values obtained from correlation analysis applied to the
select variables. This filtered and shifted input set undergoes the
feature calculation process by being multiplied to the separating
matrix, \( W \), obtained by ICA during the training phase. Then, the
features are fed to the trained ANN in order for the solar energy
estimates to be generated for the specific target day \( d \).

Once the forecast for target day \( d \) is obtained through the
ANN trained by the training subset \( \tau \), i.e., \( S^*_\tau \), the accuracy
of result must be computed in order for the training subset to
be evaluated. The performance function used in the process
of dynamic performance evaluation is defined as the ratio of
the correlation coefficient between forecast and target for each
target day \( d \) to the forecast RMSE of the same day

\[
\gamma_d^r = \frac{r_d^*}{\max(1, \text{RMSE}_d^*)} \quad (10)
\]

where \( \tau \) refers to the training set that used in order to train the
forecast engine for the specific target day \( d \). The higher the cross
analysis correlation between forecast and actual irradiance for a target
day \( d \), \( r_d^* \), and the lower the forecast error for the same day,
RMSE_d^*, and the better, i.e., higher, the performance value, \( \gamma_d^r \).

Considering the largest value of correlation coefficient, which is
1, it can be said that the highest possible value for performance
is 1, which represents the best possible case in day-ahead solar
forecasting, since the denominator of the performance function
is forced not to be smaller than 1.

For each target day \( d \), the performance values of the relevant
training subsets will form a vector called performance vector
and is added as a new row to the end of the performance matrix.

The mathematical representation is as follows:

\[
\gamma_d = \begin{bmatrix}
\gamma_1^d \\
\gamma_2^d \\
\vdots \\
\gamma_{d-1}^d
\end{bmatrix}
\gamma =
\begin{bmatrix}
\gamma_1 \\
\vdots \\
\gamma_{d-1}
\end{bmatrix} \quad (11)
\]

where \( \gamma_d \) represents the performance vector for target day \( d \)
and \( \gamma \) stands for the performance matrix. Finally, \( \gamma_{d-1} \) is the \( d \)
performance vector for the last target day, where \( d \) is the number \( 603 \)
of target day that is currently being forecast. Each row of the \( 604 \)
performance matrix includes the performance measure for all \( 605 \)
of the adaptive training subsets for a single target day.

### H. Aggregation Process and Fine-Tuning

After the performance matrix is obtained, the aggregation
coefficients need to be computed so that the weighted averaging
scheme can be applied to obtain the forecast result for target
day \( d \).

The aggregation coefficients are determined as follows and
are used to compute the forecast result for the target day:

\[
c_d^\tau = \begin{cases} 
\frac{1}{\sum_{\tau=1}^{d} v_d^\tau}, & d = 1 \\
\frac{\alpha \cdot c_d^\tau + \sum_{\tau=1}^{d-1} c_d^\tau / (d - 1)}{(\alpha + 1)}, & d > 1
\end{cases}
\quad (12)
\]

where \( c_d^\tau \) is the aggregation coefficient, i.e., the weight factor,
for the forecast result of target day \( d \) obtained using the training
subset number \( \tau \), i.e., \( S^*_\tau \), to train the ANN. This forecasting
result is denoted by \( y_d^\tau \). The total number of training subsets is \( \tau \).
The symbol \( \alpha \) stands for a constant coefficient, which is used to
give more dominance to the performance value of the last target
day, i.e., target day \( d - 1 \), with respect to the performances of
the previous target days. In our forecasts, \( \alpha \) is set to 10.

Assuming that \( y_d^\tau \) is the forecast obtained for target day \( d \) by
training set \( \tau \), the aggregated forecast can then be represented as follows:

\[
Y_d = \frac{\sum_{\tau=1}^{\tau} y_d^\tau \cdot c\tau^\tau}{\sum_{\tau=1}^{\tau} c\tau^\tau} \quad (13)
\]

where \( Y_d \) is the aggregated forecast for target day \( d \).

The relationship base used for fine-tuning of the forecast re-
sults for each target day \( d \) consists of the following relationships,
which take into account the current day’s condition and the day
types for current and past days. In the relationships mentioned
below, \( P \) and \( \frac{P}{P+h} \), respectively, refer to the atmospheric pres-
sure signal the second reference signal:

1. IF Current day is “Fall” of \( \frac{P}{P+h} \) AND “Lo” of \( P \),
   THEN Next day is “Overcast.”
2. IF Current day is “Fall” of \( \frac{P}{P+h} \) AND “Hi” of \( P \),
   THEN Next day is “Scattered.”
3. IF Current day is “Hi” of \( \frac{P}{P+h} \) AND “Hi” of \( P \),
   THEN Next day is “Scattered to Clear.”
4. IF Current day is “Hi” of \( \frac{P}{P+h} \) AND “Lo” of \( P \),
   THEN Next day is “Scattered to Overcast.”
5. IF Current day is “NOT Clear” AND “Fall” of \( \frac{P}{P+h} \)
   AND “Fall” of \( P \),
   THEN Next day is “Overcast.”
6. IF Current day is “Overcast” AND “Fall” of \( \frac{P}{P+h} \),
   THEN Next day is “Overcast.”

The rules mentioned above work as a modifier of the raw
forecast results obtained from the ANN-based forecast engine.
The forecast results are adjusted by averaging them with over-
cast and clear sky reference curves according to the conditions
mentioned in the relationships mentioned above.

### I. Day-Ahead Solar Forecast User Interface

The user interface consists of three main displays, each of
which represent specific information regarding the forecast of
solar energy and the relevant performance measures.

First display shows the forecasted irradiance curve and the
actual GHI curve for the target day along with the actual and
percent error values and average values of target and forecast.
This display is dynamically updated by time as new samples of
the actual target GHI are measured and recorded to represent fi-
nal, i.e., current state of forecast performance and the associated
error measures.

The second interface display, however, is a static display,
which represents the Monte Carlo uncertainty analysis results,
which are the two curves of 50th and 90th percentiles, for current day
and tomorrow. There is also a table represented in the display,
which summarizes the optimal market bids values for solar en-
ergy by representing hourly averaged solar irradiance forecasts.
based on three main percentiles obtained by the Monte Carlo uncertainty analysis approach. For most conservative bids, the 90th percentile, i.e., $\rho_{90}$ can be used.

The last interface display includes the forecast and actual curves for current day and tomorrow using two different approaches of 24 h ahead and two days ahead. This display focuses on comparing the different between forecast performance using two approaches already mentioned.

The user/operator also has the ability to impose changes to the process by adjusting different settings, which are offered through the interface such as the forecast horizon, the processes to be included in or excluded from the forecasting framework, etc.

IV. PERFORMANCE EVALUATION METRICS

Performance of the proposed method is evaluated for the hourly average values of day-time target and forecast GHI of the year 2012. The error indicators are the following:

Mean error (ME) or bias error (BE)

\[
BE = \frac{1}{N} \sum_{i=1}^{N} (T_i - F_i)
\]

(14)

Mean absolute error (MAE)

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |(T_i - F_i)|
\]

(15)

Root mean squared error (RMSE)

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (T_i - F_i)^2}
\]

(16)

where $N$ represents the total number of points for which the error values are calculated. The year 2012 is a leap year and is, therefore, comprised of 8784 hour points. For hourly averaged solar irradiance values, the night-time hours must be taken out from the final set in order to obtain clear understanding of system performance using only day-time samples, i.e.,

\[
T_i = \frac{1}{n} \sum_{j=(i-1)n+1}^{in} (\text{GHI}), j > 0
\]

where $T_i$ and $F_i$ are the $i$th sample of the target and forecast, respectively, and $C_i$ represents the maximum solar energy capacity, i.e., the maximum amount of solar energy that may reach the ground, at time instance $i$. The percent error will vary in the range of -100% and 100%. The mean value and the standard deviation of percent error serve as additional performance measures and are computed as follows:

\[
\mu^\% = \frac{1}{N} \sum_{i=1}^{N} e_i^\%
\]

(22)

\[
\sigma^\% = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (e_i^\% - \mu^\%)^2}
\]

(23)

The correlation coefficient represents the strength of the linear association between target and forecast. In other words, it tells us how well the forecast aligned with the actual upcoming deviations in the solar irradiance curve. The correlation coefficient is computed as follows:

\[
r = \frac{1}{N-1} \sum_{i=1}^{N} \left( \frac{T_i - \bar{T}}{\sigma_T} \times \frac{F_i - \bar{F}}{\sigma_F} \right)
\]

\[
= \frac{\sum_{i=1}^{N} (T_i - \bar{T})(F_i - \bar{F})}{\sqrt{\sum_{i=1}^{N} (T_i - \bar{T})^2} \sqrt{\sum_{i=1}^{N} (F_i - \bar{F})^2}}
\]

\[
T_i = \frac{1}{n} \sum_{j=(i-1)n+1}^{in} (\text{GHI}), j > 0
\]

(24)

where $\sigma_T$ and $\sigma_F$, respectively, stand for the standard deviation of the target samples and their associated forecast. The more closer the correlation coefficient, represented by “$r$”, is to 1, the better the changes in our forecast curve of solar energy are matched and aligned with those of the actual solar irradiance, i.e., the target, curve.
V. FORECAST RESULTS AND DISCUSSION

Forecast is done for the whole duration of the year 2012 and the forecast of the solar energy for the next day is obtained for the 366 days in that year, and the results are compared to the state of the art in forecasting industry based on standard measures of BE, MAE, RMSE, and correlation coefficient.

In Fig. 8, performance charts are represented for comparing the performance of the proposed forecast engine with the standard reference case of persistence for one-day-ahead forecast. Three performance charts are represented for each the two cases of the proposed method and the persistence as the target versus forecast scatter plot, the actual error histogram, and the percent error histogram. The measures BE, MAE, RMSE, $\mu e \%$, $\sigma e \%$, and the correlation coefficient are calculated for the hourly averaged solar energy of day-time samples only.

Fig. 9 represents the monthly performance analysis results. It can be seen from Fig. 9 that the RMSE values obtained by the proposed forecasting framework are consistently smaller than those of the persistence case for next-day prediction. The relative RMSE is obtained by dividing the RMSE values to the average measured GHI for each month.

The performance of the proposed framework is comparable to some of the dominating benchmarked solutions for solar energy forecast in European Union, which are mainly based on NWP methods. An overview on the statistical error measures for the complete Austrian dataset is given in Table I for one-day-ahead solar energy forecast [11]. Forecast methods presented by the University of Oldenburg, Meteotest, CENER, the statistical method of Blue Sky (BLUE), and the traditional synoptic method of the meteorologists of Blue Sky (SYNOP) are evaluated. The SYNOP method is spatially restricted and only available for Austria [11]. Actual and relative values for RMSE, MAE, and BE are given in Table I for the next-day forecast in Austria for European NWP-based approaches and their persistence method for duration 2007 to 2008.

For the year 2012, forecast results obtained using the method proposed in this paper, i.e., the UTSA Method, and the persistence approach for the specific NREL site at Golden, CO, USA, are also included in Table I. The actual improved RMSE values for a method with respect to persistence are the difference between the RMSE of persistence and that of the method.
The relative improvement value is defined as follows:

$$\vartheta = \frac{\text{RMSE}_{\text{Persistence}} - \text{RMSE}_{\text{method}}}{\text{RMSE}_{\text{Persistence}}} \times 100$$  \hspace{1cm} (25)

From Table I and Fig. 8, it is seen that the proposed approach strongly outperforms the persistence for one-day-ahead forecast based on the results obtained for year 2012. Considering the relative improved RMSE values, the proposed UTSA Method stands in the third place after BLUE and ECMWF-OL. This means that the proposed framework for the UTSA Method outperforms CENER, WRF-MT, and SYNOP. Noticing that SYNOP is a real-time day-ahead approach done by human meteorology experts, it is inferred that the proposed framework for solar energy forecast gives better performance compared to the predictions suggested by the experts in the field. The proposed framework “UTSA Method” also shows a promising performance using the 2012 NREL dataset of Golden, CO, compared to the methods introduced in Table I for 2007–2008 datasets of Switzerland, Germany, and Spain by mostly securing the third rank following ECMWF-OL and BLUE in RMSE improvement with respect to persistence [5]. Furthermore, our method here represents predictions of solar irradiance in a point in space. As the predictions are distributed in space over a larger area, say, in a megawatt solar plant, our forecasting method is expected to outperform others, since patterns can be recognized on a distributed manner, thus adding intelligence to the prediction, which would not be possible otherwise.

VI. CONCLUSION

An adaptive site-specific framework for next-day solar energy forecast is developed based on a combination of data-analytic approaches and artificial intelligence techniques. The models are developed and validated utilizing a raw large dataset from the NREL archive for a site in Golden, CO, USA.

A uniqueness of the methodology developed is that an integrated serial time-domain analysis coupled with multivariate analysis is used for preprocessing of the dataset. The resulting enhanced dataset is used for adaptive training of a smart forecast engine, which is implemented using an ANN structure. Standard performance measures are considered, the forecast results are obtained, and the accuracy of forecasts is studied and discussed for the year 2012. Performance of the proposed framework is comparable to that of the current state of the art in day-ahead solar energy forecasting methodologies in European Union, which are mainly based on NWP.

The capabilities of the proposed data-analytic based adaptive approach compared to physics-based models are illustrated with the fact that after just one year of research and development of a day-ahead solar forecasting tool, the method compares better than some methods in Europe and very competitive with the leading approaches. The team at the UTSA has a strategy to improve its method aiming to provide accuracies better than the benchmarks under development for decades.


ASOS/


Yashar Sahraei Manjili (GS’09) received the B.S. degree in bioeclastics from the Amirkabir University of Technology, Tehran, Iran, in 2004, the M.S. degree in automation and instrumentation engineering from the Petroleum University of Technology, Tehran, in 2007, and the Ph.D. degree in electrical engineering, system and control from University of Texas at San Antonio (UTSA), San Antonio, TX, USA, in 2015.

He is currently a Decision Science Analyst with the Innovation and Applied Analytics Division at USAA, San Antonio, TX. Before this position, he was a researcher with the Autonomous Control Engineering (ACE) Lab, UTSA and the Texas Sustainable Energy Research Institute (TSERI). He is the co-owner of two intellectual properties and has co-authored numerous journal articles, conference papers and book chapters in the areas of his research interest which include control systems and renewable energy with a focus on data analytics for smart microgrid control, solar forecast, and fuzzy systems.

Rolando Vega received the B.S. degree in civil and environmental engineering from the University of Puerto Rico at Mayaguez, Mayaguez, Puerto Rico, and the Ph.D. degree in wind science and engineering from Texas Tech University, Lubbock, TX, USA.

He has been an Energy Consultant and Researcher for the last eight years. He started and was responsible for the renewable energy business in the U.S., Mexico, Brazil, and China for a global 1700+ employee company. He is currently a Manager of Analytics and Business Insight with CPS Energy, San Antonio, TX. Before this position, he was with the Texas Sustainable Energy Research Institute, University of Texas at San Antonio, as the Director of research and innovation, and the ABS Consultant leading the technical and business development efforts of a global team to assess, predict, and mitigate effects of wind and solar power on renewable energy assets.

Dr. Vega is a Registered Professional Engineer and holds an active NCEES record for licensure in any U.S. state. In 2009, he was nominated for the American Association for Wind Engineering Award for his Ph.D. dissertation.

Mo M. Jamshidi (F’99) received the B.S. degree in electrical engineering from Oregon State University, Corvallis, OR, USA, in 1967, and the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1969 and 1971, respectively.

He is currently the Lutcher Brown Endowed Chair Professor with the University of Texas, San Antonio, TX, USA. He has been an advisor to NASA, USAF, USDOE, and EC/EU. He has authored or co-authored more than 700 technical publications including 68 books (11 text books), research volumes, and edited volumes. His current research interests include the system of systems engineering with an emphasis on cloud computing, robotics, unmanned aerial vehicles, and sustainable energy systems.

Dr. Jamshidi is a Fellow of the American Society of Mechanical Engineers, the American Institute of Aeronautics and Astronautics, the American Association for the Advancement of Science, the World Academy of Sciences, and the New York Academy of Sciences. He is the Founding Editor or Co-Founding Editor or Editor-in-Chief of five journals including the IEEE Control Systems Magazine and the IEEE SYSTEMS JOURNAL. He is also the Editor-in-Chief of AutoSoft Journal and Co-Editor-in-Chief of the Journal of Automation and Control, both published in the U.K. He is an Honorary Professor at three Chinese Universities (Nanjing and Xi’an, Deakin University (Australia), Birmingham University (U.K.), and Obuda University (Hungary). He holds honorary doctorate degrees from Ordlar Yordu University, Baku, Azerbaijan, 1998, the University of Waterloo, Waterloo, ON, Canada, 2004, and the Technical University of Crete, Chania, Greece, 2004. He was the recipient of the IEEE’s Norbert Weiner Research Achievement Award in 2005 and the IEEE-USA Career Award for Systems Engineering in 2014. He has been a member of the University of Texas System Chancellor’s Council since 2011.
Q1. Author: In order to maintain numerical order, equations have been renumbered Please check.
Q2. Author: The same symbol “τ” is used to represent both the training set number and the total number of sets in the sentence “Let us denote those subsets by . . . ” Please check.
Q3. Author: Please provide the page range in Refs. [4] and [9].
Q5. Author: Please provide the report number in Ref. [20].
Q7. Author: Please provide publisher details in Ref. [27].
Q8. Author: Please provide the year in which the author Mo M. Jamshidi became a Fellow of the IEEE.