Forecasting Solar Irradiance from Time Series with Exogenous Information

Kousuke Ariga, Kaiyu Zheng
University of Washington
{koar8470, kaiyuzh}@cs.washington.edu

Abstract

Solar irradiance forecast is the pivotal factor for power generation scheduling at photovoltaic (PV) power plants. To this end, we propose a framework to model solar irradiance that leverages both short-term dependency (several days) and dependency on time in the annual cycle. We focus on two classes of methods, autoregressive (AR), and autoregressive with exogenous variables (ARX), and investigate LASSO regression, and non-linear methods, including support vector regression (SVR) and long short-term memory (LSTM) as the underlying mechanism. Our experiments on data collected at a PV power plant for over five years show that considering external information such as weather on top of the historical solar irradiance data for longer period of time lowers the error of solar irradiance forecast.

1 Introduction

Photovoltaic (PV) power plants are energy farms that enable large-scale generation of solar power. Because power generation from solar panels is directly proportional to solar irradiance [10], it is crucial for a PV power plant to forecast the level of solar irradiance a certain time period in advance, in order to optimize the operating cost through generation scheduling [6]. Using complex physics-based models is generally considered expensive due to high computational costs [9]. Hence, in this project, we aim to use cost-effective machine learning techniques to learn models that consider weather information and are capable of forecasting solar irradiance reliably.

Solar irradiance can be naturally represented by a time series. Among the numerous methods in analyzing time series data, autoregressive (AR) models attempt to predict future values of the time series variable by only considering past observations of the variable, whereas autoregressive models with exogenous variables (ARX) is a class of methods that consider external information.

We are provided with daily solar irradiance time series data by Shizen Energy Inc., and we gathered hourly weather data near the plant location by scraping from Japan Meteorological Agency. We conducted statistical analysis of stationarity, seasonality, and trend of the solar irradiance time series. We found that the solar irradiance data exhibits seasonality for a cycle of one year. This suggests the solar irradiance on a certain day in a year is influenced by where the day lies with respect to the year. In addition, solar irradiance also depends on weather conditions that may change within a few days.

Therefore, we propose a framework for learning models to forecast solar irradiance, where both short-term dependencies and dependency on time in the annual cycle can be learned (Section 4). We experimented with linear and non-linear models for AR and ARX. For linear, we used LASSO regression. For non-linear, we used support vector regression (SVR), and long short-term memory (LSTM). We found that SVR and LSTM outperform LASSO regression with a slight gain.

Notation We use Y to denote solar irradiance variable, $\boldsymbol{X}$ to denote a set of exogenous variables. We use y_t to denote observation of solar irradiance at time t, and $\boldsymbol{x}_t$ to denote observations of

exogenous variables. For two time points t_1 and t_2 where $t_2 > t_1$, $y_{t_1:t_2} = [y_{t_1} \quad y_{t_1+1} \quad \cdots \quad y_{t_2}]^T$, and similar goes for $\boldsymbol{x}_{t_1:t_2}$. Other notations are defined in the context.

2 Related Work

Several studies have been conducted on the problem of using machine learning methods to predict solar irradiance. Prema and Rao [9] used traditional autoregressive methods such as moving average and triple exponential smoothing and were able to achieve a low error rate of 9.28% in prediction. In this work, however, such traditional statistical models perform poorly and are not recommended for use. One possible explanation is that [9] used hourly solar irradiance data, which exhibit much less variance than daily solar irradiance data used in our case.

Similar to this work, Sherma et. al. [11] proposed to learn site-specific prediction models for different power plants. Their models are based on linear least squares and support vector machine (SVM); they learn a function $y_t = f(\boldsymbol{x}_t)$ that maps, for a certain time t (day), the weather information and day in year (together as $\boldsymbol{x}_t$) to solar irradiance y_t. Different from their approach, we use ARX models that take into account values of solar irradiance and weather in the past.

Exogenous variables have been considered in predicting time series in different domains. Damon and Guillas [3] used wind speed and temperature for predicting ozone concentration, and e Silva et. al. [4] used season of year for predicting energy prices. For solar irradiance prediction, similar to this work, Alzanhrani et. al. [1] also used several weather attributes as well as day of the year as exogenous information to train a neural network; however, they only tested their model to predict solar irradiance generated from a physical model, which is far too regular compared to real observations.

3 Dataset

Our data came from two sources: solar irradiance data provided by Shizen Energy Inc. and weather data scraped from Japan Meteorological Agency. Specifically, we obtained daily observational solar irradiance data for almost 5 years starting January 2013 and collected hourly observational weather data for the corresponeding duration from http://www.jma.go.jp. The raw format of solar irradiance and weather data as well as our preprocessing methods are presented below.

time stamp (daily)	solar irradiance (kWh/m^2)

Table 1: Solar Irradiance Dataset

time stamp (hourly)	station pressure (hPa)	sea-level pressure (hPa)	rain fall (mm)	temp. (C)	dewpoint temp. (C)	vapor pressure (hPa)	humidity (%)	wind speed (m/s)
	wind direction (categorical)	sun duration (h)	station solar irr (MJ/m^2)	snow fall (cm)	snow accum. (cm)	weather (categorical)	cloud amount (%)	visibility (km)

Table 2: Weather Dataset

Note that the weather data includes missing values, some of the features are numerical and others are categorical, and the scale of the features varies. To facilitate efficient leanring by machine learing algorithms, we preprocessed in the following ways.

1. Interpolate missing values linearly in both direction in terms of time.
2. Replace categorical variables by dummy variables.
3. Normalize all the features except the dummy variables introduced in Table 2.

Moreover, weather data is originally hourly data whereas solar irradiance is daily data. LSTM described in Section 4.1.3 requires dataset to be in the same time unit, therefore, we further processed

the weather data by taking the mean over 24 hours for numerical variables and by normalizing the total counts for dummy variables.

3.1 Statistical analysis

We analyzed the stationarity, seasonality, and trend of the solar irradiance time series data. First, to study the stationarity, we employed the *Variogram* [8] which computes $G_k = \text{Var}(y_{t+k} - y_t)/\text{Var}(y_{t+1} - y_t)$, for $k \in \{1, 2, \cdots, \}$. We observe from Figure 1 (top) that the value of G_k is upper bounded with regular cycles, which suggests stationarity. The stationarity property can also be seen from Figure 1 (bottom), where the annual mean is almost constant, and the daily values vary within a bounded range. Also, from the same figure, we can see that the raw solar irradiance data exhibits seasonality for a cycle of around one year. Its value reaches maximum in the summer, minimum in the winter. However, solar irradiance has high fluctuations for a short period of time (i.e. several days). This suggests, as in [9][10], that solar irradiance is correlated with weather attributes; weather changes with regularity seasonally, but may change drastically within a shorter time frame. To investigate the correlation between solar irradiance and weather attributes, for each time step, we plotted the (averaged) weather data and the corresponding solar irradiance, shown in Figure 2. We observe that humidity and sunlight duration have strong correlation with solar irradiance. In addition, we notice that in the location chosen for the power plant, it seldomly rains, and that solar irradiance is low as the amount of rain increases.

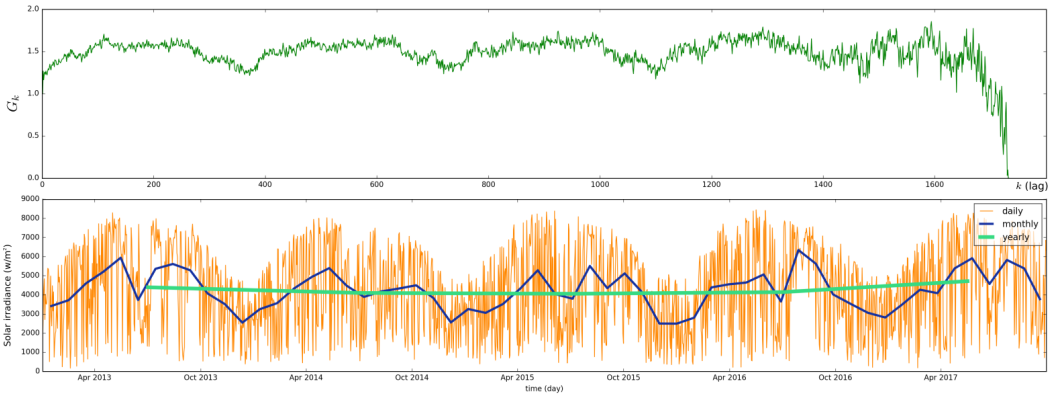


Figure 1: Plots of variogram (top), daily observations, weekly average, and yearly average (middle).

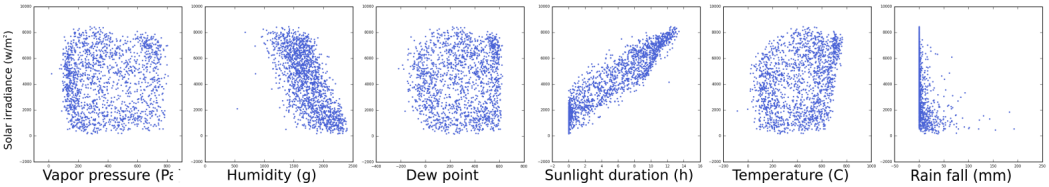


Figure 2: Correlations between solar irradiance and several weather metrics.

3.2 Smoothing models

As an extention of the statistical analysis of the dataset, we investigated several basic autoregressive time series models. These models attempt to arrive at a prediction through simply smoothing N most recent observations. Our goal is to obtain baseline results for comparison with more complex regression models. To this end, we used the simple moving average model and the n th order exponential smoothing model. Simple moving average computes a smoothed value at time T by taking the average of N most recent observations before T. The nth order exponential smoothing works by recursively applying decreasing weights to observations that are less recent, using a provided discount factor λ:

$$\hat{y}_t^{(n)} = \lambda \hat{y}_t^{(n-1)} + (1 - \lambda)\hat{y}_{t-1}^{(1)} \tag{1}$$

where $\hat{y}_t^{(0)} = y_t$, and y_0 could be specified by the user (e.g. simply 0). Figure 3 illustrates the extent that these simple models can fit to the data.

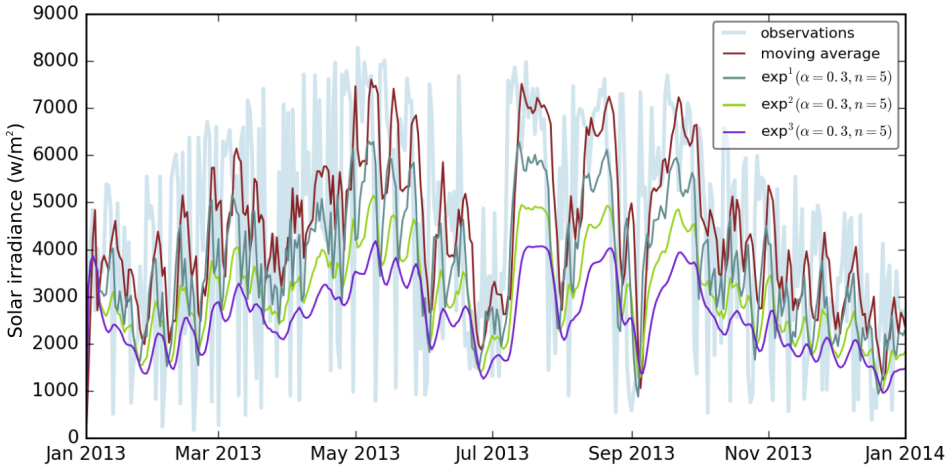


Figure 3: Solar irradiance forecasts by smoothing models.

4 Methodology

4.1 Framework for Learning Solar Irradiance Models

The modeling of solar irradiance should consider short-term dependency due to the correlation with weather attributes, as well as dependency on time in the annual cycle due to seasonality. To this end, we propose a framework for learning such models. In this framework, short-term dependency is built-in to the AR and ARX model, defined as follows.

$$\text{AR: } y_{t+h} = f(y_{t-u:t}) \tag{2}$$

$$\text{ARX: } y_{t+h} = f(y_{t-u:t}, \boldsymbol{x}_{t-u:t}) \tag{3}$$

where u defines the size of the *window* which contains recent observations, assumed to affect y_{t+h}, the forcast after temporal horizon h. The above fomulation implies the assumption that the solar irradiance y_{t+h} is only determined by past observations within a window of the target variable $y_{t-u:t}$ (as well as that of the exogenous variables $\boldsymbol{x}_{t-u:t}$ if the model is ARX). Dependency on time in the cycle is incorporated into the training data for ARX models as day in the year.

Training samples are of the form $(\boldsymbol{z}_{t-u:t}, y_{t+h})$, where

$$\boldsymbol{z}_{t-u:t} = \begin{bmatrix} \boldsymbol{z}_{t-u}^T \\ \vdots \\ \boldsymbol{z}_t^T \end{bmatrix} = \begin{cases} \begin{bmatrix} y_{t-u} \\ \vdots \\ y_t \end{bmatrix} & \text{For AR models} \\ \begin{bmatrix} \boldsymbol{x}_{t-u}^T & y_{t-u} \\ \vdots & \vdots \\ \boldsymbol{x}_t^T & y_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_{t-u}^T & d_{t-u} & y_{t-u} \\ \vdots & \vdots & \vdots \\ \boldsymbol{w}_t^T & d_t & y_t \end{bmatrix} & \text{For ARX models} \end{cases} \tag{4}$$

Here, for $i \in \{t-u, \cdots, t\}$, the exogenous variables $\boldsymbol{x}_i$ are specified as $\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{w}_i^T & d_i \end{bmatrix}$, where $\boldsymbol{w}_i$ is the weather data in the window, and $d_i \in \{1, \cdots, 365\}$ are days in the year for the time point i. The information of day in the year enables the model to learn dependency between the solar irradiance observation and its time point relative to the annual cycle, which is then reflected by $d_{t-u:t}$. During training, we can flatten $\boldsymbol{z}_{t-u:t}$ such that it becomes a feature vector of u dimensions for AR models, and $(2 + |\boldsymbol{w}_i|)u$ dimensions for ARX models. Denote the flattened $\boldsymbol{z}_{t-u:t}$ as $\boldsymbol{z}_{t-u:t}^F$.

4.1.1 LASSO Regression

The first ARX model we consider is LASSO regression, that is, linear regression with LASSO regularization. Weights $\boldsymbol{\beta}$ are learned from optimizing $\ell(\boldsymbol{\beta})$ as follows:

$$\ell(\boldsymbol{\beta}) = \frac{1}{2(T-h)} \sum_{t=1}^{T-h} (y_{t+h} - \boldsymbol{\beta}^T \boldsymbol{z}^F_{t-u:t})^2 + \lambda ||\boldsymbol{\beta}||_1 \tag{5}$$

where T equals to the total number of time points in the solar irradiance time series.

4.1.2 Support Vector Regression

Next, we consider the SVR model. We can formulate the SVR objective for our context as follows, as described in `LibSVM` [2][1]:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T K(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \epsilon \sum_{t=1}^{T-h} (\alpha_t + \alpha_t^*) + \sum_{t=1}^{T-h} y_{t+h}(\alpha_t + \alpha_t^*) \\ \text{subject to} \quad & \boldsymbol{e}^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) = 0 \\ & 0 \le \alpha_t, \alpha_t^* \le C, t = 1, \cdots, T-h \end{aligned} \tag{6}$$

where $K_{ij} = K(\boldsymbol{z}^F_{i-u:i}, \boldsymbol{z}^F_{j-u:j})$ is the kernel function. We used RBF kernel in experiments.

4.1.3 Long Short-Term Memory networks

Lastly, we consider Long short-term memory networks (LSTMs)[5], which is a class of neural network. Unlike common feedforward neural networks or convolutional neural network, LSTMs allow information to persist inside the model by having loops in their architecture. Because of this feature, LSTMs are capable of learning long-term dependencies and have been performing well on timeseries forecasting problems [7] and many more interesting problems including language modeling [12] and image captioning [13]. As other neural networks, LSTMs also have many hyperparameters such as number of layers, number of units in a layer, dropout ratio, and so on. We show the architecture selected by crossvalidation as described in the next section in Figure 4.

5 Experiments

In our experiments, we forecasted the solar irradiance of one day ahead ($h = 1$) given the solar irradiance and weather data of previous u days (window size). We examined models that span multiple levels of complexity, namely, LASSO regression, SVR with RBF kernel, and LSTM. The models were trained both autoregressively and with exogenous information, and different window sizes, $u \in \{1, 7, 30, 365\}$, were explored.

We used the first 70 percent of the timeseries data, which contains approxiamtely 1200 samples, as training dataset and the rest as testing dataset. The hyperparameters were chosen through 10-fold crossvalidation. We evaluated the models using the testing dataset by mean absolute error (MAE) between the observational data and the forecasted value.

5.1 Results and Discussion

The forecasts by autoregressive exogenous models are correlated to the observed solar irradiance as illustrated by Figure 5, and its fit is not shifted unlike the forecasts by the basic smoothing models shown in Figure 3.

For window size 1 to 30, adding exogenous information consistently reduced the error as Figure 6 shows. This result is reasonable because as we presented in the Dataset section, there are significant short-term correlation between solar irradiance and weather. Therefore, by using previous weather data on top of the solar irradiance data itself, estimators should be able to make better forecasts.

However, the error for LASSO and SVR significantly increased with weather data when window size was set to 365. We reasoned this is because of the lack of model complexity. Given a large number

[1] This objective corresponds to $\epsilon-$SVR

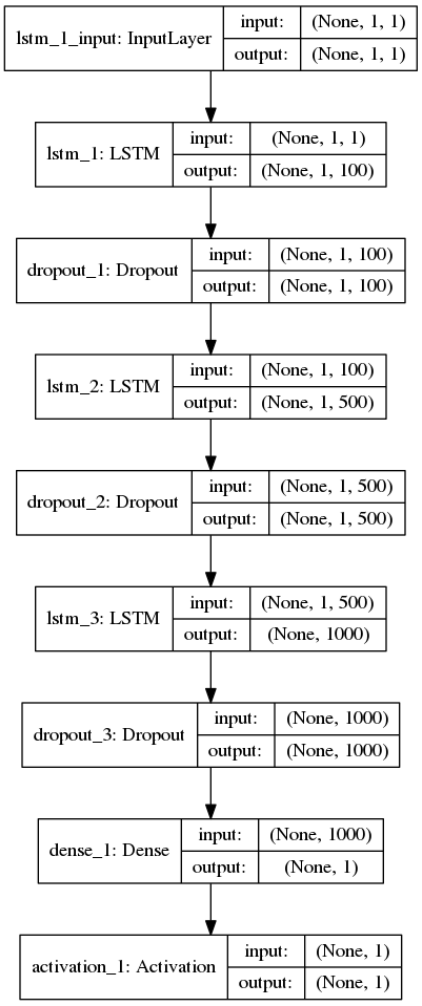


Figure 4: The architecture of the LSTM network. The first dimension is None because it varies depending on the number of samples, and the output dimension is 1 because the network outputs a scalar value prediction.

of features (365 days $\times$ 43 features $\approx 16,000$), training of the relatively simpler models converged without learning much. This claim is supported to some extent by the fact that, a more complex model, namely LSTM, outperformed the other models by far.

SVR performed the best for small window size like 1 to 7. One possible explanation is that LASSO, as a linear model, lacked the representational power to learn the relationship between solar irradiance and weather. On the other hand, LSTM, as a high dimensional non-linear model, lacked the amount of data to fit to the underlying structure that may exist. SVR is a medium-sized non-linear model, so it was just good in this particular case for smaller window size.

6 Conclusion

We have introduced a framework to train models for solar irradiance that consider both short-term dependency as well as dependency on time in the annual cycle. Our experiments demonstrate that by adding weather attributes and day in the year as exogenous information, the prediction error is reduced. Besides, the best accuracy was obtained when longer time dependency was considered with more complex model, namely the long short-term memory network. Indeed, even though our proposed methods are cost-effective and can forcast the general trend of solar irradiance, the

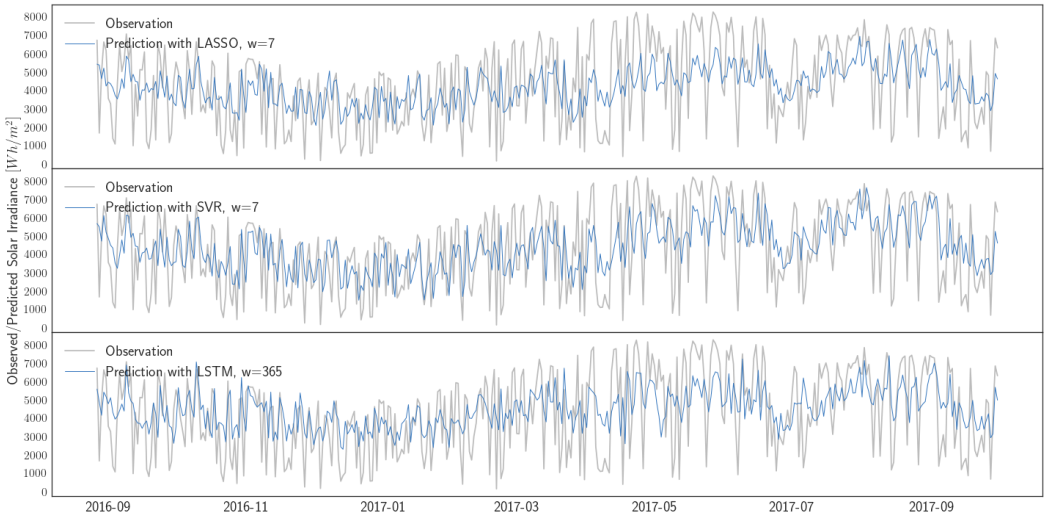
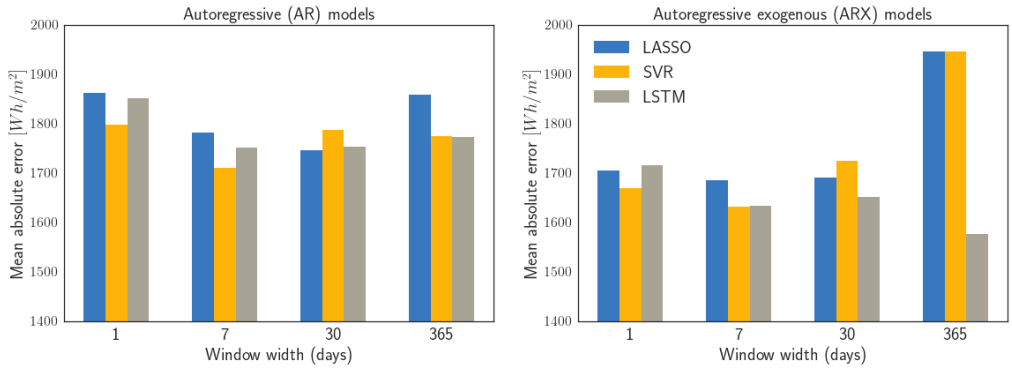


Figure 5: Solar irradiance forecasts by autoregressive exogenous models with the window size that performed the best on testing dataset. The plots are for LASSO, SVR, and LSTM from the top.



(a) MAE by autoregressive models.

(b) MAE by autoregressive exogenous models.

Figure 6: Comparison of autoregressive models vs. autoregressive exogenous models.

prediction error is significantly greater than physics-based methods which has an MAE of around 400 Wh/m^2. Nevertheless, there is plenty of room to improve. First, the model can incorporate long-term dependency (one year) by training on samples that contain observations on the same day in year from all other years. Second, currently the weather data that we obtained is only an approximation of the weather at the plant location, and we used linear interpolation to fill in missing observations. The performance can potentially benefit if weather data is more accurate. Finally, we can consider more exogenous information that may have higher correlation with solar irradiance, such as solar activity, satellite images of cloud above the plant location, and so on.

References

[1] A. Alzahrani, J. Kimball, and C. Dagli. Predicting solar irradiance using time series neural networks. *Procedia Computer Science*, 36:623–628, 2014.

[2] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[3] J. Damon and S. Guillas. The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics*, 13(7):759–774, 2002.

[4] E. C. e Silva, A. Borges, M. F. Teodoro, M. A. Andrade, and R. Covas. Time series data mining for energy prices forecasting: An application to real data. In *International Conference on Intelligent Systems Design and Applications*, pages 649–658. Springer, 2016.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. 9:1735–80, 12 1997.

[6] R.-H. Liang and J.-H. Liao. A fuzzy-optimization approach for generation scheduling with wind and solar energy systems. *IEEE Transactions on Power Systems*, 22(4):1665–1674, 2007.

[7] D. L. Marino, K. Amarasinghe, and M. Manic. Building energy load forecasting using deep neural networks. *CoRR*, abs/1610.09460, 2016.

[8] D. C. Montgomery, C. L. Jennings, and M. Kulahci. *Introduction to time series and analysis forecasting*. John Wiley & Sons, 2015.

[9] V. Prema and K. U. Rao. Development of statistical time series models for solar power prediction. *Renewable Energy*, 83:100–109, 2015.

[10] N. Sharma, J. Gummeson, D. Irwin, and P. Shenoy. Cloudy computing: Leveraging weather forecasts in energy harvesting sensor systems. In *Sensor Mesh and Ad Hoc Communications and Networks (SECON), 2010 7th Annual IEEE Communications Society Conference on*, pages 1–9. IEEE, 2010.

[11] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy. In *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*, pages 528–533. IEEE, 2011.

[12] M. Sundermeyer and y. Ralf Schlüter and Hermann Ney, booktitle=INTERSPEECH. Lstm neural networks for language modeling.

[13] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.