

***SENSOR NETWORK BASED SOLAR FORECASTING  
USING A LOCAL VECTOR AUTOREGRESSIVE RIDGE  
FRAMEWORK***

Xu, J., Yoo, S., Heiser, J., and Kalb, P.

*Published in  
Proceedings of the  
31st Annual ACM Symposium on Applied Computing  
Pisa, Italy  
April 4-8, 2016*

**Department/Division/Office**

**Brookhaven National Laboratory**

**U.S. Department of Energy  
DOE Office of Science**

Notice: This manuscript has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DE-SC0012704 with the U.S. Department of Energy. The publisher by accepting the manuscript for publication acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

This preprint is intended for publication in a journal or proceedings. Since changes may be made before publication, it may not be cited or reproduced without the author's permission.

## **DISCLAIMER**

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or any third party's use or the results of such use of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof or its contractors or subcontractors. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

# Sensor Network Based Solar Forecasting Using a Local Vector Autoregressive Ridge Framework

Jin Xu  
Electrical & Computer  
Engineering Dept.  
Stony Brook University  
jin.xu@stonybrook.edu

Shinjae Yoo  
Computational Science Center  
Brookhaven National Lab  
sjyoo@bnl.gov

John Heiser, Paul Kalb  
Environmental Sciences Dept.  
Brookhaven National Lab  
{heiser, kalb}@bnl.gov

## ABSTRACT

The significant improvements and falling costs of photovoltaic (PV) technology make solar energy a promising resource, yet the cloud induced variability of surface solar irradiance inhibits its effective use in grid-tied PV generation. Short-term irradiance forecasting, especially on the minute scale, is critically important for grid system stability and auxiliary power source management. Compared to the trending sky imaging devices, irradiance sensors are inexpensive and easy to deploy but related forecasting methods have not been well researched. The prominent challenge of applying classic time series models on a network of irradiance sensors is to address their varying spatio-temporal correlations due to local changes in cloud conditions. We propose a local vector autoregressive framework with ridge regularization to forecast irradiance without explicitly determining the wind field or cloud movement. By using local training data, our learned forecast model is adaptive to local cloud conditions and by using regularization, we overcome the risk of overfitting from the limited training data. Our systematic experimental results showed an average of 19.7% RMSE and 20.2% MAE improvement over the benchmark Persistent Model for 1-5 minute forecasts on a comprehensive 25-day dataset.

## CCS Concepts

- **Computing methodologies** → *Supervised learning by regression*;
- **Applied computing** → *Environmental sciences*;

## Keywords

Solar Forecast; Sensor Network; local vector autoregressive (LVAR)

## 1. INTRODUCTION

Solar energy is emerging as the most promising energy resource to address the world's increasing energy demand and depletion of non-renewable energy sources. In the past decade, significant technological improvements and rapidly dropping costs of solar panels have made them more affordable than ever. As of 2014, 158 GW

©2016 Association for Computing Machinery. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the United States government. As such, the United States Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.

SAC 2016, April 04-08, 2016, Pisa, Italy

©2016 ACM. ISBN 978-1-4503-3739-7/16/04... \$15.00

DOI: <http://dx.doi.org/10.1145/2851613.2853124>

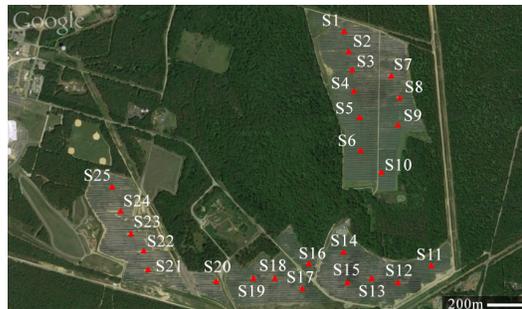


Figure 1: Locations of the 25 irradiance monitoring sensors in Long Island Solar Farm (LISF), New York, USA. Source: Google Maps, retrieved on August 30th, 2015.

of non-concentrated photovoltaic (PV) modules have been installed worldwide [1]. The U.S. solar industry grew by 34% since 2013, and is planning to double the existing solar capacity over the next two years [2]. However, unlike conventional power sources, the nature of surface solar irradiance is inherently variable. The diurnal and seasonal irradiance oscillations can be determined by the solar position, but a cloud induced variability may produce irradiance drops of 70% or more within a short period of 5 minutes [3]. If ramps of this magnitude can be predicted accurately, grid-tied PV generation will be more stable and reliable, potentially increasing its penetration into the electricity market. Recent regulations further enforce variation limits of 10% per minute on PV plants for smooth transitions to the grid in Puerto Rico and Canada [4, 5], and Australia requires minute forecasts to manage ancillary services [6]. Thus, accurate solar forecasting on the minute scale is essential to preserve power quality, reduce PV upkeep costs, and maintain reliability for smart grid integration.

The forecasting of Global Horizontal Irradiance (GHI), the main input for most solar power generation systems, has been addressed in the past by numerous methods depending on the forecast horizon, as well as the instrumentation available. Physics-based methods such as numerical weather prediction and satellite-based models are optimal for hourly and longer horizons [7]. For intra-hour forecasts, statistical methods which rely purely on historical GHI data have been well established, such as autoregressive (AR) and Artificial Neural Network (ANN) [8], but they only showed marginal improvement over the baseline Persistent Model (PM), which directly uses the present irradiance as the prediction. To better capture high frequency fluctuations of irradiance, ground sky imager-based prediction was recently developed. Local cloud movements are determined from consecutive images, and then pixel values with cloud cover information are correlated to GHI through supervised learning [9, 10]. However, forecasting accuracy suffers due to accu-

mulated errors from multiple processing steps, and the pixel cross correlation for cloud tracking is computationally intensive. Additionally, as sky imagers are costly and expensive to maintain, they are not commonly deployed together with solar panels.

A network of ground irradiance sensors with high sampling rate is an alternative to sky imagers for solar forecasting (see an example in Figure 1), where neighboring sensors can be sequentially influenced by the same passing clouds. These sensors are sparsely distributed across various ranges as they were originally designed for resource assessment purposes. Lately, researchers have started investigating the correlation of the GHI time series from along-wind and cross-wind sensor pairs in a high spatial density network [11]. In [12], by manually arranging sensors in a semicircle, the cloud motion vector is derived from the alignment and distance of the most correlated pair. It was then used for a solar forecast up to two minutes through a direct propagation of the current irradiance distribution across the plant, though with minimal gains [13].

Compared to sky imagers that visualize the cloud movement, sensor networks have a significant advantage as much of this information is inherently embedded within the spatio-temporal correlation among sensors. This strongly suggests to forecast GHI by modeling the variations of adjacent sensors, without having to explicitly determine wind field or cloud movement as an intermediate step. Therefore, it is desirable to predict the GHI of each sensor with a combination of both its previously observed values and a weighted sum of the time series from correlated neighboring sensors. It is essentially applying a previously popular univariate autoregressive (AR) model into a vector autoregressive (VAR) framework, with constant parameters representing a static spatio-temporal correlation within the sensor network.

In practice, however, the interdependency among sensors varies as both wind direction and speed can change from different days or even hours within one day, as shown in Figure 2. As the weather conditions are slowly evolving over time, it is feasible to divide the time series to smaller intervals where the weather locality holds. Case studies in [14] predicted sensor GHI using manually selected up-wind neighbors according to known wind information. Without exogenous input, the underlying correlation among sensors is better described by the most recent trends to accommodate for the changing environmental conditions, instead of global sensor correlation trends.

In this paper, we propose a local VAR (LVAR) framework to model the local sensor correlation in a highly dense network. Through learning using the very recent data, the dynamic spatial-temporal correlation can be approximated by a VAR model with constant parameters and thus underlying correlated sensor neighbors are discovered implicitly in an automated fashion. We further apply Ridge regularization on the Ordinary Least Square (OLS) solution to address the overfitting issue that occurs especially in the context of small local learning intervals or high model order, given the much smaller volume of training data. By employing LVAR, we provide an efficient solar irradiance forecasting framework that is robust to all cloud types and weather conditions.

The main contributions of this paper are summarized as follows: **1. Local Vector Autoregressive Ridge (LVAR<sub>R</sub>) Framework** We propose to utilize LVAR with Ridge regularization to model the spatio-temporal dynamics of a sensor network with only endogenous input for solar forecasting. Given the challenges of limited learning data, we introduce regularization for the learned LVAR

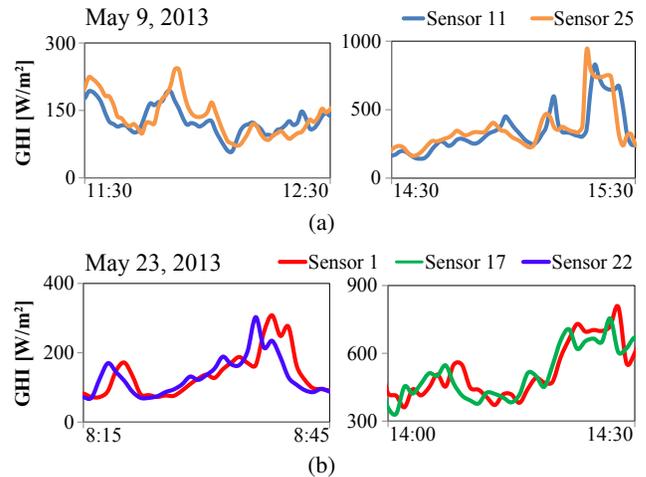


Figure 2: The locality of spatial-temporal correlations among sensors. (a) Around noon on May 9th, 2013, sensor 25 follows the trend of sensor 11, indicating an eastern wind. This pattern is reversed two hours later as the wind direction changes. (b) On May 23rd, 2013, sensor 1 follows sensor 22 in the morning, suggesting a southwestern wind, and sensor 17 in the afternoon as the wind shifts to the south.

models to ensure stable estimations and further optimize the performance. **2. Locality Study and Systematic Evaluation** We systematically study the locality and perform retrospective forecasts on a dataset consisting of 300,000 points in 2013 from a network of 25 stations. For 1-5 minute irradiance forecasts, we observe an average of 19.7% RMSE and 20.2% MAE improvement over the baseline Persistent Model, which is reported to be difficult to surpass [7, 15].

## 2. BACKGROUND

Time series forecasting methods, which relate historical GHI data to future values, have been classically employed. In this section, we start with the introduction of the most commonly used reference model in the field, and extend to more generalized time series models.

### 2.1 Persistent Model

The simplest and yet still challenging baseline in short-term irradiance prediction is the Persistent Model (*PM*). This model assumes that the irradiance at time  $t$  on the sensor  $k$ ,  $y_{k,t}$ , is best predicted with its value previously observed at time  $t - 1$ ,  $y_{k,t-1}$ :

$$y_{k,t} = A_{k,1}y_{k,t-1} + e_{k,t}, t = 1, \dots, T \quad (1)$$

where  $A_{k,1}$  is always 1 and the error term,  $e_{k,t}$ , is assumed to be zero mean. Persistent Model performs very well for clear and overcast conditions, where the irradiance variation over time is relatively small. Large errors usually occur when passing opaque clouds bring abrupt changes to the measured irradiance. As shown in the GHI analysis of [10], GHI has a very high autocorrelation in the order of minutes, and this is seen even on very cloudy days. Recent studies reported that this baseline is very difficult to beat for forecasts within 15 minutes [7, 13, 14, 15, 16]. Although it is very effective, it does not take advantage of periodic cloud patterns because it simply follows the immediate past.

### 2.2 Autoregressive (AR) Model

One of the main methods to analyze time series data is by using a

parametric approach. This assumes that there is a certain structure in the underlying stochastic process, which can be described by a small number of parameters. The Autoregressive (AR) model is such a representation that models the output variable to be linearly dependent on its own previous values and a stochastic term. In fact, the aforementioned PM can be considered as a special case of AR(1) where the coefficient is always 1. We can extend AR(1) to the more general form of AR(p) and incorporate potential  $p$  different time delaying factors caused by periodic irradiance patterns. The irradiance at time  $t$  on the sensor  $k$  is then modeled by:

$$y_{k,t} = \sum_{i=1}^p A_{k,i} y_{k,t-i} + e_{k,t}, t = 1, \dots, T \quad (2)$$

where  $p$  is called "model order", and  $e_{k,t}$  is white noise. The task then is to estimate the parameters  $A_{k,i}$  that describes the stochastic process of irradiance.

The AR model is widely used in solar and wind prediction. As investigated in the latest study of machine learning techniques for solar forecasting [16], when given only endogenous input of historical GHI, non-linear methods such as Support Vector Machine (SVR) and Neural Network (NN) do not outperform their simple counterparts (AR) for forecasting horizons less than one hour. Although AR(p) is ideal for capturing periodic irradiance patterns, it does not take into account the main cause of irradiance variability - cloud interference.

### 2.3 Vector Autoregressive (VAR) Model

While a single sensor provides measurements of temporal irradiance behavior, a network of sensors also presents spatial variability of irradiance due to cloud movement, over a distributed geographical location. In Figure 2, GHI time series from sensors aligned with the direction of cloud movement are highly correlated. In particular, the time series of the up-wind sensor is lagged by the cloud travel time.

Therefore, to effectively utilize the strengths of a sensor network, we generalize a univariate (AR) forecasting model into a vector autoregressive (VAR) framework to incorporate all available sensors in the network. Each sensor is thus predicted as a linear combination of both its own time series, and a weighted sum of the time series from other sensors. VAR framework can systematically model the underlying causality among correlated sensors, i.e. by giving a higher weight to the lagged value of the most relevant neighbors, without explicit interpretation of wind field or cloud movement. A VAR model with model order  $p$  (VAR(p)) has the form:

$$y_t = \sum_{i=1}^p A_i y_{t-i} + e_t, t = 1, \dots, T \quad (3)$$

where  $n$  is the number of sensors,  $y_t$  is a  $n \times 1$  irradiance vector,  $A_i$  is an  $n \times n$  matrix representing the coefficients of model order  $i$ , and  $e_t$  is an  $n \times 1$  error vector.

Expanding all the equations, we can get a matrix form:

$$Y = XB + E \quad (4)$$

$$Y_{(m \times n)} = \begin{bmatrix} y'_{p+1} \\ y'_{p+2} \\ \vdots \\ y'_T \end{bmatrix}, X_{(m \times np)} = \begin{bmatrix} y'_p & y'_{p-1} & \cdots & y'_1 \\ y'_{p+1} & y'_p & \cdots & y'_2 \\ \vdots & \vdots & \ddots & \vdots \\ y'_{T-1} & y'_{T-2} & \cdots & y'_{T-p} \end{bmatrix},$$

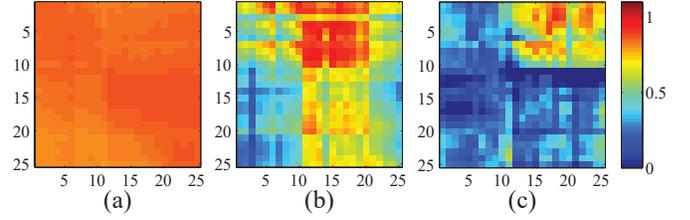


Figure 3: 2 minutes lagged cross correlation between 25 sensors on monthly data (a) and hourly data (b) and (c). For monthly data, overcast and clear periods produce time-series of relatively low variability. This results in an overall high correlation, and still retains the higher correlation along the diagonal, representing neighboring sensors. Hourly cross correlations on May 7th show two different patterns between the morning (b) and the afternoon (c), indicating the change of wind speed and direction.

$$B_{(np \times n)} = \begin{bmatrix} A'_1 \\ A'_2 \\ \vdots \\ A'_p \end{bmatrix}, E_{(m \times n)} = \begin{bmatrix} e'_{p+1} \\ e'_{p+2} \\ \vdots \\ e'_T \end{bmatrix}.$$

The matrix  $B$  can be estimated by OLS (Ordinary Least Square) as follows:

$$\hat{B} = (X'X)^{-1}X'Y \quad (5)$$

when  $X'X$  is invertible ( $m \geq np$  and non-singular). The OLS estimates are optimal with respect to being the best unbiased estimator of  $B$ , at the cost of high variance.

Since we incorporated a vector of sensors into the model with  $p$  model order, the learned model has a high risk of overfitting. To overcome the overfitting issues of VAR(p), we regularize the coefficient matrix  $A$  using Ridge regularization.

$$\operatorname{argmin}_B \|Y - BX\|_F^2 + \lambda \|B\|_F^2 \quad (6)$$

where  $\lambda$  controls the degree of regularization and  $\|\cdot\|_F$  is the Frobenius norm. We call it VAR<sub>R</sub>(p) and the least square solution can be solved as follows:

$$\hat{B} = (X'X + \lambda I)^{-1}X'Y \quad (7)$$

where  $I$  is a diagonal identity matrix.

### 3. LOCAL VECTOR AUTOREGRESSIVE (LVAR) MODEL

Although VAR<sub>R</sub>(p) introduced in Section 2 provides a flexible framework that systematically captures relationships among multiple time series from sensors, there is an inherent assumption that the spatio-temporal correlations are static and constant. However, the heteroscedastic nature of wind field and cloud movement violates this basic assumption [17]. In Figure 3(a), we can see similar correlation patterns across all sensors in the network over a month, as well as slightly stronger correlations within closely located sensors. Yet the spatio-temporal correlations show distinctive patterns when examined on an hourly scale, such as Figure 3(b) and (c), which display a correlation shift from morning to afternoon. This strongly indicates that modeling local cloud patterns instead of a global weather trend is required.

We propose a local vector autoregressive (LVAR) model for short-term GHI prediction to estimate the joint dynamics of multivariate time series from a sensor network. Instead of learning a global

model, *LVAR* allows time-varying parameters to adapt to local climate changes. To balance between time dependent parameters at each time point and one set of constant parameters like *VAR*, we apply a local homogeneity assumption that there exists an optimal time interval in the immediate past over which the current local spatio-temporal correlation can be approximated by a *VAR* model with constant parameters. This assumption can reduce computational complexity while maintaining model flexibility. It holds since most of the parameters affecting the sensor readings are slowly evolving over time, such as wind direction and speed, and large scale changes such as cloud formation and deformation can be excluded in the small scale of a dense sensor network.

Given the local homogeneity assumption, our learning objective function is adapting to local training, thus we have a different  $B$  for each training time  $t$  with the local training length  $L$ ,  $B^t$ :

$$\operatorname{argmin}_{B^{(t)}} \|Y^{(t)} - X^{(t)}B^{(t)}\|_F^2 \quad (8)$$

where

$$Y^{(t)}_{((L-p) \times n)} = \begin{bmatrix} y'_{t-L+p+1} \\ \vdots \\ y'_{t-1} \\ y'_t \end{bmatrix},$$

$$X^{(t)}_{((L-p) \times np)} = \begin{bmatrix} y'_{t-L+p} & y'_{t-L+p-1} & \cdots & y'_{t-L} \\ \vdots & \vdots & \ddots & \vdots \\ y'_{t-2} & y'_{t-3} & \cdots & y'_{t-p-1} \\ y'_{t-1} & y'_{t-2} & \cdots & y'_{t-p} \end{bmatrix},$$

$$B^{(t)}_{(np \times n)} = \begin{bmatrix} A'_p \\ \vdots \\ A'_2 \\ A'_1 \end{bmatrix}, \quad E^{(t)}_{((L-p) \times n)} = \begin{bmatrix} e'_{t-L+p+1} \\ \vdots \\ e'_{t-1} \\ e'_t \end{bmatrix}.$$

*LVAR* allows flexible and automated choice of lagged variables with the highest correlation to predict sensor irradiance. Local climate is described by the updating parameters learned from recent data. By looking at a relatively small time interval, correlations among neighboring sensors are more precisely modeled. When wind direction is unfavorable and correlation is not observed for the certain sensor, the model would switch to an autoregressive one by giving higher weight to its own lagged variables. *LVAR* is in a way similar to a stochastic model for a time series that generally reflects the fact that observations close together in time will be more closely related than observations further apart.

Compared to *VAR*, *LVAR* requires a more significant regularization because the amount of training data is limited to only  $L$  subsamples. Overfitting has a higher chance to occur as the number of parameters in the model is excessive relative to the number of observations. In other words, the OLS solution requirement,  $m \gg np$ , is much more difficult to satisfy in this situation. Again, to overcome overfitting, we add the Ridge regularization term:

$$\operatorname{argmin}_{B^{(t)}} \|Y^{(t)} - X^{(t)}B^{(t)}\|_F^2 + \lambda \|B^{(t)}\|_F^2. \quad (9)$$

We call the above model *LVAR<sub>R</sub>*( $p$ ). The  $\lambda$  controls the strength of penalty for introducing more variables. With the advantage of stability, Ridge regression is a great alternative to subsets selection

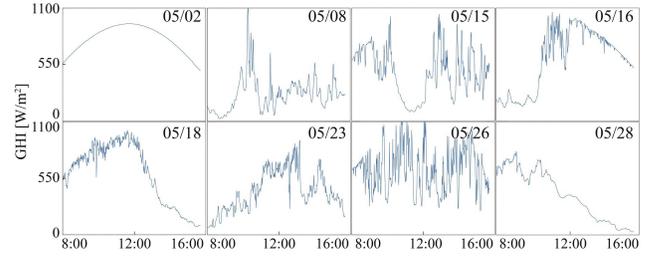


Figure 4: Representative daily GHI of the 25-day dataset.

of regressor variable for parameter shrinkage. Introducing bias to the estimate substantially reduces the high variance due to the collinearity of interdependent variables, thus ensures better predictive performance on the unseen data.

*LVAR* requires more training times than *VAR* if we use gradient based optimization methods, but since the size of training data at each time  $t$  is much smaller, it has great potential for being deployed in real-time solar forecasting.

## 4. EXPERIMENTS

In this section, we present the description of the experimental setup, results, and analysis. Models introduced in Section 2 and Section 3 were systematically evaluated. We present substantial evidence that *LVAR<sub>R</sub>* outperforms all other models including the baseline *PM*. Their respective performances are analyzed and the results are given below in terms of the three different statistical measures.

### Data and Experimental Setup

Figure 1 shows the layout of the network of 25 irradiance monitoring sensors in Long Island Solar Farm (LISF), New York, United States. Our dataset consists of 25 days of GHI time series from May 1st to May 31st, 2013, where 6 days were omitted from the month due to sensor operational downtime. GHI data was collected every 1 second from 25 sensors in the network and sampled at a 1 minute resolution. Due to the low elevation angle of the sun in the early morning and late afternoon in the northern hemisphere, we chose data from 8:00 to 16:00 on each day, when PV plants generate the most power from solar irradiance. Collectively, we have 12,000 data points per day, with a total of 300,000 points in the dataset. The GHI value was normalized using the Clear Sky Index [9] and converted back for error rate calculation. This dataset encompasses a diverse collection of weather type and cloud condition and is not biased towards any specific scenario. Representative daily GHI are shown in Figure 4, where we capture conditions such as clear sky, broken clouds, and scattered clouds.

For *VAR* and *VAR<sub>R</sub>* model parameter selection, we applied the  $K$ -fold Cross Validation approach, where dataset was partitioned into  $K$  separate sets of equal size. Of the  $K$  subsets, one single subset was retained as the validation data to test the model, while the remaining  $K - 1$  subsets were used as training data. In this study, we used  $K = 5$ , with each fold containing 5 complete days for the optimization of order  $p$  and regularization term  $\lambda$ .

In *LVAR* and *LVAR<sub>R</sub>*, since training data is local and online, the cross validation was not applicable. Thus model parameter optimization was achieved by a systematic grid search on order  $p$  (1-9), local training length  $L$  (40-400), and regularization term  $\lambda$  ( $10^{-2}$ - $10^4$ ) through the whole dataset. For completeness the first  $L$  data points of each day are predicted with *PM*.

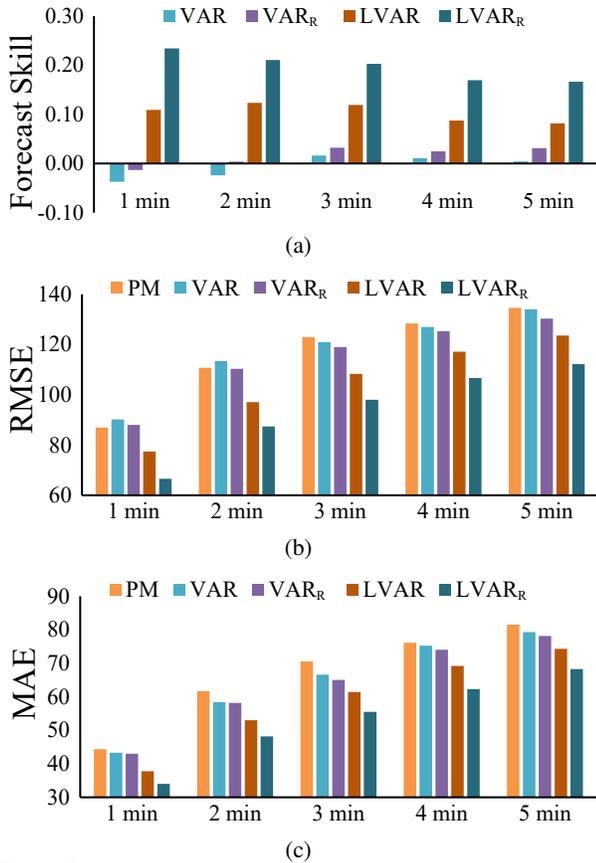


Figure 5: Comparisons of models on the full 25-day dataset using three different evaluation metrics: (a) Forecast Skill, (b) RMSE, and (c) MAE. Across all metrics and forecast horizons,  $LVAR_R$  shows the greatest improvement over baseline  $PM$ .

	1 min	2 min	3 min	4 min	5 min
Full Set FS	0.234	0.211	0.203	0.169	0.166
Subset FS	0.259	0.226	0.225	0.184	0.189

Table 1:  $LVAR_R$  forecast skill for 1-5 minutes prediction on the full dataset of 25 days and the subset of selected 5 days.

In addition to the complete 25-day dataset, we chose a subset of five days, where sensor correlation shifts were apparent by visual inspection of the raw time series data. This subset includes intra-day changes in wind speed and direction, as shown in Figure 2, and also multi-layered clouds where each layer has different wind field (i.e. May 8th) as presented in Figure 4.

### Evaluation Metrics

We used three evaluation metrics in this study: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Forecast Skill (FS). By squaring the error, RMSE gives more weight to larger errors, skewing the error estimate towards the outliers. RMSE is more useful because large errors, which lead to disproportionately high losses, are particularly undesirable in solar forecasting. Used together with MAE, variation in the errors can be diagnosed in a set of forecasts. Forecast Skill indicates the forecast improvement of a certain model over the reference  $PM$  introduced in Section 2.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (GHI_{\text{forecast},i} - GHI_{\text{measured},i})^2} \quad (10)$$

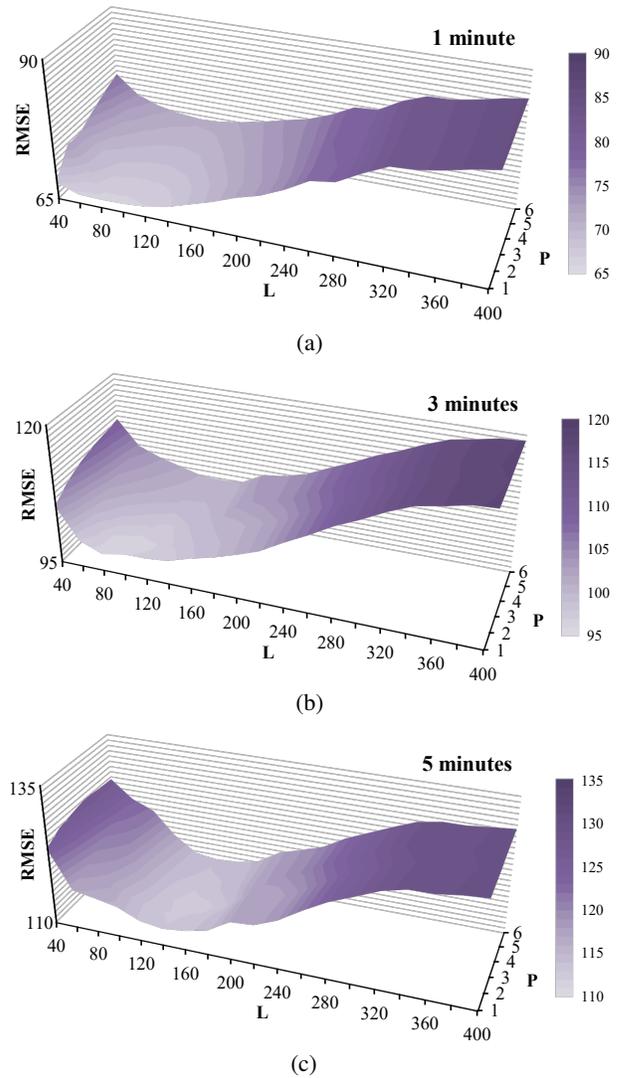


Figure 6:  $LVAR_R$  stability of 1, 3, 5 minute forecast.

$$MAE = \frac{1}{N} \sum_{i=1}^N |GHI_{\text{forecast},i} - GHI_{\text{measured},i}| \quad (11)$$

where  $N$  is the number of total points in the dataset.

$$\text{Forecast Skill} = \frac{RMSE_{PM} - RMSE}{RMSE_{PM}} \quad (12)$$

### Result and Analysis

Considering that  $PM$  in short-term irradiance prediction (1-5 minutes) is very difficult to beat, we can verify that we have achieved significant improvements using a sensor network as shown in Figure 5. In the literature, short-term forecasts that incorporate cloud motion estimated through sky images have either reported the limited application of their methods on clear sky condition, thin layer cloud, and multi-directional wind fields [7, 15], or based their validation on a manually curated cloudy dataset that is disadvantageous to  $PM$  [9].  $VAR$  models are more applicable under known constant wind conditions and showed promising performance [14]. Because  $VAR$  and  $VAR_R$  do not adapt to the local context of cloud

	1 min	2 min	3 min	4 min	5 min
P	1	1	2	2	3
L	80	100	80	140	140
$\log \lambda$	1.3	1.7	2.1	2.3	2.5

Table 2: Optimal model order, local training length, and regularization term through grid search for the full 25-day dataset

conditions, our proposed  $LVAR$  and  $LVAR_R$  showed a prediction performance boost over  $PM$ ,  $VAR$ , and  $VAR_R$  in terms of RMSE, MAE, and FS on the diverse dataset.

$VAR$  models underperformed for 1 and 2 minute predictions and were comparably worse than  $PM$  in terms of RMSE and FS due to the strong impact of autocorrelation. In particular, RMSE penalizes large errors more compared to MAE, and we could observe these effects in the collection of all three error metrics. The employment of regularization improved the performance of  $LVAR$  by a significant 10.34%, and  $VAR$  by a marginal 2.17% as the training data is sufficient to generalize the model well. As discussed in Section 3, the regularization prevented the trained model from being overfitted, which is especially critical for  $LVAR$ .

To further validate the capability of  $LVAR_R$  to adapt to changing cloud conditions, we compared the forecast skill between the full dataset and the subset in Table 1. The subset includes higher wind variability within a day, which stresses the need for a locality modeling. This is reflected in the 9.7% further improvement over the full dataset for one minute predictions. This improvement gradually decreased as we increased the prediction horizon.

In terms of model parameter sensitivity, we tested the stabilities of both model order  $p$ , and the amount of local training length  $L$  with appropriate regularization. For one minute predictions, Figure 6 shows the strong preference for a small model order  $p$  and local training size  $L$ , but the area around the global minimum is broad and has a smooth gradient. In addition, the optimal  $p$  and  $L$  values are getting larger as we increase the prediction horizon from one minute to five minutes, which is expected. Table 2 shows the optimal  $p$ ,  $L$ , and  $\lambda$  values for 1-5 minute prediction from the grid search which reaffirms this trend. As we had more parameters to tune by increasing  $p$ , the optimal  $\lambda$  was also increased to control the model complexity and avoid overfitting. For larger  $L$ , we also need to increase  $\lambda$  to get the same level of regularization, i.e. between one and two minutes and between three and four minutes.

## 5. CONCLUSION

In this paper, we proposed a short-term (1-5 minute) solar irradiance forecasting framework, which models the spatio-temporal variation of GHI by utilizing a network of sensors. To accommodate local climate condition changes, we propose a local vector autoregressive ( $LVAR$ ) model, which only uses the endogenous input of historical GHI. Since we have limited amount of data for local training, the regularization is introduced to prevent overfitting. The combination of  $LVAR$  and Ridge regularization achieved an impressive 19.7% RMSE and 20.2% MAE improvement compared to the baseline Persistent Model ( $PM$ ), as  $PM$  is known to be very difficult to outperform due to strong autoregressive effects on short-term irradiance forecasting.

## 6. REFERENCES

- [1] WWS. 100% Clean and renewable Wind, Water, and Sunlight (WWS) all sector energy roadmaps for 139

- countries of the world. <http://web.stanford.edu/group/efmh/jacobson/Articles/I/WWS-50-USState-plans.html>, 2015.
- [2] SEIA. Solar Energy Industries Association. Solar industry data. <http://www.seia.org/research-resources/solar-industry-data>.
- [3] NERC. North American Electric Reliability Corporation. Special Report: Accommodating high levels of variable generation. [http://www.nerc.com/files/ivgtf\\_report\\_041609.pdf](http://www.nerc.com/files/ivgtf_report_041609.pdf), 2009.
- [4] V. Gevorgian and S. Booth. Review of prepa technical requirements for interconnecting wind and solar generation. 11 2013.
- [5] NERC. IVGTF Task 2.2 Report: Reliability considerations for BA communications with increased variable generation. [http://www.nerc.com/comm/PC/Integration of Variable Generation Task Force I/IVGTF 2 2 Report\\_NA.pdf](http://www.nerc.com/comm/PC/Integration%20of%20Variable%20Generation%20Task%20Force%20I/IVGTF%202%20Report_NA.pdf), 2012.
- [6] B. Elliston and I. MacGill. The potential role of forecasting for integrating solar generation into the Australian national electricity market. In *Solar 2010: proceedings of the annual conference of the Australian solar energy society*, 2010.
- [7] M. Diagne, M. David, P. Lauret, J. Boland, and N. Schmutz. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews*, 27(0):65 – 76, 2013.
- [8] G. Reikard. Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Solar Energy*, 83(3):342–349, 2009.
- [9] J. Xu, S. Yoo, D. Yu, D. Huang, J. Heiser, and P. Kalb. Solar irradiance forecasting using multi-layer cloud tracking and numerical weather prediction. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing, SAC '15*, pages 2225–2230, New York, NY, USA, 2015. ACM.
- [10] J. Xu, S. Yoo, D. Yu, H. Huang, D. Huang, J. Heiser, and P. Kalb. A stochastic framework for solar irradiance forecasting using condition random field. In *Advances in Knowledge Discovery and Data Mining*, pages 511–524. Springer, 2015.
- [11] L. M. Hinkelman. Differences between along-wind and cross-wind solar irradiance variability on small spatial scales. *Solar Energy*, 88:192–203, 2013.
- [12] J. L. Bosch, Y. Zheng, and J. Kleissl. Deriving cloud velocity from an array of solar radiation measurements. *Solar Energy*, 87:196–203, 2013.
- [13] M. Lipperheide, J. Bosch, and J. Kleissl. Embedded nowcasting method using cloud speed persistence for a photovoltaic power plant. *Solar Energy*, 112:232–238, 2015.
- [14] D. Yang, Z. Ye, L. H. I. Lim, and Z. Dong. Very short term irradiance forecasting using the lasso. *Solar Energy*, 114:314–326, 2015.
- [15] S. Quesada-Ruiz, Y. Chu, J. Tovar-Pescador, H. Pedro, and C. Coimbra. Cloud-tracking methodology for intra-hour DNI forecasting. *Solar Energy*, 102(0):267 – 275, 2014.
- [16] P. Lauret, C. Voyant, T. Soubdhan, M. David, and P. Poggi. A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. *Solar Energy*, 112:446–457, 2015.
- [17] A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh. Current methods and advances in forecasting of wind power generation. *Renewable Energy*, 37(1):1–8, 2012.