Missing-data Imputation for Solar Irradiance Forecasting in Thailand

Presenter : Vichaya Layanun
Co-author: Jitkomut Songsiri
Supachai Suksamosorn

Department of Electrical Engineering,
Faculty of Engineering, Chulalongkorn University,
Bangkok, Thailand

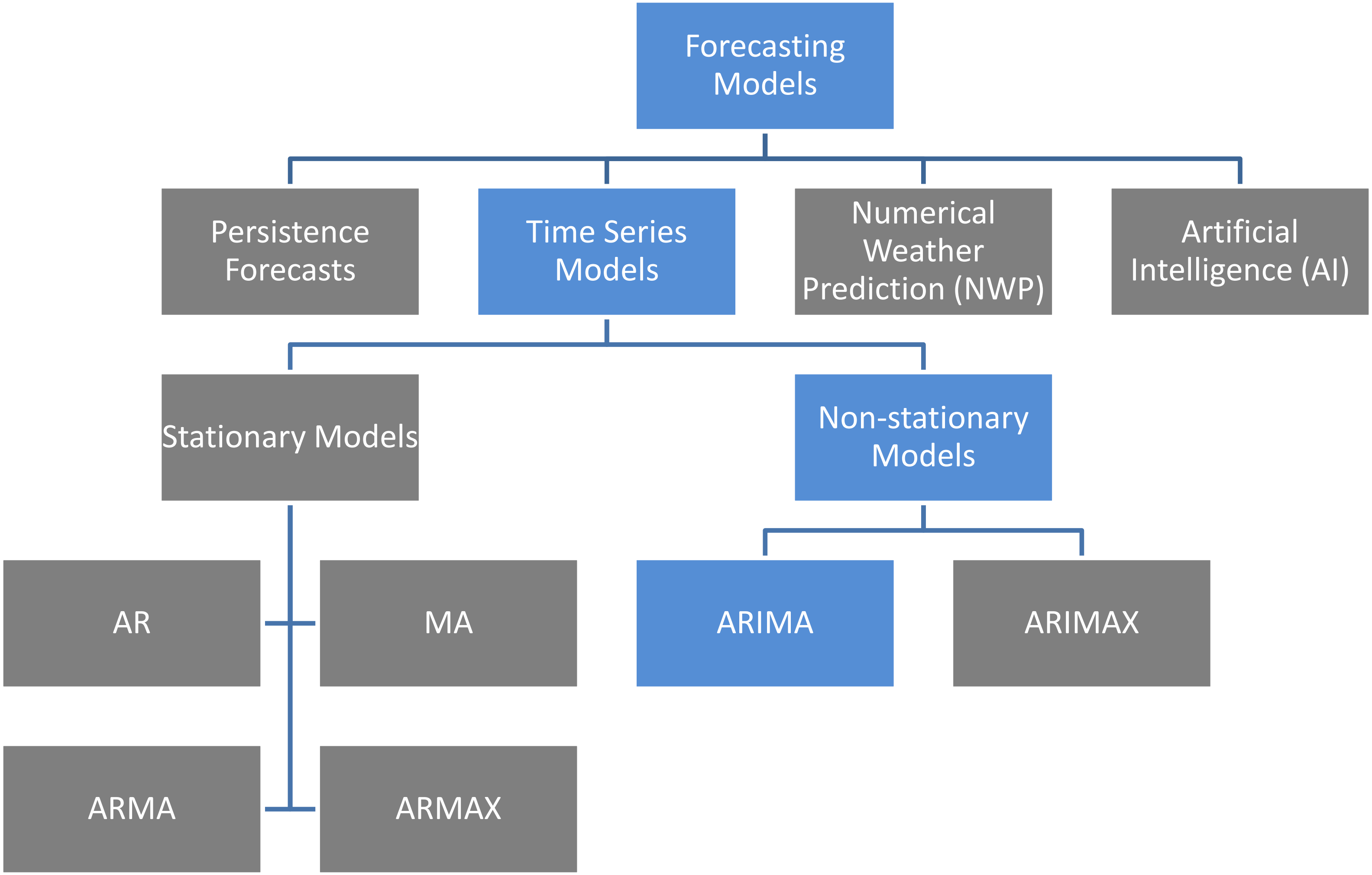
CHULA ΣNGINEERING
Foundation toward Innovation

Why solar irradiance forecasting is important?

1. **A disadvantage of the solar power** is its randomly intermittent availability and therefore **has made a power generation difficult for a power management.**

2. **PV power data are not typically available in Thailand** making it hard to predict PV power directly from historical data.

3. **Solar irradiance is highly influential to solar power**, so irradiance forecasting is a common approach.



There are many forecasting models to predict the future solar irradiance. We focus on time series models.

Objective

- **To fit** seasonal **time series models to global horizontal irradiance (GHI)** time series.
- **To impute missing data** of GHI in Thailand.
- **To propose a practical method of missing-data imputation** suitable for the condition of acquired data in Thailand.

Time Series Models

- Seasonal ARIMA models

Relevant Variables

Global Horizontal Irradiance (GHI) is the considered variable which is the geometric sum of Direct Normal Irradiance (DNI) and Diffuse Horizontal Irradiance (DHI).

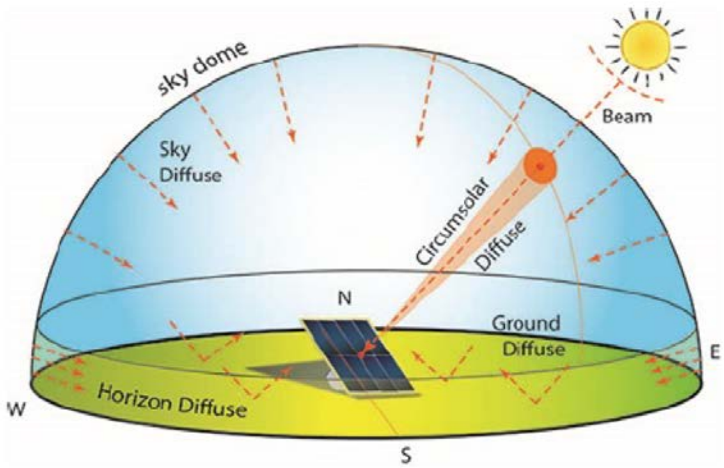


Figure 4: Solar irradiance component

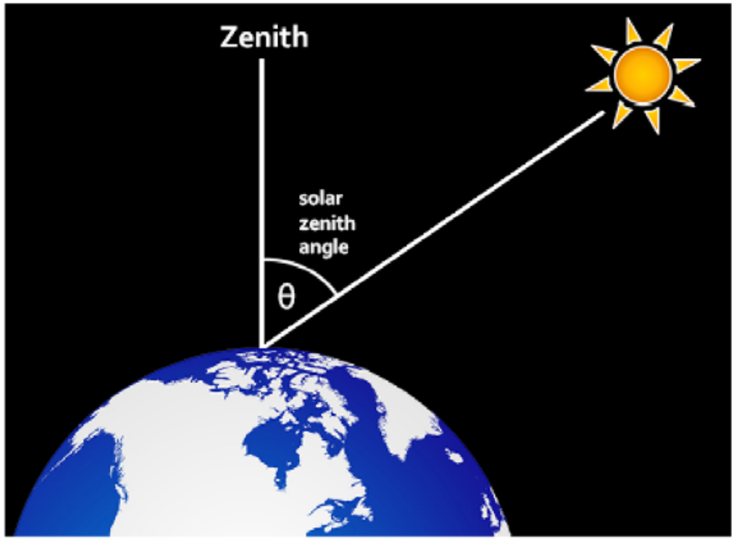
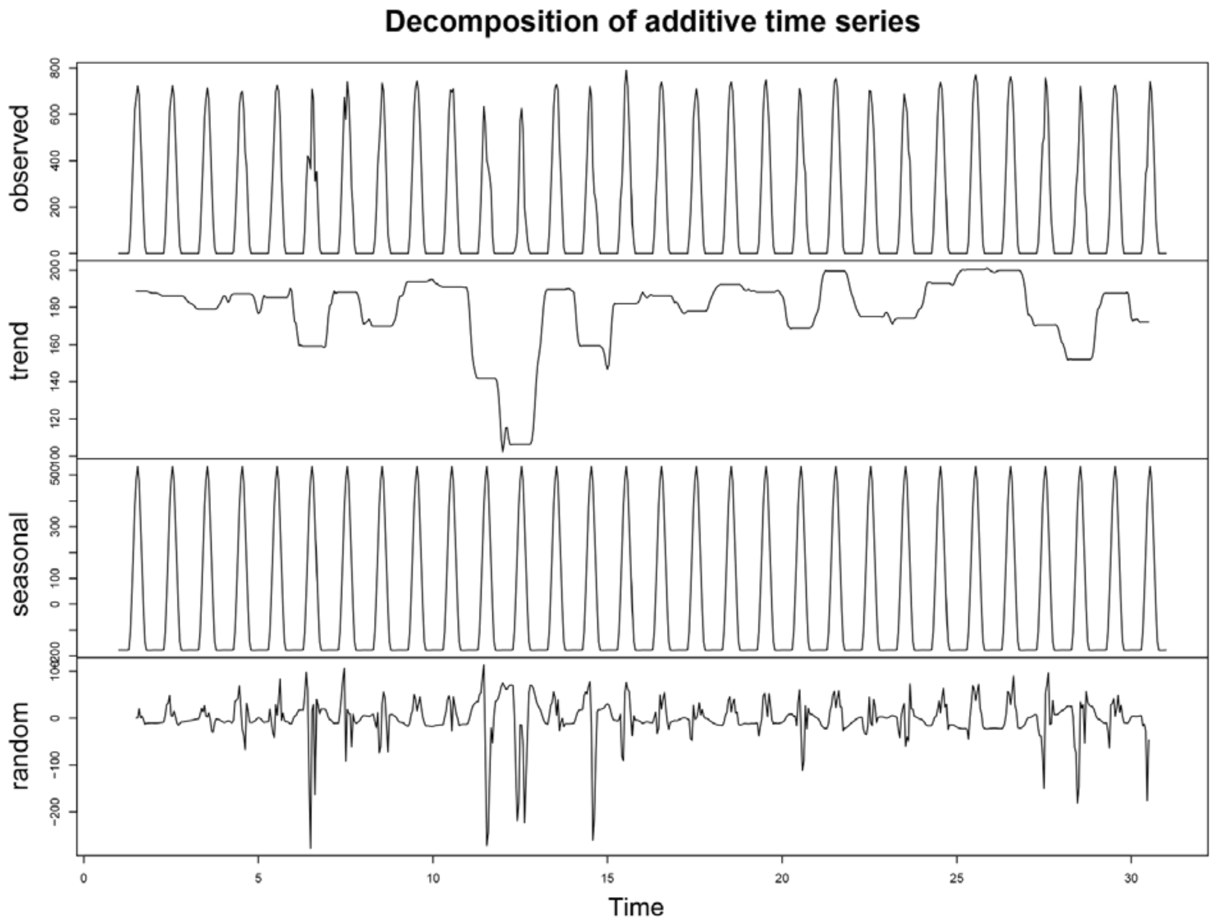


Figure 5: Solar zenith angle

This study predicted GHI.

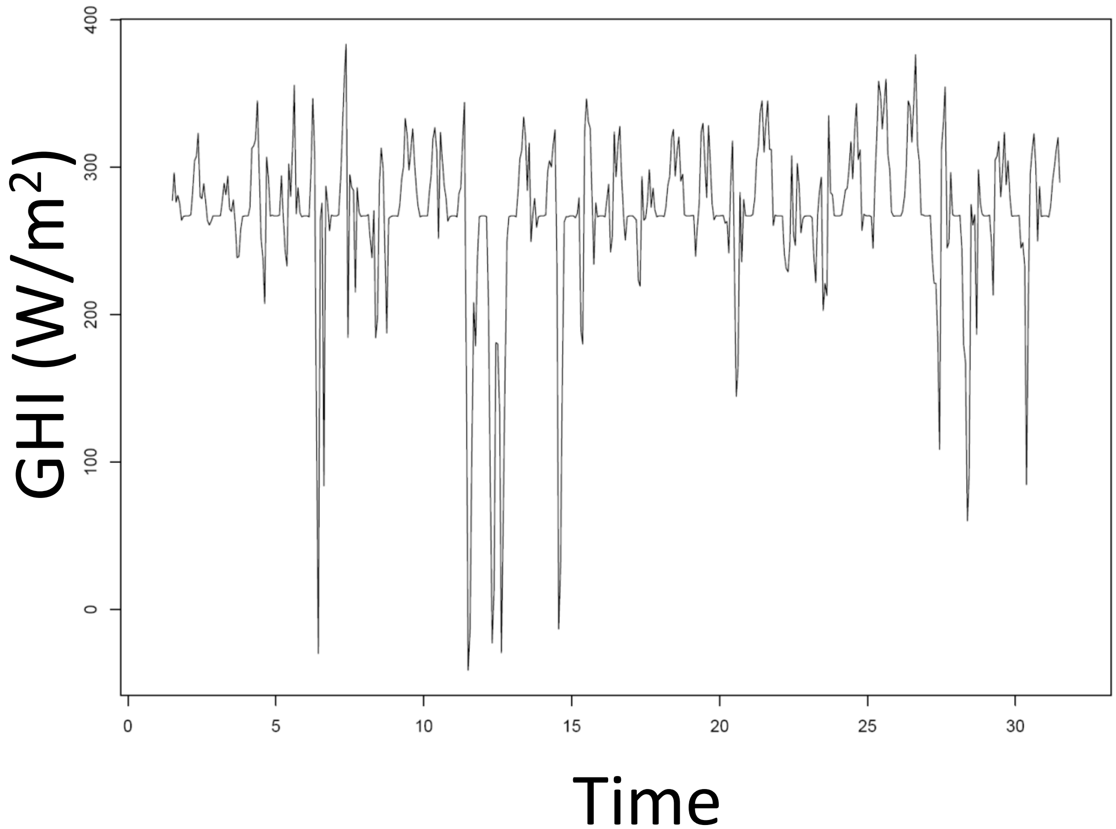
GHI data has a seasonal trend



Observed signal = trend + seasonal trend + random signal

Decomposition of Solar Irradiance in January 2014 in Bangkok.

Seasonal ARIMA model



GHI after removing seasonal trend in January 2014 in Bangkok.

We imply that GHI can be described as ARMA models containing s season:

$$A(L)y(t) = \mathrm{s(t)} + \alpha + C(L)v(t)$$

Seasonal ARIMA models

Seasonal ARIMA model

We used a seasonal ARIMA models to include the seasonal effect. The SARIMA $(p, d, q)(P, D, Q)_T$ can be defined as

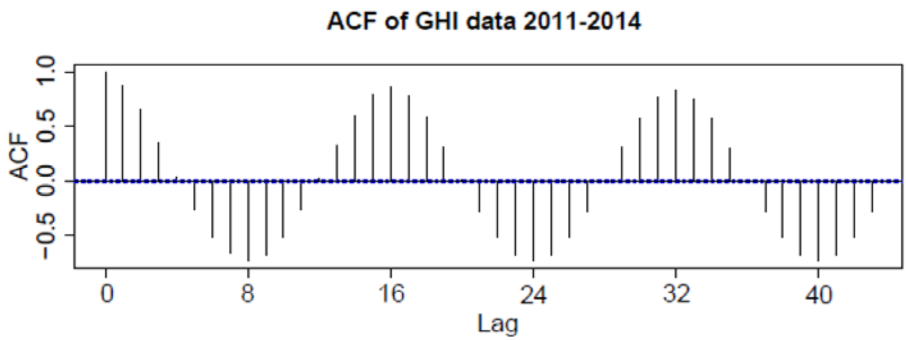
$$\boxed{\tilde{A}\left(L^T\right)A(L)(1-L^T)^D(1-L)^d y(t) = \tilde{C}\left(L^T\right)C(L)v(t)}$$

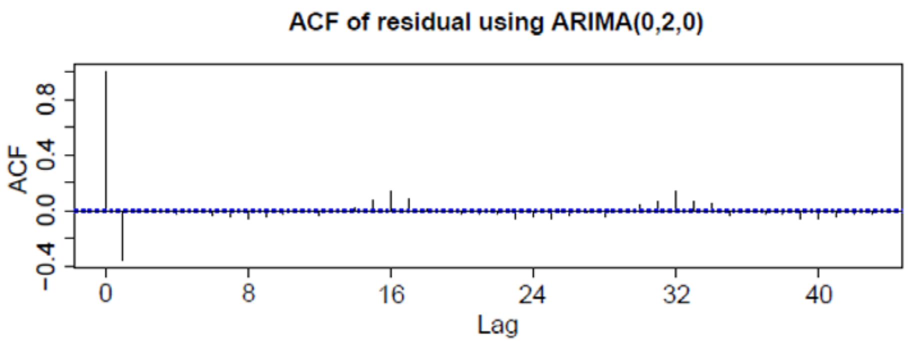
where the autoregressive and moving average polynomials are

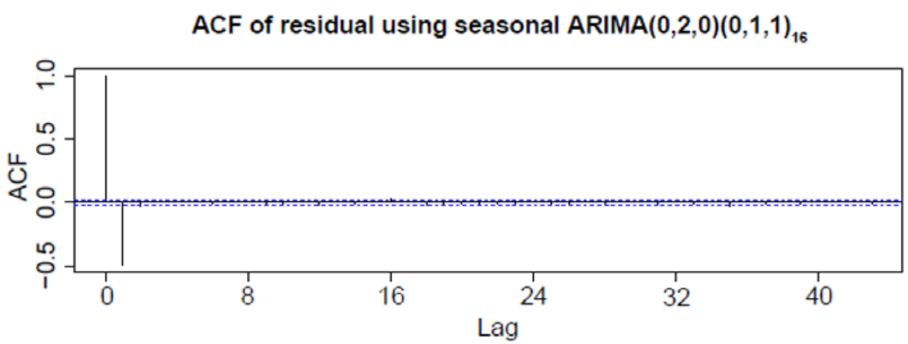
$$A(L) = I - (a_1 L + \cdots + a_p L_p)$$
$$C(L) = I + c_1 L + \cdots + c_q L_q$$
$$\tilde{A}\left(L^T\right) = I - (\tilde{a}_1 L^T + \tilde{a}_2 L^{2T} + \cdots + \tilde{a}_p L^T)$$
$$\tilde{C}\left(L^T\right) = I + \tilde{c}_1 L^T + \tilde{c}_2 L^{2T} + \cdots + \tilde{c}_q L^T$$

T is a seasonal period (setting T=16), D is integrated seasonal order, and L is a lag operator: $Ly(t) = y(t-1)$ where the time index is in an hourly basis.

Seasonal ARIMA models







$$\boxed{\textbf{SARIMA}\ (p, d, q)(P, D, Q)_T}$$

d=2 : ARIMA(0,2,0)

1. a dramatic reduction of ACF
2. some lags of ACF are not negligibly small and lie outside the confidential interval.

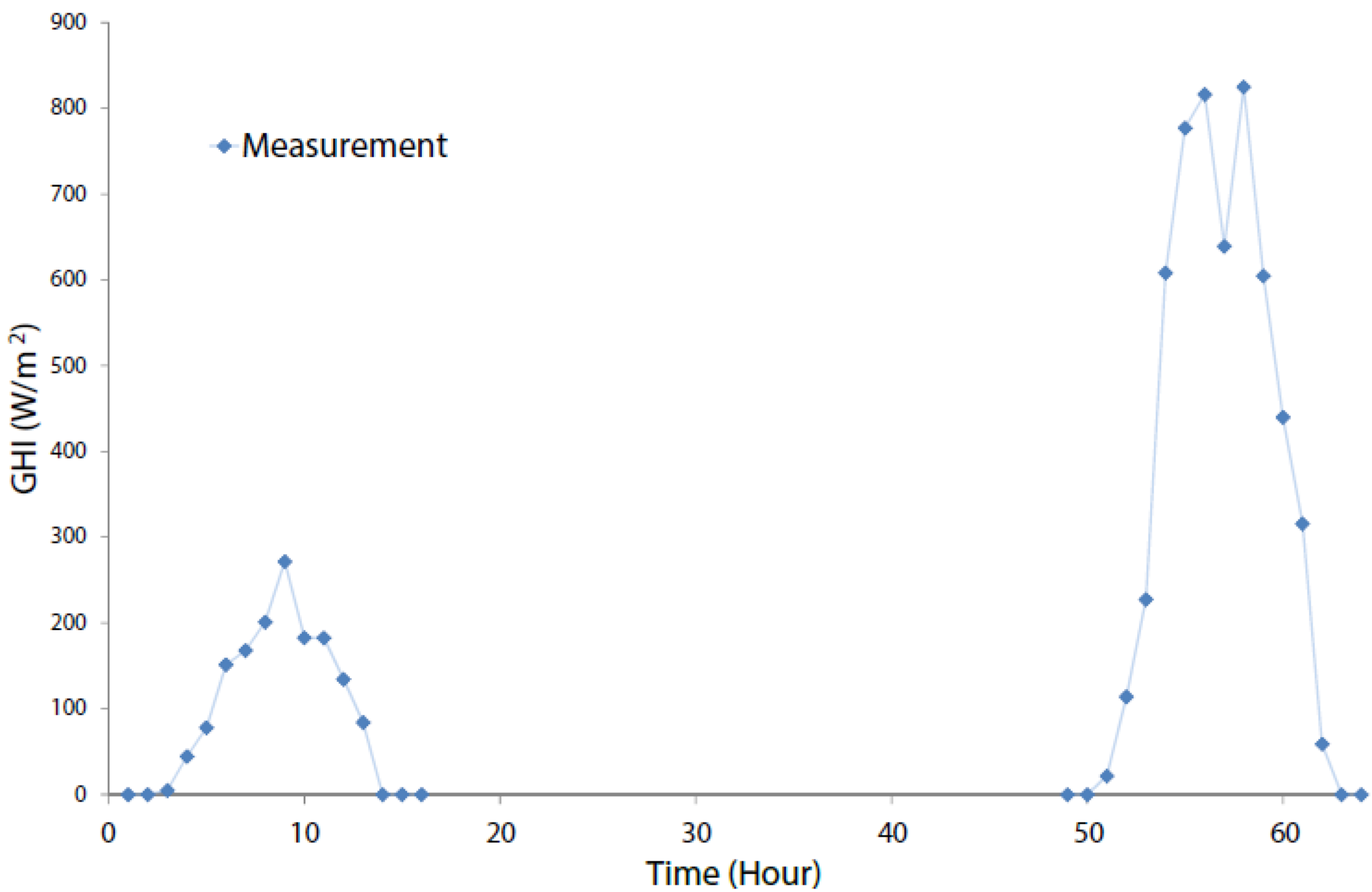
d=2 & $(P,D,Q)_T = (0,1,1)_{16}$

1. this yields only a few lags of ACF that lie outside the confidential interval

Missing data

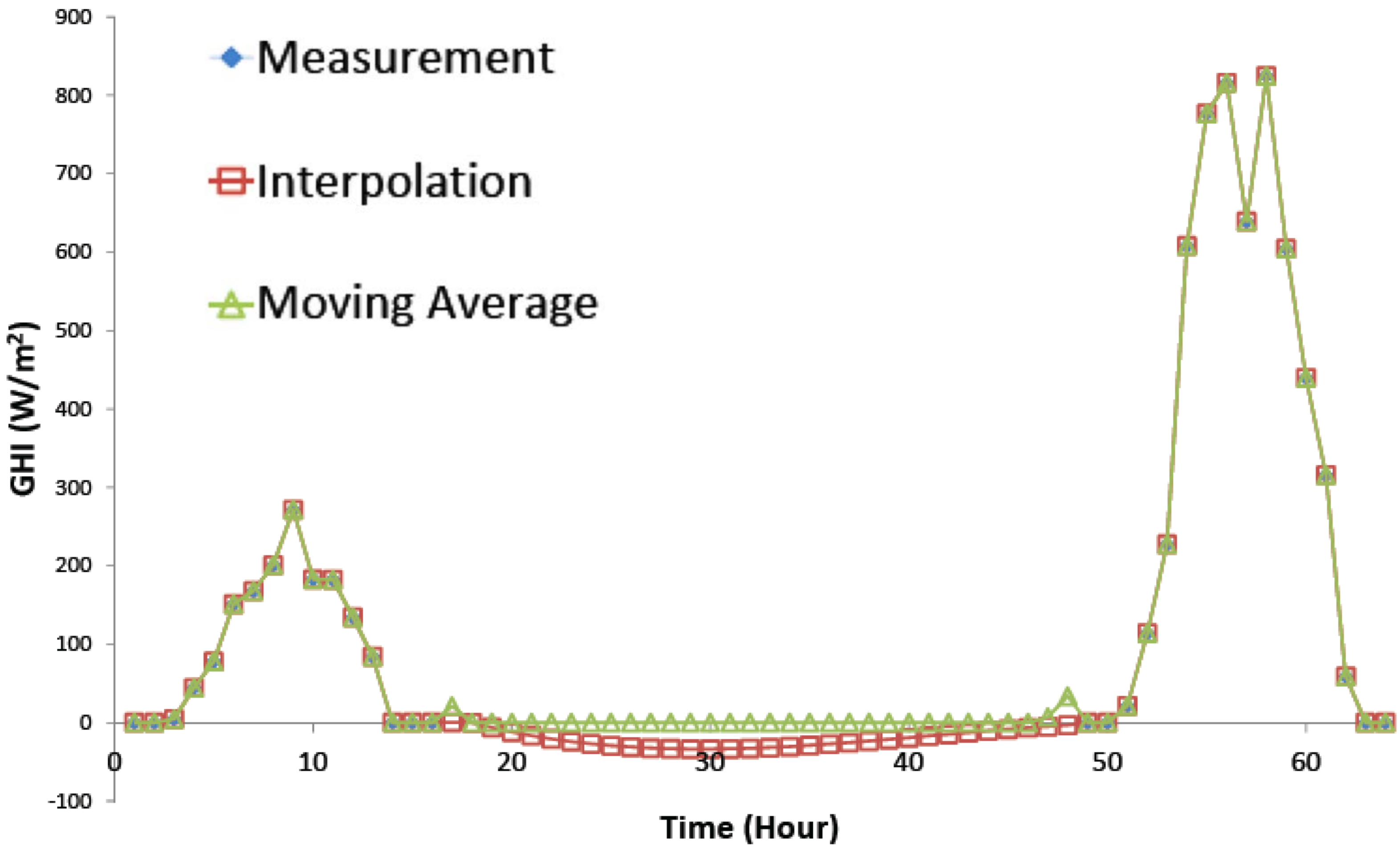
GHI data are acquired from Thailand Meteorological Department during 2011-2015

Missing data



Data missing usually holds for **several consecutive days**.

Missing-data imputation using typical methods



Typical methods such as moving average (MA) and linear interpolation that exploits the variable dynamic **cannot perform well** in this case as the imputed value is a linear combination of nearby available values.

Concept of the proposed method for missing-data

$I_{14}(t)$

Date/Time	08-09	09-10	10-11	11-12	12-13	13-14	14-15
7-Apr	284.8	596.0	771.0	801.5	850.9	807.1	673.7
8-Apr	338.0	589.3	761.6	806.3	834.4	807.4	699.5
9-Apr	NA	NA	NA	NA	NA	NA	NA
10-Apr	NA	NA	NA	NA	NA	NA	NA
11-Apr	NA	NA	NA	NA	NA	NA	NA
12-Apr	NA	NA	NA	NA	NA	NA	NA
13-Apr	NA	NA	NA	NA	NA	NA	NA
14-Apr	NA	NA	NA	NA	NA	NA	NA
15-Apr	NA	NA	NA	NA	NA	NA	NA
16-Apr	NA	NA	NA	NA	NA	NA	NA
17-Apr	NA	NA	NA	NA	NA	NA	NA
18-Apr	NA	NA	NA	NA	NA	NA	NA
19-Apr	NA	NA	NA	NA	NA	NA	NA
20-Apr	NA	NA	NA	NA	NA	NA	NA
21-Apr	NA	NA	NA	NA	NA	NA	NA
22-Apr	383.1	594.1	763.4	861.6	881.9	852.2	711.2
23-Apr	278.9	404.3	577.7	677.9	656.7	668.2	711.6
Mean	269.0	428.2	555.6	598.9	622.3	591.3	$\boxed{482.5}$

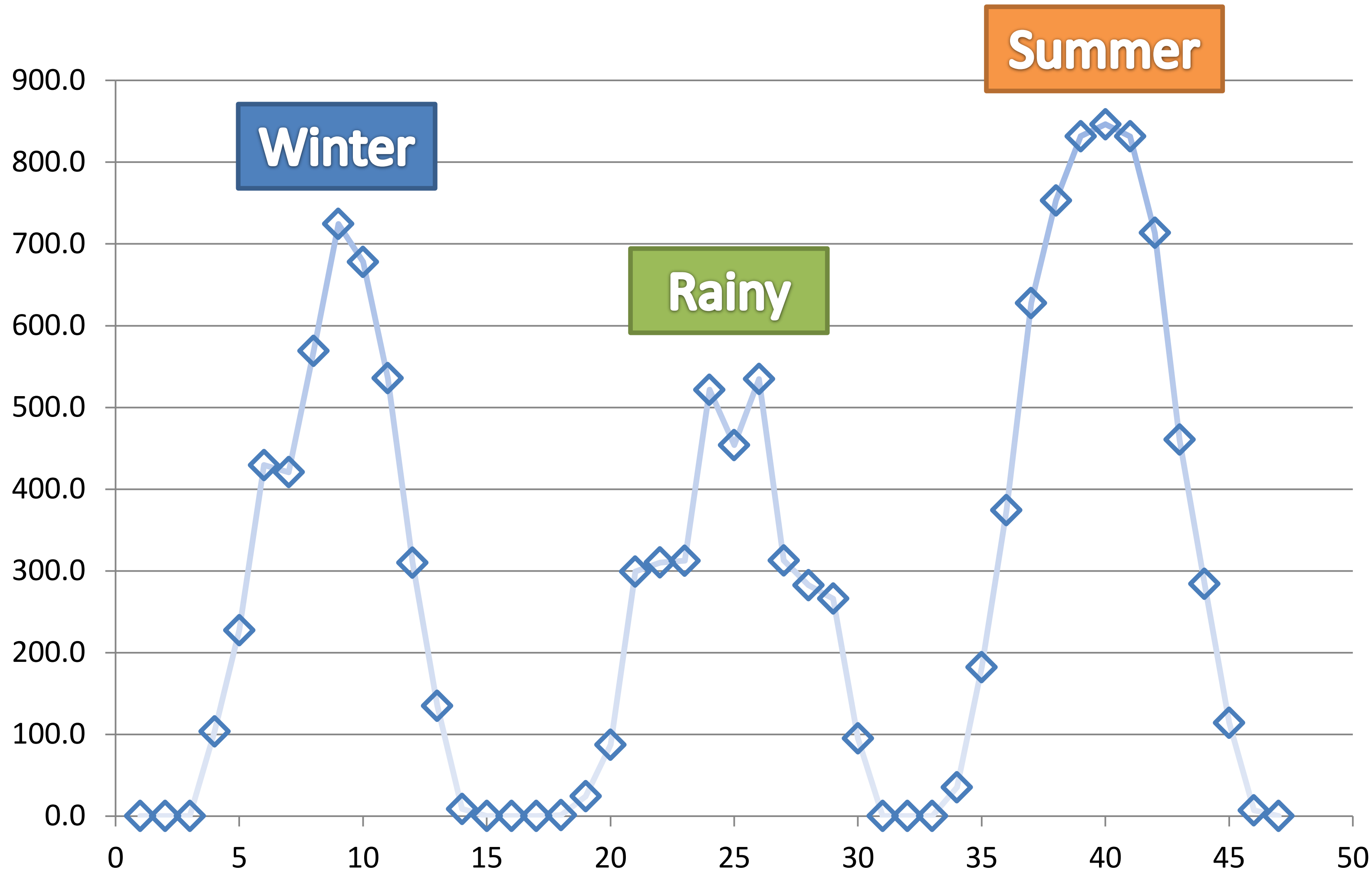
$$\bar{I}_{14} = \frac{\sum_{t=1}^{N} I_{14}(t)}{N}$$

One obvious choice is to fill the missing values with the mean.

Concept of the proposed method for missing-data

Date/Time	08-09	09-10	10-11	11-12	12-13	13-14	14-15
7-Apr	284.8	596.0	771.0	801.5	850.9	807.1	673.7
8-Apr	338.0	589.3	761.6	806.3	834.4	807.4	699.5
9-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
10-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
11-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
12-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
13-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
14-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
15-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
16-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
17-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
18-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
19-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
20-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
21-Apr	269.0	428.2	555.6	598.9	622.3	591.3	482.5
22-Apr	383.1	594.1	763.4	861.6	881.9	852.2	711.2
23-Apr	278.9	404.3	577.7	677.9	656.7	668.2	711.6
Mean	**269.0**	**428.2**	**555.6**	**598.9**	**622.3**	**591.3**	**482.5**

One obvious choice is to fill the missing values with the mean.

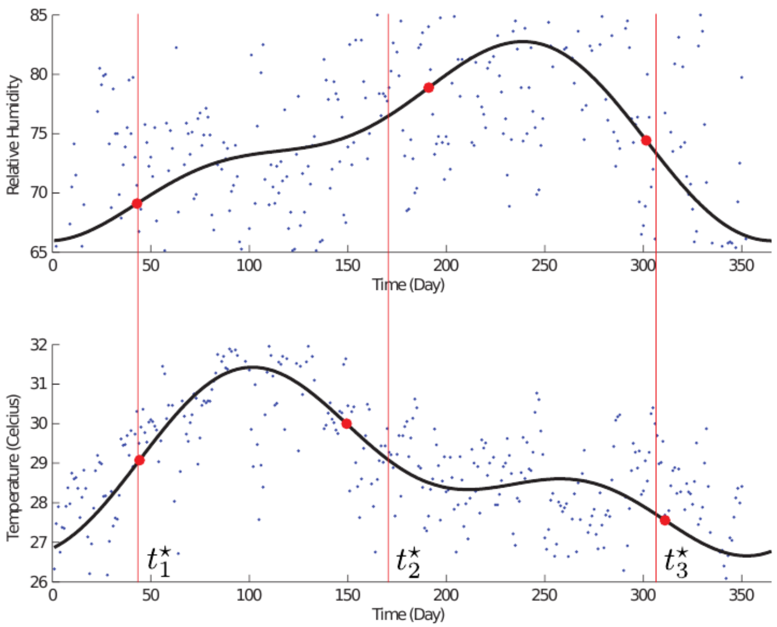


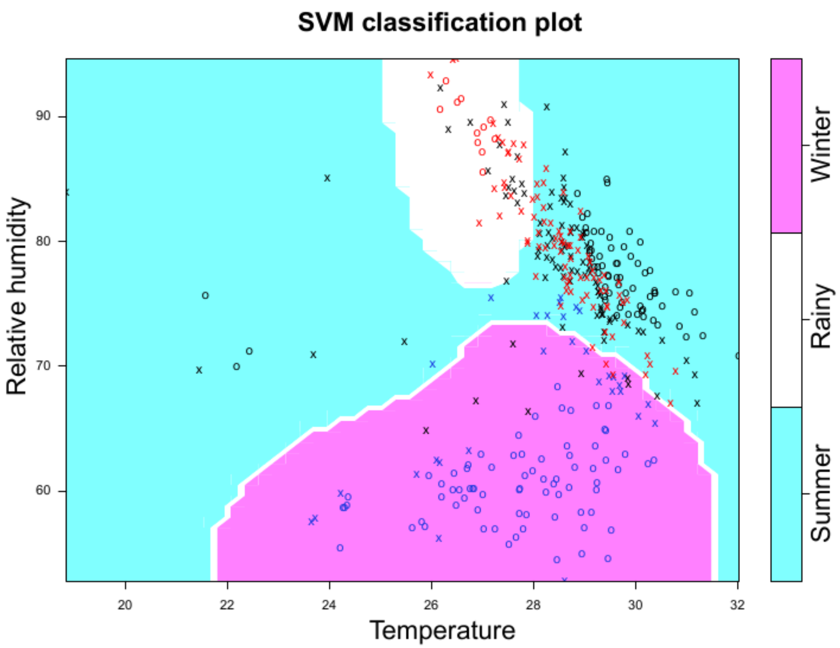
The idea is that **the mean** should be the averaged irradiance over the values **on the dates belonging to the same weather type**.

The proposed method for missing values

The required weather classification consists of two steps:

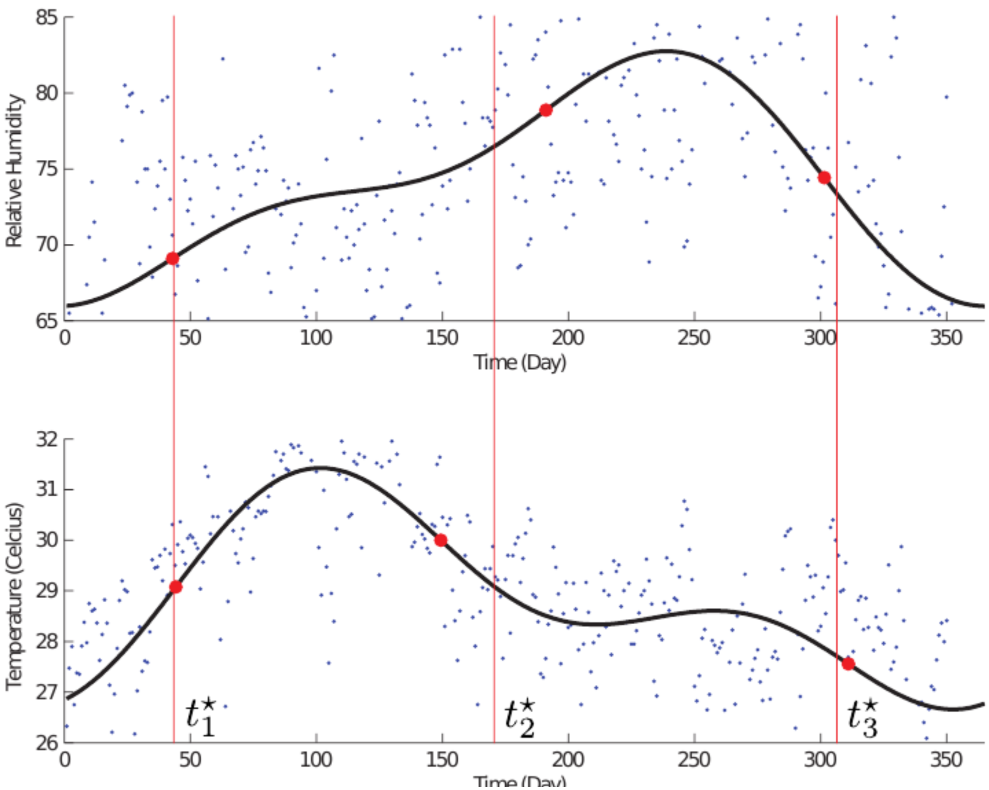
1. a seasonal segmentation based on detecting changes of monotonic properties of temperature and humidity time series
2. a nonlinear support vector machine (SVM) that uses weather labels from the previous seasonal segmentation.





The proposed method I for missing values

We propose a condition for finding time points of season changes as the monotonicity of temperature and humidity curves estimated by Fourier series.



Variables	Winter	Summer	Rainy
Relative Humidity	low	high	very high
Temperature	low	very high	high

The transition time is defined as the time point such that

$$y''(t^\star) = 0, \quad \text{and} \quad |y'(t^\star)| > \epsilon,$$

$y'(t)$ changes from being increasing to decreasing function.

Neglects too gentle slope.

Season are split by temporal order.

Weather type

1

2

3

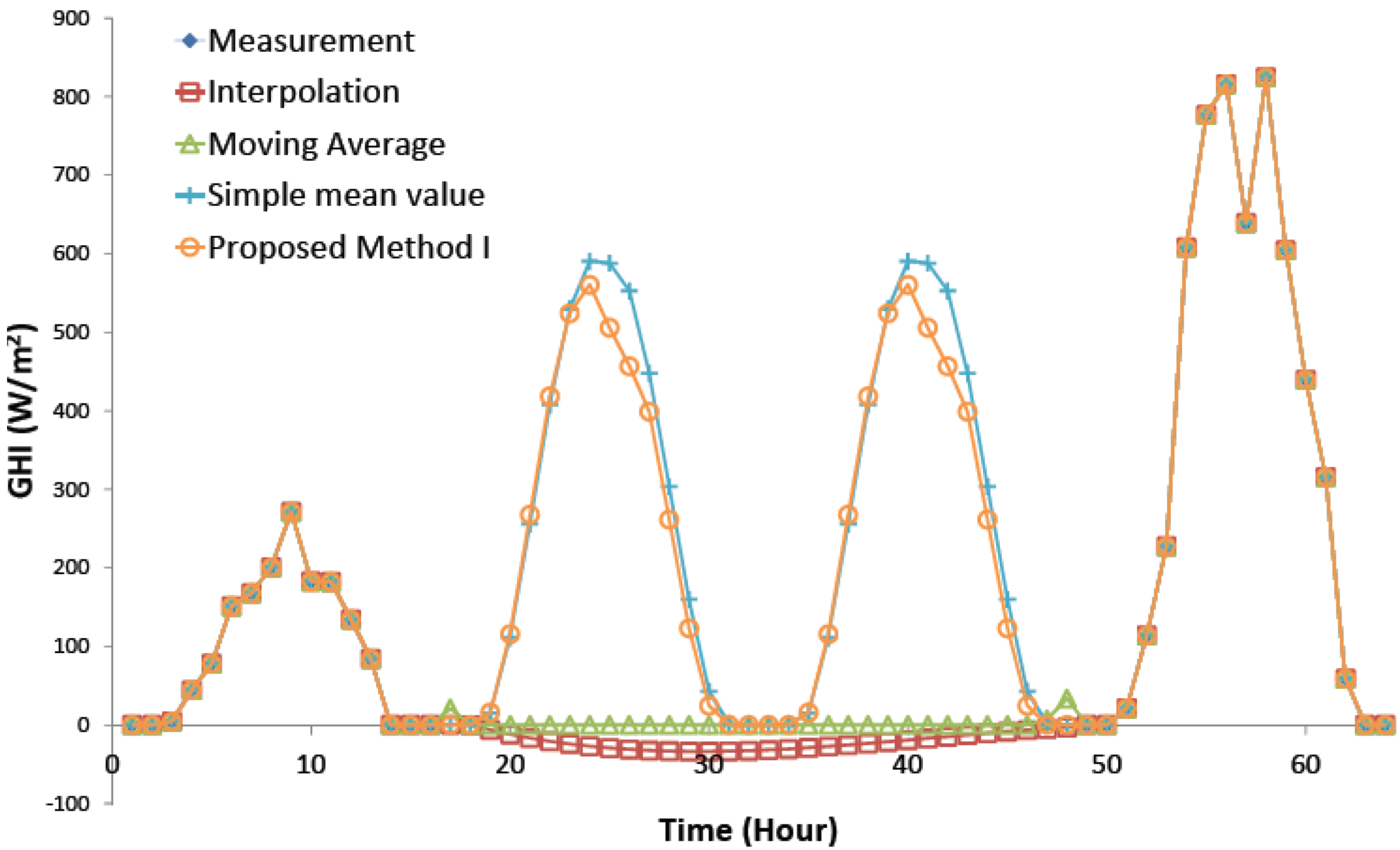
Date/Time	10-11	11-12
6-Feb	627.1	733.1
7-Feb	601.2	731.2
8-Feb	514.8	570.3
9-Feb	NA	NA
10-Feb	NA	NA
11-Feb	NA	NA
12-Feb	NA	NA
13-Feb	NA	NA
14-Feb	NA	NA
15-Feb	NA	NA
⋮	⋮	⋮
18-Jun	562.0	608.5
19-Jun	750.8	876.6
20-Jun	665.5	685.5
21-Jun	439.0	346.5
22-Jun	252.3	555.6
23-Jun	304.4	334.8

Date/Time	10-11	11-12
6-Feb	627.1	733.1
7-Feb	601.2	731.2
8-Feb	514.8	570.3
9-Feb	NA	NA
10-Feb	NA	NA
11-Feb	NA	NA
12-Feb	NA	NA
Winter	**542.8**	**604.9**

13-Feb	NA	NA
14-Feb	NA	NA
15-Feb	NA	NA
⋮	⋮	⋮
18-Jun	562.0	608.5
19-Jun	750.8	876.6
Summer	**594.8**	**665.5**

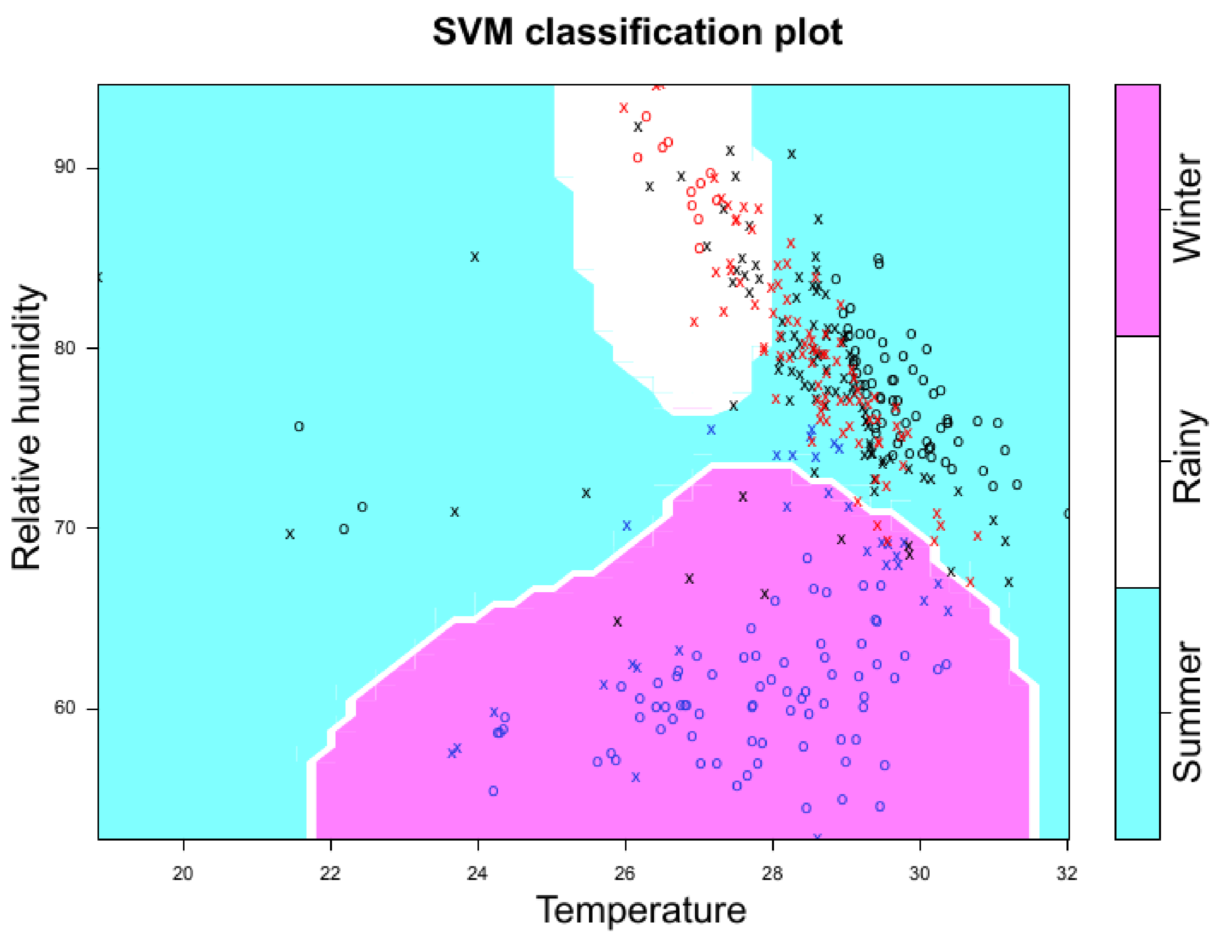
20-Jun	665.5	685.5
21-Jun	439.0	346.5
22-Jun	252.3	555.6
23-Jun	304.4	334.8
Rainny	**436.8**	**468.2**

Missing-data imputation using proposed method I



Missing data are imputed using proposed method I.

The proposed method II for missing values



Use a nonlinear SVM to classify pairs of **relative humidity and temperature data** into three classes where a prior label on the season for training is from using $t^{\star}$ to distinguish the three seasons.

Date/Time	10-11	11-12
5-Feb	386.3	278.0
6-Feb	627.1	733.1
7-Feb	601.2	731.2
8-Feb	514.8	570.3
9-Feb	NA	NA
10-Feb	NA	NA
11-Feb	NA	NA
12-Feb	NA	NA
13-Feb	NA	NA
14-Feb	NA	NA
15-Feb	NA	NA
16-Feb	NA	NA
17-Feb	NA	NA
18-Feb	NA	NA
19-Feb	NA	NA
20-Feb	NA	NA
21-Feb	NA	NA
22-Feb	NA	NA
23-Feb	687.9	783.9
24-Feb	572.4	446.6
25-Feb	585.0	508.2
26-Feb	526.1	705.8

Orange : Hot
Green : Rainy
Blue : Winter

Classified data using the proposed method.

Missing-data imputation using proposed method II

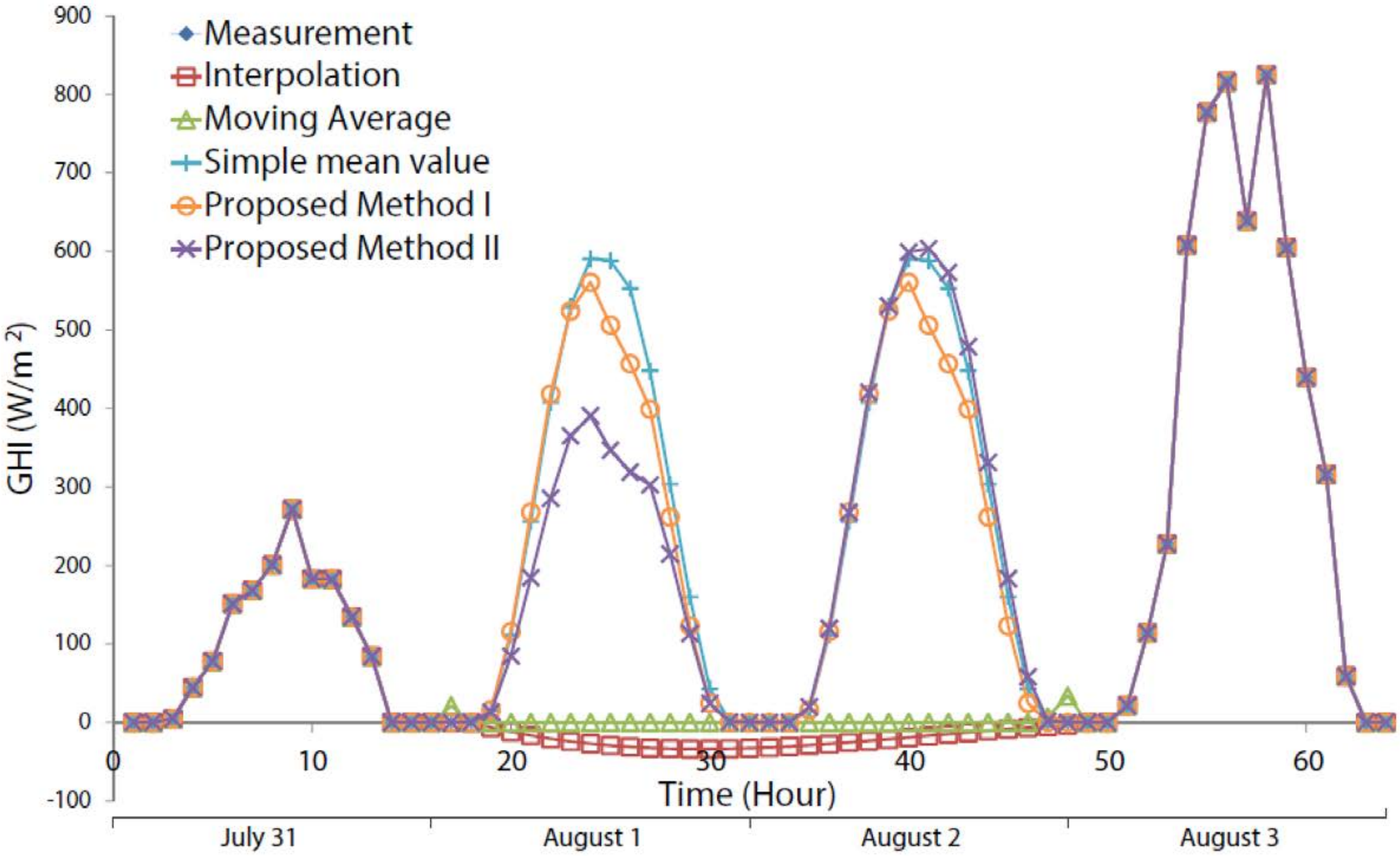
Date/Time	10-11	11-12
9-Feb	NA	NA
12-Feb	NA	NA
15-Feb	NA	NA
16-Feb	NA	NA
18-Feb	NA	NA
19-Feb	NA	NA
20-Feb	NA	NA
21-Feb	NA	NA
25-Feb	585.0	508.2
26-Feb	526.1	705.8
Summer	627.1	695.0

Date/Time	10-11	11-12
5-Feb	386.3	278.0
6-Feb	627.1	733.1
7-Feb	601.2	731.2
8-Feb	514.8	570.3
10-Feb	NA	NA
11-Feb	NA	NA
14-Feb	NA	NA
17-Feb	NA	NA
22-Feb	NA	NA
Rainny	420.4	457.3

Date/Time	10-11	11-12
13-Feb	NA	NA
23-Feb	687.9	783.9
24-Feb	572.4	446.6
Winter	563.7	634.6

Season are split by proposed method and fill the missing values with the mean belonging to the same weather type.

Missing-data imputation using proposed method II



Missing data are imputed using proposed method II.

We assess the imputation methods by presumably **deleting the recorded values** from the data sets, and then **we evaluate how well the deleted values are predicted** in a yearly basis

Forecasting Performance Evaluation Measures

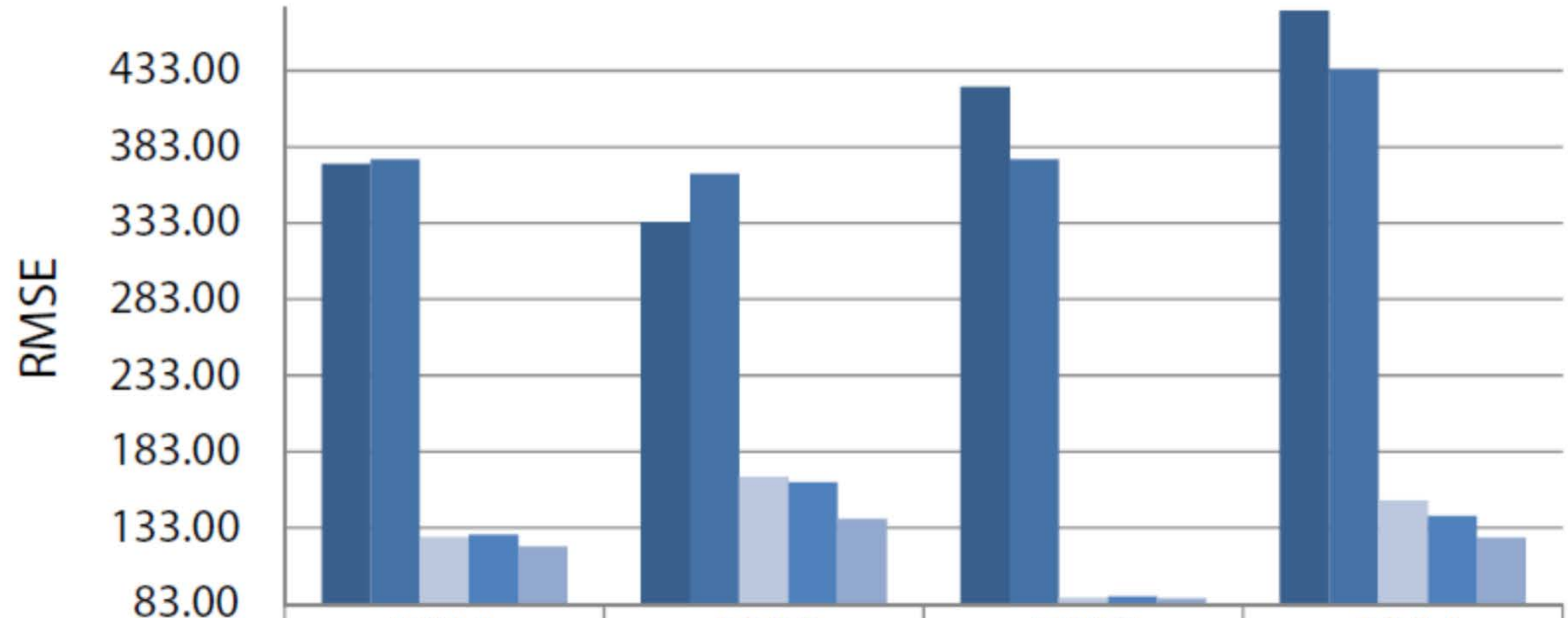
Common evaluation measures are **Root Mean Squared Error (RMSE)** and **Mean Absolute Error (MAE)** which are used to validate a forecasting method. RMSE and MAE can be defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(I(t) - \hat{I}\,(t)\right)^2}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N}\left|I(t) - \hat{I}\,(t)\right|$$

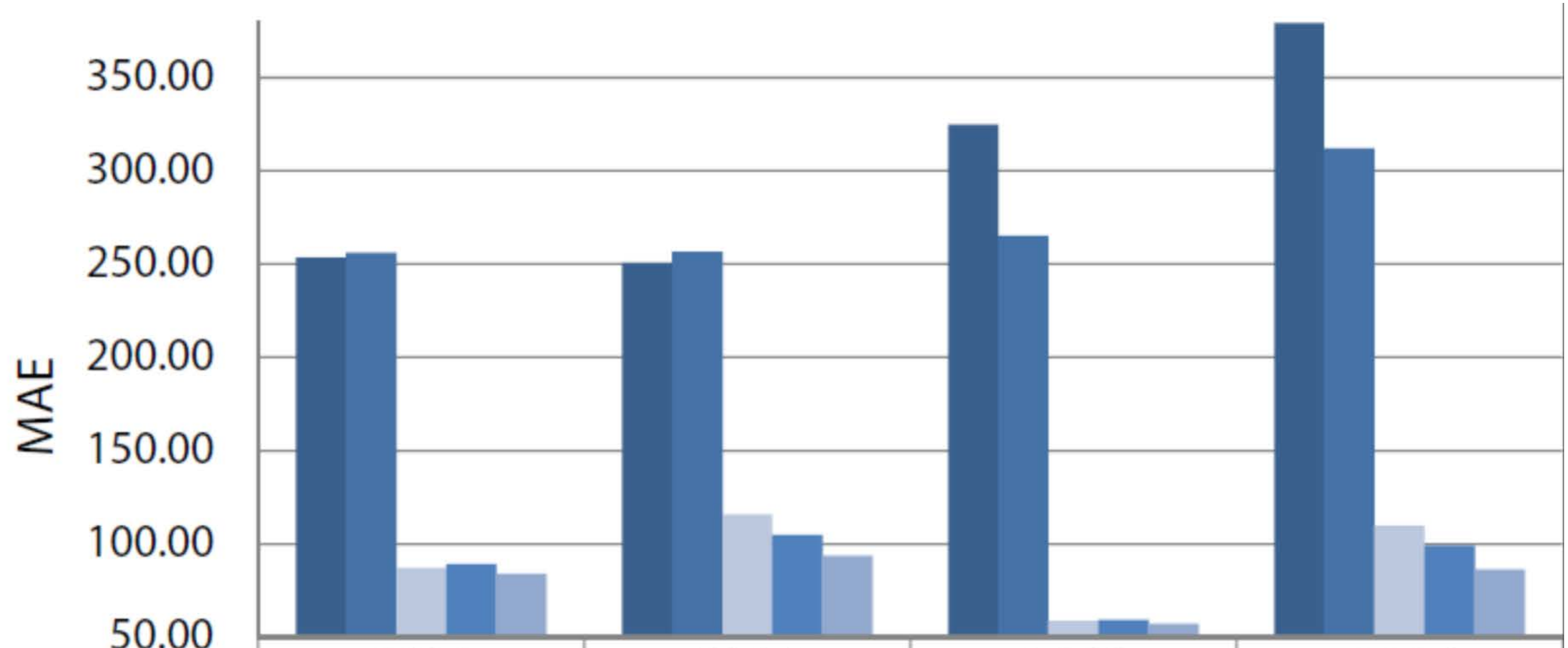
where N is the length of the time horizon.

Validation of the proposed method



	2011	2012	2013	2014
Interpolation	371.79	333.61	422.41	472.42
Moving Average	374.72	365.64	374.90	434.10
Mean in one year	126.81	166.62	87.00	150.69
Proposed Method I	128.46	162.87	87.98	141.08
Proposed Method II	120.77	138.94	86.70	126.53

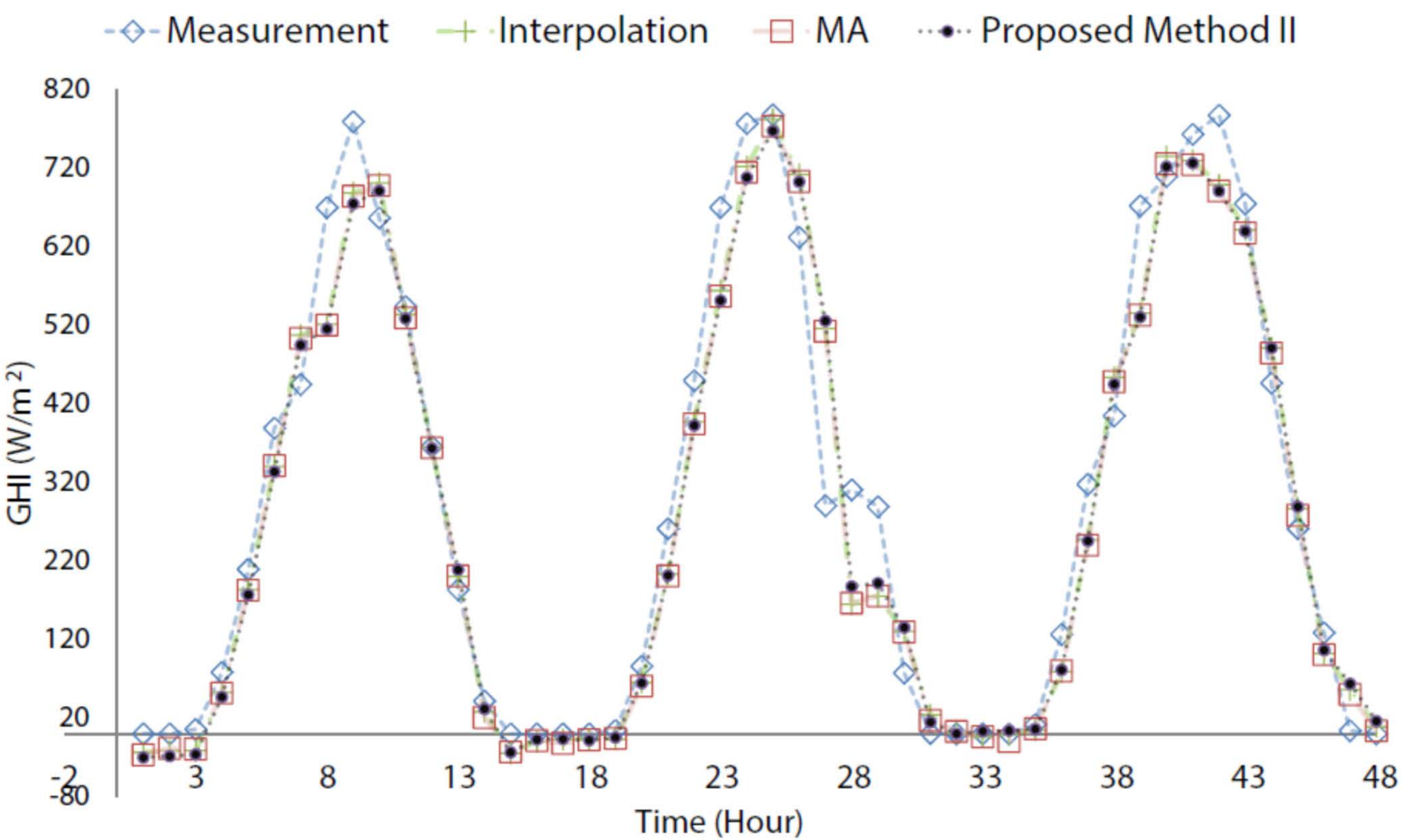
Validation of the proposed method



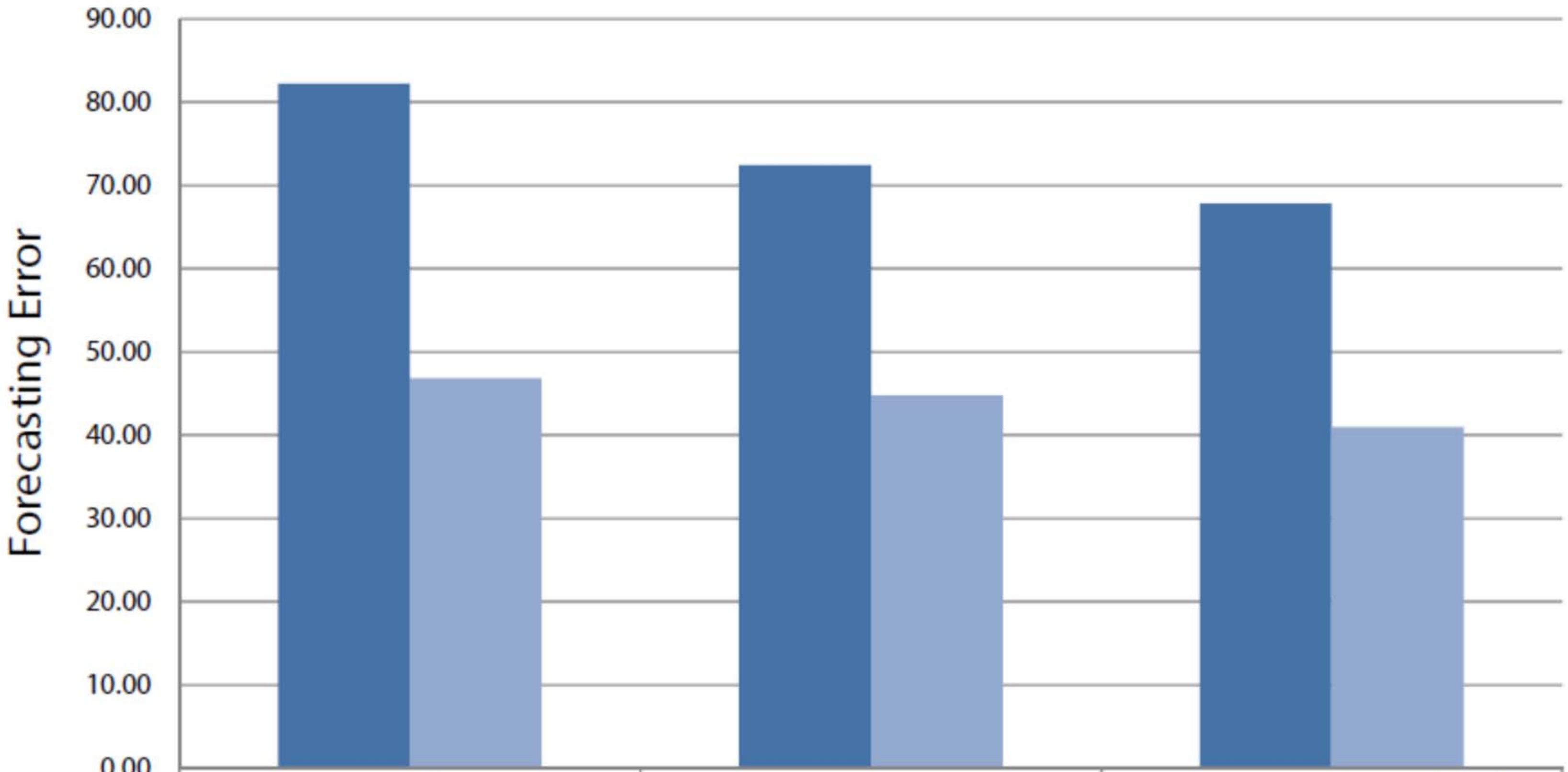
	2011	2012	2013	2014
Interpolation	253.53	250.61	324.64	379.23
Moving Average	255.98	256.58	265.15	312.16
Mean in one year	87.29	115.67	58.88	109.78
Proposed Method I	89.34	104.84	59.15	99.13
Proposed Method II	83.81	93.68	57.34	86.41

Next experiment illustrates **the forecasting results (after imputing missing-data)** by using SARIMA models (choosing SARIMA$(6,2,5)(0,1,1)_{16}$) that requires a complete set of historical data in the estimation process.

Forecasting solar irradiance (Sample)



Validation of the proposed method



	Interpolation	Moving Average	Proposed Method II
RMSE	82.20	72.45	67.77
MAE	46.83	44.79	40.97

Forecasting errors using different methods of data imputation.

Conclusions

- In Thailand, several consecutive days of missing GHI data and limited acquisition of other physical variables are prevailing conditions.
- Other imputation techniques may not perform well under these conditions
- The method acquires only the available temperature and humidity data.

Conclusions

- **The imputation using the mean value** on the dates belonging **to the same weather type.**

- The required **weather classification consists of two steps:** 1) a seasonal segmentation based on **detecting changes of monotonic properties** of temperature and humidity time series, 2) **nonlinear support vector machine (SVM)** that uses weather labels from the previous seasonal segmentation.

- The proposed method **significantly provide decreased imputation errors** and leads to a **better solar forecasting performance** compared to other imputation techniques.

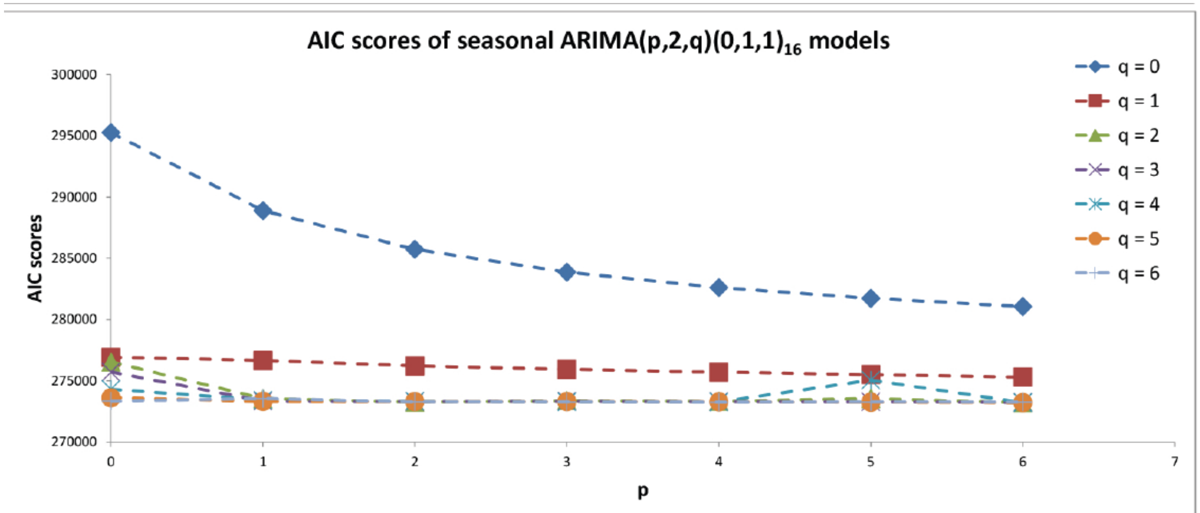
Acknowledgement

- Thank Chula Engineering for the support of research facilities.

- Thank Director Somkuan Tonjan, Numerical Weather Prediction Division, Weather Forecast Bureau, Thai Meteorological Department (TMD) for providing the GHI data and practical information.

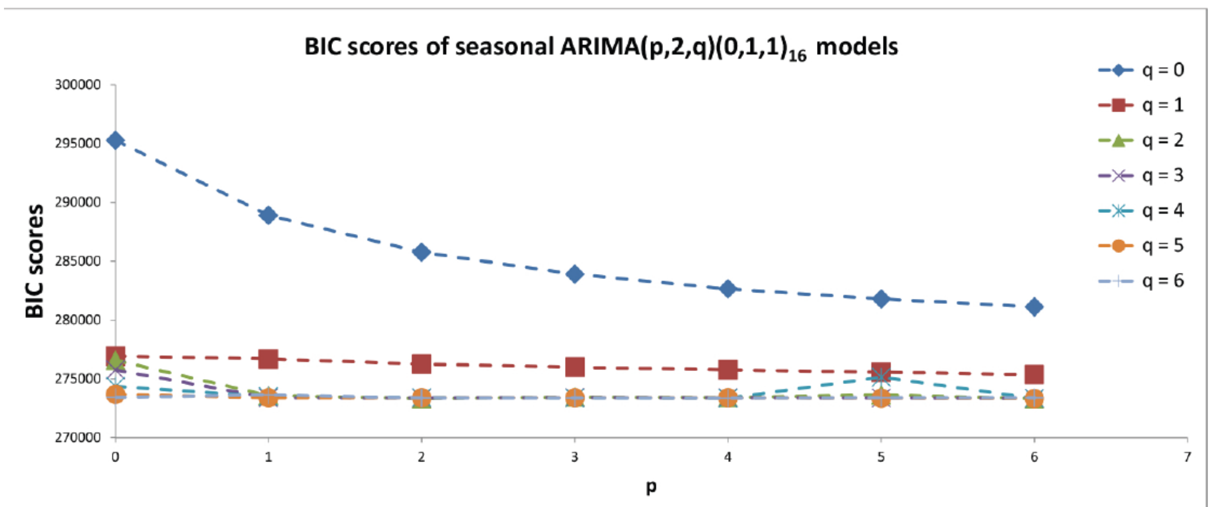
Q&A

Backup



Order : Seasonal ARIMA$(p,2,q)(0,1,1)_{16}$

Selected order : Seasonal ARIMA$(6,2,5)(0,1,1)_{16}$



Order : Seasonal ARIMA$(p,2,q)(0,1,1)_{16}$

Selected order : Seasonal ARIMA$(6,2,2)(0,1,1)_{16}$