5

ENSEMBLE FORECASTING of SOLAR IRRADIANCE USING DATA MINING TECHNIQUES

5.1 INTRODUCTION

In this chapter, we propose a novel hybrid framework based on data mining techniques which use both STL and the wavelet transform for hourly short term solar irradiance forecasting. Feedforward neural network (FFNN) is used as a predictor. Data mining through decomposition helps in better characterization of global horizontal irradiance and provides more appropriate learning of neural network to enhance the accuracy of forecast results. We use STL to decompose solar irradiance data into seasonal, trend and remainder components. Residual component of the data is then obtained by subtracting seasonal component from the data, which is also the sum of trend and remainder components. Residual component significantly contributes to changing dynamics of the data, whereas the seasonal component shows the day to day repetitions of the data and so it is a deterministic quantity. The residual series is further decomposed by using the wavelet transform. Each decomposed sub-series of the data is then used for forecasting by an appropriately fitted feedforward neural network (FFNN). The final forecast is obtained by adding the fitted FFNN results and the previously obtained seasonal data. The proposed forecasting method provides a hybrid framework for hourly solar irradiance forecasting with the following contributions:

1. A novel hybrid framework is developed that combines more than one data mining/preprocessing techniques with FFNN for solar irradiance forecasting.

2. The performance after using a preprocessing technique on forecast accuracy is evaluated.

3. The accuracy of the forecast model is compared with the competing model.

The data is hourly records of global horizontal irradiance (GHI) by Indian Meteorological Department (IMD), Jodhpur, India of the year 2015, that is, the record amounting to 744, 720 and 672 for the months having 31 days, 30 days and 28 days respectively. For the demonstration purpose, only November month data is presented, although analysis is carried out for all the months of The year 2015. In the following section, we present the proposed methodology.

The rest of the chapter is organized in the following way. Next Section 5.2 provides a complete picture of proposed modelling process adopted in this chapter. Section 5.3 is dedicated to discussing the results obtained. The chapter is completed by discussing conclusions of the work in Section 5.3.

5.2 PROPOSED METHODOLOGY

The objective of this work is to present the advantages of using a preprocessing techniques in the construction of an ensemble model and impact on the forecast accuracy. STL filtering process allows data series to decompose into a deterministic component (seasonal) and a random component (residual). The flow diagram of the whole procedure is shown in the following Figure 5.1. The proposed forecasting method can be summarized in the following step by step algorithm:

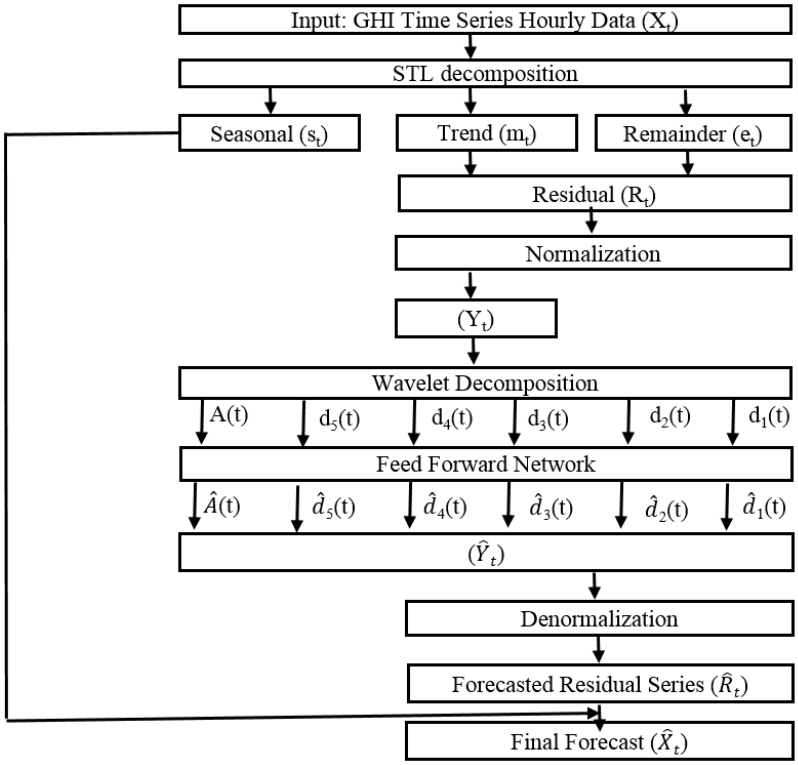

Figure 5.1 : Flowchart of the proposed methodology

1. Decompose the original GHI series (X_t) into seasonal (s_t), trend (m_t) and remainder series (e_t) using STL decomposition.

2. Subtract the seasonal component (s_t) from the original series (X_t) to get residual series (R_t) which is also the sum of trend (m_t) and remainder component (e_t).

3. The residual series is normalized by using,

$$Y_t = \frac{R_t - min(R_t)}{max(R_t) - min(R_t)}. \tag{5.1}$$

4. The normalized residual series Y_t is then decomposed using five time-steps Mallat's decomposition algorithm of wavelet decomposition. This results in a low-frequency sequence $A(t)$ and high-frequency sequences $d_1(t)$, $d_2(t)$,..., $d_5(t)$.

5. The above sequences A(t), $d_1(t)$, $d_2(t)$, $d_3(t)$, $d_4(t)$ and $d_5(t)$ are used for forecasting.

6. The forecasted values $\hat{A}(t)$, $\hat{d_1}(t)$, $\hat{d_2}(t)$,..., $\hat{d_5}(t)$ are combined to obtain $\hat{Y}_t$.

7. The data series $\hat{Y}_t$ is denormalized using,

$$\hat{R}_t = \hat{Y}_t * (max(R_t) - min(R_t)) + min(R_t), \tag{5.2}$$

where, $\hat{R}_t$ denotes the denormalized residual data series after prediction.

8. The forecasted data series $\hat{R}_t$ is added to the seasonal component of data series to obtain the final forecast $\hat{X}_t$.

To demonstrate the proposed methodology, a one-step ahead forecast of hourly solar irradiance, for November month, using STL along with the wavelet analysis and feedforward neural network is carried out. Using STL decomposition, the data is decomposed into seasonal, trend and remainder data sub-series. Figure 5.2 shows the data series and the decomposed sub-series of November month. After decomposition of the data, the residual series is obtained by subtracting the seasonal sub-series of the data from the November data.

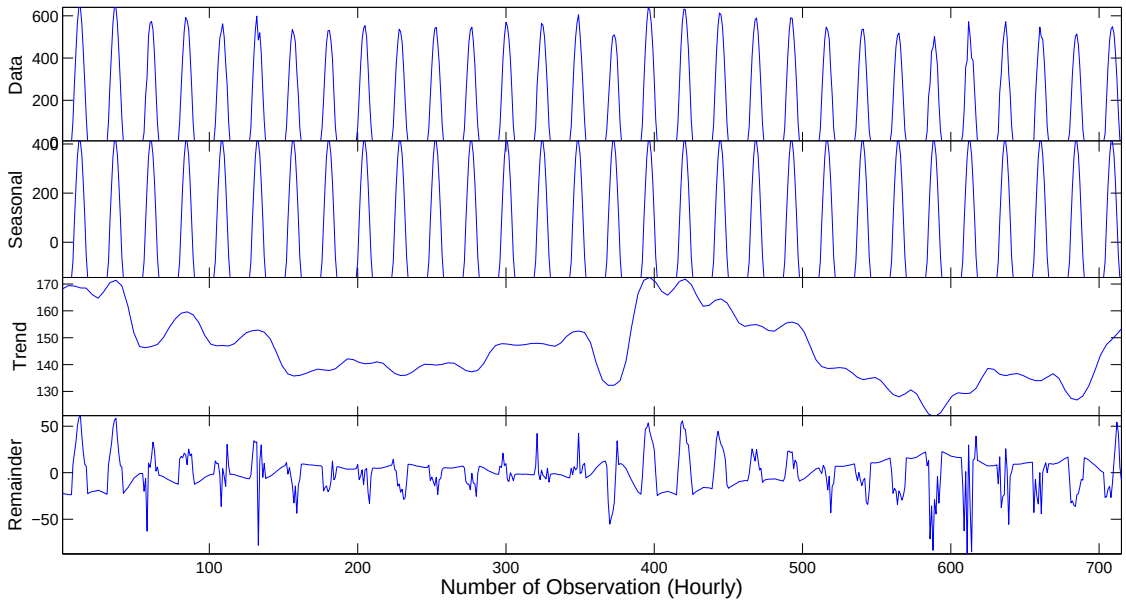

Figure 5.2 : STL decomposition to the time series of November 2015

Now the residual sub-series of the data is further decomposed using Mallet's pyramidal method with the mother wavelet of 'coif5'. The low-frequency sequence $A(t)$ and the high-frequency sequences $d_5(t)$, $d_4(t)$, $d_3(t)$, $d_2(t)$, and $d_1(t)$ of each month are obtained by the five scale wavelet decomposition. The original signal is constituted of an algebraic sum of low-frequency and high-frequency sequences. Then, $A(t)$ and $d_5(t)$, $d_4(t)$, $d_3(t)$, $d_2(t)$, $d_1(t)$ are used as inputs. Network training function, used to updates weights and bias values, is 'trainlm' which follows Levenberg-Marquardt optimization procedure.
A 3-layer FFNN is constructed in the following way.
The neural network are constituted by five input layer neurons, ten hidden layer neurons and one output layer neuron. So the FFNN can be trained using the sequence of the low-frequency of A(t), which has either 744 or 720 or 672 records depending on the month under consideration, decomposed from the irradiance sequence. For demonstration purpose, we have used 720 records of November month of 2015. In the training, the stop condition is used as the limitation of training error 0.01 along with the maximum training epochs that are set as 1000. The output of each sub-series is combined to obtain the series $\hat{Y}_t$. The series is then denormalized and added to the seasonal component to obtain the final forecast $\hat{X}_t$. Figure 5.3 shows the actual wavelet decomposed sub-series and their corresponding forecasted sub-series.

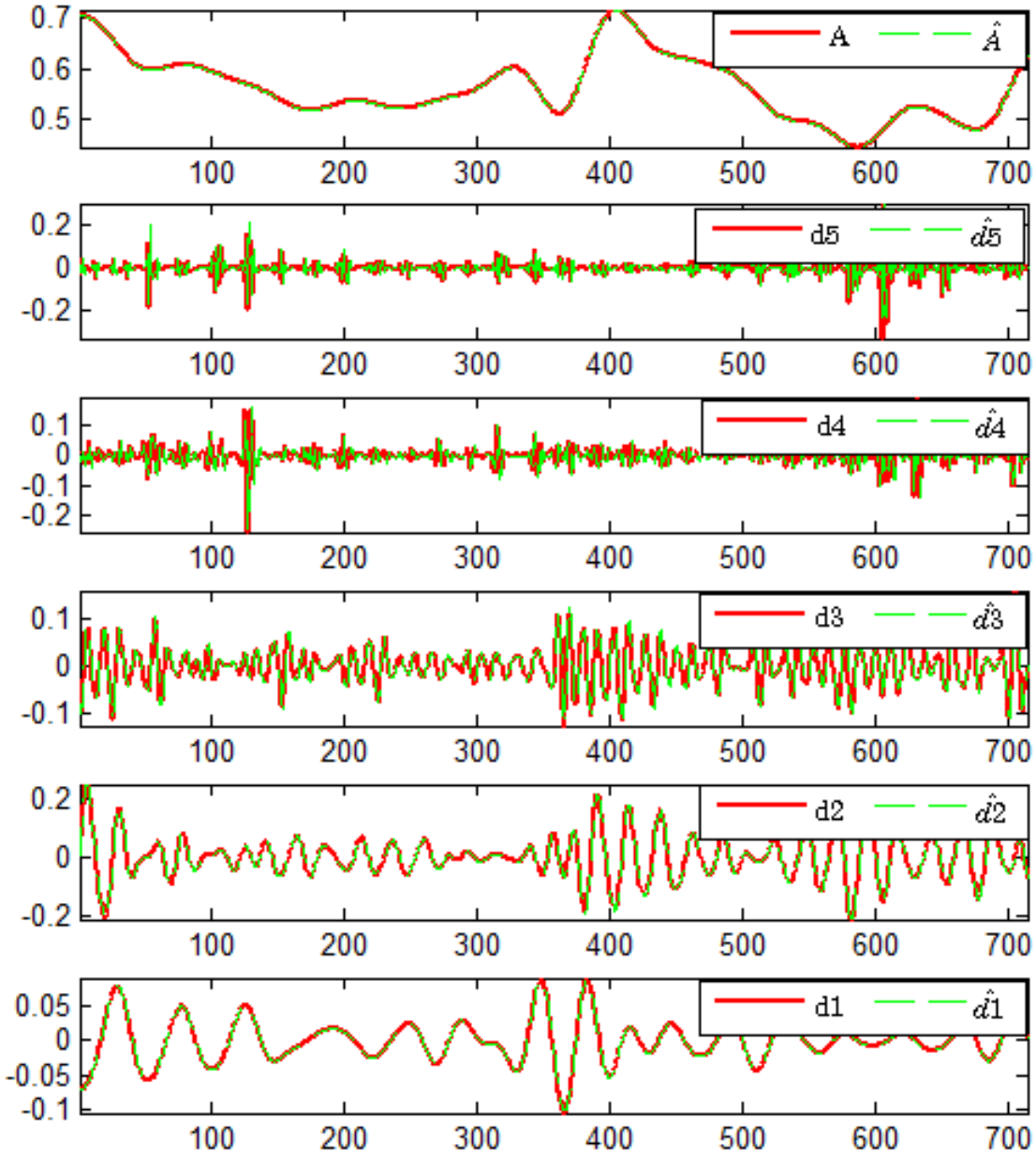

Figure 5.3 : Actual wavelet decomposed sub-series ($A(t)$, $d_1(t)$, $d_2(t)$, $d_3(t)$, $d_2(t)$, $d_5(t)$) and their corresponding predicted sub-series ($\hat{A}(t)$, $\hat{d}_1(t)$, $\hat{d}_2(t)$, $\hat{d}_3(t)$, $\hat{d}_4(t)$, $\hat{d}_5(t)$) of the November month

5.3 RESULTS AND DISCUSSIONS

To highlight the improvement in the forecast accuracy of candidate models we have calculated the forecast error Tables 5.1 and forecast skill 5.2 metric of all the candidate models respectively. As previous chapters here also we used Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as error metric to evaluate the accuracy for comparative statistical analysis of all the comparing models. The November data series and its forecast using (i) FFNN, (ii) STL along with FFNN (SFFNN) and (iii) STL, wavelet along with FFNN (SWFFNN) are shown in the Figure 5.4. For better visibility, it is scaled up and it is given shown in Figure 5.5. Along with that in 5.1 we

have calculated the RMSE and MAE error for all the model for error analysis. From the analysis it is evident that the the error gets reduce gradually as compared to persistence. SFFNN show lesser RMSE and MAE as compared to FFNN and SWFFNN has lesser error when compared to SFFNN. Further through statistical analysis we can test whether the reduction in error is random or it is significant. To test it, we have used t-test to check whether the RMSE of SWFFNN is significantly lesser than RMSE of SFFNN. We have performed one tail t-test over RMSE value of SWFFNN against RMSE value of SFFNN and calculated the p-value which is coming out to be 0.000069. So, we can conclude from the statistical test that the reduction in the error is significant. Similarly, we performed the t test over RMSE value of SFFNN and FFNN and calculated the p-value. The p-value came out to be 0.000061 in the case. Again, we can conclude from this statistical analysis that the reduction in the error is significant. So, finally we can conclude from this statistical test that our proposed model is outperforming when compared to persistent model and FFNN. The basic idea of the proposed model is two-stage decomposition of data, using STL and the

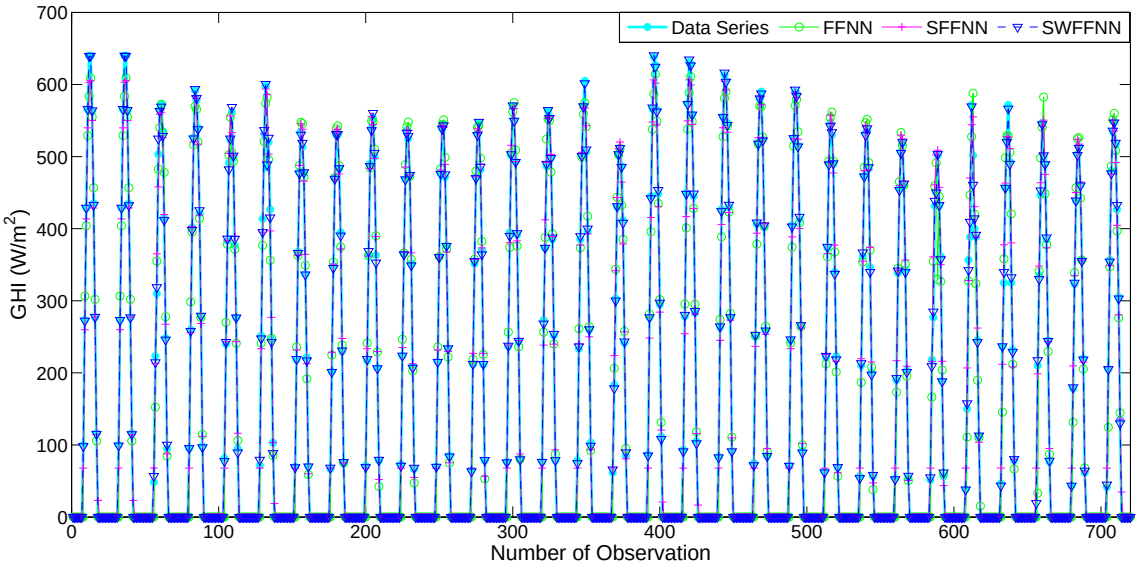

Figure 5.4 : November data series and its forecast

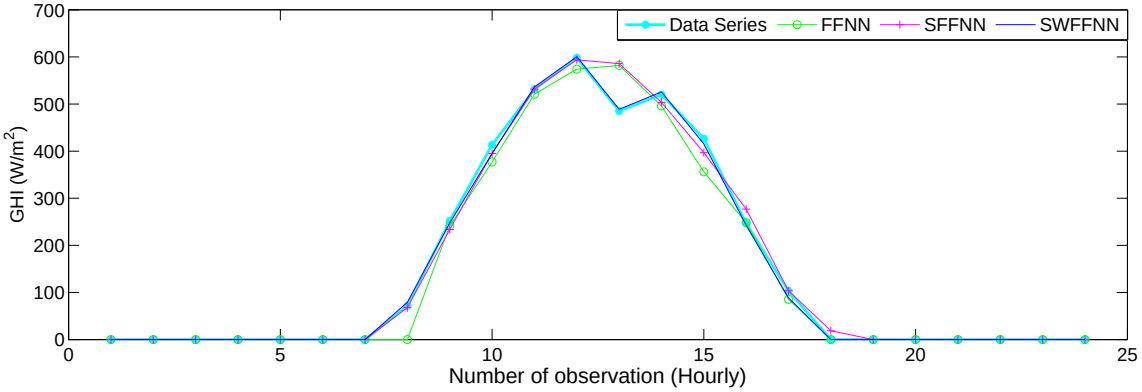

Figure 5.5 : Data series and its forecast in zoomed version

wavelet decomposition techniques to use advantages of combined model. STL decomposition along with the wavelet decomposition of data helps in reducing the complexity of the data in the

more prominent way and enhance the learning of FFNN.

Table 5.1 : Error comparison table for different models

Training	Persistence		FFNN		SFFNN		SWFFNN	
Data	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
January	50.1	85.5	26.4	44.5	15	38.1	**3.3**	**8.2**
February	57.4	92.7	19.9	30.5	14.3	29.5	**2.8**	**6.5**
March	69.9	115.5	30.3	62.7	24.1	54.6	**7.6**	**13.4**
April	70.5	107.4	25	41.8	14	32.3	**4.1**	**8.5**
May	71.5	107.4	19.8	29.3	10.6	23.3	**3.3**	**6.1**
June	71.8	114.7	52.8	78.5	32.9	64.2	**9.7**	**14.6**
July	67.7	121	63	98	45.9	80.2	**11.4**	**19.4**
August	75.1	125.7	63.9	95.8	47.2	82.7	**11.9**	**19.7**
September	68.1	111.2	40.8	65.8	26.2	51.2	**6.9**	**12.8**
October	58.1	92.6	11.6	24.6	5.2	10.6	**1.4**	**3.4**
November	47.3	77.7	11.9	21	6.1	13	**2**	**3.7**
December	45	76.8	14.9	28.8	8.3	19.5	**2.1**	**4.5**

Table 5.2 : Forecast skill comparison table for different models

Training Data	Persistence S	FFNN S	SFFNN S	SWFFNN S
January	0	0.47	0.55	**0.90**
February	0	0.67	0.68	**0.92**
March	0	0.45	0.52	**0.88**
April	0	0.61	0.69	**0.92**
May	0	0.72	0.78	**0.94**
June	0	0.31	0.44	**0.87**
July	0	0.19	0.33	**0.83**
August	0	0.23	0.34	**0.84**
September	0	0.40	0.53	**0.88**
October	0	0.73	0.88	**0.96**
November	0	0.72	0.83	**0.95**
December	0	0.63	0.74	**0.94**

5.4 CONCLUSIONS

The proposed algorithm is demonstrated for the hourly GHI dataset of each month of the year 2015 step by step for the proposed model and the developed model is compared with the persistence model, FFNN and SFFNN. Persistence model, a standard benchmark model uses the current value of the time series as the forecast value. RMSE and MAE values of each month for the proposed model is significantly low as compared to the other models throughout the year as seen in Table 5.1. The calculated RMSE and MAE values are more for FFNN and gradually decreases for SFFNN and is minimum for SWFFNN. RMSE and MAE values are minimum for

October month and maximum for August month and the same trend is followed for forecast skill in Table 5.1. August month has a maximum error because the climatic conditions of this month is more uncertain as compared to rest of the months and converse is the case for October month. Moreover, the forecast skill also shows an improvement in its value for the proposed hybrid model given in Table 5.2 which further strengthens our claim. Along with that, it re-established the fact that the preprocessing significantly reduces the complexity of the data and improves the forecast accuracy by enabling us to handle a group of data separately.

…