Inter-Hour Forecast of Solar Radiation Based on the Structural Equation Model and Ensemble Model

Tingting Zhu 1,2,*,†, Yiren Guo 1,*,†, Cong Wang 1 and Chao Ni 1

1 College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China; guoyr@njfu.edu.cn (Y.G.); c_840305481@163.com (C.W.); chaoni@njfu.edu.cn (C.N.)
2 Key Laboratory of Measurement and Control of Complex Systems of Engineering (Southeast University), Ministry of Education, Nanjing 210096, China
* Correspondence: tingting_zhu2018@163.com
† The two authors contributed equally to this work.

Received: 13 July 2020; Accepted: 31 August 2020; Published: 1 September 2020

Abstract: Given the wide applications of photovoltaic (PV) power generation, the volatility in generation caused by solar radiation, which limits the capacity of the power grid, cannot be ignored. Therefore, much research has aimed to address this issue through the development of methods for accurately predicting inter-hour solar radiation and then estimating PV power. However, most forecasting methods focus on adjusting the model structure or model parameters to achieve prediction accuracy. There is little research discussing how different factors influence solar radiation and, thereby, the effectiveness of these data-driven methods regarding their prediction accuracy. In this work, the effects of several potential factors on solar radiation are estimated using correlation analysis and a structural equation model; an ensemble model is developed for predicting inter-hour solar radiation based on the interaction of those key factors. Several experiments are carried out based on an open database provided by the National Renewable Energy Laboratory. The results show that solar zenith angle, cloud cover, aerosols, and airmass have great effects on solar radiation. It is also shown that the selection of the key factor is more important than the model structure construction for predicting solar radiation precisely. The proposed ensemble model proves to outperform all sub-models and achieves about a 12% improvement over the persistent model based on the normalized root mean squared error statistic.

Keywords: solar radiation forecast; effectiveness estimation; structural equation model; ensemble model

1. Introduction

Solar energy is an important form of clean energy and has attracted much attention because of its abundance, lack of pollution, and wide distribution. However, solar power is fluctuating and intermittent due to changes in solar radiation; these two features are, though, vital for the stability and safety of the power grid [1–3]. Therefore, it is necessary to precisely predict solar power generation in order to use it more effectively.

Since solar radiation is the key factor that causes changes in solar power, precise predictions of solar radiation can allow accurate forecasting of solar power [4]. Many methods have been developed to predict the inter-hour solar radiation and can be separated into two classes: theoretical methods and data-driven methods (or empirical methods) [5–10]. Theoretical methods usually start with the solar radiation transmission path and consider the attenuation effect of atmosphere on solar radiation [11]. These methods are complicated and involve large amounts of observation data. Data-driven methods are usually based on the statistics of historical observation data, instead of radiative transfer theory.
and are relatively simple and flexible [12,13]. Therefore, data-driven methods are more widely used to predict solar radiation in engineering applications, especially for a micro-grid or an isolated grid.

The common data-driven methods include the Regression Model (RM) [14], Support Vector Machine (SVM) [15], Artificial Neural Network (ANN) [16], and hybrid models in which several single models are combined [17–19]. However, most research on solar radiation forecasting based on data-driven methods focuses on how to improve the prediction accuracy by adjusting the model structure or model parameters [20–22]. There is little research about the effectiveness of different factors on solar radiation and how they influence the prediction accuracy of these data-driven methods. Alskaif et al. analyzed the impact of nine different meteorological variables on PV output power and used a lower-dimensional subspace of meteorological variables as input for the regression methods to calculate the PV output power [23]. However, this study focused on the stochastic factors and ignored the deterministic variables to the PV output power. On the other hand, the meteorological variables except for solar radiation can be estimated or predicted from the PV output power directly by machine learning methods, but these variables mostly influence the solar radiation and the solar radiation change causes the fluctuation and intermittence of the PV output power.

Therefore, this paper analyzes the influence of 11 variables including deterministic and stochastic factors on solar radiation and constructs an ensemble model to predict inter-hour solar radiation. The objectives of this study are (1) to estimate the effectiveness of different factors on solar radiation from various perspectives, including individual effectiveness and interactions, (2) to analyze the influence of these factors on the prediction accuracy, and (3) to construct an ensemble model to precisely predict inter-hour solar radiation.

2. Materials and Methods

2.1. Experimental Data and Evaluation

2.1.1. Data Collection and Processing

All measured data were downloaded from an open database: the NREL’s Solar Radiation Research Laboratory (SRRL) [24]. The SRRL station is located at 39.74° N and 105.18° W, 1829 m above sea level in Golden City, Colorado, USA, where there are abundant solar resources. Measured meteorological parameters are obtained with 1 min sampling frequency in the database.

In this paper, the factors discussed are mostly based on surface stations; the solar radiation data refer to the global horizontal irradiance (GHI) for photovoltaic power plants, with the aim of serving power prediction and grid connection of photovoltaic power stations. Because of the rotation of the Earth around its axis and around the Sun, the solar radiation has daily and annual cycles [25–27]. Therefore, the solar zenith angle was selected to represent the daily cycle of solar radiation, and the solar azimuth angle and the day of the year were chosen to represent the annual cycle. Considering that GHI reaches a theoretical maximum at noon, when the solar zenith angle is at its lowest value during the day, the cosine of the zenith angle rather than the zenith angle itself better describes the daily cycle of GHI; many published works have also used the cosine function to calculate the clear-sky GHI [28]. In addition, the day of the year (DOY) is set as 1 for 1 January and 365 for 31 December, so the sine function of DOY (Γ) is used instead of DOY itself in the following formulation [17].

\[ \Gamma = \sin(2\pi \frac{\text{DOY} - 105}{365}) \] (1)

Besides the length of the path of solar rays in the atmosphere, the weather or atmospheric composition is another aspect that causes changes in solar radiation. Among all components of the atmosphere, cloud may be the most direct and main factor causing short-term solar radiation fluctuations. Cloud cover and cloud motion are the two main features affecting solar radiation. Since there is no cloud motion in the open database, wind speed and wind direction were selected
instead as potential factors to represent cloud motion [29,30]. Moreover, aerosols and relative humidity have been considered as atmospheric components in some published works [31]. In addition, station pressure, airmass, and temperature are other factors affecting solar radiation.

Finally, 11 potential factors have been collected: cosine function of solar zenith angle (Z), solar azimuth angle (A), sine function of DOY (\(P\)), cloud cover (CC), wind speed (WS), wind direction (WD), aerosol optical depth (AOD), relative humidity (RH), station pressure (P), airmass (AM), and temperature (T). The daily observation data for these 11 potential factors in 2019 were selected when the solar zenith angle was smaller than 90\(^\circ\), and the 15 min averages of these observations comprised a sample in order to match the requirement of a very short-term forecast (i.e., 15 min ahead) of power dispatch. After transformation, there were 17,348 15 min samples during the daytime obtained for the whole of 2019. Eighty percent of these samples were randomly selected as the training set, 10% of them constituted the validation set, and the remaining samples were used as a testing set to evaluate the performance of the forecast models.

2.1.2. Evaluation Index

The determinant coefficient (\(R^2\)), the normalized mean bias error (nMBE), the normalized mean absolute error (nMAE), and the normalized root mean squared error (nRMSE) are used to evaluate the performance of different forecast models. They are calculated as follows:

\[
R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})(y_i - \bar{y})^2}{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2} \\
nMBE = \frac{1}{N} \sum_{i=1}^{N} \frac{(\hat{y}_i - y_i)}{\bar{y}} \times 100\% \\
nMAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| / \bar{y} \times 100\% \\
nRMSE = \frac{1}{\bar{y}} \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2} \times 100\% 
\]

where \(\hat{y}_i\) is the predicted value of a forecast model, \(\bar{y}\) is the mean of all predicted values, \(y_i\) is the target, \(\bar{y}\) is the mean of all targets, and \(N\) is the number of testing samples.

2.2. Effectiveness of Factors on GHI

2.2.1. Effectiveness of a Single Factor

The Pearson correlation coefficient (\(r\)) and the Spearman’s rank correlation coefficient (\(\rho\)) [32] are commonly used to measure the correlation between two variables. They are calculated as

\[
r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} (y_i - \bar{y})^2}} \\
\rho = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}
\]

where \(x\) and \(y\) are the two variables of interest, \(\bar{x}\) and \(\bar{y}\) are their average values, respectively, \(N\) is the number of data pairs, and \(d_i\) is the rank difference between \(x_i\) and \(y_i\).

In this paper, the effectiveness of each of the 11 factors on GHI is estimated by calculating the correlation coefficient. When the correlation coefficient between a factor and GHI is positive, this implies a positive effect on GHI and vice versa. Moreover, the greater \(r\) is, the stronger the correlation between the specific factor and GHI becomes.
2.2.2. Effectiveness of Multiple Factors on GHI

The correlation coefficient only describes the linear relationship between GHI and any of 11 factors, but neglects the effectiveness between two or more factors and GHI. The structural equation model (SEM) is a multivariate statistical method used to explain the causal relationships between variables [33]. Compared with traditional multivariate statistical methods, SEM allows the measurement of variables with errors, introduces unmeasurable latent variables, and uses a path graph model, which provides the possibility of analyzing the relationships between explicit variables and latent variables, as well as the relationships between different latent variables [34,35]. The relationships between different variables are defined as follows:

$$X = \Lambda_x \xi + \delta$$  \hspace{1cm} (8)
$$Y = \Lambda_y \eta + \epsilon$$  \hspace{1cm} (9)
$$\eta = A\eta + B\xi + \zeta$$  \hspace{1cm} (10)

where $\xi$ represents exogenous latent variables, $X$ represents observational variables of exogenous latent variables, $\Lambda_x$ is the factor loading matrix representing the relationship between the observational variable and the unobserved exogenous latent variable, and $\delta$ is the error. $\eta$ represents endogenous latent variables, $Y$ represents observational variables of endogenous latent variables, $\Lambda_y$ is a factor loading matrix representing the relationship between the observation variable and the unobserved endogenous latent variable, and $\epsilon$ is the error. $A$ and $B$ are path coefficient matrixes, and $\zeta$ is the bias of the SEM.

The analysis results of the SEM can be directly represented by the path graph. In a path graph, observational variables are represented in a rectangular box and latent variables are represented in a circular box; a line with an arrowhead is used to connect the two variables and represents a causal link between them. Usually, the path graph is constructed first, and the SEM is then used to check the validity of the path graph.

Therefore, a SEM was used to analyze the combined effectiveness of multiple factors on GHI in order to construct a more reasonable forecast model with appropriate input variables. There are four main specific steps of applying the SEM:

1. Establish the path graph model with GHI and various factors;
2. Calculate the path coefficients by SEM;
3. Evaluate the fitness of the path graph model;
4. Adjust the path graph model until it is suitable.

Finally, the interaction of multiple variables on GHI is shown in the path graph.

2.3. Ensemble Model for Inter-Hour Forecast of GHI

Considering the complexity and nonlinearity of GHI changes, accurate predictions of GHI are hard to achieve with a single common model. Therefore, an ensemble model was proposed for the inter-hour forecast of GHI based on the effectiveness of different factors mentioned above. The ensemble model was constructed with two parts, as shown in Figure 1: (1) three single forecast sub-models, named the primary model, and (2) a fusion model, which was used to automatically combine the predictions of the three sub-models. More specifically, Sub-Model 1 is a linear regression model considering the linear relationship between GHI and factors based on the individual effectiveness of a signal factor on GHI, Sub-Model 2 is a nonlinear model that considers factors with small correlation coefficients that have a nonlinear relationship with GHI, and Sub-Model 3 is a nonlinear model that considers the interaction of factors on GHI based on the analysis by SEM.

Sub-Model 1 is a multiple regression model according to the correlation analysis in terms of its linear aspect, defined as

$$GHI(t + 1) = a_0 + a_1 x_1(t) + \cdots + a_n x_n(t)$$  \hspace{1cm} (11)
where \( a_i \) are the parameters of the multiple regression model, which are calculated by the least squares method with the training set; \( x_i \) represents those factors with a greater correlation coefficient (absolute value >0.3) in Table 1 in Section 3.1.1, including Z, airmass, temperature, AOD, and relative humidity.

### Figure 1. Structure of the proposed ensemble model for global horizontal irradiance (GHI) forecast.

### Table 1. Correlation coefficients for GHI and different factors.

<table>
<thead>
<tr>
<th>Factors</th>
<th>( r )</th>
<th>( \rho )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Gamma )</td>
<td>0.19</td>
<td>0.16</td>
</tr>
<tr>
<td>Z</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>A</td>
<td>(-0.11)</td>
<td>(-0.11)</td>
</tr>
<tr>
<td>CC</td>
<td>(-0.28)</td>
<td>(-0.25)</td>
</tr>
<tr>
<td>T</td>
<td>0.37</td>
<td>0.36</td>
</tr>
<tr>
<td>AOD</td>
<td>(-0.36)</td>
<td>(-0.34)</td>
</tr>
<tr>
<td>RH</td>
<td>(-0.36)</td>
<td>(-0.34)</td>
</tr>
<tr>
<td>P</td>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td>AM</td>
<td>(-0.54)</td>
<td>(-0.75)</td>
</tr>
<tr>
<td>WS</td>
<td>0.13</td>
<td>0.17</td>
</tr>
<tr>
<td>WD</td>
<td>(-0.10)</td>
<td>(-0.0037)</td>
</tr>
</tbody>
</table>

For Sub-Model 2, the SVM, a common machine learning model, is selected to describe the nonlinear relationship between GHI and other factors, and its structure has been introduced in detail in [36]. Here, the inputs to the SVM are factors with small correlation coefficients (absolute value <0.5 and >0.1) in Table 1, except \( \cos \theta_z \), airmass, and wind direction. The kernel function of the SVM is set as the radial basis function, and its model parameters are obtained by the cross-validation method [37].

Sub-Model 3, SEM-MLP, is an Artificial Neural Network model (multilayer perceptron, MLP) with two hidden layers, but its structure is different from that of the common MLP whose input layer and the first hidden layer are in full connection. In this paper, the structure of the SEM-MLP is constructed based on the SEM as shown in Figure 2. Specifically, its inputs are 10 factors except wind direction (Table 1); the link between the input layer and the first hidden layer is based on the analysis of the path graph model, where the hidden nodes of the first hidden layer are set as latent variables of the SEM as described in Section 2.2.2.
11 factors. The GHI has the greatest positive correlation with solar zenith angle. Airmass is the second with other forecast models in Section 3.3.

The correlation coefficients between GHI and the 11 factors collected in Section 2.1.1 were calculated and are listed in Table 1. The signs of \( r \) and \( \rho \) are the same and their values are similar for all 11 factors. The GHI has the greatest positive correlation with solar zenith angle. Airmass is the second most important factor influencing GHI. Cloud cover (CC), temperature (T), and relative humidity (RH) all have a similar effect on GHI. The correlation coefficients between wind direction (WD) and GHI are different orders of magnitude. For example, cloud cover ranges from 0 to 1, while the values of temperature are from \(-20\) to \(36\) °C in the training set. Therefore, all input variables are normalized to the range from \(-1\) to \(1\) as follows:

\[
X' = 2 \times \frac{X - \min(X)}{\max(X) - \min(X)} - 1
\]

where \( X \) represents the value of one input variable, and \( X' \) is the normalized input variable; \( \min(X) \) and \( \max(X) \) are the minimum and maximum of the input variable \( X \) in the training set.

Although the average weighted method is a common method for combining the results of several sub-models of an ensemble model, it is hard to achieve the most satisfactory performance of the ensemble model because of lack of theoretical guidance. One study has demonstrated that the data-driven method, in place of the average weighted method, can achieve higher accuracy for the ensemble model \([38]\). Therefore, a fusion model is constructed to automatically weight the three sub-models based on the training set in order to achieve higher prediction accuracy. Here, the fusion model is a common three-layer MLP model with one input layer, one hidden layer, and one output layer \([39]\). The sigmoid function is selected as the activation function of the hidden layer, and the linear sum function is for the output layer.

3. Results and Discussions

In this section, the first group of experiments was carried out to evaluate the effectiveness of individual factors on GHI and the prediction accuracy in Section 3.1. The second one was constructed to assess the interaction of multiple factors on GHI based on the SEM in Section 3.2. Finally, the third group of experiments was implemented to compare the performance of the proposed ensemble model with other forecast models in Section 3.3.

3.1. Effectiveness of a Single Factor

3.1.1. Effectiveness of a Single Factor on GHI

The correlation coefficients between GHI and the 11 factors collected in Section 2.1.1 were calculated and are listed in Table 1. The signs of \( r \) and \( \rho \) are the same and their values are similar for all 11 factors. The GHI has the greatest positive correlation with solar zenith angle. Airmass is the second most important factor influencing GHI. Cloud cover (CC), temperature (T), and relative humidity (RH) all have a similar effect on GHI. The correlation coefficients between wind direction (WD) and GHI are...
the smallest, so it can be considered that the influence of wind direction on GHI is minimal and can be ignored. On the other hand, the attenuation of GHI is obviously affected by cloud cover, aerosols, relative humidity and airmass because of the negative sign of their correlation coefficients.

Figure 3 shows the correlation coefficients between every two different variables including GHI. The off-diagonal elements show the degree of dependency for every two variables. The temperature (T) has a strong positive correlation with the sine function of DOY (Γ), and the aerosols (AOD) have a strong positive correlation with the cloud cover (CC). The airmass (AM) has a negative correlation with the cosine function of the solar zenith angle (Z) indicating that when Z increases, AM decreases, which is in agreement with the two variables having strong opposite correlation with GHI in Table 1. However, there is a nonlinear relationship between AM [17] and solar zenith angle, and the AM is not calculated accurately by solar zenith angle, so these two variables are both considered in the forecast model to achieve higher predicting performance in this paper. In addition, when some of those factors are unavailable, these results provide guidance for selecting and ranking factors for constructing forecast models of inter-hour solar radiation.

![Figure 3](image_url)

**Figure 3.** Correlation coefficients for every two different variables. Z is the cosine function of the solar zenith angle, A is the solar azimuth angle, CC is cloud cover, T is temperature, AOD is aerosol optical depth, RH is relative humidity, P is station pressure, AM is airmass, WS is wind speed, and WD is wind direction. (a) Pearson correlation coefficient. (b) Spearman’s rank correlation coefficient.
3.1.2. Influence of a Signal Factor on Forecast Model Performance

A regression model with one factor (1-RM) was used to evaluate the influence of a signal factor on the prediction accuracy. The 1-RM model is a simplified version of Sub-Model 1, with a single input $x_1$ in Equation (11). Table 2 lists the results of the 1-RM model with different factors to predict GHI 15 min ahead. The $R^2$ of each 1-RM model is in agreement with the $r$ of GHI and the input factor in Table 1. For example, the absolute value of $r$ of GHI and $Z$ is the greatest among the 11 factors; the 1-RM model with $Z$ achieves the highest forecasting performance among these 11 models, which means that the $\theta_z$ is the most important factor for GHI. AM has a strong correlation with $Z$ and GHI in Figure 3, so that the 1-RM model with AM achieves the second greatest predicting performance. Although $Z$, $A$ and $\Gamma$ can be calculated by day and hour (time) with station location information [40,41], the correlation coefficients between them are small in Figure 3 and the predicted performance of the 1-RM model with each of them is different. Therefore, the three factors can be viewed as the transformational variables from time and describe the effectiveness of time on GHI from different points, and they are all considered in the proposed model. On the other hand, the $nMBEs$ of each 1-RM model are all small, but the $nMAEs$ and $nRMSEs$ are greater. Therefore, the 1-RM model does not meet the requirement of PV applications, even with the most relevant factor of $Z$. Note that the results here are analyzed in terms of linearity and the contribution of each factor may be different to GHI in terms of nonlinearity.

Table 2. Performance of regression model with 1-RM to predict GHI.

<table>
<thead>
<tr>
<th>Factors</th>
<th>$R^2$</th>
<th>$nMBE$ (%)</th>
<th>$nMAE$ (%)</th>
<th>$nRMSE$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>0.5231</td>
<td>0.62</td>
<td>40.22</td>
<td>51.13</td>
</tr>
<tr>
<td>AM</td>
<td>0.3573</td>
<td>3.08</td>
<td>49.92</td>
<td>60.82</td>
</tr>
<tr>
<td>T</td>
<td>0.1347</td>
<td>1.53</td>
<td>59.01</td>
<td>68.90</td>
</tr>
<tr>
<td>AOD</td>
<td>0.1219</td>
<td>2.37</td>
<td>57.70</td>
<td>69.44</td>
</tr>
<tr>
<td>RH</td>
<td>0.1041</td>
<td>1.66</td>
<td>58.70</td>
<td>70.11</td>
</tr>
<tr>
<td>CC</td>
<td>0.0738</td>
<td>1.77</td>
<td>59.41</td>
<td>71.31</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td>0.0391</td>
<td>1.60</td>
<td>62.87</td>
<td>72.59</td>
</tr>
<tr>
<td>P</td>
<td>0.0381</td>
<td>1.67</td>
<td>62.04</td>
<td>72.65</td>
</tr>
<tr>
<td>A</td>
<td>0.0263</td>
<td>1.84</td>
<td>62.33</td>
<td>73.08</td>
</tr>
<tr>
<td>WS</td>
<td>0.0079</td>
<td>1.98</td>
<td>63.32</td>
<td>73.77</td>
</tr>
<tr>
<td>WD</td>
<td>0.0065</td>
<td>2.19</td>
<td>63.16</td>
<td>73.82</td>
</tr>
</tbody>
</table>

3.2. Effectiveness of Multiple Factors

3.2.1. Interaction of Multiple Factors on GHI

In order to evaluate the interaction of multiple factors on GHI quantitatively, a path graph model between GHI and 10 factors, excluding wind direction, was constructed as shown in Figure 4. Air, Season, Weather, cloud cover (CC), aerosol optical depth (AOD), and solar position (Sun) were set as the 6 exogenous latent variables. For the latent Air variable, the pressure and airmass were its two observational variables; $Z$ and $A$ were two observational variables of the latent solar position variable; and temperature ($T$), relative humidity ($RH$), and wind speed ($WS$) were the three observational variables of the latent Weather variable. Partial least squares (PLS) was used to calculate the parameters of the SEM, and all path coefficients are shown in the diagram. The Sun and AOD have higher path coefficients, which means they make a greater contribution to GHI.

The comprehensive path coefficient of each factor to GHI was calculated based on the paths from the factor to GHI as shown in Figure 4, and the results are listed in Table 3. For example, there are three paths from $Z$ to GHI, so the comprehensive path coefficient of $Z$ to GHI is calculated as $0.999 \times (-0.736) \times 0.011 + 0.999 \times 0.968 + 0.999 \times 0.155 \times (-0.635) \approx 0.861$. In terms of path coefficients, the solar zenith angle, AOD, and cloud cover are the main factors influencing GHI, which is different from the results of $r$ as shown in Figure 5. That may be caused by the interaction
between these factors. The mechanisms of aerosols and cloud are similar, so the path coefficients of AOD and CC are in the same order of difference between the two methods.

![Path graph model](image)

**Figure 4.** Path graph model between GHI and 10 factors. P is station pressure, AM is airmass, Z is the cosine function of the solar zenith angle, A is the solar azimuth angle, AOD is aerosol optical depth, T is temperature, RH is relative humidity, WS is wind speed, and CC is cloud cover. The number on each line is the normalized path coefficient of each variable.

**Table 3.** Comprehensive path coefficients of 10 factors relevant to GHI.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Comprehensive Path Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>0.861</td>
</tr>
<tr>
<td>AOD</td>
<td>-0.635</td>
</tr>
<tr>
<td>CC</td>
<td>-0.462</td>
</tr>
<tr>
<td>A</td>
<td>-0.035</td>
</tr>
<tr>
<td>RH</td>
<td>0.035</td>
</tr>
<tr>
<td>T</td>
<td>-0.033</td>
</tr>
<tr>
<td>WS</td>
<td>-0.011</td>
</tr>
<tr>
<td>AM</td>
<td>0.011</td>
</tr>
<tr>
<td>P</td>
<td>-0.003</td>
</tr>
<tr>
<td>Γ</td>
<td>0.002</td>
</tr>
</tbody>
</table>

![Histogram of absolute values](image)

**Figure 5.** Histogram of absolute values of $r$ and comprehensive path coefficients of 10 factors to GHI.

### 3.2.2. Influence of Multiple Factors on Forecast Model Performance

In order to evaluate the interaction of multiple factors on forecast model performance, six RMs and MLP models were constructed for predicting GHI based on six exogenous latent variables shown in Figure 4. For example, the pressure and airmass were set as inputs to the RM and MLP models.
for the Air exogenous latent variable from the perspectives of linearity and nonlinearity. The $nMAE$ and $nRMSE$ of these RM and MLP models are shown in Figure 6. It is clear that the changes in the performance of the RMs are coincident with those of the MLP models, and that both of them achieve the greatest prediction accuracy on the exogenous latent variable of Sun in agreement with the highest path coefficient in Figure 4. On the other hand, the nonlinear MLP models do a little better than the linear RM models, especially on the exogenous latent variable of Air, which means that there is a significant nonlinear relationship between GHI and the two factors (airmass and pressure).

![Performance of Regression Model (RM) and MLP models with different observation variables of exogenous latent variables based on Figure 4.](image)

Figure 6. Performance of Regression Model (RM) and MLP models with different observation variables of exogenous latent variables based on Figure 4.

On the other hand, the performances of the sub-models and the ensemble model are compared to further analyze the importance of the 10 factors to the predicted accuracy, and the results are listed in Table 4. Sub-Model 2 is outperformed by both the linear Sub-Model 1 and the nonlinear Sub-Model 3, because the most important factors of Z and AM are not considered in Sub-Model 2. Comparing the results of the three sub-models, it can be concluded that the influence of the key factor on the prediction accuracy may be greater than that of the forecasting model. In addition, the nonlinear Sub-Model 3 outperforms the linear Sub-Model 1, and the ensemble model outperforms all sub-models, which means that the construction of a reasonable structure for a forecast model can improve the prediction accuracy to some extent.

<table>
<thead>
<tr>
<th>Models</th>
<th>$R^2$</th>
<th>$nMBE$ (%)</th>
<th>$nMAE$ (%)</th>
<th>$nRMSE$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Model 1</td>
<td>0.8733</td>
<td>1.08</td>
<td>18.31</td>
<td>26.39</td>
</tr>
<tr>
<td>Sub-Model 2</td>
<td>0.8429</td>
<td>0.66</td>
<td>19.52</td>
<td>29.41</td>
</tr>
<tr>
<td>Sub-Model 3</td>
<td>0.8944</td>
<td>0.01</td>
<td>13.69</td>
<td>24.06</td>
</tr>
<tr>
<td>Ensemble model</td>
<td>0.9146</td>
<td>−0.10</td>
<td>11.13</td>
<td>21.98</td>
</tr>
</tbody>
</table>

1 Sub-Model 1 is an RM. 2 Sub-Model 2 is a Support Vector Machine (SVM) model. 3 Sub-Model 3 is a constructed SEM-MLP model as in Figure 2.

3.3. Comparing the Performance of Different Forecast Models

Another group of experiments was carried out to evaluate the performance of the proposed ensemble model in comparison with another published method, and the results are listed in Table 5. The persistent model is usually used as a benchmark for evaluating the performance of different forecast models, and is defined as

$$GHI(t + 1) = GHI(t)$$

(13)
The evaluating index based on the persistent model, which is called the forecast skill ($Fs$), is defined as

$$Fs = \frac{nRMSE_{per} - nRMSE_f}{nRMSE_{per}} \times 100\%$$

(14)

where $nRMSE_{per}$ and $nRMSE_f$ are the nRMSEs of the persistent model and the forecast model, respectively. The other four forecast models in Table 5 have the same inputs (10 factors in Table 3). It is noteworthy that the multi-RM and SVM models in Table 5 have similar model structure but different inputs of Sub-Model 1 and Sub-Model 2 in Table 4, respectively. What is more, the RM and SVM achieved higher prediction accuracy in Table 5 by considering more factors; this proves that adding inputs (factors), especially key factors, can improve the performance of forecast models. On the other hand, the common MLP model with full connection between the input layer and hidden layer was slightly outperformed by Sub-Model 3 (a SEM-MLP with 2 hidden layers) with the same inputs, which means that considering the interaction between factors and forecast model can improve the prediction accuracy.

Table 5. The performance of different models for predicting inter-hour GHI on the testing set.

<table>
<thead>
<tr>
<th>Models</th>
<th>$R^2$</th>
<th>nMBE (%)</th>
<th>nMAE (%)</th>
<th>nRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistent</td>
<td>0.8880</td>
<td>0.27</td>
<td>15.37</td>
<td>25.14</td>
</tr>
<tr>
<td>Multi-RM</td>
<td>0.8830</td>
<td>0.98</td>
<td>16.85</td>
<td>25.37</td>
</tr>
<tr>
<td>SVM</td>
<td>0.8930</td>
<td>0.23</td>
<td>13.89</td>
<td>24.28</td>
</tr>
<tr>
<td>MLP $^1$</td>
<td>0.8935</td>
<td>0.45</td>
<td>14.15</td>
<td>24.16</td>
</tr>
<tr>
<td>Ensemble model</td>
<td>0.9146</td>
<td>−0.10</td>
<td>11.13</td>
<td>21.98</td>
</tr>
</tbody>
</table>

$^1$ MLP is an Artificial Neural Network (ANN) model with 2 hidden layers whose input nodes are fully connected to the first hidden layer, which is different from the SEM-MLP model in Table 4.

Figure 7 shows scatter plots holding the target (measured GHI) in the testing set and the GHI predicted 15 min ahead by the five forecast models in Table 5. The distribution of the predicted values by the persistent model is the most loose. The predicted values of the Multi-RM and MLP are smaller than the measured values in the large value area (greater than 800 Wm$^{-2}$), although the nMBEs of the two models are greater than zero. The predicted values of the proposed model are closer to the line with the slope of 1 than those of the other four models.

![Figure 7](image_url)

**Figure 7.** Scatter plots between the predicted GHIs and the measured GHIs in the testing set.
Figure 8 shows the error distributions of the five models in Table 5, where the red line in the blue box is the median value of the errors for each model and the top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers, with 95% confidence, extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the “+” symbol. It is clear that the maximum absolute values of the errors by the ensemble model are much smaller than those of the other four models. That is why the ensemble model achieves the smallest \( nRMSE \) in Table 5, although the top edges of its box are not the closest.

![Figure 8. Predicted error distributions of the 5 models in Table 5.](image)

The \( Fs \) of the four different forecast models in Table 5 were calculated and are shown in Figure 9. In terms of the \( nRMSE \), the proposed ensemble model shows an improvement of about 3% compared with the persistent model and does much better than the other three forecast models. From these results, the selection of key factors and the construction of a suitable model are both important for achieving high prediction accuracy; the former is more important than the latter to some extent.

![Figure 9. \( Fs \) of the last 4 models in Table 5 compared with the persistent model.](image)

4. Conclusions

In this paper, the effects of 11 factors on solar radiation were discussed and an ensemble model was constructed for predicting inter-hour GHI. Firstly, the individual effect of potential factors on GHI was estimated by correlation analysis, and the combined effects of potential factors were estimated quantitatively using a SEM. The results showed that solar zenith angle, cloud cover, aerosols, and airmass have greater effects on GHI than other factors, such as the day of the year, solar azimuth...
angle, relative humidity, temperature, wind speed, wind direction, and station pressure, where the effectiveness of wind direction was very small and almost negligible.

Secondly, the influence of these factors on the prediction accuracy of forecast models was compared by constructing different forecast models with different inputs. The results showed that the main factors with large correlation coefficients can partly represent or predict GHI, such as in Sub-Model 1 in Table 4. Factors with small correlation coefficients were also important for predicting GHI and improved the forecast performances, and considering the interaction of different factors further improved the prediction accuracy. What is more, the key factor selection was more important than the model structure for predicting GHI precisely.

Thirdly, an ensemble model was developed to predict inter-hour solar radiation considering the single and multivariate effectiveness on GHI. The results showed that the proposed ensemble model could achieve higher prediction accuracy than single sub-models and outperformed traditional data-driven methods; there was about a 3% improvement over the persistent model in terms of nRMSE.

This paper only provides a method to preliminarily discuss the effects of potential factors on GHI based on measurements of a surface station and construct a forecasting ensemble model. It is still necessary to check the accuracy and performance of the proposed model in real conditions using a field experiment in the future. What is more, the relationship between GHI and these factors may be different and the degree of the factors’ effectiveness would also change in different stations, which is another point of concern.

Next, the estimations of the effectiveness of some factors are inaccurate in this paper, such as cloud, because of the incomplete description of cloud. In future, data fusion technology could be used to represent a potential factor more precisely using measurements of several types of data. For example, sky images, such as total sky image or all sky image, could be used to extract image features to represent clouds.

On the other hand, the structure of the path graph model can be adjusted to analyze the interaction of different combinations on GHI in order to improve the prediction performance. In addition, considering the order of the inputs using historically measured data could also improve the prediction accuracy.

**Author Contributions:** T.Z. and Y.G. designed this work, and they contributed equally to this work; Y.G. and C.W. carried out the experiments and validation of this work; T.Z. wrote the original draft; C.N. reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Key Research and Development Program of China, grant number 2018YFB1500803, and the Key Laboratory of Measurement and Control of Complex Systems of Engineering (Southeast University), Ministry of Education, grant number MCCSE2020A02.

**Acknowledgments:** The authors acknowledge the National Renewable Energy Laboratory for providing the data used in this paper. We are also grateful to Dongyi Wang and reviewers for their recommendations to improve the quality of this paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Abbreviations**

Abbreviations in this manuscript are summarized as follows:

- **A**: Solar azimuth angle
- **AM**: Airmass
- **ANN**: Artificial Neural Network
- **AOD**: Aerosol
- **CC**: Cloud cover
- **DOY**: The day of the year
- **Fs**: Forecast skill
- **GHI**: Global horizontal irradiance
- **MLP**: Multilayer perceptron
NREL  National Renewable Energy Laboratory  
P  Station pressure  
PV  Photovoltaic  
$R^2$  Determinant coefficients  
RH  Relative humidity  
RM  Regression Model  
SEM  Structural equation model  
SRRL  Solar Radiation Research Laboratory  
SVM  Support Vector Machine  
T  Temperature  
WD  Wind direction  
WS  Wind speed  
$Z$  The cosine function of solar zenith angle  
$nMBE$  The normalized mean bias error  
$nMAE$  The normalized mean absolute error  
$nRMSE$  The normalized root mean squared error  
$\Gamma$  The sine function of DOY  
$r$  Pearson correlation coefficient  
$\rho$  Spearman’s rank correlation coefficient  
$\theta_z$  Solar zenith angle  

References


15. Wang, Z.; Tian, C.; Zhu, Q.; Huang, M. Hourly solar radiation forecasting using a volterra-least squares support vector machine model combined with signal decomposition. Energies 2018, 11, 68. [CrossRef]


27. Boland, J. Characterising seasonality of solar radiation and solar farm output. Energies 2020, 13, 471. [CrossRef]


38. Kim, K.; Hur, J. Weighting factor selection of the ensemble model for improving forecast accuracy of photovoltaic generating resources. Energies 2019, 12, 3315. [CrossRef]


© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).