Supplementary Materials

1. Results of data pre-processing

1) Results of Feature analysis

We employ the Pearson's correlation coefficient to analyze the correlation and impact of particular independent variables on the generated PV power. The Pearson's correlation coefficient $r_{x,y}$ between variables x and y is determined as follows:

$$r_{x,y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{cov(x,y)}{S_x S_y} \quad \text{(S1)}$$

where $\bar{x}$ and $\bar{y}$ are the means of variable x and y, S_x and S_y are the standard deviations of x and y, respectively, and n is the sample size of a particular variable.

Moreover, we pay attention to the kurtosis and skewness of every feature. The kurtosis η describes the distribution of observed data in relation to normal distribution (whose kurtosis equals 0), and the skewness ξ describes the level of asymmetry of the distribution around the average value. They are defined as:

$$\eta = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_x}\right)^4\right] - \frac{3(n-1)^2}{(n-2)(n-3)} \quad \text{(S2)}$$

$$\xi = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_x}\right)^3 \quad \text{(S3)}$$

Fig. S1 shows the correlation matrix as the form of heatmap where the stronger linear relation between a pair of variables, the greater the absolute value in the matrix, and the deeper the color of the map square. Considering the relation between p and other features, we pay attention to the first row or first column and draw a conclusion that it is r and h that have the strongest relation to p. The Pearson's correlation coefficient between them are 0.87 and -0.54 respectively. In addition, the correlation coefficient between p and t is 0.31, and other features virtually contribute little to p.

According the above results, in terms of the analysis of kurtosis and skewness, we consider p and its two strong influence factors, r and h. The results are presented in Table S1. First, the kurtosis of p, r and h are greater than 0, meaning that

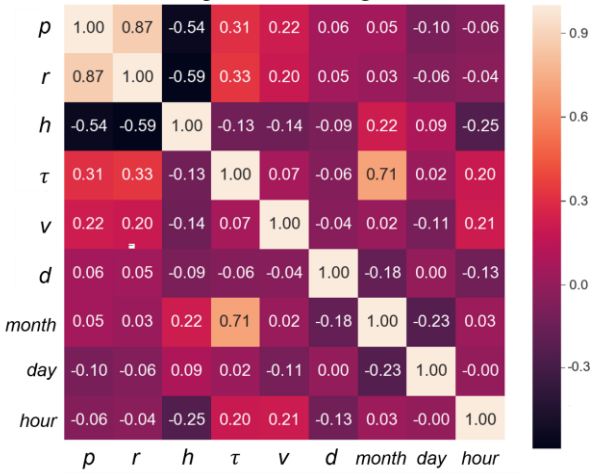


Figure S1. The heatmap of feature correlation matrix

their distributions are steeper and sharper than normal distribution. As for skewness, all of them have positive values, indicating that they have denser data whose values are lower than their mean. In a word, the distribution of PV data is complicated to be described with a probability model. Hence, we turn to a tree-based algorithm which is free of the dependency on probability.

Table S1
KURTOSIS AND SKEWNESS VALUE OF ORIGINAL DATA

Data	*p*	*r*	*h*
Kurtosis value	0.618	0.969	0.574
Skewness value	1.26	1.401	0.811

Examples of the collected data are shown in Table S2.

2) Results of Data Pre-classification

To get A dataset, we have to take maintenance records into consideration. The scheduled maintenance plan of the studied PV plant is presented in Table S3.

Table S2.
EXAMPLE OF THE COLLECTED DATA FROM A PV PLANT

p (KWh/15min)	*r* (W/㎡)	*τ* (℃)	*h* (%)	*v* (m/s)	*d* (0°-360°)	*T*(Time)
2	52.9	2.4	94.6	0	56.2	1/1 7:03
68.07	85.1	3.2	94.5	0	115.3	1/1 7:18
73.7	118.8	4.3	93.7	0	320.6	1/1 7:33
……	……	……	……	……	……	……
821.93	805.8	29.1	76.3	3	140.6	7/16 9:22
824.7	786.8	29.6	75.3	2.4	118.1	7/16 9:37
769.88	896.6	29.8	74	2.8	140.6	7/16 9:52

Table S3.
SCHEDULED PLAN OF MANUAL CLEANING

Month	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
Clean Period	1.16-1.29	2.16-2.28	3.16-3.29	4.16-4.28	5.16-5.30	6.16-6.30
Month	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Clean Period	7.16-7.30	8.16-8.30	9.16-9.30	10.16-10.30	11.16-11.30	12.16-12.30

Table S4.
WEATHER FROM 7/1 TO 7/15

Date	Weather	Date	Weather
7/1	Light rain / Overcast	7/9	Cloudy
7/2	Overcast	7/10	Cloudy
7/3	Rain Shower / Cloudy	7/11	Overcast
7/4	Light rain	7/12	Cloudy
7/5	Moderate rain	7/13	Cloudy
7/6	Light rain / Overcast	7/14	Sunny / Cloudy
7/7	Light rain	7/15	Cloudy
7/8	Overcast / Cloudy	/	/

To obtain valid B dataset, we should collect data originated from continuous sunny days. So, we check the historical weather online. Take July as an example, the weather information from 7/1 to 7/15 is listed in Table S4.

2. Results of non-continuous regression model

1) Results of clustering PV power data

We apply k-means clustering on PV power data, and then show the results in the *p*-*r* scatter plot in Fig. S2. It is evident that the power data are noncontinuous, and are clustered into discrete classes with gaps among them. As mentioned before, the gaps are caused by current-limiting. And here in the studied PV plant, its special transmission mechanism
between sensors and storage system also contributes to this noncontinuity.

In Fig. S2, each cluster is shown in different colors. Detailed statistical characteristics of every cluster are given in Table S5. We calculate the total amount (Count), the average *p* value (Mean), the maximum value (Max), and the minimum value (Min) of each cluster. In Table S5, near half of the *p* data belong to Class1 to Class4, and there are only a small amount of *p* data achieving high PV power generation. Table S5 gives a more direct demonstration that *p* data are not completely continuous, e.g., the Max of Cluster 6 is 363.52 but the Min of Cluster 7 is 397.60. Except for Cluster 1 and Cluster 2 (there are so many outliers among them that data mix and fill in the gap), every two adjacent classes have obvious blank gap between the former's Max and the latter's Min.

2) Results of Performance Evaluation, Fault Detection, and O&M Planning

As for performance evaluation, data observations located beyond the reference range (in Stages 1, 4, and 5), but at the beginning or end of the daytime, do not trigger warning. here is little sunshine at the beginning and end of a day, leading to unstable photoelectric conversion. Power generation in this period is also low and unstable, which is prone to false alarm. To avoid this, we carry out performance evaluation in the stable operation period, i.e.,7:00 to 18:00 in the summer and autumn, 8:00 to 15:00 in the spring and winter. What's more, single data that slightly deviates the baseline range does not trigger warning.

We find that the studied PV plant operates normally in most cases, as in Fig. S4 and Fig. S5. According to our fault detection method, there is no fault from 5/15 to 5/18 and from 7/9 to 7/12. Near all actual PV power data are in the expected baseline range. Accordingly, there is no need to implement O&M plan.

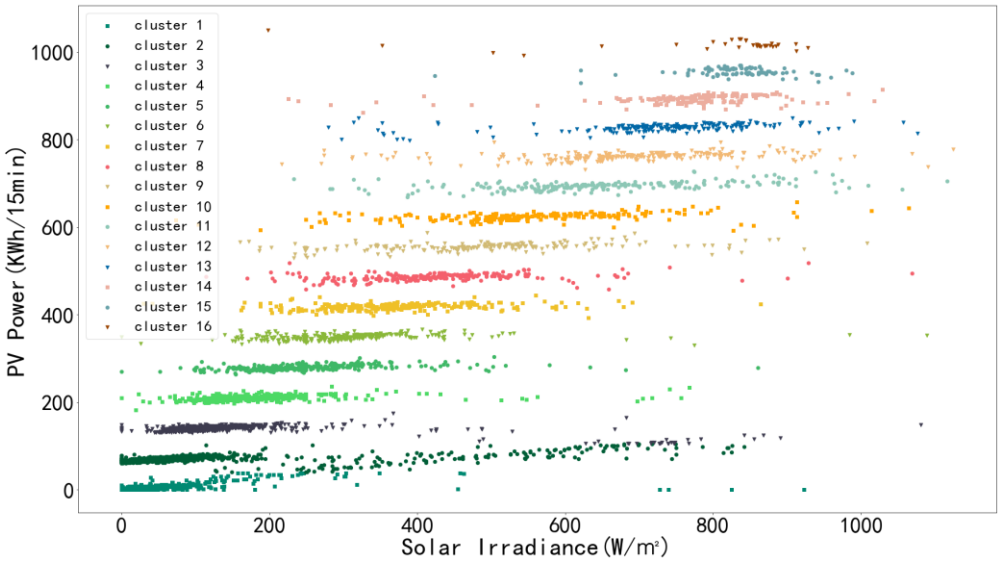
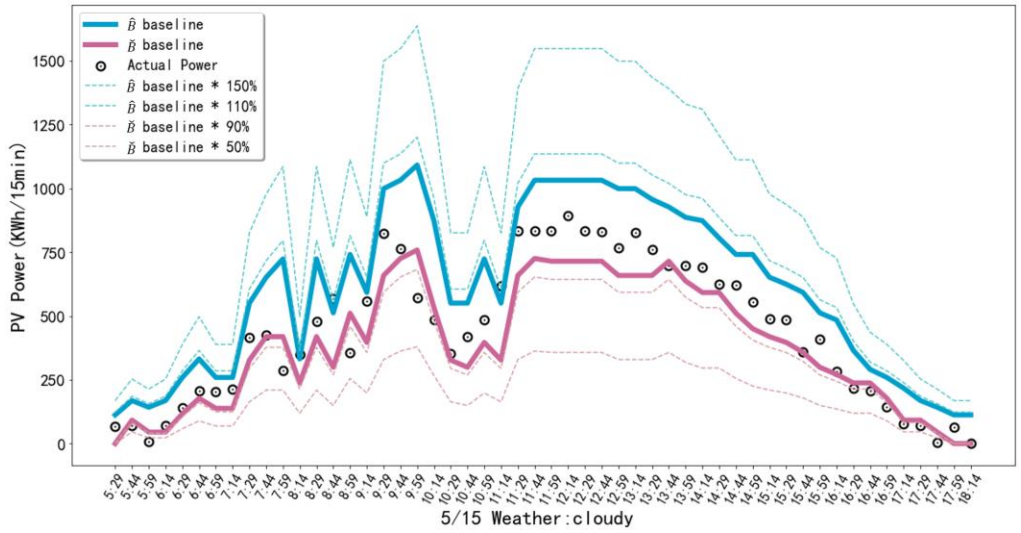


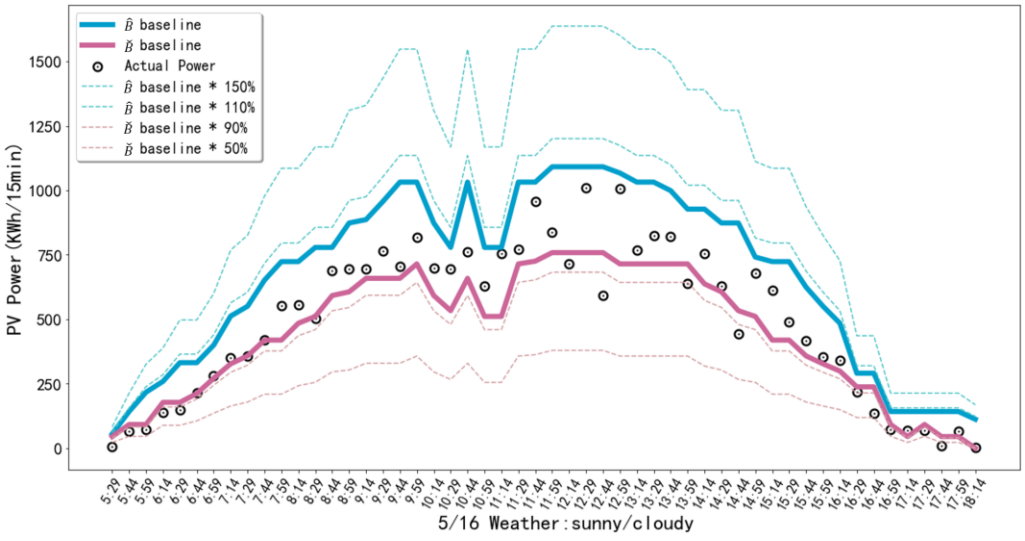
Figure S2. The k-means clustering result of the power data

Table S5.
CLUSTERING RESULT OF APPLYING K-MEANS TO PV POWER DATA

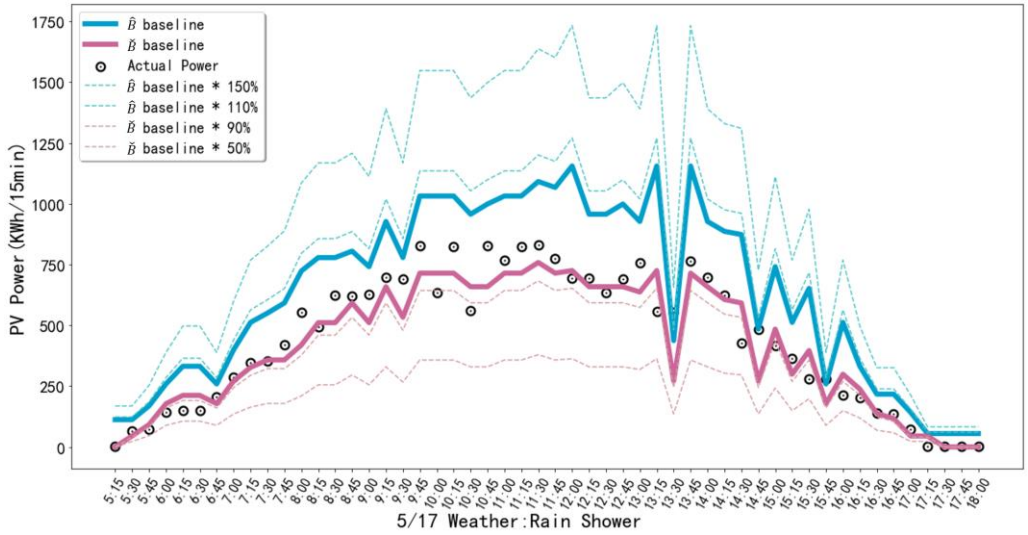
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Count	665	768	550	368	292	251	201	150
Mean	4.72	70.47	140.18	210.09	279.85	348.87	417.86	486.09
Min	0.00	50.10	129.34	197.41	264.26	332.23	397.60	465.68
Max	50.00	102.05	153.37	235.49	301.51	363.52	440.43	500.39
	Cluster 9	**Cluster 10**	**Cluster 11**	**Cluster 12**	**Cluster 13**	**Cluster 14**	**Cluster 15**	**Cluster 16**
Count	89	234	123	180	162	132	72	27
Mean	555.04	623.83	692.18	761.91	828.49	893.67	957.23	1018.46
Min	538.38	592.48	673.78	731.98	805.91	869.97	938.03	992.14
Max	567.72	657.76	707.86	793.99	842.77	908.06	970.16	1050.25



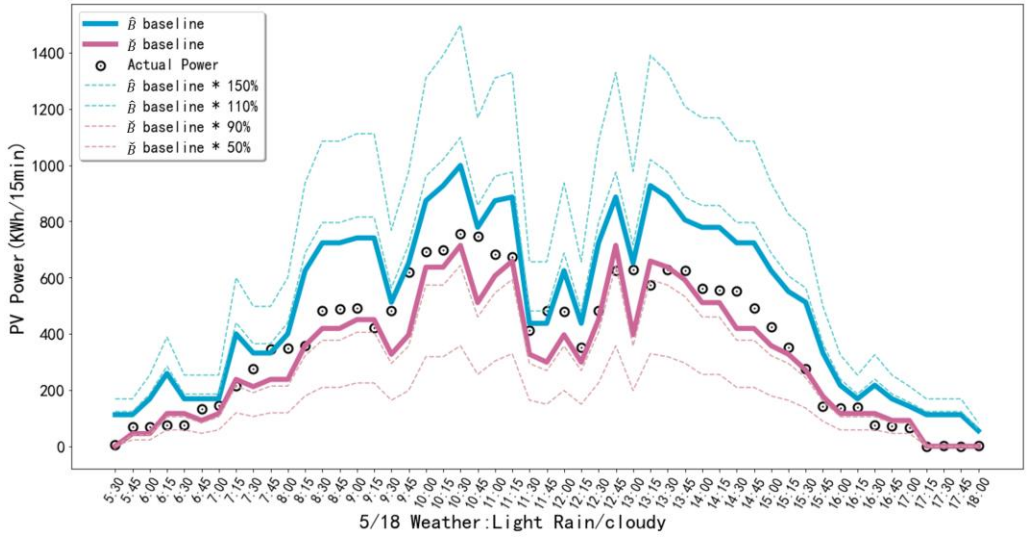
(a)



(b)

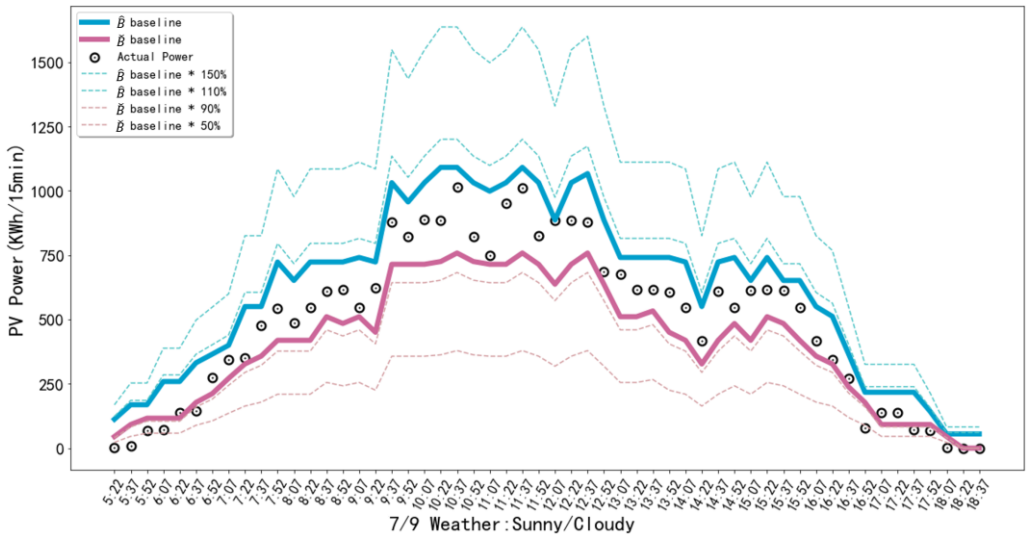


(c)

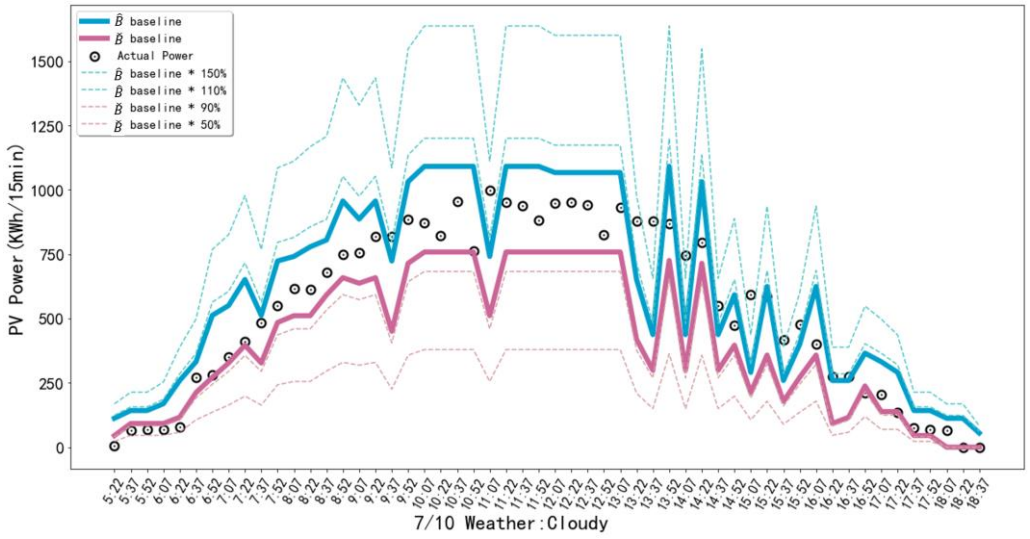


(d)

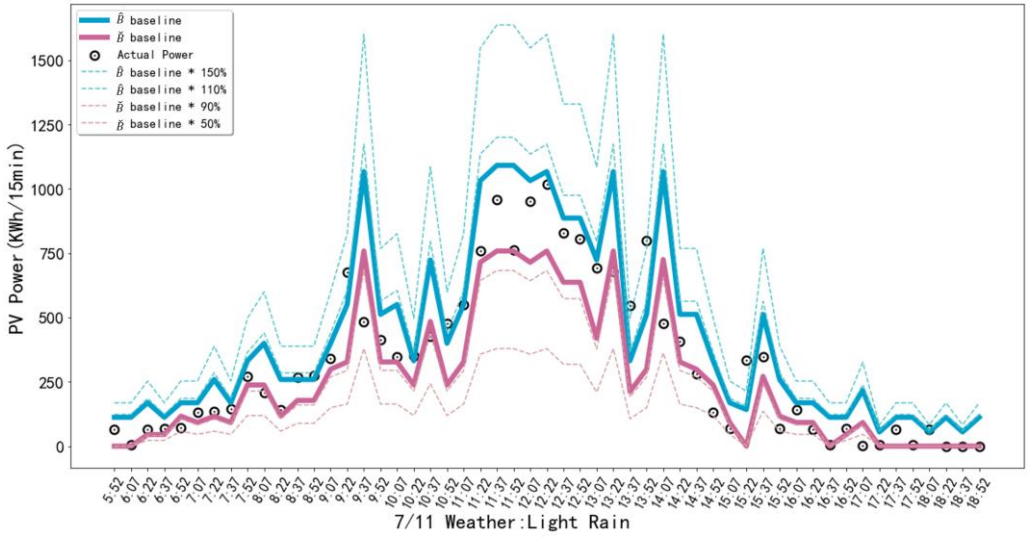
Figure S3 Performance Evaluation from 5/12 to 5/18



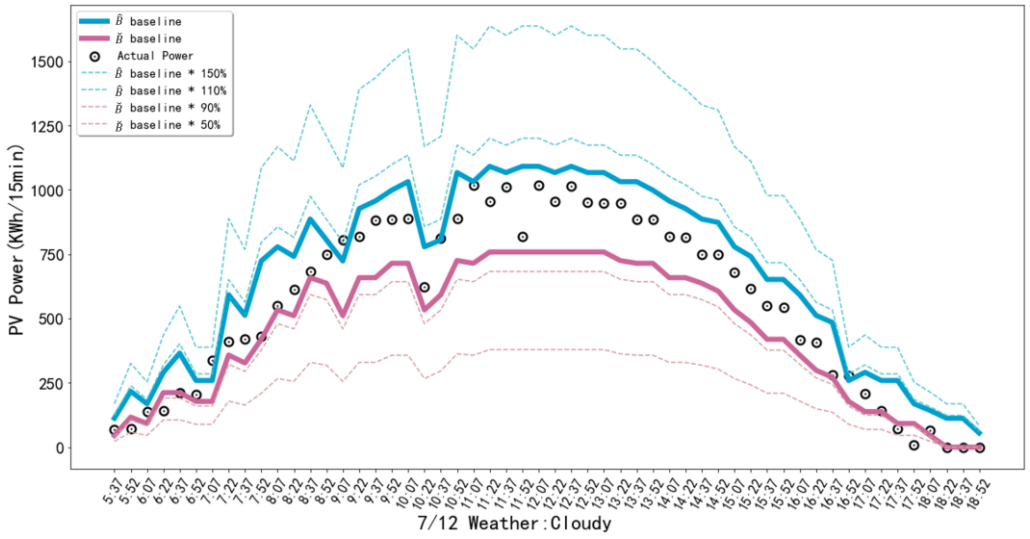
(a)



(b)



(c)



(d)

Figure S4 Performance Evaluation from 7/3 to 7/12