# Research on the error probability distribution of photovoltaic output prediction based on output fluctuation characteristics and Generalized Gaussian Mixture Model

Peng Yan[1*], Chenmeng Xiang[1], Wen Zhou[1] and Can Su[1]

[1]State Grid Hebei Electric Power Research Institute, Shijiazhuang, Hebei Province, 050021, China

**Abstract.** Photovoltaic power output forecast error exists objectively and inevitably, and it can provide a guarantee for safe and stable operation of the power system through analyzing its characteristics. In this paper, the influence of predicted output fluctuation characteristics (predicted output amplitude and power variation) on prediction error was studied based on the analysis of variance (ANOVA) method. The prediction error conditions were classified into six types based on the clustering of numerical characteristics of predicted output. Then, a Generalized Gaussian Mixture Model (GGMM) was proposed to fit the prediction error distribution of each type of photovoltaic output. The mean absolute error (MAE), coefficient of determination ($R^2$), and root mean square error (RMSE) were used as accuracy evaluation indexes. The example analysis showed that the GGMM can satisfy the asymmetry and kurtosis diversity of the error distribution after division by conditions, and the fitting result is better than that of the normal distribution, improved Laplace distribution and t Location-Scale distribution model.

## 1Introduction

Nowadays, the global energy crisis and environmental pollution are aggravating. In the face of this situation, accelerating the development of photovoltaic power generation technology as the representative of clean renewable energy has become the inevitable choice of countries around the world. Nevertheless, as large-scale photovoltaic connected to the power grid, its fluctuation and intermittency will lead to a significant increase in the uncertainty of power system operation, which undoubtedly brings new challenges to the scheduling operation of the power system. In response to this, photovoltaic output prediction is the main tool to reduce this problem. As a significant part of photovoltaic output prediction, research on the error distribution characteristics can provide statistical basis for uncertainty analysis of photovoltaic output prediction, which has a great significance to the safe and stable operation of the power system[1,2].

At present, there are only a few studies on the prediction errors of photovoltaic power generation in the literature. Moreover, among this small amount of literature, there is also some literature that describes the photovoltaic output prediction error based on the assumption that the error obeys a normal distribution[3,4]. The literature [5] argued that the normal distribution model assumptions are reasonable based on the characteristics that the expected value of the prediction error being zero and the prediction error distribution obeying the normal distribution. The literature [6] argued that the normal distribution could describe the variation pattern of photovoltaic generation prediction errors through statistical analysis, and it could be used as a random variable when studying the optimal scheduling problem of microgrid. The above assumptions lack strong theoretical proofs. It was found that it is more reasonable to use the t Location-Scale distribution model [7] than the normal distribution to describe the photovoltaic output prediction error. However, at the same time, the t Location-Scale distribution model has the problem of insufficient waist flexibility.

This paper clusters the prediction errors according to the corresponding predicted output magnitude and power variation conditions based on the clustering of the numerical characteristics of the predicted output. Then a generalized Gaussian mixture model is proposed to describe the distribution of photovoltaic power output prediction errors. The model can describe the error distribution of different kurtosis and shapes accurately. The prediction error clustering analysis method proposed in this paper can not be affected by the prediction algorithm and the geographic location of the photovoltaic plant, and its scope of application is more extensive.

*Corresponding author's e-mail: dyy_yanp@he.sgcc.com.cn

# 2Short-term photovoltaic power output prediction error model considering output fluctuation

## 2.1Effect of predicted output fluctuation on prediction error

By using the MPPT controller, the photovoltaic cell array makes the output power always maintain at the maximum power point. When the external environment such as light intensity and ambient temperature changes, the operating voltage of the cell will follow the shift, and the starting point, step, and direction of the shift may produce prediction errors, and the performance of the MPPT controller influences the magnitude of such errors. The correlation between such prediction errors and the influencing factors is investigated from a statistical point of view. The predicted output amplitude is chosen to represent the starting point of the movement, which is referred to as the E factor, and the magnitude of the adjacent predicted force difference is chosen to represent the power variation, the positive or negative of which indicates the direction of movement, which is referred to as the G factor. The E and G factors are arranged in ascending order and are set to 8 levels according to the principle of equally dividing sample size. The values of the two factors at each level are shown in Table 1, and the data in the table are standardized as taking the rated capacity as the reference value.;

**Table 1.** Values of E and G factors at each level.

| Level | Factor | |
|---|---|---|
| | E (Predicted output amplitude) | G (Power variation) |
| 1 | 0~0.1 | -0.06~-0.045 |
| 2 | 0.1~0.2 | -0.045~-0.03 |
| 3 | 0.2~0.3 | -0.03~-0.015 |
| 4 | 0.3~0.4 | -0.015~0 |
| 5 | 0.4~0.5 | 0~0.015 |
| 6 | 0.5~0.6 | 0.015~0.03 |
| 7 | 0.6~0.7 | 0.03~0.045 |
| 8 | 0.7~0.8 | 0.045~0.06 |

By counting the predicted output amplitude, step size and absolute prediction error of a photovoltaic power system in a photovoltaic power plant for three years when the short-term predicted output is not zero, the effect of E factor, G factor, E factor and G factor interaction effect (E*G) on the PV prediction error is investigated by using ANOVA method.

The basic principle of ANOVA is to divide the difference between means of different treatment groups into between-group differences and within-group differences. The difference between groups is expressed as the sum of the squared deviations of the mean value of the variable in each group from the total mean value, denoted as SSb, the degrees of freedom between groups as dfb, and the ratio of the two as mean squared MSb. The difference within groups is expressed as the sum of the squared deviations of the mean value of the variable in each group from the value of the variable within that group, denoted as SSw, the degrees of freedom within groups as dfw, and the ratio of the two as MSw. The MSb/MSw ratio constitutes the F distribution, and the difference between the means of the groups is statistically significant when the F value is much greater than 1 for comparison with 1. The probability p-value of the F-value being greater than a specific value under the condition that the test hypothesis holds was obtained by consulting the F-boundary table. If the F-value was close to 1, the difference between the means of the groups was not statistically significant. The test level was selected as 0.05 and the original hypothesis was rejected at $p<0.05$ as a significant difference, that is, the test factor had an effect on the study subject, and conversely the original hypothesis was accepted as no significant effect of the test factor on the study subject. The test parameter values corresponding to each factor can be obtained by combining the factor and level relationships in Table 1, as shown in Table 2.

**Table 2.** Test parameter values corresponding to each factor.

| Source | Sum Sq. | d.f. | Mean Sq. | F | P |
|---|---|---|---|---|---|
| E | 0.269 | 5 | 0.05389 | 8.96 | 0 |
| G | 0.302 | 3 | 0.10081 | 16.77 | 0 |
| E*G | 1.554 | 43 | 0.03615 | 6.01 | 0 |
| Error | 306.495 | 50982 | 0.00601 | | |
| Total | 318.504 | 51039 | | | |

The first column “Source” in Table 2 is the source of variance, the second column “Sum Sq” is the sum of squares corresponding to each source of variance, the third column “df” is the corresponding degrees of freedom, the fourth column “Mean Sq” is the corresponding mean square, the fifth column “F” is the observed value of the F-test statistic, and the sixth column “P” is the test p-value obtained from the distribution of the F-test statistic. It can be seen that the p data are all less than the significant level 0.05, thus determining that the E factor, G factor and E*G factor have a significant effect on the magnitude of the prediction error.

Therefore, considering both factors, the historical data of this photovoltaic plant is used to count the MAE of the samples with different levels of combination of the two factors, and the results are shown in Figure 1.

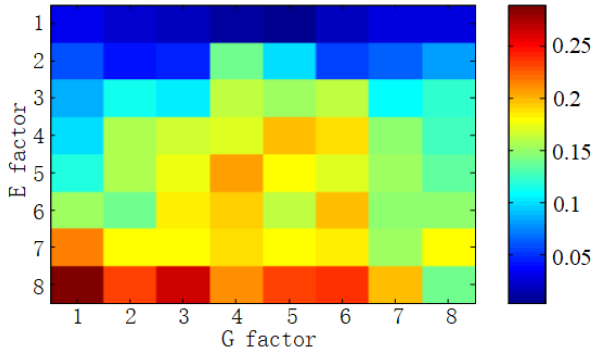


**Figure 1.** MAE statistical chart at each level of E and G factors based on historical data of the photovoltaic power station

Figure 1 consists of 64 small squares, and the color filled in each square represents the sample mean MAE under that level combination of corresponding E and G factors, and the color column on the right side indicates the correspondence between the color and the value. the values of MAE at each level of E and G factors are shown in Table 3.

**Table 3.** MAE statistical table of samples at each level of E and G factors

| E | G | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0.034 | 0.022 | 0.019 | 0.011 | 0.004 | 0.019 | 0.028 | 0.028 |
| 2 | 0.061 | 0.043 | 0.048 | 0.140 | 0.100 | 0.057 | 0.065 | 0.081 |
| 3 | 0.085 | 0.111 | 0.105 | 0.163 | 0.151 | 0.163 | 0.109 | 0.121 |
| 4 | 0.102 | 0.159 | 0.166 | 0.172 | 0.199 | 0.191 | 0.150 | 0.128 |
| 5 | 0.119 | 0.156 | 0.176 | 0.208 | 0.178 | 0.170 | 0.153 | 0.136 |
| 6 | 0.155 | 0.138 | 0.184 | 0.194 | 0.164 | 0.197 | 0.150 | 0.148 |
| 7 | 0.217 | 0.181 | 0.180 | 0.189 | 0.179 | 0.187 | 0.152 | 0.179 |
| 8 | 0.290 | 0.236 | 0.265 | 0.211 | 0.232 | 0.238 | 0.200 | 0.140 |

Based on the statistical results, the MAE was classified into six types of error conditions according to the principle of equally dividing sample size, and the results of the numerical characteristics classification of the predicted output of photovoltaic power generation based on the MAE values are shown in Table 4.

**Table 4.** Sample classification of E and G factors based on MAE values

| E | G | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 1 | 2 | 3 | 2 | 2 | 2 | 2 |
| 3 | 2 | 3 | 2 | 4 | 4 | 4 | 3 | 3 |
| 4 | 2 | 4 | 4 | 4 | 6 | 5 | 4 | 3 |
| 5 | 3 | 4 | 5 | 6 | 5 | 4 | 4 | 3 |
| 6 | 4 | 3 | 5 | 6 | 4 | 6 | 4 | 3 |
| 7 | 6 | 5 | 5 | 5 | 5 | 5 | 4 | 5 |
| 8 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 3 |

## 2.2Short-term photovoltaic power output prediction error model

*2.2.1GGMM.* There is asymmetry and kurtosis diversity in the short-term photovoltaic output prediction error distribution. Therefore, prediction error models are required to have flexible shapes and peaks. The traditional Gauss Mixture Model (GMM), is a weighted sum of multiple Gaussian functions, and the probability density function is defined as in equation (1). The number of summation terms n is usually taken as 3-5.

$$f(x|\theta)=\sum_{k=1}^{n}a_k\phi(x\mid\theta_k) \tag{1}$$

Where, $a_k$ is the weighting coefficient, $a_k\geq 0$, $\sum_{k=1}^{n}a_k=1$; $\theta_k=(\mu_k,\sigma_k^2)$; $\phi(x\mid\theta_k)$ is the Gaussian distribution density function, as in equation (2).

$$F(x\mid\theta)=\sum_{k=1}^{n}a_k\int_{-\infty}^{x}\phi(x\mid\theta_k)dt \tag{2}$$

The cumulative distribution function is:

$$F(x\mid\theta)=\sum_{k=1}^{n}a_k\int_{-\infty}^{x}\phi(x\mid\theta_k)dt \tag{3}$$

The range of the GMM is $(-\infty,+\infty)$, and the integral of each Gaussian component in this scope is 1. To ensure that the integral of the overall probability density of the GMM is also 1, it is necessary to assign a weight to each Gaussian component with a value no

greater than 1, and the sum of the weights is 1. However, the probability distribution of photovoltaic output prediction error obtained by the clustering method proposed in this paper cannot reach the range of $(-\infty,+\infty)$, and the prediction error model is required to have an integral sum of 1 in the given value range. Therefore, the GGMM is proposed in this paper. The definition formula of GGMM is basically the same as that of GMM, as in equation (1) and equation (2), only that the sum of weights is no longer required to be 1, that is, $\sum_{k=1}^{n} a_k \neq 1$, GGMM is more flexible than GMM and more suitable for describing the distribution of photovoltaic output prediction errors.

*2.2.2Model parameter estimation and accuracy evaluation index.* In this paper, the least squares method is used to estimate the model parameters, and the parameter estimates is obtained by nonlinear curve fitting function lsqcurvefit in MATLAB.

Three evaluation indexes, MAE, R2 and RMSE, were selected to evaluate the fitting effect. The MAE value characterizes the difference between the fitting function and the actual distribution function. The closer the MAE value is to 0, the better the model fitting effect and the data prediction effect is. The R2 value characterizes the closeness of the correlation. The closer the R2 value is to 1, the higher the reference value of the relevant equation is, while the closer the R2 value is to 0, the lower the reference value is. The RMSE is very sensitive to the response of extra-large or extra-small errors in a set of fits so that it can reflect the precision of the fit well. The closer the RMSE value is to 0, the higher the precision of the model fit. The specific formulas are as follows (4)-(6).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i\right| \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i)^2}{n}} \tag{6}$$

Where $y_i$ is the actual probability density value corresponding to an error interval, $y_i$ is the curve fitting value, and the subscript i indicates the i-th error interval.

## 3Example analysis

In this paper, the short-term predicted output and measured data of a photovoltaic power plant for three years are selected as training and testing samples to test the accuracy of the proposed model. The predicted and measured output data are standardized.

To illustrate the accuracy of the GGMM for describing the error distribution of short-term photovoltaic generation predicted output, the normal distribution, t Location-Scale distribution, improved Laplace distribution, and three GGMMs are used to fit the error samples of the same type, and the accuracy of each model is analyzed by comparing the fitting effects, as shown in equations (7)-(9).

(1) t Location-Scale distribution

$$f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sigma\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left[1 + \frac{\left(\frac{x-u}{\sigma}\right)^2}{v}\right]^{-\frac{v+1}{2}} \tag{7}$$

(2) Improved Laplace distribution

$$f(x) = \frac{k}{2}\lambda e^{-\lambda|x-\mu|} \tag{8}$$

(3) Normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{9}$$

By comparing and analyzing the fitting results of the six types error models, the results are shown in Figure 2 and Table 5.

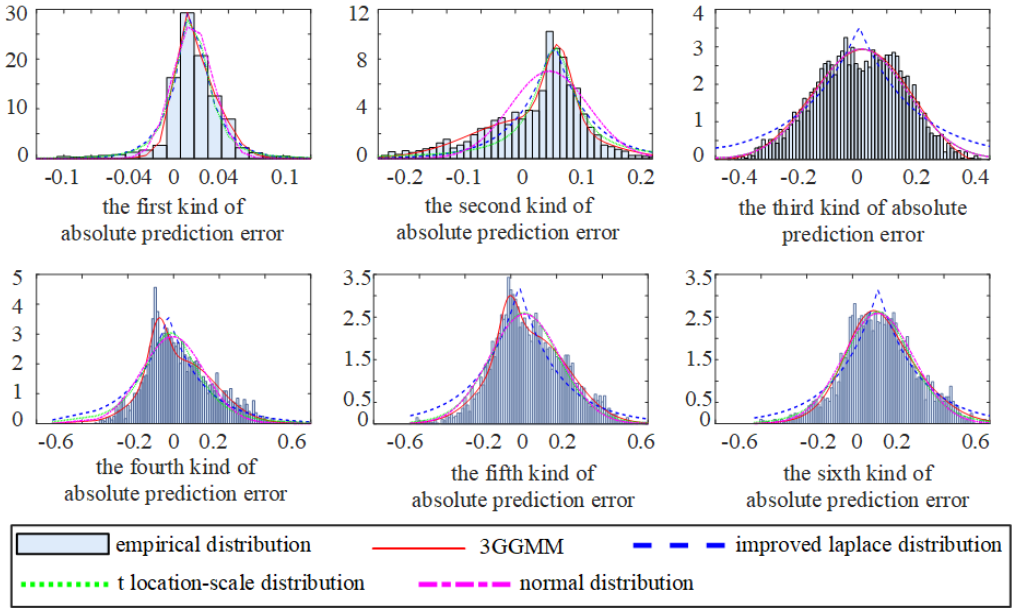


**Figure 2.** Comparison chart of distribution fitting of six groups of large errors

**Table 5.** Model fitting accuracy under different error distributions

| Error model | Evaluation index | The first kind of error | The second kind of error | The third kind of error | The fourth kind of error | The fifth kind of error | The sixth kind of error |
|---|---|---|---|---|---|---|---|
| 3GGMM | $R^2$ | 0.9234 | 0.9841 | 0.9533 | 0.9562 | 0.9528 | 0.9324 |
| | *RMSE* | 1.5094 | 0.2257 | 0.3230 | 0.2134 | 0.2092 | 0.2624 |
| | *MAE* | 0.7337 | 0.1529 | 0.2215 | 0.1445 | 0.1524 | 0.2091 |
| | *I* | 0.0330 | 0.0055 | 0.0136 | 0.0115 | 0.0106 | 0.0161 |
| t Location-Scale distribution | $R^2$ | 0.9403 | 0.9694 | 0.9612 | 0.9386 | 0.9282 | 0.9444 |
| | *RMSE* | 1.3321 | 0.3126 | 0.2946 | 0.2527 | 0.2582 | 0.2380 |
| | *MAE* | 0.6027 | 0.1821 | 0.2045 | 0.1794 | 0.1892 | 0.1727 |
| | *I* | 0.0256 | 0.0106 | 0.0112 | 0.0161 | 0.0160 | 0.0133 |
| Improved Laplace distribution | $R^2$ | 0.9294 | 0.9690 | 0.9665 | 0.9210 | 0.9214 | 0.8792 |
| | *RMSE* | 1.4493 | 0.3149 | 0.2736 | 0.2866 | 0.2700 | 0.3509 |
| | *MAE* | 0.7245 | 0.1853 | 0.1989 | 0.2046 | 0.2147 | 0.2814 |
| | *I* | 0.0304 | 0.0108 | 0.0097 | 0.0208 | 0.0177 | 0.0291 |
| Normal distribution | $R^2$ | 0.9108 | 0.9565 | 0.9341 | 0.9255 | 0.9207 | 0.9445 |
| | *RMSE* | 1.6285 | 0.3727 | 0.3839 | 0.2785 | 0.2712 | 0.2379 |
| | *MAE* | 0.7974 | 0.2405 | 0.2824 | 0.1937 | 0.2018 | 0.1727 |
| | *I* | 0.0354 | 0.0143 | 0.0181 | 0.0192 | 0.0178 | 0.0133 |

From the results in Figure 2 and Table 5, it can be seen that:

For the distribution case with prominent peak, the accuracy of the normal distribution model is the lowest, which is due to the influence of its kurtosis and cannot meet the characteristics of the high and lean distribution of the photovoltaic output prediction error. The accuracy of improved Laplace distribution is relatively low, which is due to the peak coefficient of this function is associated with its shape coefficient, therefore it makes the shape of the function limited. It cannot meet the two characteristics of high and lean peak as well as gentle waist at the same time, and cannot meet the requirement of peak. The accuracy of t Location-Scale is better than the first two, and the fitted curve at the peak basically overlaps with the GGMM curve. However, its shape is too tall and thin, and the curve at the waist is lower than the empirical distribution value. The GGMM fitting curve meets both peak and flexible waist, which has the best fit with experience distribution.

In the case of the kurtosis value slightly higher than the normal distribution, the normal distribution has a similar trend to the empirical distribution beyond the peak. It has a gentle waist curve but lacks the peak value. The waist fitting curve of the improved Laplace distribution is lower or higher than the empirical distribution when the peak value reaches the required value. The overall fitting effect of T Location-Scale distribution model is the closest to GGMM model. The GGMM accuracy is slightly less than that of the case with prominent peak, but the evaluation indexes are still

better than the other three models, so it is still the best choice.

In summary, the normal distribution model cannot express the asymmetry of the short-term photovoltaic power output prediction error, so it is not suitable to describe the photovoltaic power prediction error. The improved Laplace distribution model cannot describe the distribution condition of peak prominence. The t Location-Scale Distribution model has the problem of insufficient waist flexibility. The GGMM has a more flexible shape than other functions, and the evaluation index of GGMM is optimal for describing the prediction error of photovoltaic power generation with higher accuracy.

## 4Conclusion

This paper establishes a photovoltaic power output prediction error model based on the predicted output fluctuation, and proposes the use of GGMM to describe the photovoltaic power prediction error. Meanwhile, the GGMM is compared with t Location-Scale distribution, improved Laplace distribution and normal distribution model through an arithmetic example of a photovoltaic power plant output prediction to verify the superiority of GGMM, and the main conclusions obtained are as follows.

(1) The error value at each moment of photovoltaic power output prediction error is correlated with the predicted output amplitude and power variation at that point, which can be divided into six types of error conditions according to the principle of equally dividing sample size, and the error distribution shows asymmetry and kurtosis diversity.

(2) The GGMM proposed in this paper is flexible in shape and can satisfy both the asymmetry and kurtosis diversity of photovoltaic prediction error distribution with higher accuracy and applicability compared to normal distribution, improved Laplace distribution and t Location-Scale distribution model.

## References

1. ZHUANG Yani, YANG Xiuyuan, JIN Xincheng. Study on operation technology of wind- PV- energy storage combined power generation[J]. Power Generation Technology, 2018 39(4): 296-303.
2. TIAN Hao, ZHANG Han, FENG Wanxing, et al. CG Lightning prediction method of based on BP neural network and atmospheric electric field characteristics[J]. Insulators and Surge Arresters, 2018(6): 27-33.
3. Marquez R, Coimbra C F M. Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database[J]. Solar Energy, 2011, 85(5): 746-756.
4. Lorenz E, Hurka J, Heinemann D, et al. Irradiance Forecasting for the Power Prediction of Gird-Connected Photovoltaic Systems[J]. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 2009, 2(1): 2-10.
5. Lin Shaobo, Han Minxiao, Zhao Guopeng, et al. Capacity allocation method of distributed photovoltaic energy storage system based on stochastic prediction error [J]. Proceedings of the CSEE, 2013, 33(4): 25-33.
6. Li Pengmei, Zang Chuanzhi, Li Hepeng, et al. Energy Stochastic Optimization Scheduling of Microgrid Based on Photovoltaic Prediction [J]. Transducer and Microsystem, 2015, 34(2): 61-64.
7. Chen Yaoqi. Short-term forecasting of photovoltaic power generation based on typical weather types and taking into account stochastic forecasting error [J]. China Electric Power, 2016, 49(5): 157-162.