ISSN: 2472-3703 DOI: 10.6911/WSRJ.202604_12(4).0008

Development of a Regional Photovoltaic Baseline Dataset for Qinghai Province

Qiuping He[1, a]

[1]Meteorological Information Center of Qinghai Province, Xining, Qinghai, China

[a]993032469@qq.com

Abstract

To support the operational needs for detailed assessment of solar resources and PV power forecasting in high-altitude regions of Qinghai Province, this study conducted the standardization and development of a PV dataset, along with end-to-end quality control, based on 15-minute interval measured data from June 2021 to June 2022 at four typical PV power plants in Qinghai Province. Missing values were imputed using physical threshold constraints and linear/co-periodic mean interpolation. A combined outlier detection model utilizing DBSCAN density clustering and random forest regression was constructed to automatically identify and remove invalid values, abnormal spikes, and physically implausible data. The results show that after quality control, the valid data rate for total radiation increased from less than 30% to 87.5%, with a correlation coefficient of 0.92 relative to meteorological parameters and a root mean square error of 48.7 $W \cdot m^{-2}$. The data quality meets the requirements for solar resource assessment and power forecasting.

Keywords

Photovoltaic data, Data processing, Quality control, DBSCAN, Isolation Forest.

1. INTRODUCTION

The 2025 China Wind and Solar Energy Resource Outlook Report indicates that in 2025, solar energy resources nationwide will be below average. The national average annual horizontal total irradiance is 1,495.7 kWh/m^2, which is 25.1 kWh/m^2 lower than the average over the past 30 years. The annual optimal tilt total irradiance for photovoltaic power generation is approximately 1,742.2 kWh/m^2, which is 36.7 kWh/m^2 lower than the average over the past 30 years. There are significant regional variations in solar resources across the country, characterized by higher levels in western regions compared to central and eastern regions, and higher levels in plateau areas and arid regions with low rainfall, while lower levels are found in plains and regions with high rainfall and humidity. Nationally, the annual horizontal total irradiance shows a pattern where western regions exceed central and eastern regions. Most of Tibet, central and northern Qinghai, and western Sichuan have annual horizontal total irradiance exceeding 1,750 kWh/m^2, making these areas the most abundant in solar resources. The solar energy resources available for fixed-mount photovoltaic power generation refer to the total solar irradiance that photovoltaic modules can receive when positioned at the optimal tilt angle (the angle corresponding to maximum annual power generation), i.e., the total irradiance at the optimal tilt angle. This serves as a crucial basis for the detailed assessment of solar energy resources in the photovoltaic industry.

Leveraging advantages such as high altitude, high light intensity, and long hours of sunshine, Qinghai Province has become a major hub for clean energy in China. High-precision, long-term,

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202604_12(4).0008

and standardized PV monitoring datasets serve as the core foundation for PV power forecasting, wind and solar resource assessment and prediction, as well as high-impact weather service support and severe weather warnings [1]. Currently, PV monitoring data commonly suffer from issues such as missing time series, abnormal jumps, and poor physical consistency, which directly affect the accuracy of assessments and predictions [2].

Due to the combined effects of geographical and industry-specific constraints, publicly available research-grade PV datasets are relatively scarce. The NREL (National Renewable Energy Laboratory) dataset [3] includes PV data from 38 states in the eastern United States for the year 2006, categorized into three types: actual output, previous-day forecast, and 4-hour-ahead forecast. This dataset uses a sub-hourly total irradiance algorithm to generate 5-minute and hourly datasets for both centralized and distributed PV systems. The numerical weather prediction (NWP) data was generated by 3TIER based on NWP simulations from Phase 1 of the Western Wind and Solar Integration Study. The PVOD dataset [4] collects historical power generation records from 10 PV power plants in Hebei Province, China, comprising a total of 271,968 records. The data covers the period from September 2018 to August 2019 (365 days) with a 15-minute time resolution. This dataset includes historical power output data for each site, meteorological data (such as total direct radiation, total diffuse radiation, temperature, humidity, and wind speed), and numerical weather prediction data corresponding to the latitude and longitude of each power plant. In addition, the dataset provides basic information on the PV sites, including the capacity, area, number, material, and orientation of the PV panels. It can be applied to fields such as total radiation forecasting at different time scales, single-site PV power forecasting, and multi-site regional PV power forecasting. The 1991–2020 Chinese Ground-Based Meteorological Radiation Climate Data Set established by Xu Yongfang et al. [5] systematically standardized radiation data quality control, heterogeneity correction, and multiscale statistical methods. Liu Junjian et al. proposed a method for the fusion and evaluation of multi-source ground-based shortwave radiation data [6], providing a technical reference for the identification and correction of errors in ground-based observation data.

Based on the complex topography of the Qinghai-Tibet Plateau in Qinghai Province, this study refers to the Technical Guidelines for Quality Control of Wind and Solar Energy Resource Observation Data compiled by the National Meteorological Information Center. Taking into account the complex topography of the Qinghai-Tibet Plateau in Qinghai Province, a PV data processing workflow of “standardized processing—missing value interpolation - dual-model outlier detection" PV data processing workflow. Based on the PV characteristics of the Qinghai region, localized algorithm optimization was performed to create a Qinghai Province PV baseline dataset suitable for operational use, thereby providing data support for the development and utilization of regional solar energy resources.

2. DATA AND METHODS

2.1. Data

This study utilizes 15-minute real-time photovoltaic monitoring data from June 2021 to June 2022 for the Yixin, Cui Feng, Qingjiao, and Zhengdao photovoltaic power plants in Qinghai Province. Ground-based meteorological observation data and meteorological data such as the clear-sky index corresponding to the geographical locations of these power plants were collected from multiple sources, including big data cloud platforms, meteorological archive cloud calendars, and observational data records.

2.2. Statistical Items and Basis

The compilation and statistics of the Qinghai Province PV baseline dataset include PV power plant basic information, air pressure, wind speed, humidity, temperature, wind direction, total

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202604_12(4).0008

radiation, and actual active power output. The development of algorithms refers to the Ground Meteorological Observation Specifications compiled by the China Meteorological Administration and the Technical Guidelines for Quality Control of Wind and Solar Energy Resource Observation Data compiled by the National Meteorological Information Center.

2.3. Data Quality Control and Evaluation

Comprehensive analysis of the raw data reveals that the collected monitoring data contains a significant amount of "dirty data," including data exceeding performance parameter thresholds, duplicate data, missing data, and outliers. The development of the PV baseline dataset requires standardization of the raw data, including filling in missing values, removing outliers, and performing data quality control. The data is then re-examined and validated from the perspective of dataset application. Finally, the data quality is classified and labeled based on the validation results.

The processing workflow for the PV baseline dataset is shown in Figure 1 below:

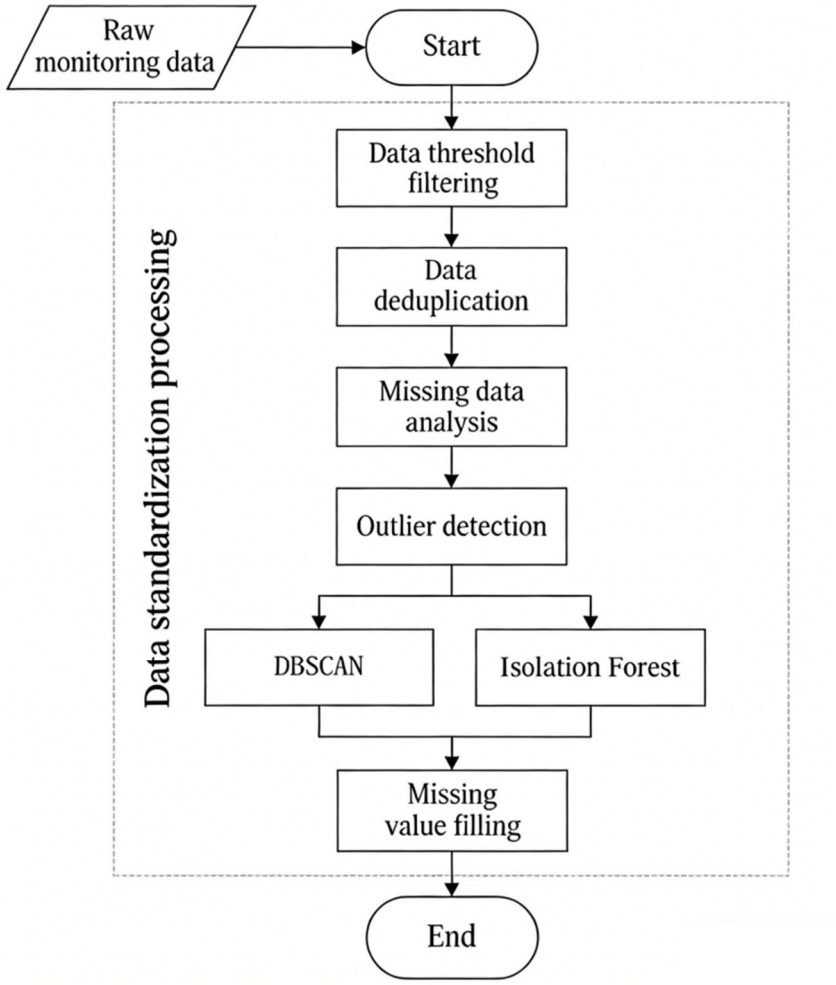

Figure 1. Data Processing Workflow for the Photovoltaic Dataset

The values and meanings of the quality identification codes are shown in Table 1 below:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202604_12(4).0008

Table 1. Values and Meanings of Quality Identification Codes

Quality ID	Meaning
0	Data correct
1	Data questionable
2	Data error
3	Data revised
6	Data revised to missing
8	Data missing or not observed

2.3.1 Standardization of Time and Units

Data is standardized to 15-minute intervals in Beijing Time, with units and precision standardized in accordance with the Specifications for Ground Meteorological Observations.

2.3.2 Threshold Checking

For monitoring data, check whether values exceed defined thresholds. For data exceeding thresholds, correct the data through manual verification of reports. If the report also contains erroneous values, mark them as missing.

2.3.3 Outlier Detection

(1) DBSCAN

DBSCAN (Density-based Spatial Clustering of Applications with Noise) [7] is a common unsupervised clustering algorithm based on density-based spatial analysis. It uses total radiation, air temperature, atmospheric pressure, and wind speed as feature vectors to identify outliers based on density. Parameters: neighborhood radius ε = 0.5, minimum number of samples min_samples = 5. Isolated points in low-density areas are marked as density anomalies.

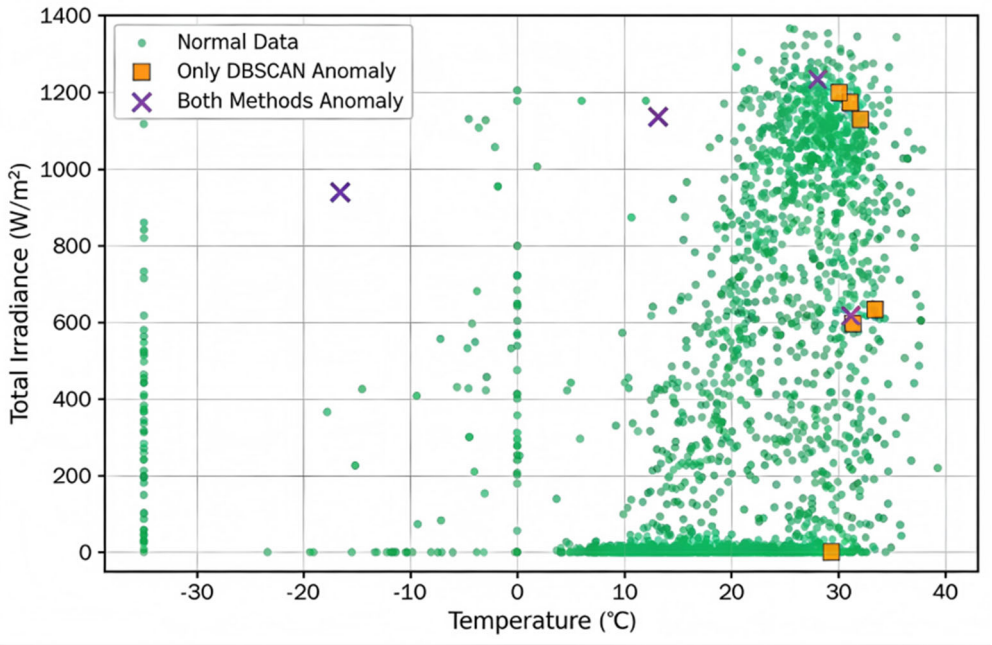

Figure 2. DBSCAN anomaly detection scatter plot

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202604_12(4).0008

(2) Isolation Forest

For the detection of outlier data values, statistical methods, clustering-based methods, and specialized outlier detection algorithms are generally used. Isolation Forest is a specialized outlier detection algorithm [8] that is often used for outlier detection in continuous data and offers high detection accuracy.

Outlier determination: Residual > 3σ (threshold ≈ 85 W/m^2)

Results: 78 anomalies detected (0.23%), primarily due to latent residual anomalies

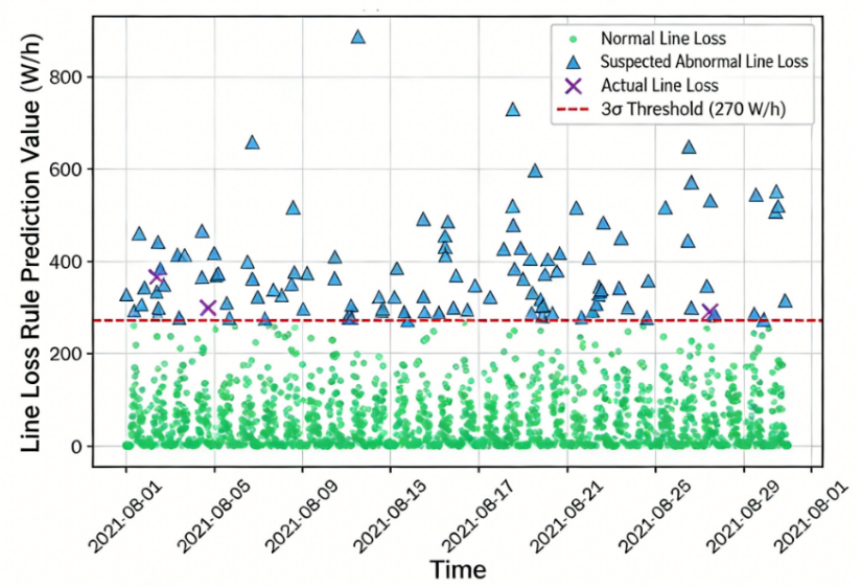

Figure 3. Residual analysis of isolated forests (including 3σ threshold lines)

2.3.3 Missing Value Imputation Strategies

Missing values in photovoltaic data can be classified into three types based on the degree of missingness: complete random missingness, random missingness, and non-random missingness [9]. Currently, common methods for imputing missing values in photovoltaic data include those based on temporal characteristics, those based on correlations, those based on spatial characteristics, and those based on multidimensional combinations. This paper adopts a strategy combining physical rules with statistical methods to effectively improve data completeness.

Table 2. Statistics on Missing Value Interpolation

Missing Type	Processing Method	Imputation Effect
Missing nighttime total radiation	Fill with 0 (before 6:00 / after 18:00)	9736 records imputed
Short time-series missing (≤45 min)	Linear interpolation	Average 9000+ records / element
Long time-series missing (>45 min)	Mean value of the same period & time	Average 700+ records / element

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202604_12(4).0008

Post-interpolation coverage rates: Total radiation 98.0%, Temperature 99.0%, Wind speed 99.4%, Air pressure 98.4%.

2.3.4 Quality Assessment

The quality of the total radiation data after quality control was evaluated using Mean Biased Error (MBE), Root Mean Square Error (RMSE), and Correlation Coefficient (CORR) as evaluation metrics [10].

Mean Biased Error (MBE) is a metric used to measure the deviation between model predictions and actual values. It indicates whether the model exhibits a positive or negative bias, as well as the magnitude of that bias. Specifically, a positive bias indicates that the model overestimates the parameter values, while a negative bias indicates that the model underestimates them.

$$MBE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)$$

Here, y is the observed value, y is the predicted value, and n is the number of samples.

The root mean square error (RMSE) is the square root of the mean square error (MSE). MSE measures the average of the squared differences between the predicted and observed values, while RMSE scales this value to match the order of magnitude of the original data, making it easier to interpret.

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

MBE = 21.3 $W\cdot m^{-2}$

RMSE = 48.7 $W\cdot m^{-2}$

CORR = 0.92

3. CONCLUSIONS

Based on PV monitoring data from June 2021 to June 2022 for the Yixin, Cui Feng, Qingjiao, and Zhengdao PV power plants in Qinghai Province, this study collected meteorological data—including ground-based meteorological observations and clear-sky indices—corresponding to the geographical locations of these power plants from multiple sources, such as big data cloud platforms, meteorological archive cloud calendars, and observational records, thereby establishing a comprehensive observation system.

(1) A PV data processing workflow comprising “standardization—missing value interpolation-dual-model outlier detection” was established. Based on the characteristics of PV systems in the Qinghai region, localized algorithm optimization was performed, ensuring compliance with meteorological industry standards and enabling replicability and operational implementation.

(2) After quality control, the effective data rate for total radiation was increased to 87.5%, with a correlation coefficient of 0.92 and an RMSE of 48.7 $W\cdot m^{-2}$, meeting the requirements for solar resource assessment and PV power forecasting.

(3) A 15-minute resolution PV baseline dataset for Qinghai Province was generated, covering four typical stations and five categories of key meteorological and radiation parameters, thereby supporting regional renewable energy meteorological services.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202604_12(4).0008

(4) This study provides a technical framework for standardizing PV monitoring data in high-altitude regions, holding significant practical value for clean energy development on the Qinghai-Tibet Plateau.

4. OUTLOOK

This paper integrates the Ground Meteorological Observation Specifications, the Technical Guidelines for Quality Control of Wind and Solar Energy Resource Observation Data, and machine learning-based anomaly detection to establish a standardized workflow for PV datasets suitable for plateau environments. DBSCAN excels at identifying density-based outliers, while Isolated Forest excels at capturing anomalies caused by inconsistent physical relationships; the combined detection of these two methods achieves higher accuracy than either method alone. Due to limitations in the number of stations, this study did not conduct multi-source fusion correction involving satellite, reanalysis, and ground-based data. Future research could incorporate FY-2 series satellite radiation products and ERA5 reanalysis data to further enhance the spatial representativeness and accuracy of the dataset.

REFERENCES

[1] Li Bo, Pan Meng, Sun Yue. A Review of the Application of Artificial Intelligence in the Development of Meteorological Datasets [J]. People's Yangtze, 2025, 56(01): 88–96. DOI: 10.16232/j.cnki.1001-4179.2025.01.012.

[2] Zhang Pei, Liu Jincheng, Zhang Bin, et al. A Review of Public Datasets for Photovoltaic Power Generation Forecasting [J]. Electric Power Information and Communication Technology, 2023, 21(08): 16–21. DOI: 10.16543/j.2095-641x.electric.power.ict.2023.08.03.

[3] Lee S G, Park S J, Lee K S, et al. Performance prediction of NREL (National Renewable Energy Laboratory) Phase VI blade adopting blunt trailing edge airfoil [J]. Energy, 2012, 47(1):47-61. DOI:10.1016/j.energy.2012.08.007.

[4] Ren Mifeng, Wang Jiahui, Ye Zefu, et al. A transferable ultra-short-term PV forecasting modeling framework applicable to single/multi-PV power plants [J]. Journal of Solar Energy, 2024, 45 (06): 359-367. DOI:10.19912/j.0254-0096.tynxb.2023-0330.

[5] Xu Yongfang, Liao Jie, Zhao Yufei. Development of a Chinese Ground-Based Meteorological Radiation Climate Data Set for 1991–2020 [J]. Atmospheric Sciences, 2024, 48(05): 2080–2094.

[6] Liu Junjian, Shi Chunxiang, Han Shuai, et al. Fusion and Evaluation of Multi-Source Ground-Based Shortwave Radiation Data [J]. Remote Sensing Technology and Application, 2018, 33(05): 850-856.

[7] Wang Dian, Chang Jun. Data Anomaly Detection Combining Deep Learning with Improved DBSCAN Clustering [J]. Journal of Dynamics and Control, 2025, 23(09): 74-84.

[8] Huang Yanjun, Zhang Bo, Zhang Yichao, et al. Research on Anomaly Detection in Power Big Data Based on an Improved Isolated Forest Model [J]. International Journal of Electronic Measurement Technology, 2025, 44(10): 88-94. DOI:10.19652/j.cnki.femt.2510015.

[9] He Jianping. A Method for Filling Missing Data in Photovoltaic Systems Based on Multi-Temporal Forecasting Models [D]. North China Electric Power University, 2025. DOI:10.27139/d.cnki.ghbdu.2025.000082.

[10] Khan Q A ,Muhammad G S ,Raza A , et al.Machine learning models for predicting carbonation depth in fly ash concrete:performance and interpretability insights[J/OL].Journal of Road Engineering, 2026, (01):74-90[2026-03-18].https://link.cnki.net/urlid/61.1520.U.20260317.1715.010..