Data Analysis Project Report

Team: I love data analysis
Andreas Heindl

January 29, 2026

1 Contributions

Solo project; all data engineering, analysis, modeling, and reporting completed by Andreas Heindl.

2 Dataset Description

- **Dataset source:** "Solar Power Generation Data" (Ani Kannal, Kaggle).
- **Suitability for time-series analysis:** Inverter data plus co-recorded weather sensors with multiple features and enough data points.
- **Time span and sampling:** 15 May 2020 00:00 17 Jun 2020 23:45 with a sampling interval of 15 minutes.
- **Key variables:** AC Power, DC Power, Total Yield, Daily Yield, Irradiation, Ambient Temperature, Module Temperature, Source Key, Plant ID.
- **Size and structure:** The generation table has 68 778 rows with 7 columns and the weather table has 3 182 rows with 6 columns.
- **Missing data summary:** Nearly no NANs were counted; some irradiation readings were interpolated while other segments stay missing for longer periods.
- **Known caveats:** DC Power needed a $\times 0.1$ scaling factor; inverter logs rotate through the day; weather data comes from one mast and was (imperfectly) assumed to represent every inverter.

3 Data Preprocessing and Data Quality

3.1 Basic statistical analysis using pandas

Table 1: Generation data grouped statistics and quantiles

Metric (unit)	mean	std	min	25%	50%	75%	max
AC Power (kW)	307.80	394.40	0.00	0.00	41.49	623.62	1 410.95
DC Power (kW)	3 147.43	4 036.46	0.00	0.00	429.00	6 366.96	14 471.13
Total Yield (MWh)	6.98	4.16	6.18	6.51	7.15	7.27	7.85
Daily Yield (kWh)	3 295.97	3 145.18	0.00	0.00	2 658.71	6 274.00	9 163.00

Table 2: Weather data grouped statistics and quantiles

Metric (unit)	mean	std	min	25%	50%	75%	max
Irradiation (kW/m^2)	0.228	0.301	0.000	0.000	0.025	0.450	1.222
Ambient Temperature (°C)	25.53	3.35	20.40	22.71	24.61	27.92	35.25
Module Temperature (°C)	31.09	12.26	18.14	21.09	24.62	41.31	65.55

3.2 Original data visual quality analysis

Following visual issues were detected in the raw data:

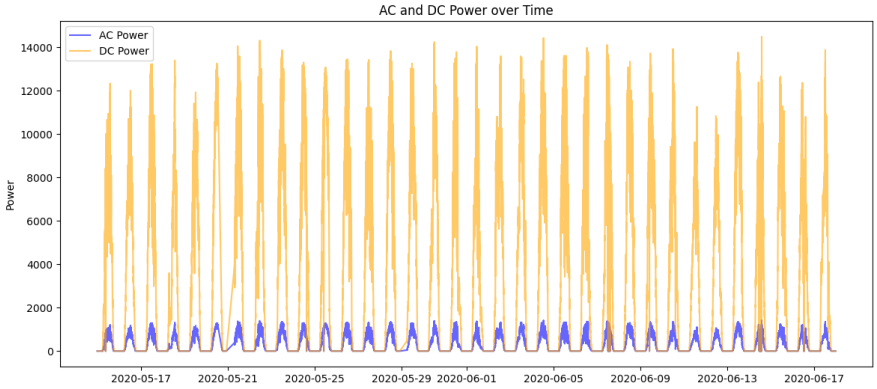


Figure 1: DC Power is about 10× higher than AC Power in the raw data, indicating a scaling error.

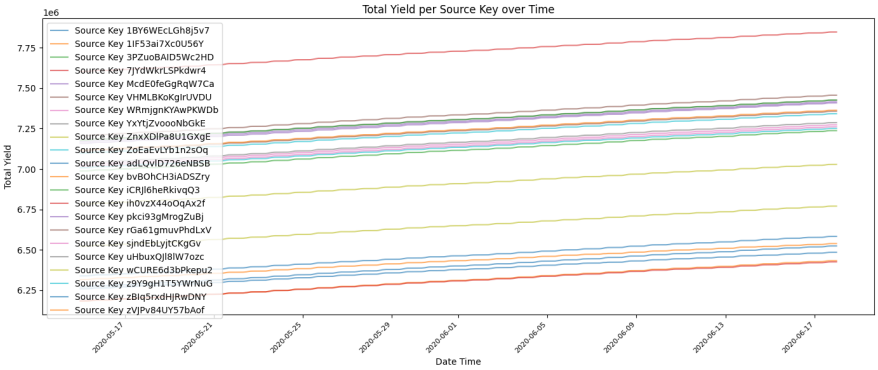


Figure 2: Total Yield over time mainly reflects different offsets per source key (not starting at 0), so it is not comparable across inverters and was excluded from later merged analysis.

3.3 Data Preprocessing

- Converted date time columns to a consistent datetime format and verified the expected 15-minute sampling over the full range (15 May–17 Jun 2020).
- Rescaled DC Power by a factor of 0.1 to match the magnitude of the co-recorded AC Power.
- Dropped Plant ID because it is constant and adds no information.
- Removed two extreme irradiation spikes and filled 44 missing irradiation values with interpolation; final merged table has no missing values (68,778 rows, 9 columns).

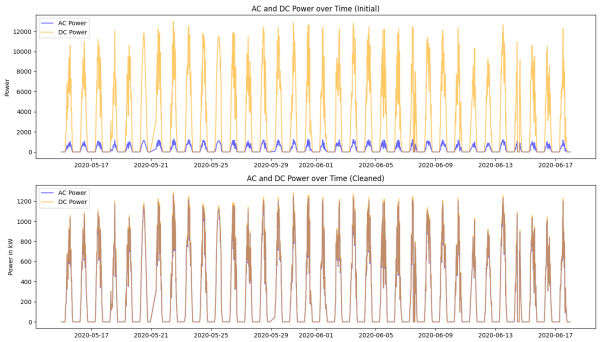
3.4 Preprocessed vs original data visual analysis

Two quick before/after plots summarize the most visible preprocessing effects.

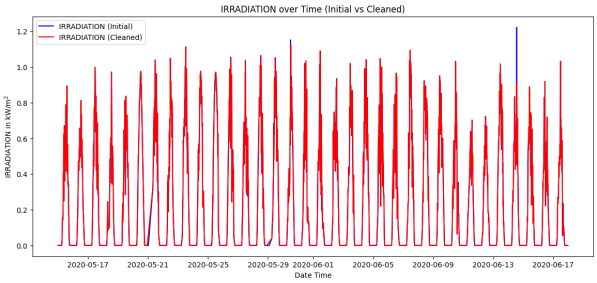
4 Visualization and Exploratory Analysis

4.1 Time-series visualizations

Figures 3a and 3b show time series visualizations.



(a) After rescaling, AC Power and DC Power align in magnitude over time.

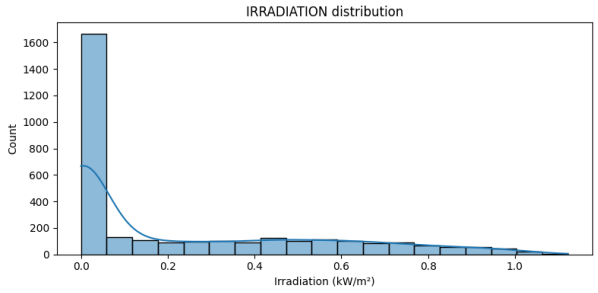


(b) Irradiation outliers removed and gaps interpolated.

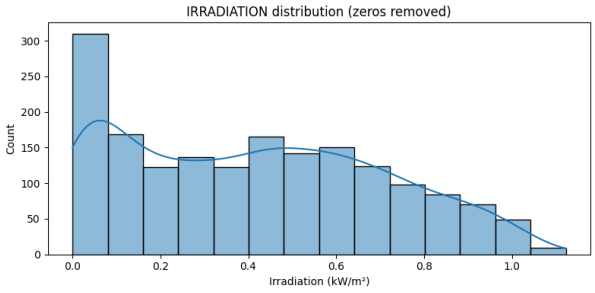
Figure 3: Preprocessed vs. original comparison for power and irradiation.

4.2 Distribution analysis

Irradiation distributions summarize typical operating conditions; zero values (nighttime) were excluded in one plot because they dominate the histogram.



(a) Including zero values



(b) Excluding zero values

Figure 4: Irradiation distribution.

4.3 Correlation analysis

The correlation matrix shows the strongest relationships between weather and power. A notable case is Daily Yield versus AC/DC Power: Pearson stays low due to the day/night plateau, while Spearman is higher because sunny hours consistently rank above nighttime values.

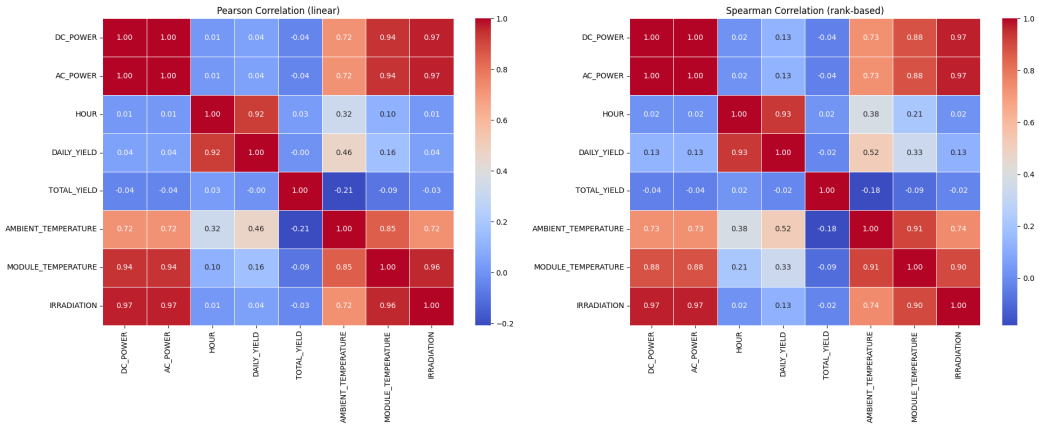


Figure 5: Correlation heatmap (Pearson and rank-based correlation) for key variables.

4.4 Daily / periodic pattern analysis

Daily patterns show a clear midday production window: irradiation and power rise after sunrise, peak around 12:00–15:00, and drop back to near zero in the evening.

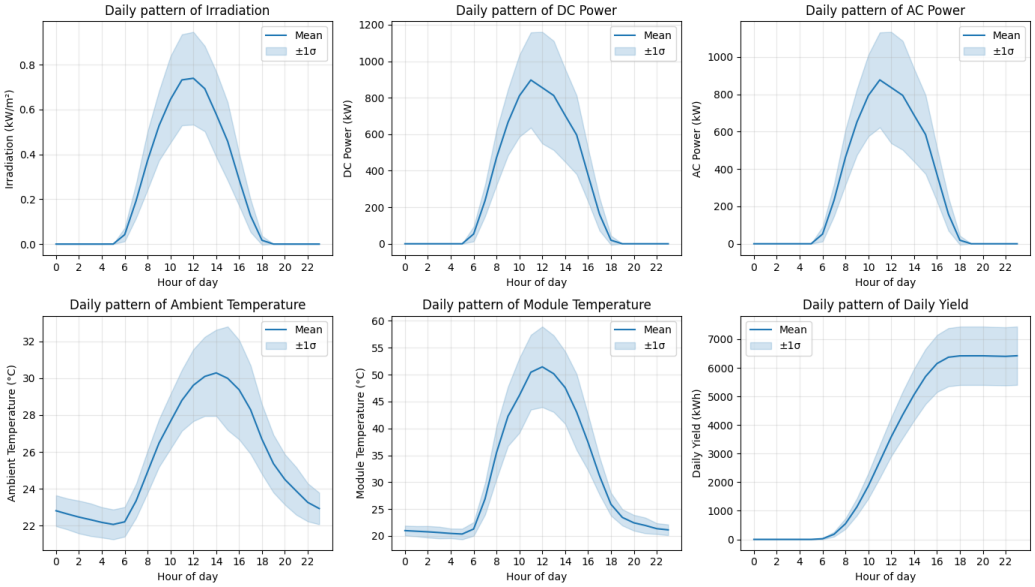


Figure 6: Daily pattern of irradiation and power across the day.

Pattern checks (True/False):

- Irradiation peaks around 12:00–15:00 each day — **True**.
- AC Power and DC Power peak together with irradiation — **True**.
- Ambient Temperature starts rising around 05:00–06:00 — **True**.
- Daily Yield end of the day variability is on the order of $\sigma \approx 2000\,\text{kWh}$ — **True**.

5 Probability and Event Analysis

5.1 Threshold-based event probability

To mark "highly productive" conditions, I used percentile thresholds (90th percentile for AC Power and 80th percentile for Irradiation). Empirical frequencies:

- $P(\text{AC Power} > 873\,\text{kW}) = 10.02\%$.
- $P(\text{Irradiation} > 0.5364\,\text{kW/m}^2) = 20.00\%$.

5.2 Cross tabulation analysis

Irradiation and Module Temperature bins co-vary strongly: zero irradiation occurs almost entirely at cooler module temperatures, while high/very high irradiation concentrates in the 40–50 °C and > 50 °C bins.

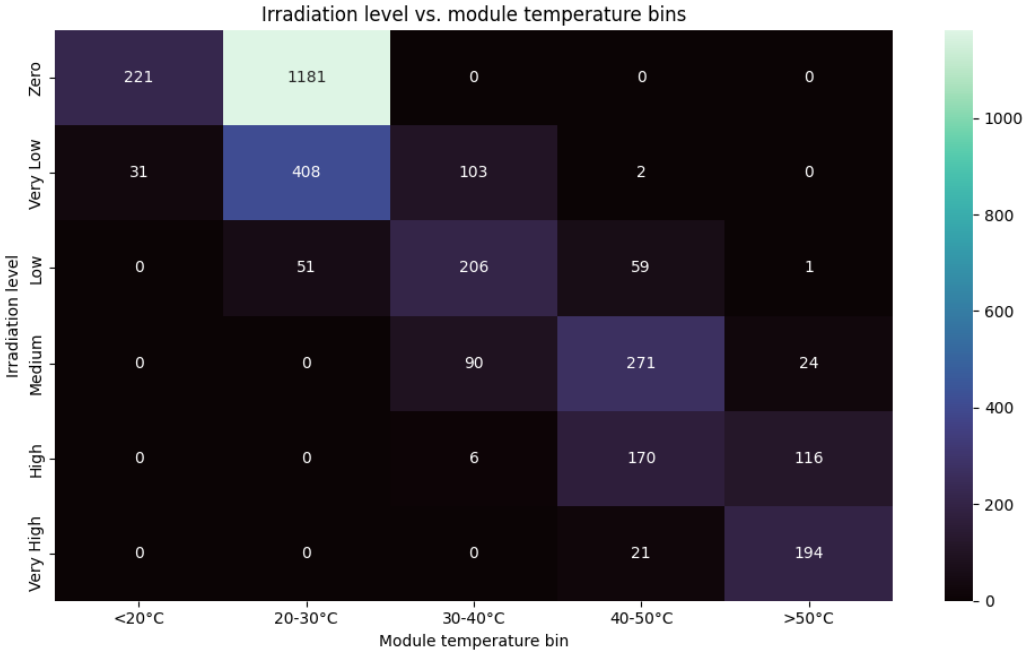


Figure 7: Cross tabulation of Irradiation bins versus Module Temperature bins (counts).

5.3 Conditional probability analysis

Conditional probabilities were computed for a high-power event (AC Power above its 90th-percentile threshold), grouped by irradiation bands and temperature bands. Medium to very high irradiation corresponds to a high likelihood of reaching the 90% power level.

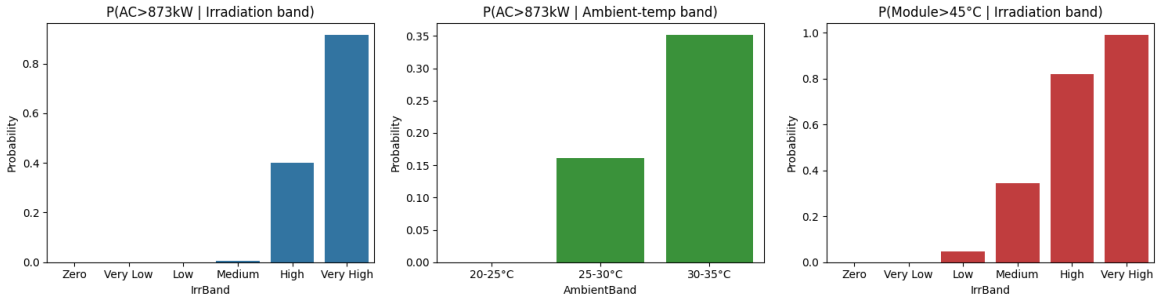


Figure 8: Conditional probabilities for high AC Power across irradiation and temperature bands.

5.4 Observations and limitations

Observations:

- 873 kW (90th percentile) is close to the practical upper output. The high irradiation bands make AC Power $>$ 873 kW very likely.
- Ambient Temperature around 25–35 °C tends to coincide with high success probability.
- Elevated module temperatures mainly appear in the highest irradiation bins.

Limitations / bias:

- One weather sensor is assumed to represent all inverters.
- Interpolated irradiation and temperatures can smooth sharp peaks.
- Equal weighting of night and day timestamps pulls down unconditional probabilities.
- Percentile thresholds (e.g., 90% / 80%) are dataset-specific and may not transfer to other sites or seasons.

6 Statistical Theory Applications

6.1 Law of Large Numbers (LLN)

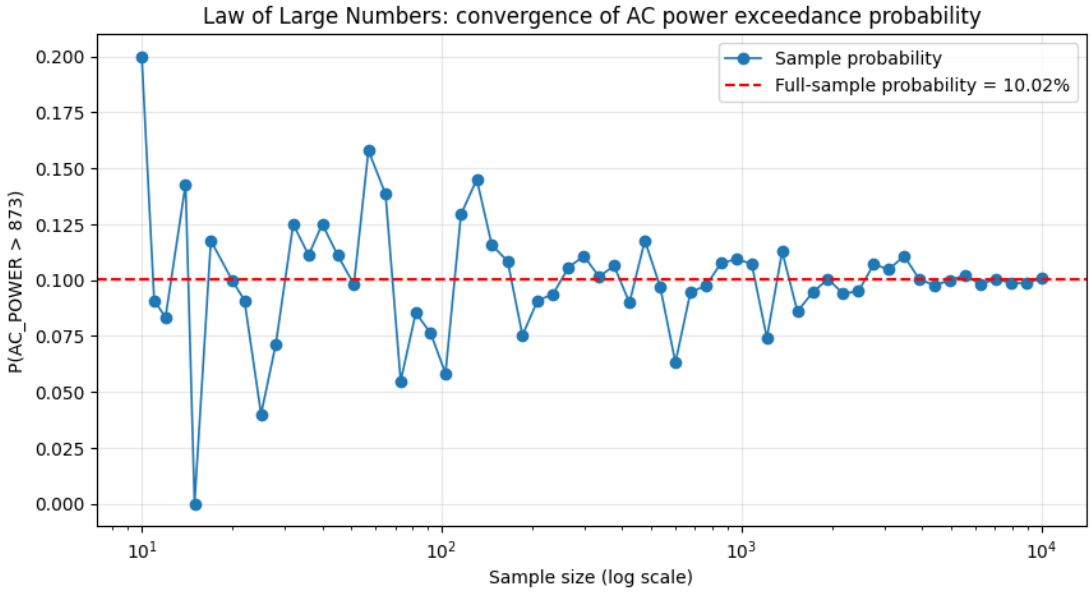


Figure 9: Law of Large Numbers.

6.2 Central Limit Theorem (CLT)

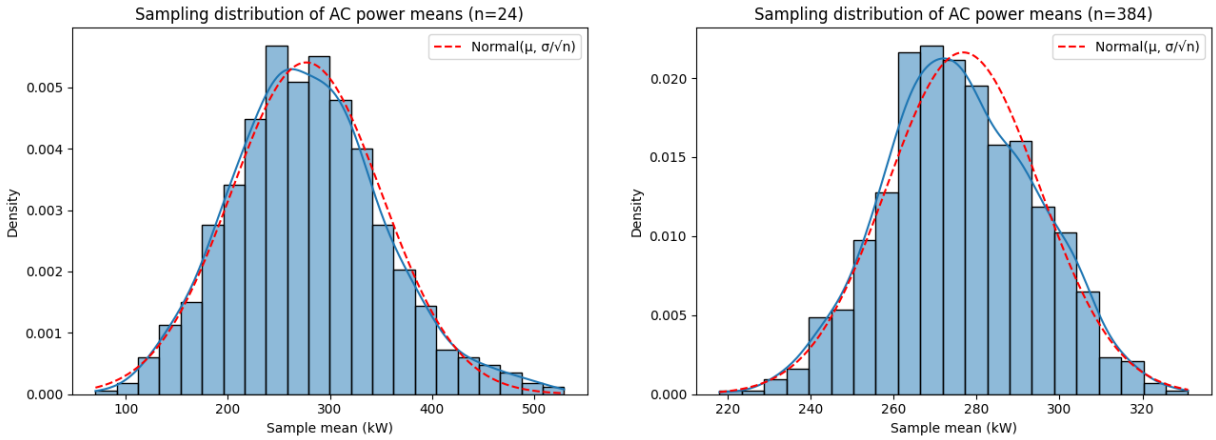


Figure 10: Central Limit Theorem.

6.3 Sanity checks and interpretation

Law of Large Numbers: The estimate starts noisy (0–0.2), settles near 10% after about 800 observations, and stays stable after about 3000 (Figure 9).
Central Limit Theorem: The sampling distribution of the mean tightens as n increases (Figure 10; left: $n = 24$, right: $n = 384$).

7 Regression and Predictive Modeling

7.1 Model definition, fitting, and validation

Prediction target: AC power.
Features: Irradiation and Hour (captures the daily sunrise/sunset cycle).
Candidates:

- Linear regression (baseline).
- Polynomial regression (degrees 2 and 3) to allow curvature and interactions.

Table 3: Validation performance for candidate models

Model	R^2_{valid}	$\text{RMSE}_{\text{valid}}$
Linear	0.880660	113.415684
Poly (deg=2)	0.881550	112.991676
Poly (deg=3)	0.865626	120.347468

Poly (deg=2) was selected as a simple best performer.
Residual summary: mean −5.8, std 112.8.

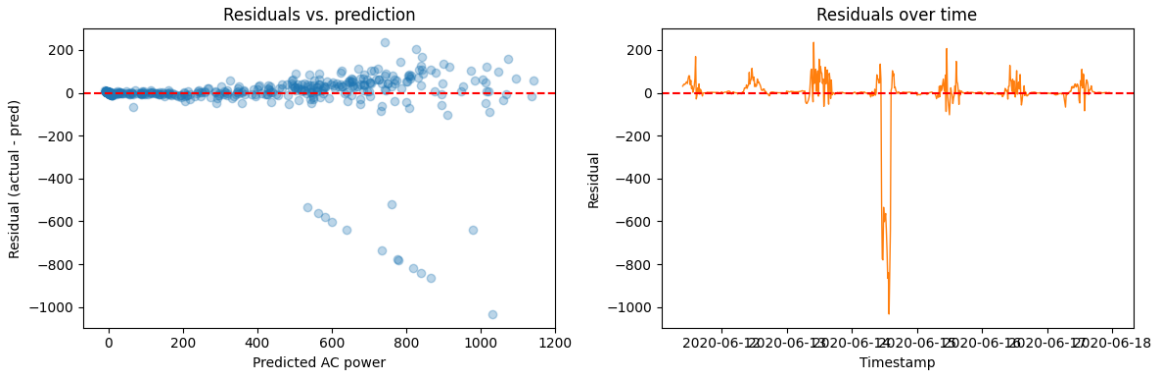
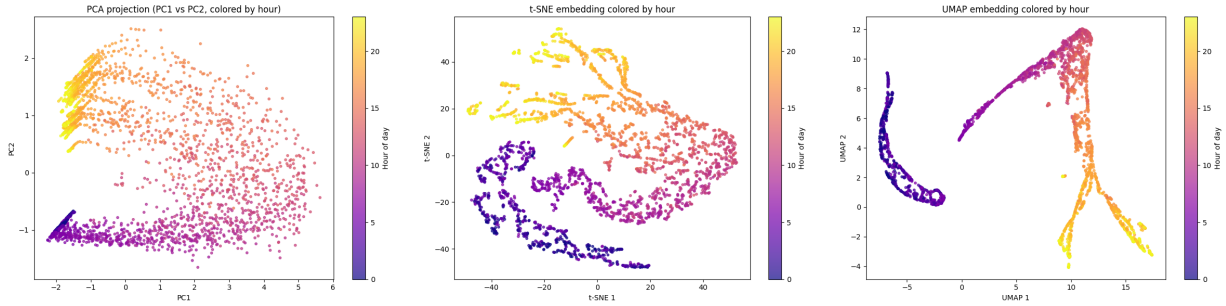


Figure 11: Residual diagnostics for Poly (deg=2): residuals vs. prediction (left) and residuals over time (right).

Residual-vs-prediction stays centered near zero with a few extreme points; the time plot shows one period with strong overestimation, likely due to unmodeled cloudy conditions.

8 Dimensionality Reduction and Statistical Tests

8.1 PCA, t-SNE, and UMAP



(a) PCA (b) t-SNE (c) UMAP

Figure 12: PCA, t-SNE, and UMAP embeddings of six standardized features, highlighting the day–night structure in the solar plant data.

PCA (variance and meaning): The first two principal components capture over 95% of the variation, which is expected because all six features are driven by the daily solar cycle. PC1 mainly represents overall solar intensity (power, irradiation, and temperature rising together). PC2 separates daily energy yield (positive) from ambient temperature (negative), which loosely tracks earlier vs. later hours based on accumulated production.
PCA (shape): In the 2D plot, points form a smooth curve that follows the sun's daily arc rather than distinct clusters: night sits near zero, then the data flows continuously through morning, noon, and evening.
t-SNE: t-SNE separates sparse night/zero-production readings from the daytime cloud. One compact lobe contains the near-zero points, while a second lobe covers timestamps with real output, with a thin gap where inputs were dropped.
UMAP: UMAP shows the same two regimes as t-SNE but spread over a wider span. It forms a quiet island for night/zero readings and a smoother continuous arc for daytime generation, with the mid-gap again aligning with missing samples.

9 Key Findings

- After cleaning, the merged data is consistent (Figures 3a–3b).
- Day/night solar cycle dominates trends and correlations (Figures 6–5).
- High AC power is uncommon overall ($\approx 10\%$) but likely under high irradiation (Figures 7–8).
- Polynomial (deg=2) predicts AC power well ($R^2_{\text{valid}} \approx 0.88$) with some cloudy-period errors (Figure 11).

10 Summary and Conclusions

- **Conclusion:** Irradiation + time-of-day explain most variability.
- **Limits/next:** One weather sensor + interpolation; add lag/rolling features and time-aware validation if extending.