

BengalGrid: A Decade-Scale Hourly Dataset of Bangladesh Power Supply by Source Type, Including Cross-Border Imports from India and Nepal

Md. Abir Rahman

abirrahman.sakin@gmail.com

East West University <https://orcid.org/0009-0006-3013-1249>

Rafi Ahamed

East West University <https://orcid.org/0009-0003-4543-8719>

Munaia Jannat Easha

East West University

Tasnia Tarannum Roza

East West University

Research Article

Keywords: Electricity generation, Bangladesh, PGCB, time series, fuel mix, cross-border electricity imports, web scraping, open dataset

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11381576/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

BengalGrid: A Decade-Scale Hourly Dataset of Bangladesh Power Supply by Source Type, Including Cross-Border Imports from India and Nepal

Md. Abir Rahman^{1*}, Rafi Ahamed¹, Munaia Jannat Easha¹, Tasnia Tarannum Roza¹
¹Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh.
*Corresponding author(s). E-mail(s): abirrahman.sakin@gmail.com;
Contributing authors: fidaahamed15@gmail.com; jannateasha004@gmail.com;
tasniatarannumroza28@gmail.com;

Abstract

Long, consistent, and machine-readable records of electricity supply are a prerequisite for forecasting, planning, and policy analysis, yet such records are scarce for Bangladesh. This paper presents BengalGrid, an hourly dataset of Bangladesh's electricity supply collected from the public Power Grid Company of Bangladesh (PGCB) ERP portal through a Python-based web scraping pipeline. The dataset contains 104,134 records covering April 17, 2015 to October 7, 2026, and reports total generation, fuel-wise generation (gas, liquid fuel, coal, hydro, solar, and wind), cross-border imports (Bheramara HVDC, Tripura, and Adani from India, and Nepal), and operational peak-hour remarks. We describe the complete collection and cleaning workflow, including Bangla-to-ASCII numeral conversion, timestamp standardization, and duplicate resolution, in which 438 duplicate timestamps were removed. Furthermore, we conduct a systematic technical validation covering temporal integrity, completeness, value plausibility, internal consistency, calendar and physical plausibility, and inter-variable structure. The validation shows that total generation equals the sum of domestic sources and imports to within 1 MW in 98.1% of the records where all components are available, and that the series reproduces expected diurnal, weekly, and seasonal patterns. It also documents structural missingness in newer sources and a small number of implausible values. Missing values are retained and outliers are not handled, so that users can apply their own preprocessing. BengalGrid supports research on generation forecasting, fuel-mix assessment, import analysis, peak-hour behavior, and anomaly detection, and it is openly available through Mendeley Data (version 2) under a CC BY 4.0 license.

Keywords: Electricity generation, Bangladesh, PGCB, time series, fuel mix, cross-border electricity imports, web scraping, open dataset

1 Introduction

Electricity is a foundation of modern economic and social life, and the way a country generates and imports its power shapes its costs, emissions, and energy security. Bangladesh has experienced a steady rise in electricity supply over the last decade, and its supply mix has been changing at the same time. Gas has traditionally dominated generation, while coal, imported electricity, and renewable sources have gained importance in recent years [2]. Understanding this transition requires long, consistent, and machine-readable records of generation by source.

The Power Grid Company of Bangladesh (PGCB) publishes generation records through its public ERP portal [1]. However, these records are presented as paginated HTML tables, with Bangla numerals in the date and time fields, mixed time formats, and repeated entries. Therefore, they cannot be directly used for time-series analysis, forecasting, or machine learning. To the best of our knowledge, a curated, openly documented, and analysis-ready version of these records has not been widely available to the research community.

This paper introduces **BengalGrid**, a dataset that converts the public PGCB records into a clean and well-documented time series [2]. The dataset spans more than eleven years, from April 2015 to October 2026, and contains 104,134 records after cleaning. It provides total generation, generation by fuel type, imports through four interconnection channels, and the peak-hour

remarks published by the source. Missing values are retained instead of being replaced, and no outlier correction is applied, so that researchers can make their own informed modeling decisions. The documentation follows the FAIR principles [9] and the datasheet recommendations [10].

Beyond releasing the data, this paper also reports a comprehensive technical validation of BengalGrid. We examine its temporal integrity, completeness, value plausibility, internal consistency, calendar and physical plausibility, and statistical relationships between variables. In this way, we aim to give future users a clear picture of both the opportunities and the limitations of the data.

The contribution of our paper is as follows.

- **Dataset Release:** Presents BengalGrid, an hourly record of Bangladesh’s total generation, fuel-wise generation, and cross-border imports from April 2015 to October 2026, publicly available on Mendeley Data under a CC BY 4.0 license.
- **Reproducible Pipeline:** Documents the full collection and cleaning workflow, including the scraping script, Bangla numeral conversion, timestamp standardization, and duplicate handling, so that the dataset can be reproduced and extended.
- **Technical Validation:** Provides a systematic analysis of timestamp regularity, missing values, outliers, and component consistency, and quantifies the agreement between total generation and its components.
- **Usage Guidance:** Identifies the data-quality issues that require attention, such as extreme values and structural missingness, and offers practical recommendations and a short screening example for analysis.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the data collection and processing pipeline. Section 4 describes the structure and content of the dataset. Section 5 presents the technical validation. Sections 6 and 7 provide usage notes, limitations, and potential applications, and Section 8 concludes the paper.

2 Related Works

Open and well-curated datasets have been a major driver of progress in energy analytics. The Global Energy Forecasting Competitions provided widely used benchmark data for load, price, wind, and solar forecasting, and encouraged the development of probabilistic forecasting methods [3]. Similarly, the Open Power System Data initiative collected, cleaned, and published time-series data on power systems in several European countries, demonstrating the value of transparent processing and documentation [4]. These efforts show that the usefulness of an energy dataset depends not only on its coverage but also on how clearly its provenance and processing are described.

Time-series forecasting and electricity market analysis are mature research areas [5, 6]. However, such methods depend on consistent historical records. In many regions, the available data is published in formats that are suitable for display but not for analysis. Furthermore, real-world operational data often contains missing values, entry errors, and structural changes in reporting. Statistical treatments of missing data emphasize that the mechanism behind the missingness should be understood before any imputation is applied [7]. Likewise, the anomaly detection literature stresses that outliers can represent either errors or genuine events, and that this distinction should be made carefully [8].

Dataset documentation practices have also received attention. The FAIR principles promote findable, accessible, interoperable, and reusable data [9], while datasheets for datasets recommend that creators describe the motivation, composition, collection process, and recommended uses of a dataset [10]. Exploratory data analysis, in the tradition of Tukey [11], remains an essential first step for understanding the shape and quality of a dataset.

Previous efforts mainly focus on regions where operators publish structured data. In this research, the focus is on the Bangladesh power sector, where public generation records exist but

are not available in a ready-to-use form. Furthermore, this work follows the documentation practices described above by reporting the collection process, the cleaning rules, and an extensive validation of the final dataset.

3 Data Collection and Processing

In this section, we thoroughly discuss how BengalGrid was collected from the PGCB ERP portal and how the raw records were converted into the final dataset. Figure 1 summarizes the complete workflow.

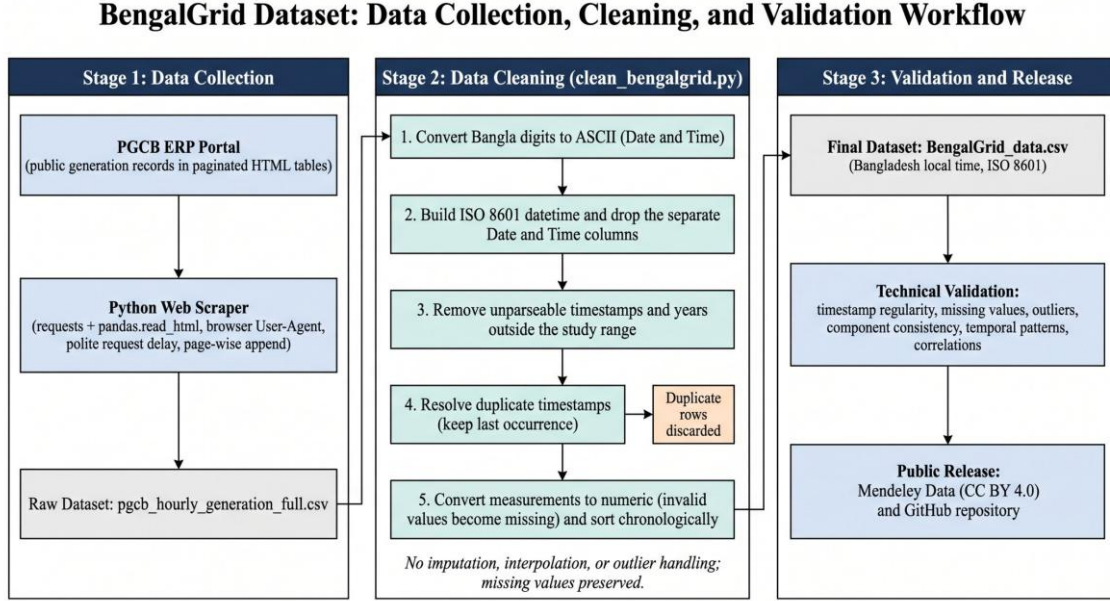


Figure 1: Workflow of the BengalGrid dataset, from web scraping of the PGCB ERP portal to cleaning, duplicate handling, validation, and publication.

3.1 Data Source

The records were collected from the official PGCB ERP portal [1], which publicly presents generation records as a paginated table¹. The portal covers April 2015 to October 2026 in the version used for this study, and the table spans 2,051 pages. Each row of the table contains the date, the time, the total generation, the generation of each fuel type, the imports through each interconnection, and an optional remark.

3.2 Web Scraping

A Python 3 script (`scrape_bengalgrid.py`) was developed using the `requests` library and `pandas.read_html()`. The script requests each page from 1 to 2,051 within a persistent session, extracts the first HTML table of the page, and appends it to a single CSV file (`pgcb_hourly_generation_full.csv`). To reduce the load on the server, a browser-like User-Agent header and a one-second delay between requests were used. If a request failed or a page contained no table, the script reported the page number and continued. The raw collection contained 104,582 rows and 14 columns.

3.3 Data Cleaning

The raw file was processed by a second script (`clean_bengalgrid.py`), which applies the following steps in order.

¹ https://erp.powergrid.gov.bd/w/generations/view_generations

- **Numerical Conversion:** Bangla (Bengali-script) digits in the `Date` and `Time` fields were converted to ASCII digits using `str.maketrans()` and `str.translate()`.
- **Datetime Construction:** Times written in `HH:MM` format were extended with `:00`, and the date and time were combined into a single timestamp, which was stored in ISO 8601 format (`YYYY-MM-DDTHH:MM:SS`) in the `datetime` column. The original `Date` and `Time` columns were then removed, which reduces the number of columns from 14 to 13.
- **Date Filtering:** Rows with unparseable timestamps and rows whose year falls outside 2015–2026 were removed.
- **Duplicate Resolution:** Where several rows shared the same timestamp, only the last occurrence in file order was retained. A total of 438 duplicate rows were removed in this step.
- **Numeric Conversion:** All measurement columns were converted to numeric types. Entries that could not be parsed became missing values.
- **Sorting:** The records were sorted chronologically and written to `BengalGrid_data.csv`.

Table 1 summarizes the number of records at each stage. The 10 rows removed before duplicate resolution ($104,582 - 438 = 104,134$) correspond to unparseable timestamps or years outside the configured range.

Table 1: Record counts at each stage of the processing pipeline.

Stage	Rows	Columns
Raw scraped table (<code>pgcb_hourly_generation_full.csv</code>)	104,582	14
Invalid timestamps or years outside 2015–2026 removed	10	–
Duplicate timestamps removed (last occurrence kept)	438	–
Final dataset (<code>BengalGrid_data.csv</code>)	104,134	13

3.4 Design Decisions

Several design decisions were made deliberately. First, missing values are preserved and are not replaced with zero, because the mechanism behind a missing value should be understood before any imputation [7], and a blank value in a variable such as `Solar` or `Nepal` may mean that the source was not yet reported rather than that its output was zero. Second, no imputation, interpolation, or outlier handling is performed, so that the dataset reflects the source as published and every correction remains a transparent, user-level decision. Third, half-hour readings (at 18:30 and 19:30) are retained as published. Researchers who need a strictly hourly series can exclude records whose minute component is not 00.

4 Dataset Description

4.1 Overview

BengalGrid contains 104,134 records and 13 columns, covering the period from April 17, 2015 18:30 to October 7, 2026 00:00. Timestamps are in Bangladesh local time (UTC+6). All timestamps are valid and unique. There are no negative values, no infinite values, and no constant columns. Table 2 gives the key characteristics of the dataset, and Table 3 describes each column.

Table 2: BengalGrid at a glance.

Property	Value
Geographic scope	Bangladesh (national level)
Original source	PGCB ERP portal (public generation records)
Temporal coverage	17 April 2015 18:30 to 7 October 2026 00:00 (Bangladesh local time, UTC+6)
Temporal resolution	Hourly, with additional half-hour readings (4.5% of records)
Records and variables	104,134 records; 13 columns (timestamp, total generation, six generation sources, four import channels, and one remark)
Variable groups	Total generation; six generation sources; four import channels; peak-hour remarks
Missing cells	364,059 (26.89%), concentrated in five columns
File format	Comma-separated values (CSV), UTF-8
License	CC BY 4.0
Repository	Mendeley Data, version 2 (V2), DOI: 10.17632/xx9dhmy9rj.2

Table 3: Description of the columns of BengalGrid. Non-missing counts refer to the final dataset of 104,134 records.

Column	Unit	Description	Non-missing
datetime	ISO 8601	Date and time of the observation (Bangladesh local time)	104,134
Generation_MW	MW	Total national electricity generation dispatched to the grid at the recorded time	104,134
Gas	MW	Generation from gas-fired plants	104,134
Liquid Fuel	MW	Generation from liquid-fuel-based plants	104,134
Coal	MW	Generation from coal-fired plants	104,134
Hydro	MW	Generation from hydropower	104,134
Solar	MW	Generation from solar power	82,237
Wind	MW	Generation from wind power	30,562
India_Bheramara_HVDC	MW	Import through the Bheramara HVDC interconnection with India	104,134
India_Tripura	MW	Import through the Tripura interconnection with India	104,134
India_Adani	MW	Import through the Adani power supply connection with India	19,239
Nepal	MW	Electricity imports attributed to Nepal	17,254
Remarks	text	<i>Day Peak</i> or <i>Evening Peak</i> marks the daytime or evening peak-demand hour; blank otherwise	7,319

4.2 Data Records

The dataset is deposited in Mendeley Data [2] (version 2, DOI: 10.17632/xx9dhmy9rj.2) under a CC BY 4.0 license, and the code is also available on GitHub². The files are listed in Table 4. Providing the raw file and both scripts allows users to audit and reproduce every processing step.

Table 4: Files included in the BengalGrid repository.

File	Description
BengalGrid_data.csv	Final cleaned dataset (104,134 records, 13 columns)
pgcb_hourly_generation_full.csv	Original scraped dataset (104,582 rows, 14 columns)
scrape_bengalgrid.py	Data collection script
clean_bengalgrid.py	Data cleaning script

5 Technical Validation and Exploratory Analysis

This section presents a systematic validation of BengalGrid, in the tradition of exploratory data analysis [11], along six dimensions: (i) temporal integrity, (ii) completeness, (iii) value plausibility, (iv) internal consistency between total generation and its components, (v) calendar and physical plausibility of the temporal patterns, and (vi) the statistical structure between variables. All statistics are computed directly from the released file without any correction. Where a figure is drawn to show typical behavior despite extreme values, this is stated in the caption, and the underlying data remain unchanged.

5.1 Timestamp Regularity

The dataset is mainly an hourly series. Among the 104,133 intervals between consecutive timestamps, 94,602 (90.9%) are 60 minutes and 9,293 (8.9%) are 30 minutes. The 30-minute intervals arise from 4,651 additional half-hour readings (4.5% of all records), which occur almost exclusively at 19:30 (4,127 records) and 18:30 (524 records). The remaining 238 intervals are longer gaps, ranging from 90 minutes to about 21.6 days (Table 5). Only three gaps exceed one day, and the longest occurs between January 18, 2017 11:00 and February 9, 2017 01:00 (31,080 minutes), which is visible as a break in Figure 7b. Gaps are most frequent in the early part of the record: 155 of the 238 (65%) fall in 2015–2019, and the yearly count declines to 7 in 2025 and 2 in 2026. Figure 2 shows the distribution of the dominant gaps.

Table 5: Distribution of the time between consecutive timestamps.

Interval	Count	Share (%)
30 minutes	9,293	8.92
60 minutes	94,602	90.85
Longer than 60 minutes	238	0.23
Total intervals	104,133	100.00

² <https://github.com/abirrahman26/BengalGrid-Dataset>

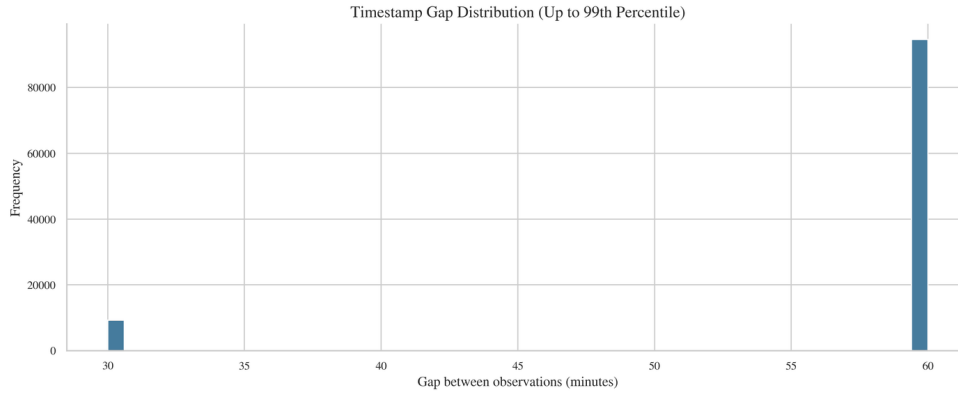


Figure 2: Distribution of gaps between consecutive observations up to the 99th percentile. Hourly readings dominate, and half-hour readings form the second group.

5.2 Missing Values

In total, 364,059 cells (26.89%) are missing. The missingness is concentrated in five columns (Table 6), whereas `datetime`, `Generation_MW`, `Gas`, `Liquid Fuel`, `Coal`, `Hydro`, `India_Bheramara_HVDC`, and `India_Tripura` are complete. Figure 3 presents these results.

Table 6: Missing values by variable, with the first timestamp at which each variable is reported.

Variable	Missing	Missing (%)	Non-missing	First reported
Remarks	96,815	92.97	7,319	2015-04-17
Nepal	86,880	83.43	17,254	2024-11-14
India_Adani	84,895	81.53	19,239	2024-08-28
Wind	73,572	70.65	30,562	2023-06-01
Solar	21,897	21.03	82,237	2017-09-20

The missingness is not random, and the missingness mechanism determines how it should be treated [7]. The missingness pattern (Figure 3b) and the yearly missingness (Figure 4) show that it is structural: each of `Solar`, `Wind`, `India_Adani`, and `Nepal` is entirely missing until a particular date and then becomes available almost continuously. `Solar` is first reported on September 20, 2017, `Wind` on June 1, 2023, `India_Adani` on August 28, 2024, and `Nepal` on November 14, 2024, and each of them is reported in more than 95% of the records from the following month onward (June 2023 for `Wind`). From January 2025 onward, all of these columns are fully populated (16,092 records). The `Remarks` column is empty for most records by design, because it marks only the peak-demand hours. Therefore, a missing value in these columns most often indicates that the source was not yet reported by the portal, and it should not be treated as zero output.

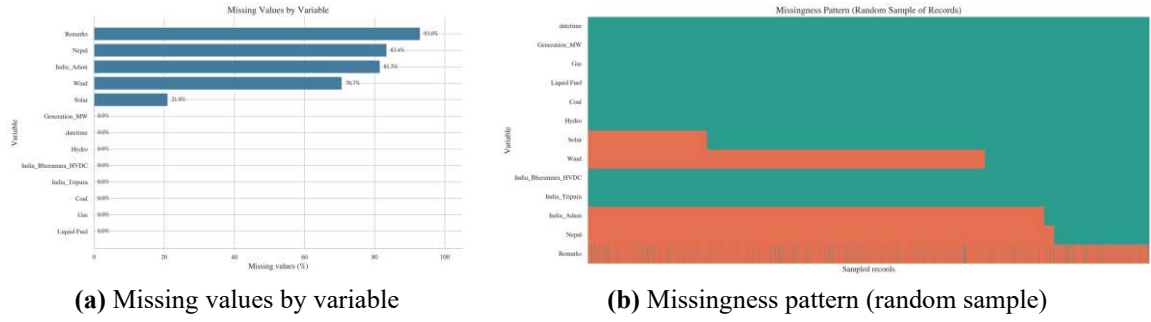


Figure 3: Missing values in BengalGrid. Missingness is structural and is concentrated in Remarks, Nepal, India_Adani, Wind, and Solar.

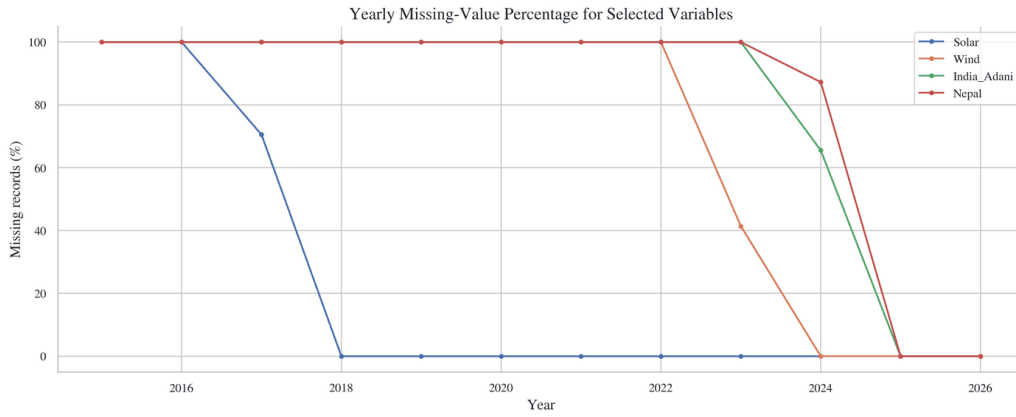


Figure 4: Yearly percentage of missing records for Solar, Wind, India_Adani, and Nepal.

5.3 Descriptive Statistics

Table 7 summarizes the numerical variables. The medians are in line with what is expected for the national grid, for example a median total generation of 8,823 MW, a median gas generation of 5,021 MW, and a median Bheramara HVDC import of 731 MW. However, several maxima and standard deviations are extremely large. For example, the maximum of `Generation_MW` is 64,526,500 MW and the maximum of `Liquid Fuel` is 29,222,897 MW. These values are physically implausible for the national grid, and they dominate the means, standard deviations, skewness, and kurtosis. Consequently, robust statistics such as the median and the percentiles in Table 8 are more informative for these variables. Figure 5 shows the distribution of each variable after limiting the displayed range to the 0.1st–99.9th percentiles, and this range restriction is applied to the figure only.

Table 7: Descriptive statistics of the numerical variables (MW).

Variable	N	Mean	Median	Std.	Min	Max
Generation_MW	104,134	9,788.2	8,823	199,949.7	73	64,526,500
Gas	104,134	5,093.3	5,021	1,102.3	0	74,818
Liquid Fuel	104,134	1,947.3	1,483	90,629.3	0	29,222,897
Coal	104,134	1,331.3	550	1,550.1	0	31,687
Hydro	104,134	101.1	80	73.2	0	5,623
Solar	82,237	61.2	0	132.2	0	2,998
Wind	30,562	9.9	4	15.6	0	922
India_Bheramara_HVDC	104,134	683.6	731	337.6	0	76,292
India_Tripura	104,134	94.2	98	47.4	0	1,565
India_Adani	19,239	1,003.9	1,008	351.5	0	13,609
Nepal	17,254	12.5	0	17.9	0	48

Table 8: Percentiles of the numerical variables (MW). The 1st–99th percentile range describes the bulk of the data and is not affected by the isolated extreme values.

Variable	P1	P25	Median	P75	P99
Generation_MW	4,232	7,023	8,823	11,201	15,388
Gas	3,308	4,415	5,021	5,743	7,071
Liquid Fuel	0	675	1,483	2,402	4,971
Coal	0	166	550	2,492	5,323
Hydro	0	42	80	156	238
Solar	0	0	0	36	592
Wind	0	0	4	14	58
India_Bheramara_HVDC	134	476	731	908	953
India_Tripura	0	68	98	128	176
India_Adani	413	759	1,008	1,276	1,493
Nepal	0	0	0	38	38

The percentiles reveal the characteristic shape of each variable. Gas is approximately symmetric around 5,000 MW, whereas Coal is bimodal, with a large group of low values (the years before coal expanded) and a second group around 3,000 to 5,000 MW (Figure 5). The Bheramara HVDC import is concentrated near its upper range (the 75th and 99th percentiles are 908 and 953 MW), which suggests that the link often operates near a capacity ceiling. Solar and Wind are strongly right-skewed with many zero values, and Nepal takes only two dominant levels, 0 and about 38 MW.

The validity checks show that no variable contains negative or infinite values and that none is constant. Zero values are natural for some variables: 53.1% of the non-missing Solar values are zero, which is expected at night, and 67.0% of the non-missing Nepal values are zero, indicating that this import channel was inactive for most of the period in which it is reported.

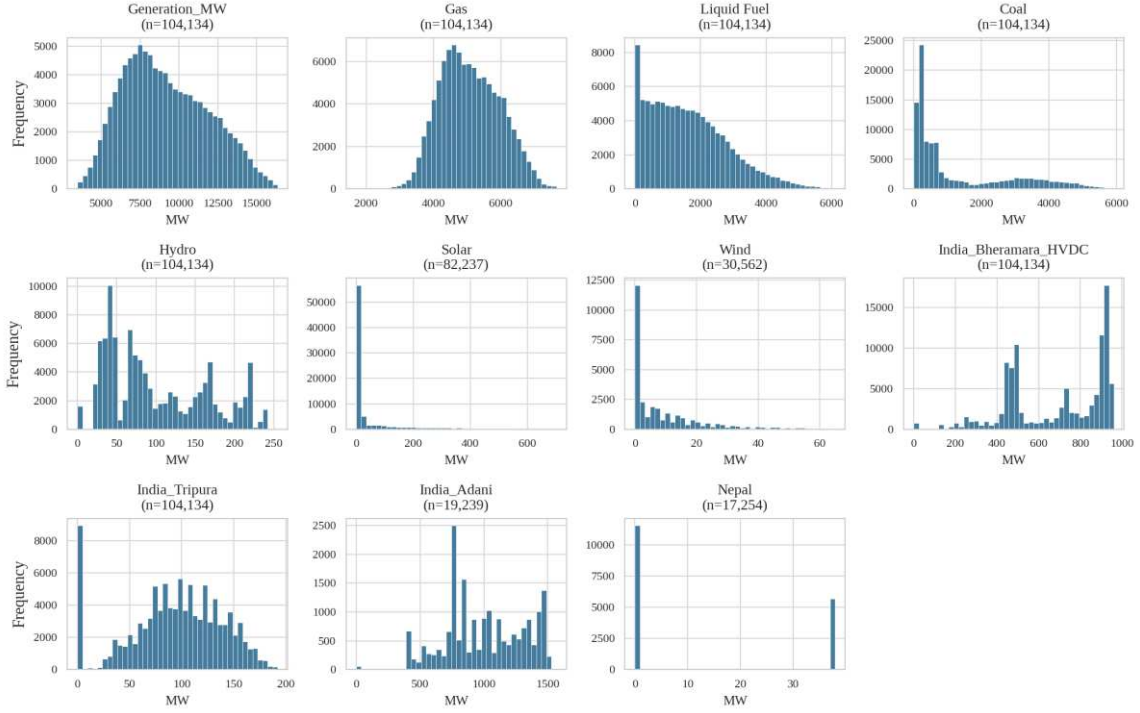


Figure 5: Distribution of each numerical variable. For display only, the horizontal range of each panel is limited to the 0.1st–99.9th percentiles of the non-missing values; no records are removed from the dataset.

5.4 Outlier Analysis

We screened every variable with the interquartile range (IQR) rule [11], flagging values outside $[Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR}]$. The results are given in Table 9. For `Generation_MW`, only 7 records are flagged: one very high value of 64,526,500 MW (January 15, 2023 at 05:00) and six unusually low values between 73 and 746 MW (on 2016-09-16, 2016-12-09, 2017-10-16, 2018-01-04, 2018-01-18, and 2020-12-24). The highest total-generation value that is in line with the rest of the series is 17,200 MW, recorded on May 20, 2026 at 21:00 during an evening peak.

Table 9: IQR-based outlier screening. Bounds and maxima are in MW.

Variable	Lower	Upper	Flagged	Flagged (%)	Observed max
Generation_MW	756.0	17,468.0	7	0.007	64,526,500
Gas	2,423.0	7,735.0	243	0.233	74,818
Liquid Fuel	-1,915.5	4,992.5	1,004	0.964	29,222,897
Coal	-3,322.6	5,980.4	71	0.068	31,687
Hydro	-129.0	327.0	85	0.082	5,623
Solar	-54.0	90.0	15,483	18.827	2,998
Wind	-21.0	35.0	1,984	6.492	922
India_Bheramara_HVDC	-172.0	1,556.0	19	0.018	76,292
India_Tripura	-22.0	218.0	50	0.048	1,565
India_Adani	-16.5	2,051.5	5	0.026	13,609
Nepal	-57.0	95.0	0	0.000	48

The flagged proportions for Solar (18.8%) and Wind (6.5%) should be interpreted carefully. These variables are strongly right-skewed with many zero values, so the IQR bounds are very tight, and ordinary daytime solar output or windy periods are flagged. In contrast, the isolated

values far beyond the normal range, such as 74,818 MW for Gas, 31,687 MW for Coal, 29,222,897 MW for Liquid Fuel, and 76,292 MW for India_Bheramara_HVDC, are physically implausible. The boxplots in Appendix A illustrate these cases for each variable. The extreme values affect only a very small share of the records, but they strongly distort moment-based statistics, and whether such values are errors or genuine events must be judged case by case [8]. As stated above, outliers are not handled in this release, and they require validation by the user before analysis.

5.5 Consistency of Generation Components

An important question is how the reported total relates to the individual sources. Table 10 compares `Generation_MW` with two sums: the domestic sources alone, and the domestic sources plus imports. When imports are not included, the total exceeds the domestic source sum by 1,556 MW on average (median 1,655 MW), and essentially none of the records agree within 1 MW. When imports are included, the residual is centered at zero (median 0 MW, mean -3.1 MW), and 98.1% of the records agree within 1 MW and 98.7% within 10 MW. Therefore, `Generation_MW` is the total supply dispatched to the grid, i.e., domestic generation plus imports. The positive residual of roughly 500 to 2,500 MW in Figure 6 is the contribution of the imports. The complete comparison is possible only for 17,254 records, for which all components (including Wind, India_Adani, and Nepal) are available. The comparison with domestic sources alone uses the 30,562 records with a reported Wind value. This result also provides an independent check of the data: the columns were scraped separately from a single table, and their arithmetic agreement indicates that the scraping and cleaning steps preserved the alignment between columns.

Table 10: Consistency between total generation and the sum of its components (MW).

Comparison	N	Mean	Median	RMSE	≤ 1 MW (%)	≤ 10 MW (%)
Total – domestic sources	30,562	1,556.4	1,655	1,760.3	0.00	0.00
Total – (domestic + imports)	17,254	-3.1	0	202.5	98.12	98.74

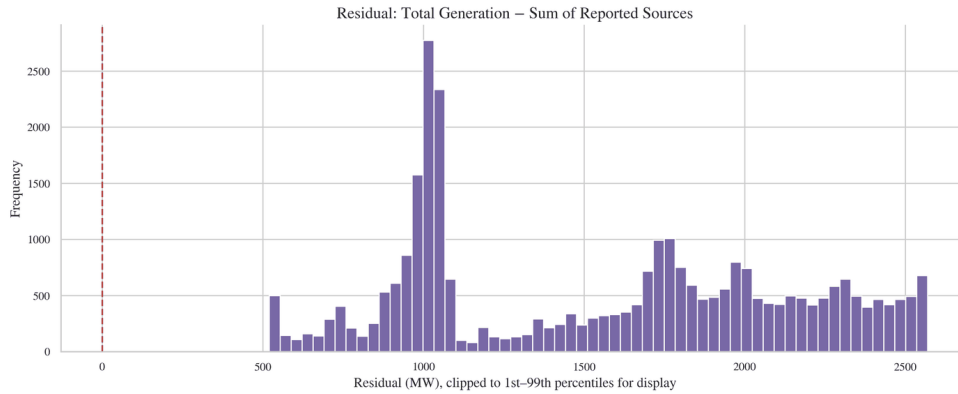


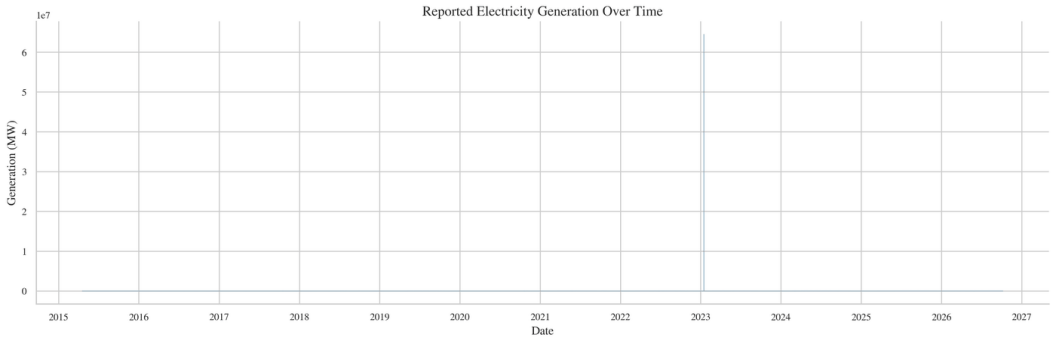
Figure 6: Residual between total generation and the sum of the reported domestic sources. The residual is always positive because imports are not part of the domestic sum.

5.6 Long-Term Trend of Total Generation

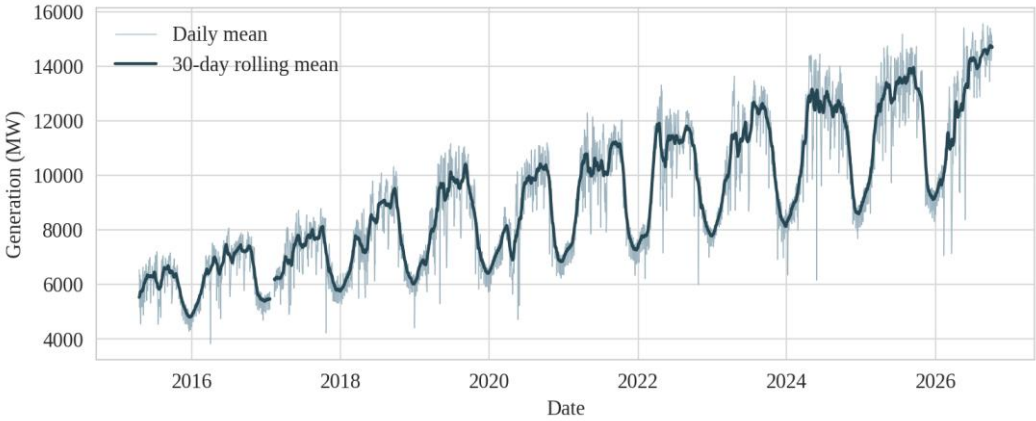
Figure 7a shows the complete series on its raw scale. It is dominated by the single extreme record of January 2023, so that the rest of the series appears flat. For this reason, Figure 7b shows the same series after the seven records flagged in Table 9 are hidden for display only. It reveals three features of the data: a strong long-term growth, a regular annual cycle, and a small number of short breaks, such as the one in early 2017.

Robust statistics quantify the growth (Table 11). The annual median of total generation increases from 6,664 MW in 2016 to 12,282 MW in 2025, the two complete calendar years at the ends of

the record, which corresponds to a compound annual growth rate of about 7.0%. The partial years 2015 (from April 17) and 2026 (until October 7) have medians of 6,063 and 13,200 MW, respectively. The monthly median (Figure 8) follows the same pattern, whereas the monthly mean shows a spike in January 2023. This difference illustrates why medians are preferable to means for this dataset.



(a) Raw series



(b) Daily mean and 30-day rolling mean (seven flagged records hidden for display only)

Figure 7: Total generation over time. (a) The extreme record in January 2023 compresses the scale. (b) Without it, growth, seasonality, and a data break in early 2017 are visible.

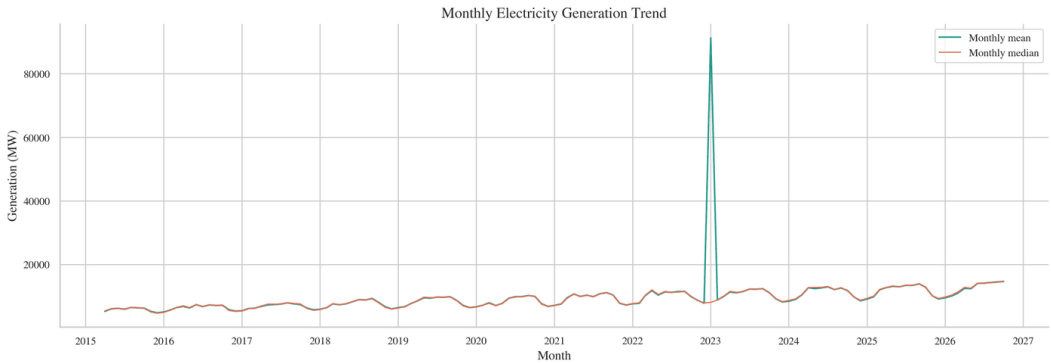


Figure 8: Monthly mean and median of total generation. The median is robust to the extreme record, whereas the mean is not.

Table 11: Annual summary of the main variables. Values are annual medians in MW; N is the number of records. 2015 and 2026 are partial years, and “–” means that fewer than 500 values are available.

Year	N	Total	Gas	Liquid Fuel	Coal	Bheramara	Adani
2015	6,574	6,063	4,253	1,031	129	445	–
2016	9,384	6,664	4,336	1,519	160	455	–
2017	8,421	7,081	4,575	1,628	141	492	–
2018	9,039	7,704	5,022	1,640	187	484	–
2019	9,030	8,369	5,965	1,066	222	840	–
2020	9,133	8,409	5,653	1,252	580	799	–
2021	9,112	9,477	6,087	2,114	559	730	–
2022	9,109	10,279	5,212	3,033	875	913	–
2023	9,102	10,724	5,621	1,559	2,256	907	–
2024	9,138	11,370	5,073	1,202	3,264	900	818
2025	9,118	12,282	5,270	906	3,628	894	971
2026	6,974	13,200	4,922	1,170	4,520	895	1,104

5.7 Daily, Weekly, and Seasonal Patterns

A useful test of whether a dataset captures real system behavior is whether it reproduces known calendar and physical regularities. Diurnal, weekly, and seasonal cycles are standard characteristics of electricity time series [3, 6]. Figure 9 and Figure 10 summarize the diurnal, weekly, and seasonal structure using medians, which are not affected by the extreme records.

Diurnal pattern. The median total generation is lowest at about 06:00 (about 7,500 MW), increases during the day, and shows a distinct step at 19:00 to a plateau of about 10,100 to 10,200 MW between 19:00 and 21:00 (Figure 9a). This evening plateau is consistent with the evening peak-demand hours marked in the `Remarks` column (Section 5.11).

Weekly pattern. Using the complete years 2016–2025, the median is almost identical from Sunday to Thursday (8,998 to 9,009 MW), lower on Saturday (8,689 MW, about 3.5% below), and lowest on Friday (8,253 MW, about 8.4% below) (Figure 9b). This is consistent with Friday being the weekly holiday in Bangladesh. The hour–weekday heatmap (Figure 10a) shows that the Friday reduction is concentrated in the daytime hours from about 06:00 to 18:00, while the evening peak remains visible on all days.

Seasonal pattern. Dividing each month’s median by the annual median of the same year and averaging over 2016–2025 gives a seasonal index (Figure 9c). The index is lowest in January (0.76) and December (0.78) and highest in August (1.13) and September (1.15), so the median supply in the warmer months is about 50% higher than in the winter trough. The year–month heatmap (Figure 10b) shows the same cycle superimposed on the long-term growth, with the highest values in the summer months of 2025 and 2026 (above 14,000 MW).

Physical plausibility of solar output. The median `Solar` value is zero from 18:00 through 06:00, rises from 07:00, and reaches its maximum of 133 MW at 12:00 (Figure 9d). This bell-shaped profile centered on noon is the expected signature of photovoltaic generation, and it supports the correctness of the `Solar` column and the timestamps (Bangladesh local time).

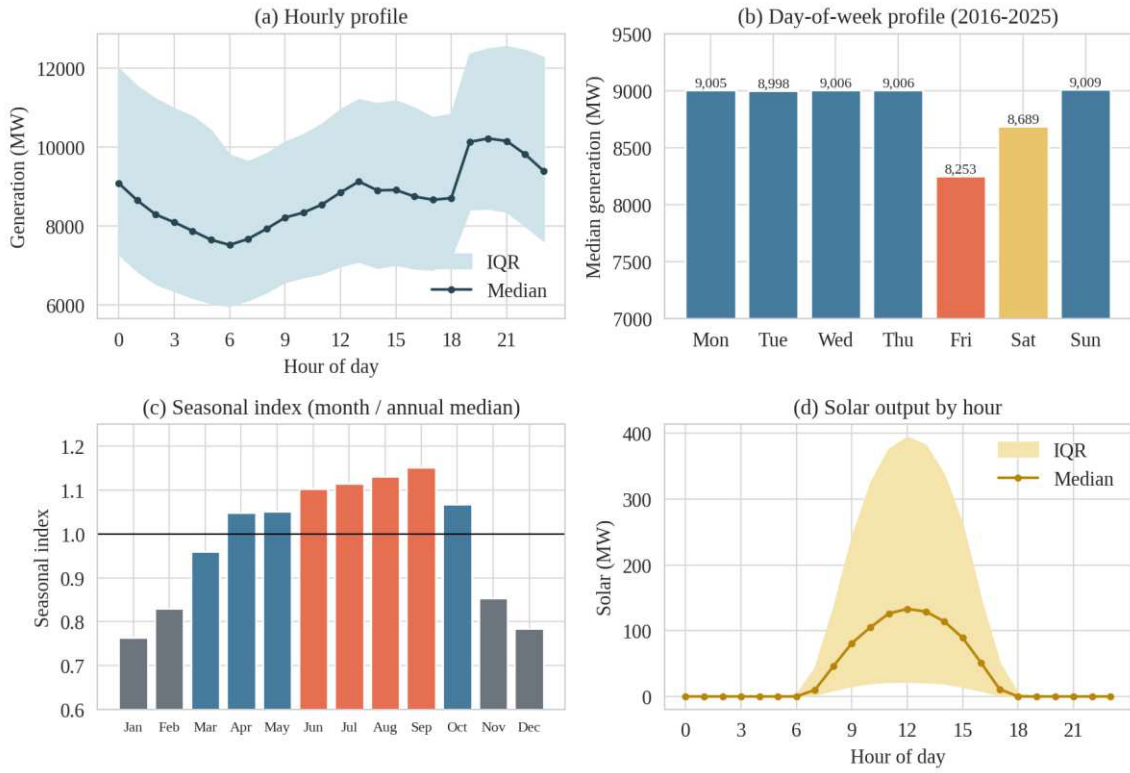


Figure 9: Median profiles of the dataset. (a) Hourly profile of total generation with the interquartile range (IQR). (b) Day-of-week profile for 2016–2025. (c) Seasonal index of total generation. (d) Hourly profile of solar generation. Records flagged in Table 9 are hidden for display only.

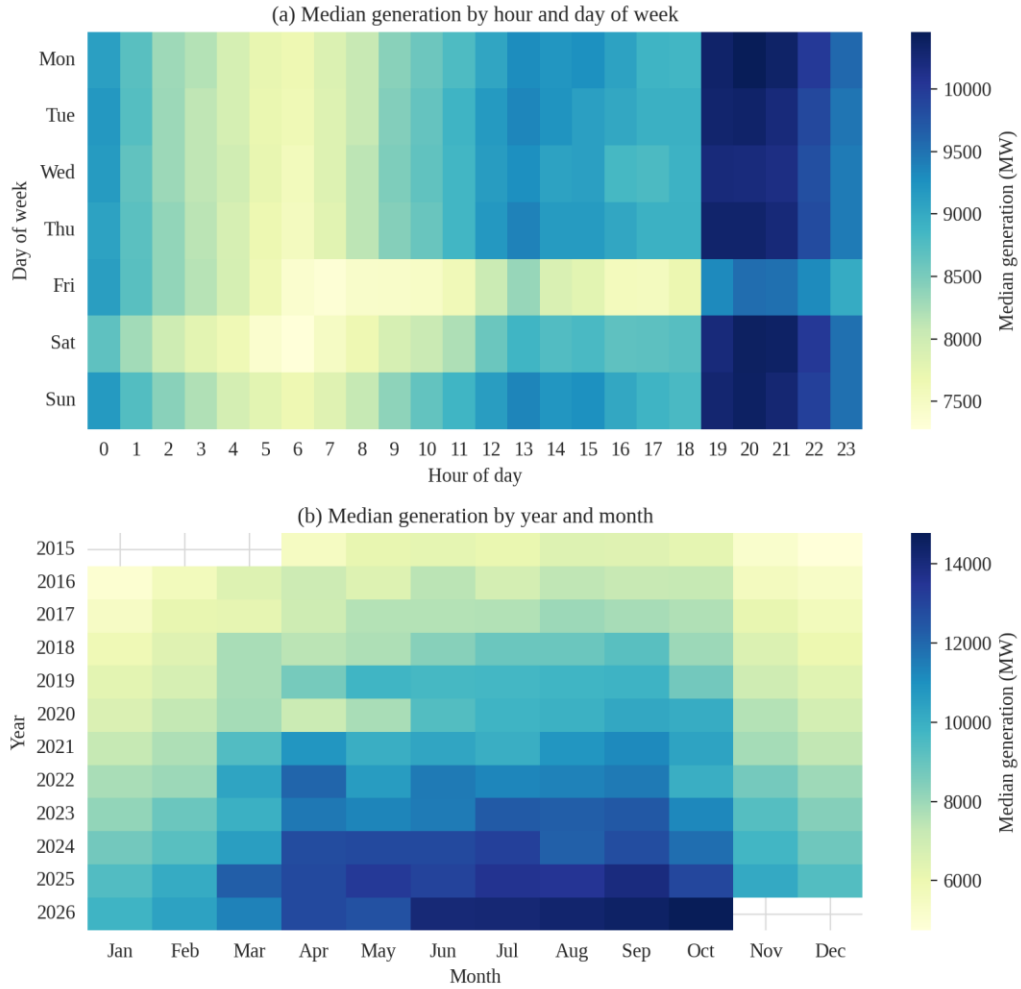


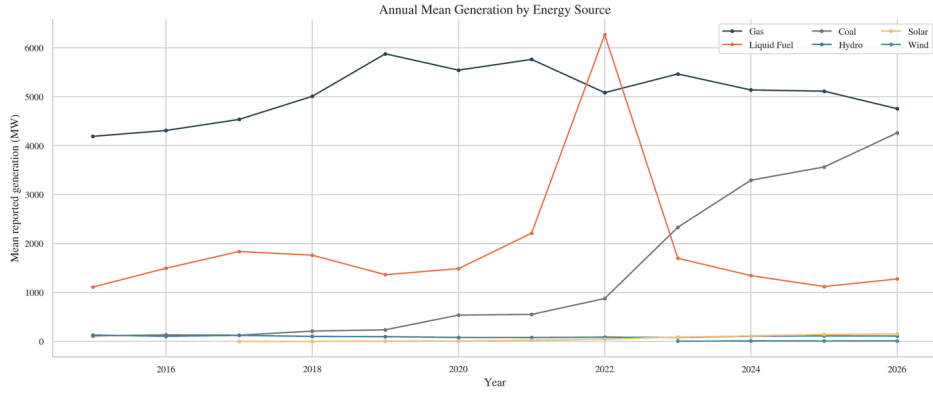
Figure 10: Heatmaps of median total generation by (a) hour and day of week and (b) year and month. Records flagged in Table 9 are hidden for display only; blank cells in (b) are months outside the coverage period.

5.8 Evolution of the Source Mix

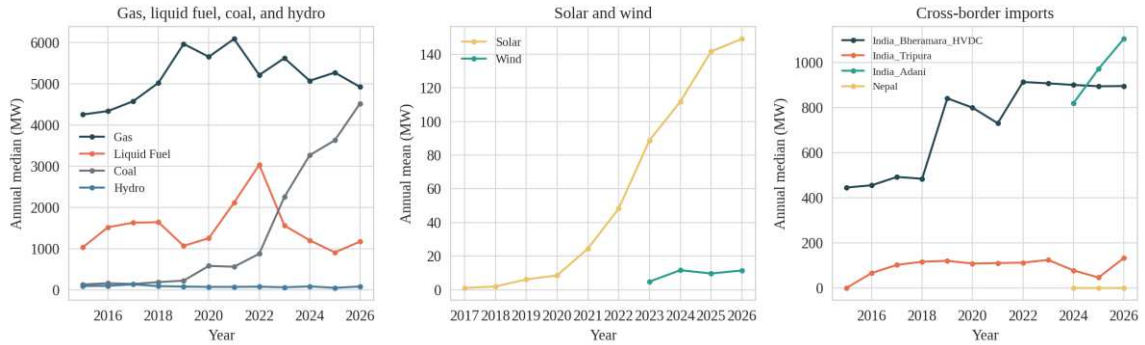
The fuel-wise columns reveal a clear structural change in Bangladesh's supply (Figure 11 and Table 11). The annual median of gas generation rises from 4,253 MW in 2015 to 6,087 MW in 2021 and then declines to 4,922 MW in 2026. Coal grows from a very small contribution (129 MW in 2015) to 4,520 MW in 2026, and the largest increase occurs between 2022 and 2024. In the most recent months of the source-share chart (Figure 12), the share of coal is comparable to that of gas: from mid-2023 to 2026, gas falls from about 60% to about 38% of the observed source total, while coal rises from about 17% to about 38%.

Liquid fuel requires a careful reading. The annual mean peaks at about 6,300 MW in 2022 (Figure 11a), but this value is inflated by implausibly large records. The annual median also peaks in 2022 (3,033 MW, compared with 2,114 MW in 2021 and 1,559 MW in 2023; left panel of Figure 11b), which suggests a genuine temporary increase in liquid-fuel generation. This increase coincides with a decline in the median gas generation (6,087 MW in 2021 and 5,212 MW in 2022). We report this as an observation and do not attempt to explain its cause with the available data.

Solar and wind grow from a very small base (middle panel of Figure 11b). The annual mean of `Solar` (with values above the 99.9th percentile excluded for display) increases from about 1 MW in 2017 to about 149 MW in 2026, and the annual mean of `Wind` is about 5 MW in 2023 and about 11 MW in 2026. Together, they remain at roughly 1 to 2% of the observed source total, and hydro stays small, with an annual median between 46 and 132 MW.



(a) Annual mean generation by source (affected by extreme values)



(b) Robust annual statistics by source group and import channel

Figure 11: Annual generation by source and import channel. (a) The liquid fuel peak in 2022 is partly caused by implausibly large values. (b) Left: annual medians of gas, liquid fuel, coal, and hydro; middle: annual means of solar and wind, with values above the 99.9th percentile excluded for display only and years with fewer than 1,000 values omitted; right: annual medians of the import channels, shown for years with at least 500 values.

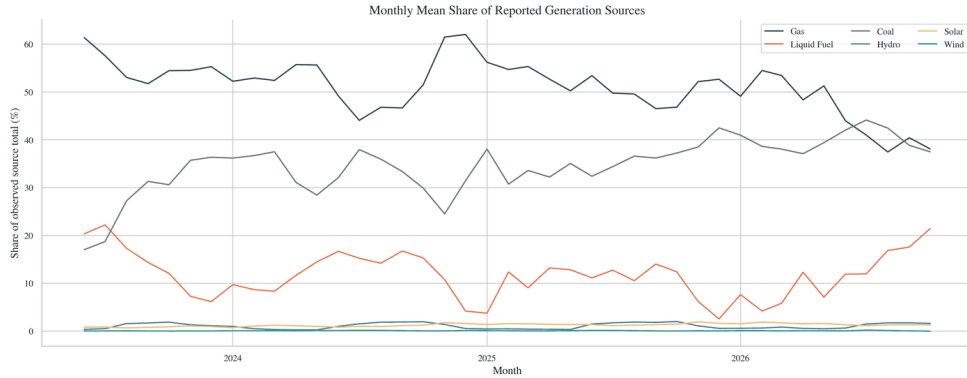


Figure 12: Monthly mean share of the reported generation sources (from the period in which all sources are reported).

5.9 Cross-Border Imports

The right panel of Figure 11b and Table 11 describe the imports using annual medians. The Bheramara HVDC link is the most significant long-running channel. Its median is between 445 and 492 MW in 2015–2018, rises to 840 MW in 2019, and then remains in the range of about 730 to 913 MW until 2026 (895 MW), which indicates that the link has operated near a stable level since 2019. The Adani connection is reported from 2024 and increases from 818 MW in 2024 to 971 MW in 2025 and 1,104 MW in 2026. The Tripura interconnection is smaller, with a median of about 100 to 124 MW between 2017 and 2023, a decline to 78 MW in 2024 and 46 MW in 2025, and a recovery to 132 MW in 2026. The Nepal channel has a median of 0 MW in all years because it is inactive for most of the time, and it supplies about 38 MW when active. In

2026, the three Indian channels together have a combined median of roughly 2,100 MW, which is approximately one-sixth of the median total generation (13,200 MW). This confirms that the import columns are an essential part of the supply picture and justifies their inclusion in the dataset.

5.10 Correlation Analysis

Pearson and Spearman correlation matrices are shown in Figure 13. The two matrices disagree strongly for the total generation. For example, the Pearson correlation between `Generation_MW` and `Coal` is 0.01, whereas the Spearman correlation is 0.76, and for `India_Bheramara_HVDC` the values are 0.01 and 0.72, respectively. This difference is a typical symptom of extreme outliers, which dominate the linear correlation but have little effect on rank correlation. Therefore, the Spearman matrix is the more reliable summary for this dataset. According to it, total generation is strongly associated with coal (0.76) and Bheramara imports (0.72), and moderately associated with gas (0.59), Adani imports (0.59), liquid fuel (0.52), Nepal (0.46), and Tripura (0.46). Part of these associations reflects the common upward trend over time, so they should not be interpreted as causal relationships. Hydro and Nepal are strongly correlated with each other (Spearman 0.69), which is plausible given the seasonal pattern of hydropower. We caution that the correlations involving Wind, India_Adani, and Nepal are computed on pairwise-available records from the later years only.

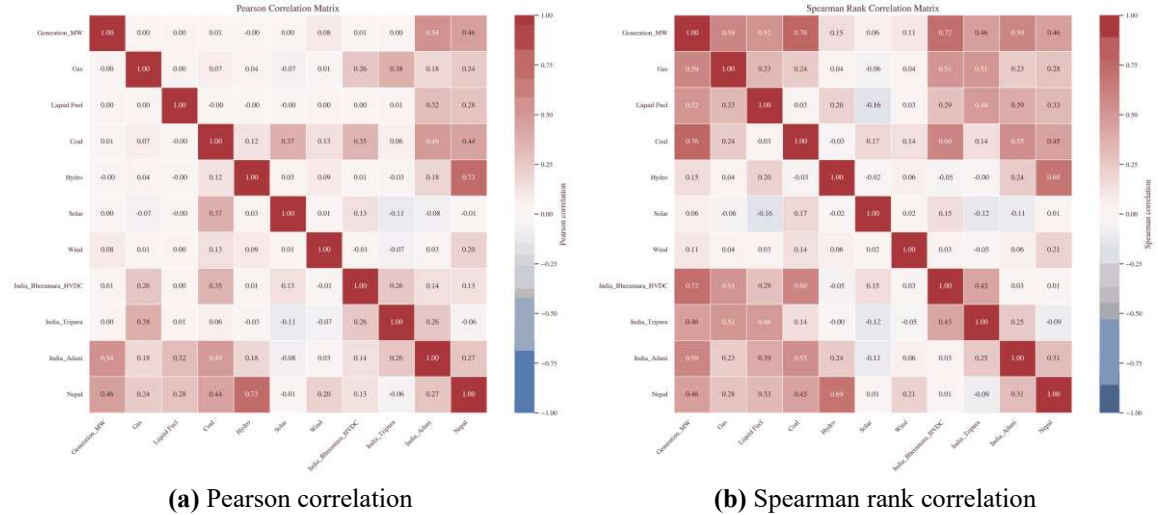


Figure 13: Correlation matrices of the numerical variables. The disagreement between the two matrices reflects the influence of outliers.

5.11 Peak-Hour Remarks

The Remarks column marks 7,319 records as either *Day Peak* (3,755 records) or *Evening Peak* (3,564 records), denoting the daytime and evening peak-demand hours, and it is blank for all other hours. Day Peak readings are concentrated at 12:00 (3,448 of 3,755), while Evening Peak readings fall mainly at 19:00 (1,515) and 21:00 (1,390), with smaller groups in the 18:00 hour (251, including the 18:30 readings) and at 20:00 (405). As summarized in Table 12 and Figure 14, evening-peak readings have a higher level of generation than day-peak readings (mean 10,987 MW versus 9,304 MW, and median 10,834 MW versus 9,204 MW). This agrees with the evening plateau in the hourly profile (Figure 9a). The largest total-generation value that is consistent with the rest of the series, 17,200 MW, occurs under an evening peak.

Table 12: Total generation (MW) by source-provided remark.

Remark	Count	Mean	Median	Std.	Min	Max
Day Peak	3,755	9,303.6	9,204.3	2,547.7	3,634	15,449
Evening Peak	3,564	10,986.8	10,833.5	2,597.8	4,554	17,200

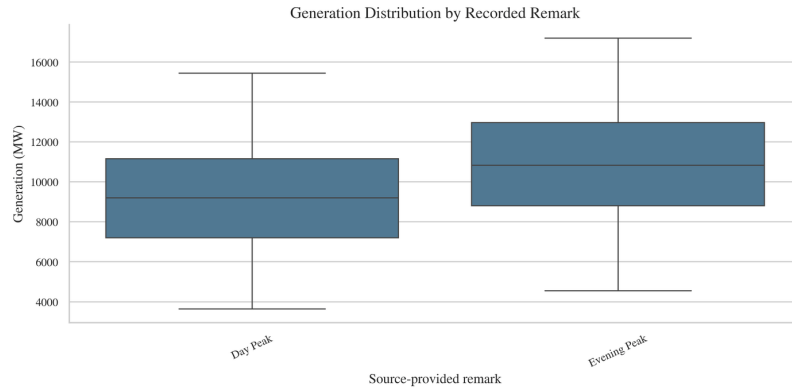


Figure 14: Distribution of total generation by recorded remark.

5.12 Summary of Validation Findings

Table 13 summarizes the validation results and their implications for users.

Table 13: Summary of the technical validation.

Check	Result	Implication for users
Timestamp validity	All 104,134 timestamps are valid and unique	Timestamps can be used directly as a unique index
Temporal regularity	90.9% hourly intervals, 4.5% half-hour records, 238 longer gaps (3 longer than one day)	Resample or filter to exact hours if a regular frequency is required
Completeness	Eight columns are complete; five are structurally missing in early years	Do not replace blanks with zero; use the period in which a variable exists
Value plausibility	One implausibly high and six implausibly low total-generation values; extreme values in several source columns	Validate and treat outliers before analysis
Internal consistency	Total equals domestic sources plus imports within 1 MW for 98.1% of complete records	The identity can be used to screen records
Calendar and physical plausibility	Solar output peaks at noon; Friday is lowest; evening plateau and summer peak are present	Indicates that the timestamps and values capture real system behavior
Inter-variable structure	Pearson and Spearman correlations differ strongly	Prefer rank-based or robust methods

6 Usage Notes and Limitations

Based on the validation above, we offer the following recommendations to future users of BengalGrid.

- **Validate Outliers Before Use:** Outliers are not handled in this release. Extreme values exist in `Generation_MW`, `Gas`, `Liquid Fuel`, `Coal`, `Hydro`, `India_Bheramara_HVDC`, and `India_Adani`. A practical check is to compare `Generation_MW` with the sum of its components (Table 10) or to apply domain-informed upper limits. Users should decide whether to remove, cap, or replace such values.
- **Treat Missingness as Structural:** Blank values in `Solar`, `Wind`, `India_Adani`, `Nepal`, and `Remarks` should not be replaced by zero. Time-series models may need to be trained only on the period in which the relevant variables are available.
- **Handle Irregular Timestamps:** The series includes half-hour readings and a few long gaps, so that resampling or filtering to exact hours (minute component 00) may be required for models that assume a regular frequency.
- **Use Robust Statistics:** Medians, rank correlations, and robust scalers are preferable to means, Pearson correlations, and standard scaling until the outliers are handled.
- **Interpret Total Generation Correctly:** `Generation_MW` is the total supply dispatched to the grid, and it does not represent electricity demand. Studies that require demand or load-shedding measurements need additional datasets, since these are not variables in BengalGrid.

- **Scope and Source Revisions:** The dataset is aggregated at the national level and does not provide plant-level or regional detail. The records are reproduced as published by PGCB at the time of collection, and values may differ from later revisions of the source portal.

As an example, the following Python code reads the dataset, keeps exact-hour records, and uses the consistency identity of Section 5.5 to select records whose components agree with the total within 10 MW.

```
import pandas as pd

df = pd.read_csv("BengalGrid_data.csv", parse_dates=["datetime"])
hourly = df[df["datetime"].dt.minute == 0] # exact hours only

dom = df[["Gas", "Liquid Fuel", "Coal", "Hydro", "Solar", "Wind"]]
imp = df[["India_Bheramara_HVDC", "India_Tripura",
         "India_Adani", "Nepal"]]
resid = (df["Generation_MW"]
        - dom.sum(axis=1, min_count=6)
        - imp.sum(axis=1, min_count=4))

consistent = df[resid.abs() <= 10] # complete and internally consistent
```

7 Potential Research Applications

BengalGrid can support a wide range of studies. First, it can be used for electricity generation forecasting with statistical and machine learning models [3, 5], including multivariate models that use fuel-wise and import features. Second, it supports trend and seasonality analysis and the investigation of the long-term shift in the fuel mix, for example the growth of coal and imports against the decline of the gas share. Third, the import columns enable studies of cross-border electricity trade and its relationship with domestic generation. Fourth, the `Remarks` column allows the study of day-peak and evening-peak behavior. Fifth, the documented outliers and structural gaps make the dataset a realistic testbed for anomaly detection [8], missing-data handling [7], and data-quality assessment methods. Finally, the solar and wind columns allow an early investigation of renewable generation patterns in Bangladesh, although their coverage is limited to the later years.

8 Conclusion

This paper presented BengalGrid, a public dataset of Bangladesh’s electricity generation by source and cross-border imports, containing 104,134 records from April 2015 to October 2026. We described the scraping and cleaning workflow and carried out a systematic technical validation. The analysis showed that the dataset has a consistent structure: total generation matches the sum of domestic sources and imports within 1 MW for 98.1% of the records where all components are available, and the series reproduces the expected diurnal, weekly, and seasonal regularities, including a solar profile centered on noon and a lower Friday demand. It also reveals meaningful long-term trends, such as a compound annual growth of about 7% in median generation, the rising role of coal, and the growth of imports. At the same time, we documented the issues that users must consider, namely a small number of extreme outliers that are not handled in this release, structural missingness in the newer sources, and irregular timestamps. In conclusion, by releasing the raw data, the scripts, and the cleaned dataset without silent corrections, we aim to provide a transparent and reproducible foundation for future research on the Bangladesh power sector.

Declaration

Data Availability Statement

The BengalGrid dataset (version 2) is publicly available in Mendeley Data at <https://doi.org/10.17632/xx9dhmy9rj.2> under a CC BY 4.0 license. The source records are publicly accessible through the PGCB ERP portal.

Code Availability

The data collection script (`scrape_bengalgrid.py`) and the data cleaning script (`clean_bengalgrid.py`) are included in the dataset repository and are also available at <https://github.com/abirrahman26/BengalGrid-Dataset>.

Ethics and Data Governance

BengalGrid is built entirely from aggregated, system-level power generation records that PGCB makes publicly accessible. It contains no personal or confidential information, and no individual can be identified from it. The records were collected with sequential requests and a one-second delay between requests, to avoid placing an undue load on the portal. PGCB is the original source of the data and is acknowledged as such. BengalGrid is an independent compilation and is not an official PGCB product, so users who need official or operational figures should consult PGCB directly. The dataset is released under CC BY 4.0, which requires attribution to the dataset and to the original source.

Funding Information

Not applicable.

Competing Interests

Not applicable.

Research Involving Human and/or Animals

Not applicable. The dataset contains only aggregated, publicly available power system records and no personal information.

Author Contributions

Md. Abir Rahman: Conceptualization, Methodology, Software, Data Collection, Data Cleaning, Writing-Original Draft; Rafi Ahamed: Data Collection, Validation, Formal Analysis, Visualization; Munaia Jannat Easha: Formal Analysis, Visualization, Writing-review and Editing; Tasnia Tarannum Roza: Investigation, Validation, Writing-review and Editing.

References

- [1] Power Grid Company of Bangladesh (PGCB), Generation records, ERP portal. https://erp.powergrid.gov.bd/w/generations/view_generations
- [2] M.A. Rahman, R. Ahamed, BengalGrid: A Decade-Scale Hourly Dataset of Bangladesh Power Supply by Source Type, Including Cross-Border Imports from India and Nepal, Mendeley Data, V2 (2026). <https://doi.org/10.17632/xx9dhmy9rj.2>
- [3] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, R.J. Hyndman, Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting* 32(3), 896–913 (2016)
- [4] F. Wiese, I. Schlecht, W.-D. Bunke, C. Gerbaulet, L. Hirth, M. Jahn, F. Kunz, C. Lorenz, J. Mühlenpfordt, J. Reimann, W.-P. Schill, Open Power System Data – Frictionless data for electricity system modelling. *Applied Energy* 236, 401–409 (2019)
- [5] R.J. Hyndman, G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd edn. (OTexts, Melbourne, 2021)

- [6] R. Weron, Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting* 30(4), 1030–1081 (2014)
- [7] R.J.A. Little, D.B. Rubin, *Statistical Analysis with Missing Data*, 3rd edn. (Wiley, Hoboken, 2019)
- [8] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey. *ACM Computing Surveys* 41(3), 1–58 (2009)
- [9] M.D. Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018 (2016)
- [10] T. Gebru, J. Morgenstern, B. Vecchione, J.W. Vaughan, H. Wallach, H. Daumé III, K. Crawford, Datasheets for datasets. *Communications of the ACM* 64(12), 86–92 (2021)
- [11] J.W. Tukey, *Exploratory Data Analysis* (Addison-Wesley, Reading, 1977)

A Boxplots of the Numerical Variables

This appendix presents the boxplot of each numerical variable. The isolated points far from the main body of the data illustrate the outliers discussed in the technical validation. Boxplot outliers are marked in red.

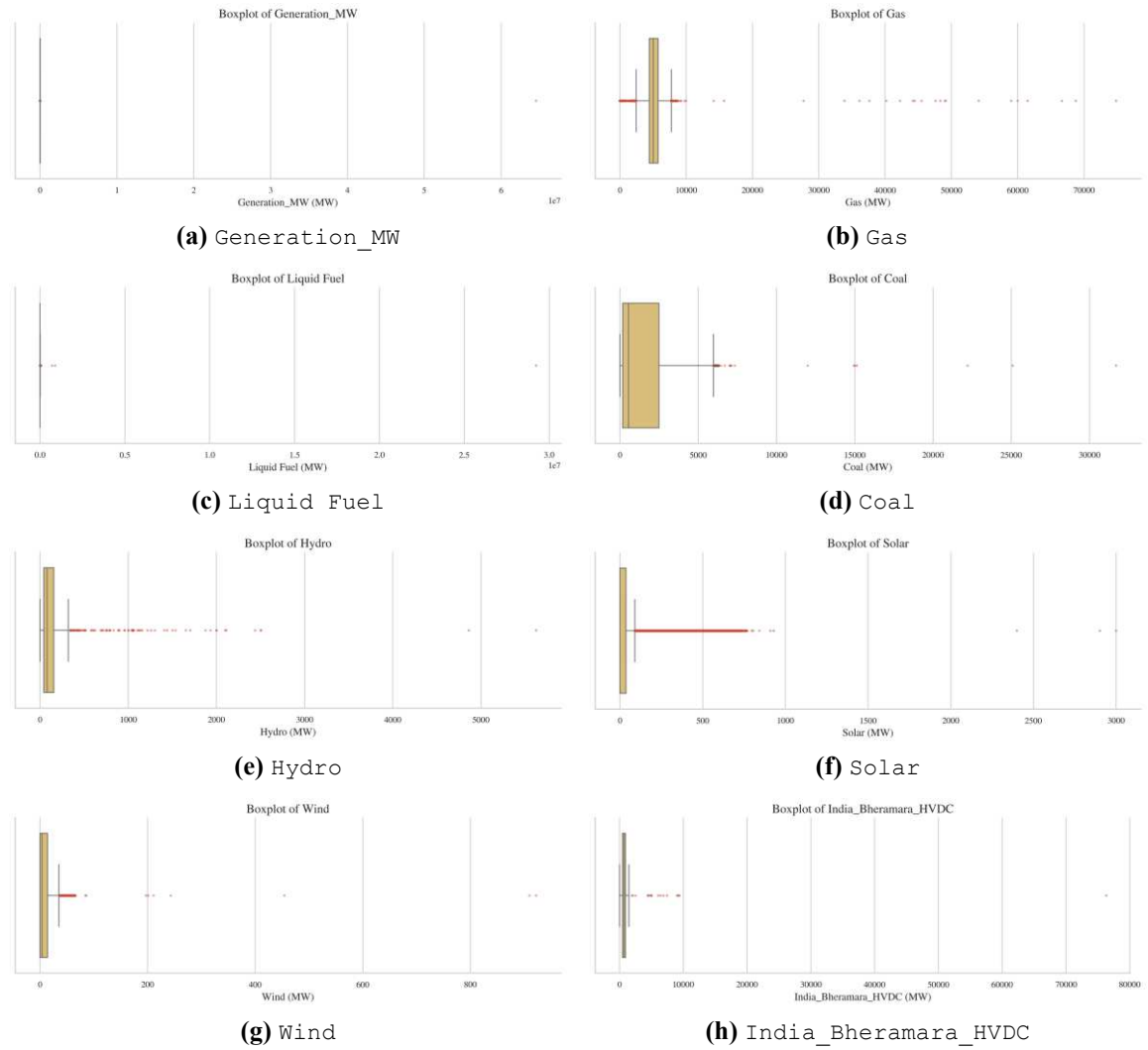


Figure 15: Boxplots of total generation, domestic sources, and Bheramara imports.

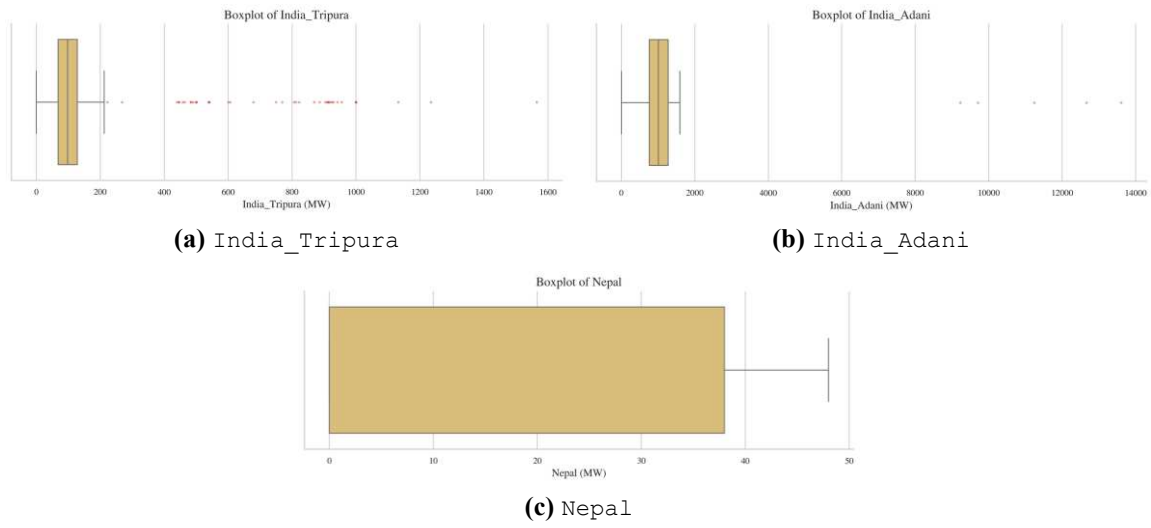


Figure 16: Boxplots of the remaining import channels.