FFG Forschung wirkt.
Bundesministerium Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie
AI for Green – Ausschreibung 2023

This project is funded by the FFG.
www.ffg.at

transpAIrent.energy

Transparent AI Forecasts for Green Energy in Austria

Project number: 910239

D2.1 Documentation of data validation

WP2 – Data collection, evaluation, and documentation

28.05.2025

Authors	Organisation
Fabian LEIMGRUBER	AIT
Peter WIDHALM	AIT

Funding Disclaimer

This project is being carried out as part of the 2023 “AI for Green” call for proposals from the Federal Ministry for Climate Protection, Environment, Energy, Mobility, Innovation and Technology (BMK). The processing is carried out on behalf of the BMK by the Austrian Research Promotion Agency (FFG). The project is funded as part of the topic of digital technologies, an initiative of the BMK, under the grant agreement number FO999910239. The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the texts, or in the decision to publish the results.

Table of Contents

AI for Green – Ausschreibung 2023

This project is funded by the FFG.
www.ffg.at

Executive Summary

Exploratory data analysis and proof-of-concept models from WP3 identified relevant input variables. Identification and collection of public data sets containing these variables are the main outcomes of this deliverable.

In previous projects, subsets of the identified datasets have been used. Data fetching of those subsets was ad-hoc, manual, and only semi-automated if at all. Another issue is that datasets fetched at different points in time (but for the same time interval) might contain different values due to data changes in the source system (for example corrections or updates) which hinders reproducible analysis of forecasts based on these datasets.

To solve those issues, a continuous data collection system named “historian” has been designed to collect and store the identified data sets in a central location.

A preliminary version of metadata describes the identified data sets in terms of sources and collection jobs. Sources represent API endpoints and jobs represent API requests with specific parameters for each job.

The collected data is provided in the “historian” database as the main source for the subsequent stages of the data pipeline (for example Task 2.2).

The derived data is provided in the “librarian” database as a fast analytical database optimized for forecasting and data validation tasks. This database is updated continuously via data pipeline jobs, so it represents the state of data availability in the course of time.

The state of data availability is the basis for the data consuming services such as forecasting and data validation.

Basic and advanced data validation and verification steps have been carried out and are being put into operation to serve as real time data quality indicators.

AI for Green – Ausschreibung 2023

This project is funded by the FFG.
www.ffg.at

1. Introduction

A review of public data sets, exploratory data analysis and proof-of-concept ML models (WP3) enabled identification of relevant data sets.

Data sets were acquired and shared initially in a manual way. That data was obtained from manual download on the respective webpages. Downloads were made with custom time ranges and custom file names. To make data acquisition reproducible, traceable and more automated, data APIs were researched, tested and implemented.

Relevant data sets are provided via REST APIs where data is obtained via HTTP requests/responses.

REST API endpoints are called “sources”, concrete endpoint requests with relevant HTTP headers and query parameters (for instance relevant time intervals for historic or forecasted data respectively) are called “jobs”.

2. Metadata

To prepare the way for re-usability of data according to FAIR data principles[1], metadata is already needed at the data collection and storage layers. Globally unique persistent identifiers are implemented as UUID4 data types (see also Table 2). Relations between identifiers with these properties share those properties as well. This enables a basic lineage of data for subsequent data tasks that can use the metadata to satisfy the respective use-case, for example forecasting or optimization.

Metadata is open for extension in general, and also selectively for example in case of specific variables that need to be universally addressed and related to sources and jobs or other parts of the metadata.

The human-readable parts of the metadata with respect to data collection as mentioned in Section 1 (see also Section 8) for a fetch job are (see also Table 1):

- id: unique identifier in UUID4 format for referencing the job.
- source_id: identifier in UUID4 format of source, see Table 2.
- name: Human readable description of job.
- function: Name of the implementation function.
- args: Position arguments for resource (REST API).
- kwargs: Keyword-argumente for resource (REST API) for example country codes or direction of import/export of energy.

Where `args` and `kwargs` are parameters for a specific source and job combination. Examples are country codes, direction of import/export of energy, geographic longitude/latitude pairs and other custom source-specific information.

3. ENTSO-E Transparency

Following EU regulation Nr. 543/2013, ENTSO-E publishes electricity market data (provided by the individual EU countries’ TSOs) on their Transparency website[2]. In the following, the data categories with the respective numbering following the EU regulation are listed.

[1] https://en.wikipedia.org/wiki/FAIR_data
[2] https://transparency.entsoe.eu/

This project is funded by the FFG.
www.ffg.at

License: CC-BY 4.0[3,4]

3.1. Actual Generation per Production Type [16.1.B&C]

UUID: `46d74416-5cf8-424f-8941-bbd8b3455a81`

https://transparency.entsoe.eu/content/static_content/Static%20content/knowledge%20base/data-views/generation/Data-view%20Actual%20Generation%20per%20Production%20Unit.html

3.2. Generation Forecast - Day ahead [14.1.C]

UUID: `327a5b27-57fc-41e8-964b-ab6a35351502`

https://transparency.entsoe.eu/content/static_content/Static%20content/knowledge%20base/data-views/generation/Data-view%20Generation%20Forecast%20-%20Day%20Ahead.html

3.3. Day-ahead Prices [12.1.D]

UUID: `8cbc7f8e-0281-4388-a96f-541ea76b5a7a`

https://transparency.entsoe.eu/content/static_content/Static%20content/knowledge%20base/data-views/transmission-domain/Data-view%20Day-ahead%20prices.html

3.4. Day-ahead Generation Forecasts for Wind and Solar [14.1.D]

UUID: `f909928b-6110-4274-84fb-9dcfbad90bfc`

https://transparency.entsoe.eu/content/static_content/Static%20content/knowledge%20base/data-views/generation/Data-view%20Generation%20Forecasts%20-%20Day%20Ahead%20for%20Wind%20and%20Solar.html

3.5. Physical Flows [12.1.G]

UUID: `7eef72cd-1d0a-4dee-a404-e8b9e5c0d8ba`

https://transparency.entsoe.eu/content/static_content/Static%20content/knowledge%20base/data-views/transmission-domain/Data-view%20Cross%20Border%20Physical%20Flows.html

3.6. Imbalance: Imbalance Volumes and Prices [17.1.G] & [17.1.H]

Volumes UUID: `74cacad5-185a-46b4-8e1e-ace8ced5860e`

Prices UUID: `7a8bc5ae-f3b3-42a1-92a3-02ebf693c455`

https://transparency.entsoe.eu/content/static_content/Static%20content/knowledge%20base/data-views/balancing/Data-view%20Imbalance.html

4. APG Transparency

Following EU regulation Nr. 543/2013, APG publishes electricity market data on their Transparency website[5]. In the following, the data categories with the respective numbering following the EU regulation are introduced.

License: Probably also CC BY 4.0 as ENTSO-E Transparency, pending written consent by APG to be on the safe side (no reply from APG was received yet as of 28/05/2025).

3 https://transparency.entsoe.eu/content/static_content/download?path=/Static%20content/terms%20and%20conditions/231018_List_of_Data_available_for_reuse.pdf

4 https://creativecommons.org/licenses/by/4.0/legalcode

5 https://transparency.apg.at

FFG Bundesministerium Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie
AI for Green – Ausschreibung 2023

This project is funded by the FFG.
www.ffg.at

4.1. Activation of aFRR

UUID: `bcde1296-f3e9-4166-879f-a7ed768b4368`

Automatic Frequency Restoration Reserve (aFRR) activation is published by APG Transparency.

4.2. Imbalance (Delta der Regelzone)

UUID: `f3d7b2e2-5e6d-499f-9e46-2eea95a96038`

Control Area Imbalance is published by APG Transparency.

Where “15m” means 15 minutes resolution and “1m” means 1 minute resolution. Columns are not identical for all parameter combinations.

4.3. Generation (Erzeugung pro Produktionstyp)

UUID: `03d553fc-47c2-4ee1-8d68-56edc642a4d0`

Generation is published by APG Transparency

4.4. Generation forecast (Prognose über die Erzeugung)

UUID: `a82d40df-e56a-426e-af9b-d739c7edd3e9`

Generation forecast is published by APG Transparency

5. Electricity Maps

5.1. CO_2 intensity

UUID: `2686cfc4-7d7b-47f0-9cfc-9d7f7b175fe3`

CO_2 intensity

API docs, relevant endpoints are:

- `/v3/carbon-intensity/latest`
- `/v3/carbon-intensity/history`

Where `history` covers past 24 hours and `latest` is as of now(?).

https://docs.electricitymaps.com/#live-carbon-intensity

License: ODbL[6].

6. UBIMET

Project internal access to UBIconnect weather data REST API. Listed for completeness and comparison to publicly available data. Documentation: https://docs-ubiconnect-eu.ubimet.com/api/pinpoint.html

6.1. Regions

UUID: `40948d66-c332-4535-a3b8-8802376c7e12`

- Every hour on the hour for the next 24 hours.
- Every six hours on the hour for the next 72 hours.

List of variables:

[6] https://opendatacommons.org/licenses/odbl/1-0/

- WIND_U_1H: wind speed, vector east-west [m s^{-1}]
- WIND_V_1H: wind speed, vector north-south [m s^{-1}]
- INC_SFC_RAD_1H: Radiation onto an inclined surface. [W m^{-2}]
- PREC_CONVECTIVE_1H: Convective Precipitation [kg m^{-2}]
- N: cloud cover [Intervall von 0 bis 8]
- R: amount of precipitation; in case of snow: liquid equivalent [mm]
- T: temperature (2 m), also called dry-bulb temperature [°C]

These variables are available for pre-defined geographic locations, see Figure 1.

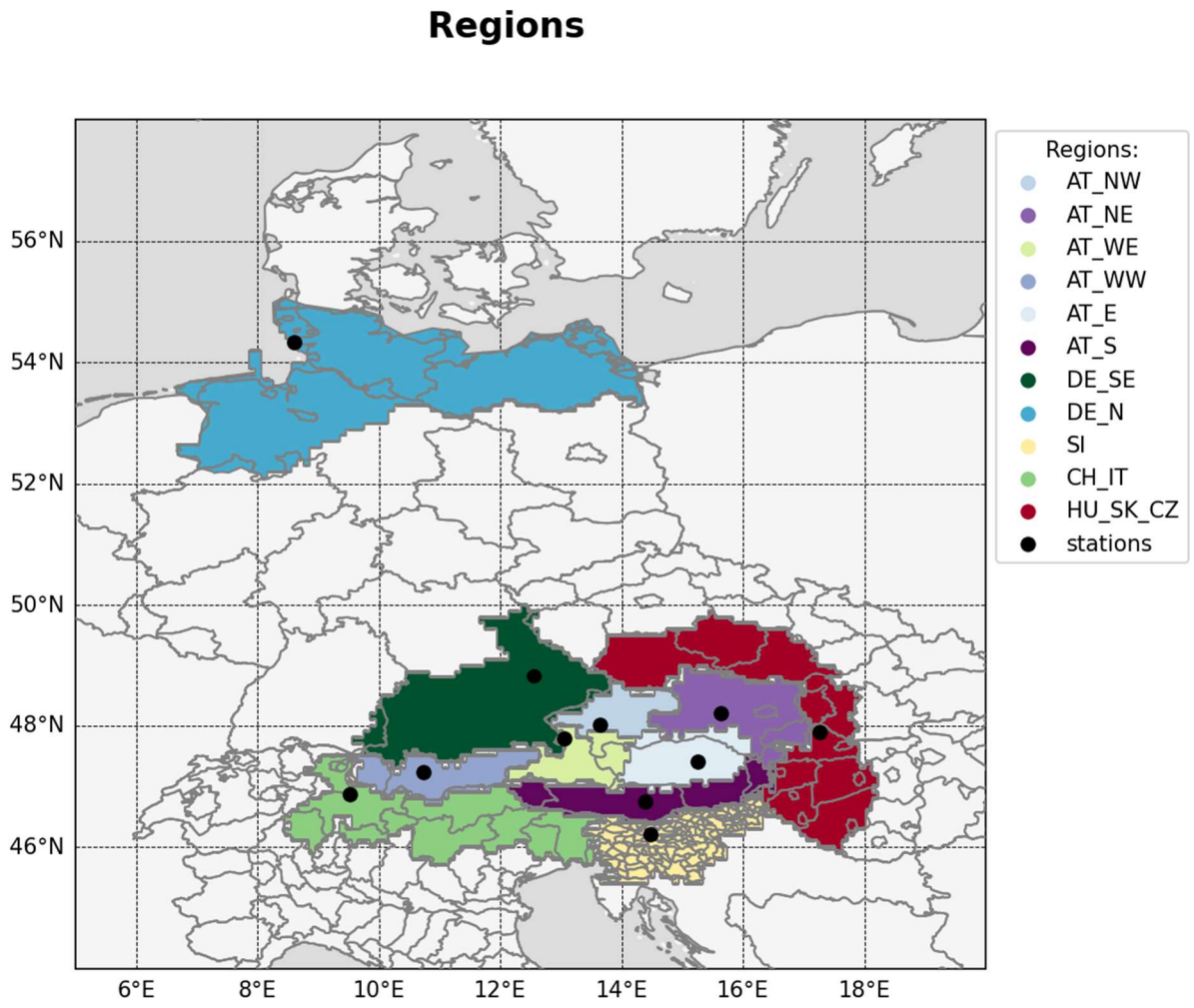

Figure 1 Geographic locations for the weather forecasts which were predefined together with UBIMET.

6.2. Strompool

UUID: `3b26f5ef-8fc8-472d-a3ac-564e49beab6b`

- Every hour on the hour for the next 24 hours.
- Every six hours on the hour for the next 72 hours.

Specific geographic locations must be supplied to the API request.

List of variables:

- WIND_U_1H: wind speed, vector east-west [m s^{-1}]
- WIND_V_1H: wind speed, vector north-south [m s^{-1}]
- INC_SFC_RAD_1H: Radiation onto an inclined surface. [W m^{-2}]

- PREC_CONVECTIVE_1H: Convective Precipitation [kg m^{-2}]
- N: cloud cover [Intervall von 0 bis 8]
- R: amount of precipitation; in case of snow: liquid equivalent [mm]
- T: temperature (2 m), also called dry-bulb temperature [°C]

7. Additional data sources

Additional data sources that were identified or will be identified in due course during the project (but have not yet been collected) can be added to the “historian” system by creating respective UUIDs for the sources and jobs and implementations of the API endpoints as functions to be used with the respective fetching jobs. Examples of additional data sources are data from JAO via the Publication Tool[7].

8. Data architecture

The data architecture comprising data collection, data flow, and data serving is shown in Figure 2. APIs are queried via HTTP requests and put into the “historian” database. The historian database is then queried by data pipeline processors who pick up any new collected data, normalize the data to a bi-temporal schema, and save the data to the “librarian” database. This librarian database represents the main project-internal API for providing data to the other use cases such as data quality checking, visualization, optimization, and forecasting.

[7] https://publicationtool.jao.eu/core/api

This project is funded by the FFG.
www.ffg.at

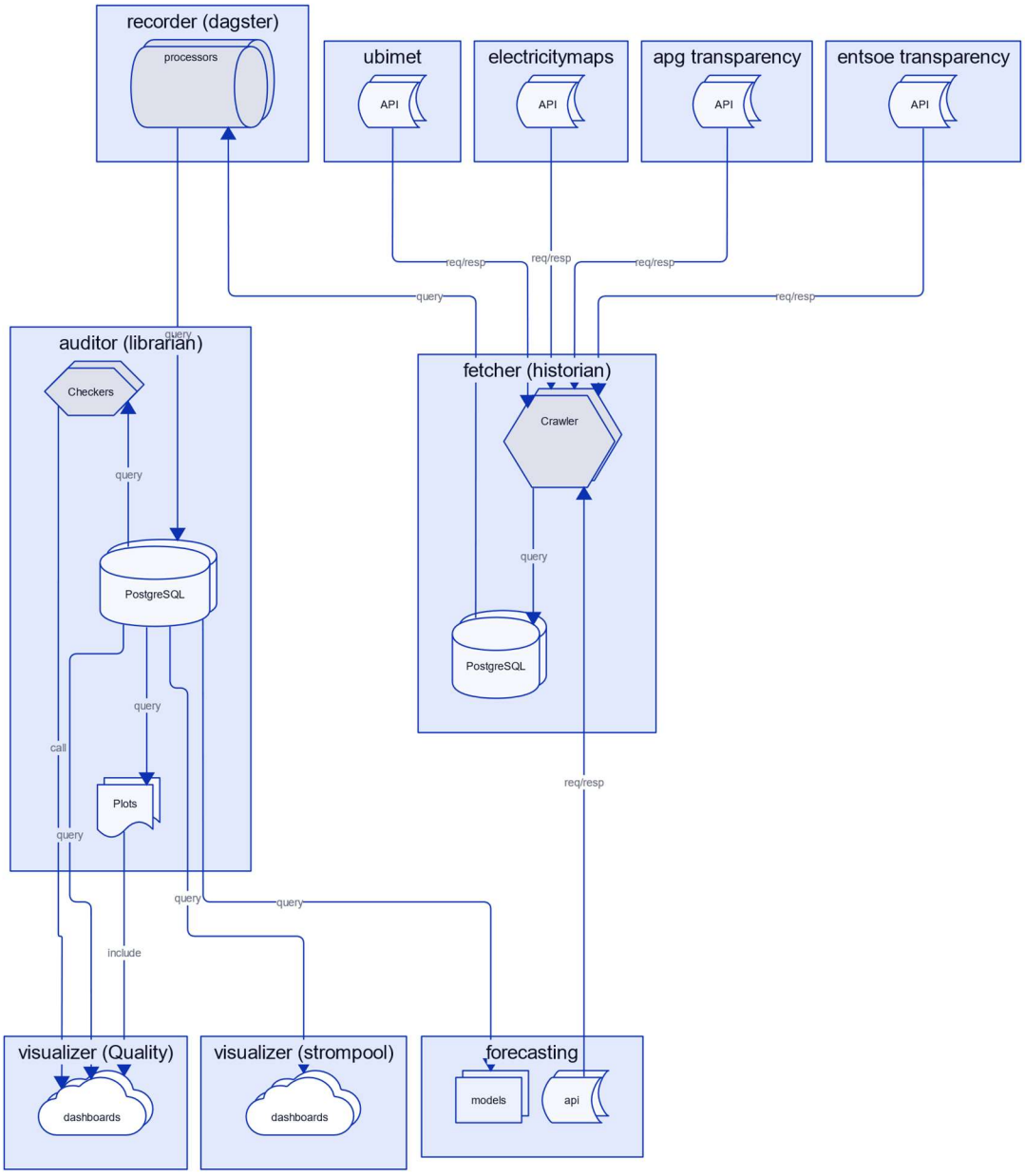

Figure 2 Data flow architecture

The following focuses on the historian component which is the core of the data collection process and is composed of a scheduler and a database. Scheduler and database are decoupled with respect to deployment and run on different machines. Communication between the two sub-components happens via SQL.

8.1. Historian

A time-synchronized scheduler runs fetch jobs (see Table 1) periodically every 5 minutes. Some jobs have specific requirements for query frequency so their schedule can be set individually to once every hour or every 4 hours for example. The scheduled fetcher uses the Sentry[8] system for monitoring, alerting and error tracking. This system ensures that runtime

[8] https://sentry.io

errors do not pass unnoticed and provides a way to inspect exceptional failures with all the context at the point of failure. Other failures (mainly unavailability of APIs) are treated the same as successful API responses and are also stored in the database. The scheduler converts the API response data to JSON payloads and stores that payload with derived data to the database.

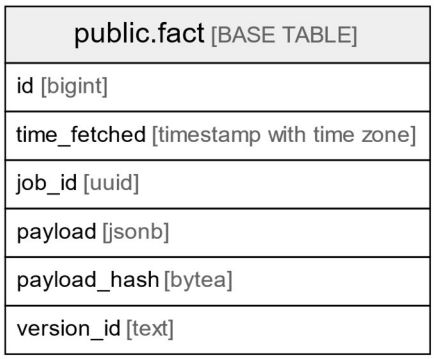

Figure 3: historian "fact" table schema

The historian database table schema is shown in Figure 3. The columns are:

- id: auto-increment integer for each execution of a fetch job.
- time_fetched: Time-zone aware timestamp of the point in time when the API request is initiated (jobs that query forecast data determine the time intervals for API requests based on this timestamp).
- job_id: See section 2.
- payload: The binary encoded representation of the API response JSON data. The resulting BJSON is used for storage efficiency.
- payload_hash: The SHA-256 hash of the JSON payload content, used for caching, see below.
- version_id: The git SHA1 commit string representing the development version of the scheduler component (context for debugging).

This database schema represents a trade-off in terms of downstream usage of the data in the dependent data services. The schema has the following desireable properties:

- Intentionally does not differentiate between “live” and “historic” data (the interpretation of which is contained in the payload table column and the timestamps within it).
- Point-in-time information for each recorded request/response pair, enabling “as of” queries to simulate historical data fetching.
- Open for extension: job_ids and their respective JSON payloads can be added without changes to the schema. UUID4s of job_ids are given by the metadata schema (which is also extensible by design and as a general property of metadata) and it is assumed that the payloads of all future APIs can be serialized to JSON as it is a de-facto fallback standard for machine-to-machine data exchange, similar to CSV for human-to-human data exchange.
- Caching: If the response JSON payloads of two consecutive API requests are the same (byte for byte) then their SHA-256 hash is the same (hash is used for fast comparison) and the payload is not saved into the database.

This project is funded by the FFG.
www.ffg.at

- Additional processes can be added that use other logic than the implemented polling scheduler, for example event based asynchronous data fetching such as messaging queues.

This trade-off also implies other less desirable properties. If the raw data schema changes upstream (for example day-ahead price related sources change the time resolution from 1 h to 15 minutes or the name mapping for power generation types (B01, ..., B20) changes) the data JSON payload schema changes as well, because just the JSON representation is saved, deliberately without schema checks to guarantee the desirable properties above. To mitigate this problem, the raw data schemas per source have to be versioned in the downstream data consumer libraries, for example in librarian component source code. Due to the usage of UUIDs, that mapping could even be made persistent in another event log where the entries are upstream schema changes with associated timestamps of when the change occurred.

The database is implemented using PostgreSQL.

For reference, the inspiration for the chosen database schema is the concept of an event log used as the basis for event driven architectures.

8.2. Librarian

Using the periodically fetched data from the historian database, Dagster pipeline tasks normalize the fetched data into a database that is optimized for analytical queries, called the “librarian” database.

Decoupling of the two databases allows for reproducible recreation of a traceable history of data availability that is the basis for the data consuming services:

- Development and operation of accurate forecasting methods.
- Optimization of assets
- Continuous and historic data validation

The librarian database table schema is shown in Figure 3. The columns are:

- time_valid: Time stamp for the time period during to which a value is associated.
- time_transaction: Time stamp of when the valid time and value were recorded in the database.
- job_id: See section 2.
- variable: The name of the variable as named in source_id (see section 2) or a normalized name depending on and prepared for the data consuming services.
- value: Numeric value of the variable during time_valid and known since time_transaction.
- lat: Latitude of geographic point coordinate in GPS.
- lon: Latitude of geographic point coordinate in GPS.

This project is funded by the FFG.
www.ffg.at

public.history [BASE TABLE]
time_valid [timestamp with time zone]
time_transaction [timestamp with time zone]
job_id [uuid]
variable [text]
value [real]
lat [double precision]
lon [double precision]

Figure 4 librarian "history" table schema

The librarian database is filled continuously from the historian database via a data orchestration pipeline implemented in Dagster.[9] The completion status and timing information of one of these librarian data jobs is shown in Figure 5. The pipeline enables parallel execution of the data import as well as traceability per fetch job as shown with a specific job UUID as an example.

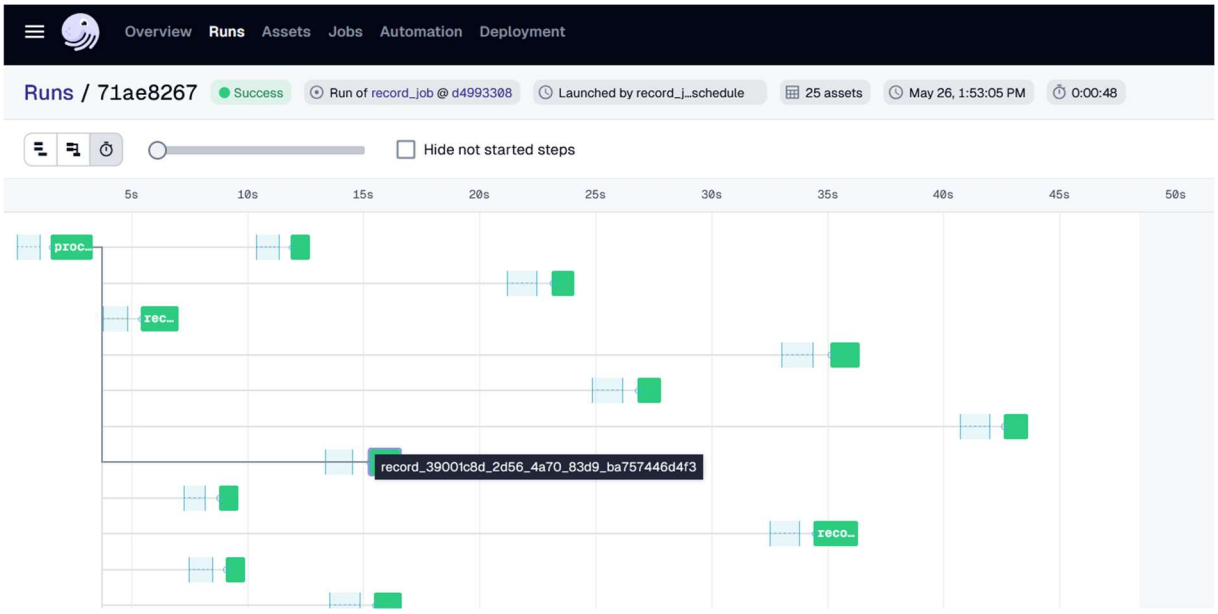

Figure 5 Librarian pipeline job run example for ingestion of recent historian data

The librarian database is implemented using PostgreSQL with the TimescaleDB extension optimized for timeseries data.

The bi-temporal librarian database enables continuous validation of data flowing through the pipeline using the methods described in section 9.

[9] https://dagster.io/

This project is funded by the FFG.
www.ffg.at

Table 1 Metadata of historian jobs

id	source_id	name	function	args	kwargs
83b91ba7-cc28-4ca4-ade6-d50aba2730fc	46d74416-5cf8-424f-8941-bbd8b3455a81	entsoe_generation_at_tier1	query_generation_entsoe_tier1		{"country": "AT"}
6f8e0060-a2ee-4fb3-bb5b-ddc47e1556b8	46d74416-5cf8-424f-8941-bbd8b3455a81	entsoe_generation_de_tier1	query_generation_entsoe_tier1		{"country": "DE"}
a8fa9051-75a6-4393-b623-69b3638c56d1	327a5b27-57fc-41e8-964b-ab6a35351502	entsoe_generation_forecast_at_tier1	query_generation_forecast_entsoe_tier1		{"country": "AT"}
97020d0b-f5a8-4978-be0c-e654a8e7e2dd	327a5b27-57fc-41e8-964b-ab6a35351502	entsoe_generation_forecast_de_tier1	query_generation_forecast_entsoe_tier1		{"country": "DE"}
37340216-e4df-4709-b0a7-3afaf7f69edc	f909928b-6110-4274-84fb-9dcfbad90bfc	entsoe_generation_wind_and_solar_forecast_at_tier1	query_generation_forecast_wind_and_solar_entsoe_tier1		{"country": "AT"}
408f0eff-b76c-45f3-aa9e-14b1395066da	f909928b-6110-4274-84fb-9dcfbad90bfc	entsoe_generation_wind_and_solar_forecast_de_tier1	query_generation_forecast_wind_and_solar_entsoe_tier1		{"country": "DE"}
50c62f7b-8d77-407d-ad43-3e7d413b9180	8cbc7f8e-0281-4388-a96f-541ea76b5a7a	entsoe_dayahead_prices_at_tier1	query_dayahead_prices_entsoe_tier1		{"country": "AT"}
6b097d62-0b74-417b-ac67-e20838bca199	7eef72cd-1d0a-4dee-a404-e8b9e5c0d8ba	entsoe_crossborder_flows_at_export_tier1	query_crossborder_flows_entsoe_tier1		{"country": "AT", "export": true}
064de41b-d13f-41a6-a21f-1ebcdf9fdb77	7eef72cd-1d0a-4dee-a404-e8b9e5c0d8ba	entsoe_crossborder_flows_at_import_tier1	query_crossborder_flows_entsoe_tier1		{"country": "AT", "export": false}
5018c54e-90d8-4fd7-aabd-a6fc5fe31e76	7a8bc5ae-f3b3-42a1-92a3-02ebf693c455	entsoe_imbalance_prices_at_tier1	query_imbalance_prices_entsoe_tier1		{"country": "AT"}
f2bcdc62-bc66-41fb-89e3-849025e47ec4	74cacad5-185a-46b4-8e1e-ace8ced5860e	entsoe_imbalance_volumes_at_tier1	query_imbalance_volumes_entsoe_tier1		{"country": "AT"}
7b4657ba-86f5-44d7-918f-5f92d88494c3	f3d7b2e2-5e6d-499f-9e46-2eea95a96038	apg_imbalance_tier1	query_imbalance_apg_tier1		{"quality": "betrieblich"}
58078718-58f1-4695-9223-383100fe5bc9	bcde1296-f3e9-4166-879f-a7ed768b4368	apg_afrr_tier1	query_afrr_apg_tier1		{}
39001c8d-2d56-4a70-83d9-ba757446d4f3	03d553fc-47c2-4ee1-8d68-56edc642a4d0	apg_generation_act_tier1	query_generation_act_apg_tier1		{}
b9a27fe1-554f-4a24-b0c9-c67ce9e2685f	a82d40df-e56a-426e-af9b-d739c7edd3e9	apg_generation_fc_tier1	query_generation_fc_apg_tier1		{}
c49634e7-fb1b-4b75-bb51-5722219bb0ae	a82d40df-e56a-426e-af9b-d739c7edd3e9	apg_generation_fc_tier2	query_generation_fc_apg_tier2		{}
95193cdd-bae3-4dfc-a83f-81ab868e9a6d	a82d40df-e56a-426e-af9b-d739c7edd3e9	apg_generation_fc_tier3	query_generation_fc_apg_tier3		{}
2b0408f6-7b3a-4666-90d4-ddcdd3ee5e34	ed3407f9-8cc7-4210-8b6f-a33eeed5e553	apg_dayahead_prices_tier1	query_dayahead_prices_apg_tier1		{}
32c53bc0-f712-430e-a22e-55ab82a7a184	ed3407f9-8cc7-4210-8b6f-a33eeed5e553	apg_dayahead_prices_tier2	query_dayahead_prices_apg_tier2		{}
7a6b093e-07ef-4295-b360-47954e640d1f	ed3407f9-8cc7-4210-8b6f-a33eeed5e553	apg_dayahead_prices_tier3	query_dayahead_prices_apg_tier3		{}
0b7922ec-c29c-4a71-966f-5c255bc61d02	7355bb3e-77b5-437a-83c6-e872f7eebae5	apg_imbalance_prices_tier1	query_imbalance_prices_apg_tier1		{}
4b63ce05-71bd-4782-9650-d37737c8eb08	7355bb3e-77b5-437a-83c6-e872f7eebae5	apg_imbalance_prices_tier2	query_imbalance_prices_apg_tier2		{}
50a0dfb1-dda4-4999-9919-f11e1edc54d0	2686cfc4-7d7b-47f0-9cfc-9d7f7b175fe3	electricitymaps_carbon_intensity_latest	query_co2_intensity_latest_electricitymaps		{"zone": "AT"}

4a3d2d3b-80cd-495f-beed-3effa738d5ed	40948d66-c332-4535-a3b8-8802376c7e12	UBIMET 'Regions' 1h	query_regions_ubimet	{"preset": "AIT_Maggauer_1d_1h"}
6550fadf-c77f-42c3-a60f-016870946af6	40948d66-c332-4535-a3b8-8802376c7e12	UBIMET 'Regions' 6h	query_regions_ubimet	{"preset": "AIT_Maggauer_3d_1h"}
1ab6db1a-9fa5-4497-bc5e-891f0c02f639	3b26f5ef-8fc8-472d-a3ac-564e49beab6b	UBIMET 'Strompool' 1h	query_strompool_ubimet	{"preset": "Strompool_1d_1h", "coordinates": [[10.294186, 47.142505]]}
8906f453-03d8-4bb8-90f7-cbc20b50d197	3b26f5ef-8fc8-472d-a3ac-564e49beab6b	UBIMET 'Strompool' 6h	query_strompool_ubimet	{"preset": "Strompool_3d_1h", "coordinates": [[10.294186, 47.142505]]}

Table 2 Metadata of historian sources

id	name	upstream_id	upstream_name
8cbc7f8e-0281-4388-a96f-541ea76b5a7a	dayahead_prices_entsoe_transparency	12.1.D	Day-ahead Prices
46d74416-5cf8-424f-8941-bbd8b3455a81	generation_historic_entsoe_transparency	16.1.B&C	Actual Generation per Production Type
327a5b27-57fc-41e8-964b-ab6a35351502	generation_forecast_entsoe_transparency	14.1.C	Generation Forecast - Day ahead
f909928b-6110-4274-84fb-9dcfbad90bfc	generation_forecast_wind_and_solar_entsoe_transparency	14.1.D	Generation Forecasts for Wind and Solar
7eef72cd-1d0a-4dee-a404-e8b9e5c0d8ba	crossborder_flows_entsoe_transparency	12.1.G	Physical Flows
7a8bc5ae-f3b3-42a1-92a3-02ebf693c455	imbalance_prices_entsoe_transparency	[17.1.G] & [17.2.F] ([17.1.G] & [17.1.H])	Imbalance Prices (Imbalance Volumes and Prices)
74cacad5-185a-46b4-8e1e-ace8ced5860e	imbalance_volumes_entsoe_transparency	[17.1.H] & [17.2.G] ([17.1.G] & [17.1.H])	Total imbalance volume (Imbalance Volumes and Prices)
f3d7b2e2-5e6d-499f-9e46-2eea95a96038	imbalance_historic_apg_transparency	DRZ	Deltaregelzone
bcde1296-f3e9-4166-879f-a7ed768b4368	afrr_historic_apg_transparency	SCP	Sekundärregelreserve (SRR)
03d553fc-47c2-4ee1-8d68-56edc642a4d0	generation_historic_apg_transparency	AGPT	Erzeugung pro Produktionstyp
a82d40df-e56a-426e-af9b-d739c7edd3e9	generation_forecast_apg_transparency	DAFTG	Prognose über die Erzeugung
7355bb3e-77b5-437a-83c6-e872f7eebae5	imbalance_prices_apg_transparency	AE	Ausgleichsenergiepreise
ed3407f9-8cc7-4210-8b6f-a33eeed5e553	dayahead_prices_apg_transparency	EXAAD1P	Day-ahead Preise
2686cfc4-7d7b-47f0-9cfc-9d7f7b175fe3	latest_carbon_intensity_electricitymaps	/v3/carbon-intensity/latest	Live carbon intensity
40948d66-c332-4535-a3b8-8802376c7e12	UBIconnect 'Regions' pinpoint data	/pinpoint-data	Get Pinpoint Data

FFG Bundesministerium Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie
AI for Green – Ausschreibung 2023
This project is funded by the FFG.
www.ffg.at

3b26f5ef-8fc8-472d-a3ac-564e49beab6b	UBIconnect 'Strompool' POI data	/poi-data	Get POI Data

AI for Green – Ausschreibung 2023

This project is funded by the FFG.
www.ffg.at

9. Validation and verification

Data validation was carried out with three classes of methods.

Firstly, data type checks were performed to find problems due to data type conversions. This also includes expected time intervals, time resolution and completeness of data with respect to missing time stamps.

Secondly, basic value range checks were carried out to check for non-negativity (for example, strictly non-negative solar power generation) and non-zero values (for example total baseline power generation is never exactly or close to zero).

Thirdly, the above checks were augmented with checks for the distribution of values. The question of a baseline empirical distribution of values was answered by using a custom AI tool that also incorporates timeseries information such as time of day and seasonal variations.

The data verification has been done and will continue periodically using the mentioned AI tool. This tool employs a Transformer-based Autoencoder that learns the quantiles of its own reconstruction errors to measure deviations between expected and actual time series shapes and distributions. The tool checks two types of value ranges: 1) A broad range based on the 1st to 99th percentiles of the modelled distribution. 2) A narrower interquartile range (IQR), defined as IQR = Q3 – Q1, and robust outlier thresholds defined canonically as $1.5 \times IQR$.

The tool then compares observed values in each time series to the learned empirical value distributions to identify anomalies. Historical data checks were carried out in this manner for each day from 2019 to mid 2024, see also Figure 6.

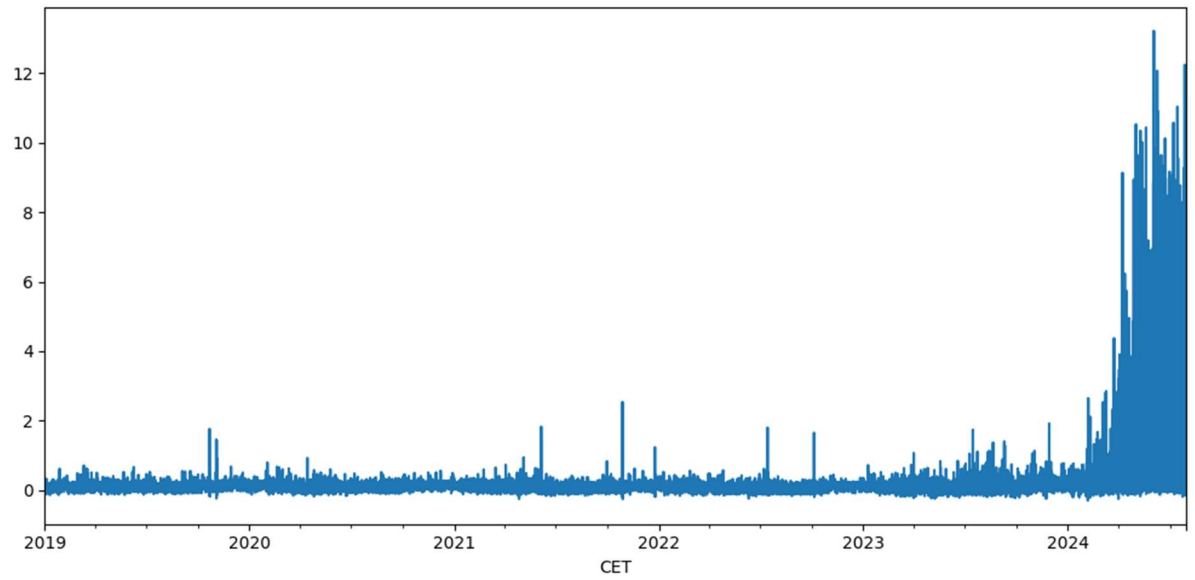

Figure 6 Anomaly scores for each day of the year (full data range)

Data errors such as zero solar power production throughout a particular day (see Figure 7) or unrealistic exactly zero total power generation in the control area (see end of day in Figure 8 or a longer period of the day in Figure 9) were successfully identified using the tool.

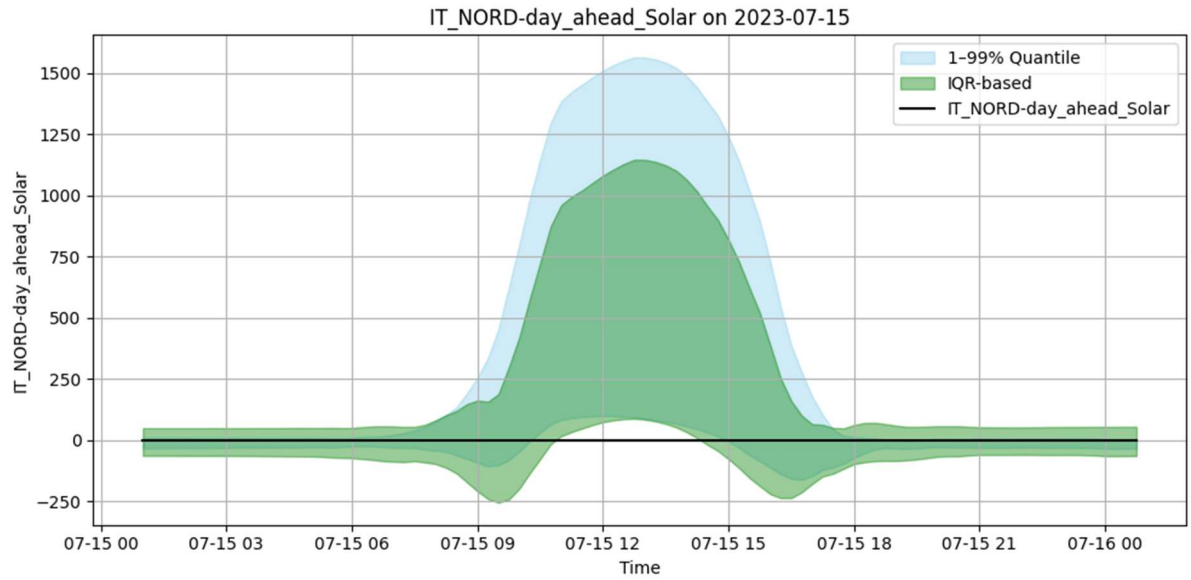

Figure 7 Expected day-ahead solar production for 15th of July in Italy vs. actual solar production

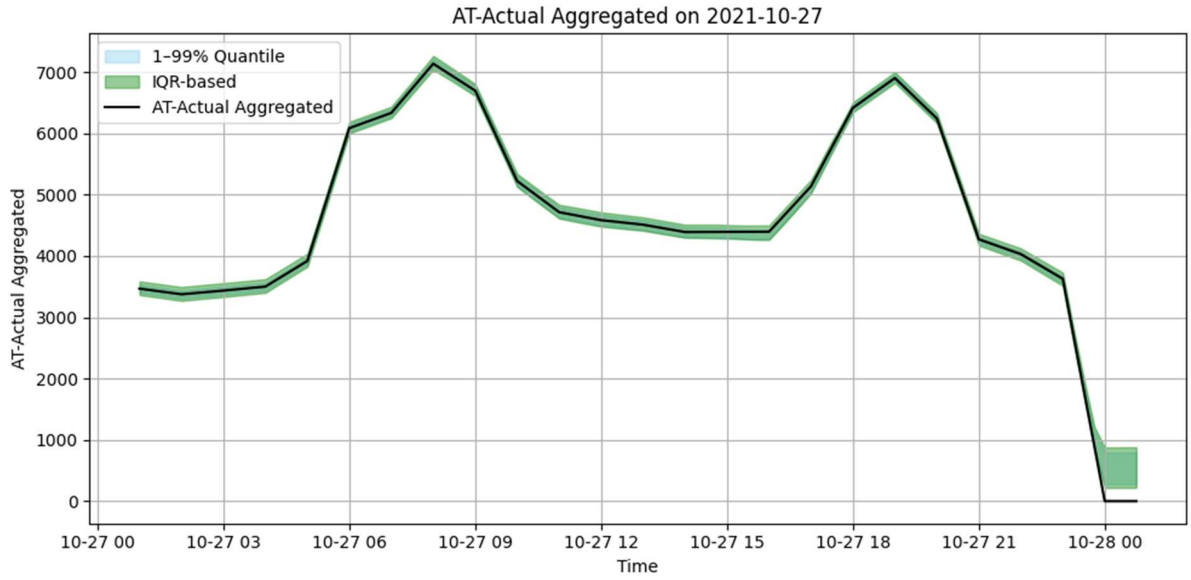

Figure 8 Expected non-zero base load and actual zero base load at midnight

AI for Green – Ausschreibung 2023

This project is funded by the FFG.
www.ffg.at

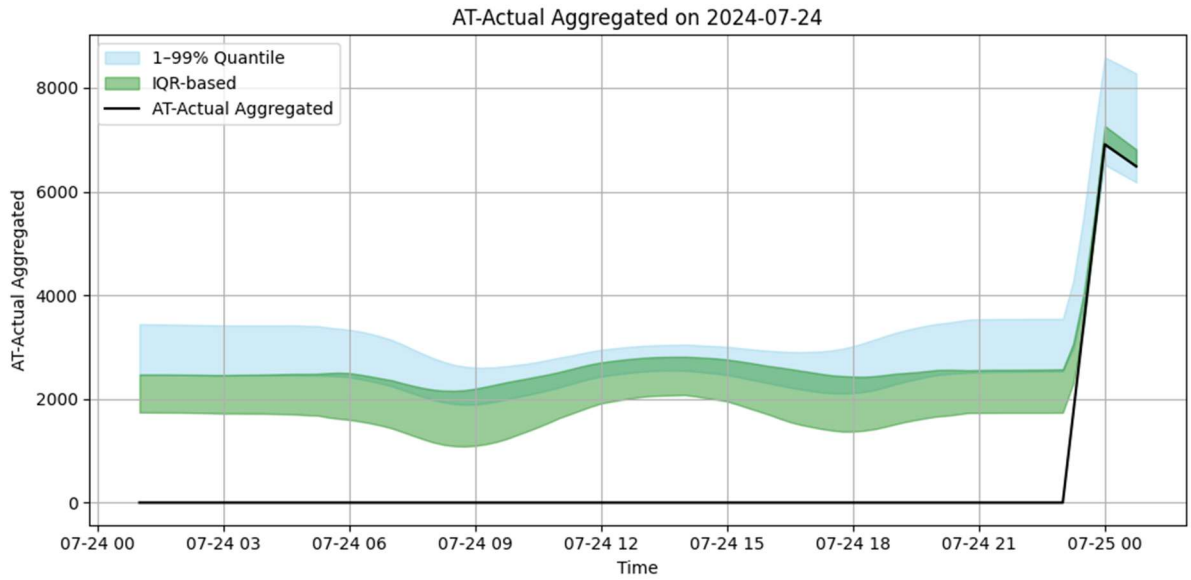

Figure 9 Expected aggregated generation and actual zero values and values returning to expected value ranges

The tool also found false positives, for example the “anomaly” of the increased solar power generation in 2024 (see Figure 11) compared to the learned history from the years 2019 to 2023 (see Figure 10) which can be explained by a significantly higher installed PV capacity in 2024 compared to previous year-by-year changes.

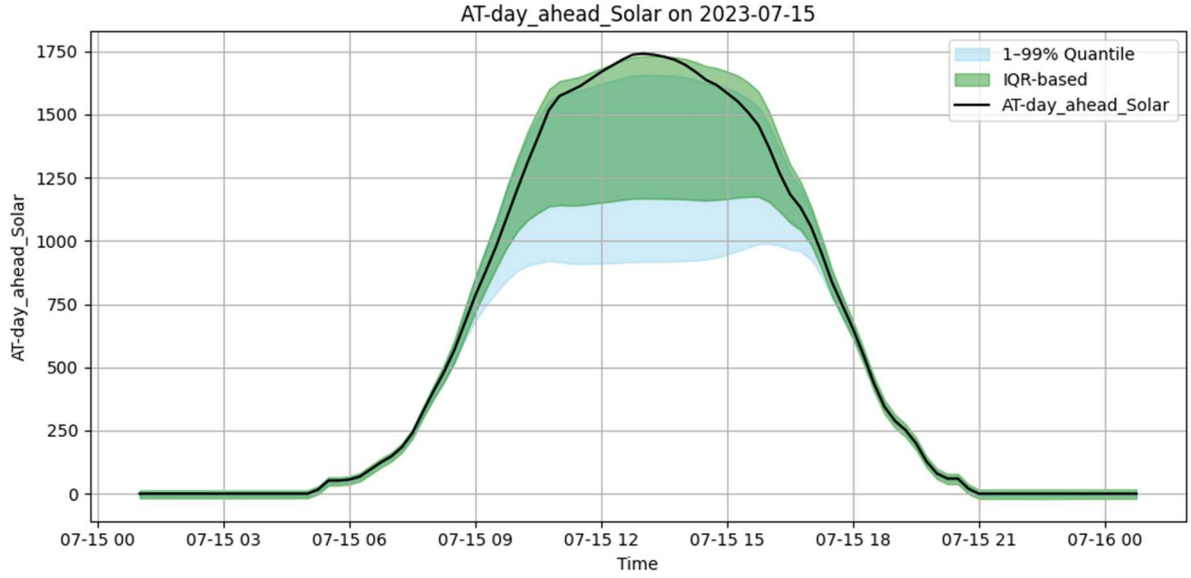

Figure 10 Expected value ranges of solar generation up until the year 2023

This project is funded by the FFG.
www.ffg.at

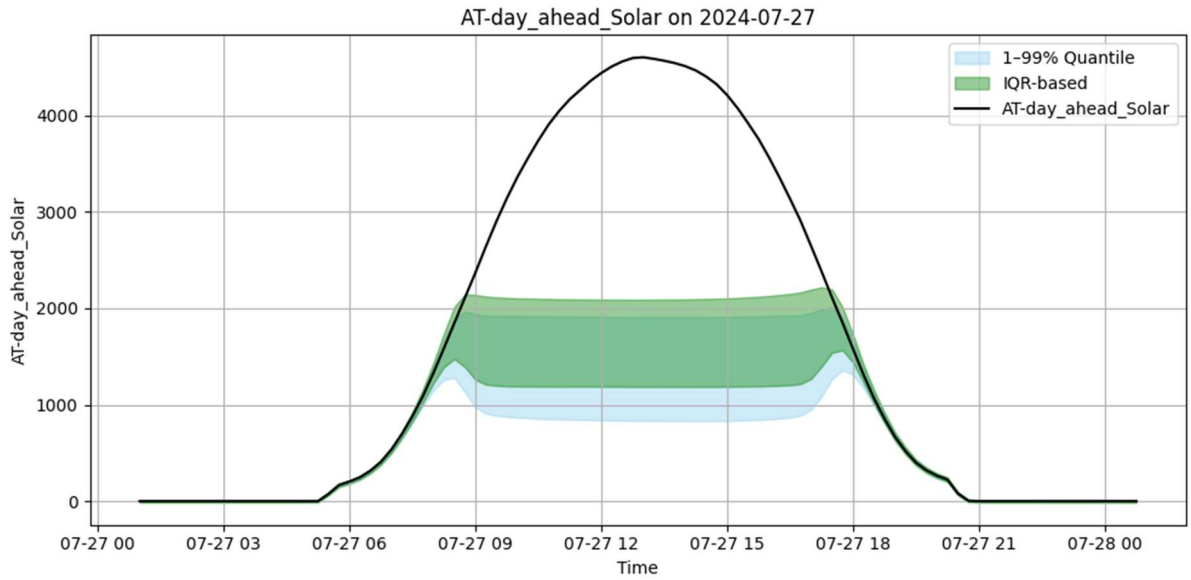

Figure 11 Expected solar generation ranges and actual values for a sunny day in 2024

AI for Green – Ausschreibung 2023

This project is funded by the FFG.
www.ffg.at

10. Conclusion

The components developed and described in this deliverable are operational and continuously collecting data for the documented data sources since February 2025.

The data is provided for internal quality checks (Task 2.1, Task 2.2) and successfully used for forecasting (Task 3.2).

The codebases of the components are undergoing a process of documentation and code review to be eventually published in Task 2.4.

Project coordinator:

Klara Maggauer, M. Sc., B. Sc.

AIT Austrian Institute of Technology GmbH

Center for Energy

Giefinggasse 4

A-1210 Vienna

E-mail: klara.maggauer@ait.ac.at

This project is funded by the FFG.
www.ffg.at