Plan Overview

A Data Management Plan created using DMPonline.be

Title: Leveraging Data-driven Weather Models for Enhanced End-to-End Forecasting in Renewable Energy Systems

Creator: Ada Canaydin

Principal Investigator: Hussain Syed Kazmi

Data Manager: Ada Canaydin

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Hussain Syed Kazmi

Data Manager: Ada Canaydin

Project abstract:

Significant investments in renewable energy generation, primarily solar and wind, are being made to meet national and European climate objectives. However, the transition poses challenges in grid stability due to renewable generation's inherent intermittency and variability, leading to increased uncertainty in power system operation driven by forecast errors. This underscores the urgent need for more accurate forecasting tools to ensure system balance, prevent congestion events, and optimize energy strategies. Nevertheless, renewable forecasting remains a challenge due to factors such as inadequate weather data and historical observational data, limiting the full utilization of modern data-driven architectures.

To address this, the Ph.D. aims to develop a novel end-to-end approach using advanced machine learning techniques to directly integrate global weather information for better renewable energy forecasts. The model will produce forecasts both nationally and locally, focusing on refining accuracy at specific sites and installations with sparse observational data. The closer weather integration is expected to lead to more accurate models, which will be validated against established state-of-the- art methods, using real-world data. The reduced computational expense at inference time enables applications like nowcasting, advancing the field of renewable energy forecasting and enhancing grid reliability and efficiency in the face of increasing renewable energy integration.

ID: 213051

Start date: 01-11-2024

End date: 31-10-2028

Last modified: 16-04-2025

Leveraging Data-driven Weather Models for Enhanced End-to-End Forecasting in Renewable Energy Systems

DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

Leveraging Data-driven Weather Models for Enhanced End-to-End Forecasting in Renewable Energy Systems

GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Global Physic-based Weather Model Data	Historical large-scale meteorological forecast data from sources; (1) Copernicus Open Access Hub: ERA5 reanalysis data and IFS data (2) NCEP NOAA: GFS data (3) DWD: ICON-EU, ICON-Global data (4) KNMI: Harmonie Arome Europe data (5) Rebase.energy platform	Reuse existing data	Digital	Software	.nc, .grib2,.csv,.parquet	<10TB	
Global Data-driven Weather Model Data	Historical forecast data generated from locally implemented data-driven weather models	Generate new data	Digital	Software	.nc, .grib2,.csv,.parque	<10TB	

Local Raw Weather Observations	Historical site-specific meteorological raw data from Lidar measurements, and national/regional weather agencies	Reuse existing data	Digital	Compiled/aggregated data, Observational	.nc, .grib2,.csv,.parquet	<10TB	
Processed Weather Data	Intermediate weather data derived from raw global and local weather datasets (i.e. extracted features, cleaned observations, and interpreted data)	Generate new data	Digital	Software	.csv, .parquet	<100GB	
Renewable Energy Generation Data	Observed historical wind and solar power generation data (SCADA or aggregator data) from utility/TSO or energy providers and publicly available national and site-specific renewable energy forecasts. National-level generation data sources; (1) Elia Open data platform (2) ENTSO-E Transparency platform Site-level generation data sources; (3) SmarThor platform (4) Ørsted (private) data	Reuse existing data	Digital	Compiled/aggregated data, Observational	.csv,.parquet	<100GB	

Distributed Solar/Wind Systems Physical Characteristics Data	Data describing physical characteristics of energy systems (i.e. location coordinates, installed capacity, system orientation and tilt, technology type, and installation dates) from sources; (1) ENTSO-E Transparency platform (2) Ørsted (private) data	Reuse existing data	Digital	Compiled/aggregated data, Observational	.csv,.parquet,.xlsx	<100GB	
Model Code / Software Artifacts	Packaged python file and Jupyter notebooks, used to train and run the ML energy forecasting models	Generate new data	Digital	Software	.py,.ipynb	<100 GB	
Experimental Logs	Digital records of optimal hyperparameters, performance metrics, model weights, and computational times for each tested method	Generate new data	Digital	Experimental	.txt,.csv	<100 GB	
Experimental Forecasting Outputs	Case studies data confirming algorithm validation	Generate new data	Digital	Experimental	.nc, .grib2,.csv,.parquet	<100 GB	
Reports	Presentation and discussion of results	Generate new data	Digital	Software	.ppt,.doc(x),.pdf,.	<100 GB	
Figures and Graphs	Figures and graphs of processed and generated data	Generate new data	Digital	Software	.xls, .jpeg, .pfzx, .png	<100 GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

Elia Open data platform: https://www.elia.be/en/grid-data/open-data
ENTSO-E Transparency platform: https://transparency.entsoe.eu/

- Installed Capacity per Production Type [14.1.A]
- Actual Generation per Production Type [16.1.B&C]
- Generation Forecasts for Wind and Solar [14.1.D]
- Installed Capacity Per Production Unit [14.1.B]

Copernicus Open Access Hub: https://cds.climate.copernicus.eu/

- ERA5 reanlaysis DOI: 10.24381/cds.e2161bac

NCEP NOAA: https://registry.opendata.aws/noaa-gfs-bdp-pds/

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, ...)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

Renewable Generation Data, (4) Ørsted (private) data: Yes, there are restrictions specifically for the exploitation and dissemination of Ørsted's private site-level generation data. A unilateral confidentiality agreement has been signed between Ørsted and KU Leuven (acting on behalf of its research unit Electrical Energy Systems and Applications). Under this agreement:

- The data may only be used within the confines of the project.
- The data cannot be stored on any server that allows access to individuals outside the project.
- If I wish to publish work based on this data, Ørsted must be notified at least 30 days prior to the publication deadline and grant approval.

Global Physic-based Weather Model Data, (5) Rebase.energy data platform: Historical weather forecasts are publicly available, but not all are archived. To address this gap, we will use the Rebase.energy data platform to collect historical forecasts. As these data are sold commercially, they will not be shared.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

All weather datasets—Global Data-driven Weather Model Data, Local Raw Weather Observations, and Global Physic-based Weather Model Data—will be stored in a centralized repository. Given the expected terabyte-scale volume, we will use specialized preservation tools (e.g., ICTS at KU Leuven) and an organized folder structure. Each dataset will be divided into three versions:

- **Raw:** Unmodified, generated, or fetched data.
- **Intermediate:** Processed (cleaned/transformed) data.
- **Final:** Data directly used in the experimental forecasting models.

Each dataset folder will include following documentation:

- **Python Scripts:** All Python scripts used for preprocessing weather data or generating weather forecasts will be managed via a version control system in GitHub.
- **README Files:** Each Python script will be accompanied by a well-structured README.md file. This file will outline the preprocessing, fetching, or inference procedures, as well as software versions, parameters, file formats, and quality control measures.
- **Data Dictionaries:** A dedicated .txt file will list all weather variables, descriptions, units, and data coverage, including detailed latitude and longitude points. It will also specify the required input data for newly generated weather data. Related metadata will be recorded using standardized formats (e.g., UTC for time-zone data, ISO 19115 for geospatial information).
- **Data Collection or Generation Log Files:** Comprehensive logs documenting the data collection, processing steps, and analysis procedures will be maintained.

Energy datasets will follow the same documentation process as the weather datasets, with additional folder categorization into national-level and site-level data. Metadata for 'Distributed Solar/Wind Systems Physical Characteristics Data' will be maintained together in dedicated .xlsx and .txt files, accompanied by a README file detailing the data structure, variables, and usage guidelines.

Finally, trained models and the associated Python scripts for model training and testing will be managed via GitHub. Each repository will include a README.md file outlining the training and testing procedures, and experimental logs documenting digital records of optimal hyperparameters, performance metrics, model weights, and computational times for each tested method.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

For Python scripts, no strict standards currently exist for writing README.md files. Therefore, user-friendly documentation will be created to ensure ease of understanding.

3. Data storage & back-up during the research project

Where will the data be stored?

Scripts and models will be managed using GitHub repositories for version control and stored locally in the project owner's OneDrive folder.

Active research data will be stored on the portable HDD for quick access, while archived/inactive data will be compressed and stored on the ELECTA servers with automatic back-up procedures.

Both storage solutions are provided and managed by KU Leuven's ICTS. Published data will be made available following the publisher's standards.

How will the data be backed up?

Each commit and push to the GitHub repository automatically backs up the model and script via version control. In addition, OneDrive performs daily backups to the ELECTA network drive, where KU Leuven's ICTS then manages multiple backups each day for both active and archived data.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

For active data, we currently have a 5 TB HDD for storage, while ELECTA server ensures sufficient storage and backup capacity for archived data. As data collection continues, we can expand our capacity using additional resources from KU Leuven's ICTS or Energyville if needed. All solutions are scalable, allowing capacity to be expanded as needed.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The HDD is stored in a locked drawer at the ELECTA office. The network drive is secured by KU Leuven's ICTS service, and access to the owner's OneDrive folders on the ELECTA drive is limited to authorized users only. In addition, the GitHub repositories are managed solely by the author and secured with Multi-Factor Authentication (MFA).

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The project anticipates no additional costs for data storage and backup.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data obtained during this FWO project, except for Orsted’s site-level generation data due to contractual restrictions, will be retained for five years after the project ends. After this period, the data will remain accessible to ELECTA members.

Where will these data be archived (stored and curated for the long-term)?

Throughout the project, all archived data will be securely stored on ELECTA's server provided by KU Leuven's ICTS. Upon project completion, the data will be permanently transferred and archived on ELECTA's server. Additionally, all scripts and trained models will be maintained on GitHub to ensure proper version control and long-term accessibility.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

The project anticipates no additional costs for data preservation.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

All generated weather data, including its metadata, as well as the energy forecast model and its experimental results, will be made available.

If access is restricted, please specify who will be able to access the data and under what conditions.

All data mentioned above will be available without restrictions.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Intellectual Property Rights

There are restrictions specifically for the dissemination of Ørsted's private site-level generation data. A unilateral confidentiality agreement has been signed:

- The data may only be used within the confines of the project.
- The data cannot be stored on any server that allows access to individuals outside the project.
- If I wish to publish work based on this data, Ørsted must be notified at least 30 days prior to the publication deadline and grant approval.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Software (model and scripts) will be continually pushed to GitHub. The data and metadata will be stored and available through KU Leuven Research Data Repository and ELECTA network drive.

When will the data be made available?

Upon publication of the research results.

Which data usage licenses are you going to provide? If none, please explain why.

To be specified later.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

MIT License will be provided.

What are the expected costs for data sharing? How will these costs be covered?

There are no expected costs for data sharing.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Ada Canaydin - FWO fellow

Who will manage data storage and backup during the research project?

Ada Canaydin - FWO fellow

Who will manage data preservation and sharing?

Ada Canaydin - FWO fellow, prof. Hussain Kazmi- PI and ICTS

Who will update and implement this DMP?

Ada Canaydin - FWO fellow