Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D184B2FB2F

Title: Ultra-short-term irradiation forecasting combining traditional image processing with deep learning based on ground-based all-sky imagery

Creator: Lilla Barancsuk - **ORCID:** 0000-0002-3036-0133

Affiliation: Hun-ren Center For Energy Research

Principal Investigator: Veronika Groma, Dalma Günter, Lilla Barancsuk

Data Manager: Lilla Barancsuk

Project Administrator: Veronika Groma, Lilla Barancsuk

Contributor: Dalma Günter

Funder: Hungarian Research Network (hun-ren.hu)

Template: Digital Curation Centre

Project abstract:

Data Management Plan (DMP) for ground-based remote sensing all-sky imagery recorded using a wide lens all-sky camera, meteorological measurements, and sky-parameter dataset calculated from the images using image processing and neural network architectures pre-trained with the datasets. The data collected during the project is recorded at the HUN-REN Center for Energy Research (HUN-REN CER) as part of the internal research project "Ultra-short-term irradiation forecasting using a combination of deep learning and image processing". The research aims to support the proper quality of power grid services, by forecasting solar radiation in the ultra-short-term, leading to improved operation of photovoltaic plants. The forecast is based on sky camera images and weather data, using a combination of deep learning and traditional image processing methodologies.
The total expected data volume for the project is approximately 250 GB. This includes raw images, processed data, and derived datasets.

Keywords: all-sky-imagery, meteorological data, deep neural network, image processing, cloud

Start date: 11-01-2021

End date: 11-01-2026

Last modified: 05-28-2024

Copyright information:

Ultra-short-term irradiation forecasting combining traditional image processing with deep learning based on ground-based all-sky imagery

Data Collection

What data will you collect or create?

The following datasets are collected and/or created during the project:

1. **All-sky imagery**
   Approximately 400,000 all-sky images are recorded with a ground-based all-sky camera at the HUN-REN Center for Energy Research from 2021 November to 2024 November with a 1-minute recording frequency during daylight hours. The images are 370x370 pixel resolution, three-channel, 16-bit color depth, PNG format. The expected volume of images at the end of the project is 240 GB in total.

2. **Meteorological measurements**
   1-minute time resolution weather data recorded by a meteorological station. File format: .csv (comma-separated values), 3500 files, totaling 90 MB.

3. **Sky-parameter dataset**
   The dataset contains sky parameters extracted from all-sky images using traditional image processing algorithms, stored in tabular form. File format: .csv (comma-separated values), 3500 files, in total 90 MB. Contains the characteristic features of the sky and cloud, like cloud cover, and cloud homogeneity.

4. **Pre-trained neural network architectures**
   The architecture and trained weight parameters of the deep neural networks used for forecasting. The network architectures are developed during the project. Stored in the form of binary files. Approximately 40 TensorFlow HDF5 binary files, in total 400 MB.

5. **Forecast dataset**
   Datasets containing the forecasts produced by each network type, for multiple time horizons, for global and diffuse irradiation, the main product of the research project. The data is stored in tabular format, 1-miute resolution. Format: .csv (comma-separated values) files, totaling 1 GB.

6. **Research software**

   Four codebases were developed during the research project:

   C# codebase for camera control and image capture, including image quality control and preprocessing. Approximately 20 C# code files, totaling 60 KB.

   C++ Code based on the OpenCV library: Implementing image processing tasks; approximately 100 code files, totaling 300 KB.
   Python codebase for data analysis: Covering weather data preprocessing and cleaning, and temporal alignment of various data sources. Approximately 20 Python scripts (.py files), totaling 70 KB.

   Python codebase for Deep Artificial Neural Network training and evaluation. Implements the training, validation, and testing of deep learning models developed during the research project.

   Approximately 50 Python scripts (.py files), totaling 170 KB.
   The codebases will be open source through a public BitBucket repository.

The total expected data volume for the project is approximately 250 GB. This includes raw images, processed data, and derived datasets.

The amount of data justifies the storage of the datasets on institutional cloud storage infrastructures.

How will the data be collected or created?

1. **All-sky imagery**
   The images are recorded using a ground-based total-sky imager (TSI) located at the HUN-REN Center for Energy Research. The TSI is a high-resolution, 180° wide-angle color Starlight Xpress Oculus all-sky camera, with its optical axis positioned vertically upwards. For controlling the imaging process, custom software has been developed so that the frequency of photography and exposure time can be adjusted. The imaging has been ongoing since November 2021, and since then the camera has been continuously recording sky images during daylight hours. The collection of the images involves the calibration of the camera, capturing, recoloring, high dynamic range merging, and quality improvement of the images. The collection is automated, and the process is checked weekly.

   The images are stored in a nested folder structure that encodes the time of recording in the following manner: **year->month->day**

   Image names code the exact time of recording: **capture_<%Y%m%d%H%M>_<number of image in daily sequence>**

   Dataset quality is ensured by thorough testing of the capturing process and visual validation during the collection phase of the project. Image quality is also ensured by post-processing of already captured data.

2. **Meteorological measurements/weather data**
   Meteorological measurements are recorded by a weather monitoring system located in direct proximity to the camera, ensuring a high time-correspondence between the two recordings. The station measures the air temperature, surface wind speed, atmospheric pressure, relative humidity, and solar irradiance (including global and diffuse irradiation). These measurements are recorded at five-second intervals. The system logs one-minute average values and their corresponding one-minute standard deviations. The measurements are stored in a tabular format, in .csv files.

   Each file consists of an entire year of measurement data, and is named accordingly: **meteorological_measurements_<year>.csv**. All files are stored in a single folder.

   Dataset quality is ensured by using a standard, calibrated weather station, as well as data cleaning by removing invalid measurements, like negative humidity values and outliers from the dataset..

3. **Sky parameter dataset**
   The sky parameter dataset is a derived, secondary dataset extracted from the all-sky images by a cutom image processing software implemented in C++. The resulting .csv files contain the following sky-characteristics: cloud coverage, largest clear sky area, number of individual clouds, cloud inhomogeneity, degree of cloud periodicity, and average intensity for each image in the all-sky image dataset.

   The resulting files each contain an entire month of data with the filename referring to the encoded month: **sky_parameters_<%Y%m>.csv.**

   Quality is ensured by calibrating the image processing algorithms based on a ground-truth dataset representative of Hungarian cloud types and weather conditions, containing all-sky images and respective cloud masks created manually. The dataset represents an ideal segmentation, and serves as a ground truth for optimizing image processing algorithm parameters. The sky parameter dataset is created using the fine-tuned parameter image processing algorithms.

4. **Pre-trained neural network architectures**
   During the research, neural network architectures are trained using the sky-parameter dataset and the meteorological measurements to forecast global and diffuse solar irradiance in multi-time-horizon. The resulting trained deep neural networks are coded as binary files containing the neurons and the weight parameters.

   The network architectures are stored in a folder that encodes the type of network, and the file is named based on the input data combination used for training (i.e., scenario) in the following manner: **<architecture type> -> file:network_scenario**

Quality is ensured by thorough validation and testing of the performance of each network on an independent dataset. The performance indicators are made publicly available as well.

5. **Forecast dataset**
   Datasets containing the forecasts produced by each network type, for multiple time horizons, for global and diffuse irradiation.

   Stored in a folder structure that corresponds to that of the networks, that encodes the type of network the forecast was produced by, and the scenario (input data combination): **architecture type -> forecast_<scenario>.csv**

6. **Software**
   Research software will be organized according to the best practices of object oriented software and the requirements and best practices of the specific programming language.

Data collection, preprocessing steps and processing and training are described in detail in the following paper DOI: 10.3390/en17020438.

Documentation and Metadata

What documentation and metadata will accompany the data?

The following metadata and documentation will accompany the data:

Documentation:

The dataset will be published in the form of a data paper (in a peer-reviewed, international journal), detailing the measurement, methodology including the image capturing process, the image processing, and deep neural network training. The resulting datasets' main properties (such as format and file types) are also detailed in the paper. The documentation will contain the data storage and access as well.

Data collection, preprocessing steps, and processing and training are described in detail in the following paper DOI: 10.3390/en17020438.

Each research codebase will be accompanied with an in-code documentation, and a readme detailing the method of running, configuring and input creating process for the software.

In addition, brief documentation of each dataset will be provided in the form of readme files at the HUN-REN Data Repository Platform, where the datasets will be retained for long-term access and storage.

Metadata:

The **Dublin Core** standard will be used as the default schema for metadata, with the following fields:
Creator, Contributor, Publisher, Title, Date, Language, Format, Subject, Description, Identifier, Relation, Source, Type, Coverage, and Rights.

Some datasets deviate from this when necessary:

1. **All-sky imagery**
   Metadata for sky imagery will also contain custom fields that serve as a descriptor of the sky condition, as this is the most relevant information for filtering the dataset based on. The additional fields are:
   - sky condition (list): clear sky, partially cloudy, overcast
   - cloud type (list): altocumulus,cumulus,high-level,stratus,multi,clear sky

2. **Meteorological measurements**
   Default schema applies.

3. **Sky parameter dataset**
   Default schema applies. Additional field refers to the folder which served as a basis for the dataset.
   - based on: <folder name>

4. **Pre-trained neural network architectures**
   Additional fields describing the network architecture, commonly used in the deep learning community (based on the huggingface website for network models: https://huggingface.co/models) are:

   - task (multiple of the following): multimodal,computer vision,tabular
   - libraries (list): tensorflow,pytorch
   - language: english
   - dataset: custom
   - licences: CC 4.0

5. **Forecast dataset**
   Default schema applies.

6. **Software**
   Default shcema applies.

Most files already contain the metadata required, additional information will be added automatically via Python script.

Ethics and Legal Compliance

How will you manage any ethical issues?

No personal or sensitive data is handled during the project, and no ethical issues are associated with the data collection.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

The data is owned by the HUN-REN Center for Energy Research.

Data will be licensed under the Creative Commons licence: CC BY-NC-SA 4.0 Deed Attribution-NonCommercial-ShareAlike 4.0 International.

This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, for noncommercial purposes only. If others modify or adapt the material, they must license the modified material under identical terms.

Software will be licensed under the GNU Public Licence 3.0 (GPLv3). The licence permits the usage of the software for any purpose, changing and sharing the software and sharing the changes, but prohibits the licensing of the software for commercial use.

Storage and Backup

How will the data be stored and backed up during the research?

Each dataset described above is stored on an institutional cloud storage (NextCloud), and has a bacup on a local institutional server as well. Only one backup copy is created, due to the high reliability of the cloud storage. The backup is manually refreshed each month.

Backup and recovery is managed by Lilla Barancsuk.

In case of an incident, the data is recovered by the backup server manually.

Research software is stored in separate BitButcket Git repositories. Git repositories are distributed version control systems, storing the code files on computers with access to the repository, thus implementing a multi-backup system, providing a reliable storage space for software.

How will you manage access and security?

Access and security is managed by the institutional cloud infrastructure the data is stored on.

Software security is managed by BitBucket.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

All data described above will be retained for the long-term, as they can all contribute to research in the field:

- As benchmark datasets for all-sky imagery and weather:
Althoug already existing, similar all-sky imagery datasets are available, there is no dataset recorded specifically for Hungary (although this is a highly climate-dependent problem), or one with a joint weather dataset recorded at the same location and numerical sky parameters available.

- Pre-trained deep learning models:
Very few deep learning models pre-trained exist in the field, thus these can bring additional value to the research cmmunity

The amount of data justifies the retention as well, as it does not exceed 200 GBs, and can be made available through research data repositories.

For sharing and preservation, additional metadat will be added to the files.

- Software:
The implementing codebases can provide valuable basis for researchers in the field to validate our results and further the methodology. The codebases will be preserved on the long-term in BitBucket public repositories.

What is the long-term preservation plan for the dataset?

All datasets will be made openly available through the **HUN-REN Data Repository Platform**, which is free of charge for researchers working at the HUN-REN Center for Energy Research.

Website: https://researchdata.hu/en

The HUN-REN Data Repository Platform is selected due to the following characteristics:

- OpenAIRE compliant
- Wide metadata customizability
- Open access to researchers
- Catalogued in re3data
- Generalist repository
- Free of charge for institutional research data
- Long-term storage
- Provides persistent identifier (DOI)
- Data paper can be linked

Research software will be shared publicly via separate BitBucket Git repositories.

https://bitbucket.org

Data Sharing

How will you share the data?

The data and code will be openly available and free for research purposes.

The data will be shared through the HUN-REN Data Repository Platform in the long-term, the platform provides a persistent link and DOI for the dataset. The link and DOI will also be shared through the dataset paper documenting the dataset.

The data will be shared under the CC 4.0 licence, and made freely available for non-commercial purposes with crediting the authors via the DOI provided by the Repository.

Research software will be shared publicly via separate BitBucket Git repositories. The software will be open source under the GNU General Public License v3.0 (GPLv3).

Are any restrictions on data sharing required?

The data does not require an exclusive sharing period, and will be openly available for non-commercial use as soon, as it is uploaded to the Repository.

The usage is regulated under the CC 4.0 licence, permitting non-commercial usage by crediting the authors and sharing further under the same restrictions.

Software will be available throughout the course of the research on BitBucket under the GNU General Public License v3.0. that permits the usage and even redistribution of the software freely.

Responsibilities and Resources

Who will be responsible for data management?

Lilla Barancsuk is responsible for all activities of data management, including data capture, metadata production, data quality, storage and backup, data archiving, and data sharing.

Dalma Günter will support the above processes, in particular, metadata production and data quality tasks in the role of research assistant.

What resources will you require to deliver your plan?

All necessary technical and infrastructural requirements are already available.

The data management costs associated with the project are the following:

1. Compensation for Lilla Barancsuk in the role of data manager and Dalma Günter, the assistant researcher. They are responsible for data capture, metadata production, and storage during the data repositing time period (from April 2024 to November 2024).

Their work on these tasks is expected to require up to approximately 10 hours per week, and the costs include the salary of an assistant researcher in 10 hours (approximately €2000 in total). This is covered by the responsibles' respective salaries.

2. Cost of FAIR data publishing. This includes the publication fee in an open source data journal, which can cost up to €2000 (e.g. Nature Scientific Data open access article processing charge). APC will be covered by the institution, or by national grant.

Planned Research Outputs

Dataset - "All-sky imagery"

Approximately 400,000 all-sky images are recorded with a ground-based all-sky camera at the HUN-REN Center for Energy Research from 2021 November to 2024 November with a 1-minute recording frequency during daylight hours. The images are 370x370 pixel resolution, three-channel, 16-bit color depth, PNG format. The images amount to 120 GB in total.

The images are recorded using a ground-based total-sky imager (TSI) located at the HUN-REN Center for Energy Research. The TSI is a high-resolution, 180° wide-angle color Starlight Xpress Oculus all-sky camera, with its optical axis positioned vertically upwards. For controlling the imaging process, custom software has been developed so that the frequency of photography and exposure time can be adjusted. The imaging has been ongoing since November 2021, and since then the camera has been continuously recording sky images during daylight hours. The collection of the images involves the calibration of the camera, capturing, recoloring, high dynamic range merging, and quality improvement of the images. The collection is automated, and the process is checked weekly.

Dataset - "Weather dataset"

Meteorological measurements are recorded by a weather monitoring system located in direct proximity to the camera, ensuring a high time-correspondence between the two recordings. The station measures the air temperature, surface wind speed, atmospheric pressure, relative humidity, and solar irradiance (including global and diffuse irradiation). These measurements are recorded at five-second intervals. The system logs one-minute average values and their corresponding one-minute standard deviations. The measurements are stored in a tabular format, in .csv files.

1-minute time resolution weather data in tabular format. File format: .csv (comma-separated values), 3500 files, totaling 90 MB.

Dataset - "Sky parameter dataset"

The sky parameter dataset is a derived, secondary dataset extracted from the all-sky images by a custom image processing software implemented in C++. The resulting .csv files contain the following sky characteristics: cloud coverage, largest clear sky area, number of individual clouds, cloud inhomogeneity, degree of cloud periodicity, and average intensity for each image in the all-sky image dataset. File format: .csv (comma-separated values), 3500 files, in total 90 MB.

Model representation - "Pre-trained deep learning models for irradiation forecasting"

The architecture and trained weight parameters of the deep neural networks used for forecasting. The network architectures are developed during the project. The models are stored in the form of binary files, they amount to approximately 40 TensorFlow HDF5 binary files, in total 400 MB.

The models are trained using the sky-parameter dataset and the meteorological measurements to forecast global and diffuse solar irradiance in multi-time-horizons. The resulting trained deep neural networks are coded as binary files containing the neurons and the weight parameters.

Dataset - "Multi-time-horizon irradiation forecasts"

Datasets containing the forecasts produced by each network type, for multiple time horizonts, for global and diffuse irradiation, the main product of the research project. The data is stored in tabular format, 1-miute resolution. Format: .csv (comma-separated values) files, totaling 1 GB.

Data paper - "All-sky imagery with a joint weather dataset and derived sky parameters"

The description of the datasets will be published in the form of a data paper (in a peer-reviewed, international data journal, e.g. Earth System Science Data or Nature Scientific Data), detailing the measurement process, preprocessing and cleaning methodology including the meteorological data recording, image capturing, the image processing pipeline, and the deep neural network training. The resulting datasets' main properties (such as number, format and file types and statistics) are also detailed in the paper. The documentation will contain the data access, and metadata properties as well.

Software - "Software for ultra-short term irradiation forecast based on ground-based all-sky images, combining traditional image processing with deep learning"

Four codebases were developed during the research project:

1. C# codebase for camera control and image capture, including image quality control and preprocessing. Approximately 20 C# code files, totaling 60 KB.
2. C++ Code based on the OpenCV library: Implementing image processing tasks; approximately 100 code files, totaling 300 KB.
3. Python codebase for data analysis: Covering weather data preprocessing and cleaning, and temporal alignment of various data sources. Approximately 20 Python scripts (.py files), totaling 70 KB.
4. Python codebase for Deep Artificial Neural Network training and evaluation. Implements the training, validation, and testing of deep learning models developed during the research project. Approximately 50 Python scripts (.py files), totaling 170 KB.

The codebases will be open source through a public BitBucket repository.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
All-sky imagery	Dataset	2024-08-31	Open	HUN-REN Data Repository Platform		Creative Commons Attribution Non Commercial Share Alike 4.0 International	Dublin Core	No	No
Weather dataset	Dataset	2024-08-31	Open	HUN-REN Data Repository Platform		Creative Commons Attribution Non Commercial Share Alike 4.0 International	Dublin Core	No	No
Sky parameter dataset	Dataset	2024-08-31	Open	HUN-REN Data Repository Platform		Creative Commons Attribution Non Commercial Share Alike 4.0 International	Dublin Core	No	No
Pre-trained deep learning models for irradiation f ...	Model representation	2024-08-31	Open	HUN-REN Data Repository Platform		Creative Commons Attribution Non Commercial Share Alike 4.0 International	Dublin Core	No	No
Multi-time-horizon irradiation forecasts	Dataset	2024-08-31	Open	HUN-REN Data Repository Platform		Creative Commons Attribution Non Commercial Share Alike 4.0 International	Dublin Core	No	No
All-sky imagery with a joint weather dataset and d ...	Data paper	2024-09-30	Open	arxiv.org Nature Scientific Data		Open Publication License v1.0	None specified	No	No
Software for ultra-short term irradiation forecast ...	Software	2024-08-31	Open	Bitbucket		GNU General Public License v3.0 or later	None specified	No	No