Weather Generator

D1.1 Data Management Plan

Due date of deliverable	31/07/2025
Submission date	29/07/2025
File Name	D1.1 Data Management Plan
Work Package /Task	WP1/T1.1
Organisation Responsible of Deliverable	ECMWF
Author name(s)	Ilaria Luise, Christian Lessig, Tim Hunter, Tanya Warnaars, Peter Dueben
Revision number	1.1
Status	Final
Dissemination Level	Public

Funded by the European Union

The WeatherGenerator project (grant agreement No 101187947) is funded by the European Union.
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them.

1 Executive Summary

The WeatherGenerator Data Management Plan describes the specifications for data, quality control, metadata generation, data access, data stewardship and how data will be maintained and preserved. The types of data that will be used in the project include reanalyses and forecast data at different resolutions, geostationary and polar orbiting satellites and in-situ observations, radar data, and climate simulations. The data produced by the project will be primarily global and local forecast predictions at multiple resolutions and at multiple time scales for a wide range of variables depending on the application. The data of the project will comply with the FAIR data principles 'as open as possible and as closed as necessary'[1]. The produced data will be accessible using existing data portals that will be defined during the project.
This document is a living document which will be updated and amended during the lifetime of the project to follow and share the developments of the WeatherGenerator project.

[1] European Commission, Directorate-General for Research and Innovation, *Horizon Europe, open science – Early knowledge and data sharing, and open collaboration*, Publications Office of the European Union, 2021, https://data.europa.eu/doi/10.2777/18252

Table of Contents

2 Introduction

2.1 Background

The WeatherGenerator project will build the world's best generative Foundation Model of the Earth system – that will serve as a new Digital Twin for Destination Earth. The WeatherGenerator will be based on representation learning and create a general and versatile tool that models the dynamics of the Earth system based on a large variety of Earth system data. At the same time, it will integrate observations and simulations at a previously unseen level and scale.

This project brings together multiple Europe's leading scientific groups and research institutes as well as Small and Medium-sized Enterprises (SMEs) in the area of Earth system modelling, high-performance computing (HPC) and machine learning to build the WeatherGenerator as a new Digital Twin of DestinE. Once trained, the WeatherGenerator will be applied for selected high-impact applications in the energy, food, water and health sectors.

The WeatherGenerator will lead to key innovations in weather and climate science and machine learning to enable Europe to defend its leadership in Earth system modelling. The WeatherGenerator will define a new state-of-the-art in both machine learning and weather and climate sciences. Through its vastly improved efficiency and flexibility compared to current Earth system models, the WeatherGenerator will create new opportunities for fast DestinE services that allow testing of many different management options. This will enable new levels of interactivity for a large user base, including, for example, city planners, regional and national authorities, architects, and engineering companies.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverables

This deliverable contains a summary of all the datasets used in the WeatherGenerator project. It also describes the foreseen metadata handling and dissemination strategies.

2.2.2 Work performed in this deliverable

In this deliverable the work as planned in the Description of Action (DoA, WP1 T1.1) was performed.

2.2.3 Deviations and counter measures

No deviations have been encountered.

2.2.4 Reference Documents

[1] [RD 1] 101082194-WeatherGenerator-HORIZON-INFRA-2024-TECH-01-03 - New digital twins for Destination Earth- Description of the Action

[2] European Commission, Directorate-General for Research and Innovation, *Horizon Europe, open science – Early knowledge and data sharing, and open collaboration*, Publications Office of the European Union, 2021, https://data.europa.eu/doi/10.2777/18252

2.2.5 WeatherGenerator Project Partners:

ECMWF	EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS
FZJ	FORSCHUNGSZENTRUM JUELICH
MetNor	NORWEGIAN METEOROLOGICAL INSTITUTE
MPG	MAX-PLANCK-GESELLSCHAFT ZUR FÖRDERUNG DER WISSENSCHAFTEN E.V.

KNMI	ROYAL NETHERLANDS METEOROLOGICAL INSTITUTE
MetFrance	MÉTÉO-FRANCE
SMHI	SWEDISH METEOROLOGICAL AND HYDROLOGICAL INSTITUTE
UKMO	UK METOFFICE
CMCC	CENTRO EURO-MEDITERRANEO SUI CAMBIAMENTI CLIMATICI
eScience	NETHERLANDS ESCIENCE CENTER
Buluttan	BULUTTAN
KAJO	KAJO SERVICES
LT	LATEST THINKING
Statkraft	STATKRAFT
ETHZ	EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE
MetSwiss	METEOSWISS

3 Data Summary

Our Data Management Plan (DMP) is developed following the standard approach the Data.europa.eu data quality guidelines[2] whereby it sets out the specifications for data, quality control, metadata generation, data access, data stewardship and how data will be maintained and preserved. It is developed to provide guidelines to adhere to article 17 to the Grant Agreement. As with scientific peer-reviewed publications, datasets generated by the project will be deposited in repositories and made Open Access. Data will be made freely available for use where possible. To facilitate the exploitation and monitoring of the Data Management Plan a specific Task 1.1 (WP1) is responsible for this activity.

3.1 Definitions related to the approach to Open Science:

The Horizon Europe programme guide states[3]: "*Open science is an approach based on open cooperative work and systematic sharing of knowledge and tools as early and widely as possible in the process.*" In this regard we clarify for WeatherGenerator the vocabulary on open access below:

Open Access Data: Open access refers to unrestricted access to research results. Commonly, the open access characterization is given to open-source peer-reviewed publications, datasets, tools and source code. Open access focuses on building a community and enables scientists, researchers, interest groups and individuals to:

- Build and enhance existing research results
- Avoid redundancy
- Participate in Open Innovation activities
- Benefit from the results of the WeatherGenerator project

Open Research Data: Open research data refers to the disclosure of the linked research data which are needed to assess, validate and replicate the results presented in research publications. Complementary to the concept of open access, open research data enables the online availability of data resources towards promoting research.

The open research data concept focuses on enabling researchers and individuals to:

- understand, assess, reconstruct and further expand scientific publications
- build innovative concepts on top of existing research data
- establish a continuous improvement mechanism of research

3.2 Approach

The general strategy for data management sets out the specifications for data, quality control, metadata generation, data access, data stewardship and how data will be maintained and preserved. The types of data that will be used or produced in the project are reanalysis,

[2] Data.europa.eu data quality guidelines https://op.europa.eu/en/publication-detail/-/publication/023ce8e4-50c8-11ec-91ac-01aa75ed71a1/language-en

[3] Guidelines on FAIR Data Management in Horizon Europe (Version 2.0, 01 April 2022), https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf

forecast data, satellite and in-situ[4] observations, radar data as well as climate simulations. . This includes output generated by the WeatherGenerator model itself.

4 FAIR Data

The data of the project will comply with the FAIR data principles, as much as possible. Most of the data are accessible using existing public data portals, such as the Copernicus Data Store, the EUMETSAT data store and the NASA data store. All these data portals have been designed to support interoperability and include clear licensing information as well as tools to make best use of the data.
Each participating organization will examine whether open access of the WeatherGenerator output data can be granted without affecting any legal and ethical requirements, including the Intellectual Property Rights as per the dissemination access level of each dataset produced as output of the WeatherGenerator evaluation.
This DMP follows the EU guidelines[1] and describes the data management procedures according to the FAIR principles[5]. The acronym FAIR identifies the main features that the project research data must have in order be findable, accessible, interoperable and reusable.

4.1 Making data findable, including provisions for metadata

Importance is placed on enhancing the discoverability of the collected and generated data. Metadata links information and data across the web and constitutes a powerful tool that helps individuals (researchers, developers, citizens, etc.) to discover, identify, and manage digital resources. The metadata are defined as the unique set of information used to describe and identify the data collected and/or generated. It is usually structured as textual information that describes the creation, content, or context of a digital resource. The most notably known types of metadata are names, dates, location, data types, relations and interdependencies to other data sets.

In order to make data findable, searchable and to expose the metadata in a unified way, we use a database based on STAC[6] with an entry for each dataset used in the project and unique numerical identifiers to facilitate discovery. STAC is an open standard to expose collections of spatio-temporal data in the domain. The STAC database is exposed through a web browser to increase discoverability and readability. The STAC database contains one file per dataset, which is in json format and therefore easy to further process or to present in human-readable form. Each file of the database describes the content, status and versioning of a dataset. The WeatherGenerator data catalogue (Figure 1) can be found at this link. At the moment the database is hosted on the EERIE cloud at DKRZ. This limits the flexibility in terms of rendering and page layout. In the future we thus plan to host the database on a different server, to be identified later, to have full control of the page layout for an optimized user experience.

[4] In the current EU Space Regulation, in-situ observations are defined as follows: ‘Copernicus in-situ data’ means observation data from ground-based, seaborne or airborne sensors, as well as reference and ancillary data licensed or provided for use in Copernicus
[5] The FAIR data principles (GO FAIR https://www.go-fair.org/fair-principles/)
[6] The STAC library: https://stacspec.org/en

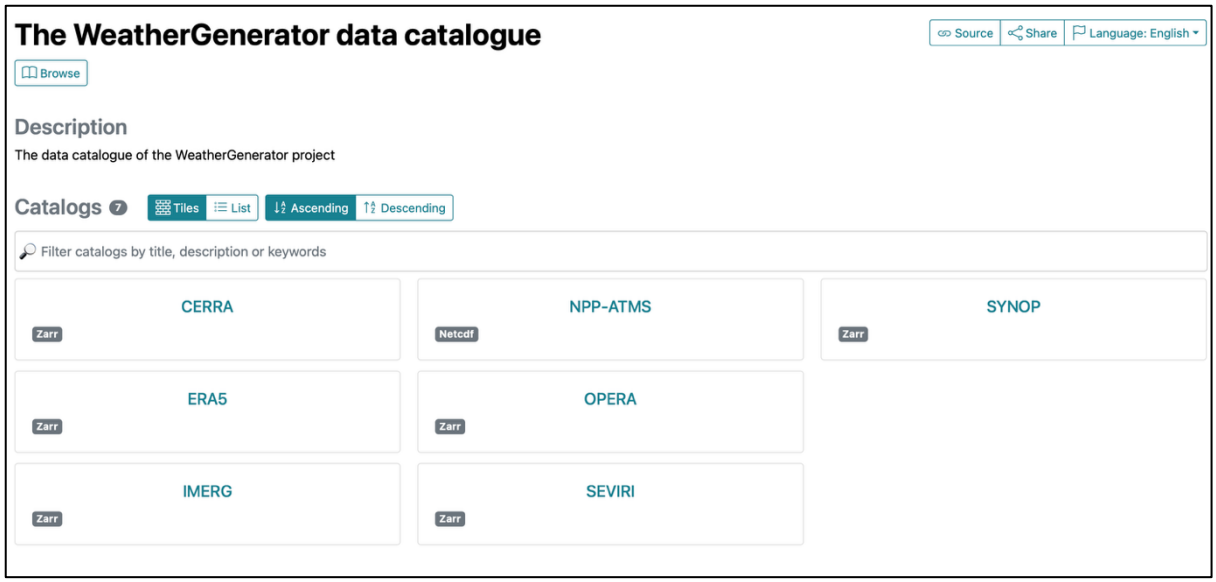

Figure 1: Homepage of the WeatherGenerator STAC catalogue.

4.2 Making data accessible

FAIR open access to the data guide refers to making data accessible to all project partners, researchers and the public, following the privacy and anonymity guidelines of the EU and national regulations. Accessibility in the context of Horizon Europe, means that all data generated and used, if possible, are publicly open and available. The WeatherGenerator partnership will ensure the integrity of personal data and sensitive information prior to the dissemination of the datasets.

The project does not aim to replicate any data. The project partners will maintain the list of datasets utilised for all WeatherGenerator activities up-to-date on the above mentioned STAC catalog. In this way the accessibility of the data will be ensured at two levels: internally to the project, and to the general public, i.e. regarding where the data are stored in the various HPCs and, for the public and when available, a link pointing to the provider to download the data autonomously.

During the execution of the project, each partner will provide detailed information on privacy/confidentiality and the procedures that will be implemented for data collection, storage, access, sharing policies (especially when third party countries are concerned), protection, retention and destruction. The consortium will confirm that the project complies with national and EU legislation throughout its lifetime and after its completion.

As a guiding principle, the WeatherGenerator seeks to ensure open access to research data, via repositories, as soon as possible and within the limits and deadlines set out in the DMP, in order to allow dissemination, validation and re-use of research results. During the project, the model outputs, the scores produced during evaluation and in general all relevant output data will be publicly stored on an OpenStack S3-style bucket hosted at the European Weather Cloud (EWC), as shown in Figure 2. The European Weather Cloud is a federated cloud computing infrastructure developed by ECMWF and EUMETSAT to provide secure, scalable, and collaborative access to weather and climate data and services for European meteorological organizations. The use of the S3 bucket at the EWC ensures accessibility by all project partners. Data deposition in repositories will facilitate the project reporting procedures and guarantee long time preservation and accessibility to datasets.

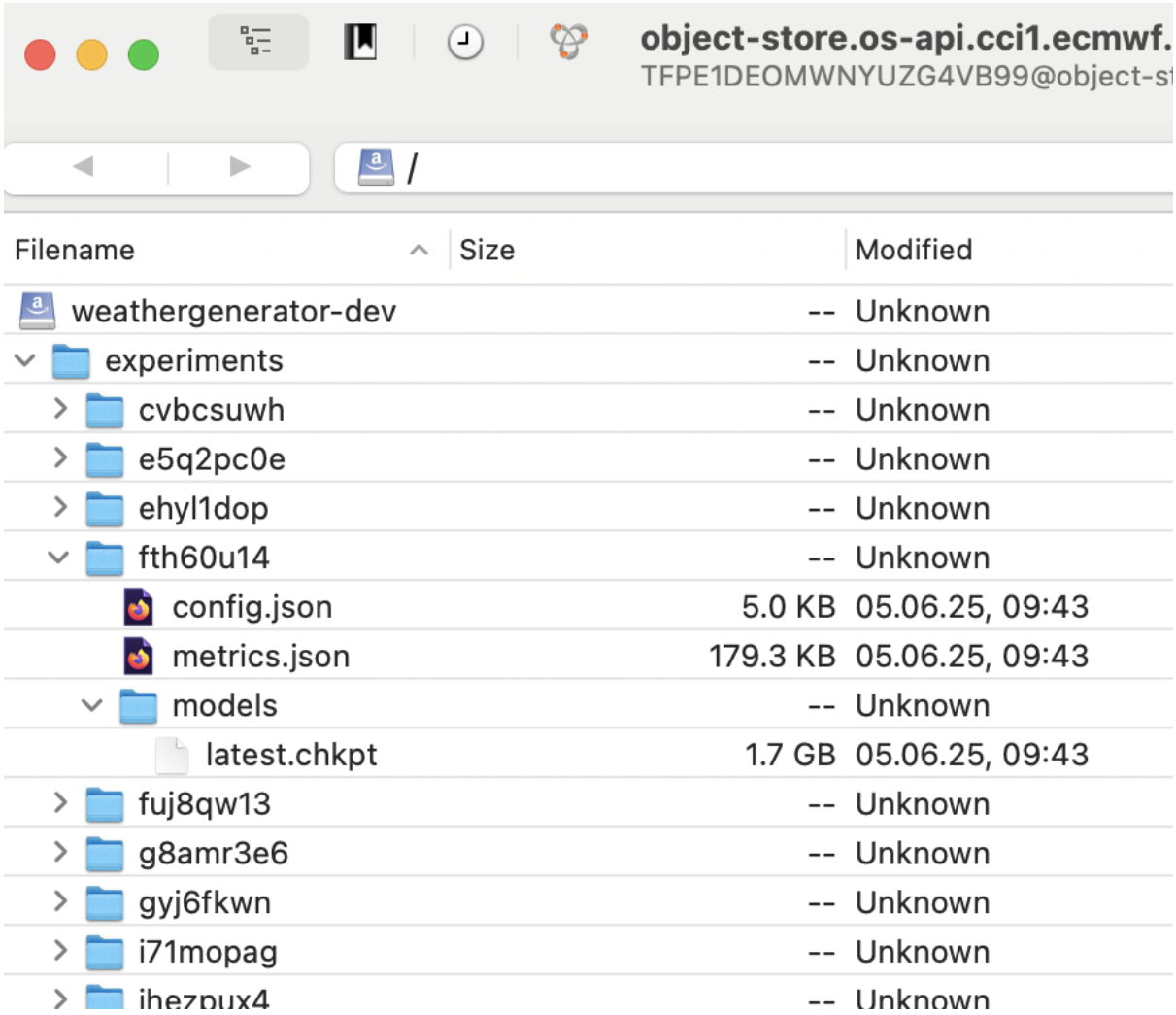

Figure 2: screenshot of the OpenStack S3-style bucket used to store the data generated by the WeatherGenerator. Each folder refers to a different train or inference run and contains a config.json file with all the settings to reproduce it. Please note the data deposition in repositories to facilitate readability and findability.

Restrictions to access are applied only in the following cases:

- when collected data belongs to third party which have denied permission for sharing them;
- on account of confidentiality and proprietary issues;
- protection of personal data of subjects involved in the research
- when availability of the data would mean that the project's main aim might not be achieved.

For data that falls under any of the restrictions described above and for which it is not possible to take any action to make them shareable, EU allows complete closure or restricted access to them. Nevertheless, even in such cases we aim to make at least the existence of the dataset known so that individual negotiations and non-disclosure-agreements can be sought.

The WeatherGenerator DMP indicates the versions or parts of the data sets that can(not) be freely shared providing the specific details in Annex II. The specific repositories for data set publication and preservation will be further expanded during the project.

4.1 Making data interoperable

Data interoperability refers to the ability of systems and services to access readable and editable data, in terms of their content, context and meaning. To achieve it, the WeatherGenerator project will incorporate suitable standards and vocabularies for data and metadata creation. In the case of WeatherGenerator, the primary end user of the data is the weather and climate community as well as national weather services. The level of integration with existing services is therefore a driver for the project, as WeatherGenerator products need to be interoperable with the existing applications and workflows of the partners involved in the project, e.g. applicable WMO standards.

To allow data exchange and re-use among researchers, institutions, organisations, countries, etc., partners will make them available in well-known and documented open formats, as much as possible compliant with available (open) applications.

In particular, the current model output is in Zarr format, which is increasingly becoming a new community standard for machine learning in weather and climate. In addition, we will also provide tools that allow to access the model output in the standard formats used by the community, such as the grib(2) and/or netCDF. The format will be chosen in agreement with the project partners to maximise the compatibility with their existing tools and workflows.

4.2 Increase data re-use

The GO FAIR principles state “FAIR is to optimise the reuse of data”. Data availability after the end of the project depends highly on the type and content of data, taking into account sensitivity and specific licences. Data should be available for public reusability after being granted permission from their respective contributors, following the proposed legal and ethics requirements.

Rich metadata will enable proper discovery and identification of the data along with the appropriate licensing schemes facilitating their re-usability. In principle, it is expected that data will become available after the publication of the respective deliverables and will remain available after the completion of the project.

To safeguard the transparency, consistency, quality, completeness and accuracy of the data, the WeatherGenerator adopts a data quality assurance procedure. For each new dataset the project partners are required to follow a standardized set of guidelines:

- Add the scripts used to download the dataset inside the `$WEATHERGEN-PRIVATE/data/download` folder[7], adding a `DOWNLOAD_README.md` file containing the instructions to use the scripts.
- Check for NaNs and document it e.g. in a dedicated document in OnlyOffice
- Check for data quality flags and document them
- Check for unreasonably large values and document them
- If needed, compute means, standard deviations and tendencies, like in these scripts: `$WEATHERGEN-PRIVATE/data/preprocessing`
- Add an entry in the STAC database following the instructions in the WeatherGenerator code repository (link).

[7] The `$WEATHERGEN-PRIVATE` folder is a *private* GitLab project hosted by JSC to store private paths and settings (e.g. access keys) for the different HPCs. It is also the main archive for standalone scripts related to data handling, including downloading and pre-processing.

5 Other research outputs

Other research data will be stored on trusted servers (e.g. ecmwf sites) or cloud (e.g. B2Drop and OnlyOffice). This is particularly relevant to project documents, presentations and deliverables.

6 Allocation of resources

The WeatherGenerator is a multi-site project where the different partners run on different HPC facilities. This is a challenge in terms of data distribution, as the data need to be maintained and transferred to multiple sites. A protocol to move data across HPCs is currently under development. When possible, the datasets are distributed to the other sites through the European Weather Cloud, while in the other cases the consortium is exploring alternatives like uftp or globus.

At the moment, we have the current allocated storage resources across different sites:

- HPC2020: 1PB – 50M inodes
- JSC: 400 TB (on SCRATCH disk) – 35M inodes
- ALPS: 1.6 PB – 10M inodes
- Leonardo: 10T

The data are stored in a common data folder per site, accessible to all project partners. The location of these folders differs among the HPC sites, so the paths are documented within the `$WEATHERGEN-PRIVATE` repository. The workflow is organized such as the WeatherGenerator code automatically detects the HPC and the path to the input data folder. Dedicated PMs from WP1 and WP2 are allocated to supervise and maintain such pipeline in place, as well as the dataset folders at each site.

The resources required for making the data generated by WeatherGenerator FAIR have been included in the budget of the project. In general, the WeatherGenerator, and in particular the partners responsible for WP1, will decide and contribute to relevant aspects of the data management cycle for the duration of this project. Monthly meetings dedicated to data management are in place to ensure regular updates among the partners on the status of the different datasets and increase data findability and re-use. A specific table summarising the research team leaders responsible for each dataset will be added in the future release of the DMP. The data management plan will be reassessed and updated every 6-months after regular re-iterations of the survey campaign added to Annex II within all project partners.

At this state, the chosen repository for long term deposition and preservation of searchable data intended for public use, does not apply fees for archiving and data curation. Peer-reviewed publications costs related to open-access research data are eligible in Horizon Europe and will be covered by the WeatherGenerator budget.

7 Data security

The WeatherGenerator consortium places a strong emphasis on ensuring the security of all the datasets, safeguarding them from unauthorized access and loss.

7.1 Input data

When possible, the input data are stored on long-term storage partitions and at multiple sites to prevent losses. In some cases, like e.g. at JSC, the data are temporarily stored on short-term storage ($SCRATCH). Chrono jobs have been put in place to touch the data at regular intervals to prevent data deletion. This is considered as a temporary solution while waiting for JUPITER to be operational. The data access is restricted only to the members of the consortium.

7.2 Output data and other project artifacts

To make the data produced by the WeatherGenerator project publicly accessible in dedicated public repositories, the data will be stored in S3-style buckets through the OpenStack platform deployed on the European Weather Cloud (hosted by ECMWF). Other output data, e.g. datasets generated for public dissemination could be published through the MeteoCloud (hosted by JSC). All these platforms ensure a robust and rigorous data security system. The physical security it is attained at high standards on a best effort basis to guarantee 24/7 monitoring. It also includes fire suppress and, where available, power backup systems.

The S3 buckets follow standard security protocols to minimize the risk of accidental information leakage. Two buckets have been created:

- weathergen-bucket-dev : read/write access to developers only. This is the bucket used throughout the course of developing new models. It is used by all the developers to upload their experiments and communicate among each other. A retention policy of 1 year is in place to limit space usage and the risk of accidentally storing unnecessary information. Each partner institute has its own access key and these keys are expected to be rotated on a regular basis. All data upload goes through automated scripts, but developers also have the possibility to directly edit and remove data if necessary. Furthermore, following the principle of least privilege, all automated scripts that only need to read data use a read-only key (LIST and GET permissions) for all operations.

- weathergen-bucket-release: public read access, write access through release scripts. This is the main storage area for sharing models and results with the wider public audience. It also serves as an archive of the public releases of the models. To limit risks of (1) accidental deletion or (2) accidental upload, all data upload goes through automated scripts. The public has GET access but no LIST access.

7.3 Sensitive data

All the relevant personal protection protocols, such as GDPR, ECMWF's Personally Identifiable Information Protection and relevant national legislation, will be applied on information of an individual and any reference to personal data or sensitive information will be fully masked in any printed materials, project reports or dissemination activities. Personal data, such as personal information from project partners' members, will be treated confidentially, taking into consideration all the proper technical means. General and personal data will be stored separately. All personal data not needed for the final report, will be destroyed at the end of the project and not retained after the completion of the final report.

8 Ethics

All details about ethics and legal compliance in terms of current EU legislative initiatives have been considered and are not of relevance at this point for the data arising from WeatherGenerator. Additionally, the Grant Agreement and the WeatherGenerator Consortium Agreement are to be referred to for further details on the ownership and management of intellectual property and access.

No ethics or legal issues are foreseen in the project apart from the respect of the GDPR rules, ECMWF's Personally Identifiable Information Protection and relevant national legislation when gathering personal information.

9 Conclusion

In this deliverable, the WeatherGenerator Data Management Plan has been initiated.
Whilst this provides a good starting point for the FAIR data activities of the WeatherGenerator project, it nevertheless needs careful further reflection and updating when appropriate to ensure that new developments (technical as well as strategic) within the WeatherGenerator project and beyond are well reflected by the Data Management Plan. The WeatherGenerator Consortium will ensure that all generated datasets do not infringe either partner IPR rules or regulations related to personal data protection.

Document History

Version	Author(s)	Date	Changes
1.0	Tanya Warnaas, Ilaria Luise, Christian Lessig, Timothee Hunter	17/06/2025	Initial version
1.1	Tanya Warnaas, Ilaria Luise	15/07/2025	Added comments from reviewers

Internal Review History

Internal Reviewers	Date	Comments
Tanya Warnaars	24/06/2025	Question on clarification, figure formatting and layout.
Martin Schultz,	09/07/2025	Clarifications requested
Tom Dunsten	09/07/2025	Clarifications and editing

10 ANNEX I

Annex I includes the text of the questionnaire that was shared with each Work Package to gather, in table format, the data sets by WPs. The table below shows what was asked in order to describe if data:

- is available, or
- will be generated, or
- will be collected

Work Package X

<Data set reference and name>	
Data set description	*Description of the data that will be generated or collected (or is already available to the project), its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.* *Limitations?* *Constraints?*
Standards and metadata	*Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.* *Will you generate proper metadata for you data?* *If yes: how do they look like?* *If no: why?* *Data format?* *Will there be a review process to quality- check the data?*
Data Sharing	*Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).* *In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).* *License?* *Access URL?*

Archiving and preservation (including storage and backup)	*Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.* *At which Data Center do you want to store your data?* *Is there an established workflow for your requested DOI process in place?* *According to which standards*

11 ANNEX II

Annex II includes an extensive list of the datasets, already available or to be developed in the context of the project's research and implementation activities. The WeatherGenerator Theme 3 (WP 5 and WP6) is planned to start only in August 2025. The WP1 is responsible for the data handling of both the WeatherGenerator Theme 1 (WP1 and WP2) and Theme 2 (WP 3 and WP4). Thus, at this stage in the project the list includes contributions from WP1 only. The table below shows each data set that:

- is available, or
- will be generated, or
- will be collected

(Note that this is a living document and the information included here may be subject to change throughout the lifetime of the project).

Work Package 1:

<Data set reference and name>	ERA5 reanalysis
Data set description	The ERA5 reanalysis dataset, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) under the Copernicus Climate Change Service (C3S), provides consistent and high-resolution global atmospheric, land, and oceanic data. Covering the period from 1950 to the present, ERA5 offers hourly estimates at a spatial resolution of approximately 30 km (0.25° grid). *Limitations? No Limitations* *Constraints? No Constraints*
Standards and metadata	ERA5 complies with established climate and geospatial data standards by providing data in .grib2 format following CF (Climate and Forecast) metadata conventions, ensuring interoperability, consistent variable naming, units, and geolocation metadata in line with community best practices for climate science. *Will you generate proper metadata for your data?* A STAC-based metadata catalogue to expose the metadata of all the datasets used by the WeatherGenerator is under development (see above). *Data format?* Original data format is .grib2, and the dataset has been converted into .zarr for better compatibility with the current machine learning standards *Will there be a review process to quality- check the data?* The data already undergo extensive quality-checks by the Copernicus Climate Change Service before release.
Data Sharing	ERA5 is openly available via the Copernicus Climate Data Store (CDS), with robust documentation and support for reproducible climate research and long-term data accessibility. We plan to publish the .zarr dataset for dissemination purposes. No specific software or other tools will be necessary to use these data.

	The final storage repository will be available by the end of 2025 in a WeatherGenerator large data storage project at the Juelich Supercomputing Center. For the moment, the project partners are using the data through copies on standard repositories at the Juelich Supercomputing Center, Leonardo and Levante. *License?* The dataset is open source under the Copernicus Products Licence *Access URL?* Link to Copernicus Data Store
Archiving and preservation (including storage and backup)	The final archiving strategy will be finalised at a later stage. For the moment different copies of the dataset are available at multiple locations (ECMWF's ATOS, Leonardo, MareNostrum5, Lumi) as well as at the European Weather Cloud through a different project. The data has been copied at JSC and Levante and made available to all project partners. *At which Data Center do you want to store your data?* Juelich Supercomputing Center *Is there an established workflow for your requested DOI process in place?* The dataset already has an assigned DOI: 10.24381/cds.adbb2d47

<Data set reference and name>	CERRA, Copernicus European Regional Reanalysis
Data set description	The Copernicus European Regional Reanalysis (CERRA), developed under the Copernicus Climate Change Service (C3S), provides a high-resolution reanalysis of atmospheric conditions over Europe. Covering the period from 1984 onward, CERRA delivers hourly data at a spatial resolution of 5.5 km, capturing fine-scale climate and weather patterns with improved detail compared to global reanalyses. *Limitations? No limitations* *Constraints? No constraints*
Standards and metadata	The CERRA dataset adheres to recognized climate science data standards by being distributed in .grib2 or NetCDF format and conforming to the Climate and Forecast (CF) metadata conventions, which ensures compatibility with widely used analysis tools, supports data interoperability, and facilitates long-term accessibility and reuse within the geosciences community. *Will you generate proper metadata for you data?* A STAC-based metadata catalogue to expose the metadata of all the datasets used by the WeatherGenerator is under development (see above). *Data format?* Original data format is .grib2, and the dataset has been converted into .zarr for better compatibility with the current machine learning standards *Will there be a review process to quality- check the data?* The data already undergo extensive quality-checks by the Copernicus Climate Change Service before release.
Data Sharing	CERRA is openly accessible through the Copernicus Climate Data Store (CDS), offering detailed documentation and ensuring reproducibility, transparency, and long-term data accessibility.

	The final storage repository will be available by the end of 2025 within a WeatherGenerator large data storage project at the Juelich Supercomputing Center. For the moment, the project partners are using the data through copies on standard repositories at different HPC facilities (ECMWF's HPC, Leonardo, MareNostrum). *License?* The dataset is open source under the Copernicus Products Licence *Access URL? Link to Copernicus Data Store*
Archiving and preservation (including storage and backup)	The final archiving strategy will be finalised at a later stage. For the moment the data are available on ECMWF's ATOS, Leonardo and MareNostrum5 as well as the European Weather Cloud through a different project. *At which Data Center do you want to store your data?* Juelich Supercomputing Center *Is there an established workflow for your requested DOI process in place?* The dataset already has an assigned DOI: 10.24381/cds.a39ff99f

<Data set reference and name>	IMERG
Data set description	The IMERG (Integrated Multi-satellitE Retrievals for GPM) dataset is a NASA product that provides high-resolution global precipitation estimates by combining data from multiple satellites from the Global Precipitation Measurement (GPM) and the Tropical Rainfall Measuring Mission (TRMM) with rain gauge analysis products. It offers half-hourly rainfall data at 0.1° resolution, between 1998 and the present. Three IMERG products of varying quality and different latency times are available. In the WeatherGenerator project, level 3 data from the Final Run version 07 is used, since it provides research-quality gridded precipitation estimates undergoing several adjustment processing steps. *Limitations?* The data quality of the IMERG dataset is reduced in polar regions due to sparse surface observations and the inherent challenges in accurately estimating precipitation from satellite data, particularly where frozen precipitation predominates. *Constraints?* Data access requires a NASA Earthdata account (see below).
Standards and metadata	IMERG data comply with established Earth Science and geospatial data standards. The IMERG data is available in various standard formats such as GeoTIFF, HDF5, netCDF and OPenNDAP. *Will you generate proper metadata for you data?* Metadata provision follows CF Metadata Conventions, particularly when provided in netCDF. *Data format?* For better compatibility with current machine learning standards, the half-hourly netCDF data files are converted to zarr-format. *Will there be a review process to quality- check the data?* The data already undergoes extensive quality-checks and adjustments before the release of its Final Run product.

Data Sharing	IMERG data are openly available through NASA's GPM Data Directory and the NASA GES DISC archive, with multiple formats and access options including direct download and cloud access. Users can register for free with NASA Earthdata for full access, and extensive documentation is provided to support reproducible research and long-term data accessibility. *License?* The IMERG data are distributed by NASA under an open data policy. *Access URL?* Main access.
Archiving and preservation (including storage and backup)	The final archiving strategy will be finalised at a later stage. The complete dataset has been downloaded at ECMWF's ATOS. For the time being, the data was converted to zarr-format, and the temporal resolution was coarsened to 6 hours. A copy of the zarr 6h-dataset is being stored at the Juelich Supercomputing Center (JSC) and made available to all project partners. *At which Data Center do you want to store your data?* Juelich Supercomputing Center *Is there an established workflow for your requested DOI process in place?* The following DOI is available for the native dataset in use: https://doi.org/10.5065/7DE2-M746

<Data set reference and name>	NPP-ATMS, Advanced Technology Microwave Sounder (ATMS) on the National Polar-orbiting Partnership (NPP) satellites.
Data set description	The NPP-ATMS (Advanced Technology Microwave Sounder) dataset is derived from the ATMS instrument onboard the NOAA/NASA National Polar-orbiting Partnership (NPP) satellites. It has 22 channels (23.8 GHz to 183.3 GHz) providing near-global coverage twice/day of atmospheric temperature, moisture, and pressure profiles, crucial for weather forecasting and climate monitoring. With a spatial resolution of 16 km, NPP-ATMS data are collected at high temporal frequency (6 min, 1.5 hours) and span multiple microwave channels, offering detailed atmospheric observations. The NPP-ATMS data are available from 2012 to present. *Limitations?* Large number of files *Constraints?* No Constraints
Standards and metadata	The data are distributed in NetCDF or HDF5 formats, adhering to climate science metadata standards and ensuring compatibility with common atmospheric science tools. *Will you generate proper metadata for you data?* A STAC-based metadata catalogue to expose the metadata of all the datasets used by the WeatherGenerator is under development (see above). *Data format?* The original data format is .netCDF

	Will there be a review process to quality- check the data? The data undergo extensive quality-checks before release.
Data Sharing	The NPP-ATMS dataset is available through the EUMETSAT Data Store as well as through NASA's Earth Science Data and Information System (ESDIS) and NOAA's National Centers for Environmental Information (NCEI) providing long-term data accessibility. *License?* The data are available without charge under EUMETSAT CC-BY-4.0 licence. *Access URL? https://user.eumetsat.int/catalogue/EO:EUM:DAT:0345*
Archiving and preservation (including storage and backup)	The final storage repository will be available by the end of 2025 as part of a WeatherGenerator large data storage project at the Juelich Supercomputing Center. For the moment the data are being downloaded on ECMWF's ATOS and will be transferred at JSC at a later stage. *At which Data Center do you want to store your data?* Juelich Supercomputing Center *Is there an established workflow for your requested DOI process in place?* The dataset already has an assigned DOI: *https://doi.org/10.15770/EUM_SEC_CLM_0034*

<Data set reference and name>	OPERA, Operational Programme for the Exchange of RADAR information
Data set description	The OPERA radar dataset is produced by the EUMETNET OPERA program, which coordinates and harmonizes European weather radar observations. It provides quality-controlled, pan-European radar composites and individual radar data from national meteorological services. These data include measurements such as radar reflectivity and derived precipitation products, with high spatial and temporal resolution (typically 1 km × 1 km grids at 5–15 minute intervals). Data are available from 2013 to 2023. *Limitations?* Data access requires approval by EUMETNET *Constraints?* No constraints
Standards and metadata	OPERA data follow the OPERA Data Information Model (ODIM) and are delivered in HDF5-based formats (ODIM_H5), which are compatible with WMO standards and commonly used radar processing tools. *Will you generate proper metadata for you data?* OPERA radar data are accompanied by metadata structured according to the ODIM_H5 standard, ensuring clear documentation of spatial and temporal coverage, radar parameters, quality indicators, and processing history. data already contain pre-defined metadata. A STAC-based metadata catalogue to expose the metadata of all the datasets used by the WeatherGenerator is under development (see above). *Data format?* Original data format is ODIM, compliant with BUFR and HDF5 formats, which are common standards in the community. The dataset has been converted into .zarr for better compatibility with the current machine learning standards *Will there be a review process to quality- check the data?*

	The data already undergo extensive quality-checks by the *OPERA* program before release.
Data Sharing	The data are hosted and distributed via EUMETNET and selected data portals (e.g., EUMETView or national repositories), ensuring long-term data accessibility and interoperability within the European radar network. *License?* Free to use under a research account license at EUMETNET *Access URL?* Available by contacting support.opera[at]eumetnet.eu
Archiving and preservation (including storage and backup)	The final archiving strategy will be finalised at a later stage. For the moment different copies of the dataset are available at multiple locations (ECMWF's ATOS, and MareNostrum5) as well as at the European Weather Cloud through a different project. The data has been copied at JSC and made available to all project partners. *At which Data Center do you want to store your data?* Juelich Supercomputing Center *Is there an established workflow for your requested DOI process in place?* The **dataset** does not currently have a single, centralized DOI for the entire dataset. The DOI strategy is independent from the WeatherGenerator Project

<Data set reference and name>	RADKLIM
Data set description	The RADKLIM dataset is a radar-based precipitation climatology developed by the German Weather Service (DWD). It provides precipitation estimates at a high spatial resolution of 1 km over Germany and is available in two forms – as hourly precipitation sums that have been adjusted with rain gauge measurements and a five-minute precipitation rates that are quasi-adjusted with the help of the hourly product. The dataset covers the time span from 2001 onwards with regular yearly updates In late spring. The current dataset version 2017.002 applies comprehensive quality control and correction algorithms to remove radar artefacts and to improve the representation of precipitation patterns, particularly for strong rainfall events and orographically influenced regions. RADKLIM is widely used in climate research for its consistency and suitability for long-term precipitation analysis in Germany. *Limitations?* Despite the comprehensive correction methods, the data still contains some artefacts due to orographic shading in some areas and may underestimate the amplitude of extreme precipitation events. Constraints? *No constraints.*
Standards and metadata	RADKLIM data adheres to established geospatial and climate data standards and is available as GIS-readable ASCII data, in Binary data and in NetCDF format. *Will you generate proper metadata for you data?* The data provision in netCDF format follows following CF (Climate and Forecast) Metadata Conventions. *Data format?* The data is currently stored in NetCDF format.

	Will there be a review process to quality- check the data? The dataset undergoes a comprehensive chain of quality checks. In addition to the operational RADOLAN procedure to calibrate and denoise radar measurements in real-time, several corrections methods to remove radar artefacts and to improve its calibration on surface observations are applied.
Data Sharing	RADKLIM data is openly available via the German Weather Service (DWD) Open Data Portal, with comprehensive documentation and support for reproducible climate research and long-term data accessibility. A copy of the dataset is in Juelich Meteocloud which is a joint data repository for meteorological reanalyses, model data, and satellite observations. Juelich Meteocloud is partially operated by the Juelich Supercomputing Center. *License?* The data is open source according to DWD Open Data policy. *Access URL?* DWD provide the data through their Open Data Server. Link to DWD download page for RADKLIM data
Archiving and preservation (including storage and backup)	The final archiving strategy will be finalised at a later stage. In the meanwhile, the data will continue to be stored at Juelich Meteocloud. *Is there an established workflow for your requested DOI process in place?* *The rain-gauge adjusted hourly precipitation sums have the following DOI:* 10.5676/DWD/RADKLIM_RW_V2017.002. The quasi-adjusted 5-min precipitation rates have the following DOI: 10.5676/DWD/RADKLIM_YW_V2017.002.

<Data set reference and name>	SEVIRI, Spinning Enhanced Visible and InfraRed Imager
Data set description	The Spinning Enhanced Visible and InfraRed Imager (SEVIRI) is an onboard sensor of the Meteosat Second Generation (MSG) satellites operated by EUMETSAT. SEVIRI provides high-frequency geostationary observations of the Earth's atmosphere, land, and ocean surfaces over Europe, Africa, and parts of the Atlantic. With a temporal resolution of 15 minutes and a spatial resolution of approximately 3 km (1 km for the high-resolution visible channel), SEVIRI delivers multi-spectral data across 12 channels. SEVIRI data are available from 2004 to present. For storage reasons we plan to use 10 years from 2015 to 2024. *Limitations?* Large number of files, data format not suitable for machine learning. See below for coping strategies. *Constraints?* No constraints

Standards and metadata	The data are provided in HDF5 or native MSG format, conforming to EUMETSAT metadata standards and WMO conventions, facilitating interoperability with common Earth observation processing tools. *Will you generate proper metadata for your data?* A STAC-based metadata catalogue to expose the metadata of all the datasets used by the WeatherGenerator is under development (see above). *Data format?* The data are available in MSG native format. The dataset will be converted into .zarr for better compatibility with the current machine learning standards *Will there be a review process to quality- check the data?* The data already undergo extensive quality-checks by EUMETSAT before release.
Data Sharing	SEVIRI data are accessible via EUMETSAT's data portals, ensuring long-term data accessibility. *License?* Open source under the EUMETSAT Meteosat >3hr latency & Metop Policy *Access URL?* *https://navigator.eumetsat.int/product/EO:EUM:DAT:MSG:HRSEVIRI*
Archiving and preservation (including storage and backup)	The final archiving strategy will be finalised at a later stage. For the moment the data are available on ECMWF's HPC facility. *At which Data Center do you want to store your data?* Juelich SuperComputing Center *Is there an established workflow for your requested DOI process in place?* *The data already have a DOI: 10.5676/EUM_SAF_CM/SARAH/V002_01*

<Data set reference and name>	SYNOP, Surface Synoptic Observation
Data set description	SYNOP (surface synoptic observation) data consist of standardized meteorological observations collected from land-based weather stations worldwide, typically at 6-hourly or hourly intervals. These observations include key atmospheric variables such as temperature, wind speed and direction, pressure, humidity, cloud cover, and precipitation. *Limitations?* No limitations *Constraints?* No constraints
Standards and metadata	SYNOP data follow the World Meteorological Organization (WMO) metadata code standards (FM-12) and are disseminated in structured formats such as BUFR or CSV, depending on the source. *Will you generate proper metadata for you data?* A STAC-based metadata catalogue to expose the metadata of all the datasets used by the WeatherGenerator is under development (see above). *Data format?* The data format depends on the source, BUFR or CSV are the most common. The dataset has been converted into .zarr for better compatibility with the current machine learning standards

	Will there be a review process to quality- check the data? The data already undergo extensive quality-checks before release
Data Sharing	The final storage repository will be available by the end of 2025 in a WeatherGenerator large data storage project at the Juelich Supercomputing Center. For the moment, the dataset is stored on ECMWF's ATOS. *License?* Mixed, depends on original source *Access URL?* Mixed, depends on original source
Archiving and preservation (including storage and backup)	The final archiving strategy will be finalised at a later stage. For the moment different copies of the dataset are available *at ECMWF's HPC facility.* *At which Data Center do you want to store your data?* Juelich Supercomputing Center *Is there an established workflow for your requested DOI process in place?* The dataset does not currently have a single, centralized DOI for the entire dataset. The DOI strategy is independent from the WeatherGenerator Project

<Data set reference and name>	Land Model dataset
Data set description	The dataset contains remote sensing observational data which are distributed freely by, for example, CDSE, MPC, EarthData as individual images. The data is reformatted to a data cube format for efficient access and storage according to our sampling design. The data cubes contain reflectance values from optical sensors of Sentinel-2 (ABC) and VIIRS (Suomi NPP), with various bands in the visible to near infrared wavelengths, and SAR microwave backscatter data of Sentinel-1 (ABC). Due to the scale of tens to hundreds of meters, we aim for a representative subset of the Earths landcover. The data cubes contain information relevant to earth surface changes with sensors sensitive to vegetation status, surface roughness and are therefore useful to a broad range of remote sensing applications. While similar data sets are published (DeepExtremes Project Nature data paper), our sampling approach, sensor combination and handling of the time dimension require custom solutions. With all data provided as openly available, publication of the generated cubes for re-use is possible. The process script is public on Github. Further data streams are planned and will be included for their specific downstream task (for example Land Surface Temperature by thermal IR sensors). *Limitations?* *Due to vast data sizes, a global dataset is not feasible with uncompressed data at the highest possible spatial resolution. Due to the unlabelled nature of EO data, connection to a certain use-case requires further reference data (GEDI, ECOSTRESS as high-value point samples, ALS for structural properties, weather station data for LST reference)* *Constraints? Data size requires subsampling, access is limited by some providers.*

Standards and metadata	*STAC metadata are becoming standard in EO data. For the data cubes we will adhere to this to leverage spatial and temporal as well as tabular requests to the datasets.* *Metadata will contain a list of sensors and bands, including sensor characteristics, spatio-temporal extent and process steps carried out during preprocessing (e.g. resampling)* *Data format?* *The Datasets will be formatted as per-sample zarr, limiting the number of files for the server by zipping. The format is likely to change with zarr 3.0 which supports sharding inside of chunks.* *Will there be a review process to quality- check the data?* *Planned are certain metadata checks within the processing to assure correct grid, datatype of the datasets. As the process aims only to reformat the data with no changes to the actual observations, visual inspection is only possible for small subset where outliers are apparent, most we rely on the providers quality check.*
Data Sharing	*The processed data cubes will be initially hosted on our premises as the land model is its prior use case. Due to the open data usage policy, sharing can be achieved initially by our own S3 buckets. No embargo.* *Tools for accessing or reproducing the datasets are fully open access and rely on python, xarray, and zarr as framework.* *A permanent data repository is TBD, which is only limited by the amount of data to be shared. As scripts for reproducing will be available, direct access might not be necessary.* *License? TBD* *Access URL? TBD*
Archiving and preservation (including storage and backup)	*The provided processing scripts will guarantee the reproducibility of our datasets.* *The actual data will be stored at the MPCDF during initial phase of the project. Further large-scale training of the Land Model are planned on requested compute, for which the data will be moved/reprocessed.* *DOI: TBD.*