Funded by the European Union under the
HORIZON-EUROHPC-JU-2023-INCO-06 programme.
Grant Agreement No. 101196247

D1.3 – Data Management Plan

WP1: Management & Dissemination

Under Review

Funded by
the European Union

Document Information

Deliverable Number	D1.3
Deliverable Title	Data Management Plan
Due Date	2025-07-31 (PM6)
Deliverable Lead	KTH
Authors	Rossen Apostolov, Cathrine Bergh
Keywords	Data management
WP	WP1
Nature	Report
Dissemination Level	Public
Final Version Date	2025-07-29
Reviewed by	UU, CINECA, BSC, UMA
MGT Board Approval	2025-07-30

Document History

Partner	Date	Comments	Version
KTH	2025-07-24	First draft	0.1
KTH	2025-07-29	Final draft	0.2

Executive Summary

This document describes the general data management principles and strategy for GANANA. Detailed data management plans are given for data that is expected to be produced as part of the software development and use cases. These include how the data will be stored and made findable and accessible, and preserved for future use beyond the lifespan of the current project.

Table of Contents

1 Introduction

GANANA is committed to open data following FAIR principles and aims to facilitate long-term data archival. This will be achieved with tools for metadata harvesting and deposition and, as a general policy that is applicable to the data produced as part of the GANANA software development and use cases, by storing relevant data in open repositories such as Zenodo. Where data size exceeds limits on repository size, in institutional repositories, possibly linked to global or federated ones such as OpenAire.

This document describes the Data Management Plan (DMP) of the GANANA software development and use cases, which constitute the primary activities that are expected to generate research data, and which are therefore clearly in scope of the Open Research Data Pilot. The data management plans are structured in a similar way to a DMP for research projects, with the caveat that the software and use cases are not projects per se but rather research activities aimed at showcasing GANANA capabilities in the context of real-world challenges. Work is expected to evolve in response to user and community needs as well as project-internal software and workflow developments and will be reviewed regularly.

The plan is based on the DMP template for the UK's Engineering and Physical Sciences Research Council (EPSRC), which was used previously in other projects such as the CoEs. The plan is formulated to contain enough information to ensure that data and metadata from the work undertaken will be Findable, Accessible, Interoperable, and Reusable (FAIR).

2 Data Management Plan – Template

The data management plan was formulated following the following template.

2.1 Data Collection

What data will be collected/created - distinguish between raw and processed/curated data? How much data is expected to be collected/created?

Are there common data types e.g. experimental data, modeling data, etc.?

How will data be collected/created?

2.2 Documentation and Metadata

What documentation and metadata will accompany the data?

2.3 Ethics and Legal Compliance

How will you manage any ethical issues?

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

2.4 Storage and Backup

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

How will you manage access and security?

2.5 Selection and Preservation

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

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

2.6 Data/Metadata Access and Sharing

How will you (or others) access the data?

What is the policy for sharing data incl. publication?

Are any restrictions on data sharing required?

2.7 Responsibilities and Resources

Who will be responsible for data management?

What resources will you require to deliver your plan?

3 Life Sciences

3.1 Description

Herein we describe the multiple streams of external experimental and modelling/simulations raw data that will be needed as input for the development of

GANANA software and use cases, and how resulting output data from simulations will be handled.

3.2 Data Collection

The use cases will involve executing many different applications, which will generate a broad variety of data in many different formats. Table 1 summarizes the types of data produced by the experiments, with the possible formats associated.

Whenever possible, the generated data types will be formatted using standards in the field, such as the PDB [1] or mmCIF [2] file format for 3D macromolecular structures, SDF/MOL2 for small molecules or XTC file format for MD trajectories (this latter is specifically related to GROMACS), and other databases such as SKEMPI [3] collecting binding affinity changes upon mutations. Binding affinity changes for antibody-antigen complexes are collected in the publicly available databases ATLAS [4] and AB-BIND [5].

[1] http://www.wwpdb.org/documentation/file-format
[2] http://mmcif.wwpdb.org
[3] https://life.bsc.es/pid/skempi2/
[4] http://atlas.wenglab.org/web/index.php
[5] https://github.com/sarahsirin/AB-Bind-Database

Table 1. Data types and formats produced by the different experiments in use case 3

Experiment	Type	Format
Structure modeling MD setup Protein ligand docking Trajectory clustering	Macromolecule 3D Structure	PDB mmCIF MOL MOL2 SDF GROMACS structure file GRO
MD simulation	Simulation trajectory	XTC
Modified Nucleic acids parametrization	Nucleic acids force-field parameters	HADDOCK topo and param files (text files)
Ligand parameterization Free Energy Perturbation	Small compound force-field parameters	GROMACS topology file ITP, TOP
Macromolecular flexibility	Time series plots	DAT XVG
Essential Dynamics (Principal Component Analysis - PCA)	Time series plots Compressed Trajectory	DAT XVG PCZ
Pathological Mutation Prediction	Protein SNV/SAV pathological report	ASCII Text CSV JSON
Classical Molecular Interaction Potentials	3D grid representing molecular properties	GRD CUBE

Data description:

- **Macromolecular 3D Structures:** Experimental or theoretical models of structural macromolecules obtained from public databases (PDB, DUD-e, DrugBank, Protein Model Portal, etc.) or from modeling, docking and simulation methods.

- **Simulation Trajectories**: Trajectories and metadata for molecular dynamics trajectories of macromolecules.

- **Modified Nucleic Acids Parameters:** Force-field parameters associated with modified nucleic acids for docking simulations in HADDOCK.

- **Ligand Parameters:** Force-field parameters associated to small compounds, needed to run MD simulations or FEP of ligands or protein-ligand complexes.

- **Macromolecular Flexibility Analysis:** Results of flexibility analysis done on macromolecular structures and simulation trajectories.

- **Essential Dynamics:** Results of principal component analysis done on simulation trajectories, extracting the correlated motions of proteins which are the most fundamental to the activity of the protein.

- **Pathological Mutation Prediction:** Results of a pathological mutation prediction from a trained classifier using sequence conservation and physico-chemical properties (PMut).

- **Classical Molecular Interaction Potentials:** Results of a classical molecular interaction potential (CMIP).

3.3 Documentation and Metadata

For macromolecular 3D structures, data will follow Protein Data Bank (PDB) recommendations and formats for documentation and metadata. Simulation metadata and molecule description for simulation trajectories will follow data ontologies as described in EDAM ontology and in [Ten simple rules on how to create open access and reproducible molecular simulations of biological systems]. Ligand parameters and macromolecular flexibility analysis (including essential dynamics) documentation and metadata will follow data and formats terms as described in EDAM ontology (Missing EDAM terms would be eventually proposed to EDAM maintainers). Pathological mutation predictions documentation and metadata will be included in the generated JSON files. Classical Molecular Interaction Potentials data will be accompanied by a documentation file with information about the content of the grid.

The HADDOCK3 workflows and parameters will be described in the associated TOML (ASCII format) and in the corresponding parts of scientific publications. For describing the HADDOCK server data and parameter settings, JSON files are used, which ideally should be provided with any publication resulting from the use of the HADDOCK server.

The simulation protocols and parameters used for alchemical free energy calculations and molecular dynamics simulations with GROMACS will be described in the corresponding parts of scientific publications. The protocols and simulation parameters together with versioning information for the software used will be deposited in a publicly accessible repository.

Trajectory data consists of the popular portable XDR format and can thus be accessed on multiple computing platforms. The simulations models, consisting of simulation parameters (timestep, methods for computing interactions, etc.), starting structures, topologies (i.e. the interactions) force field files and QM/MM models, are stored as ASCII text. The format of these files follows the GROMACS standard. In the case of GROMACS, all new algorithms will be coded in the C++ (.cpp) or Python (*.py) languages and made publicly available as part of the open- and distributed under LGPL.

Any other simulation data generated by this use case will include a README file in each sub-directory describing the content of the data.

3.4 Ethics and Legal Compliance

N/A

3.5 Storage and Backup

Data storage and backup will be provided by partners' local servers during the research. Long-term archive file systems and databases such as the Starlife infrastructure will be used. Access and security are coupled to the research center's user's credentials.

We will also make all input required to reproduce our trajectories and results, as well as source codes, publicly available at the consortium (https://github.com/GANANA-EU-INDIA) and/or code repositories: gitlab.com/gromacs, www.virtualchemistry.org, github.com/haddocking or github.com/GANANA-EU-INDIA.

Zenodo will be used for final data storage, under the umbrella of the GANANA community.

Data are not sensitive. No special measures are required beyond standard authenticated access procedures for local and remote systems.

3.6 Selection and Preservation

Setup and analysis scripts, together with the relevant, software-specific, generated data will be stored and shared, ideally through open repositories. The preserved data and metadata should allow third parties to repeat the computations with proper settings and compare obtained results with the deposited data to ensure reproducibility.

The long-term preservation plans are as follows:

- For HADDOCK: Datasets associated with a publication are usually deposited in the SBGrid data repository (https://data.sbgrid.org/labs/32/) or Zenodo, with an associated DOI.
- For GROMACS: Data will reside in long-term storage facilities at the PDC - Center for High Performance Computing at KTH. Datasets associated with a publication are deposited on Zenodo.
- We will also possibly make use of the Indian ICE-CLOUD infrastructure

Molecular dynamics trajectories used to extract relevant information for the use case will be stored and made available following the simple rules from [Ten simple rules on how to create open access and reproducible molecular simulations of biological systems], allowing reproduction of trajectories directly from the input configuration files.

Because we expect to generate in total well over a TB of molecular dynamics trajectories, it will be technically difficult to share these trajectories directly with others. Instead, we will make all that is needed to reproduce these trajectories on any high-performance computing system publicly available, including starting structures, force fields parameters, simulation settings and source codes.

3.7 Data/Metadata Access and Sharing

Data and metadata will be made available through open repositories such as Zenodo with proper FAIR identifier or DOI. GANANA has its own community on Zenodo (https://zenodo.org/communities/ganana). Local database identifiers will be communicated through the identifiers.org service. Where data size exceeds limits on repository size, in institutional repositories, possibly linked to global or federated ones such as OpenAire.

For any cases where the raw data produced by the molecular dynamics simulations exceeds the limit imposed by Zenodo or institutional repositories, raw data will be available upon request from the relevant use case partner.

No restrictions on data sharing are required unless proprietary industry data is used, in which case access will be limited to project partners and will not be made public.

The trajectories will be analyzed and the result of these analyses, as well as modelling data, will be published in scientific journals. In cases where high amounts of data are generated, we will instead provide users with the input files needed to reproduce the results. These will be stored in a publicly accessible repository such as Zenodo. Publications will include a link to the repository where the data is stored.

3.8 Responsibilities and Resources

Responsible partners for data management are:

HADDOCK: Alexandre Bonvin (UU)
GROMACS: Alessandra Villa (KTH)

No resources are expected to be needed beyond those already available.

4 Natural Hazards

4.1 Data Collection

The use cases related to the Natural Hazards work package involve many different applications and therefore utilize and produce a diversity of data and formats. The different data requirements of the applications will be described in the following sections and are also summarized in Table 2.

UCIS4EQ (Urgent Computing Integrated Services for Earthquakes) is a high-performance computing (HPC)-based system designed to provide rapid, automated, physics-based assessments of seismic events. Its primary goal is to support emergency response efforts by delivering fast and accurate estimates of earthquake ground motion impacts.

UCIS4EQ utilizes:

- Large-scale 3D full waveform simulations, driven by HPC.
- Automated seismic workflows that quickly assess earthquake impacts.
- Massively parallel solvers, enabling results within minutes to hours.
- Containerized microservices, enhancing cloud deployment and interoperability.

- Synthetic shaking map generation, supporting seismic digital twins.
- Uncertainty modeling, accounting for unknowns like source parameters and local soil effects.

The input data requirements of UCIS4EQ include seismic source data (such as historical earthquakes or synthetic events), 3D and 1D velocity models (including parameters such as Vp, Vs, density, Qp, Qs), topographic and soil data (Vs30 layers and topo-bathymetry), receiver locations (seismic stations where motion is recorded), and observed ground motion data (for validation). Output data include a simulation catalog where full waveform simulations and synthetic maps are stored in a database, shaking maps from urgent or non-urgent scenarios, and time series data of ground motion metrics. The data volume is primarily dependent on the number of scenarios, spatial/temporal resolution, and the number and type of variables stored, and is expected to be in the terabyte range, especially for high-resolution 3D simulations. Data will primarily be collected through:

- Open-access datasets (e.g. global velocity models, seismic catalogs) collected by partners (e.g. NCS and potentially shared with Project 3).
- Restricted or high-resolution regional data sourced through partner institutions or national agencies.
- Modeling data created using the UCIS4EQ workflow on HPC resources, based on physics-based solvers.

Next, SAIPy is an open-source Python package designed for fast, automated processing of seismic waveform data using deep learning for single-station earthquake monitoring, offering a robust suite of deep learning solutions for various seismological tasks. It provides comprehensive capabilities for (i) earthquake signal detection, (ii) seismic phase picking, (iii) first motion polarity identification, and (iv) magnitude estimation. SAIPy introduces the first fully automated deep learning-based pipeline for processing continuous waveforms, streamlining the entire monitoring workflow from event detection to the estimation of P- and S-arrival times, magnitudes, and first-motion polarity. This makes SAIPy a powerful tool for real-time earthquake monitoring, enabling timely actions to mitigate seismic impacts.

The earthquake monitoring use cases involve continuous and triggered seismic recordings across multiple stations, resulting in various types of raw and processed seismic data. Seismic waveform data is obtained from both single and multiple earthquake stations, either as raw counts (preferred) or acceleration with sampling rates up to 100 Hz and pre-processed with bandpass filtering (1-45 Hz). The event catalogue is in CSV format and includes earthquake magnitude, P- and S-wave arrival times, first-motion polarity, and fault type (if available). After successful training of the models, high-quality waveform segments of earthquakes and earthquake catalogs will be generated. The data volume depends on the number of recorded earthquakes and the number of seismic stations reporting those events, as provided by nodal agencies. The data is

expected to be in the Gigabyte range. Common formats include observational recorded waveform data (e.g. MiniSEED or ASCII) and corresponding earthquake information (e.g. CSV). Data will primarily be collected through:

- Open-access datasets such as seismic catalogs with magnitude of the earthquake and origin time.
- Waveforms data of the earthquakes are planned to be sourced through partner national institutions from India.
- Modeling data created using the UCIS4EQ workflow on HPC resources, based on physics-based solvers will be obtained from partners (BSC).
- Recorded continuous data will be analyzed using the SAIPy framework using custom scripts.

Further, Tsunami-HySEA (Hyperbolic Systems and Efficient Algorithms) is a numerical code developed by the EDANYA Group of the University of Málaga for simulating the generation, propagation, and inundation of tsunamis. It is designed to be both accurate and computationally efficient, making it suitable for early warning systems, hazard assessment, and research applications. Tsunami-HySEA is part of the HySEA family of codes.

This use case requires input data in the form of bathymetric data for propagation simulations, finer-resolution coastal bathymetry and topography for inundation simulations, seismic source data from historical and synthetic events for simulation catalogue creation, and observed data for verification and benchmarking. The resulting output data will be in the form of a simulation database with propagation and eventually high-resolution inundation, and time series data for selected locations. The volume of data will be highly dependent on the number of scenarios and variability produced, number of variables stored, and the temporal-spatial resolution of the simulations.

Finally, the wildfire spread forecasting system code is based on the coupling between a forest-fire model (WRF-SFire) and a pollutant dispersion model (FALL3D) and will provide the following products:

- Fire area (modelling data)
  Gridded data from WRF-SFire model with spatial resolution of ~40 m and temporal frequency of 1h over the typical scales involved in areas affected by forest fires
  Formats: netCDF and GeoTIFF
- Emission factors
  Gridded data from WRF-SFire model with spatial resolution of ~40 m and temporal frequency of 1h over the typical scales involved in areas affected by forest fires
  Formats: netCDF and GeoTIFF
- Concentration of Particulate Matter (PM), such as PM2.5 or PM10, and black carbon.

Gridded data from FALL3D model with spatial resolution of ~1-10 km and temporal frequency of 1h over continental scales (~1000 km)
Formats: netCDF, GeoTIFF, Cloud Optimized GeoTIFF (COG)

- Concentration of relevant gaseous pollutants released during forest fires, including carbon monoxide (CO) and nitrogen oxides (NOx), such as NO2
  Gridded data from FALL3D model with spatial resolution of ~1-10 km and temporal frequency of 1h over continental scales (~1000 km)
  Formats: netCDF, GeoTIFF, Cloud Optimized GeoTIFF (COG)

4.2 Documentation and Metadata

The earthquake and tsunami output data will be accompanied by both metadata and documentation to support reproducibility, version-tracking, and cross-system interoperability. Simulation metadata will include experiment description, initialization definition (e.g. Okada parameters), earthquake source parameters, veolcity model specifications, simulation configuration files (e.g. json and 1D files), and software version and execution environment. Data formats will follow standardized scientific conventions such as NetCDF and json. All generated datasets will be properly documented with a clear README file stored in each data repository, describing all the steps taken to generate the data.

Outputs from the wildfire/smoke dispersion system will include metadata and associated documentation. Those products stored in netCDF format will include both global and variable attributes holding metadata about data and files. GeoTIFF files will be collected within a STAC catalog, short for SpatioTemporal Asset Catalog, allowing a standardized way to describe and organize geospatial data. Documentation about the simulations leading to these products will be reported in multiple log files with different running information, such as computing time, errors and warnings, and simulation parameters will be gathered in json files which are automatically generated by the models in every run.

4.3 Ethics and Legal Compliance

Ethical risks are minimal as no personal data is used, but the consortium will be in compliance with the General Data Protection Regulation (GDPR: Regulation (EU) 2016/679). Care will be taken to ensure responsible communication of simulated risk scenarios, especially in public or policy settings, to avoid misinterpretation.

Intellectual Property Rights (IPR) issues will be handled as follows:

- Open-access data will be clearly cited with appropriate attribution.
- Restricted data will be used under agreements with proper licenses and acknowledgments.

- All codes developed within this work package will follow their respective IPR and licensing policies.
- Data generated may be shared under open licenses (e.g., CC-BY), depending on project guidelines.
- All the codes, results and documentation will be released under open-source license.

4.4 Storage and Backup

Data from the earthquake and tsunami use cases will be backed up and maintained by each partner with local infrastructure. Additionally, centralized or distributed storage might be used depending on simulation volume and sensitivity. Output data from the wildfire/smoke dispersion system will be kept in separate storage units from local HPC clusters at C-DAC and CSIC in order to ensure storage mirroring and backup data. Some products for visualization will be stored in remote repositories and cloud storage locations as well to make them accessible to users through a consistent API. Access will be regulated by keeping role-based permissions, secure user authentication protocols, and ensuring project-wide access policies. Sensitive or unpublished data will remain internal until released through official channels.

4.5 Selection and Preservation

Data with long-term value suitable for preservation include synthetic shaking maps, high-quality processed waveform segments, validated simulations, and time series from benchmark scenarios. Additionally, simulation catalog outputs that contribute to digital twin repositories or decision support tools will be kept as well. These data will be archived in institutional repositories or cloud-based storage. Published data will be made available in open research data platforms (e.g. Zenodo, SDL, EUDAT) using standard formats for long-term accessibility and interoperability.

4.6 Data/Metadata Access and Sharing

Data for the earthquake and tsunami use cases will be kept at local project infrastructures such as HPC systems or cloud-based solutions, public portals, and open-access repositories (e.g. B2Drop). Data will be openly shared after it has been validated and internally approved, and when there is proper documentation and licensing. High-value outputs (e.g. shaking maps for past events) might be published in academic papers or released through public platforms. Data will be restricted that is based on restricted or proprietary inputs (e.g. local velocity models) or has not yet been validated or peer-reviewed. Sharing will follow legal agreements and data ownership protocols.

Regarding the information from the wildfire/smoke dispersion system, data stored in HPC clusters and local storage units will be accessible to external users under request according to the data access policy requirements defined by C-DAC and CSIC institutes. In addition, part of the data will be made public as spatiotemporal assets within a STAC catalog enabling users to browse, search, and retrieve data from cloud storage locations through a consistent API.

4.7 Responsibilities and Resources

Each project partner will be responsible for the data they collect or generate, but the work package leader will serve as a data management coordinator to oversee consistency and ensure metadata standards and sharing policies are followed.

The following resources will be needed:

- HPC resources to run simulations and train AI models.
- Data storage infrastructure (local and/or cloud-based).
- Personnel for training the AI models, maintaining metadata and backups, and coordinating access, sharing, and compliance with policies.

4.8 Summary

All data management aspects discussed in this section are summarized for the four different codes in the table below.

Table 2. Summary of data management strategies for work package 4.

Codes/data management aspect	**UCIS4EQ**	**SAIPy**	**Tsunami-HySEA**	**FALL3D**
Purpose of code	Rapid, physics-based seismic impact assessment using HPC	Real-time earthquake monitoring using deep learning	Tsunami simulation: generation, propagation, inundation	Forecasting fire spread & pollutant dispersion using coupled models
Raw data collection	Seismic sources, velocity models, topography, soil	Continuous waveform data, event metadata	Bathymetry, topography, seismic sources,	WRF-SFire & FALL3D outputs: fire area, emissions, PM,

	data, observed motions		historical/synthetic event data	gaseous pollutants
Processed data creation	Full 3D simulations, synthetic shaking maps, time series	Processed waveform segments, final earthquake catalogs	Propagation and inundation simulation databases, time series	Gridded outputs (NetCDF, GeoTIFF, COG) for fire/pollutant parameters
Data volume	Terabytes (due to 3D HPC simulations)	Gigabytes	Variable; potentially large depending on scenario complexity	High-resolution data over large scales; likely terabytes
Data types	Modeling and observational	Observational waveform + event metadata	Modeling data	Modeling data
Data sources	Open datasets, national institutions, HPC-generated simulations	Open catalogs, national seismic networks	UMA open data, restricted high-res data, generated by Tsunami-HySEA	C-DAC & CSIC simulations
Metadata & documentation	Simulation configs, versioning, JSON/NetCDF formats	README, event metadata (origin, magnitude, polarity)	Experiment details, configuration files, software versioning	Metadata in NetCDF/GeoTIFF/STAC, simulation logs, JSON config files
Ethics & legal aspects	GDPR-compliant, careful with public interpretation, data under proper licenses	GDPR-compliant, open licenses where applicable	GDPR-compliant, responsible risk communication	Same as consortium (GDPR, IPR-respecting)

Storage & backup	Distributed/local backups, centralized storage depending on data volume	Local backups, possible centralized storage	Local HPC sites; central strategy	Local + cloud storage, mirrored across sites (e.g. C-DAC & CSIC)
Access & security	Double authentication, staged data releases	Double authentication, staged data releases	Double authentication, staged data releases	API access for public data (STAC), internal access via policy
Long-term data	Validated simulations, shaking maps, benchmark time series	Earthquake catalogs, processed segments, SAIPy configs	Digital twin data, inundation results	Fire/pollutant gridded data products
Preservation plan	Institutional/cloud repositories (Zenodo, SDL), open formats	Institutional/cloud repositories (Zenodo, SDL), open formats	Institutional/cloud repositories (Zenodo, SDL)	Public STAC catalog + internal HPC backups
Data sharing policy	Public after validation	Open after review; follows data ownership/licensing protocols	Public after validation	Partly public via API/STAC; access request policy for rest
Data restrictions	Yes – for restricted models, unvalidated results	Yes – for restricted network data	Yes – for restricted models, unvalidated results	Yes – governed by C-DAC/CSIC data policies
Data management responsibilities	Each partner manages their own data; DMP coordinator ensures consistency	Each partner manages their own data; DMP coordinator ensures consistency	Each partner manages their own data; DMP coordinator ensures consistency	Each partner manages their own data; DMP coordinator ensures consistency

Resources needed	HPC, cloud storage, personnel for simulation, metadata, compliance	HPC for AI training, storage, staff for training, backups, metadata	HPC, cloud storage, personnel for simulation, metadata, compliance	HPC, local + cloud storage, STAC catalog infrastructure

5 Weather & Climate

This work package being mostly focused on model development, deployment, performance analysis and model optimisation, the management of the numerical data itself is very limited. We mostly document here the software repository used that will ensure part of the FAIRness of the data produced.

5.1 Data Collection

The only external data used by the models are the initial conditions. For EC-Earth (Task 5.7), they come from ERA5 and ORAS5 and will be stored in a C-DAC supercomputer. For ICON-CLM (Task 5.10), the data come from ERA5 for the initial evaluation of the setup and from a GCM for the final climate projections. The selection of the most suitable GCM to be downscaled over the Indian domain has still to be performed. The data will be stored on a C-DAC supercomputer.

For the profiling analysis (Task 5.2), temporary data (traces) will be generated in the HPC where the simulations are run, but they will not be preserved after the writing of the deliverables.

For Task 5.7 “Creation of climate use-cases”, the data generated by the EC-Earth model will not be kept after the execution of the model, as the purpose of the task is only to demonstrate the viability of the deployment.

The Task 5.5 “Autosubmit delployment on a C-DAC supercomputer”, will include by default automatic provenance tracking mechanism (more or less advanced depending of the version of Autosubmit installed) that will also ensure the FAIRness (especially the Reusability) of the data generated and the workflow itself.

The numerical data generated by Task 5.10 will be publicly available in NetCDF format through the Data Delivery System (DDS) at CMCC (https://dds.cmcc.it). The CMCC DDS provides a unique, consistent and seamless access point for all data produced and used by CMCC through a unified API interface. The user can browse the Catalog and the

available datasets through the DDS Web Portal and access and download data through the DDS API Python client.

The performance data generated by task 5.13 will be published and if possible, will be also published on the HPCW website and part of the planned ranking.

5.2 Documentation and Metadata

The documentation of the model developments, deployments, performance analyses and model optimisations will be done in the deliverables associated to each task.

As there are no numerical data produced for most of the tasks of this work package, the metadata of the output is not relevant here but, by default, the data generated by the models are following international conventions of format and metadata (NetCDF, CMOR), that would ensure their FAIRness in the case they should be disseminated later on.

The data produced by ICON-CLM are in NetCDF format and can also be CMORized, ensuring FAIRness compliance for future dissemination.

5.3 Ethics and Legal Compliance

No ethics and legal compliance have been identified for the weather and climate work package.

5.4 Storage and Backup

The data generated during GANANA for Projects 1, 2 and 4 of WP5, even if temporary, will be generated in the dedicated partition (usually scratch) of each supercomputer (C-DAC platforms and EuroHPC platforms). The numerical data themselves not being the main outcome of the project, no backup strategy has been put in place. However, the models and tools themselves are made permanently available in their own version-controlled repositories, ensuring that they can be easily redeployed or redownloaded.

The list of repositories of models and tools used in the Projects 1, 2 and 4 is shown below:

- Extrae (BSC tools): https://github.com/bsc-performance-tools/extrae
- EC-Earth4: https://git.smhi.se/ec-earth/ecearth4
- Autosubmit: https://github.com/BSC-ES/autosubmit
- HPCW: https://gitlab.dkrz.de/hpcw/hpcw

The dataset generated at the end of the GANANA project for Project 3 of WP5 will be stored at CMCC and made available through the DDS service. The backup policy will be the one currently implemented at the CMCC supercomputing center.

The list of repositories of models and tools used in the Project 3 is shown below:

- ICON-CLM: https://gitlab.dkrz.de/icon/icon-model
- SPICE (workflow to manage ICON-CLM runs): https://zenodo.org/records/6838984

5.5 Selection and Preservation

This section does not apply.

5.6 Data/Metadata Access and Sharing

The data and code used during the project are only made accessible to the partners of the project and the Indian partners as well, there are no plans to share them outside when not already done (for example for open-source codes).

5.7 Responsibilities and Resources

Due to the very small data management part of the project, no additional resources to those already mentioned have been assigned to this task. The software management is handled in the other work packages, with associated human resources. However, the work package leader of this project will be ultimately responsible for any data management matters within the work package.