CopERnIcus climate change Service Evolution

D8.4 Data Management Plan

Due date of deliverable	June 2023
Submission date	June 2023
File Name	CERISE-D8.4-V1.0
Work Package /Task	WP8/ T8.4
Organisation Responsible of Deliverable	ECMWF
Author name(s)	Rhona Phipps, Tanya Warnaars, Patricia de Rosnay and CERISE project partners
Revision number	1.0
Status	Issued
Dissemination Level	Public

Funded by the European Union

The CERISE project (grant agreement No 101082139) 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 CERISE Data Management Plan 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 satellite and in-situ observations, reanalysis prototypes, and seasonal forecast demonstrators. The data of the project will comply with the FAIR data principles, adhering to the principle 'as open as possible and as closed as necessary'*.

The data will be archived on the ECMWF MARS system and be made accessible using existing data portals (e.g. Zenodo).

This document is a living document which will be developed during the lifetime of the project to follow and share the developments of the CERISE project.

* 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

The following provides the plans for how the project will set up, administer and archive the legacy of data arising from CERISE. This deliverable aims at supporting partners' in their efforts and responsibilities in making project data that is FAIR (Findable, Accessible, Interoperable, Reusable) and 'as open as possible, as closed as necessary'. It will also ensure consistency across the project.

This deliverable is primarily targeted at the consortium partners and should serve as a reference for the management of data products in the relevant deliverables. It also serves to support the cross-cutting activity on data integration and data products, which will interact with all WPs throughout the duration of the project to maximize benefits of the data generated by CERISE.

This CERISE data management plan describes the data management life cycle for all datasets to be collected, processed and generated in the project. It constitutes the first version of the DMP and provides the baseline of the policy that will be followed by the CERISE consortium with respect to the data management related activities. More specifically, it covers the following activities:

- What types of data will be collected and/or generated?
- What standards will be used?
- How will this data be exploited, shared, processed and made accessible?
- How will this data be curated, stored and preserved?
- Which tools and methodologies will be used to store this data and for how long?
- How are data restriction levels managed?

This DMP outlines how research data will be handled throughout the life cycle of the project.

2.1 Background

The scope of CERISE is to enhance the quality of the Copernicus Climate Change Service (C3S) reanalysis and seasonal forecast portfolio, with a focus on land-atmosphere coupling.

It will support the evolution of C3S, over the project's 4 year timescale and beyond, by improving the C3S climate reanalysis and the seasonal prediction systems and products towards enhanced integrity and coherence of the C3S Earth system Essential Climate Variables.

CERISE will develop new and innovative ensemble-based coupled land-atmosphere data assimilation approaches and land surface initialisation techniques to pave the way for the next generations of the C3S reanalysis and seasonal prediction systems.

These developments will be combined with innovative work on observation operator developments integrating Artificial Intelligence (AI) to ensure optimal data fusion fully integrated in coupled assimilation systems. They will drastically enhance the exploitation of past, current, and future Earth system observations over land surfaces, including from the Copernicus Sentinels and from the European Space Agency (ESA) Earth Explorer missions, moving towards an all-sky and all-surface approach. For example, land observations can simultaneously improve the representation and prediction of land and atmosphere and provide additional benefits through the coupling feedback mechanisms. Using an ensemble-based approach will improve uncertainty estimates over land and lowest atmospheric levels.

By improving coupled land-atmosphere assimilation methods, land surface evolution, and satellite data exploitation, Research and Innovation inputs from CERISE will improve the

representation of long-term trends and regional extremes in the C3S reanalysis and seasonal prediction systems.

In addition, CERISE will provide the proof of concept to demonstrate the feasibility of the integration of the developed approaches in the core C3S (operational Service), with the delivery of reanalysis prototype datasets (demonstrated in pre-operational environment), and seasonal prediction demonstrator datasets (demonstrated in relevant environment).

CERISE will improve the quality and consistency of the C3S reanalysis systems and of the components of the seasonal prediction multi-system, directly addressing the evolving user needs for improved and more consistent C3S Earth system products.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverables

This D8.4 Data Management Plan provides the initial outline of the data management plan including information on which data sets will be created in the project and how they will be made available. It will also include information on the datasets that will be used in the project.

This document represents only the initial version where details may not be available yet, and it will be further developed over the course of the project.

2.2.2 Work performed in this deliverable

In this deliverable the work outlined in The Description of Action [RD1] (WP8 T8.4) is performed. An initial assessment of the data used and produced is collected from WP leaders via a questionnaire. The draft questionnaire is Annex 1, the feedback from the each of the WP leaders is presented in Annex 2.

2.2.3 Deviations and counter measures

No deviations have been encountered.

2.2.4 Reference Documents

[1] Project 101082139- CERISE-HORIZON-CL4-2021-SPACE-01 Grant Agreement

[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

2.2.1 CERISE Project Partners:

ECMWF	European Centre for Medium-Range Weather Forecasts
Met Norway	Norwegian Meteorological Institute
SMHI	Swedish Meteorological and Hydrological Institute
MF	Météo-France
DWD	Deutscher Wetterdienst
CMCC	Euro-Mediterranean Center on Climate Change
BSC	Barcelona Supercomputing Center

DMI	Danish Meteorological Institute
Estellus	Estellus
IPMA	Portuguese Institute for Sea and Atmosphere
NILU	Norwegian Institute for Air Research
MetO	Met Office

3 Data Summary

Our Data Management Plan (DMP) is developed following the standard approach to the European Monitoring and Evaluation Programme (EMP) 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 8.4 (WP8) is responsible for this activity.

The products of CERISE will comprise reports, graphical displays, datasets and improved methods, algorithms and code. All these elements have their own important role. Graphical displays, where applicable, are targeted at all users as supportive information for the various model runs, method comparisons, and input datasets. The datasets will also target a wide user community to support them with parallel or alternative studies. Improved methods, algorithms and code are meant to form the basis for follow-on development after the CERISE project has finished. Some data and prototypes developed will be archived at ECMWF and be available on request.

Data products arising from the project:

- Global and regional scale land and coupled land-atmosphere reanalysis pre-prototypes and prototype datasets
- Seasonal forecasts demonstrators with balanced land-atmosphere in initial conditions
- Time varying datasets of lake cover, LAI and land cover back to 1925.

The CERISE developments will enhance the relevance of the C3S products to European public institutions, policymakers, industry, researchers, media and the general public. They will also support the evolution of C3S towards additional sectors, such as arctic regions and Strategic Development Goal (SDG) indicators monitoring. A key driver for developing coupled assimilation in CERISE is to fill the gap in the exploitation of current and future satellite observations, especially those coming from the Sentinels missions. CERISE's main scope is to improve the C3S product quality and consistency based on coupled assimilation and initialisation, with methodological improvement, demonstration of the proof of concept and its validation.

CERISE will develop and produce data to support C3S services:

- ➔ it will develop and produce prototype datasets for future C3S operational global and regional climate reanalysis services, with a focus on improved consistency and long-term trends of land-atmosphere Essential climate Variables (ECV's)

- ➔ CERISE will also develop and produce seasonal forecast demonstrator datasets designed to test options for land surface initialisation in the operational systems participating in the C3S seasonal multi-system

- ➔ CERISE will use satellite and in-situ observations. CERISE will prepare for the usage of new satellite retrieval observations, e.g. from the Metop-SG SCA instrument and in the longer term from the Copernicus Sentinels, e.g. CIMR. As a result, CERISE will provide valuable feedback to space agencies and data providers enabling them to further improve the exploitation of satellite data from the Copernicus Sentinels and the EPS-SG programmes.

CERISE's methodologies, prototype and demonstrator datasets, and diagnostics tools will be developed in strong collaboration with the existing C3S operational and development teams and will use the C3S infrastructure. These strong links between the CERISE consortium and the Copernicus operational Service will ensure efficient research to operation transition of the CERISE developments, with staged integration of the CERISE developments in C3S during the project and beyond during the current and future Copernicus phases.
The integration of the CERISE developments in the core Service will directly enhance the C3S impacts on society. The output from CERISE once implemented in C3S, will directly be available to a wide range of users and stakeholders, including WMO, IPCC and research centres.

A further positive impact can be expected on Numerical Weather Prediction (NWP) applications as the new couped land-atmosphere assimilation and ML/AI-based novel observation operator developments will be part of the ECMWF Integrated Forecasting System (IFS). They will thus contribute to improving ECMWF NWP performances, through enhanced consistency at land-atmosphere interface and improved satellite data usage over land surface, contributing to the ECMWF all-sky and all-surface strategy and benefitting EU member states.
The land initialisation activities will facilitate the uptake of C3S land reanalysis products by the seasonal forecasting centres and the wider community. It is expected that the improvement will translate into better quality of the climate information from reanalyses and more reliable forecasts from days to seasons ahead.
New forecast products of land-related quantity (soil, snowpack) are expected to benefit C3S sectoral applications.

3.1 Definitions related to the approach to Open Science:

The Horizon Europe programme guide states[1]: "*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 CERISE 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 CERISE 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

[1] 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

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 satellite and in-situ[2] observations, prior emissions, and results from inversion studies.

CERISE has a strong link to the Copernicus Climate Change Service. The close collaboration will ensure that the CERISE activities are complementary to what is done elsewhere in other projects/ initiatives e.g. the EU project CONFESS and CoCO2.

4 FAIR Data

The data of the project will comply with the FAIR data principles, as much as possible.

The data will be archived on the ECMWF MARS system and will be made accessible using existing data portals (e.g. Zenodo).

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 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.

This DMP follows the EU guidelines[2] and describes the data management procedures according to the FAIR principles[3]. 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. Metadata refers to information about 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.

Datasets that will be uploaded to open access repositories will be deposited in a searchable resource and listed on our project website. The naming conventions for the project's data files can significantly increase their searchability. Towards this, CERISE will design consistent data file names that properly describe their content, status and versioning, with a view on increasing their discoverability.

During the course of the project, and at least at the moment of publication of the project results, each research team will deposit and describe the relative underlying data sets.

[2] 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

[3] The FAIR data principles (GO FAIR), https://www.go-fair.org/fair-principles

Trusted data repositories can attribute persistent unique identifiers (PIDs) to the deposited items (e.g. Zenodo).

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 for the Horizon Europe, which states that all data generated and used, if possible, are publicly open and available. The CERISE 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 and will maintain a list of data sets it accesses for the purposes of CERISE activities on the project website. The accessibility of the data will be ensured at two levels: internally to the project, and to the general public. The strong connection to the C3S community strengthens the use and accessibility of CERISE outputs.

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, CERISE 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, trusted repositories will be chosen such as Zenodo. The public project data sets will be visible via the OpenAIRE portal, facilitating project reporting procedures. Data deposition in repositories will guarantee long time preservation and accessibility to datasets.

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 some 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.

The CERISE DMP Annex 2 provides the specific information indicating the versions or parts of the data sets that can(not) be freely shared. The repositories for data set publication and preservation may be further defined during the project.

4.3 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, CERISE will incorporate suitable standards and vocabularies for data and metadata creation. However, in the case of CERISE, the primary end user of the data is the C3S community. To this end the level of integration to those existing services is a driver for the project as CERISE products need to be interoperable with the applications and workflows of the C3SS services.

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.

4.4 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, CERISE adopts a data quality assurance procedure. Peer-reviews of the data generation methods and/or data summaries are inherent in the work of the project and will be applied to assess the quality of the dataset and identify any need for improvement.

5 Other research outputs

Other research data will be stored and backed up regularly through existing back-up mechanisms in place at Sharepoint and the internal Confluence pages. This is particularly relevant to project documents, reports, internal data sharing between consortium partners and web content.

6 Allocation of resources

The resources required for making the data generated by CERISE FAIR have been included in the budget of the project. In general, the CERISE consortium as a whole will decide and contribute to relevant aspects of the data management cycle during and after the completion of this project. The research team leaders responsible for each dataset will be added in the future release of the DMP.

At this state, the chosen repository for long term deposit 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 CERISE budget.

7 Data security

The CERISE consortia place a strong emphasis on ensuring the security of all the produced datasets, safeguarding them from unauthorized access and loss. All the information will be stored in a private and secure storage area. The data will be backed up on a regular basis and access will be restricted only to the members of the consortium.

In case of personal data collections, it is crucial that this data can only be accessible by those authorized to do so.

To make the data publicly accessible in dedicated public repositories or storage environments, we will investigate in depth options such as Zenodo.

For what concerns ECMWF, a robust and rigorous data security system is available, including backups. The physical security includes 24/7 monitoring, fire suppress and power backup systems.

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 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 CERISE. Additionally, the Grant Agreement and the CERISE 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 when gathering the personal information.

9 Conclusion

In this deliverable, the CERISE Data Management Plan has been initiated.

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

Annex 1:

Annex I includes the template used to collect the information from WP leaders regarding data to be used or produced. The completed tables are in Annex 2

WP leaders to complete the list of the datasets, already available or to be developed in the context of the project's research and implementation activities. The list is defined for each workpackage of CERISE. The table below shows each data set that:

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

Workpackage 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?* *Usage 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 are you aiming to store your data?* *Is there an established workflow for your requested DOI process in place?* *According to which standards*

Annex 2:

Annex 2 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 list is defined for each workpackage of CERISE. 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:

Completed by: Pete Weston with input from WP partners

Dataset	**Land data assimilation methodology experiments**
Data set description	Generated data: NWP experiments within the IFS with different methodologies applied to the land-surface data assimilation. Time span: sampling a few months for summer and winter seasons for a recent year post 2015. Horizontal resolution: TCo399 (25km), global domain The data underpins scientific publications related to land data assimilation methodology developments.
Standards and metadata	Data is produced in GRIB-1 and GRIB-2 format and is stored in the ECMWF MARS archive. It can also be downloaded in NetCDF4 format. The dataset will be reviewed internally in WP1 and by WP6.
Data Sharing	Data will be available via the ECMWF MARS archive and CERISE consortium members can access this using their ECMWF member-state account.
Archiving and preservation (including storage and backup)	The data in the MARS tape library are backed up.

Dataset	**Training database used to generate the ML-based observation operator**
Data set description	Generated data: Low-frequency passive microwave observations (1.4GHz-36GHz from SMOS, SMAP, AMSR2 & SSMIS) and collocated NWP model variables. Time span: 15 months from July 2020 to September 2021. Horizontal resolution: TL511 (40km), global domain The data underpins scientific publications related to ML-based observation operator developments.

Standards and metadata	Data is produced in NetCDF4 format and is stored in the ECMWF ECFS archive. The dataset will be reviewed internally in WP1
Data Sharing	Data will be available via the ECMWF CERISE WP1 FTP site (bolftp.ecmwf.int)
Archiving and preservation (including storage and backup)	The data in the ECFS archive are backed up.

Work Package 2:

Completed by: Patricia de Rosnay with input from WP partners

Dataset	**Coupled reanalysis experiments**
Data set description	Generated data: NWP experiments with the IFS for different configurations of land-atmosphere data assimilation coupling. Time span: sampling a few months for summer and winter seasons for a recent year post 2015. Horizontal resolution TCo399 (25km), global domain The data underpins scientific publications related to coupled land-atmosphere data assimilation developments.
Standards and metadata	Data is produced in GRIB-1 and GRIB-2 format and is stored in the ECMWF MARS archive. The dataset will be reviewed internally in WP2 and by WP6.
Data Sharing	Data will be available via the ECMWF MARS archive and CERISE consortium members can access this using their ECMWF member-state account.
Archiving and preservation (including storage and backup)	The data in the MARS tape library are backed up.

Work Package 3:

Completed by: Jean-Christophe and Tim Stockdale with input from WP partners

Dataset	**Set of land surface initial conditions for 1993-2022 for use in seasonal forecast demonstrators**
Data set description	State land variables (e.g. soil moisture, LAI, snow fractional cover) at a global scale, produced in different service provider environments (CMCC, DWD, Meteo-France) ; open-loop and analysis.
Standards and metadata	Metadata in Netcdf header. Data format: Netcdf.
Data Sharing	Zenodo (TBC)

Archiving and preservation (including storage and backup)	Zenodo (TBC)

Work Package 4:

Completed by: Hans Hersbach, Harald Schyberg with input from WP partners

Dataset	**Reanalysis global demonstrators**
Data set description	Two global land reanalysis prototypes from 1939 to 2022, at TCo319 (about 35 km) horizontal grid for several parameters at the surface, soil-moisture and snow levels. Two prototype reanalyses with full two-way coupling of ocean, sea ice, upper-air and land for three one-yearly periods each. Spatial resolution is TBD.
Standards and metadata	Data is produced In GRIB-1 and GRIB-2 format and is stored in the ECMWF MARS archive. These dataset will be reviewed by WP6 (Evaluation of results).
Data Sharing	Data will be available via the ECMWF MARS archive and CERISE consortium members can access this using their ECMWF member-state account.
Archiving and preservation (including storage and backup)	The data in the MARS tape library are backed up.

Dataset	**Reanalysis regional demonstrators**
Data set description	Two Arctic offline land surface prototype reanalyses for two up to five-year time periods providing several surface and soil parameters at a resolution of 2.5 km or finer. In addition one European offline land reanalysis prototype for 1993 to 2019 at 5.5 km resolution. Two coupled 3D regional reanalysis prototype data sets, both produced for a full year post 2015: One for a pan-Arctic domain, indicatively at a spatial resolution of 3.75 km, and one with an ensemble reanalysis system for a pan-European domain, indicatively at a spatial resolution of 3.75 km (7.5 km for ensemble members).
Standards and metadata	Data is produced In GRIB-1 and/or GRIB-2 (and possibly NetCDF) formats and are all stored in the ECMWF MARS archive.
Data Sharing	Data will be available via the ECMWF MARS archive and CERISE consortium members can access this using their ECMWF member-state account.
Archiving and preservation (including storage and backup)	The data in the MARS tape library are backed up.

Work Package 5:

Completed by: Núria Pérez-Zanón with input from WP partners

Dataset	CERISE seasonal forecast demonstrators
Data set description	Ensemble seasonal forecast simulations will be produced according to an experimental design agreed between WPs 5 and 6. The purpose is to determine whether new ways of initialising the land surface in the simulations have an impact on the quality of the forecasts. Three rounds of demonstrators (referred to as 'phases') are anticipated - pending availability of initial conditions from WPs 4 or 3. All five participants in WP5 will deliver such demonstrators: for phases 1 and 2, all will produce new integrations, with new initial conditions; for phase 0 – the benchmark for the analysis – three participants (ECMWF, DWD and CMCC) will also produce new integrations, using new version of models, while the other two (MetO and MF) will offer the most recent version of the operational systems participating in C3S seasonal operations. These simulations will only be performed in retrospective mode (to make it possible to compare with observations,) which leads to the **main limitation** for climate service users: no real-time predictions. The value of this data is in understanding land-surface mechanisms as they operate in (these) models, and improving these models, and thus the reanalysis and prediction products they are used for. Data will be made available for use by all partners in the project, as well as other research groups working on these topics, on request.
Standards and metadata	Qualitatively, this data will be virtually identical to the C3S operational seasonal predictions, in which all partners participate. This makes compliance with pre-existing standards and metadata conventions straightforward: producers of this data will encode the outputs using the tools created for C3S participation; ECMWF will convert this data in the format required by the archiving platform, after checking completeness of metadata and its match with the experiment design. The archiving format is GRIB; for this formatting, some of the metadata is external to the files. Any such information will be documented appropriately for the purposes of this project. The use of standardised ECMWF tools for the creation of the GRIB data will go a significant way towards ensuring uniform quality.

Data Sharing	Data will be shared from the main archive (MARS, at ECMWF), using either the standard MARS interface or a version of the interface provided by the C3S Climate Data Store. ECMWF tools for decoding and operating on GRIB data will be available to all partners and other users. No embargo periods are anticipated. The data will be accessible to partners in this project and scientists from other research groups working on these topics, on request. The licence will allow open, free access and use of this data. This data has no direct value in societal decision making. Details on access will be provided once the data is archived.
Archiving and preservation (including storage and backup)	Data will be archived in ECMWF's MARS archive. It is a technology tested over decades, widely used in research (including collaborative, international research projects), services and operational activities. It ensures reliable access, long-term preservation policies and practices, availability of DOIs and of tools to decode and process the data. The storage costs are supported by ECMWF.

Dataset	**Common assessment scores**
Data set description	Generated data: A set of scores (deterministic and probabilistic) assessing the general performance of the output of the seasonal forecast demonstrators. They will have global coverage for selected atmospheric variables and circulation indices initially expected to be monthly frequency and 1 x 1 degree of spatial resolution. The purpose is to determine whether the new ways of initialising the land surface in the simulations show an impact on the quality of the forecasts. Available data: In order to perform the assessment, reference datasets already available at MARS will be downloaded to be used, such as ERA5. Other datasets, such as GPCP, available at CDS and https://psl.noaa.gov/data/gridded/data.gpcp.html, may be used.
Standards and metadata	Up to our knowledge, CF Metadata Conventions doesn't exist for verification scores. We will include basic metadata and documentation on the storage structure. Data format is NetCDF Basic checks and visual inspection of the plotted scores will be performed.
Data Sharing	Data will be shared from the main BSC archive, using either the B2DROP interface (https://b2drop.bsc.es/) or sent to ECMWF to be read within a private interface of the C3S Climate Data Store No embargo periods are anticipated.

	The data will be accessible to partners in this project. The licence will allow open, free access and use of this data. This data has no direct value in societal decision making. Details on access will be provided once the data is archived.
Archiving and preservation (including storage and backup)	Data will be store at the main BSC archive or sent to ECMWF to be read within a private interface of the C3S Climate Data Store. No DOI is planned at present, but the data could be made accessible through a platform like Zenodo (https://zenodo.org/) that will also act as a backup at the end of the project.

Work Package 6:

Completed by: Frederic Vitart with input from WP partners

Dataset	**Offline hydrological simulations for seasonal forecast demonstrators**
Data set description	Offline hydrological simulations with the CTRIP model, using daily surface and subsurface runoff from seasonal forecast systems as input, at 1° resolution for the period 1993-2022. Daily discharge at 1/2° resolution of a selection of rivers around the world will be provided. CTRIP model was described in the following publication: Decharme, B., Delire, C., Minvielle, M., Colin, J., Vergnes, J.-P., Alias, A., et al. (2019). Recent changes in the ISBA-CTRIP land surface system for use in the CNRM-CM6 climate model and in global off-line hydrological applications. Journal of Advances in Modeling Earth Systems, 11, 1207– 1252. https://doi.org/10.1029/2018MS001545
Standards and metadata	Metadata will be provided in NetCDF header The Data format will be NetCDF There will be a review process to quality- check the data
Data Sharing	Data available on MARS, accessible from web API
Archiving and preservation (including storage and backup)	MARS archive at ECMWF

Dataset	**Offline hydrological simulations for regional reanalysis demonstrators**
Data set description	Offline hydrological simulations with the CTRIP model, using daily surface and subsurface runoff from regional demonstrator CERRA-Land as input, at 1/12° resolution for the period 1993-2003. Daily discharge of a selection of rivers over Europe will be provided. High-resolution CTRIP model was described in the following publication: Munier, S. and Decharme, B.: River network and hydro-geomorphological parameters at 1/12° resolution for global hydrological and climate studies, Earth Syst. Sci. Data, 14, 2239–2258,

	https://doi.org/10.5194/essd-14-2239-2022, 2022.
Standards and metadata	Metadata will be provided in NetCDF header The Data format will be NetCDF There will be a review process to quality- check the data
Data Sharing	Data available on MARS, accessible from web API
Archiving and preservation (including storage and backup)	MARS archive at ECMWF

Dataset	**River discharges**
Data set description	A subset of daily river discharges from the Global Runoff Data Centre will be used to evaluate the hydrological simulations using as input seasonal forecasts and regional reanalysis demonstrators.
Standards and metadata	Metadata will be provided in NetCDF header The Data format will be NetCDF
Data Sharing	GRDC data policy is available here: https://www.bafg.de/GRDC/EN/01_GRDC/12_plcy/policy_guidelines.pdf?__blob=publicationFile. River discharges will be provided to all partners via a ftp server.
Archiving and preservation (including storage and backup)	Data will be archived in the ECMWF ECFS storage system.

Dataset	**Sensitivity experiments for Task 6.2**
Data set description	Seasonal forecast hindcast experiments will be performed to assess the impacts of coupled land data assimilation and time-varying vegetation. Limitations: configuration to be finalized but might not cover more than 4 decades.
Standards and metadata	metadata will be included in the Grib header. Data will be produced in Grib1. Netcdf conversion will be straightforward. There will be a review process to quality- check the data
Data Sharing	Data will be archived in the ECMWF MARS archiving system., which will be accessible from https://apps.ecmwf.int/research-experiments/expver/. Data can be accessible using webapi.
Archiving and preservation (including storage and backup)	Archived in MARS at ECMWF. The data will be preserved during the duration of the project and preservation beyond will be envisaged in case of publications.

Work Package 7:

Completed by: Carla Cardinali, Patricia de Rosnay with input from WP partners

Dataset	**CERISE verification dataset**
Data set description	In situ and satellite datasets used to support verification and diagnostic activities across the CERISE project. It is a collection of existing datasets from various institutions that will be interpolated to 1/4 degree resolution. Satellite products: Snow data: - IMS snow cover from NOAA NESDIS, 1997-present - ESA CCI MODIS Snow cover 2000-2019 and snow water equivalent 1979-2019 - CryoRisk/CryoClim Norwegian Space Center/ESA data 1982-2015 - In situ snow depth from SYNOPand from the Danish GEUS 2008-present LAI: - GLASS NASA LAI based on AVHRR-MODIS, 1981-present - Copernicus-Global Land Service based on SPOT-VGT and PROBA-V, 1999 – Jun 2020 Soil moisture: - ESA CCI soil moisture, 1978-2000 - SMOPS NOAA multi-sensor product. 2017-present Land surface - Copernicus Global Land Service 2010-present - EUMETSAT LSA SAF AVHRR/Metop, 2007-present Albedo: - Copernicus Global Land SPOT/VGT and PROBA-V, 1998-2014 and 2014-2020 - LSA SAF, 2007-present In situ data - Fluxes from ICOS and EUFluxes Cluster, AmeriFlux + NEON, TERN - Snow depth from the SNOTEL network and other sources - International Soil moisture network data Time span: multi-decadal with periods covered depending on the data-source. The data underpins scientific publications related to coupled land-atmosphere data assimilation and initialisation R&I in CERISE
Standards and metadata	Data format: NetCDF4. The dataset will be used internally for evaluation purposes

Data Sharing	Data will be shared internally to the project using a password -protected ftp account
Archiving and preservation (including storage and backup)	Each partner will use its own system for long-term data storage

Dataset	**CERISE land cover, LAI and lakes**
Data set description	Time-varying Land Cover and LAI datasets (CERISE-LC and CERISE-LAI) from 1925 to 1993, and lake cover (CERISE-Lake) 1925-2020. Resolution: 1km Domain: global land surfaces Input data: LUH2 historical land-use data, CCI-land cover, CONFESS-LAI, GEOV2-AVHRR, GEOV2-CGLS, LUH2, GSWE (Global Surface Water Explorer) Time span: 1925-1993 for LAI and land cover datasets, and 1925-2020 for lake cover dataset. Temporal resolution: yearly for CERISE-Lake and CERISE-LC, 10-day for CERISE-LAI The input data will be data will be merged in CERISE and used for the reanalysis prototypes and seasonal forecasts demonstrators
Standards and metadata	Data format: NetCDF4. Metadata according to the Climate and Forecast (CF) conventions
Data Sharing	Data will be available via the ECMWF MARS archive
Archiving and preservation (including storage and backup)	The data in the MARS tape library are backed up

Document History

Version	Author(s)	Date	Changes
0.1	Rhona Phipps, Tanya Warnaars, Patricia de Rosnay and CERISE team	June 2023	Initial version
1.0	Rhona Phipps	27/6/23	Updated after project reviewers comments.

Internal Review History

Internal Reviewers	Date	Comments
Jean-Christophe Calvet (MF) and Núria Pérez-Zanón (BSC)	June 2023	Initial version