Climateurope2

Task 2.2

D2.7 Good practices in provenance of processing methodologies and ways to communicate in a user-friendly manner

Authors: Alessandro Spinuso, Fabrizio Antonio, Xavier Domingo, Sandro Fiore, Julien Lenhardt, Ian van der Neut, Christian Pagé

Key contributors: Alaa Al Khourdajie, Alessia Cattozzo, Birgit Hassler, Christine Träger, Emily Lane, Francesco Lumpp, Gerard van der Schrier, Jared Lewis, John Douros, Marcus Zanacchi, Marlies van der Schee, Massimiliano Drudi, Paul van Schayck, Rob Roebelling, Simone Mantovani

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101056933.

Document Information

Grant agreement	No 101056933
Project title	Supporting and standardising climate services in Europe and beyond
Project acronym	Climateurope2
Project start date	01/09/2022
Related work package	WP2
Related task(s)	Task 2
Lead organisation	KNMI
Authors	Alessandro Spinuso, Fabrizio Antonio, Xavier Domingo, Sandro Fiore, Julien Lenhardt, Ian van der Neut, Christian Pagé
Submission date	31/03/2026
Dissemination level	Public

History

Date	Submitted by	Reviewed by	Vision (notes)
3/12/2025		All Task2 Partners	First D2.7 Outline review
9/01/2026		All Task 2 Partners	Preliminary set of recommendation extracted from the interviews
19/02/2026		All Task 2 Partners	Internal review of technical section and private sector interviews.
27/02/2026	Alessandro Spinuso	Submitted to Pierre Antoine Bretonnière for Internal Review	
27/03/2026	Alessandro Spinuso	Comments addressed by all authors .	Final version ready for submission

Please cite this report as: A. Spinuso, F. Antonio, X. Domingo, S. Fiore, J. Lenhardt, I. van der Neut, C. Pagé, Good practices in provenance of processing methodologies and ways to communicate in a user-friendly manner, D2.7, Climateurope2 Project, 2026.

Table of Contents

Table of Acronyms

Acronym	Full name
AI	Artificial Intelligence
API	Application Programming Interface
AR6 / AR7	Sixth / Seventh IPCC Assessment Report
CAN	Climate Normals & Anomalies
CEN-CENELEC	European Committee for (Electrotechnical) Standardization
C3S	Copernicus Climate Change Service
CDMS	Climate Data Management System
CF	Climate and Forecast (metadata conventions)
CMCC	Euro-Mediterranean Center on Climate Change
CMIP	Coupled Model Intercomparison Project
CMIP6 / CMIP7	Sixth / Seventh CMIP phase
CORDEX	Coordinated Regional Climate Downscaling Experiment
CSRD	Corporate Sustainability Reporting Directive
DDC	Data Distribution Centre (IPCC)
DEM	Digital Elevation Model
DEP	Data Exploitation Platform
DOI	Digital Object Identifier
DQA	Data Quality Assurance
ESGF	Earth System Grid Federation
ESMValTool	Earth System Model Evaluation Tool
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable

Acronym	Full name
FinOps	Financial Operations
GreenOps	Green Operations
HVD	High Value Dataset
IPCC	Intergovernmental Panel on Climate Change
ISO	International Organization for Standardization
KNMI	Royal Netherlands Meteorological Institute
LLM	Large Language Model
Med-MFC	Mediterranean Monitoring and Forecasting Centre
MIM	Metamodel voor Informatiemodellen
MY	Multi-Year
NORA	Dutch Government Reference Architecture
NRT	Near Real-Time
OGC	Open Geospatial Consortium
PID	Persistent Identifier
PROV-O	W3C Provenance Ontology
REF	Rapid Evaluation Framework
RODEO	EUMETNET initiative on High Value Datasets
RO-Crate	Research Object Crate
STAC	SpatioTemporal Asset Catalog
TSU	Technical Support Unit
WCRP	World Climate Research Programme
WGI / WGIII	IPCC Working Group I / III
WMO	World Meteorological Organization

Acronym	Full name
W3C	World Wide Web Consortium

About Climateurope2

Timely delivery and effective use of climate information is fundamental for a green recovery and a resilient, climate neutral Europe, in response to climate change and variability. Climate services address this through the provision of climate information for use in decision-making to manage risks and realise opportunities.

The market and needs for climate information has seen impressive progress in recent years and is expected to grow in the foreseeable future. However, the communities involved in the development and provision of climate services are often unaware of each other and lack interdisciplinary and transdisciplinary knowledge. In addition, quality assurance, relevant standards, and other forms of assurance (such as guidelines, and good practices) for climate services are lagging behind. These are needed to ensure the saliency, credibility, legitimacy, and authoritativeness of climate services, and build two-way trust between supply and demand.

Climateurope2 aims to develop future equitable and quality-assured climate services to all sectors of society by:

● Developing standardisation procedures for climate services

● Supporting an equitable European climate services community

● Enhancing the uptake of quality-assured climate services to support adaptation and mitigation to climate change and variability

The project will identify the support and standardisation needs of climate services, including criteria for certification and labelling, as well as the user-driven criteria needed to support climate action. This information will be used to propose a taxonomy of climate services, suggest community-based good practices and guidelines, and propose standards where possible and appropriate. A large variety of activities to support the communities involved in European climate services will also be organised.

Executive summary

This deliverable examines good practices in provenance and traceability for climate services. It develops a set of evidence-based recommendations to support more transparent, interoperable, and user-oriented climate services, with particular attention to how data origins and processing methodologies can be documented and communicated. The intended audience is both the wider climate-services community and Climateurope2's internal work aimed at initiating future standardisation pathways.

Building on earlier Climateurope2 work, this report combines insights from a broad interview campaign involving institutional actors, national services, research infrastructures, and private-sector providers, complemented by recent literature and technical developments. The analysis shows that current challenges relate less to the absence of technical standards than to the practical interpretation, governance, granularity, and usability of provenance information across different user groups. For readers interested in the evidence gathered through the interviews, conducted using the script provided in **Appendix 1**, **Section 2** presents institutional perspectives, while **Section 3** covers private-sector experiences.

Based on the analysis of these interviews, **Section 4** introduces **fourteen recommendations**, which represent the **main output of this document**. These are formulated as stand-alone but combinable principles that climate services can adopt progressively, depending on their priorities, maturity, and operational preferences. Together, they address shared data stewardship terminology, the structuring and communication of provenance, versioning and persistence, publication readiness and maturity-levels, automation and long-term preservation, as well as emerging needs related to AI, regulation, and trust.

To support informed prioritisation, **Section 4.1** positions the recommendations according to their relevance for upstream and downstream users, and for technical reproducibility versus interpretative support, helping readers identify the guidance most relevant to their context. **Section 4.3** highlights key research gaps and future priorities, emphasising the need for layered provenance and consistent terminology to improve its usability across user contexts. It also points to the importance of advancing interoperable solutions and community coordination. **Section 5** complements this by reviewing recent technical solutions and infrastructure developments that also extend and update the technical work introduced in D2.2 [1].

Keywords

Trustworthy climate services; Provenance and traceability; Reproducibility; Usability; Data stewardship; Standardisation pathways.

1. Introduction

Provenance and traceability are becoming central components of the evolution of climate services, as datasets, processing workflows, and user expectations continue to grow in complexity. Climate information now integrates observational records, modelling outputs, operational services, and increasingly AI-supported analyses, creating new demands for transparency, interoperability, and long-term stewardship. While earlier work within Climateurope2 (Deliverable D2.2) [1] explored provenance mainly from the perspective of data producers and technical infrastructures, this updated deliverable expands the scope by incorporating downstream needs and operational realities gathered through an extensive interview campaign and recent literature.

The interviews, conducted with institutional actors, national services, research infrastructures, and private-sector providers (see Appendix 1 for the interview framework and guiding questions), represent a major methodological effort for the task team and form the empirical backbone of this report. Across these contexts, challenges in provenance emerge less from missing technical standards than from their practical interpretation, governance, and usability for diverse user groups. Institutional perspectives highlight increasing maturity and regulatory expectations, while operational and private-sector experiences emphasise resource constraints and the need to balance technical traceability with clear, user-oriented explanations. The analysis also connects climate-service practices with broader developments in data stewardship and governance, drawing on institutional updates, lineage frameworks, and emerging technical solutions outlined in Section 5.

Building on these insights, the deliverable presents a set of recommendations (Section 4) framed as guiding principles rather than prescriptive technical standards. The aim is to inform community and policy discussions by clarifying priorities for future standardisation efforts, as well as identifying areas where further research is needed. These principles highlight key priorities, trade-offs, and user-oriented requirements, supporting ongoing community-driven dialogue on traceability and data stewardship. In this context, the recommendations are positioned along two complementary dimensions, upstream versus downstream use and technical reproducibility versus interpretative guidance, illustrating how traceability must address both expert workflows and non-expert decision-making needs. By clarifying what matters for different users and operational contexts, the deliverable provides a structured basis that can support initiatives led by CEN–CENELEC[1] and other standardisation bodies.

[1] https://www.cencenelec.eu/about-cenelec/

2. Institutional perspectives: evolving traceability practices and challenges

This section presents updated institutional perspectives on provenance and traceability, drawing on interviews, literature, and operational developments across organisations such as the IPCC, WMO, Copernicus Climate Change Service (C3S), national meteorological services, and European infrastructure initiatives including EUMETNET. These institutions were selected because they represent complementary roles across the climate-service value chain, ranging from global assessment and intergovernmental coordinating bodies to operational data services, national providers, and infrastructures supporting long-term accessibility and interoperability of climate information. These representative, high-impact, and complementary institutions balance between maturity levels: combining highly mature frameworks with evolving practices. Their experience illustrates how traceability practices are evolving in response to increasing expectations for transparency, FAIR implementation, and long-term stewardship, while also highlighting persistent challenges related to documentation effort, scalability, and user guidance. Rather than focusing solely on technical solutions, the institutional experiences emphasise governance, interoperability, and practical implementation pathways that inform the recommendations outlined later in this report.

2.1 IPCC and WMO updates

2.1.1 IPCC: TG-Data guidance and TSUs operations

As an update to the work reported in D2.2 [1], which already documented preliminary traceability approaches implemented during the IPCC's 6th Assessment Report (AR6), the interview with IPCC scientist Alaa Al Khourdajie, involved in the current AR7 cycle, highlights that traceability is becoming increasingly essential for ensuring transparency, credibility, and reproducibility in climate assessments. This emphasis is driven by the adoption of FAIR principles and by strengthened AR7 requirements, which build on lessons learned from previous assessment cycles [2] and require that all underlying data remain publicly accessible throughout the full assessment process. By underlying data we refer to the digital assets directly supporting IPCC report outputs, including input datasets for figures and tables, intermediate processed data, final published data packages, and the associated analysis code and software. Figure 1 illustrates the technical solutions implemented by the IPCC Data Distribution Centres (DDCs) to support the FAIRness of assets underpinning the reports, including input data, intermediate and final datasets, as well as analysis code and software.

While the need for clear lineage, metadata, methods, and intermediate datasets is widely recognised, implementation is hampered by the growing complexity of data handled across Working Groups, inconsistent availability of underlying scripts or inputs, and limited Technical Support Units (TSUs) capacity to manage detailed provenance at scale. To help address these limitations in AR7, TG-Data[2] also developed a set of general recommendations which we report in Figure 2. They reinforce that these efforts must be supported by sustained capacity in TSUs, clear guidance and engagement with authors, and continuity across assessment cycles to track methodological evolution over time.

[2] https://www.ipcc.ch/data/

Together, these measures highlight that effective traceability in AR7 depends not only on technical solutions, but on organisational commitment, adequate resources, and governance practices that make provenance an integral part of the assessment process rather than a late-stage add-on. As an immediate follow-up, confirmed by the interview with Alaa, minimum expectations now include full traceability for all Summary for Policymakers figures, with additional chapter-level data packages desirable where feasible. Alaa emphasised recurring user difficulties in locating or interpreting scattered data and stressed the need for harmonised terminology and structured TSU guidance, as authors cannot act as data experts. Positive examples from AR6, such as the WGIII mitigation figure (SPM.7[3]) whose reusable data enabled widespread replication[4], demonstrate the value of transparent workflows, while gaps in the traceability of scenario data processing highlight remaining challenges.

	Material		
Principle	**Input data**	**Intermediate/Final data**	**Software/Code**
Findable	• DOI issued for datasets or data collections • Rich metadata • Indexed in the DDC catalog • indexed in DataCite and further portals harvesting DataCite • WGI input data included in the official WGI GitHub repository [16]	• DOI issued for datasets or data collections • Rich metadata • Indexed in the DDC catalog • Indexed in DataCite and further portals harvesting DataCite • WGI final datasets linked from the IPCC figure page • WGI final datasets included in the official WGI GitHub repository (https://github.com/IPCC-WG1)	• DOI issued by Zenodo • Indexed in DataCite and further portals harvesting DataCite • WGI software integrated in the official WGI GitHub repository (https://github.com/IPCC-WG1)
Accessible	• Catalog entries link to public data holdings at the DDC or an external provider	• Catalog entries link to public data holdings at the DDC	• Public code repositories
Interoperable	• Community data formats • Community metadata standards and vocabularies • Common DDC metadata schema for DDC catalog • Documentation and references accompany the data	• Well known data formats (CSV, netCDF) • Common DDC metadata schema for the DDC catalog • Documentation and references accompany the data	• Recommended use of open end-to-end frameworks: Climate4R and ESMValTool [17]
Reusable	• Permissible data usage license • Rich metadata • Metadata-data consistency checks • Provenance information and data-report interlinking provided as available	• Permissible data usage license • Rich metadata • Provenance information provided as available	• Permissible software license • Rich documentation of repository content and software usage including instructions on figure reproducibility

https://doi.org/10.1371/journal.pclm.0000533.t001

Figure 1: Overview of FAIR solutions adopted in the IPCC AR6 for the different type of assets used in the report [2]

Regarding AI, Alaa noted its potential for literature search but warned that synthesis via LLMs is unsuitable due to risks of distortion, making explicit documentation of algorithmic bias and clear links back to source information essential.

[3] https://www.ipcc.ch/report/ar6/wg3/figures/summary-for-policymakers/figure-spm-7/
[4] https://ipcc-browser.ipcc-data.org/browser/dataset?id=447

Outcome	Recommendation
Start early	Provide guidance and training to authors for the organization of code and data at the beginning of the assessment period. Empathy with the challenge of finding a balance between the practices of authors and the requirements for long term archival.
Flexibility	Show empathy with authors to find a reasonable balance between the practices and time constraints of authors and the requirements for long term archival.
Enhanced integration	Integrate the FAIR guidelines into the IPCC assessment process with clear data policies, enhanced internal collaboration, effective communication and outreach, and improved tool support.
Fully linked data provenance	Apply FAIR data practices to connect input data, through intermediate data to the final data plotted in the figures.
More author engagement	Inform authors about mandatory requirements and provide clear timelines.
More engagement with external partners	Engage and collaborate with external partners providing input data or technical solutions in workflow improvements, FAIR guidelines implementation and training and outreach activities
Reduce time pressures	Provide sufficient resources at the end of the cycle, when the substantial part of archival and curation work will be completed, including the inter-connection of data and code with the report itself.
License compatibility	Encourage authors to pay attention to the licenses that are used to publish their data and strongly encourage the use of open and accessible licenses.
Continuity between assessments	Improve data workflows between assessment cycles and provide hands-on assistance to authors.
Additional positions in TSU	Hire dedicated TSU staff supporting authors and working with the IPCC DDC for the duration of the assessment cycle.

https://doi.org/10.1371/journal.pclm.0000533.t002

Figure 2: Synthesis of IPCC TG-Data recommendations fostering FAIRness and traceability in AR7 [2].

2.1.2 WMO - Climate Data Management System (CDMS) specifications

According to the updated WMO Specification (Section 4.3.3 of WMO-No. 1131, 2025) [3], the CDMS provides end-to-end provenance and lineage capabilities that includes mandatory capture of: (1) what changed, (2) when, (3) what it was derived from, (4) what was done, (5) how and why, (6) who or what performed the change, and (7) the authority under which the change was made, see Figure 3.

The WMO specifications expects that CDMS shall maintain full lineage traceability from raw observations through all transformations to published outputs, shall support creation, maintenance, quality control, and monitoring of provenance metadata, and shall embed lineage descriptions within dataset discovery metadata. For the implementation, the recommendations refer to W3C PROV [5] as a data model, in alignment with ISO 19115 lineage structures.

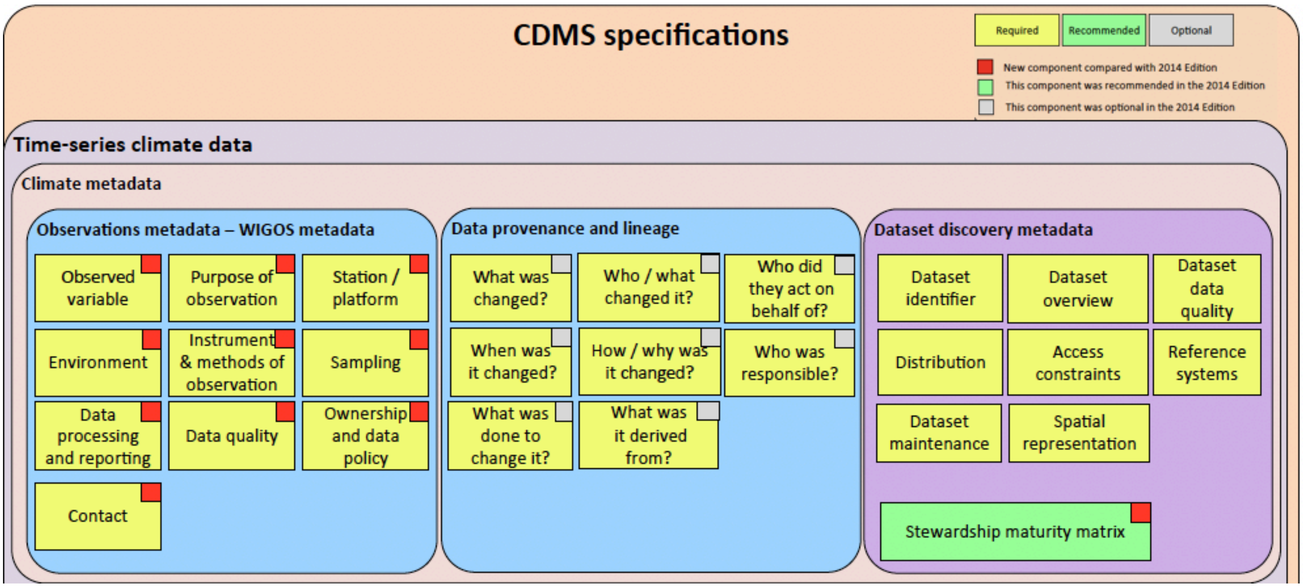

Figure 3 : Excerpt from the updated CDMS figure depicting *Time-series climate data components* showing that the provenance and lineage component is now classified as a required element, whereas in the previous version it was only optional.

Besides indicating the main expected components associated with provenance information, WMO acknowledges that maintaining high-quality provenance metadata is particularly challenging, so priority should be given to data of high importance and visibility such as high-quality climate monitoring datasets. Notably, the notion of “high-priority” datasets increasingly aligns with the EU concept of High Value Datasets (HVD). Reflections on traceability and metadata emerging from HVD-related initiatives are therefore further explored through the interview with the EUMETNET RODEO project, which is discussed later in this document. Finally, preliminary guidance on traceability aspects associated with increased quality and stewardship maturity level, has also been introduced by WMO[5], drawing on earlier literature in this field [4]. We will refer to maturity levels again in the Derived Recommendations section (See R11 - Align traceability requirements with publication readiness).

2.2 EUMETNET RODEO: Traceability of High Value Datasets

The interview with Paul van Schayck, Data Engineer at KNMI, focused on the traceability needs within RODEO[6], a European EUMETNET[7] initiative aimed at improving access to and usability of observational High Value Datasets from National Meteorological Services. Because the service delivers long-term station-based datasets, the interviewee highlighted that users depend heavily on accurate station metadata, clear documentation of changes in measurement conditions, and transparent information about how observations are processed, aggregated, or quality-controlled. The most recurrent user need relates to understanding station history, interpreting quality flags, and

[5] https://library.wmo.int/idurl/4/68238
[6] https://rodeo-project.eu/about/
[7] https://www.eumetnet.eu/

assessing the reliability of extremes and trends. Key challenges include incomplete or inconsistent historical metadata, fragmented legacy systems, limited standardisation across Europe, and resource constraints that hinder systematic provenance capture. Despite these challenges and the complexity associated with implementing formal provenance models, the interviewee emphasised the value of practical, human-readable documentation (see Recommendation R5) and better-structured metadata as immediate enablers for user understanding, and also as a foundation for future AI-assisted data exploration. The latter will particularly benefit from structured and formal metadata including provenance details. Thereby, with the purpose of incrementally adopting formal representation and model in support of traceability, training on standards (DCAT[8], W3C PROV [5] , LinkedData[9]) would be highly beneficial.

2.3 EUMETSAT Climate Normals & Anomalies Service

In collaboration with other tasks of WP2 we also gathered insights from representatives of EUMETSAT[10], which acknowledges both the importance, as well as the complexity of implementing traceability for satellite-based climate data. Provenance is recognised as an important objective, with a clear ambition to achieve end-to-end traceability. This is particularly challenging for global and long-term climate data records, which combine observations from multiple satellites, instruments, agencies, and international partners. To address this, EUMETSAT aims to provide comprehensive provenance tracking within their future Climate Normals & Anomalies (CAN) service, documenting all processing steps from raw observations through successive processing levels to derived and anomaly products, and linking each step to persistent identifiers for datasets, documentation, validation reports, and algorithm descriptions. The availability of the CAN service and its associated information is expected later in the year. While full reproducibility will be supported through the public release of software, the EUMETSAT representative emphasised that detailed provenance may not be equally accessible to all users. For this reason, technical documentation will be complemented by clearer, more accessible product descriptions tailored to the broader geoscience community.

2.4 C3S perspective on traceability and associated challenges

The interview with a representative of C3S[11] highlighted that traceability is a core operational requirement for maintaining quality, authority, and trust in climate services delivered to a wide and diverse user base. Operating within the Copernicus framework, C3S treats traceability as an integral part of data stewardship rather than an optional enhancement. This is reflected in their systematic adoption of international standards such as CF conventions[12], ISO 19115[13]/1939[14] metadata, controlled vocabularies, DOIs[15], and consistent data access mechanisms across APIs and user interfaces. Traceability is closely tied to their quality-control model, which combines contractual

[8] https://www.w3.org/TR/vocab-dcat-1/
[9] https://www.w3.org/wiki/LinkedData
[10] https://www.eumetsat.int/
[11] https://climate.copernicus.eu/
[12] https://cfconventions.org/
[13] https://www.iso.org/standard/53798.html
[14] https://www.iso.org/standard/67253.html
[15] https://www.doi.org/the-identifier/what-is-a-doi/

pre-delivery checks with continuous monitoring and escalation mechanisms involving data providers as third-line support.

A key strength of C3S is its mature versioning and user notification strategy, where datasets are explicitly tagged, older versions are retained under deprecation policies, and users are informed of changes through catalog banners, forums, and known-issues tables. This ensures transparency and supports reproducibility over time. However, the interview also highlights several challenges. While C3S achieves a high FAIR maturity[16] and applies external quality control through independent third parties, fully leveraging traceability information for end users remains difficult. There are acknowledged gaps in user-facing documentation explaining how to interpret lineage and versioning details, and a recognised need for consensus on terminology across the climate services community to avoid fragmentation or competing standards.

Looking ahead, C3S identifies scalability and emerging technologies as major challenges for traceability. Upcoming datasets such as CMIP7 and ERA6 will significantly increase data volumes, requiring innovative serving approaches while preserving lineage and version transparency. In addition, while machine learning is increasingly explored in data production, traceability requirements for AI-generated products are not yet fully defined, representing an open area for future development. Overall, the C3S interview illustrates a highly mature institutional approach to traceability, combined with pragmatic recognition that usability, documentation, and AI-related traceability remain key challenges to address.

2.5 WCRP - CMIP Rapid Evaluation Framework (REF)

The interview explored the crucial role of traceability and provenance in climate data science through the experiences of Jarred Lewis and Birgit Hassler in developing the REF framework[17] and the ESMValTool software[18]. The central theme was the need to guarantee scientific integrity in high-impact reports such as those of the IPCC. Hassler emphasizes that, in order to maintain public and scientific trust, every published figure or metric must be fully reproducible. This means not only presenting the result, but also documenting the entire chain of custody: which raw datasets were used (tracked via DOIs or the handle system), which exact software version was employed, and which computational steps were carried out. An interesting policy for ESMValTools extensions is that "without these records, ESMValTool does not accept new diagnostics into its codebase".
An important social aspect that emerged is the usefulness of these tools for researchers in the Global South. Through REF, which can compute metrics on centralized nodes (such as ESGF), researchers with limited computational or bandwidth resources can directly access analyzed results or intermediate files, without having to download and process massive raw datasets.
However, implementing traceability faces both practical and cultural obstacles. Lewis and Hassler noted that, although there is broad agreement on the importance of transparency, scientists often perceive the completion of provenance records as an additional burden. Furthermore, there is technical complexity in managing version control for millions of intermediate files produced during model development.

[16] https://www.ecmwf.int/en/newsletter/183/news/towards-fairer-data-stores-service
[17] https://wcrp-cmip.org/cmip-phases/cmip7/rapid-evaluation-framework/
[18] https://esmvaltool.org/

In conclusion, the interviewees agreed that provenance is not merely a best practice but a necessity, one that will become even more relevant with the growing use of Artificial Intelligence. In a future shaped by synthetic data and increasingly complex models, knowing exactly how a dataset was generated will be the only way to maintain trust in climate science.

2.6 KNMI Climate Scenarios, E-OBS and ECA&D

We conducted an interview with John Douros, the data steward of the KNMI Climate Scenarios[19] project and highlighted that the primary motivation for implementing traceability was to ensure scientific integrity and reproducibility, especially since earlier editions of the climate scenarios had suffered from unclear workflows and missing documentation. To address this, the team undertook a substantial effort to document the data sources, processing steps, and software versions used to generate all figures and tables in both the scientific and user reports, creating detailed provenance tables and managing software pipelines through GitLab versioning. This work significantly improved transparency but also revealed challenges, such as uncertainty about what level of detail to document, the large workload involved, the inconsistent storage of intermediate datasets, and the difficulty of engaging scientists who naturally prioritized their scientific tasks over documentation. The climate scenarios rely heavily on RACMO model outputs and CMIP data [6], and the interviewee noted that external metadata for these sources was generally adequate and did not create major issues. While ISO-based metadata standards were partially applied through the KNMI Data Platform[20], the provenance materials produced in the project were intended for internal reproducibility rather than external publication. Looking ahead, the interviewee stressed that broader exposure of traceability information would require clearer guidance, more training for data stewards, and machine-readable approaches, particularly in light of growing Open Science expectations. AI tools were not used in the project, and although the interviewee sees potential benefits, he emphasized that any use of AI must avoid adding new uncertainties or reducing transparency.

We also collected additional insights from ECA&D[21], and its gridded product E-OBS. This is another KNMI hosted service which provides validated observational climate datasets compiled from multiple national meteorological services across Europe. While each contributing service applies its own quality-control protocols, these procedures are not always transparently documented, and provenance information is not consistently exposed to users. In principle, E-OBS allows each data point to be traced back to its associated metadata and data provider, but navigating this information requires significant expertise and familiarity with the system. As a result, most users do not attempt to trace issues themselves and instead report problems directly to the E-OBS team, creating a substantial support burden. This experience highlights that making provenance technically available is not sufficient on its own; it must also be accessible and understandable if users are to effectively trace and interpret the data they use.

[19] https://www.knmi.nl/kennis-en-datacentrum/achtergrond/knmi-23-klimaatscenario-s
[20] https://dataplatform.knmi.nl/
[21] https://www.ecad.eu/

2.7 SMHI Climate Scenario dashboard

The Climate Scenario dashboard (Klimatscenariotjänsten)[22] hosted at SMHI is an evolving service, with its first iteration published in 2008/2009, and is still seen as a not clearly mature service as of today. Traceability in the service does not come as a requirement from users, but is seen as the right thing to do, especially considering the FAIR principles among other best practices. In the current state of the service, the accessible information corresponding to traceability covers different aspects. Details about latest changes to the data or processing are included in accessible changelogs. Explanatory videos and text provide details about the underlying data (climate models, scenarios, bias adjustment, ensembles), and recommendations on how to use the data or read the graphs (uncertainty, robustness of results). Displayed data products are downloadable, but still lack metadata about processing and data sources. The users of the service are not really identified as of now or merely through feedback, even though some local counties are known to be using the dashboard for climate adaptation activities. For now, only the code and the data from the last step are archived. As a result, a consolidation of the code base and processing behind the dashboard is ongoing. Efforts are focused on creating a data production pipeline that would ensure the reproducibility of the end product, the inclusion of metadata and provenance information during processing, alongside software versioning. The large part of this effort is aimed at internally improving the current infrastructure with an additional motivation to use some building blocks of this pipeline for scientific analysis.

2.8 CMCC and the Mediterranean Monitoring and Forecasting Centre (Med-MFC)

The insights on the implementation of data provenance and traceability within the Copernicus Marine Service are drawn from an interview with Massimiliano Drudi, Computer Engineer at the Euro-Mediterranean Center on Climate Change (CMCC). Drudi outlines the operational complexity of the Mediterranean Sea Monitoring and Forecasting Centre (Med-MFC)[23] where his team handles the physical component of the service (analysis and forecast of temperature, salinity, circulation), dealing with both Near Real-Time (NRT) and Multi-Year (MY) products from the Copernicus data catalogue. This process involves a vast supply chain of data that is provided by other production centers or services, and integrates inputs from satellite observations, in-situ instruments, and other models for the boundary conditions and atmospheric forcing.
A significant portion of the discussion revolves around the motivations for implementing data provenance. Drudi emphasizes that beyond scientific reproducibility, traceability is also important for planning and economic impact assessment. By tracking which data sources contribute to the final product, institutions can assess the overall economic sustainability of the entire infrastructure, justifying the heavy investments required to maintain it. Furthermore, provenance provides additional useful information for end-users to verify the quality of the inputs, ensuring the final data is suitable for their specific needs.

However, the interview also highlights the gap between theory and practice. While the scientific community recognizes the importance of standards like W3C PROV [5] practical implementation is

[22] https://www.smhi.se/klimat/framtidens-klimat/klimatscenariotjansten
[23] https://marine.copernicus.eu/about/producers/med-mfc

often stalled by a lack of mandatory requirements in operational contracts and the sheer difficulty of managing metadata for massive, continuous data streams. Currently, Persistent Identifiers (PIDs) are effectively used to reference NRT[24] and MY[25] products, but distinct versioning for updated products is less mature. A key challenge lies in the use of PIDs for version management: input data and derived products are not always consistently linked, and would benefit from being traced through rich metadata that explicitly capture the relationships among their respective PIDs.

The conversation concludes with a look at Artificial Intelligence, agreeing that robust data provenance is the foundational layer required to build trustworthy AI models in the climate sector, distinguishing high-quality scientific data from less reliable sources. On the other hand, AI could provide strong support in analyzing large graphs that include dependencies across heterogeneous data, while accounting for the temporal dimension and downstream data derived from the final product.

2.8 National-scale workflow to model flood inundation across New Zealand

During the discussion, Emily Lane from Earth Sciences New Zealand described a national-scale initiative aimed at modeling flood inundation across the entire country. The project operates on a catchment-based domain structure, where each hydrological basin is modeled individually. A core component of the workflow is the construction of high-resolution Digital Elevation Models (DEMs), generated by integrating multiple heterogeneous data sources, including Lidar surveys, river network datasets, and OpenStreetMap information. These DEMs provide the topographic foundation necessary for reliable flood simulation at national scale. The modeling workflow itself combines meteorological and hydrological components. Design storm data are derived from the High Intensity Rainfall Design System (HIRDS), while hydrological processes are simulated using the TopNet model. This integrated pipeline allows the team to assess flood probabilities under different return periods and to incorporate climate change projections, expressed in terms of degrees of warming. As a result, the system can simulate both current and future flood risk scenarios in a consistent computational framework.

A particularly innovative aspect of their approach is the explicit incorporation of data provenance within the Digital Elevation Model structure. Each pixel of the elevation model is associated with an additional provenance layer that specifies the precise origin of the underlying data. For example, identifying the specific Lidar flight, survey campaign, or estimated bathymetric source. This pixel-level traceability enables a transparent reconstruction of how each model input was generated and significantly strengthens reproducibility and accountability.

The importance of strict version control emerged as a critical lesson from their experience. At one point, inconsistencies were detected between present-day and future climate simulations. The issue was eventually traced back to the inadvertent use of an outdated Digital Elevation Model version in the future scenario pipeline. This episode clearly demonstrated that even minor lapses in version tracking can compromise the validity of comparative analyses, particularly in climate change studies. It

[24] https://doi.org/10.48670/mds-00359
[25] https://doi.org/10.25423/CMCC/MEDSEA_MULTIYEAR_PHY_006_004_E3R1

reinforced the need for rigorous management of input data versions throughout the modeling lifecycle.

From an infrastructure perspective, the team uses Cylc[26] to orchestrate the complex workflow, ensuring reliable scheduling and execution of interdependent modeling tasks. For data publication, they rely on a custom-developed Data Hub[27] rather than external repositories such as Zenodo, due to security constraints. This platform allows them to issue DOIs and align with FAIR principles, promoting findability and reusability of outputs. Despite funding and operational challenges, the project adopts an open data policy. Results are released under a CC-BY-4.0-NC[28] license to maximize accessibility and practical usability for stakeholders, including regional councils and infrastructure providers. This openness is considered essential for translating scientific modeling into actionable policy and planning decisions. Tracking provenance in such a national-scale workflow is an important need and would represent a key extension to this work to further enhance transparency of the simulations as well as their reproducibility.

Finally, Lane reflected on the relationship between traceability and artificial intelligence. While AI tools may assist in summarizing complex provenance information for human users, she emphasized that trustworthy AI-driven predictions, particularly for extreme weather events, depend fundamentally on rigorous traceability. Knowing exactly which software version, parameter configuration, and input dataset produced a given result is indispensable for maintaining scientific credibility and stakeholder trust in climate risk assessments.

2.9 Lineage at NORA (Dutch Government Reference Architecture)

The Data Lineage Guidelines developed by NORA (Dutch Government Reference Architecture) Expert Group Data Management[29], provide a comprehensive framework for establishing trust, transparency, and accountability in government data workflows. They define data lineage as the complete record of data origins, movements, and transformations, including who performed which actions, when, how, and for what purpose. The guidelines distinguish between horizontal lineage (data movement across organizations and systems) and vertical lineage (relationships between data, models, concepts, rules, and technical artifacts) and emphasize that both are needed for robust understanding of data provenance. They outline multiple granularity levels, from dataset-level lineage to detailed element-level lineage and even instance-level lineage, depending on user needs and regulatory obligations.

The guidelines highlight several key benefits, including improving data quality, supporting impact/error analysis, enabling reproducibility, ensuring compliance with laws such as GDPR and the AI Act, and increasing public trust through transparency. They identify a diverse set of stakeholders: citizens, auditors, data stewards, architects, scientists, and describe how each benefits differently from lineage, reinforcing that lineage views must be tailored to user roles.

---

[26] https://cylc.github.io/
[27] https://data.niwa.co.nz/
[28] https://creativecommons.org/licenses/by-nc/4.0/deed.en
[29] https://www.noraonline.nl/wiki/Handreiking_Datalineage

Implementation guidance stresses the importance of open standards for interoperability across government organizations, with PROV-O[30], Dublin Core[31], DCAT,[32] MIM[33], and OpenLineage each covering complementary aspects of provenance, cataloguing, and transformation documentation. A set of lineage “patterns” provides reusable architectural solutions, from simple object-level metadata to full reproducibility patterns required for scientific workflows or policy-relevant datasets (pp. 20–27).

2.9.1 Similarity with the climate services context

Just as government data requires public accountability, climate services require users to trust scientific datasets, models, and derived products. Interviewees from IPCC and climate-tech organizations repeatedly stressed the need to understand what data was used, trace how transformations were applied, and assess whether processing choices affect interpretability or uncertainty. This aligns directly with the guideline’s emphasis on transparency and traceability for public trust.

Horizontal and vertical lineage are both critical: Climate services routinely integrate data from multiple sources (e.g., observations, reanalysis, CMIP, emissions inventories). Horizontal lineage helps track inter-organizational flows (e.g., C3S → national services → downstream climate providers), especially addressing the delivery value chain of climate service-products. Vertical lineage is needed to clarify data and model-level transformations, especially where scientific rules or algorithms encode domain knowledge. Interviews emphasised this duality, particularly in workflows where users must explain dataset choices or methodological differences.

Lineage granularity linked to stakeholder needs: Climate-service interviews showed a similar pattern with policy-facing users needing summary-level explanations (dataset used, method applied, version), technical users needing algorithm-level or code-level lineage for reproducibility, and the private sector being prepared with internal provenance chains to make sure they can keep-up with changes affecting their offer, as well as providing scientific evidence backing up their portfolios, to increase trust and credibility.

---

[30] PROV-O https://www.w3.org/TR/prov-o/
[31] Dublin Core; https://www.dublincore.org/
[32] DCAT; https://www.w3.org/TR/vocab-dcat-3/
[33] MIM; https://www.noraonline.nl/wiki/MIM_%28Metamodel_voor_informatiemodellen

3. Private sector approaches to traceability

This section presents additional private-sector perspectives on provenance and traceability, based on focused interviews conducted using the script in Appendix 1 and on joint interviews carried out in collaboration with other Climateurope2 WP2 tasks. The insights have been organised around specific market sectors to highlight shared challenges, benefits, and operational priorities emerging from different application contexts. The views and practices described reflect the experiences shared during these discussions and do not necessarily represent the positions of all companies in the sector, which may evolve over time. Given the exploratory and qualitative nature of this work, the findings should not be considered statistically representative of the broader climate-services market or any specific sector, nor interpreted as the official position of the authors or of the Climateurope2 project.

Acknowledgements
We would like to acknowledge the valuable contributions of all participating organisations including, among others, representatives from Hypermeteo, MEEO, Hydroclimat, Lobelia, MeteoClim, Krateo and Geoskop, whose perspectives helped inform the analysis presented here. We also extend our sincere thanks to those companies and representatives who preferred to remain anonymous in accordance with internal policies or personal preferences; their insights have equally contributed to shaping the findings and recommendations of this report.

3.1 Carbon sequestration

The insights on climate services in the carbon-sequestration and reforestation market sector are drawn from an interview with a provider that uses ecological modelling, national forest inventories, climate datasets (e.g., ERA5), and remote-sensing information to produce carbon projections and monitor forest resilience. In the interview, the team emphasized that internal reproducibility and model versioning are essential for credibility, particularly during due-diligence assessments. However, traceability is mostly maintained for internal quality assurance rather than as a response to user demand, since clients and investors seldom request detailed metadata or understand scientific standards like peer review. The company relies on their own data lake, Git-based software versioning, and strong internal documentation, but keep proprietary elements opaque for commercial reasons. External climate data is generally trusted and perceived as stable, though they note that clearer versioning, update notifications, and API-level provenance metadata would improve reliability. According to the interviewee, standardization frameworks (FAIR, ISO, WMO guidelines) are largely unknown in their sector and not currently seen as a priority or competitive advantage, although formal publications and open datasets are used strategically to signal scientific credibility. AI is used cautiously for internal tasks, with human fact-checking and no external disclosure, reflecting awareness of risks but a lack of formal organizational policy.

3.2 Climate risk assessment and energy transition support

We report here the interview with a climate-oriented service provider supporting the ecological and energy transition, particularly through tools designed for financial actors such as banks and home-insurance providers. Provenance is managed primarily through detailed metadata capturing data sources, processing steps, and methodological choices, allowing the team to track data lineage and distinguish between open datasets and model-derived information. These practices serve internal quality assurance and enable transparent communication. For example, explicitly stating that object detection is based on certain imagery processed through a machine-learning model, with a validated 90% true-positive rate.

External climate data sources are selected not only for scientific validity but also for user trust and recognizability, illustrated by their decision to abandon a private provider of post-processed CMIP6 data in favour of nationally endorsed DRIAS datasets[34], albeit from CMIP5. While formal standards (ISO, AFNOR[35]) are viewed as impractical at their scale, the company follows widely recognized best practices and undergoes indirect audits in financial partnerships. Overall, provenance communication aims to remain accurate yet accessible: users rarely demand technical metadata, but the company prioritizes clear, user-friendly explanations of methods, data origins, and known limitations to support informed decision-making without overwhelming non-expert audiences.

In a separate interview with a climate data scientist working for another company contributing to this sector, we explored the role of traceability within Climate Sentinel Index and Scenario services. These services translate complex scientific datasets, such as Copernicus reanalysis products and CMIP6 projections, into high-resolution, actionable indicators for the energy, finance, and infrastructure sectors. The interviewee highlighted that the main objective is to support risk assessment and adaptation planning, while aligning with emerging regulatory frameworks such as the EU-Taxonomy[36] and Corporate Sustainability Reporting Directive (CSRD)[37].

A recurring theme in the discussion is the necessity of traceability for building trust. The scientist notes that as the climate service market grows, it is crucial to distinguish scientifically robust services from "black boxes." Clients need to know the provenance of the data to validate their decision-making processes. However, according to the interviewee, the company faces the specific challenge of finding a balance between offering full transparency on data lineage and protecting their proprietary methods and intellectual property.

Although reproducibility is supported through detailed documentation and ISO 9001–aligned quality practices, and FAIR principles are partially addressed via OGC-compliant APIs and standard data formats, formal adoption of metadata standards such as W3C PROV [5] and ISO 19115 has not yet been fully achieved. Future improvements are anticipated within the EU-funded RI-SCALE project, which aims to provide stakeholders in this sector with modern Data Exploitation Platforms (DEPs), where persistent identifiers and more rigorous provenance tracking are expected to become core components. The interviewee also highlights the need for better upstream data management from scientific providers (like C3S and ESGF), requesting standardised and easily accessible provenance information, clearer versioning and change logs, and machine-readable metadata in order to improve transparency and reproducibility of climate products.

---

[34] *French national portal for regionalised climate scenarios for the impact and adaptation of our societies and environments,* https://www.drias-climat.fr

[35] *French standardisation association,* https://www.afnor.org

[36] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R2139

[37] https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32023R2772

The interview concludes with a focus on Artificial Intelligence, asserting that it is essential to establish robust and transparent traceability measures to build trust in AI-generated climate information, especially when it is used to support operational or strategic decisions. More specifically, end-users need to clearly understand where data comes from, how the AI model was built, and how the final output was produced. Combining climate data provenance with AI model features would provide a unified and accessible description of the entire workflow.

3.3 Earth observation data analytics and environmental monitoring

The interview involved two representatives from a private-sector organisation specialising in the development and implementation of products and services based on remote sensing of the Earth-Atmosphere system.

The core of the discussion gravitated toward the concept of data provenance and traceability, focusing on the "state of the art." In this regard, they argued that while the scientific community has established robust standards for the source data itself, tracing the "processing chain is significantly lagging and still remains a complex challenge. They emphasized that provenance is not just a bureaucratic requirement but a fundamental necessity for trust and reproducibility. Indeed, in a scenario lacking automated lineage, if a user contests a historical result, the provider is forced into a costly "Human Intelligence" process—digging through old emails and relying on memory to determine which version of a dataset was used. Conversely, with a structured data lineage system, resolving such a dispute becomes a matter of minutes, as the system can instantly identify that a specific output was generated using a specific version for both input data and algorithm.

However, implementing these systems presents substantial technical challenges. It was noted that while strict tracking is burdensome and often unnecessary during the internal R&D phase, it becomes mandatory from "Day Zero" once a dataset enters the Public Domain. Furthermore, the shift toward Cloud Native technologies complicates matters: since data is often processed directly in the cloud without being downloaded, the lineage must track any remote data transformation processes rather than simply keeping track of the data used. In addition, it was acknowledged that no single standard (ISO, OGC, W3C PROV) covers everything perfectly: they adopt an "ecosystemic" approach trying to map metadata at a high level to ensure interoperability across different domains (e.g., marine vs. atmosphere).

A significant portion of the conversation was then dedicated to the user perspective, highlighting a key criticality: users frequently demand metadata and traceability (often driven by rigid project requirements), yet they often lack the expertise to interpret this complex information.

Interestingly, the discussion ended on a positive paradoxical note regarding AI. While Generative AI is often criticized for being a "black box," they observed that LLMs are actually excellent consumers of provenance data. They have the capacity to navigate the massive documentation that humans find tedious, effectively bridging the gap between complex technical metadata and the user's need for simple answers.

3.4 Infrastructure-Level Traceability: Insights from Cloud Platform Engineering

Among the interviews, we also include cloud platform engineering startups. The idea to approach a cloud company was to provide an infrastructural perspective to provenance in a climate service context. By combining this view on provenance with others more related to the climate domain can positively complement and provide an added value to this report. During the interview, it was highlighted the strategic importance of data traceability and provenance in managed cloud environments. The company coordinates DevOps and FinOps processes for diverse clients, including financial institutions and manufacturing companies. Although provenance implementation is still at an early stage, mainly due to cost and complexity, the company considers it a high-priority requirement. Without standardized traceability mechanisms, organizations struggle to reconstruct responsibilities during failures or security incidents and to accurately associate infrastructure costs with specific application activities. A key area of application discussed is GreenOps. The company pointed out that current CO_2 emission data used for sustainability reporting often lacks transparency and verifiable origins. A structured provenance framework would enable validation of environmental data sources, increasing trust and reliability in carbon accounting practices. The conversation also explored the relationship between provenance and Artificial Intelligence. While AI systems depend on trusted, high-quality data, Krateo reported that AI is also essential for analyzing the massive volume of traceability information generated by complex infrastructures. Because provenance data spans multiple layers—from low-level infrastructure to high-level application logic—human analysis alone is insufficient. AI-driven analytics are therefore necessary to process this “vertical” data and transform it into actionable insights and comprehensive system observability.

3.5 Additional insights from interviews

Additional interviews with private companies were conducted in collaboration with other WP2 tasks, addressing broader aspects of data and process management. This section presents a combined analysis of these interviews, with a specific focus on insights related to traceability.

We note that while some providers apply and value elements of the FAIR principles mainly for internal efficiency and quality control, often relying on structured metadata, dataset identifiers, and software versioning, they are generally willing to increase the level of transparency when operating under institutional contracts such as Copernicus. Other services delivering climate intelligence for sectors such as energy and finance highlight a critical downstream challenge associated with the limited transparency on climate model selection, assumptions, and ensemble composition, which constrains informed interpretation and decision-making.

Particularly, two companies, involved in sectors such as engineering design and climate risk assessment, traceability is driven less by formal client requirements and more by internal needs for quality assurance, reproducibility, and the ability to explain results credibly over time. This is reflected in systematic dataset and software versioning, with the additional technical solutions, in support of reproducibility, which include containerised processing environments, storage of intermediate checkpoints or quality statistics, accompanied by detailed documentation of methods, assumptions, and uncertainty management. These interviews highlight how methodological transparency is treated as a competitive advantage that builds trust with clients, even though full openness of code or exhaustive metadata is not always commercially viable. At the same time, both interviews highlight emerging challenges, including limited upstream information on model assumptions and ensemble construction, as well as concerns about explainability when advanced statistical methods or AI are introduced.

Representatives from a national private company operating in the climate domain explained that while the company maintains rigorous internal tracking of data origins, transformations, and versions, essential for their internal quality control and ISO certification, this information is rarely exposed to the client. In their view, the market currently prioritizes the final accuracy of the forecast over the "story" of the data (provenance). Consequently, standard metadata protocols or persistent identifiers (PIDs) have not yet been implemented because clients have not requested them. However, they also acknowledged a shift in the landscape, particularly driven by the rise of Artificial Intelligence and argued that traceability would be vital for AI, as the reliability of AI models depends heavily on the quality and known origin of the input data. Therefore they concluded that implementing standardized, interoperable traceability could serve as a certification of data quality and process transparency.

4. Derived recommendations

The recommendations presented in this section combine concise rationale with actionable guidance, directly linking insights gathered from institutional and private-sector interviews, direct project involvement, and recent literature to proposed actions. They reflect how provenance in climate services spans multiple technical dimensions, including observational, processing, and modelling aspects, while also addressing the broader need to support the credibility and interpretability of downstream, user-facing climate information. By integrating perspectives from public institutions, operational infrastructures, and market-oriented providers, the recommendations aim to strengthen user confidence, enable trusted adoption, and support informed decision-making across diverse stakeholder groups. While the recommendations can be combined, most can also be implemented independently and progressively, depending on priorities. Section 4.1 positions them according to their relative contribution to technical reproducibility versus interpretative value.

As a general note, when institutional perspectives are referenced in the introducing rationale and motivations of each recommendation, they reflect the experiences and viewpoints shared by individual interview participants. These insights are included to illustrate practical lessons learned and should not be interpreted as official organisational positions, unless supported by documented implementations or publicly available evidence.

R1. Adopt consistent terminology for data stewardship concepts across climate services

Across almost all interviews, ambiguity in terminology repeatedly surfaced as a hidden but persistent barrier to effective traceability. IPCC contributors noted that even within a single assessment cycle, terms such as provenance, lineage, uncertainty, and methods are interpreted differently across Working Groups, complicating documentation and reuse. KNMI Scenarios and SMHI dashboard interviews showed uncertainty about what “counts” as provenance and where to draw boundaries between metadata, documentation, and methodology. In private-sector interviews, it became clear that these terms are often unfamiliar to non-scientific users, leading to misaligned expectations between providers and clients. RODEO and CMCC further highlighted how inconsistent terminology across institutions makes interoperability and automation difficult.

Recommendation 1 (R1): Adopt consistent terminology for data stewardship concepts across climate services
• Establish and maintain a shared, controlled vocabulary for traceability and data stewardship concepts, aligned with existing international standards where possible. This vocabulary should be used consistently across technical documentation, user guidance and governance processes, and treated as a living resource that evolves with community input. For instance, an initial selection of general stewardship concepts could be summarised as follows: • **Provenance** – Information describing the origin of assets (data, software, derived products),

Recommendation 1 (R1): Adopt consistent terminology for data stewardship concepts across climate services
including sources and attribution. ● **Traceability** – Ability to follow data and products through decision processes. ● **Reproducibility** – Ability to recreate results using the same inputs and processes. ● **Data lineage** – The chain of transformations applied to data. ● **Workflow provenance** – Documentation of processing pipelines and algorithms. ● **Versioning** – Tracking changes to datasets, models, or software.

R2. Define multi-dimensional provenance (data, processing, service interpretation)

Interviews consistently indicated that provenance is often treated as a single, undifferentiated concept, without distinguishing between its different technical and interpretative components. In practice, however, provenance consists of multiple dimensions that require separate consideration. Interviews in sectors such as earth-observation data analytics and environmental monitoring, as well as CMCC, explicitly noted that provenance of input datasets is relatively mature, while provenance of services, workflows, and processing algorithms remains fragmented and poorly standardised. CMCC described the complexity of coordinating provenance across multiple coupled components (physics, biogeochemistry, waves), while IPCC contributors highlighted difficulties tracing figures back through intermediate processing steps. Without separating these dimensions, provenance information becomes either incomplete or unusable. EUMETSAT's experience with satellite-based Climate Data Records clearly illustrates that provenance must span multiple dimensions: observations from different satellites and instruments, processing chains, as well as responsibilities across agencies. Pixel-level Digital Elevation Model provenance from Earth Sciences New Zealand shows how data lineage can exist simultaneously at various dimensions: observational level (Lidar source), processing level (Digital Elevation Model generation), and modelling level (flood simulations).

Recommendation 2 (R2): Define multi-dimensional provenance (data, processing, service interpretation)
Implement provenance as a **multi-dimensional framework** reflecting the different ways climate data are created, processed, and interpreted. Traceability practices should explicitly identify and categorise the assets associated with each dimension, and assign relevant metadata in alignment with the corresponding stewardship concepts (R1) and climate-service uses. Below, we propose a set of general dimensions. ● **Data and Models (origins and context)**. Covers dataset sources, observational systems and instruments, models and derived products. It includes information about licensing, attribution, domain metadata and version history.

Recommendation 2 (R2): Define multi-dimensional provenance (data, processing, service interpretation)
Relevant for: validating input data quality, understanding observational and modeling context, and ensuring FAIR reuse across services. • **Process (transformation and workflow).** Describes algorithms, software environments, automated pipelines and parameter choices that transform inputs into climate products. **Relevant for:** reproducibility, auditing, debugging, and operational monitoring. • **Service and interpretation (meaning and application).** Documents assumptions, methodological decisions, uncertainty framing, scenario selection, and limitations communicated to users. **Relevant for:** downstream interpretation, guidance, policy relevance, and decision support.

R3. Provide clear dataset and software versioning

The lack of explicit versioning emerged as one of the most concrete and costly problems across interviews. Insights from earth-observation and environmental monitoring described cases where requests for traceability information from users led to extensive manual investigation, because dataset or algorithm versions were not clearly recorded. CMCC highlighted the challenge of implementing distinct versioning of near-real-time services, where updates are frequent and incremental. KNMI Scenarios, as well as private sector interviews, stressed that internal debugging and operational reliability depend on tight coupling between datasets and the software versions that produced them. IPCC contributors emphasised that figure-level reproducibility is impossible without precise version references. Finally, a real incident reported by Earth Science New Zealand, which was caused by using an outdated Digital Elevation Model version, demonstrates operational risks.

Recommendation 3 (R3): Provide clear dataset and software versioning
Provide explicit versioning of datasets, software, and configurations using persistent identifiers where possible. Dataset versions should be uniquely identifiable and resolvable, and software versions should be managed through version control systems. Links to these assets should be recorded in the provenance information.

R4. Store and document intermediate data when feasible

Insights from the interviews indicate that the handling of intermediate data products is a recurring challenge, but the issue is raised primarily in the context of national scenario production and private-sector service development, rather than large Copernicus operational centres. The KNMI Climate Scenarios interview highlighted that intermediate datasets were sometimes essential for understanding how figures and indicators were produced, yet were not consistently stored due to uncertainty about their long-term value and storage constraints. Similarly, the earth-observation and environmental monitoring sector, described cloud-native processing workflows in which data is transformed remotely, making it impractical to archive every intermediate dataset, while still recognising the need to preserve enough information to support reproducibility and resolve disputes. Across these interviews, the common issue is not the absence of intermediate data per se, but the lack of a clear and pragmatic strategy for deciding what to store and how to document what is not preserved.

Recommendation 4 (R4): Store and document intermediate data when feasible
Preserve intermediate data products where feasible, particularly when they are critical for understanding, validating, or reproducing high-impact results. When storage of intermediate datasets is impractical due to volume, redundancy or operational constraints, retain detailed metadata, configuration files, software version references, and representative reference values that enable debugging, validation, and assessment of reproducibility without requiring full data preservation.

R5. Provide user-oriented guidance on methods, assumptions, and limitations

Users often do not benefit from technical metadata alone and require explicit guidance on how to read and correctly apply climate products. Across the interviews, participants consistently stressed that clear explanations of assumptions, methodological choices, and uncertainties are essential for users to correctly understand and confidently apply climate information. Contributions from IPCC, KNMI Climate Scenarios, carbon-sequestration stakeholders, RODEO, and E-OBS all highlight the value of concise, human-readable guidance that links provenance details to practical decision-making, as illustrated by usage-oriented guidance pages such as those provided through the KNMI Data Platform, see Figure 4. The energy-transition sector interviews show that users value clear, concrete explanations of how outputs are produced, including model choices, machine-learning methods, and performance metrics such as accuracy rates. Importantly, explanations should remain accessible and avoid excessive technical detail. The SHMI dashboard places significant emphasis on interpretative support through explanatory text and videos[38]. These materials explain climate models, scenarios, bias adjustment, ensembles, uncertainty, and robustness of results, as well as how users should read graphs

[38] *SHMI Climate Scenario Service – Overview and Purpose* https://www.smhi.se/klimat/framtidens-klimat/klimatscenariotjansten/om-klimatscenariotjansten

and use the data. This directly supports the recommendation to require human-readable explanations of assumptions and methodological choices, even in the absence of formal provenance standards.

KNMI Data Platform > Open Data

Documentation hourly, daily, monthly and annual in-situ ground-based meteorological observations in the Netherlands

Time

Timestamps in the file and filename are stored in UTC. For the Netherlands, local time is UTC+1 (CET) and during daylight saving times local time is UTC+2 (CEST).

The timestamp in the files indicates the end of the measuring interval preceding that timestamp. The filenames use an abbreviated timestamp indicating the total interval. See the table below for examples.

Individual variables may be measured and aggregated over a different time interval. This can be shorter than the interval of the file. This is denoted in the description of the variable and the `measurementType.duration` field of the EDR metadata.

Table 1: Examples of how different timestamps are displayed

Type	Aggregation period	Filename	NetCDF and EDR
Hourly	2025-01-02T23:00:00Z 2025-01-03T00:00:00Z	hourly-observations-20250102-23.nc	2025-01-03T00:00:00Z
Daily	2025-01-02T00:00:00Z 2025-01-03T00:00:00Z	daily-observations-20250102.nc	2025-01-03T00:00:00Z
Monthly	2024-12-01T00:00:00Z 2025-01-01T00:00:00Z	monthly-observations-202412.nc	2025-01-01T00:00:00Z

Figure 4: Example of user-oriented guidance supporting the interpretation of metadata and dataset limitations, illustrating how explanatory documentation can complement technical provenance by clarifying, among others, timestamps for meteorological observations[39].

Recommendation 5 (R5): Provide user-oriented guidance on methods, assumptions, and limitations
Provide human-readable explanations of assumptions, for instance, concerning, methods, scenario choices, and known limitations for each climate-service product. These explanations should complement technical metadata and directly support guidance for interpretation of metadata and fitness-for-purpose assessments.

[39] *KDP User Guidance for Aggregated In Situ Meteorological Data*
https://english.knmidata.nl/open-data/hourly-daily-monthly-and-annual-in-situ-meteorological-observations

R6. Provide layered provenance suitable for different user groups

The interviews showed that interpretative guidance alone is not enough to meet the needs of all climate-service users. Technical experts, infrastructure operators, and automated systems require more structured provenance than simplified explanations can provide. Institutional interviews (IPCC, C3S, KNMI Scenarios) and infrastructure-focused perspectives (RODEO, cloud-platform engineering, and carbon sequestration) highlighted the need for provenance to be available at different levels of detail, from concise summaries for decision-makers to detailed workflow lineage for reproducibility and machine-actionable use, while also accounting for confidentiality requirements. While R2 addresses the need to distinguish different provenance dimensions, and R5 emphasises clear human-oriented explanations of assumptions, uncertainties, and methods, R6 responds to the need for a provenance model that can connect these aspects coherently. Its purpose is therefore not only to clarify what is communicated, but also how provenance is structured and delivered across audiences. R6 promotes a layered approach that links human-readable guidance, expert-level technical documentation, and machine-readable metadata within a common framework. Sections 4.5 and 5 point to emerging approaches in the literature that support multi-level and domain-contextualised provenance, often grounded in existing provenance models, which also motivates R7.

Recommendation 6 (R6): Provide layered provenance suitable for different user groups
Provide provenance in layered forms that address different levels of expertise and use. Climate services should combine concise, human-readable explanations for downstream users with more detailed technical documentation for expert users, while also exposing machine-readable metadata for automation, interoperability, and reuse. These layers should remain consistent and linked within a common provenance framework, so users and systems can navigate them according to context and need, while allowing transparency to be adjusted to confidentiality requirements.

R7. Use open, interoperable metadata standards

Interviews and established practice in the climate data community show that traceability relies on the combined use of multiple systems and metadata schemas rather than a unified solution, with some companies not seeing standardisation and shared models directly giving advantages over competitors. RODEO described the value of adopting managed vocabularies and standard data-models, such as PROV-O and DCAT. Several private-sector stakeholders interviewed in this report recognised the relevance of standards such as ISO 19115 and OGC APIs, while noting that no single standard sufficiently addresses the diversity of data types and services they manage. Interoperability of provenance information therefore depends on the coordinated use of complementary frameworks, including DCAT-AP, STAC, ISO 19115, and PROV-O, as also reflected in NORA's reference architecture. Section 5 highlights current implementations and research initiatives that build on these community standards to advance traceability in practice.

Recommendation 7 (R7): Use open, interoperable metadata standards

Adopt and combine open, community-recognised metadata standards and vocabularies to support interoperable provenance and data exchange. Climate services should align their metadata practices within established frameworks (e.g. ISO 19115, DCAT-AP, PROV-O), ensuring that datasets, services, and provenance information can be consistently described, discovered, and integrated across platforms. Where necessary, mappings between standards should be defined to maintain coherence across different provenance layers and user needs.

R8. Ensure stable, persistent, machine-readable endpoints

While R3 addresses versioning and the unique identification of datasets and software, interviews showed that traceability often breaks down when referenced resources are not persistently accessible or machine-resolvable. Stakeholders from IPCC, C3S, RODEO, and KNMI highlighted issues with unstable links, fragmented catalogues, and documentation that is difficult to reuse programmatically, leading users to rely on direct support instead of provenance records. R8 therefore focuses on ensuring stable, persistent, and machine-readable endpoints so that versioned assets remain discoverable, accessible, and usable across infrastructures and automated workflows. The FAIR data literature consistently emphasises that persistent identifiers and machine-readable access are essential for findability, automation, and long-term reuse, particularly in distributed and federated infrastructures. Insights from interviews with IPCC, CMCC, and the New Zealand Data Hub further illustrate how the use of DOIs supports stable access and fosters reproducibility in practice. This approach is also reflected in recently published software archived for long-term preservation and citation, such as the IPCC Atlas DataLab [7], as illustrated in Figure 5.

Recommendation 8 (R8): Ensure stable, persistent, machine-readable endpoints

Provide stable, persistent, and machine-readable access points for datasets, services, and software to ensure that assets remain resolvable and usable over time, especially when referenced through provenance information. Persistent identifiers should be associated with long-term archival, link to structured metadata describing the resource, and, where possible, enable programmatic access to the underlying assets. This recommendation complements and further extends R3.

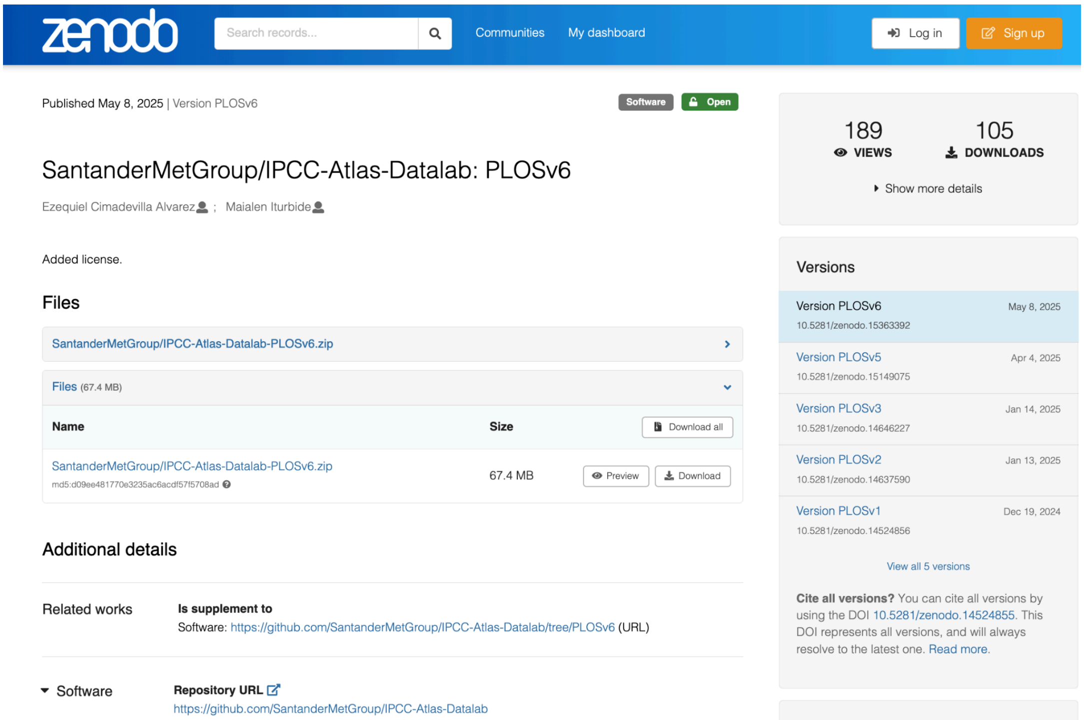

Figure 5: Zenodo release of the IPCC Atlas DataLab software, illustrating how versioned code, licensing, and persistent identifiers (DOI) support reproducibility, traceability and long term archival, of climate-assessment workflows and data-analysis tools. [7]

R9. Automate provenance capture where possible

Interviews show that manual provenance documentation does not scale in operational or near-real-time environments. Experiences from KNMI, CMCC and private providers demonstrate that gaps emerge when provenance depends on post-hoc documentation rather than being embedded in workflows. Automation therefore becomes a foundational enabler for reproducibility, sustainability and long-term governance. Also REF and ESMValTool enforce structured recording of inputs, software, and processing steps, even rejecting diagnostics lacking provenance, while Cylc orchestration used for Digital Elevation Models illustrates automated provenance capture embedded in pipelines. Finally Cloud DevOps/FinOps environments require automated traceability for observability and incident response.

Recommendation 9 (R9): Automate provenance capture where possible
Embed provenance capture directly into data-processing pipelines and workflow orchestration systems so that inputs, transformations, software versions, parameters and outputs are recorded automatically. Automation should prioritise essential lineage elements that support reproducibility

Recommendation 9 (R9): Automate provenance capture where possible
and layered user communication (R6), while remaining flexible enough to integrate evolving standards and tools rather than enforcing a single technical implementation.

R10. Ensure long-term sustainability of provenance information

Across all interviews, long-term sustainability of provenance emerged as a critical but often under-addressed requirement. IPCC contributors stressed that provenance must persist across multi-year assessment cycles to support reproducibility and accountability long after publication. KNMI and SMHI highlighted that project-based or ad hoc storage of provenance information risks losing context over time, especially when staff or infrastructures change. RODEO and CMCC emphasized that operational services evolve continuously, making it essential to retain historical provenance for reanalysis, audits, and troubleshooting. Private-sector interviews further showed that traceability may be needed years later to respond to regulatory reviews. The FAIR literature consistently reinforces that metadata and provenance must remain accessible even when data products are updated or deprecated.

Recommendation 10 (R10): Ensure long-term sustainability of provenance information
Treat provenance information as a long-lived digital asset that is preserved, versioned, and governed alongside data and services. Store provenance in sustainable repositories with clear ownership and maintenance responsibilities, ensure that it remains accessible across service updates and organisational changes, and design stewardship practices that support long-term reuse, verification, and trust in climate services.

R11. Align traceability requirements with publication readiness

Interviews across institutional and private-sector contexts consistently highlighted a clear distinction between internal development and the point at which climate data or services are published or delivered to users. C3S requires datasets to meet defined documentation, versioning, and quality criteria before publication in the Climate Data Store, while private-sector providers describe how traceability expectations increase once products become public or client-facing. IPCC and REF set clear expectations for the traceability of published products, making this a very important requirement; in particular, REF emphasises exhaustive traceability of published outputs (e.g. IPCC figures) as fundamental for reproducibility purposes.

Maturity Scale / Key Component	Level 1 Ad Hoc Not Managed	Level 2 Minimal Managed Limited	Level 3 Intermediate Managed Defined, Partially Implemented	Level 4 Advanced Managed Well-Defined, Fully Implemented	Level 5 Optimal Level 4 + Measured , Controlled , Audit
Data Quality Assurance	Data quality assurance (DQA) procedure unknown or none	Ad Hoc and random DQA procedure not defined and documented	DQA procedure defined and documented and partially implemented	DQA procedure well documented, fully implemented and available online with master reference data Limited data quality assurance metadata	Level 4 + DQA procedure monitored and reported Conforming to community quality metadata & standards External review
Data Quality Control/ Monitoring	None or Sampling unknown or spotty Analysis unknown or random in time	Sampling and analysis are regular in time and space Limited product-specific metrics defined & implemented	Level 2 + Sampling and analysis are frequent and systematic but not automatic Community metrics defined and partially implemented Procedure documented and available online	Level 3 + Anomaly detection procedure well-documented and fully implemented using community metrics, automatic, tracked and reported Limited quality monitoring metadata	Level 4 + Cross-validation of temporal & spatial characteristics Physical consistency check Conforming to community quality metadata & standards Dynamic providers/users feedback in place
Data Quality Assessment	Algorithm/method/model theoretical basis assessed (methods and results online)	Level 1 + Research product assessed (methods and results online)	Level 2 + Operational product assessed (methods and results online)	Level 3 + Quality metadata assessed Limited quality assessment metadata	Level 4 + Assessment performed on a recurring basis Conforming to community quality metadata & standards External ranking
Transparency /Traceability	Limited product information available Person-to-person	Product information available in literature	Algorithm Theoretical Basis Document (ATBD) & source code online Dataset configuration managed (CM) Unique Object Identifier (OID) assigned (dataset, documentation, source code) Data citation tracked (e.g., utilizing Digital Object Identifier (DOI) system)	Level 3 + Operational Algorithm Description (OAD) online, OID assigned, and under CM	Level 4 + System information online Complete data provenance online

Figure 6: Data Stewardship Maturity Matrix, showing different levels of maturity levels across different aspects, including among others, Traceability and Transparency [4].

Finally, the earth observation and environmental monitoring sector emphasised that, while strict provenance may be unnecessary during internal R&D, it becomes essential from "day zero" once data enter the public domain. At the same time, interviews showed that users generally expect clear and consistent traceability summaries rather than full technical detail. Mapping publication readiness to data stewardship maturity levels, as investigated in the literature [4] could provide a practical way to scale traceability over time while ensuring transparency and trust at the moment of publication, see Figure 6.

Recommendation 11 (R11): Align traceability requirements with publication readiness

Align traceability requirements with the publication readiness of climate services by defining standardised traceability profiles that apply at the point of public or operational release. Each published product should include a concise traceability profile, stating clear expectations concerning availability of key data sources, versions, processing steps, assumptions, and uncertainties depending on agreed stewardship maturity levels.

R12. Prepare traceability for AI-enabled climate services

The increasing use of machine-learning methods in climate services reinforces the need for explicit traceability requirements for AI-generated outputs. The IPCC interview highlighted risks when AI systems do not preserve uncertainty framing or source attribution. The energy transition sectors adopt machine-learning models for rooftop detection for solar suitability, confirming that AI outputs must clearly state training data sources, model type, validation metrics, and limitations in order to remain interpretable and auditable. The interview with WCRP stresses provenance as essential in a future involving synthetic data and AI, linking risks directly to scientific integrity concerns. Significant work has been done in the Skills4EOSC project regarding AI/ML and FAIR[40], including provenance and traceability, with recommendations listed in Osmenaj et al., 2025 [8]. Other efforts in the same area like yProv4ML address traceability of AI workloads (with a strong emphasis on tracking energy consumption and application-specific metrics) as well as reproducibility of AI models [9].

Recommendation 12 (R12): Prepare traceability for AI-enabled climate services
Ensure that AI-enabled climate services include transparent, layered traceability describing training data, model purpose and version, preprocessing steps, validation metrics, and known limitations. Provenance for AI components should remain interpretable for non-experts while providing machine-readable documentation for automation and auditability. Rather than prescribing specific technical standards, services should integrate AI traceability into existing provenance frameworks to maintain continuity across traditional and AI-driven workflows.

R13. Anticipate emerging regulatory, FAIR, and Open Science requirements

Institutional interviews (IPCC, WMO, Copernicus) highlight increasing regulatory and governance expectations around transparency, FAIRness, and accountability, while GreenOps and sustainability reporting point to emerging compliance pressures. The interviews confirm that early alignment with traceability practices helps organisations adapt to evolving requirements and reduce future compliance burdens. However, many organisations still struggle to operationalise these expectations consistently, particularly under resource constraints.

40 https://www.skills4eosc.eu/resources/events/paris-10-june-2025/52-13-fair-practices-in-artificial-intelligence

Recommendation 13 (R13): Anticipate emerging regulatory, FAIR, and Open Science requirements
Adopt a forward-looking approach by monitoring emerging Open Science, FAIR, AI-governance and regulatory developments, and translating them into practical traceability principles rather than rigid technical requirements. Organisations should support data stewards through training on FAIR and OS methodologies[41,42], provide a clear mandate for the development of tailored implementation guidance, and foster spaces for community coordination to ensure that provenance practices remain fit for purpose while sustaining interoperability across infrastructures and initiatives.

R14. Prioritise recognisability and trust in provenance communication

Evidence from interviews and existing climate data infrastructures shows that trust in climate-service outputs is shaped not only by technical quality but also by the visibility of authoritative sources, recognisable institutional and scientific endorsement. Major services such as Copernicus Climate Data Store and IPCC Data Distribution Centres routinely highlight institutional provenance, scientific references, and methodological documentation to signal credibility and enable reuse. Private-sector interviews confirm a similar dynamic: downstream users rarely inspect full technical lineage but rely heavily on trusted data providers, scientific publications, and clear explanations of data origin to assess reliability.

Recommendation 14 (R14): Prioritise recognisability and trust in provenance communication
Traceability practices and communication should explicitly reference and highlight authoritative sources, such as peer-reviewed literature and endorsed material associated with the particular product or asset. These should be made visible, for instance, through high-level summaries designed for non-expert users. Provenance information should be communicated in ways supporting transparency, while recognising the social and institutional dimensions of trust in climate services.

4.1 Positioning of traceability recommendation, stakeholder needs

The interviews and derived recommendations confirm that climate data are used across diverse scientific, operational, and decision-support contexts, each associated with different expectations toward provenance and traceability. Institutional actors such as IPCC, WMO, Copernicus services, and national meteorological organisations emphasise reproducibility, long-term stewardship, and methodological transparency, particularly where complex processing chains and multi-model

[41] https://learning.skills4eosc.eu/course/view.php?id=%2019
[42] https://fair-impact.eu/events/fair-implementation-workshops

workflows are involved. Examples such as REF and national-scale modelling pipelines illustrate how provenance supports auditing, error tracking, and scientific credibility.

These insights reveal that provenance must serve both scientific functions (reproducibility, auditability) and stakeholder functions (interpretation, credibility, policy relevance). Thus, to help clarify the scope and intended impact of the recommendations, in Figure 7, we positioned them along two complementary dimensions that reflect the climate-service value chain and the nature of traceability needs. It should be noted that this positioning represents an expert-based interpretation of the recommendations and their intended scope, and is therefore presented as a conceptual framework rather than a result derived from quantitative or statistical analysis.

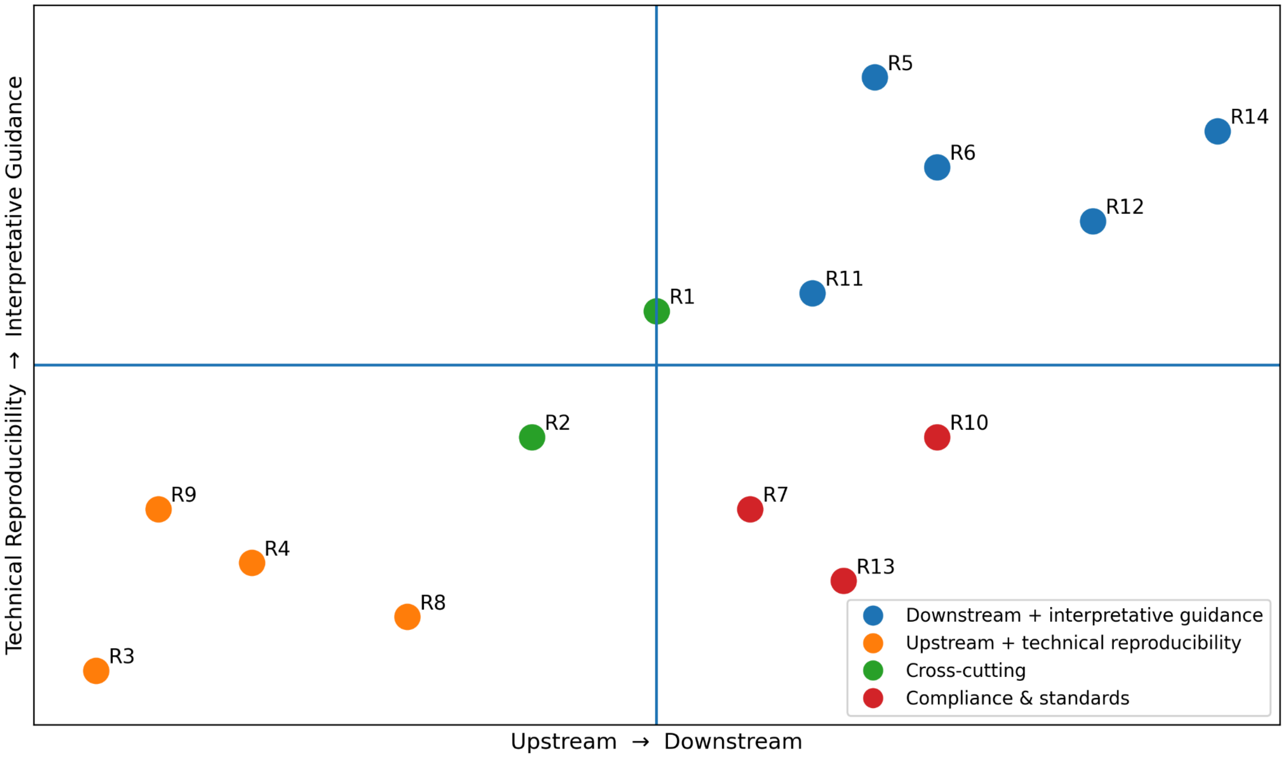

Figure 7: Conceptual positioning of traceability recommendations, grouped by colour to reflect their primary impact on upstream technical reproducibility, downstream interpretation and guidance, cross-cutting concerns over terminology, and standards.

The horizontal axis distinguishes between upstream actors and downstream users. Upstream actors are represented by data producers, model developers, and infrastructure providers. Downstream users cover service operators, decision makers, and end users. The scientific community spans across upstream and downstream based on its involvement in the chain, domain, and expertise. The vertical axis reflects the primary aim of each recommendation, ranging from supporting technical reproducibility and auditability to strengthening sound interpretation, communication, and guidance.
This conceptual positioning highlights how some recommendations primarily address upstream technical foundations, while others focus on downstream understanding and responsible use, with

several spanning both dimensions. Together, the chart illustrates that effective traceability in climate services requires a balanced approach, combining robust technical practices with clear interpretative support across the full data-to-decision pathway.

The colour coding highlights four functional groups that help readers navigate the recommendations according to their primary purpose. The interpretation associated with each group is described below.

The bottom-left cluster (upstream + technical reproducibility) groups recommendations that strengthen operational robustness and workflow transparency (e.g., versioning, automated lineage capture, persistent endpoints). These recommendations mainly target data producers, infrastructure developers, and technical stewards responsible for ensuring reproducibility and auditability of climate workflows.

R3. Provide clear dataset and software versioning
R4. Store and document intermediate data when feasible
R8. Ensure stable, persistent, machine-readable endpoints
R9. Automate provenance capture where possible

The **top-right cluster (downstream + interpretative guidance)** focuses on how provenance supports understanding and responsible use of climate information. Recommendations in this area emphasise layered explanations, assumptions, uncertainty framing, AI transparency, and recognisable sources, helping decision-makers and non-expert users interpret outputs with confidence.

R5. Provide user-oriented guidance on methods, assumptions, and limitations
R6. Provide layered provenance suitable for different user groups
R11. Align traceability requirements with publication readiness
R14. Prioritise recognisability and trust in provenance communication
R12. Prepare traceability for AI-enabled climate services

A set of **cross-cutting** recommendations sits near the centre of the diagram. These address conceptual alignment and governance aspects such as terminology, multidimensional provenance, and stewardship practices, which underpin both upstream and downstream implementation.

R1. Adopt consistent terminology for data stewardship concepts across climate services
R2. Define multi-dimensional provenance (data, processing, service interpretation)

Finally, the **compliance and standards** group spans the upstream-to-downstream axis, reflecting recommendations related to interoperability, FAIR alignment, and long-term governance. These elements provide a structural backbone that enables technical solutions and interpretative guidance to remain coherent across infrastructures and organisational contexts.

R7. Use open, interoperable metadata standards
R10. Ensure long-term sustainability of provenance information
R13. Anticipate emerging regulatory, FAIR, and Open Science requirements

Rather than representing a strict classification, the positioning illustrates how different recommendations contribute to complementary objectives, helping users of the recommendations to identify which principles are most relevant depending on their role, expertise level, and stage in the climate-service lifecycle.

4.2 Provenance services and EU infrastructure perspectives

Two recently funded EU projects will support the setup of a provenance service infrastructure to serve the climate community. In particular the Horizon Europe FUTURA and ENES-RISe.
They will specifically address:

1) FUTURA: community aspects like the definition of a blueprint architecture for provenance management as well as the collection of community needs to support major climate experiments like CMIP7 and CORDEX;
2) ENES-RISe: infrastructural and operational aspects related to services that will address climate community needs at European level and more internationally within the Earth System Grid Federation.

In particular, ENES-RISe will specifically support the operation of the yProv ecosystem. At its very core the setup hosted at the University of Trento will consist of the yProvStore (to manage provenance documents), a PID service (to issue unique identifiers for each document), and the yProvExplorer (to deliver a visualization framework). Sections 5.1 and 5.2 provide more detailed information about the yProv ecosystem.

4.3 Traceability Challenges and Research Directions.

The interviews and institutional analyses indicate that advancing provenance and traceability in climate services requires addressing a number of targeted research gaps, which are less related to the absence of technical standards and more to their practical interpretation, governance, and usability across diverse operational contexts. A first priority concerns the need for shared terminology and stronger conceptual alignment as reflected in **R1** and **R2**, which promote a consistent vocabulary for data stewardship concepts and a multi-dimensional understanding of provenance covering data origins, processing workflows, modelling components, and the varied interpretative contexts associated with different stakeholder groups served by climate services.

A second gap relates to operational implementation across heterogeneous infrastructures. Providers reported difficulties in managing versioning, intermediate data, and scalable provenance capture in complex workflows, informing **R3**, **R4**, **R8**, and **R9**, which promote persistent identifiers, pragmatic storage strategies, and automated lineage capture embedded within operational pipelines.

Bridging technical traceability with interpretative guidance emerges as a key priority. Interviews with institutional and private-sector stakeholders highlight that users benefit from accessible explanations of assumptions, uncertainties, and methodological choices alongside technical metadata. This perspective is reflected in **R5** and **R6**, which promote layered provenance and user and domain oriented contextualisation tailored to different stakeholder groups, opening ways for further research.

Although earlier literature, particularly in the context of scientific workflow systems and scientific databases [10][11][12], has explored related concepts, there is now a clear opportunity to advance these approaches pragmatically by engaging sectoral communities and supporting their adoption and evaluation in operational settings.

Long-term sustainability and governance represent another critical area where further work is needed. Interviews consistently show that provenance practices risk becoming fragmented when they depend on short-term projects or ad hoc documentation. This challenge is addressed through **R10** (long-term sustainability of provenance) and **R13** (alignment with Open Science, FAIR, and regulatory developments), which position traceability as an integral component of data stewardship rather than an optional technical feature. The recommendation set also recognises that provenance requirements evolve throughout the lifecycle of a service, informing **R11** (alignment of traceability with publication readiness) as a mechanism to scale expectations over time [4s. The comparison with government data-lineage initiatives such as NORA shows that similar challenges arise beyond the climate services domain, reinforcing the need for long-term stewardship models and interoperable approaches.

Emerging technologies introduce additional challenges. The growing use of AI and machine learning can obscure assumptions, uncertainties, and source attribution, highlighting the need for clear documentation of training data, model behaviour, and validation processes. This reinforces the importance of further research on **R12** (traceability for AI-enabled services), building on work already explored in sectors such as healthcare [13] and more recently in the Skills4EOSC project [8] and the yProv ecosystem [14][9]. In this evolving landscape, the recommendations serve as a stable reference point to ensure that traceability principles remain aligned with transparency, accountability, and trust, while allowing technical standards to evolve in response to emerging needs and technologies.

Building on this perspective, the following section provides an overview of recent technical, as well as Research & Developments updates relevant to provenance and traceability, illustrating how existing frameworks, tools, and implementations are adapting to support these principles in practice.

5. Technical updates on provenance-enabling solutions

This section provides an overview of recent technical developments that support the implementation of provenance and traceability in climate services. Rather than introducing new standards, it highlights how existing models, specifications, and tools are evolving to better operationalise the principles outlined in the recommendations.

5.1 Provenance Exploration Systems

Provenance Exploration Systems (PES) are advanced platforms designed not only to store provenance information, but to explore, query, and analyze it interactively and at scale. Unlike traditional provenance repositories, whose primary goal is persistence of PROV graphs, a PES integrates causal querying, multi-level abstraction, pattern discovery, and visual analytics, transforming provenance into an operational instrument for debugging, auditing, optimization, and decision support. PES architectures are typically grounded in the W3C PROV data model, which relies on three concepts: Entities, Activities, and Agents.

During the interviews, the need for tools to "consume" provenance documents has emerged as a cross requirement jointly with the need for visualization tools able to provide a high-level understanding of such information.

Some work in the area addresses relevant aspects of Provenance Exploration Systems like active provenance and advanced provenance visualization frameworks.

The Active Provenance [10] approach introduces a paradigm shift from passive, post-hoc provenance recording to a model in which provenance becomes an active, runtime component of computational systems. Instead of merely documenting execution history for later inspection, provenance is captured as a continuous event stream that can be queried and analyzed during execution. Complementing this perspective, the Visualization of Data Provenance [15] work focuses on the human-centered challenges of exploring complex provenance graphs. Provenance structures are often large, deeply nested, and difficult to interpret through naïve node-link diagrams. The work analyzes visualization strategies to mitigate cognitive overload, including hierarchical abstraction, semantic grouping, filtering, temporal layering, and interactive exploration techniques. The work emphasizes that provenance visualization must go beyond static diagrams and instead support interactive, task-driven visual analytics.

An example of PES to explore provenance documents is provided by the yProvExplorer[43], see Figure 8 , developed in the context of the yProv [14] research project. More information about the yProv ecosystem is reported in the next section.

[43]yProvExplorer: https://explorer.yprov.disi.unitn.it/

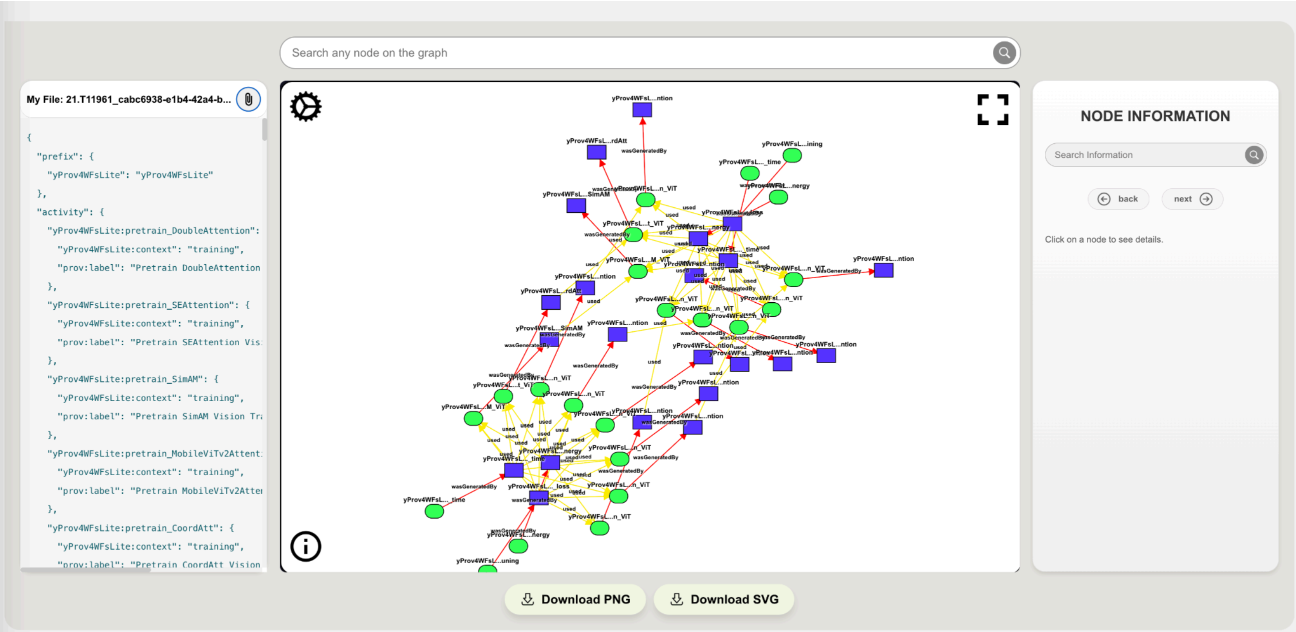

Figure 8: Provenance visualization via the yProvExplorer framework.

5.2 yProv ecosystem and multi-level provenance management

This section deals with two key concepts that emerged during the interviews, namely the need for (i) a structured approach to provenance (ecosystem) and for (ii) different levels of details according to different needs (multi-level support). The two aspects are addressed by the yProv ecosystem which we briefly introduce in the remainder of this section.

The yProv ecosystem is a modular software framework designed to enable comprehensive provenance management for large-scale scientific workflows, particularly in High-Performance Computing (HPC) and AI-driven applications. As presented in [14] yProv addresses the need for scalable, fine-grained, and semantically structured provenance capture in environments where experiments involve distributed infrastructures, heterogeneous hardware (CPUs/GPUs), complex workflows, and AI training processes.

At its core, yProv provides a set of interoperable components that support the entire provenance lifecycle: instrumentation, collection, storage, indexing, querying, and analysis. The ecosystem includes libraries for workflow-level and AI-level instrumentation, services for ingesting and transforming execution traces into structured provenance graphs, and back-end integrations with graph databases for scalable storage and querying. The architecture is designed to operate in HPC, cloud, and distributed environments, minimizing overhead while preserving detailed execution semantics.

A defining feature of the yProv ecosystem is its multi-level provenance management model. Rather than treating provenance as a single flat graph, yProv organizes provenance information across multiple abstraction layers (the finest one is named micro-provenance). This layered approach enables users to navigate seamlessly from fine-grained execution traces to high-level conceptual representations of computational studies. As highlighted in [16] the micro provenance concept can contribute towards (i) stronger traceability, transparency and trust, through a richer and detailed set of metadata; (ii) multidimensional exploration/navigation of provenance metadata (by definition), (iii) knowledge extraction, through data mining techniques and (iv) fine-grained tasks understanding as a means/tool to address the explainability of complex processes (i.e., Machine Learning/Deep Learning).

5.3 RO-Crate for provenance and traceability in climate services

RO-Crate[44] (Research Object Crate) provides a practical framework and format for packaging assets in a single machine-readable package or "crate" (JSON-LD using schema.org) aligned with the FAIR principles. The overarching goal is to provide reproducibility, transparency, and user-friendly communication.

RO-Crate is a valuable asset in terms of structural support for climate services that aim to directly incorporate FAIR principles and notably multi-dimensional provenance in their service. RO-Crate can simultaneously describe data sources, processing workflows, software and models, alongside the contextual information needed for interpretation. A beneficial aspect of RO-Crate is the packaging of machine-readable metadata and human-readable explanations, enabling both technical traceability and relevant communication to users (experts or not). RO-Crate provides climate services with:

- Clear linkage between datasets, models, workflows, and outputs across the climate service value chain.
- Records of versions, software, and processing steps in a structured way.
- Layered provenance in the shape of short summaries for decision-makers or technical lineage for specialists.
- Organic alignment with FAIR, Open Science, and interoperability requirements without restricting the usage to a single domain-specific standard.

As an example of implementation, [17] describes Workflow Run RO-Crate, an extension of RO-Crate to encapsulate multi-layer provenance for process, workflow, and provenance information. Applied to use cases of machine learning for image analysis, the low-level RO-Crates describe the different inference tasks with their input, output, and inference model, while high-level RO-Crates describe the general workflow of the processing pipeline. The granularity of the provenance is encoded by layered provenance profiles to ensure an efficient and global representation of the overarching RO-Crate through several interoperable RO-Crates. This granularity further provides flexibility for the implementation of a RO-Crate object as initial general provenance and workflow information can gradually be refined up to process-level. Furthermore, the fairly basic structure of each of these profiles supports the versatility of the framework and its adaptability to diverse workflow management systems (e.g. Autosubmit).

However, RO-Crate does not compose a complete traceability solution. For example, there is no definition or prescription of controlled vocabularies, and no requirements on the completeness of provenance. No specifications are provided for how assumptions, uncertainties, and trust indicators should be communicated. Furthermore, there is no enforcement of governance, no assessment or certification of regulatory compliance, and no thresholds for the publication-readiness of content.

[44] https://www.researchobject.org/ro-crate/

Overall, RO-Crate supports a lot of provenance-oriented elements, but it is not a tool that enforces compliance in any way. It should therefore be seen as an enabling framework for climate services rather than a prescriptive standard. It provides the pertinent structure to implement and use good provenance and communication practices. Its efficiency depends on how climate services define what represents "good enough" traceability resources at release, publication, and use.

5.4 Autosubmit - Climate Data Workflows System

In essence, Autosubmit[45] is a workflow manager, a software tool designed to automate, coordinate, and monitor a sequence of tasks. In particular, Autosubmit is a high-performance, lightweight workflow manager and meta-scheduler specifically designed to handle the complexities of climate and weather research [18]. Unlike other workflow solutions in the domain, it functions as a self-contained application that integrates experiment management, workflow orchestration, and real-time monitoring.

Autosubmit ensures research transparency by adhering to the **FAIR principles** through several mechanisms:

- **Unique Experiment Persistent Identifiers (PIDs):** Every experiment is assigned a unique experiment ID. This ID serves as a persistent internal identifier that links all generated data, logs, and configurations.
- **RO-Crate Integration:** Research Object Crate (RO-Crate) is a community standard to package research data with their metadata through a lightweight approach. Autosubmit specifically conforms to: Process Run Crate, Workflow Run Crate, and Workflow RO-Crate.
- **Temporal standardisation:** To ensure interoperability, it utilises ISO-8601 standards (Part 1 and 2) for all date and time representations.

Traceability is procured by capturing every "trace" of the experiment's lifecycle, from configuration to execution error handling:

- Version Control & Configuration: If a project uses Git, the system automatically clones the repository inside the experiment's directory. The cloned repository may contain YAML configuration files with settings for models and applications, Autosubmit jobs, as well as the template scripts.
- System Logs and Scripting:
    - **Log Pairs:** Every command generates timestamped .log (standard output) and _err.log (standard error) files.
    - **Job Traces:** Autosubmit archives the final scripts generated from templates, ensuring the exact code executed on the HPC is retrievable.
    - **Status auditing:** The system tracks start/end times and completion status for every task, aggregating this into statistical summaries for performance analysis.

To support long-term academic reproducibility, Autosubmit facilitates the transition from "active experiment" to "published record":

- **Repository Integration:** Experiments archived as RO-Crates can be directly uploaded to Zenodo or WorkflowHub.

[45] https://autosubmit.readthedocs.io/en/master/#

- **Persistent Identification:** Through these platforms, experiments receive Digital Object Identifiers (DOI). These allow researchers to cite the exact state of an Autosubmit experiment in academic papers, enabling other researchers to resolve the reference and inspect the underlying workflow.

6. Conclusions and way forward

The extensive interview campaign conducted for this deliverable represented a major methodological effort and an unprecedented exercise for the task team. Engaging institutional actors, national services, infrastructures, and private-sector providers required careful adaptation to diverse practices, terminology, and varying levels of maturity in how traceability is understood and implemented. This diversity strengthened the analysis by grounding the recommendations in operational realities and highlighting cross-community dialogue as a key enabler for advancing traceability. The effort invested in collecting operational experiences enabled the task to move beyond purely technical discussions and to identify concrete priorities reflecting real user needs, resource constraints, and evolving governance expectations. These priorities also point to further opportunities for development and infrastructure sustainability in upcoming EU initiatives such as FUTURA and ENES-RISe, which will focus on community-driven designs for provenance solutions tailored to climate research.

Overall, the interviews confirm that current challenges are less related to the absence of technical standards and more to their practical interpretation, governance, and usability. Although frameworks such as FAIR, ISO 19115, or W3C PROV already exist, organisations often struggle to translate them into operational data management, decide the appropriate level of detail, and present provenance in ways that different user groups can actually understand and use. Institutional perspectives (IPCC, WMO CDMS, C3S) demonstrate growing technical maturity alongside a need for clearer guidance and harmonised terminology, In parallel, national and infrastructure initiatives (KNMI Climate Scenarios, SMHI dashboards, EUMETNET RODEO) highlight operational constraints related to documentation effort, metadata usability, and long-term stewardship. Private-sector contributions reinforce the need to balance technical traceability with accessible explanations that support decision-making. This is reflected in recommendations that provide strong strategic value by enhancing credibility (R14) and regulatory or quality expectations at the point of release (R11). Adopting layered provenance (R6) further enables companies to maintain detailed internal traceability while exposing only the level of transparency appropriate for clients, helping balance openness with commercial constraints.

The derived recommendations emphasise guiding principles that support both technical reproducibility and interpretative clarity across upstream and downstream users. Their positioning in Section 4.2 provides a practical framework to navigate these needs, highlighting priority areas such as layered provenance, consistent terminology, automated lineage capture, and clearer communication of assumptions and limitations. Rather than proposing new technical standards, the deliverable aims to inform community and discussions by identifying where alignment and coordination are most needed, while reflections from NORA and cloud-platform engineering and GreenOps discussions further underline the growing need for scalable provenance management across contexts.

Looking ahead, the outcomes of this deliverable provide a foundation for advancing standardisation through a balanced pathway that combines shared principles with evolving technical implementation. While the interview-driven recommendations identify priorities for community alignment and governance, Section 5 provides updates on emerging tools that can help operationalise these concepts. Future efforts should focus on coordinated experimentation and collaboration to ensure that standardisation remains flexible, evidence-based, and responsive to both upstream and downstream user needs.

References

[1] A. Spinuso et al., "Preliminary best practices in provenance of processing methodology to guarantee traceability and reproducibility of climate data and ways to communicate it in a user-friendly manner," Zenodo, 2024. Available: https://doi.org/10.5281/zenodo.14036054

[2] M. Stockhause et al., "Implementing FAIR data principles in the IPCC seventh assessment cycle: Lessons learned and future prospects," PLOS Climate, vol. 3, no. 12, 2024. Available: https://doi.org/10.1371/journal.pclm.0000533

[3] World Meteorological Organization, Climate Data Management System Specifications (WMO-No. 1131), 2025. Available: https://doi.org/10.59327/WMO/CDMS/1131

[4] G. Peng, J. L. Privette, E. J. Kearns, N. A. Ritchey, and S. Ansari, "A unified framework for measuring stewardship practices applied to digital environmental datasets," Data Science Journal, vol. 13, pp. 231–253, 2015. Available: https://doi.org/10.2481/dsj.14-049

[5] P. Missier, K. Belhajjame, and J. Cheney, "The W3C PROV family of specifications for modelling provenance metadata," in Proc. 16th Int. Conf. Extending Database Technology (EDBT), Mar. 2013, pp. 773–776. Available: https://doi.org/10.1145/2452376.2452478

[6] K. van der Wiel et al., "KNMI'23 climate scenarios for the Netherlands: Storyline scenarios of regional climate change," Earth's Future, vol. 12, e2023EF003983, 2024. Available: https://doi.org/10.1029/2023EF003983

[7] E. Cimadevilla et al., "The IPCC Interactive Atlas DataLab: Online reusability for regional climate change assessment," PLOS Climate, vol. 4, no. 6, e0000644, 2025. Available: https://doi.org/10.1371/journal.pclm.0000644

[8] E. Osmenaj, C. Sharma, U. Moschini, L. Berberi, and V. Pasquale, "FAIR principles implementation in ML/AI - Findings from Skills4EOSC Delphi Study," International Journal of Digital Curation, vol. 19, p. 6, 2025. Available: https://doi.org/10.2218/ijdc.v19i1.1085

[9] G. Padovani, V. Anantharaj, and S. Fiore, "Provenance tracking in large-scale machine learning systems," in Proc. ICPP Workshops '25: 54th Int. Conf. Parallel Processing Workshops, New York, NY, USA, 2025, pp. 167–174. Available: https://doi.org/10.1145/3750720.3757295

[10] A. Spinuso, M. Atkinson, and F. Magnoni, "Active provenance for data-intensive workflows: Engaging users and developers," in Proc. 2019 15th Int. Conf. eScience (eScience), San Diego, CA, USA, 2019, pp. 560–569. Available: https://doi.org/10.1109/eScience.2019.00077

[11] D. de Oliveira, V. Silva, and M. Mattoso, "How much domain data should be in provenance databases?" in Proc. 7th USENIX Conf. Theory and Practice of Provenance (TaPP), 2015, p. 9. https://www.usenix.org/conference/tapp15/workshop-program/presentation/de-oliveira

[12] Y. Gil et al., "Mind your metadata: Exploiting semantics for configuration, adaptation, and provenance in scientific workflows," in Lecture Notes in Computer Science, vol. 7032, LNCS, Part 2, pp. 65–80, 2011. Available: https://doi.org/10.1007/978-3-642-25093-4_5

[13] D. Babushkina and A. Votsis, "Epistemo-ethical constraints on AI-human decision making for diagnostic purposes," Ethics and Information Technology, vol. 24, no. 2, pp. 1–15, 2022. Available: https://doi.org/10.1007/s10676-022-09629-y

[14] G. Padovani et al., "A software ecosystem for multi-level provenance management in large-scale scientific workflows for AI applications," in Proc. SC '24 Workshops of the Int. Conf. High Performance Computing, Network, Storage, and Analysis (SC-W '24), 2024, pp. 2024–2031. Available: https://doi.org/10.1109/SCW63240.2024.00253

[15] A. Spinuso, M. Atkinson, and F. Magnoni, "S-ProvFlow: Storing and exploring lineage data as a service," Data Intelligence, vol. 4, no. 2, pp. 226–242, 2022. Available: https://doi.org/10.1162/dint_a_00128

[16] S. Fiore, M. Rampazzo, D. Elia, L. Sacco, F. Antonio, and P. Nassisi, "A graph data model-based micro-provenance approach for multi-level provenance exploration in end-to-end climate workflows," in Proc. 2023 IEEE Int. Conf. Big Data (BigData), Sorrento, Italy, 2023, pp. 3332–3339. Available: https://doi.org/10.1109/BigData59044.2023.10386983

[17] S. Leo et al., "Recording provenance of workflow runs with RO-Crate," PLoS ONE, vol. 19, no. 9, e0309210, 2024. Available: https://doi.org/10.1371/journal.pone.0309210

[18] D. Manubens-Gil, J. Vegas-Regidor, C. Prodhomme, O. Mula-Valls, and F. J. Doblas-Reyes, "Seamless management of ensemble climate prediction experiments on HPC platforms," in Proc. 2016 Int. Conf. High Performance Computing & Simulation (HPCS), Innsbruck, Austria, 2016, pp. 895–900. Available: https://doi.org/10.1109/HPCSim.2016.7568429

Appendix 1 - Interview Questions

Interviews questions from WP2 Task 2 - Traceability in Climate Services

1 Introduction. Scope and Expectations of the interview

This interview explores the role of traceability in climate services, focusing on why it matters, its priority level, the barriers affecting how it is perceived and applied, and the challenges faced by both providers and users. We will discuss motivations for adopting traceability measures, practical experiences with standards and tools, and how users interpret and benefit from traceability information. The aim is not to evaluate specific organizations, but to collect insights and perspectives that can guide future best practices, common requirements, and support for more transparent and trustworthy climate services.

GDPR requires interviewee's consent, see template consent_form_2025_05_20_xds.docx

2 Interview Questions

2.1 Contextual information about the interviewee

- **Personal Context** – Background and connection to climate services
- **Service Context** – Description, purpose, and role in the service.

- **Participants & Users** – Service participants developing, maintaining, or delivering the service and final users.
- **Data & Workflow** – Data handled, value chain, and stakeholder expectations (data, guidance, metadata).

2.2 General questions on the relevance and understanding of traceability practices.

- What are the **main reasons that should motivate a climate service provider (like yours)** to include **traceability measures** in their data and production processes?
    - Are these measures generally seen as **a priority that's worth investing in**?
        - If not, what are the blockers?

- Would it be **beneficial** to agree on a set of **minimum requirements** that an operational climate service should meet in terms of traceability?
    - If yes, could you think of a short list of **minimum requirements/key components** to guarantee reliable reuse of data and derived products? *e.g., provenance chain, metadata completeness, versioning, uncertainty annotation.*
    - Do you think there's still a need to **reach consensus on terminology?** For instance, is the **difference between Provenance and Data Lineage** clear to you?

- **As a Climate Service using data** from Climate Data Centers (e.g., C3S, ESGF, National Portals), are there traceability details or other requirements concerning transparency and reproducibility that you wish were available to help make your products and assessments more reliable?

2.3 Implementation: methods and challenges

- Have you **implemented** traceability measures in your CS?

- If yes, could you provide details? What are the impediments otherwise?
- Have you received **feedback from external audits or peers/stakeholders** on your traceability practices (or the lacking of)?
- Are you aware or use **existing standards, principles, recommendations, tools?**
    - FAIR principles
    - ISO 19115,
    - W3C PROV and/or climate dedicated extensions.
    - Handle system, Global Handle Registry
    - DOI foundation, DOI standard
    - DOI registration agencies (datacite, crossref)

- If yes, **do you find them suitable?**
- Do you think there is enough information on how to implement them? Is there a need for **more training or specialised personnel?**

2.4 Interpretation and uptake of traceability Information by the users of Climate Services

- Can you think of any example, or report one, where a **lack of traceability, or unclear guidance** on how to access and interpret it, caused problems for stakeholders trying to use the data?
- On a more positive note, any real-world examples where **downstream users (including your organisation) benefitted from good traceability?**

- How do you think users should be **supported in interpreting/using** this information in their work?
- Looking ahead, which **traceability measures would help users trust AI-generated** climate information, especially when AI-generated recommendations **contribute to decision-making processes?**
    - In your opinion, which **AI tools in the field of climate services** should be required to include traceability measures as a standard?