IS-ENES3 White Paper on provenance handling in the model evaluation process (Version 1)

Bouwe Andela, Joaquin Bedia, Antonio Cofino, Rémi Kazeroni, C. Pagé, Daniel San Martin, Stéphane Sénési, Jérôme Servonnat, Alessandro Spinuso, Mats Veldhuizen, et al.

IS-ENES3 White Paper

Framework for the propagation of documentation and provenance information in the model evaluation process

Bouwe Andela (NLeSC), Joaquin Bedia (University of Cantabria), Antonio Cofiño (CSIC), Rémi Kazeroni (DLR), Christian Pagé (CERFACS), Daniel San Martín (Predictia Intelligent Data Solutions), Stéphane Sénési (IPSL), Jérôme Servonnat (IPSL), Alessandro Spinuso (KNMI), Mats Veldhuizen (KNMI), Klaus Zimmermann (SMHI)

This white paper was elaborated in the context of IS-ENES3 WP3 task 3.2, and contributes to the deliverable D3.3 "Standards Synthesis".

December 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Table of contents

Framework Summary

Documenting provenance across the model evaluation process is an instrumental step for instance in activities where decision making is built on such results, such as for the IPCC Assessment Reports. It allows to build confidence in evaluation results, by ensuring traceability regarding the processing steps, their inputs and their computing context. It is a also a valuable cornerstone for results reproducibility

However, provenance is a very general concept which can be instantiated in extremely various ways, and with numerous tools, even in the restricted context of model evaluation process. Selecting a Web standard such as the PROV family of specifications is of course a necessary step but this alone still allows for semantically non-compatible implementations of provenance handling. The first sensitive item in this respect is the vocabulary, or ontology, to use; this is clearly demonstrated by e.g. the number of entries in the directory of vocabularies and ontologies in Earth, Space and Environmental Sciences maintained by the Research Data Alliance's relevant group. Another key issue is providing tools for instrumenting diagnostic code provided by climate experts with provenance without putting significant additional constraints on such code.

In the context of IS-ENES3 WP3 Task 3.2, besides implementing provenance handling in ESMValTool, handling provenance in model evaluation has been jointly addressed during a series of discussions

As a result, the proposed framework for the propagation of documentation and provenance information in the model evaluation process is based on Web standards and includes : a general setup exemplified by ESMValTool's implementation for provenance handling, an ontology based on METACLIP, the optional use of provenance templates as experimented in Climate4Impact, and possibly the services of the PROV API and the METACLIP viewer. These elements are further described hereafter, together with explanations about how ESMValTool and icclim could take advantage of such tools.

1. Recording of the provenance in ESMValTool

The handling of the provenance framework in ESMValTool has been revised for the release of the version 2.0 (Righi et al., 2020) and continuously improved in subsequent releases. Provenance information is stored in the W3C PROV format and collected at run time. The provenance recording is done for each diagnostic script used in a recipe and saved to disk once it has finished running.

The scientific provenance information recorded includes items listed in the documentation section of the recipe as well as items from the diagnostic scripts. The recipe's documentation[1] contains several tags that are recorded in the provenance such as the authors of the recipe, its title and description, and a list of scientific references in BibTeX format. For each diagnostic section of a recipe, additional tags such as realms, list of high-level components (atmosphere, ocean, sea ice, etc.), and themes (aerosols, clouds, physics, etc.) can be defined and recorded in the provenance. Another set of tags can be used inside each implemented diagnostic script to record scientific provenance, such as a list of authors, a list of BibTeX references, a caption text for each plot generated, the types of plot (error bar plot, histogram, time series, etc.) and statistics used (anomaly, climatology, correlation, mean, etc.), and a list of domains regarding the spatial coverage of the datasets considered (mid-latitudes, regional, tropics, etc.). The complete list of tags for recipes and diagnostic scripts can be found in the config-references file[2] of ESMValTool.

In addition to the scientific provenance items, the tool also collects the names and global NetCDF attributes of all input files used to create a result.

For each output file produced in a recipe run, for example, plots and netCDF files, a provenance file is generated in W3C PROV-XML format containing the data information and scientific information related to the produced output files.

ESMValTool provides support to process the recorded .xml provenance files via a utility script[3] based on the prov library.

The ESMValTool documentation[4] contains all the necessary information for contributors to record the scientific provenance of their recipes and diagnostics. In particular, guidelines and examples are provided for all four types of supported open-source programming languages for diagnostic scripts: Python, NCL, R, and Julia.
While provenance was originally embedded into output files in ESMValTool 2.0, this feature has been discontinued in version 2.4[5] . The reason is that provenance information of large ESMValTool recipes accounts for a significant fraction of the output file attributes and data size,

1 https://docs.esmvaltool.org/projects/esmvalcore/en/latest/recipe/overview.html#recipe-section-documentation
2 https://github.com/ESMValGroup/ESMValTool/blob/main/esmvaltool/config-references.yml
3 https://docs.esmvaltool.org/en/latest/utils.html#extracting-a-list-of-input-files-from-the-provenance
4 https://docs.esmvaltool.org/en/latest/community/diagnostic.html#recording-provenance
5 https://github.com/ESMValGroup/ESMValTool/releases/tag/v2.4.0

making those difficult to read by the user or exceeding the maximum length supported by certain file formats such as png. The plotting of provenance information in an SVG file for human inspection has similarly also been disabled in version 2.4 since the typically very large graphs were not usable in practice and their generation was a source of runtime errors.

A possible enhancement of the current provenance recording framework of the ESMValTool could be to link the output files to the online documentation[6] available for each recipe in the ESMValTool documentation. This could be done by adding a URL to the recipe documentation in the metadata of the plots and attributes of the netCDF files.

2. Recording of the provenance in icclim

The icclim[7] open-source software package, developed by CERFACS, calculates standard climate indices and outputs CF-compliant files. Complete documentation is available on readthedocs[8]. In the current 4.x version, provenance information is basic and it currently uses the NetCDF file metadata global attribute history to record basic provenance information. The icclim processing step is appended to the existing history global metadata attribute. If the history attribute does not exist, it is created.

An example of icclim provenance information is shown below. It is appended in the history metadata global attribute of the output file containing the results of the calculation of the SU climate indice:
"2021-05-11 15:46:04 Calculation of SU indice (summer time series) from 1980-1-1 to 2019-12-31."
It is non-machine readable information, intended only for human reading.

3. METACLIP's ontology for climate products

METACLIP (METAdata for CLImate Products) is a language-independent framework envisaged to tackle the problem of climate product provenance description, offering a solution for identifying, extracting, linking and assembling the pieces of information needed to fully describe the provenance of a climate product. METACLIP also includes a tool (METACLIP Viewer[9]) allowing for interactive medatada discovery and visualization. The current version (Bedia et al. 2019) allows basic navigation through the metadata graph. A new version with enhanced visualization modes is under development. The METACLIP framework builds upon semantics exploiting the web standard Resource Description Framework (RDF, RDF Working Group 2014), building on domain-specific extensions of standard vocabularies (e.g., PROV-O, Provenance Working Group 2013a) describing the different aspects involved in climate product generation (see e.g. Fig. 2). The METACLIP ontologies (and the rest of its components, including the web interpreter) are open source and available in GitHub (https://github.com/metaclip). The project is open to collaboration and contributions to increase the value of the vocabularies to the community.

6 https://docs.esmvaltool.org/en/latest/recipes/
7 https://github.com/cerfacs-globc/icclim
8 https://icclim.readthedocs.io/
9 A prototype of the METACLIP Viewer can be accessed at www.metaclip.org

The METACLIP approach was initially developed in the project QA4Seas (Manubens et al., 2017), framed in the Copernicus Climate Change Service (C3S, https://climate.copernicus.eu). Since the beginning, it proved a successful approach for dealing with seasonal forecast product provenance description (there are seasonal forecast examples in the METACLIP gallery at http://www.metaclip.org), responding to specific project's needs such as linking low-level processing steps with calibration/verification activities and with known community datasets and organizations. These needs are common to other projects and research contexts which require tools for climate product provenance tracking, such as the more recent IPCC AR6 Interactive Atlas (Iturbide et al. 2021), that has also adopted METACLIP for provenance representation of the main products delivered (maps of ECVs and climate indices under different future climate scenarios, Fig. 2). The granularity of the approach allows for a very high level schematic representation of provenance, aimed at providing a general overview of product generation to a broad audience, as well as a low-level technical provenance description that in this case can be downloaded for each specific visualization as a JSON-LD file. In this case, the visualization is possible through the METACLIP Viewer, although the viewer is not yet tailored to such large provenance descriptions entailing complex RDF graphs with possibly thousands of nodes and arcs.

Therefore METACLIP has a broad scope within climate science, and is envisaged to adequately describe scientific outcomes in various related disciplines (short, medium and long range predictions, climate change projections, observational studies ...). A number of climate science-related vocabularies have been developed to this aim, conceived as domain-specific extensions of the agnostic PROV data model (Provenance Working Group, 2013b) and other standard vocabularies. For instance, METACLIP relies on the verification framework defined in the COST Action VALUE (Maraun et al. 2015), establishing a European Network for a comprehensive validation and development of statistical downscaling methods, which has currently a continuity under the CORDEX-ESD activities. In this initiative, a vocabulary has been developed in order to describe statistically downscaled regional climate projections, leading to the development of a expert-community inspired vocabulary that represents a taxonomy of the main bias correction and downscaling techniques currently in use by the community (Bedia et al. 2018). A brief overview of the METACLIP vocabularies is provided in Table 1. Both datasource and ipcc_terms vocabularies, that deal with the primary data sources, have specific annotation properties linking their own features with the CMIP5, CMIP6 and CORDEX Data Reference Syntax, taking as reference their respective controlled vocabularies.

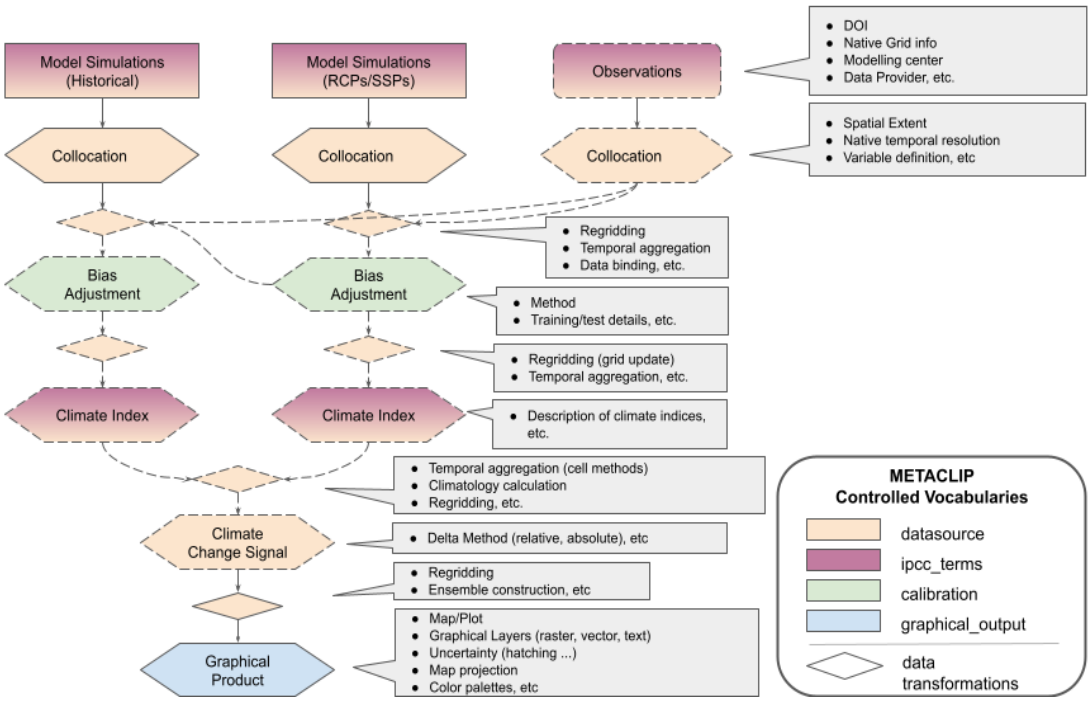

Fig 2. Schematic representation of an Atlas product generation workflow (a Bias adjusted Climate index Delta change similar to the map displayed in Fig. 3). The main steps involve database description, subsetting and data transformation, and finally graphical product generation. METACLIP specifically considers the different intermediate steps consisting of various data transformations, statistical adjustment, climate index calculation and graphical product generation, providing a semantic description of each stage and the different elements involved. The different controlled vocabularies describing each stage are indicated by the colors. The gradient indicates that both ipcc_terms and datasource vocabularies are involved, usually meaning that specific individual instances have been defined in ipcc_terms extending generic classes of the datasource vocabulary (Table 1).

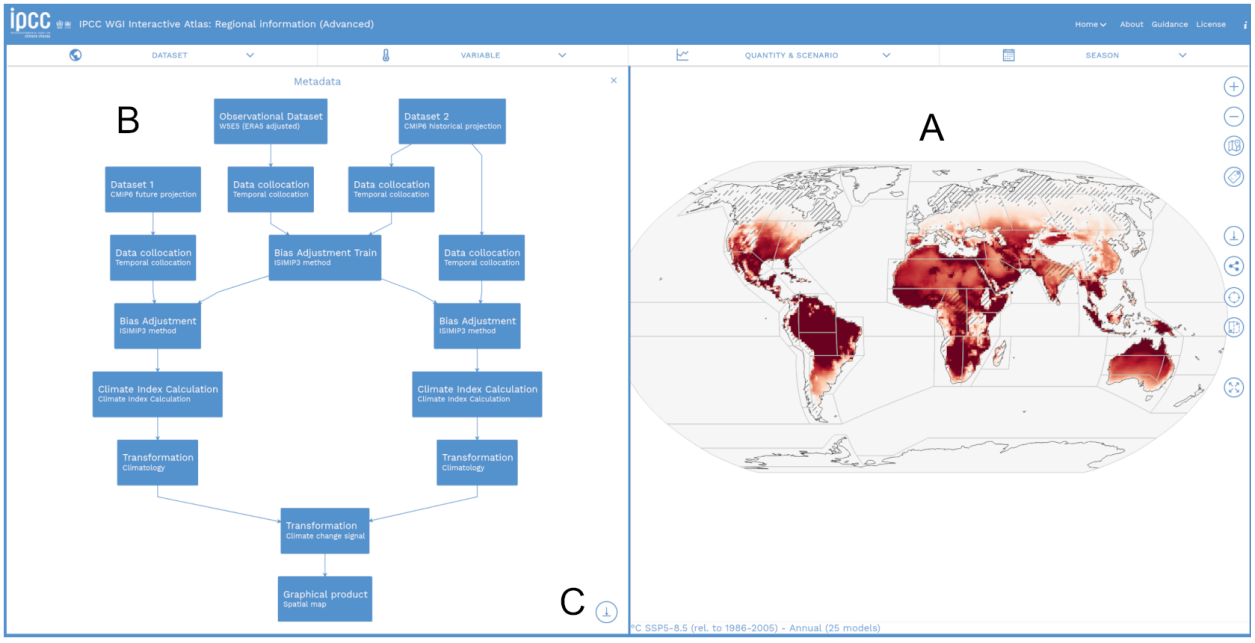

Fig 3. Screenshot of the IPCC AR6 Interactive Atlas with provenance information activated. Right panel (A) depicts an (absolute) delta change map of a bias-adjusted climate index (TX35, i.e., number of days with maximum daily temperature exceeding 35 degC), considering the CMIP6 ensemble simulations (25 models) and the SSP5-8.5 scenario, for the 1986-2005 baseline period and the +2degC global warming level. The left Panel (B), displays a high-level representation of the different steps involved in map creation. Each box contains further metadata for a complete description of the ensemble members (models, datasets, modelling centers, data transformations -regridding, aggregation, climate index calculation, warming level periods for each model, etc.-, as well as a precise definition of the bias adjustment technique employed, calibration period etc...), and the meaning of graphical elements expressing uncertainty, such as hatching. Finally, the download button (C) allows for the obtention of a full provenance description in JSON-LD format.

Name (prefix)	Description	Vocabulary IRI
datasource (ds:)	Describes the origin of the input data and subsequent transformations (subsetting, aggregation, anomalies, PCA, climate indices etc.). It also establishes the links between the different transformation commands and arguments in each step (source code).	http://metaclip.org/datasource.owl
calibration (cal:)	encodes the metadata describing bias correction, downscaling and other forms of statistical adjustment (variance inflation, ensemble recalibration etc.)	http://metaclip.org/calibration.owl
Verification (veri:)	encodes the metadata related with the verification of seasonal forecast products, providing a description of the verification measures applied as well as a description of the verification aspect addressed by each measure. It also includes more general validation procedures, indices and measures, and uncertainty aspects.	http://metaclip.org/verification.owl
graphical_output (go:)	graphical product description (charts, maps), including a characterization of uncertainty types represented and how these are communicated.	http://metaclip.org/graphical_output.owl
Ipcc_terms (ipcc)	provides a definition of individual entities used by the Working Group 1 of the IPCC-AR6, extending some of the classes of the datasource vocabulary (e.g. ds:ClimateIndex, ds:GCM, ds:RCM, ds.ModellingCenter, ds:Variable...)	http://metaclip.org/ipcc_terms.owl

Table 1. Summary of the METACLIP vocabularies

Besides, the use of tags and provenance vocabulary in ESMValTool diagnostics is currently left to the contributors when implementing their code. This results in different levels of detail and accuracy when the provenance of diagnostics is recorded in ESMValTool. A unified framework for provenance vocabulary could be added to the technical standards for including diagnostic scripts into ESMValTool.

Considering the careful design of METACLIP's set of vocabularies and semantics, and its successful uses in the above mentioned projects, the recommended provenance framework includes converging the vocabularies used in ESMValTool, ICCLIM, and METACLIP. Adopting this ontology for the longer term however requires setting up a governance, and some support for it. Its current open source and GitHub setups represent a first step.

Further, visualizing provenance information in the case of real-life documents has proven to be a very difficult issue (see above ESMValTool's feedback). The METACLIP viewer, which allows for dynamic exploration of such graphs, with summary and zooming capability, should be considered for further support.

4. Provenance templates for easy provenance description, and architecture for a service

PROV is a generic provenance model, which can be further extended to represent provenance relations and specific semantics associated with a domain or a computational system. For instance, ProvOne and METACLIP use and extend PROV to represent provenance information associated with generic scientific workflow and provenance information of seasonal forecast verification products, respectively.

In order to generate provenance information by code pieces (such as diagnostic scripts from experts) without having to explicitly handle such information in these codes, one may use the recent specification of the PROV-Template (Moreau, 2017), for the automatic generation of provenance in PROV format. Templates are documents encoded in PROV-N[10] syntax. They describe recurrent provenance patterns, which are instantiated by passing the template and a template's variables binding document to a template-expansion algorithm. Templates facilitate the modelling of re-usable provenance scenarios (generalisation vs tailoring), and remove the burden to hardcode formal provenance editing in the computational tools. This task is delegated to template expansion tools and services, such as the ProvToolbox[11] and the ProvTemplateCatalogue[12], respectively. Thereby, changes to the template might not affect the tool's source code, except in those circumstances when the updated template introduces new variables, which are usually relatively easy to add to the binding document.

This conceptual and technical PROV framework is already adopted in IS-ENES3 (WP7, WP10) to represent and manage the provenance associated with the workspaces of Climate4Impact v2[13]. This is controlled via the SWIRRL API[14]. The provenance management technology used in SWIRRL is further developed in the context of the KNMI EWC (Early Warning Center), aiming at tracing the generation of Warning Products of different types.

The recommended provenance framework includes an optional use of a similar approach, pending on its compatibility with the constraints specific to each component of the IS-ENES CDI and software stack, and pending on the availability of open, reliable and long lasting services for the PROV-template catalog feature and for provenance information storage. Within IS-ENES KNMI owns and/or contributes to the software components represented as PROV-Template Catalogue, Provenance API and Neo4J database and may in the future release these software components to the public domain with a relevant license. This would pave the way to setting up such services in the future IS-ENES CDI.

With respect to icclim, it is planned to fully support the provenance framework concept in the current version 5.x development plan (version 5.0 is currently at beta stage). As icclim is already

10 https://www.w3.org/TR/prov-n/
11 https://lucmoreau.github.io/ProvToolbox/
12 https://github.com/EnvriPlus-PROV/ProvTemplateCatalog
13 https://dev.climate4impact.eu
14 https://zenodo.org/record/4264852#.X6lvNdv_pSw

tightly integrated with Climate4Impact, the implementation of the framework will complete the proper integration of provenance information and leverage icclim. This will contribute to make icclim closer to FAIR software.

Regarding ESMValTool, the scheme could ease provenance information management for contributed diagnostic scripts. This would allow contributors of diagnostics to refine the provenance information granularity down to elementary processing steps by their scripts without explicitly handling provenance structures in their code. ESMValTool development team would then have to work with diagnostic script contributors in building relevant templates (or selecting it from the catalogue) and to allow for the generation of the template's bindings. The ESMValTool core software, ESMValCore, would then set up the communication for the template expansion and storage of the provenance documents via the API. In Figure 4 we show the corresponding architecture of the Provenance and Lineage management services.

However, this would generate a dependency of ESMValTool to external services, which would be a major change with respect to its current status, namely a stand-alone, self contained, system easily installed. Also, for ESMValTool users, exporting provenance information to an external system would require confidentiality management.

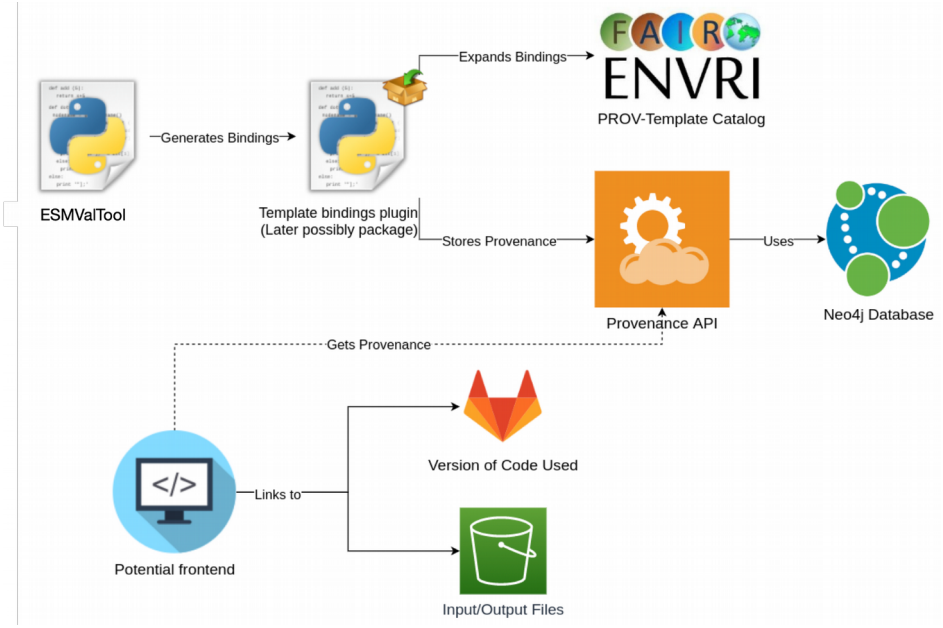

Figure 4. Software Architecture for the provenance and lineage management services.

References

Bedia, J., San-Martín, D., Iturbide, M., Herrera, S., Manzanas, R., Gutiérrez, J.M., 2019. The METACLIP semantic provenance framework for climate products. Environmental Modelling & Software 119, 445–457. https://doi.org/10.1016/j.envsoft.2019.07.0050

Bedia, J., San-Martín, D., Herrera, S., Iturbide, M., Gutiérrez, J.M., 2018. A proposal for a bias correction metadata model in the framework of METACLIP (METAdata for CLImate Products). 2nd Workshop on Bias Correction in Climate Studies, Santander, Spain. URL: http://www.meteo.unican.es/en/node/73439

Iturbide, M. et al., 2021. Repository supporting the implementation of FAIR principles in the IPCC-WGI Atlas. https://doi.org/10.5281/zenodo.3691645

Manubens, N., Hunter, A., Bedia, J., Bretonnière, P.A., Bhend, J., Doblas-Reyes, F.J., 2017. Evaluation and Quality Control for the Copernicus Seasonal Forecast Systems, in: AGU Fall Meeting Abstracts. Presented at the AGU Fall Meeting 2017, American Geophysical Union, New Orleans, Louisiana, USA.

Moreau, Luc, et al. "A templating system to generate provenance." *IEEE Transactions on Software Engineering* 44.2 (2017): 103-121.

Maraun, D., et al., 2015. VALUE: A framework to validate downscaling approaches for climate change studies. Earth's Future 3, 2014EF000259

Provenance Working Group, 2013a. PROV-DM: the PROV Data Model. W3C Recommendation, W3C. URL: https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-provenance.

Provenance Working Group, Apr. 2013b. PROV-O: the PROV Ontology. W3C Recommendation, W3C. URL: https://www.w3.org/TR/2013/REC-prov-o-20130430/

RDF Working Group, Feb. 2014. RDF - Semantic Web Standards. URL: https://www.w3.org/RDF

Righi, M. et al., 2020. Earth System Model Evaluation Tool (ESMValTool) v2.0 - technical overview. Geosci. Model Dev., 13, 1179-1199, https://doi.org/10.5194/gmd-13-1179-2020