INTERGOVERNMENTAL PANEL ON climate change

Task Group on Data Support for Climate Change Assessments (TG-Data)

TG-DATA RECOMMENDATIONS FOR AR7

Version 1.0, 13 October 2023

This guidance document is an agreed product of the IPCC Task Group on Data Support for Climate Change Assessments (TG-Data). The mandate of the TG-Data was approved by the IPCC, but this does not imply the Panel's endorsement or approval of documents emanating from the TG-Data or any recommendations contained herein. Furthermore, this document has not been subject to the procedural IPCC review processes.

Table of Contents

Executive summary **1**
Recommendation on objectives 1
Recommendation on implementation 1
Introduction **2**
Recommendations on AR7 objectives **3**
Data curation 3
Final datasets for SPMs, TSs and all Chapters 3
Intermediate datasets 3
Input datasets 3
Interactive IPCC products 4
Software and scripts 4
Citation and licensing 5
Citation for third-party data and software 5
Licensing for third-party data and software 5
DDC stability 5
Implementation recommendations **6**
AR timeline 6
Author support 6
Author documentation on FAIR practices 6
Software tools for data and metadata preparation 6
Communication regarding common datasets 6
Data support staff 7
TSU 7
Chapter teams 7
Digital interactive products 7
Scoping 8
Staffing 8
Outreach 8
Conclusion **9**
References **9**

Executive summary

The Task Group on Data Support for Climate Change Assessments (TG-Data) works to make the IPCC exemplary in its scientific data handling practices. Based on consultations with all three Working Group (WG) Technical Support Units (TSU) and Data Distribution Centre (DDC) following the sixth Assessment Report (AR6), TG-Data has prepared recommendations for AR7.

Recommendation on objectives

- Aim to ensure final datasets and provenance metadata underlying all report figures are archived either at the DDC or another trustworthy long-term data repository;
- Identify key intermediate datasets with high potential for reuse and ensure their long-term curation;
- Ensure key input datasets are preserved in a trusted long-term archive;
- Ensure interactive IPCC products and the data they generate comply with FAIR (Findable, Accessible, Interoperable, Reusable) principles;
- Prepare guidance material and workshops on the proper use of online interactive IPCC products;
- Ensure authors publish and cite software code and data used in the preparation of the report;
- License code and data produced by IPCC authors to facilitate reuse and the creation of derivative products;
- Ensure DDC and TSUs have resources over the full AR cycle to fulfill the objectives above.

Recommendation on implementation

- Account for code, data and metadata preparation tasks in the AR timeline;
- Prepare documentation for IPCC authors on data preparation best practices, data curation steps they're responsible for, and practical examples of how to prepare data using different tools and programming languages;
- Establish communication channels across chapters and WGs to make sure the same dataset version is used by everyone;
- Ensure TSUs have a data team in place to tailor FAIR data processes and procedures to the WG's needs and requirements, support authors and chapter scientists and liaise with DDCs;
- Embed data stewards in data intensive chapters to support authors and liaise with TSU;
- Plan the development of digital interactive tools early in the AR cycle, including a detailed memorandum of understanding with software developers, and outreach activities on the use of IPCC digital interactive products.

The publication of the first order draft, second order draft and final draft are intense and stressful moments for authors, chapter scientists and TSU staff. Careful planning, efficient tools, clear communications and workflows will be essential to carry out the recommendations above without unduly adding to that toll.

Introduction

The Task Group on Data Support for Climate Change Assessments (TG-Data) has a mandate to *facilitate the availability and use of climate change related data resulting from the activities of the IPCC*[1]. In cooperation with the IPCC Data Distribution Centre (DDC) and Technical Support Units (TSU), it offers guidance to *provide curation, transparency, traceability and stability of data and scenarios related to the reports of the IPCC, and strives to facilitate the availability and consistent use of climate change related data and scenarios in support of the implementation of the work programme of the IPCC*.

Many scientific journals now require authors to publish supporting data and results alongside papers. Experience suggests that storage devices fail, data formats change and authors retire; without proper curation practices, valuable scientific data is eventually lost to history. As the highest scientific authority on climate change and a pillar of national and international policies, IPCC reports are expected to be exemplary in their data curation practices. Making sure IPCC data adheres to the FAIR principles (Findability, Accessibility, Interoperability, Reusability) adds value to the work of authors, facilitates reuse of IPCC results – including for authors of the next cycle, and adds to the credibility of the reports.

Before the Sixth Assessment Report (AR6), data underlying key results of IPCC reports were only available by contacting authors directly. Thanks to the concerted efforts of TSUs, DDC and TG-Data, data for all the figures from the AR6 Summary for Policymakers (SPM) and, to different extents, Technical Summary (TS) of Working Groups (WG) I, II and III are now stored at the DDC and indexed in the DDC catalog. Each catalog entry includes the names of authors, a link to the actual data, the chapter and figure it appears in, a recommended citation, and a Digital Object Identifier (DOI). Data for some individual chapters are also available, as well as key intermediate datasets with potential for reuse, for example on WGI assessed sea level rise. In continuity with previous cycles, climate model simulations, greenhouse gas emissions scenarios and other core input datasets are also curated and stored in long-term archives.

These achievements are impressive considering TG-Data started its activities in 2019, in the middle of the AR6 cycle, and the sheer amount of unplanned work this required from authors, TSU staff and DDC members. The lack of dedicated personnel, the specialized expertise required, and the addition of new tasks in the busy IPCC schedule all contributed to limit what could reasonably be done for AR6.

For AR7, we hope to build on these achievements, go further in the curation of IPCC data, while making the process more efficient and less time consuming for all those involved. Based on lessons learned from the AR6 experience, this document synthesizes TG-Data recommendations to ensure the IPCC is exemplary in its handling of data. It also provides implementation suggestions to streamline the process. We believe that making early decisions on data curation policies will allow authors, TSUs and DDC to better plan resources and time, and deliver the exceptional reports that the world now expects.

The document groups recommendations in two broad categories: objectives for AR7, and implementation suggestions regarding staffing and timelines.

[1] See the TG-Data terms of reference and mandate

Recommendations on AR7 objectives

Data curation

Apply FAIR guidelines as widely as possible in IPCC products

During AR6, the TG-Data, the DDC, the three WGs, and the Synthesis Report, adopted the FAIR data guidelines (Pirani, et al. 2022). The motivation is to increase transparency and accessibility of the assessment, the implementation of the IPCC Error Protocol, and the long-term curation of the assessed digital information that otherwise cannot be traced easily to specific sources. Although limited in what could be achieved at such late notice, hundreds of datasets were archived and cataloged. We believe AR7 can go further and apply FAIR guidelines to most of the material found in the reports.

Final datasets for SPMs, TSs and all Chapters

Apply FAIR guidelines to all final data products produced by IPCC authors

The FAIR principles should be incorporated as a core practice within the assessment process from the start of the AR7. For successful incorporation, TSU staff and the DDC partners need to be strengthened during the assessment process to guarantee adequate support to authors. This would allow for datasets from all chapters and summary products to be published simultaneously with the report. Close TSU-DDC staff collaboration will be required in final days of report preparation.

The incorporation of the FAIR principle in the AR6 summary products is already increasing the transparency and accessibility of the assessment. During the first year after adopting the FAIR data guidelines, the DDC recorded more than 10,000 users accessing archived data. Some cases of intense usage have also been recorded, for example, the dataset for AR6 WGI Figure SPM.2 has been downloaded more than 2,000 times.

Intermediate datasets

Apply FAIR guidelines to key intermediate datasets with high potential for reuse

Intermediate datasets are those datasets that result as an outcome of the authors' assessment that build on the authors' assessment of multiple lines of evidence in the literature to 'constrain' available datasets. They can also be datasets resulting from non-trivial post-processing, in an intermediate step to the final data in the report. There is an expected high reuse potential of these datasets for further research. Examples from AR6 are ocean heat content, global surface temperature, sea level rise projections assessed by WGI, the WGI Interactive Atlas datasets and the WGIII Scenario database.

Such highly relevant intermediate datasets should be identified and prioritized by the TSU as soon as possible during the assessment process, together with the related author team. Datasets should be published during the AR, issued a DOI and properly referenced in the report as well as within the DDC catalog, following IPCC FAIR Guidelines. This would enable continuous traceability of IPCC results, their long-term accessibility and reusability, and availability if they're needed for the IPCC Error Protocol. Downloads are recorded and documented in statistics as part of the DDC Annual Reports.

Input datasets

Ensure key input datasets are preserved in a trusted long-term archive

Input datasets used by authors in their assessment and for creation of report figures are an important component of the traceability of IPCC results back to the underlying research. Some input datasets are not long-term preserved in a trustworthy repository. Therefore the DDC partners have coordinated the transfer of key input datasets with the data owners into the DDC Partners' repositories. Examples from AR6 are datasets from the Coupled Model Intercomparison

Project Phase 6 (CMIP6) and the INFORM Global Risk Index. The long-term preservation keeps IPCC results traceable, accessible and reusable in the long-term. Download and reuse of datasets are recorded and documented in statistics as part of the DDC Annual Reports.

The curation of key input datasets as part of the FAIR Guidelines should be incorporated as core practice within the assessment process from the start of the AR7. An author team should be made responsible for the dataset usage across chapters, facilitate communications with the data provider, ensure the appropriate version is used and that it is properly referenced across IPCC products, and assist the data curation process along with the TSU and DDC. In the case of datasets already archived outside of the DDC, the DDC will record relevant metadata in its catalog.

Interactive IPCC products

Ensure interactive IPCC products and the data they generate follow FAIR principles

Interactive or static products accompanying the AR, such as the WGI Interactive Atlas or the WGII Annex I Global to Regional Atlas, facilitate access and use of datasets that have been assessed by the AR. Their exposure to a wider public makes it especially important that they comply fully with FAIR standards. This implies that static products are recorded in the DDC catalog, and interactive tools export provenance information, including references to input or intermediate data as well as source code. Essentially, all material required to review these IPCC products should be made available, so that they can be reviewed as part of the AR process.

Provide guidance on proper use of interactive tools and data generated by them

In view of their usefulness within and outside of the IPCC process, interactive products should be accompanied by rich documentation and information on their use and interpretation. Also, outreach activities should be planned and delivered at the end of the AR7 cycle, to steer the scientific and practitioner community in the proper use of these interactive tools and data generated by them.

Software and scripts

Require authors to version control and publish code used in the preparation of the report

For some chapters, the production of the AR relies heavily on software code written by authors or chapter scientists. Code is needed to process data, extract statistics, and generate figures, among other uses. The efforts pursued in AR6 for the curation and distribution of this source code highlight the need to outline additional guidelines to foster FAIRness of the research software developed for the IPCC reports (Jiménez et al. 2017; Lamprecht et al. 2020).

All the software (analysis scripts and notebooks, libraries, interactive tools) developed in each WG should be organized in repositories managed by the TSU. Prior to the release of each report draft version, authors should release the code used for the generation of data and graphics, and issue a DOI. TSUs should provide templates to collect metadata on attribution and purpose of the software (e.g. WGI AR6 template). This may be complemented by tools to link the code, data and the figures published in the report (e.g. the IPCC Figure and Data Manager). It is recommended that authors adopt an open source language for the code developed for the report.

Ancillary data (e.g. aggregated dataset, configuration files and parameterization), instrumental to the code execution should be stored in the source code repository. This is especially important when these particular data are not available at the DCC because the mechanisms ensuring their consistent curation and access would add impractical complexity.

Regarding licensing, authors should follow the recommendations provided in the IPCC Data and Code Licensing Guidelines (Huard et al. 2022).

Citation and licensing

The citation of data and software used in a publication in order to give credit to their creators have become a widely accepted practice in scientific publications supported by many organizations, publishers and individuals[2].

Citation for third-party data and software

Include citations for data and software used in the preparation of the report

It is recommended that IPCC reports cite data and software from third parties, e.g. input data or third party software/code, in the body of the report with corresponding references in the reference list. Data and software might have a separate section in the reference list. The reference should include the URL of the software, as indicated by the software authors (when a DOI or a public repository is not available), and the version used. In case different versions of the same software are used across the chapters, the particular release should be mentioned in the text of the report.

Licensing for third-party data and software

Apply IPCC Data and Code licensing guidelines

TG-Data published IPCC Data and Code Licensing Guidelines making specific recommendations on the licensing of input, intermediate and final data, as well as source code. We recommend AR7 follows these recommendations to facilitate access and reuse of IPCC content.

DDC stability

Ensure the IPCC DDC is properly resourced for the full cycle

The services provided by the DDC are currently funded out of the generous contribution of countries and institutions. This ad hoc funding model has been strained during AR6 due to the significant increase in workload asked of DDC members, and a decreasing willingness of long time DDC members to contribute. With heightened expectations regarding data access for AR7, we recommend the IPCC includes data curation among its responsibilities, and dedicates a budget to it for the next cycle. Budgeted activities should minimally cover software infrastructure development & maintenance for data and figure preparation workflows; author training and support; liaison with TSUs and TG-Data; data and metadata review; long-term archival; web catalog development & maintenance; and user outreach. Different levels of funding and services are detailed in IPCC-LVII/INF. 7. Achieving the objectives recommended in this document would require a level of funding between Option 3 (continuity of current service) and Option 4 (optimal service).

Funding for DDC is especially critical in the lull in-between AR6 and AR7, when TG-Data is losing its TSU ex-officio members. Without adequate resources, TG-Data will have little capacity to prepare for the next cycle, and risks carrying its AR6 backlog into AR7.

[2] See the Commitment statement in the Earth, space, and environmental sciences

Implementation recommendations

The following are recommendations meant to help achieve the objectives laid out in the previous section. They are drawn from feedback collected from IPCC authors and TSUs (including Pirani, et al. 2022), DDC managers and TG-Data members.

AR timeline

Integrate data and code preparation timeline into the AR schedule

Data underlying key results and graphics should be considered an integral part of the assessment text and graphics, on par with how citations that have been assessed are provided. This implies that data identification, processing and publication milestones be integrated in the AR timeline to ensure resources are allocated adequately and work spread uniformly over time. An internal management planning timeline should include agreement on standards and workflows (DDC, TSU ,TG-Data), author training (TSU), data preparation (authors, chapter scientists), cross WG handshake mechanisms (e.g. WGII and III relying on WGI data), author feedback following releases (TSU, DDC), archival (DDC), DOI assignment (DDC) and outreach activities (authors, TG-Data, TSU) following the publication of datasets. An overview of this timeline with key milestones needs to be provided to authors at the start of the assessment process as part of the guidance material prepared for authors. As with the report, data preparation should follow staged releases to reach maturity, and undergo external review.

Author support

Author documentation on FAIR practices

Provide authors documentation on data preparation best practices

The time until the start of AR7 should be used to prepare documentation to help authors meet FAIR principles. The documentation should lay out expectations towards authors, tips to work efficiently and links to valuable external resources. References to software tools would be especially valuable to handle datasets heavily used in the AR, such as CMIP data and the emission scenario database. A requirement for such author guidance is a clear understanding between TSUs and DDCs of the data preparation and publishing requirements, metadata conventions, quality-control responsibilities, and communication channels.

Software tools for data and metadata preparation

Prepare practical examples of how to prepare data using different tools and languages

In the AR6 cycle, some chapters used advanced software libraries to streamline the production of graphics and analysis for the report (e.g. ESMValTool, Climate4R, pyam). Coordination with the software developers of these tools would be useful to make sure they support the FAIR guidelines that authors are expected to follow. The repositories found on the WGI Github repository contain different approaches used by authors to prepare data for the report, and could serve as templates for AR7.

Communication regarding common datasets

Establish communication channels to ensure datasets are used consistently across chapters and WGs

Challenges were encountered in AR6 in the consistent use of datasets across the reports. For input datasets (e.g. CMIP6), not all authors systematically kept track of which versions were used to create final data products. Regarding final datasets, another challenge was the transfer of data across WGs. In some cases, those handshakes between WGI and II and WGI and III occurred very late in the timeline.

We recommend establishing systematic communication channels between TSUs, DDCs, authors and data providers. Throughout the assessment process, authors need to be made aware whenever any of the datasets they are working with is updated by the data provider and/or whenever the same dataset is being used by other authors. The Figure Management System (FMS) could be a tool for facilitating these communications and tracking the evolution of datasets.

Data support staff

TSU

Include data team in TSUs

We recommend TSUs include a data team component. This team would be responsible for designing the FAIR data processes and procedures tailored to the Working Group's needs and requirements, in close cooperation with the related DDC Partners and in liaison with the TG-Data. Personnel could even be shared between TSU and DDC Partners. These processes and procedures would facilitate the full compliance of FAIR Data Principles across the entire Working Group. The TSU data team would liaise with chapters' Data Steward (see below). The TSU data team would ensure harmonised level of compliance across the entire Working Group.

The TSU data team would be responsible to provide metadata and data to the related DDC partners in the agreed format, and review data and metadata jointly with the DDC. It would operate the Figure Management System (FMS), a key tool for figure preparation and communication (see recommendation no.12).

The data team should include staff with a background in computer science, or a closely related field (e.g. data science or statistics). It is essential to have demonstrable experience in handling data, metadata and software, adopting standards and interoperable solutions. The data team could thus support chapters with no Data Steward.

Chapter teams

Embed data steward in the data intensive chapters

We recommend creating a new chapter-level role called Chapter Data Steward, which could be assumed by Chapter Scientists. The aim of this role is to support the author team in following the TG-Data processes as set out by the TSUs and the TG-Data and implemented and carried out by the TSU and the DDC. These TG-Data processes are designed to ensure the full curation of the underlying and input data of each figure in each chapter as well as code scripts when applicable. The Chapter Data Steward role could also include data provision to authors in an accessible format, as long as this falls within the Steward's domain of expertise. The Chapter Data Steward is responsible for liaising between the TSU data team, and the author team to ensure full compliance of the TG-Data processes within the chapter. Some data-intensive chapters may require more than one Chapter Data Steward, and others such as the introductory chapters may not need one at all. The level of need is subject to consultation between the TSU and the CLAs at the beginning of the cycle, noting that it may evolve during the assessment cycle. For this position we suggest a profile with a background close to the domain, to help bridging communication and terminology gaps between the authors and the TSU data team. However the appointed Chapter Data Steward should have demonstrable experience in using technological solutions to handle data and metadata, as well as analytical software.

Digital interactive products

Digital products can add substantial value to the AR, for example by displaying more material with greater levels of details than static publications, or by allowing stakeholders to focus on data covering specific regions. Given the success of the WGI Interactive Atlas, we encourage the IPCC to explore frameworks to continue such activities in AR7 for all WGs.

Scoping

Plan the development of digital interactive tools early in the AR cycle

Creating such products however requires advance planning, considerable resources during development, and continued maintenance beyond the publication date. In particular, the content of these interactive products should be scoped with the report outline, and be under the responsibility of an author team large enough to ensure redundancy. This implies that relevant experience is considered at the author selection stage, as well as in the TSU staffing plan and budget. For example, WGII' "Appendix 1: Regional to Global Atlas" did not have dedicated authors, making it considerably challenging to produce.

Staffing

Sign detailed MoU with developers

In AR6, digital products have so far been developed using two mechanisms. The WGI Interactive Atlas has been developed as an official IPCC product, meaning it had an author team and was branded as an IPCC product through a memorandum of understanding. In contrast, the NASA Sea level explorer leverages IPCC intermediate data developed in the course of the AR to build an external application whose development, operation and maintenance is not tied to the IPCC.

In both cases, the development of the product will most likely rely on a team of software developers external to the IPCC. It is thus important to clarify from the onset how these software teams will interact with the author team, the TSU and the Secretariat, set clear expectations, and make sure constraints specific to the IPCC are well understood. An MoU between the software development team and the IPCC should describe the development timeline and milestones, intellectual property and data licensing agreements, how the IPCC logo can be used, and describe an end-of-maintenance procedure for eventual ownership transfer.

The data displayed should comply with FAIR principles, and be included in the DDC data catalog. Software meaningful to the data analysis should be made public and ideally undergo review.

Outreach

Plan outreach activities on the use of IPCC digital interactive products

Interactive products can help communicate the outcome of the report by customizing the content to the audience, either by region, topic or time horizon. In AR6, outreach activities organized around the WGI Interactive Atlas WGI and WGIII Scenario Explorer presented results from the report, explained how to access data, and engaged with regional research practitioner communities. For AR7, we recommend that these outreach activities are planned by TSUs early in the cycle, involve DDC members to connect the products to the DDC data catalog, and are broadcasted to focal points and other IPCC stakeholders.

Conclusion

One of the purposes of TG-Data is to facilitate the availability and use of climate change related data and scenarios that result from IPCC activities. This was partially achieved in the AR6 cycle, although at the cost of considerable unplanned efforts from authors, TSU and DDC staff. Indeed, while the feedback on the outcome has been unanimously positive, the additional workload for authors, chapter scientists, TSU and DDC staff took a toll that we should take care to avoid in AR7.

Given the positive feedback received, a recommendation from TG-Data is to widen the scope of the IPCC data and code covered by the FAIR principles. This recommendation is consistent with broad trends in scientific publishing, and with the intent of the IPCC to be exemplary in its scientific practice.

The amount of work involved to achieve this is however too great to be improvised during the AR cycle. IPCC data and code curation need to be planned at every stage, from the scoping of the report to its publication, and accounted for in the author selection process, TSU staffing and Secretariat IT resources. Training material for authors needs to be prepared early in the cycle, IT tools need to be deployed to streamline data preparation and curation, and milestones included in the AR timeline to smooth the workload over time. This can only be achieved with the active support and contribution of WG co-chairs, and through a collaboration between authors, TSUs, DDCs and the IPCC Secretariat.

References

Huard, David, Anna Pirani, Robert Chen, José Manuel Gutiérrez, Martin Juckes, Volker Krey, Alessandro Spinuso, and Martina Stockhause. 2022. 'IPCC Data and Code Licensing Guidelines'. Zenodo. https://doi.org/10.5281/zenodo.7431834.

Jiménez, Rafael C., Mateusz Kuzak, Monther Alhamdoosh, Michelle Barker, Bérénice Batut, Mikael Borg, Salvador Capella-Gutierrez, et al. 2017. 'Four Simple Recommendations to Encourage Best Practices in Research Software'. F1000Research 6. https://doi.org/10.12688/f1000research.11407.1.

Lamprecht, Anna-Lena, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, et al. 2020. 'Towards FAIR Principles for Research Software'. Data Science 3 (1): 37–59. https://doi.org/10.3233/DS-190026.

Pirani, Anna, Andrés Alegria, Alaa Al Khourdajie, Wawan Gunawan, José Manuel Gutiérrez, Kirstin Holsman, David Huard, et al. 2022. The Implementation of FAIR Data Principles in the IPCC AR6 Assessment Process. https://doi.org/10.5281/zenodo.6504469.

Pirani, Anna, Diego Cammarano, Ellie Fisher, Beate Krüss, Robin Matthews, Charlotte Pascoe, Lina Sitz, and Martina Stockhause. 2022. 'Experience in the Implementation of FAIR Data Principles in the WGI AR6 Assessment'. Zenodo. https://doi.org/10.5281/zenodo.6992173.