

AR6 Working Group I FAIR Supplementary Material

Compiled by Anna Pirani,
12 April 2022

WGI-SM1: IPCC Server-side Author Workspaces	2
WGI-SM2: IPCC Metadata Form to archive data	2
Table 1: IPCC Metadata Form including descriptive explanation of what information is required.	3
Table 2: An example based on a dataset from the AR5 WGII report	5
WGI-SM3: Applying FAIR data principles in data files	6
Data on a spatial mesh	6
Time series data	7
WGI-SM4: Figure Documentation	8
WGI-SM4.1 Figure caption guidelines	8
WGI-SM4.2 Chapter Data Tables	10
Template	10
TSU Guidance on completing data tables	10
Example completed table	12
Notes on completing each column	14
WGI-SM4.3 Model Metadata files	17
WGI-SM5: Recommended documentation for code	19
WGI-SM6 Step-by-Step Process for Archiving Data	21

WGI-SM1: IPCC Server-side Author Workspaces

The IPCC DDC has made the following resources available for authors.

CEDA [JASMIN](#):

- Apply to become a JASMIN user at this [link](#). Once your petition is accepted, you can sign in [here](#)
- You will need to generate an SSH key pair (e.g. using the “ssh-keygen” command: see http://bit.ly/JASMIN_keygen) and paste the public key into the application form
- Apply for access to JASMIN services: http://bit.ly/JASMIN_get_login
- Apply for access to DDC disk cache: http://bit.ly/CEDAddc_disk
- Apply for access to DDC server: http://bit.ly/CEDAddc_get_server
- [User Guide to the CEDA Interactive Server](#):
https://cedadev.github.io/ipcc_ddc/Documents/IPCC_DDC_CEDA_Interactive_Server_UserGuide_v1-1.pdf

DKRZ [Mistral](#):

- Create an account at <https://luv.dkrz.de/projects/newuser>
- After you confirm your email address sign in at <https://luv.dkrz.de>
- Join the group ‘1088: DKRZ_MIP_POOL_Analysis’ (your petition will be sent to the administration team). Once your petition is accepted, a new username will be sent to you by email.
- Documentation of DKRZ’s Virtual Workspace:
http://bit.ly/IPCC_DKRZ_Virtual_Workspace

IFCA [IPCC-Hub](#):

- Email your GitHub username to Maialen Iturbide (miturbide@ifca.unican.es) so she can give you access to the platform, which is available via a web browser at <http://hub.ipcc.ifca.es>. You will then be able to sign in by using your GitHub credentials.
- This environment runs based on Jupyter Notebooks on [IFCA cloud](#) resources, building on the [climate4R](#) R framework for data access and post-processing. Worked examples (R notebook which creates some of the figures included in the FOD version of the Atlas chapter) is available in the user space. This notebook may serve as a starting point for your own analysis (see also this [collection of R notebooks](#) for further examples on how to use climate4R).

WGI-SM2: IPCC Metadata Form to archive data

The IPCC Metadata Form is designed to collect the information needed by the DDC to catalogue and curate data, and to do this with minimal effort. Table 1 describes what is required, Table 2 provides an example from AR5. For archival of final data for WGI (e.g., plotted figure data), metadata is collected via the WGI Figure Manager system.

Table 1: IPCC Metadata Form including descriptive explanation of what information is required.

Label	Description of desired information from authors.	
1. Hierarchy Link	A reference to the upper hierarchy (sorting level) the dataset belongs to. These entries need to be present before dataset metadata can be entered.	Project/Experiment, or IPCC chapter
2. Title	A concise description that distinguishes this from other datasets. [Title for data reference]	<i>Title</i>
<i>2.1 Catalogue Long Name</i>	<i>Name to appear in portal search result list</i>	<i>Entry Name</i>
3. Abstract	Description for data with aspects 2a-2f	<i>Summary [Experiment]</i>
3a: What	Describe what it is in terms of variables archived.	
3b: When	Specify the time period covered, and any temporal averaging or other temporal statistics used, including the calendar reference.	temporal coverage
3c: Who	Include name, institution, email address and (if available) web address (home page or ORCID page) for lead author, plus name and institution for others involved. [optional: provide additional contacts with other roles like funder, research Group, author etc.; ORCID recommended for persons]	contact (investigator)/person/person_externalid/institute [Experiment]
3d: Why	What are the objectives of this dataset?	

3e: Where	Specify the region covered and any spatial averaging or other spatial statistics used.	spatial coverage
3f: How	Describe the procedures used to produce the data: Figure caption to catalogue records, and provided links to parts of the report with additional. info.	
4. Findable Descriptors		
4a. Short title	A short and unique title, e.g. "ar6wg1_f5-3b_precip"	entry acronym
4b. Keywords	A set of words to help users find the data. The words used should be chosen consistently across datasets within a working group.	keywords
4c. Reference	Add the reference for the stored data: authors, title	standard citation [Experiment]
4d. License	Add a license statement for the data	
5. Data References/Links	Provide links for (1) Documentation, (2) Source data, (3) Software, (4) Related data or cite these resources (providing authors, title, publisher, publication year in addition to the access link) The relation to the archived data should be defined with a standard, e.g. DataCite metadata standard, http://schema.datacite.org/	Reference/Citation [Experiment]
6. File format	Specify the file format	Format [Dataset]

7. Variables	<p>For each variable, list one or more names. If possible provide a name from a well-known standard (e.g. CF standard name for physical climate variables -- if you have a physical climate variable which does not currently have a CF standard name we can try to register one).</p> <p>We usually collect short_name, long_name (recommendation for CF Standardname), unit</p>	Parameter/Topic/Unit [Dataset]
---------------------	---	--------------------------------

Table 2: An example based on a dataset from the AR5 WGII report

Label	Information from author
1. Hierarchy Link	AR5 WGII report, Chapter 21
2. Title	IPCC AR5 Seasonal Mean Temperature and Precipitation in IPCC Regions for CMIP3 and CMIP5
3. Abstract	
3a: What	Projected regional average change in seasonal and annual mean temperature and precipitation for the IPCC SREX regions for CMIP5 and CMIP3
3a: When	Seasonal and annual mean temperature and precipitation for the period 2071-2100 are compared to a baseline of 1961-1990 for CMIP5 and CMIP3 General Circulation Model (GCM) projections.
3c: Who	Carol McSweeney, UK Met. Office, enquiries@metoffice.gov.uk , https://www.metoffice.gov.uk/research/people/carol-mcsweeney
3d: Why	The data compare the range of projections from 35 Coupled Model Intercomparison Project Phase 5 (CMIP5) ensemble members under four Representative Concentration Pathway (RCP) scenarios compared with GCM projections from 22 CMIP3 ensemble under three Special Report on Emission Scenarios (SRES) scenarios.
3e: Where	Regional averages are based on SREX regions defined by the IPCC Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (IPCC, 2012: also known as "SREX").

3. How	Regional averages over specified regions.
4. Findable descriptors	
4a: Short title	ar5wg2_fsm21-1
4b: Keywords	IPCC, AR5, IPCC-DDC, regional, CMIP, CMIP3, CMIP5, SREX, RCP25, RCP45, RCP60, RCP85, A1B, A2, B1
4c: Links	https://www.ipcc.ch/site/assets/uploads/2018/02/WGIIAR5-Chap21_FINAL.pdf Hewitson, B.C., A.C. Janetos, T.R. Carter, F. Giorgi, R.G. Jones, W.-T. Kwon, L.O. Mearns, E.L.F. Schipper, and M. van Aalst, 2014: Regional context. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Barros, V.R., C.B. Field, D.J. Dokken, M.D. Mastrandrea, K.J. Mach, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.
4d:	
5. Data references/ Links	
6. File Formats	CSV
7. Variables	long_name: Precipitation [mm/day], standard_name: precipitation_flux

WGI-SM3: Applying FAIR data principles in data files

A user will be disappointed if, having found your data in the archive, downloaded the files and set to work, they find the files full of impenetrable jargon and code words. Ideally, we would like information to be accessible in a format which is recognised by common software packages, though this may not be possible in all cases.

Data on a spatial mesh

When data represents values on a grid of latitude and longitude points, we strongly recommend the use of NetCDF files and compliance with the CF Conventions (see <http://cfconventions.org/>). The CF Conventions provide a huge range of options: the key to implementing it cleanly is to start with the basic requirements. In this case, we want to identify the data variable and the temporal and spatial coordinates.

```

double time(time);
    time:long_name = "time";
    time:units = "days since 1998-04-19 06:00:00";
    time:calendar = "standard";
float lon(lon) ;
    lon:standard_name = "longitude";
    lon:long_name = "longitude of grid cell centre";
    lon:units = "degrees_east";
float lat(lat) ;
    lat:standard_name = "latitude";
    lat:long_name = "latitude of grid cell centre" ;
    lat:units = "degrees_north";

```

BOX: Illustration of metadata needed to specify time, longitude and latitude coordinates for data on a simple grid.

1. The attributes “standard_name” and “units” should be set on all spatial coordinates, and the value of the “standard_name” must come from the approved vocabulary list on the CF Conventions web pages. The “units” must be physically consistent with the canonical units of the “standard_name” (e.g. a “height” coordinate can have units “m” or “km”, but not “hPa”).
2. The “long_name” attribute may be used to give some additional information about the coordinate.
3. The time “units” attribute specifies both a unit of measure and a reference time. The reference time is a string of the form “YYYY-MM-DD hh:mm:ss” giving year, month, day, hour, minute and second. The second value may be decimal. SI time units are recommended. Values of “months” or “years” are technically valid within the CF Convention, but are likely to cause confusion and should be avoided (when used in the time “units” attribute, “years” refers to the mean orbital period of the Earth around the sun, not calendar years, and “months” refers to one twelfth of this year, not a calendar month).

The time axis should carry a “calendar” attribute. For data from climate models this may be an artificial calendar used in the model simulations, but for observations it should be the standard calendar (i.e., a mixed Gregorian/Julian calendar).

Time series data

Time series data (such as annual global temperature) can be provided as NetCDF files, but it may be more convenient to provide spreadsheets. In order to avoid delays and confusion, we want to work with files which are convenient for the chapter scientists to work with. See the [guide to archiving data at CEDA](#), including instructions on creating badc-csv data files (which contain FAIR supporting metadata).

WGI-SM4: Figure Documentation

WGI-SM4.1 Figure caption guidelines

These best practice guidelines should be followed for **figure captions** in the report.

- The first line of the caption should be written in the form of a short title. The aim is to clarify upfront what the main message of the figure is, it helps readability and facilitates subsequent figure data management, indexing etc.
- Write out acronyms and **abbreviations** in full. Once an acronym is defined, it can be used later in the caption.
- **Variable name** and **units** should be clearly stated for **each** variable in the figure.
 - Observed OR simulated/modeled/projected should be stated.
 - Specifiers should be included: ocean/land/air, surface/bottom, global/regional, etc
 - Non-standard naming should be avoided, e.g., warming is not a variable name, temperature is a variable name. It can be the trend in temperature, or temperature anomaly that defines the warming.
- **Time period** and **spatial coverage** should be clearly stated: from when to when and where?, mean over what time/space?, anomaly with respect to what?
- **Type of analysis** on the variable should be stated: Mean, Anomaly, Trend, Bias, etc.
- All **visual cues** (colours, symbols, points, arrows, etc.) on the figure should be explained.
- Ensure **accurate wording** is used to describe what is plotted, e.g:
 - **Plots over time.** *Wording in past reports: change, evolution, response, variability, variation, time variation, trend, signal, patterns of...*

‘Variability’ is misleading to describe a plot over time as it is a calculated statistical parameter (statistical dispersion). Use ‘Time variation of X’, ‘Temporal variation of X’, ‘X over time’, or just ‘X over 1970–2015’ instead.
 - **Trends** (e.g., trend lines). *Wording in past reports: change, rate, rate of change*

The word ‘change’ is used throughout the captions to describe trends and plots over time. Change is not the same as trend. Change can be the difference between any two values while trend is a calculated statistical parameter. Use

‘trend’ if that’s what’s meant. In other cases the word ‘change’ can be used as long as the calculation method is also stated to avoid confusion (e.g., ‘temperature change, expressed as an average anomaly’).

- **Anomaly plots** (e.g., anomaly time-series, global anomaly maps). *Wording in past reports: change, rate of change, response, signal, departure from baseline, ...*

Use ‘anomaly’ if that’s what’s meant.

- Not essential, but type of graph is very useful to state: Temporal, Spatial, Zonal/Meridional, Vertical etc. distribution, histogram, X vs Y (scatter/line), etc.
- For traceability, the **source (input) datasets** used should be clearly stated in the caption and their citations provided and appearing in the chapter reference list. Use in-text citations of the form ‘Jones et al. (2013)’ (inserted using Mendeley for WGI) and not direct weblinks. Provide the citations to the datasets themselves where these are available, else provide citations for related webpages (e.g., metadata pages) or publications. The latter can also be provided in addition to the dataset citations. Figure code can also be cited. In all cases, the reference entries should contain a DOI (or URL if no DOI). Figure code will be hosted on the IPCC GitHub, with the corresponding citations and DOIs generated using Zenodo. If it is a very long list of source datasets (e.g., a multi-model ensemble), the reader should be directed to the chapter’s Supplementary material where these are documented in a table, for instance: ‘See Table 1.SM.1 for source datasets’. Source code used to generate the figure can also be cited in this table. A template for such a dataset table is provided in section A4.2 below. Information on CMIP and CORDEX datasets can be documented in ‘model metadata files’ (text files with metadata designed by the WGI TSU). See section A4.3 on Model Metadata files below for further information.
- Any methods of calculation or choice of a specific analysis applied to the data (e.g., filter described in Somebody et al. is used or only values exceeding X are taken into account) should be stated. For trends, both the calculation method and temporal resolution of the data used should be stated. If too long, there should be a reference to the text where the method is explained (e.g., in a chapter Appendix).
- Where spread is plotted (e.g., for a multi-model ensemble) and this is not the standard 5-95% percentile range, the range used should be stated.
- No Discussion / Interpretation:
 - No interpretation in the caption. For instance, don’t put “Importance of X”, “Sensitivity of X to Y is high” unless a parameter called importance or sensitivity was calculated and plotted against something, as this is subjective.
 - Discuss in the text instead: “This figure shows that the increase in X can be explained by the decrease in Y” should go in the text.

WGI-SM4.2 Chapter Data Tables

Source datasets and code for figures/tables/calculations can be documented in a table in a separate section of each chapter's Supplementary Material. The template for this documentation table is provided below, along with a detailed explanation of how to complete it.

Template

X.SM.X: Data Table

Table X.SM.1: Input Data Table. Input datasets and code used to create chapter figures.

Figure number / Table number / Chapter section (for calculations)	Dataset / Code name	Type	Filename / Specificities	License type	Dataset / Code citation	Dataset / Code URL	Related publications / Software used	Notes
Figure X.1								
Figure X.2								

TSU Guidance on completing data tables

The input dataset/code tables will go at the end of the Supplementary Material for each chapter and should contain a complete list of all the source datasets used to generate each data-driven figure and table. Source data used to calculate key values in report text can also be documented. Typically only observational/reanalysis data should be manually entered in the tables - the TSU will generate rows for model data (e.g., CMIP5, CMIP6, CORDEX) from the model metadata files provided by chapters. If you only have a few model datasets, you

can enter information on these manually, rather than providing model metadata files. See the guidance under ‘Manually documenting model datasets’ below.

If the input datasets used for a figure/table are identical to those used for another figure/table in the report, the table can just state ‘See data table for [Figure X.X or Table X.X]’ in the ‘Dataset / Code name’ column, with the other columns left blank. Feel free to provide this template to CAs who worked on figures so they can add information.

Code used to create the report figures and tables can also be documented. For instance, plotting or data processing routines developed by CLAs/LAs/CAs. If the code is run in a climate research software (e.g., Climate4R, ESMValTool), this can be cited in the ‘Related publications’ column. Code from external sources can also be documented in a table row. Don’t put entire software (e.g., Climate4R, ESMValTool) in its own row though – the code should be more specific.

If a figure has subpanels, they can be documented in turn:

Figure number / Table number / Chapter section (for calculations)	Dataset / Code name	Type	Filename / Specificities	License type	Dataset / Code citation	Dataset / Code URL	Related publications / Software used	Notes
Figure X.1, panel a								
Figure X.1, panel b								

Example completed table

Figure number / Table number / Chapter section (for calculations)	Dataset / Code name	Type	Filename / Specificities	License type	Dataset / Code citation	Dataset / Code URL	Related publications / Software used	Notes
Figure X.1	Hadley Centre Sea Surface Temperature v3 (HadSST3) [Name in full, abbreviation and version number. If no version number, put the accessed date in the 'Dataset / code URL' column]	Input dataset	HadSST.3.1.1.0_annual_global_ts.txt	CC BY-SA 4.0 [License as stated by the data provider]	[If available, put the citation specific to the dataset. Leave blank if none - don't put publications here]	https://www.metoffice.gov.uk/hadobs/hadsst3 [If the dataset doesn't have a version number add the date the webpage was accessed in this column, e.g., '(accessed 29/10/2020)']	Kennedy et al. (2011a, 2011b) [Should be Mendeley links]	
	ESA Sea Surface Temperature Climate Change Initiative (SST_cci): GHRST Multi-Product ensemble (GMPE), v2.0	Input dataset	1981–2016, monthly	CEDA license: http://licences.ceda.ac.uk/image/data_access_condition/esacci_sst_terms_and_conditions.pdf	Good (2020) [Should be Mendeley links]	https://catalogue.ceda.ac.uk/uuid/a963d9415bb74247830f8704f825aa90	Merchant et al., (2019) [Should be Mendeley links]	

	Figure 1.1 Code	Code	ch1_fig1_ plotting_ code.r (plotting routine) [Original plotting code for the figure. State the purpose of the code in brackets]	[Leave blank if original code - to be completed by the TSU after archival]	[Leave blank if original code - to be completed by the TSU after archival]	https://github.com/authorgithub/Ch1 [Links to private repositories will only be used for code archival by the TSU]	Climate4R: Iturbide et al. (2019) [Can cite own publications for which the code was originally developed. Can also cite the research software the code is run in, like Climate4R]	
		Code	processin g_code.r (data processin g routine) [Code used for data processin g. Don't put an entire software like ESMValT ool]		Yelekci (2018) [Citation for the code]	https://github.com/externalgithub/routines [Webpage for the code]	Matthews and Yelekci (2018) [Can cite own or external publications the code was developed for]	

Manually documenting model datasets

If you want to manually enter info on model datasets (rather than submitting model metadata files to the TSU), put the short form names of the model(s) and the model experiment(s) in the 'Dataset / Code name' column (e.g., 'BCC-CSM2-MR: hist-aer, hist-ghg, historical'). As usual, also provide the dataset citation (if available), the URL where the data can be downloaded and related publications in the appropriate columns.

Notes on completing each column

Figure number / Table number / Chapter section (for calculations)

For figures and tables, complete as shown in the above example tables. For computed values in chapter text (which is what 'Chapter section (for calculations)' refers to), provide the section number and the title of the computation in brackets.

Dataset / Code name

Datasets:

Provide the name in full, along with the version number and the acronym/abbreviation. If not already included in the name, you can put the data provider at the end in brackets, e.g., '(NOAA)'. For observational datasets the name should be consistent with that used in Annex I: Observations. If the input data used for a figure/table is identical to that used for another figure/table in the report, you can just state 'See data table for [Figure X.X or Table X.X]' in the 'Dataset / Code name' column and leave the other columns blank.

Example: 'Hadley Centre Sea Surface Temperature v3 (HadSST3)'

Code:

Give a description of the code.

Example: 'Figure 1.1 Code'. Put something more specific if necessary (e.g., 'Figure 1.1 temporal averaging code')

Type

Datasets:

Put 'Input dataset'

Code:

Put 'Code'

Filename / Specificities

Datasets:

Put the name(s) of the data file(s) that were used (i.e., those downloaded from the data website). In case there's too many filenames to list, you can just put a text summary like '1981–2016, monthly'.

Example: HadSST.3.1.1.0_annual_globe_ts.txt

When listing data files, you can provide specificities like which columns were used from the data files. Also, if not already in the filenames, you can also state which temporal and spatial versions of the dataset were used, e.g. 'Annual data', 'Monthly data', 'JJA average', '1000 and 500 hPa'.

Code:

Put the name(s) of the code file(s) and state the purpose of the code in brackets (i.e., is it for data processing or plotting?).

Example: ch1_fig1_code.m (plotting routine)

License type

Datasets:

Provide the license type of the input dataset if specified by the data provider. This will help the TSU determine if we can make related plotted figure data publically available.

Example: CC BY-SA 4.0

Code:

If original code leave blank – will be added by the TSU. If external code, put the license given by the code provider (if specified).

Dataset / Code citation

Datasets:

If a separate citation is available for the data, add it here in the form of a Mendeley link. The Mendeley entry should include the names of the creators or data provider and a DOI or URL. Often datasets don't have their own citation, in which case leave the field blank. Don't put publications in this column - they should go in the 'Related publications' column.

Example of a dataset citation: Good, S.A. (2020): ESA Sea Surface Temperature Climate Change Initiative (SST_cci): GHR SST Multi-Product ensemble (GMPE), v2.0. Centre for Environmental Data Analysis, 05 August 2020.
doi:10.5285/a963d9415bb74247830f8704f825aa90.

In the table, this would appear as the Mendeley link 'Good, 2020'

Code:

Leave blank if it's original code you want the TSU to archive - the TSU will add the citation when they archive the code. Else put the citation for the external code (if available) as a Mendeley link.

Dataset / Code URL

Datasets:

Provide the webpage where the data can be downloaded from, or at least, the webpage with the metadata for the dataset. If no version number is provided, put the date you accessed the data after the link.

Example: <https://www.metoffice.gov.uk/hadobs/somedataset> (accessed 29/10/2020)

Code:

If the code is stored online put the webpage here. Links to private author repositories (e.g., on GitHub) won't be put in the report - they are just to help the TSU with code archival.

Example: <https://github.com/authorgithub/Ch1>

Related publications

Datasets:

Provide the citations for related journal articles and other scientific documents (e.g., dataset documentation) here in the form of a Mendeley links. Publications already cited in the corresponding figure or table caption can simply be dragged and dropped into the table.

Example: Kennedy et al. (2011a, 2011b)

Code:

If there are publications relating to the code, add them here as Mendeley links. If applicable, cite the publication for which the code was originally developed (by IPCC authors in the case of original code, or by external researchers for external code). Any climate research software used to generate the figure (for data processing, plotting etc.) can be cited under 'Related publications' (e.g., ESMValTool, Climate4R). No need to cite R, Python or Matlab though.

Example: Climate4R: Iturbide et al. (2019)

Notes column

Add any additional information you think would be useful in this column. For instance, you could add info on data processing such as conversions undertaken to bring data from multiple sources together (e.g., stating the value used to shift climate variables relative to the 1850-1900 reference period). To keep the tables short, such information should be concise – more detailed information should go elsewhere (e.g., Supplementary Material).

WGI-SM4.3 Model Metadata files

Model metadata files are text files designed by the WGI TSU for documenting the data from CORDEX, CMIP5 and CMIP6 used in a particular figure

- Create one file for each model ensemble in a figure (different format for CORDEX, CMIP5, CMIP6)
- List which subpanels the datasets are used for in the subpanel column
- How to name the files:
Fig1-1_md (for Figure 1.1)
CCB1-1_Fig1_md (for Cross-Chapter Box 1.1, Figure 1)
Box1-1_Fig1_md (for Box 1.1, Figure 1)
FAQ1-1_Fig1_md (for FAQ1.1, Figure 1)

FigTS-1_md (for Figure TS.1)

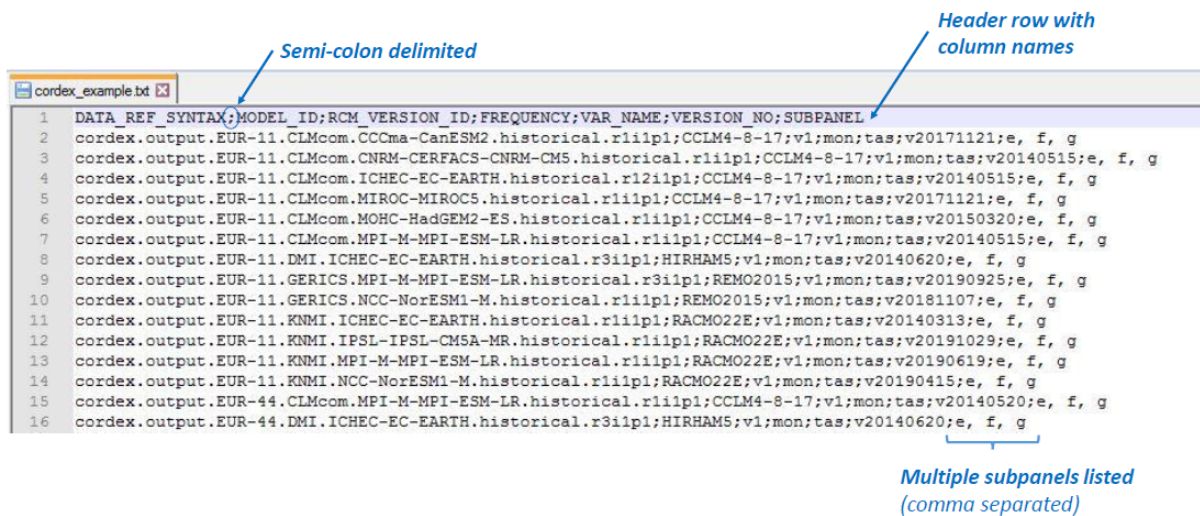
BoxTS-1_Fig1_md (for Box TS.1, Figure 1)

FigSPM-1_md (for Figure SPM.1)

Add ‘_cordex’, ‘_cmip5’, ‘_cmip6’ to the end of the filename to indicate the ensemble
‘_md’ stands for metadata

The following slides outline the different formats of these files for CMIP5, CMIP6 and CORDEX metadata.

What they look like



Columns in the CMIP6 metadata files

- 1) Data reference syntax to experiment level (i.e., mip_era.activity_id.institution_id.source_id.experiment_id) [DATA_REF_SYNTAX]
e.g., CMIP6.CMIP.CNRM-CERFACS.CNRM-CM6-1.1pctCO2
- 2) Sub-experiment ID (sub_experiment_id) [SUB_EX_ID]
e.g., s1960 (will be 'none' in many cases)
- 3) Ensemble member (variant_label) [ENS_MEMBER]
e.g., r1i1p1f2
- 4) Table ID (table_id) [TABLE_ID]
e.g., Amon
- 5) Variable name (variable_id) [VAR_NAME]
e.g., pr
- 6) Grid label (grid_label) [GRID_LABEL]
e.g., gn
- 7) Version number on ESGF [VERSION_NO] *Version number for ESGF dataset. Not the same as NetCDF file creation_date*
e.g., v20180626
- 8) OPTIONAL: Handle in NetCDF file (tracking_id) [HANDLE] *NetCDF file handle (not ESGF dataset handle)*
e.g., hdl:21.14100/f445d60e-6afc-4c0c-8efc-9221be2b315d
- 9) Subpanel [SUBPANEL]. (e.g., 'a', 'b'). Can leave column blank if no subpanels.

All attributes can be extracted from NetCDF files except
version numbers, which must be looked up on the ESGF
(netcdf_file_attribute) [NAME_FOR_METADATA_FILE_HEADER]

Columns in the CMIP5 metadata files

- 1) **Data reference syntax to experiment level** (i.e., project_id.product.institute_id.model_id.experiment_id) [DATA_REF_SYNTAX]
e.g., cmip5.output1.CCCma.CanESM2.1pctCO2
- 2) **Frequency** (frequency) [FREQUENCY]
e.g., day
- 3) **Modeling realm** (modeling_realm) [MODELING_REALM]
e.g., atmos
- 4) **MIP table** (table_id) [TABLE_ID]
e.g., cfDay
- 5) **Ensemble member** [ENS_MEMBER]
e.g., r1i1p1
- 6) **Version number on ESGF** [VERSION_NO]. Not the same thing as a NetCDF file creation_date.
e.g., v20130114
- 7) **Variable name** [VAR_NAME]
e.g., albiscpp
- 8) **OPTIONAL: Handle in NetCDF file** (tracking_id) [HANDLE].
- 9) **Subpanel** [SUBPANEL]. e.g., 'a', 'b'. Can leave column blank if no subpanels.

Columns in the CORDEX metadata files

- 1) **Data reference syntax to ensemble member** (includes regional domain and driving GCM) [DATA_REF_SYNTAX]
 project_id.product.CORDEX_domain.institute_id.driving_model_id.driving_experiment_name.driving_model_ensemble_member

Institution of the RCM
CMIP5 GCM
CMIP5 experiment
Ensemble from the CMIP5 GCM

 e.g., CORDEX.output.EUR-11.CLMcom.CCCma-CanESM2.historical.r1i1p1
- 2) **RCM model name** (model_id) [MODEL_ID]
e.g., CCLM4-8-17
- 3) **RCM version ID** (rcm_version_id) [RCM_VERSION_ID]
e.g., v1
- 4) **Frequency** (frequency) [FREQUENCY]
e.g., mon
- 5) **Variable name** [VAR_NAME]
e.g., tas
- 6) **Version number on ESGF** [VERSION_NO]. Not the same thing as a NetCDF file creation_date.
e.g., v20150609
- 7) **OPTIONAL: Handle in NetCDF file** (tracking_id) [HANDLE].
- 8) **Subpanel** [SUBPANEL]. e.g., 'a', 'b'. Can leave column blank if no subpanels.

WGI-SM5: Recommended documentation for code

The **documentation and content that code repositories should minimally contain** is provided here:

- A README including references to supporting material (datasets, libraries, algorithms, etc).

- A file describing the software environment (libraries and their versions), for example a conda “environment.yml” file.
- Code and scripts (whenever possible, authors are encouraged to use free software tools).
- Please, follow the basic good practices for writing code: be clean, assign descriptive names to variables, document your scripts as much as possible.
- A test suite (ideally).
- Final output (generated figures/tables/values).
- Please try to organize your repo in a clear, clean way. For instance, there might be one folder for each figure (“fig_1”).
- Final and analysis data (if reasonably small; let’s say < 1 Gb).
- If the analysis uses local data, provide a file mapping the local data to publicly available data sources (e.g., in Figshare). Otherwise, at least provide a link to the data provider.
- Example of a possible directory structure:
 - README (.txt, .rst, .md)
 - environment.yml (software environment description)
 - src/ (analysis and utilities code)
 - scripts/
 - § fig_1.py (running this script creates fig_1,)
 - § fig_2.R
 - § ...
 - figs/
 - § fig_1 (.eps, .pdf)
 - § ...
 - tables/
 - § table_1 (.csv, .xlsx)
 - stats/
 - § ?
 - data/
 - § input/
 - Small files (under version control)
 - Large files
 - README (detailed source information on data)
 - § tmp/
 - § output

WGI-SM6 Step-by-Step Process for Archiving Data

IPCC Working Group I with CEDA
Robin Matthews, Lina Sitz, Working Group I TSU
v0.8, March 22nd 2022

N.B. The yellow highlights indicate who does that task.

Data Archival Process

We have created the following spreadsheets used as communication tool to work in collaboration:

1. TSU internal usage: [Summary dataset and code - AR6 - WGI](#). It gathers all the information about the status of every figure, issues and contact authors.
2. TSU-CEDA usage: [AR6 Figures CEDA Status](#). It contains information about the status of the dataset on CEDA (in arrivals, in processing, in archive, published, DOI), links to the catalogue records and links to code.

TSU does initial check of provided data and metadata

1. Go to the 'Data and Code archival' folder on the DMS (AR6 instance):
<https://wg1.ipcc.ch/apps/dm/filemanager.php?q=4429>. Go into the data and code folders for each chapter ('Chapter 01', 'Chapter 02' etc.) and identify figure folders recently 'Updated/Created'. Such folders should contain files added by authors (data or code files depending on the location). If you click on the file names, you can see who last updated each file ('Last Updated By').
2. For each figure folder with files, document this on the corresponding Dataset page on Figure Manager as follows:
 - Tick the 'Data on DMS' check box.
 - Note that the 'Archival Status' dropdown refers to data archival only (there's a separate dropdown for code archival). Check if metadata has been provided for the Figure. If none has been provided, tick 'Missing metadata' under 'Archival Sub Status'.
 - For code, under the Code Archival section, set the 'Code Archival' dropdown to 'Yes' and tick the 'Code on DMS' check box.
3. Classify files, adding this information to the TSU internal usage spreadsheet:
 - a. final data: Data used to create the figure (these datasets have priority in the archival process).
 - b. input data: original data used (external data - e.g., a global temperature dataset) without related publication (second in the archival priority process)
 - c. intermediate data: data that's been processed by the authors - something in between the input data and what's shown on the plot (final data).

- d. data coming from a publication Datasets that already have their own citation. We don't need to archive these datasets. We only need to cite them in the catalogue records if there will be an associated dataset (based on those data but after some processing) archived on CEDA.
 - e. code.
4. Identify newly completed Figure Manager dataset pages. To see which dataset pages have been updated, click on 'Metadata updated by author' in the right hand panel. If no data files have been provided on the DMS for a figure, tick the box 'No datafile(s)' under 'Archival Sub Status'.

TSU check files and metadata

1. Check completed metadata on Figure Manager and adjust as needed:
 - a) Adjust text and structure to standardize the Abstracts
 - b) Check for List of data provided
 - c) Check for Data provided in relation to figure
 - d) Check source of additional info: Check if there is information related to the figure in the input data table.
 - e) Other relevant online documentation or links added manually: Check URLs and DOIs and convert to weblink format if not already in that format. For instance, you may need to append <https://doi.org/> to the start of DOIs.
 - f) Check keywords
2. Check provided data and readme files, and adjust them as needed. Use Check-in/Check-out on Figure Manager when updating files.
 - 2.1. Do a spell check of each file.
 - 2.2. Check file names, replace spaces and symbols by underscore (" _").
 - 2.3. Check files structures using the following test:

BADC-CSV files

- Run the BADC-CSV file checker:

<http://archive.ceda.ac.uk/cgi-bin/badccsv/csvcchecker>

NetCDF files

- Check if you can open with Panoply.
- Check if there is information about (complete the missing information):
 - a. Author name
 - b. Date of creation
 - c. Units
 - d. Missing value
 - e. Variable names

Files with formats different from **BADC-CSV** and **NetCDF**:

- Always recommend to use non-proprietary format files, but also to choose formats common to your field (to ensure interoperability and reusability).
- Request an additional readme file containing at least the following information (we have used a [readme template](#) in order to facilitate the way of providing this information):
 - a. Author name
 - b. Date of creation
 - c. Units
 - d. Missing value
 - e. Variable names

Note: for all the file formats: compare variable names with labels and figure caption. If names are different, adjust.

2.4. Check data values:

- a) See if values are consistent with plot (reproducing the plot)
- b) Check data format (if needed change to dot-decimal notation)

2.5 Create a general readme file: if there are several readme files for the various dataset unify information, add missing information.

3. **If an issue is identified in step 1 or 2**, Add a comment in the TSU internal usage spreadsheet. Set the archival status to '**Issue with TSU Check [TSU]**'. Tick the appropriate box(es) that appear and add a '**Note on Current status**'.

The screenshot shows a form with three main sections:

- Archival status**: A dropdown menu with 'Issue with TSU Check [TSU]' selected. Below it, the text 'To be completed by the TSU' is visible.
- Archival Sub Status**: A list of checkboxes:
 - ☐ Issue with data format
 - ☐ Issue with provided metadata
 - ☐ Missing datafile(s)
 - ☐ Missing metadata
 - ☐ Missing input data Information
 - ☐ No datafile(s)
- Note on Current status**: A text area with a rich text editor toolbar (bold, italic, link, unlink, bulleted list, numbered list, indent, outdent, undo, redo, source) and a 'Format' dropdown. The text area is currently empty.

At the bottom right of the form, there is a small '0'.

4. Export a YAML file from Figure Manager.
5. Upload the YAML files, data, license and readme files to CEDA. Go to <https://arrivals.ceda.ac.uk>. Click Choose existing directory? Once complete, set the

archival status to '**Data/metadata given to CEDA [TSU]**'. Add a comment in the TSU internal usage spreadsheet.

CEDA check files and metadata

6. Conduct standard CEDA checks on the provided YAML files, data and readme files
7. **If an issue is identified in step 6**, send a message to TSU outlining the issues for the TSU to follow up.¹

TSU followup on issues flagged by CEDA

8. **If an issue is identified in step 7**, contact the appropriate persons to fix the issue. Add a comment in the TSU internal usage spreadsheet.²

CEDA create Catalogue Record³

9. Create the CEDA Catalogue Record. Add the link in the TSU-CEDA usage spreadsheet.
10. Send a message to TSU with the link to the record.⁴
11. Move the data into the CEDA archive.⁵

TSU get contributors to check Preview Catalogue Record

12. Add the link to the catalogue record in the TSU internal usage spreadsheet.
13. Contact all the contributors to send them the link to the Catalogue record.⁶
14. **If contributors respond with issue(s)**, Update the metadata fields on Figure Manager as necessary.⁷ When ready to ask CEDA to update the catalogue record (using slack).^{8,9}
15. Update the TSU internal usage spreadsheet.

CEDA Finalise and Publish Catalogue Record

16. **If contributors approve the Preview (no issues)**, change the catalogue record status to Working mode.¹⁰
17. Update the TSU-CEDA usage spreadsheet.
18. If the authors have found issues, then update the catalogue record following the instructions outlined by the TSU (using slack). Once the changes have been made to the catalogue record, send a message to TSU (using slack)¹¹
19. After the report press launch, publish the record (i.e. make it public). Create the DOI for the Catalogue record¹²
20. Update the TSU-CEDA usage spreadsheet.

Adding links to Published Catalogue Records

Once the report website is live, the following as Related documents:

- Chapter/SPM/TS webpage on report microsite (once website is live)

- If applicable, Chapter Supplementary Material download on report microsite (once website is live). The figure may be documented in the data table and/or may have info about the figure.
- If applicable, corresponding input data archived with DKRZ
- If applicable, code archived with Zenodo

TSU prepare the links and text to add to abstracts

21. Prepare the links for CEDA to add to all Catalogue Records.
22. Add the CEDA DOIs to the figure popups on the report microsite.

CEDA add the links and abstract text

23. Add the URL links in Related Documents and the following the text provided by the TSU to the abstract field:

Links

Links to Chapter X in the Working Group I Contribution to the IPCC Sixth Assessment Report, [the Supplementary Material for Chapter X ([supporting information in Section X.SM.Y] [and documentation of the input data used in Table X.SM.Z]) [and the code used to generate the figure] can be found in the Related Documents section of this catalogue record. [Input data used for this figure has been archived with the German Climate Computing Center (DKRZ) - see the link in the Related Documents section.]

CEDA update the citation in the abstract

24. Once the report is published with a DOI, update the citation in the 'Citing the dataset' part of the abstract.

Error Handling

TSU prepare new files and metadata

25. TSU checks the new files and identifies any changes needed to the catalogue record.
26. Once ready, the TSU emails CEDA the new files and informs them of changes to the catalogue record. The 'creation date' will be different in the data files.

CEDA

27. CEDA checked the new files and requested changes to the catalogue record.
28. CEDA creates a directory for the new data and a new catalogue record with a new DOI. The old catalogue record will be linked to the original one.

Useful links

Author and Chapter Scientists names and emails can be found in the Excel files in this DMS folder (TSU instance): <https://wg1.ipcc.ch/apps/dm/filemanager.php?q=343>
Use Email1 as the primary way to contact authors (CLAs, LAs) and chapter scientists.

The list of WGI report authors can be found here by chapter:

<https://archive.ipcc.ch/report/authors/report.authors.php?q=35&p=&p>

Guide to Document Manager:

https://wg1.ipcc.ch/sites/default/files/documents/Document_Manager_User_Guide.pdf

Link to FGD chapter files:

<https://wg1.ipcc.ch/apps/dm/filemanager.php?q=5759>

Link to SPM files:

<https://wg1.ipcc.ch/apps/dm/filemanager.php?q=5901>

Archival of Input data with DKRZ:

http://bit.ly/IPCC_DKRZ_inputdata

Figure Manager statuses

Statuses relevant to TSU

Data/code does not require archival [TSU]

Issue with TSU Check [TSU]

Passed TSU Check [TSU]

Data/metadata given to CEDA [TSU]

Issue with CEDA check [CEDA] - *TSU to follow up on issue*

Data in archive [CEDA] - *TSU to share catalogue record preview*

Preview with authors [TSU]

Issue with Preview [TSU]

Passed Preview Check [TSU]

CEDA to update catalogue record [TSU]

Record published (DOI added here) [CEDA]

Statuses relevant to CEDA

Data/metadata given to CEDA [TSU] - *CEDA to start archiving these*

Issue with CEDA check [CEDA]

Passed CEDA Check [CEDA]

Record created (UUID added here) [CEDA]

Data in archive [CEDA]

CEDA to update catalogue record [TSU] - *Update catalogue record as directed by TSU in the notes field*

Record finalised (working mode) [CEDA]

Record published (DOI added here) [CEDA]

N.B. The following procedure was planned but we haven't executed it in practice:

⁰If no issues are found, set the archival status to '**Passed TSU Check [TSU]**'. Add a '**Note on Current status**' if you wish.

¹ set the archival status to '**Issue with CEDA Check [CEDA]**' and add a '**Note on Current status**' outlining the issues for the TSU to follow up.

²Add who you have contact at the top of '**Note on Current status**' added by CEDA (want to keep the CEDA's list of issues in the note). Once you have followed up on the issue, set the archival status to '**Data/metadata given to CEDA [TSU]**' and update the TSU message at the top of the '**Note on Current status**'.

³N.B. The following procedure was planned but we haven't executed it in practice: **If no issues are found**, set the archival status to '**Passed CEDA Check [CEDA]**'. Add a '**Note on Current status**' if you wish.

⁴Add the URL for the catalogue record on the dataset page on Figure Manager in the field 'CEDA UUID'. Set the archival status to '**Record created (UUID added here) [CEDA]**'.

⁵Set the archival status to '**Data in archive [CEDA]**'.

⁶N.B. The following procedure was planned but we haven't executed it in practice: Once done this, set the archival status to '**Preview with authors [TSU]**'.

⁷N.B. The following procedure was planned but we haven't executed it in practice: set the archival status to '**Issue with Preview [TSU]**'. Add the issues to the '**Note on Current status**'.

⁸Set the archival status to '**CEDA to update catalogue record [TSU]**'. Outline all the changes to be made in the '**Note on Current status**'.

⁹**If contributors approve the Preview (no issues)**, set the archival status to '**Passed Preview Check [TSU]**'.

¹⁰Set the archival status to '**Record finalised (working mode)**'.

¹¹Set the archival status to '**Record finalised (working mode)**'.

¹²Add the DOI to the Figure Manager fields '**CEDA DOI**' and '**CEDA Short DOI**', then set the archival status to '**Record published (DOI added here) [CEDA]**'.