Wiki »

Table of content:

Table of contents

Correct cmorised data with the cmor-fixer
Status EC-Earth data at standard resolution
Status EC-Earth HighresMIP data

Correct cmorised data with the cmor-fixer

Under Published CMIP6 data and responsible people the EC-Earth ESGF publishers are listed in the table.

Because the amount of cmorised ocean data compared to the rest of the cmorised data is less than 4%, and because applying the script to all data is easier and less error prone and mostly because publishing the errata for the entire dataset is much easier, it is agreed that we will apply the cmor-fixer script on all EC-Earth published data. That means that we will republish the ocean data as well with a new `tracking_id`.

What needs to be done in the correct order by each of the ESGF publishers in charge, is:

1. A list of all published EC-Earth data per ESGF node has to be made before unpublishing. Each ESGF publisher can create this list by running: `esglist_datasets cmip6`, as mentioned in #731-21. Each list will be visible under the tab "AFFECTED DATASETS" of the published erratum once the erratum is published at the ESGF Errata Service.

2. One erratum text has to be written for the ES-DOC Dataset Errata database, I propose the following text:

Title	Incorrect shift of the longitude values in EC-Earth consortium KNMI-data
Description	The longitude values of the EC-Earth atmospheric data have been accidentally shifted half a grid cell eastward, this affects lon & lon_bnds (this error applies to all EC-Earth atmospheric data published until December 1st 2019). The error will be corrected as soon as possible. Note that this error does not affect any global and zonal means which might have been taken from these data by scientific studies. We decided to unpublish all EC-Earth3 data including the ocean data which is not affected by this error, as this is the easiest and most robust approach for this major operation.
Severity	High

A list of currently created errata:

- CMIP6 - Incorrect shift of the longitude values in EC-Earth consortium KNMI-data
- CMIP6 - Incorrect shift of the longitude values in EC-Earth consortium SMHI-data
- CMIP6 - Incorrect shift of the longitude values in EC-Earth consortium ICHEC-data
- CMIP6 - Incorrect shift of the longitude values on BSC ESGF data node

Note that the erratum tab "AFFECTED DATASETS" lists each variable (each table - variable combination) which will be unpublished & republished (including in our case the ocean variables which do not have the error but will obtain a new `tracking_id` for republishing).

3. The data has to be unpublished on the (primary) ESGF data node, use the `--retract` flag during unpublishing. And preferable move it to another destination at the same filesystem, so unintended access is prevented.

4. An Erratum as above (see point 2) has to be published for each ESGF node separately for all the EC-Earth3 data on that node via the Errata Service. This erratum has to be linked with the list of unpublished data (see point 1).

5. For each ESGF node publisher: As soon the EC-Earth data is unpublished and the erratum is published for your ESGF node, please contact Ruth Petrie (ruth.petrie@stfc.ac.uk) from CDNOT, so she will inform the CDNOT network to take action.

Only when the steps above are completed, one can continue with the next points, point 6 and further.

An ESGF node exists of a data node and a server. For instance, the server which is connected to the ESGF node of NSC has one node with 8 cores. When unpublishing the data the data will remain at this same disk, but it is not longer visible and accessible via the ESGF database. It is possible in that situation to apply a script on the unpublished data without moving the data to another disk (which would be painful for such an amount of data). Therefore we developed a separate package cmor-fixer to fix the longitude error. The release of `cmor-fixer v2.0` should be used to apply the fixes. The `cmor-fixer` can use for instance eight cores at this node with the `--npp 8` argument.

6. Apply cmor-fixer to the root dir of the CMIP6 data of the ESGF node:

1. Install `cmor-fixer` at the ESGF node server by following these install instructions which use `miniconda3`.
2. First run `cmor-fixer` in the save `dry-run` mode (use the `--dry` argument), it will show what will happen without changing anything. It is advisable to use the `--olist` option. With the `--olist` option each file which will be subject of modification (even if only a new `trackin_id` is given), is listed in the `list-of-modified-files.txt` file. The other important argument to include as well in our strategy is the `--forceid`. With this `--forceid` argument `cmor-fixer` will also replace the `tracking_id` in files without the longitude error, like the ocean files. This is

necessary because these files have been unpublished in our strategy as well and therefore they need a new `tracking_id` for republishing.

```
# Take care any list-of-modified-files*.txt files are moved or removed before start
./cmor-fixer.py --dry --verbose --forceid --olist --npp 1 CMIP6/
```

3. Run `cmor-fixer`, now without the `--dry` argument, so the longitude error will be fixed. One can `diff` the `list-of-modified-files.txt` files. They should be identical (at least after a sort).

```
./cmor-fixer.py --verbose --forceid --olist --npp 1 CMIP6/
diff list-of-modified-files-1.txt list-of-modified-files.txt
```

4. Run `cmor-fixer` once more with the `--dry` argument and without the `--forceid`, if everything went fine the new `list-of-modified-files-2.txt` will be empty.

```
./cmor-fixer.py --dry --verbose --olist --npp 1 CMIP6/
wc list-of-modified-files-2.txt
```

5. Run the script `./versions.sh` for instance to set all version directory names to January 20 2020 (see also here):

```
./versions.sh -v v20200120 -m CMIP6/
```

Important is that the version differs from the previous published version.

6. *Optional but not recommended by us*: Suggestion from CDNOT: move the data to a separate flat directory (so get all files in one directory and thus removing the CMIP6 directory structure). Run the `cmor-fixer` on that flat directory and then use the `esgdrs` tool to remake the correct CMIP6 directory structure. In this way the version directories will be recreated by the `esgdrs` tool itself. In this case the `versions.sh` script needs not to be run to apply new version directory names, although it is still useful to check whether everything went fine. However, we see risks in this approach, because the prior non flat directory might for whatever reason contain two versions for the same file (also to be checked: are all experiment names unique across all MIPs?).

7. After the corrections have been applied, some basic validation is recommended (see here):
   1. Run the script `files-per-year.sh` which counts the number of files per year in your dataset.
   2. Run `nctime` to check for glitches.
   3. You might want to check the `lon` & `lon_bnds` values themselves (see here).

8. When all is fine and in place, the new checksums have to be created, this is probably by far the most time consuming step.

9. Republish the EC-Earth3 CMIP6 data at your ESGF node with the corrected longitudes, the renewed `tracking-id` and the new `version`-directory names.

10. Update the Erratum for your ESGF node, by marking it as solved.

11. For each ESGF node publisher: as soon the EC-Earth data at your ESGF node is republished and the erratum is marked solved, please contact Ruth Petrie (ruth.petrie@stfc.ac.uk) from CDNOT, so she will inform the CDNOT network to take action. They will start to copy the corrected EC-Earth data to the non-primary image disks.

Note that step 6.2 - 6.4 can be replaced by using the `cmor-fixer-save-mode-wrapper.sh` wrapper script, which will check more intensively and will message in case there might have been an interruption during the operation. The `cmor-fixer-save-mode-wrapper.sh` wrapper script first does a check on the entire dataset and will stop immediately in case no file has the longitude or longitude boundaries error. It can be run like:

```
./cmor-fixer-save-mode-wrapper.sh 1 CMIP6/
```

Where the first argument is the number of cores to be used by `cmor-fixer` and the second argument is the path of the directory with the cmorised data.

After step 5 is complete the sites in the UK, Germany, America etc (the non-primary nodes) will simultaneously withdraw their copies of the EC-Earth data but they need completed Errata to do this. In the meanwhile, I understood from Ruth Petrie (from CDNOT), that after we informed the CDNOT network, several of the non-primary nodes already started to remove the erroneous EC-Earth data.

Note that if you want to apply other fixes at parts of the data as well (for instance correcting the metadata attributes like `branch_time_in_child` or `branch_time_in_parent`), a separate erratum for this correction and this part of the data (probably the data of one experiment) has to be published with a separate file list for this part of the data. This erratum needs then to be marked solved as well after publishing. It can go in the same unpublishing & republishing effort, which is beneficial in reducing work as it will save additional unpublishing, checksums & republising efforts. However, note that using the `--meta some-metadata-changes.json` option, needs the highest care! We absolutely recommend to apply this in a separate `cmor-fixer` run, after a prior `cmor-fixer` run in which the longitude issue has been fixed. We added this `--meta some-metadata-changes.json` to the `cmor-fixer` on request, but be aware that if you make just a stupid mistake with this option, you can damage largely your data. You probably never should apply `cmor-fixer` with this option on the root tree of your data, because the fix will usually only apply on one MIP experiment. The risk with this option is that you unintended modify just by mistake the metadata of other experiments, and the way back will be painful. The help (`cmor-fixer -h`) mentions this. Note that it is recommended (though not obligated) to use a `metadata.json` file which only list the fields you want to change. Note also that after running `cmor-fixer` once more the `trackin_id` will change once more as well for those files, but that is not a problem. Please contact us if you feel unsure about this option.

In general, it is good to know that it is not problematic if the `tracking_ids` are changed more than once during the unpublished intermezzo. This means that in case the check in point 6.4 indicates

there has been an interruption, one can reapply 6.3: Files which always have been correct and the ones which have been corrected will just obtain another new `tracking_id`, and files which were not treated due to the interruption will be corrected now and obtain a new `tracking_id` now.

For any further questions about Errata Service contact:
Guillaume Levavasseur (glipsl@ipsl.fr)
Mark A. Greenslade (momipsl@ipsl.fr)

Status EC-Earth data at standard resolution

All EC-Earth consortium data at standard resolution which has been published on the various ESGF nodes is not fixed yet at January 31 2020. The `cmor-fixer v2.0` is tested and will be released at January 31 2020. This release is the recommended version to use for applying the fix to the data at the ESGF nodes.

Status EC-Earth HighresMIP data

All EC-Earth consortium HighResMIP data which has been published on an CEDA's ESGF node is correct.

Those who directly download data from the PRIMAVERA group workspaces at JASMIN need to be careful and check each dataset because parts of the data at JASMIN still have the error and other parts of the data at JASMIN is correct.

▸ Files