How to use ESMValTool connection to Earth system data cube (ESDC) cloud data

ESMValTool installation

The latest ESMValTool release version is available as a conda-forge package, making it easily accessible to users of the Conda[1] or Mamba[2] package managers. It is recommended to use Mamba:

```
mamba create --name esmvaltool esmvaltool
```

ESMValTool can also be installed directly from its GitHub repository[3]. Detailed installation instructions can be found on the ESMValTool documentation[4] and tutorial[5].

ESMValTool configuration

Once installed, ESMValTool has to be configured to the system where it will be used. For the connection to the ESDC cloud data, the focus will be on the output data folder and the raw data input folder. More information about the configuration is available on the documentation[6] and the tutorial[7].

To install the default configuration file in the default location, run:

```
esmvaltool config get_config_user
```

This command stores the configuration by default in:

```
~/.esmvaltool/config-user.yml
```

The output data folder is the folder where ESMValTool will write the output of its operations. It is defined in the line:

```
output_dir: ~/esmvaltool_output
```

Although the ESDC data is available on the cloud, for legacy reasons it requires an input folder to be available locally. There are several input folders for ESMValTool, depending on the dataset. The ESDC data is considered “raw data”, so its location is defined in the “rootpath” definitions “RAWOBS”:

```
rootpath:
  RAWOBS: ~/rawobs_inputpath
```

ESMValTool expects input data to be sorted by “Tiers”, depending on their level of privacy, with ESDC being Tier 2. Thus, for ESMValTool to recognize the ESDC dataset, a folder has to be created on:

---

[1] https://docs.conda.io/en/latest/index.html
[2] https://mamba.readthedocs.io/en/latest/index.html
[3] https://github.com/ESMValGroup/ESMValTool
[4] https://docs.esmvaltool.org/en/latest/quickstart/installation.html
[5] https://esmvalgroup.github.io/ESMValTool_Tutorial/02-installation/index.html
[6] https://docs.esmvaltool.org/en/latest/quickstart/configuration.html
[7] https://esmvalgroup.github.io/ESMValTool_Tutorial/03-configuration/index.html

```
mkdir ~/rawobs_inputpath/Tier2/ESDC
```

For ESMValTool to access ESDC on the cloud, this folder must not contain data. If it does, ESMValTool will use that data instead.

Use ESMValTool to access ESDC cloud data

The connection to the ESDC dataset was implemented as a "CMORizer" script. "CMORizers" are an integral part of the ESMValTool pre-processing infrastructure and used to make a dataset compliant with the CMOR standard, the required format for input data to be used with ESMValTool. Currently, ESMValTool has CMORizer scripts for almost 90 datasets, the full list of which, including ESDC, is available online on the documentation[8].

To confirm that the ESMValTool version installed has the ESDC dataset and see more information about the dataset and how to access it, we can run the command:

```
esmvaltool data info ESDC
```

A CMORizer script can be executed with the ESMValTool instruction "data format":

```
esmvaltool data format ESDC
```

While normally ESMValTool

The CMORized ESDC data is saved to disk. ESMValTool automatically generates a new output directory with every run. The location is determined by the output_dir option in the config-user.yml file, the name of the operation ("data_formatting" in this case), and the date and time, using the format: YYYYMMDD_HHMMSS. The specific folder is also shown on the ESMValTool console output:

```
INFO    input_dir  = /home/user/rawobs_inputpath
INFO    output_dir = /home/user/esmvaltool_output/data_formatting_20230418_135029
```

Development of ESDC CMORizer in ESMValTool

It is possible to modify and extend the behavior of the ESDC CMORizer to use a different version of the dataset or add support for more variables. This involves modifying the ESMValTool source code. It is possible to install ESMValTool on "developer mode"[9] by installing the source code on the environment with:

```
pip install --editable '.[develop]'
```

This allows any changes made to the source code to be immediately be available in the installed version.

Configure dataset version

By default, the script uses the version 3.0.1 of the ESDC dataset, with a resolution of 0.25 degrees and chunking 1x720x1440. These settings can be modified on the file:

/ESMValTool/esmvaltool/cmorizers/data/cmor_config/ESDC.yml[10]

[8] https://docs.esmvaltool.org/en/latest/input.html#supported-datasets-for-which-a-cmorizer-script-is-available
[9] https://docs.esmvaltool.org/en/latest/quickstart/installation.html#install-from-source
[10] https://github.com/ESMValGroup/ESMValTool/blob/main/esmvaltool/cmorizers/data/cmor_config/ESDC.yml

Add support for more variables

Before ESMValTool can work with ESDC data, it has to be converted to the CMOR standard. This process may involve several changes depending on the original data, for example changing of units, coordinates, variable names or frequencies. This is a very involved process, so in this document only a superficial introduction will be given.

The CMORization process is defined on the file

/ESMValTool/esmvaltool/cmorizers/data/formatters/datasets/esdc.py[11]

For ESMValTool to attempt to CMORize a variable, it first has to be defined on the ESDC configuration file:

/ESMValTool/esmvaltool/cmorizers/data/cmor_config/ESDC.yml

under the "variables" list. E.g.:

```
variables:
  tas:
    mip: Amon
    raw: air_temperature_2m
  tasmax:
    mip: Amon
    raw: max_air_temperature_2m
  tasmin:
    mip: Amon
    raw: min_air_temperature_2m
```

Full information on how to develop for a CMORizer script is available on the documentation[12]. There is also a lesson available in the tutorial[13].

Process data with recipes

A "recipe" is a file that defines the input data and preprocessing steps applied to be analyzed by ESMValTool diagnostic scripts. Detailed information about how to run a recipe[14] and understand its output[15] is available on the ESMValTool documentation and in the tutorial[16].

Make ESDC data available to ESMValTool

To access the CMORized ESDC data, ESMValTool expects to find it on the "OBS" folder. This folder is defined in the configuration:

```
rootpath:
  OBS: ~/obs_inputpath
```

[11] https://github.com/ESMValGroup/ESMValTool/blob/main/esmvaltool/cmorizers/data/formatters/datasets/esdc.py
[12] https://docs.esmvaltool.org/en/latest/develop/dataset.html
[13] https://esmvalgroup.github.io/ESMValTool_Tutorial/09-cmorization/index.html
[14] https://docs.esmvaltool.org/en/latest/quickstart/running.html
[15] https://docs.esmvaltool.org/en/latest/quickstart/output.html
[16] https://esmvalgroup.github.io/ESMValTool_Tutorial/04-recipe/index.html

The CMORized ESDC data has to be moved or copied to this folder. The data is already output to the correct "Tier" folder:

```
mv ~/esmvaltool_output/data_formatting_20230418_135029/Tier2 ~/obs_inputpath/
```

or

```
mv ~/esmvaltool_output/data_formatting_20230418_135029/Tier2/ESDC
~/obs_inputpath/Tier2/
```