

# EOSC Future Science Project 9

## Climate Neutral and Smart Cities

Joachim Wackerow

Integrate climate data from Copernicus ERA5 and air quality data from the European Environmental Agency (EEA) with data from the European Social Survey (ESS)



Climate data

Social scientific survey

European Air Quality



**European  
Environment  
Agency**

# Metadata Requirements for Cross-Domain

- The integration of data from multiple domains has **higher demands** on metadata than reuse of data in a single domain
  - It **increases** the requirements on the quality and scope of the metadata
- This applies especially for **provenance** information
- *Following slides show screenshots of the process description application prototype*

# Overview of Process Activities

## CDI-Workflow description of the EOSC Future WP6 Task 3, Science Project 9 'Climate Neutral and Smart Cities'

### Main Process Sequence

**Description:** Main Sequence of the process

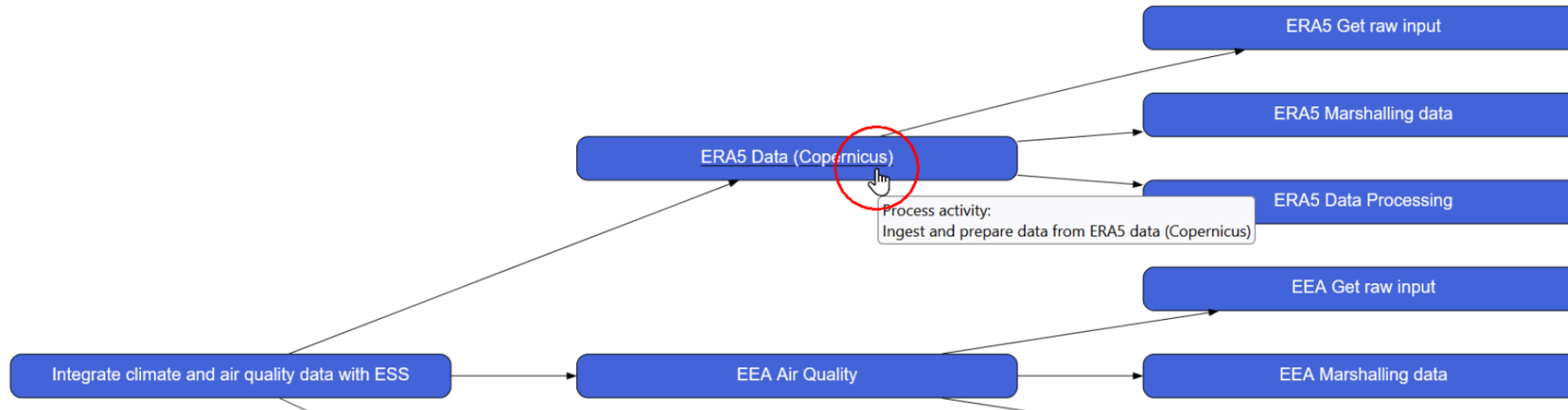
Processing Agent: EOSC project team at Sikt - Norwegian Agency for Shared Services in Education and Research

Purpose: Integrate climate data from ERA5 and air quality data from the EEA with the ESS survey data

Production Environment: Sikt - Norwegian Agency for Shared Services in Education and Research acting as a participant of SP9

### Overview Diagram of the Process Activities (in sequential order)

**Note:** Move the mouse cursor over an activity to see more information. Click on an activity to go to the corresponding page.



# ERA5 Data Processing

ESS Labs Process Search

Go

Contents

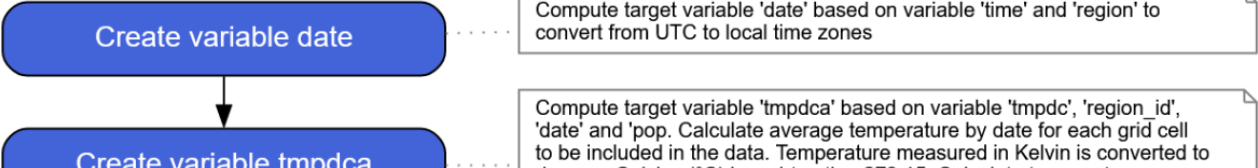
- Integrate climate and air quality data with ESS
  - ERA5 Data (Copernicus)
    - ERA5 Get raw input
    - ERA5 Marshalling data
    - ERA5 Data Processing
      - Create variable date
      - Create variable tmpdca
      - Create variable tmpdcmx
      - Create variable tmpdcmn
      - Create variable tmpdcaw
      - Create variable tmpdcam
      - Create variable tmpdca3m
      - Create variable tmpdcay
      - Create variable tmpdcacm
      - Create variable tmpdcamb
      - Create variable tmp95pacmb
      - Create variable tmpanod
      - Create variable tmpanocm
      - Create variable paccta**
      - Create variable pacctaw
      - Create Process step:
        - Compute target variable 'paccta' based on variable 'pac', 'region\_id', 'date' and 'pop'.
        - Calculate total precipitation by date in mm for each grid cell to be included in the data.
        - The total precipitation over 24 hours is the sum of the individual total precipitation values for each hour. Precipitation measured in m converted to mm. Calculate total precipitation by date for each region as an average of the grid cell value, weighted by variable 'pop' that is based on global human settlements statistics.
    - Create variable iwg10mxam
    - Create variable iwg10mxa3m
    - Create variable iwg10mxay
    - Create variable iwg10mxamb
  - EEA Air Quality
  - Merging of ERA5, EEA and ESS data

**Description:** The process involves creating a date variable from timestamps based on the time zone of each region, considering that the data is recorded hourly. It also addresses unit differences, converting Kelvin to Celsius and meters to millimeters. The data is then grouped by date, variable, and region, and temperature is averaged while also obtaining maximum and minimum values, accumulating precipitation by date, and identifying the maximum wind gust value. Moving averages are calculated for variables using different time windows (7-day, 30-day, 90-day, 365-day). Baseline values for temperature, precipitation, wind gust, and deviations from the baseline (anomalies) are determined based on the period from 1991 to 2020. Data older than 2015 is removed, and a group-by operation is performed, collapsing the data by region using population-weighted averages. It is important to note that the ERA5 data may contain imputed and missing values. In memory, each row corresponds to a region, with mesh-blocks aggregated per day to calculate region-level values by taking the average of all variables weighted by the population of each block. The resulting data is stored to disk in CSV, SAV, or other suitable formats, as the data size remains manageable.

**Diagram of the Process Activity**



**Sub-Activities (in sequential order)**



# Create Variable ,paccta'

## Create variable paccta

### Process Step

**Description** Compute target variable 'paccta' based on variable 'pac', 'region\_id', 'date' and 'pop'. Calculate total precipitation by date in mm for each grid cell to be included in the data. The total precipitation over 24 hours is the sum of the individual total precipitation values for each hour. Precipitation measured in m converted to mm. Calculate total precipitation by date for each region as an average of the grid cell value, weighted by variable 'pop' that is based on global human settlements statistics.

This step uses a [script](#) written in Python3.

### Diagram of the Process Step



Following process step:  
Compute target variable 'pacctcm' based on variables 'paccta', 'date' and 'region\_id'. Calculate 'pacctcm' as an average of total precipitation date values for the calendar month.

**Hint:** Move the mouse cursor over a parameter to see more information. Click on a parameter or a related step to go to the corresponding page.

**Legend:** Input parameter → used by → Process step → produces → Output parameter

# Create Variable 'pacctcm'

## Create variable pacctcm

### Process Step

**Description** Compute target variable 'pacctcm' based on variables 'paccta', 'date' and 'region\_id'. Calculate 'pacctcm' as an average of total precipitation date values for the calendar month.

This step uses a **script** written in Python3.

### Diagram of the Process Step



**Hint:** Move the mouse cursor over a parameter to see more information. Click on a parameter or a related step to go to the corresponding page.

**Legend:** Input parameter (yellow box) — used by —> Process step (blue rounded rectangle) — produces —> Output parameter (purple box)

# Code for the Creating Variable 'pacctam'

Files

master

Go to file

- .gitignore
- config\_RENAME\_ME.py
- eea-download.py
- eea-prepare.py
- era5-download.py
- era5-prepare.py
- merge.py
- requirements.in
- requirements.txt
- utils.py

```
ess-labs-data-sp9 / era5-prepare.py
```

```
82 def calculating_anomalies(df_in):
96
97     df["tspanod"] = df["tmpdca"] - df["tmpdcamb"]
98     # more_than_95p = df["tmpdca"] > df["tmp95pacmb"]
99     # df["tmp95p3d"] = more_than_95p.rolling(3, min_periods=3).sum()
100
101     df["tmpdcacm"] = df.groupby("year_month")["tmpdca"].transform("mean")
102     df["tspanocm"] = df["tmpdcacm"] - df["tmpdcamb"]
103
104     # paccta
105     # sum daily to year_monthly first
106     df["pacctcm"] = df.groupby("year_month")["paccta"].transform("sum")
107     df_baseline["pacctcm"] = df_baseline.groupby("year_month")["paccta"].transform(
108         "sum"
109     )
110     # calculate monthly for baseline, join back in
111     pac_cal_month_groupby = df_baseline.groupby("cal_month")["pacctcm"]
112     pacctmb = pac_cal_month_groupby.aggregate("mean")
113     pacctmb.name = "pacctmb"
114     df = df.join(pacctmb, on="cal_month")
115     # anomalies
116     df["paccdc"] = (df["pacctcm"] / df["pacctmb"]) * 100
117
118     iwg_cal_month_groupby = df_baseline.groupby("cal_month")["iwg10mx"]
119     iwg10mxamb = iwg_cal_month_groupby.aggregate("mean")
120     iwg10mxamb.name = "iwg10mxamb"
121     df = df.join(iwg10mxamb, on="cal_month")
122
123     df = df.set_index(["region", "date"])
124     df = df[
125         [
126             "tmpdcamb",
127             "tmp95pacmb",
128             "tspanod"
```

Symbols

Find definitions and references for functions and other symbols in this file by clicking a symbol below or in the code.

Filter symbols

- func create\_date\_column
- func fix\_measurements
- func groupby\_date
- func calculating\_moving\_averages
- func calculating\_anomalies
- func weighted\_average
- func collapse\_grid
- func order\_columns
- func do\_for\_region
- func check\_labeled
- func main



# Created Variable 'pacctcm' in the Data Repository

European Social Survey Data Portal Data Wizard About Search questions/variables

DATASET: [EOSC Future Science Project Climate Neutral and Smart Cities](#)

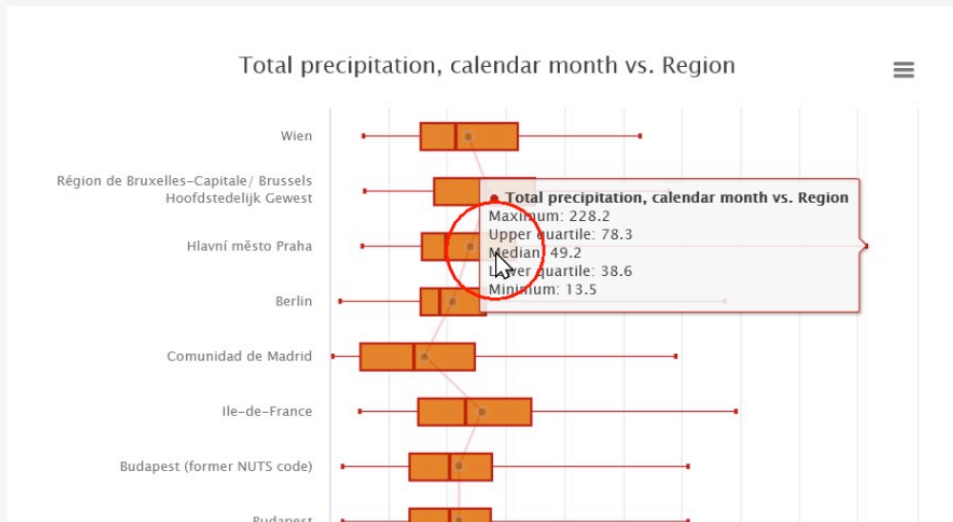
Subject

era5-regions

Variables

pacctcm - Total precipitation, calendar mo

pacctcm - Total precipitation, calendar month



# Process Information in Standardized Form

- The description of the process/provenance information is made in DDI-CDI
- DDI Cross Domain Integration (DDI-CDI) is an emerging standard for the domain-independent description of data
- The DDI-CDI Process description is aligned with
  - Prov Ontology of World Wide Web Consortium (W3C)
  - Business Process Model and Notation (BPMN) of the Object Management Group (OMG)



Further Project Information at ESS Labs

<https://europeansocialsurvey.org/eslabs/>