EUMETSAT

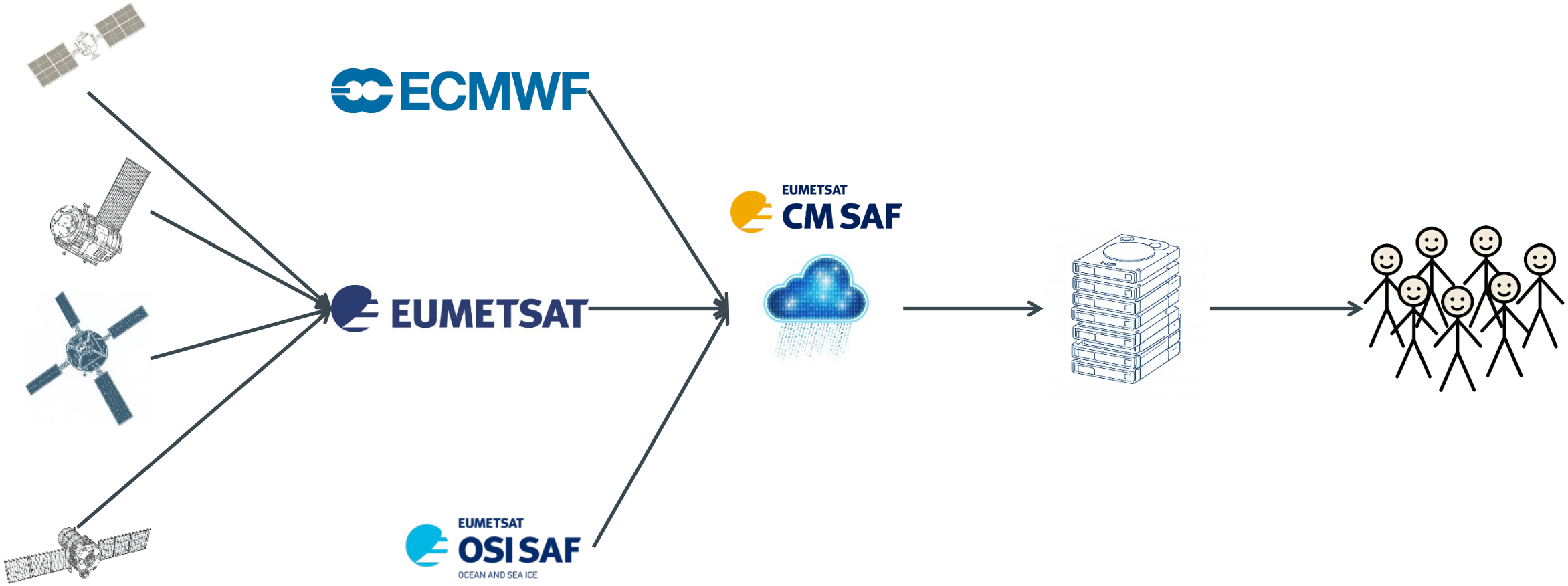
Creation of satellite-based climate data records on the EWC

Marco Marquard and Yannis Eigenbrodt
CM SAF / DWD

EWC User Workshop / 2025-09-18

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Satellites | Data provider | Processing | Data Store | User

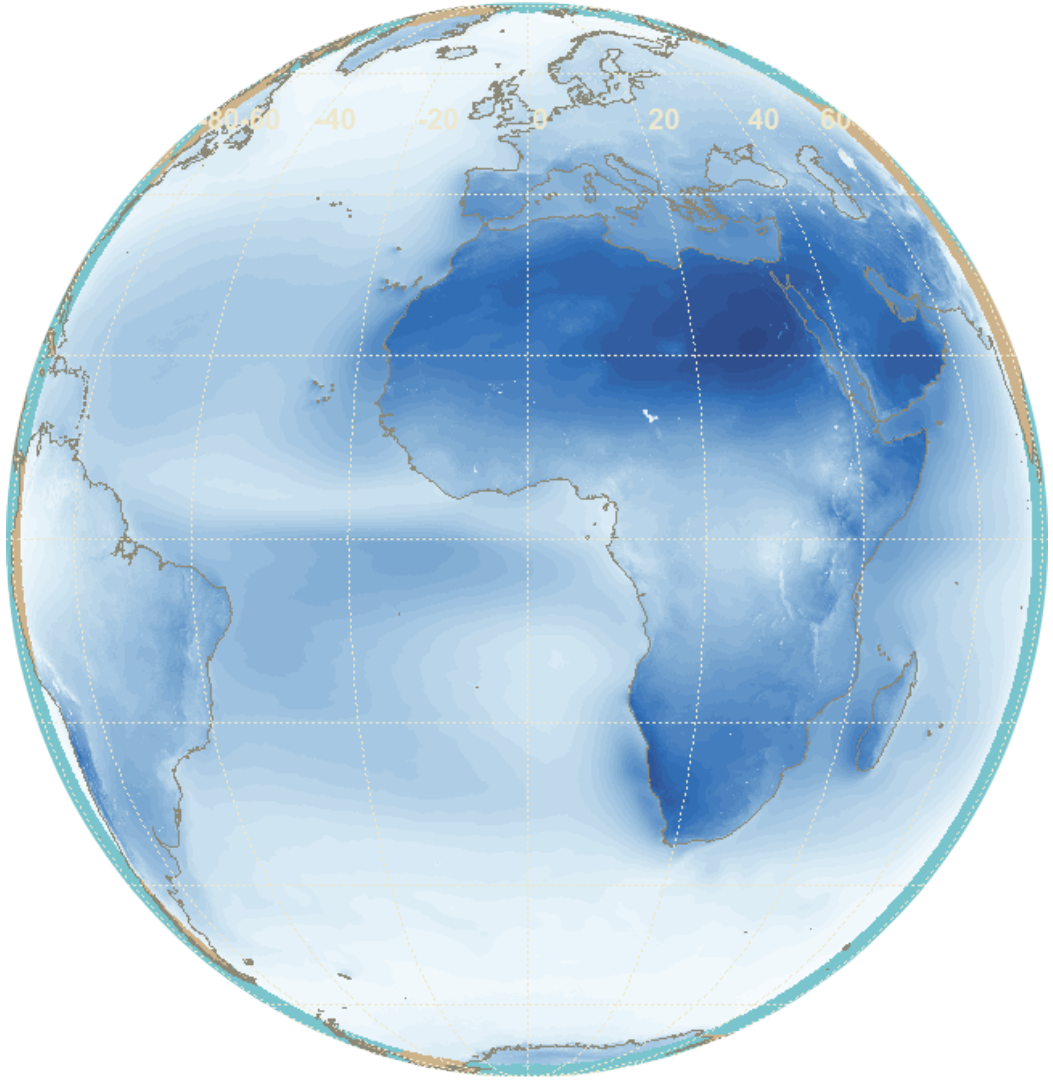


CLAAS-4 CDR

EUMETSAT CM SAF

www.cmsaf.eu

- Climate dataset of cloud properties
- Covers SEVIRI disk of Meteosat Prime (0°)
- New edition, planned release 2026
- Temporal coverage: 2004 – 2025
- Total size of dataset ~ 150 TB

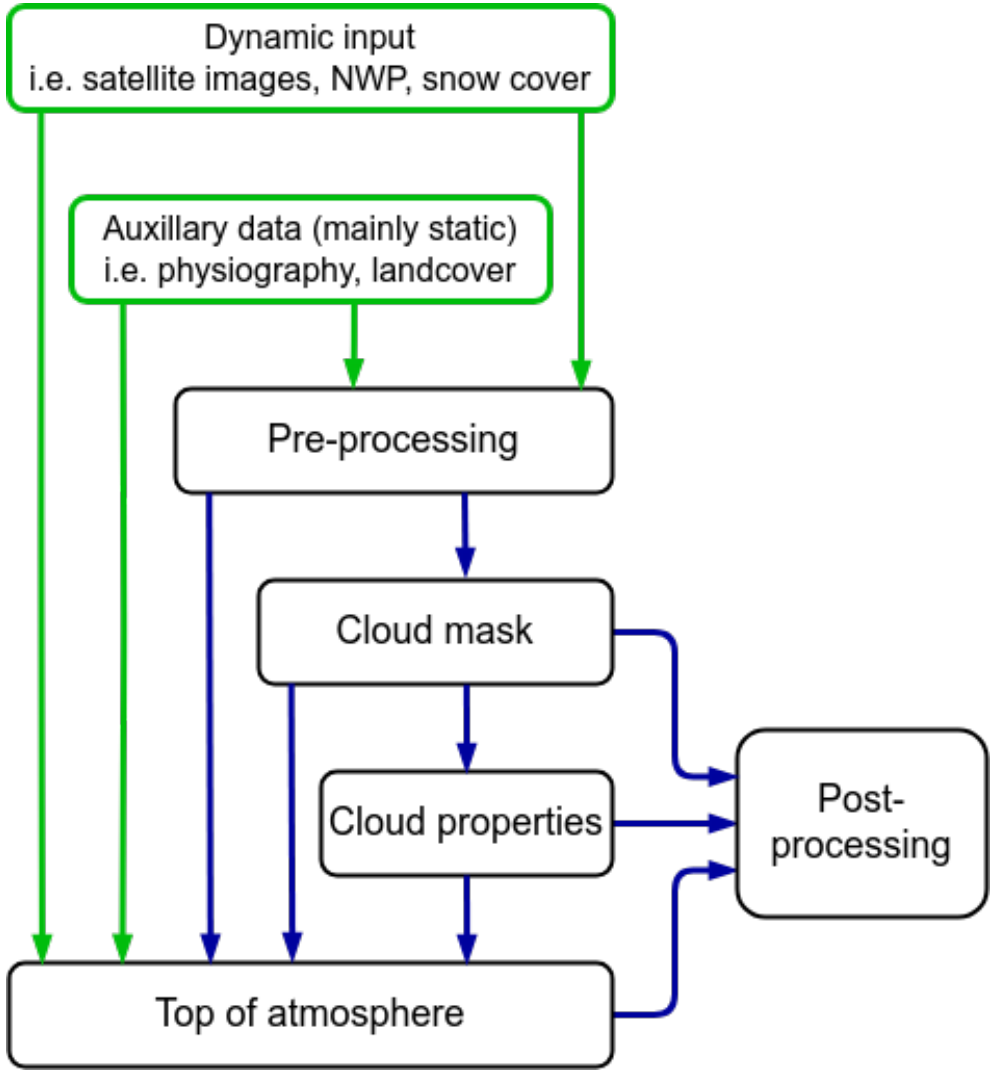


DOI:10.5676/EUM_SAF_CM/CLAAS/V003
Cloud Fractional Coverage (CFC)

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

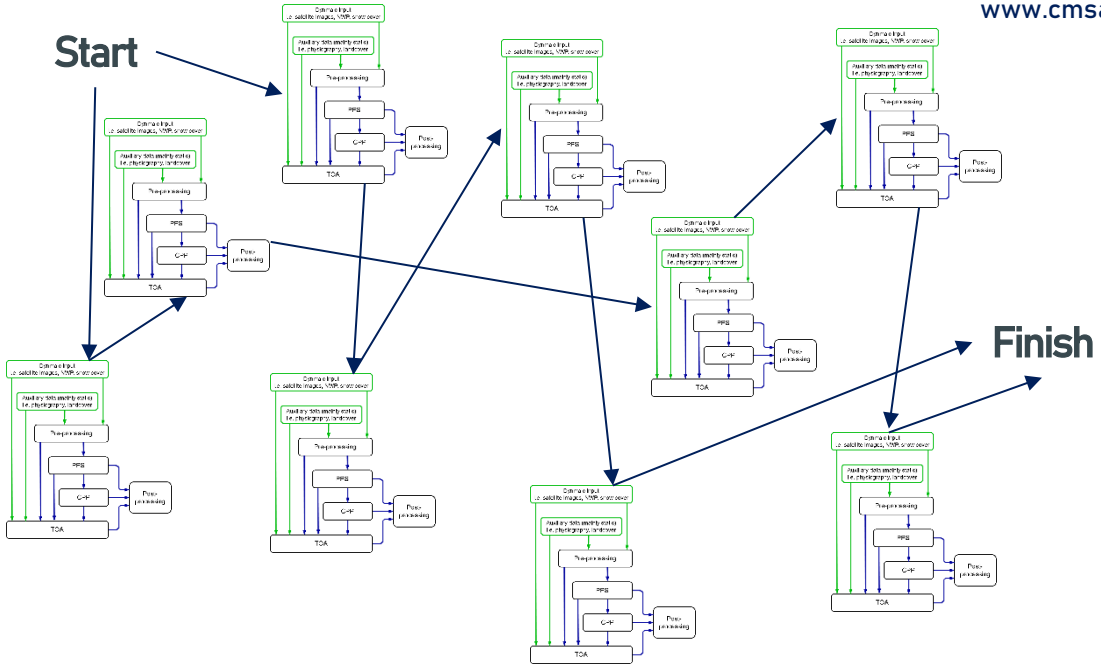
CLAAS-4 Production

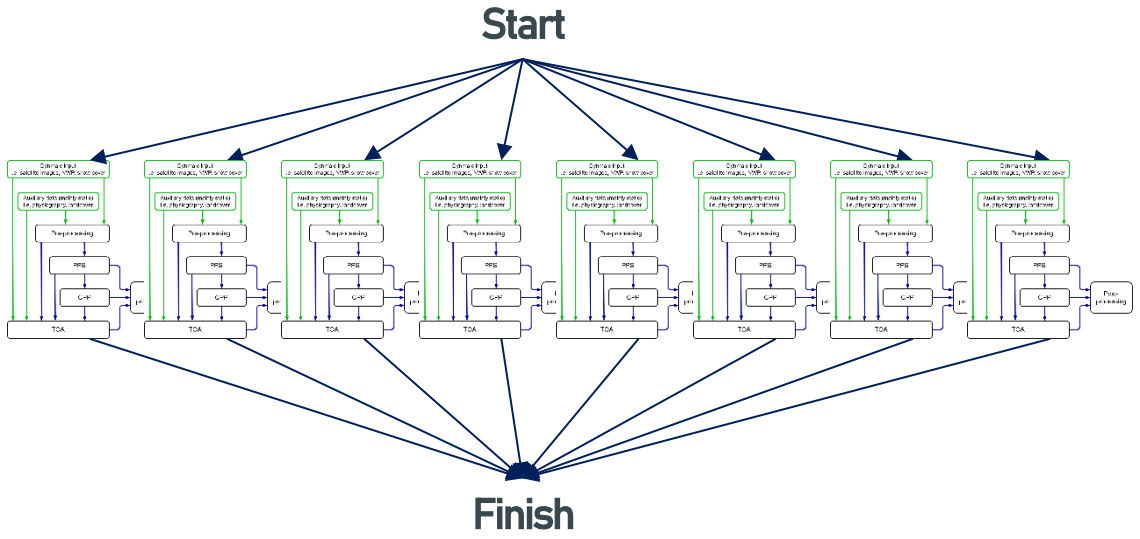
- Software components
  - provided by CM SAF partners
- Dependencies inherited from previous CLAAS editions
- Partially under active development
- Input to be processed for 20+ years:
  - 96 SEVIRI images per day
  - hourly ERA-5 input
  - daily ice map
  - auxiliary data



Workflow management 1

- Wishlist for *scientific* workflow management software
  - monitoring
  - profiling
  - run bash scripts
  - HTCondor integration
  - works with container in HTCondor
  - active project
- Collection of workflow engines from
  - https://workflows.community/systems
  - https://s.apache.org/existing-workflow-systems

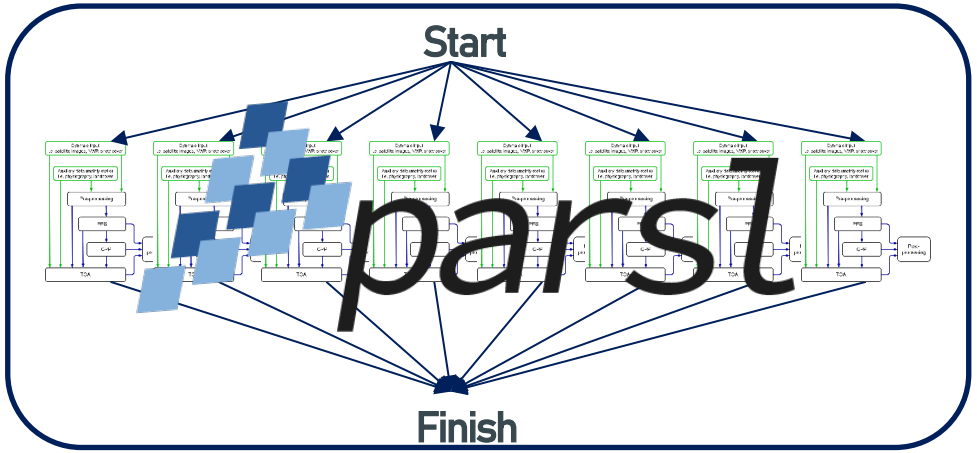




Workflow management 2

No “one fits all” solution found

- Daily workflow



- parsl
  - provides workflow management, parallel execution, monitoring, profiling

- Days for full time period

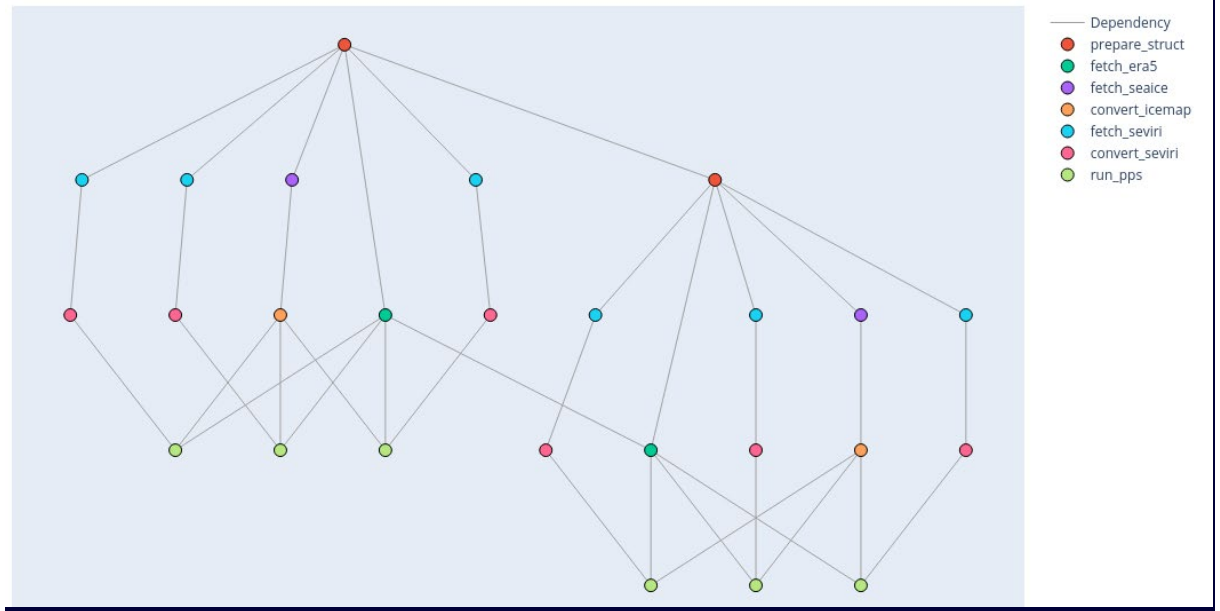
- HTCondor
  - provides integration to batch system, monitoring, workflow management

Workflow Implementation

- Implement processing steps as python functions and bash scripts
- Handle dependencies in parsl
- parsl HighThroughputExecutor provides monitoring and profiling capabilities

```python
@bash_app(executors=["local_htex"])
def convert_icemap(mamba_env: str, d_path: str, dy_path: str, stdout=parsl.AUTO_LOGNAME, stderr=parsl.AUTO_LOGNAME, walltime = 300, **kwargs):
    """
    App to unzip and convert into the right netcdf file type for the processing.
    - mamba_env: The name of the environment which contains the netcdf tools for the conversion.
    - d_path: Path of the data directory where all folders for the different dates are located
    - dy_path: Path of the folder for the current processing date
```

Workflow DAG




```python
    logger = logging.getLogger("fetch_seaice")

    seaice_dir = Path(d_path, dy_path, "import", "ICEMAP_data")
    seaice_download = Path(seaice_dir, "download")
    seaice_nh_file = f"ice_conc_nh_polstere-100_multi_{fetch_date:%Y%m%d}1200.nc*"
    seaice_sh_file = f"ice_conc_sh_polstere-100_multi_{fetch_date:%Y%m%d}1200.nc*"
    seaice_files_n = list(seaice_dir.rglob(seaice_nh_file))
    seaice_files_n.extend(list(seaice_download.rglob(seaice_nh_file)))
    seaice_files_s = list(seaice_dir.rglob(seaice_sh_file))
```

EWC Tools 1

Data input / output / storage

- Dynamic input
  - EUMETSAT data store
  - MARS
  - OSI SAF ftp Server
- SFS
  - shared storage between VMs
  - storage for sharing files and test data
- S3
  - cache for semi-static, pre-processed and final data
  - manage access through different keys
  - enable access to test user without EWC access

EWC Tools 2

Data processing

- VMs
  - selection of VM size according to requirement
  - deployment of larger / more VMs for processing tasks
- HTCondor batch
  - flexible resource usage
  - burst capacity (planned)

- Container
  - portable processing environment
  - integration of static data
  - container size of multiple GB so far no problem
  - container registry

Status

- Test data production
  - usage of multiple VMs with shared SFS storage
  - production of test data covering multiple months
  - automatic download and upload of data
  - successful porting of code to container and batch farm
- Lessons learned
  - data input is critical
  - easy deployment of VMs for different purposes / services can be very helpful
  - docker for HTCondor needs some special settings

Outlook

CDR Processing

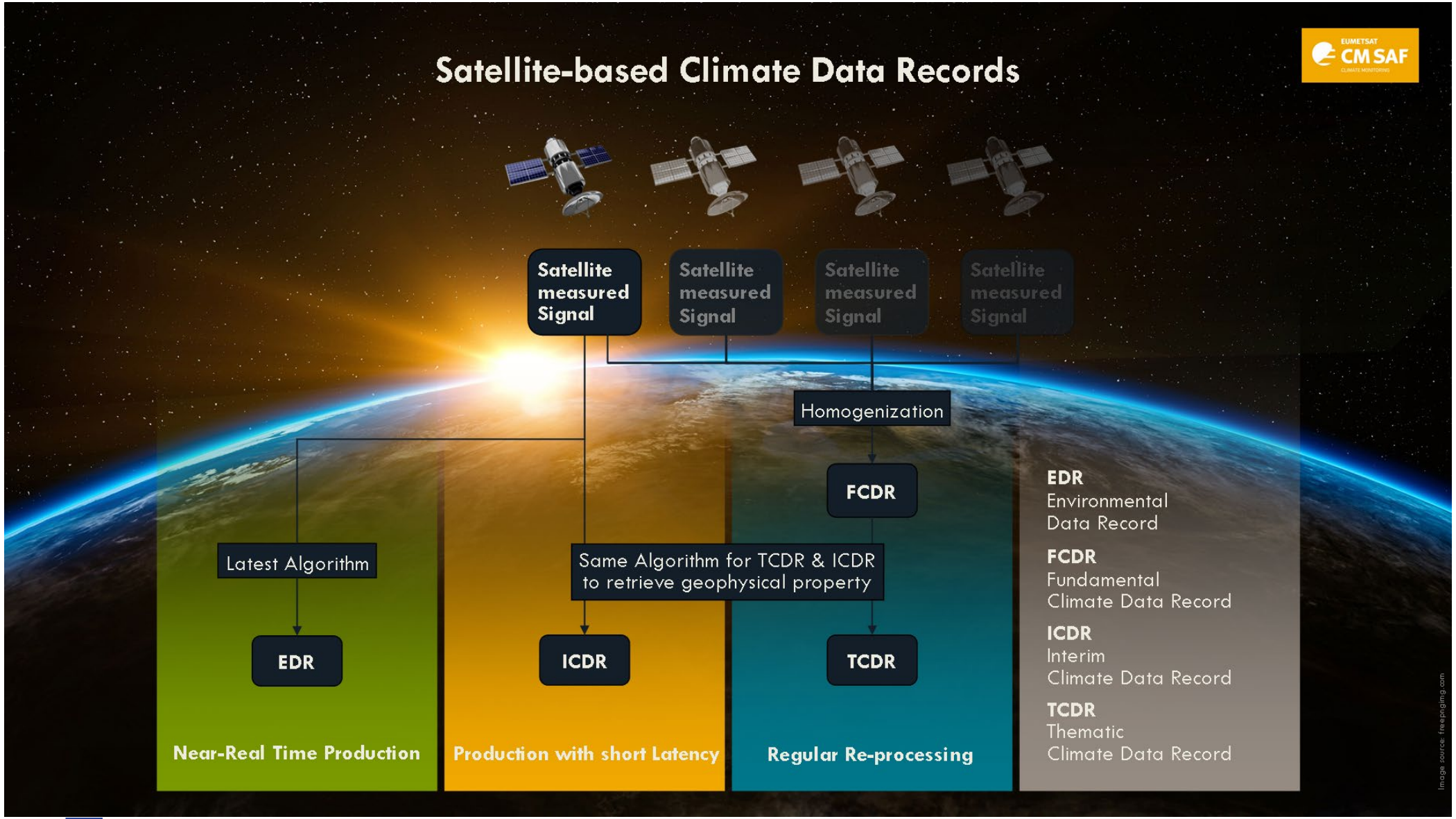
- Integrate all software components
- Scale processing up
  - check for bottlenecks

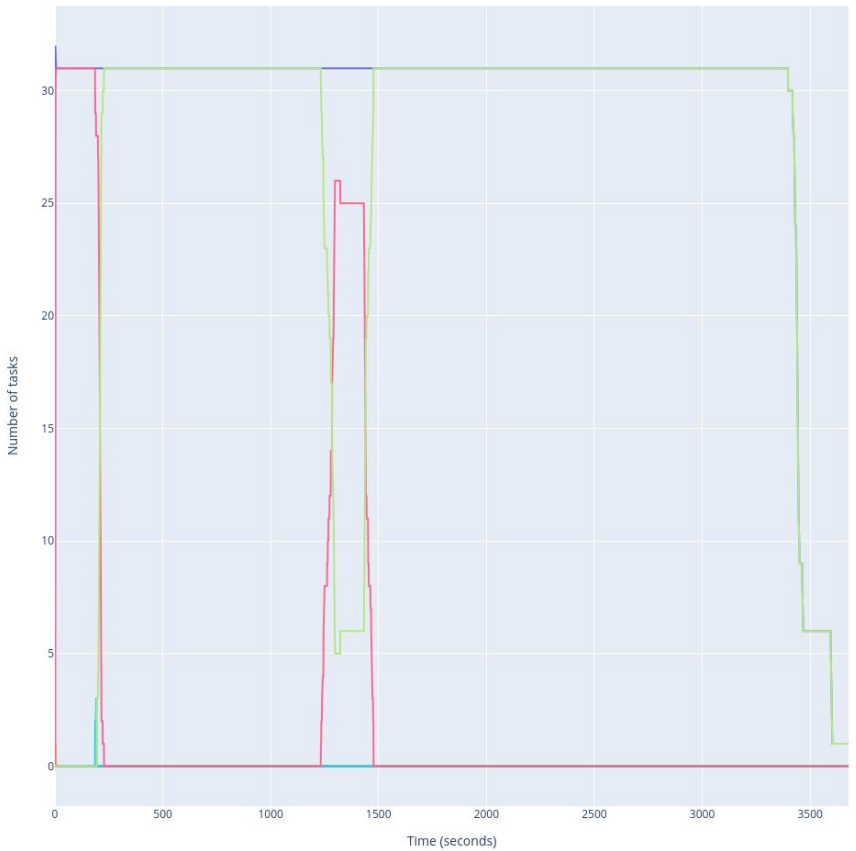
Development

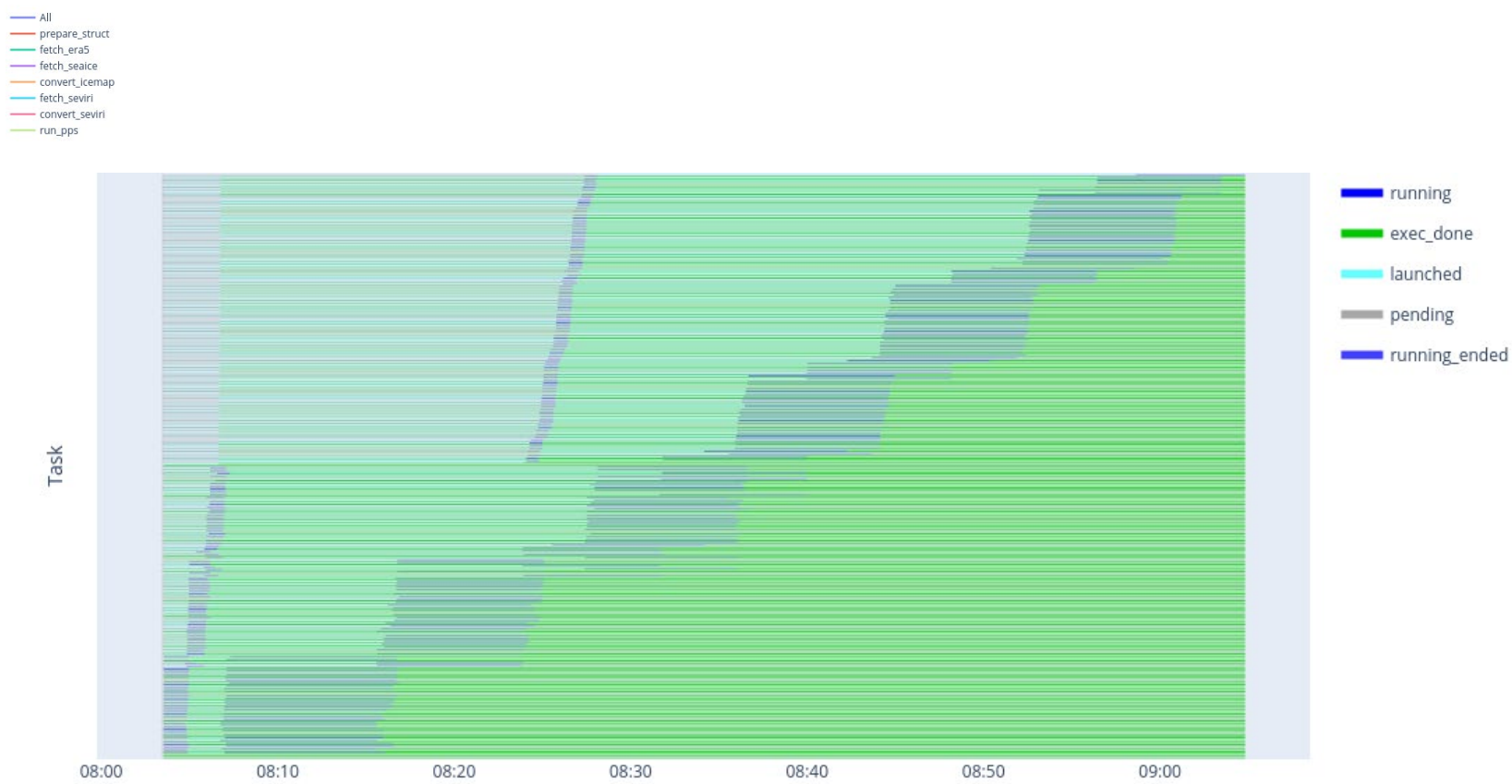
- Better data exchange with other SAFs
- Evaluate possibilities to process interim products
  - data processed in a daily manner
  - kubernetes?
- Evaluate additional scientific workflow manager (data-centric)

Questions?

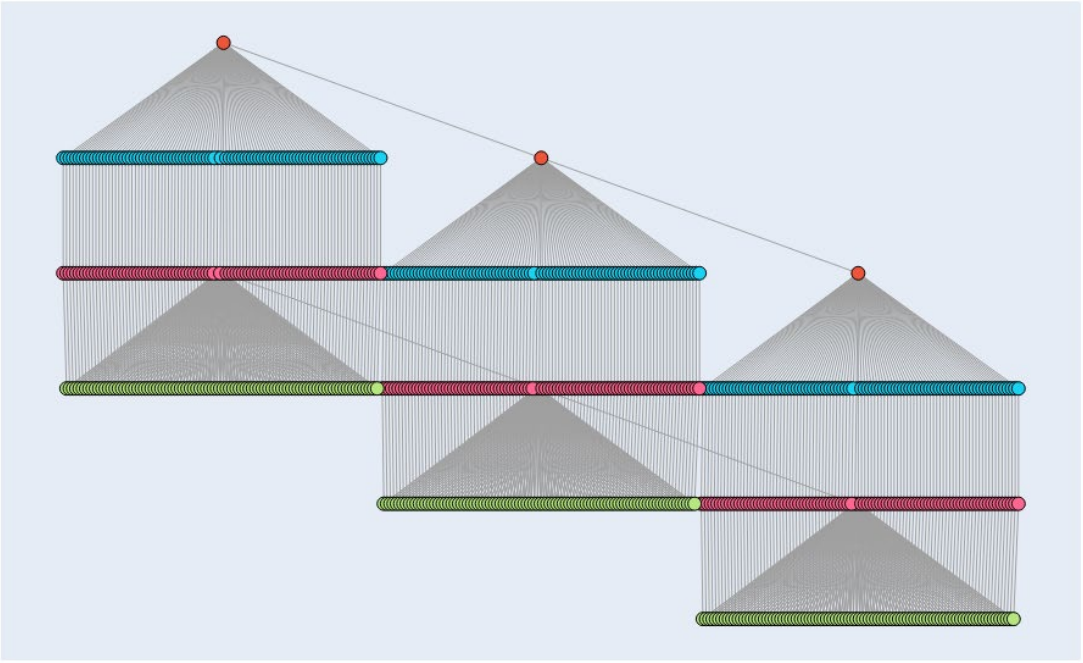
Thank you for your attention!


Satellite-based Climate Data Records
Satellite measured Signal
Satellite measured Signal
Satellite measured Signal
Satellite measured Signal
Homogenization
FCDR
Latest Algorithm
Same Algorithm for TCDR & ICDR to retrieve geophysical property
EDR
ICDR
TCDR
Near-Real Time Production
Production with short Latency
Regular Re-processing
EDR Environmental Data Record
FCDR Fundamental Climate Data Record
ICDR Interim Climate Data Record
TCDR Thematic Climate Data Record
Image source: freepngimg.com


Number of tasks
Time (seconds)
0
5
10
15
20
25
30
0
500
1000
1500
2000
2500
3000
3500

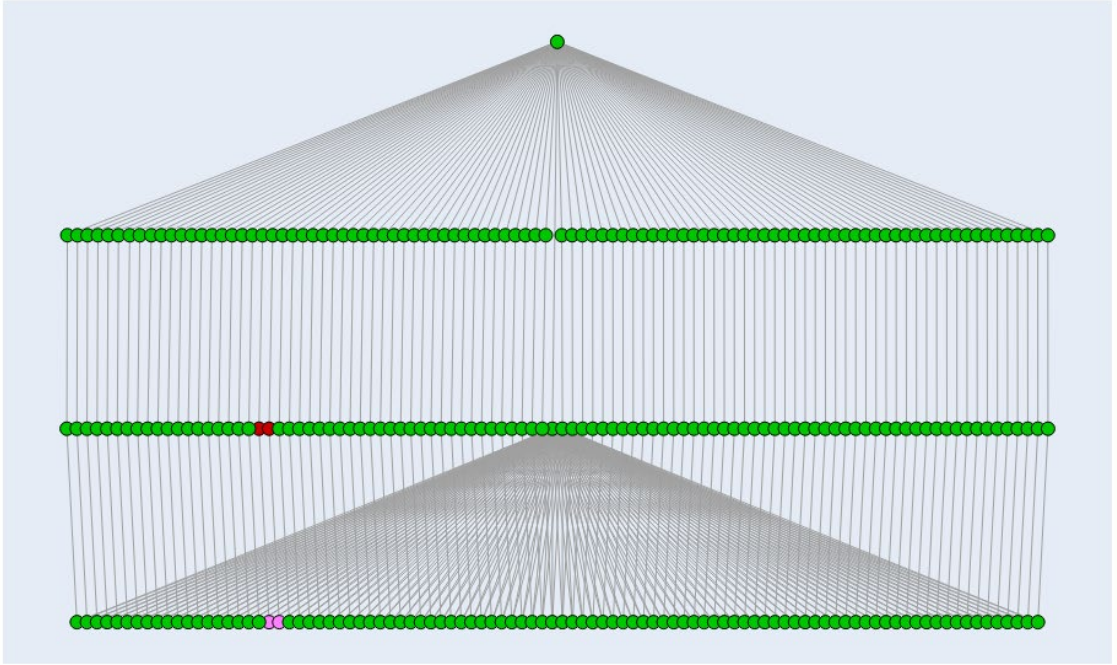

All
prepare_struct
fetch_era5
fetch_seaice
convert_icemap
fetch_seviri
convert_seviri
run_pps
Task
08:00
08:10
08:20
08:30
08:40
08:50
09:00
running
exec_done
launched
pending
running_ended

Workflow DAG

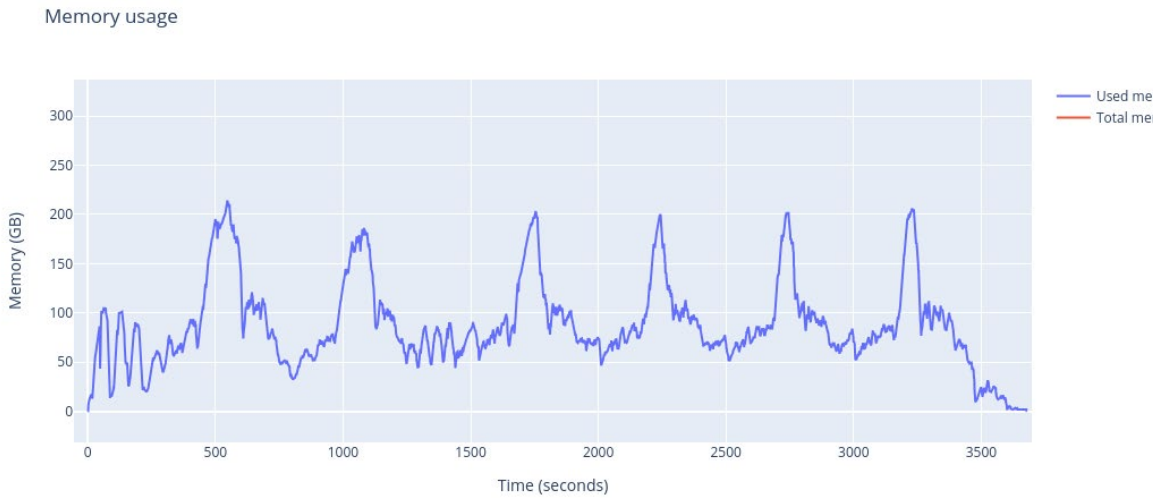


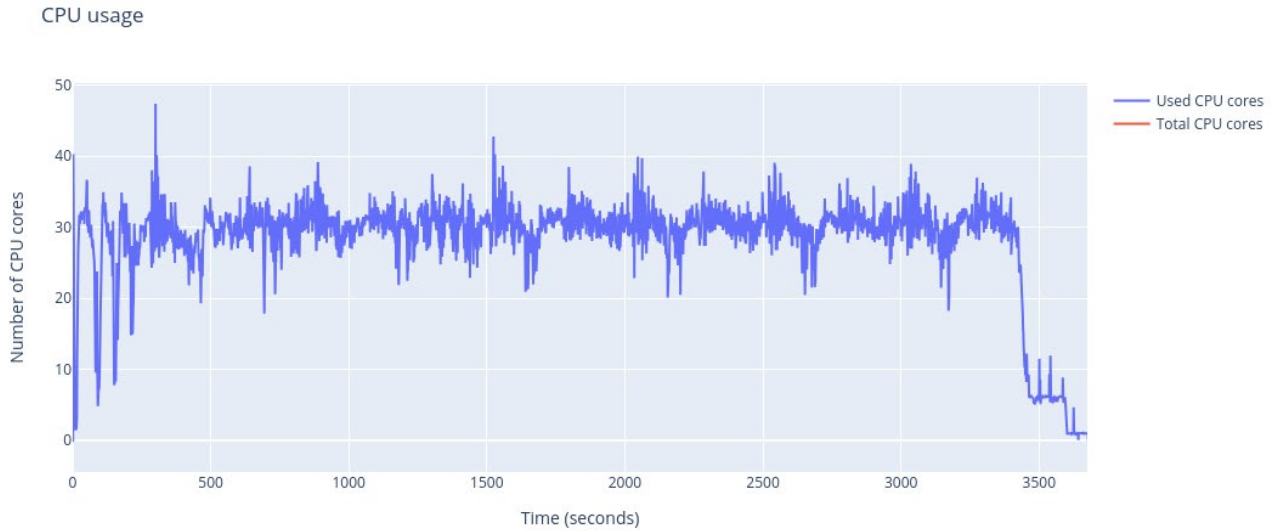
Dependency
prepare_struct
fetch_era5
fetch_seaice
convert_icemap
fetch_seviri
convert_seviri
run_pps

Workflow DAG

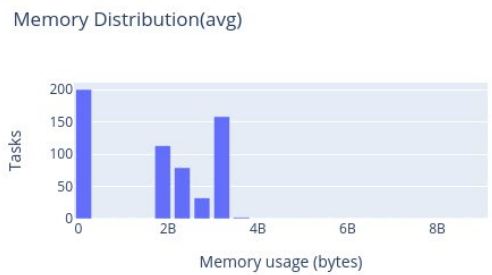


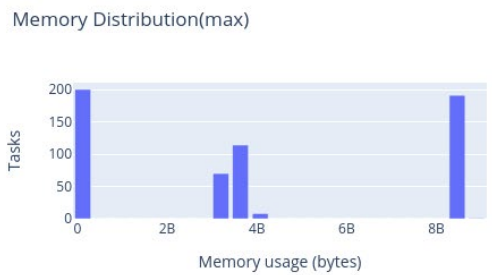
Dependency
dep_fail
failed
exec_done



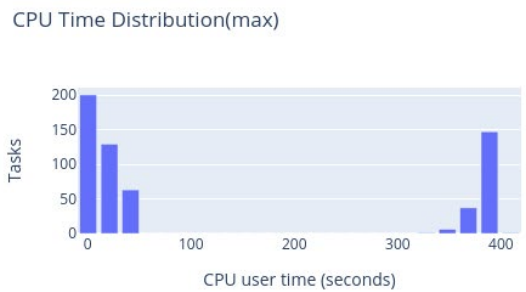


Memory Usage





CPU Usage



Docker + HTCondor

- container in HTCondor run with unprivileged user `nobody` in HTCondor specific workdir
  - all required files and folders need to be world readable and world executable (folders)
  - workdir `/tmp` in dockerfile helps testing outside HTCondor
  - use `_entrypoint.sh` as `ENTRYPOINT`
    - copy all required files to HTCondor workdir `cp /a/x /b/y /c/z .`
    - use HTCondor workdir to have access to condor I/O