# Data analytics workflows with the Ophidia Framework

**D. Elia**[1,2], **F. Antonio**[1]

[1] Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), Lecce, Italy
[2] University of Salento, Lecce, Italy

**ESiWACE2 online training course on High-Performance Data Analytics and Visualisation**

3rd session

22 October 2020

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

cmcc
Centro Euro-Mediterraneo sui Cambiamenti Climatici

# Session outline

- ✔ *Introduction to scientific analyses workflows and motivations*
- ✔ *Data analytics workflows in Ophidia*
- ✔ *Real-world examples of analytics workflow with the Ophidia framework*
  - ✔ *Multi-model climate experiment*
- ✔ *DEMO: Practical examples of simple workflow creation and integration with PyOphidia (Fabrizio)*
- ✔ *HANDS-ON: on data analytics workflows (Fabrizio)*

# Large scale climate analysis

Complexity of the analysis leads to the need for ***end-to-end workflow support***

- Typical approaches (mostly based on bash-like scripts) requires climate scientists to take care of implement and replicate workflow-like control logic
- Analyses can require the execution of *tens/hundreds of analytics operators*
  - *Efficient orchestration of the tasks is critical*
  - *Parallelism has to be handled both at intra-task and inter-task level*
  - *Task failure should also taken into account*

Workflows can represent a way to define ***portable*** and ***re-usable*** analyses (targeting FAIR principles)



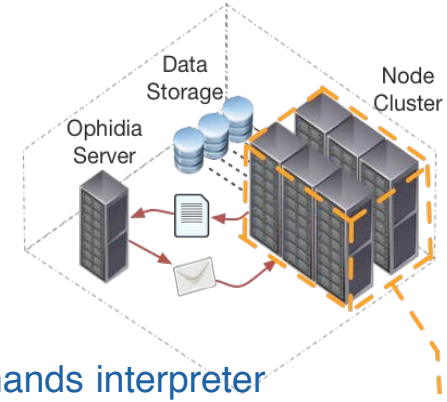
# Ophidia High-Performance Data Analytics Framework

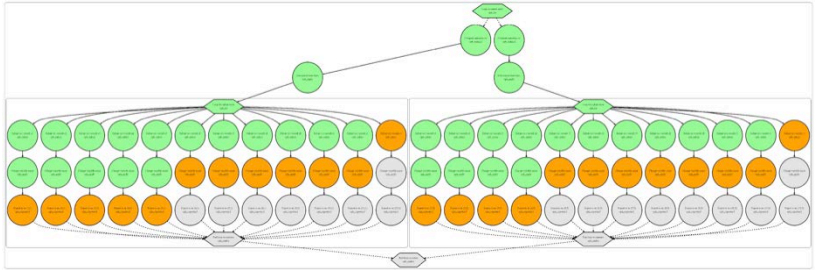
**Ophidia** (http://ophidia.cmcc.it) is a CMCC Foundation research project addressing data challenges for eScience

- a *High-Performance Data Analytics* (HPDA) framework for multi-dimensional scientific data joining HPC paradigms with scientific data analytics approaches
- in-memory and server-side data analysis exploiting parallel computing techniques and database approaches
- a multi-dimensional, array-based, storage model and partitioning schema for scientific data leveraging the datacube abstraction
- end-to-end mechanisms to **support complex experiments and large workflows on scientific datacubes**, primarily in climate domain

Ophidia

# Server-side paradigm and the datacube abstraction

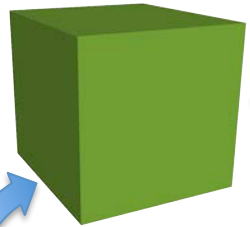



**Oph_Term**: a terminal-like commands interpreter serving as a client for the Ophidia framework

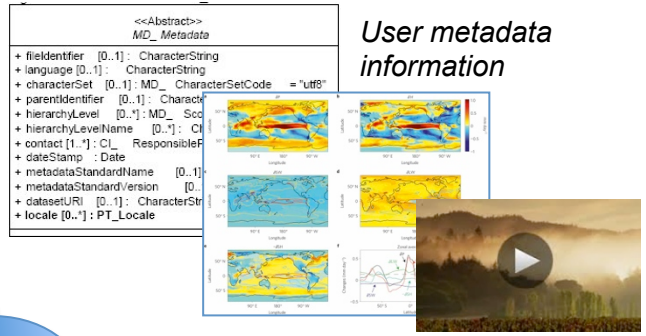
**PyOphidia**: a Python interface for datacube management & analytics with Ophidia

Through **oph_term/PyOphidia** the user run ("send") commands ("operators") to the Ophidia framework to manipulate datasets ("datacubes")

Three interaction modes:
*Operators, **Workflows**, Python Apps*



*System metadata of the datacube (size, distribution, etc.)*




*User metadata information*

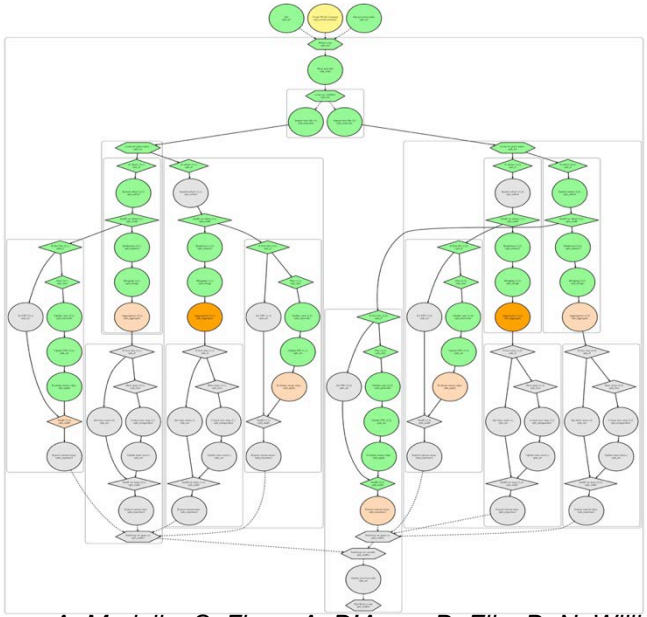
***Metadata provenance***

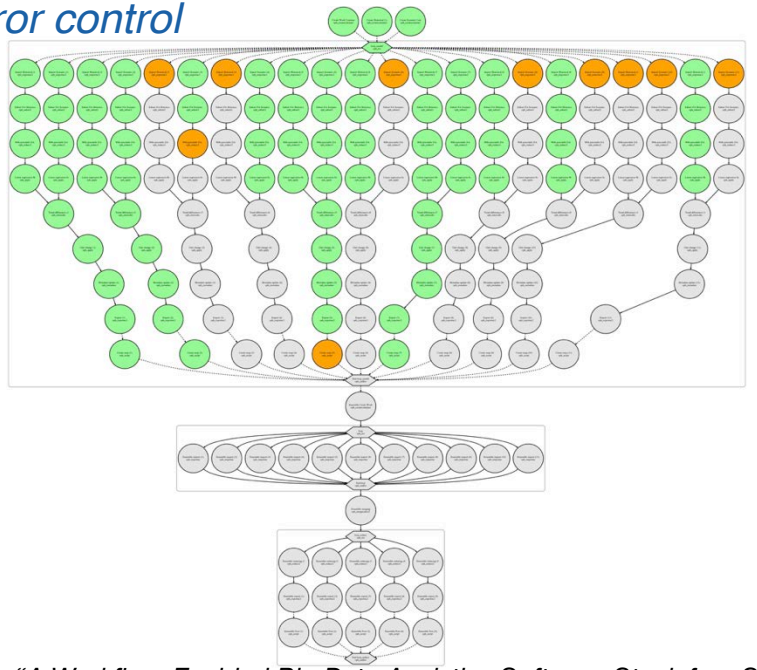
```
--> https://ophidia.cmcc.it:8443/162/169 (ROOT)
  ├─ https://ophidia.cmcc.it:8443/162/170 (oph_reduce)
  │    └─ https://ophidia.cmcc.it:8443/162/171 (oph_merge)
  │         ├─ https://ophidia.cmcc.it:8443/162/172 (oph_aggregate2)
  │         └─ https://ophidia.cmcc.it:8443/162/173 (oph_rollup)
  │              ├─ https://ophidia.cmcc.it:8443/162/174 (oph_reduce)
  │              └─ https://ophidia.cmcc.it:8443/162/175 (oph_reduce)
  ├─ https://ophidia.cmcc.it:8443/162/176 (oph_aggregate)
  └─ https://ophidia.cmcc.it:8443/162/177 (oph_aggregate)
```

# Analytics workflows

Ophidia supports the execution of complex workflows of operators.

- It defines a **JSON representation** for the workflow DAG specification
- Supports different constructs: *dependencies; massive tasks; iterative (group of) tasks; parallel (group of) tasks; flow and error control*

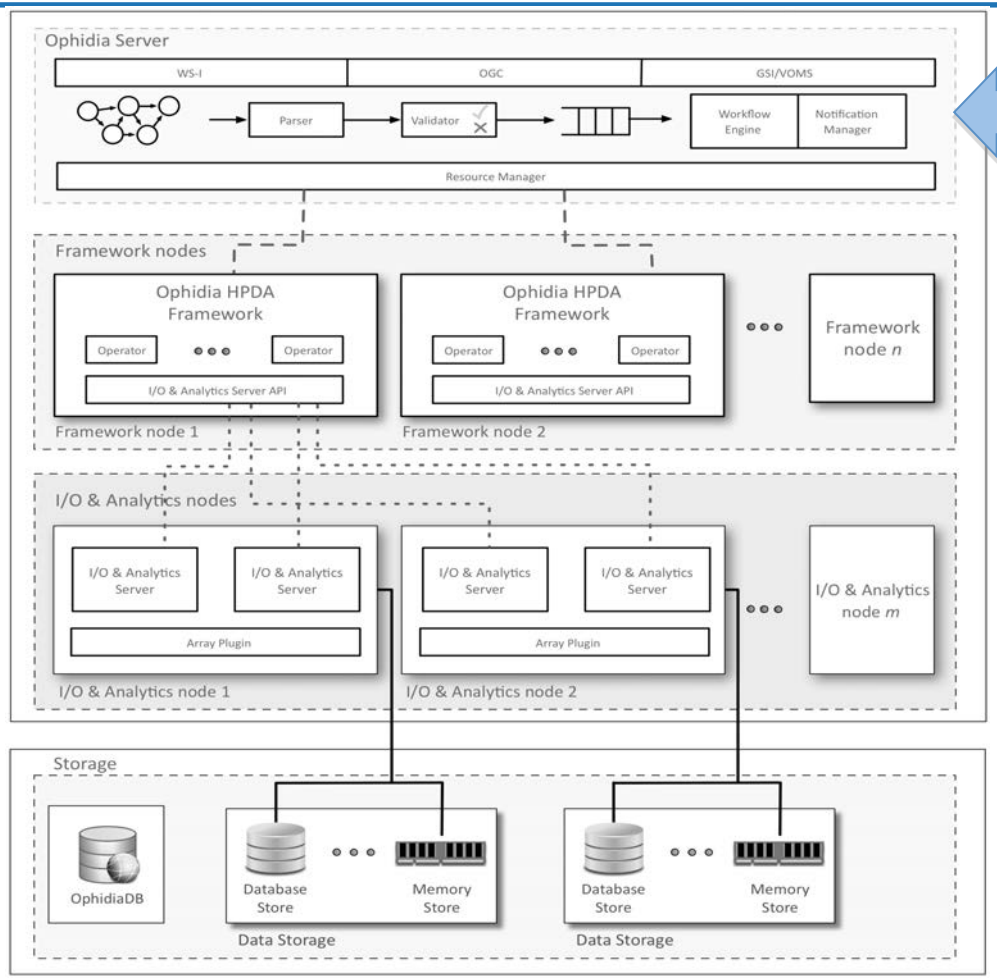




*C. Palazzo, A. Mariello, S. Fiore, A. D'Anca, D. Elia, D. N. Williams, G. Aloisio, "A Workflow-Enabled Big Data Analytics Software Stack for eScience", HPCS 2015, pp. 545-552*

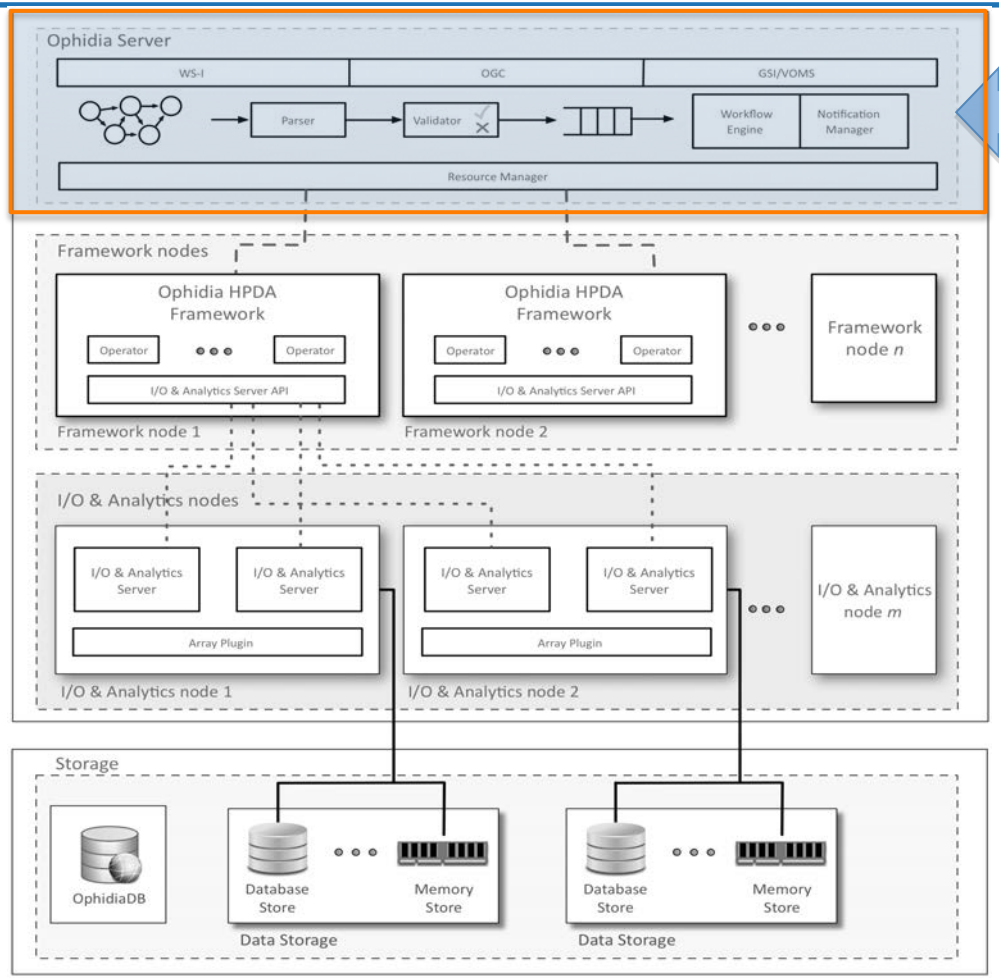
# Ophidia architecture: overview

- **Workflow** support on the server side
- Interoperable interface (OGC WPS)
- Modular and extensible software stack
- In-memory support
- Tasks: from single operators to complex analyses (workflows/apps)




# Ophidia architecture: overview

- **Workflow** support on the server side
- Interoperable interface (OGC WPS)
- Modular and extensible software stack
- In-memory support
- Tasks: from single operators to complex analyses (workflows/apps)




# The Ophidia Terminal

The **Ophidia Terminal**, a CLI bash-like client for the Ophidia framework:

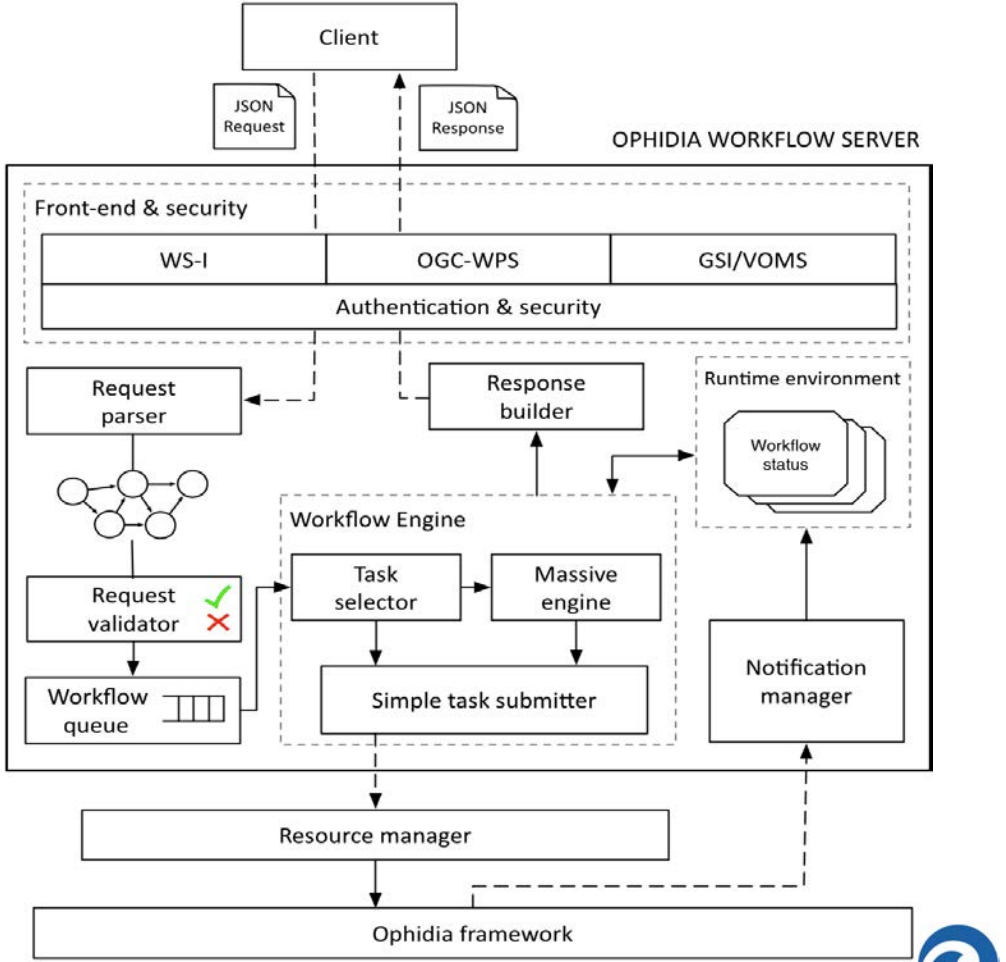
- Executing *interactive* data analytics sessions;
- Executing *batch* data analytics tasks of *workflows*;
- Experiment and operators *debugging*;
- *File system exploration* and *environment management*.

```
[11..4495] >> oph_list level=2;
[Request]:
operator=oph_list;path=;level=2;sessionid=http://127.0.0.1/ophidia/sessions/11123869522950595227155862
1818154495/experiment;exec_mode=sync;cdd=/;

[JobID]:
http://127.0.0.1/ophidia/sessions/11123869522950595227155862181815449 5/experiment?2#45

[Response]:
Ophidia Filesystem: /
--------------------
+===+====================+================================================+=============+
| T | PATH               | DATACUBE PID                                   | DESCRIPTION |
+===+====================+================================================+=============+
| c | test               | http://127.0.0.1/ophidia/2917/374976           |             |
+===+====================+================================================+=============+
```

# The Ophidia Server

The **workflow management system (WMS)** is a core component of the Ophidia Server:

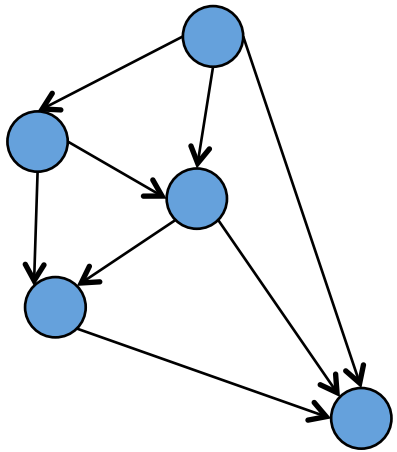
- manages **user request**
- **formats the commands** for the analytics framework
- handles task **dependencies** and **execution** flow
- **submits** the **tasks** to the resource manager
- **manages task status** updates
- provides the proper **response messages**



# Analytics Workflow Schema

Ophidia **workflows schema**:

- based on **JSON representation** for requests/responses
- defines application-level **semantic** and **syntactic rules**
- models scientific computations as **DAG**



**Main supported abstractions:**

- *Shared properties*
- *Flow/data dependencies*
- *Simple/massive tasks*
- *Iterative (group of) tasks*
- *Parallel (group of) tasks*
- *Flow and error control*
- *Interleaving and interactive tasks*

# Behind the scene: workflow JSON representation

ophrpm@ophidiarpm:~/workflow

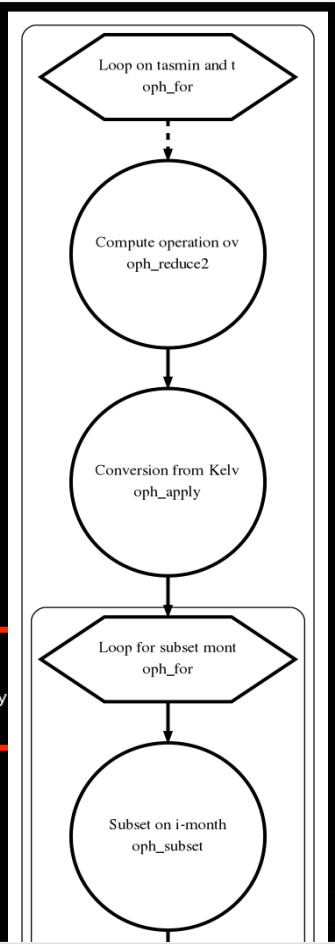
ophrpm@ophidiarpm:~/workflow

```
    "tasks": [
        {
            "name": "Loop on tasmin and tasmax cubes",
            "operator": "oph_for",
            "arguments": [ "name=cube","counter=1:2", "values=${1}|${2}", "parallel=yes" ]
        },
                {
                        "name": "Compute operation over time",
                        "operator": "oph_reduce2",
                        "arguments": [
                                "cube=@{cube}",
                                "dim=time",
                                "concept_level=M",
                                "midnight=00",
                                "operation=$3",
                                "container=tmp"
                        ],
                        "dependencies": [
                                { "task": "Loop on tasmin and tasmax cubes" }
                        ]
                },
                {
                        "name": "Conversion from Kelvin to Celsius degrees",
                        "operator": "oph_apply",
                        "arguments": [
                                "query=oph_sum_scalar('oph_float','oph_float',measure,-273.15)"
                        ],
                        "dependencies": [{
                                "task": "Compute operation over time",
                                "type": "single"
                        }]
                },
                {
                        "name": "Loop for subset months",
                        "operator": "oph_for",
                        "arguments": [ "name=index","counter=1:12", "values=Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec", "parallel=yes" ],
                        "dependencies": [ { "task":"Conversion from Kelvin to Celsius degrees", "type": "single" } ]
                },
                        {
                                "name": "Subset on i-month",
                                "operator": "oph_subset",
                                "arguments": [
                                        "subset_dims=time",
                                        "subset_filter=&index:12:end"
                                ],
                                "dependencies": [
--More--(65%)
```

# Behind the scene: workflow JSON representation

ophrpm@ophidiarpm:~/workflow

```
"tasks": [
  {
      "name": "Loop on tasmin and tasmax cubes",
      "operator": "oph_for",
      "arguments": [ "name=cube","counter=1:2", "values=${1}|${2}", "parallel=yes" ]
  },
  {
          "name": "Compute operation over time",
          "operator": "oph_reduce2",
          "arguments": [
                  "cube=@{cube}",
                  "dim=time",
                  "concept_level=M",
                  "midnight=00",
                  "operation=$3",
                  "container=tmp"
          ],
          "dependencies": [
                  { "task": "Loop on tasmin and tasmax cubes" }
          ]
  },
  {
          "name": "Conversion from Kelvin to Celsius degrees",
          "operator": "oph_apply",
          "arguments": [
                  "query=oph_sum_scalar('oph_float','oph_float',measure,-273.15)"
          ],
          "dependencies": [{
                  "task": "Compute operation over time",
                  "type": "single"
          }]
  },
  {
          "name": "Loop for subset months",
          "operator": "oph_for",
          "arguments": [ "name=index","counter=1:12", "values=Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec", "parallel=y
          "dependencies": [ { "task":"Conversion from Kelvin to Celsius degrees", "type": "single" } ]
  },
          {
                  "name": "Subset on i-month",
                  "operator": "oph_subset",
                  "arguments": [
                          "subset_dims=time",
                          "subset_filter=&index:12:end"
                  ],
                  "dependencies": [
--More--(65%)
```



# Analytics workflows constructs

## Workflow Management

This group includes a number of flow control operators that could be used within an Ophidia workflow to implement complex data processing in batch mode. In particular, they implement several advanced features: setting of run-time variables, iterative and parallel interface, selection interface, interactive workflows, interleaving workflows, etc.

| NAME | DESCRIPTION |
|---|---|
| OPH_ELSE | Start the last sub-block of a selection block "if". |
| OPH_ELSEIF | Start a new sub-block of a selection block "if". |
| OPH_ENDFOR | Close a loop "for". |
| OPH_ENDIF | Close a selection block "if". |
| OPH_FOR | Implement a loop "for". |
| OPH_IF | Open a "if" selection block. |
| OPH_INPUT | It sends commands or data to an interactive task. |
| OPH_SET | Set a parameter in the workflow environment. |
| OPH_WAIT | Wait until an event occurs. |

# Iterative Interface

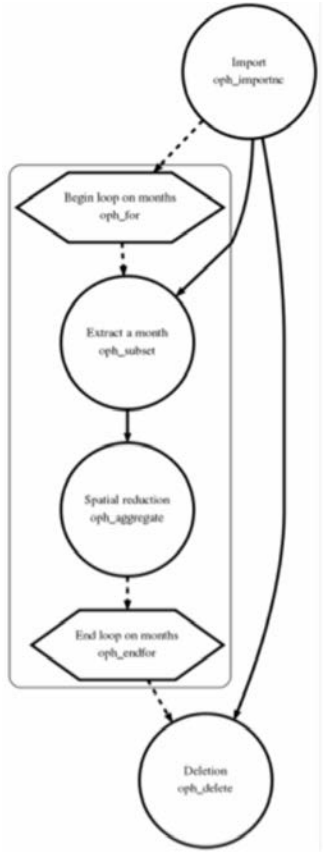
Allows to repeat the execution of a block of workflow tasks over different input data or over the result of the previous iteration.

Selection interface operators:

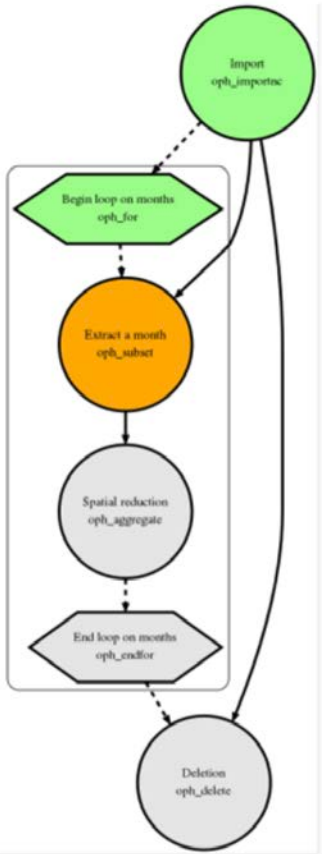
- *OPH_FOR*
- *OPH_ENDFOR*

*The statement can be used in nested fashion*

**AT DEFINITION TIME**




**AT RUNTIME**



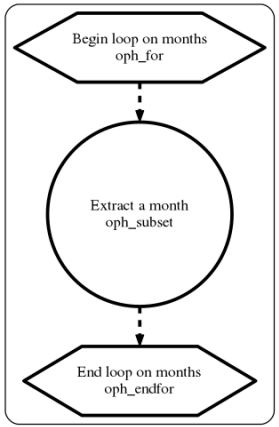

*Workflow iterative interface documentation:* http://ophidia.cmcc.it/documentation/users/workflow/workflow_for.html

# Parallel Interface

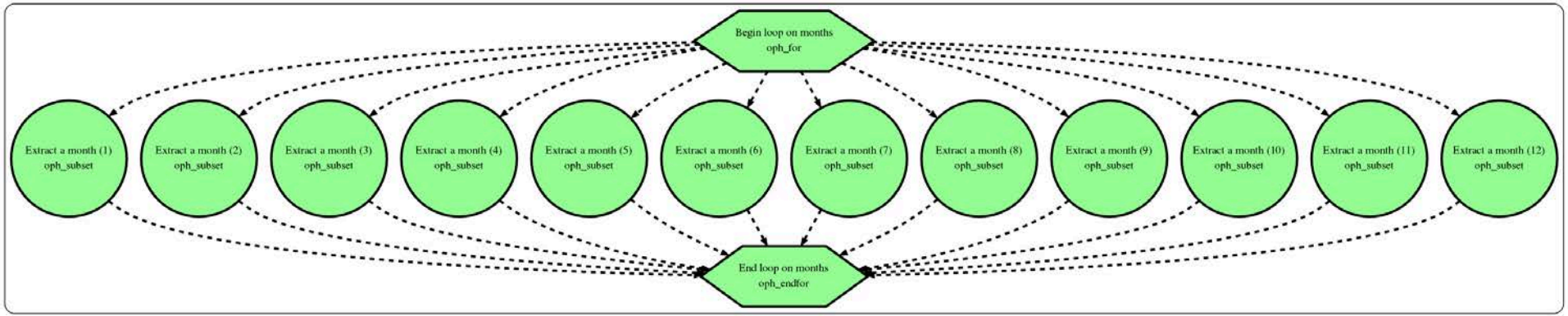
Extension of the OPH_FOR interface for parallel (concurrent) execution of the loop iterations.

```
{
        "name": "Begin loop on months",
        "operator": "oph_for",
        "arguments":
        [
                "parallel=yes",
                "name=index",
                "counter=1:12",
                "values=Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec"
        ]
}
```

## AT DEFINITION TIME




## AT RUNTIME




*Workflow parallel interface documentation:* http://ophidia.cmcc.it/documentation/users/workflow/workflow_for.html#parallel-interface

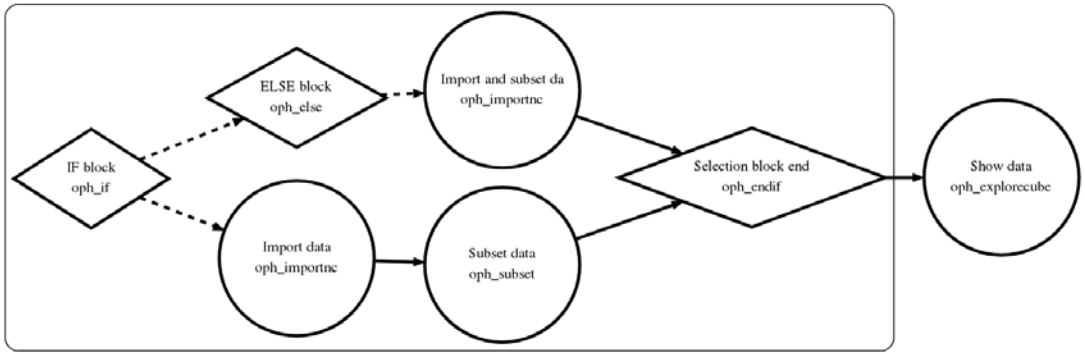
# Selection Interface

Enables the workflow manager to dynamically execute a block of tasks based on boolean conditions evaluated at run-time.

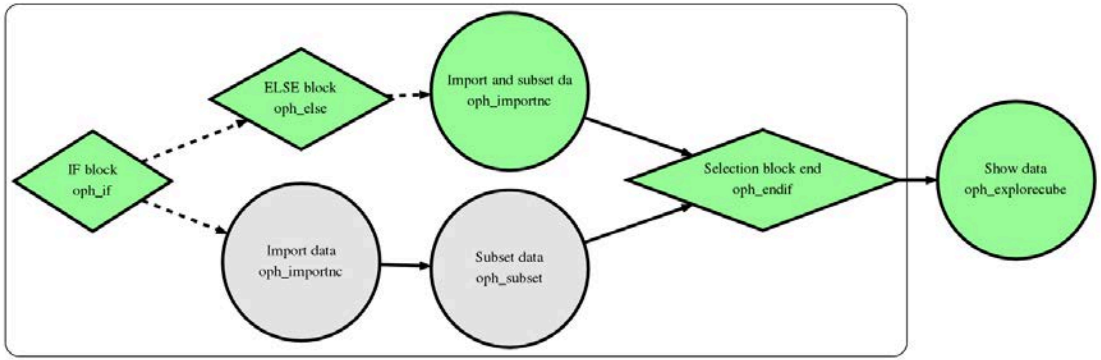
Selection interface operators:

- *OPH_IF*
- *OPH_ELSEIF*
- *OPH_ELSE*
- *OPH_ENDIF*

## AT DEFINITION TIME

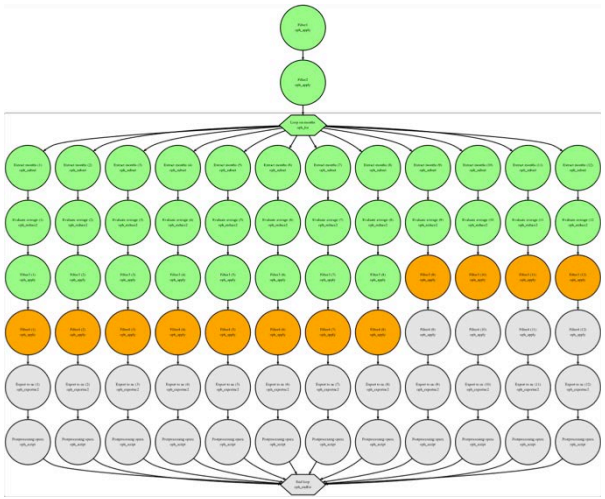


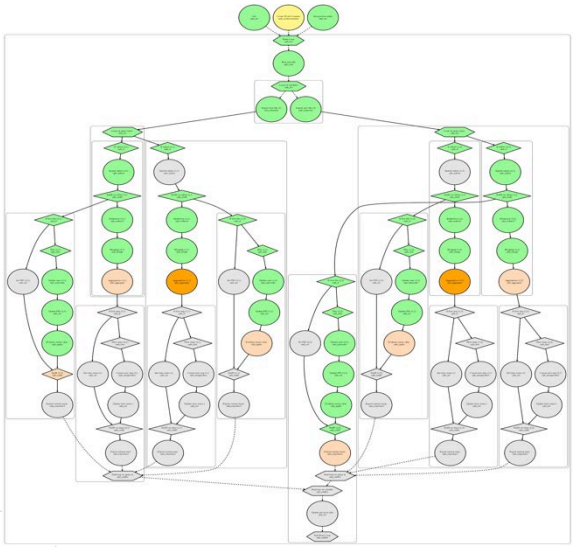

## AT RUNTIME

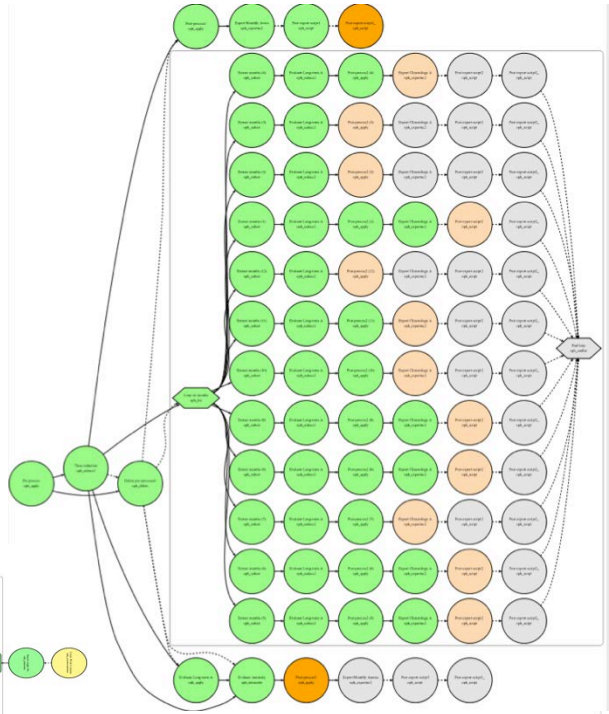


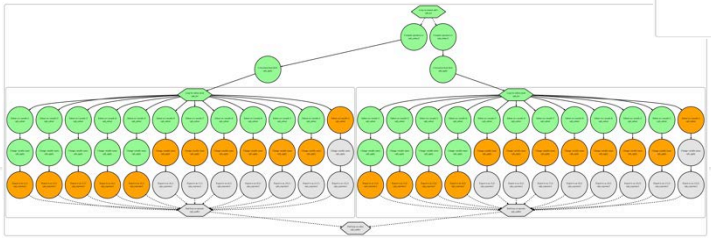

*Workflow selection interface documentation:* http://ophidia.cmcc.it/documentation/users/workflow/workflow_if.html

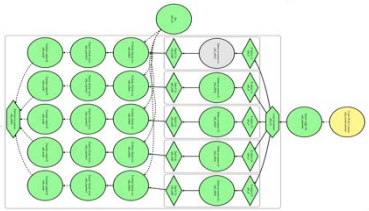
# Analytics workflows support and interfaces

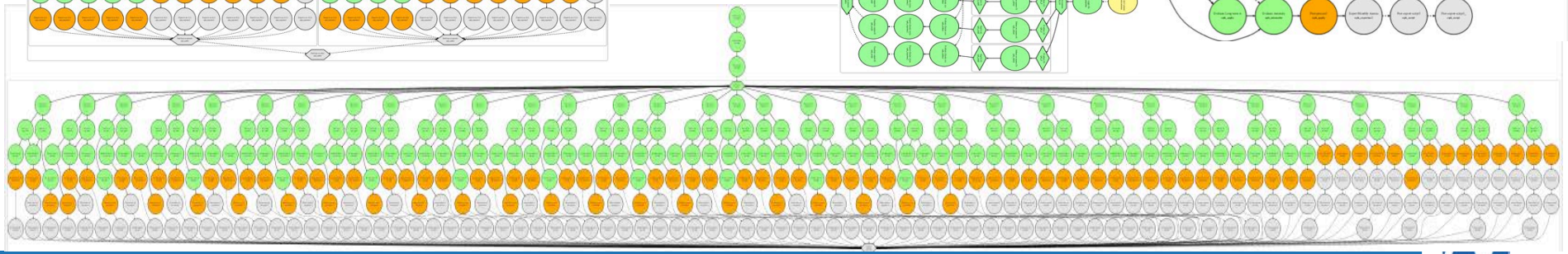




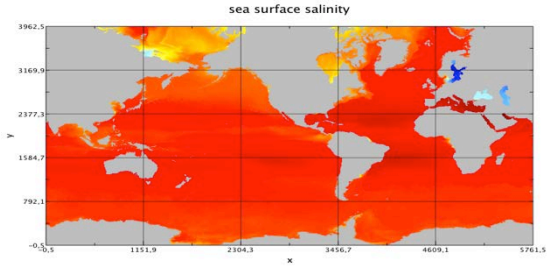


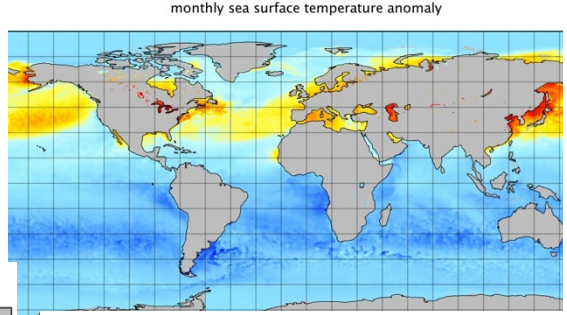


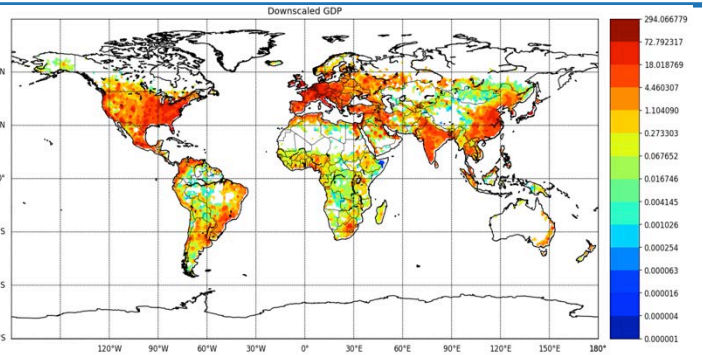


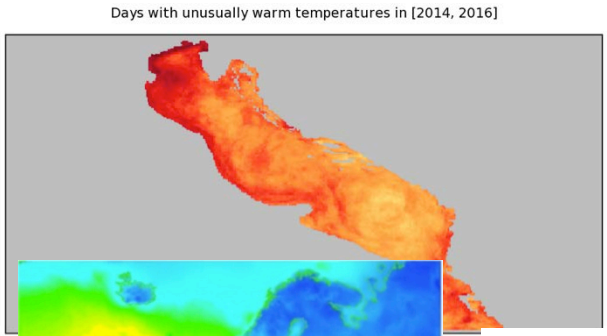


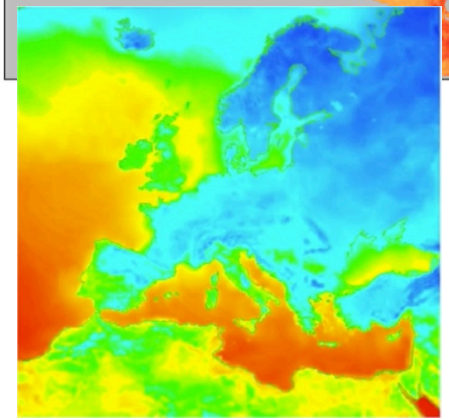
# Efficient support for advanced analytics experiments

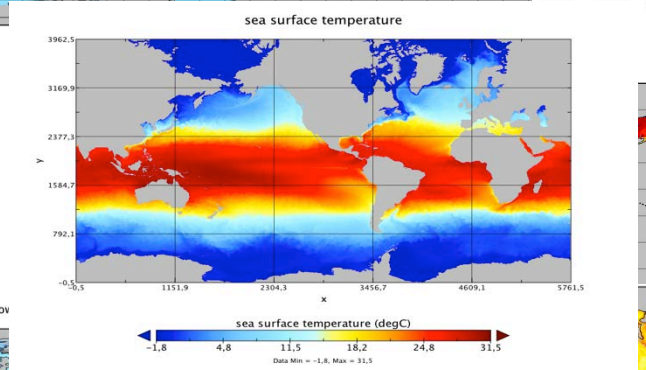


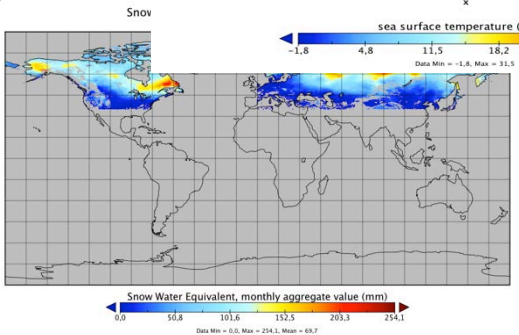



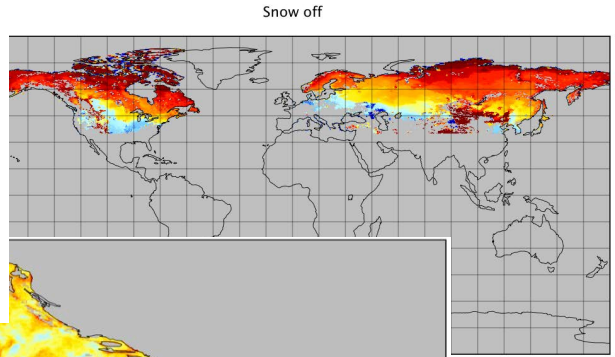



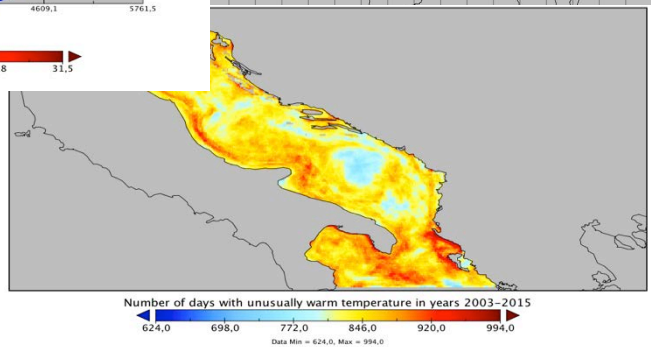



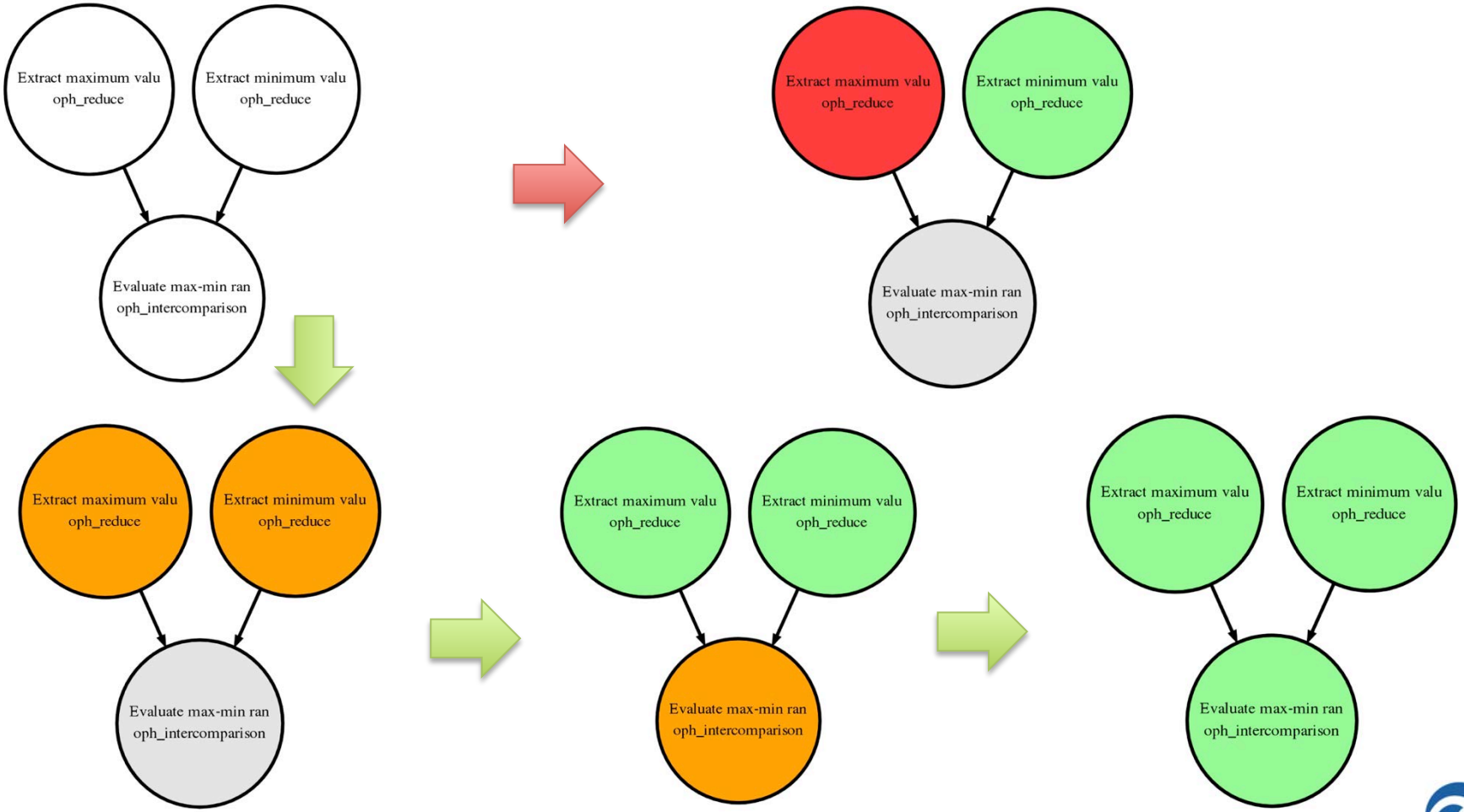


# Workflow status monitoring



# Workflow submission

ophrpm@ophidiarpm:~/devel/oph-client/res | ophrpm@ophidiarpm:~/workflow

```
[37..6380] >>
[37..6380] >> ./Tind_loop.json http://193.204.199.174/ophidia/29/2046 http://193.204.199.174/ophidia/30/2047 max
[JobID]:
http://193.204.199.174/ophidia/sessions/376699238311302232511449455166146380/experiment?247#3144

[37..6380] >> view 247
[247] ./Tind_loop.json http://193.204.199.174/ophidia/29/2046 http://193.204.199.174/ophidia/30/2047 max [http://193.204.199.174/ophidia/sessions/376699238311302
232511449455166146380/experiment?247#3144]

[Response]:
Workflow Status
---------------

OPH_STATUS_COMPLETED

Workflow Progress
-----------------

+==========================+======================+
| NUMBER OF COMPLETED TASKS | TOTAL NUMBER OF TASKS |
+==========================+======================+
| 82                       | 82                   |
+==========================+======================+

Workflow Task List
------------------

+=====================================================+=====================+========+=======+===========+============================+=====+=============+
| OPH JOB ID                                          | SESSION CODE        | WORKFL | MARKE | PARENT MA | TASK NAME                  | TYP | EXIT STATUS |
|                                                     |                     | OW ID  | R ID  | RKER ID   |                            | E   |             |
+=====================================================+=====================+========+=======+===========+============================+=====+=============+
| http://193.204.199.174/ophidia/sessions/37669923831130 | 37669923831130223251 | 247    | 3145  | 3144      | Loop on tasmin and tasmax  | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3145          | 1449455166146380    |        |       |           | cubes                      | PLE | COMPLETED   |
|-----------------------------------------------------|---------------------|--------|-------|-----------|----------------------------|-----|-------------|
| http://193.204.199.174/ophidia/sessions/37669923831130 | 37669923831130223251 | 247    | 3146  | 3144      | Compute operation over ti  | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3146          | 1449455166146380    |        |       |           | me (1)                     | PLE | COMPLETED   |
|-----------------------------------------------------|---------------------|--------|-------|-----------|----------------------------|-----|-------------|
| http://193.204.199.174/ophidia/sessions/37669923831130 | 37669923831130223251 | 247    | 3147  | 3144      | Compute operation over ti  | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3147          | 1449455166146380    |        |       |           | me (2)                     | PLE | COMPLETED   |
|-----------------------------------------------------|---------------------|--------|-------|-----------|----------------------------|-----|-------------|
| http://193.204.199.174/ophidia/sessions/37669923831130 | 37669923831130223251 | 247    | 3148  | 3144      | Conversion from Kelvin to  | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3148          | 1449455166146380    |        |       |           | Celsius degrees (1)        | PLE | COMPLETED   |
|-----------------------------------------------------|---------------------|--------|-------|-----------|----------------------------|-----|-------------|
| http://193.204.199.174/ophidia/sessions/37669923831130 | 37669923831130223251 | 247    | 3149  | 3144      | Conversion from Kelvin to  | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3149          | 1449455166146380    |        |       |           | Celsius degrees (2)        | PLE | COMPLETED   |
|-----------------------------------------------------|---------------------|--------|-------|-----------|----------------------------|-----|-------------|
```

# Workflow submission

ophrpm@ophidiarpm:~/devel/oph-client/res | ophrpm@ophidiarpm:~/workflow

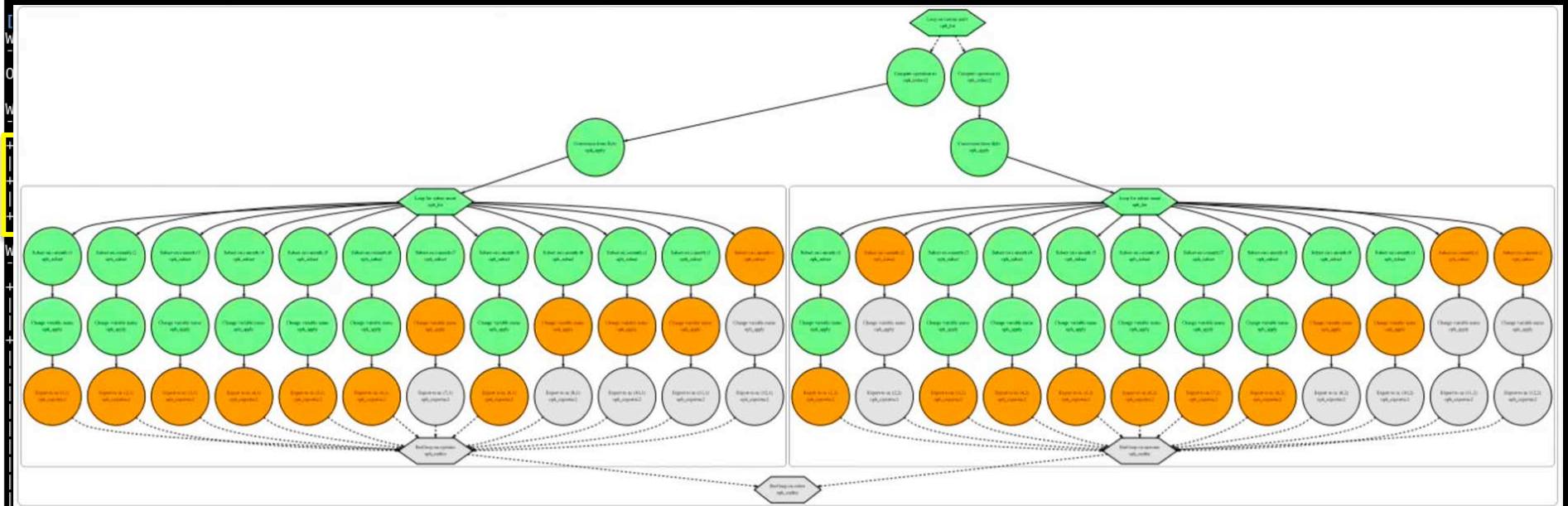
```
[37..6380] >> ./Tind_loop.json http://193.204.199.174/ophidia/29/2046 http://193.204.199.174/ophidia/30/2047 max
[JobID]:
http://193.204.199.174/ophidia/sessions/37669923831130223251144945516614638O/experiment?247#3144

[37..6380] >> view 247
[247] ./Tind_loop.json http://193.204.199.174/ophidia/29/2046 http://193.204.199.174/ophidia/30/2047 max [http://193.204.199.174/ophidia/sessions/37669923831130223251144945516614638O/experiment?247#3144]

[Response]:
Workflow Status
---------------

OPH_STATUS_COMPLETED

Workflow Progress
-----------------

+==========================+======================+
| NUMBER OF COMPLETED TASKS | TOTAL NUMBER OF TASKS |
+==========================+======================+
| 82                       | 82                   |
+==========================+======================+

Workflow Task List
------------------

+==========================================================+======================+========+=======+===========+============================+=====+=============+
| OPH JOB ID                                               | SESSION CODE         | WORKFL | MARKE | PARENT MA | TASK NAME                  | TYP | EXIT STATUS |
|                                                          |                      | OW ID  | R ID  | RKER ID   |                            | E   |             |
+==========================================================+======================+========+=======+===========+============================+=====+=============+
| http://193.204.199.174/ophidia/sessions/37669923831130   | 37669923831130223251 | 247    | 3145  | 3144      | Loop on tasmin and tasmax  | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3145               | 1449455166146380     |        |       |           | cubes                      | PLE | COMPLETED   |
|----------------------------------------------------------|----------------------|--------|-------|-----------|----------------------------|-----|-------------|
| http://193.204.199.174/ophidia/sessions/37669923831130   | 37669923831130223251 | 247    | 3146  | 3144      | Compute operation over ti  | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3146               | 1449455166146380     |        |       |           | me (1)                     | PLE | COMPLETED   |
|----------------------------------------------------------|----------------------|--------|-------|-----------|----------------------------|-----|-------------|
| http://193.204.199.174/ophidia/sessions/37669923831130   | 37669923831130223251 | 247    | 3147  | 3144      | Compute operation over ti  | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3147               | 1449455166146380     |        |       |           | me (2)                     | PLE | COMPLETED   |
|----------------------------------------------------------|----------------------|--------|-------|-----------|----------------------------|-----|-------------|
| http://193.204.199.174/ophidia/sessions/37669923831130   | 37669923831130223251 | 247    | 3148  | 3144      | Conversion from Kelvin to  | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3148               | 1449455166146380     |        |       |           | Celsius degrees (1)        | PLE | COMPLETED   |
|----------------------------------------------------------|----------------------|--------|-------|-----------|----------------------------|-----|-------------|
| http://193.204.199.174/ophidia/sessions/37669923831130   | 37669923831130223251 | 247    | 3149  | 3144      | Conversion from Kelvin to  | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3149               | 1449455166146380     |        |       |           | Celsius degrees (2)        | PLE | COMPLETED   |
|----------------------------------------------------------|----------------------|--------|-------|-----------|----------------------------|-----|-------------|
```

# Workflow submission

ophrpm@ophidiarpm:~/devel/oph-client/res

```
[37..6380] >> ./Tind_loop.json http://193.204.199.174/ophidia/29/2046 http://193.204.199.174/ophidia/30/2047 max
[JobID]:
http://193.204.199.174/ophidia/sessions/37669923831130223251144945516614638O/experiment?247#3144

[37..6380] >> view 247
[247] ./Tind_loop.json http://193.204.199.174/ophidia/29/2046 http://193.204.199.174/ophidia/30/2047 max [http://193.204.199.174/ophidia/sessions/37669923831130223251144945516614638O/experiment?247#3144]
```

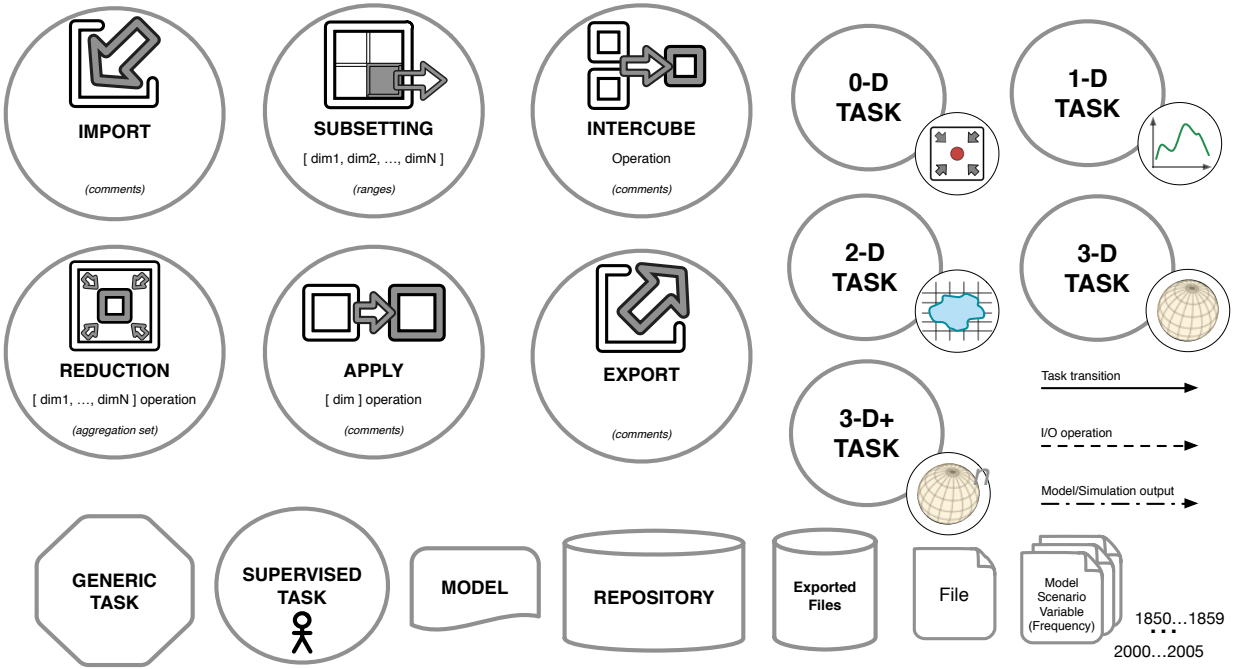
ophrpm@ophidiarpm:~/workflow



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| http://193.204.199.174/ophidia/sessions/37669923831130 2232511449455166146380/experiment?247#3148 | 37669923831130223251 1449455166146380 | 247 | 3148 | 3144 | Conversion from Kelvin to Celsius degrees (1) | SIMPLE | OPH_STATUS_COMPLETED |
| http://193.204.199.174/ophidia/sessions/37669923831130 2232511449455166146380/experiment?247#3149 | 37669923831130223251 1449455166146380 | 247 | 3149 | 3144 | Conversion from Kelvin to Celsius degrees (2) | SIMPLE | OPH_STATUS_COMPLETED |

# Analytics Workflow modeling

- *A Data Analytics Workflow Modelling Language (DAWML) has been defined*
- ***Extensible** schema jointly defined with application-domain scientists*
- *Provides an **abstraction** for the definition of workflows*



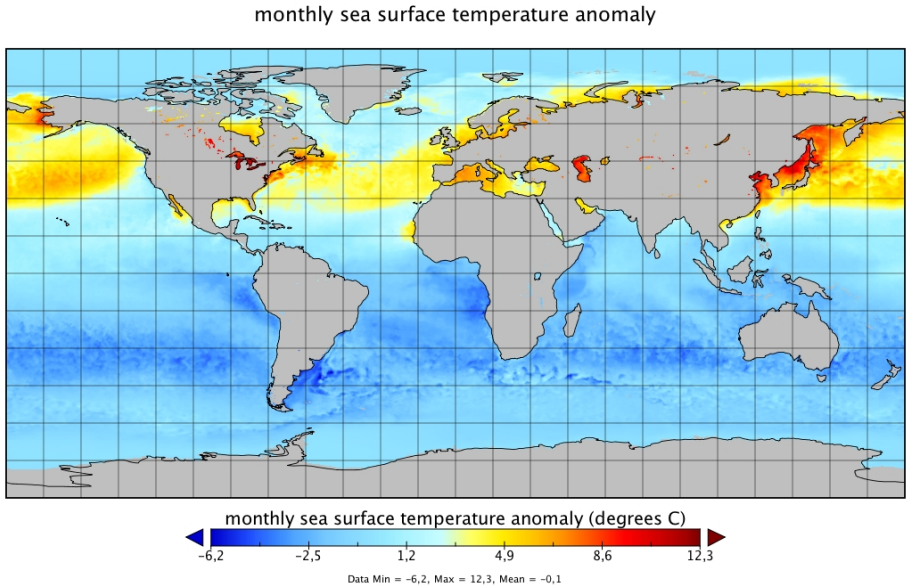

*C. Palazzo, A. Mariello, S. Fiore, A. D'Anca, D. Elia, D. N. Williams, G. Aloisio, "A Workflow-Enabled Big Data Analytics Software Stack for eScience", HPCS 2015, pp. 545-552*

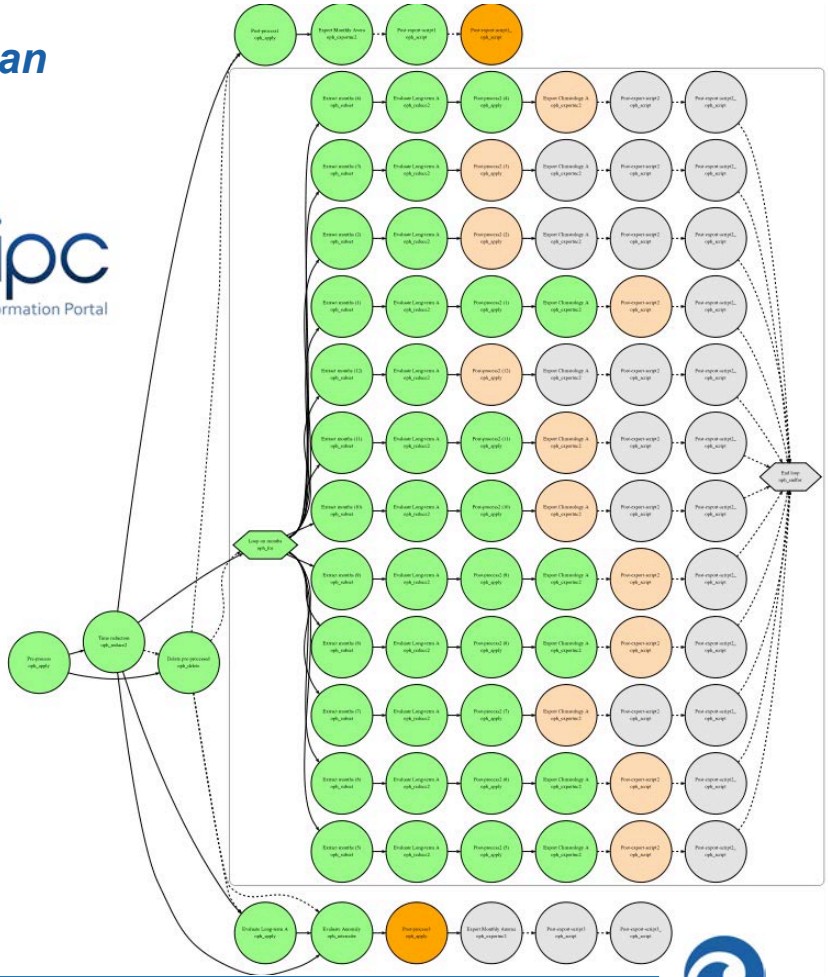
# *Some real-world analytics workflows examples*

# Workflow example I: climate indicators processing

## *SST (monthly) mean, anomaly, climatological mean*

- *Dataset time range: 1991-2010*
- ***7062** nc files*
- ***350GB** of input data*
- ***87 tasks** performed*
- *12x51MB + 2x12GB of output files*

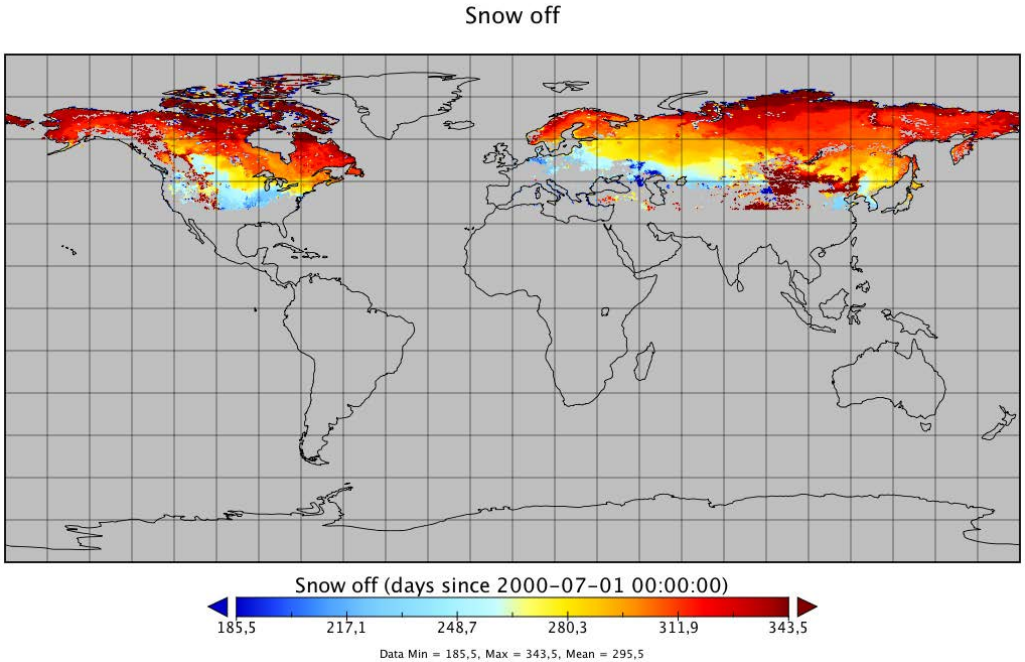


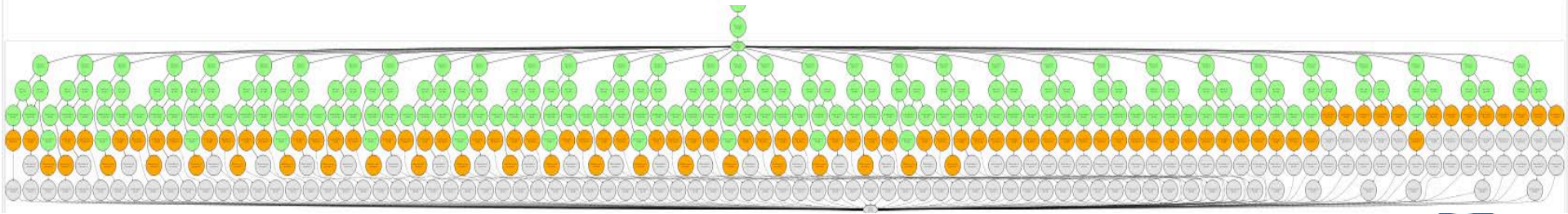
# Workflow example II: climate indicators processing

***Snow on/off – Length of snow season***
*(single workflow for 3 indicators)*

- *Dataset time range: 1979-2012*
- ***6341** nc files*
- ***50 GB** of input data*
- ***599 tasks** performed*
- ***99** NetCDF output files (6MB each)*
- ***21** tasks in the exp. description*

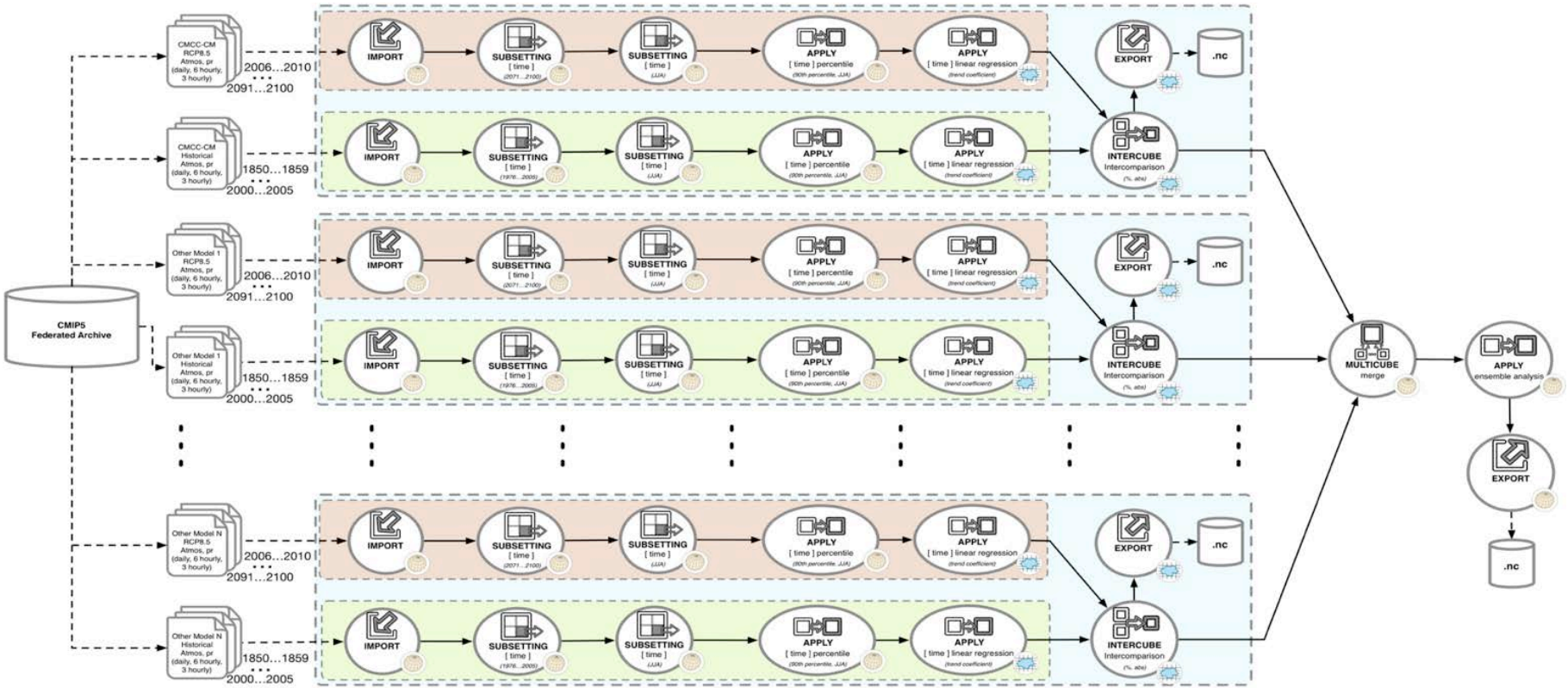



Climate Information Portal



# Workflow example III: Multi-model experiment design

**Precipitation Trend Analysis use case** implemented as an Ophidia workflow

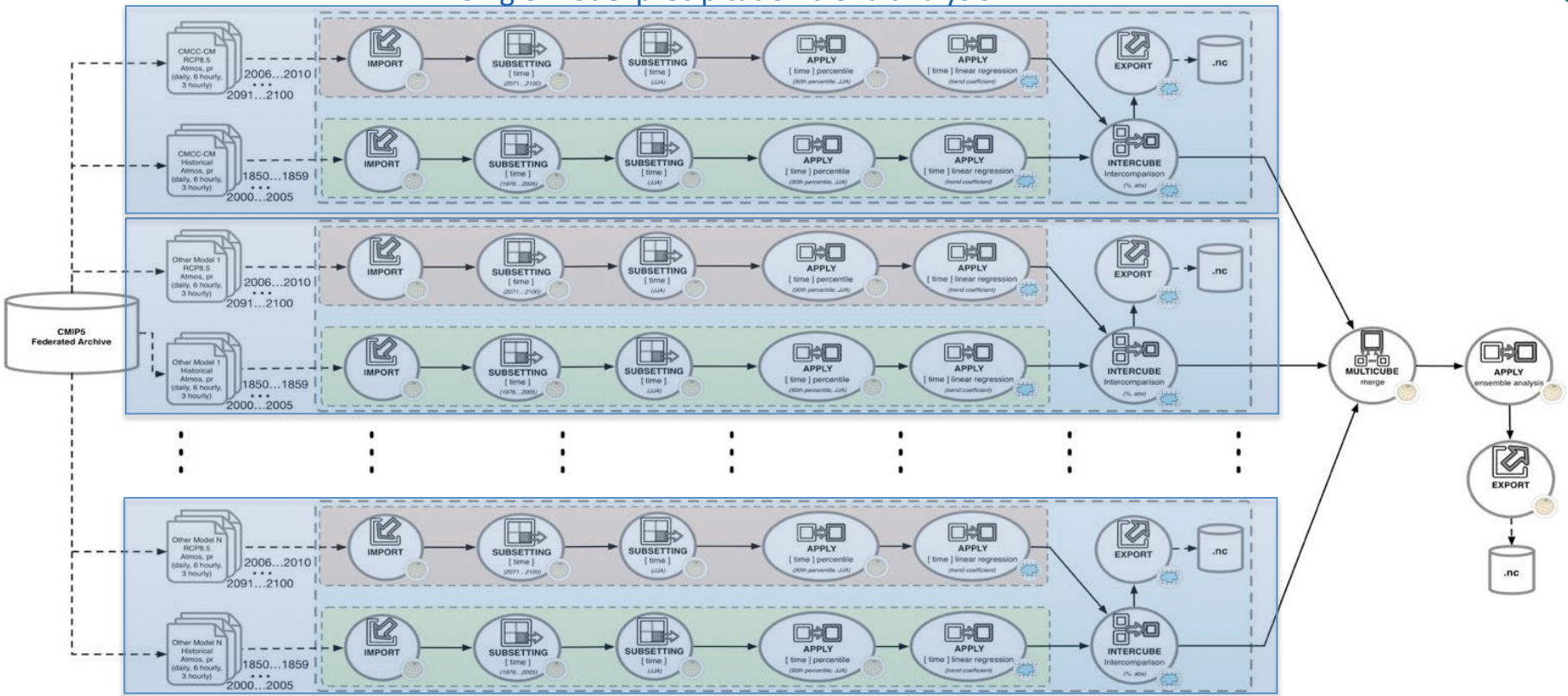



*S. Fiore, et al., "Distributed and cloud-based multi-model analytics experiments on large volumes of climate change data in the earth system grid federation eco-system". In Big Data (Big Data), 2016 IEEE Int. Conference on. IEEE, 2016. pp. 2911-2918*

# Workflow example III: Multi-model experiment design

**Precipitation Trend Analysis use case** implemented as an Ophidia workflow

Single model precipitation trend analysis

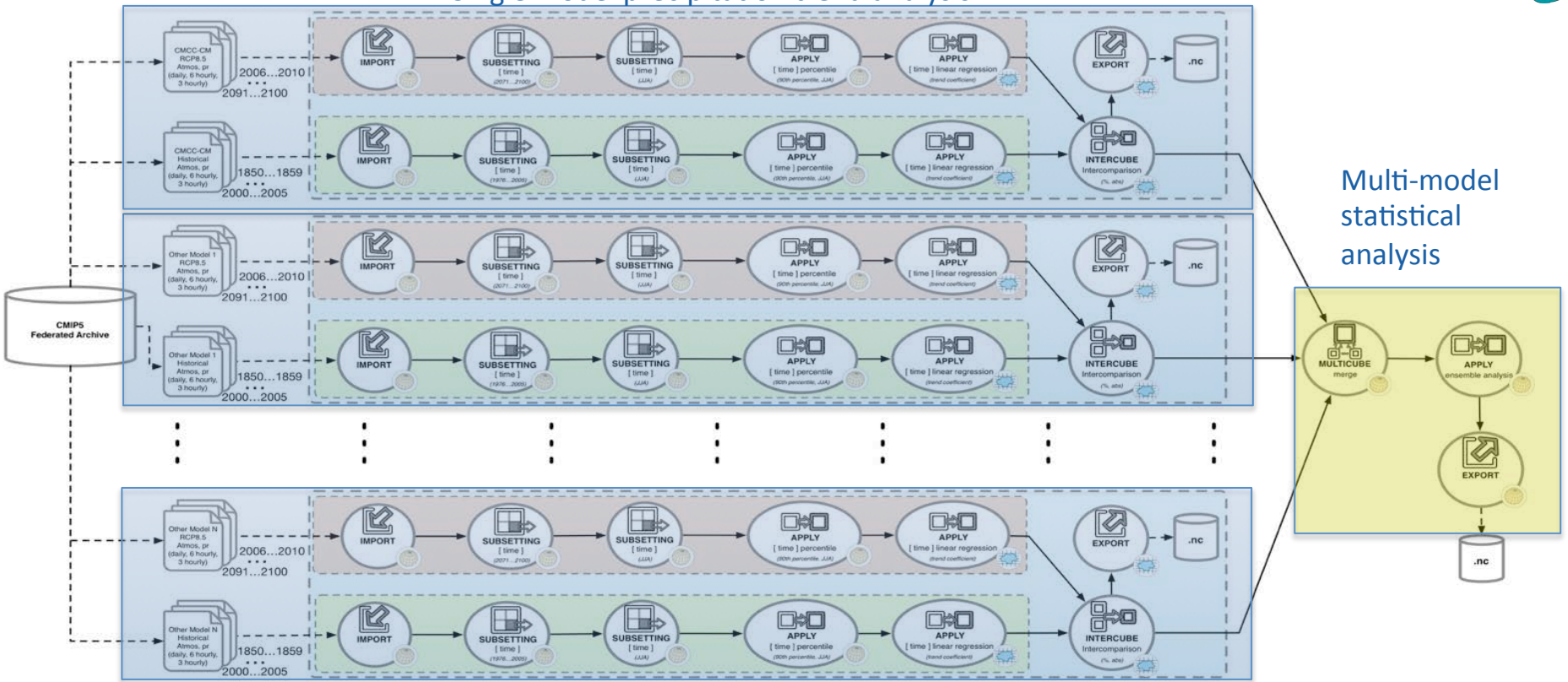



*S. Fiore, et al., "Distributed and cloud-based multi-model analytics experiments on large volumes of climate change data in the earth system grid federation eco-system". In Big Data (Big Data), 2016 IEEE Int. Conference on. IEEE, 2016. pp. 2911-2918*

# Workflow example III: Multi-model experiment design

**Precipitation Trend Analysis use case** implemented as an Ophidia workflow

Single model precipitation trend analysis

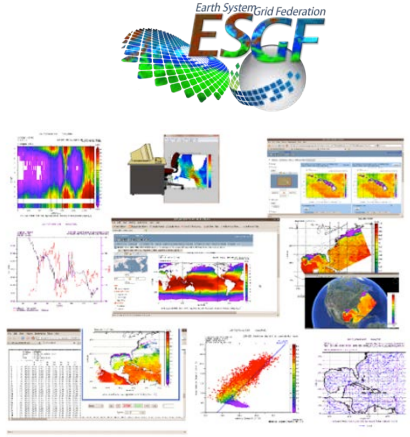



Multi-model statistical analysis

*S. Fiore, et al., "Distributed and cloud-based multi-model analytics experiments on large volumes of climate change data in the earth system grid federation eco-system". In Big Data (Big Data), 2016 IEEE Int. Conference on. IEEE, 2016. pp. 2911-2918*

# Multi-model experiment input data

**ESGF[1] is a coordinated multiagency, international collaboration of institutions that continually develop, deploy, and maintain software needed to facilitate and empower the study of climate.**

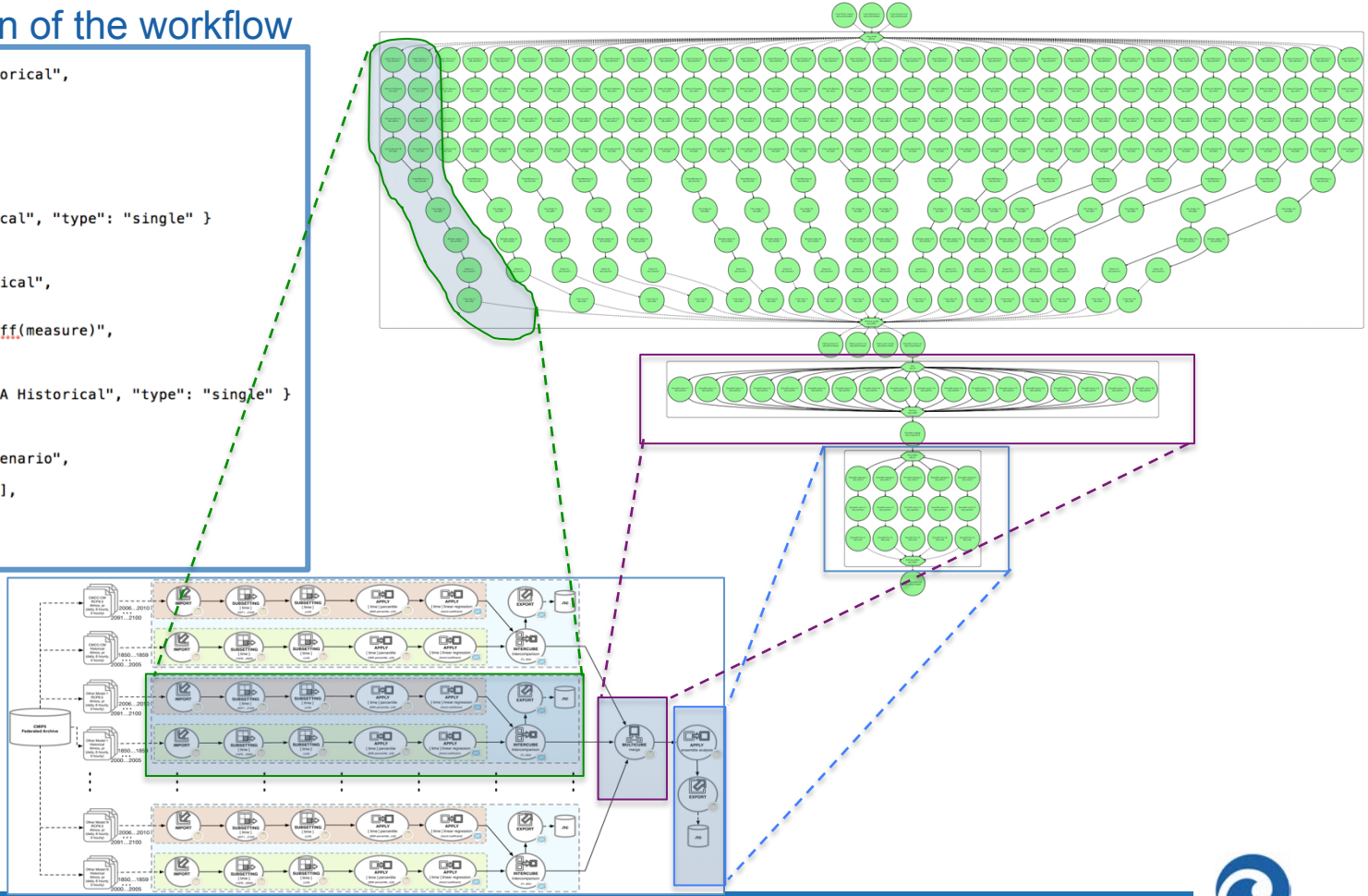


| Model acronym | Model expansion | Institute |
|---|---|---|
| *CCSM4* | Community Climate System Model, v4 | National Center for Atmospheric Research (NCAR) |
| *CMCC-CESM* | CMCC - Community Earth System Model | Euro-Mediterranean Center on Climate Change (CMCC) |
| *CMCC-CMS* | CMCC - Coupled Modeling System | Euro-Mediterranean Center on Climate Change (CMCC) |
| *CMCC-CM* | CMCC - Climate Model | Euro-Mediterranean Center on Climate Change (CMCC) |
| *CNRM-CM5* | CNRM - Coupled Global Climate Model, v5 | Centre National de Recherches Météorologiques (CNRM)/Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS) |
| *CSIRO Mk3.6.0* | CSIRO Mark, v3.6.0 | Commonwealth Scientific and Industrial Research Organisation (CSIRO) in collaboration with Queensland Climate-Change Centre of Excellence (QCCCE) |
| *CanESM2* | Second Generation Canadian Earth System Model | Canadian Centre for Climate Modelling and Analysis (CCCma) |
| *GFDL-CM3* | GFDL Climate Model, v3 | National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL) |
| *GFDL-ESM2G* | GFDL Earth System Model with Generalized Ocean Layer Dynamics (GOLD) component | National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL) |
| *GFDL-ESM2M* | GFDL Earth System Model with Modular Ocean Model 4 (MOM4) component | National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL) |
| *HadGEM2-CC* | Hadley Centre Global Environment Model, v2 (Carbon Cycle) | Met Office (UKMO) Hadley Centre (HC) |
| *HadGEM2-ES* | Hadley Centre Global Environment Model, v2 (Earth System) | Met Office (UKMO) Hadley Centre (HC) |
| *INM-CM4.0* | INM Coupled Model, v4.0 | Institute of Numerical Mathematics (INM) |
| *IPSL-CM5A-MR* | IPSL Coupled Model, version 5, coupled with NEMO, mid resolution | L'Institut Pierre-Simon Laplace (IPSL) |
| *MIROC5* | Model for Interdisciplinary Research on Climate, v5 | Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology |
| *MPI-ESM-MR* | MPI Earth System Model, medium resolution | Max Planck Institute for Meteorology (MPI-M) |
| *MRI-CGCM3* | MRI Coupled Atmosphere - Ocean General Circulation Model, v3 | Meteorological Research Institute (MRI) |
| *NorESM1-M* | Norwegian Earth System Model, v1 (intermediate resolution) | Norwegian Climate Centre (NCC) |

# Multi-model experiment implementation & execution

## JSON implementation of the workflow

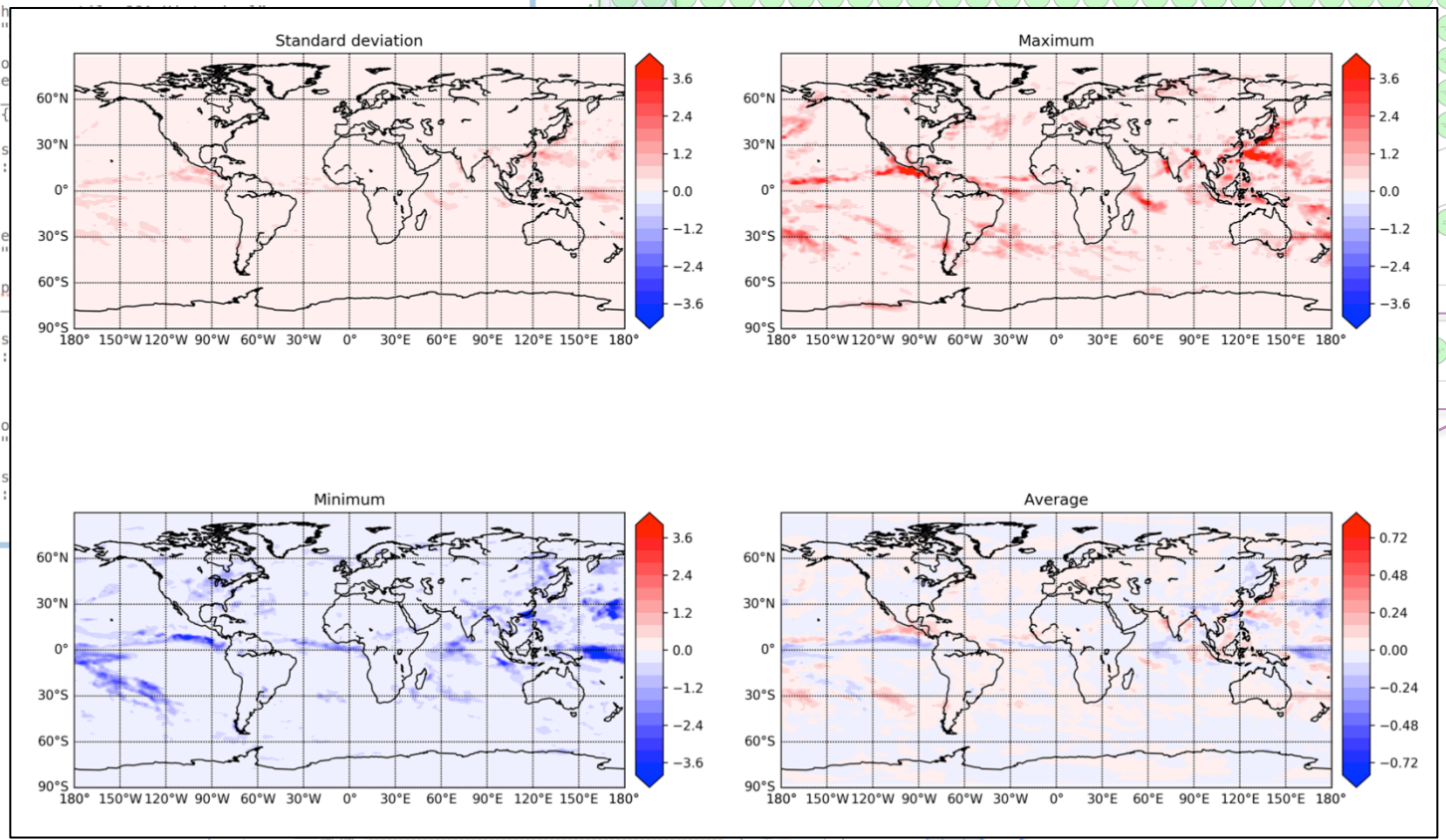
```
    {
        "name": "90th percentile JJA Historical",
        "operator": "oph_reduce2",
        "arguments": [
            "operation=quantile",
            "dim=time",
            "concept_level=y",
            "order=${5}"
        ],
        "dependencies": [
            { "task": "Subset JJA Historical", "type": "single" }
        ]
    },
    {
        "name": "Linear regression Historical",
        "operator": "oph_apply",
        "arguments": [
            "query=oph_gsl_fit_linear_coeff(measure)",
            "measure_type=auto"
        ],
        "dependencies": [
            { "task": "90th percentile JJA Historical", "type": "single" }
        ]
    },
    {
        "name": "Import Type Selection Scenario",
        "operator": "oph_if",
        "arguments": [ "condition=${10}" ],
        "dependencies": [
            { "task": "loop_model" }
        ]
    },
```



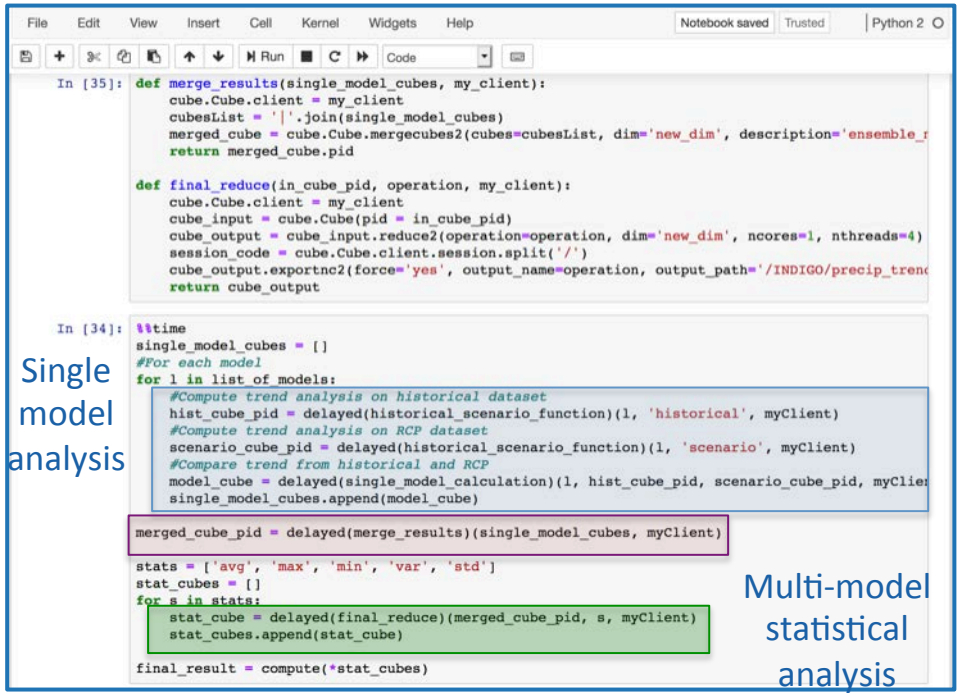
# Multi-model experiment implementation & execution

## JSON implementation of the workflow

```
{
    "name": "90th
    "operator": "
    "arguments":
        "operatio
        "dim=time
        "concept_
        "order=${
    ],
    "dependencies
        { "task":
    ]
},
{
    "name": "Line
    "operator": "
    "arguments":
        "query=op
        "measure_
    ],
    "dependencies
        { "task":
    ]
},
{
    "name": "Impo
    "operator": "
    "arguments":
    "dependencies
        { "task":
    ]
},
```



# Two approaches for the implementation




```
{
    "name": "90th percentile JJA Historical",
    "operator": "oph_reduce2",
    "arguments": [
        "operation=quantile",
        "dim=time",
        "concept_level=y",
        "order=${5}"
    ],
    "dependencies": [
        { "task": "Subset JJA Historical", "type": "single" }
    ]
},
{
    "name": "Linear regression Historical",
    "operator": "oph_apply",
    "arguments": [
        "query=oph_gsl_fit_linear_coeff(measure)",
        "measure_type=auto"
    ],
    "dependencies": [
        { "task": "90th percentile JJA Historical", "type": "single" }
    ]
},
{
    "name": "Import Type Selection Scenario",
    "operator": "oph_if",
    "arguments": [ "condition=${10}" ],
    "dependencies": [
        { "task": "loop_model" }
    ]
},
```

| | Approach | Mode | Library | Code | ExecTime |
|---|---|---|---|---|---|
| **Workflow** | SS - SI* | Batch | Ophida WF | JSON | **~170s (1.35x)** |
| **Notebook** | SS - MI* | Interactive | PyOphidia | Python | ~230s |

** SS: Server Side; SI: Single Interaction, MI: Multiple Interactions*

# Summary

- ✔ *Climate data analysis can be very complex and requires workflow support*
- ✔ *The **Ophidia HPDA framework** provides **workflow management features:***
  - *Target large-scale analysis*
  - *Parallel execution of tasks*
  - *Support for different constructs*
  - *Integrated job orchestration, management and monitoring features*
- ✔ *Real case studies can be modeled as (complex) workflows composed of hundreds of tasks*
  - ***Multi-model climate analysis example***

# References and further readings


- *Luca Cinquini, et al. (2014). The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. Future Gener. Comput. Syst. 36: 400-417.*
- *S. Fiore, A. D'Anca, C. Palazzo, I. T. Foster, D. N. Williams, G. Aloisio (2013). Ophidia: Toward Big Data Analytics for eScience. ICCS 2013, volume 18 of Procedia Computer Science, pp. 2376-2385.*
- *E. Deelman, et al. (2018) 'The future of scientific workflows', The International Journal of High Performance Computing Applications, 32(1), pp. 159–175.*
- *S. Fiore, A. D'Anca, D. Elia, C. Palazzo, I. Foster, D. Williams, G. Aloisio (2014). "Ophidia: A Full Software Stack for Scientific Data Analytics", proc. of the 2014 Int. Conference on High Performance Computing & Simulation (HPCS 2014), pp. 343-350.*
- *S. Fiore, D. Elia, C. Palazzo, F. Antonio, A. D'Anca, I. Foster and G. Aloisio (2019), "Towards High Performance Data Analytics for Climate Change", ISC High Performance 2019. Lecture Notes in Computer Science, vol. 11887, pp. 240-257.*
- *D. Elia, S. Fiore, A. D'Anca, C. Palazzo, I. Foster, D. N. Williams, G. Aloisio (2016). "An in-memory based framework for scientific data analytics". In Proc. of the ACM Int. Conference on Computing Frontiers (CF '16), pp. 424-429.*
- *C. Palazzo, A. Mariello, S. Fiore, A. D'Anca, D. Elia, D. N. Williams, G. Aloisio (2015), "A Workflow-Enabled Big Data Analytics Software Stack for eScience", HPCS 2015, pp. 545-552*
- *A. D'Anca, et al. (2017), "On the Use of In-memory Analytics Workflows to Compute eScience Indicators from Large Climate Datasets," 2017 17th IEEE/ACM Int. Symposium on Cluster, Cloud and Grid Computing (CCGRID), pp. 1035-1043.*
- *S. Fiore, et al. (2016). "Distributed and cloud-based multi-model analytics experiments on large volumes of climate change data in the earth system grid federation eco-system". In Big Data (Big Data), 2016 IEEE Int. Conference on. IEEE. pp. 2911-2918.*

# Thank you!

*This training has been organised in the context of the ESiWACE2 project:*

*ESiWACE2 has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823988.*

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

Ophidia website: *http://ophidia.cmcc.it*

Contact: *ophidia-info AT cmcc.it*

cmcc
Centro Euro-Mediterraneo sui Cambiamenti Climatici