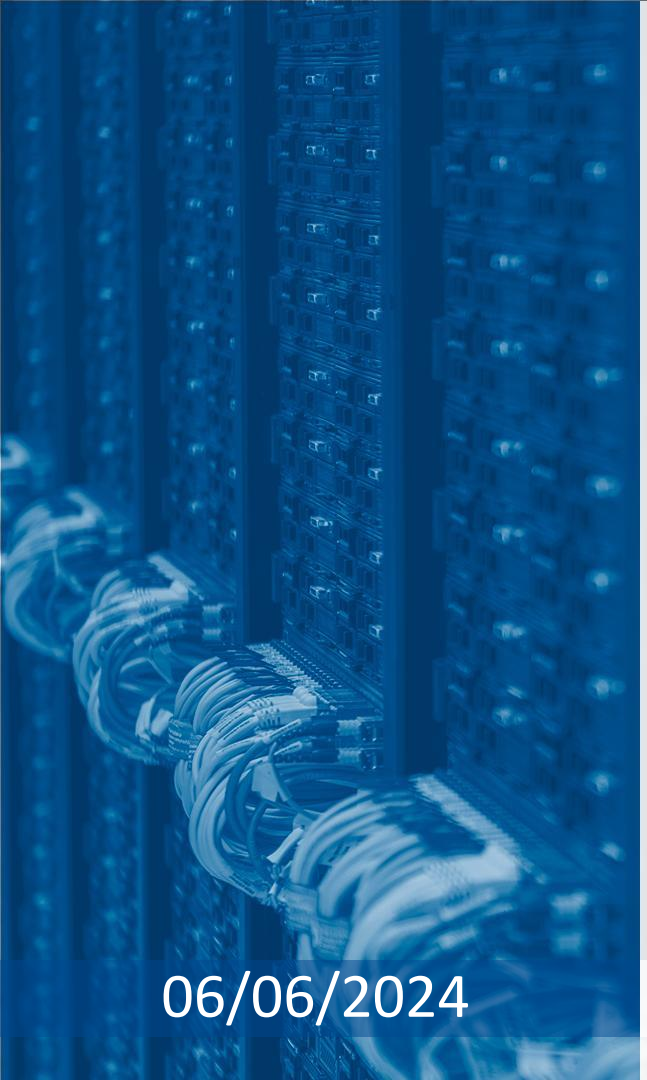


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# R user meeting

06/06/2024

Victòria Agudetse, Ariadna Batalla

# Agenda

1. Ice-breaker: R environment poll
2. News
    - General R
    - startR
    - s2dv
    - CSTools
    - esviz
    - SUNSET
3. Presentation: METACLIP provenance for SUNSET workflows
4. Q&A

# Ice-breaker:
# Poll: What tools do you use to write R code?


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# General R


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# SUNSET conda environment on MN5 (experimental)

A conda environment for SUNSET, with R/4.2.2 and CDO/2.1.0 has been installed on MN5.
It contains the latest-released versions of our in-house R packages, as well as several external R packages. The complete list is available here: SUNSET environment YAML file

If you need to run R scripts on MN5, you can try this environment. No need to install, just run:

```
source /gpfs/projects/bsc32/software/suselinux/11/software/Miniconda3/4.7.10/etc/profile.d/conda.sh
conda activate /gpfs/projects/bsc32/repository/apps/conda_envs/SUNSET-env_2.0.0
```

NOTE: We will not be adding additional external packages to the environment at this time. We are waiting for updates on the official MN5 software stack.

# startR


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# Start(): Retrieve correct time steps when time is _across

There are cases in which some inner dimension of the data is **spread across different files.**
For example, this can be the case for the 'time' dimension in some decadal models in /esarchive:

| **file** | **forecast time** |
|---|---|
| tas_Amon_HadGEM3-GC31-MM_dcppA-hindcast_s1991-r3i1p1f2_gn_199111-199112.nc: | 1, 2 |
| tas_Amon_HadGEM3-GC31-MM_dcppA-hindcast_s1991-r3i1p1f2_gn_199201-199212.nc: | 3, 4, …, 14 |
| tas_Amon_HadGEM3-GC31-MM_dcppA-hindcast_s1991-r3i1p1f2_gn_199301-199312.nc: | 15, …, 26 |
| ... | |

In this case, Start() has the special '`*_across`' specification for inner dimensions, so it can retrieve the correct indices from each file.

Issue: https://earth.bsc.es/gitlab/es/startR/-/issues/198

status: in branch develop-correct_timesteps_across_dim

# Start(): Retrieve correct time steps when time is _across

A bug was found in the case where **the indices do not start at the first file** and the file has to be skipped when loading the data.

★ For the previous case, if we load **time indices 1 to n**:

```
# The first time step is November 1991 as expected
attr(exp, "Variables")$common$time[1]
[1] "1991-11-15 UTC"
```

★ If we request **indices 3 to n**, Start() returns the wrong data:

```
# The first time step is actually forecast time 5! and not forecast time 3.
attr(exp, "Variables")$common$time[1]
[1] "1992-03-16 UTC"
```

Find a reproducible example in the issue:

Issue: https://earth.bsc.es/gitlab/es/startR/-/issues/198

status: fixed in branch develop-correct_timesteps_across_dim

# Start(): Retrieve correct time steps when time is _across

This bug has been fixed in the branch develop-correct_timesteps_across_dim and will be merged to the master branch soon.

**IMPORTANT REMINDER!**

Always **check the data and metadata** retrieved by Start() to make sure they are what you expect; especially if you are working with a new dataset or new code.

If you find anything wrong, please open an issue to report it. Even if it's a "mistake" and not a real bug, the information can still be helpful for others!

Issue: https://earth.bsc.es/gitlab/es/startR/-/issues/198

status: in branch develop-correct_timesteps_across_dim

# Compute(): Allow chunking over “across” dimension

In some cases, when an inner dimension is defined as going `_across` a file dimension, Compute() does not allow chunking along the inner dimension. For example, in the previous issue, we need to define a file dimension (e.g. “period”) for the names of the different files:

```
path_list <-  paste0("$ensemble$/Amon/$var$/gn/v20200417/",
                     "$var$_Amon_*_dcppA-hindcast_s$syear$-$ensemble$_gn_$period$.nc")
```

Trying to chunk along the time dimension results in the following error (see the complete example in the GitLab issue):

```
Error in Start(dat =
"$ensemble$/Amon/$var$/gn/v20200417/$var$_Amon_*_dcppA-hindcast_s$syear$-$ensemble$_gn_$period$.nc",   :
  Chunk over dimension 'time' is not allowed because 'time' is across 'period'.
```

Issue: https://earth.bsc.es/gitlab/es/startR/-/issues/196

status: work in progress

# Compute(): Allow chunking over “across” dimension

We are working to enhance Compute() by enabling this feature.

Question:

Does anyone have more scripts that use the `*_across` parameter in Start()?

If so, please add a comment in the issue! They would be very useful for testing if our solutions work for all cases.

They are useful even if you are not using Compute()!

Issue: https://earth.bsc.es/gitlab/es/startR/-/issues/196

status: work in progress

# s2dv


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# Significance testing in Bias() and SprErr()

The SprErr() function has been modified to return the statistical significance of the values. The development includes the new parameters `sign`, `alpha` and `na.rm`.

Issue: https://earth.bsc.es/gitlab/es/s2dv/-/issues/115

status: in branch dev-spread_error_ratio

The Bias() function has been modified to return the statistical significance of the values, using Welch's test. It is only available when computing the absolute bias. The parameter `alpha` can be specified to select the significance level.

Issue: https://earth.bsc.es/gitlab/es/s2dv/-/issues/118

status: in branch dev-sigBias

# N.eff for RandomWalkTest()

This development adds the option to use the effective number of degrees of freedom (N.eff) in RandomWalkTest() to account for time series autocorrelation when assessing the statistical significance of skill scores.

It affects the following s2dv functions: MSSS(), RMSSS(), RPSS() and CRPSS().

The N.eff parameter can be:

- `NA` (and it will be computed with s2dv:::.Eno()). This is the default value;
- `FALSE` (autocorrelation is not considered);
- a single numeric value to be applied to all the skill array (`skill_A`);
- A numeric array with the same dimensions as `skill_A` except for `time_dim`.

MR: https://earth.bsc.es/gitlab/es/s2dv/-/merge_requests/182

status: in branch develop-RandomWalk-N.eff

# Inconsistent missing values in Histo2Hindcast()

Histo2Hindcast() returns unexpected results when the initial date requested in the function call is not present in the data.

For example, if:

```
sdatesin <- '199001' # My data starts in January 1990
sdatesout <- paste0(as.character(c(seq(1989, 2014))), '0501') # I want start dates from May 1989 to May 2014
```

The function does not expect this case, so the resulting array contains incorrect data for some of the time steps of the “empty” start date.

Question: Should the function return an array with NA padding, or should it raise an error?

Issue: https://earth.bsc.esitlab/es/s2dv/-/issues/117

status: potential bugfix in branch dev-histo2hindcast-bugfix

# CSTools


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# CST_Start(): updated documentation

Documentation update:

```
#'CST_Start() uses as.s2dv_cube() to transform the output into an s2dv_cube
#'object. The as.s2dv_cube() function is designed to be used with data that
#'has been retrieved into memory. To avoid errors, please ensure that
#'CST_Start(..., retrieve = TRUE) is specified.
```

```
# do not run
...
#'  data <- CST_Start(dat = path,
#'                    var = 'prlr',
#'                    ensemble = indices(1:6),
#'                    sdate = sdates,
                     ...
#'                    retrieve =FALSE)
#'                    retrieve =TRUE)
```

Issue: https://earth.bsc.es/gitlab/external/cstools/-/issues/151

status: in master

# as.s2dv_cube(): added specific error

startR::Start(..., retrieve = FALSE) and CSTools::CST_Start(..., retrieve = FALSE) creates an object of class startR_cube, which as.s2dv_cube() does not support.
as.s2dv_cube() now incorporates a specific error to alert of this, informing users to set "retrieve = TRUE":

```
...
  } else if (inherits(object, 'startR_cube')) {
    stop("Unsupported object class: 'startR_cube'. ",
         "When using startR::Start() or CSTools::CST_Start(), set ",
         "'retrieve = TRUE' to ensure the data is retrieved into "
         "memory and can be converted into a 's2dv_cube' object.")
...
```

Issue: https://earth.bsc.es/gitlab/external/cstools/-/issues/151

status: in master

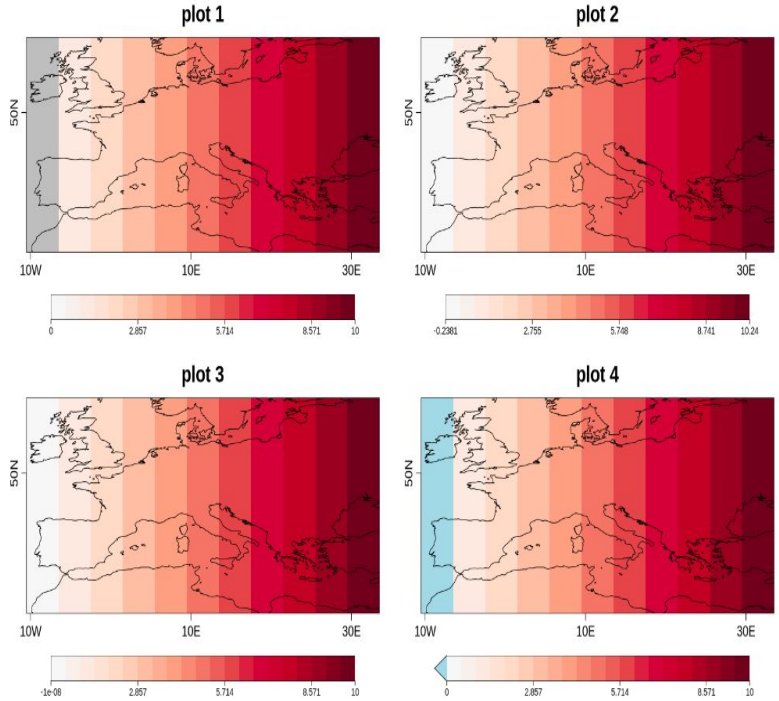
# esviz

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# Color bar boundaries in ColorBarContinuous()

Currently the boundaries of color bar are "(, ]": the lower bound is **not included** and the upper bound **is included**.

When there are values that are exactly the same as the lower bound and the lower triangle is disabled, the color bar looks like **plot 1**:



# Color bar boundaries in ColorBarContinuous()

★ **Action:** The parameter `include_boundaries`, a vector of two logical elements, is being developed. The documentation will be improved.

★ **Questions:** is “include_boundaries” a good name? Would you prefer two parameters, one for each boundary? What should the default values be?

Issue: https://earth.bsc.es/gitlab/es/esviz/-/issues/15

status: in branch dev-ColorBarContinuous_boundaries

# SUNSET


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# New Time Aggregation module

A new module for temporal aggregation has been included in the master branch.

An example of a script using time aggregation is available on GitLab.

```
Workflow:
  Time_aggregation:
    execute: yes # Either yes/true or no/false. Defaults to false. (Mandatory, bool)
  method: average # Aggregation method. Available methods: 'average, 'accumulated'. (Mandatory, string)
    # ini and end: list, pairs initial and final time steps to aggregate.
    # In this example, aggregate from 1 to 2; from 2 to 3 and from 1 to 3
    ini: [1, 2, 1]
    end: [2, 3, 3]
    # user_def: List of lists, Custom user-defined forecast times to aggregate.
    user_def:
    DJF: !expr sort(c(seq(1, 120, 12), seq(2, 120, 13), seq(3, 120, 14))# aggregate 1,2,3,13,14,15,...
```

MR: https://earth.bsc.es/gitlab/es/sunset/-/merge_requests/137

status: in master

# Recipe template available on GitLab

A new recipe template is available on GitLab. This template is a complete recipe with all the possible parameters.

If you add new parameters in a development, please include them in this file:
https://earth.bsc.es/gitlab/es/sunset/-/blob/master/recipe_template.yml

If you find problems or missing parameters, please open an issue!

MR: https://earth.bsc.es/gitlab/es/sunset/-/merge_requests/137

status: in master

# Visualization: Mask and/or dots for metric visualization

Currently, Visualization() has the parameter `significance = TRUE/FALSE`, where `TRUE` marks grid points that are not statistically significant with dots.

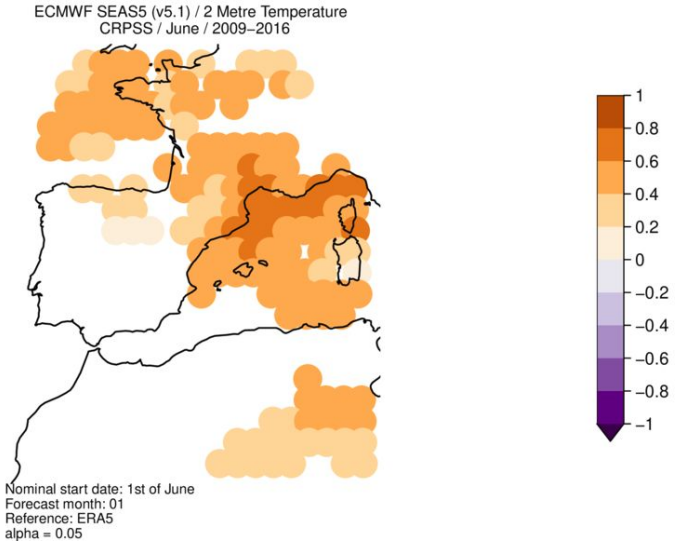
A new development will include two options:

- `'mask'`: Masking non-significant grid points; only available when using PlotRobinson()
- `'dots'`: Dots over non-significant grid points

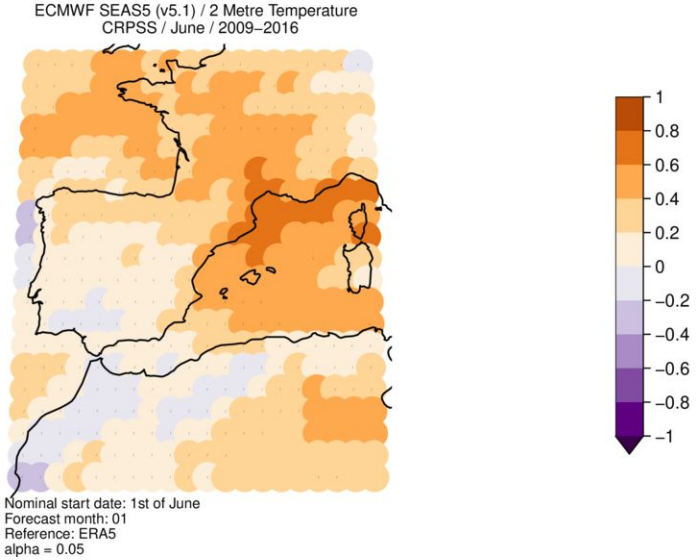
MR: https://earth.bsc.es/gitlab/es/sunset/-/merge_requests/140

status: in branch dev-vis-mask

# Visualization: Mask and/or dots for metric visualization




Option ‘mask’: only display statistically significant values



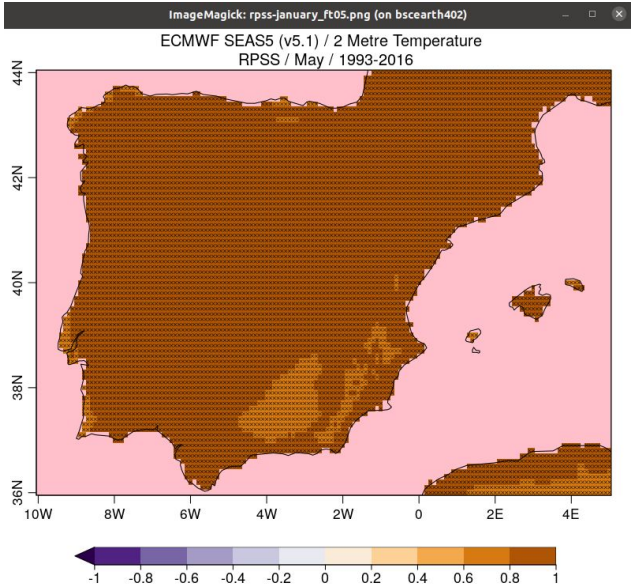

Option ‘dots’: dots over non-statistically significant values

MR: https://earth.bsc.es/gitlab/es/sunset/-/merge_requests/140

status: in branch dev-vis-mask; to be tested

# Visualization: Choose the color of NA values in plots

New parameter NA_color to choose the color of NA values in plots. The current default in PlotEquiMap()/VizEquiMap() is pink:




```
Workflow:
  Visualization:
    plots: skill_metrics forecast_ensemble_mean most_likely_terciles
    NA_color: 'pink'
```

Issue:

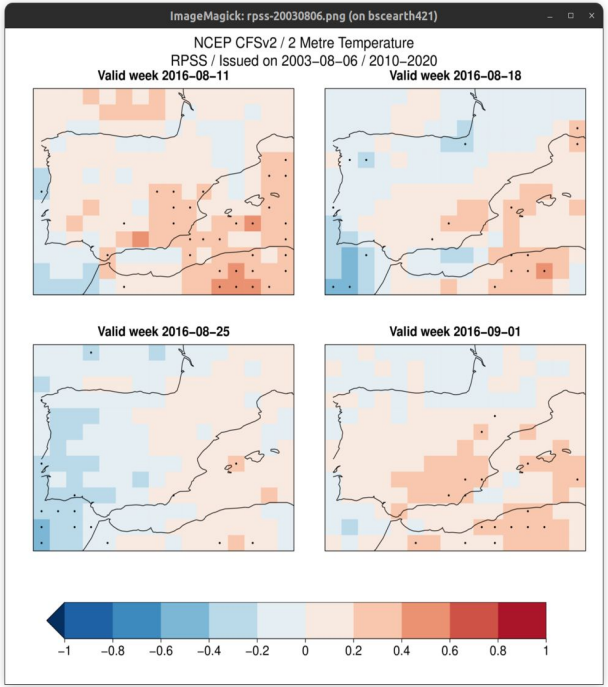
status: in branch dev-nan_color; to be tested

# Development of subseasonal in SUNSET

The processing of subseasonal data is being developed in SUNSET.

The first module to support this data will be the Visualization module.

Feel free to check out the development branch and add any opinions or suggestions in the MR.




MR: https://earth.bsc.es/gitlab/es/sunset/-/merge_requests/134

status: in branch dev-subs_vis

# User presentation:

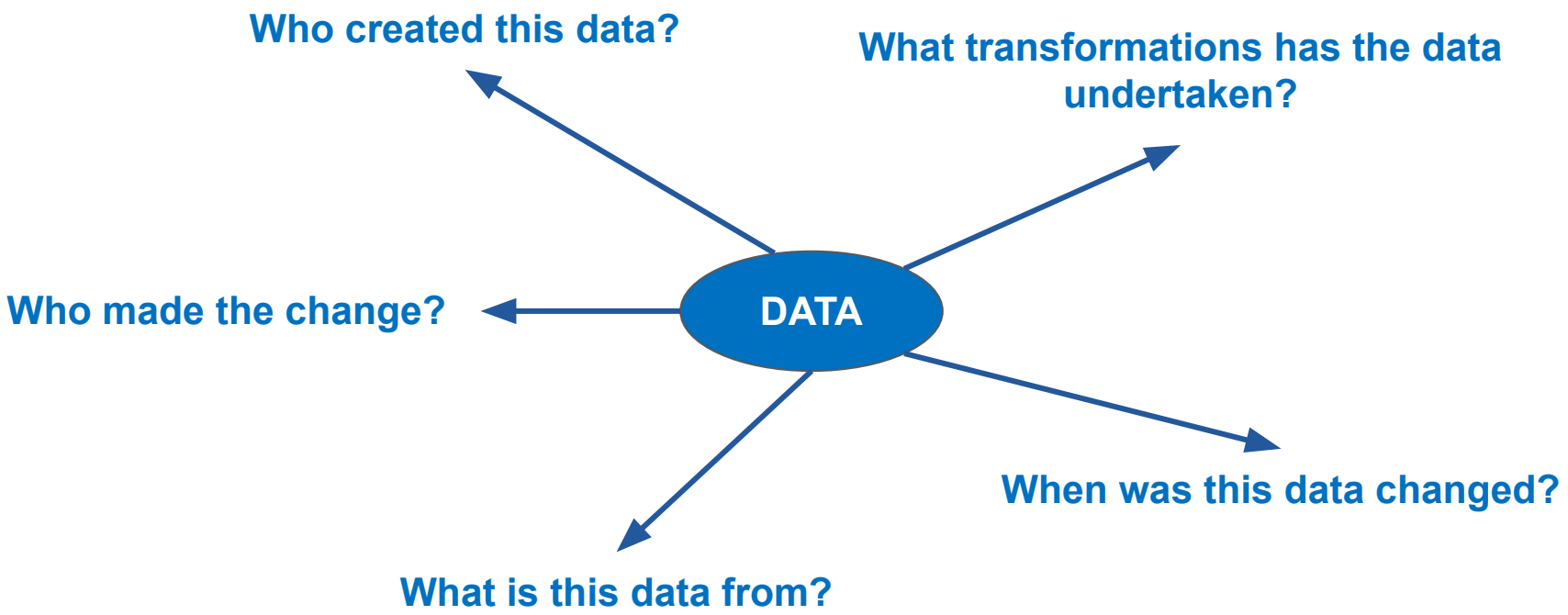
# METACLIP Provenance for SUNSET (Albert Puiggròs)


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

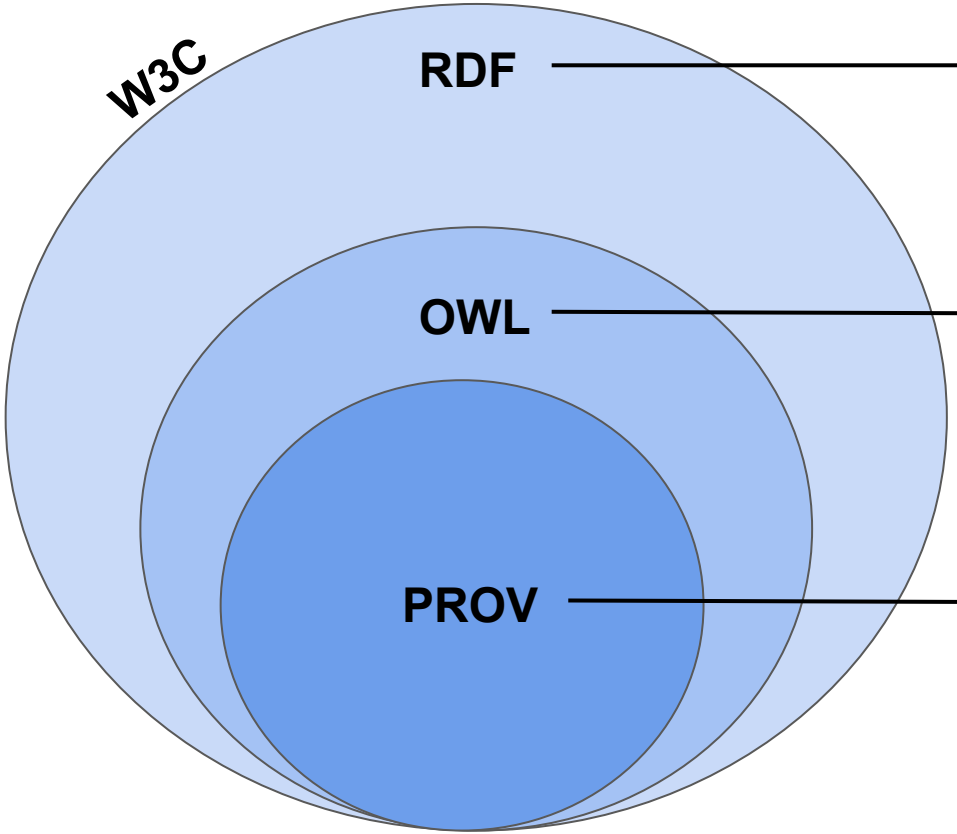
# Data Provenance

*“Record which specifies the people, institutions, entities, and activities involved in the creation and influence exercised on data”*

# Data Provenance



# Data Provenance: Languages and ontologies




**Resource Data Framework**

Describe, codify, recover and interchange information about web resource. Generally used in the Web. Foundation of structuring data. (IRI)

**Onology Web Language**

Extension of RDF, includes more complex relations to create ontologies.

**W3C Provenance**

Based on RDF and OWL, describes provenance of data.

# Data Provenance: OWL basic elements

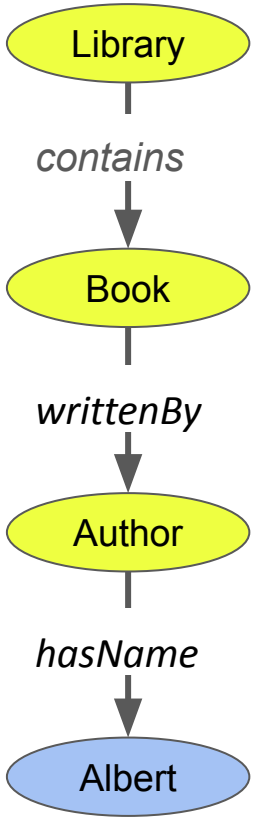
-**Classes**: represents a group of elements that share common properties
(*Library*, *Book*, *Author*…)

-**Individuals**: specific examples of classes
(*BSC Library*, *TheGreatGatsby*, *FScottFitzgerald*…)

**-Object properties**: define relationships between classes
(*writtenBy*, *contains*…)

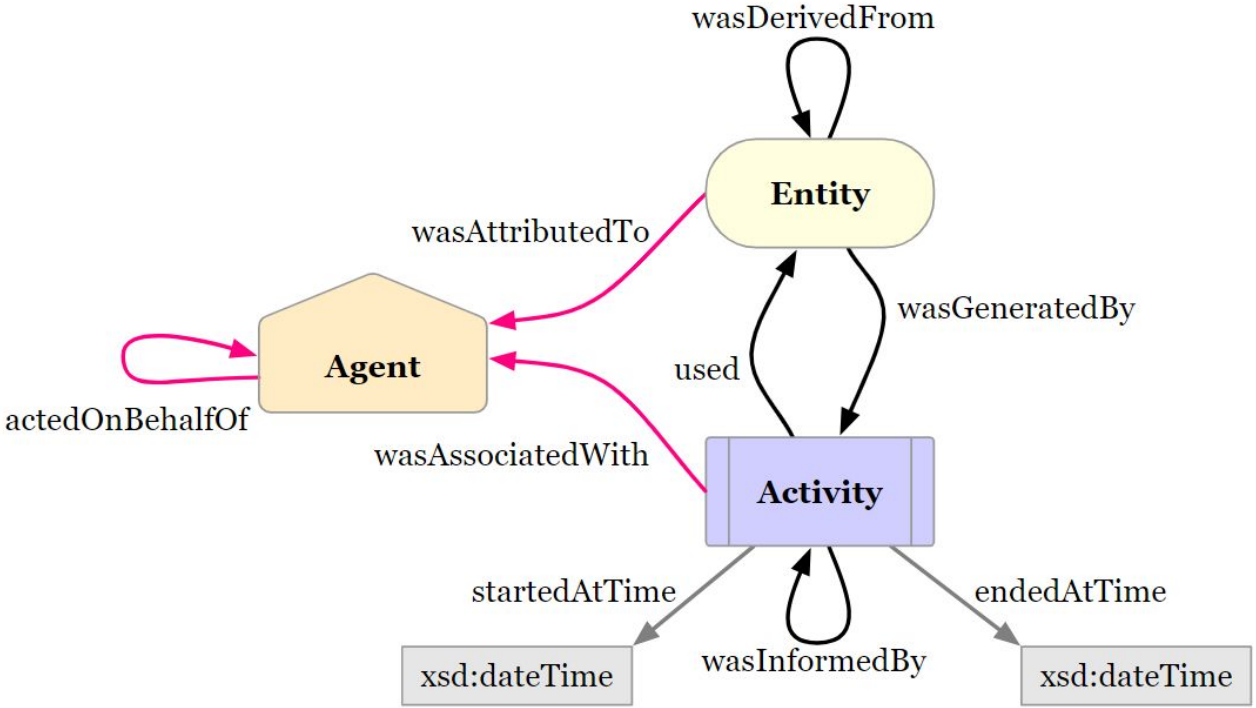
**-Data Properties**: attributes linking classes or individuals to data values
(*hasTitle, hasName*, *has PublicationYear*…)

**-Annotation properties**: non-logical annotations
(*rdf:comment, rdf:referenceURL*…)

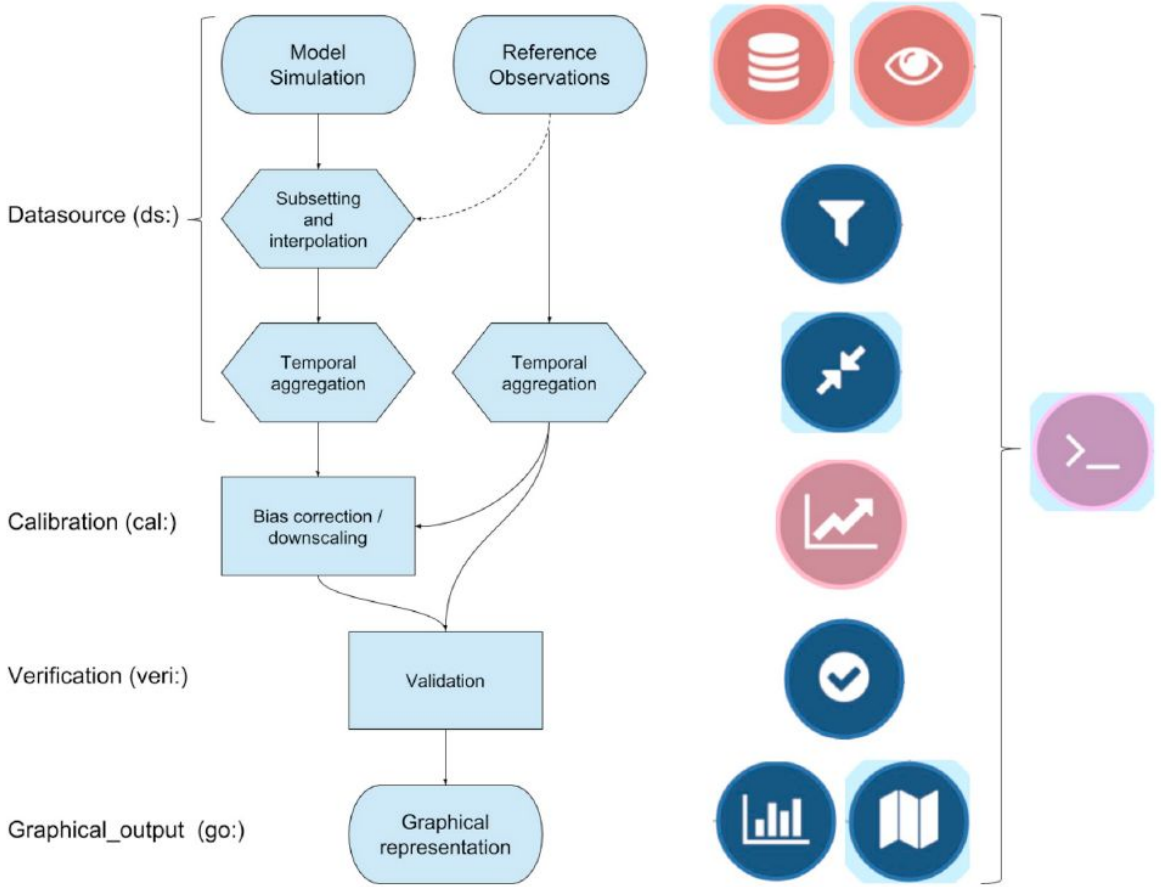


# Data Provenance: Languages and ontologies

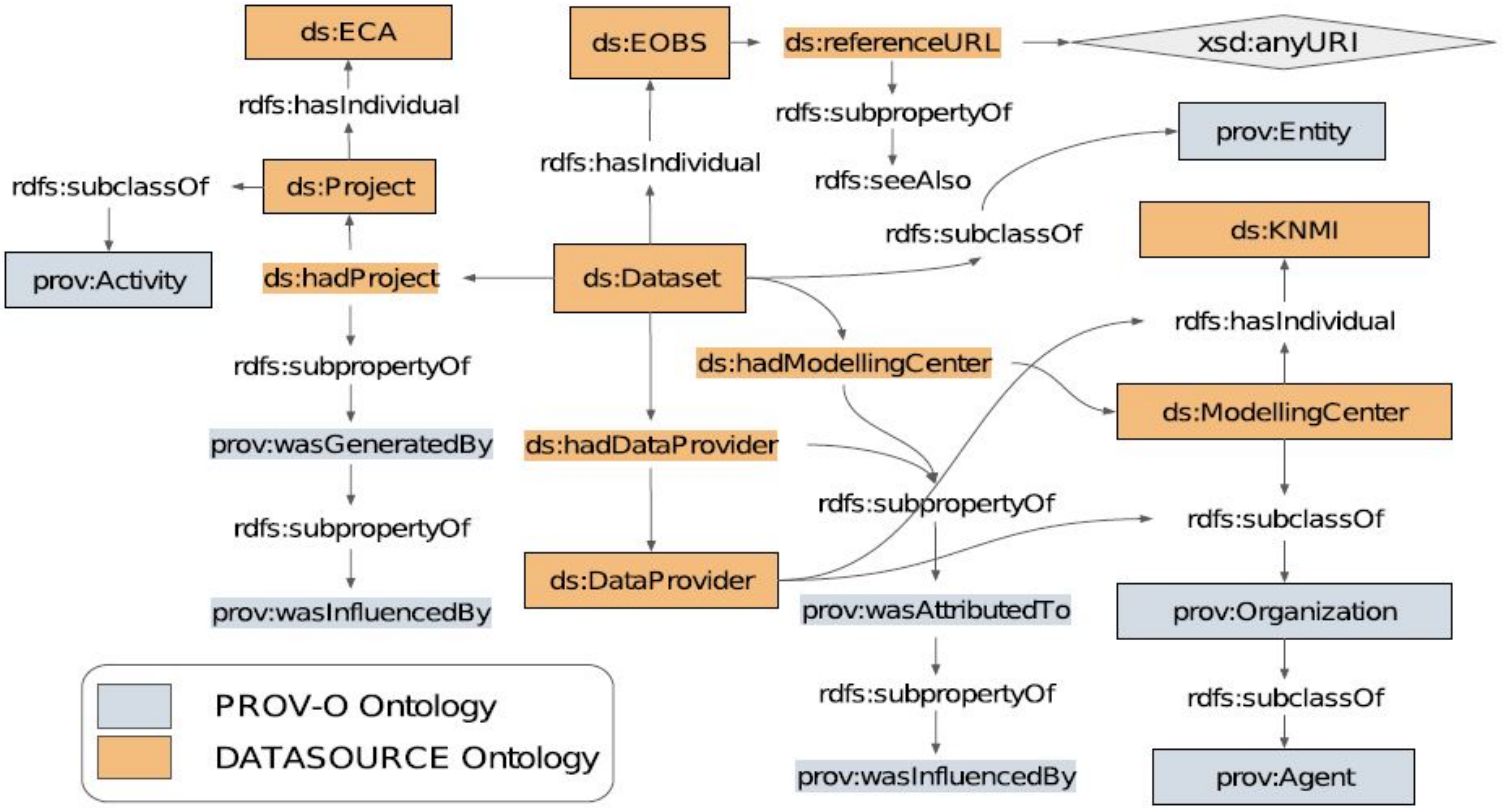
**-PROV Data Model**



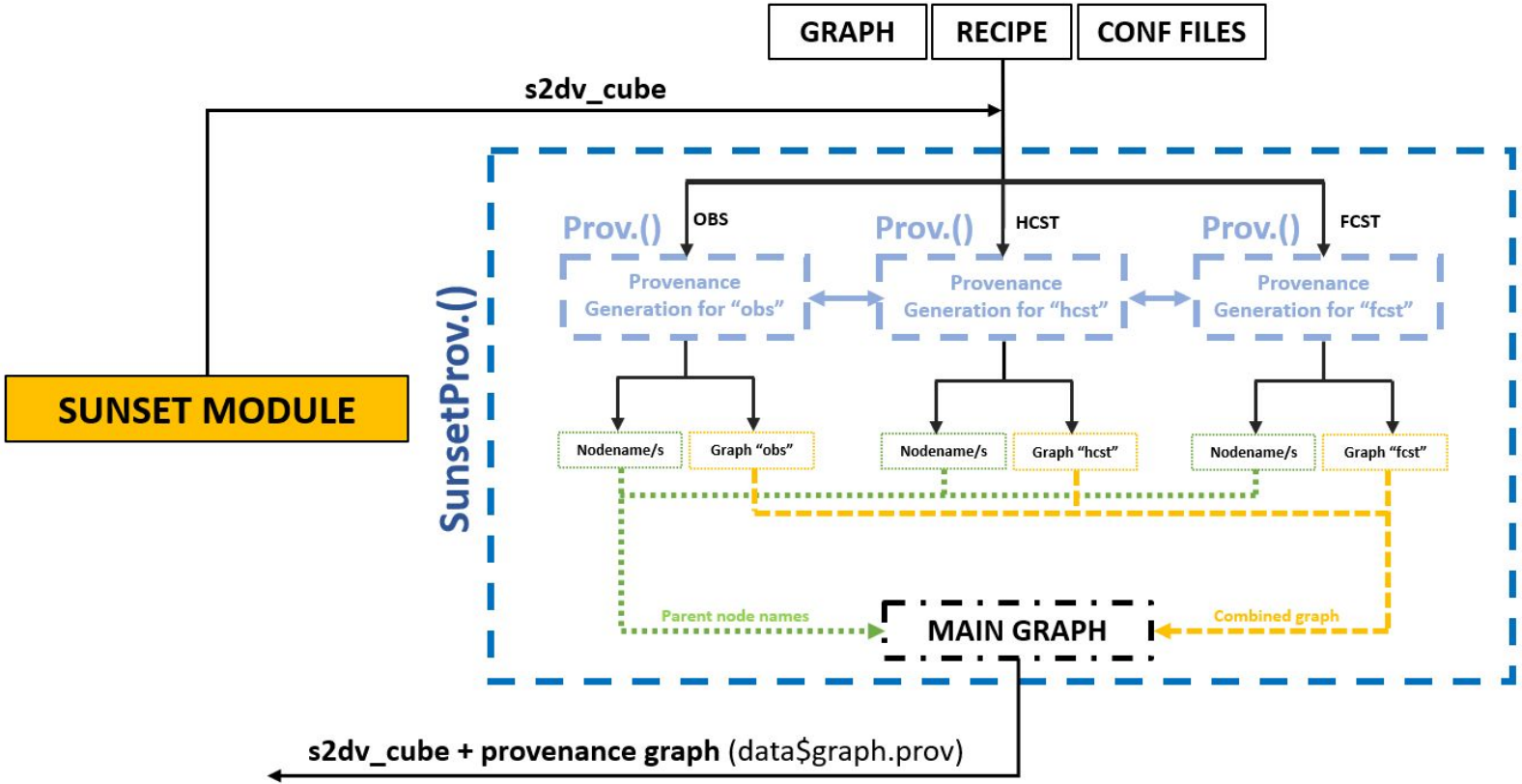
# METACLIP: METAdata for CLImate Products



# METACLIP: METAdata for CLImate Products

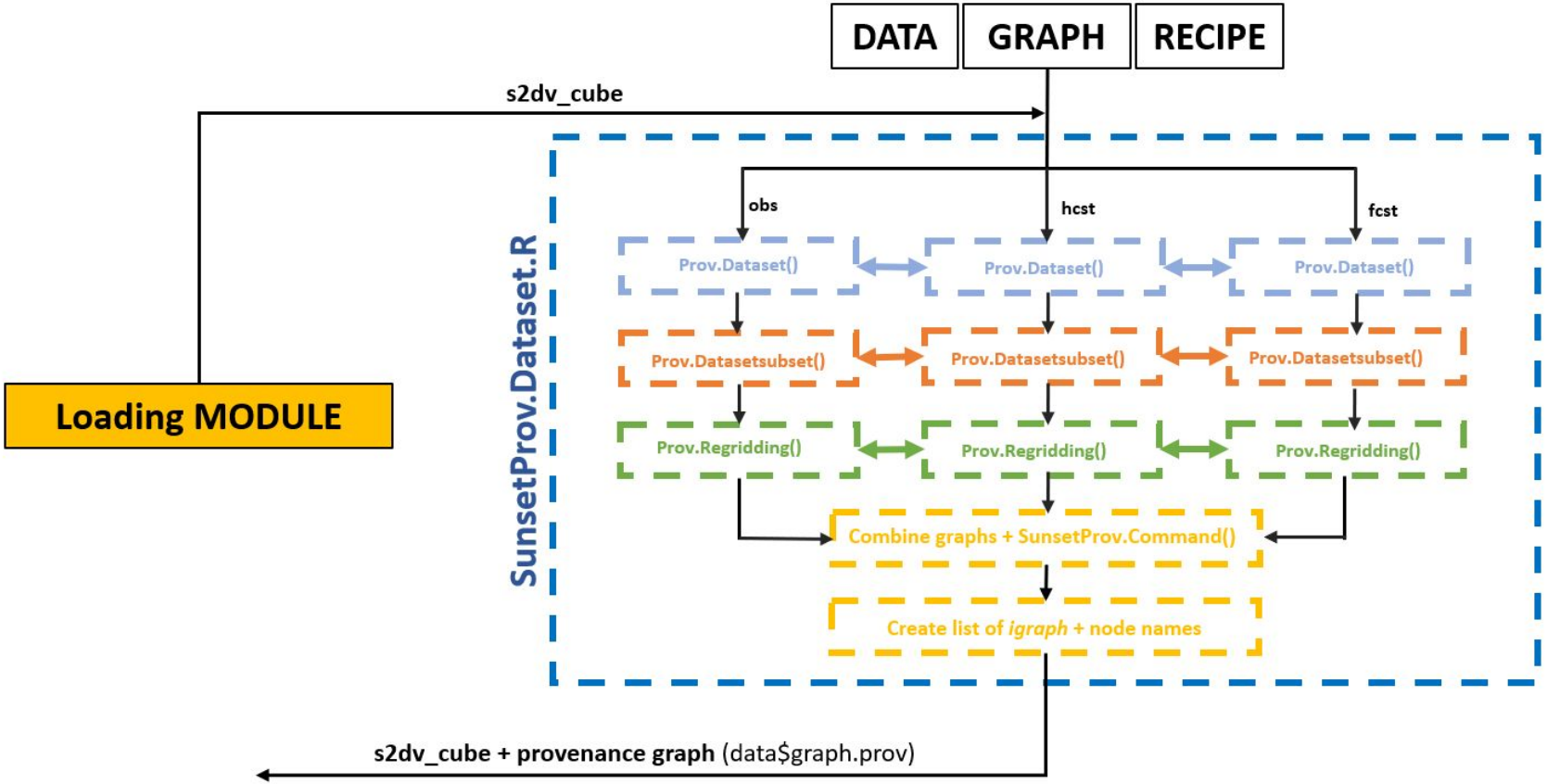


# SUNSET Provenance



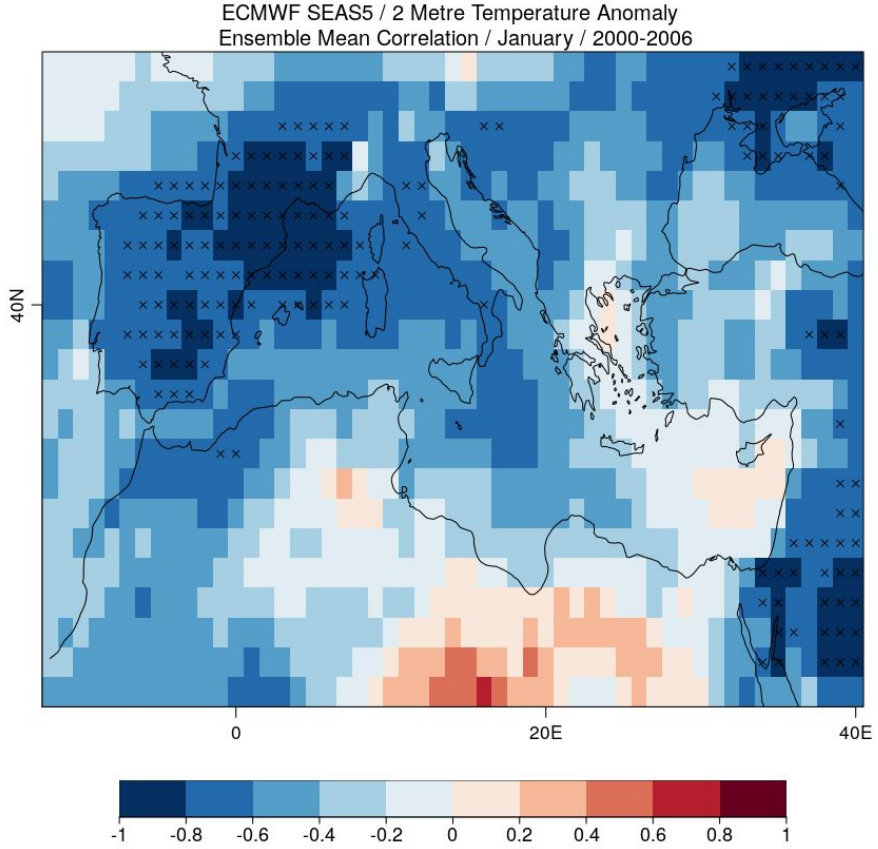

https://earth.bsc.es/gitlab/es/sunset/-/tree/dev-test-provenance?ref_type=heads

# SUNSET provenance: Loading Module

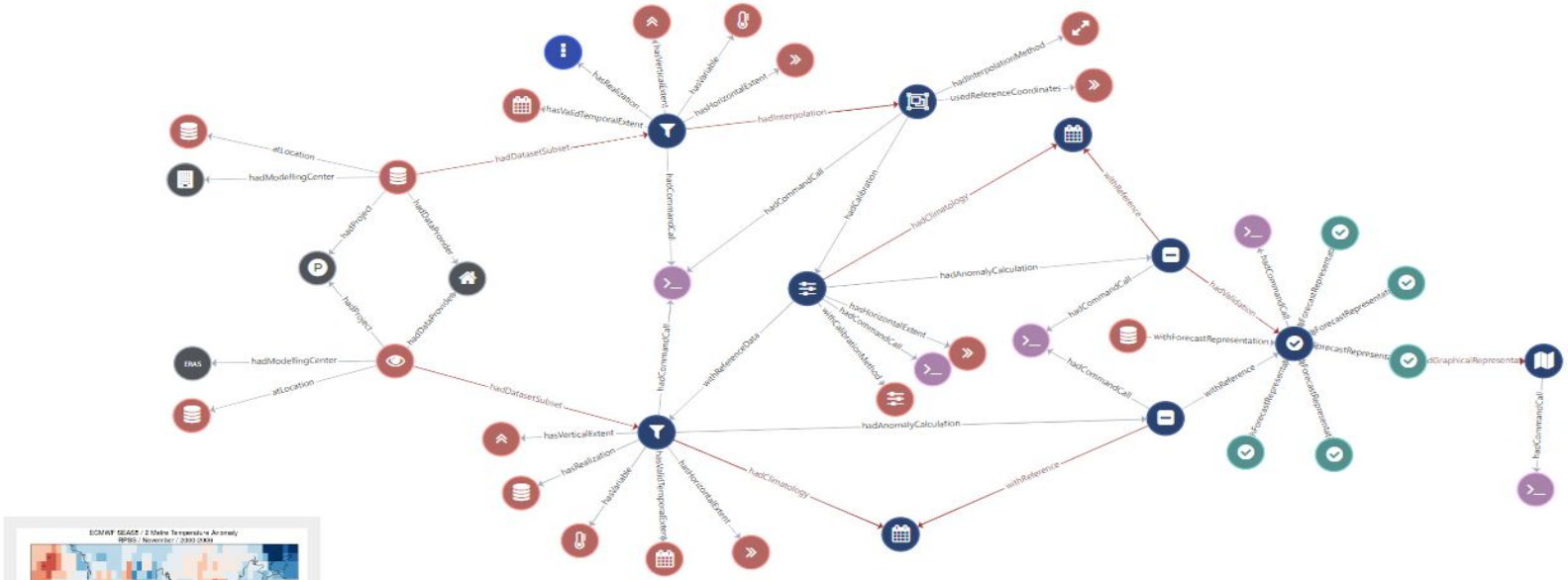


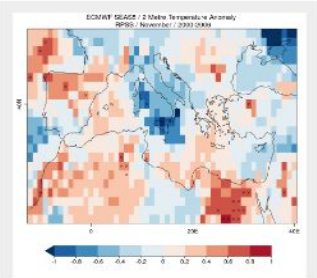
# SUNSET PROVENANCE: Example

```
#Recipe
recipe_file <- "SUNSET_PROV/test_provenance/recipes/recipe_test_provenance1.yml"

#Loading recipe
recipe <- prepare_outputs(recipe_file)

#Loading module
data <- Loading(recipe)

#Units module
data <- Units(recipe, data)

#Calibration
data_cal <- Calibration(recipe, data)

#Anomalies
data_ano <- Anomalies(recipe, data_cal)

#Compute skill metrics
skill_metrics <- Skill(recipe, data_ano)

#Compute percentiles and probability bins
probabilities <- Probabilities(recipe, data_ano)

#Define provenance graph for Visualization
graph.prov <- skill_metrics$graph.prov

#Visualization
Visualization(recipe, data_ano, graph.prov,
              skill_metrics = skill_metrics,
              probabilities = probabilities,
              significance = TRUE)
```



# SUNSET PROVENANCE: Example



# SUNSET PROVENANCE: Ontology expansion

*prov:Entity -> prov:Influence -> prov:EntityInfluence -> ds:Step -> ds:Transformation -> cal: Calibration ->* **cal:Downscaling**

```
<!-- http://www.metaclip.org/calibration/calibration.owl#Downscaling -->

<owl:Class rdf:about="http://www.metaclip.org/calibration/calibration.owl#Downscaling ">
        <rdfs:subClassOf
        rdf:resource="http://www.metaclip.org/calibration/calibration.owl#Calibration"/ >
        <dc:description>A downscaling operation refers to the computational procedure
        applied to transform coarser-resolution climate data into finer-resolution
        representations, typically achieved through statistical or dynamical
        methodologies. This process involves the utilization of mathematical algorithms or
        numerical simulations to extrapolate localized climate phenomena and detailed
        spatial patterns from broader-scale atmospheric or oceanic information, thus
        facilitating more precise analyses and projections at regional or local
        scales.</dc:description>
       <dc:title>Downscaling</ dc:title>
       <rdfs:seeAlso>https://en.wikipedia.org/wiki/Downscaling </rdfs:seeAlso>
</owl:Class>
```

https://github.com/metaclip/vocabularies/pull/6

# SUNSET PROVENANCE: Future work

-**Expand the ontology**, adding new classes and individuals that can be suitable to describe SUNSET (Unit conversion and probabilities)

-**Expand the capabilities** of the current functions (multimodel and other scenarios)

-Incorporate checks to **verify with the METACLIP ontologies.**

# Thanks for joining

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Next meeting: July 4th