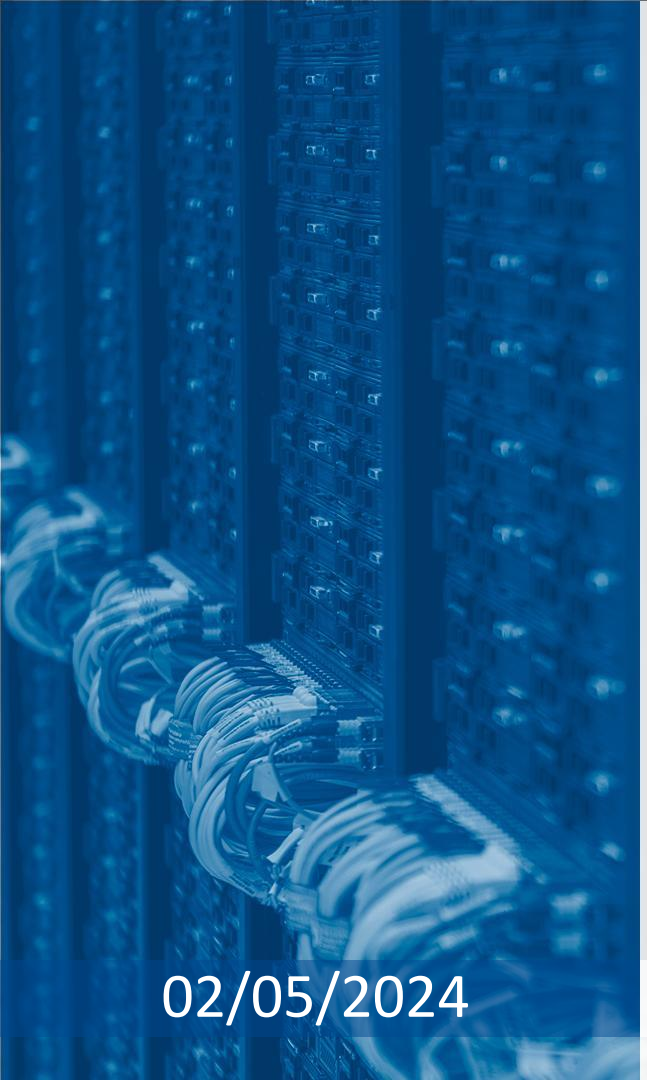


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

R user meeting

02/05/2024

Victòria Agudetse, Ariadna Batalla

Agenda

1. Ice-breaker: String manipulation
2. News
    - General R
    - s2dv
    - CSTools
    - esviz
    - SUNSET
3. Presentation: BSC-ES Infrastructure
4. Q&A

Ice-breaker


BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

String manipulation

Base R

- To change the the string or the characters in a vector or data frame to lower or upper case:
  - tolower(x): lower case

```
> tolower("Hello World!")
[1] "hello world!"
```

  - toupper(x): upper case

```
> toupper("Hello World!")
[1] "HELLO WORLD!"
```

String manipulation

Base R

- To substitute the string or the characters in a vector or data frame with a specific string:
  - sub(pattern, replacement, x): only substitutes the first occurrence.

```
> sub("world", "R user", "Hello world! The world is great")
[1] "Hello R user! The world is great"
```

  - gsub(pattern, replacement, x): applies a global substitution to all matches.

```
> gsub("world", "R user", "Hello world! The world is great")
[1] "Hello R user! The R user is great"
```

where:

➔ pattern: The pattern or string which you want substituted. Can be in regex form.

➔ replacement: A input string to substitute the pattern string.

➔ x: A vector or a data frame to substitute the strings.

String manipulation

stringr package: a set of internally consistent tools for working with character strings

```
> > x <- c("abcdef", "ghifjk")
> str_sub(x, 3, 3) <- "X"
> x
[1] "abXdef" "ghXfjk"
```

- Character manipulation:
  - str_length(), str_sub(), str_dup()
- Whitespace tools:
  - str_pad(), str_trim(), str_trunc(), str_wrap()
- Locale sensitive operations: function(x, locale= “en”):
  - str_to_upper(x) *similar to toupper()*, str_to_title(x), str_to_lower(x) *similar to tolower()*, str_order(x), str_sort(x)
- Pattern matching functions:
  - str_detect(), str_subset(), str_count(), str_locate(), str_locate_all(), str_extract(), str_extract_all(), str_match, str_match_all, str_replace() *similar to sub()*, str_replace_all() *similar to gsub()*, str_split(), str_split_fixed()
  - Four main engines to describe patterns: regex, fixed(), coll(), boundary()

More information: https://cran.r-project.org/web/packages/stringr/vignettes/stringr.html

Cheat sheet: https://evoldyn.gitlab.io/evomics-2018/ref-sheets/R_strings.pdf

General R

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Using Jupyter notebooks with R on the Hub

JupyterLab can be used to write and run code with Jupyter Notebooks in R.

- Previously it was available in the Workstations, **but the software is now outdated**: it only works with R/3.6.1 and the latest versions of the packages are not available.
- **In the Hub with R/4.2.1**: You can use the IDE feature following the instructions in the wiki, but the Notebook feature is not available. There is an open issue about it: https://earth.bsc.es/gitlab/es/requests/-/issues/2446

s2dv

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

CDORemap() irregular grid interpolation error

Depending on the dimensions of the input array, CDORemap() can raise an error when interpolating from an irregular grid to a gaussian grid.
This was due to a bug in the code that was causing incorrect matching of the dimensions.

This error was fixed and the function has been tested to ensure that the interpolation results are the same as before the bugfix.

Issue: https://earth.bsc.es/gitlab/es/s2dv/-/issues/114

status: in master

CSTools

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

CST_MergeDims(): Dates returned as numeric values

CST_MergeDims() returns the $Dates element as numeric values instead of the POSIXt/POSIXct date format. For example:

```
> test_hcst$attrs$Dates
, , 1
           [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
[1,] 1438074000 1469696400 1501232400 1532768400 1564304400 1595926800
```

This bug is fixed in the branch develop-fix_CST_MergeDims_Dates.

Issue: https://earth.bsc.es/gitlab/external/cstools/-/issues/149

status: in branch develop-fix_CST_MergeDims_Dates

CST_Start(): use startR functions explicit namespace

Previously, startR functions used in the call to CST_Start() had to have the `startR::` namespace unless the startR library had been loaded beforehand, otherwise the function would rise an error:

```
# do not run
library(CSTools)
...
res <- CST_Start(dat = list(list(name = 'system4_m1', path = repos2),
                            list(name = 'system5_m1', path = repos1)),
                 var = c('tas', 'sfcWind'),
                 sdate = c('20160101', '20170101'),
                 ensemble = startR::indices(1:2),
                 time = startR::indices(1:2),
                 lat = startR::indices(1:10),
                 lon = startR::indices(1:10),
                 ...
                 retrieve = TRUE)
```

CST_Start(): use startR functions explicit namespace

Now, the functions work even if the namespace is not explicitly added:

```
# do not run
library(CSTools)
...
res <- CST_Start(dat = list(list(name = 'system4_m1', path = repos2),
                            list(name = 'system5_m1', path = repos1)),
                 var = c('tas', 'sfcWind'),
                 sdate = c('20160101', '20170101'),
                 ensemble = indices(1:2),
                 time = indices(1:2),
                 lat = indices(1:10),
                 lon = indices(1:10),
                 ...
                 retrieve = TRUE)
```

Issue: https://earth.bsc.es/gitlab/external/cstools/-/issues/140

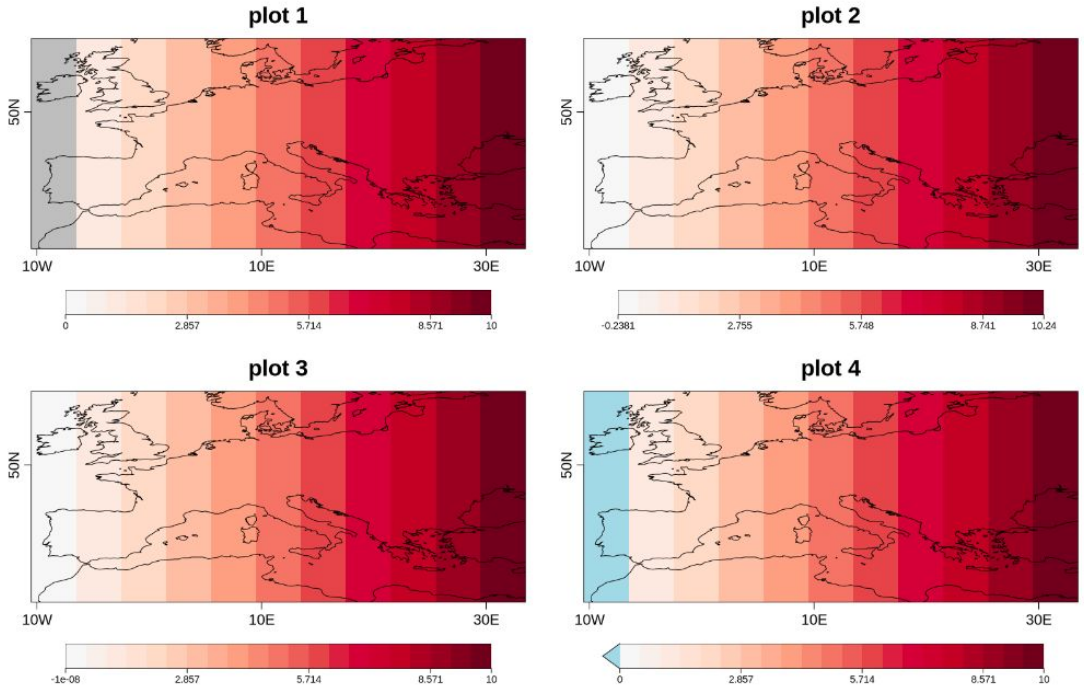
status: in master

esviz

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Color bar boundaries in ColorBarContinuous()

In ColorBarContinuous(), there is a condition showing that 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, the color bar looks like this:



Color bar boundaries in ColorBarContinuous()

This condition is not clear in the documentation of VizEquiMap() and other similar functions, which can lead to confusion.

★ **Action:** The documentation will be improved.

★ **Question:** should we consider adding a parameter to specify if the value right of the lower boundary should go to the lower triangle or be included in the lowest interval?

Feel free to add your opinion or suggestions in the issue.

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

status: pending

SUNSET

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Downscaling: New methods and downscaling forecasts

New features for the Downscaling module:

- ★ Downscaling the forecast
- ★ Downscaling through the large scale variables in analogs
- ★ 4nn method in `Intlr` option has been changed to the 9nn method with principal component pre-filtering.

The information about the new features is provided in the merge request and in the Downscaling section of the SUNSET wiki:
https://earth.bsc.es/gitlab/es/sunset/-/wikis/home#downscaling-module

The Anomalies module still needs to be adapted to the case where hcst and fcst have a different grid from the reference dataset.

MR: https://earth.bsc.es/gitlab/es/sunset/-/merge_requests/123

status: in branch dev-Downscaling

Scorecards: requirements to plot ‘enscorr’

In the Scorecards module, the value of the ensemble correlation ‘enscorr’ needs to be recalculated in order to do a mathematically correct aggregation, so the values provided by the Skill module can’t be used directly.

The Statistics module should be called, and the metrics ‘cov’, ‘std’ and ‘n_eff’ should be requested:

```
Statistics:
   metric: cov std n_eff
   save: 'all'
```

A check has been included that will alert the recipe if ‘enscorr’ is requested in the Scorecards but the statistics are missing from the recipe.

MR: https://earth.bsc.es/gitlab/es/sunset/-/merge_requests/133

status: in branch dev-add_enscorr_scorecards_check

Visualization of subseasonal data

The Visualization module is being modified to be able to produce plots for weekly subseasonal datasets. This is the first step to adapt SUNSET to work for full subseasonal workflows.

For now, subseasonal data can be loaded and processed with our regular R functions and SUNSET can be used to plot the results. An example script will be added when the development is merged.

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

status: in branch dev-subs_vis

Autosubmit: username change

After the general HPC maintenance, the Nord3v2 usernames have changed from **bsc32xxx** to **bsc032xxx**.

The Autosubmit templates have been modified to be able to work with this change. The development will be tested once /esarchive is available on Nord3v2. Until the changes are merged to the master branch, the Autosubmit configurations provided by the launcher will not work.

MR: https://earth.bsc.es/gitlab/es/sunset/-/merge_requests/135

status: in branch dev-update_bsc32xxx_to_bsc032xxx

What do we know about BSC infrastructure?

Authors: Núria Pérez-Zanón and An-Chi Ho (December 2021)
Updated by Victòria Agudetse (May 2024)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

What do we know about BSC-ES infrastructure?

Aside from the data and software in our personal laptops, we all have access to common BSC infrastructure.

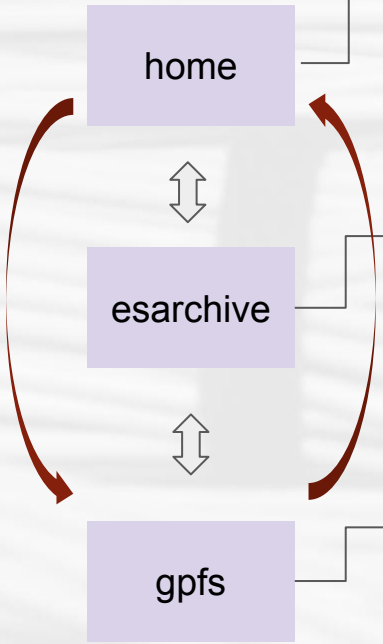
We **access the BSC infrastructure**:

- ★ When we connect to the BSC-ES Hub
- ★ When we use the workstations in the office
- ★ When we connect remotely via ssh to a workstation (bscearthXXX.int.bsc.es)
  - To ssh from windows: https://earth.bsc.es/wiki/doku.php?id=computing:sshwindows
  - To set up passwordless ssh connection: https://earth.bsc.es/wiki/doku.php?id=computing:sshkeyautologon
- ★ When we connect to one of the servers or HPC machines in BSC (MN5, Nord3v2, etc.)

PARTITIONS

What do we know about BSC-ES infrastructure?

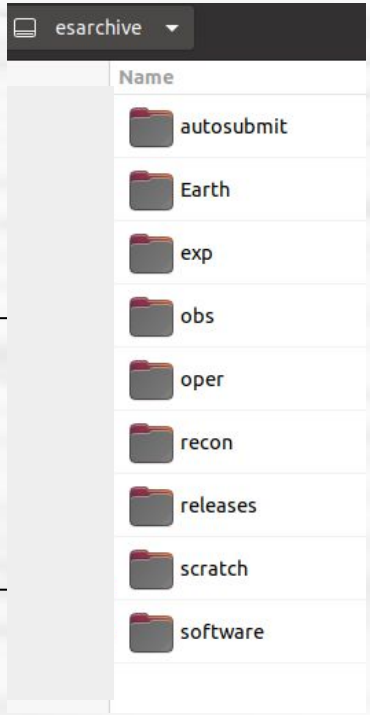
When we connect to the BSC infrastructure, we find several **partitions**. A disk partition, or simply 'partition', is a segment of a hard drive that is separate and independent from other segments. Each partition serves a different purpose and is accessible from different machines.



First entry point
/home/Earth/<username>/
Internet connection available
Daily back-up, limited space: 50 GB per user
Tips: https://earth.bsc.es/wiki/doku.php?id=computing:workstations
https://earth.bsc.es/wiki/doku.php?id=computing:sharedstorage

Data storage for BSC-ES department
Personal folder: /esarchive/scratch/<username>/
Internet connection available (from WS and Hub only)
NO back-up, limited total space for all users
Tip: add esarchive partition in a laptop (forbidden in Windows):
https://earth.bsc.es/wiki/doku.php?id=computing:addesarchive

General storage accessible from all HPC machines (Nord3v2, CTE-AMD, MN5)
Personal folder: /gpfs/scratch/bsc32/bsc032XXX/
No internet connection
Limited total space for all users in ES



PARTITIONS

What do we know about BSC-ES infrastructure?

It is also possible to connect to BSC infrastructure through **servers** (physical machines), which have different uses:

- ★ **bscearth000.int.bsc.es and bscearth001.int.bsc.es**
  - Download data
  - run the automatic package tests (GitLab CI/CD, see e.g.: https://earth.bsc.es/gitlab/es/s2dv/-/pipelines)
- ★ **transfer1.bsc.es (formerly dt01.bsc.es and dt02.bsc.es)**
  - Internal transfer of data, e.g. from esarchive to GPFS and vice versa.
- ★ **bscesshiny01.bsc.es**
  - Shiny server, hosts shiny apps.
- ★ **bscesftp.bsc.es**
  - Share files externally, see: https://earth.bsc.es/wiki/doku.php?id=computing:public_ftp
- ★ **bscesautosubmit01.bsc.es and bscesautosubmit02.bsc.es**
  - Launch workflows with the Autosubmit workflow manager https://earth.bsc.es/wiki/doku.php?id=tools:autosubmit

SERVERS

What do we know about BSC-ES infrastructure?

A **software stack** is the collection of programs and modules (including the operating system, architectural layers, protocols, runtime environments, …) that are installed in a machine.

- ★ The software stack at BSC can be different among different machines and departments
- ★ We have access to:
  - BSC software stack (not managed by CES)
  - BSC-ES software stack (managed by CES)
    - Workstations, Nord3v2 and CTE-AMD already using it
    - Hub has a slightly different software stack (more updated, but on testing status)
    - In some machines, we should edit the **bashrc** to use it (instructions are always in the wiki: https://earth.bsc.es/wiki/doku.php?id=library:computing)
    - It is built on **modules**, some useful commands are:
      - *module list* # show all loaded modules
      - *module load ** # load the '*' module
      - *module av ** # show all available modules matching '*'
    - other software programs like mendeley can be open in the workstation: */shared/earth/software/mendeley/latest/bin/mendeleydesktop*
- ★ Open an issue in **the Requests GitLab** to ask for new software or R packages

SOFTWARE STACK

What do we know about BSC-ES infrastructure?

What information do we need to know for each machine?

- Does it have BSC-ES software?
- is /esarchive/ mounted?
- Internet access?
- Job scheduler: slurm, lsf…?
- Memory per node, cores per node….

Hub
Workstations (WS)
Marenostrum 5
AMD cluster
Nord3_v2

Find the information here: https://earth.bsc.es/wiki/doku.php?id=library:computing

SOFTWARE STACK

What do we know about BSC-ES infrastructure?

Workstations

- R/4.1.2
- To be used for debugging code (small data) or running startR workflows in remote machines
- Internet connection
- BSC-ES software stack
- /esarchive is mounted

Hub (testing phase)

- R/4.2.1
- To be used for debugging code (small data) or running small jobs. Will replace workstations.
- Internet connection
- BSC-ES software stack
- /esarchive is mounted

Nord3_v2

- R/4.1.2
- To be used to run more memory-intensive jobs
- job scheduler: **slurm**
- No internet connection
- BSC-ES software stack
- /esarchive is mounted
- **will be decommissioned (when?)**

Marenostrum 5

- 'Pre-pre-production' status
- To be used to run more memory-intensive jobs
- **BSC-ES software stack currently not available**, conda environments can be installed
- internet access in login node 4
- no access to /esarchive (non-negotiable)

CTE-AMD

- R/4.1.2 or R/4.3.3 (for R-INLA)
- To be used to run more memory-intensive jobs
- job scheduler: **slurm**
- BSC-ES software stack
- no access to /esarchive (for now?)

Nord4

- Coming soon?
- /esarchive?

MACHINES

What do we know about BSC-ES infrastructure?

RECOMMENDATIONS

Recommendations

★ **Save your scripts in GitLab (intermediate and final versions)**
- In an existing GitLab project
- In a personal project
- Documentation: https://earth.bsc.es/wiki/doku.php?id=library:computing#git
- If you have internet connection, you can source your code directly from GitLab
- Clone repositories under /esarchive/scratch/<username>/
  - You will have internet connection to push your changes
  - The code will be accessible from workstations, hub and Nord3v2
  - There is no back-up copy of /esarchive (another good reason to use gitlab)

★ **Don't install local versions of R packages**
- If you do, we cannot debug the code and reproduce the errors
- Better to open an issue in Requests to ask for the installation: it's easier to debug and everyone can use it

★ **Infrastructure in the wiki:**
https://earth.bsc.es/wiki/doku.php?id=library:best_practices#network_infrastructure

What do we know about BSC-ES infrastructure?

Q&A: What else do we need to know? What questions do we have?

- Q: When will we be able to use Nord4?
  A: There is no official date yet.
- Q: Will Nord4 and/or CTE-AMD have /esarchive mounted?
  A: It is currently being negotiated, it is likely that at least one of the two may have it, but we do not know for sure.
- Q: Can we use RStudio in the Hub?
  A: Not right now, but CES is working on it. Requests issue:
  https://earth.bsc.es/gitlab/es/requests/-/issues/2154

Thanks for joining

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Next meeting: June 6th