Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# R user meeting

An-Chi Ho, Eva Rifà, Victòria Agudetse

contributor: Eren Duzenli

05/10/2023

# Agenda

1. Ice-breaker: Package “cli”
2. News
   - s2dv
   - startR
   - CSTools
   - CSIndicators
   - SUNSET
   - New plotting package
3. Presentation:  randomForest [Eren]
4. Q&A

# Ice-breaker

# cli

Helpers for Developing Command Line Interfaces

From [README](): 

A suite of tools to build attractive command line interfaces (CLIs), from semantic elements: headers, lists, alerts, paragraphs, etc. Supports theming via a CSS-like language. It also contains a number of lower level CLI elements: rules, boxes, trees, and Unicode symbols with ASCII alternatives. It supports ANSI markup for terminal colors and font styles.

Example:

https://earth.bsc.es/gitlab/aho/aho-testtest/-/blob/master/script_cli.R

# s2dv

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# New release 2.0.0 (coming soon)

Compatibility break! Please update your scripts accordingly.

- The original functionality remains the same but some parameters are changed, name or default value.
- If you're sourcing the functions from master → Remove the source line in the scripts
- If you're using the installed version (1.4.1) → Modify the function usage if error appears.

Check NEWS.md:
https://earth.bsc.es/gitlab/es/s2dv/-/blob/master/NEWS.md

# obs is not required to have member dimension

ACC(), Ano_CrossValid(), RMS(), Corr(), and RatioSDRMS() parameter "memb_dim" is optional for obs.

E.g.,

```
exp <- array(rnorm(20), dim = c(sdate = 5, member = 4))
obs <- array(rnorm(5),  dim = c(sdate = 5, member = 1))
res <- RMS(exp, obs, memb_dim = 'member')
```

# Choose different test type of RandomWalkTest()

AbsBiasSS(), RPSS(), CRPSS() have parameter “sig_method.type” to choose different types of Random Walk Test. It is corresponding to parameter “test.type” in RandomWalkTest().

[Recap]

“test.type”: Whether forecaster A and forecaster B are significantly different in terms of skill with…

- "two.sided.approx": a two-sided test
- "two.sided": an exact two-sided test
- "greater": an one-sided test for negatively oriented scores
- "less": an one-sided test for positively oriented scores).

# Decide NA threshold

New parameter "`na.rm`" in RPS() and RPSS() to decide the acceptable fraction of NA values in the data. It can be:

- TRUE: NAs are allowed
- FALSE: No NA is allowed (Default)
- A numeric value between 0 and 1: The lower limit for the fraction of the non-NA values. 1 is equal to FALSE (no NA is acceptable), 0 is equal to TRUE (all NAs are acceptable).

If the fraction of non-NA values in the data is less than the threshold, the function returns NA.

# startR

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# Use case of Autosubmit

In use case ex2_1_timedim.R, the snippet of Compute() with Autosubmit is added.
https://earth.bsc.es/gitlab/es/startR/-/blob/master/inst/doc/usecase/ex2_1_timedim.R?ref_type=heads#L94

Reminder: You can find details in practical_guide.md:
https://earth.bsc.es/gitlab/es/startR/-/blob/master/inst/doc/practical_guide.md#4-3-3-compute-on-hpcs-with-autosubmit

# CSTools

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# Accept spatial NAs in CST_MultiEOF

**Accept spatial NAs:** If one grid point has NAs for each variable, all time steps at this point should be NAs → consistent along `time_dim`

**Other changes**

- New parameters: `time_dim`, `sdate_dim`, `var_dim` and `ncores`.
- Changed parameter name from `time` to `dates`.
- CST output: Ilist of 's2dv_cubes', keeping the metadata of the original data.
- Improved checks.

```
Before:
> result <- CST_MultiEOF(datalist = list(exp1, exp2), neof_composed = 2)
Error in CST_MultiEOF(datalist = list(exp1, exp2), neof_composed = 2) :
  Input data contain NA values.
```

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

status: In branch master
Check issue: https://earth.bsc.es/gitlab/external/cstools/-/issues/127

# Correct Calibration warning for ncores > 1

**Problem**

- Calibration must return a warning when there aren't enough samples, but when multiple cores are used it doesn't appear.
  → The warning message was inside the atomic function of Apply

**Solution**

- Replace the warning outside Apply (as it was in the original code)

```
NOTE: Avoid warning / error messages inside atomic functions.
Instead, use: tryCatch(), writting log files, save a value and
check it outside Apply, …
```

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

status: In branch master
Check issue: https://earth.bsc.es/gitlab/external/cstools/-/issues/130

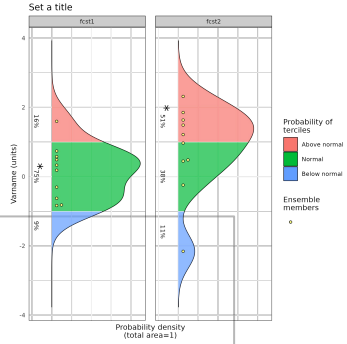
# Correct PlotForecastPDF background in Nord3

**Problem**

- When using PlotForecastPDF() in Nord3 the background was transparent.
  → Because the graphic devices are different in each machine.

**Solution**

- Added some lines specifying the background color:




```
plot <- plot +
        theme_minimal() +
        theme(
              panel.background = element_rect(fill = "white"),
              plot.background = element_rect(fill = "white", color = NA))
        )
```

**NOTE:** Function to find information about the graphic device: capabilities()

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

status: In branch master
Check issue: https://earth.bsc.es/gitlab/external/cstools/-/issues/131

# Improve CST_SaveExp warnings and metadata

## Changes

1. Corrected saved metadata in the netCDF that contained character vectors

```
s2dv_cube:
[...]
$metadata
$metadata$region$lats_range
[1] "20" "80"
```

→

```
(Error) NetCDF:
  int region(region) ;
        region:name = "NAO region" ;
        region:lats_range = "20" ;
```

```
(Now) NetCDF:
  int region(region) ;
        region:name = "NAO region" ;
        region:lats_range = "20, 80" ;
```

2. New parameter `drop_dims` to drop dimensions of length 1.

```
dim(data$data)
# region    syear ensemble      dat      var     sday    sweek     time
#      1       24       25        1        1        1        1        1
```

**Example:**

```
drop_dims = c('dim1', 'dim2')
```

```
dim1    dim2
  1       1
```

3. Removed warnings regarding not having longitude and latitude dimensions.

```
Warning messages:
1: In SaveExp(data = data$data, destination = destination, Dates = data$attrs$Dates,  :
  Spatial coordinates not found.
```

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

status: In branch develop-SaveExp_warnings_and_metadata
Check issue: https://earth.bsc.es/gitlab/external/cstools/-/issues/132

# New function CST_Start: changes in the package

- **General changes:**
  - Update all the vignettes with substituting CST_Load() by CST_Start()
  - New sample data: `lonlat_temp_st`, `lonlat_prec_st`

```
> lonlat_prec_st$dims
dataset     var  member   sdate   ftime     lat     lon
      1       1       6       3      31       4       4
```

status: In branch master
Check issue: https://earth.bsc.es/gitlab/external/cstools/-/issues/126

- **Other changes:**
  - CST_Analogs(): Add `sdate_dim` parameter and improve initial checks

    status: In branch develop-CST_Analogs_improve
  - CST_Anomaly(): Allow `memb_dim` to be NULL in obs (remove checks)

    status: In branch develop-CST_Anomaly_checks

# Next in CSTools

**Next developments:**

- Improve as.s2dv_cube(): add `time_frequency` attribute
- Improve s2dv_cube(): remove unnecessary warnings and change dates class
- Improve CST_MultiMetric() ?

**Next release:**

- As soon as possible: planned on **October 16th** (or a bit later if possible)

# CSIndicators

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# New functions: PeriodMax, PeriodMin, PeriodVariance

- New functions in CSIndicators to compute **Bioclimatic Indicators**:
  - BIO1 = Annual Mean Temperature --> PeriodMean()
  - BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp)) --> PeriodMean()
  - BIO3 = Isothermality (BIO2/BIO7) (×100) --> vignette
  - BIO4 = Temperature Seasonality (standard deviation ×100) --> NEW PeriodVariance (?)
  - BIO5 = Max Temperature of Warmest Month --> NEW PeriodMax() (?)
  - BIO6 = Min Temperature of Coldest Month --> NEW PeriodMin() (?)
  - …

## How to use them?

- They compute the max (min, mean, variance or accumulation) through `time_dim`
- We can specify a period subset with `start` and `end` parameters
- The output is an array. The **CST** version keeps the original metadata and `time_bounds` are added

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

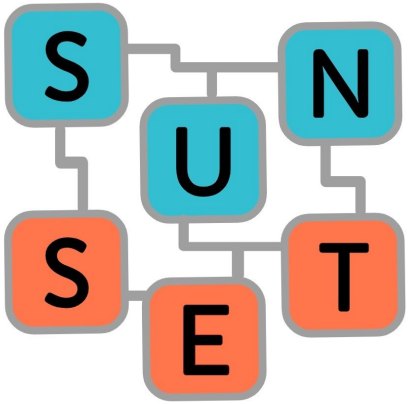
status: In branch master
Check issue: https://earth.bsc.es/gitlab/external/cstools/-/issues/127

# SUNSET

# New logo!

SUNSET has a new logo design!

Feel free to use it in presentations, documents or anywhere else :)




Design by An-Chi Ho

# New module: Downscaling

The new Downscaling module integrates the in-house CSDownscale package into SUNSET, allowing the user to downscale a hindcast and its corresponding observations.

After downscaling, the data can be saved and/or the workflow can continue with other SUNSET modules, such as Skill or Visualization.

You can find an example script here: https://earth.bsc.es/gitlab/es/sunset/-/blob/master/example_scripts/example_downscaling.R

status: in master

# New module: Downscaling

An example of the Downscaling module section of the recipe:

```
    Downscaling:
      type: intbc # 'none', 'int', 'intbc', 'intlr', 'analogs', 'logreg'.
      int_method: conservative # regridding method accepted by CDO.
     bc_method: bias # 'bias', 'calibration', 'quantile_mapping', 'qm', 'evmos', 'mse_min',
'crps_min', 'rpc-based'.
      lr_method: # If type=intlr. Options: 'basic', 'large_scale', '4nn'
              log_reg_method:  #  If  type=logreg.  Options:  'ens_mean',  'ens_mean_sd',
'sorted_members'
      target_grid: r1440x760 # path to nc file or grid description accepted by CDO
      nanalogs: # If type=analogs. Number of analogs to be searched. Default is 3.
      save: 'all' # 'all'/'none'/'exp_only'
```

status: in master

# SUNSET launcher

A script to split a recipe into atomic recipes and then help launch the verifications for each atomic recipe as an independent, parallel job.
If the autosubmit parameter is set to ‘yes/True’ in the recipe, it will provide instructions to launch the Autosubmit experiment. If not, it will send the jobs to be executed on the machine directly through sbatch.

```
bash launch_SUNSET.sh <path_to_recipe> <path_to_script>
```

Additional flags:

`--wallclock=<HH:MM:SS>`: The wallclock time for the jobs

`--custom_directives=<custom_directives>`: Custom directives for SLURM

`--ncpus=<ncpus>`: Number of CPUs to be requested by SLURM for each job

status: in branch dev-test_CERISE

# Scorecards: plot Scorecards for multiple datasets

A bug in the Scorecards module has been fixed and now the Scorecards job will plot the Scorecards for all available datasets i.e. for every variable, system and reference requested in the recipe.

Remaining bug in CSScorecards: If the Scorecards are plotted in a loop and more than 1 core is used in the computation, the R session ends with an error:

```
Error while shutting down parallel: unable to terminate some child processes
```

Any ideas? Please report in the issue:
https://earth.bsc.es/gitlab/ess/csscorecards/-/issues/7

status: in branch dev-test_CERISE

# New plotting package

# ShapeToMask new parameter

- **New parameter:**
  - Added parameter `find_min_dist` in the function **ShapeToMask**. If it is TRUE, it searches for the closest coordinate when there is no intersection.

- **Next developments:**
  - Substitute for loop for each region with multiApply
  - Add initial checks
  - Add unit test
  - Add saving option



status: In branch main
Check issue: https://earth.bsc.es/gitlab/es/climplot2/-/issues/4

# Name Proposal

**ESViz**

**esviz**

PlotEquiMap → **Viz**EquiMap
PlotRobinson → **Viz**Robinson
PlotForecastPDF → **Viz**ForecastPDF
PlotMostLikelyQuantileMap → **Viz**MostLikelyQuantileMap

Previous discussion…

**CSPlot**: Not only Climate Services; Confusing with s2dv_cube concept

**PICS(R)**: Sounds too broad; Sound the same as PIXAR® 🚨🚓

# [User presentation] ‘randomForest’ by Eren

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# ‘randomForest’ package

*The Random Forest is a popular machine learning technique that brings together the results of regression trees to arrive at a final outcome. It can effectively deal with both classification and regression issues.

Functions: randomForest, predict.randomForest, plot.randomForest etc.

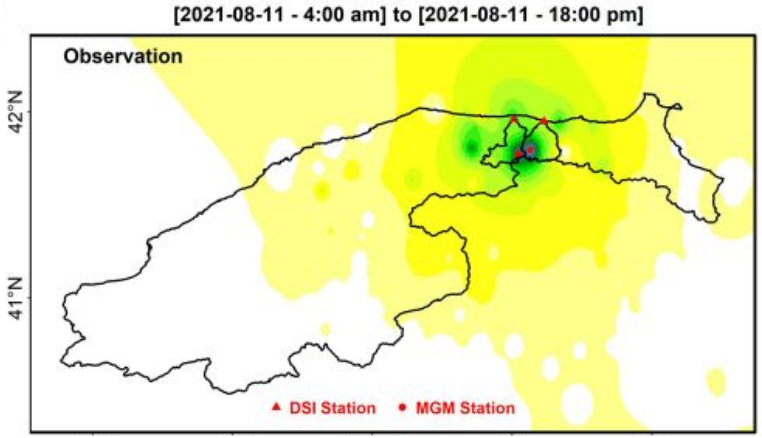
```
arguments:
  data: data.frame including both predictors and predictand
  ntrees: number of multiple decision trees
  importance: Should importance of predictors be assessed?
  replace: Should sampling of cases be done with or without replacement?
```

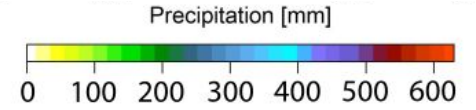
Why this method has been used?

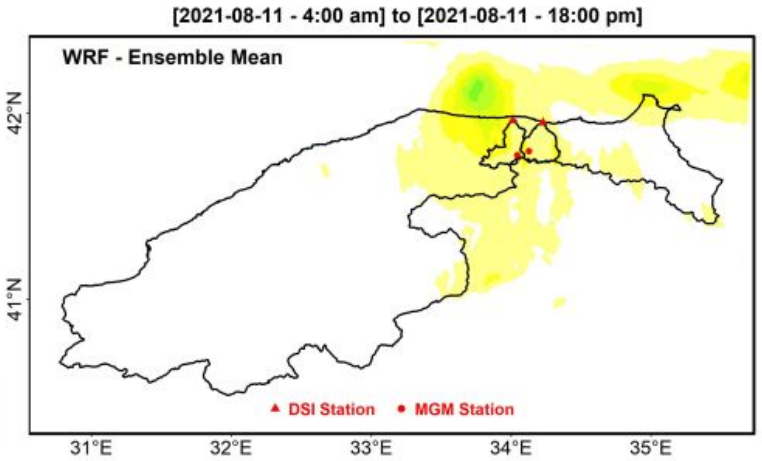
to merge 24 precipitation products obtained from WRF models initiated with different parameterization combinations.

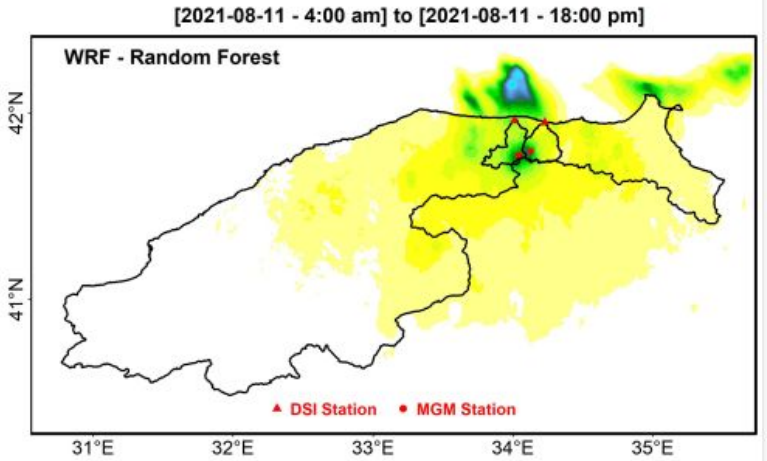
*https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

# Contribution of the randomForest

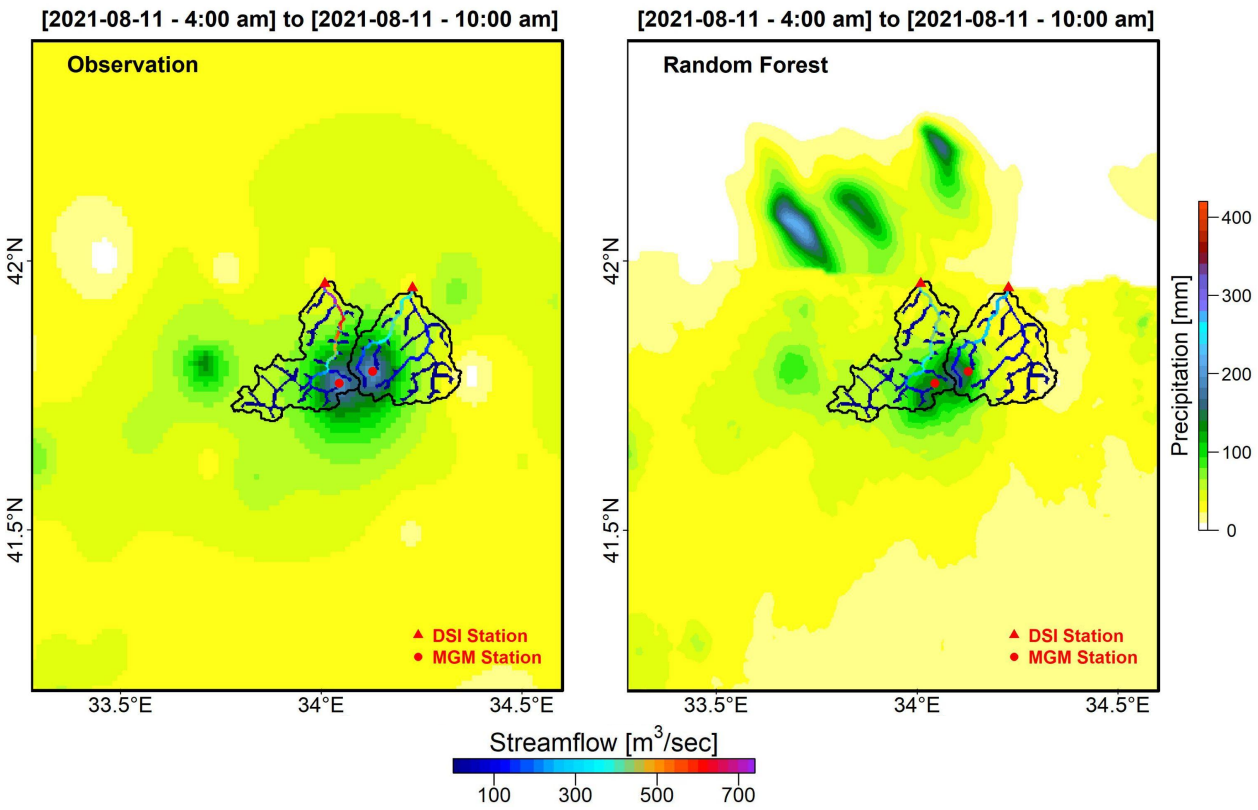


# Contribution of the randomForest



# Thanks for joining

Next meeting: 2nd November, 12h