in C. Chen et al. (Eds.), 2018: Proceedings of the 8th International Workshop on Climate Informatics (CI2018). NCAR Technical Note NCAR/TN-550+PROC, doi: 10.5065/D6BZ64XQ
***including corrections in Sec. III and Figure 2.***

8th International Workshop on Climate Informatics
September 19-21, 2018
CI 2018 Hosted by the National Center for Atmospheric Research in Boulder, CO

# Climate Research Reproducibility with the climate4R R-based Framework

José Manuel Gutiérrez[1] , Joaquín Bedia[2,3], Maialen Iturbide[1], Sixto Herrera[2], Rodrigo Manzanas[1], Jorge Baño-Medina[1], María Dolores Frías[2], Daniel San-Martín[3], Jesús Fernández[2], Antonio Cofiño[2]

More information at `http://www.meteo.unican.es/climate4R`

***Abstract*—Climate driven sectoral applications in a variety of domains (such as hydrology, agriculture, energy, or health) typically require elaborated data processing workflows involving multiple data access, collocation, harmonization and postprocessing (e.g. bias correction) steps. This is a time-consuming and error-prone task which, in many cases, is performed with different tools and lack of appropriate metadata for reproducibility. The R-based `climate4R` framework provides a solution to this problem, building on open source software and standards. `climate4R` allows accessing, postprocessing and visualizing local and remote (OPeNDAP) data sources, providing also full provenance information via METACLIP (semantic MEtadata for CLIimate Products). As a result, `climate4R` provides a unique comprehensive framework for end-to-end fully reproducible sectoral studies favoring open science.**

## I. Motivation

Research transparency and reproducibility is an issue of major concern in all experimental disciplines (see e.g. go.nature.com/huhbyr and [1]). In climate science, data access and post-processing (e.g. regridding, aggregation, index calculation) are common steps of the data workflow which are often not appropriately documented, thus hampering the reproducibility of the results. Moreover, recent popular postprocessing techniques, such as bias correction, are very technical and requiere community-driven specialized vocabularies for full reproducibility. Here we describe `climate4R` (*climate for R*) [2], an R-based framework for climate studies where most common tasks can be performed using a few lines of code, allowing end-to-end experimental reproducibility and facilitating the description (metadata) and documentation of the whole workflow — from data access and postprocessing, to climate product generation (dataset or graphic).—This is done through an extension of the METACLIP (semantic METAdata for CLImate Products; http://www.metaclip.org) RDF-based provenance framework, which automatically generates semantic modular metadata for `climate4R` through domain-specific extensions of standard vocabularies [3]. An up-to-date description of `climate4R` and METACLIP, including information on the available packages and datasets is provided in the wiki page http://www.meteo.unican.es/climate4r

## II. The climate4R Framework

The `climate4R` framework consists on three layers (see a schematic representation in Fig. 1): (a) Data services building on NetCDF-Java and THREDDS to provide access to local or remote data, including datasets from the in-house User Data Gateway (UDG); (b) The `climate4R` R bundle for data access and post-processing, formed by four core packages for data loading, transformation, downscaling (including bias correction) and visualization; (c) the metadata layer, based on METACLIP integrated with the four core packages, which are described in further detail below:

- `loadeR` loadeR is the central building-block of `climate4R` and allows to transparently access local and remote climate datasets building on NetCDF-Java. `loadeR` goes beyond the file-oriented concept for data access, supporting reading (and writing) CDM datasets, i.e. "collections" of NetCDF files, instead of individual files, so users do not need to worry about a particular directory tree structure or file naming schema, and a single URL pointing to the dataset is need. Besides local and remote OPeNDAP datasets, `climate4R` is transparently connected to the User Data Gateway (UDG), a climate data service hosted by University of Cantabria (http://meteo.unican.es/udg-wiki) providing state-of-the-art global and regional climate projections such as those from the CMIP5 [4] and CORDEX [5].

Correspondence: J.M. Gutiérrez, gutierjm@ifca.unican.es
[1]Institute of Physics of Cantabria (CSIC-UC), Santander, Spain
[2]Dept. of Applied Mathematics and Computation. UC, Santander
[3]PREDICTIA Intelligent Data Solutions, Santander

- `transformeR` transformeR performs common data processing tasks such as regridding/interpolation, subsetting or spatio-temporal aggregation, among others.
- `downscaleR` implements several statistical downscaling and bias correction methods [6]. The latter adjust directly the target variable predicted by the climate model, using as reference the corresponding local observations. Due to their simplicity, these methods have become very popular during the last decade. However, it is important to understand their assumptions and limitations in order to avoid the misuse of these techniques [7]. The `biasCorrection` function is the workhorse to apply several standard bias correction techniques, including the popular empirical quantile mapping (EQM) [8].
- `visualizeR` [9] is an R package for climate data visualization, implementing basic visualization functionalities for gridded and point-based data, time series, and a set of advanced tools for forecast visualization in a form suitable to communicate the underlying uncertainty.

Besides these core packages, `climate4R` extends its capabilities by integrating the functionalities of other external packages via wrapping packages. For instance, the wrapper `climate4R.climdex` allows to transparently compute the 27 ETCCDI indices [10] for extremes implemented in the publicly available package `climdex.pcic` [11].

## III. An Illustrative Example

In this example we showcase the main functionalities of `climate4R` by describing the complete workflow to compute and postprocess an ETCCDI climate index. In particular, we consider summer days (SU) —defined as the number of days with maximum temperature $> 25°C$,— over a Mediterranean domain and use data from a EURO-CORDEX Regional Climate Model (RCM) [12] to obtain and adjust future SU projections. The different steps of the example are shown in Figure 2 (panels a to e), which shows the code (left) and resulting figures (right).

First (panel a), observational data for daily maximum temperature is obtained remotely for a particular geographical domain (Southern Europe) and temporal period (1971-2000) from the EOBS ECA opendap server, using the function `loadGridData`. The function `climdexGrid` allows to easily compute the annual values of the SU index from this data. The results can be easily plot with the function `spatialPlot`. The same procedure is applied to CORDEX historical data for the same period (panel b), which is directly available from `climate4R` via the UDG defined datasets (details for each simulation are provided along with the URLs needed to load this information). Since CORDEX data is originally in rotated coordinates, the third block of code (panel c) illustrates how regridding and masking can be easily performed, so the bias of the model can be computed in the regular E-OBS grid (panel d).

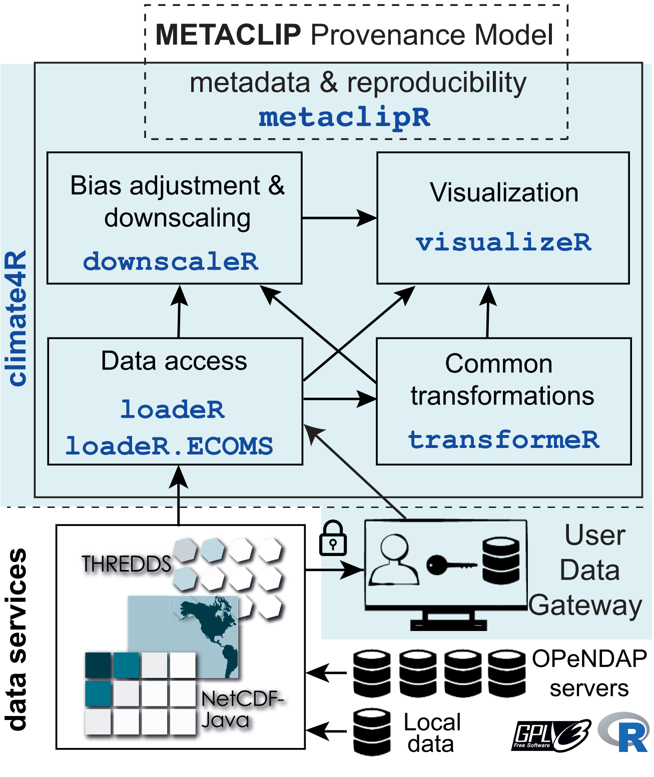
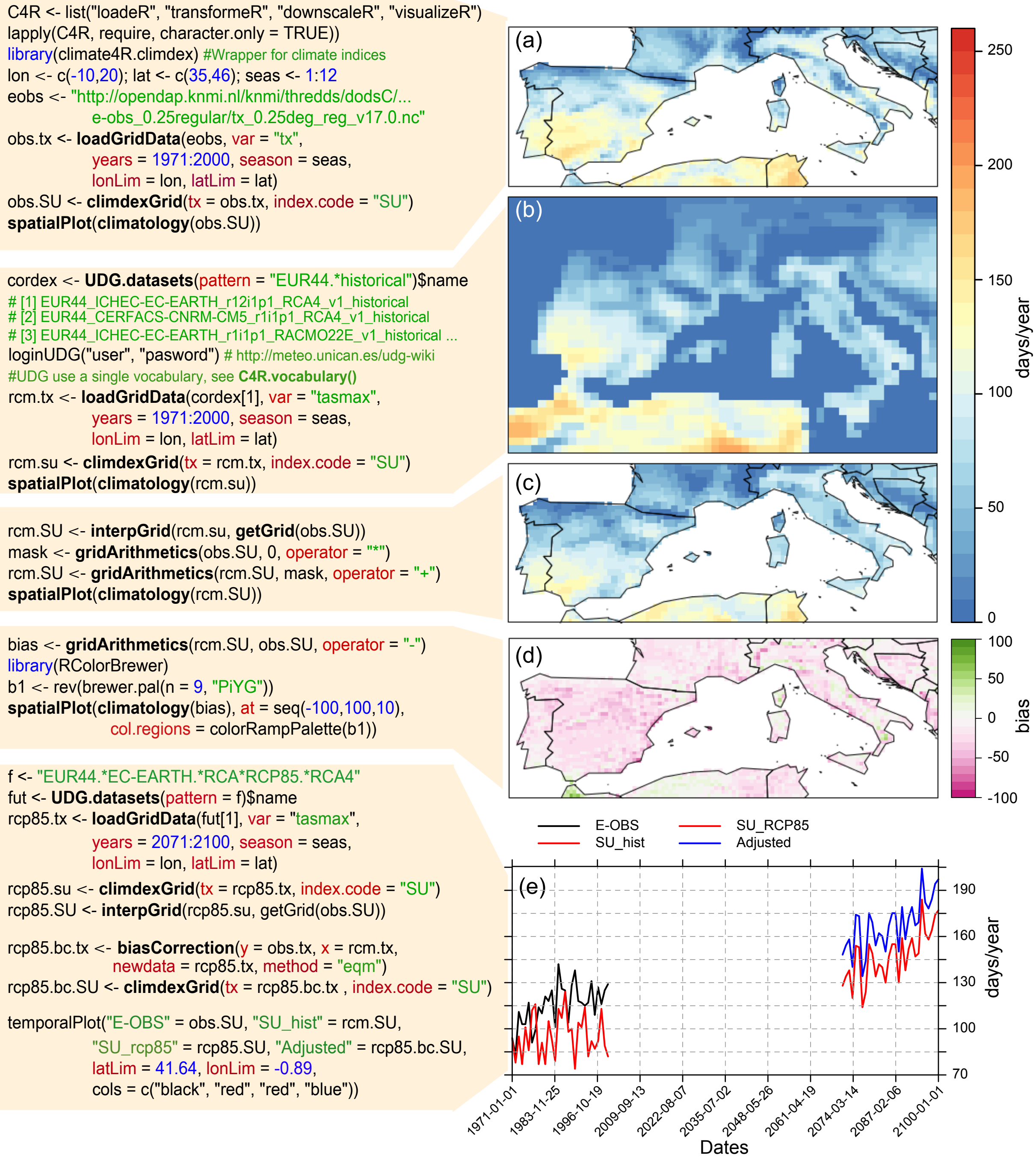


Fig. 1. Schematic illustration of the layers and components of the `climate4R` R-based framework. All components are distributed under GNU General Public License.

Finally (panel e), future projections from the same model are loaded from the RCP8.5 scenario (2071-2100) and bias adjusted using the `biasCorrection` function. In order to compare the differences in the resulting SU index due to bias adjustment, it is computed both before and after the adjustment of maximum temperature. The function `temporalPlot` allows to visualize the resulting time series (before and after the adjustment, in blue and red color, respectively) for a particular gridbox (the closest to Zaragoza, within a region with high model bias). This figure illustrates the high sensitivity of the results to the model biases, particularly relevant for threshold-dependent indices as in thre present case.

All the figures generated in the above example have attached information with the full metadata description,



```
C4R <- list("loadeR", "transformeR", "downscaleR", "visualizeR")
lapply(C4R, require, character.only = TRUE))
library(climate4R.climdex) #Wrapper for climate indices
lon <- c(-10,20); lat <- c(35,46); seas <- 1:12
eobs <- "http://opendap.knmi.nl/knmi/thredds/dodsC/...
            e-obs_0.25regular/tx_0.25deg_reg_v17.0.nc"
obs.tx <- loadGridData(eobs, var = "tx",
            years = 1971:2000, season = seas,
            lonLim = lon, latLim = lat)
obs.SU <- climdexGrid(tx = obs.tx, index.code = "SU")
spatialPlot(climatology(obs.SU))
```

```
cordex <- UDG.datasets(pattern = "EUR44.*historical")$name
# [1] EUR44_ICHEC-EC-EARTH_r12i1p1_RCA4_v1_historical
# [2] EUR44_CERFACS-CNRM-CM5_r1i1p1_RCA4_v1_historical
# [3] EUR44_ICHEC-EC-EARTH_r1i1p1_RACMO22E_v1_historical ...
loginUDG("user", "pasword") # http://meteo.unican.es/udg-wiki
#UDG use a single vocabulary, see C4R.vocabulary()
rcm.tx <- loadGridData(cordex[1], var = "tasmax",
            years = 1971:2000, season = seas,
            lonLim = lon, latLim = lat)
rcm.su <- climdexGrid(tx = rcm.tx, index.code = "SU")
spatialPlot(climatology(rcm.su))
```

```
rcm.SU <- interpGrid(rcm.su, getGrid(obs.SU))
mask <- gridArithmetics(obs.SU, 0, operator = "*")
rcm.SU <- gridArithmetics(rcm.SU, mask, operator = "+")
spatialPlot(climatology(rcm.SU))
```

```
bias <- gridArithmetics(rcm.SU, obs.SU, operator = "-")
library(RColorBrewer)
b1 <- rev(brewer.pal(n = 9, "PiYG"))
spatialPlot(climatology(bias), at = seq(-100,100,10),
              col.regions = colorRampPalette(b1))
```

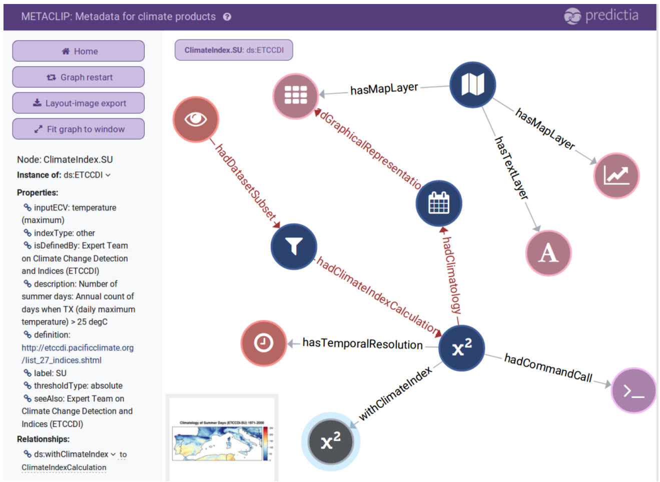
```
f <- "EUR44.*EC-EARTH.*RCA*RCP85.*RCA4"
fut <- UDG.datasets(pattern = f)$name
rcp85.tx <- loadGridData(fut[1], var = "tasmax",
            years = 2071:2100, season = seas,
            lonLim = lon, latLim = lat)
rcp85.su <- climdexGrid(tx = rcp85.tx, index.code = "SU")
rcp85.SU <- interpGrid(rcp85.su, getGrid(obs.SU))

rcp85.bc.tx <- biasCorrection(y = obs.tx, x = rcm.tx,
              newdata = rcp85.tx, method = "eqm")
rcp85.bc.SU <- climdexGrid(tx = rcp85.bc.tx , index.code = "SU")

temporalPlot("E-OBS" = obs.SU, "SU_hist" = rcm.SU,
            "SU_rcp85" = rcp85.SU, "Adjusted" = rcp85.bc.SU,
            latLim = 41.64, lonLim = -0.89,
            cols = c("black", "red", "red", "blue"))
```

Fig. 2. Southern Europe summer days (ETCCDI SU index) climatology for the reference period 1971-2000 according to: (a) 0.22° E-OBS gridded observations dataset, (b) 0.44° RCA regional climate model (driven by EC-EARTH GCM, historical scenario), (c) same as (b), but after regridding onto the regular E-OBS grid and (d) RCM bias (days/year) w.r.t. E-OBS. Panel (e) shows the raw and bias corrected data for a particular gridbox (near Zaragoza, Spain) for a historical and future (2071-2100, RCP8.5) period. The model (red) is corrected in the future (blue) based on the relationships with observations in the present (black).

in RDF format (compression is applied to minimize the file size overhead). The METACLIP Interpreter (http://metaclip.org/interpreter) allows to extract and explore this embedded information in a user-friendly way, making use of the semantic description of the metadata to provide modular access to the information with incremental levels of detail. It has a drag-and-drop area where all products (e.g. images) with attached METACLIP information can be dropped for metadata interactive visualization. Figure 3 shows an example of this facility for the map shown in Fig. 2c, which is shown in the bottom left part of the interpreter.



Node: ClimateIndex.SU
**Instance of:** ds:ETCCDI
**Properties:**
- inputECV: temperature (maximum)
- indexType: other
- isDefinedBy: Expert Team on Climate Change Detection and Indices (ETCCDI)
- description: Number of summer days: Annual count of days when TX (daily maximum temperature) > 25 degC
- definition: http://etccdi.pacificclimate.org/list_27_indices.shtml
- label: SU
- thresholdType: absolute
- seeAlso: Expert Team on Climate Change Detection and Indices (ETCCDI)

**Relationships:**
- ds:withClimateIndex to ClimateIndexCalculation

Fig. 3. A snapshot of the METACLIP Interpreter displaying the provenance representation of Figure Fig. 2c. Here, the metadata of the climate index node (the grey node labelled $X^2$) is displayed on the left panel, providing the information of the definition and code for reproducibility. The information of the different nodes can be interactively queried by the user. Double-clicking each node will expand it to further nodes displaying other sub-properties and their corresponding annotations, until the lowest representation level is reached.

## IV. Conclusions

This work presents the open `climate4R` R-based framework for accessing and post-processing climate data and describes its main components —data services, core packages, metadata generation and external packages— and functionalities via an illustrative case study. This provides a unique comprehensive open framework for end-to-end sectoral reproducible applications. All the packages, data and documentation for reproducing the experiments in this paper are available from http://www.meteo.unican.es/climate4r

## Acknowledgments

Funding for the authors was provided by the project MULTI-SDM (CGL2015-66583-R, MINECO/FEDER).

## References

[1] M. Baker, “1,500 scientists lift the lid on reproducibility,” *Nature News*, vol. 533, no. 7604, p. 452, 2016.

[2] M. Iturbide, J. Bedia, S. Herrera, J. Bano, J. Fernández, M. Frías, R. Manzanas, D. San-Martín, E. Cimadevilla, A. Cofino, and J. M. Gutiérrez, “climate4R: An R-based Open Framework for Reproducible Climate Data Access and Post-processing,” *Submitted to Environmental Modelling and Software*.

[3] J. Bedia, D. San-Martín, M. Iturbide, S. Herrera, and J. M. Gutiérrez, “METACLIP: A semantic provenance framework for climate products,” *Submitted to Environmental Modelling and Software*, submitted.

[4] K. E. Taylor, R. J. Stouffer, and G. A. Meehl, “An overview of CMIP5 and the experiment design,” *Bull. Amer. Meteor. Soc.*, vol. 93, pp. 485–498, Oct. 2011.

[5] F. Giorgi and W. J. Gutowski, “Regional dynamical downscaling and the CORDEX initiative,” *Annual Review of Environment and Resources*, vol. 40, no. 1, pp. 467–490, 2015.

[6] J. M. Gutiérrez, M. Maraun, D. abd Widmann, R. Huth, E. Hertig, R. Benestad, R. Roessler, T. Wibig, R. Wilcke, S. Kotlarski, D. San-Martín, S. Herrera, J. Bedia, A. Casanueva, R. Manzanas, M. Iturbide, and M. Vrac, “An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment,” *International Journal of Climatology*, 2018.

[7] D. Maraun, T. G. Shepherd, M. Widmann, G. Zappa, D. Walton, J. M. Gutiérrez, S. Hagemann, I. Richter, P. M. M. Soares, A. Hall, and L. O. Mearns, “Towards process-informed bias correction of climate change simulations,” *Nature Climate Change*, vol. 7, pp. 764–773, Nov. 2017.

[8] M. Déqué, “Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values,” *Global and Planetary Change*, vol. 57, pp. 16–26, May 2007.

[9] M. D. Frías, M. Iturbide, R. Manzanas, J. Bedia, J. Fernández, S. Herrera, A. S. Cofiño, and J. Gutiérrez, “An R package to visualize and communicate uncertainty in seasonal climate prediction,” *Environmental Modelling & Software*, vol. 99, pp. 101–110, Jan. 2018.

[10] T. R. Karl, N. Nicholls, and A. Ghazi, “CLIVAR/GCOS/WMO Workshop On Indices And Indicators For Climate Extremes. Workshop Summary,” *Climatic Change*, vol. 42, pp. 3–7, 1999.

[11] D. Bronaugh, *climdex.pcic: PCIC Implementation of Climdex Routines*, 2015. R package version 1.1-6.

[12] D. Jacob, J. Petersen, B. Eggert, A. Alias, O. B. Christensen, L. M. Bouwer, A. Braun, A. Colette, M. Déqué, G. Georgievski, E. Georgopoulou, A. Gobiet, L. Menut, G. Nikulin, A. Haensler, N. Hempelmann, C. Jones, K. Keuler, S. Kovats, N. Kröner, S. Kotlarski, A. Kriegsmann, E. Martin, E. van Meijgaard, C. Moseley, S. Pfeifer, S. Preuschmann, C. Radermacher, K. Radtke, D. Rechid, M. Rounsevell, P. Samuelsson, S. Somot, J.-F. Soussana, C. Teichmann, R. Valentini, R. Vautard, B. Weber, and P. Yiou, “Euro-cordex: new high-resolution climate change projections for european impact research,” *Regional Environmental Change*, vol. 14, pp. 563–578, Apr 2014.