Manipulation of netCDF data with R for climate change research: Multi-model analysis for CMIP5 models

Bruno Lopes de Faria[1], Hugo Prado Amaral[2]
[1, 2] Grupo de pesquisa em Geoinformática
Instituto Federal do Norte de Minas Gerais (IFNMG)
Pirapora – MG – Brazil
bruno.lopes@ifnmg.edu.br, hpifnmg@gmail.com

Abstract. Geoscientists now live in a world with an exponential growth in digital data and methods. Climate change studies usually describe computational methods informally. Climate scientists seek to share their information, the justification of reproducible research has received increasing attention in geosciences. To have it in an open-source format makes it easier to interchange not only with fellow scientists but also a variety of sources including funders, publishers, and journalists. R is a open-source computer language powerful and highly extensible that can promotes reproductive science techniques in a easier way. R is highly accessible for non-computational scientists when coupled with packages like 'raster', 'netcdf', ´rgdal`and 'rasterVis', R enables scientists to make sense of their data and to carry out complex data analysis. In this paper we have assessed the power of R language for manipulating climate data from a huge dataset: the Coupled Model Intercomparison Project Phase 5 (CMIP5). Moreover we have proposed an example of best practices to handle model ensembles. This is the first study to our knowledge to promote best practices for CMIP5 ensemble. The NetCDF data accessible to R via raster package capabilities provides efficient access to the multi-model, with crucial applications in climate change research. In recent years more than 100 peer-reviewed scientific publications have used the CMIP5 data sets. We envision that in the near future (5-10 years), scientists will use radically new tools to author papers and disseminate information about the process and products of their research.

keywords: *climate change, CMIP5, R, reproducible research*

1. Introduction

The exponential increase in the generation and collection of data has led us in a new era of data analysis and information extraction. In the past few years, R has increasingly become an important tool for scientific investigation. Before R, scientists might be required to master low-level computer languages like 'C' or 'Fortran' in order to analyze their data. Because much of the complexity found in these traditional programming languages has been abstracted away, R is highly accessible for non-computational scientists. When coupled with packages like 'raster', 'netcdf', ´rgdal`and 'rasterVis', R enables scientists to make sense of their data and to carry out complex data analysis in a practical manner. The netCDF file format is broadly used in the atmospheric sciences for data archival. Until recently, many scientists using the netcdf4 enhanced data model did not enjoy a reliable R solution when working with this data.

NetCDF is a widely used file format in atmospheric and oceanic research – especially for weather and climate model output – which allows storage of different types of array based data, along with a short data description. The NetCDF format (Network Common Data Format, http://www.unidata.ucar. edu/software/netcdf/) has been developed since 1988 by Unidata (a programme sponsored by the United States National Science Foundation) with the main goal of making best use of atmospheric and related data for education and research (Rew et al., 2011; Rew and Davis, 1990). NetCDF files are

stored as machine-independent binary data, such that files can be exchanged between computers without explicit conversion (Rew and Davis, 1990). Until version 3.6.0, only one binary data format was used. This is the default format for all NetCDF versions and is also named NetCDF classic format (Rew et al., 2011). Version 3.6.0 of the NetCDF library introduced the 64-bit offset format, which allowed addressing of much larger datasets; version 4.0.0 introduced also the HDF5 format, in a way that the NetCDF library can use HDF The classic model of netCDF represents data as a set of multi-dimensional arrays, with sharable dimensions, and additional metadata attached to individual arrays or the entire file. In netCDF terminology, the data arrays are variables, which may share dimensions, and may have attached attributes. Attributes may also be attached to the file as a whole. One dimension may be of unlimited length, so data may be efficiently appended to variables along that dimension. Variables and attributes have one of six primitive data types: char, byte, short, int, float, or double. Generally used to store large, multi- dimensional arrays A netCDF file includes metadata as well as data: names of variables, data locations in time and space, units of measure, and other useful information.

The study of climate change is important in decision-making for public policies, due to their social and economic impact. However, this analysis deals with a considerable mass of data. Handle effective tools to support this process is highly recommended, for increase productivity and reliability, especially if they provide analysis and visualization of climate change effects.

The goal of this paper is demonstrate an wide range application in climate change study how netcdf4-R can be used to read, subset, analyze, visualize and data stored in many formats by accessing NetCDF4 Data in R. Thus, providing tools and techniques for CMIP5 data analysis becoming more reproducible and reliable.

2. Methods

The justification of reproducible research has received increasing attention, particularly in climate science (Santer et al 2011). The latest Coupled Model Intercomparison Project Phase 5 (CMIP5) provides vast amounts of model simulations useful for scrutinizing the past and future climate change (Taylor, 2012). The computational expense and size of outputs for CMIP5 are much larger than its previous phase, CMIP3, due to the high resolution and complicated processes included in CMIP5 models. As more models are publicly available for intercomparison projects, it is expected that major climate science journals require sharing the data analysis procedure in publications and making analysis results reproducible and applicable to similar datasets.

2.1 Coupled Model Intercomparison Project Phase 5 (CMIP5).

The fifth phase of the Coupled Model Intercomparison Project (CMIP5) produce a state-of-the- art multimodel dataset designed to advance our knowledge of climate variability and climate change. Researchers worldwide are analyzing the model output and will produce results likely to underlie the forthcoming Fifth Assessment Report by the

Intergovernmental Panel on Climate Change. Unprecedented in scale and attracting interest from all major climate modeling groups, CMIP5 includes "long term" simulations of twentieth-century climate and projections for the twenty-first century and beyond.

The availability of such large amounts of data and the multi-dimensional nature induces an acute need for development of novel approaches to assist with the near real-time processing, visualization, and analysis by end users. (see Taylor et al. 2012).

2.2 Data analysis with R Raster package

The raster package provides classes and functions to manipulate geographic (spatial) data in 'raster' format. Raster data divides space into cells (rectangles; pixels) of equal size (in units of the coordinate reference system). Such continuous spatial data are also referred to as 'grid' data, and be contrasted with discrete (object based) spatial data (points, lines, polygons). The package should be particularly useful when using very large datasets that can not be loaded into the computer's memory. Functions will work correctly, because they process large files in chunks, i.e., they read, compute, and write blocks of data, without loading all values into memory at once to create a RasterBrick. NetCDF files are easy organized and accessed as rasterStack as a group, referred to as Raster objects. manipulations (Figure 1).

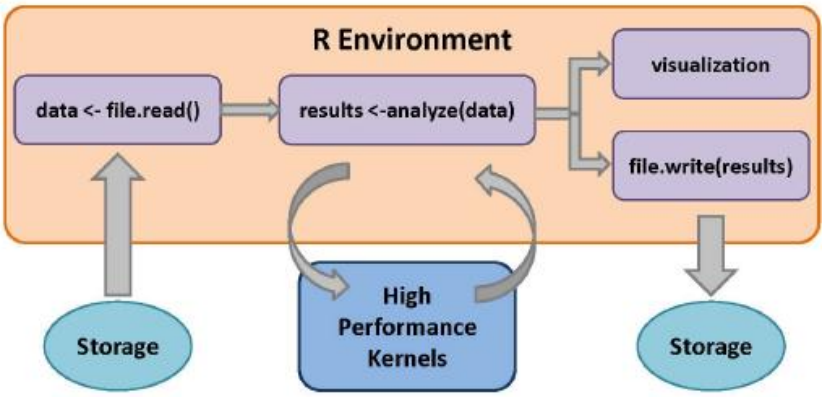

Figure 1 - Dataflow in the R enviroment

These data structure is more simple to processing, analyzing and comparing simulations by multiple climate models, an wide range application for climate change studies is create a multimodel mean results from 35-models of CMIP5. But first we need to handle with different resolution problem (table II). To address this question we have used Resample native R-raster package function. Resample transfers values between non matching Raster* objects (in terms of origin and resolution). Aggregate a Raster* object to create a new RasterLayer or with a lower resolution. Aggregation groups are rectangular areas to create larger cells. The value for the resulting cells is computed with a user-specified function.

This method is capable of expressing a variety of dynamic spatial models and spatial data manipulations within a common framework spatial data that promote map algebra, ideal for advanced analysis.

Another data manipulation including mask and subseting. The First Create a new Raster* object that has the same values, except for the cells that are "no data"(NA) (or other maskvalue) in a 'mask'. These cells become NA (or other updatevalue). The mask can be either another Raster* object of the same extent and resolution, or a Spatial* object (e.g. SpatialPolygons) in which case all cells that are not covered by the Spatial object are set to updatevalue. You can use inverse=TRUE to set the cells that are not NA (or other maskvalue) in the mask, or not covered by the Spatial* object, to NA (or other updatvalue).Secondly the subset function that extract a set of layers from a RasterStack or RasterBrick object. Here we intentionally detail with a tutorial language by showing all the computational steps. Provenance depend on knowing all the logical steps.

3. Results

Data analysis under CMIP5 multi-model ensemble is recurrent used by the scientific community such as Statistical downscaling of CMIP5 multi-model ensemble for projected changes of climate in the Indus River Basin (Su et. al. 2016) and The projection of temperature and precipitation over China under RCP scenarios using a CMIP5 multi-model ensemble. (Chong-Hai & Ying, 2016).

In this work our case study were obtained from our current research about future patterns of fire-induced forest degradation in Amazonia towards a reproducing a set of results by sharing all the computational steps that have been applied. Our example analyzing the spatio-temporal variability through use of subsetting on future scenarios from CMIP5 data. In the purpose of performing numerical experiments (e.g. we denote these subsets as future scenarios by near future (2010-2039), middle future (2040-2069), distant future (2070-2099) for vapor pressure deficit in Amazonia mask) moreover these data are easy visualization com rasterVis package (Figure 2).

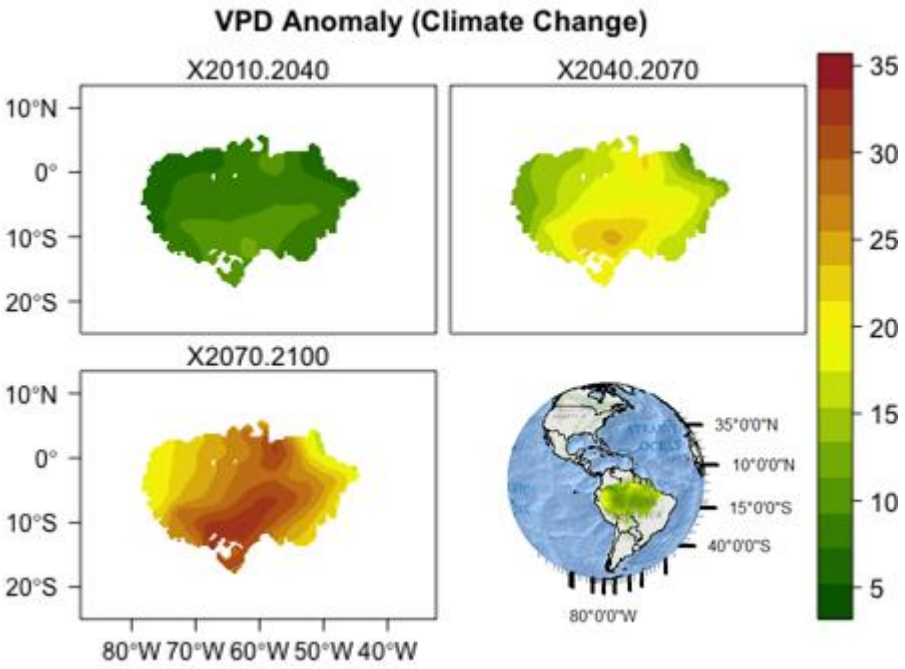

Figure 2- Example of application in a big volume of data. Projections of percent of vapor pressure deficit (VPD) increases (Increases of air dryness in Amazonia) for the CMIP5 multi-model ensemble. Future scenarios: near future (2010-2039), middle future (2040-2069), distant future (2070-2099)

This new approach allows scientists to visualize, analyze, serve multi-dimensional gridded data as well as easily fuse scientific data from different sources in a common coordinate system in desktop application or Web application. Moreover able to exporting to Geographic Information System (GIS) that is a system for handling geospatial information. ArcGIS, as a leading GIS software platform, has been widely and successfully applied to many research and application fields.

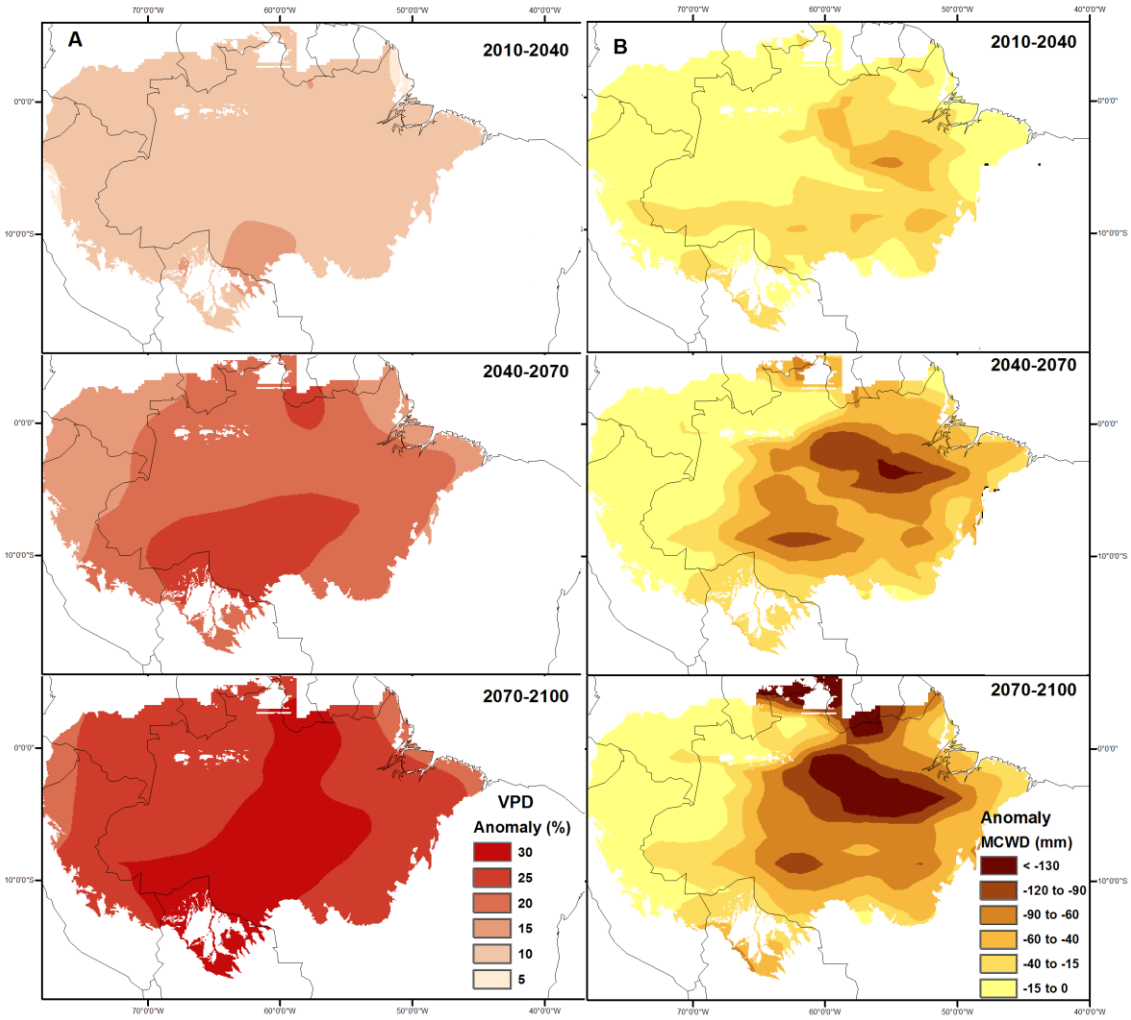

Figure 3- Projected dry-season changes in VPD (kPa) based on the multi-model ensemble average for RCP8.5 scenario (a). Variation of MCWD for future (b).

The following examples show the results of using multi-dimensional mosaic dataset in visualization, from CMIP5 multimodel ensemble aggregation, and analysis of multi-dimensional data using R Raster package for analysis and ArcGIS Software for visualization.

Figure 3 shows how multiple slices (times periods e.g: 2010-2040) are returned, the mosaic dataset model has the capability to aggregate the multiple slices into one based on any of the aggregation methods: Minimum, Maximum, Average, Sum (Figure 4). This flexible query capability allows answering typical questions such as "what is the temperature of particular layer? What is the average temperature of a year? What is the maximum temperature of a particular month, or what is the total rainfall of a year, and so on. Just need to use stackApply from raster package.

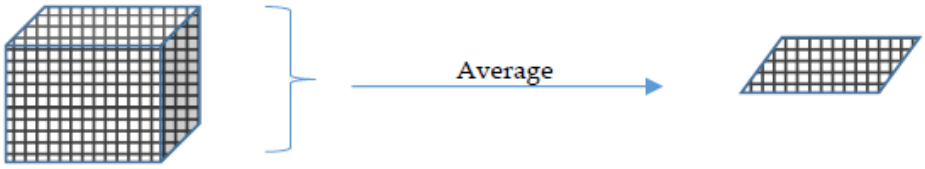

Figure 4- Example of a stackApply method

4. Discussions

The powerful tools provided by R language today are very useful to improve open science. Prior work has documented the crucial importance of rethinking about sharing research in the future of geosciences in order to reach best practices for documenting code and simulations. (Gil et. al., 2016). Climate change studies usually describe computational methods informally, often requiring a significant effort to understand and to reuse similar approaches. Attempts to replication of published work naturally reveal uncertainties, which enable further scientific progress. Jasny et al (2011). The publication of research papers is slowly changing to adapt to the digital age.

In this paper we have assessed the versatility of R language for manipulating digital climate data, and share your results. Our results empowers scientists to analyze and manipulation their research products.

The methods shown in this study can aid to manipulation and data analysis in a more simple and reproducible way. Increasingly, scientists are asked to share their data, software, and other results of their research. These requests, which used to come only from fellow scientists, are now coming from a variety of sources including funders, publishers, and journalists (Gil et. al., 2016). R is open-source and can promotes information exchange among scientists. As it has been explained in this work, NetCDF is a flexible source of information to help climatic change examination. It also has the intention to prove how easy it can be to use R when you have a huge set of data. The final objective is to increase the usage of NetCDF.

Most notably, this is the first study to our knowledge to promote best practices for CMIP5 ensemble. The NetCDF data accessible to R via raster package capabilities provides efficient access to the multi-model database used by the Intergovernmental Panel on Climate Change (IPCC) to write its five climate assessment report (IPCC, 2014). The CMIP5 mandated that models adhere to the NetCDF format. More than 100 peer-reviewed scientific publications have used the CMIP5 data sets as a result of this forethought, coordination, and open access (https://cmip-publications.llnl.gov). The widespread use of these internationally shared climate data demonstrates the potential for producers and users of other environmental modeling software to leverage their models and data. By understanding the data analysis practices and principles illustrated in this paper, environmental scientists can learn to create and manipulate gridded data sets.

5.Conclusion

We envision that in the near future (5-10 years), scientists will use radically new tools to author papers and disseminate information about the process and products of their research. These tools will document and publish the workflow as well as all the associated digital objects (data, software, etc.) that form the basis of a paper.

There are several major forces that push scientists to make their data and research open and accessible. A free software like R, that is gaining prominence in the scientific community can help. For example The American Geophysical Union (AGU) does include research code in its definition of "data" that must be shared for publication in its journals (AGU 2013; Hanson 2014) Data analysis with R Raster package, can easy clean and subsetting data for share enable further scientific progress.

We believe that the multi-dimensional data within netCDF file can efficiently with R packages established the link between science and processing data to effectively aid scientific research in many fields and facilities sharing the scientific results in a easier way to users and public. In the polarized context of climate research, making code available to public holds the potential to improve trust.

.

6. References

AGU 2013. AGU Publications Data Policy. American Geophysical Union, December 2013. Available from http://publications.agu.org/author-resource-center/publication-policies/datapolicy/

B. Eaton, J. Gregory, B. Drach, K. Taylor, and S. Hankin. NetCDF Climate and Forecast (CF) Metadata Conventions, Version 1.6, 2011. [p31]

Chong-Hai, X. U., and Xu Ying. "The projection of temperature and precipitation over China under RCP scenarios using a CMIP5 multi-model ensemble." Atmospheric and Oceanic Science Letters 5.6 (2012): 527-533.

D. Pierce. ncdf: Interface to Unidata netCDF data files, 2011. URL http://CRAN.R-project.org/package= ncdf. R Package Version 1.6.6.

Hanson "AGU 's Data Policy: History and Context." 2014. Brooks Hanson. Eos, 95(37), 337.16 September 2014.

Hijmans, R. J., van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J. A., ... & Hijmans, M. R. J. (2015). Package 'raster'. *R package*.

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again…. *Science*, *334*(6060), 1225-1225.

IPCC, 2014: Summary for Policymakers. In: *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea,T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L. White (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA P. Michna.

RNetCDF: R Interface to NetCDF Datasets, 2012. URL http://CRAN.R-project.org/package= RNetCDF. R Package Version 1.6.1-2. [p29]

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., Goodall, J. L., ... & Pierce, S. A. (2016). Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science*.

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*,*93*(4), 485.

Santer, B. D., Wigley, T. M. L., & Taylor, K. E. (2011). The reproducibility of observational estimates of surface and atmospheric temperature change. *Science*, *334*(6060), 1232-1233.

Su, B., Huang, J., Gemmer, M., Jian, D., Tao, H., Jiang, T., & Zhao, C. (2016). Statistical downscaling of CMIP5 multi-model ensemble for projected changes of climate in the Indus River Basin. Atmospheric Research, 178, 138-149.