Austrian Journal of Statistics
March 2016, Volume 45, 81–96.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i1.98

From Climate Simulations to Statistics – Introducing the wux Package

Thomas Mendlik
Wegener Center

Georg Heinrich
Wegener Center

Andreas Gobiet
ZAMG

Armin Leuprecht
Wegener Center

Abstract

We present the R package **wux**, a toolbox to analyze projected climate change signals by numerical climate model simulations and the associated uncertainties. The focus of this package is to automatically process big amounts of climate model data from multi-model ensembles in a user-friendly and flexible way. For this purpose, climate model output in common binary format (NetCDF) is read in and stored in a data frame, after first being aggregated to a desired temporal resolution and then being averaged over spatial domains of interest. The data processing can be performed for any number of meteorological parameters at one go, which allows multivariate statistical analysis of the climate model ensemble.

Keywords: climate research, climate uncertainty, multi-model ensembles, data processing, R.

1. Introduction

The human influence on the climate system is often assessed using numerical climate simulations (General Circulation models or GCMs). These are models representing physical processes in the atmosphere, ocean, cryosphere and land surface. However, due to their relative coarse resolution in the order of several hundred kilometers, they are not able to cover important processes at smaller spatial scales. For a more regional analysis, the models can be refined using regional climate models (RCMs) (Giorgi and Mearns 1991) and statistical downscaling techniques (Maraun, Wetterhall, Ireson, Chandler, Kendon, Widmann, Brienen, Rust, Sauter, Themeßl, Venema, Chun, Goodess, Jones, Onof, Vrac, and Thiele-Eich 2010). Under certain assumptions of future greenhouse gas (GHG) emissions, those models can project climate into future periods. We define a climate change signal of a particular meteorological parameter (from climate simulations) as the measure of change between a future climate projection and the past climate.

However, those projected climates are subject to different sources of uncertainty stemming from the natural variability of the climate system, unknown future GHG emissions, and errors and simplifications in GCMs and from regionalization methods. The resulting uncertainties can be partly assessed by analyzing so-called multi-model ensembles (i.e. climate projections which are generated by various GCMs and RCMs), which aim to sample the various sources of uncertainty. However, those ensembles do not systematically sample components of model uncertainty (e.g. physical parametrizations), and thus do not stem from an experimental design in a statistical sense (Knutti, Furrer, Tebaldi, Cermak, and Meehl 2010). They cannot be expected to represent unbiased distributions of possible future climate states. Also, interdependence between GCMs may induce additional biases in the sample, which makes a proper statistical analysis even more difficult. Several publications address those problems, for example Tebaldi and Knutti (2007); Smith, Tebaldi, Nychka, and Mearns (2009); Pirtle, Meyer, and Hamilton (2010); Bishop and Abramowitz (2012); Collins, Chandler, Cox, Huthnance, Rougier, and Stephenson (2012); Fischer, Weigel, Buser, Knutti, Künsch, Liniger, Schür, and Appenzeller (2012); Kang, Cressie, and Sain (2012); Stephenson, Collins, Rougier, and Chandler (2012); Chandler (2013); Rougier, Goldstein, and House (2013); Stephenson *et al.* (2012); Mendlik and Gobiet (2015).

This paper introduces the R package **wux** (Wegener Center Climate Uncertainty Explorer) (Mendlik, Heinrich, and Leuprecht 2015), a toolbox which enables multi-model handling for statistical analysis of climate scenarios. It is intended to be used to interpret climate model output and provide uncertainty information for the end-user of the climate simulations. Having in mind the heterogeneous target audience, we want this tool to perform following tasks:

1. Enable easy statistical *descriptive analysis* of user-defined climate model ensembles.

2. Be *expandable* to any kind of statistical analysis (to push the development of new statistical methods for climate multi-model analysis).

3. Easily *process climate simulations* to a common data format usable for statistical analysis. This enables reproducing data for any analysis needed.

Descriptive statistics of climatic changes from ensembles (point 1) are crucial to understand the underlying data. In practice people sometimes tend to forget this important step and prefer to directly address their complex research questions without having an overview of the data beforehand. A lot of valuable information lies in this analysis. Having some ready-to-use tools already implemented in **wux** should encourage users to perform this sort of analysis more often.

However, such a tool should not restrict the user to a pre-defined set of standard methods, on the contrary, development of new methods for statistical inference on climate simulations should be strongly supported, as this is still ongoing research (Knutti *et al.* 2010). Having set up this tool directly in R, allows to explore an extremely broad pool of ready-to-use methods, also from other disciplines using different approaches (point 2).

One of the most time consuming and frustrating tasks when analyzing climate simulations can be the step of processing data (point 3). The user of this tremendously big amount of datasets will find him-/herself challenged, when trying to aggregate them to the desired format (typically some sort of data frame) or get the desired statistics of the ensemble for certain geographical regions of interest. The challenge here is definitely a technical one: Processing ensembles of data in a binary-format usually requires dedicated programming work. The upside is that the data comes in the handy NetCDF file format[1], where a lot of meta-information about the data is stored in its header, however, life is more complicated in practice. Quite often it happens that meta-information between individual climate simulation output files differ substantially. For this reason it quickly becomes a nuisance when treating large samples of these files in an automated way. Up to now, no such tool is available which processes user-defined climate simulations in an automated way and which allows sophisticated statistical analysis. Furthermore, it is very difficult to reproduce statistical analysis from the scientific community when either the data set from the publication is not available, or the user wishes to apply the method with his/her own climate data. Providing a software which takes this burden, allows the user to solely focus on the interpretation of the climate model output

[1] http://www.unidata.ucar.edu/software/netcdf/

without spending too many resources on technicalities. We consider it a great strength of this package to perform this task in an automated way.

Several powerful tools already exist to process climate model outputs, such as CDO[2], NCO (Zender 2008), climate explorer[3] (van Oldenborgh, Drijfhout, van Ulden, Haarsma, Sterl, Severijns, Hazeleger, and Dijkstra 2009) and NCL[4]. All of those tools are designed to perform some sort of descriptive analysis and/or process the data to a desired format, however, none of those tools combines both easy multi-model handling and flexibility in statistical analysis. For example the climate explorer allows very straight forward processing of multi-model ensembles without any programming work. The user specifies what climate models to analyze simply by clicking on their names and the desired statistics. Such web-based tools however, being simple to use, lack of flexibility for a real programming interface. In addition it is not possible to extend those tools for own climate simulations which are not implemented. Also, statistical analysis is restricted to available methods. More programming-oriented tools like CDO and NCO also provide possibilities to analyze ensembles of climate simulations. However, the user has to specify the location of the data each time when calling a function and the data have to be pre-formatted for the program to understand its meaning. Changing local NetCDF files too much is a restriction to reproducible research. Even though programming is possible, we are restricted to pre-defined CDO statistics operators. The main difference of **wux** compared to those tools is the easy way it can read in a multitude of climate simulations and simply the fact that this tool is embedded in R, which allows to apply a very broad range of sophisticated statistical tools and is not restricted only by methods implemented in the toolbox itself.

The structure of this paper is as follows. Section 2 gives an overview of the functionalities of the **wux** package. Section 3 describes how climate data are being processed to a suitable data frame step-by-step. We introduce the statistical functionalities implemented in the package in Section 4 and provide an example application in Section 5 to show possible extensions of the implemented statistical functionalities. We conclude in Section 6.

2. Package overview

wux is meant to be an interfacing toolbox for scientists performing statistical analysis on climate models. Its focus is to provide a simple data frame for the user to make statistical inference on the ensemble. In particular, this package performs following actions, which are depicted in Figure 1 and described in Table 1:

Climate data processing. The function `models2wux` reads output of climate model simulations from NetCDF files, extracts subregions of interest, and writes climate change signals or time series to a data frame. Specific meta-information, like file locations, are stored in a `modelinput` input argument, which allows to simple processing of the simulations. For any new climate simulation it is enough to specify those meta-information without having to actually program a new input routine.

Statistical analysis of climate change signals. Based on the data frame returned by `models2wux`, we implemented various plotting options and summarizing utilities for a descriptive analysis of the projected climate change signals (e.g. scatterplots of temperature and precipitation). In addition, reconstruction tools allow to fill up missing climate simulations by multiple imputation methods. Based on such a reconstructed data frame (here termed as `rwux.df`), the user can assess for variance components via the implemented ANOVA tools or perform exploratory data analysis.

[2] CDO 2014: Climate Data Operators. Available at: https://code.zmaw.de/projects/cdo

[3] http://climexp.knmi.nl

[4] The NCAR Command Language (Version 6.2.1) [Software]. (2014). Boulder, Colorado: UCAR/NCAR/-CISL/VETS. http://dx.doi.org/10.5065/D6WD3XH5

I. Climate Data Processing (Section 3)

Function	Input	Output	Description
`models2wux`	NetCDF files	`wux.df`	Reads NetCDF climate model output, processes it, and writes the results to a data frame which is the backbone of all further **wux** analyses.
`read.wux.table`	`wux.df` files	`wux.df`	Reads data frame files produced by `models2wux`.

II. Statistical Analysis of Climate Change Signals (Section 4)

Function	Input	Output	Description
a) Descriptive analysis (Section 4.1)			
`summary`	`wux.df/rwux.df`	summary statistics	Summary statistics of the **wux** data frame (`wux.df` object).
`plot`	`wux.df/rwux.df`	figure	Scatter plot
`plotAnnualCycle`	`wux.df/rwux.df`	figure	Annual cycle plot
`hist`	`wux.df/rwux.df`	figure	Density plot
b) Reconstruction tools (Section 4.2)			
`reconstruct`	`wux.df`	`rwux.df`	Filling missing values of an unbalanced climate model design matrix in order to avoid biased ensemble estimates. Currently, the underlying reconstruction technique is based on an ANOVA using various methods for estimation. Returns reconstructed `wux data.frame` of class `rwux.df`.
c) Analysis of variance components (Section 4.2)			
`aovWux`	`rwux.df`	`wux.aov`	Extracts variance components of multiple climate model simulations using an ANOVA. Data must be balanced, so a reconstruction preprocessing is necessary.
`plot`	`wux.aov`	figure	Barchart for `aovWux` output.

Table 1: Most important functionalities of the **wux** package.

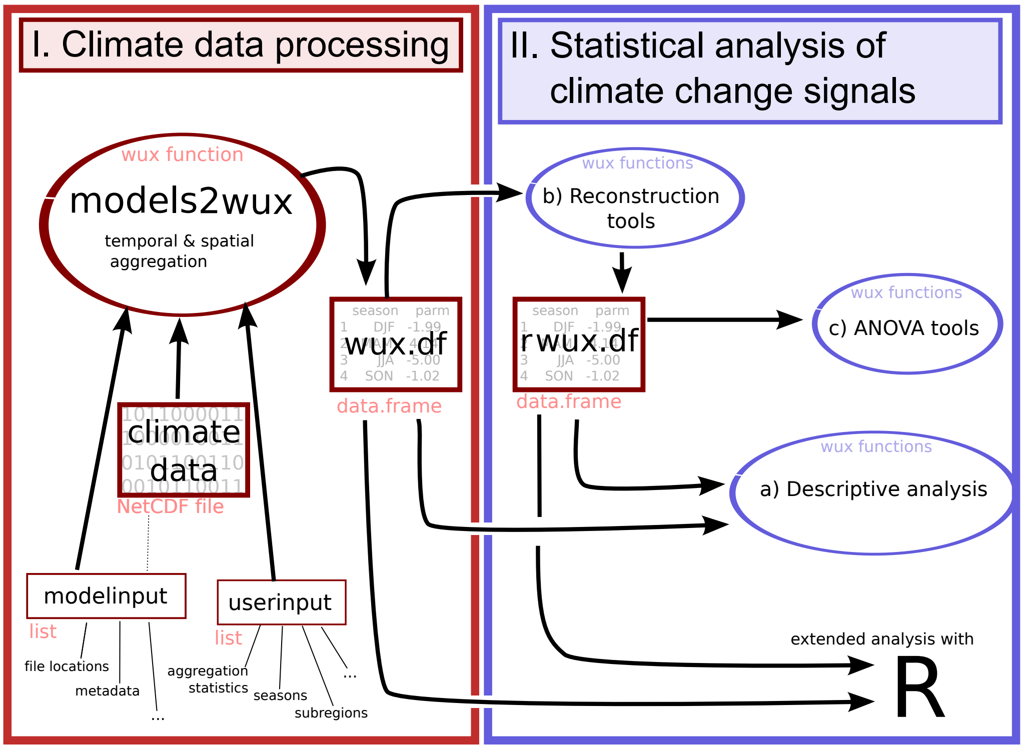

Figure 1: Basic functionalities of the **wux** package.

3. Climate data processing

The central role of the **wux** package is to automatically read in binary climate model output data from NetCDF files and process them to a data frame for statistical analysis. This task is performed by the function `models2wux`. The resulting data frame (further called `wux.df`, as it is technically a `wux.df` object) contains the climate change signals for user-specified periods, regions, seasons, and parameters for each of the climate models. One example `wux.df` is shown at the end of Section 3.1. Alternatively, also time series data can be obtained.

3.1. From climate model output to wux data frame

This is what `models2wux` is doing for each specified climate model:

1. Read in a three dimensional array (longitude, latitude, time) from binary climate model output.

2. *Temporal aggregation* of the fields according to user-specified climate periods and seasons. Aggregation statistics can also be specified by the user.

3. *Spatial aggregation* (arithmetic mean) over geographical domain.

4. Computing climate change signal for specified periods.

The resulting climate change signals for each climate model are returned to a data frame.

Temporal aggregation can be performed several times serially, going from fine temporal resolution to coarser resolution, each time using another statistic for aggregation. For example, daily temperature of a climate model output could first be aggregated to monthly resolution using the `mean` function and as a second step the warmest month in the year can be calculated with `max`. This would result in a climate change signal of the warmest monthly averages. We can thus calculate a vast amount of sufficient statistics to explore the climate data. Also, the user has the possibility to retrieve the full time-series of the climate model instead of the climate change signal. This can, however, result in quite a large data frame. The lowest time resolution currently implemented for time-series data is on monthly basis.

Being able to flexibly perform spatial aggregation over a specified domain is one of the key strengths of this program. Several ways exist for the user to identify the region of interest. For example a rectangular region defined by the longitude-latitude corners can be specified. For more flexibility, polygons can be defined using ESRI shapefiles[5] to cut out and aggregate over the desired subregion domain. The spatial aggregation is always performed using the arithmetic mean over geographical regions of any complexity. However, this process is not as trivial as it first may seem. One problem lies in the geographical projection of the climate model. Averaging over pixels of a model on a Mercator projection (angle preserving) will result in a different value than averaging over pixels in an area-preserving projection. GCMs usually do not come on an area-preserving projection. Therefore, the pixels should be weighted by the cosine of their latitudes, otherwise areas near the poles would gain much more weight then areas near the equator. When aggregating over a certain subregion, another problem arises from the gridpoints which are associated with the subregion. Instead of either considering a gridpoint to be within a region or not (0 and 1 weight), we may want to weight all the model cells that contribute even partly to the considered subregion, i.e. seize the fraction of the cell corresponding to the area covered by the subregion.

3.2. Setting up `models2wux`

To process a climate multi-model ensemble of your choice, `models2wux` needs two input arguments `userinput` and `modelinput`, each being a named list object or a file containing a named list.

`modelinput` stores general information about your climate data, i.e. the locations of the NetCDF files and their filenames. It also saves certain meta-information for the specific climate simulations (e.g. a unique acronym for the simulation, the developing institution, the radiative forcing). Usually the `modelinput` information should be stored in a single file on your system and should be updated when new climate simulations come in. It is advisable to share this file with your colleagues if you work with the same NetCDF files on a shared IT infrastructure.

The second input argument, `userinput`, defines which meteorological parameters of which climate simulations defined in `modelinput` should be analyzed. This is simply done by calling the models acronym, as all meta-information is already stored in the `modelinput` file. Also the geographical regions of interest and the temporal statistics are specified in this file. This file typically changes depending on the type of analysis performed.

3.3. Getting started

We explain `models2wux` in more detail by considering an example of a typical workflow for climate data processing. We start with downloading a couple of global climate simulations (GCMs) from the CMIP5 project (Taylor, Stouffer, and Meehl 2012), then we specify their meta-information and the output statistics and finally we run `models2wux` to process the binary data to an object of class `wux.df`.

To obtain CMIP5 climate simulations you can get started with downloading some example NetCDF files directly from an ESGF (Earth System Grid Federation) node[6] or using the `CMIP5fromESGF` function from the **wux** package (Linux only).

```
> ## I) Load wux functions and example datasets...
> library("wux")

> ## II) obtain some climate simulations
> CMIP5fromESGF(save.to = "~/tmp/CMIP5/",
```

[5] `http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf`

[6] e.g. from the data node `http://pcmdi9.llnl.gov`

```
                models = c("NorESM1-M", "CanESM2"),
                variables = c("tas", "pr"),
                experiments= c("historical", "rcp85"))
```

Here, we download the 2 m air temperature and surface precipitation files (`tas` and `pr`) from two simulations `NorESM1-M` and `CanESM2` for the `historical` period (here 1850–2005) and the future projection (2006–2100), assuming a strong change in future radiative forcing (`rcp85`, see Taylor *et al.* (2012)). The data will be downloaded into a temporary directory `~/tmp/CMIP5/` which can take a while. You need a valid account at any ESGF node for this function to run.

In order to run `models2wux`, you need to specify the two input arguments explained above: A `modelinput` file to define which climate simulations you have on your hard-disk and a `userinput` file which controls `models2wux` itself. An example for the model specification can be obtained in the package itself:

```
> ## III) Meta-information on downloaded data for models2wux.
> data(modelinput_test)
> str(modelinput_test)
List of 2
 $ CanESM2-r1i1p1_rcp85  :List of 11
 ..$ rcm              : chr ""
 ..$ gcm              : chr "CanESM2"
 ..$ gcm.run          : num 1
 ..$ institute        : chr "CCCma"
 ..$ emission.scenario: chr "rcp85"
 ..$ file.path.alt    :List of 2
 .. ..$ air_temperature :List of 2
 .. .. ..$ historical      : chr "~/tmp/CMIP5/CanESM2/historical"
 .. .. ..$ scenario        : chr "~/tmp/CMIP5/CanESM2/rcp85"
 .. ..$ precipitation_amount:List of 2
 .. .. ..$ historical      : chr "~/tmp/CMIP5/CanESM2/historical"
 .. .. ..$ scenario        : chr "~/tmp/CMIP5/CanESM2/rcp85"
 ..$ file.name          :List of 2
 .. ..$ air_temperature  :List of 2
 .. .. ..$ historical      : chr "tas_Amon_CanESM2_historical_r1i1p1_
     185001-200512.nc"
 .. .. ..$ scenario        : chr "tas_Amon_CanESM2_rcp85_r1i1p1_200601-210012.nc"
 .. ..$ precipitation_amount:List of 2
 .. .. ..$ historical      : chr "pr_Amon_CanESM2_historical_r1i1p1_
     185001-200512.nc"
 .. .. ..$ scenario        : chr "pr_Amon_CanESM2_rcp85_r1i1p1_200601-210012.nc"
 ..$ gridfile.path    : chr "~/tmp/CMIP5/CanESM2/historical"
 ..$ gridfile.filename: chr "tas_Amon_CanESM2_historical_r1i1p1_185001-200512.nc
     "
 ..$ resolution       : chr ""
 ..$ what.timesteps   : chr "monthly"
 $ NorESM1-M-r1i1p1_rcp85:List of 11
...
```

This input specifies the simulations which have just been downloaded. It is a named list with the name being an unique acronym of the climate simulation. The example input here specifies two simulations, but for the sake of brevity we only display the first one, being the `CanESM2-r1i1p1_rcp85` model. As this is a GCM, the `rcm` tag has no entry. The other tags specify the model in more detail: This simulation is run number 1 of the GCM `CanESM2` and has been developed by the `CCCma` institution[7]. The corresponding anthropogenic forcing is `rcp85`. `file.path.alt` defines the file locations for both temperature and precipitation files as well as for historical runs and future scenario projections. In this case the historical and the future scenario runs are located in different directories, whereas both meteorological parameters are saved in the same path. `file.name` gives information for the corresponding file names. The files which are necessary to define the geographical longitude and latitude information are specified in `gridfile.path` and `gridfile.filename`. The data is on a monthly timescale, which is defined in `what.timesteps` and the horizontal resolution is not specified here as it is optional.

It is advisable to store this list as a single file on your system. You should share this file with colleagues using the same IT infrastructure to use synergies. Such a file can also be

[7] Canadian Centre for Climate Modelling and Analysis (`www.ec.gc.ca/ccmac-cccma`)

created in an automated way using the function CMIP5toModelinput, for data obtained with CMIP5fromESGF (see the manual for more details).

Next, we want to tell models2wux to get climate change signals of both simulations we just defined above. In this example we are specifically interested in the temperature changes for the Alpine area at the end of the 21st century. Therefore we specify a user input file which contains a named list with all the necessary information:

```
> ## IV) Input argument controlling models2wux.
> data(userinput_CMIP5_changesignal)
> str(userinput_CMIP5_changesignal)
List of 9
 $ parameter.names     : chr "air_temperature"
 $ area.fraction       : logi TRUE
 $ reference.period    : chr "1971-2000"
 $ scenario.period     : chr "2071-2100"
 $ temporal.aggregation:List of 1
  ..$ stat.level.1:List of 3
  .. ..$ period     :List of 4
  .. .. ..$ DJF: chr [1:3] 12 1 2
  .. .. ..$ MAM: chr [1:3] 3 4 5
  .. .. ..$ JJA: chr [1:3] 6 7 8
  .. .. ..$ SON: chr [1:3] 9 10 11
  .. ..$ statistic  : chr "mean"
  .. ..$ time.series: logi FALSE
 $ subregions          :List of 1
  ..$ AL: num [1:4] 5 15 48 44
 $ plot.subregion      :List of 4
  ..$ save.subregions.plots: chr "/tmp/"
  ..$ xlim                 : num [1:2] 0 20
  ..$ ylim                 : num [1:2] 40 50
  ..$ cex                  : num 10
 $ save.as.data     : chr "/tmp/wuxexample"
 $ climate.models   : chr [1:2] "CanESM2-r1i1p1_rcp85", "NorESM1-M-r1i1p1_rcp85"
```

The userinput argument tells models2wux to process air_temperature (parameter.names) for both models CanESM2-r1i1p1_rcp85 and NorESM1-M-r1i1p1_rcp85 (climate.models tag). We define our base period (tag reference.period) to be 1971–2000 and the projected future period of interest (tag scenario.period) for the climatic change to be 2071–2100. We want the data to be aggregated to seasons summer (June, July, August: JJA), autumn (SON), winter (DJF) and spring (MAM). For each of those seasons models2wux returns the climate change signal defined by the user by calculating scenario.period minus reference.period (for precipitation, changes are in addition calculated relative to reference.period). When setting the attribute time.series to TRUE, the output would be a transient time series instead of climate change.

We want to aggregate over the spatial extend of the Alpine area (AL, see Christensen and Christensen (2007)), which is defined in the subregions tag. Here it is a named vector of longitude and latitude coordinates and it defines a rectangular region (western, eastern, northern and southern coordinates of the corners). There are plenty of other ways to define a subregion, like reading in shapefiles. To analyze which model grid cells lie within the specified region, we can specify plot.subregion (see Figure 2). We usually want to aggregate all model cells which lie within the specified region, however, sometimes we would like to down-weight those cells which only partly contribute to the considered region. Setting area.fraction as TRUE weights the cells corresponding to the area covered by the subregion (Figure 2). Furthermore, area.fraction=TRUE is necessary, if the size of the subregion is in the same order of magnitude as the grid cell. Such cases should be handled with care, since the grid point interpretation of climate models is problematic. In most cases, the analyzed subregions should be much larger than the grid size of the models and the error produced by setting area.fraction to FALSE is negligible and processing gains a massive speed up. The data frame will also be saved as a comma-separated file to /tmp/wuxexample.

Finally we run models2wux with the input arguments explained above to obtain the temperature climate change signals (delta.air_temperature) for both simulations aggregated over the Alpine region and four seasons. Columns besides subreg, season and the temperature

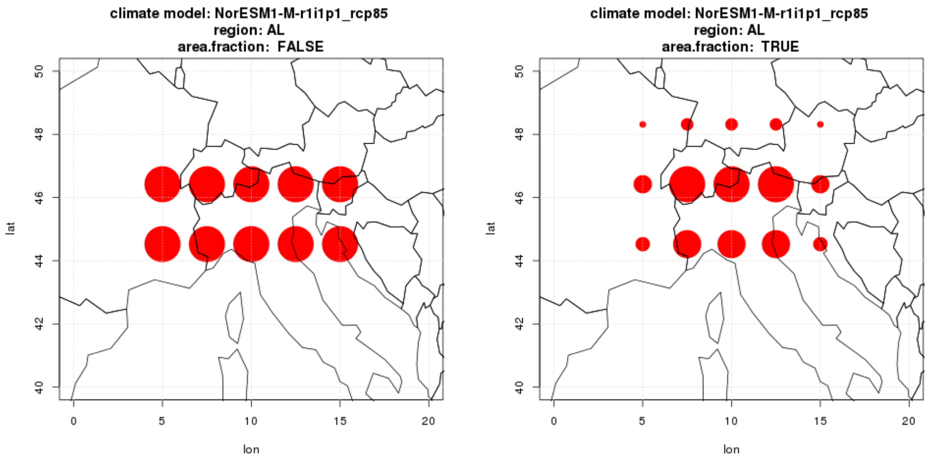

Figure 2: Grid cells of the NorESM1-M climate model being aggregated. On the left side `area.fraction` is switched off, taking all cells with their centroids lying within the `AL` region and weight them equally. The right figure has `area.fraction` on: The smaller the circles, the smaller the coverage of the model cells and the smaller their weight.

change parameter are meta-information of the climate data and derived from the `modelinput` input argument.

```
> ## V) Process NetCDF files
> climchange.df <- models2wux(userinput = userinput_CMIP5_changesignal,
>                             modelinput = modelinput_test)
> climchange.df
   subreg season                acronym institute       gcm gcm.run em.scn
1      AL    DJF   CanESM2-r1i1p1_rcp85     CCCma   CanESM2       1  rcp85
2      AL    JJA   CanESM2-r1i1p1_rcp85     CCCma   CanESM2       1  rcp85
3      AL    MAM   CanESM2-r1i1p1_rcp85     CCCma   CanESM2       1  rcp85
4      AL    SON   CanESM2-r1i1p1_rcp85     CCCma   CanESM2       1  rcp85
13     AL    DJF NorESM1-M-r1i1p1_rcp85       NCC NorESM1-M       1  rcp85
14     AL    JJA NorESM1-M-r1i1p1_rcp85       NCC NorESM1-M       1  rcp85
15     AL    MAM NorESM1-M-r1i1p1_rcp85       NCC NorESM1-M       1  rcp85
16     AL    SON NorESM1-M-r1i1p1_rcp85       NCC NorESM1-M       1  rcp85
      period ref.per resolution corrected delta.air_temperature
1  2071-2100      no         NA        no              4.066630
2  2071-2100      no         NA        no              8.041165
3  2071-2100      no         NA        no              4.261498
4  2071-2100      no         NA        no              5.686222
13 2071-2100      no         NA        no              3.336806
14 2071-2100      no         NA        no              5.378479
15 2071-2100      no         NA        no              3.922325
16 2071-2100      no         NA        no              3.787082
```

4. Statistical analysis of climate change signals

Several functions are available to analyze the processed climate change signals created by `models2wux`.

4.1. Descriptive analysis

The `summary` function gives a descriptive overview of the climate model ensemble which has been processed. On the one hand it calculates categorical statistics (counting climate models, emission scenarios, RCM-GCM cross-tables, . . .) and on the other hand it returns statistics of continuous climate change signals (mean, standard deviation, coefficient of variation and quantiles) split by season, emission scenario, meteorological parameters and subregions. Let us consider the climate change signals from 1961–1990 until 2021–2050 in the Greater Alpine Region (GAR) of a multi-model ensemble consisting of 22 RCMs from the ENSEMBLES project (van der Linden and Mitchell 2009).

```
> ## VI b) Analyze climate change data - summary statistics
> data(ensembles)
> # consider Greater Alpine Region (GAR) only
> wuxtest.df <- droplevels(subset(ensembles, subreg == "GAR"))
> ## summary statistics
> summary(wuxtest.df)
    ---------------------------------------------------------------------
    ----------------------- FREQUENCIES BY SCENARIO ---------------------
    ---------------------------------------------------------------------
A1B:
  8 GCMs (disregarding runs)
  22 models total
  Number of GCMs used:
       ARPEGE   BCCR-BCM2.0         CGCM3 ECHAM5/MPI-OM      HadCM3Q0
            3             3             1             5             5
    HadCM3Q16      HadCM3Q3      IPSL-CM4
            2             2             1
  Number of RCM runs:
   CLM   CRCM HIRHAM HadRM3 PROMES  RACMO    RCA   RCA3   REMO  RM4.5  RM5.1
     2      1      5      3      1      1      3      1      1      1      1
  RRCM  RegCM
     1      1
  Number of RCMs: 13

    ---------------------------------------------------------------------
    ---------------- CLIMATE MODEL STATISTICS BY SUBREGION --------------
    ---------------------------------------------------------------------

------------ GAR ------------
perc.delta.precipitation_amount:
 [A1B]
        n       mean    sd      coefvar min     max     med     q25     q75
   DJF: 22      2.88    5.09    1.77    -8.96   10.25   3.81    1.54    5.8
   JJA: 22      -2.82   6.87    2.44    -12.42  10.71   -3.7    -7.19   1.61
   MAM: 22      -0.64   4.99    7.83    -9.41   6.61    0.7     -5.52   2.87
   SON: 22      0.76    5.7     7.51    -12.16  12.46   0.77    -2.09   3.65

delta.air_temperature:
 [A1B]
        n       mean    sd      coefvar min     max     med     q25     q75
   DJF: 22      1.66    0.51    0.31    0.92    2.41    1.56    1.19    2.13
   JJA: 22      1.7     0.65    0.38    0.47    2.79    1.88    1.31    2.18
   MAM: 22      1.25    0.53    0.43    -0.02   2.26    1.21    0.91    1.55
   SON: 22      1.57    0.55    0.35    0.61    2.88    1.64    1.27    1.8
```

For the sake of brevity, we do not show all parts of the output. The `FREQUENCIES` output shows that $n = 22$ climate simulations driven by 8 GCMs forced with one emission scenario (A1B) have been processed and shows the count of the specific RCMs and GCMs used in the analysis. The `CLIMATE MODEL STATISTICS` output shows a descriptive analysis of the continuous variables in the data set based on all $n = 22$ climate simulations available. In this case the continuous variables are the relative change of precipitation (`perc.delta.precipitation_amount`) in percent and the absolute change of temperature (`delta.air_temperature`) in °C. The precipitation change in the GAR is not significant for either season, but there is a tendency in DJF for a slight increase of total precipitation. In contrast to that, the change signal for temperature is significant for all seasons showing quite an uniform warming, where MAM seems to have the smallest trend.

Also, functions for a graphical overview of the climate model ensemble are available in **wux**. The method `plot` for a `wux.df` object draws one or more scatterplots containing climate change signals of selected meteorological parameters.

```
> ## VI b) Analyze climate change data - scatterplots
> plot(ensembles, "perc.delta.precipitation_amount",
>      "delta.air_temperature", boxplots = TRUE,
>      xlim = c(-40,40), ylim = c(0, 4),
>      xlab = "Precipitation Amount [%]", ylab = "2-m Air Temperature [K]",
>      main = "Scatterplot", subreg.subset = c("GAR"))
```

This draws a simple scatterplot which accounts for certain meta-information of the climate change data frame and allows to highlight certain models. One of the scatterplots produced by this call is shown on the left side of Figure 3. This is a very useful plot as it gives a good overview on the model behavior and the climate change uncertainty. In our example, some models project an increase in precipitation change, whereas some project a decline. No correlation between temperature and precipitation change is visible on this small spatial scale.

4.2. Data reconstruction methods

Due to limited computational capacities, even in large-scale climate modeling projects such as CMIP5 or CORDEX (Jacob, Petersen, Eggert, Alias, Christensen, Bouwer, Braun, Colette, Déqué, Georgievski, Georgopoulou, Gobiet, Menut, Nikulin, Haensler, Hempelmann, Jones, Keuler, Kovats, Kröner, Kotlarski, Kriegsmann, Martin, Meijgaard, Moseley, Pfeifer, Preuschmann, Radermacher, Radtke, Rechid, Rounsevell, Samuelsson, Somot, Soussana, Teichmann, Valentini, Vautard, Weber, and Yiou 2013) only a limited number of climate simulations can be realized and it is a question of the experimental design which uncertainty components are primarily tackled within the ensemble. Therefore, missing realizations within climate projection ensembles are a common problem and even simple ensemble estimates such as mean and variability for e.g. temperature changes are potentially biased due to unequal sampling of the uncertainty components. In order to avoid such biases, Déqué, Rowell, Lüthi, Giorgi, Christensen, Rockel, Jacob, Kjellström, Castro, and Hurk (2007) introduced an iterative data reconstruction method which assumes additivity between uncertainty components in order to estimate the missing climate change signals. This reconstruction method was further applied in several studies in order to obtain a balanced design for the analysis of variance components (Déqué *et al.* 2007; Heinrich, Gobiet, and Mendlik 2014; Prein, Gobiet, and Truhetz 2011; Déqué, Somot, Sanchez-Gomez, Goodess, Jacob, Lenderink, and Christensen 2011; Mendlik and Gobiet 2015). In **wux**, we implemented the method of Déqué *et al.* (2007) for a two-factorial design (`reconstruct`) such as realized in the ENSEMBLES project (van der Linden and Mitchell 2009). In ENSEMBLES, a set of 21 high resolution RCM simulations with a horizontal grid spacing of about 25 km was produced. The ensemble consists of 8 GCMs and 16 RCMs only forced by the A1B emission scenario, but due to limited computational resources, only a small fraction (16.4 % of the possible GCM-RCM combinations) could be realized. The result of such a reconstruction is shown in Figure 3. In that case, filling up the missing GCM-RCM combinations does not alter the distribution of temperature and precipitation change. However, as the method relies on an implicit formulation of the uncertainty components, it cannot be used to extend the ensemble to GCMs that have not been used as driver for any RCM in the ensemble. Further reconstruction methods which are able to extend the ensemble to GCMs outside of the original design are investigated in Heinrich *et al.* (2014).

5. Example: Further statistical analysis

It is one of the key strengths of this package to be directly implemented in R and for that reason to have direct access to a huge magnitude of statistical methods to analyze climate data. We provide an example application in this section to show possible extensions based fully on the `wux.df`. We use a linear mixed effects model from the **lme4** package (Bates, Mächler, Bolker, and Walker 2015) to estimate the average summer temperature trend over the Greater Alpine Region based on individual time-series of 16 GCMs from the CMIP5 ensemble under a moderate stabilization scenario (RCP 4.5).

To generate the appropriate `wux.df`, the `timeseries` tag in the `userinput` file was set `TRUE` (see Section 3). The aim here is to get an average linear trend while accounting for the unbalanced model design. Several of the GCMs were run a couple of times (up to 10 times) with different initial conditions, which induces a dependency structure in the data set. We assess for this dependency by putting random effects in the linear model:

$$Y_{ijk} = \beta_0 + \beta_1 \text{year}_{jk} + b_{0i} + b_{1i} \text{year}_{jk} + \epsilon_{ijk}$$

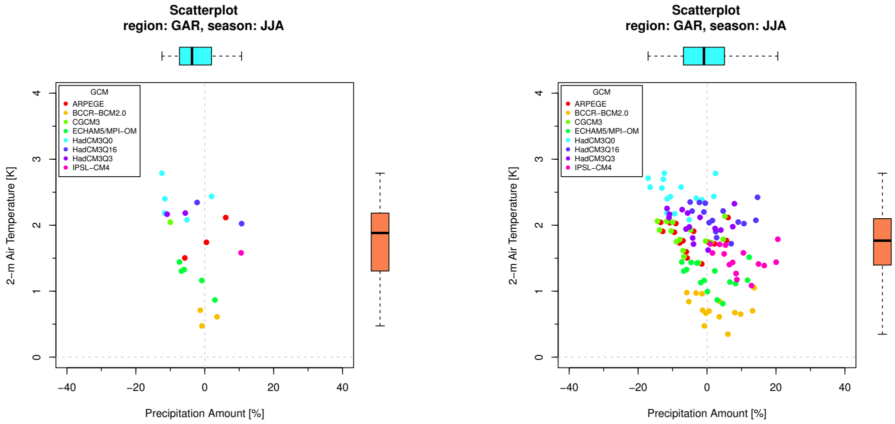

Figure 3: Projected changes of summer precipitation and temperature of the ENSEMBLES models from 1961–1990 to 2021–2050 in the Greater Alpine Region. The left plot shows the originally available 22 RCMs, whereas the right plot depicts a reconstructed dataset filled up with the function `reconstruct`.

where Y_{ijk} is the average summer temperature projected by $i = 1, \ldots, 16$ GCMs with $j = 1, \ldots, n_i$ runs per GCM and $k = 1, \ldots, 130$ yearly time steps. The random effects are defined as

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{gcm} & 0 \\ 0 & \sigma^2_{gcm.t} \end{pmatrix} \right) \quad \text{and} \quad \epsilon_{ijk} \overset{iid}{\sim} N\left(0, \sigma^2_y\right).$$

We use the `lmer` function from the **lme4** package for our analysis to estimate the fixed effects $\hat{\beta}_0, \hat{\beta}_1$ and to predict the individual random effects $\hat{\boldsymbol{b}}_0 = (\hat{b}_{0,1}, \ldots, \hat{b}_{0,16})'$, $\hat{\boldsymbol{b}}_1 = (\hat{b}_{1,1}, \ldots, \hat{b}_{1,16})'$. The time-series data and the trends are shown in Figure 4 plotted with the **lattice** package (Sarkar 2008).

```
> data(alpinesummer)
> ## pick just a few GCMs for this example - for a more compact display
> gcms.sub <- c("ACCESS1-3", "BCC-CSM1-1", "CESM1-CAM5", "CMCC-CM",
>               "CNRM-CM5", "CSIRO-Mk3-6-0", "EC-EARTH", "FGOALS-g2",
>               "GFDL-CM3", "HadGEM2-ES", "INM-CM4", "IPSL-CM5A-LR",
>               "MIROC5", "MPI-ESM-LR", "MRI-CGCM3", "NorESM1-M")
> alpinesummer.sub <- droplevels(subset(alpinesummer, gcm %in% gcms.sub))
> ## transform for better convergence
> alpinesummer.sub$time <- alpinesummer.sub$year - 1971
> lmm.fit <- lmer(air_temperature ~ 1 + time  + (1 |gcm) + (0 + time|gcm),
+                 data = alpinesummer.sub)
> summary(lmm.fit)
Linear mixed model fit by REML ['lmerMod']
Formula: air_temperature ~ 1 + time + (1 | gcm) + (0 + time | gcm)
   Data: alpinesummer.sub

REML criterion at convergence: 16472.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.0410 -0.6150 -0.0321  0.5766  4.5612

Random effects:
 Groups   Name        Variance  Std.Dev.
 gcm      (Intercept) 2.5671124 1.60222
 gcm.1    time        0.0001318 0.01148
 Residual             1.2482244 1.11724
Number of obs: 5330, groups:  gcm, 16

Fixed effects:
            Estimate Std. Error t value
(Intercept) 16.49168    0.40257   40.97
time         0.03443    0.00292   11.79
```

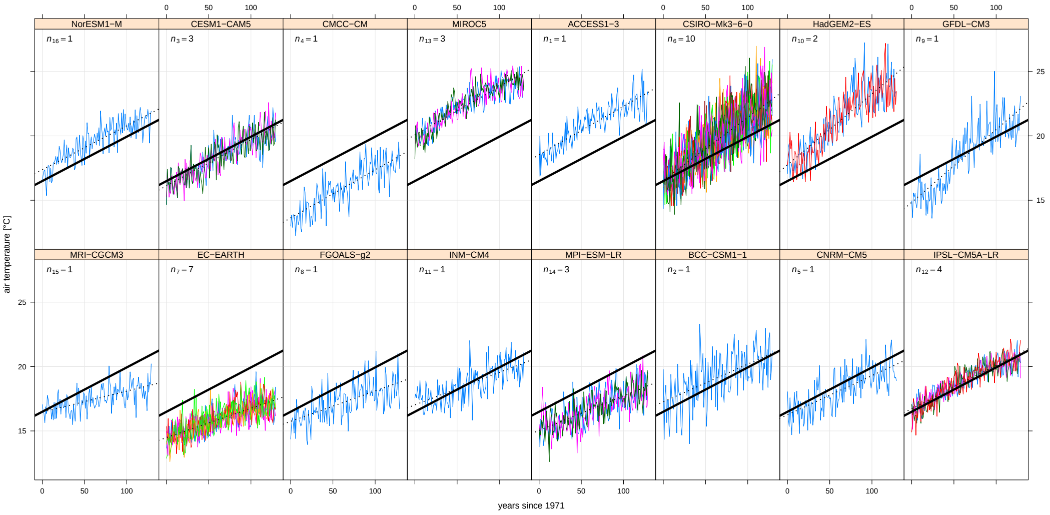

Figure 4: Time-series of GCMs from the CMIP5 ensemble for summer temperature in the Alpine region. The estimated average trend $\hat{\beta}_1$ is shown as a bold line, the predicted random effects trends are shown as a dashed line. The simulations are ordered from low trend (lower left panel) to high trend (upper right panel).

```
Correlation of Fixed Effects:
     (Intr)
time -0.016
> ## prints the first random effects
> head(coef(lmm.fit)$gcm)
              (Intercept)       year
ACCESS1-3        18.53855 0.03755274
BCC-CSM1-1       17.26063 0.02928145
CESM1-CAM5       16.11973 0.03574971
CMCC-CM          13.62953 0.03733811
CNRM-CM5         16.25376 0.03042184
CSIRO-Mk3-6-0    16.55908 0.04872848
```

The average slope $\hat{\beta}_1 = 0.34\,°\text{C/decade}$ ($0.034\,°\text{C/y}$) is highly significant and the individual slopes of the GCMs reach from slowly warming simulations $\hat{b}_{1,1} = 0.16\,°\text{C/decade}$ to very sensitive simulations $\hat{b}_{1,16} = 0.56\,°\text{C/decade}$ assuming linear temperature evolution over 130 years from 1971–2100. The residual standard deviation is $\hat{\sigma}_y = 1.12\,°\text{C}$, which in this case can be interpreted as the average year-to-year natural variability.

6. Conclusion

It is crucial in climate research not only to analyze outcomes of single climate models, but to consider entire multi-model ensembles, as it is virtually demanded in every climate impact related study to assess the associated uncertainties of the projected changes. There is, however, definitely a technical challenge to process large amount of climate simulations at once and not many tools exist to assess this problem. Another more general problem arises from the measure of uncertainty in multi-model ensembles. It is somewhat uncomfortable to make statistical inference on multi-model ensembles, as they do not stem from a designed experiment (Knutti *et al.* 2010), are utterly unbalanced (Déqué *et al.* 2007), and are known to be biased (Maraun *et al.* 2010; Themeßl, Gobiet, and Leuprecht 2011).

The focus here is not to show solutions for sophisticated statistical analyses of climate datasets, but merely to present a flexible and easy-to-use tool which is able to pre-process the datasets for further statistical analysis. This way, the user can focus on solving the grand challenges of statistical inference of multi-model datasets and does not need to spend valuable resources on technical data issues. The function `models2wux` fulfills exactly this task by processing magnitudes of binary climate model data to a R data frame of climate change signals. Subsequently, the user can take advantage of the vast amount of methods available in R, to analyze this data set.

However, this package also provides some functions for a first exploratory data analysis, as e.g. a `summary` function and some plotting routines. Such simple analysis provide very valuable information on the multi-model ensemble. In addition, we also provide a couple of methods to address the issue of unbalanced experiment designs. Several methods from literature are implemented to fill up the incomplete data matrix (Déqué *et al.* 2007; Heinrich *et al.* 2014).

It should be kept in mind, that also other software packages exist which partly fulfill similar tasks (e.g. climate explorer, CDO, NCL). The climate explorer can be a very convenient way to have a quick descriptive analysis of a multi-model ensemble. It is easy to use, but it is also restricted to a non-programming environment. Also, one can analyze only models which are implemented in the system, and the statistical methods are restricted as well. It should be noted, that no spatial analysis is currently possible within **wux**, as the emphasize lies on averaged domains. For spatial maps, tools as CDO or NCL are far better suited. Another limitation can be the hardware needed to process large datasets. R is not the most memory-efficient environment and one can run into trouble when reading climate simulations with a very high spatial resolution.

To sum it up, **wux** is a very flexible tool dealing with different aspects of climate model uncertainty in climate change impact investigations and enables a quick analysis of climate scenario uncertainty, which typically demands a considerable technical effort as well as fundamental knowledge about climate modeling. It can be used to achieve a quick overview on the involved uncertainties to identify the most important sources of uncertainty or to select representative sub-ensembles to be used as input for impact studies. **wux** is fully flexible regarding the meteorological parameter and region under consideration and is able to assess uncertainties based on multiple user-defined parameters.

Acknowledgments

We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. The ENSEMBLES data used in this work was funded by the EU FP6 Integrated Project ENSEMBLES (Contract number 505539) whose support is gratefully acknowledged.

References

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using **lme4**." *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.

Bishop CH, Abramowitz G (2012). "Climate model dependence and the replicate Earth paradigm." *Climate Dynamics*, pp. 1–16. doi:10.1007/s00382-012-1610-y.

Chandler RE (2013). "Exploiting strength, discounting weakness: combining information from multiple climate simulators." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **371**(1991). doi:10.1098/rsta.2012.0388.

Christensen JH, Christensen OB (2007). "A summary of the PRUDENCE model projections of changes in European climate by the end of this century." *Climatic Change*, **81**(1), 7–30. ISSN 0165-0009. doi:10.1007/s10584-006-9210-7.

Collins M, Chandler RE, Cox PM, Huthnance JM, Rougier J, Stephenson DB (2012). "Quantifying future climate change." *Nature Climate Change*, **2**(6), 403–409. doi:10.1038/nclimate1414.

Déqué M, Rowell DP, Lüthi D, Giorgi F, Christensen JH, Rockel B, Jacob D, Kjellström E, Castro M, Hurk B (2007). "An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections." *Climatic Change*, **81**(S1), 53–70. doi:10.1007/s10584-006-9228-x.

Déqué M, Somot S, Sanchez-Gomez E, Goodess CM, Jacob D, Lenderink G, Christensen OB (2011). "The spread amongst ENSEMBLES regional scenarios: regional climate models, driving general circulation models and interannual variability." *Climate Dynamics*. doi:10.1007/s00382-011-1053-x.

Fischer AM, Weigel AP, Buser CM, Knutti R, Künsch HR, Liniger MA, Schür C, Appenzeller C (2012). "Climate change projections for Switzerland based on a Bayesian multi-model approach." *International Journal of Climatology*, **32**(15), 2348–2371. ISSN 1097-0088. doi:10.1002/joc.3396.

Giorgi F, Mearns LO (1991). "Approaches to the simulation of regional climate change: A review." *Reviews of Geophysics*, **29**(2), 191–216. ISSN 1944-9208. doi:10.1029/90RG02636.

Heinrich G, Gobiet A, Mendlik T (2014). "Extended regional climate model projections for Europe until the mid-twentyfirst century: combining ENSEMBLES and CMIP3." *Climate Dynamics*, **42**(1-2), 521–535. doi:10.1007/s00382-013-1840-7.

Jacob D, Petersen J, Eggert B, Alias A, Christensen OB, Bouwer LM, Braun A, Colette A, Déqué M, Georgievski G, Georgopoulou E, Gobiet A, Menut L, Nikulin G, Haensler A, Hempelmann N, Jones C, Keuler K, Kovats S, Kröner N, Kotlarski S, Kriegsmann A, Martin E, Meijgaard E, Moseley C, Pfeifer S, Preuschmann S, Radermacher C, Radtke K, Rechid D, Rounsevell M, Samuelsson P, Somot S, Soussana JF, Teichmann C, Valentini R, Vautard R, Weber B, Yiou P (2013). "EURO-CORDEX: new high-resolution climate change projections for European impact research." *Regional Environmental Change*, pp. 1–16. ISSN 1436-3798. doi:10.1007/s10113-013-0499-2.

Kang EL, Cressie N, Sain SR (2012). "Combining outputs from the North American Regional Climate Change Assessment Program by using a Bayesian hierarchical model." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**(2), 291–313. ISSN 1467-9876. doi:10.1111/j.1467-9876.2011.01010.x.

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010). "Challenges in Combining Projections from Multiple Climate Models." *Journal of Climate*, **23**. doi:10.1175/2009JCLI3361.1.

Maraun D, Wetterhall F, Ireson AM, Chandler RE, Kendon EJ, Widmann M, Brienen S, Rust HW, Sauter T, Themeßl M, Venema VKC, Chun KP, Goodess CM, Jones RG, Onof C, Vrac M, Thiele-Eich I (2010). "Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user." *Reviews of Geophysics*, **48**(3), 1–34. doi:10.1029/2009RG000314.

Mendlik T, Gobiet A (2015). "Selecting climate simulations for impact studies based on multivariate patterns of climate change." *Climatic Change*. Submitted.

Mendlik T, Heinrich G, Leuprecht A (2015). ***wux**: Wegener Center Climate Uncertainty Explorer*. R package version 1.2-3, URL http://CRAN.R-project.org/package=wux.

Pirtle Z, Meyer R, Hamilton A (2010). "What does it mean when climate models agree? A case for assessing independence among general circulation models." *Environmental Science*, **13**(5), 351 – 361. ISSN 1462-9011. doi:10.1016/j.envsci.2010.04.004.

Prein AF, Gobiet A, Truhetz H (2011). "Analysis of uncertainty in large scale climate change projections over Europe." *Meteorologische Zeitschrift*, **20**(4), 383–395. doi:10.1127/0941-2948/2011/0286.

Rougier J, Goldstein M, House L (2013). "Second-Order Exchangeability Analysis for Multimodel Ensembles." *Journal of the American Statistical Association*, **108**(503), 852–863. doi:10.1080/01621459.2013.802963.

Sarkar D (2008). ***lattice**: Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5.

Smith RL, Tebaldi C, Nychka D, Mearns LO (2009). "Bayesian modeling of uncertainty in ensembles of climate models." *Journal of the American Statistical Association*, **104**(485), 97–116. doi:10.1198/jasa.2009.0007.

Stephenson DB, Collins M, Rougier JC, Chandler RE (2012). "Statistical problems in the probabilistic prediction of climate change." *Environmetrics*, **23**(5), 364–372. ISSN 1099-095X. doi:10.1002/env.2153.

Taylor KE, Stouffer RJ, Meehl GA (2012). "An Overview of CMIP5 and the Experiment Design." *Bulletin of the American Meteorological Society*, **93**(4), 485–498. ISSN 0003-0007. doi:10.1175/BAMS-D-11-00094.1.

Tebaldi C, Knutti R (2007). "The use of the multi-model ensemble in probabilistic climate projections." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **365**(1857), 2053–2075. doi:10.1098/rsta.2007.2076.

Themeßl M, Gobiet A, Leuprecht A (2011). "Empirical-statistical downscaling and error correction of daily precipitation from regional climate models." *International Journal of Climatology*, **31**(10), 1530–1544. ISSN 1097-0088. doi:10.1002/joc.2168.

van der Linden P, Mitchell JFB (2009). *ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project.* Met Office Hadley Centre.

van Oldenborgh GJ, Drijfhout S, van Ulden A, Haarsma R, Sterl A, Severijns C, Hazeleger W, Dijkstra H (2009). "Western Europe is warming much faster than expected." *Climate of the Past*, **5**(1), 1–12. doi:10.5194/cp-5-1-2009.

Zender CS (2008). "Analysis of Self-describing Gridded Geoscience Data with netCDF Operators (NCO)." *Environmental Modelling & Software*, **23**(10), 1338–1342. doi:10.1016/j.envsoft.2008.03.004.

Affiliation:

Thomas Mendlik
Wegener Center for Climate and Global Change
University of Graz
Brandhofgasse 5, 8010 Graz, Austria
E-mail: thomas.mendlik@uni-graz.at

Submitted: 2015-01-28
Accepted: 2015-09-21