SoftwareX 21 (2023) 101267

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

ag5Tools: An R package for downloading and extracting agrometeorological data from the AgERA5 database

David Brown [a,b,*], Kauê de Sousa [c,d], Jacob van Etten [c]

[a] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Droevendaalsesteeg 3, 6708 PB, Wageningen, The Netherlands
[b] Digital Inclusion, Bioversity International, 30501, Turrialba, Costa Rica
[c] Digital Inclusion, Bioversity International, Parc Scientifique Agropolis II, 34397, Montpellier Cedex 5, France
[d] Department of Agricultural Sciences, Faculty of Applied Ecology, Agricultural Sciences and Biotechnology, Inland Norway University of Applied Sciences, 2318 Hamar, Norway

ARTICLE INFO

Article history:
Received 24 August 2022
Received in revised form 6 November 2022
Accepted 15 November 2022

Keywords:
Agriculture
Climate
Crop variety evaluation
Field trials

ABSTRACT

Agrometeorological data is important in agricultural research, especially in agronomy and crop science, for investigating genotype by environment interactions. The AgERA5 dataset from the Copernicus Climate Data Store provides free and public access to global gridded daily agrometeorological data, from 1979 to present, with ready to use variables tailored for agricultural and agro-ecological studies. We developed the R package *ag5Tools*, which provides a simplified interface for downloading and extracting AgERA5 data. The package facilitates extracting time-series data for sets of geographic points in a format that can be conveniently used in statistical models applied in agricultural research. The use of the package is demonstrated with a synthetic dataset of multi-location trials in Arusha, Tanzania.

Code metadata

Current code version	*0.0.1*
Permanent link to code/repository used for this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-22-00257
Permanent link to reproducible capsule	
Legal code license	MIT
Code versioning system used	git
Software code languages, tools and services used	R
Compilation requirements, operating environments and dependencies	R
If available, link to developer documentation/manual	https://github.com/AgrDataSci/ag5Tools
Support email for questions	david.brownfuentes@wur.nl

1. Motivation and significance

The use of climatic data as model covariates in the analysis of multilocation trials enables extracting location-specific insights, such as targeted recommendations of crop varieties [1,2]. Several statistical and machine learning models allow incorporating climatic data as model covariates. The lack of accessibility to climatic data from local weather stations at the required temporal and spatial resolution has been an obstacle for its application [3]. Recently, several climatic datasets with global coverage have been made freely available to the public, enabling agricultural researchers to incorporate this kind of data in their analysis.

The AgERA5 dataset provides ready-to-use agrometeorological indicators from 1979 to present for agricultural and agro-ecological research studies [4]. It is derived from on the European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric re-analyses of the global climate (ERA5) data [5,6]. It is a gridded reanalysis data product, which has a global coverage, with a temporal coverage from 1979 to present at a daily temporal resolution, and a spatial resolution of 0.1° × 0.1° (approximately 11 km × 11 km at the equator).

The AgERA5 dataset provides 22 variables (Table 1) tailored for agronomic research [5]. It allows the users to get all the required

* Corresponding author at: Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Droevendaalsesteeg 3, 6708 PB, Wageningen, The Netherlands.
E-mail address: david.brownfuentes@wur.nl (David Brown).

https://doi.org/10.1016/j.softx.2022.101267

Table 1
Variables and statistics available for download from the AgERA5 dataset.
Source: Information retrieved from: https://doi.org/10.24381/cds.6c68c9bb.

Variable	Statistic	Time	Unit
10 m wind speed	24 h mean		m s^{-1}
2 m dewpoint temperature	24 h mean		K
2 m relative humidity		06:00	%
		09:00	
		12:00	
		15:00	
		18:00	
2 m temperature	24 h maximum		K
	24 h mean		
	24 h minimum		
	Day time maximum		
	Day time mean		
	Nighttime mean		
	Nighttime minimum		
Cloud cover	24 h mean		
Liquid precipitation duration fraction			
Precipitation flux			mm day^{-1}
Snow thickness	24 h mean		cm
Snow thickness LWE	24 h mean		cm
Solar radiation flux			J m^{-2} day^{-1}
Solid precipitation duration fraction			
Vapor pressure	24 h mean		hPa

variables from a single climate dataset with a homogeneous spatial resolution. When different climatic products from different sources are used, it is often the case that they are in different spatial resolutions and coordinate reference systems. Since the AgERA5 dataset provides a large number of variables tailored for agricultural research, the need for mixing datasets from different sources, and hence potential disagreement among them, is largely reduced.

The AgERA5 dataset is freely available online for downloading from the Copernicus Climate Data Store (CDS). The data can be downloaded using the CDS web interface but depending on the amount of data required, this interface might become unpractical. For example, there is a limit of 100 items, which means that only around 3 months of daily data can be downloaded in each request. Functionality for programmed downloading data from the CDS is provided by the CDS Application Programming Interface (API) (https://cds.climate.copernicus.eu/api-how-to). The CDS API is developed and supported by the ECMWF. Currently, the officially-supported API client is available only as a Python library (https://pypi.org/project/cdsapi/). The CDS API can also be used with the online CDS Toolbox. However, even when the official Python CDS API is used, the previously-mentioned restrictions for downloading the data still apply. R [7] users can access a wide range of ECMWF datasets, including AgERA5, through the package ecmwfr [8]. The wide range of accessibility to ECMWF products provided by the package ecmwfr is indeed convenient for users that require several datasets in their modelling workflows. To provide this cross-dataset compatibility, several parameters are available in the package ecmwfr. However, for users whose main interest lies in only one climatic dataset, this large number of parameters available in the package ecmwfr may be confusing. For instance, new users might feel overwhelmed by just finding if a parameter is indeed required for a download request of the AgERA5 dataset or not. Therefore, when the modelling workflow relies on mainly one climatic data product, such as the AgERA5 dataset, a product-specific tool might be more convenient. Furthermore, the data limit of 100 items also applies to download requests using the ecmwfr package.

The AgERA5 data is provided by the CDS as Network Common Data Form (NetCDF-4) files. This type of file can be easily read and handled in R by packages like terra [9], especially if the data will be used in raster format, either as a single or multilayer SpatRaster object. However, when data is required as a point-based time series for the locations of interest, the corresponding files should be searched by date and climatic variable, which can be a tedious task, especially when the required workflow includes several time-series of different meteorological variables and statistics.

2. Software description

We developed the R package *ag5Tools* to facilitate agricultural researchers to download and extract AgERA5 data. The package is aimed at supporting data analytics and synthesis workflows, such as the analysis and modelling of on-farm crop variety trials data, to assess the effect of climatic factors on a trait of interest (e.g., yield or disease resistance). In many of these workflows, the data is often required in a point-based format, such as R numeric vectors or *data.frame* objects.

2.1. Software architecture

The *ag5tools* package was developed following the R add-on packages guidelines and applying the S3 methods style [10]. Fig. 1 presents the architecture diagram of the *ag5Tools* package. The package file structure consists of seven main sub-directories: data, dev, docs, inst, man, R, and vignettes. The root directory contains the files DESCRIPTION, LICENSE, NAMESPACE and NEWS.

For the development of the package *ag5Tools*, we have used several open and free software, such as R and packages *devtools*, *fs*, *terra*, *reticulate*, and *sf* [7,9,11–14]. The downloading functionality of package *ag5Tools* uses the Python library cdsapi [15].

Since the package is published in *The Comprehensive R Archive Network* (CRAN), installing it can be made from R by executing `install.packages("ag5Tools")`. The source code is available in the GitHub repository https://github.com/AgrDataSci/ag5Tools and the development version can be installed from there by executing:

```
devtools::install_github("agrdatasci/ag5Tools",
  build_vignettes = TRUE).
```

Once installed, it can be loaded into a typical R session by executing `library(ag5Tools)`. The package automatically checks and configures the local environment to fulfill the requirements for downloading data from the CDS services. The only pre-requisite is that the user should be registered with the CDS and has

Table 2
Description of the methods, input parameters and outputs for the function *ag5_extract*.

Method	Parameters	Output
ag5_extract.numeric	***coords*:** numeric vector of length = 2 of the form (lon, lat), or a data.frame with required columns ***path*:** character indicating the path for the folder containing the AgERA5 files ***dates*:** character the dates for extracting the specified variable, a vector of length 1 extracts a single date, while a vector of length 2 indicates the start and end dates. ***variable*:** character indicating the AgERA5 variable to extract, see details for available options ***statistic*:** character, required only for some variables. ***time*:** only for variable Relative-Humidity-2 m ***celsius*:** logical, if TRUE converts the temperature values from the degrees Kelvin to degrees Celsius. Only for variables "Temperature-Air-2m" and "2m_dewpoint_temperature".	A numeric vector with length equal to the number of dates between the first and second date of the input parameters *dates*. The vector names correspond to the requested dates. If only one date is provided the function returns a numeric vector of length = 1.
ag5_extract.data.frame[a]	***coords*:** a *data.frame* with required columns ***start_date*:** character indicating the column name for the start of period of time to be extracted. ***end_date*:** character indicating the column name for the end of period of time to be extracted. ***lon:*** character indicating the name of the column containing the longitude values in the input *data.frame* ***lat:*** character indicating the name of the column containing the latitude values in the input *data.frame*	A list of named numeric vectors, each one corresponding to the rows in the input *data.frame*.

[a]Parameters path, variable, statistic, time and celsius are also required but omitted for brevity.

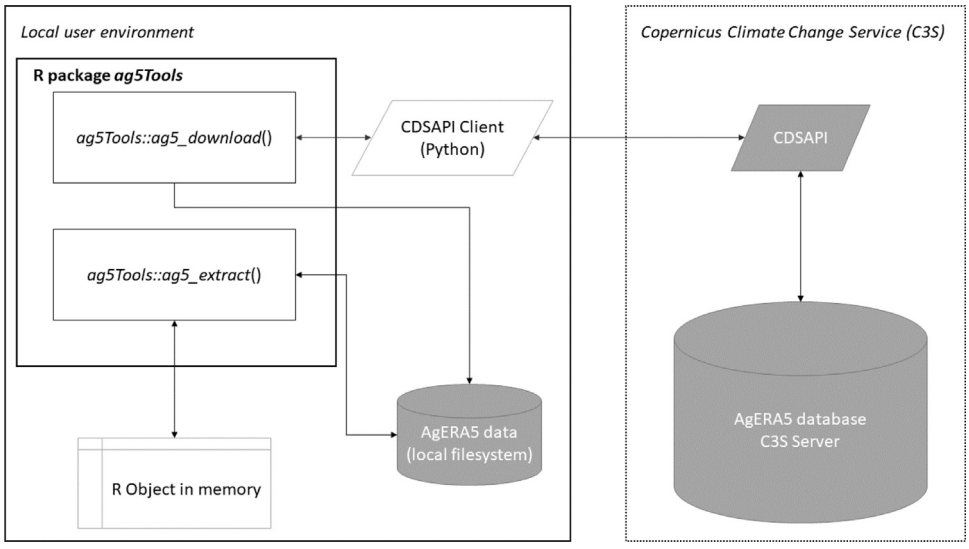

Fig. 1. Architecture diagram of the *ag5Tools* package.

retrieved his or her user key. After that, the user should store the key in a file in a local hard drive, which will be retrieved automatically by the *ag5Tools* package. Instructions to retrieve the CDSAPI key can be found on the official website of C3S https://cds.climate.copernicus.eu/api-how-to.

2.2. Software functionalities

Download data

The package *ag5Tools* provides functionality for downloading the full set of variables and statistics available from the AgERA5 dataset (Table 1). The users can make a download request through the function *ag5_download*, which is internally parsed by the R package *reticulate* to the Python library cdsapi. Those dependencies are internally managed by the *ag5Tools* package and do not require the intervention of the user. One advantage of the *ag5Tools* package is that it also sidesteps the current limitation of download request of the CDS platform, which does not allow requesting more than 100 elements. Therefore, *ag5Tools* users can request one or more years of data, without worrying about this limitation. Since *ag5_download* is specific for the AgERA5 dataset, it requires less input parameters from the user compared to other tools. For instance, parameters such as dataset name, dataset type and file format are handled internally by the function *ag5_download*, providing a simplified programming interface.

Extract data

Each NetCDF files (file extension .nc) downloaded from the Copernicus Climate Change Service contains AgERA5 data for a specific day. These files can be easily read with the R package *terra* [9]. When the data is required as numeric vector or as a *data.frame*, for multiple point locations and different time frames, extracting the data could be a complex task for non-expert users. The *ag5_extract* function provides a simple interface that facilitates the extraction of AgERA5 data, automatically searching each of the required files in the local hard drive. The *ag5_extract* is a generic function, which encapsulates different methods depending on the input parameters and the corresponding output (Table 2).

3. Illustrative examples

3.1. Download data

"Example code 1" below shows the code required for downloading the maximum daytime temperature data for years 2000 to 2005 using the function *ag5_download*. The request required six parameters, whereas twelve would be required using the ecmwfr package.

3.1.1. Example code 1

```
library(ag5Tools)

ag5_download(variable = "2m_temperature",
       statistic = "day_time_maximum",
       day = "all",
       month = "all",
       year = 2000:2005,
       path = "C:/custom_target_folder/")
```

The data are downloaded to the location indicated by the path argument in the function call. Within this path, a subfolder is created for each year contained in the download request. The data is downloaded as a temporary zip file named agera5_download.zip which is automatically uncompressed and deleted by the *ag5Tools* package after copying the files to the corresponding folder. The downloaded and extracted files after uncompressing the zip file are already named by the CDS using their nomenclature system. In the example above, we explicitly indicated that we wanted to download all days and months for each of the selected years, but specific days or months can also be requested. Depending on the variable, some arguments need to be specified while others do not. In the previous example, the variable 2m_temperature needs specification of the statistic day_time_maximum. In the case of relative humidity (2m_relative_humidity), a statistic should not be indicated, but indicating the time is mandatory. Example code 2 shows how to download relative humidity for times 6:00 and 18:00 for the same years as in the previous example.

3.1.2. Example code 2

```
ag5_download(variable = "2m_relative_humidity",
       time = c("06_00", "18_00"),
       day = "all",
       month = "all",
       year = 2000:2005,
       path = "C:/custom_target_folder/")
```

3.2. Extracting data

To demonstrate the functionality of *ag5_extract*, we use a synthetic dataset of 100 locations randomly generated across Arusha, Tanzania (Fig. 2).

For this example, we will focus only on variable maximum daytime temperature. Table 3 presents the first 10 data points of the example dataset.

If the data presented in Table 3 is stored in an *R data.frame* object, the following code can be used to extract the maximum daytime temperature data with the *ag5_extract* function.

Table 3
First 10 data points of the synthetic data example, presenting geographic coordinates, along with planting and harvest dates.

Longitude	Latitude	Planting date	Harvest date
35.726	−2.197	4/22/1991	8/20/1991
36.102	−2.851	1/24/1990	5/24/1990
35.463	−3.603	3/6/1991	7/4/1991
36.292	−3.856	10/10/1990	2/7/1991
35.453	−3.616	1/22/1990	5/22/1990
35.401	−3.216	10/19/1990	2/16/1991
35.170	−3.356	3/22/1990	7/20/1990
35.601	−2.502	10/14/1990	2/11/1991
36.537	−3.645	3/6/1990	7/4/1990
35.488	−2.981	4/27/1991	8/25/1991

Table 4
Climatic variables daily maximum daytime temperature (maxDT), minimum night temperature (minNT), precipitation (prec), solar radiation flux (srf), and relative humidity at time 09:00 a.m. (rhum_09), extracted for the trial data points of the example dataset and averaged for the corresponding planting to harvest period.

maxDT (°C)	minNT (°C)	prec (mm)	srf (J)	rhum_09 (%)
24.04	14.59	1.36	18778210.69	74.62
27.44	19.34	3.74	19964845.93	71.71
25.77	16.43	1.71	21830741.07	65.16
27.47	16.82	2.61	21344995.40	60.05
25.32	17.03	3.74	21727487.53	69.16
20.60	11.02	11.48	21106727.64	70.30
25.73	18.80	1.58	22175333.94	68.14
28.97	17.85	0.73	23368231.39	60.19
24.81	16.47	2.50	17724179.31	77.39
24.89	14.72	1.00	20505541.25	56.97

3.2.1. Example code 3

```
arusha_maxDT <- ag5_extract(coords = arusha_data,
                path = "D:/agera5_data/",
                variable = "Temperature-Air-2m",
                statistic = "Max-Day-Time",
                start_date = "planting_date",
                end_date = "harvest_date",
                celsius = TRUE)
```

AgERA5 temperature data is provided in degrees Kelvin. In our example, we set the parameter *celsius* = *TRUE*, to extract the data in degrees Celsius. We set the parameters *start_date* and *end_date*, as the column names in the *data.frame* have different column names, *planting_date* and *harvest_date* respectively. However, if the column names of the *data.frame* corresponding to dates are named as *start_date* and *end_date*, those parameters could be omitted. In Example code 3, the parameters *lon* and *lat* were omitted from the function call, because the column names match the default function parameters. If the column names in the input *data.frame* do not match the parameters, the column names corresponding to *lon* and *lat* should be provided as parameters in the function call. When the *coords* parameter is provided as a *data.frame*, the function returns a list of *data.frames*, each one containing a time series for each of the data points (the rows in the original *data.frame*), where column names are each of the dates from start_date to end_date. If the data is aimed at being used directly as model covariates, we need to compute the required aggregate metric (e.g., mean) for each time series. Following the same example trial dataset, Table 4 shows the data extracted for variables maximum daytime temperature (maxDT), minimum night temperature (minNT), precipitation (prec), solar radiation (srf), and relative humidity at time 09:00 a.m. (rhum_09). Since the variables are downloaded as daily observations, we computed the mean corresponding for time from planting to harvest of each trial data point of the synthetic dataset.

The data shown in Table 4 is ready to be used as model covariates in a statistical model. Also, the data extracted using *ag5_extract* can be used to calculate additional climatic variables

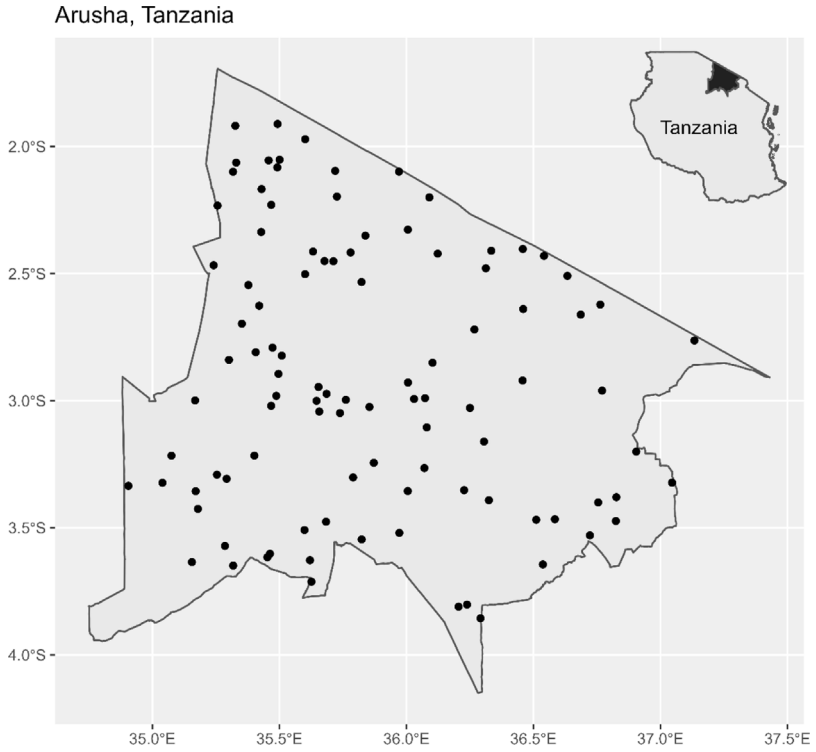

Fig. 2. Location of the data points randomly generated in Arusha, Tanzania.

or indices not directly available from Ag5ERA. The package *climatrends* [16] provides functionality for computing a range of climatic indices.

The *ag5_extract* function can also be used to extract data for one point location and one date or a time series for one point location. In the case of one point location, the argument *coords* should be provided as a vector of length = 2, in the form c(lon, lat). For example, using the coordinates of the first row in Table 3, the argument *coords* would be c(35.726, −2.197). This functionality might be useful in the case where climatic characterization of a single site is required. For instance, the example code 4 shows an example of daily precipitation data extracted for the first location of the synthetic dataset. If the data is extracted for only one date, the argument dates should be a vector of length = 1, and either a character or *Date* object. On the other hand, if a time series is required for just one location, the argument dates should be a vector of length = 2, where the first value indicates the start date and the second the end date of the series.

3.2.2. Example code 4

```
arusha_prec_01 <- ag5_extract(coords = c(35.726, -2.197162),
                    dates = c("1991-04-22", "1991-08-20"),
                    variable = "Precipitation-Flux",
                    path = "D:/agera5_data/")
```

4. Impact

The study of the effects of environmental factors on any genotype's performance is important in agronomy and crop science research. For instance, in breeding trials, the environment represents the main source of yield variability [17]. The use of climatic data as model covariates can support the generation of location-specific insights in crop variety evaluations [1,2,18]. The AgERA5 dataset provides an alternative data source when climatic data have not been collected in the field trials or when it is not available from local weather stations. This is even more relevant when a study involves several locations at regional or global scale, with disparities in terms of climate data availability. Given its wide time span (1979 to present) the AgERA5 provides an important source of climatic information for modelling purposes. Repurposing and reanalyzing legacy crop variety evaluation data, as described by Brown, Van den Bergh [19], is an example in which this wide time span is useful. For instance, the *ag5Tools* package was used by Brown, de Bruin [20] for downloading and extracting climatic data, which were used as covariates for modelling and predicting genotype performance. The *ag5Tools* package has been released in the Comprehensive R Archive Network (CRAN) and currently has more than 2500 downloads.

5. Conclusions

In this software paper, we have described the functionalities of the R package *ag5Tools*, for downloading and extracting AgERA5 data. As far as we know, this is the only R package with tools for downloading and extracting data exclusively designed for the AgERA5 dataset. The package is available freely for downloading at CRAN: https://cran.r-project.org/package=ag5Tools. We provided examples on how to download and extract AgERA5 data. Additional examples and information can be found on the package website https://agrdatasci.github.io/ag5Tools/. Since the development version is hosted in GitHub, current functionality problems or new feature requests can be managed by opening an issue in the GitHub repository.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The source code and example data are available in: https://github.com/AgrDataSci/ag5Tools

Acknowledgement

We acknowledge the constructive criticism and useful suggestions made by Dr. Sytze de Bruin. We also acknowledge Vincent Johnson (Science Writing Service of the Alliance of Bioversity International and CIAT) for English editing of this manuscript.

References

[1] van Etten J, et al. Crop variety management for climate adaptation supported by citizen science. Proc Natl Acad Sci 2019;116(10):4194–9.
[2] Buntaran H, Forkman J, Piepho H-P. Projecting results of zoned multi-environment trials to new locations using environmental covariates with random coefficient models: accuracy and precision. Theor Appl Genet 2021;134(5):1513–30.
[3] Ramirez-Villegas J, Challinor A. Assessing relevant climate data for agricultural applications. Agricult Forest Meteorol 2012;161:26–45.
[4] Boogaard H, van der Grijn G. Agrometeorological indicators from 1979 to present derived from reanalysis. In: Wageningen Environmental Research, editor. Copernicus climate change service. 2020.
[5] Boogaard H, van der Grijn G. Data stream 2: AgERA5 historic and near real time forcing data. In: ECMWF, editor. Product user guide and specification. Wageningen Environmental Research; 2020.
[6] Hersbach H, et al. The ERA5 global reanalysis. Q J R Meteorol Soc 2020;146(730):1999–2049.
[7] RCore Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.
[8] Hufkens K, Stauffer R, Campitelli E. The ecwmfr package: an interface to ECMWF API endpoints. 2019.
[9] Hijmans RJ. Terra: Spatial data analysis. R package version. 2021.
[10] RCore Team. Writing R extensions. 2022, Available from: https://cran.r-project.org/doc/manuals/r-release/R-exts.html.
[11] Wickham H, et al. Devtools: tools to make developing R packages easier. CRAN; 2021.
[12] Ushey K, Allaire J, Tang Y. Reticulate: Interface to 'Python'. 2022.
[13] Hester J, Wickham H, Csárdi G. Fs: cross-platform file system operations based on 'Libuv'. CRAN: CRAN; 2021.
[14] Pebesma E. Simple features for R: Standardized support for spatial vector data. R J 2018;10(1):439–46.
[15] ECMWF. Python API to access the copernicus climate data store. ECMWF; 2019.
[16] de Sousa K, van Etten J, Solberg SØ. Climatrends: climate variability indices for ecological modelling. 2020.
[17] Chenu K. Chapter 13 - characterizing the crop environment – nature, significance and applications. In: Sadras VO, Calderini DF, editors. Crop physiology. second ed.. San Diego: Academic Press; 2015, p. 321–48.
[18] de Sousa K, et al. Data-driven decentralized breeding increases prediction accuracy in a challenging crop production environment. Commun Biol 2021;4(1):944.
[19] Brown D, et al. Data synthesis for crop variety evaluation. a review. Agron Sustain Dev 2020;40(4):25.
[20] Brown D, et al. Rank-based data synthesis of common bean on-farm trials across four central American countries. Crop Sci 2022.