FluxnetLSM R package (v1.0): A community tool for processing FLUXNET data for use in land surface modelling

5   Anna M. Ukkola[1], Ned Haughton[1], Martin G. De Kauwe[2], Gab Abramowitz[1,3], Andy J. Pitman[1,3]

[1]ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney, NSW 2052, Australia

10   [2]Macquarie University, Sydney, NSW 2109, Australia

[3]Climate Change Research Centre, University of New South Wales, Sydney, NSW 2052, Australia

Correspondence to: A. M. Ukkola (a.ukkola@unsw.edu.au)

Abstract. Flux towers measure ecosystem-scale surface-atmosphere exchanges of energy, carbon dioxide and water vapour. The network of flux towers now encompasses ~900 sites, spread across every continent. Consequently, these data have become an essential benchmarking tool for land surface models (LSMs). However, these data as released are not immediately usable for driving, evaluating and

20   benchmarking LSMs. Flux tower data must first be transformed into a LSM-readable file format, a process which involves changing units, screening missing data and varying degrees of additional gap-filling. All of this often leads to an under-utilisation of these data in model benchmarking. To resolve some of these issues, and to help make flux tower measurements more widely used, we present a reproducible, open-source R package that transforms the latest FLUXNET2015 release into community

25   standard NetCDF files that are directly usable by LSMs.

1 Introduction

30   Land surface models (LSMs) provide the lower boundary condition for climate and weather forecast models, simulating the exchange of carbon, water and energy fluxes between the soil, vegetation and the atmosphere (Pitman, 2003). Flux towers measure ecosystem-scale exchanges of carbon dioxide, water vapour fluxes and energy (Baldocchi, 2014) and have proven invaluable for LSM evaluation and benchmarking (Abramowitz et al., 2008; Best et al., 2015; Blyth et al., 2010; Haughton et al., 2016;

35   Luo et al., 2012; Williams et al., 2009). Flux towers are particularly useful for modelling applications as they provide simultaneous observations of the meteorological data needed for forcing offline models as well as the key ecosystem variables against which models may be evaluated (e.g. sensible and latent heat) at time intervals similar to those used by LSMs, often over multiple years. As such, they are ideal

for characterising the interactions between climate and ecosystem processes and allow the evaluation of LSMs over time periods ranging from sub-daily through to seasonal and inter-annual time scales (e.g. Blyth et al., 2010; Bonan et al., 2011; Mahecha et al., 2010; Matheny et al., 2014; Powell et al., 2013; Ukkola et al., 2016; Wang et al., 2011; Whitley et al., 2016). The investment in flux tower

5 measurements is considerable and there are multiple benefits to these data being more widely used. First, the use of these data for LSM evaluation and benchmarking helps realise the value of existing investments. Second, where flux tower measurements identify biases in how LSMs represent processes, the potential exists to improve how well these models simulate the surface energy, water and carbon balances. Since LSMs are central to the simulation of key phenomenon including droughts, water

10 resource availability, carbon storage and feedbacks on heatwaves this has direct policy implications. Thirdly, greater use of flux tower measurements by the LSM and climate science community could help with the argument in support of on-going resourcing of flux tower measurements. In short, the effective and widespread use of flux tower measurements is beneficial across the science and policy communities.

Before data from flux tower sites can be used in models they commonly require significant pre-processing. In principle, flux towers provide near-continuous observations of ecosystem fluxes but, in practice, the measurements often include discontinuities due to instrument failure or unfavourable weather conditions (Reichstein et al., 2005). As LSMs must be provided with continuous

20 meteorological forcing data, flux tower datasets require varying degrees of gap-filling of missing time steps. This also poses challenges for using these data for model evaluation and benchmarking. Ideally, models should be evaluated against high-quality observations. Due to data gaps, as well as measurement biases (e.g. Leuning et al., 2012), flux tower measurements do not provide reliable observations representative of the true ecosystem dynamics in all circumstances. Arguably therefore,

25 the full breadth of flux tower data available across the entire network is unlikely to be suitable to the role of evaluating LSMs.

FLUXNET, an international network of flux tower sites, comprises of >900 sites globally (http://fluxnet.fluxdata.org/). The latest FLUXNET data release (FLUXNET2015;

30 http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/) provides flux tower measurements for 212 sites. The available data overcome some of the limitations of raw eddy covariance measurements through significant post-processing and gap-filling. Despite this, these datasets cannot be employed directly by LSMs. Critically, the FLUXNET data are not provided with temporally continuous observations of all essential meteorological variables (e.g. precipitation and wind speed) for forcing LSMs. For example,

35 across 155 Tier 1 sites reporting half-hourly observations, nearly all sites include gaps in rainfall and 77% of the sites have missing air temperature observations with up to 61% (median 5%) of the time series missing despite this variable being nominally gap-filled. Further, evaluation variables, such as latent and sensible heat, are generally gap-filled but to vastly different extents depending on the site and variable. For example, between 0% and 89% (median 31%) of the latent heat time series and 0%

40 and 83% (median 25%) of the sensible heat time series have been gap-filled across the 155 sites. This

poses a challenge for utilising these data for LSM applications and additional post-processing is necessary. A specific concern is that individual land surface modellers are very likely to post-process flux data in different ways, with different assumptions and varying levels of acceptance on how many gaps represent a worthwhile data set. When the gap-filled data are subsequently used and published, the

5   detail of how all the possibilities around post-processing the data are resolved is rarely documented. This leads to difficulties in interpreting model evaluation studies, a lack of reproducibility and, given many groups process data individually, wasted effort.

In an effort to resolve some of these problems and to connect the flux tower researchers with the LSM
10   researchers more strongly, we present the R package "`FluxnetLSM`" to facilitate the processing of FLUXNET datasets for use in LSMs. The package serves several important functions. Firstly, it enables the creation of fully gap-filled meteorological forcing datasets for running LSMs. Past studies have relied on various (often ad-hoc) gap-filling methods that are rarely fully documented in the literature. Worryingly, it would be virtually impossible to reproduce many existing LSM evaluation
15   and benchmarking studies although we note some exceptions (Best et al., 2015). The R package provides a community tool for creating LSM forcing datasets in a fully citeable and reproducible framework. Secondly, the package assists with the quality controlling of the data. It enables the selection of good-quality measurement periods and sites through automated screening of heavily gap-filled or missing data periods according to user-defined thresholds. To complement the automated
20   quality controlling, the package also provides tools for creating diagnostic plots to visualise output data periods using the Protocol for the Analysis of Land Surface models (PALS) R package. This facilitates detection of unusual data periods and encourages screening of flux tower sites for model applications. Finally, the package converts the flux tower data into the community standard NetCDF format used by the climate modelling and LSM community and collates metadata on data variables and flux tower sites
25   in the output files.

The package offers a useful tool for post-processing eddy covariance datasets for modelling applications and encourages better documentation of data processing methods to enhance the reproducibility of studies. In the following sections, we describe the different functionalities of the
30   package.

2 Package description

The `FluxnetLSM` package (v1.0) was developed to serve as a community tool to facilitate the use of
35   flux tower measurements in LSMs. It is written in the open-source R language (https://www.r-project.org/) and is freely accessible in a version-controlled repository (see Code Availability for full details). Instructions for installation are provided in the following section.

The package has two processing streams: the collection of site metadata and processing of high
40   frequency temporally varying variables. These are described in sections 2.3 and 2.4, respectively. The

package outputs a separate NetCDF file for meteorological and evaluation variables, with metadata stored in each file. Additionally, a log file is produced detailing output file names, potential warnings and errors. The package also provides the option to produce diagnostic plots for further data exploration. Figure 1 illustrates the general workflow with each component described in detail below.

2.1 Installation and requirements

`FluxnetLSM` requires an R version ≥3.1.0. It relies on base R functions as well as four additional packages: `pals`, `R.utils`, `ncdf4` and `rvest`. These packages should be installed prior to the

10   installation of `FluxnetLSM`. The `devtools` package is also recommended to aid installation.

The `R.utils`, `ncdf4`, `rvest` and `devtools` packages can be installed directly in R with the command `install.packages("package_name")`. The Protocol for the Analysis of Land Surface Models (`pals`) R package can be obtained from the Github repository at

15   https://github.com/dudek313/palsR and installed in R with the command:

```
devtools::install_github('dudek313/palsR')
```

Finally, the `FluxnetLSM` package can be downloaded from the Github repository at

20   https://github.com/aukkola/FluxnetLSM and installed within R by typing:

```
devtools::install_github("aukkola/FluxnetLSM")
```

Alternative installation methods are provided in the package github repository. After installation, the

25   `FluxnetLSM` package can be loaded into the R session by typing `library(FluxnetLSM)`. Other required packages are loaded automatically by the `FluxnetLSM` package.

2.2 Running FluxnetLSM

30   The package is run by invoking a single R function called `convert_fluxnet_to_netcdf`. This function accepts fifteen input arguments:

```
convert_fluxnet_to_netcdf(infile, site_code, out_path, ERA_file,
                          ERA_gapfill, datasetname, datasetversion,
                          missing, gapfill_all, gapfill_good,
                          gapfill_med, gapfill_poor,
                          include_all_eval, min_yrs, plot)
```

The user must set the arguments `infile`, `site_code`, `out_path` and `ERA_file`, with all other

40   arguments being optional. Each argument and its default value is described in Table 1 and discussed in

detail in the following sections. A full example for usage is provided in Section 3. Three example scripts are also provided with the package and are stored in `examples`. The `example_conversion_single_site.R` file shows an example for processing a single site. The `example_conversion_multiple_sites.R` and

5 `example_conversion_multiple_sites_parallel.R` files show an example for processing multiple sites using serial and parallel programming, respectively.

2.3 Collation of site metadata

10 The package collates metadata on the flux tower sites and stores these as attributes in the output NetCDF files. These include information required for modelling such as site coordinates, elevation and vegetation type. The primary source for metadata is a site attribute file provided with the package (stored in `data/Site_metadata.csv`). This file includes metadata detailed in Table 2 for the Tier 1 sites of the FLUXNET2015 November 2016 release (see http://fluxnet.fluxdata.org/data/fluxnet2015-

15 dataset/ for more information). The metadata were collated by the code developers from the site information provided on the FLUXNET website as well as individual flux tower network websites (see `data/README.md` for full details). The metadata file can be edited by the user to include additional sites or to modify existing data. The code first extracts site metadata from the CSV file. If any metadata cannot be found in the provided file, the code attempts to retrieve missing metadata from the

20 FLUXNET website (http://fluxnet.fluxdata.org), followed by the Oak Ridge National Laboratory (ORNL) FLUXNET website (https://fluxnet.ornl.gov/) by using functions for reading html webpages provided in the `rvest` library.

Additionally, the code stores the dataset name and version (as set by the `datasetname` and

25 `datasetversion` arguments to the main function), as well as the processing time and date as attributes in the output files. The code also calculates the mean annual precipitation for the output period when precipitation is outputted. It is stored as an attribute in the meteorological output file and can be useful particularly for rescaling precipitation for LSM spin-up so that each year's precipitation during the spin-up matches the site average.

30

This processing step connects key site metadata directly to each model forcing files. It can be extended to include additional metadata, such as site soil or vegetation properties, with minimal code modifications.

35 ### 2.4 Processing of high frequency data variables

2.4.1 Output variables

The package is supplied with a suggested list of output variables (stored in `data/Output_variables.csv`) that will be processed by the package for each site, where

40 available. The output variables are categorised as meteorological or evaluation variables, and a separate

NetCDF output file is produced for each category. Where possible, the output variables are named using the Assistance for Land-surface Modelling Activities (ALMA) convention (http://www.lmd.jussieu.fr/~polcher/ALMA/convention_output_3.html) commonly employed by LSMs. The package also performs common unit conversions between the original FLUXNET and ALMA convention units (see section 4.4). The output variables are fully customisable according to user requirements by removing or adding variables to the output variable list. The information required for each output variable is shown in Table 3.

2.4.1.1 Meteorological variables

The meteorological variables include the data variables typically required to force LSMs. The meteorological variables processed by the package by default are detailed in Supplementary Table 1. The user can also nominate essential meteorological variables that must be available and processed by modifying the `Essential_met` field in the output variable list (see Table 3). By default, these include air temperature, downward shortwave radiation, vapour pressure deficit, precipitation and wind speed. If any of these variables are not provided in the input data file, the code will terminate and the site will not be processed. The code gap-fills meteorological variables by default (see Section 2.2.3 for details).

2.4.1.2 Evaluation variables

The evaluation variables include the data variables typically predicted by land surface models and used to evaluate model outputs. The default evaluation variables processed by the package are provided in Supplementary Table 2. The evaluation variables are not gap-filled by the package. The user can nominate preferred evaluation variables by modifying the `Preferred_eval` field in the output variable list (see Table 3). By default these include net radiation, latent (LE) and sensible (H) heat and net ecosystem exchange (NEE). If none of the preferred variables are available in the input data file, the site will not be processed.

In addition to common evaluation variables, the package also processes and outputs uncertainty estimates provided with the FLUXNET2015 release by default. These include uncertainty bounds for LE, H and NEE, as well as error estimates for gross primary productivity (GPP). Several estimates for NEE and GPP are also included to reflect the inherent uncertainties in deriving these variables from eddy covariance data (Papale et al., 2006; Reichstein et al., 2005; Supplementary Table 2).

2.4.2 Gap-filled and missing values

The code produces NetCDF files with whole years of data only, to ensure LSM automated spin-up procedures remain relatively unbiased. It determines which years are included in its output according to user-defined thresholds for gap-filled and missing values as detailed below.

A threshold must be set for the maximum percentage of missing values per year (argument `missing`, 15% by default). The code checks for the percentage of missing values for each data variable during each year. If *any* essential meteorological variables or *all* preferred evaluation variables have missing

5 values in excess of this threshold, the year is not processed.

Additionally, thresholds can be set for the maximum percentage of all gap-filling (default option; set by argument `gapfill_all` using 20% as the default) or separately for "good", "medium" and "poor" quality gap-filling (arguments `gapfill_good`, `gapfill_med` and `gapfill_poor`, respectively;

10 see section 4.3). The percentage of gap-filled values is then checked for each data variable with a corresponding quality control flag during each year. If *any* essential meteorological variable or *all* preferred evaluation variables include gap-filled values in excess of the threshold(s), the year is not processed. Note the November 2016 FLUXNET2015 release has gaps in quality control flags for some evaluation variables even when data are present. In this case, data quality cannot be ascertained from

15 the flags (D. Papale, pers. comm.) and is treated as poor-quality gap-filling.

If a threshold for gap-filling is set, the percentage of both gap-filled and missing values must not exceed their respective thresholds for a year to be processed. If no years fulfilling the criteria are found, or the time period is shorter than the user-defined minimum number of consecutive years (set by

20 argument `min_yrs`, by default 2 years), the site it not processed. If several, non-consecutive, time periods fulfilling the criteria are found, these are written to separate output files.

Provided that at least one evaluation variable has fewer gaps than the user-defined thresholds, all evaluation variables are written to the output file by default, with the exception of any variables that

25 only contain missing values. An option is provided to discard any evaluation variables with gaps exceeding the user-defined thresholds by setting the argument `include_all_eval` to `FALSE`.

2.4.3 Gap-filling meteorological variables

30 A number of essential variables for forcing LSMs, such as rainfall and wind speed, are not gap-filled in the FLUXNET2015 release. The package uses the downscaled meteorological data from the ERA-Interim reanalysis (Dee et al., 2011; Vuichard and Papale, 2015) provided as part of the FLUXNET2015 dataset for gap-filling meteorological variables. The ERA-Interim estimates are available only in the "FULLSET" version of the FLUXNET2015 release

35 (http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/fullset-data-product/).

The ERA-Interim variable corresponding to each meteorological variable is set in the output variable list (`ERAinterim_variable` field; Table 2). If an ERA-Interim estimate is available for a given variable, the code gap-fills any missing time steps with the corresponding ERA-Interim data value. The

40 package saves information on the gap-filled time steps in quality control flag variables (see Section

2.2.4 for details). The gap-filling functionality can be disabled if no additional gap-filling is required by setting the argument `ERA_gapfill` to `FALSE`.

2.4.4 Quality control flags

The code retains and outputs the original FLUXNET2015 quality control (QC) flags, when these are included in the output variable list. These flags are set to 0 for measured data, and 1, 2 and 3 for good, medium and poor quality gap-filling, respectively (Reichstein et al., 2005; http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/).

Additionally, the code produces QC flags for meteorological variables when they are gap-filled using ERA-Interim data. The QC flag is set to 4 when a time step is gap-filled with ERA-Interim data. If a QC flag does not exist for a given variable, the code creates a QC flag variable with measured time steps set to 0 and ERA-Interim gap-filled time steps set to 4. This flag is automatically stored as a

15 variable in the meteorological data output file and is named as the output variable plus the extension "_qc" (e.g. Precip_qc).

2.4.5 Unit conversions

20 The package uses ALMA convention units for outputs by default where possible (as indicated in Supplementary Tables 1 and 2). These differ from the original FLUXNET units for a number of variables and a conversion is performed in each case. Available conversions are detailed in Table 4. If a conversion is not available for the specified units, the code will produce an error and terminate. Additionally, the package provides functions for converting i) vapour pressure deficit to relative

25 humidity and ii) relative humidity to specific humidity.

For these conversions, saturated vapour pressure (e_{sat}) is first calculated from air temperature (T_{air}; °C) (Jones, 1992) at each time step as

30
$$e_{sat} = 613.75 * \exp[17.502 * T_{air} / (240.97 + T_{air})] \tag{1}$$

Relative humidity is then determined from e_{sat} and vapour pressure deficit (D; Pa) as

$$R_h = 100 * (1 - (D * 100) / e_{sat}) \tag{2}$$

35

To calculate specific humidity (Q_{air}; kg kg^{-1}), specific humidity at saturation (w_s; kg kg^{-1}) is derived from e_{sat} and air pressure (ρ_{air}; Pa) as

$$w_s = 0.622 * e_{sat} / (\rho_{air} - e_{sat}) \tag{3}$$

Q_{air} is then calculated as

$$Q_{air} = (R_h / 100) * w_s \qquad (4)$$

2.4.6 Visualisation of outputs

The package provides an option to visualise outputs variables. Three types of plots can be produced: a mean annual cycle, a mean diurnal cycle by season and a time series figure. This is controlled by the
10   argument `plot` that can be set to any combination of `annual`, `diurnal` and `timeseries` for the three plot options, respectively. Examples of each plot are provided in Figure 2.

The outputs are retrieved from the output NetCDF files and all data variables are plotted with separate figures produced for meteorological and evaluation variables. Any missing values are ignored during
15   plotting, but their presence is noted in the figure, when applicable. The data are plotted in their output units, with the exception of air temperature (converted from Kelvin to Celsius) and rainfall (converted from mm/s to mm/time step). It is envisaged the plots will complement the automated quality control performed during data processing and enable further detection of unsuitable data periods or sites.

20   ### 3 Example application

Here we present an example application using "FluxnetLSM" for processing data at the Howard Springs (Australia) flux tower site. This example is provided in full with the package and stored in `examples/example_conversion_single_site.R`. It is also reproduced in Supplementary
25   section S.1 for convenience. Meteorological data is gap-filled using ERA-Interim estimates in this example but this functionality can be disabled if desired by setting `ERA_gapfill` argument to `FALSE` (see below). The user must provide four inputs, with the following inputs used in this example:

```
infile   <- "FLX_AU-How_FLUXNET2015_FULLSET_ HH_2001-2014_1-3.csv"
ERA_file <- "FLX_AU-How_FLUXNET2015_ERAI_HH_1989-2014_1-3.csv"
site_code <- "AU-How"
out_path  <- "~/FluxnetLSM/Outputs"
```

All other arguments shown below are optional and set to their default values in this example (see Table
35   1 for argument descriptions). The data can then be processed by invoking:

```
convert_fluxnet_to_netcdf(infile, site_code, out_path, ERA_file,
                          ERA_gapfill=TRUE,
                          datasetname="FLUXNET2015",
                          datasetversion="1-3", missing=15,
```

```
gapfill_all=20, gapfill_good=gapfill_NA,
gapfill_med=NA, gapfill_poor=NA,
include_all_eval=TRUE, min_yrs=2,
plot=c("annual", "diurnal","timeseries"))
```

5   The package automatically selects output years based on the thresholds (as detailed in Section 2.3.2). Figure 3 shows the full time series of essential meteorological variables and two example evaluation variables at Howard Springs. The code helps exclude time periods with extensive missing periods, such as the first year (2001) of the time series, as well as heavily gap-filled time periods (e.g. around January
10 2007). Extended periods with missing QC flags (see Section 2.2.3) are also excluded for evaluation variables due to unknown data quality (Figure 3b). Based on the default thresholds, the time period 2010-2014 is chosen and outputted, indicated by grey shading in Figure 3. The rest of the data are discarded. Thresholds can of course be modified by the user to change this result.

15 Once the data have been processed and outputted, they can be visualised. Three types of plots are produced by default: mean annual and diurnal cycles and a time series plot. Figure 2 shows an example of each type of output plot produced by the package. These plots can be used for further quality controlling to detect any anomalous data periods not automatically excluded by the package.

20 **4 Discussion and Conclusions**

Efforts to better utilise existing observational data has multiple benefits including bringing research communities together, evaluating models against broader data, and providing further support to groups seeking to maintain primary observations. To maximise the use of observed data, it is helpful that these
25 data are easily used by communities other than those that collect the data. In the case of the FLUXNET data, one major community is the land surface modelling sciences that are linked to the climate modelling field.

To enhance transparency, to help reproducibility and as a platform for further community efforts we
30 have presented an R package that transforms FLUXNET data into a form directly useable by LSMs. As released, FLUXNET data cannot be directly employed in LSMs due to data gaps, incompatible units and non-standard (land surface community) file format (CSV rather than NetCDF). The R package also collates metadata on data processing steps and the flux tower sites and stores these in the output files for easy access, and to permit more reliable reproducibility for modelling experiments. Finally, the
35 package generates visualisations of outputs to facilitate further quality control of flux tower data and to help inform appropriate site selection, an important step in applying these data to modelling studies.

The package is open source, fully documented and simple to use, requiring minimal input from the user. It allows multiple sites to be processed into a form usable by LSMs in a short R script.
40 Simultaneously, it provides optional settings for an advanced user to produce flux tower datasets suited

for specific applications. Importantly, the package provides a tool for producing flux tower datasets for modelling applications in a fully citeable and reproducible framework. The package is stored in a publicly available repository and is being actively developed with community contributions encouraged.

Code availability

The `FluxnetLSM` code can be downloaded from the Github repository at https://github.com/aukkola/FluxnetLSM. The code relies on the Protocol for the Analysis of Land

10 Surface Models (PALS) R package. This package can be downloaded from the Github repository at https://github.com/dudek313/palsR. Other required packages (`R.utils`, `ncdf4` and `rvest`) can be installed directly in R with the command `install.packages("package_name")`. See section 2.1 for further details on installation.

15 **Acknowledgements**

We acknowledge the support of the Australian Research Council Centre of Excellence for Climate System Science (CE110001028). M.G. De Kauwe was supported by Australian Research Council Linkage grant LP140100232. This work used eddy covariance data acquired and shared by the

20 FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management

25 Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices.

References

30 Abramowitz, G., Leuning, R., Clark, M. and Pitman, A.: Evaluating the performance of land surface Models, J. Clim., 21, 5468–5481, doi:10.1175/2008JCLI2378.1, 2008.

Baldocchi, D.: Measuring fluxes of trace gases and energy between ecosystems and the atmosphere - the state and future of the eddy covariance method, Glob. Chang. Biol., 20, 3600–3609,

35 doi:10.1111/gcb.12649, 2014.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. ., Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. a., Stevens, L. and Vuichard, N.: The plumbing of land

40 surface models: benchmarking model performance, J. Hydrometeorol., 16, 1425–1442,

doi:10.1175/JHM-D-14-0158.1, 2015.

Blyth, E., Gash, J., Lloyd, A., Pryor, M., Weedon, G. P. and Shuttleworth, J.: Evaluating the JULES Land Surface Model Energy Fluxes Using FLUXNET Data, J. Hydrometeorol., 11, 509–519,
5   doi:10.1175/2009JHM1183.1, 2010.

Bonan, G. B., Lawrence, P. J., Oleson, K. W., Levis, S., Jung, M., Reichstein, M., Lawrence, D. M. and Swenson, S. C.: Improving canopy processes in the Community Land Model version 4 (CLM4) using global flux fields empirically inferred from FLUXNET data, J. Geophys. Res., 116, 1–22,
10   doi:10.1029/2010JG001593, 2011.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach,
15   H., Holm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J. N. and Vitart, F.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137, 553–597, doi:10.1002/qj.828, 2011.

20   Haughton, N., Abramowitz, G., Pitman, A. J., Or, D., Best, M. J., Johnson, H. R., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Santanello, J. A., Stevens, L. E. and Vuichard, N.: The plumbing of land surface models: is poor performance a result of methodology or data quality?, J. Hydrometeorol., JHM-D-15-0171.1, doi:10.1175/JHM-D-15-0171.1, 2016.
25

Jones, H. G.: Plants and microclimate: A quantitative approach to environmental plant physiology, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., 1992.

Leuning, R., van Gorsel, E., Massman, W. J. and Isaac, P. R.: Reflections on the surface energy
30   imbalance problem, Agric. For. Meteorol., 156, 65–74, doi:10.1016/j.agrformet.2011.12.002, 2012.

Luo, Y. Q., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P., Dalmonech, D., Fisher, J. B., Fisher, R., Friedlingstein, P., Hibbard, K., Hoffman, F., Huntzinger, D., Jones, C. D., Koven, C., Lawrence, D., Li, D. J., Mahecha, M., Niu, S. L., Norby, R., Piao, S. L., Qi,
35   X., Peylin, P., Prentice, I. C., Riley, W., Reichstein, M., Schwalm, C., Wang, Y. P., Xia, J. Y., Zaehle, S. and Zhou, X. H.: A framework for benchmarking land models, Biogeosciences, 9, 3857–3874, doi:10.5194/bg-9-3857-2012, 2012.

Mahecha, M. D., Reichstein, M., Jung, M., Seneviratne, S. I., Zaehle, S., Beer, C., Braakhekke, M. C.,
40   Carvalhais, N., Lange, H., Le Maire, G. and Moors, E.: Comparing observations and process-based

simulations of biosphere-atmosphere exchanges on multiple timescales, J. Geophys. Res.
Biogeosciences, 115, 1–21, doi:10.1029/2009JG001016, 2010.

Matheny, A. M., Bohrer, G., Stoy, P. C., Baker, I. T., Black, A. T., Desai, A. R., Dietze, M. C., Gough,
C. M., Ivanov, V. Y., Jassal, R. S., Novick, K. a., Schäfer, K. V. R. and Verbeeck, H.: Characterizing
the diurnal patterns of errors in the prediction of evapotranspiration by several land-surface models: An
NACP analysis, J. Geophys. Res. Biogeosciences, 1458–1473, doi:10.1002/2014JG002623, 2014.

Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B., Rambal,
S., Valentini, R., Vesala, T. and Yakir, D.: Towards a standardized processing of Net Ecosystem
Exchange measured with eddy covariance technique: algorithms and uncertainty estimation,
Biogeosciences, 3, 571–583, doi:10.5194/bg-3-571-2006, 2006.

Pitman, A. J.: The evolution of, and revolution in, land surface schemes designed for climate models,
Int. J. Climatol., 23, 479–510, doi:10.1002/joc.893, 2003.

Powell, T. L., Galbraith, D. R., Christoffersen, B. O., Harper, A., Imbuzeiro, H. M. a, Rowland, L.,
Almeida, S., Brando, P. M., da Costa, A. C. L., Costa, M. H., Levine, N. M., Malhi, Y., Saleska, S. R.,
Sotta, E., Williams, M., Meir, P. and Moorcroft, P. R.: Confronting model predictions of carbon fluxes
with measurements of Amazon forests subjected to experimental drought, New Phytol., 200, 350–365,
doi:10.1111/nph.12390, 2013.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C.,
Buchmann, N., Gilmanov, T., Granier, A., Grünwald, T., Havránková, K., Ilvesniemi, H., Janous, D.,
Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J.
M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M., Tenhunen, J., Seufert, G., Vaccari, F., Vesala,
T., Yakir, D. and Valentini, R.: On the separation of net ecosystem exchange into assimilation and
ecosystem respiration: Review and improved algorithm, Glob. Chang. Biol., 11, 1424–1439,
doi:10.1111/j.1365-2486.2005.001002.x, 2005.

Ukkola, A. M., De Kauwe, M. G., Pitman, A. J., Best, M. J., Abramowitz, G., Haverd, V., Decker, M.
and Haughton, N.: Land surface models systematically overestimate the intensity, duration and
magnitude of seasonal-scale evaporative droughts, Environ. Res. Lett., 11, 104012, doi:10.1088/1748-
9326/11/10/104012, 2016.

Vuichard, N. and Papale, D.: Filling the gaps in meteorological continuous data measured at
FLUXNET sites with ERA-Interim reanalysis, Earth Syst. Sci. Data, 7, 157–171, doi:10.5194/essd-7-
157-2015, 2015.

Wang, Y. P., Kowalczyk, E., Leuning, R., Abramowitz, G., Raupach, M. R., Pak, B., van Gorsel, E.

and Luhar, A.: Diagnosing errors in a land surface model (CABLE) in the time and frequency domains, J. Geophys. Res., 116, G01034, doi:10.1029/2010JG001385, 2011.

5 Whitley, R., Beringer, J., Hutley, L., Abramowitz, G., De Kauwe, M. G., Duursma, R., Evans, B., Haverd, V., Li, L., Ryu, Y., Smith, B., Wang, Y.-P., Williams, M. and Yu, Q.: A model inter-comparison study to examine limiting factors in modelling Australian tropical savannas, Biogeosciences, 13, 3245–3265, doi:10.5194/bgd-12-18999-2015, 2016.

10 Williams, M., Richardson, A. D., Reichstein, M., Stoy, P. C., Peylin, P., Verbeeck, H., Carvalhais, N., Jung, M., Hollinger, D. Y., Kattge, J., Leuning, R., Luo, Y., Tomelleri, E., Trudinger, C. M. and Wang, Y.-P.: Improving land surface models with FLUXNET data, Biogeosciences, 6, 1341–1359, doi:10.5194/bg-6-1341-2009, 2009.

25

30

35

40

Tables

Table 1: Input arguments to the main `convert_fluxnet_to_netcdf` function.

Argument	Description	Default value
infile	FLUXNET2015 file containing data variables	-
site_code	FLUXNET site ID (see Table 2)	-
out_path	User-defined output path	-
ERA_file	FLUXNET2015 file containing ERA-interim variables	-
ERA_gapfill	Should meteorological variables be gap-filled using ERA-interim estimates?	TRUE
datasetname	User-defined dataset name. Stored as metadata in output files	FLUXNET2015
datasetversion	User-defined dataset version. Stored as metadata in output files	1-3
missing	Maximum percentage of time steps allowed to be missing in any given year	15
gapfill_all	Maximum percentage of time steps allowed to be gap-filled (any quality) in any given year	20
gapfill_good	Same as above for good-quality gap-filling	NA
gapfill_med	Same as above for medium-quality gap-filling	NA
gapfill_poor	Same as above for poor-quality gap-filling	NA
include_all_eval	Should all evaluation values be outputted, regardless of data gaps? If set to FALSE, any evaluation variables with missing or gap-filled values in excess of the thresholds will be discarded.	TRUE
min_yrs	Minimum number of consecutive years to process	2
plot	Output plots to be produced. Set to NA if not required.	c("annual", "diurnal", "timeseries")

Table 2: Site metadata provided with the package. All attributes are provided for each Tier 1 site, with the exception of tower and canopy height.

Attribute	Description
SiteCode	FLUXNET site ID*, e.g. AU-How
Fullname	FLUXNET site name*, e.g. Howard Springs
SiteLatitude	Latitude (degrees north)
SiteLongitude	Longitude (degrees east)
SiteElevation	Elevation (metres)
IGBP_vegetation_short	International Geosphere-Biosphere Programme (IGBP) vegetation type (short name), e.g. WSA
IGBP_vegetation_long	IGBP vegetation type (long name), e.g. Woody Savannas
TowerHeight	Height of measurement tower (metres)
CanopyHeight	Height of canopy at site (metres)
Tier	FLUXNET site tier*
Exclude	Should site be excluded? Allows sites with known problems to be excluded *a priori*. Set to TRUE or FALSE.
Exclude_reason	Reason why site should be excluded (user-defined)

*See http://fluxnet.fluxdata.org/sites/site-list-and-pages/

25

30

35

Table 3: Attributes required for each output variable and stored in `data/Output_variables.R`.

Field name	Description	Value
Fluxnet_variable	Original FLUXNET variable name[1]	e.g. TA_F_MDS
Fluxnet_unit	Original FLUXNET variable unit[1]	e.g. C
Fluxnet_class	Variable data type. Used to define the `colClasses` argument in the R `read.csv` function when reading the input data file. Set to "numeric" if not known.	"numeric" or "integer"
Output_variable	Output variable name	User-defined, e.g. Tair
Output_unit	Output unit (note section 2.4.5 for unit conversions)	User-defined, e.g. K
Longname	Long variable description. Written as a variable attribute in the output file.	User-defined, e.g. Near surface air temperature
Standard_name	Climate and Forecast (CF) convention standard name[2]. Written as a variable attribute in the output file.	User-defined, e.g. air_temperature
Data_min	Minimum acceptable data value. Used to check data ranges. Output units	User-defined, e.g. 200
Data_max	Maximum acceptable data value. Used to check data ranges.	User-defined, e.g. 333
Essential_met	Sets variable as essential when set to `TRUE` (see section 2.4.1.1)	"TRUE" or "FALSE"
Preferred_eval	Sets variable as preferred when set to `TRUE` (see section 2.4.1.2)	"TRUE" or "FALSE"
Category	Determines if the variable is written in the meteorological or evaluation NetCDF output file.	"Met" or "Eval"
ERAinterim_variable	Name of ERA-interim variable[1]	e.g. "TA_ERA"

[1]Must match naming conventions on http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/

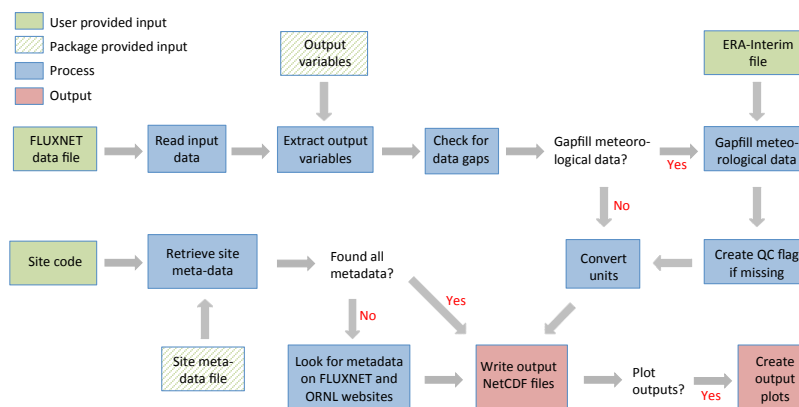
[2]see http://cfconventions.org/standard-names.html

Table 4: Available unit conversions.

Variable	FLUXNET variable name	Original unit	Converted unit
Air temperature	TA_F_MDS	C	K
Rainfall	P	mm	$kg\ m^{-2}\ s^{-1}$
Air pressure	PA	kPa	Pa
Atmospheric CO_2 concentration*	CO_2_F_MDS	$\mu mol\ CO_2\ mol^{-1}$	ppm

*Note these units are equal and the conversion is included to allow different notations

25

30

Figures

5   **Figure 1:** General workflow of the FluxnetLSM R package.

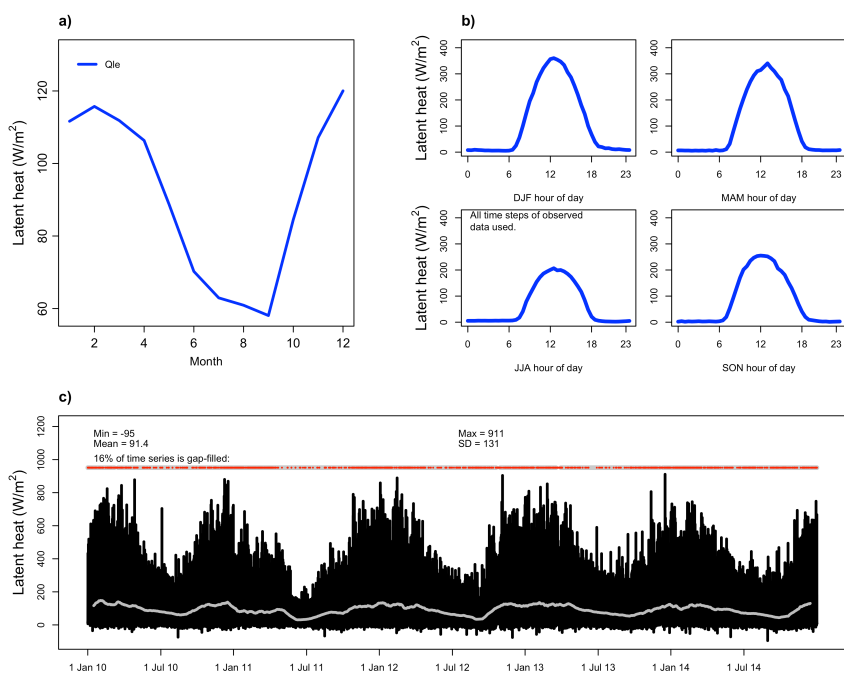



Figure 2: Examples of output plots produced by the package. Mean annual cycle by month is shown in panel a) and mean diurnal cycle by season in panel b). A time series is plotted in panel c), with the full time series shown in black and a smoothed 14-day running mean in grey. Gap-filled periods are indicated in red.

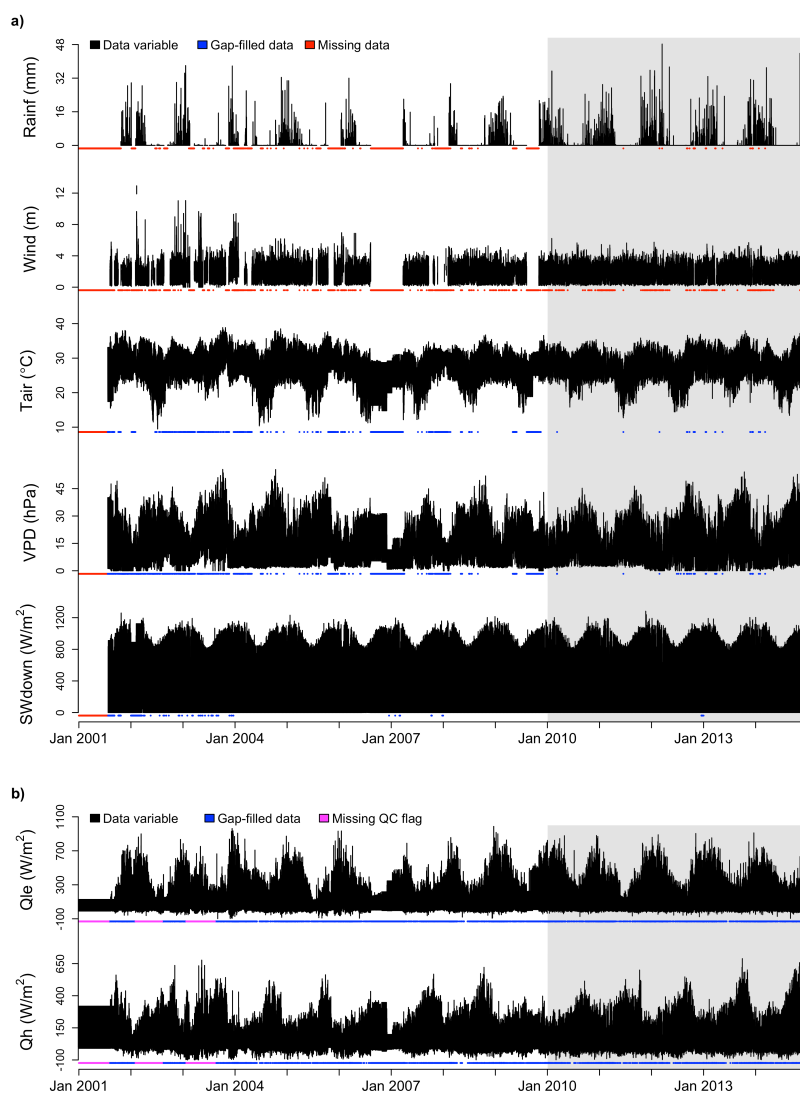



Figure 3: Time series of (a) essential meteorological variables and (b) select evaluation variables in Howard Springs. Meteorological variables include precipitation (Rainf), wind speed (Wind), air temperature (Tair), vapour pressure deficit (VPD) and incoming shortwave radiation (SWdown). Latent heat (Qle) and sensible (Qh) are shown as examples of evaluation variables. Gap-filled periods are indicated in blue and missing periods in data variables in red. For evaluation variables, periods with missing quality control (QC) flags are shown in pink.