UK Centre for Ecology & Hydrology

Guide to produce soil moisture product

Step-by-step guide for producing remote sensing/modelled soil moisture grids for Great Britain

Maliko Tanguy

Project: Hydro-JULES WP5.2

Version 3

Date 19/07/2021

Title Guide to produce soil moisture product

Project HydroJULES WP5.2

Confidentiality, copyright and reproduction For UKCEH internal use only

UKCEH reference version 3

UKCEH contact details Maliko Tanguy

t: 01491692342
e: malngu@ceh.ac.uk

Author Tanguy, Maliko

Approved by

Signed

Date 19/07/2021

Version	Date	Changes in version
v2	15/07/2020	Added code to regrid to 1km instead of 9km
v3	19/07/2021	Version using ASCAT instead of SMOS. Version created after RSE paper was accepted

Contents

1 Introduction 2
2 Code on GitHub 4
3 Download the data 5
3.1 SMAP Data 5
3.2 SMOS Data 9
3.3 CHESS Data 10
3.4 ASCAT Data 11
4 Pre-process the Data 13
4.1 Introduction 13
4.2 SMAP data 13
4.2.1 Reformat from GeoTIFF to NetCDF 13
4.2.2 Average AM and PM soil moisture data, and merge into one large file. 14
4.2.3 Re-grid to a 12.5km grid or a 1km grid 14
4.3 SMOS data 15
4.3.1 Merge all files into one large NetCDF file 16
4.3.2 Crop to study area (GB) and regrid to 9km or 1km 16
4.4 CHESS data 18
4.4.1 Reformatting 1D to 2D 18
4.4.2 Convert NetCDF into GeoTIFF 19
4.4.3 Reproject data 20
4.4.4 Convert back GeoTiff to NetCDF 21
4.4.5 Create one large NetCDF file 22
4.4.6 Re-grid to 12.5km or 1km 23
4.5 ASCAT data 24
4.5.1 Convert raw data into a gridded dataset 24
4.5.2 Convert to a grid 26
4.5.3 Convert degree of saturation to volumetric water content 30
5 Merging the Data: Triple Collocation technique 36
6 Gap-filling the merged soil moisture product ..... **Error! Bookmark not defined.**
7 Formatting the Output Data 44
8 Future work 45

1 Introduction

This document was created to serve as a guide to produce Soil moisture (SM) gridded product for Great Britain based on Jian Peng's work in Hydro-JULES (Peng et al., 2021, accepted for Remote Sensing of Environment).

The initial SM product is the result of merging three datasets using triple collocation technique. The three datasets merged are:

- SMAP
- ASCAT (this was originally SMOS, but was replaced by ASCAT after the paper was reviewed)
- CHESS modelled Soil moisture

The overlapping period for these three datasets currently goes from April 2015 to December 2017. Therefore the initial version of the SM product only convers that period. How to extend and update future versions is open to discussion.

This document covers the following points:

- Where to find the code to produce the grids (section 2)
- Where and how to acquire the input data (section 3)
- Steps needed to pre-process and format the input data (section 4)
- Merging the data (section 5)
- Gap-filling the data (section 6)
- Formatting output data (section 7)

Figure 1 shows a flow chart of the steps needed.

NOTE: Highlighted in blue are the paths that might need updating at a later stage if things get moved around.

All the codes and scripts described in this document have been developed and tested on Linux OS (except step described in section 3.4.3). Therefore, it is advised to run all the steps on UKCEH Linux system.

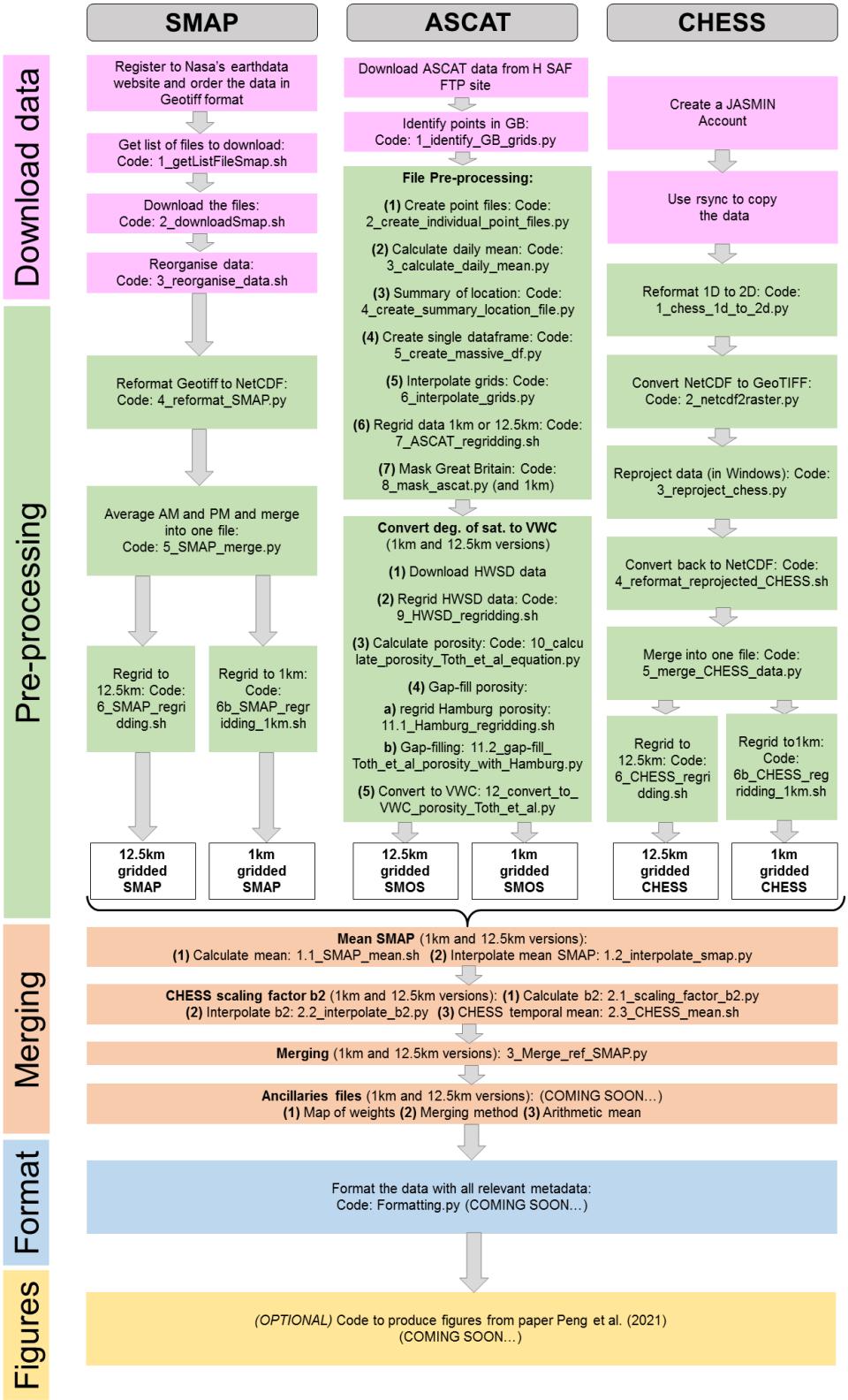


Figure 1: Flow chart of steps needed to produce soil moisture product

2 Code on GitHub

The code to produce the soil moisture grids can be found on HydroJULES private GitHub repository: https://github.com/hydro-jules/soil_moisture

If you don't have access to the repository, request access to Thibault Hallouin (thibault.hallouin@ncas.ac.uk), Rich Ellis (rjel@ceh.ac.uk) or Matt Fry (mfry@ceh.ac.uk).

You can make a clone of the repository wherever you want on one of UKCEH Linux machines, but one sensible place would be to save the code in ***/prj/hydrojules/users/<your_username>/***

The description of the codes in all following sections will assume the code is saved in this directory.

To clone the repository, from the directory where you want it to be saved, from a Linux machine, type the command:

```
git clone https://github.com/hydro-jules/soil_moisture
```

If you already have a copy of the repository, to make sure you have the latest version, check if there any changes between your copy and the original repository with the command:

```
git fetch
```

If there are some differences and you want the latest version, use the command:

```
git pull
```

3 Download the data

The first step is to download the data, which requires registration to different web pages.

3.1 SMAP Data

Citation for this data:

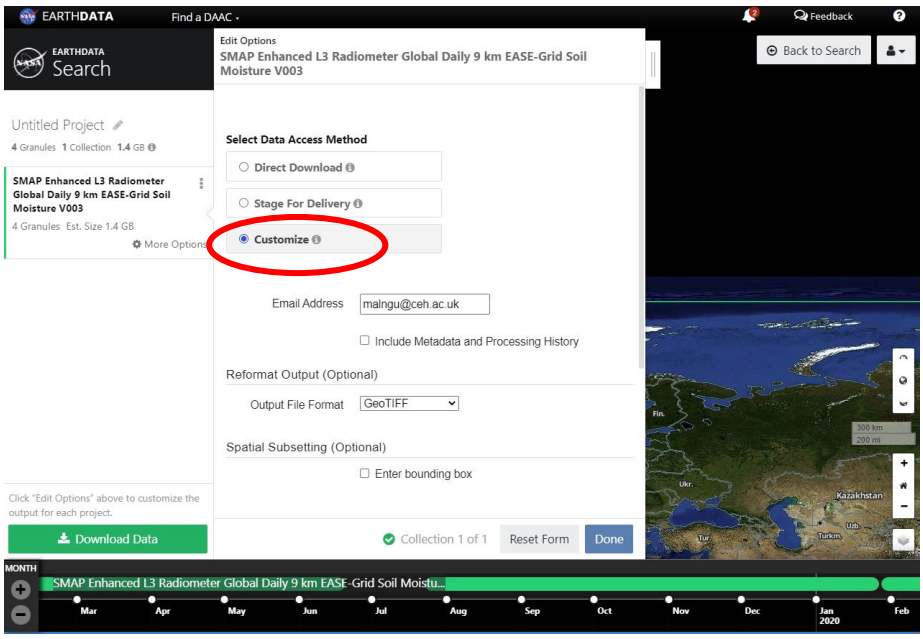
O'Neill, P. E., S. Chan, E. G. Njoku, T. Jackson, R. Bindlish, and J. Chaubell. 2019. *SMAP Enhanced L3 Radiometer Global Daily 9 km EASE-Grid Soil Moisture, Version 3.* [Indicate subset used]. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. doi: https://doi.org/10.5067/T90W6VRLCBHI. [Date Accessed].

STEPS to acquire SMAP data are:

1) Register for free to https://earthdata.nasa.gov/

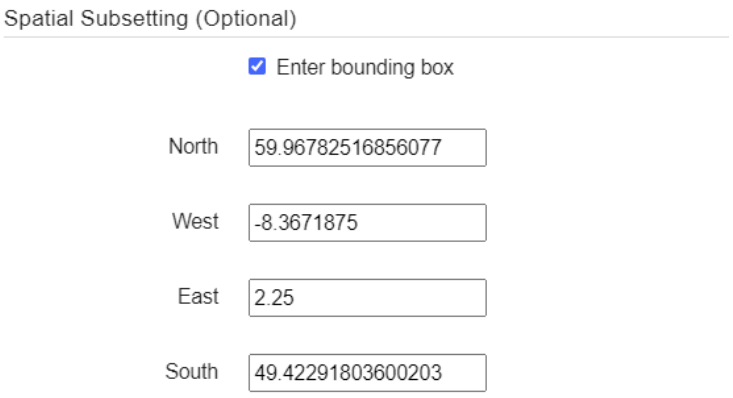
2) Decide the range of dates you want to download data for. The records start on 2015/03/31. To check the latest date available, check https://n5eil01u.ecs.nsidc.org/DP4/SMAP/SPL3SMP_E.003 (usually available up to 2 days ago).
3) Customise start and end dates (high-lighted in yellow) from link below: https://search.earthdata.nasa.gov/projects?p=C1625717578-NSIDC_ECS!C1625717578-NSIDC_ECS&q=SMAP%20Enhanced%20L3%20Radiometer%20Global%20Daily%209%20km%20EASE-Grid%20Soil%20Moisture%20V003&sb=-8.3671875%2C49.42291803600203%2C2.25%2C59.96782516856077&m=76.64339153591293!-12.375!2!1!0!0%2C2&qt=2015-03-31T00%3A00%3A00.000Z%2C2020-02-28T23%3A59%3A59.999Z&tl=1565780368!4!!

4) Copy the link with the updated dates into your browser. You will get to a page similar to the screenshot below. Choose the option “Customize” (red circle below).

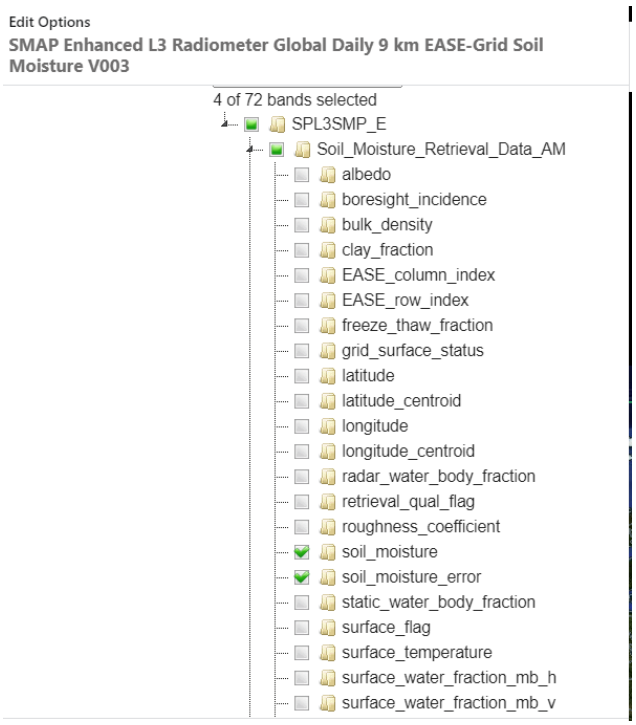


Choose the following options:

- Email address: provide the address where you want your order to be sent.
- Format: GeoTIFF (there is an option to download in netcdf format as well, but the file needs reformatting to match Jian's code anyway so it is easier to download them in Geotiff format)
- Spatial Subsetting: (it should populate automatically with the link in point 3, but if not, enter the following coordinates)

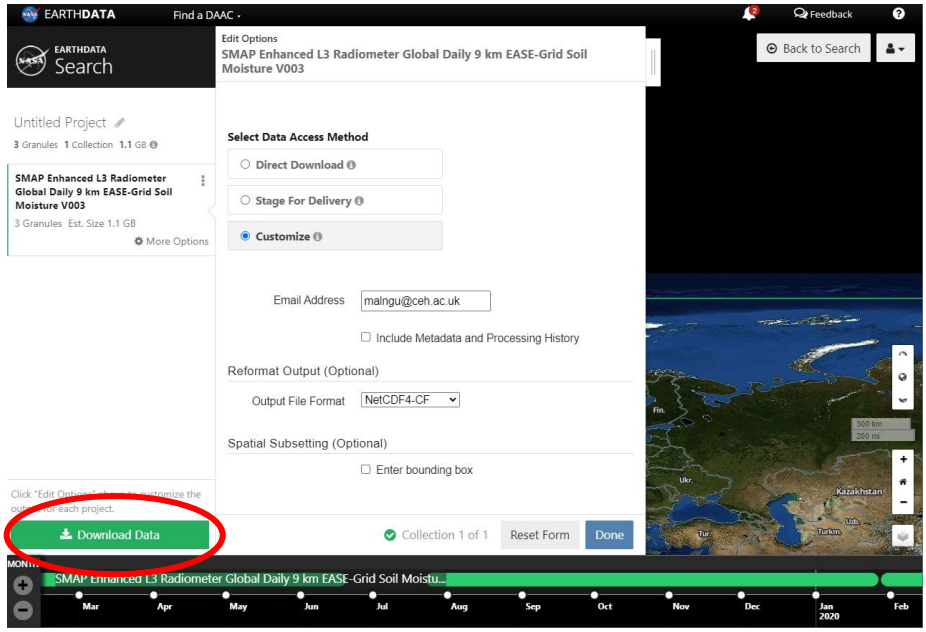


- Re-projection options: Geographic
- Band Subsetting: We don't need all the bands, so select only the following four bands:
    - In "Soil_Moisture_Retrieval_Data_AM":
        - soil_moisture
        - soil_moisture_error
    - In "Soil_Moisture_Retrieval_Data_PM":
        - soil_moisture_PM
        - soil_moisture_error_PM



NOTE: we are currently not using the “soil_moisture_error” band, so we might not need it.

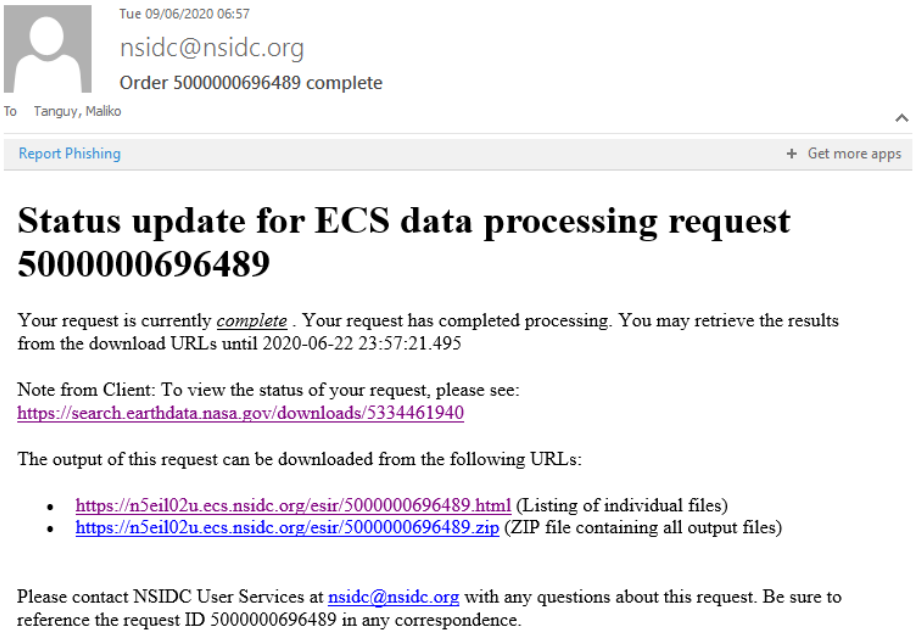
5) Once you have customized your order, click on “Download Data” (red circle below)



When you click that button, you will be redirected to a page informing you of the Status of your order. You can close that window and wait until you receive the notification emails.

6) You will receive a first email shortly after placing your order to notify you that the order is being processed. Once the order is ready, you will receive a second

email informing you that the order is ready. The email will look similar to the following screenshot:



Tue 09/06/2020 06:57

nsidc@nsidc.org

Order 5000000696489 complete

To Tanguy, Maliko

Report Phishing + Get more apps

Status update for ECS data processing request 5000000696489

Your request is currently *complete* . Your request has completed processing. You may retrieve the results from the download URLs until 2020-06-22 23:57:21.495

Note from Client: To view the status of your request, please see:
https://search.earthdata.nasa.gov/downloads/5334461940

The output of this request can be downloaded from the following URLs:

- https://n5eil02u.ecs.nsidc.org/esir/5000000696489.html (Listing of individual files)
- https://n5eil02u.ecs.nsidc.org/esir/5000000696489.zip (ZIP file containing all output files)

Please contact NSIDC User Services at nsidc@nsidc.org with any questions about this request. Be sure to reference the request ID 5000000696489 in any correspondence.

If you have only ordered a small amount of files, you can download directly the zip file in your email. However, if you have ordered a large amount of data (more than a few days), it is not recommended that you download the zip file, as the download can take a long time and fail halfway through the process.
If you need to download a large amount of files, follow the steps 7 and 8.

The scripts mentioned in steps 7, 8 and 9 are all located in the following directory:
/prj/hydrojules/users/<your_username>/soil_moisture/1_preprocessing/smap/

7) First, you will need to download a text file with the list of all files to be downloaded. For this, use the script: ***1_getListFileSmap.sh***

   Usage is:
   `1_getListFileSmap.sh <request_ID>`
   You can find the request ID in the email you have received.
   You will be prompted your EarthData username and password, so have them ready.

   In the example above, the request ID is 5000000696489.
   This will download a text filed called 5000000696489.txt in the same directory.

8) Next, use the following script to download all the files: ***2_downloadSmap.sh***

   Usage is:
   `2_downloadSmap.sh <file_name>`
   <file_name> is the name of the file you have downloaded in point 7.
   In the previous example, the command would be:
   `2_downloadSmap.sh 5000000696489.txt`

You will be prompted your EarthData username and password again.
By default, the files you download will be saved in:
/prj/hydrojules/data/soil_moisture/preprocessed/smap/smap_tif/
If you want to change that, you will have to manually edit the script and change the variable ***targetFolder***.

If the download fails halfway through, you can make a copy of the text file with the list of files to download, and manually edit it deleting all files that have already successfully been downloaded. You can then rerun the ***2_downloadSmap.sh*** script using this new text files to download the remaining files.

9) Reorganise the data:
Lastly, we need to reorganise the data within the main folder to separate error data (currently not used) and soil moisture AM and PM data.
Run the following code: ***3_reorganiseData.sh***

The default main folder is:
/prj/hydrojules/data/soil_moisture/preprocessed/smap/ smap_tif/
If you want to change that, you will have to manually edit the script and change the variable ***targetFolder***.

3.2 SMOS Data

NOTE: SMOS data is not used in the latest version of the SM dataset, and has been replaced by ASCAT following concerns from reviewers of lack of independence between SMAP and SMOS (both are passive microwave products). However, this section is left in this document in case it is useful for future work.

STEPS to acquire SMOS data are:

1) First, register to this website: http://bec.icm.csic.es/bec-ftp-service-registration/

2) You will receive an email with your credentials.

3) First, check what the latest date available on the server is (First date available: 2010-06-01). Use the following commands (replace "2020" by current year):

```
sftp -P27500 <username>@becftp.icm.csic.es
<username>@becftp.icm.csic.es's password:
sftp> cd
/data/LAND/SM/SMOS/EUROPE_and_MEDITERRANEAN/v5.0/L4/REPROCESSED/daily/ASC/2020/
sftp> ls
```

This will show the list of files available. Data is loaded usually once a week.
If you need to download just a few files, you can use the "get" command to copy them:

```
sftp> get <file_name>
```

Exit the sftp site using the "exit" command.
Change the permission on the files that you have downloaded using:

```
chmod 664 *.nc
```

If you need to download a large amount of files, you can use the script described below in point 4).

4) If you need more than a few files, you can use the following script (in Linux) to download the data:
/prj/hydrojules/users/<your_username>/soil_moisture/1_preprocessing/smos/1_downloadSmos.sh

   You will be prompted your BEC username and password (from your registration email), and the start and end dates (both inclusive) of the period you want to download data for.

   At the moment, there is no clever syntax checking or error handling bits in the code.

   The data will be saved in:
/prj/hydrojules/data/soil_moisture/preprocessed/smos/smos_nc/

   If you want to change the path, you have to modify the variable ***targetFolder*** in the code (first line of code)

3.3 CHESS Data

STEPS to download CHESS data:

1) Create a JASMIN account:

   CHESS data is stored on JASMIN.

   If you do not have a JASMIN account, you will need to:

   - Create a JASMIN account: https://accounts.jasmin.ac.uk/services/group_workspaces/?query=hydro_jules
   - Set up everything on your machine to be able to access JASMIN: https://help.jasmin.ac.uk/article/189-get-started-with-jasmin
   - Request access to the Group workspace hydro_jules: (https://accounts.jasmin.ac.uk/services/group_workspaces/?query=hydro_jules).

2) Transfer the data using *rsync*:

   Soil moisture data can be found in:

   /gws/nopw/j04/hydro_jules/data/uk/jules_outputs/chess/

If you need to copy the data to your local machine, follow these steps (on a Linux terminal):

```
eval   $(ssh-agent -s)

ssh-add ~/.ssh/id_rsa_jasmin

#(you will be prompted for your JASMIN passphrase)

rsync -avzh <jasmin_username>@jasmin-xfer1.ceda.ac.uk:
/gws/nopw/j04/hydro_jules/data/uk/jules_outputs/chess/<file_name>
<path_on_your_local_machine>
```

NOTE: there is no space after `<jasmin_username>@jasmin-xfer1.ceda.ac.uk:`

Change the permissions on these files by using the command:

```
chmod 664 <file_name>
```

3.4 ASCAT Data

Citation for this data:

H SAF (2020): ASCAT Surface Soil Moisture Climate Data Record v5 12.5 km sampling - Metop, EUMETSAT SAF on Support to Operational Hydrology and Water Management, DOI: 10.15770/EUM_SAF_H_0006.
http://doi.org/10.15770/EUM_SAF_H_0006

STEPS to acquire ASCAT data are:

1) Register for free to http://hsaf.meteoam.it/User/Register

2) Identify all grids located in the UK:

   Grid point locator: https://dgg.geo.tuwien.ac.at/
   You can export list of points for a given country on WARP grid → gives list of cell_id needed later (for mget command, point three below)

3) Download ASCAT soil moisture H115 product using the following commands (the data is currently stored here: /prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/):

   ```
   ftp -i ftphsaf.meteoam.it
   ```
   (you will be prompted your username and password)
   ```
   cd h115
   get H115_<cell_id>.nc #if you want one specific cell
   mget *.nc #if you want all files (you can use wildcards and use list of
   points generated in point 2)
   ```

4) Check which files correspond to Great Britain (point 2 identifies the data for the UK, but here we narrow it down to GB):
   The code to run can be found in this folder:
   /prj/hydrojules/users/<username>/shared_code/soil_moisture/1_preprocessing/ascat/

Use code: ***1_identify_GB_grids.py***

This code reads in a control file, which specifies the path to the folder where the files are located.

An example of the control file is: ***1_CONTROL_FILE_ASCAT.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to folder with ASCAT raw data
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***1_CONTROL_FILE_ASCAT.txt***

Usage:

```
python 1_identify_GB_grids.py <control_file_name>
```

4 Pre-process the Data

4.1 Introduction

To be able to use Jian Peng's code, the three datasets each need to be converted into a single NetCDF file, with the grid matching the 12.5km grid from ASCAT data. The three files must have the same temporal and spatial extend.

The code was initially developed to produce a 9km gridded merged product, then to 12.5km when SMOS was replaced by ASCAT. It was subsequently modified to be able to downscale it to a 1km product. There is now the option to produce either 12.5km or 1km spatial resolution.

This section describes the steps required to achieve this.

4.2 SMAP data

SMAP data is acquired in GeoTIFF format.

The data needs reformatting to NetCDF, averaging between AM and PM values, merging into one large file, and re-gridding to a 12.5km grid or a 1km grid.

All the code used in this section are located here:
/prj/hydrojules/users/<your_username>/soil_moisture/1_preprocessing/smap/

4.2.1 Reformat from GeoTIFF to NetCDF

Use code: ***4_reformat_SMAP.py***

This code reads in a control file, which specifies the path of input and output files.

An example of the control file is: ***4_CONTROL_FILE_SMAP.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Main input folder (there should be subfolders 'smap_AM' and 'smap_PM' in
it)
/prj/hydrojules/data/soil_moisture/preprocessed/smap/smap_tif/

# Main output folder (subfolders 'smap_AM' and 'smap_PM' will be created if
they don't exist)
/prj/hydrojules/data/soil_moisture/preprocessed/smap/smap_nc/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***4_CONTROL_FILE_SMAP.txt***

Usage:

```
python 4_reformat_SMAP.py <control_file_name>
```

4.2.2 Average AM and PM soil moisture data, and merge into one large file.

NOTE: At the moment dates are hardcoded in the code.

Use code: ***5_SMAP_merge.py***

This code reads in a control file, which specifies the path of input and output files.

An example of the control file is: ***5_CONTROL_FILE_SMAP.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Main input folder with NetCDF files (there should be subfolders 'smap_AM'
and 'smap_PM' in it)
/prj/hydrojules/data/soil_moisture/preprocessed/smap/smap_nc/

# Output folder
/prj/hydrojules/data/soil_moisture/preprocessed/smap/smap_merged/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***5_CONTROL_FILE_SMAP.txt***

Usage:

```
python 5_SMAP_merge.py <control_file_name>
```

4.2.3 Re-grid to a 12.5km grid or a 1km grid

Finally, the output grids are cropped to the study area and re-gridded to either 12.5km grid or 1km grid. Which grid you decide to produce will affect which code you need to run in subsequent steps.

1) <u>Produce a 12.5km grid</u>

Use code: ***6_SMAP_regridding.sh***

This code needs **cdo** to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***6_CONTROL_FILE_SMAP.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to the merged SMAP netCDF file:
/prj/hydrojules/data/soil_moisture/preprocessed/smap/smap_merged/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***6_CONTROL_FILE_SMAP.txt***

USAGE:

```
6_SMAP_regridding.sh <control_file_name>
```

2) <u>Produce a 1km grid</u>

Use code: ***6b_SMAP_regridding_1km.sh***

This code needs **cdo** to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***6b_CONTROL_FILE_SMAP.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to the merged SMAP netCDF file:
/prj/hydrojules/data/soil_moisture/preprocessed/smap/smap_merged/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***6b_CONTROL_FILE_SMAP.txt***

USAGE:

```
6b_SMAP_regridding_1km.sh <control_file_name>
```

4.3 SMOS data

NOTE: This section was written before SMOS was replaced by ASCAT. These steps are thus not needed anymore, but the section is kept here in case SMOS data is used in future for another purpose.

SMOS data is already in NetCDF. The only steps needed is to merge all the files into a single NetCDF file, crop to the study area and regrid to 9km or 1km grid.

The codes for this section are located in the following directory:

/prj/hydrojules/users/<your_username>/soil_moisture/1_preprocessing/smos/

4.3.1 Merge all files into one large NetCDF file

Use code: ***2_SMOS_merge.py***

This code reads in a control file, which specifies the path of input and output files.

An example of the control file is: ***2_CONTROL_FILE_SMOS.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# start date (format YYYY-MM-DD):
2015-04-01

# end date (format YYYY-MM-DD):
2017-12-31

# Main input folder with NetCDF files
/prj/hydrojules/data/soil_moisture/preprocessed/smos/smos_nc/

# Output folder
/prj/hydrojules/data/soil_moisture/preprocessed/smos/smos_merged/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***2_CONTROL_FILE_SMOS.txt***

Usage:

```
python 2_SMOS_merge.py <control_file_name>
```

NOTE: This step takes some time to run.

4.3.2 Crop to study area (GB) and regrid to 9km or 1km

Finally, the output grids are cropped to the same size as the rest of data. There are two options: produce either a 9km grid or 1km grid. Which grid you decide to produce will affect which code you need to run in subsequent steps.

1) Produce a 9km grid

Use code: ***3_SMOS_regridding.sh***

This code needs cdo to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***3_CONTROL_FILE_SMOS.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to the merged SMOS netCDF file:
/prj/hydrojules/data/soil_moisture/preprocessed/smos/smos_merged/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***3_CONTROL_FILE_SMOS.txt***

USAGE:

```
3_SMOS_regridding.sh <control_file_name>
```

2) Produce a 1km grid

Use code: ***3b_SMOS_regridding_1km.sh***

This code needs cdo to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***3b_CONTROL_FILE_SMOS.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to the merged SMOS netCDF file:
/prj/hydrojules/data/soil_moisture/preprocessed/smos/smos_merged/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***3b_CONTROL_FILE_SMOS.txt***

USAGE:

```
3b_SMOS_regridding_1km.sh <control_file_name>
```

4.4 CHESS data

NOTE: IMPROVEMENT NEEDED.

The pre-processing of CHESS data needs improving, as it is currently very inefficient.

The original code from Jian Peng did not work for some reason.

New code was developed to achieve the same result, but involves many unnecessary steps, switching between Linux and Windows and moving files around.

CHESS data is provided in NetCDF files, but are not gridded (1D land-only vector). The following pre-processing steps are necessary:

NOTE: For files from 1961 to 2015, the “time” variable in the original CHESS NetCDF files are in the unit “seconds from 1961-01-01”. From 2016 onwards, the unit changes to “seconds from <current_year>-01-01”. So for example for the 2016 file, the time units are “seconds from 2016-01-01”.

If this changes, it might produce some errors, which will require some tweaking in the codes in future versions.

All the codes and control files in this section are found in the following directory:

/prj/hydrojules/users/<your_username>/soil_moisture/1_preprocessing/chess/

4.4.1 Reformatting 1D to 2D

The original CHESS data is in one dimension. This needs to be reformatted to a 2D grid.

This is done using the following python code: ***1_chess_1d_to_2d.py***

To run this code, you need to define the variables in a control file. You can find an example of a control file here: ***1_CONTROL_FILE_CHESS.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# start year:
```

```
2015

# end year:
2017

# path to folder with input 1D CHESS data
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_1d/

# path to folder with output 2D CHESS data
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***1_CONTROL_FILE_CHESS.txt***

Usage:

```
python 1_chess_1d_to_2d.py <control_file_name>
```

Note that this code also needs the file ***CHESS_pdef_jasmin.gra*** to be in the same folder.

4.4.2 Convert NetCDF into GeoTIFF

This is done using the following code: ***2_netcdf2raster.py***

To run this code, you need to define the variables in a control file. You can find an example of a control file here: ***2_CONTROL_FILE_CHESS.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# start year:
2015

# end year:
2017

# path to folder with input 2D NetCDF CHESS files
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/

# path to folder with output GeoTiff CHESS files
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/geotiff/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***2_CONTROL_FILE_CHESS.txt***

Usage:

```
python 2_netcdf2raster.py <control_file_name>
```

4.4.3 Reproject data

NOTE: THIS STEP PARTICULARLY NEEDS IMPROVING.

In this step, the files produced in the previous point need to be copied across to the local computer, in Windows. The reprojection is made through a python (2.7) code, which uses arcpy module (from ArcGIS). This is slow and only works on Windows, and requires ArcGIS to be installed.

Once the files have been reprojected, the new files need to be copied again on the Linux system to continue with the next step.

CHESS data is on the British National Grid, whereas SMAP and SMOS data are in geographical coordinates (WGS 1984).

This paragraph explains how to reproject the data to match the projection of the other two datasets.

NOTE: You need to have ArcGIS installed on your computer. If you do not have it, contact CCS helpdesk (wlhelp@ceh.ac.uk).

Follow the steps below:

1) Copy python code

Make a copy on your local computer (Windows) of the following code: ***3_reproject_chess.py***

To serve as an example for the following instruction, we will assume the code was copied to the following path: ***C:\code\3_reproject_chess.py***

Define start_year and end_year on line 5 and 6 of the code.

2) Create local directory and copy CHESS Geotiff files.

Create a directory on your C drive called: ***C:\chess_tif***

Copy the Geotiff files created in step 3.4.2 in this new directory from the following folder: \\nercwlsmb01\prj\hydrojules\data\soil_moisture\preprocessed\chess\chess_2d\geotiff

3) Identify python executable linked to ArcGIS

To run the code, you need to use python which is linked to ArcGIS. Check the path on your computer. It is usually similar to the following path: ***C:\Python27\ArcGIS10.6\python.exe***

4) Run the code from the command line

Open a command prompt (search “Command Prompt” in Search Windows).

Run the following command:

```
<path_to_python_executable> <path_to_python_reproject_code>
```

So in our example, the command will be:

```
C:\Python27\ArcGIS10.6\python.exe C:\code\3_reproject_chess.py
```

5) Copy back the files to Linux

Once the code has finished running, the output files need to be copied back to UKCEH Linux drives to continue with the rest of the steps.

Unfortunately, this needs to be done in two steps, as you cannot copy files directly from Windows to the ‘prj’ folder on Linux.

- First, create a directory called ‘chess_reprojected’ in your personal workspace on Linux
- Copy the reprojected files from Windows (in C:/chess_tif/temp/) to your newly created directory on your personal workspace on Linux (\\nercwlsmb01\<your_username>\chess_reprojected)
- We only need the Geotiff files (ending in ‘.tif’) so we will delete all the others, using the following commands (in Linux):

```
cd ~/chess_reprojected/
rm *.tfw
rm *.xml
rm *.ovr
```

- Move the files from your personal workspace to the HydroJULES project folder using the following command:

```
cp ~/chess_reprojected/*.tif
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/reprojected/
```

- Change the permissions on the files copied from Windows. In the folder where the files are located, use the following command:

```
chmod 664 *.tif
```

4.4.4 Convert back GeoTiff to NetCDF

Now that the grids have been reprojected, they need to be converted back to NetCDF format.

Use the following code: ***4_reformat_reprojected_CHESS.sh***

To run this code, you need to define the variables in a control file. You can find an example of a control file here: ***4_CONTROL_FILE_CHESS.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Start date (Format YYYY-MM-DD):
2015-1-1

# End date (Format YYYY-MM-DD):
2017-12-31

# Path to input CHESS reprojected Geotiff files:
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/geotiff/temp
/

# Path to output CHESS reprojected NetCDF files:
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/reprojected/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***4_CONTROL_FILE_CHESS.txt***

Usage:

```
4_reformat_reprojected_CHESS.sh <control_file_name>
```

4.4.5 Create one large NetCDF file

Next, a large NetCDF file is created, as this is the format expected in the codes in the next steps.

Use the following code: ***5_merge_CHESS_data.py***

To run this code, you need to define the variables in a control file. You can find an example of a control file here: ***5_CONTROL_FILE_CHESS.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# start date (format: YYYY-MM-DD):
2015-4-1

# end date (format: YYYY-MM-DD):
2017-12-31

# path to folder with input reprojected files
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/reprojected/

# path to folder with output merged CHESS files
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***5_CONTROL_FILE_CHESS.txt***

Usage:

```
python 5_merge_CHESS_data.py <control_file_name>
```

NOTE: This step takes some time to run.

4.4.6 Re-grid to 12.5km or 1km

Finally, the output grids are cropped to the study area, and are regridded to either 12.5km or 1km. Which grid you decide to produce will affect which code you need to run in subsequent steps.

1) Produce a 12.5km grid

Use code: ***6_CHESS_regridding.sh***

This code needs cdo to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***6_CONTROL_FILE_CHESS.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to the merged SMAP netCDF file:
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***6_CONTROL_FILE_CHESS.txt***

USAGE:

```
6_CHESS_regridding.sh <control_file_name>
```

2) Produce a 1km grid

Use code: ***6b_CHESS_regridding_1km.sh***

This code needs cdo to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***6b_CONTROL_FILE_CHESS.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to the merged SMAP netCDF file:
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***6b_CONTROL_FILE_CHESS.txt***

USAGE:

```
6b_CHESS_regridding_1km.sh <control_file_name>
```

4.5 ASCAT data

ASCAT data downloaded from H SAF ftp site has a unique structure (time series at point data on the Swath) which needs to be converted into a regular grid.

Furthermore, ASCAT soil moisture is provided in degree of saturation, which needs to be converted into volumetric water content to match the unit from the other datasets.

The pre-processing steps for ASCAT data are thus: (i) to convert raw data into daily time series; (ii) convert into a regular grid and re-grid the data to match the other datasets spatial resolution and extension; (iii) convert SM to volumetric water content. These steps are described in detail in this section.

The codes for this section are located in the following directory:

/prj/hydrojules/users/<your_username>/soil_moisture/1_preprocessing/ascat/

4.5.1 Convert raw data into daily time series

The conversion into a gridded dataset consists of the following sub-steps:

1) **First step:** to indentify GB grids. This point is described in section 3.4.1.

2) **Create individual point files:**

ASCAT data is structured into various netcdf files. Each netcdf files have a series of point data, but all lumped together. The first step is to convert this complicated netcdf file into a set of simpler time series files: one file per point.

Use code: ***2_create_individual_point_files.py***

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***2_CONTROL_FILE_ASCAT.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to folder with ASCAT raw data:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***2_CONTROL_FILE_ASCAT.txt***

USAGE:

```
python 2_ create_individual_point_files.py <control_file_name>
```

This code creates two sub-folders (if they don't exist already) in the same folder where the raw data is stored:

- *Subfolder locFolder:* where some information on the coordinates of each individual points is stored
- *Subfolder timeSeriesFolder:* where the actual time series for each individual point are stored in csv format

3) Calculate daily mean

The second step is to calculate the daily mean from the time series data.

Use code: ***3_calculate_daily_mean.py***

This code reads in a control file, which specifies start and end dates, and the path to the input file.

An example of the control file is: ***3_CONTROL_FILE_ASCAT.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# start date:
2015-04-01

# end date:
2017-12-31

# path to folder with ASCAT raw data:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***3_CONTROL_FILE_ASCAT.txt***

USAGE:

```
python 3_calculate_daily_mean.py <control_file_name>
```

This code creates one sub-folders (if it doesn't exist already) in the same folder where the raw data is stored:

- *Subfolder timeSeriesFolder/dailyMean/:* where the daily mean time series for each individual point are stored in csv format

4) Create location summary file

This step is to summarise the location information for the individual points.

Use code: ***4_create_summary_location_file.py***

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***4_CONTROL_FILE_ASCAT.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to folder with ASCAT raw data:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***4_CONTROL_FILE_ASCAT.txt***

USAGE:

```
python 4_create_summary_location_file.py <control_file_name>
```

This code will create some files in the subfolder *locFolder/*.

4.5.2 Convert to a grid

5) Create one single dataframe

To make subsequent calculation easier, all the data is summarised into a single large dataframe:

Use code: ***5_create_massive_df.py***

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***5_CONTROL_FILE_ASCAT.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to folder with ASCAT raw data:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***5_CONTROL_FILE_ASCAT.txt***

USAGE:

```
python 5_create_massive_df.py <control_file_name>
```

The large dataframe will be created in the subfolder *locFolder/*.

6) Interpolate grids

Linear interpolation was adopted after testing a few options. This is hardcoded in the code, but can be changed if required (line 7 of code).

NOTE: all subsequent code uses “linear” interpolation as default. The only option to use a different interpolation method at the moment is to modify the hardcoded options in the successive codes.

Use code: ***6_interpolate_grids.py***

This code reads in a control file, which specifies start and end dates, and the path to the input file.

An example of the control file is: ***6_CONTROL_FILE_ASCAT.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# start date:
2015-04-01

# end date:
2017-12-31

# path to folder with ASCAT raw data:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***6_CONTROL_FILE_ASCAT.txt***

USAGE:

```
python 6_interpolate_grids.py <control_file_name>
```

7) Regrid data

This is to make sure all datasets have the same resolution and extension

a) Produce a 12.5km grid

Use code: ***7_ASCAT_regridding.sh***

This code needs cdo to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***7_CONTROL_FILE_ASCAT.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to the merged ASCAT netCDF file:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/gridded_data/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***7_CONTROL_FILE_ASCAT.txt***

USAGE:

```
7_ASCAT_regridding.sh <control_file_name>
```

b) Produce a 1km grid

Use code: ***7b_ASCAT_regridding_1km.sh***

This code needs cdo to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***7b_CONTROL_FILE_ASCAT.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to the merged ASCAT netCDF file:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/gridded_data/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***7b_CONTROL_FILE_ASCAT.txt***

USAGE:

```
7b_ASCAT_regridding_1km.sh <control_file_name>
```

8) Mask GB

We use CHESS data to mask out areas not included in GB.

a) 12.5km version

Use code: ***8_mask_ascat.py***

This code reads in a control file, which specifies the path to the input and output files.

An example of the control file is: ***8_CONTROL_FILE_ASCAT.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to 12.5km CHESS file
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/chess
_12.5km.nc

# path to folder with ASCAT data
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/gridded_data/

# path to output folder:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/gridded_data/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***8_CONTROL_FILE_ASCAT.txt***

USAGE:

```
python 8_mask_ascat.py <control_file_name>
```

NOTE: if you need to mask ASCAT using a CHESS file having a different temporal extent to the ASCAT file, you can use a workaround: just uncomment the text following the note in the code after line 74.

b) 1km version

Use code: ***8b_mask_ascat_1km.py***

This code reads in a control file, which specifies the path to the input and output files.

An example of the control file is: ***8b_CONTROL_FILE_ASCAT.txt***

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to 1km CHESS file
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/chess
_1km.nc

# path to folder with ASCAT data
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/gridded_data/

# path to output folder:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/gridded_data/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***8b_CONTROL_FILE_ASCAT.txt***

USAGE:

```
python 8b_mask_ascat_1km.py <control_file_name>
```

4.5.3 Convert degree of saturation to volumetric water content

ASCAT provides SM estimates as degree of saturation. This has to be converted to volumetric water content in order to be able to merge the data with the other two datasets.

Volumetric water content = degree of saturation (%) x porosity

To estimate porosity, we use the Harmonized World Soil Database (HWSD, Wieder et al., 2014) and the pedo-transfer function from Toth et al. (2015) - equation 22 from the supplementary material:

$\theta_s = 0.63052 - 0.10262 * BD^2 + 0.0002904 * pH^2 + 0.0003335 * Cl$

θ_s is the saturated soil moisture content (≈porosity), BD = bulk density (T_BULK_DEN from HWSD), pH is pH in water, Cl is Clay content (%)

HWSD data is downloaded from: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1247

The data is stored here:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/HWSD/

References:

Toth, B., Weynants, M., Nemes, A., Mako, A., Bilas, G., Toth, G. (2015): New generation of hydraulic pedotransfer functions for Europe, Eur. J. Soil Sci., 66 (1), pp. 226-238, https://doi.org/10.1111/ejss.12192

Wieder, W.R., J. Boehnert, G.B. Bonan, and M. Langseth. 2014. Regridded Harmonized World Soil Database v1.2. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA. http://dx.doi.org/10.3334/ORNLDAAC/1247

9) Regrid HWSD data

This is to make sure all datasets have the same resolution and extension. There are two versions: one for the 12.5km and the other for the 1km grids (these are shown in parenthesis).

Use code:

9_HWSD_regridding.sh *(**9b_HWSD_regridding_1km.sh** for the 1km version)*

This code needs cdo to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***9_CONTROL_FILE_HWSD.txt*** *(**9b_CONTROL_FILE_HWSD.txt** for the 1km version)*

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to the HWSD netCDF file:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/HWSD/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***9_CONTROL_FILE_HWSD.txt*** *(**9b_CONTROL_FILE_HWSD.txt** for the 1km version)*

USAGE:

```
9_HWSD_regridding.sh <control_file_name>
```

```
9b_HWSD_regridding_1km.sh <control_file_name>
```

10)Calculate porosity

Using Toth et al. (2015) pedo-transfer function. There are two versions: one for the 12.5km and the other for the 1km grids (these are shown in parenthesis).

NOTE: The default bulk density data used is the “T_BULK_DEN” data from HWSD (there is another option “T_REF_BULK”, which is calculated in a slightly different way). All subsequent code uses this bulk density as default. The only option to use change this at the moment is to modify the hardcoded options in the successive codes.

Use code:

10_calculate_porosity_Toth_et_al_equation.py
*(**10b_calculate_porosity_Toth_et_al_equation_1km.py** for the 1km version)*

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***10_CONTROL_FILE_HWSD.txt***
*(**10b_CONTROL_FILE_HWSD.txt** for the 1km version)*

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to HWSD data:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/HWSD/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***10_CONTROL_FILE_HWSD.txt*** *(**10b_CONTROL_FILE_HWSD.txt** for the 1km version)*

USAGE:

```
python 10_calculate_porosity_Toth_et_al_equation.py <control_file_name>
```

```
python 10b_calculate_porosity_Toth_et_al_equation_1km.py
<control_file_name>
```

11)Hamburg dataset for gapfilling porosity

The porosity data derived in the previous point has some gaps in it. To fill these gaps, we use the dataset created by Hamburg university available to download from here: https://www.cen.uni-hamburg.de/en/icdc/data/land/ascat-soilmoisture.html

The porosity calculated by Hamburg is also based on HWSD, and is very similar (but not identical) to the one calculated in point 10. The reason we haven't used Hamburg's porosity directly is because there are some problems with the data along the coast. However, it is used to gap-fill where data is missing in our grids.

Hamburg dataset is saved here:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/HWSD/ASCAT_Hamburg_porosity.nc

11.1. Regrid Hamburg porosity

There are two versions: one for the 12.5km and the other for the 1km grids (these are shown in parenthesis).

Use code:

11.1_Hamburg_regridding.sh *(**11.1b_Hamburg_regridding_1km.sh** for the 1km version)*

This code needs cdo to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***11.1_CONTROL_FILE_Hamburg.txt*** *(**11.1b_CONTROL_FILE_Hamburg.txt** for the 1km version)*

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to Hamburg file:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/HWSD/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***11.1_CONTROL_FILE_Hamburg.txt*** *(**11b_CONTROL_FILE_Hamburg.txt** for the 1km version)*

USAGE:

```
11.1_Hamburg_regridding.sh <control_file_name>
```

```
11.1b_Hamburg_regridding_1km.sh <control_file_name>
```

11.2. Gap-fill porosity using Hamburg data

The porosity data is gap-filled with Hamburg data, and masked using chess data.

There are two versions: one for the 12.5km and the other for the 1km grids (these are shown in parenthesis).

Use code:

11.2_gap-fill_Toth_et_al_porosity_with_Hamburg.py *(**11.2b_gap-fill_Toth_et_al_porosity_with_Hamburg_1km.py** for the 1km version)*

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***11.2_CONTROL_FILE_HWSD.txt*** *(**11.2b_CONTROL_FILE_HWSD.txt** for the 1km version)*

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to 12.5km (1km) CHESS file
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/

# path to folder with HWSD data
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/HWSD/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***11.2_CONTROL_FILE_HWSD.txt*** *(**11.2b_CONTROL_FILE_HWSD.txt** for the 1km version)*

USAGE:

```
python 11.2_gap-fill_Toth_et_al_porosity_with_Hamburg.py
<control_file_name>
```

```
python 11.2b_gap-fill_Toth_et_al_porosity_with_Hamburg_1km.py
<control_file_name>
```

12)Convert degree of saturation into VWC

There are two versions: one for the 12.5km and the other for the 1km grids (these are shown in parenthesis).

Use code:

12_convert_to_VWC_porosity_Toth_et_al.py
*(**12b_convert_to_VWC_porosity_Toth_et_al_1km.py** for the 1km version)*

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***12_CONTROL_FILE_ASCAT.txt***
*(**12b_CONTROL_FILE_ASCAT.txt** for the 1km version)*

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to 12.5km (1km) CHESS file
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/

# path to folder with HWSD data
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/HWSD/

# path to folder with ASCAT data
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/gridded_data/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***12_CONTROL_FILE_ASCAT.txt*** *(**12b_CONTROL_FILE_ASCAT.txt** for the 1km version)*

USAGE:

```
python 12_convert_to_VWC_porosity_Toth_et_al.py <control_file_name>
```

```
python 12b_convert_to_VWC_porosity_Toth_et_al_1km.py <control_file_name>
```

5 Merging the Data: Triple Collocation technique

Before you can run the code to merge the data, you will need to install a series of python packages required, which are not already installed on UKCEH Linux system.

Type the following commands into your Linux terminal:

```
pip install pandas==0.16.2 --user
pip install setuptools==12.4 --user
pip install pygeogrids==0.2.6 --user
pip install pygeobase==0.3.18 --user
pip install pynetcf==0.1.17 --user
pip install pyscaffold==2.5.11 --user
pip install ascat==1.0 --user
pip install ismn==0.3 --user
pip install configparser==3.7.5 --user
pip install pykdtree --user
pip install pytesmo==0.7.1 --user
```

Once all the required packages are installed, you will find all the codes needed for the merging of the three datasets using triple collocation technique in the following folder:

/prj/hydrojules/users/<your_username>/soil_moisture/2_merging/

NOTE: the whole time series is used to perform the triple collocation analysis (TCA), derive temporal mean, calculate the weights and scaling factors. This means that if the dataset was updated in future, the overlapping period might differ slightly, as the added timesteps might affect the results of the TCA or mean calculation.

For future updates, the following approaches can be taken:

- Re-run everything and update the whole dataset (which will be slightly different from the previous version). This has the advantage of being more rigorous and consistent, but the disadvantage of ending up with multiple versions, which are slightly different.
- Append the new timesteps to the original dataset. The advantage is that we end up with a unique dataset rather than multiple versions. However, the dataset is not fully consistent, as the temporal mean and TCA will change when new timesteps are incorporated.

5.1.1 SMAP interpolation

All data is rescaled against SMAP data before merging. In order to be able to rescale the data everywhere (even where SMAP data is missing), the first step is to interpolate mean SMAP value everywhere.

1) Calculate mean SMAP

There are two versions: one for the 12.5km and the other for the 1km grids (these are shown in parenthesis).

Use code:

1.1_SMAP_mean.sh *(**1.1b_SMAP_mean_1km.sh** for the 1km version)*

This code needs cdo to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***1.1_CONTROL_FILE_Merge.txt*** *(**1.1b_CONTROL_FILE_Merge.txt** for the 1km version)*

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to the merged SMAP 12.5km netCDF file:
/prj/hydrojules/data/soil_moisture/preprocessed/smap/smap_merged/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***1.1_CONTROL_FILE_Merge.txt*** *(**1.1b_CONTROL_FILE_Merge.txt** for the 1km version)*

USAGE:

```
1.1_SMAP_mean.sh <control_file_name>
1.1b_SMAP_mean_1km.sh <control_file_name>
```

2) Interpolate mean SMAP

There are two versions: one for the 12.5km and the other for the 1km grids (these are shown in parenthesis). CHESS data here is only used for masking purposes.

Use code:

1.2_interpolate_smap.py *(**1.2b_interpolate_smap_1km.py** for the 1km version)*

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***1.2_CONTROL_FILE_Merge.txt*** *(**1.2b_CONTROL_FILE_Merge.txt** for the 1km version)*

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to 12.5km (or 1km) CHESS file
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/

# path to 12.5km (or 1km) SMAP folder
/prj/hydrojules/data/soil_moisture/preprocessed/smap/smap_merged/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***1.2_CONTROL_FILE_Merge.txt*** *(**1.2b_CONTROL_FILE_Merge.txt** for the 1km version)*

USAGE:

```
python 1.2_interpolate_smap.py <control_file_name>
```

```
python 1.2b_interpolate_smap_1km.py <control_file_name>
```

5.1.2 CHESS scaling factor (b2) interpolation and CHESS mean

Where all other merging options fail (see table 1 below, in section 5.1.4), rescaled CHESS data will be used to gap-fill. Therefore, the scaling factor for CHESS also needs interpolating.

1) Outputting CHESS scaling factor (b2)

There are two versions: one for the 12.5km and the other for the 1km grids (these are shown in parenthesis).

Use code:

2.1_scaling_factor_b2.py *(**2.1_scaling_factor_b2_1km.py** for the 1km version)*

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***2.1_CONTROL_FILE_Merge.txt*** *(**2.1b_CONTROL_FILE_Merge.txt** for the 1km version)*

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to folder with chess file:
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/

# path to folder with smap file:
```

```
/prj/hydrojules/data/soil_moisture/preprocessed/smap/smap_merged/

# path to folder with ascat file:
/prj/hydrojules/data/soil_moisture/preprocessed/ascat/h115/GB/gridded_data/

# path to folder where scaling factor b2 file should be created:
/prj/hydrojules/data/soil_moisture/merged/weights/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***2.1_CONTROL_FILE_Merge.txt*** *(****2.1b_CONTROL_FILE_Merge.txt*** *for the 1km version)*

USAGE:

```
python 2.1_scaling_factor_b2.py <control_file_name>
```

```
python 2.1b_scaling_factor_b2_1km.py <control_file_name>
```

2) Interpolating CHESS scaling factor (b2)

There are two versions: one for the 12.5km and the other for the 1km grids (these are shown in parenthesis).

Use code:

2.2_interpolate_b2.py *(****2.2_interpolate_b2_1km.py*** *for the 1km version)*

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***2.2_CONTROL_FILE_Merge.txt*** *(****2.2b_CONTROL_FILE_Merge.txt*** *for the 1km version)*

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to folder with chess file:
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/

# path to folder with the scaling factor b2:
/prj/hydrojules/data/soil_moisture/merged/weights/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***2.2_CONTROL_FILE_Merge.txt*** *(****2.2b_CONTROL_FILE_Merge.txt*** *for the 1km version)*

USAGE:

```
python 2.2_interpolate_b2.py <control_file_name>
```

```
python 2.2b_interpolate_b2_1km.py <control_file_name>
```

3) Calculate CHESS temporal mean

There are two versions: one for the 12.5km and the other for the 1km grids (these are shown in parenthesis).

Use code:

2.3_CHESS_mean.sh *(**2.3b_CHESS_mean_1km.sh** for the 1km version)*

This code needs cdo to be set up.

Before running the code, type the following command in your Linux command line:

```
setup cdo
```

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***2.3_CONTROL_FILE_Merge.txt*** *(**2.3b_CONTROL_FILE_Merge.txt** for the 1km version)*

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# Path to the merged CHESS 12.5km netCDF file:
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/
```

The path must be on the second line of the control file. If you think you have accidentally changed the configuration of the control file, you can make a new copy from the original ***2.3_CONTROL_FILE_Merge.txt*** *(**2.3b_CONTROL_FILE_Merge.txt** for the 1km version)*

USAGE:

```
2.3_CHESS_mean.sh <control_file_name>
2.3b_CHESS_mean_1km.sh <control_file_name>
```

5.1.3 Merging

This is the step in which the actual merging of the dataset is performed.

The reference dataset is SMAP.

There are two versions: one for the 12.5km and the other for the 1km grids (these are shown in parenthesis).

NOTE: The version for 1km takes a long time to run (around 24 hours). For future development, it might be worth considering to parallelise this step. Each pixel is being calculated separately, so this code could be optimized for parallel computing.

Use code:

3_Merge_ref_SMAP.py *(**3b_Merge_ref_SMAP_1km.py** for the 1km version)*

This code reads in a control file, which specifies the path to the input file.

An example of the control file is: ***3_CONTROL_FILE_Merge.txt*** *(**3b_CONTROL_FILE_Merge.txt** for the 1km version)*

Make a copy of this file with a name of your choice and edit it to define the variables. The file looks like this:

```
# path to folder with chess file:
/prj/hydrojules/data/soil_moisture/preprocessed/chess/chess_2d/merged/

# path to folder with the scaling factor b2:
/prj/hydrojules/data/soil_moisture/merged/weights/
```

It is important that you keep the empty lines and the lines with comments. If you accidentally delete one of them, you can make a new copy from the original ***3_CONTROL_FILE_Merge.txt*** *(**3b_CONTROL_FILE_Merge.txt** for the 1km version)*

USAGE:

```
python 3_Merge_ref_SMAP.py <control_file_name>
python 3b_Merge_ref_SMAP_1km.py <control_file_name>
```

5.1.4 Map of weights

Triple collocation is a weighted average. The following code produces map of the weights used in the merging.

[Coming soon…]

5.1.5 Merging method

Where triple collocation analysis (TCA) is considered unreliable, alternative merging methods are applied, following rules detailed in Peng et al. (2021), which are summarised in Table 1 below:

Table 1: *Merging scheme for non-collocated grids. X, Y, and Z refer to different soil moisture products.*

p-value < 0.05 (X--Y)	p-value < 0.05 (X--Z)	p-value < 0.05 (Y--Z)	Merging scheme
yes	yes	yes	TCA weighted mean(X, Y, Z)*
yes	yes	no	X
no	yes	yes	Z
yes	no	yes	Y
yes	no	no	Arithmetic mean (X, Y)
no	yes	no	Arithmetic mean (X, Z)
no	no	yes	Arithmetic mean (Y, Z)
no	no	no	Disregard

*where only two datasets are available, the least-squared-based weights were derived from the uncertainties and a weighted averaged between the two products used in the merging (Equations 11 and 12 from Yilmaz et al., 2012).

The following code produces a map detailing which method was used for the merging in each pixel.

[Coming soon…]

5.1.6 Arithmetic mean

In Peng et al. (2021), TCA merging is compared to a simple arithmetic mean. The following code calculates this mean.

[Coming soon…]

6 Formatting the Output Data

[Coming soon…]

7 Figure production *(optional)*

[Coming soon…]

This section lists the codes used to produce the figures for Peng et al. (2021) paper. This is not needed to produce the dataset.

All the code uses the 12.5km version of the dataset, and uses a shortened version (finishing on the 2017/10/12) to match the data available at the time the analysis was done.

8 Future work

The current workflow could be improved in multiple ways:

- Increase the level of automation
- Proper error handling and messaging
- Migrate code from python 2.7 to python 3 (this might be tricky as some of the required packages are only available on python 2.7 at the moment)
- Eliminate the need of editing code and automate production of control files.
- Implement version control system for better traceability, auditability and reproducibility
- Improve data pre-processing by reducing the number of re-formatting steps. In particular, redesign the pre-processing steps for CHESS data, especially the need to reproject the data in Windows using ArcGIS.
- Extend the codes to allow periodical updates of the dataset (e.g. monthly)
- Investigate if parallelisation would be worth implementing to speed up production, and if yes, develop code
- Implement on JASMIN (maybe not necessary as dataset is relatively small)
- NOTE: all the scripts have the chmod command implemented to set the read and write permission to group users (chmod 664) for any new file created, but this has not been properly tested.

BANGOR
UK Centre for Ecology & Hydrology
Environment Centre Wales
Deiniol Road
Bangor
Gwynedd
LL57 2UW
United Kingdom
T: +44 (0)1248 374500
F: +44 (0)1248 362133

EDINBURGH
UK Centre for Ecology & Hydrology
Bush Estate
Penicuik
Midlothian
EH26 0QB
United Kingdom
T: +44 (0)131 4454343
F: +44 (0)131 4453943

LANCASTER
UK Centre for Ecology & Hydrology
Lancaster Environment Centre
Library Avenue
Bailrigg
Lancaster
LA1 4AP
United Kingdom
T: +44 (0)1524 595800
F: +44 (0)1524 61536

WALLINGFORD (Headquarters)
UK Centre for Ecology & Hydrology
Maclean Building
Benson Lane
Crowmarsh Gifford
Wallingford
Oxfordshire
OX10 8BB
United Kingdom
T: +44 (0)1491 838800
F: +44 (0)1491 692424

enquiries@ceh.ac.uk

www.ceh.ac.uk