Documentation of Data Processing and Compilation for the NSSN FMC Monitoring Project

Author: Porntipa Poonpolsub
Date: 2025-06-06

The code repository used for the processing is https://github.com/DiaPorntipa/nssn-fmc-data-compilation#

Table of Contents

Code repository structure

- Data (containing all input data and downloaded remote data)
- Notebooks
    - in-situ_data_preparation
        - pcs.ipynb
        - phd.ipynb
        - fc.ipynb
    - remote_data_compilation
        - afdrs_dfmc.ipynb
        - barra2_soil_mois.ipynb
        - barra2_vpd.ipynb
        - eratos_sdi.ipynb
        - eratos_vpd.ipynb
        - sentinel2_dfmc.ipynb
        - silo_vpd.ipynb
    - topography
        - topography_calculation.ipynb
        - topography_insertion.ipynb
- Output (containing all outputs from scripts except from 'in-situ_data_preparation' outputs)
    - csv
    - geojson
    - tif
- Utils (containing custom helper python functions for the processing scripts and an Australia shape file for segmenting dem data)

Overall flow for compiling an in-situ / remote pair dataset

Below are the designed steps to produce an in-situ / remote pair dataset with this codebase.

1. Put the in-situ (.csv), remote (.csv), dem (.tif), and vegetation cover (.nc) data in the 'Data' directory.
2. Run one corresponding script in the 'Notebooks/in-situ_data_preparation' to clean and format the in-situ data.
3. Prepare slope, relief, and aspect .tif files of the bbox covering all in-situ sites using Australia dem data with topography_calculation.ipynb
4. Use topography_insertion.ipynb fill each observation of in-situ data with slop, aspect, relief, and vegetation cover.
5. Run one corresponding script in the 'Notebooks/remote_data_comilation' to map the remote observations with the in-situ data.
6. The output in-situ / remote pair dataset will be in the 'Output/csv' directory with the naming convention of {remote source}_{focused variable}_{in-situ source (with conditions, if any)}.

Important detail of each script

Scripts for cleaning and formatting in-situ data to the desired design (Directory: Notebooks/in-situ_data_preparation)

All the scripts clean and format in-situ data to have the following condition.

- Consistent column names
- Consistent datetime format
- NaN representation of invalid data
- No relative humidity exceeding 100
- No irregular observations (ex. Datetime comes before installation datetime)
- Clear presence of VPD, DFMC, or soil moisture values

pcs.ipynb

- DBSCAN is used to clustered inconsistent PCS stations' latlon. The parameter 'eps' (the maximum distance between two samples for one to be considered as in the neighbourhood of the other) and 'min_sample' used are 50 metres and 10 respectively. TODO: Ask Nick whether this number is too large.
- The mean location of each cluster is used as representative latlon.

Scripts for adding topography to all in-situ observations (Directory: Notebooks/topography)

topography_calculation.ipynb

- Australia DEM data is cropped to a desired bbox and saved. An intersection of the bbox and Australia shapefile is used for cropping.
- Negative DEM values is replaced with 0.0 and invalid values with NaNs.
- Slope and aspect .tif files of the bbox are generated from Gdal.
- Relief is calculated comparing to the median elevation within 5 km tiles. (Some relief values near the `bbox` border may be inaccurately filled for convenience in coastal calculations. This should not affect results as long as selected points are not near the `bbox` border.)

topography_insertion.ipynb

- This script can only be used to filled in vegetation cover value that are saved individually for each site. (At the moment, only PCS and Forestry Corp sites have their vegetation values saved separately. While the data for Nick's PhD sites are saved together as 'veg_cover_phd_rounded.nc'.)
- For slope, aspect, and relief, values of the nearest grid cells to the sites are used. Their grids follow the original Australia DEM's.
- For site observations that the corresponding vegetation cover data in DEA Fractional Cover product are NaNs, the closest non-NaN vegetation cover values in time within 2 months are used.
- The vegetation cover value that is over 100 is treated as 100.

Scripts for mapping remote observations to all in-situ observations (Directory: Notebooks/remote_data_compilation)

When mapping remote data to in-situ observations, if the remote data is available hourly, times of in-situ observations are rounded to the nearest hour and matched with data of the nearest remote product's grid cell.

Otherwise, when the remote date is daily or less frequent, an observation with minimum or maximum of certain variables of each day is selected to match with remote data. The first and last observation dates of each site are disregarded as there is no data for the whole day for those dates.

In addition, observations without in-situ or remote target variable of each dataset are removed.

VPD scripts

All VPD values are calculated from temperature and relative humidity (RH).

barra2_vpd.ipynb

- BARRA2 temperature ('tas') and RH ('hurs') data is instantaneous readings and available hourly.
- BARRA2 'tas' and 'hurs' data is downloaded and saved in 'Data/barra2'

eratos_vpd.ipynb

- ERATOS temperature and RH data is available hourly.

silo_vpd.ipynb

- SILO provides daily maximum and minimum temperatures, each with a corresponding RH. A SILO "day" starts at 9:00 am and ends at 8:59 am the next day. (For example, observations between 8:59 am on 1 April 2020 and 9:00 am on 2 April 2020 are recorded as 2 April 2020.)
- In-situ observations are compared to SILO data using the same daily time window. The in-situ values for daily minimum and maximum temperatures are matched with SILO values. If there are multiple in-situ records with the same min/max temperature, the one with the lowest RH is selected (TODO: to be confirmed).

Soil moisture scripts

barra2_soil_mois.ipynb

- AFDRS data is available hourly.
- BARRA2 'mrsos' data is downloaded and saved in 'Data/barra2'

eratos_sdi.ipynb

- ERATOS SDI is available daily.
- In-situ observations with the daily (UTC time) minimum soil moisture values are matched with ERATOS SDI observations. If there are multiple observations that has the daily minimum soil moisture values, the earliest observation is selected.

DFMC scripts

afdrs_dfmc.ipynb

- AFDRS data is available hourly.
- The remote DFMC value is calculated from AFDRS temperature and RH.

sentinel2-dfmc.ipynb

- Sentinel-2 data is available every 2 - 5 days with frequent cloud cover.
- In-situ observations with the daily (UTC time) minimum DFMC values are matched with nearest non-NaN (TODO: Confirm) Sentinel-2 observations in time. Select the earliest observations if there are multiple observations that has the daily minimum DFMC values.
- TODO:

Other important details

- A summary of all site data is available in 'Data/all_sites_latlon_dates.csv'.
- Australia DEM (Digital Elevation Model) data is stored in 'Data/aus_dem/Ancillary'.
- Detail of vegetation cover data
    - The vegetation cover values are from the DEA Fractional Cover product.
    - The data has been cloud-filtered and smoothed using a weighted rolling average of PV (photosynthetic vegetation) and NPV (non-photosynthetic vegetation).
    - All downloaded vegetation cover files have an incorrect crs (coordinate reference system). The correct CRS is EPSG:32754.
- When converting Canberra local time to UTC time. In-situ observations with ambiguous Canberra local timestamps (e.g., during the daylight-saving transition from AEDT to AEST) will be excluded from the output.
- The coding scripts have not been fully refactored due to time limitations.