

Overview of Data Cleaning Pipeline

Overview of Theory and Instructions

FLUXNET-CH4 V2.0: Towards a more global characterization of methane-emitting sites workshop
21st-23rd October 2024

Rosie Howard, Zoran Nesic*, Sara Knox, Paul Moore, June Skeeter*****

*University of British Columbia, Vancouver BC, Canada

**McMaster University, Hamilton ON, Canada

***NRCan, Edmonton AB, Canada

Outline

A. Overview

1. Motivation
2. Software Installation
3. Data Cleaning Principles
4. Project Directory Structure
5. Create Database
6. Create INI files to clean data including output for Ameriflux
7. Data Visualization (led by Paul Moore and Sara Knox)

B. Live DEMO with working example

Downloads:

1. [Sample dataset](#)
2. [Template INI/configuration files](#)

1. Motivation

- Eddy covariance data presents many challenges...
- Our approach standardizes and allows reproducibility for:
 - Data post-processing QCQA;
 - Gap-filling;
 - CO2 flux partitioning.
- Minimized errors enhance reliability;
- Overall, enables integration of diverse datasets into large (global) networks.

*Note: high frequency data processing was covered in an earlier session

1. Motivation

- Focus on *post-processing of 30-min frequency data* (not high-frequency; and other averaging periods will work).
- Pipeline users need *minimal coding skills*, only need to edit some input and configuration files.

Pipeline Documentation is on [EcoFlux Lab Website](https://ecoflux-lab.github.io/PipelineDocumentation/PipelineDocumentation.html):

<https://ecoflux-lab.github.io/PipelineDocumentation/PipelineDocumentation.html>

1. Motivation

- While there are many online resources that can help with all of these different challenges of eddy covariance measurements, here we focus on the approach that the EcoFlux Lab uses for flux data post-processing and QCQA.

Examples of additional resources:

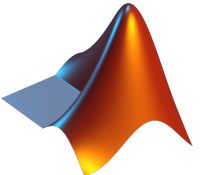


- [Flux Data Post-Processing and QA/QC](#)
- [Fluxcourse Educational Materials](#)
- [Videos on the FLUXNET website](#)

See [Documentation: Section 2](#)
and DEMO for details

2. Software

See also [Software Versions](#)

Required:

-  **Matlab**: pipeline scripts are in Matlab. For now, all users need Matlab installed (future versions will hopefully be open source).
-  **Biomet.net** library: download git repository and configure.
-  **R/RStudio** for third stage and visualization tools.

Optional:

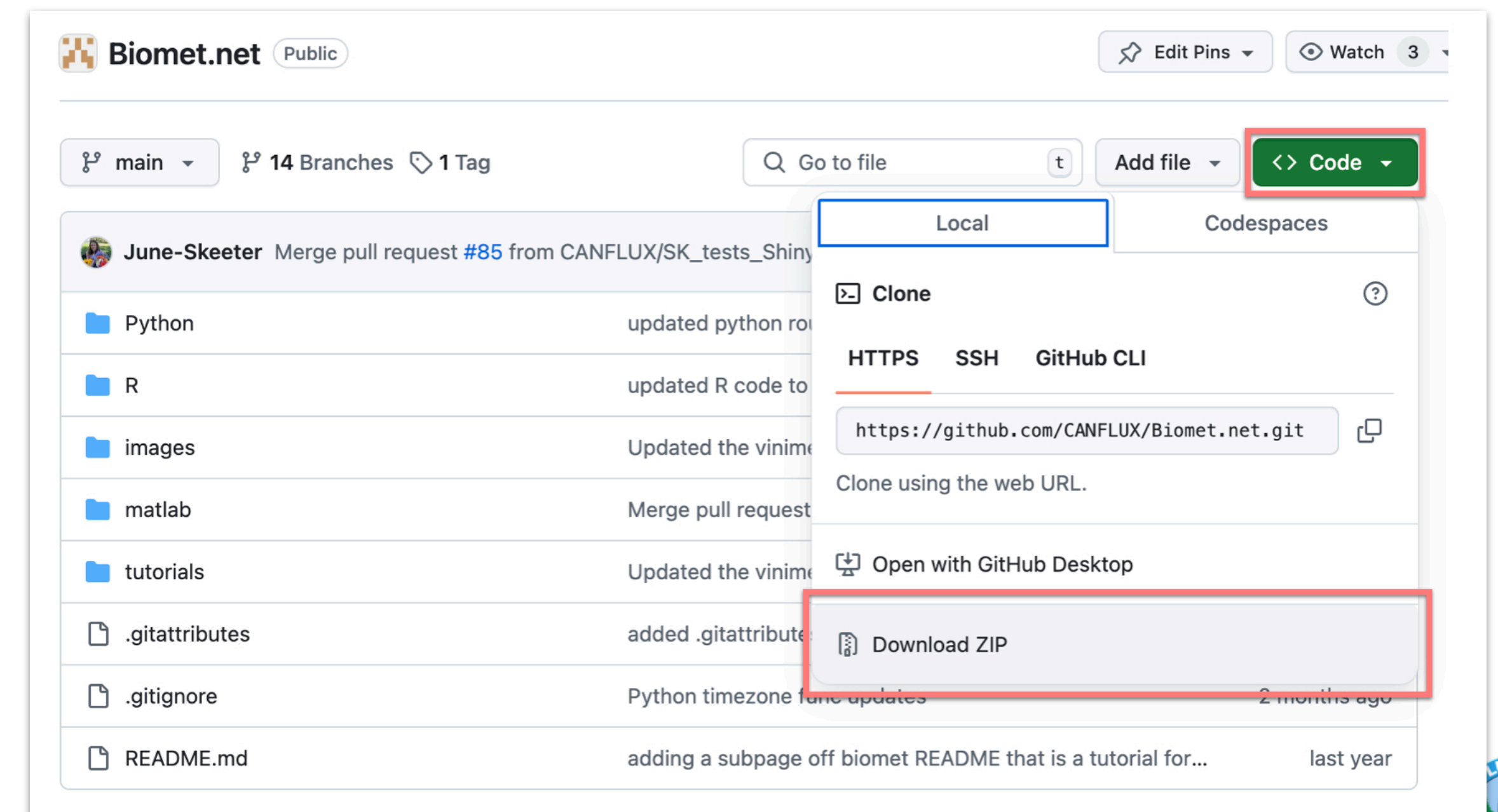
- Install Git/create GitHub account (for contributing to [Biomet.net](#)).
- Python (for [high frequency data processing](#) and [CH4 gap-filling](#)).

2. Software

Note on **Biomet.net** library: (1) download git repository and (2) configure:

(1) Go to <https://github.com/CANFLUX/Biomet.net>; click on “Code”, and “Download ZIP” file (or, use `git clone` command with HTTPS or SSH);

(2) will be covered later in the tutorial.



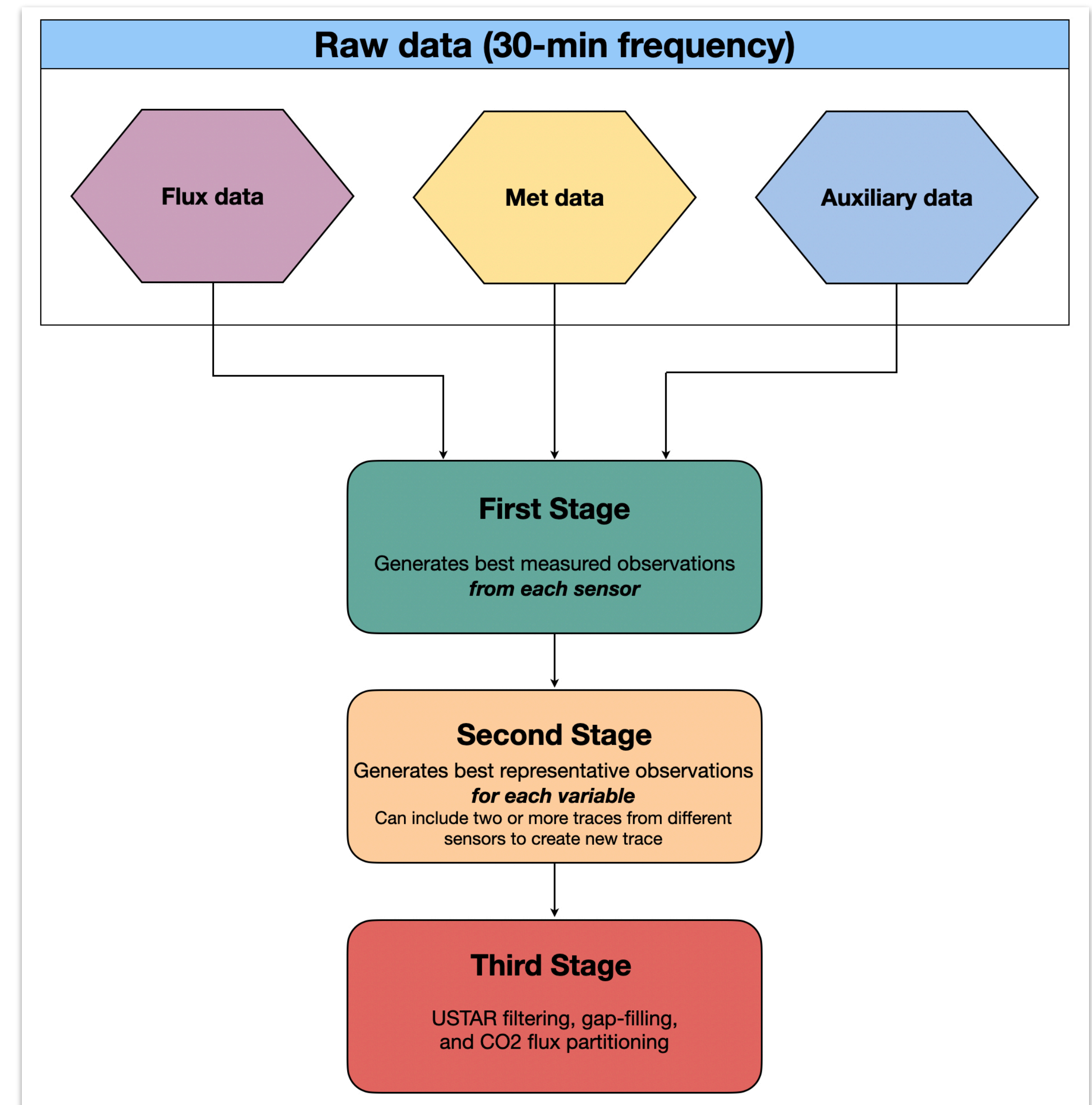
3. Data Cleaning Principles

Inputs:

- **30-min flux** data from, e.g., LiCor/CSAT3 system output by EddyPro;
- **30-min meteorological** data, e.g., output from Campbell Scientific (CS) system;
- Auxiliary data, e.g. metadata, field notes and comments.

Important:

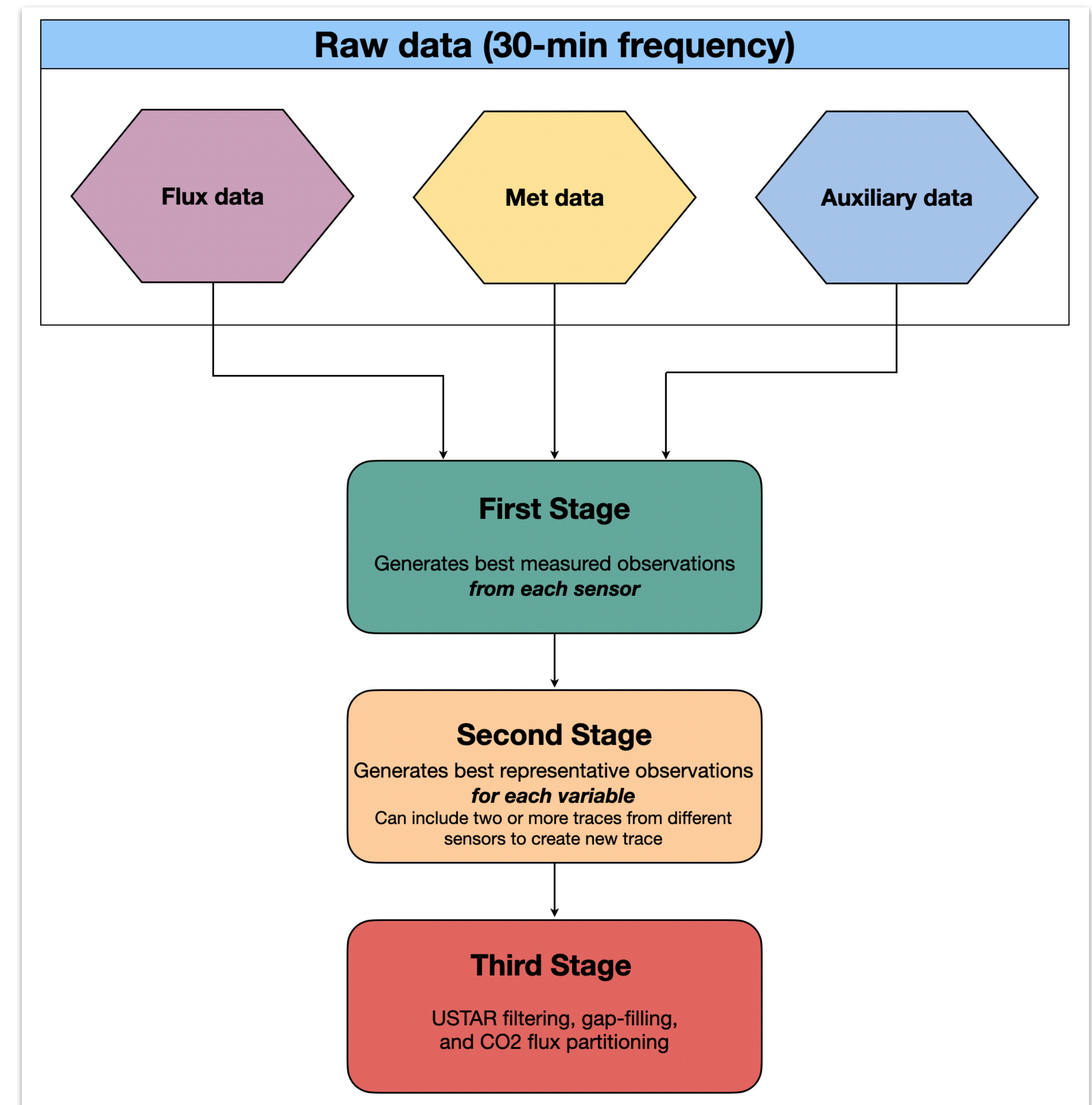
- Other averaging periods (e.g., 1 hour) also work as input for the pipeline;
- Any kind of data can be input;
- Our tutorial uses 30-min data, with EddyPro and CS output.



3. Data Cleaning Principles

First Stage:

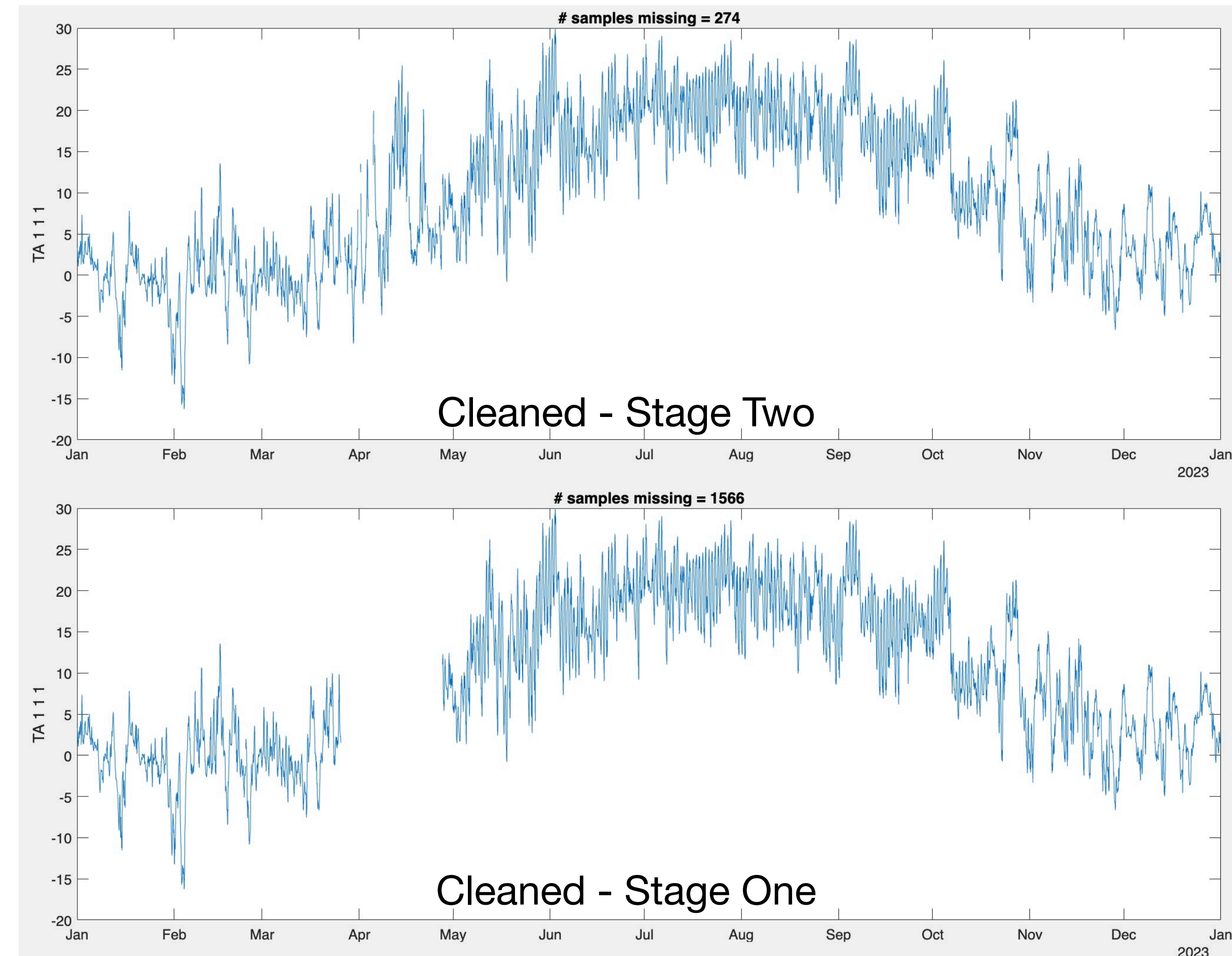
- Keep best *measured data from each sensor*
- Includes min/max filtering of raw data, calibration, spike-filtering
- Standardizes variable names in line with Ameriflux guidelines
- No gap-filling at this stage
- Prepares output for stage two, data saved as binary files (float32)



3. Data Cleaning Principles

Second Stage:

- Keep best *measured* data that *represents each property or variable*
- Can combine two or more (stage-one cleaned) variables to create new stage-two trace
- Averaging; gap-filling met data from nearby climate station/other measurement site
- Prepares output for stage three



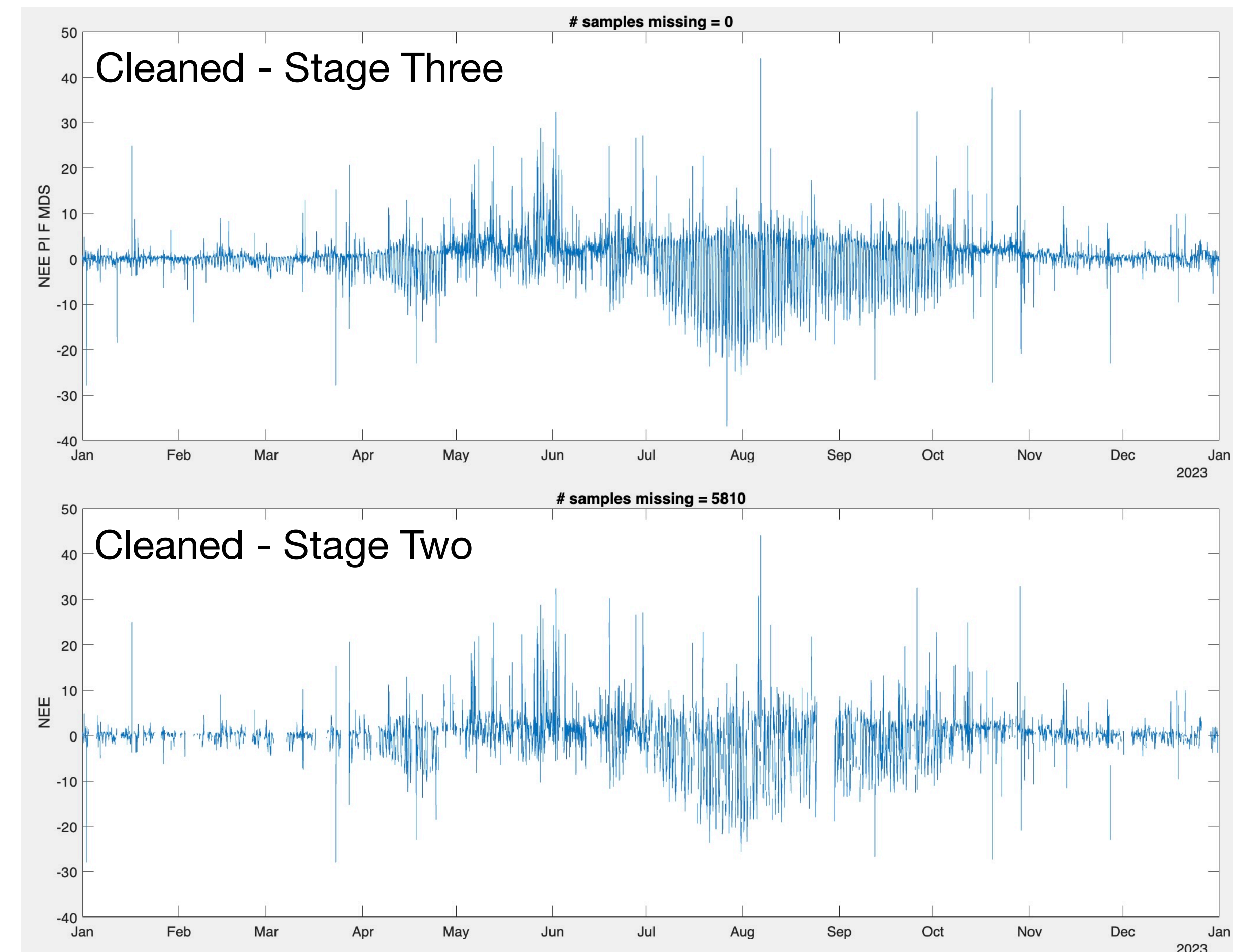
3. Data Cleaning Principles

Third Stage:

- USTAR filtering, gap-filling, and CO2 flux partitioning
- Uses REddyProc package (Wutzler et al. 2018) adapted for Matlab
- Gap-filling*: marginal distribution sampling (Reichstein et al. 2005), random forest approach (Kim et al. 2020)

Ameriflux Output:

- Relevant data is formatted as required for Ameriflux (CSV file)



*Wednesday's tutorial will go over how to methane gap-filling using machine learning, using the data you already formatted and cleaned

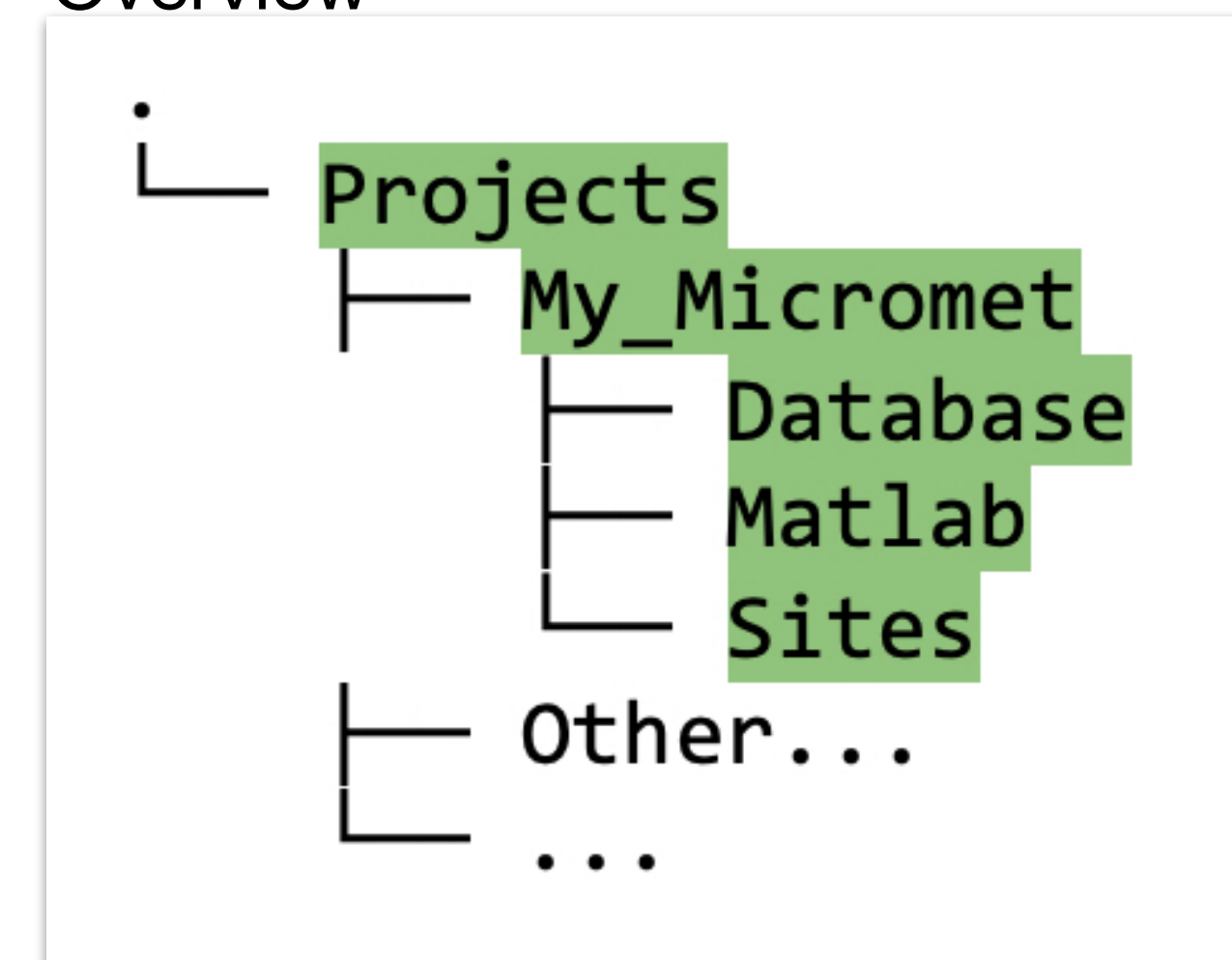
4. Project Directory Structure

- Main project folder: data cleaning, analysis, and research for group of similar flux sites
- Project name, e.g., My_Micromet
- Functions to create directory structure and configure it to work with Matlab are provided in Biomet.net library:
 1. `create_TAB_ProjectFolders`
 2. `set_TAB_project`

See [Quick Start: Section 4](#)
and DEMO for details

result →

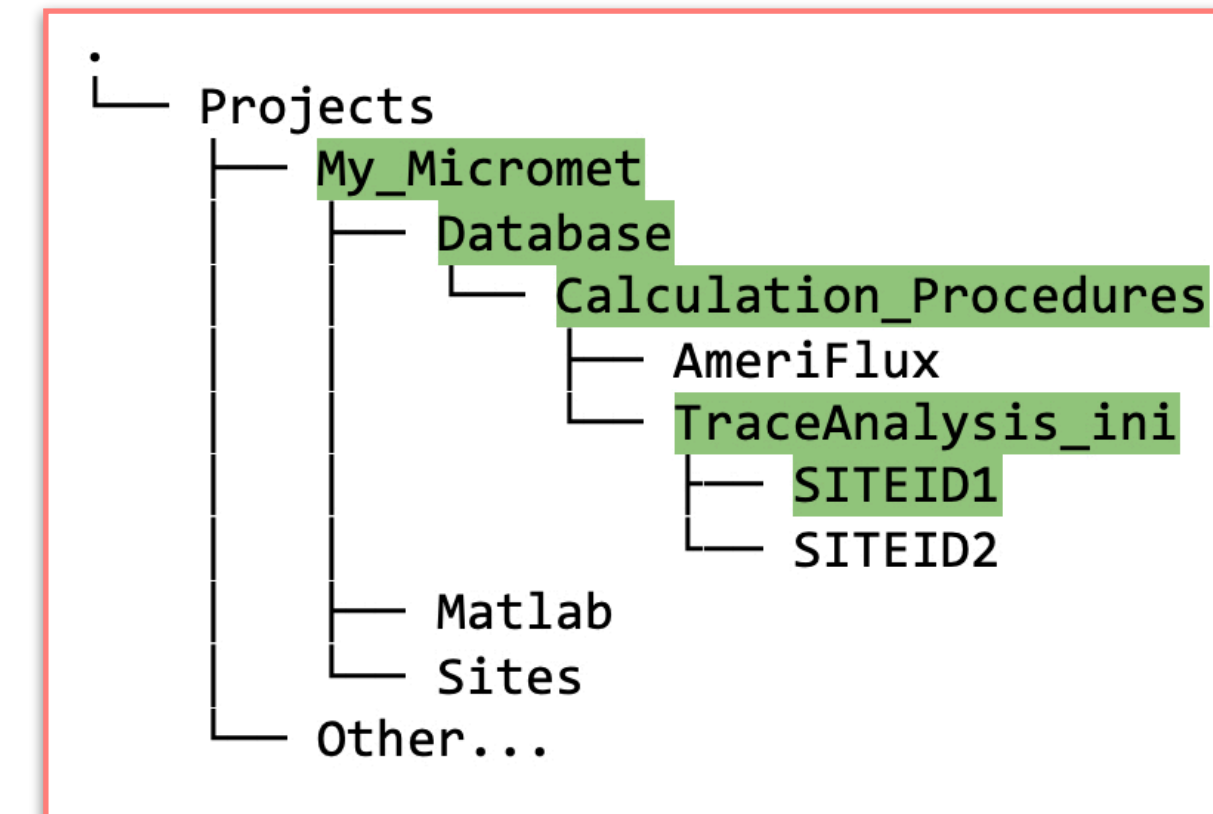
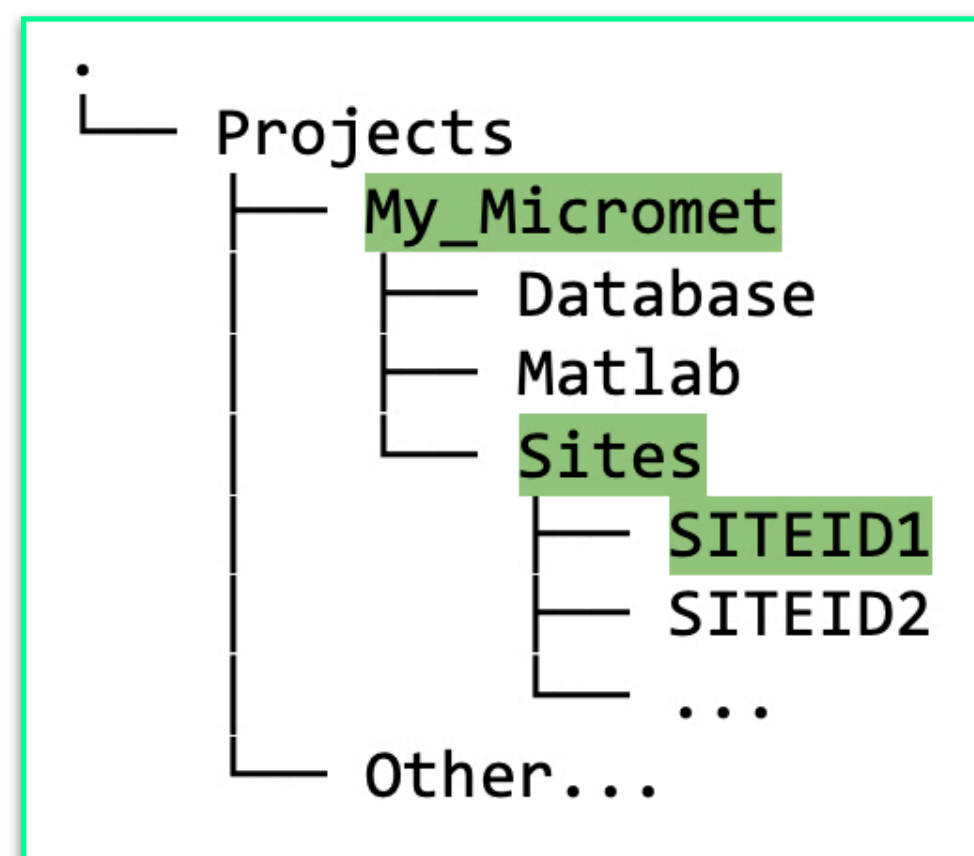
Overview



4. Project Directory Structure: Details

Sites:

- Raw, *uncleaned* data goes here



Database:

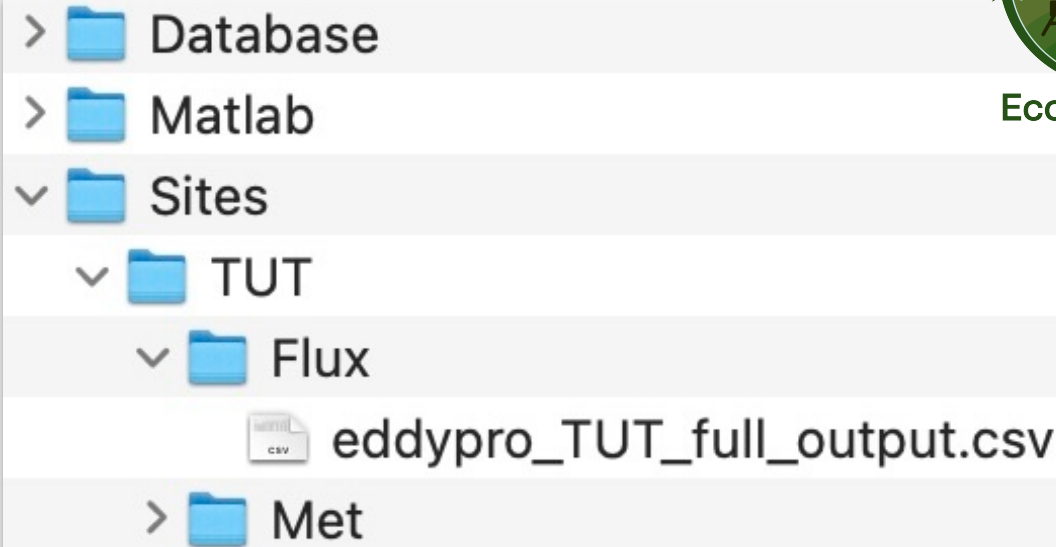
- Site-specific INI configuration files
- Initial database created from your raw data
- Cleaned data

Summary of Steps

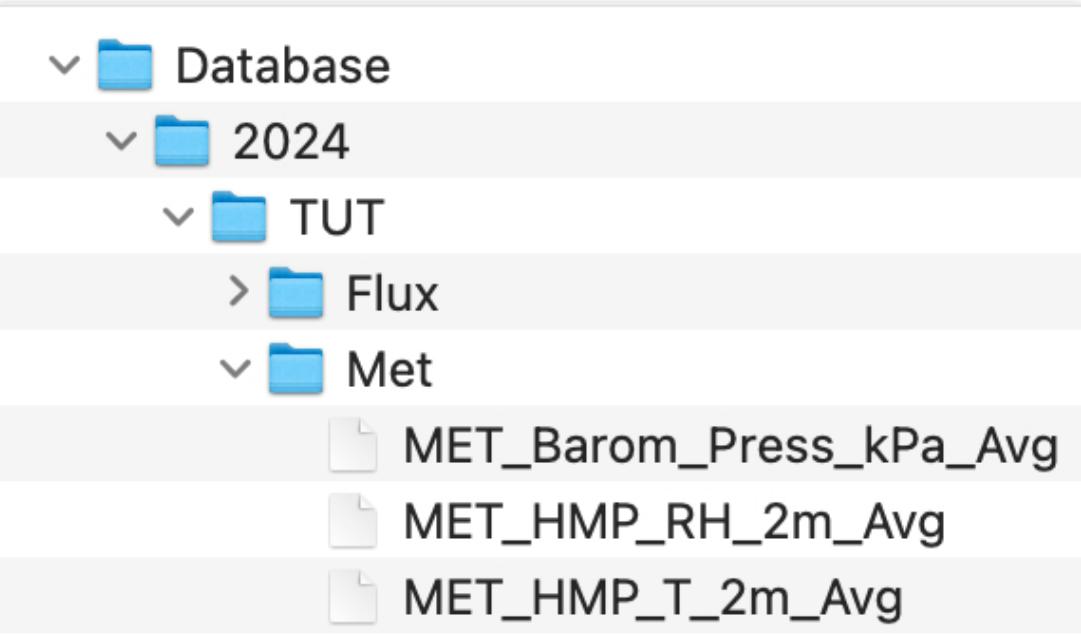
TOA5	TUT_CR1000	CR1000
TIMESTAMP	RECORD	MET_Barom_Press_kPa
TS	RN	kPa
		Avg
2024-07-16 1:30		1896
2024-07-16 2:00		1897
2024-07-16 2:30		1898
2024-07-16 3:00		1899

Raw data files in Sites folder

Sites



Database



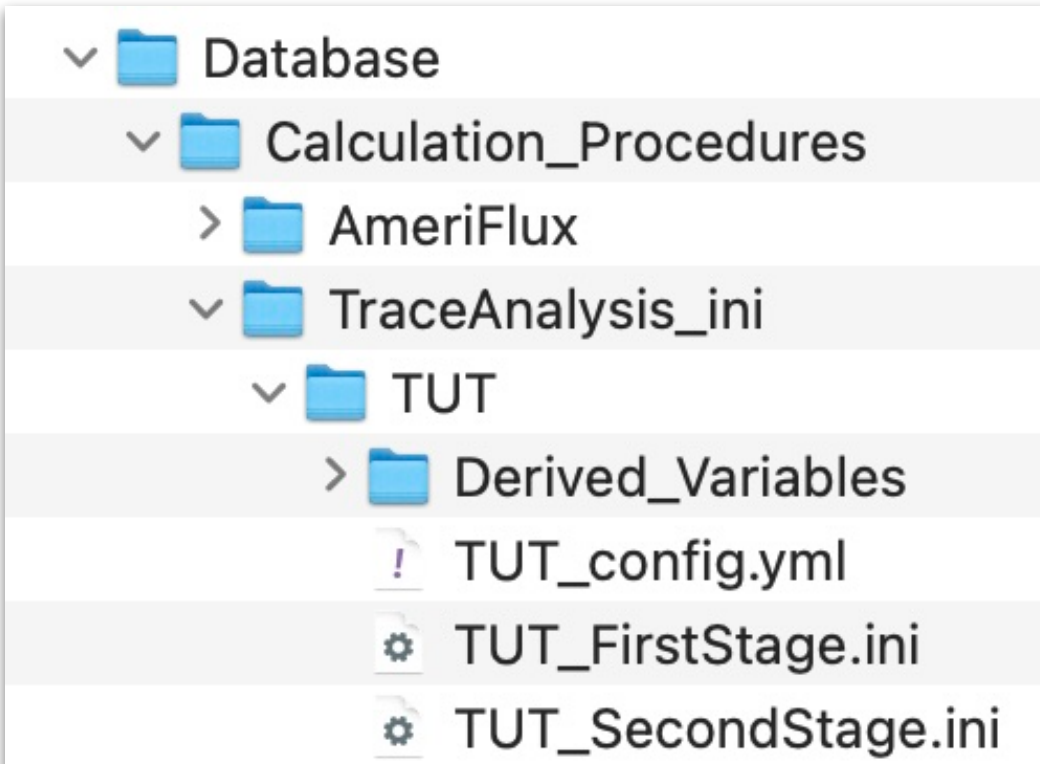
Edit and run Matlab script

projectPath
siteID
raw data filenames

Yellow highlighted code must be edited

```
%% Main function for MyMicrometSites data processing
% Created by <author> on <date>
%
% =====
% Setup the project and siteID
projectPath = '/Users/<username>/Project/My_MicrometSites/';
structProject=set_TAB_project(projectPath);
siteID = 'SITEID1';
```

INI files



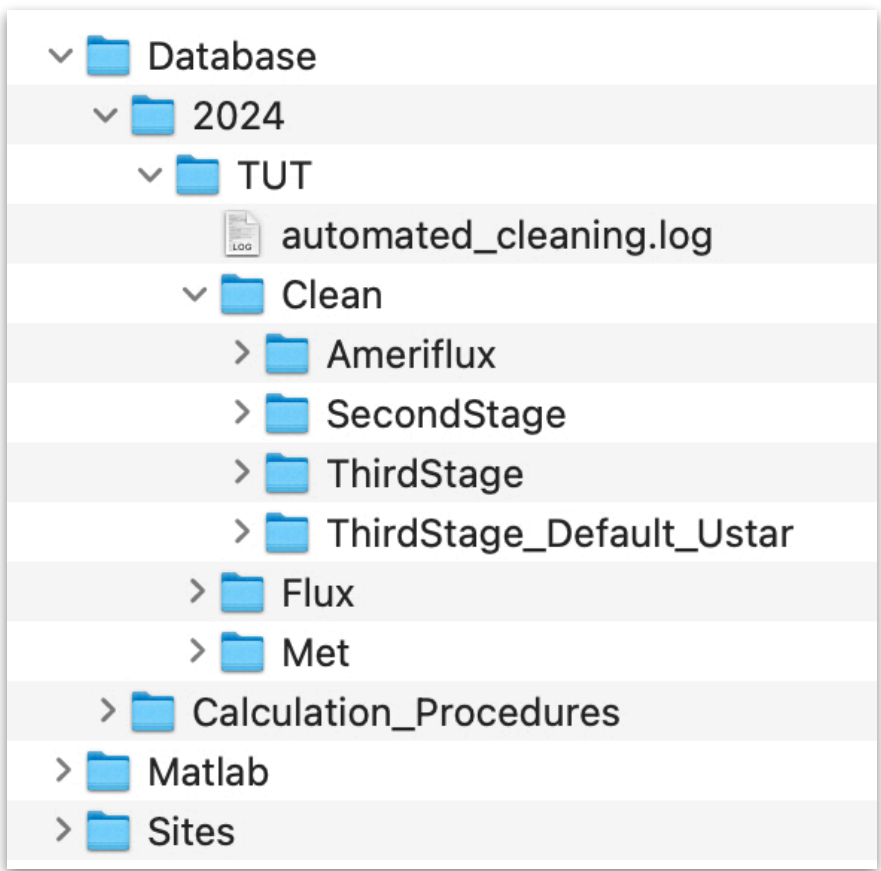
Rename and edit INI templates

site metadata
original variable names
related parameters...

```
[Trace]
variableName = 'TA_1_1_1'
title = 'Air temperature at 2m (HMP)'
inputFileName = {'MET_HMP_T_2m_Avg'}
inputFileName_dates = [datetime(1900,1,1) datetime(2999,12,31)]
measurementType = 'Met'
units = '°C'
instrument = 'HMP155A'
instrumentType = ''
```

Run one command for data cleaning

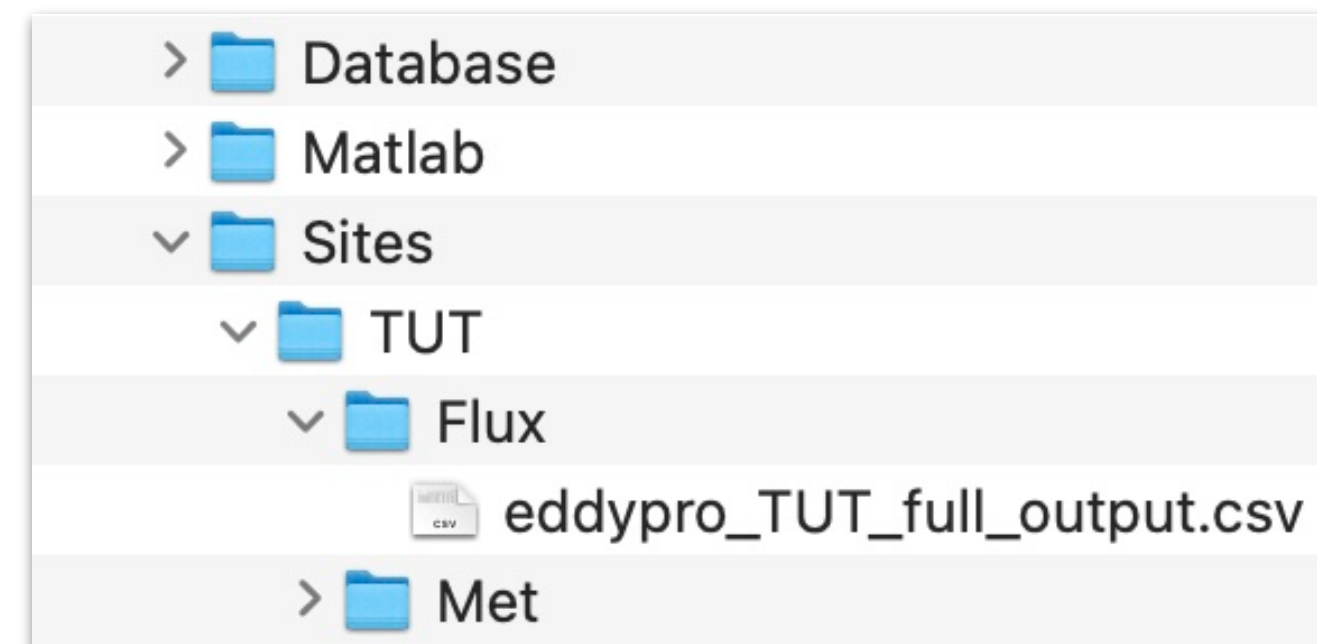
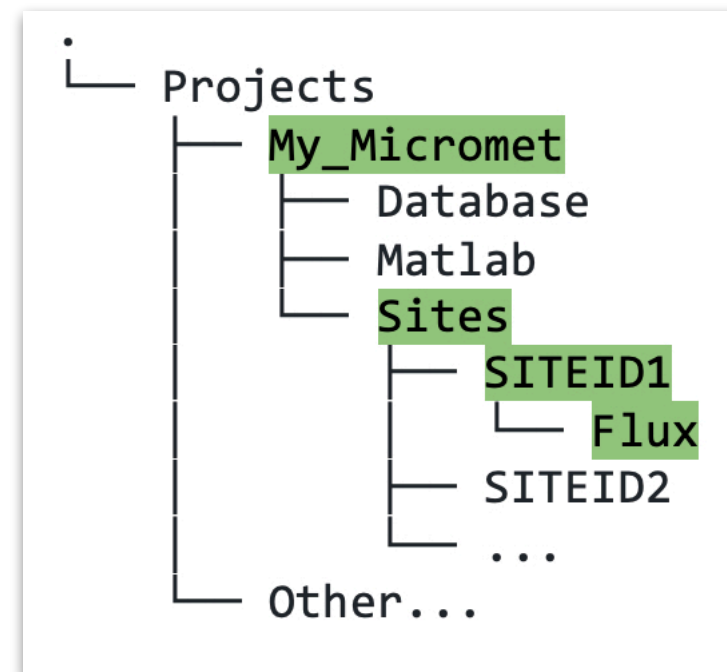
fr_automated_cleaning(YYYY, 'SITEID', 1)



Cleaned and standardized data ready for Ameriflux

5. Create Database

- Put your *raw* 30-min data in Sites folder, organized by site ID and then data type (Flux/Met)



Yellow highlighted code must be edited

```
%% Main function for MyMicrometSites data processing
% Created by <author> on <date>
%
% =====
% Setup the project and siteID
projectPath = '/Users/<username>/Project/My_MicrometSites/';
structProject=set_TAB_project(projectPath);
siteID = 'SITEID1';

% Create database from raw data
%% Flux data from EddyPro output files
%
% Input file name
fileName = fullfile(structProject.sitesPath,siteID,'Flux','MY_EDDYPRO_OUTPUT.csv');

% Read the file
optionsFileRead.flagFileType = 'fulloutput'; % select fulloutput, biomet, or summary
[~,~,tv,outStruct] = fr_read_EddyPro_file(fileName,[],[],optionsFileRead);

% set database path
databasePath = fullfile(db_pth_root,'yyy',siteID,'Flux');

% Convert outStruct into database
missingPointValue = NaN;
timeUnit= '30MIN';
structType = 1;
db_struct2database(outStruct,databasePath,0,[],timeUnit,missingPointValue,structType,1);

%% Met data from Campbell Scientific TOA5 output files
%
% Input file name
fileName = fullfile(structProject.sitesPath,siteID,'Met','MY_CS_TOA5_OUTPUT.csv');

% Read the file
[~,~,~,outStruct] = fr_read_TOA5_file(fileName);

% set database path
databasePath = fullfile(db_pth_root,'yyy',siteID,'Met');

% Convert outStruct into database
missingPointValue = NaN;
timeUnit= '30MIN';
structType = 1;
db_struct2database(outStruct,databasePath,0,[],timeUnit,missingPointValue,structType,1);
```

- Copy Matlab code to create *new* “main” script, within your Matlab folder:

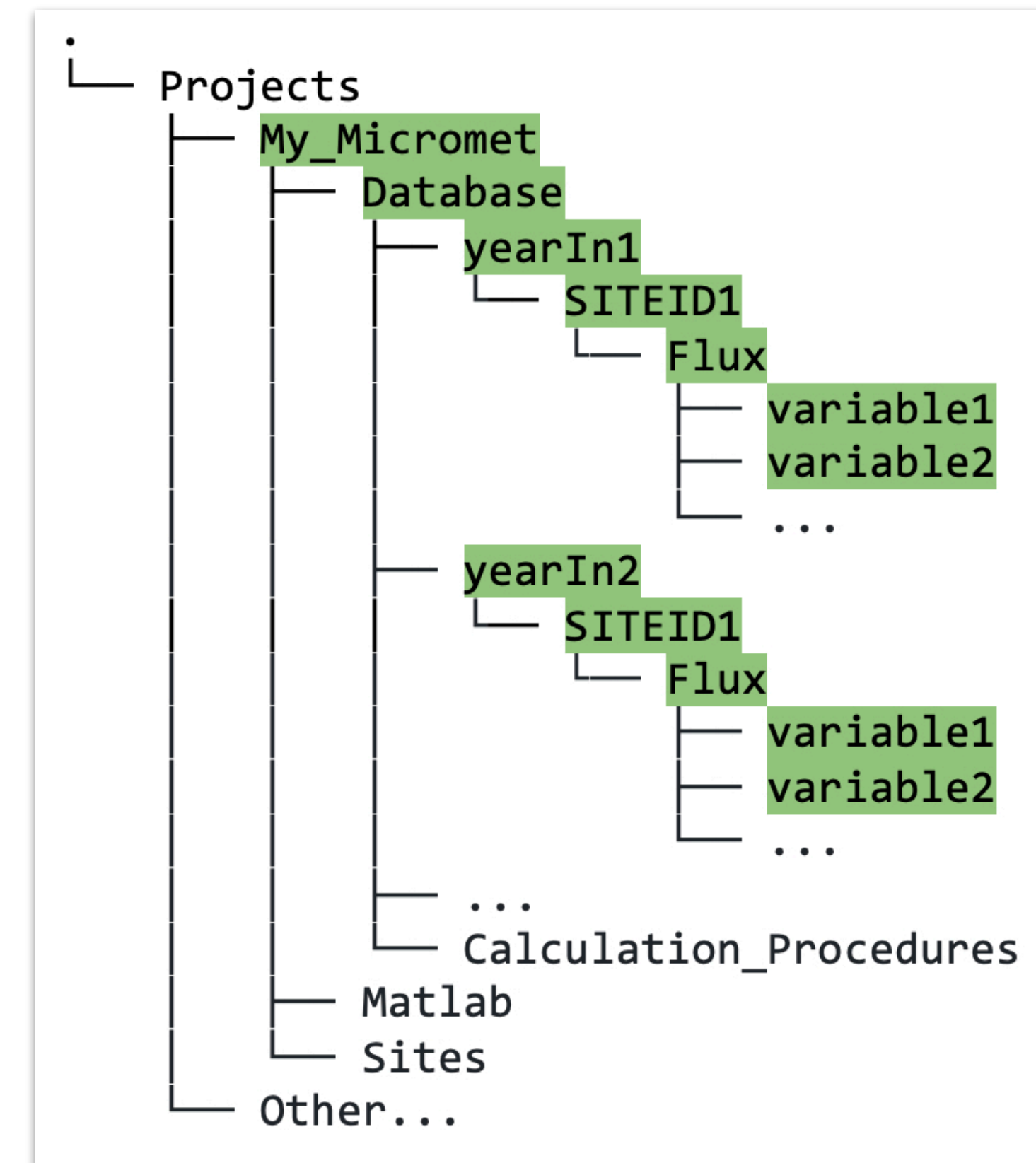
- Creates database
- Carries out all three cleaning stages (eventually)

See [Quick Start: Section 5.1](#) and DEMO for Matlab script and further details

5. Create Database and Inspect Contents

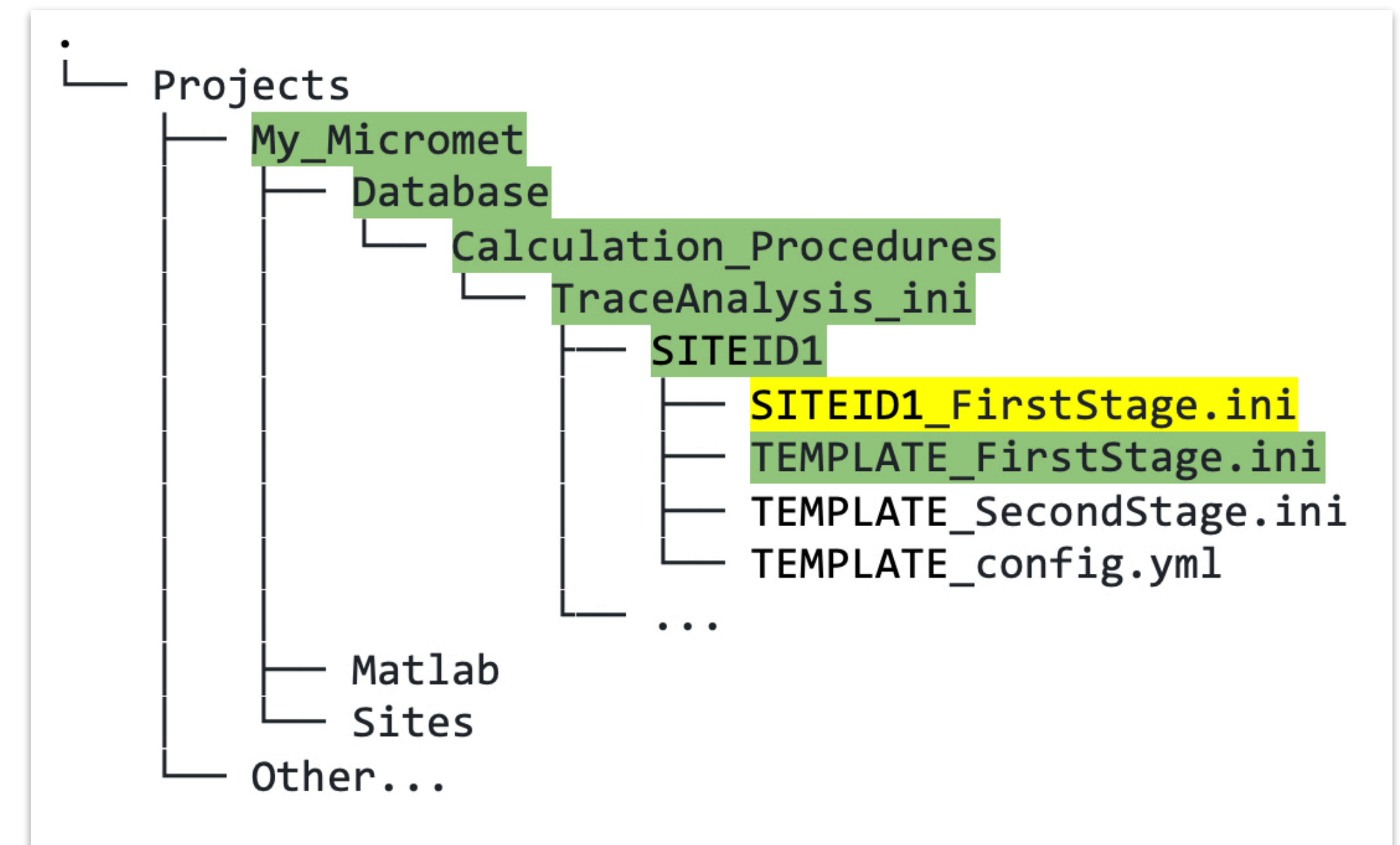
- Data output in your Database directory is grouped by year, then by site, then by data type
- We recommend inspecting your database at this point, and at every stage of cleaning, using visualization tools provided

See [Quick Start: Section 5.1](#) and [DEMO](#) for further details



6. Create INI Files to Configure Data Cleaning

- Obtain TEMPLATE ini files and put them in Database site-specific folder
- Make copies and rename with SITEID



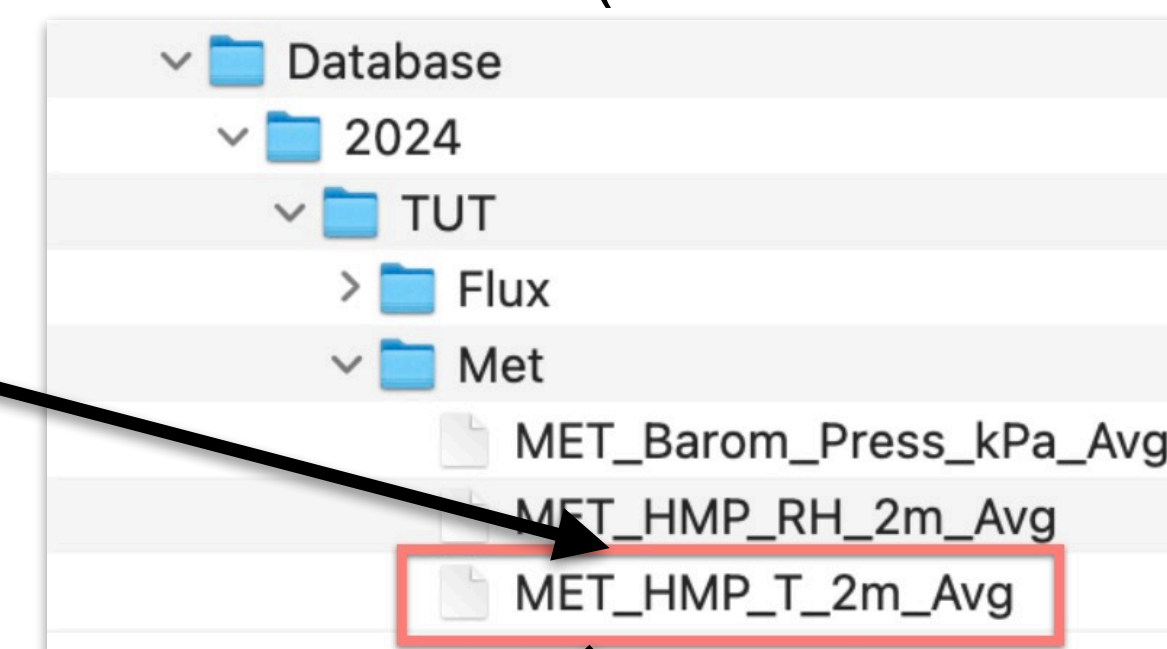
See [Quick Start: Section 6](#)
and [DEMO](#) for template files
and further details

6. Create INI Files: First Stage

Raw Met CSV file (Sites folder)

TOA5	TUT_CR1000	CR1000	87365	CR1000.Std.32.02	C
TIMESTAMP	RECORD	MET_Barom_Press_kPa_Avg	MET_HMP_T_2m_Avg	MET_HMP_RH_2m_Avg	M
TS	RN	kPa	Deg C	%	D
		Avg	Avg	Avg	A
2024-07-16 1:30	1896	101.1671	14.21723	92.39651	
2024-07-16 2:00	1897	101.1561	13.92406	93.8927	
2024-07-16 2:30	1898	101.17	13.64951	94.74481	
2024-07-16 3:00	1899	101.1677	13.85005	95.00518	

New database (Database folder)



A. *Input meteorological data
in SITEID1_FirstStage.ini*

SITEID1_FirstStage.ini

Ameriflux
format

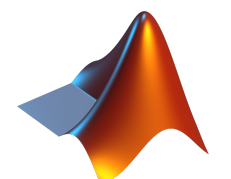
```
[Trace]
variableName = 'TA_1_1_1'
title = 'Air temperature at 2m (HMP)'
inputFileName = {'MET_HMP_T_2m_Avg'}
inputFileName_dates = [datetime(1900,1,1) datetime(2999,12,31)]
measurementType = 'Met'
units = '°C'
instrument = 'HMP155A'
instrumentType = ''
instrumentSN = 'N4520546'
loggedCalibration = []
currentCalibration = []
comments = ''
minMax = [-20,50]
zeroPt = [-9999]
dependent = ''

[End]
```

manually copy
variable name

6. Create INI Files: First Stage

B. Once you have added a few variables, run first stage cleaning in Matlab using Biomet.net function:

 `fr_automated_cleaning(YYYY, 'SITEID', 1)`

See [Quick Start: Section 6](#)
and DEMO for template files
and further details

C. Add “include” files — what are “include” files?

- INI files provided by us that already contain core information needed for relevant data variables from:
 1. Four-component radiometer, e.g., CNR4;
 2. EC system EddyPro output, e.g., from IRGA and CSAT3.

6. Create INI Files: First Stage

How do we use “include” files?

- Bottom of first stage INI, uncomment the relevant lines of code according to your own measurement system
- Remove “%” only! Leaving “#”
- Add your raw variable names, dates, ...

Near top of First Stage INI

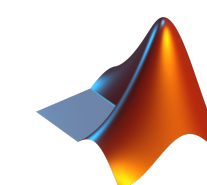
```
%-----
% Global variable specification (trace-specific)
%-----

globalVars.Trace.SW_IN_1_1_1.inputFileName = {'MET_CNR4_SWi_Avg'}
globalVars.Trace.SW_OUT_1_1_1.inputFileName = {'MET_CNR4_SWo_Avg'}
globalVars.Trace.LW_IN_1_1_1.inputFileName = {'MET_CNR4_LWi_Avg'}
globalVars.Trace.LW_OUT_1_1_1.inputFileName = {'MET_CNR4_LWo_Avg'}
```

Bottom of First Stage INI

```
%-----
% Call #include ini files
%-----
%--> Must be at end of .ini file
%--> Comment out include files that are not needed
#include EddyPro_Common_FirstStage_include.ini
#include EddyPro_LI7200_FirstStage_include.ini
#include EddyPro_LI7500_FirstStage_include.ini
#include EddyPro_LI7700_FirstStage_include.ini
#include RAD_FirstStage_include.ini
```

*After adding RAD include, run cleaning,
then after adding EddyPro include files,
again run first stage cleaning:*



```
fr_automated_cleaning(YYYY, 'SITEID', 1)
```

See [Quick Start: Section 6.1](#)
and [DEMO](#) for further details

See [Quick Start: Section 6.2](#)
and DEMO for further details

6. Create INI Files: Second Stage

Second Stage INI file

- Template provided by us (SITEID1_SecondStage.ini) already contains fundamental climate variables and minimum required variables to go on to run third stage

Features:

- Combine multiple traces into one using `calc_avg_trace` function
- Use `Evaluate` to operate on your traces:

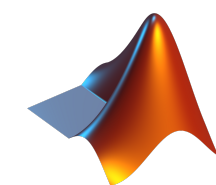
```
%-----  
% Main climate variables - temperature, pressure etc.  
%-----  
  
[Trace]  
    variableName    = 'TA_1_1_1'  
    Evaluate        = 'TA_1_1_1 = TA_1_1_1;'  
    title           = 'Air temperature at ...'  
    units           = '°C'  
  
[End]  
  
[Trace]  
    variableName    = 'RH_1_1_1'  
    Evaluate        = 'RH_1_1_1 = RH_1_1_1;'  
    title           = 'Relative humidity at ...'  
    units           = '%'  
  
[End]  
  
[Trace]  
.  
.
```

```
Evaluate = 'TA_1_1_1 = calc_avg_trace(clean_tv,TA_1_1_1,TA_OTHER_SOURCE,-1);'  
%(TA_OTHER_SOURCE variable must already be defined in First Stage INI)
```

See [Quick Start: Section 6.2](#)
and DEMO for further details

6. Create INI Files: Second Stage

Once happy with second stage INI file, you can run second stage cleaning using the Biomet.net function:

 `fr_automated_cleaning(YYYY, 'SITEID', 2)`

```
%[Trace]
    variableName      = 'TA_1_1_1'
    Evaluate          = 'TA_1_1_1 = calc_avg_trace(clean_tv,TA_1_1_1,TA_OTHER_SOURCE,-1);'
    % TA_OTHER_SOURCE variable must already be defined in First Stage INI
    title             = 'Air temperature at ...'
    units             = '°C'

[End]

[Trace]
    variableName      = 'RH_1_1_1'
    Evaluate          = 'RH_1_1_1 = RH_1_1_1;'
    title             = 'Relative humidity at ...'
    units             = '%'

[End]
```

6. Create INI Files: Third Stage

Third Stage configuration file

- Add site metadata to site-specific configuration file provided by us (SITEID1_config.yml)
- Most settings are in a global configuration file obtained during directory set up - do not edit!
- You can update/override settings if needed in your site-specific config file

```
# Written by June Skeeter
# March 2024
# Modified by <author>
# Date: <date>

Metadata:
  siteID: SITEID1
  estYear: <YYYY>
  lat: <latitude>           % North is positive
  long: <longitude>         % West is negative
  TimeZoneHour: <timezone> % (e.g., for Pacific standard time, GMT - 8)

# Optional parameters to modify default third stage parameters can be added here
```

See [Quick Start: Section 6.3](#)
and DEMO for further details

6. Create INI Files: Third Stage

Note on gap-filling CH₄ flux

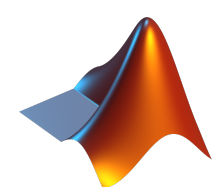
- Predictors for random forest models are all set to:

Predictors: SW_IN_1_1_1, TA_1_1_1, VPD_1_1_1

- For FCH₄, inputs should prioritize soil variables: soil temperature, soil moisture, water table depth
- Change these in the “Optional parameters” in your site-specific config file

```
# Optional parameters to modify default third stage parameters can be added here
Processing:
  ThirdStage:
    Storage:
      Apply_Correction: True
    REddyProc:
      Run: True
    RF_GapFilling:
      Run: True
    Models:
      FCH4_PI_F_RF:
        var_dep: FCH4
        Predictors: <add variables here>
```

Next, test third stage cleaning using the Biomet.net function:



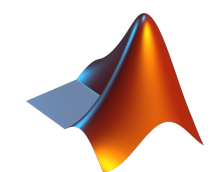
```
fr_automated_cleaning(YYYY, 'SITEID', 7) % note 7, not 3!
```

See [Quick Start: Section 6.3](#) and DEMO for further details

6. Create INI Files: Ameriflux Output and Running All Stages

For Ameriflux submission

- Reminder: *inspect your data at each stage* using the provided visualization tools
- Once happy with INI/config files and their output, convert the data to CSV file formatted for Ameriflux submission:



```
fr_automated_cleaning(YYYY, 'SITEID', 8)
```

- You can also run multiple cleaning stages at once:

```
fr_automated_cleaning(YYYY, 'SITEID', [1 2 7 8])
```

- And, multiple years and sites, for example:

```
fr_automated_cleaning(2019:2024, {'SITEID1', 'SITEID2'}, [1 2 7 8])
```

Running multiple years,
sites, and stages

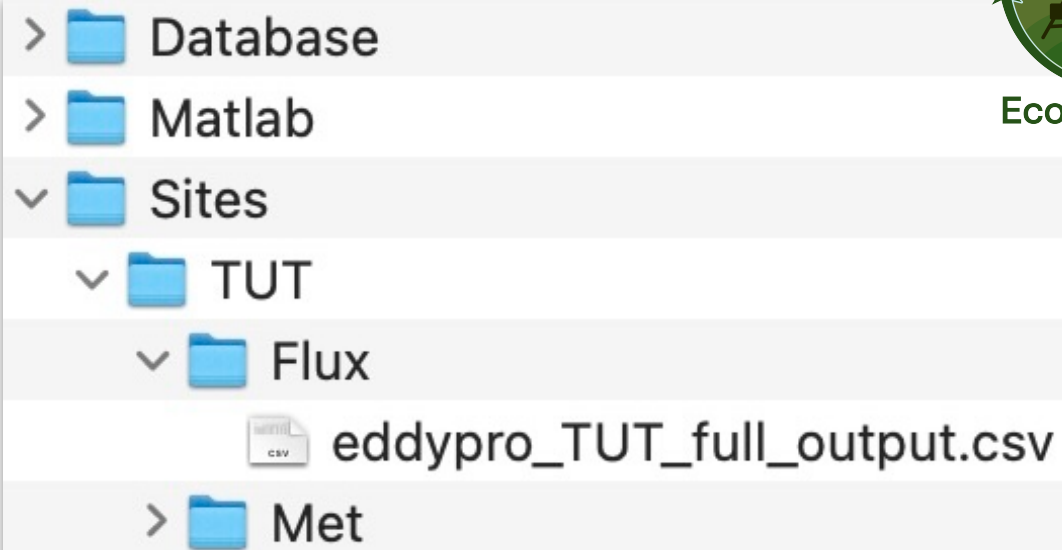
See [Quick Start: Section 6.3](#)

Summary of Steps

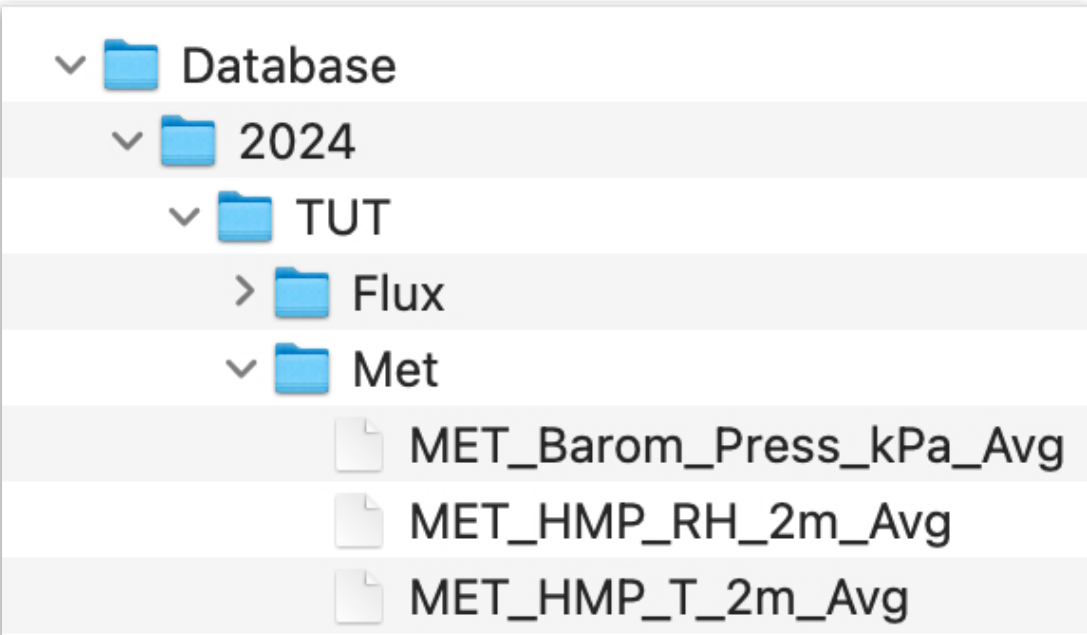
TOA5	TUT_CR1000	CR1000
TIMESTAMP	RECORD	MET_Barom_Press_kPa
TS	RN	kPa
		Avg
2024-07-16 1:30		1896
2024-07-16 2:00		1897
2024-07-16 2:30		1898
2024-07-16 3:00		1899

Raw data files in Sites folder

Sites



Database



Edit and run Matlab script

projectPath
siteID
raw data filenames

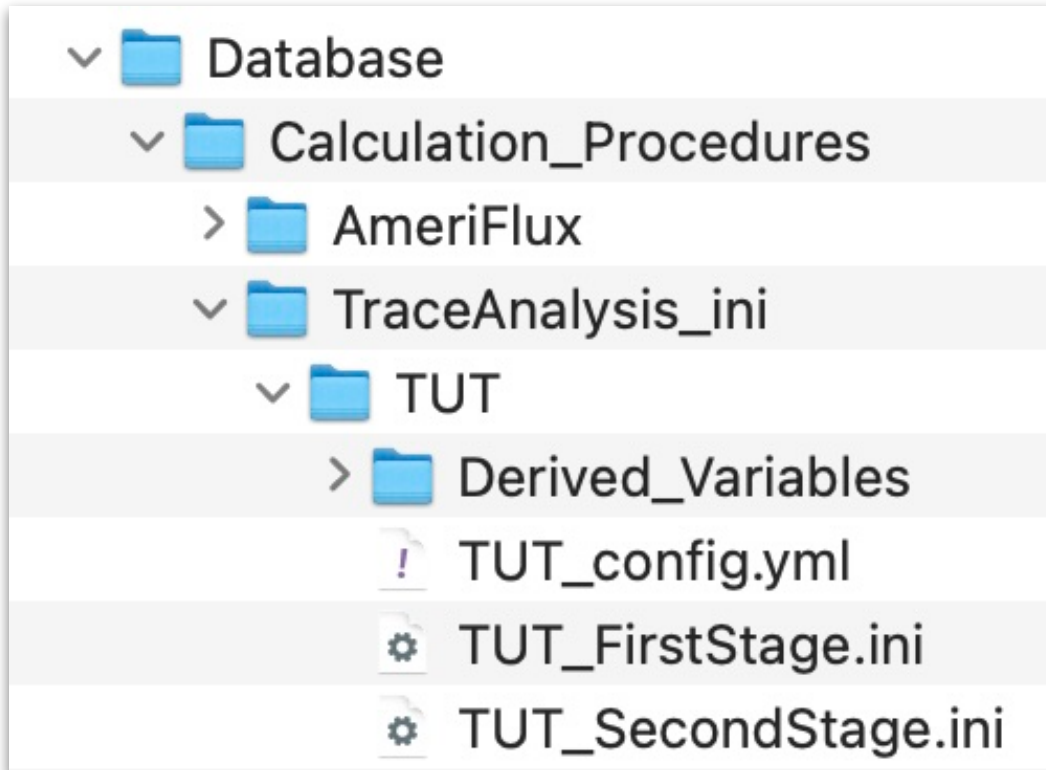
Yellow highlighted code must be edited

```
%% Main function for MyMicrometSites data processing
% Created by <author> on <date>
%
% =====
% Setup the project and siteID
projectPath = '/Users/<username>/Project/My_MicrometSites/';
structProject=set_TAB_project(projectPath);
siteID = 'SITEID1';
```

Rename and edit INI templates

site metadata
original variable names
related parameters...

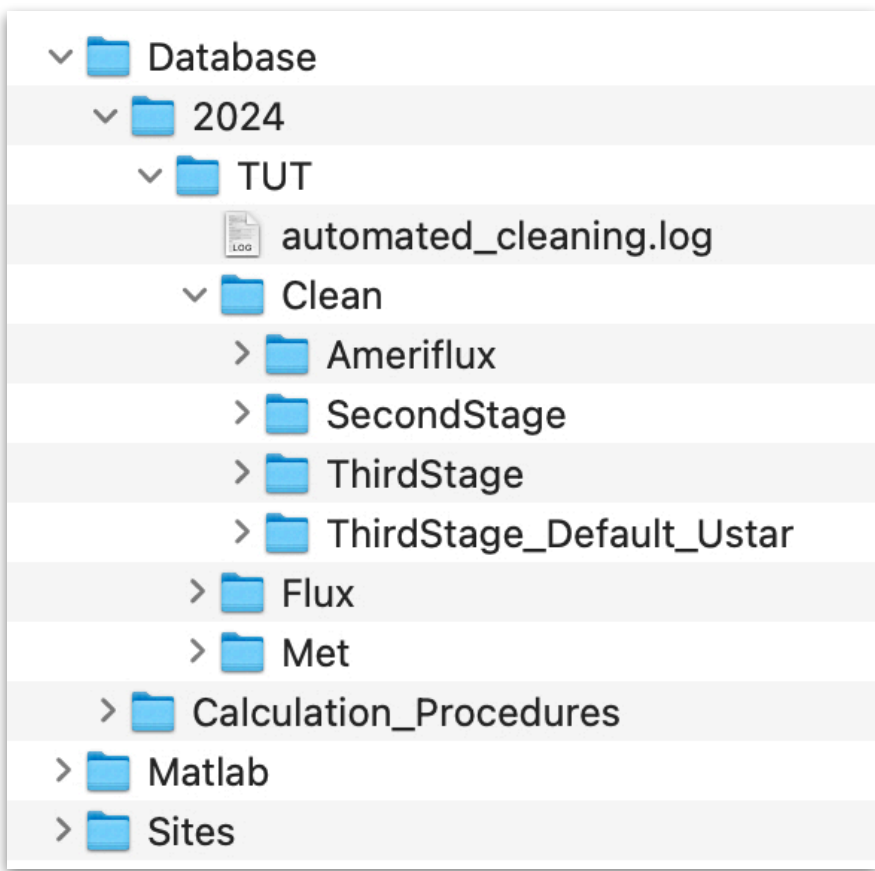
INI files



*Run one command
for data cleaning*


```
fr_automated_cleaning(2019:2024,{ 'SITEID1' , 'SITEID2' }, [1 2 7 8])
```

**Cleaned and
standardized data
ready for Ameriflux**



7. Data Visualization...

 plotApp: led by Paul Moore

 RShiny app: led by Sara Knox

Next...

B. Live DEMO with working example

- Follow along using sample data**
- Follow along using your own data**