**Crucial Needs:**

1. **Average blanks**
2. **Uncertainty for aerosol sample types.**

**1 Overview**

Ice Nucleating Particle (INP) workflow.

**2 Definitions**

The **Pyce Tools** workflow has a very strict naming convention. For this reason, it’s important to first define some terms.

INP samples come in various **TYPES**, which can be collected from a range of **LOCATIONS,** and are subject to different **PROCESSES** before analysis with the LINDA instrument.

**TYPE:** *[aerosol, seawater, mq]* the category of the INP sample. This is most generally either seawater or aerosol. Note that if the sample is a blank, seawater types are defined as ‘mq’.

**LOCATION:** *[uway, wkbt\_sml, wkbt\_ssw, bubbler, coriolis] ]*where the sample was collected. For instance, seawater type samples can come from the boat’s underway or from workboat deployments. Workboat measurements are further specified as either wkbt\_sml or wbt\_ssw depending on where in the water column the sample is coming from. Aerosol type INP samples can come from a bubbler, coriolis, or ambient measurements. For blanks samples, the options are bubbler, Coriolis, mq, mq\_wboat. This is kind of a weird naming convention. Sorry.

**PROCESS:** *[uf, f]* a technique used on an INP sample to gather more information. Possible processes include HEATING the sample or FILTERING it. Others include H2O2 but we do not do that in these samples. The way the code is written currently, is that the process of either filtering or leaving samples unfiltered is described in the sample name, while heated and unheated is actually assumed to have been conducted during initial sample analysis. For this reason, only UF/F is specified in sample names.

These three terms defined above are used throughout the Pycetools workflow, which is described below.

**3 Workflow**

Throughout the workflow of an INP sample, several file types are made. They are termed raw, calculated, and cleaned. The analysis workflow and how each file type fits into it is generalized as follows:

1. An INP sample of a specific type is collected from a specific location.
2. The sample then undergoes whatever post-collection processes the experimenter decides.
3. The sample is analyzed with LINDA, and **raw** data files are saved to their appropriate folder with the appropriate naming convention.
4. Raw data files are then pre-processed into **calculated** report files. This step involves, among other things, subtraction of the blank (See Blank Correction section). Calculatedreport files are then saved to their appropriate folder.
5. Calculated Report files from the calculation step are then cleaned and saved as **cleaned** output files, ready for analysis. During this step, error bars can be calculated and saved in separate files as well.

The specifics of these steps are further described in the following sections.

**3.1 Saving raw data files**

It is imperative that raw files are saved in the correct location and with the correct names. The rules are described for the two sample types (i.e., aerosol or seawater)

Seawater Sample Type Naming Convention:

seawater\_[location]\_[process]\_[DDMMYY]\_[HHmm]

These files are saved into the following folder \\[PROJECT\_ROOT]\\data\\raw\\IN\\[SAMPLE\_TYPE]\\[FILE]

Aerosol Sample Type Naming Convention:

**IMPORTANT NOTE:** When I initially began creating this workflow, I named files with dayXX. This is bad and I should not have done it. For this reason, there’s a section of code that changes dayXX into specific dates and times. I will need to revamp the code in the aerosol section to directly collect date/time instead of referencing a day from a dictionary. Going forward, files should be saved as follows, or at least corrected in some way.

aerosol\_[location]\_[process]\_[size]\_[DDMMYYYY\_HHhmm]

These raw files are saved into the following folder \\[PROJECT\_ROOT]\\data\\raw\\IN\\[SAMPLE\_TYPE]\\[FILE]

**3.2 Creating calculated sample report files from raw data**

Calculated report files are created by processing the raw data using either calculate\_raw\_blank or calculate\_raw functions. Raw blank calculation is described in the Blank Correction section below. Here we describe the process of creating a calculated sample report file for raw data, assuming that calculated blank data files are ready for use.

Raw file is loaded. Metadata is calculated and listed for raw data source, type, location, process, sample source name, sample collection date, sample analysis date, number of tubes, ml per tube, issues, and sigma. Sigma is used for error bar calculation and should be left at 1.96 for confidence intervals of 95%.

If the sample is of aerosol type, additional metadata is calculated and used in INP calculation. These parameters are rinse volume, size, average flow, sample collection time (calculated from start and stoptime) and total sampled air volume (calculated from sample collection time and average flow).

Blank data file is loaded into the template file and used to subtract from raw data. Important Note: Right now, only a single blank file is used. It is necessary to add functionality to average all relevant blank files into one file which is then subtracted from the data.

Calculated report file is then saved to its appropriate location according to the following convention: \\[PROJECT\_ROOT]\\data\\interim\\IN\\calculated\\[SAMPLE\_TYPE]\\[SAMPLE\_LOCATION]\\[type]\_[location]\_[process]\_[date]\_[time]\_calculated.xlsx

**3.3 Blank Correction**

INP raw data files from the LINDA need to be blank corrected. To do this, blanks are collected, analyzed with LINDA, calculated, and finally subtracted from experiment raw data files. The following functions are used in the preparation of blank files.

**calculate\_raw\_blank** function to calculate raw data files of blanks. When conducting LINDA experiments, the name of the blank file should follow this naming template:

For seawater type samples:

[type]\_blank\_[process]\_[collection\_date]

For aerosol type samples:

[type]\_blank\_[process]\_[size]\_[collection\_date]

It is crucial that naming conventions are followed, so see specific function documentation for more details.

The calculated report file will include metadata on the following: raw data source, type, location, process, sample source name, sample collection date, sample analysis date, number of tubes, volume (ml) per tube.

If the sample is type aerosol, the metadata will also include rinse volume and size.

Calculated report file is then saved to the appropriate folder. See function documentation for specifics.

**IMPORTANT NOTE: As of 02/11/2020 there is not yet any functionality to average the various blank files. This is obviously something that needs to be implemented.**

**3.4 Cleaning**

Cleaning function loads and cleans the calculated report file into a format that’s easier to use in Python. Some of the steps include renaming column names, combining all data into a single file, and reshaping the dataframe. Ouput is a CSV where each row is an observation and each column is a temperature.

Seawater sample types will also have columns for datetime, time, process, type, location, and filtered/unfiltered indicator. The IN values are given in IN/mL of water.

Aerosol sample types from the bubbler or Coriolis will have date, hour, start\_date, stop\_date, size (not for Coriolis), process.

**IMPORTANT NOTE: aerosol sample files should be changed to have similar format as seawater sample files.**

Cleaned files are saved to \\[PROJECT\_ROOT]\\data\\interim\\IN\\calculated\\[SAMPLE\_TYPE]\\[location]+[start\_date]\_[end\_date].csv

**3.4.1 (sort of) Optional: Calculation of error bars (TO DO)**

Calculation of Wilson Errors (link to paper) is done using the calculated report file. Output goes to the same location as the cleaned file.

Right now the code works for seawater samples.