This document outlines the workflow of the AHI folder to go from NAPS raw pollutant data to the final AHI data (gap-filled and daily interpolation)

1 File Organization

We begin with an overview of the file organization. The folder contains seven sub-folders; getCatnaps, getMortMorb, getTemp, getAP and getAPInterp, Outputs and metadata. Description of the contents follows;

- **metadata**: This folder contains all the metadata files used including; mapping from NAPS station to census division (`AHTI CD napsid.xlsx` and `assign_to_cd.csv`) and the list of 53 census divisions used in the project (`List of 53 selected CDs_2021.xlsx`).
- **getAP**: The scripts in this folder manage the downloading of NAPS PM data and transformation from raw data to a large data frame of observations
  - Note: In November 2021, when downloading the large NAPS files, the connection to the website would close and data from the file ended up being skipped. As a workaround, each file was downloaded 3 times, differences in the files were compared and checks were used to ensure the data was valid and complete. As of January 2022, the website is working, and files were able to be downloaded without the workaround.
- **getCatnaps**: This folder contains all the code to run catnaps (gap-fill or 10 hr interpolate the naps data). Code was obtained from Wesley Burr with few modifications (takes approx 14 hr to run with 53 CD, if I remember correctly. I recommend submitting the script as a job). Catnaps preforms a series of checks on the data, outlined in Section 3.2 of the interim report. Relevant outputs are located in `/AHI/getCatnaps/Hourly_Out/` and consists of pollutant specific csv files named as; `pollutant_original.csv` and `pollutant_screen_10hr_int.csv`
- **getAPInterp**: This folder is used for the post catnaps air pollution data interpolation. The scripts read in the catnaps output and organize the data into a list of census divisions, it then interpolates daily observations. The daily interpolation takes approximately 1 day to complete.

- **getTemp**: The script in the folder (`getTempDat.R`) accesses Environment Canada climate database to download temperature data from all weather stations in our list of CD's (`metadata/List of 53 selected CDs_2021.xlsx` obtained from Hwashin). Temperature data is deleted after downloading as it seemed to bog down the server (not sure why, but RStudio would stop responding when trying to access the Raw data folder).
  - The weather station to CD mapping was obtained from Hwashin (`getTemp/cd147_weather_stations.xlsx`)
- **getMortMorb**: This folder contains the script to organize mortality and morbidity csv files from Hwashin and James. A script reads the files and expects the following files (which have been removed for distribution);
  - `AHIdat_Morb.Circ_1996-2019.csv`
  - `AHIdat_Morb.Pulm_1996-2019.csv`
  - `AHIdat_Mort.Circ_1984-2015.csv`
  - `AHIdat_Mort.Pulm_1984-2015.csv`

  It then organizes the data into a list of CD's. There are 107 CD's and spans from 1984 – 2019. Note that the morbidity data starts in 1996 and the mortality data ends in 2015.
- **Outputs**: This folder contains the final AHItools data outputs, organized by a list of census divisions, including;
  - `AHIdat_ap.rda`: air pollution data, columns have prefix n (raw aggregated data), c (catnaps 10 hour interpolated data) and k (daily interpolated data)
  - `AHIdat_m.rda`: mortality and morbidity health data
  - `AHIdat_default.rda`: temperature data
  - `AHIdat_all.rda`: all above data combined

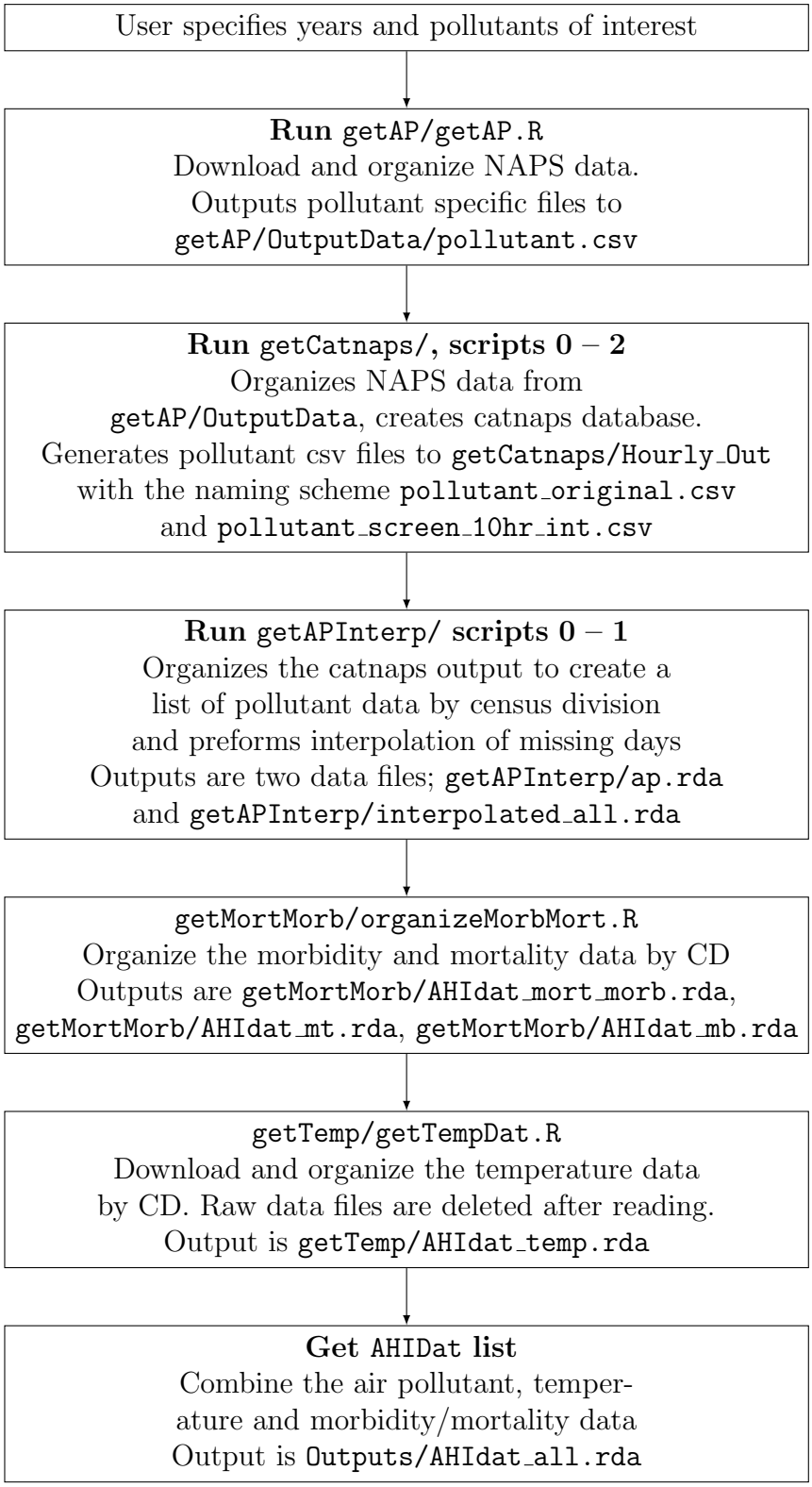
2 Description of the main file

The main file to preform the entire process is: `1-createAHIDat.R`. The start of the script contains parameters for the user to modify based on the time period and pollutant of interest. The parameters are;

- start_year: The first year of PM data to use. It was set to 1980
- end_year: The last year of PM data to use. It was set to 2019 (as of November 2021, the 2020 PM data has not been uploaded to EC NAPS database)
- types: A vector of the pollutant codes the user wishes to obtain. Possible pollutants are: PM10, O3, NO2, NO, SO2, PM25, CO and NOX
- dir: the file path of the AHI folder

2.1 Workflow

A diagram of the workflow of `AHI/1-createAHIDat.R` script is shown below;


User specifies years and pollutants of interest
Run getAP/getAP.R
Download and organize NAPS data.
Outputs pollutant specific files to
getAP/OutputData/pollutant.csv
Run getCatnaps/, scripts 0 – 2
Organizes NAPS data from
getAP/OutputData, creates catnaps database.
Generates pollutant csv files to getCatnaps/Hourly_Out
with the naming scheme pollutant_original.csv
and pollutant_screen_10hr_int.csv
Run getAPInterp/ scripts 0 – 1
Organizes the catnaps output to create a
list of pollutant data by census division
and preforms interpolation of missing days
Outputs are two data files; getAPInterp/ap.rda
and getAPInterp/interpolated_all.rda
getMortMorb/organizeMorbMort.R
Organize the morbidity and mortality data by CD
Outputs are getMortMorb/AHIdat_mort_morb.rda,
getMortMorb/AHIdat_mt.rda, getMortMorb/AHIdat_mb.rda
getTemp/getTempDat.R
Download and organize the temperature data
by CD. Raw data files are deleted after reading.
Output is getTemp/AHIdat_temp.rda
Get AHIDat list
Combine the air pollutant, temper-
ature and morbidity/mortality data
Output is Outputs/AHIdat_all.rda

3 Updating the AHItools and AHIdatm package

The purpose of the code is to update the data in the AHItools and AHIdatm package. Thus, after running the code and obtaining the output data (located in the Outputs folder), the user must update the corresponding data in the packages.

3.1 AHItools

The process is as follows;

- Copy the output data (Outputs/AHIdat_all.rda and Outputs/AHIdat_default.rda) to the data directory of AHItools package
- Compile the package;
  Navigate to the terminal with the working directory set to the location of the AHItools package and run the following commands, replacing the date and version as appropriate
  ```
  R CMD build AHItools
  R CMD INSTALL AHItools_YEAR.MONTH-VERSION.tar.gz
  ```
- Reload the AHItools library by entering the following line into the console;
  ```
  library(AHItools)
  ```

3.2 AHIdatm

The same process follows to update the AHIdatm package. This package is separate as it contains the sensitive morbidity and mortality data. The package should only be distributed to authorized persons. The following steps are used;

- Copy the output data (Outputs/AHIdat_m.rda, Outputs/AHIdat_mt.rda, Outputs/AHIdat_mb.rda) to the data directory of AHIdatm package
- Compile the package;
  Navigate to the terminal with the working directory set to the location of the AHIdatm package and run the following commands, replacing the date and version as appropriate;
  ```
  R CMD build AHIdatm
  R CMD INSTALL AHIdatm_YEAR.MONTH-VERSION.tar.gz
  ```

- Reload the `AHIdatm` library by entering the following line into the console;

```
library(AHIdatm)
```

4 Overview of steps and imputation method

The following provides an overview of the steps used to impute Environment Canada's Air Pollution Data.

1. After reading the raw data, replaced negative hourly air pollutant concentrations with NA

2. Ran CATNAPS screening and interpolation to obtain 10-hour gap-filled data for air pollutant metrics. The following screening steps were implemented and values failing the conditions were flagged:

   - Ten or more zeroes in a row
   - Baseline shift detected; shifts in daily minima concentrations
   - Truncated daily maxima values detected (saturation); more daily maxima concentrations than expected
   - More than 3 observations with the the same daily maxima concentration in a row
   - Outlier zeroes; zeroes that are more than 20 away from the daily median concentration

   After screening, gaps of 10 or more hours were interpolated. Negative concentrations were replaced with 0.

3. Aggregated hourly pollutant station measurements to daily census division (CD) metrics

4. Performed daily interpolation on air pollutant metrics. Replaced all negative concentrations with 0.