Overview: June 13th, 2023

1. New Data
    - [WIP] Soil moisture data
    - [WIP] Hourly weather station data
    - [WIP] Copernicus satellite data
    - [ADDED] Ergot data
    - [ADDED] Soil data
2. Documentation Updated
3. Environment Changes
4. Quality of Life Features Added to the Weather Station Pipeline
5. Model Research and Experimentation

New Data

Soil moisture data – *[Dharmit|Jay|Remmy]*

- Data pulled as averages for Manitoba, Alberta, Saskatchewan
- NetCDF

Hourly weather station data – *[Josef]*

- Parallelizing the same process – different attributes

Copernicus satellite data – *[Dane]*

- 280 requests total, 30 minutes each
- Google earth engine – problem after problem
- GRIB file reader – slow
    - Took 12 hour to pull an hour of data for one day (all attributes)
    - 12 hours x 24 different hours x 280 different year/months = long time
- Pooled NETCDF reader
    - 3 hours? Starting to use screens and need to slow down pull requests
    - Learnt more of the GeoPandas library

Ergot data – *[Dane]*

- Rough data validator – goal was that for any other CSV files could import similarly

Soil data – *[Dane|Dharmit]*

- Soil attributes

Documentation updated

Updated setting up the environment – *[Dane|Josef]*

Data Sources – *[Dane]*

- Make it easier to find data sources and automatic vs manual

Database Tables – *[Dane]*

- Brief data descriptions

Useful Links – *[Dane|Josef]*

- Manuals/download sites mostly

New environment setup

- Synchronizing containers with the workspace – *[Josef]*
    - Attaching to containers
- Migrating our database for larger database processes – *[Josef]*

Quality of life features added to the weather station pipeline

- Check which data is pulled and set stations to inactive – *[Dane]*
    - If active always pulls at minimum the last year but discards duplicates
    - Last updated is stored automatically
- Lower case names which means queries need less formatting – *[Dane]*
- Better update messages – *[Dane]*:

```
[205/292] Pulling data for station 3035206 between 1992-2011
          updated 141 rows
[206/292] Pulling data for station 3035208 between 1996-2022
          updated 671 rows
[207/292] Pulling data for station 3045100 between 2018-2022 …
```

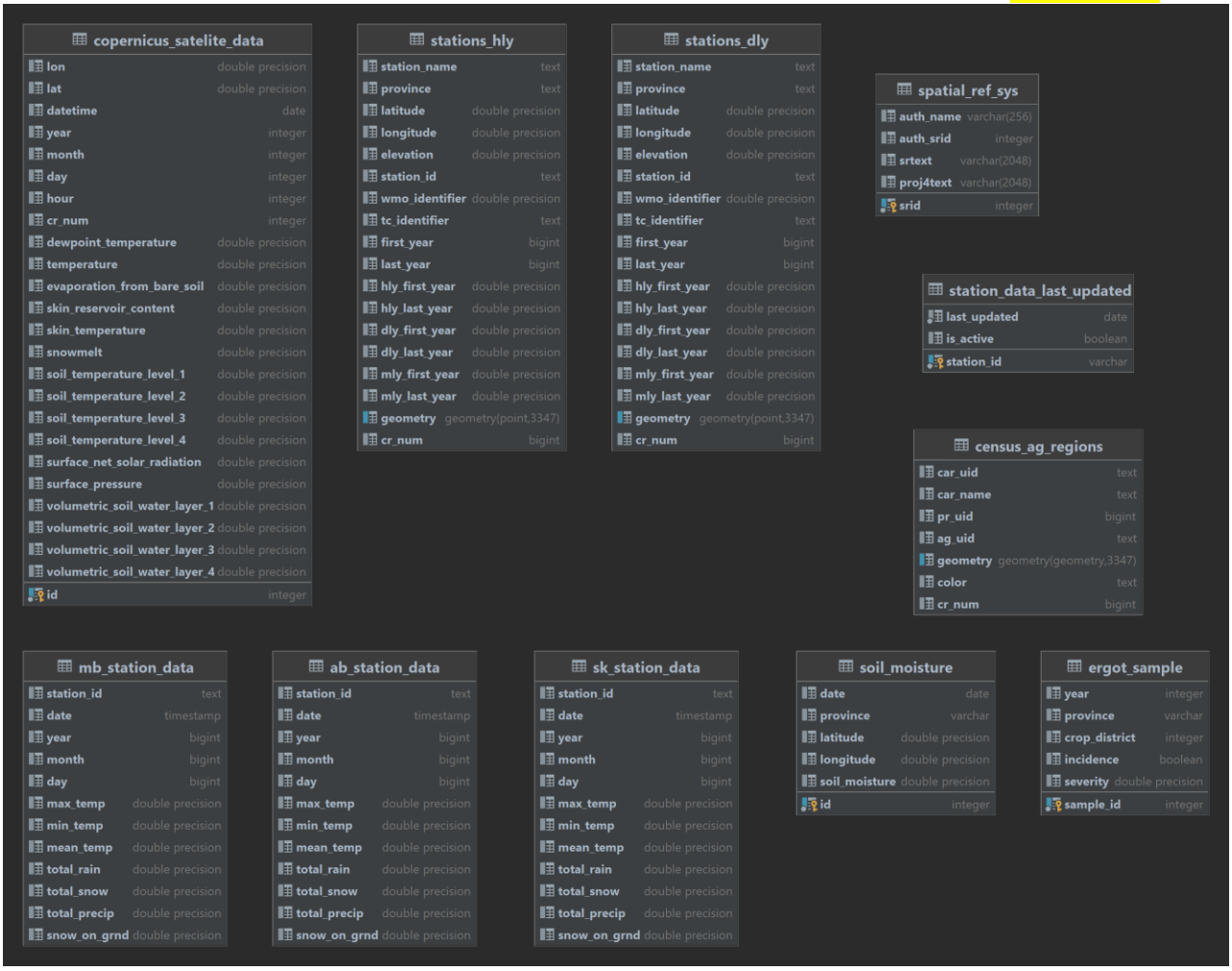
Model research and Experimentation

- Supervised learning models – *[Daniel]*:
    - Autoencoder
    - Lstm
    - Logistical regression

- Random forest
- Decision tree
- Gradient boost
- Svc
- Linear svc

Questions

- **Dane** – lots of time spent collecting data so far – moving towards processing but is there anything else of interest? As per last meeting:
    - ~~Soil Type Data~~ (not shown below)
    - ~~Soil PH~~ (not shown below)
    - ~~Ergot levels~~
    - ~~Soil temperature~~
    - **Growth session length?** – data source was not what I expected but we can probably infer this from weather conditions? **README**



- **Dane** – is a master pipeline valuable in the future? Could run one command and have all data updated: less granular, data validator, JSON configuration
    - Create folders
    - Download files and store them in the folders
    - Run it through the pipeline ft. data validator
    - Automatically move these folders to a used folder to indicate it's been added
    - Run all automated pipelines
    - Save last updated for all automated data sources

Goals for July 5th

Currently behind schedule by an entire block. Fortunately,/unfortunately, our next check in will not be for another 3 weeks – some time to catchup and condensed classes will be finished!

Josef

- Finish setting up the hourly stations
- Aggregate data
- Data visualization/exploratory analysis
- DNN model implementation
    - Cross the date: 365 x current feature numbers = # of attribute columns
    - Engineer a week of year x attribute to reduce the column count

Dane

- Data visualization/exploratory analysis
- Model research
- Feature engineer
    - growing sessions
- Research into finding connections and read papers gathered in project infancy
- Familiarize myself with data:
    - Start by looking at where most ergot cases have occurred
    - Check percentage of district ergot samples against all district samples taken
    - Data cleaning?

Remmy:

- Label satellite moisture data
- Model research

Jay

- Data visualization/exploratory analysis

Dharmit

- Data visualization/exploratory analysis
- Data cleaning

Daniel

- Data visualization/exploratory analysis
- Data cleaning
- Write lstm model
- Write modified version of the current autoencoder