**1. Background**:

* For a given date, we have one **compressed file**. This compressed file contains several grib2 files. These grib2 files are recorded at different times of that date. Each compressed file typically contains 30~50 grib2 files.
* Grib2 files: each time point has 4 corresponding files. For example, at time 1
  + File 1: ruc2.t00z.bgrb13f01.grib2, each such file contains information for 750 layers for the US continent.
  + File 3: ruc2.t00z.bgrb13anl.grib2, each such file contains information for 750 layers for the US continent.
  + File 3: ruc2.t00z.pgrb13f01.grib2, each such file contains information for 317 layers for the US continent.
  + File 3: ruc2.t00z.pgrb13anl.grib2, each such file contains information for 317 layers for the US continent.
  + Colby suggested **using only the two anl files for each time point**. So, for each time point, we have one instance1067 (750+317) layers. It means that each instance has **1067 attributes + other related attributes**.
  + What does the 13 mean in the file name? One pixel is 13cm X 13cm
  + The “anl” files have the actual measured raw data.
  + The “f01” files have the predicted measures using some prediction model (the predication of the next hour).
* The different layers record information for different physical values (e.g., temperature, air pressure).
* Each **layer** has a name such as temperature\_alt1, temperature\_alt2, humidity\_alt1 …ect.
* Each layer is partitioned in 337 (rows) x 451 (columns) cells.
* **Each cell** has a (longitude, latitude) coordinate, which is the center point of this cell

**2. Raw data description**:

* Folder named ***RUC13*** (given by Josue) contains 26% of the data in Dr. DuBoise’s lab. This folder contains compressed grib2 files.
* **training\_data\_index.csv**: this file shows the dust events happened from 2002.03.01 to 2012.05.27. The detailed description of this file see below.
* We need to extract the actual data from the grib2 files for the given training data index.
* TODO: All the data (~700G) check whether they contain all the dust events in the training\_data\_index.csv file
  + Where is the data is located ???
  + Among all the data, only one file on date matches the dust events
  + The detailed information can be found for data/ tarfile\_with\_date.csv

**2.1 training\_data\_index.csv** **File description:**

It contains 4 columns

Longitude/Latitude: the location of a recorded dust event, which may not be the center of a cell (calculations are need in order to find the cells that have this dust event).

Date: date of a dust event in the format of yyyy.mm.dd

ID: does not have any real meaning. It is generated in the original system. Can be deleted.

For example:



**3. How to extract dust events**

**3.1 Procedure to extract training data from grib2 files and training data index**

**Input**: (1) training\_data\_index\_db.csv, (2) compressed grib2 files in RUC13, (3) requested event time periods

**Output**: actual training data

**Procedure**:

* For each row in training\_data\_index.csv which gives a date (2002.03.01), the program first finds the compressed file for that date.
* From the grib2 files in that compressed file, we extract the values at coordinate (-107.40643,31.594833) from all 317 layers.
  + Which grib2 files to use? Set this to be a parameter.
    - E.g., -1 means from all the files.
    - {10, 11, 13} means from the grib2 files that are recorded at 10am, 11am, 1 pm.
  + Which cells for this coordinate (-107.40643,31.594833)?
    - Colby suggested to use the closest cell to this coordinate.
    - Note: (not used, for documentation purpose) Yuhao used a parameter threshold (say delta) to determine whether a cell contains this dust event. Specifically, if abs(event. longitued-cell.longitued)< delta and abs(event. latitude -cell. latitude)< delta, then, this cell is counted as having this dust event.

**3.2 Format of actual dust training data**

**Each row of the training data has the following attribute:**

1. dust\_event\_date
2. dust\_event\_longitude,
3. dust\_event\_latitude,
4. timepoint (e.g., 0, 1, … representing t00, t01, …)
5. cell\_id (0 for dust event, 1-8 for non-dust events)
6. cell\_longitude,
7. cell\_latitude,
8. to (1074) var\_L1, var\_L2, …, var\_L1067

**Programs to write:**

1). A program that checks to see if we have data for a given date based on the training data index.

Text file

One row of dust event csv file, grib2 file

2). A program that extracts all the variables and their values from the specified location (lat, long) this program should also export data into a csv file.

**4. How to choose non-dust events**

Suggested by Colby:

For a given cell with dust, randomly choose another cell that is a neighbor of this dust cell. In total, we can choose from 8 neighboring cells.

**Reading one compress file: about 1-2 hour**.

**5. How to organize training data**

- Put all the dust events to one csv file. The cell id is zero.

- Put all the non-dust events to one csv file. => nondust\_8cells.csv (Needs to have column with cell id (from 1 to 8))

**Progress**:

+ Finished code to get the non-dust event data and the dust event data.

+ By 11/27/2017, Get all the available data; but only one day’s data (06/04/2011) matches one dust event; Already notified Josue and Dr. Colby.

+ By 12/8/2017, add one source file “src/Check\_matched\_file.py” to check the matching files with the dust event. The results are in “src/Training\_data\_match.csv”.

+ By 12/8/2017, write code to “src/Extract\_data\_from\_grib2.py”

nondust\_8cells.csv => nondust\_1cell.csv

**TODO**:

Put the data folder in Qixu’s computer here

Add the readme information of src/Check\_matched\_file.py to README file.

About the data: get clarification from Josue for Sean’s data

Add one test case for src/Extract\_data\_from\_grib2.py (i.e., add one compressed file and the generated csv files)

Put a readme file for the data and properly organize them

Get all the training (dust, non-dust) data (up in the air)

Suggested by Dr. Colby (Use another data set in https://nomads.ncdc.noaa.gov/data/rucanl/. So I need to write a python crawler to find the date folder that matches the given dust events and download the files from this website. For each folder, we only use the grib2 at 14:00 pm, so each folder we can only find one file. At most, we can have (334 dust event instances and 334 non-dust event instances), each instance has 310 attributes. I am not sure these data are enough for training a model even if we use cross validation.

* Write the code
* Add the readme
* Put the data
* Extract the information

Start to work on the RNN algorithm (NO)