**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.
* Each **grib2 file** contains information for 317 layers for the US continent.
* 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 row\_num x col\_num cells.
* **Each cell** has a (longitude, latitude) coordinate, what is for? Center point of a cell

**2. Description of input from Josue**:

* 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 decription of this file see below.
* We need to extract the actual data from the grib2 files for the given training data index.

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

It contains 4 columns

Longitude/Latitude: the location of a cell

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

ID: How is this ID created? (unique number for individual cells)

ID is for differentiate the generated files. ID doesn’t any special meaning.

For example:



1. **Procedure to extract training data from grib2 files and training data index**

**Input**: training\_data\_index.csv, compressed grib2 files in RUC13, requested event times

**Output**: actual training data

**Procedure**:

* For each row 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.

1. **Format of actual training data**

**3.1 Each row of the training data:**

Date, latitude, longitude, values for the 317 layers from each grib2 file

**3.2 Format of training data (“||” is a seperator):**

date(1) **||** lat **||** long **||** variable 1 **||** variable 2 **||** . . . . . **||** variable n-1 **||** variable n **||**

date(2) **||** lat **||** long **||** variable 1 **||** variable 2 **||** . . . . . **||** variable n-1 **||** variable n **||**

date(3) **||** lat **||** long **||** variable 1 **||** variable 2 **||** . . . . . **||** variable n-1 **||** variable n **||**

date(4) **||** lat **||** long **||** variable 1 **||** variable 2 **||** . . . . . **||** variable n-1 **||** variable n **||**

date(5) **||** lat **||** long **||** variable 1 **||** variable 2 **||** . . . . . **||** variable n-1 **||** variable n **||**

**.**

**.**

**.**

date(m-1) **||** lat **||** long **||** variable 1 **||** variable 2 **||** . . . . . **||** variable n-1 **||** variable n **||**

date(m) **||** lat **||** long **||** variable 1 **||** variable 2 **||** . . . . . **||** variable n-1 **||** variable n **||**

**3.3 # of instances:**

548 rows of Training\_Data.csv \* number of grib2 we choose

**Programs to write:**

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

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.