Group Project: Proposal

Hao Tong (htong25), Jiayi Shen (jshen226), Junxia Zhao (jzhao347), Yuanyou Yao (yyao93)

November 13th, 2020

Data Set

GHCN (Global Historical Climatology Network)-Daily includes records from over 100,000 stations in 180 countries areas. This data set provides diverse daily variables, including maximum and minimum temperature, total daily precipitation, snowfall, and snow depth. The range of the record and the time of the record vary by station and coverage interval, changing from less than one year to more than 175 years. (`https://www.kaggle.com/noaa/noaa-global-historical-climatology-network-daily?select=by-year-status.txt`)

Statistical questions

Can we use the climate data in past days to accurately predict the climate in coming days? Based on the climate data in past ten years, we might.

Code

```
> require(data.table)
> Year1980 =fread("1980.csv")
> head(Year1980)[ ,c(1,2,3,4)]
```

Variables

There are several variables including the following meteorological elements:

- Station Identifier: IDs for each station
- Date (yyyymmdd; where yyyy=year; mm=month; and, dd=day)
- Daily Minimum Temperature: the lowest temperature in a day
- Daily Maximum Temperature: the highest temperature in a day
- PRCP : daily precipitation
- SNOW: amount of snow fall
- SNWD: snow depth

Station ID	Date	Variables	Observations
USC00011512	19800101	TMIN	67
USC00011512	19800101	TMAX	28
USC00011512	19800101	PRCP	28
USC00011512	19800101	SNOW	0
USC00011512	19800101	SNOW	0

Statistical Methods

- Based on online research, we found out that we can use data in around ten days as a measure, and conduct the prediction based on these data. For example, since we have recorded the climate data in the past 9 days, we can set these data as our scheme. Then we can search for 5 9-day climate data that fits our scheme best, and use these data to predict the climate in the coming days.

Computation

We would mainly use Shell, CHTC, and R. Since our data set is pretty large so that we have to use CHTC and do parallel computation for data each year.