CS 670/470 Team Project (Part 2)

-- Long-lead Forecasting of Extreme Precipitation

Yong Zhuang, Xin Shu

Why we need long lead extreme precipitation forecasting?

Warning responses that are often seen are denial and mistrust, both resulting in total lack of response. This may be mitigated by careful provision of information, by avoiding surprises, and by facilitating social support. Finally, there are certain people who may be less likely to respond to warnings for various reasons, including isolation, age, disability and language. In these cases, support groups may hold the key to ensuring an effective response. For all of these potential improvements, there is a need for more lead time: time to prepare and disseminate more information; time for that information to be read and understood; time for support groups to mobilize and contact vulnerable people; time for people to seek confirmatory information from multiple sources.

--- B. W. Golding in "Review Long lead time flood warnings: reality or fantasy?"

Lead Time

the latency between the initiation and execution of a process.

Given a week's lead time, the sequence of information actions might be:

- ❖ 3–5 days ahead: issue 'advisory' or 'period of heightened risk'; engage in awareness raising activities through the media, mobilize support organizations for the vulnerable; initiate 'participatory' information sharing by local flood response organizations.
- ❖ 1–2 days ahead: issue 'early warning' or 'watch'; activate mitigation measures for flood minimization and protection of critical infrastructure; provide active support to vulnerable groups; move to a 'consultative' engagement with those in the most vulnerable areas.
- ❖ Hours ahead: issue 'flood warning'; activate emergency response; evacuate most vulnerable groups if appropriate; provide 'prescriptive' advice to individuals.

The challenge of long lead extreme precipitation forecasting

- While a short-term prediction of certain location depends only on variables in near spatial and temporal neighborhood, predictions with long lead time must consider variables in a long time window and large spatial neighborhoods, this means an enormous amount of potentially influencing variables and only a subset of them strongly relate to prediction. Processing a deluge of variables and discovering strongly relevant features pose a significant challenge for big data analytics.

The challenge of long lead extreme precipitation forecasting

- ❖ Extreme precipitation rarely occurred in a year, so the total number of positive samples (periods of extreme precipitation) in the experimental data set is much less than the number of negative samples.

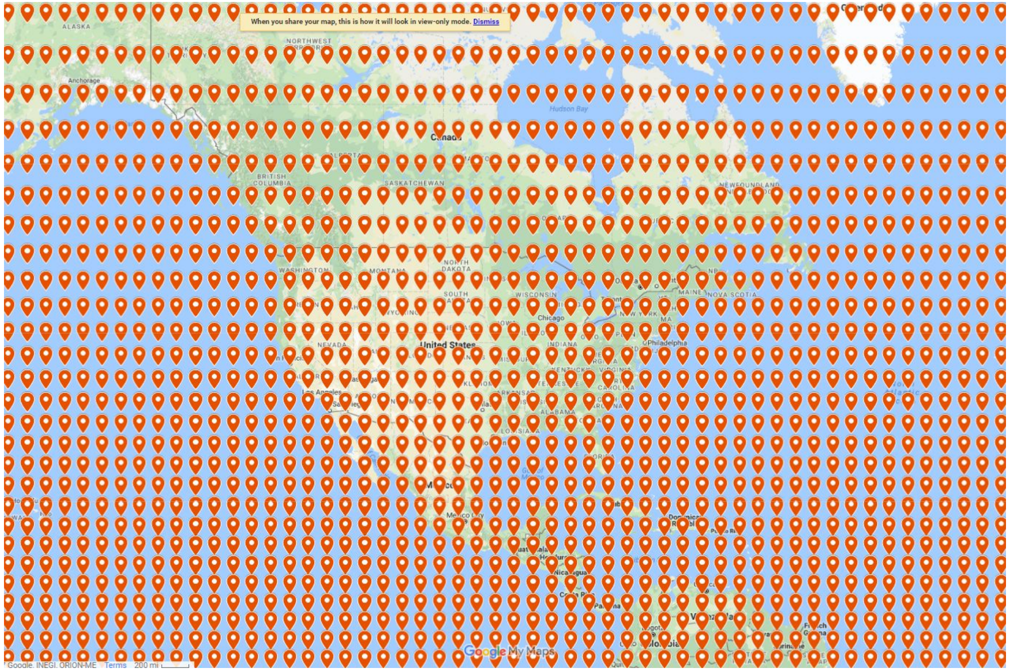
- ❖ How to deal with the imbalance problem is another challenge.

Problem Formulation

Let S be a set of locations, m be a vector of meteorological variables, M be a set of m, and C be the precipitation risk level. For each location s, its precipitation risk level at a specific time t is denoted as C_t^s, and its historical climate information in a fixed time-period q is given as $M_{(t-q+1)\sim t}^s = \{m_{t-q+1}^s, m_{t-q+2}^s, \ldots, m_t^s\}$, where $m_{t_i}^s$ presents the vector of variables collected in the location s at the time t_i. And the historical climate information of all locations in the fixed time-period q is denoted as $M_{(t-q+1)\sim t}^S = \{M_{(t-q+1)\sim t}^{s_1}, M_{(t-q+2)\sim t}^{s_2}, \ldots\}$. Given the lead time as p, then the long lead extreme precipitation forecasting can be formulated as to predict C_{t+p}^s , based on the history $M_{(t-q+1)\sim t}^S$

Locations, S

There are totally 5,328 locations in the Northern Hemisphere evenly distributed between the equator and the North pole (37 latitudes and 144 longitudes).

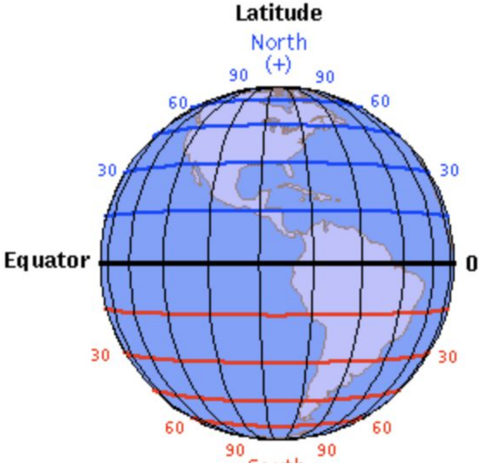


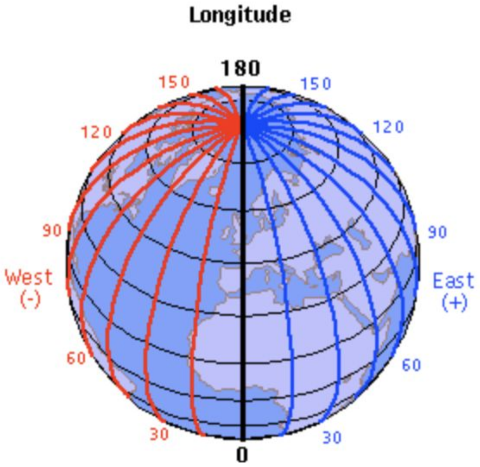
When you share your map, this is how it will look in view-only mode. Dismiss

A grid system of 2.5 degree in latitudes and longitudes.

Locations, S

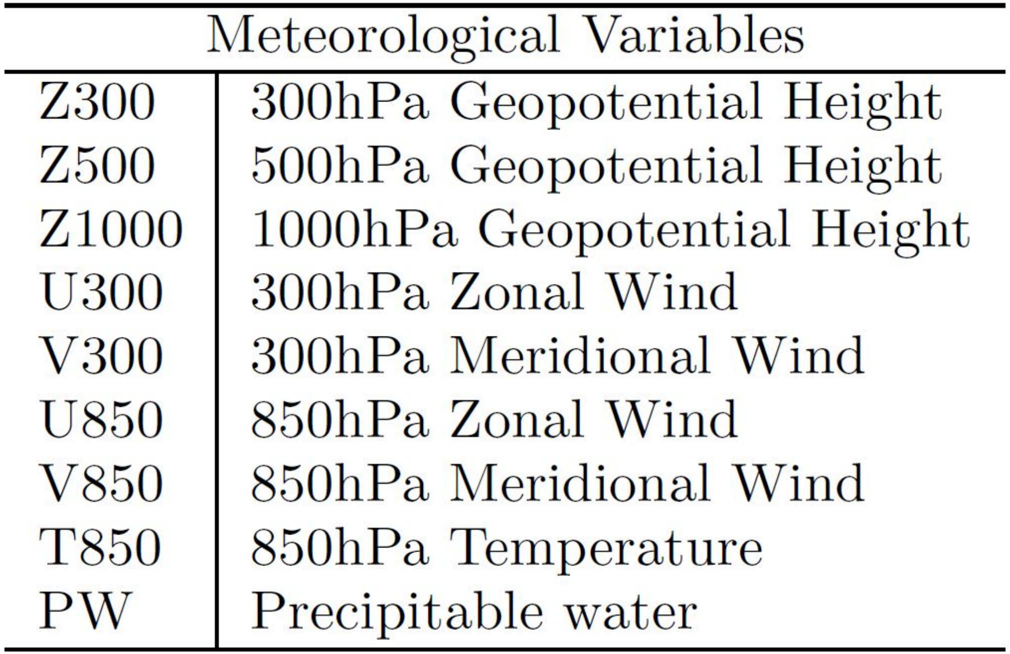
There are totally 5,328 locations in the Northern Hemisphere evenly distributed between the equator and the North pole (37 latitudes and 144 longitudes).





Meteorological Variables, m

We choose 9 variables from the historical meteorological data of the Northern Hemisphere as features.

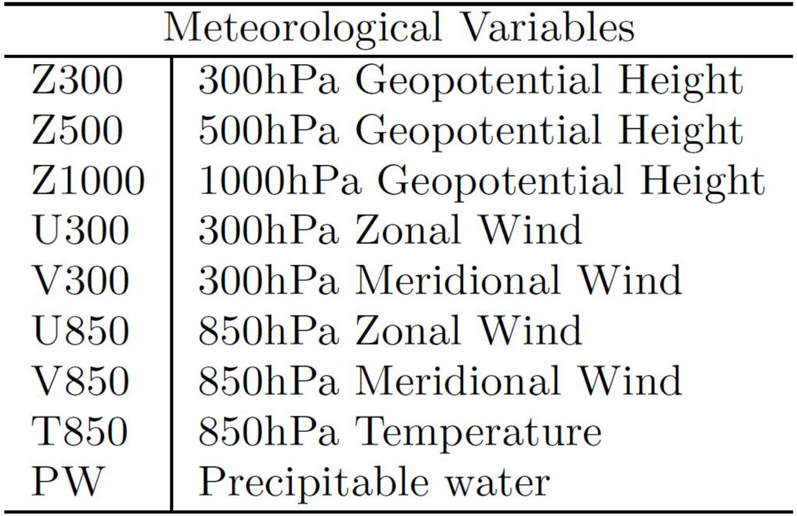


Meteorological Variables	
Z300	300hPa Geopotential Height
Z500	500hPa Geopotential Height
Z1000	1000hPa Geopotential Height
U300	300hPa Zonal Wind
V300	300hPa Meridional Wind
U850	850hPa Zonal Wind
V850	850hPa Meridional Wind
T850	850hPa Temperature
PW	Precipitable water

Meteorological Variables, m

- Z~: Geopotential Height: a vertical coordinate referenced to Earth's mean sea level — an adjustment to geometric height (elevation above mean sea level) using the variation of gravity with latitude and elevation.

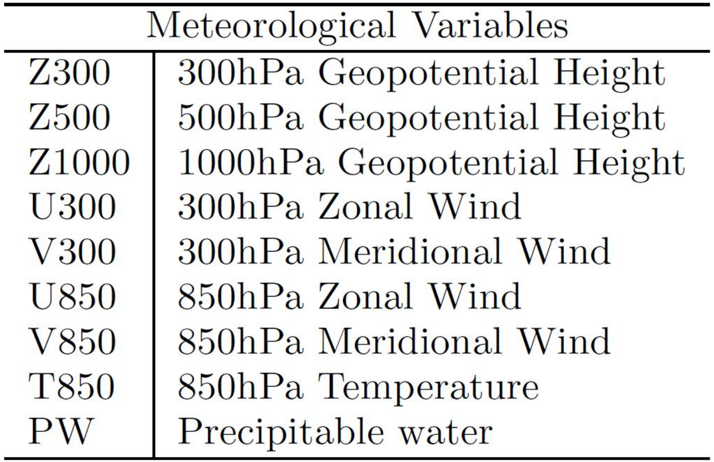
- a “gravity-adjusted” height



Meteorological Variables	
Z300	300hPa Geopotential Height
Z500	500hPa Geopotential Height
Z1000	1000hPa Geopotential Height
U300	300hPa Zonal Wind
V300	300hPa Meridional Wind
U850	850hPa Zonal Wind
V850	850hPa Meridional Wind
T850	850hPa Temperature
PW	Precipitable water

Meteorological Variables, m

- Compared with the geometric height, the height of the potential does not take into account the changes in gravity with height.
- One usually speaks of the geopotential height of a certain pressure level, which would correspond to the geopotential height at which that pressure occurs.
- To reflect the changes of the pressure field.



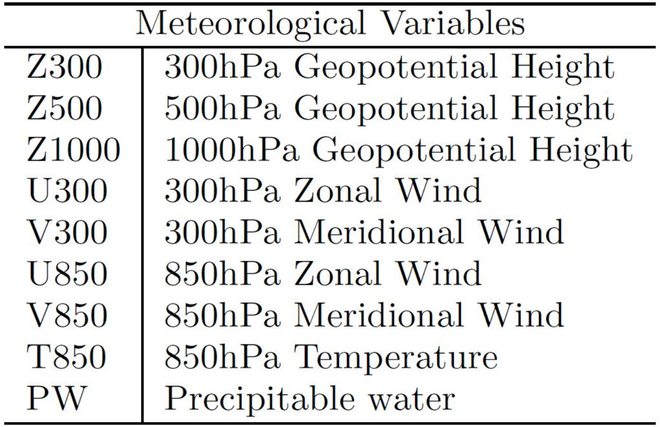
Meteorological Variables	
Z300	300hPa Geopotential Height
Z500	500hPa Geopotential Height
Z1000	1000hPa Geopotential Height
U300	300hPa Zonal Wind
V300	300hPa Meridional Wind
U850	850hPa Zonal Wind
V850	850hPa Meridional Wind
T850	850hPa Temperature
PW	Precipitable water

- E.g. Z300 corresponds to the geopotential height at which that pressure occurs at 300 hectopascal (hpa).
(1 hPa = 100 Pa)

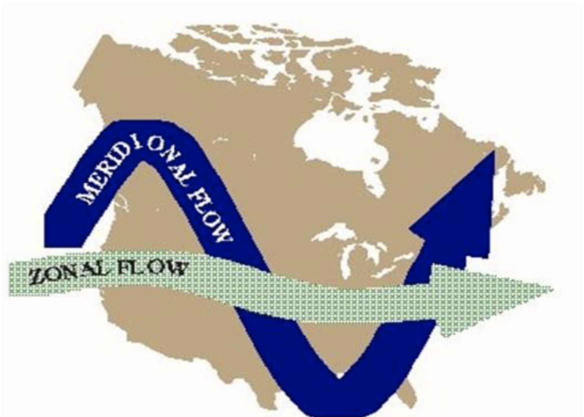
Meteorological Variables, m

- U~: Zonal wind means "along a latitude circle" or "in the west–east direction"; which reflects the changes in upper atmospheric east-west winds.

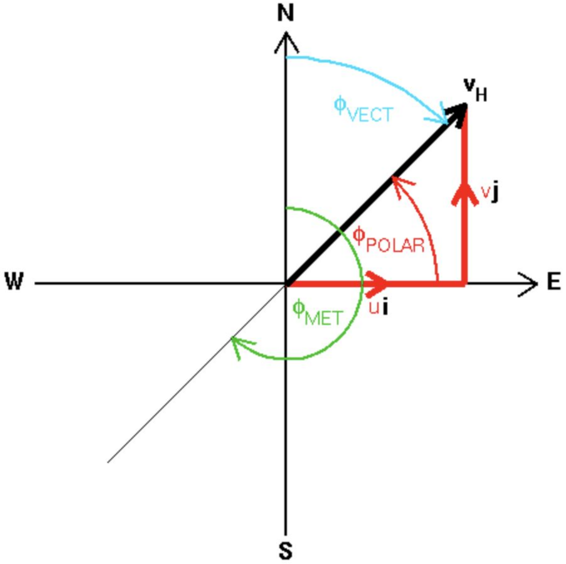
- V~: a meridional flow is a general air flow pattern from north to south, or from south to north, along the Earth's longitude lines (perpendicular to a zonal flow), which reflects the changes in upper atmospheric north-south winds.

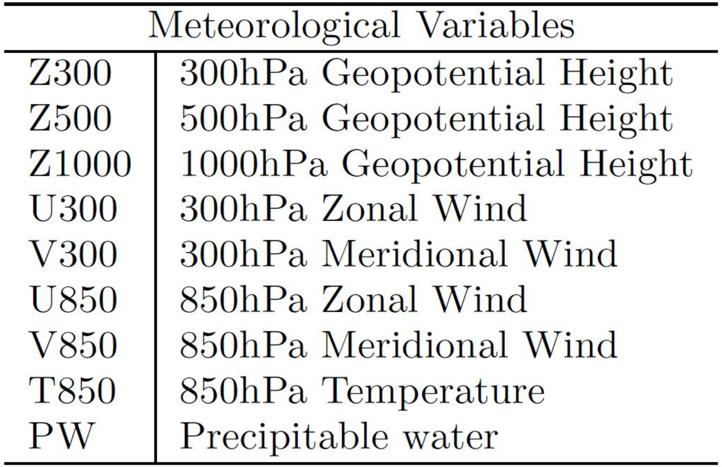


Meteorological Variables	
Z300	300hPa Geopotential Height
Z500	500hPa Geopotential Height
Z1000	1000hPa Geopotential Height
U300	300hPa Zonal Wind
V300	300hPa Meridional Wind
U850	850hPa Zonal Wind
V850	850hPa Meridional Wind
T850	850hPa Temperature
PW	Precipitable water



Meteorological Variables, m





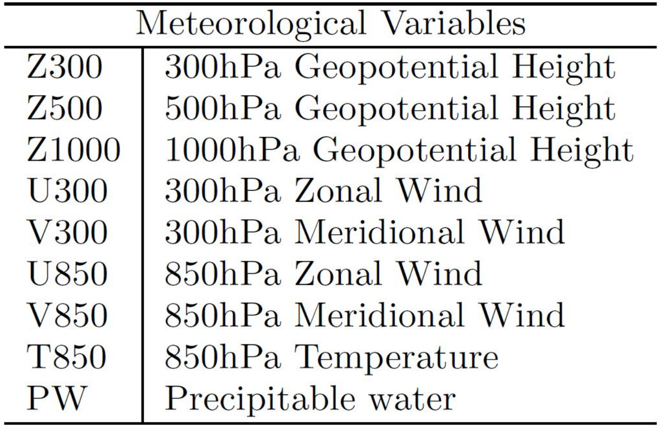
Meteorological Variables	
Z300	300hPa Geopotential Height
Z500	500hPa Geopotential Height
Z1000	1000hPa Geopotential Height
U300	300hPa Zonal Wind
V300	300hPa Meridional Wind
U850	850hPa Zonal Wind
V850	850hPa Meridional Wind
T850	850hPa Temperature
PW	Precipitable water

Vh: the horizontal wind vector
I: unit vector towards East
J: unit vector towards North

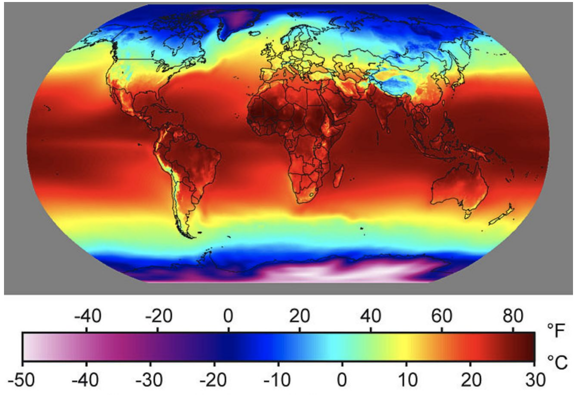
Meteorological Variables, m

- T~:The changes in temperature as measured above each location at the point where the pressure is 850 hpa.

- PW: Total precipitable water. Precipitation is any product of the condensation of atmospheric water vapor that falls under gravity; the depth of water in a column of the atmosphere, if all the water in that column were precipitated as rain.



Meteorological Variables	
Z300	300hPa Geopotential Height
Z500	500hPa Geopotential Height
Z1000	1000hPa Geopotential Height
U300	300hPa Zonal Wind
V300	300hPa Meridional Wind
U850	850hPa Zonal Wind
V850	850hPa Meridional Wind
T850	850hPa Temperature
PW	Precipitable water



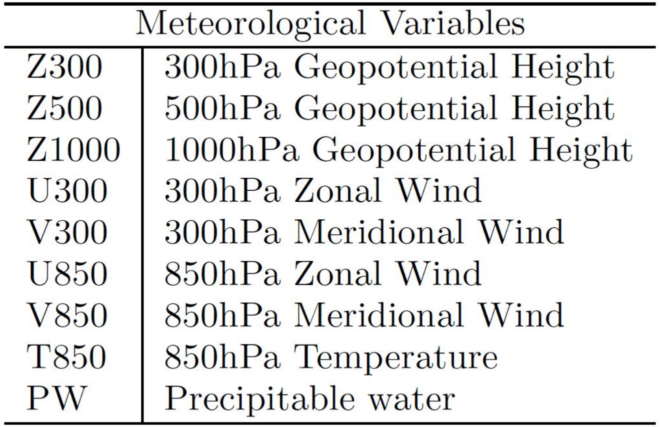
Annual Mean Temperature

Meteorological Variables, m

- **The atmosphere is a continuous layer of gases**
- **Pressure changes with Height**

pressure decreases with increasing altitude

As altitude increases, pressure diminishes, as the weight of the air column decreases. The pressure at any level in the atmosphere may be interpreted as the total weight of the air above a unit area at any elevation. At higher elevations, there are fewer air molecules above a given surface than a similar surface at lower levels.



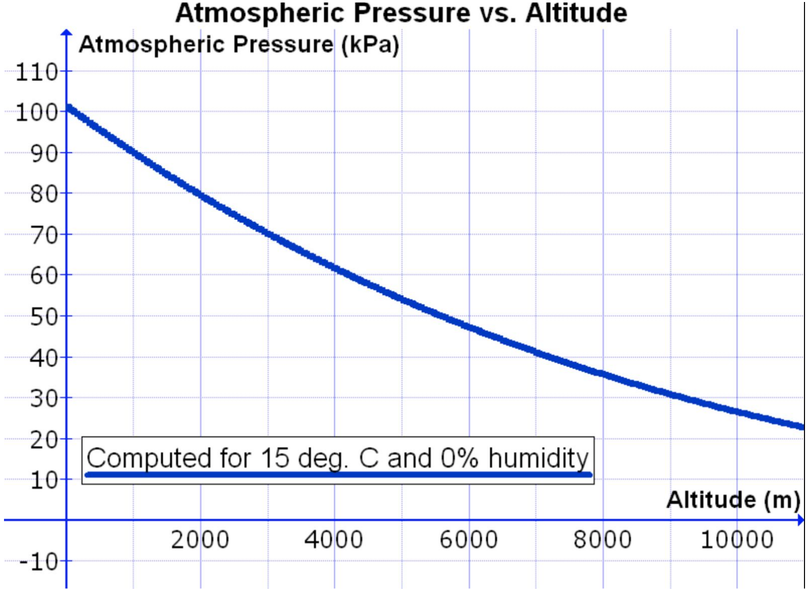
Meteorological Variables	
Z300	300hPa Geopotential Height
Z500	500hPa Geopotential Height
Z1000	1000hPa Geopotential Height
U300	300hPa Zonal Wind
V300	300hPa Meridional Wind
U850	850hPa Zonal Wind
V850	850hPa Meridional Wind
T850	850hPa Temperature
PW	Precipitable water

Meteorological Variables, m

- **Pressure changes with Height**
- **Static equations**：the atmospheric static equation applied to the weather.

$$h = \frac{R_d T}{P g} = \frac{8000}{P}(1 + \frac{t}{273})(\frac{m}{hPa})$$

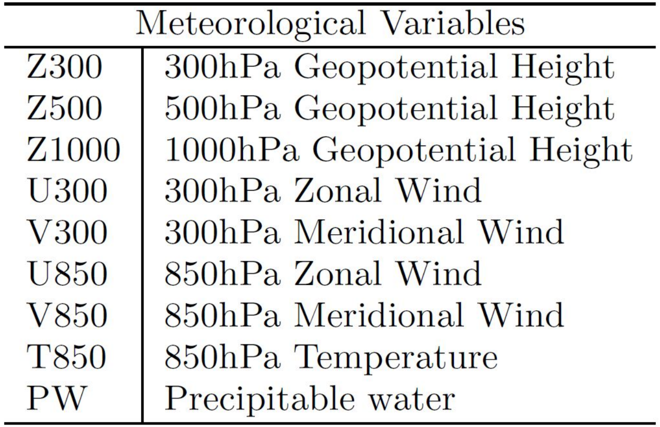
elevation (m)	pressure (h Pa)
30000	12
16000	100
11000	250
5500	500
3000	700
1500	850
0	1000



Meteorological Variables, m

❖ **Pressure with Height**

The isoparametric views of the daily work of the meteorological station have 850 hPa , 700 hPa , 500 hPa and 300 , 200 , 100 hPa and so on. They represent the horizontal pressure field near 1500m , 3000m , 5500m and 9000m , 12000m and 16000m respectively . Sea level pressure field is generally used to see the contours (zero height) to analyze, if necessary, with 1000 hPa isobaric surface map instead.



Meteorological Variables	
Z300	300hPa Geopotential Height
Z500	500hPa Geopotential Height
Z1000	1000hPa Geopotential Height
U300	300hPa Zonal Wind
V300	300hPa Meridional Wind
U850	850hPa Zonal Wind
V850	850hPa Meridional Wind
T850	850hPa Temperature
PW	Precipitable water

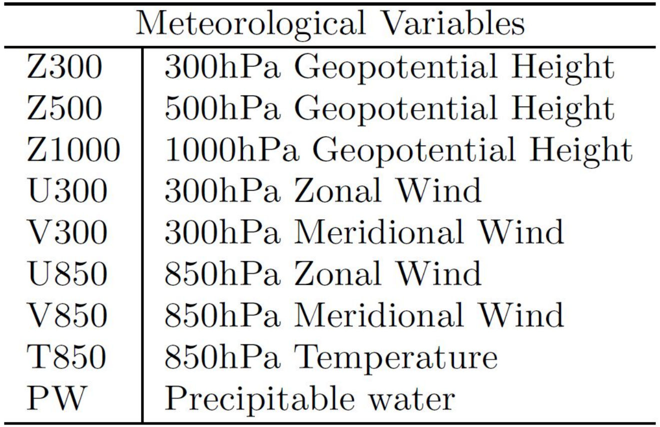
Meteorological Variables, m

- **Pressure with Height**

As for why we choose 300 hPa, 500 hPa, 850 hPa, 1000 hPa, is because :
In meteorology, we use these isobaric surface at these height to represent:

Upper atmosphere;	300 hPa
Mesosphere;	500 hPa
planetary boundary layer (PBL)	850 hPa
Surface ;	1000 hPa

which are of great typical significance.



Meteorological Variables	
Z300	300hPa Geopotential Height
Z500	500hPa Geopotential Height
Z1000	1000hPa Geopotential Height
U300	300hPa Zonal Wind
V300	300hPa Meridional Wind
U850	850hPa Zonal Wind
V850	850hPa Meridional Wind
T850	850hPa Temperature
PW	Precipitable water

Meteorological Variables, m

- **Pressure with Height**

The geopotential heights of 1000 hPa and 500 hPa are chosen because, taken together, the fields will contain information that allows us to infer where large-scale rising motion (and therefore precipitation) is likely to take place.

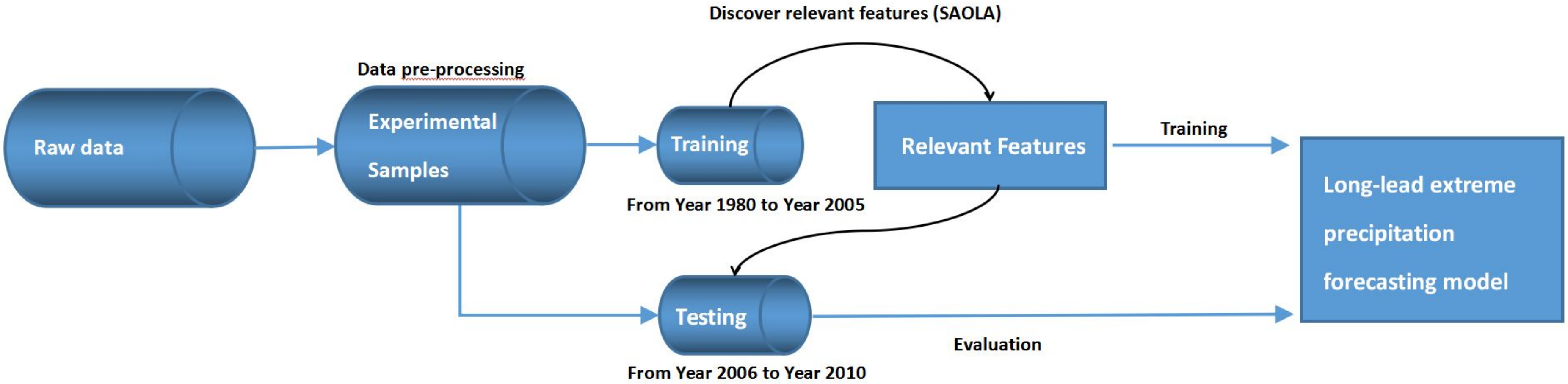
The 850 hPa meridional (i.e. North- South) wind is chosen because it is extremely important for the transport of heat and moisture from the tropics into the mid-latitudes.

Precipitation Risk Level, $\mathcal{C}$

Binary risk level:

extreme precipitation ($\mathcal{C}$ =1) and non-extreme precipitation ($\mathcal{C}$ =0).

Experiment - Flow Chart



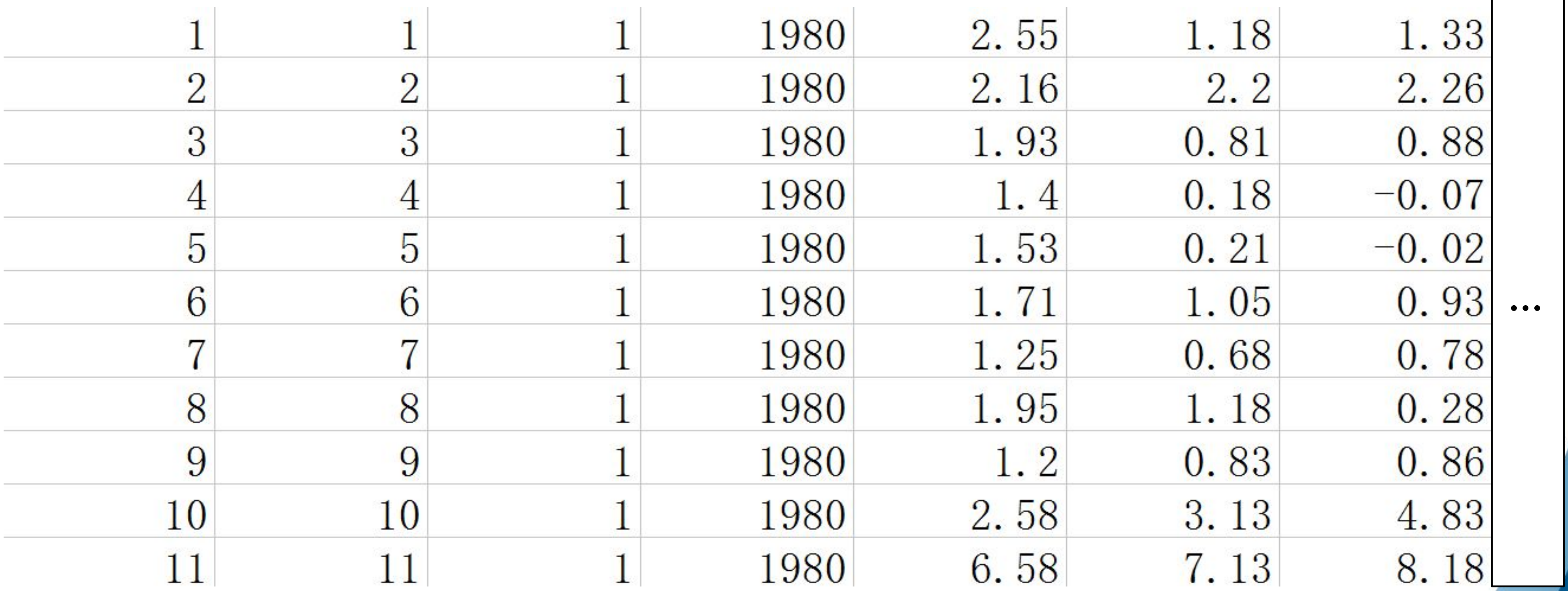
Raw Data

The data we used for experiment is the historical meteorological data (including 9 variables) of the whole Northern Hemisphere (5,328 locations) over 30 years (1980-2010) and the historical spatial average precipitation data of the state Iowa from the same time period.

Each variable has 31 csv files (1 file per year).

CSV file example

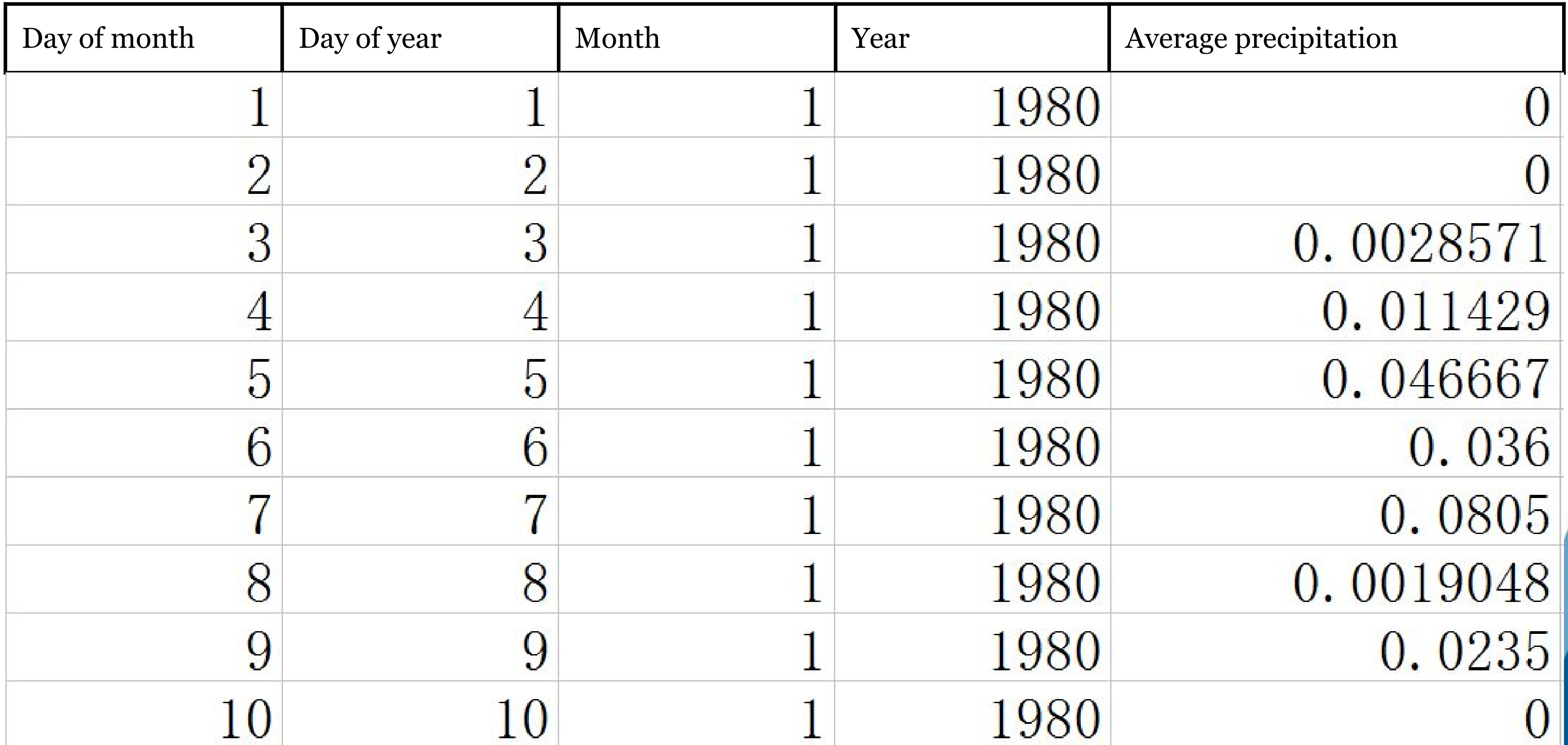
pw_1980.csv



Day of month	Day of year	Month	Year	Measurements in 5,328 locations ...			
1	1	1	1980	2.55	1.18	1.33	
2	2	1	1980	2.16	2.2	2.26	
3	3	1	1980	1.93	0.81	0.88	
4	4	1	1980	1.4	0.18	-0.07	
5	5	1	1980	1.53	0.21	-0.02	
6	6	1	1980	1.71	1.05	0.93	…
7	7	1	1980	1.25	0.68	0.78	
8	8	1	1980	1.95	1.18	0.28	
9	9	1	1980	1.2	0.83	0.86	
10	10	1	1980	2.58	3.13	4.83	
11	11	1	1980	6.58	7.13	8.18	

CSV file example

iowa_1980.csv



Day of month	Day of year	Month	Year	Average precipitation
1	1	1	1980	0
2	2	1	1980	0
3	3	1	1980	0.0028571
4	4	1	1980	0.011429
5	5	1	1980	0.046667
6	6	1	1980	0.036
7	7	1	1980	0.0805
8	8	1	1980	0.0019048
9	9	1	1980	0.0235
10	10	1	1980	0

Data Pre-Processing: Features

S: 5328 locations.
m: 9 variables.

So the historical climate information in a fixed time period q (10 days) is:

$$M^{S}_{(t-9)\sim t} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|}\hline m^{S}_{t-9} & m^{S}_{t-8} & m^{S}_{t-7} & m^{S}_{t-6} & m^{S}_{t-5} & m^{S}_{t-4} & m^{S}_{t-3} & m^{S}_{t-2} & m^{S}_{t-1} & m^{S}_{t} \\ \hline \end{array}$$

5328* 9 * 10 features

Data Pre-Processing: Labels

We label a time window as being a period of extreme precipitation if the total rainfall in that period reaches a historically high level (above the 95th percentile of the historical record). For this project, we'll be using a window size of 15 days.

Since each window has a unique start date, we can assign the label of the window starting on that date to the date itself. So if the period from July 1st to July 15th is above the 95th percentile for rainfall, we mark July 1 as a positive example, as the beginning of an extreme precipitation period.

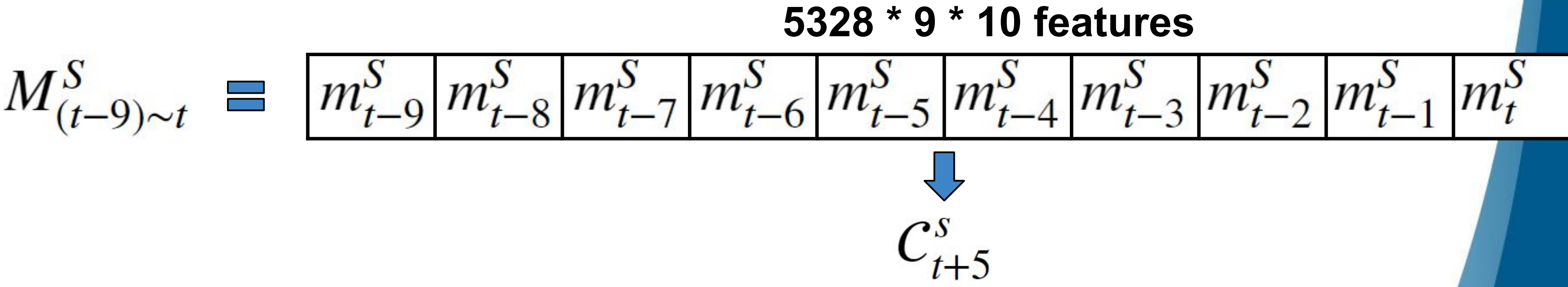
Data Pre-Processing: Labels

Lead time p: 5 days.

S : 5328 locations.

Target area s: Iowa

Data : The historical spatial average precipitation data of the state Iowa .



5328 * 9 * 10 features

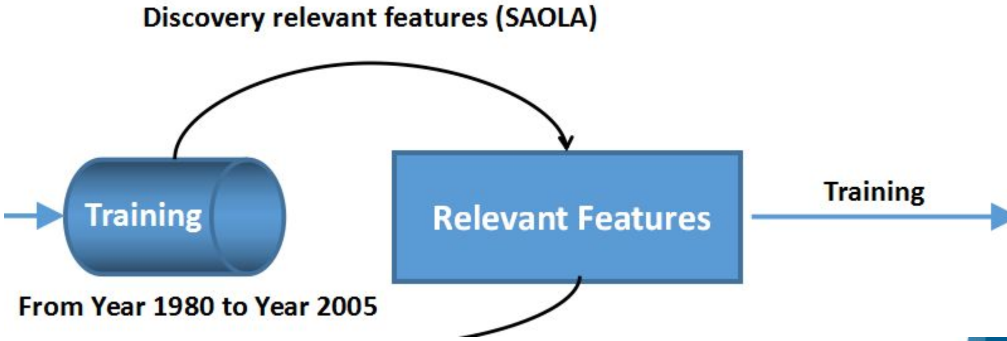
$$M^{S}_{(t-9)\sim t} = \boxed{m^S_{t-9}}\boxed{m^S_{t-8}}\boxed{m^S_{t-7}}\boxed{m^S_{t-6}}\boxed{m^S_{t-5}}\boxed{m^S_{t-4}}\boxed{m^S_{t-3}}\boxed{m^S_{t-2}}\boxed{m^S_{t-1}}\boxed{m^S_{t}}$$

$$\downarrow$$

$$C^{s}_{t+5}$$

C^{s}_{t+5} =1, if $\sum_{t+5}^{t+19} PW_{Iowa}$ is above 95% percentile of any sum of 15 days' precipitations in the historical records. Otherwise, C^{s}_{t+5} = 0.

Training: Class Imbalance

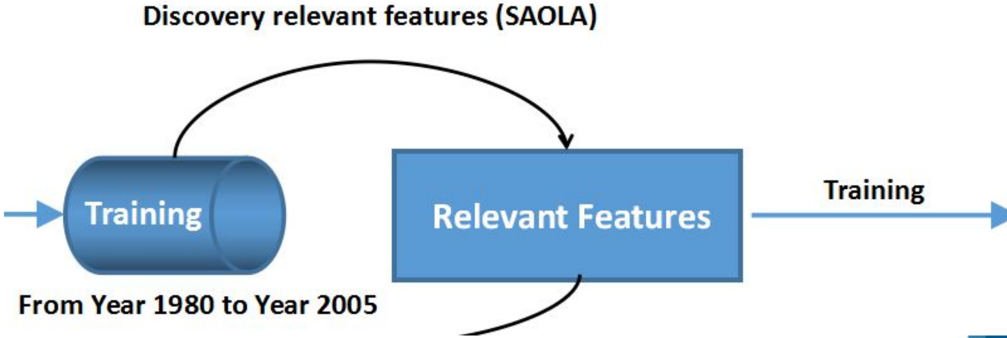

Discovery relevant features (SAOLA)
Training
Relevant Features
Training
From Year 1980 to Year 2005

Because extreme precipitation rarely occurred, we have only 5% positive samples and 95% negative samples. We should solve the class imbalance problem before we train the prediction model.

Task: Solve the class imbalance problem

Bonus points: Can you modify SAOLA to improve its performance on unbalanced data?

Training: Normalization


Discovery relevant features (SAOLA)
Training
Relevant Features
Training
From Year 1980 to Year 2005

Because different variables have different scales, We should adjust values measured on different scales to a notionally common scale (e.g. 0~1) before we train the prediction model

Task: Do Normalization

Training: Classification Models

Long-lead extreme precipitation forecasting model

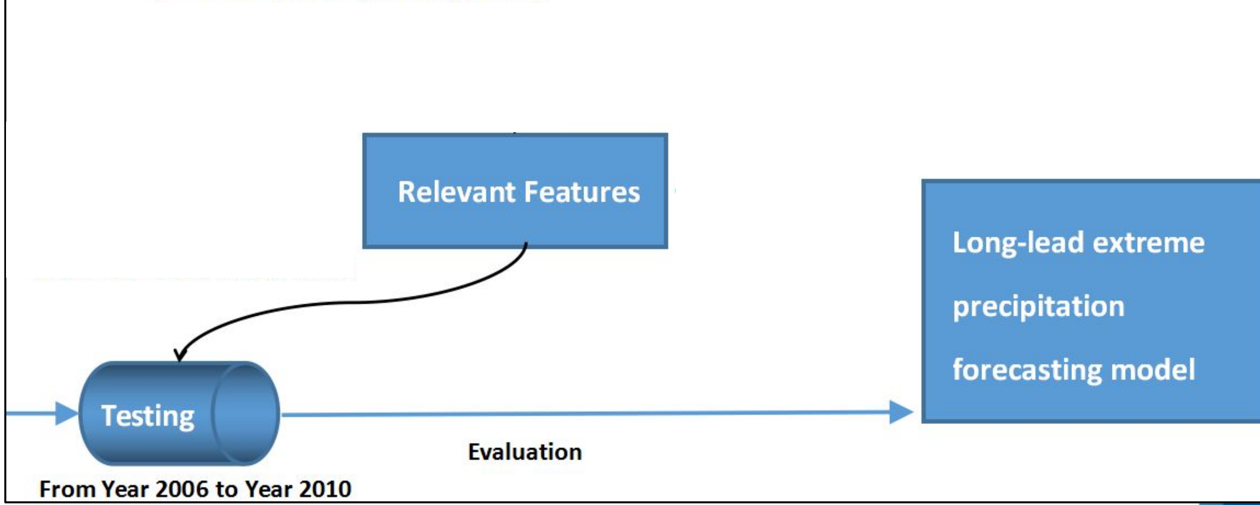
Task:

Do 10-fold cross-validation on the classification models.

Classification models:

Use KNN, Support Vector Classification(SVC), Random Forests, Linear Discriminant Analysis, Multilayer Perceptron. And then compare the results.

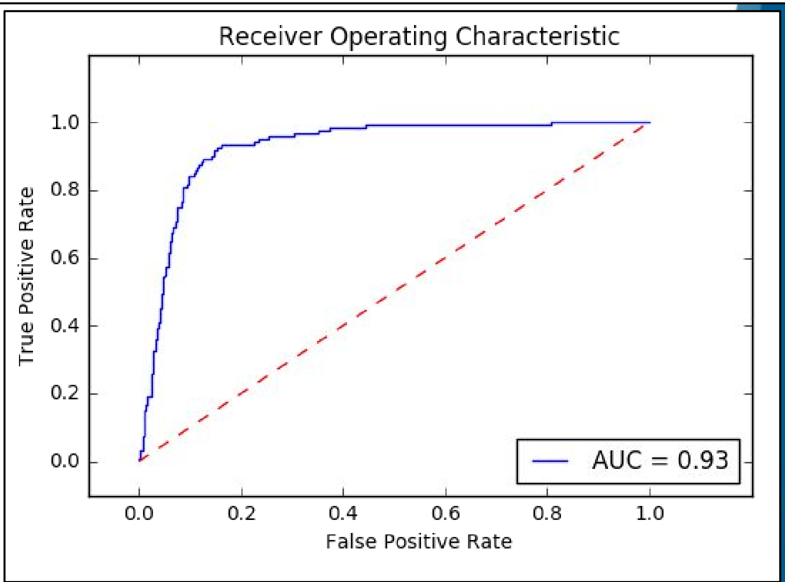
Testing: Metrics


Relevant Features
Long-lead extreme precipitation forecasting model
Testing
Evaluation
From Year 2006 to Year 2010

Task:

Calculate the accuracy.

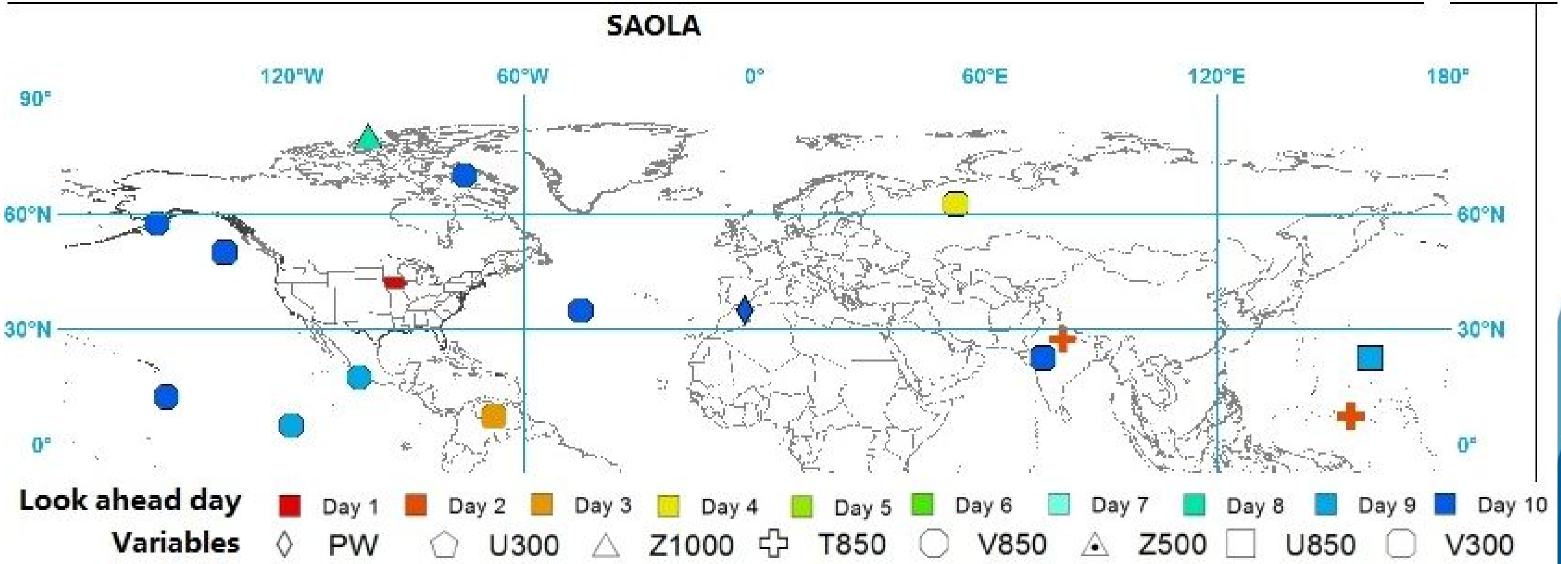
Draw the receiver operating characteristic curve (ROC) and calculate the Area Under Curve (AUC).



Evaluate the relevant features on a map

Task:

Bonus points: Show your selected relevant features on a map and briefly discuss the results.



Tasks

1. Implement SAOLA in Python
2. Do Pre-Processing to create the experimental samples
3. Solve the class imbalance problem

   --Bonus points: Modify SAOLA to improve its performance on unbalanced data.
4. Do normalization on the experimental samples.
5. Train the classification Models.
6. Evaluate the classification models on testing set, calculate accuracy and AUC, and draw ROC curve.
7. Evaluate the relevant features on a map (Bonus points).

Reference

- KNN: http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- SVC: http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- Random Forests: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- Linear Discriminant Analysis: http://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html#sklearn.discriminant_analysis.LinearDiscriminantAnalysis
- Multilayer Perceptron: http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html#sklearn.neural_network.MLPClassifier
- AUC: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html

Thank You!