**Presentation Draft**

Good morning, ladies and gentlemen. It’s an honor to be here.Thank you for attending my project talk. So, let’s start it. My project is conducted by the GateWay Antarctica, a research center in UC. And you can see, my project topic is: Application of a Machine Learning Algorithm,Self-Organizing Map, for Examining the consistency between Atmosphere reanalyses about near-surface horizontal wind fields.So, why GateWay Antartica ,and why we do this project? And what is Atmosphere reanalysis? I can explain for your guys. Atmosphere reanalysis provide the best estimate of the atmospheric state at any one time, usually by combining in-situ and satellite observations with forecast model data from a fixed version of a numerical weather prediction (NWP) system. Here, I have summary it, that is :Atmosphere reanalysis = forecast model + satellite observations.It means, reanalysis is equal forecast model, which is built from the previous data and plus daily observations, which may from adiosonde、satellite、buoy、aircraft and ship. But, usually, the reanalysis products are not accurate, because both of the models and observational data may carry biases and errors. So, the different reanalysis may have some differences between each other. MY Job is to exam how far the differences between each other, or in other words, that is the consistency between them. And my goals for this topic :

1. ---- Which regions the quality of consistency for different atmospheric reanalyses is best、worst or others?
2. Which years the quality of consistency for different atmospheric reanalyses is best、worst or others?
3. How these atmospheric reanalyses characterize the atmospheric state in Antarctic regions, is it bad? And if so, how bad?

First step, Data Acquisition, usually, the atmospheric reanalysis products has been divided into three generations. In first generations, three centres ,European Centre for Medium-Range Weather Forecasts(ECMWF) 、National Aeronautics and Space Administration Data Assimilation Office and National Centers for Environmental Prediction(NCEP)and National Center for Atmospheric Research(NCAR).they provided three products; In second generation, Japan Meteorological Agency conducted first Asian long-term global atmospheric reanalysis project and In USA, Department of Energy(DOE) cooperated with NCEP to develop their products; In third generation, more centres joined to develop products. Here, in this project, we used the ERA-Interim and ERA20C, these two products from ECMWF to exam their consistency. And all of the data are three dimensions: longitude、latitude and time , one is from one of January, 1979 to the end of 2015, and another is from first day in 1900 to the end of 2010. And the data format is netCDF, which is called network common data form, which is based on array. For example, when using python xarray library to open it , we can see this topography data is three coordinates, same as three dimensions , and here is the data variables. And here are some libraries that we used in this project, like numpy、pandas、xarray、cartopy and sompak and so on.

Then, when we got the wind field data, actually it was given as a matrix form, like this. The rows in matrix represents the space(here because each space point can be determined by longitude and latitude, so we complied these two dimensions together, just using space to represent) , and the column in matric represents the time. Here, We using EOF( Empirical Orthogonal Function technique), it also called PCA in mathematics. We using EOF analysis to decompose the Time and Space from the original data. Here, V is m by m matric, which is a space function and Z is m by n matric, which is a time function.

So, how can we do it in mathematics, here I have divided it into six steps.first step is to Select the data, and preprocessing data to obtain the data matrix, That is orginal data X. then Computing the cross product of X with its transpose, here is the Capital A and it is a m by m matric.from the knowledge in linear algebra, we can know if a matric is a square matrix, maybe we can get its eigenvalues ,λ, and eigenvectors V and make sure A multiplied by eigenvectors V is equal to eigenvectors V multiplied by eigenvalues ,λ. Then, after we got the eigenvalues ,we arrange the eigenvalues λ, from the largest ones to the smallest ones, because each eigenvalue corresponds to an eigenvector, V, so according to the eigenvalues ,λ, we can get the related eigenvecots. And all of them join together, that is the capital V. And this is what we want in the beginning. Then, step5, computing principal components Z. from this formal, we can know, if X is equal to V multiplied Z, then the corss product of X, A can also be used V、Z and their transposed matrix . So, according it,By projecting V onto the original data X , the corresponding time coefficients of all eigenvectos, here we name them are the principal components, can be obtained. Lastly, Computing contribution rate, in this project, threshold is set as 0.9. this means the first 90% of eigenvalues , λ, are what we want and we removed the lowest 10%.

Until now, we have got the principal components, capital Z. then we do the data modelling. We use the Self-Organizing Maps algorithms to get the SOM patterns from these PCs. So, what is Self-Organizing Maps. Yeh, it is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional(typically two-dimensional),discretized representations of the input training samples. So , everyone, please notice here. Its job is to create the representation for the orginal data. Here is a picture to show some common machine learning algorithms and the differences between their functions. please remember that, Self-Orhganizing maps is used for pattern classifications.

Actually , this algorithm has some similarities with K-means, which we have learnt in data mining course. So, understaning k-means can extremely helpful for us to understand Slef-Organizing Maps.But because the time limitation, I can not give your guys more details about K-means. Just review the course notes if you fogot it. The basic logic of K-means is changing the central points and calculate the distances and adjust each point to create the new central points and repeating again and again until satisfying the requirements.

So, how the Self-Organizing Map works? Let’s start it now.

This picture shows the whole process of Self-Organizing Maps. Frist, giving your guys an example.

In this dataframe, it has three variables, it means three dimensions,x1,x2,x3.And in Self-Organizing Map, we call each column as a map. Here,9 different output nodes have been selected.But in the project, we choose the 12 different output nodes. As we can find, a weight has been assigned to each of these connections. However, the meaning of weight is total different with those weights used in artificial or convolutional neural networks.In ANN, multiplying the input node’s value by the weight and applying an activation function ( = Σ𝑊𝑒𝑖𝑔ℎ𝑡∗𝑖𝑛𝑝𝑢𝑡)+𝑏𝑖𝑎𝑠 ). But in SOM, no this function. These weights do not separate from the output, and we can view these weights as the output’s coordinates.Then, calculate the distance.we can find in this example, the close node is node3,0.4. and we call the node is the best matching unit in Self-Organizing Map. But in real ,we need to adjust the weights to get the best BMU.Here,we use a 9\* 12 output nodes. Suppose that if the green circle is the best matching unit for the Map1.Here, is a picture from Google. And we can see, if we drag the BMU(the yellow one) closer to the data point, the nearby nodes also pulled closer to the data point.So, to determine which nearby nodes need to be adjusted,we need set a radius.And by change the length of radius, we can update the BMUs.And in the final , it can get the SOM patterns. Here is just an introduction about SOM.

Before Using SOM, we have divide the world into 36 equal-size regions. then by using the same methods, we apply the Self-Organizing Maps to each different regions and got the 12 different patterns.Here, the color stands for the strength of wind. And u is the zonal(east-west) and v is meridional (north-south) wind patterns.

Because when I divided the world, I just choose a specific longitude as the starting point, but for

Removing the randomness,we need to offset the longitude and make sure the all of the region boxes can overlap. So we choose 10 degree each time to adjust the model. Because each region is 60 degree longitude, so just 6 times it can be form a cycle.And finally, there are 6 differet states totally.

Then, by using the Self-Organizing Maps, we can know an important information: which nodes does each day belong to ? For example, in state 1, region1, the 1979-01-01 is belong to node1. And we can change the form of data into a datafram. Like time, node number, distance. And by mergeing the data, we can know for each year, how many days are belong to each modes, or their percentages.

Of course, we can use another way to analysis the result. One useful way is contingency table. Here is an example.in state1, region1, 1980. Both of the number of each nodes that the two datasets common have. And if we check the result with the histogram, we can find it is correct.

And how can we analysis the contingency table. One common method that w have learn in statistics is Chi-Square. And we use the Cramer’s V to parameterize the Chi-Square. If Cramer’s V lies into 0 and 1. Close to 0, the two reanalyses have no association,close to 1, more association.Another one is called entropy. It is created by Claude Shannon, the father of information theory. Also , for reparametrization of entropy, we use a measurement called “dependency”. And it also lies to 0 and 1 . Close to 0, no association; close to 1, more association.

So, according it ,we can get the time series of dependency for each region, from region1 to region36.each point in the line is a dependency value of one specific year in specific region from a specific state.

And also we can use heatmap to do this. One heatmap contains 36 different blocks,and each block stands for each region. For example, this is region1, region2.and the color stands for the strength of value. Green is Cramer’V, and the blue stands for the denpendency.

Until now, w have got all of what we want. So we can answer the beginning questions. Here, because for each region, it has the different values in different year, so we can sum of them and calculate the average value of them. And then Grade the different regions according to the final average value. Like Grade A , if the value is bogger than or equal to 0.7, and so on. And according to the values, we can answer the question1, more bigger value, the quality of consistency for the reanalyses are better.

And similar method, for each state, because we get the time series of dependency in the 36 different regions, we can put some of together according to the result from question1,like here, Grade A is the region 1、7、9、10、11、12、13、15、17. And we merge their time series picture together. and we can find their common trend, and according to the trend , we can answer the question2

For question3, base on question1 and question2, we can know in Antarctic regions, the Cramer’V valus and dependency value and their Grade is much lower than other regions.

Ok, Here , I have give your guys all of what I gave done in this project, hope you can understand and enjoy it . And here is a link , one intereting video about entropy, and you can watch it.

Ok , So, does anyone have any questions? And I can try to answer .