The long-term variation of the frontal system in Taiwan with machine learning-based classifier

Chiao-Wei Chang[1], Shih-Hao Su[1*], Ting-Shuo Yo[2], and Jung-Lien Chu[3]
Chinese Culture University, Taipei, Taiwan[1]
National Taiwan University, Taipei, Taiwan[2]
National Science and Technology Center for Disaster Reduction, Taipei, Taiwan[3]

Abstract

In Taiwan, the precipitation associated with frontal system not only serves as one of the major water resource in Taiwan but also the cause of natural disasters. Chang et al. (2019) performed three different objective methods to detect the frontal system in all seasons in Taiwan region. The first one is the traditional objective diagnostic method based on the spatial variability of thermal parameters, it can provide the intensity and position of frontal system while diagnosing. Its hit rate was low (10-20 %) as it missed most of the frontal systems that passed through Taiwan, while the false alarm rate was very low as well. The second method be applied is the self-organizing map (SOM)-based classifier. SOM is an unsupervised machine learning method that groups the patterns with similar characteristics into clusters. The SOM-based method had much higher hit rate (70-80 %) than the traditional objective methods, however, the false alarm rate was also high (20-60 %) between different seasons, which indicated that SOM over-diagnosed the pattern s of frontal systems. The third method is the machine learning-based diagnostic tool. The hit rate of machine learning-based method was approximately 60-70% while its false alarm rate was only 20-30 %. Moreover, both the linear and non-linear kernel were tested and showed that the long-term statistical properties of the front frequency can be well represented.

In this study, we applied the machine learning-based diagnostic method to select CMIP6 model datasets for future climate projection and to investigate the long-term statistical properties of frontal systems near Taiwan. The preliminary findings are that the derived results can be affected by model features, therefore the original data were replaced by the climatological anomalies to eliminate the difference among models.

Key word: frontal system, objective classifier, machine learning method, climate projection.

[*]Corresponding author: Shih-Hao Su, ssh3@g.pccu.edu.tw

1. Introduction

The precipitation associated with the fronts not only serves as the major water resource but also the cause of floods in Taiwan. Therefore, the prediction of the front system is a crucial issue in the management of water resources and flood protection. The subjective diagnostic analysis is one of the major methods for the prediction. However, it is time-consuming, and the subjective biases cannot be amended systematically. Objective diagnostic methods were developed in the early1960s. The spatial gradient of the thermodynamic parameters was used as the tracker of the front system. (e.g., Renard et al., 1965; Clark et al., 1966; Steinacker, 1992). Hewson (1998) indicated that merely using the thermal or dynamical variables would cause bias on the identification of the fronts. Hence it is essential to enhance the prediction skill by adding the threshold. Nevertheless, the setting of the threshold value could also lead to subjective biases.

Benefit from advances in computer science and the statistical methods; the clustering analysis was also applied to classify the varied types of weather patterns since clustering algorithms advances in resolving the non-linear problems. The self-organizing map (SOM) proposed by Kohonen (1982) is a clustering algorithm based on artificial neural networks (ANNs) and was taken as the diagnostic tool for the synoptic weather patterns associated with the Mei-Yu/Baiu fronts. (e.g., Nguyen et al., 2017) In more recent years, benefitting from the thriving development in the field of data science, the machine learning techniques are showing us a new path in the detection of weather systems (Su et al. 2018). In this study, following the work of Chang et al. (2019), we are introducing the traditional objective diagnostic method, SOM, and machine learning-based diagnostic tools for the identification of the fronts affecting Taiwan and further explore the potential of the machine learning methods in for future climate projection. The data and methodology we sued are introduced in section 2, and the results are presented in section 3. The summary and the future work are discussed in section 4.

2. Data and Methodology

The record of the front events over 1980-2016 was acquired from Taiwan Atmospheric event Database (TAD) (Su et al. 2018). The front events in TAD were

identified with the subjective analysis of the surface weather map launched by CWB at 00Z on a daily basis.

The National Centers for Environmental Prediction (NCEP) Climate Forecast System Reanalysis (CFSR) datasets (Saha et al., 2010; Saha et al., 2014) from 2001 to 2016 were utilized for the objective analysis and the training of the models. The CFSR reanalysis datasets provide detailed information of high temporal (6h) and spatial ($0.5^{o}\times0.5^{o}$) resolution. The variables we selected are listed in table 1, including 500 hPa geopotential height (HGT), mean sea level pressure (MSLP), the zonal wind (U), the meridional wind (V), relative humidity (RH), and temperature (T) on 925 hPa, 850 hPa, and 700 hPa.

In this study, we proposed three objective diagnostic tools to identify the front systems over 2011-2016 and evaluated the performance of each method.

2.1 Frontal thermal parameter (TFP) diagnostic tool

This tool took the TFP in the lower atmosphere as the indicator of the frontal system affecting Taiwan. The earlier study used the gradient of the wet-bulb equivalent potential temperature (θ_w) to track the frontal line (Hope et al., 2014). However, the gradient of the θ_w is insignificant near Taiwan, since the average relative humidity near Taiwan is generally higher than 75 % even under the influence of global warming (Su et al., 2012) and hence the cooling effect of evaporation is limited. On the other hand, the equivalent potential temperature (θ_e) can quantities the variations in both temperature and the effect of the latent heat of evaporation, which is more representative of the atmospheric environment in the Taiwan area. Here we used the gradient of the equivalent potential temperature ($\nabla\theta_e$) on 925 hPa in replace of θ_w and multiply it with the precipitation rate ($\nabla\theta_e\times$PR) as the TFP to detect the potential frontal systems and remove the non-precipitation ones. Only the data points with its height exceed 750 m were kept; as a result, the geographical compensation was performed to fill the filtered data. Moreover, to eliminate the weak systems, we analyzed the cumulative distribution function (CDF) of $\overline{\nabla\theta_e \times \mathrm{PR}}$ (unit: 10·K·kg/km·s) of each candidate and set 200 (10·K·kg/km·s) as the threshold value. The operational flow is summarized in figure 1 (adapted from Chang et al, 2019).

2.2 Self-organized map (SOM)-based diagnostic tool

The SOM is a non-linear algorithm for clustering and classification. The SOM is a single-layer neural network, and the training of the SOM model is based on unsupervised learning by moving its prototype to fit the data according to the distribution of the input data. (Sá et al., 2012) From the earlier study, we have noticed that the features of the frontal systems affecting Taiwan are distinct from those in the mid-latitudes. (e.g., Ninomiya et al., 2007) In this study, we utilized the SOM as an objective strategy to classify the typical patterns of the fronts affecting Taiwan by taking the long-term data in May and June (MJ) from 1980 to 2016 as the inputs. We selected U, V, T, and RH on 850 hPa, the variables that can describe the most significant features of the Mei-Yu front. (Ninomiya et al., 2007) To reduce the data size and enhance the computational efficiency, EOF is applied for all the input variables, and the modes that can explain 90 % of total variance were kept as the input feature set.

2.3 The machine learning-based diagnostic tool

The machine learning-based diagnostic tool (ML) was designed on the basis of a supervised learning strategy. The structure of the ML consists of the training stage and the application stage. In the training stage, the binomial label (true or false) of whether the frontal events occur and the corresponding atmospheric features are essential for the ML model learns the mathematical relationship classification and the data features, and establish the classification model for the frontal systems. We combined the frontal event data logs and the CFSR 6-hourly reanalysis data from 2010 to 2010 to establish a golden standard as the training datasets. The variables we selected contained the dynamical and thermodynamical fields on the standard levels and were summarized in detail in table 1. Also, we applied PCA in the data pre-processing to reduce the dimension of the data and retained the first 20 modes as the inputs. For the classification algorithm, we applied the Support Vector Machine (SVM) (Chang et al., 2001) with a polynomial kernel as the primary pattern classifier. The SVM has been widely used in the field of data science. It projects the inputs data to high-dimensional feature space and acquires a hyperplane in the space that can separate the "true or false" conditions. We also applied the general linear model (GLM) as a baseline of the classification results.

2.4 Model evaluation

The performance of the three objective analysis methods was evaluated by the confusion matrix (Fawcett, 2006). The difference between the predicted event and the actual event can be categorized into True positive (TP or hits), False positive (FP or false alarm), False negative (FN or missing), and True negative (TN). From the confusion matrix, the accuracy, the hit rate, the false alarm rate, and F1 score can be evaluated by the formulas below:

Accuracy = (TP + TN)/(TP + FP + FN + TN)
Hit rate = TP/(TP + FN)
False alarm rate = FP/(FP + TN)
F1 score = 2×TP/(2×TP+FP+FN)

The scores of the models are demonstrated in the next section.

3. Results

3.1 TFP diagnostic tool

If $\nabla\theta_e$ was used as the tracker of the fronts, the hit rates in DJF, MA, and MJ (Mei-Yu) would be 0.16, 0.16, and 0.09, and the false alarm rate would be 0.18,

0.15, and 0.05, respectively. This suggested that $\nabla\theta_e$ is more sensitive to the frontal systems in winter and spring seasons while it did not have the skill to diagnose the Mei-Yu front since the temperature gradient is more potent in winter and spring seasons than in the Mei-Yu season. Replacing $\nabla\theta_e$ with $\nabla\theta_e \times$ PR, the hit rate in DJF, MA, and MJ were 0.15, 0.21, and 0.29, and the false alarm rates were 0.14, 0.10, and 0.12, respectively. (Summarized in figure 1) The systems with no precipitation would be discarded if PR=0, as a result, the hit rate in MA and MJ seasons increased, since the frontal rainfall is significant in these seasons. However, the hit rate in DJF descended slightly due to the precipitation associated with the fronts is weaker in the season. Note that the false alarm rate in both DFJ and MA decreased while it increased in the Mei-Yu season as some linear systems with weak precipitation would be recognized as fronts. Besides PR, we applied the threshold value to eliminate the candidates with weak precipitation in a similar way to Hewson (1998), and there was a loss of 31.5 % of the hitting events in DJF while 79.5 % of the false alarms were improved; in MA, the loss in hitting events was 25.8 %, and 68.2 % of the false alarms were improved; In MJ, the loss in hitting events was 8.6 %, and 28.6 % of the false alarms were improved. Overall, $\nabla\theta_e \times$ PR with the use of the threshold value can reduce the false alarm rate in the detection of the front systems.

3.2 SOM diagnostic tool

The hit rates of SOM in DJF, MA, and MJ (Mei-Yu) were 0.85, 0.92, and 0.79, and the false alarm rates were 0.20, 0.5,9 and 0.46, respectively. (Summarized in figure 1) The clusters of the weather pattern feathers in MJ classified by SOM is shown in figure 2. The weather pattern in MJ can be classified into nine types. When the Mei-Yu fronts influenced Taiwan, there is a warm tongue pointing to Taiwan on 850 hPa θ_e. (figure 1a). On the other hand, there is a significant convergence zone above Taiwan. (figure 1b) These results suggested that SOM was capable of detecting the front systems in all seasons. Yet, the false alarm rates were higher than that of the TFP methods due to the loose criterion of the clustering analysis. Among all the seasons, the false alarm rate in MA was the highest. In MA, the front-like rainbands over the southern China area that moved eastward to Taiwan were attributable to such a bias. Nevertheless, in terms of the F1-score, the skill in the Mei-Yu season, which scored 0.6 outperformed the skill in DJF (F1-score = 0.52) and MA (F1-score = 0.45).

In this study, we used the identical variables for the detection of the fronts in each season due to the limited computational resources. Nevertheless, the previous studies had shown that there were distinctive differences in the features of the synoptic-scale environment. It would reduce the false alarm rate by using the representative variables which best describe the seasonal features of the atmospheric environments as the inputs.

3.3 Machine learning-based (ML-based) diagnostic tool

The label of the front events and CFSR reanalysis data from 2001 to 2010 was used as the training dataset and trained the model on the basis of both GLM and SVM. In the training process, the training dataset was split into ten equal portions; the model was trained with 90 % of the dataset and validated with the other 10 %. In the training process, the cross-validation was repeated until the model converged. We applied the model to the identification of the front events through 2011-2016; the hit rate of GLM (SVM) in DJF, MA, and MJ were 0.61 (0.65), 0.71 (0.70), and 0.69 (0.71). The false alarm rate of GLM (SVM) in DJF, MA and MJ were 0.15 (0.17), 0.26 (0.26) and 0.22 (0.28) (figure 2). Overall, the ML-based diagnostic tool outperformed the TFP and the SOM in terms of both the hit rate and false alarm rate.

The identified and observed frontal days were summarized in figure 4. The correlation coefficient of the historical and GLM-identified frontal days was 0.8, while that of SVM was 0.74 due to its higher false alarm rate. The results suggested that there were some biases between the identified conditions and the true conditions. We further examined the bias between the hit events and the false alarm events (figure 5). In winter, the bias of GLM displayed a band of a positive anomaly in 850 hPa temperature with a strong zonal temperature gradient and 850 hPa specific humidity extending from Japan to Taiwan (figure 6). The bias pattern of GLM in MA was similar to DJF, while in MJ, the GLM tended to identify the environment with stronger cold air north to Taiwan and the moist condition near Taiwan. The bias patterns of SVM were similar to GLM, except the 850 hPa temperature showed a band of positive anomaly similar to the patterns in DJF and MA. These results indicated that the ML-based diagnostic tool tends to identify the systems with the structure similar to the mid-latitude baroclinic fronts as the front events. Similar to the SOM, the ML-based diagnostic tool follows the principle of subjective analyses while it allows exceptions.

4. Summary and future works

In this study, we have tested three objective diagnostic tools to identify the front systems affecting Taiwan. The TFP method based-on the spatial variations of the thermodynamic parameters was the most ineffective one among these methods we have tested, which is attributable to the weak gradients of the thermodynamic parameters near Taiwan that led to the low sensitivity of the tool. However, the TFP method could provide the position and the intensity of the frontal system in the analytic process. The SOM was established on the basis of the clustering algorithm, and it clustered the inputs data into groups depending on the similarity of the features. It not only identified the targets but also classified the varied types of the front systems. For the ML-based tool, both the GLM and the SVM algorithm can not only identify the typical frontal

systems but also effectively diagnose the long-term variations of the occurrence frequency of the front.

For the future work, we are attempting to apply the ML-based weather classifier on the front systems under the future climate scenario by analyzing the CMIP6 model output. Currently, we are selecting the models by analyzing the similarities between the historical simulations and the reanalysis datasets via hierarchical clustering. To standardize the effects of model properties, the climatology anomalies were used as inputs to identify the weather events. For further study, the ML-based weather classifier will be applied to estimate the changes and the variations in frontal precipitation and properties under difference climate change scenarios.

Reference

Chang, C. C., and Lin, C. J., 2001, LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/cjlin/libsvm.

Chang, C. W., Chiang, C. T., Liu, K. Y., and Su, S. H., 2019: The comparison of objective diagnose methods for Taiwan frontal system classification. 大氣科學, **47(1),** 1-29.

Clarke, L.C., and Renard, R. J., 1966, The U.S. Navy numerical frontal analysis scheme: further development and a limited evaluation. *J. Appl. Meteorol.*, **5**, 764-777.

Fawcett, T., 2006, An introduction to ROC analysis. *Pattern Recognition Lett.*, **27**, 861-874.

Hewson, T. D., 1998, Objective Fronts. *Meteorol. Appl.*, **5**, 37-65.

Hope, P., Keay, K., Pook, M., Catto, J., Simmonds, I., Mills, G., McIntosh, P., and Berry, G., 2014, A comparison of automated methods of front recognition for climate studies: a case study in southwest Western Australia. *Mon. Weather Rev.*, **142**, 343-363.

Kohonen, T., 1982, Self-organized formation of topologically correct feature maps. *Biol. Cybern*, **43**, 59-69.

Nguyen-Le, D., Yamada, T. J., and Tran-Anh, D., 2017, Classification and forecast of heavy rainfall in northern Kyushu during Baiu season using weather pattern recognition. *Atmos. Sci. Lett.*, **18**, 324-329.

Ninomiya, K., and Shibagaki, Y., 2007: Multi-scale features of the Meiyu-Baiu front and associated precipitation systems. *J. Meteor. Soc. Japan*, **85B**, 103–122.

Renard, R. J., and Clarke, L.C., 1965, Experiments in numerical objective frontal analysis. *Mon. Weather Rev.*, **9**, 547-556.

Saha, S., Moorthi, S., Pan, H.L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.T., Chuang, H.Y., Juang, H.M.H., Sela, J., Iredell, M., Treadon, R., Kleist, D., van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., van den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R.W., Rutledge, G. and Goldberg, M., 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91(8)**, 1015-1057.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.T., Chuang, H.Y., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M.P., van den Dool, H., Zhang, Q., Wang, W., Chen, M. and Becker, E. (2014) The NCEP climate forecast system version 2. Journal of Climate, 27, 2185 – 2208.

Steinacker, R. A., 1992, Dynamic aspects of frontal analysis. *Meteorol. Atmos. Phys.*, **48**, 93-103.

Su, S.-H., Kuo, H.-C., Hsu, L.-H., and Yang, Y.-T., 2012, Temporal and spatial characteristics of typhoon extreme rainfall in Taiwan. *J. Meteor. Soc. Japan*, **90**, 721-736.

Su, S. H., Chu, J. L., Yo, T. S., & Lin, L. Y. (2018). Identification of synoptic weather types over Taiwan area with multiple classifiers. *Atmospheric Science Letters*, e861.

Sá, J., Rocha, B., Almeida, A., and Souza, J. R., 2012, Recurrent self-organized map for severe weather patterns recognition. Recurrent Neural Networks and Soft Computing, Intech Open, 151-175.

Levels	200 hPa			500 hPa	700 hPa / 850 hPa / 925 hPa				925 hPa	Surface level	
Variables / Methods	Zonal wind (U)	Meridional wind (V)	Temperature (T)	Geopotential height (H)	Zonal wind (U)	Meridional wind (V)	Temperature (T)	Relative Humidity (RH)	Equivalent potential temperature (Θ_e)	Mean sea level pressure (MSLP)	Precipitation rate (Pr)
Traditional objective analysis									×		×
SOM clustering method					850 hPa only	850 hPa only	850 hPa only	850 hPa only			
Machine learning method	×	×	×	×	×	×	×	×		×	
Unit	m s^{-1}	m s^{-1}	K	gpm	m s^{-1}	m s^{-1}	K	%	K	Pa	$K \cdot kg \cdot kg^{-3} \cdot s^{-1}$

Table 1. The input NCEP-CFSR variables of the objective diagnostic tool. (The variables used are marked by "x".)

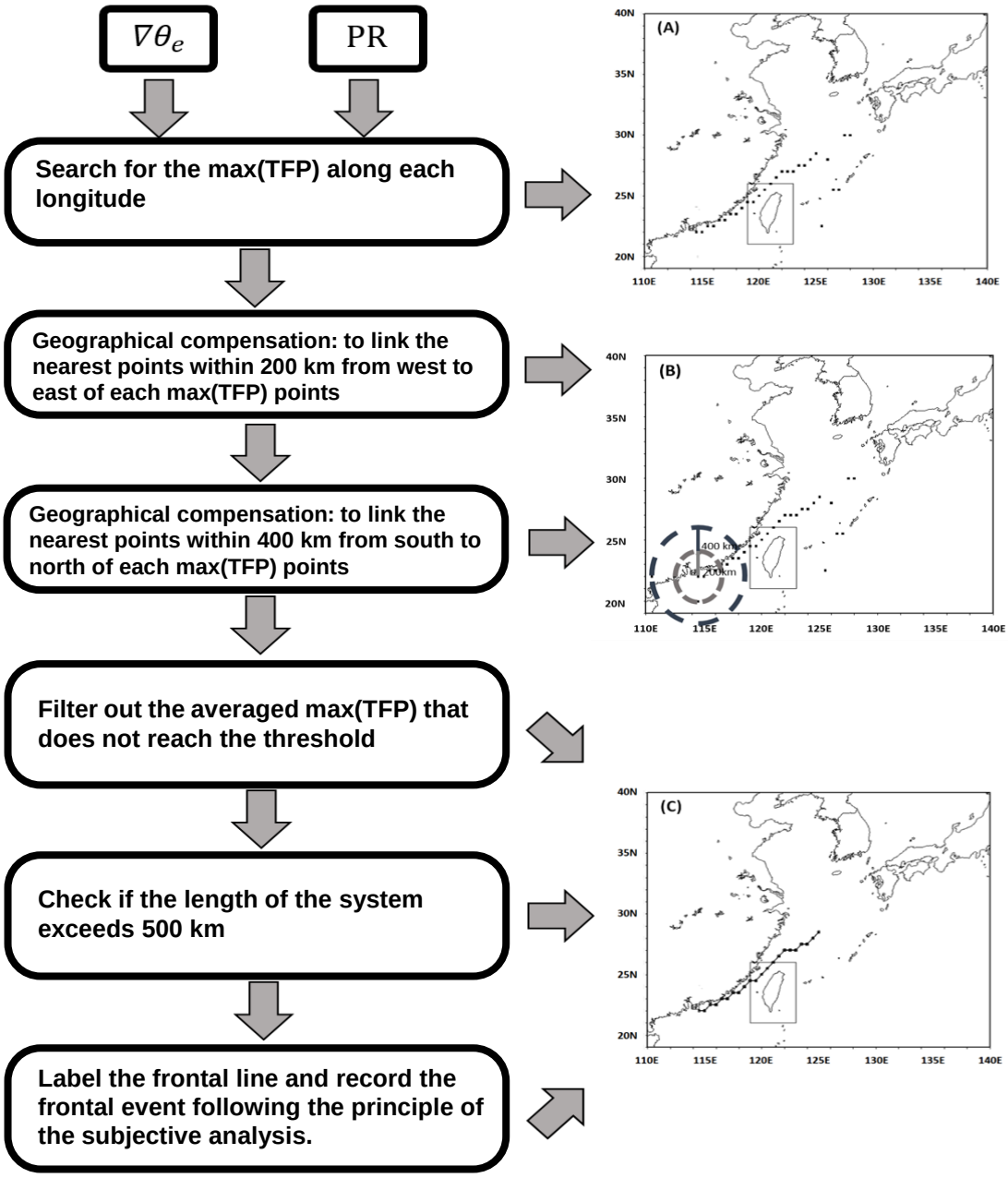

Figure 1. The work flow of TFP front identification. (a) calculating $\nabla\theta_e$, $\nabla\theta_e \times PR$ and max(TFP), (b) operating the geographical compensation, (c) the target front line filtered by the threshold value. (Adapted from Chang et al., 2019)

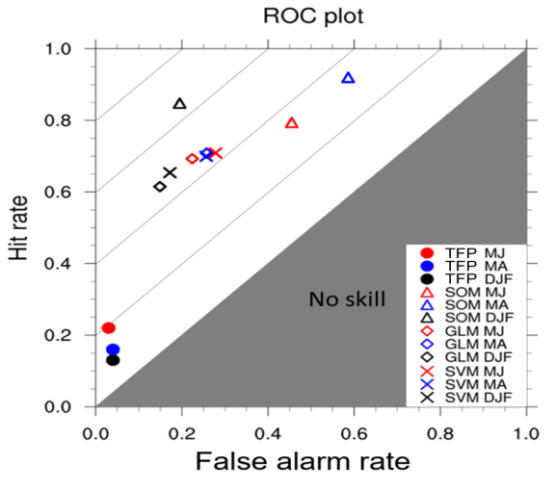

Figure 2. Relative Operating Characteristic (ROC) plot. The solid dots are the scores of the TFP diagnostic tool, the triangles represent the SOM, the diamonds and "X" represent the GLM and SVM, respectively. DJF, MA and MJ are colored by black, blue and red, respectively. The diagonal lines mark the diagnostic skill, the shaded mean "No skill". (Adapted from Chang et al., 2019)

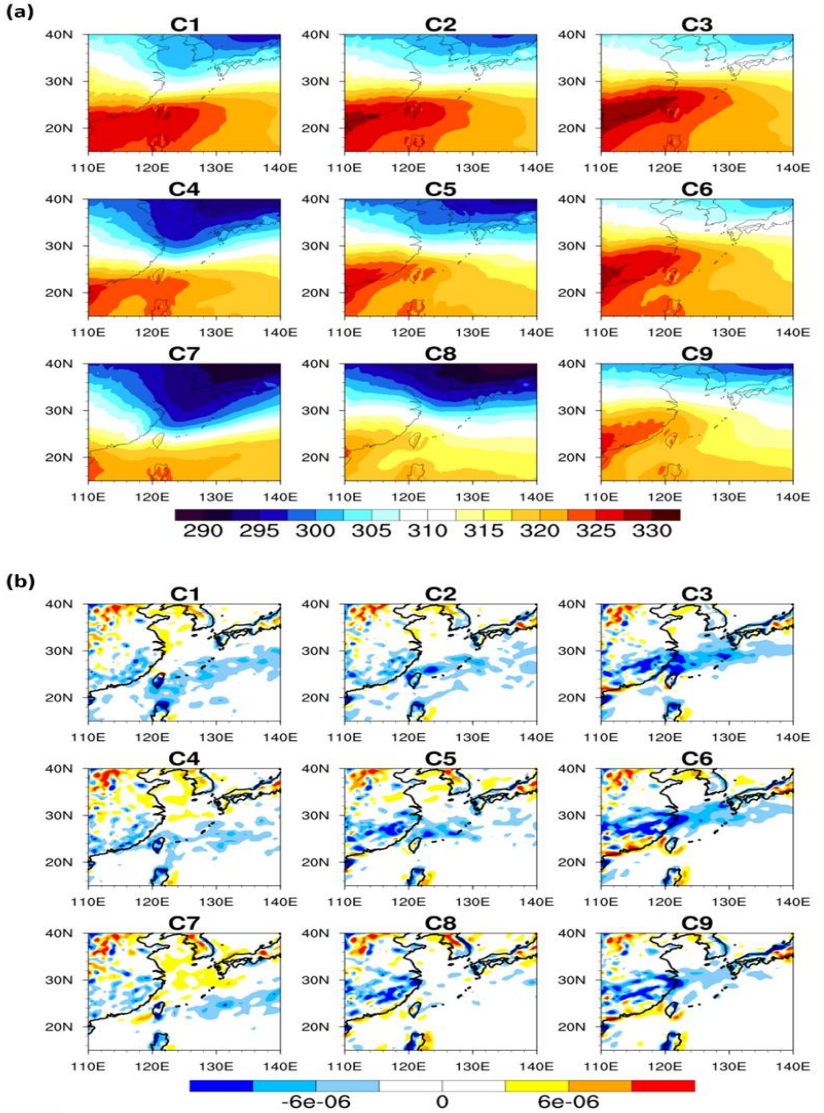

Figure 3. The clustering of the Mei-Yu fronts by SOM. (a) 850 hPa θ_e (units: K), (b) 850 hPa divergence field (units: s^{-1}). C1-C5 represents the primary configuration of Mei-Yu fronts. (Adapted from Chang et al., 2019)

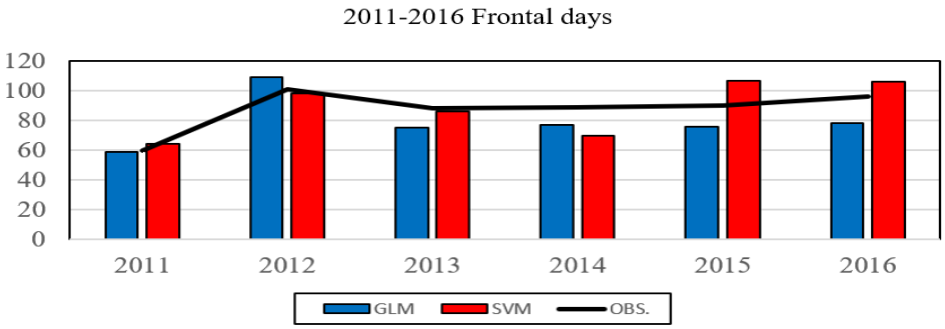

Figure 4. The number of frontal events in 2011-2016 analyzed by ML-based diagnostic tool. The black line denotes the number of frontal events observed by Central Weather Bureau, the blue and red bars are the number of the events identified by GLM and SLM. In this analysis, the frontal days is recognized as long as one of the 6-hourly surface weather maps or the ML-tool report the front system. The average (standard deviation) recognized by the observation and the ML models are 79 (16.3), 88.5 (18.4) and 98.3 (14.3) days. (Adapted from Chang et al., 2019)

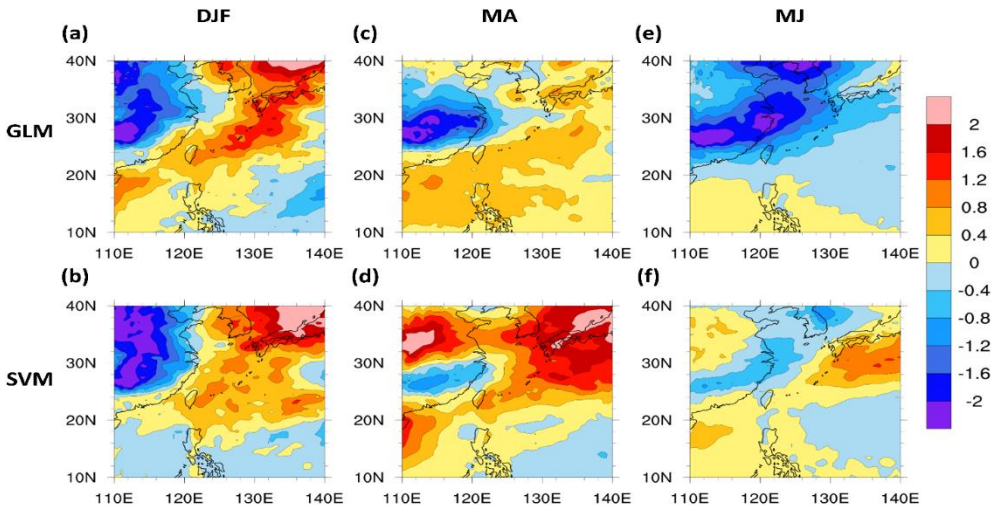

Figure 5. The temperature bias on 850 hPa. From the left to the right are DJF, MA and MJ. (a), (c), (e) are the temperature bias diagnosed by GLM and (b)、(d)、(e) are the temperature bias diagnosed by SVM. (units: K) (Adapted from Chang et al., 2019)

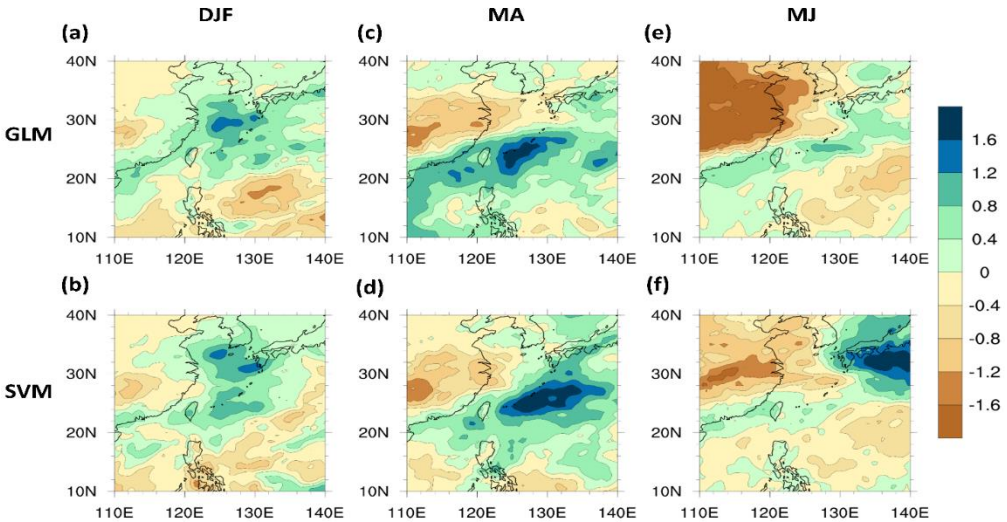

Figure 6. The specific humidity bias on 850 hPa. From the left to the right are DJF, MA and MJ. (a), (c), (e) are the temperature bias diagnosed by GLM and (b)、(d)、(e) are the specific humidity bias diagnosed by SVM. (units: g kg^{-1}) (Adapted from Chang et al., 2019)