Scientific Research Publishing

Atmospheric and Climate Sciences, 2021, 11, 284-301
https://www.scirp.org/journal/acs
ISSN Online: 2160-0422
ISSN Print: 2160-0414

Classified Early Warning and Forecast of Severe Convective Weather Based on LightGBM Algorithm

Xinwei Liu[1], Haixia Duan[2], Wubin Huang[1]*, Runxia Guo[1], Bolong Duan[1]

[1]Lanzhou Central Meteorological Observatory, Lanzhou, China
[2]Lanzhou Institute of Drought Meteorology, China Meteorological Administration, Lanzhou, China
Email: *hwb0707@sina.com

How to cite this paper: Liu, X.W., Duan, H.X., Huang, W.B., Guo, R.X. and Duan, B.L. (2021) Classified Early Warning and Forecast of Severe Convective Weather Based on LightGBM Algorithm. *Atmospheric and Climate Sciences*, **11**, 284-301.
https://doi.org/10.4236/acs.2021.112017

Received: February 5, 2021
Accepted: April 11, 2021
Published: April 14, 2021

Abstract

Severe convective weather can lead to a variety of disasters, but they are still difficult to be pre-warned and forecasted in the meteorological operation. This study generates a model based on the light gradient boosting machine (LightGBM) algorithm using C-band radar echo products and ground observations, to identify and classify three major types of severe convective weather (*i.e.*, hail, short-term heavy rain (STHR), convective gust (CG)). The model evaluations show the LightGBM model performs well in the training set (2011-2017) and the testing set (2018) with the overall false identification ratio (FIR) of only 4.9% and 7.0%, respectively. Furthermore, the average probability of detection (POD), critical success index (CSI) and false alarm ratio (FAR) for the three types of severe convective weather in two sample sets are over 85%, 65% and lower than 30%, respectively. The LightGBM model and the storm cell identification and tracking (SCIT) product are then used to forecast the severe convective weather 15 - 60 minutes in advance. The average POD, CSI and FAR for the forecasts of the three types of severe convective weather are 57.4%, 54.7% and 38.4%, respectively, which are significantly higher than those of the manual work. Among the three types of severe convective weather, the STHR has the highest POD and CSI and the lowest FAR, while the skill scores for the hail and CG are similar. Therefore, the LightGBM model constructed in this paper is able to identify, classify and forecast the three major types of severe convective weather automatically with relatively high accuracy, and has a broad application prospect in the future automatic meteorological operation.

Keywords

Severe Convective Weather, Machine Learning, LightGBM, Early Warning

and Forecast

1. Introduction

Severe convective weather usually refers to kinds of disastrous weather generated by deep moist convections, such as hail, gale, tornado and heavy precipitation [1]. Although there is no unified criterion for the definition of severe convective weather, the severe convective weather defined by the Central Meteorological Observatory of China Meteorological Administration refers to the event with any or several following weather conditions: hail with a diameter of 5 mm or above on the ground, tornado at any level on land, convective wind gust (CG) of more than 17 $m \cdot s^{-1}$ and short-term heavy rainfall (STHR) of 20 $mm \cdot h^{-1}$ or above [2]. Since the severe convective weather has strong destructiveness and often brings great harm to industry, agriculture and people's safety, its nowcasting and early warning play a great important role in the meteorological disaster prevention and mitigation. Moreover, severe convective weather occurs abruptly and locally with short duration, so it is still difficult to be early warned and forecasted in the meteorological operation. For example, the probability of detection of human forecasts of STHR, hail and CG are all less than 35%, while the false alarm ratio is even higher than 90% for the CG and hail in 2015-2017 [3]. Therefore, it is urgent to improve the forecast and early warning skills of severe convective weather in China and enhance the services of disaster prevention and mitigation.

With the continuous densifying of the Doppler weather radar network, the radar-echo products have been playing a key role in the monitoring, analysis and short-term early warning of severe convective weather [4] [5]. At present, most of the new generation Doppler weather radars in China are S-band and C-band radars. The S-band weather radar stations mainly are located in eastern China, and their data are the majority of radar products in researches with advantages such as the weak echo attenuation (He, 2012). The C-band radar stations mainly are located in western China, but their data are not fully studied and applied in researches and operations. Although the C-band radar products have problem of echo attenuation, studies have shown that the research results based on the products of S-band and C-band radar have good consistency in detecting rain [6], hail [7] and so on. Due to the complex topography in northwestern China with plateaus, mountains and deserts, disasters such as mountain torrent and debris flow are easy to occur in extreme severe convective weather, such as the heavy mountain torrent and debris flow disaster in Zhouqu, Gansu Province on August 8, 2010 and the disaster of heavy hail, mountain torrent and debris flow in Min County, Gansu Province on May 10, 2012. Therefore, it is very important for the meteorological operation in northwestern China to explore the application of C-band radar in the early warning and forecast of severe convective weather and enhance its warning and forecast skill.

At present, the early warning and nowcasting (0 - 2 hours) of severe convective weather are carried out mainly through the manual recognition of radar-echo and satellite images. Traditional methods of the early warning and forecast of severe convective weather mainly include the extrapolation forecast, the experimental forecast, the statistical forecast [8] [9] and the probability forecast [10], which have certain limitations, such as hysteresis, and low accuracy. Artificial intelligence and machine learning can classify and identify severe convective weather automatically, rapidly and systematically without artificial deviation. Therefore, the application of advanced big data, machine learning and artificial intelligence techniques in the short-term forecast of severe convective weather is one of the groundbreaking hotspots in meteorological research. For example, McGovern *et al.* [11] showed the applications of modern artificial intelligence techniques in forecasting a wide variety of high-impact weather phenomena, including storm duration, severe wind, severe hail, precipitation classification, renewable energy and aviation turbulence. Czerneckia *et al.* [12] applied machine learning to large hail predictions using the ERA5 data. Zhou *et al.* (2019) used a deep learning algorithm to forecast severe convective weather including hail, short-duration heavy rain, convective gust (CG) and thunderstorm, and found that the deep learning algorithm has a higher classification accuracy than support vector machine and random forest. Machine learning methods are also used to forecast the damage of straight-line wind [13], nowcast the 0 - 2 h storms [14] and lighting occurrence [15], diagnose aviation turbulence [16], map storm structures in advance [4] and even map the spatial distribution of soil organic matter [17].

The light gradient boosting machine (LightGBM) algorithm is a research hotspot in the data mining and classified prediction in recent years, and is widely used in the classification problems in all walks of life. For example, LightGBM is used in human activity recognition such as the safe driving [18], and the medical research such as the protein-protein interactions [19]. LightGBM is also widely used in the prediction of economics. Ma *et al.* [20] generated a prediction of peer-to-peer (P2P) network loan default based on the LightGBM and the extreme gradient boosting algorithms. Jiang *et al.* [21] predicted the directions of stock-price index using four machine-learning methods including the LightGBM. Sun *et al.* [22] used LightGBM to forecast the price trend of cryptocurrency market and found that the robustness of the LightGBM model is better than other methods. LightGBM has wide applications in atmospheric science as well, such as the prediction of air quality [23] [24] and wind power [25]. Moreover, Fan *et al.* [26] evaluated the LightGBM, random forest, the tree-based M5 Model Tree and four empirical models (Hargreaves-Samani, Tabari, Makkink and Trabert) to estimate daily reference evapotranspiration with local and external meteorological data, and pointed out that LightGBM generally performs better than other models.

Although some studies have applied machine learning in the classification and identification of a single type of severe convective weather [12] [13] [27], various types of severe convective weather often occur concomitantly. Therefore, this study develops a LightGBM model to identify and classify three major types of severe convective weather using C-band radar-echo data and ground observations of severe convective weather. The LightGBM model is tested and evaluated in the training set and the testing set of independent samples, respectively, and then applied in the forecast of severe convective weather. The LightGBM algorithm and its model construction are introduced in Section 2. Section 3 introduces the datasets used in this study, including the radar products and ground observations of severe convective weather, and the methods of model evaluation. The main results are given in Section 4, and the main conclusions are summarized and discussed in Section 5.

2. Methodology and Data

2.1 LightGBM Algorithm and Model Construction

The LightGBM is a gradient boosting decision tree (GBDT) algorithm framework proposed by Microsoft in 2017 [28], which aims to solve the problems of large time consumption and poor scalability in the calculation of high-dimensional large-sample data. It is essentially an ensemble learning algorithm to boost a weak learner to a strong one by combining many low-accuracy trees [25]. Through the continuous iteration and gradient descent method, LightGBM makes the loss function smaller and smaller by moving toward the negative gradient direction of the loss function in each iteration, and finally a superior tree is obtained and used as the prediction model [20] [29]. Before the LightGBM, there have been many efficient algorithms to achieve the GBDT [30], such as the extreme gradient boosting (XGBoost) algorithm. However, these algorithms show relatively low efficiency and cost much time when the data is high-dimensional with large sample size [25]. This is mainly because these algorithms need to iterate over all the data samples and then estimate the information gain of all the possible divide points. To solve this problem, the LightGBM adopts two innovative sampling algorithms, the exclusive feature bundling (EFB) and gradient-based one-side sampling (GOSS). The EFB algorithm reduces the number of features by binding mutually exclusive features, so the data feature scale is reduced and the model's training speed is improved. The GOSS algorithm excludes most samples with low gradient and estimates the information gain with the rest samples. The training amount is reduced while the information gain is guaranteed, and the model's generalization ability is enhanced. Therefore, compared with other traditional GBDT frameworks, the LightGBM has the advantages of high speed, memory saving and better generalization ability. Detailed information of the calculation procedures of LightGBM can be found in related literatures [19] [24] [26].

The construction steps of LightGBM model are shown in **Figure 1**. It is divided into five steps: data collection, feature engineering, model training, cross validation and model evaluation. The data collected in this study includes the observations of classified severe convective weather (dependent variable), radar-echo products and other ground meteorological observations during the period of severe convection (independent variables). Then the collected data are screened preliminarily by manual work to select the characteristic variables, which have the possible ability to identify the dependent variables. Then the collected data are put into the next step of feature engineering. The feature engineering is the most important part in constructing the LightGBM model, and finds the features of independent variable which can best reflect the essence of the dependent variables for further identification and classification. The feature engineering is applied over and over accompanied by the further data screening, which eliminates the useless samples with incomplete characteristic variables or singular values, and finally selects the most useful independent variables and also improves the data qualities of training and testing set. The processed data are then put into step 3 and 4 to establish the LightGBM model through the repetitive training, parameter adjustment (step 3) and cross validation (step 4) using the LightGBM algorithm (Figure 1). In the cross validation, the ratio of the model-training samples to the cross-validation samples is 8:2. After repetitive training and parameter adjustment, if the LightGBM model reaches the expected optimal performance, the model is then applied l evaluation in the training and testing set, respectively. If the LightGBM model into the next step of applications, such as the forecast, and also gives the results of mode does not reach the expected performance, the model needs to be re-built and back to step 2 (Figure 1).

2.2. Datasets

2.2.1. Independent Variables of LightGBM Model

Previous studies have shown that the Doppler weather radar products, such as reflectivity factor, echo intensity, top height and vertical liquid water content,

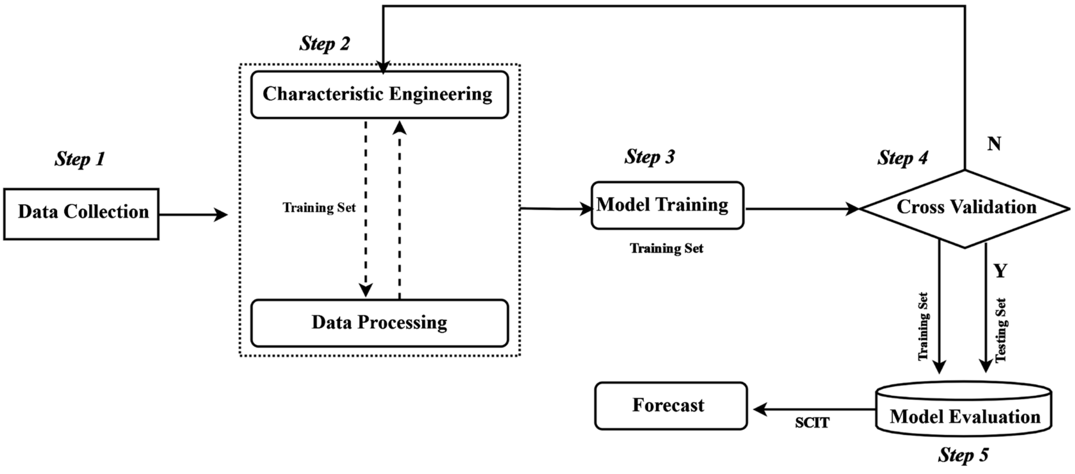


Figure 1. Flow map of the construction steps of LightGBM model.

can provide quantitative criteria for the severe convective weather [31]. Therefore, radar products are the major independent variables of the LightGBM model to classify and identify severe convective weather. The radar products include the reflectivity factor (R), combined reflectivity (CR), average radial velocity (V), echo top height (ET), storm top height (TOP), maximum echo height (HT) and vertical integrated liquid-water content (VIL). The quality control has been used for the R and CR by filtering the ground clutter through the membership function [32]. Due to the close relationship between the CG and the ground elements, the ground observations used in this paper include pressure (PRS), air temperature (TEM), relative humidity (RHU) and instantaneous wind speed (WIN), which are all from the ground-observation stations of CMA (China Meteorological Administration). The ground-observation stations in the study area include 49 national basic-reference stations, and 1112 intensified stations which are the unattended automatic meteorological stations measuring very limited variables such as precipitation and temperature.

The radar products used in this paper are from three C-band radars located in Lanzhou City, Tianshui City and Qingyang City of Gansu Province (**Figure 2**). The effective scanning radius of C-band radar is 150 km, but the top of the storm cannot be scanned if the distance is less than 30 km and the bottom of the storm cannot be scanned if the distance is greater than 120 km. Therefore, the optimal distance for the C-band radar is 30 - 120 km, which is adopted in this paper. The extraction methods of radar products are as follows.

The radar products R, CR and V are all from the precipitation model, with the spatial resolution of 1 km×1˚ (polar coordinates). The data at 3 elevation angels of 0.5˚, 1.5˚ and 2.4˚ are adopted for the R and the V (*i.e.*, 0.5˚ R, 1.5˚ R, 2.4˚ R, 0.5˚ V, 1.5˚ V and 2.4˚ V), so there are three sets of products for each. In the

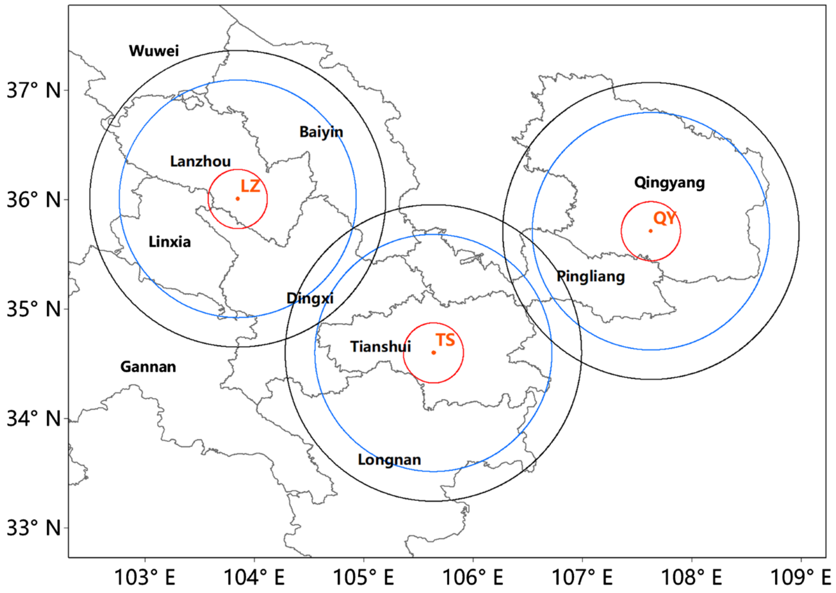


Figure 2. The scan area of Lanzhou, Tianshui and Qingyang Radar. The blue, red and black circle are the areas in 30 km, 120 km and 150 km radius, respectively.

extraction of product feature data, the pixel nearest to the observation point of the severe convective weather is taken as the center, and the maximum value within the surrounding 5 × 5 pixels is extracted as the storm characteristic value of the severe convective weather. The specific calculation formula is as follows.

$$\text{Value } n = \max\left(\text{Value } n\left(i, j\right)\right) \tag{1}$$

where n is the storm cell number, and the values of i and j are from −2 to 2.

The similar method is also used in the extraction of ET. The ET is the height of the echo with R ≥ 18.3 dBZ, which is the average value within the surrounding 5 × 5 pixels. The TOP is the maximum height of the storm with R ≥ 30 dBZ, which reflects the height of the strong echo top of the storm. The HT is the strong echo height calculated from the vertical difference of the strongest echo at each elevation angle, which is more accurate and credible than the maximum reflectivity height found at a single elevation angle.

The calculated variables based on the radar products include the core thickness (H), the centroid height with R above 45 dBZ (H_45) and the strong echo (R ≥ 45 dBZ) duration (Time). The VILD (Vertically Integrated Liquid Water Content Density) is the ratio of the VIL to the TOP, and the H is the height difference between the upper and lower boundaries of the area with R ≥ 45 dBZ in the storm.

In the meteorological operation, it is also necessary to track and forecast the classified severe convective weather. The storm cell identification and tracking (SCIT) [33] product of radar echo is one of the most representative strong convection identification algorithms at present, which can correctly track and identify 70% - 90% of storm cells, so this product is used to forecast the moving position of severe convective weather. In this paper, the forecast duration of the SCIT is the product of the forecast interval time (15 minutes) and the number of forecast intervals (1 - 4), which is 15 - 60 minutes.

2.2.2. Dependent Variables

The cases of severe convective weather and non-severe convective weather within the scanning radius of the three radars from 2011 to 2018 are collected in this paper as the dependent variables of LightGBM model. The collected severe convective weather cases are classified into three types firstly: hail, CG and STHR. Tornado rarely occurs in the study area, so it is not considered here. Then, according to the locations and occurring time of the severe convective weather, the corresponding radar-scan data in the same period of the severe convective weather are extracted as the samples of severe convective weather. For example, ground observations show that a station had hail during 14:28-14:58 (Beijing time, the same below), while the radar scan starts at 14:26 with an interval of 6 minutes, so five complete radar-scan data within the next 30 minutes (14:32, 14:38, 14:44, 14:50 and 14:56) are extracted as five hail samples. For the non-severe convective weather, samples are selected randomly from the radar-scan results during the period without severe convective weather, including various situations with no echo, weak echo and strong echo. In this way, the datasets of 5,741 samples of severe convective weather and 14,001 samples of non-severe convective weather during 2011-2018 are established (Table 1). Since the LightGBM model needs a large training set, 17,749 samples during 2011-2017 are taken as the training set (Table 1), which has 14,200 samples for the model training and 3,549 samples for the cross validation. The samples in 2018 are used as the testing set for independent validation including 541 samples of severe convective weather and 1452 samples of non-severe convective weather (Table 1). To facilitate the machine-language recognition, the events of hail, CG, STHR and non-severe convection are labeled. The classification labels for the non-severe convection, hail, CG and STHR are 0, 1, 2 and 3, respectively (Table 1).

2.3. Model Evaluation Methods

In the classification problem, the contingency table is often used to compare the observation and the prediction and using the FIR to show the evaluation results. The FIR used in this paper refers to the ratio of the number of false identifications of LightGBM model to the total sample number of the observed three types of severe convective weather, as shown in Equation (2). The overall FIR used in this paper is defined the same, which is the ratio of the total number of false identifications of all types of severe convective weather to the total number of samples.

$$\text{FIR}=\frac{\text{false identification number}}{\text{total sample number}}\times 100\% \tag{2}$$

The widely used probability of detection (POD), critical success index (CSI) and false alarm ratio (FAR) in the meteorological field are also applied to evaluate the identification and forecast effect of the severe convective weather (Lu *et al.* 2018). Their calculations in this study are as follows.

$$\text{POD}=\frac{\text{correct identification number}}{\text{total sample number}}\times 100\% \tag{3}$$

$$\text{CSI}=\frac{\text{correct identification number}}{\text{total sample number}+\text{false identification number}}\times 100\% \tag{4}$$

$$\text{FAR}=\frac{\text{false identification number}}{\text{correct identification number}+\text{false identification number}}\times 100\% \tag{5}$$

The inter-comparison between the model results based on radar products and

Table 1. The sample numbers of severe convective weather (SCW) for the training set (2011-2017) and the testing set (2018) used in the LightGBM model and their label values.

	Non-SCW	hail	CG	STHR
Training set	12,549	834	231	4135
Testing set	1452	51	12	478
Label value	0	1	2	3

the observations of severe convective weather is used to determine the identification of LightGBM model is correct or false. The identification time range is the duration of the observed events of severe convective weather, and the spatial range is 5˚ × 5 km (radar extraction radius, polar coordinates). For example, if an STHR occurs somewhere during 15:00-15:30, then the LightGBM model is used to identify all the radar products in this period within the 5˚ × 5 km area around the station. When an STHR is identified by the model, it is a correct classification. When other severe convective weather or non-severe convective weather is identified, it is a false identification. Therefore, the detecting time of LightGBM model for the samples in the training set and the test set is synchronous with the occurrence time of the observations, so it can be easily applied to the early warning system in the operation. For the accuracy of forecasts, the radar products used in LightGBM model are 15 - 60 minutes in advance. Therefore, for the forecast results of the model, the longest detection time is one hour. The identification criteria for the CSI and the FAR are similar to that of the POD.

3. Results

3.1. Characteristic Value Analysis

The feature engineering in the LightGBM modeling will find the characteristics of independent variable that can best reflect the essence of the dependent variables to classify samples, and allocate the weights of the independent variables in the calculation according to their importance [30]. Therefore, the importance of each independent variable to the classification result can be obtained. According to the results from LightGBM model, the characteristic values, namely, the importance scores of the independent variables, are ordered. The larger value indicates the greater importance (**Figure 3**).

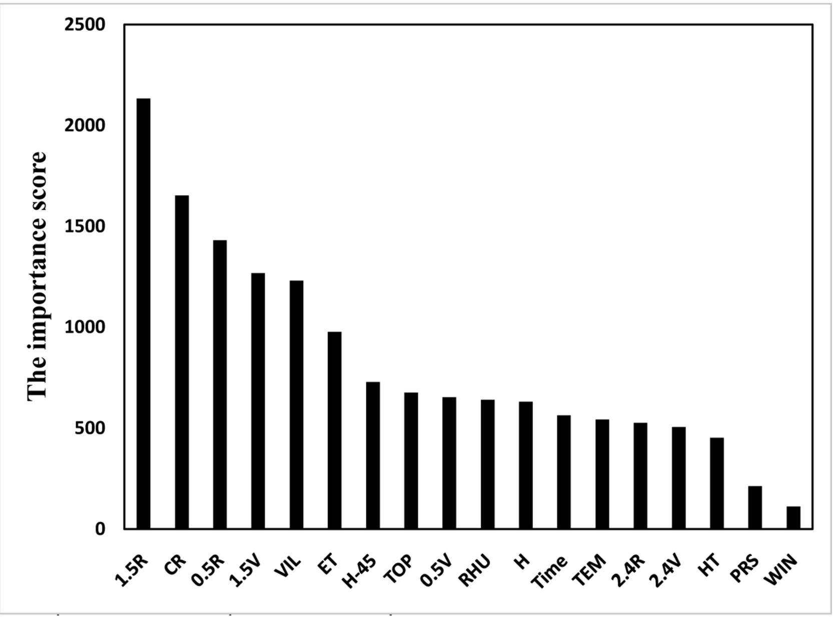


Figure 3. The importance scores of the independent variables of LightGBM model.

As shown in **Figure 3**, the importance of 1.5˚ R is highest, indicating that it contributes the most to the identification and classification of severe convective weather. The top six factors are 1.5˚ R, CR, 0.5˚ R, 1.5˚ V, VIL and ET, respectively, indicating that these characteristic variables are significantly different in different severe convective weather and can be easily identified and classified. These factors are also the most widely used radar products to identify severe convective weather in the meteorological operation and research. Thus, those results of LightGBM model are consistent with the actual forecast experience. The importance of the factors' characteristics below the sixth rank has similar order of magnitude, but it does not mean that these independent variables are not important. The permutation and combination of these independent variables are used comprehensively in the LightGBM model.

3.2. Model Evaluation in Training Set

The samples of severe convective weather in the training set from 2011 to 2017 are used to train the LightGBM model, and the model results for the training set is obtained, as shown in **Table 2**. For the three types of severe convective weather of hail, CG and STHR, the minimum and maximum FIRs are 6.2% (STHR) and 14.4% (hail), respectively. The STHR is mainly misclassified as non-severe convection weather and hail, and the hail is mainly misclassified as STHR. The FIR for the CG is 13.0%, and it is mainly misclassified as hail and STHR. The FIR for the non-severe convective weather is only 3.6%. If the total false identifications for all the types are divided by the total number of training set samples, the overall FIR of the LightGBM model for the severe convective weather is obtained, which is only 4.9%.

The skill scores of LightGBM model in the training set are shown in **Figure 4**. Among the three types of severe convective weather (*i.e.*, hail, CG and STHR), the STHR shows the highest POD and CSI, which are 93.8% and 84.4%, respectively, while the POD and CSI for the hail and the CG are similar. The STHR also has the lowest FAR of 10.6%. In summary, the average POD, CSI and FAR of the LightGBM model are 88.8%, 73.9% and 18.8%, respectively, indicating that the LightGBM model can achieve high accuracy and satisfactory result after sufficient training.

Table 2. The occurrence number of severe convective weather (SCW) in the training set (2011-2017) for the ground observations and the results of LightGBM model, and their false identification rates (FIR) and the overall FIR.

Obs.	Classification of LightGBM model					
	Hail	CG	STHR	Non-SCW	FIR	Overall FIR
Hail	143	5	19	0	14.4%	4.9%
CG	4	40	2	0	13.0%	
STHR	18	5	775	28	6.2%	
Non-SCW	17	3	71	2419	3.6%	

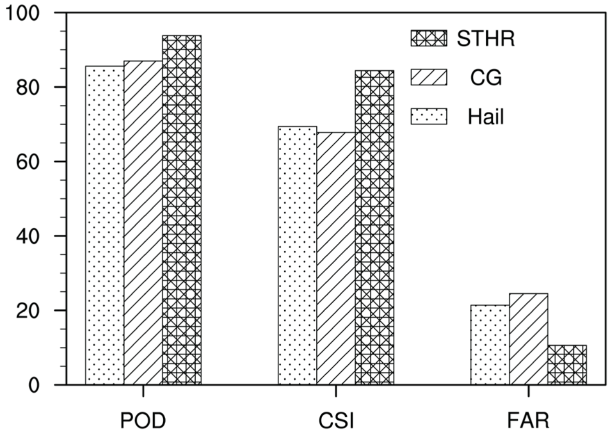


Figure 4. The POD, FAR and CSI scores for the results of LightGBM model in the training set (2011-2017).

Table 3. The occurrence number of severe convective weather (SCW) in the testing set (2018) for the ground observations and the identification results of LightGBM model, and their false identification rates (FIR) and the overall FIR.

Obs.	Classification of LightGBM model					
	Hail	CG	STHR	Non-SCW	FIR	Overall FIR
Hail	43	2	6	0	15.7%	7.0%
CG	1	10	1	0	16.7%	
STHR	12	2	438	26	8.4%	
Non-SCW	8	2	80	1362	6.2%	

3.3. Model Evaluation in Testing Set Using Independent Samples

The constructed LightGBM model is then applied to the testing set with 1993 independent samples in 2018 for further evaluations. Results in Table 3 show that, in the independent validation, the FIRs for hail, CG and STHR are 15.7%, 16.7% and 8.4%, respectively, and the non-severe convective weather shows the lowest FIR of 6.2%. Therefore, the overall FIR is 7.0% for the testing set. Similar to the results of the training set, the hail is mainly misclassified as STHR, and the STHR is mainly misclassified as non-severe convection and hail. The skill scores for three types of severe convective weather are shown in Figure 5. Although the skill scores of the independent testing set are slightly lower than those in the training set, the average POD still reaches 86.4% (Figure 5). The average CSI and FAR are 64.3% and 29.0%, respectively. In addition, the POD and CSI for the STHR are also the highest with the lowest FAR, and the skill scores of the CG and the hail are similar (Figure 5).

3.4. Model Application in Forecast

Based on the fact that LightGBM model can identify three types of severe convective weather well, the SCIT products are then used in LightGBM model to forecast the type and position of the severe convective weather in advance of 15 -

60 minutes. The forecast samples are 45 severe convective weather events in 2020. In addition, although the SCIT products can only forecast the moving position of the storm centroid, the spatial range of the forecast is 5˚ × 5 km (polar coordinates), which is the radar extraction radius and the detection range of the model.

According to the skill scores of forecast (**Figure 6**), three major types of convective weather in the future 15 - 60 minutes can generally be forecasted based on the LightGBM model and the SCIT products. In the classified forecast of severe convective weather, the STHR shows the highest POD of 63.8% and CSI of 64.5%, followed by the CG. The STHR also shows the lowest FAR of 30.9%, followed by the hail. The average POD, CSI and FAR for the three types of severe convective weather are 57.4%, 54.7% and 38.4%, respectively. Although the forecast of severe convective weather carried out by LightGBM model is less accurate than the classification and identification in the training and testing set, it can still meet the need of classified nowcasting for the severe convective weather in the operation of weather forecast with the advantages of higher efficiency and automation.

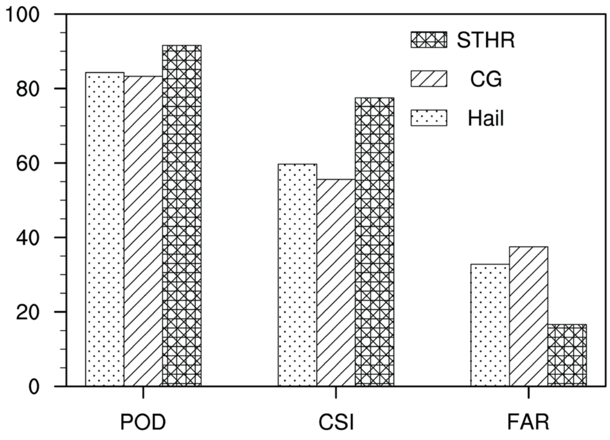


Figure 5. The POD, FAR and CSI scores for the results of LightGBM model in the testing set (2018).

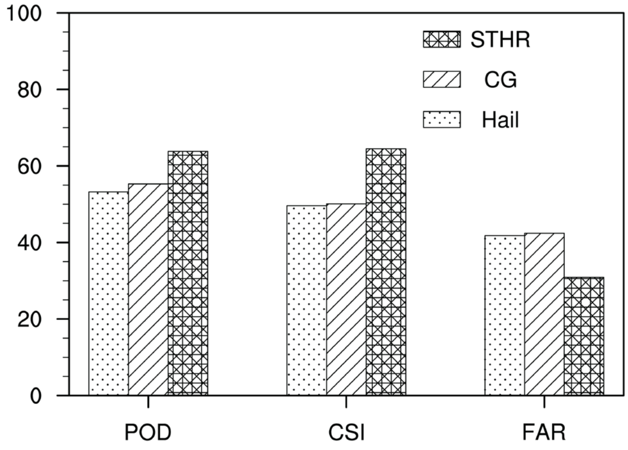
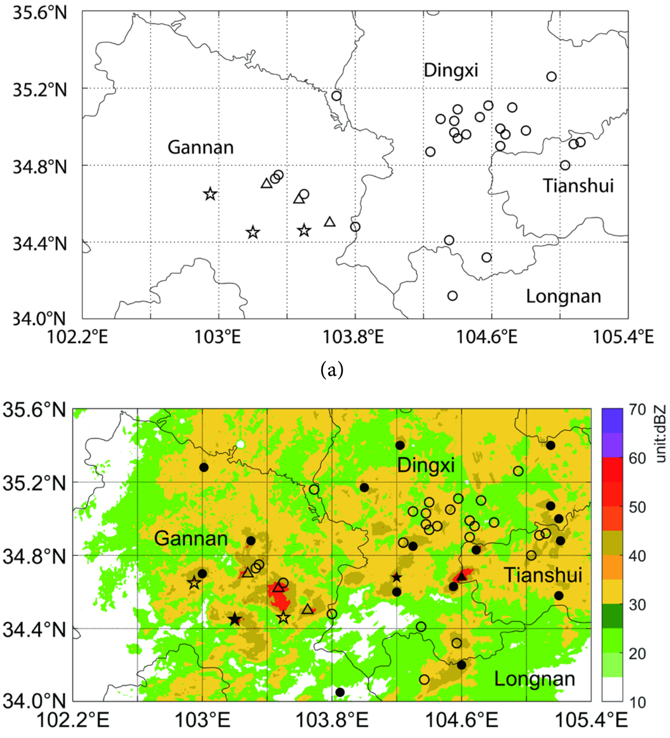


Figure 6. The POD, FAR and CSI scores for the results of LightGBM model in 2020.

In order to further illustrate the forecast performance of the LightGBM model and the SCIT products, a case of severe convection episode in 2020 is demonstrated in Figure 6. In the observation, scattered convections begin to develop in Gannan and Linxia areas at 16:00 on May 6th 2020. From 17:00 to 19:00, three stations in the southeast of Gannan had hail and CG, and 26 stations in Gannan, Dingxi, Tianshui and the northwest of Longnan had STHR (Figure 7(a)). Thus, the LightGBM model and SCIT products are used since 17:00 to forecast and classify the severe convective weather 15 - 60 minutes in advance. According to the forecast results (Figure 7(b)), LightGBM model correctly forecasts all the events of severe convective weather, but falsely alarms one record of hail in the south of Dingxi, one record of CG in the south of Gannan and 16 records of STHR in Gannan, Dingxi, Tianshui and Longnan. Although there are false alarms in LightGBM model for all the three types of severe convective weather, there is no missing report.

4. Discussion

The LightGBM model generated in this study can classify and identify three major types of severe convective weather rapidly and automatically, and its performances in accuracy and false alarm are also relatively better than other systems. For



(a)

(b)

Figure 7. The ground observations (a) and the forecast results of LightGBM model 15 - 60 minutes in advance (b) for the severe convective weather on May 6th 2020. The triangle, circle and pentagram are the hail, STHR and CG, respectively, and their hollow patterns represent the right classification, and the solid patterns represent the false classification. The colorful shadow is the CR value (unit: dBZ).

example, the STHR and hail identified by LightGBM model in the training and testing set have higher POD and lower FAR than those of the logistic regression, the random forest, the support vector machine algorithm, the multilayer perceptron and the deep convolutional neural network based on numerical weather prediction data [3]. The forecast POD of hail, CG and STHR of LightGBM model are 53.2%, 55.3% and 63.8%, respectively, which are all higher than those of the 12-h in advance forecast of the deep convolutional neural network (around 25%, 30% and 55%, respectively) [3]. The mean forecast FARs of LightGBM model is lower than 40%, which is over 50% for the STHR and over 90% for the CG and hail in the study of Zhou *et al.* (2019). In general, the LightGBM model is satisfactory for the severe convective weather classification, and has the advantages of high speed and less computing resource consumption, which can significantly improve the early warning and forecast skill of severe convective weather, and has a broad application prospect in the future automatic meteorological operation.

The generally ideal performance of LightGBM model is mainly because it is trained by a large number of radar products and observations. The application of extensive radar products and intensified ground-observation data is the key and main feature of this paper. Although there are plenty of Doppler weather radars in the central and western regions of China, the secondary development and application of the radar products in the actual meteorological operation are still relatively insufficient, so the abundant radar products accumulated during the historical period have not been fully utilized. This paper makes full use of these precious data, and extends the application value of the C-band radar products. However, there are still some deficiencies in the LightGBM model generated in this study, such as the limited sample number of severe convective weather and the insufficient quality control of the radar products, which need more effort in the future for further improvements.

5. Conclusions

This study constructs a LightGBM model based on the C-band radar-echo data to identify, classify and forecast three major types of severe convective weather (hail, CG and STHR). The model performances are evaluated by the training set and testing set, and then applied in the forecast of severe convective weather. The main conclusions are as follows.

In the training set during 2011-2017, the overall FIR of the LightGBM model for three types of severe convective weather is 4.9%. The FIRs for the STHR and the hail are 6.2% and 14.4%, respectively, which are the minimum and the maximum. The FIR for the non-severe convective weather is only 3.6%. In terms of the skill scores, the average POD, CSI and FAR for the three types of severe convective weather are 88.8%, 73.9% and 18.8%, respectively. Therefore, the LightGBM model constructed based on the C-band radar products and the ground observations of severe convective weather is satisfactory to the accuracy of meteorological operation.

The evaluations of LightGBM model using the testing set with independent samples in 2018 show that the overall FIR for the three types of severe convective weather and non-severe convective weather is 7.0%. The FIRs for the hail, the CG and the STHR are 15.7%, 16.7% and 8.4%, respectively. The skill scores for the testing set are lower than those in the training set, but the average POD is still up to 86.4%, with the average CSI of 64.3% and the average FAR of 29.0%. The STHR shows the highest POD and CSI and the lowest FAR, and the skill scores for the CG and the hail are similar. Therefore, the LightGBM model demonstrates relatively accurate classification and identification of severe convective weather.

The SCIT products and the LightGBM model are used to forecast the types and locations of severe convective weather in advance 15 - 60 minutes. The results show that the average POD, CSI and FAR for the three types of severe convective weather are 57.4%, 54.7% and 38.4%, respectively. Similar to the results of training set and testing set, the STHR shows the highest POD and CSI, which are 63.8% and 64.5%, respectively, and it also shows the lowest FAR (30.9%).

In general, the classification of severe convective weather based on the LightGBM model is ideal. The LightGBM algorithm used in this study is more advanced than neural network, multiple linear regression and other methods commonly used in meteorological field, and has more advantages than SVM, logistic and other machine learning methods. Therefore, this method has a broad application prospect in the future automatic identification and early warning of severe convective weather. However, there are still some deficiencies in the LightGBM model generated in this study, such as the limited sample number of severe convective weather and the insufficient quality control of the radar products, which need more effort in the future for further improvements.

Data Availability

The radar products and ground observations used to support the findings of this study were supplied by the Lanzhou Central Meteorological Observatory, so cannot be made freely available. Requests for access to these data should be made to the Lanzhou Central Meteorological Observatory, hwb0707@sina.com.

Funding Statement

This research was funded by the National Key R&D Program of China, grant number 2017YFC1502101, the Innovation-team Project of Gansu Meteorology Administration, grant number GSQXCXTD-2020-01 and the Arid Meteorological Science Research Foundation, and grant number IAM202011.

Acknowledgements

We thank the founding support of Dr. Yanhong Gao (Department of Atmospheric and Oceanic Sciences & Institute of Atmospheric Sciences, Fudan University, Shanghai, China), and Yongjie Pan (Key Laboratory of Land Surface Process and Climate Change in Cold and Arid Regions, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou, China). We also thank the help of our colleagues in data collection, and the free access of the LightGBM algorithm.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Doswell, C.A. (2015) Severe Convective Storms in the European Societal Context. *Atmospheric Research*, **158-159**, 210-215. https://doi.org/10.1016/j.atmosres.2014.08.007

[2] Yu, X. and Zheng, Y. (2020) Advances in Severe Convection Research and Operation in China. *Journal of Meteorological Research*, **34**, 189-217. https://doi.org/10.1007/s13351-020-9875-2

[3] Zhou, K., Zheng, Y., Li, B., Dong, W. and Zhang, X. (2019) Forecasting Different Types of Convective Weather: A Deep Learning Approach. *Journal of Meteorological Research*, **33**, 797-809. https://doi.org/10.1007/s13351-019-8162-6

[4] Han, L., Sun, J.Z. and Zhang, W. (2019) Convolutional Neural Network for Convective Storm Nowcasting Using 3-D Doppler Weather Radar Data. *IEEE Transactions on Geoscience and Remote Sensing*, **58**, 1487-1495. https://doi.org/10.1109/TGRS.2019.2948070

[5] Prudden, R., Adams, S., Kangin, D., Robinson, N., Ravuri, S., Mohamed, S. and Arribas, A. (2020) A Review of Radar-Based Nowcasting of Precipitation and Applicable Machine Learning Techniques. arXiv:2005.04988.

[6] Aydin, K. and Giridhar, V. (1992) C-Band Dual-Polarization Radar Observables in Rain. *Journal of Atmospheric and Oceanic Technology*, **9**, 383-390. https://doi.org/10.1175/1520-0426(1992)009%3C0383:CBDPRO%3E2.0.CO;2

[7] Féral, L., Sauvageot, H. and Soula, S. (2003) Hail Detection Using S- and C-Band Radar Reflectivity Difference. *Journal of Atmospheric and Oceanic Technology*, **20**, 233-248. https://doi.org/10.1175/1520-0426(2003)020%3C0233:HDUSAC%3E2.0.CO;2

[8] Seed, A.W. (2003) A Dynamic and Spatial Scaling Approach to Advection Forecasting. *Journal of Applied Meteorology and Climatology*, **42**, 381-388. https://doi.org/10.1175/1520-0450(2003)042%3C0381:ADASSA%3E2.0.CO;2

[9] Fox, N.I. and Wikle, C.K. (2005) A Bayesian Quantitative Precipitation Nowcast Scheme. *Weather and Forecasting*, **20**, 264-275. https://doi.org/10.1175/WAF845.1

[10] Mecikalski, J.R., Williams, J.K., Jewett, C.P., Ahijevych, D., LeRoy, A. and Walker, J.R. (2015) Probabilistic 0-1-h Convective Initiation Nowcasts That Combine Geostationary Satellite Observations and Numerical Weather Prediction Model Data. *Journal of Applied Meteorology and Climatology*, **54**, 1039-1059. https://doi.org/10.1175/JAMC-D-14-0129.1

[11] McGovern, A., Elmore, K.L., Gagne, D.J., Haupt, S.E., Karstens, C.D., Lagerquist, R., Smith, T. and Williams, J.K. (2017) Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather. *Bulletin of the American*

Meteorological Society, **98**, 2073-2090. https://doi.org/10.1175/BAMS-D-16-0123.1

[12] Czernecki, B., Taszarek, M., Marosz, M., Półrolniczak, M., Kolendowicz, L., Wyszogrodzki, A. and Szturc, J. (2019) Application of Machine Learning to Large hail Prediction—The Importance of Radar Reflectivity, Lightning Occurrence and Convective Parameters Derived from ERA5. *Atmospheric Research*, **227**, 249-262. https://doi.org/10.1016/j.atmosres.2019.05.010

[13] Lagerquist, R., McGovern, A. and Smith, T. (2017) Machine Learning for Real-Time Prediction of Damaging Straight-Line Convective Wind. *Weather and Forecasting*, **32**, 2175-2193. https://doi.org/10.1175/WAF-D-17-0038.1

[14] Łoś, M., Smolak, K., Guerova, G. and Rohm, W. (2020) GNSS-Based Machine Learning Storm Nowcasting. *Remote Sensing*, **12**, Article No. 2356. https://doi.org/10.3390/rs12162536

[15] Mostajabi, A., Finney, D.L., Rubinstein, M. and Rachidi, F. (2019) Nowcasting Lightning Occurrence from Commonly Available Meteorological Parameters Using Machine Learning Techniques. *npj Climate and Atmospheric Science*, **2**, Article No. 41. https://doi.org/10.1038/s41612-019-0098-0

[16] Williams, J.K. (2014) Using Random Forests to Diagnose Aviation Turbulence. *Machine Learning*, **95**, 51-70. https://doi.org/10.1007/s10994-013-5346-7

[17] Wiesmeier, M., Barthold, F., Blank, B. and Kögel-Knabner, I. (2011) Digital Mapping of Soil Organic Matter Stocks Using Random Forest Modeling in a Semi-Arid steppe Ecosystem. *Plant and Soil*, **340**, 7-24. https://doi.org/10.1007/s11104-010-0425-z

[18] Gao, X., Luo, H., Wang, Q., Zhao, F., Ye, L. and Zhang, Y. (2019) A Human Activity Recognition Algorithm Based on Stacking Denoising Autoencoder and LightGBM. *Sensors*, **19**, Article No. 947. https://doi.org/10.3390/s19040947

[19] Chen, C., Zhang, Q., Ma, Q. and Yu, B. (2019) LightGBM-PPI: Predicting Protein-Protein Interactions through LightGBM with Multi-Information Fusion. *Chemometrics and Intelligent Laboratory Systems*, **191**, 54-64. https://doi.org/10.1016/j.chemolab.2019.06.003

[20] Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q. and Niu, X. (2018) Study on a Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms According to Different High Dimensional Data Cleaning. *Electronic Commerce Research and Applications*, **31**, 24-39. https://doi.org/10.1016/j.elerap.2018.08.002

[21] Jiang, M., Liu, J., Zhang, L. and Liu, C. (2020) An Improved Stacking Framework for Stock Index Prediction by Leveraging Tree-Based Ensemble Models and Deep Learning Algorithms. *Physica A*: *Statistical Mechanics and its Applications*, **541**, Article ID: 122272. https://doi.org/10.1016/j.physa.2019.122272

[22] Sun, X., Liu, M. and Sima, Z. (2020) A Novel Cryptocurrency Price Trend Forecasting Model Based on LightGBM. *Finance Research Letters*, **32**, Article ID: 101084. https://doi.org/10.1016/j.frl.2018.12.032

[23] Zhang, C., Wu, M., Chen, J., Chen, K., Zhang, C., Xie, C., Huang, B. and He, Z. (2019) Weather Visibility Prediction Based on Multimodal Fusion. *IEEE Access*, **7**, 74776-74786. https://doi.org/10.1109/ACCESS.2019.2920865

[24] Zhang, Y., Zhang, R., Ma, Q., Wang, Y., Wang, Q., Huang, Z. and Huang, L. (2020) A Feature Selection and Multi-Model Fusion-Based Approach of Predicting Air Quality. *ISA Transactions*, **100**, 210-220. https://doi.org/10.1016/j.isatra.2019.11.023

[25] Ju, Y., Sun, G., Chen, Q., Zhang, M., Zhu, H. and Rehman, M.U. (2019) A Model Combining Convolutional Neural Network and LightGBM Algorithm for Ul-

tra-Short-Term Wind Power Forecasting. *IEEE Access*, **7**, 28309-28318. https://doi.org/10.1109/ACCESS.2019.2901920

[26] Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X. and Zeng, W. (2019) Light Gradient Boosting Machine: An Efficient Soft Computing Model for Estimating Daily Reference Evapotranspiration with Local and External Meteorological Data. *Agricultural Water Management*, **225**, Article ID: 105758. https://doi.org/10.1016/j.agwat.2019.105758

[27] Gagne, D.J., McGovern, A., Haupt, S.E., Sobash, R.A., Williams, J.K. and Xue, M. (2017) Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Weather and Forecasting*, **32**, 1819-1840. https://doi.org/10.1175/WAF-D-17-0010.1

[28] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and, Liu, T.-Y. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, CA, December 2017, 3149-3157.

[29] Yasser, K. and Hemayed, E. (2017) Novelty Detection for Location Prediction Problems Using Boosting Trees. 2017 *International Conference on Computational Science and Its Applications*, Trieste, 3-6 July 2017, 173-182. https://doi.org/10.1007/978-3-319-62395-5_13

[30] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203450

[31] Kaltenboeck, R. and Ryzhkov, A. (2013) Comparison of Polarimetric Signatures of Hail at S and C Bands for Different Hail Sizes. *Atmospheric Research*, **123**, 323-336. https://doi.org/10.1016/j.atmosres.2012.05.013

[32] Marzano, F.S., Scaranari, D., Montopoli, M. and Vulpiani, G. (2008) Supervised Classification and Estimation of Hydrometeors from C-Band Dual-Polarized Radars: A Bayesian Approach. *IEEE Transactions on Geoscience and Remote Sensing*, **46**, 85-98. https://doi.org/10.1109/TGRS.2007.906476

[33] Johnson, J.T., Mackeen, P.L., Witt, A., Mitchell, E.D., Stumpf, G.J., Eilts, M.D. and Thomas, K.W. (1998) The Storm Cell Identification and Tracking Algorithm: An Enhanced WSR-88D Algorithm. *Weather and Forecasting*, **13**, 263-276. https://doi.org/10.1175/1520-0434(1998)013%3C0263:TSCIAT%3E2.0.CO;2