Characteristics of Hail in Inner Mongolia and Research on Machine Learning-Based Hail Identification Methods: Postprint

Authors: Xin Yue, Su Lijuan, Xucheng Zheng, Li Hui, Yi Nana, Jin Yuchen

Date: 2025-02-27T00:00:00+00:00

Abstract

Using manual hail observation records from Inner Mongolia for the period 1959–2021, this study analyzes the spatiotemporal characteristics of hail distribution and constructs a hail identification method based on machine learning algorithms. The results show that: (1) In terms of temporal distribution, both the number of stations experiencing hail events and the station-day count exhibit a decreasing trend; in terms of spatial distribution, hail is mainly concentrated in the Yinshan Mountains and the Greater Khingan Range, with hail-prone areas extending along the mountain ranges. (2) Hail occurrence exhibits distinct seasonal and diurnal variation characteristics; May–September is the period of frequent hail activity each year, accounting for 91.79% of annual hail days, and within hail days, 12:00–19:00 is the primary time period for hail occurrence. (3) Using four machine learning algorithms—Random Forest, LightGBM, K-Nearest Neighbors, and Decision Tree—through steps including data preprocessing, predictor selection, model training, and model optimization, hail weather processes in Inner Mongolia are modeled and evaluated. Evaluation results indicate that the machine learning approach can effectively identify hail weather processes, with TS scores of all models reaching above 0.83 and hit rates exceeding 92%; the Random Forest algorithm demonstrates the best identification performance on the test set. The research findings can provide references for hail forecasting and warning as well as artificial hail suppression operations in Inner Mongolia.

Full Text

Hail Characteristics and a Machine Learning-Based Hail Identification Method in Inner Mongolia

XIN Yue[1], SU Lijuan[1], ZHENG Xucheng[1], LI Hui[1], YI Nana[1], JIN Yuchen[2] [1]Inner Mongolia Weather Modification Center, Hohhot 010051, In-

ner Mongolia, China [2]Inner Mongolia Meteorological Science Institute, Hohhot 010051, Inner Mongolia, China

Abstract

Based on manual hail observation records from Inner Mongolia, China, spanning 1959–2021, this study analyzes the spatiotemporal distribution characteristics of hail and constructs a hail identification method using machine learning algorithms. The results reveal three key findings. First, regarding temporal distribution, both the number of hail-affected stations and station-days exhibit a declining trend. Spatially, hail events are predominantly concentrated in the Yinshan Mountains and Greater Hinggan Mountains, with hail-prone areas extending along these mountain ranges. Second, hail occurrence shows pronounced seasonal and diurnal variations, with May–September accounting for 91.79% of annual hail days. The most frequent hail period occurs between 12:00 and 19:00 Beijing Standard Time. Third, four machine learning algorithms—random forest, LightGBM, K-nearest neighbors, and decision tree—were employed to model and evaluate hail weather processes in Inner Mongolia through systematic steps of data preprocessing, predictor selection, model training, and hyperparameter tuning. Evaluation results demonstrate that machine learning methods effectively identify hail weather processes, with all models achieving threat scores above 0.83 and hit rates exceeding 92%. Among them, the random forest algorithm delivers the optimal identification performance on the test set. These findings provide valuable references for hail forecasting, early warning, and artificial hail suppression operations in Inner Mongolia.

Keywords: hail station-days; spatiotemporal characteristics; machine learning; hail identification

Introduction

Hail represents a short-duration, high-impact weather phenomenon that inflicts severe damage on agricultural production, transportation infrastructure, buildings, and even human lives and property. To better understand hail formation patterns and enable advanced forecasting and early warning, Chinese scholars have conducted extensive research on hail' s climatic spatiotemporal evolution, identification methodologies, and formation mechanisms. Zhang et al. analyzed hail observation data from 2,351 stations nationwide, revealing that China' s primary hail-prone regions are the Tibetan Plateau, northern North China, and Northeast China, with a significant decreasing trend in hail days over recent decades. Their study also identified marked seasonal and diurnal variations, with summer and winter representing the peak and minimum hail seasons, respectively, and hail occurrence concentrated in the afternoon to evening hours. Tang et al. investigated the spatiotemporal distribution characteristics of hail disaster events across China from 2012 to 2020, obtaining results consistent with Zhang et al. Wei et al. conducted a statistical analysis of hail characteristics in the Tianjin region over 11 years, comparing environmental conditions

across different months, weather patterns, and hail sizes to establish representative environmental parameter thresholds for hail forecasting. Zhong et al. analyzed 143 hail events in Hubei Province using NCEP GFS data, employing dichotomous and continuous probability methods to develop a hail probability forecasting model for different synoptic situations that has been implemented operationally. Liu et al. utilized random forest and LightGBM algorithms to classify and nowcast severe convective weather events including hail using radar products.

Hail is a highly localized hazardous weather phenomenon, with varying characteristics and activity patterns across different regions, necessitating region-specific identification parameters. Inner Mongolia spans a vast east-west territory with diverse climate conditions and complex topography, where large-scale severe convective weather events including hail, thunderstorms, and strong winds frequently occur. For Inner Mongolia, hail disasters rank second only to drought in terms of socioeconomic impact. Therefore, this study employs statistical methods to investigate the variation patterns and spatiotemporal distribution characteristics of hail disasters in Inner Mongolia based on manual hail observation records. Through modeling and hail identification using multiple machine learning approaches, we compare the recognition effectiveness of different methods to further improve hail forecasting accuracy and provide references for hail forecasting, early warning, and artificial hail suppression operations in the region.

1.1 Study Area Overview

Inner Mongolia is located in northern China' s border region, extending from 37°24 N to 53°23 N and 97°12 E to 126°04 E, covering a total area of 1.18$\times$10^{6}$ km^2. The region comprises 12 leagues and municipalities, with Alxa League, Wuhai City, Ordos City, Bayannur City, and Baotou City constituting western Inner Mongolia; Hohhot City, Ulanqab City, and Xilingol League representing central Inner Mongolia; and Chifeng City, Tongliao City, Hinggan League, and Hulunbuir City forming eastern Inner Mongolia. The geomorphological pattern features alternating belts of plains, mountains, and plateaus, dominated by plateaus surrounded by the Greater Hinggan and Yinshan mountain ranges. Influenced by mid-latitude westerly flows and characterized by a temperate continental monsoon climate, Inner Mongolia frequently experiences the convergence of cold and warm air masses, making it a region prone to severe convective weather such as hail.

1.2 Data Sources

The study utilizes three main data sources: (1) Special weather observation records from surface meteorological stations during 1959–2021; (2) Hail records from hail suppression operation sites in Bayannur City during 2014–2021; and (3) Hourly rainfall data from surface observations during 2014–2021 for screen-

ing hail and non-hail cases. Additionally, ERA5 reanalysis data are employed to calculate atmospheric thermodynamic, dynamic, and moisture parameters under various weather conditions, which serve as predictors for machine learning models.

1.3 Construction of Machine Learning Label Dataset

Positive hail samples were primarily derived from hail processes recorded at surface meteorological stations and hail suppression operation sites, cross-validated with Doppler radar products. Given that severe convective weather events often occur simultaneously, cases were classified based on disaster severity. When both hail and thunderstorms were observed, the event was recorded as a hail process; when hail and strong winds co-occurred, it was generally classified as hail. Since hail suppression operation records only documented township-level hail occurrence times without specific coordinates, the location with the strongest radar composite reflectivity factor within the administrative region at the hail time was designated as the hail point. This yielded 1,120 positive samples. Negative samples comprised thunderstorms, strong winds, short-duration heavy precipitation, and single weather-type cases. Thunderstorm and strong wind observations came from special weather records at surface stations, while short-duration heavy precipitation cases were selected from typical rainfall processes in weather modification operation records, with hourly rainfall exceeding 20 $mm \cdot h^{-1}$ defined as short-duration heavy precipitation. This process collected 1,120 negative samples.

1.4 Machine Learning Methods

To compare the hail identification capabilities of different algorithms, we selected four commonly used and effective machine learning algorithms from previous studies: random forest, LightGBM, K-nearest neighbors, and decision tree. The hail forecasting problem was formulated as a binary classification task. The overall training approach involved: (1) Using labeled hail and non-hail data from the training set; (2) Establishing feature engineering with different algorithms; (3) Conducting iterative training, parameter tuning, and cross-validation to determine optimal hyperparameter combinations; (4) Optimizing model algorithms continuously; and (5) Evaluating models on the test set using hit rate, false alarm rate, miss rate, and threat score metrics.

2.1.1 Temporal Distribution Characteristics of Hail

Following previous definitions of hail occurrence frequency, a station-day was counted when hail was observed at a station on a given day, regardless of duration. Inner Mongolia averaged 45 stations with hail annually, totaling 1,120 hail station-days. The year 1969 recorded the maximum number of hail-affected stations (101), while 2021 had the minimum (13). Although peak and valley years for station-days and station numbers did not perfectly coincide, 1969 also

saw the highest number of hail station-days (254), with 1971 recording the lowest (20). Both metrics show a declining trend, with station-days decreasing at an average rate of 2.3 per decade and station numbers decreasing by 0.6 per decade, consistent with the nationwide hail frequency reduction trend.

Seasonally, hail is concentrated in summer (June–August), accounting for 62.32% of annual hail station-days, followed by autumn (19.41%) and spring (16.33%). Winter has the fewest hail events (1.94%). Monthly analysis reveals that May–September encompasses 91.79% of hail processes, with July, June, and August ranking as the top three months, representing 25.45%, 20.54%, and 16.33% of annual hail processes, respectively. No hail was observed in February.

Diurnally, hail occurrence exhibits a single-peak structure [Figure 2: see original paper]. Hail is rare between 00:00–08:00 and 21:00–24:00, with frequencies increasing rapidly after 10:00, peaking at 15:00 (16.3% of daily occurrences). The primary hail period occurs between 12:00–18:00, representing 62.3% of daily hail events. This pattern aligns with previous studies on hail diurnal characteristics in Hohhot and Chifeng, attributed to enhanced solar radiation, rising temperatures, decreasing pressure, and increased thermal instability in the near-surface atmosphere during afternoon hours, which favors the development of mesoscale convective systems.

2.1.2 Spatial Distribution Characteristics of Hail

Among the 119 surface meteorological stations, the top three stations by total hail days are Qahar Right Front Banner in Ulanqab City (254 days), Wuchuan County in Hohhot City (212 days), and Taibus Banner in Xilingol League (245 days)—all located in central Inner Mongolia. Statistical analysis of average annual hail days shows that 43 stations (36.13%) experience more than 1.0 hail day per year, while 76 stations have fewer than 1.0 day. Hail-prone areas are primarily situated along the Yinshan Mountains and Greater Hinggan Range, extending along mountain ranges. Grassland and desert regions show lower hail frequencies, with the 23 stations in Alxa League averaging only 0.32 hail days annually. This pattern aligns with national studies identifying high-hail regions in high-altitude, complex terrain areas such as the Tibetan Plateau, Tianshan Mountains, Greater and Lesser Hinggan Mountains, Qilian Mountains, Helan Mountains, and Yinshan Mountains. The enhanced hail frequency in mountainous areas results from dramatic daytime heating over uneven terrain with heterogeneous vegetation, facilitating valley breeze circulation when cold air passes over mountains, thereby promoting convective development under sufficient moisture conditions.

2.2.1 Data Preprocessing

Before model training, positive and negative samples in the dataset were randomly shuffled and split into training and test sets at a 7:3 ratio. The training

set contained 1,568 samples (784 hail positives, 784 non-hail negatives), while the test set contained 672 samples (336 hail positives, 336 non-hail negatives). To prevent magnitude differences among predictors from affecting training outcomes, all feature variables were standardized prior to model training.

2.2.2 Predictor Selection and Importance Analysis

Based on the machine learning label dataset, we calculated various thermodynamic, dynamic, moisture conditions, and severe convective indices from ERA5 reanalysis data as potential predictors. After data screening and cleaning to remove samples with incomplete variables or outliers, 20 physical parameters were selected as candidate predictors . Using default parameters, the four machine learning algorithms were trained and evaluated on the test set . Random forest achieved the best performance, correctly identifying 313 of 336 hail cases with a hit rate of 93.11%, false alarm rate of 6.88%, miss rate of 7.82%, and threat score of 0.87. LightGBM and K-nearest neighbors also achieved hit rates above 88%, while decision tree ranked last but still attained an 88.34% hit rate and 0.78 threat score.

Decision tree, random forest, and LightGBM calculate information entropy to establish feature engineering, identifying predictors most closely related to forecast outcomes while outputting each predictor' s contribution to model performance. The top contributors across all models were total index and hail occurrence time, consistently ranking first and second. Other important factors included moisture conditions (column water vapor, column total water), dynamic factors (850–500 hPa wind speed), and instability parameters (850–500 hPa height, convective available potential energy). The physically meaningful predictors selected by machine learning align well with subjective forecasting experience, lending credibility to the hail identification models for operational application.

Based on contribution rankings, the most influential predictors—K index, total index, 850 hPa pseudo-equivalent potential temperature, 500 hPa pseudo-equivalent potential temperature, 850–500 hPa pseudo-equivalent potential temperature difference, 850–500 hPa dewpoint depression, 500 hPa temperature, convective available potential energy, column total water, and column water vapor—were selected as model input variables, along with spatiotemporal information. Testing confirmed that adding or removing any of these variables reduced model performance.

2.2.4 Model Hyperparameter Tuning

Hyperparameter optimization significantly impacts model training performance. Different hyperparameter combinations can substantially affect model behavior, making tuning essential for improving performance, preventing overfitting, and accelerating convergence. In K-nearest neighbors, the n_{neighbors} parameter controls the number of neighboring samples considered; too small

a value increases sensitivity to outliers and noise (risking overfitting), while too large a value over-smooths the classification (risking underfitting). In tree-based models (decision tree, random forest, LightGBM), n_{estimators} and max_{depth} are critical hyperparameters. n_{estimators} represents the number of base learners—typically, larger values improve performance but increase training time and overfitting risk. max_{depth} controls tree depth; greater depth increases model complexity and overfitting potential. min_{{{sample}}_{{leaves}}} specifies the minimum samples per leaf node, where smaller values increase model flexibility.

The hyperparameter search ranges are shown in . Using grid search, we tested various combinations to optimize each model. The final configurations were: random forest (n_{estimators}=161), LightGBM (n_{estimators}=152), K-nearest neighbors (n_{neighbors}=4), and decision tree (max_{depth}=8, min_{{{sample}}_{{leaves}}}=3). After tuning, all models showed improved performance on the test set , with threat scores increasing by 0.01-0.04. Decision tree showed the largest improvement, with hit rate increasing by 4.78% and threat score improving by 0.07. The model ranking remained unchanged: random forest, LightGBM, K-nearest neighbors, and decision tree, in descending order of comprehensive performance.

2.2.5 Model Application Verification

Using hail records from surface stations in Inner Mongolia during May–August 2023, we verified the optimized models' performance . Among 23 observed hail events, random forest correctly identified 19 (82.61% accuracy), LightGBM identified 17 (73.91%), K-nearest neighbors identified 15 (65.22%), and decision tree identified 17 (73.91%). The four missed cases (Tumote Right Banner, Horqin Left Middle Banner, etc.) involved isolated convective cells. For instance, the Tumote Right Banner event only reached 30 dBZ in radar reflectivity, indicating that the models have slightly weaker identification capability for weaker convective systems. Future work will incorporate radar parameters such as reflectivity factor, maximum echo top height, and vertically integrated liquid water to enhance short-term nowcasting performance.

4 Conclusions

Hail is a common severe convective weather phenomenon in Inner Mongolia, characterized by its sudden onset, strong locality, and destructive potential, posing serious threats to socioeconomic development and public safety. The region' s elongated geography, complex terrain, and varied climate conditions create distinct spatial patterns in hail distribution, with significantly higher frequencies in mountainous areas like the Yinshan and Greater Hinggan ranges compared to other regions. This aligns with Tang et al.' s national hail distribution study, which found that hail-prone areas generally extend along mountain systems, with central and eastern Inner Mongolia ranking among the nation' s

highest hail frequency zones.

Long-term trend analysis reveals a decreasing pattern in hail station-days in Inner Mongolia, consistent with trends in Xinjiang, Shaanxi, and most regions of China. This decline is closely related to global warming, rising average temperatures, and improved artificial hail suppression capabilities. Hail occurrence exhibits strong seasonality and a concentrated monthly distribution, with May-September representing the peak period—a conclusion consistent with studies on seasonal hail distribution in North China by Hu et al. and Wei et al. The underlying mechanism involves increasingly active circulation patterns and enhanced moisture transport from summer onward, combined with intensified solar radiation and surface heating that destabilize the boundary layer, favoring convective development.

However, accurate hail identification and forecasting remain among the most challenging tasks in weather prediction. While national-level severe convective weather subjective forecast products achieve threat scores of only 0.01–0.07 for 6–24h hail forecasts, this study demonstrates that machine learning algorithms can effectively identify hail events. All four algorithms achieved threat scores above 0.83 and hit rates exceeding 92% after hyperparameter tuning, with performance ranking from highest to lowest as: random forest, LightGBM, K-nearest neighbors, and decision tree.

Predictor importance analysis revealed that K index, total index, 850 hPa pseudo-equivalent potential temperature, 500 hPa pseudo-equivalent potential temperature, 850–500 hPa pseudo-equivalent potential temperature difference, 850–500 hPa dewpoint depression, 500 hPa temperature, convective available potential energy, column total water, and column water vapor are critical environmental diagnostic parameters for hail development. These physically meaningful variables align with subjective forecasting experience, lending credibility to the machine learning models for operational application.

Despite these advances, limitations remain. The current study relies primarily on manually observed hail cases from stations, which cannot capture hail events occurring beyond observation networks. Future work will incorporate additional observational data from satellites and radar to expand monitoring coverage and further optimize the sample dataset, thereby enhancing model forecasting performance.

References

[1] Yao Zhanyu, Tu Qi, An Lin, et al. Review of advances in hail formation process and hail suppression research[J]. Acta Meteorologica Sinica, 2022, 80(6): 835-863.

[2] Tao Tao, Zhang Lixin, Sang Jianren, et al. A case analysis of microphysical characteristics of atypical hail formation over Liupan Mountain, China[J]. Arid Land Geography, 2020, 43(2): 299-307.

[3] Wang Yun, Xie Xiangyang, Ma Yu, et al. Moving paths and nowcasting indicators of radar of hail cloud in northern Tianshan Mountains[J]. Arid Land Geography, 2017, 40(6): 1152-1164.

[4] Zhang Fanghua, Gao Hui. Temporal and spatial features of hail days in China[J]. Journal of Nanjing Institute of Meteorology, 2008, 31(5): 687-693.

[5] Tang Xingzhi, Huang Zhiyong, Zhang Rong, et al. Temporal and spatial distribution characteristics of hail disaster events in China from 2012 to 2020[J]. Torrential Rain and Disasters, 2023, 42(2): 223-231.

[6] Wei Yinghua, Hua Jiajia, Wang Ying, et al. Statistical characteristics and convection indicators of hailstorm over Tianjin in recent 11 years[J]. Meteorological Monthly, 2023, 49(2): 213-223.

[7] Zhong Min, Guo Yinglian, Chen Xuan, et al. Study on hail probability forecast method based on objective classification[J]. Plateau Meteorology, 2022, 41(4): 934-944.

[8] Liu Xinwei, Wenjia, Huang Wubin, et al. Study of the classified identification of the strong convective weathers based on the LightGBM algorithm[J]. Plateau Meteorology, 2021, 40(4): 909-918.

[9] Liu Xinwei, Jiang Yingsha, Huang Wubin, et al. Classified identification and nowcast of hail weather based on radar products and random forest algorithm[J]. Plateau Meteorology, 2021, 40(4): 898-908.

[10] Luo Xiping, Liao Bo, Zhang Xiaojuan, et al. Climatic characteristics of hail in Guizhou from 1961 to 2020[J]. Journal of Arid Meteorology, 2022, 40(6): 1024-1032.

[11] Tang Xingzhi, Huang Xingyou. Doppler radar identification parameters and their effect on early warning of hail clouds[J]. Torrential Rain and Disasters, 2009, 28(3): 261-265.

[12] Han Jingwei, Wang Haimei, Wu Lan, et al. The analysis and assessment on thunderstorm and hail disasters and the countermeasures in Inner Mongolia[J]. Journal of Arid Land Resources and Environment, 2009, 27(7): 31-38.

[13] Gu Ruiyuan, Sun Yonggang, Han Jingwei, et al. Weather forecast manual of Inner Mongolia[M]. Beijing: China Meteorological Press, 2012: 277-281.

[14] Li Wenjuan, Zhao Fang, Li Minjie, et al. Forecasting and classification of severe convective weather based on numerical forecast and random forest algorithm[J]. Meteorological Monthly, 2018, 44(12): 1555-1564.

[15] Zhou Kanghui. Convective weather forecasting with convolutional neural networks[D]. Beijing: University of Chinese Academy of Sciences, 2021.

[16] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[17] Zhu Sihua, Luo Ji, Qu Lianglu. The spatial temporal distribution and radar echo signatures of hail in Aksu, Xinjiang[J]. Desert and Oasis Meteorology, 2021, 15(2): 81-88.

[18] Hu Yaqiong, Bian Yuxuan, Huang Mengyu, et al. Characteristics of hailstone distribution based on disaster in Beijing from 1981 to 2017[J]. Journal of Meteorological Science, 2019, 30(6): 710-721.

[19] Tang Wenyuan, Zhou Qingliang, Liu Xinhua, et al. Analysis on verification of national severe convective weather categorical forecasts[J]. Meteorological Monthly, 2017, 43(1): 67-76.

[20] Friedman J H. Greedy function approximation: A gradient boosting machine[J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[21] Zhang Yong, Liu Hui, Zheng Yingfei, et al. Effect validation and analysis of classified products outputted by artificial intelligent nowcasting model[J]. Desert and Oasis Meteorology, 2023, 17(1): 115-121.

[22] Liu Ruiliang, Jia Keli, Li Xiaoyu, et al. Inversion of soil salt content by combining optical and microwave remote sensing in cultivated land[J]. Arid Land Geography, 2024, 47(3): 433-444.

Note: Figure translations are in progress. See original paper for figures.