Доклади на Българската академия на науките
Comptes rendus de l'Académie bulgare des Sciences

Tome 79, No 11, 2026

GEOPHYSICS

Meteorology

DATA-DRIVEN THUNDERSTORM DETECTION FOR SOFIA AIRPORT USING RANDOM FOREST MODEL

Denislav Bonchev[1]✉, Nikolay Penov[2], Guergana Guerova[1]

Received on January 10, 2026
Presented by K. Stoykova, Corresponding Member of BAS, on October 27, 2026

Abstract

Hazardous phenomena such as strong winds, hail, heavy rainfall, and tornadoes are often associated with severe thunderstorms. Although relatively rare, these events account for a substantial portion of the damage caused by meteorological hazards. Numerical weather prediction models frequently miss those phenomena due to their small spatial scale and the complex interplay of factors governing their development. Consequently, data-driven machine-learning approaches for thunderstorm nowcasting have a potential to improve their forecasting. In this study, a Random Forest (RF) classifier is applied to surface METAR observations and convective indices derived from the ERA5 reanalysis. Seven convective indices were calculated for Sofia Airport during the warm seasons (April–September) from 2014 to 2024. Among the examined indices, the K Index most effectively discriminates between thunderstorm and non-thunderstorm cases, outperforming indices such as CAPE-SHEAR and MUCAPE. RF performance is evaluated using three input datasets: (1) surface-based observations only, (2) convective indices only, and (3) a combination of surface observations and convective indices. The combined dataset yields the best performance for thunderstorm diagnosis. However, the RF classifier has a tendency to overestimate storm duration, reflecting limitations of the selected environmental predictors.

Key words: thunderstorms, instability indices, METAR surface observation, Random Forest

This study is financed by the European Union – NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No BG-RRP-2.004-0008-C01.

https://doi.org/10.7546/CRABS.2026.11.09

Introduction. Thunderstorms and the hazardous phenomena associated with them – such as hail, strong wind gusts, and heavy precipitation – remain among the most challenging weather events to diagnose and forecast. Their localized spatial extent, short life cycle, and strong dependence on both thermodynamic instability and atmospheric dynamics limit the predictive skill of traditional Numerical Weather Prediction (NWP) models. As a result, there is a growing interest in complementary approaches that can improve the diagnosis and nowcasting of convective storms using data-driven methods [1]. In recent years, Machine Learning (ML) has become an increasingly important tool in meteorology, particularly for nowcasting of hazardous weather. Ensemble-based algorithms, such as Random Forest, have demonstrated strong performance in diagnosing convective hazards by learning nonlinear relationships between atmospheric predictors and observed events [2, 3]. Random Forest models have been successfully applied to the probabilistic prediction of hail, severe wind gusts, and tornadoes, often outperforming traditional operational guidance and exhibiting improved reliability, especially in environments characterized by strong instability and weak wind shear [3, 4]. These advances highlight the ability of deep learning methods to extract meaningful spatial patterns from large and complex meteorological datasets. Despite these promising results, machine learning approaches are not a substitute for physically based NWP models, as their performance depends strongly on the quality and availability of the training data. Rather, ML methods should be viewed as complementary tools that can support the diagnosis and nowcasting of convective hazards by identifying complex relationships within large meteorological datasets. The aim of this study is to diagnose the occurrence of thunderstorms at Sofia Airport with an RF machine learning model using convective instability indices derived from atmospheric reanalysis (ERA5), and surface observations from METAR reports as model features.

Methodology. ***Surface observations and ERA5 data.*** Surface observations for this study were obtained from the METAR reports of Sofia Airport. These reports provide high-frequency measurements of near-surface meteorological variables, including 2 meter air temperature (T 2m), 2 meter relative humidity (RH 2m), pressure reduced to sea level, and 10-m wind speed and direction. Thunderstorm occurrence was extracted from the present-weather codes following standard WMO conventions. The METAR dataset offers a consistent, quality-controlled observational basis for diagnosing near-surface atmospheric conditions relevant to convective environments. All METAR reports were decoded using the Python library MetPy [5]. Upper-air thermodynamic and kinematic information was derived from the ERA5 global reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) [6]. ERA5 provides hourly profiles of temperature, humidity, and wind on a fixed vertical grid, enabling the calculation of convective instability indices not available from surface observations alone. These data were extracted for the grid point closest to Sofia Airport and inter-

polated vertically when required. The high temporal resolution of ERA5 makes it suitable for investigating the evolution of convective environments on sub-daily timescales.

Instability indices. All instability indices were computed from the ERA5 vertical profiles using the MetPy library for atmospheric thermodynamic calculations. To maintain reproducibility and consistency, MetPy standard algorithms and parcel methods were used for all indices. Convective Available Potential Energy (CAPE) and Convective Inhibition (CIN) were calculated using both surface and most-unstable parcel definitions, but the analysis employs MUCAPE and MUCIN, which are widely regarded as more representative of elevated instability layers. All the temperatures used are virtual temperatures, calculated using Wallace and Hobbs [7]. The Lifted Index (LI) [8] is defined as the temperature difference between the environment and a lifted parcel at 500 hPa; negative values indicate strong instability favourable for deep convection. The K Index [9] quantifies thunderstorm potential by combining lapse rates and mid-level moisture. Total Totals (TT) [10] evaluates the combined effect of thermal lapse rates and low-level moisture. The Severe Weather Threat (SWEAT) index [10] combines instability, moisture, and wind-shear terms to estimate the likelihood of severe thunderstorms. The bulk wind shear was computed as the vector wind difference between 750 m and 6 km. The CAPE–shear index [11] is intended for diagnosing severe convective environments.

Data preparation and pre-processing. The dataset used in this study consists of METAR observations and ERA5 vertical profiles covering the 11-year period 2014–2024. Only data from April to September were included to maintain a more balanced and seasonally relevant sample. To capture the temporal evolution of the atmospheric environment preceding a convective event, lagged variables were introduced into the feature set. Specifically, for each predictor, values from preceding timesteps (t-1, t-2, t-3, t-4) were appended as additional, distinct features for the current observation at time t with timestep of 30 min. The selection of 4 lagged timesteps was determined by evaluating model performance across various maximum lag lengths tested from t-1 up to t-6. Validation metrics – specifically Probability of Detection (POD) and Balanced Accuracy (BA) – showed that performance improved up to 4 lags and plateaued thereafter, capturing the short-term temporal evolution of atmospheric predictors. As the Random Forest classifier requires numerical inputs, all predictor variables were scaled to the interval [0,1] using the MinMaxScaler [12]. For the dependent variable, the present-weather information in the METAR reports was simplified to a binary classification – namely observations containing thunderstorm codes were assigned a value of 1, while all remaining observations were assigned 0. The resulting dataset is strongly imbalanced: out of 96 609 observations for the period, only 1.46% contain thunderstorms.

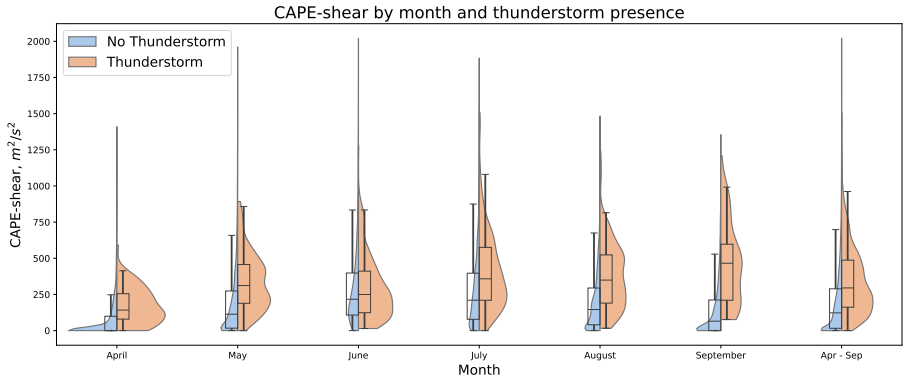
Random Forest classification model and hyperparameter optimization. At each tree, a random subset of the predictors and training samples is chosen with the bootstrapping method, and splits are selected to minimize Gini impurity. The final classification is obtained through majority voting across all trees in the ensemble. Standard Random Forest implementation utilises independent and identically distributed observation-level bootstrapping with replacement. While meteorological time series exhibit temporal autocorrelation, ordinary bootstrapping is employed here because preceding environmental evolution is explicitly embedded directly into each sample's feature vector via lagged variables (t-1 to t-4). Thus, each feature row encapsulates its own local temporal context independently of adjacent rows. To test the sensitivity of our model to temporal dependencies, auxiliary evaluations were performed using time-series resampling alternatives – specifically Moving Block Bootstrap and Circular Block Bootstrap with a 4-hour block size (8 timesteps). Although performance metrics and top feature rankings remained largely invariant to the resampling technique, the findings in this study should be interpreted with caution, as observation-level bootstrapping inevitably breaks the chronological sequence of the training data Hyperparameters of the Random Forest model – including the number of trees, maximum depth, minimum samples required for splitting a node, minimum samples per leaf, and the maximum number of features considered at each split – were optimized through systematic testing. To address the strongly imbalanced sample, class weights were applied as:

$$w_i = \frac{N}{K \times n_i},$$

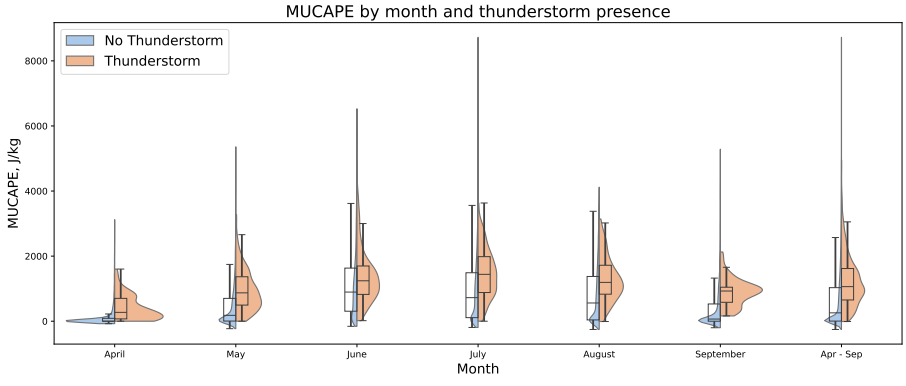
where w_i is the weight for class i, n_i is the number of observations in that class, N is the total sample size, and K is the number of classes. All model training and evaluation were performed using the scikit-learn implementation of Pedregosa et al. [12].

Results. ***Convective indices for Sofia Airport April – September 2014–2024.*** For the period April–September 2014–2024, seven convective indices were calculated for Sofia Airport. All the calculated indices show peak (minimum for LI) values in the months June–July. SWEAT and TT show smaller seasonal variation, with mean values closer to the overall seasonal peak in June–July compared to the other instability indices. Bulk shear 0–6 km shows minimal variation throughout the period, indicating its weak seasonality. Thus SWEAT, TT and Bulk shear 0–6 km are not shown in Fig. 1 with violin plots with overlaid quartile diagrams of three of the calculated instability indices are presented.

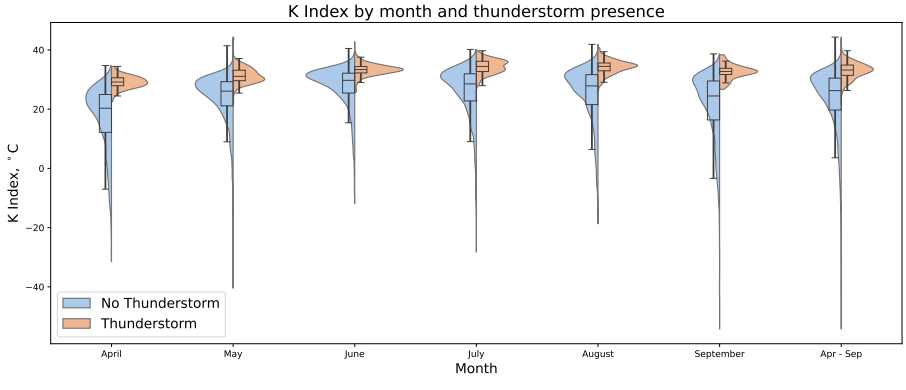
Figure 1 shows that CAPE-SHEAR does not distinguish well the cases with thunderstorm (TH) and without thunderstorm (NTH). The same is true for the other instability indices. Only during the transitional months April and September there is almost no overlap of P75 and P25 of the TH and NTH samples. As with the other indices, this is likely due to the more pronounced dynamic nature of



(a)



(b)



(c)

Fig. 1. Violin plot with overlaid quartile diagrams of three of the calculated instability indices: CAPE-shear (a), MUCAPE (b), and K Index (c) by month and for the entire study period. Blue represents cases without thunderstorm, while orange represents cases with a thunderstorm

thunderstorms in the transitional months. As can be seen in Fig. 1 there is an overlap between the 75th and 25th percentiles (P75 and P25) of the two samples for the months of May, June, July and August. This is due to the combination of 0–6 km wind shear and CAPE in the index. The index is used for forecasting extreme events and, similar to the 0–6 km wind shear, the index does not forecast well, only the presence of thunderstorms.

Figure 1b shows that MUCAPE does not clearly distinguish between TH and NTH cases, in contrast to the other convective indices. In most months with the exception of September, there is an overlap of the P75 and P25 percentiles of the two samples – respectively the TH (orange) and NTH (blue). The largest overlap between P75 and P25 between the NTH and TH samples is during the months with the highest atmospheric instability, namely June and July. During the transition months April, May, and September smallest overlap is seen in Fig. 1b. This is due to the weaker instability of the atmosphere during these months and the more pronounced dynamic character of convection in the transition months. For all months, the peaks of the MUCAPE distributions in the TH and NTH samples differ distinctly. The peak at low MUCAPE values in the NTH sample is due to prolonged periods with stable atmospheric stratification. However, because of the extreme values, even in the absence of thunderstorms, the median and P75 are high. All of this makes MUCAPE an excellent indicator of atmospheric instability, but not such a good indicator of instability leading to thunderstorms.

Figure 1c shows that the K Index distinguishes well the TH and NTH samples. For all months there is no overlap between P75 and P25 of the two samples. It can also be seen that during the transition months April, May, and September the peaks of the TH and NTH distributions and the medians have a greater difference compared to the other months. This is due to the weaker atmospheric instability during these months and the more pronounced dynamic character of convection during the transition months. For these months, the maxima of the distributions and the median of the K Index for the TH cases do not differ substantially from the means for the period April–September. In comparison, the corresponding values for the NTH cases are lower than the means for the April–September period. Nevertheless, the K Index distinguishes better the TH and NTH samples during the summer months than the other calculated indices.

Machine learning models for convective storms at Sofia Airport. Three different tests have been conducted in which three different ML models using the RF method have been developed. In Test 1 the model was trained solely on METAR surface observations. In Test 2 the RF model is trained exclusively on all convective indices namely K Index, Total Totals, SWEAT, Lifted Index, MUCAPE, MUCIN, CAPE-shear, Bulk Shear 0–6 km. In Test 3 we use both METAR data and convective indices. The performance of the models in the different tests is compared in Table 1. It can be seen that in Test 2 and Test 3, the model diagnoses thunderstorms well, with a POD above 0.8 in both cases.

Table 1

Model performance in the different tests

Metric	Test 1	Test 2	Test 3
POD	0.777	0.826	0.856
FAR	0.969	0.957	0.952
CSI	0.031	0.043	0.048
TSS	0.519	0.631	0.675
BA	0.760	0.815	0.838
HSS	0.040	0.063	0.072

Test 1, which is trained only on ground observations, has a POD of 0.777, and consequently, its performance is worse than that of Test 2 and Test 3. In all tests, the model has a high FAR, mostly due to the many cases where the model diagnoses thunderstorms but no thunderstorms occurred in unstable atmospheric conditions (see Fig. 2). Tests 2 and 3 have slightly lower False Alarm Ratio (FAR) values compared to Test 1. In Test 2, the values of the Critical Success Index (CSI) and the True Skill Statistic (TSS) are significantly higher compared to Test 1, which means that the model performs significantly better when trained only with the convective indices compared to being trained only with METAR data. In Test 3, the CSI and TSS values are higher than in Test 2, although the difference is small. The calculated BA and Heidke Skill Score (HSS) [13] also further imply that the model performs better if it is trained on convective indices or a combination of ground observations or convective indices. Also both BA and HSS show that in all tests the model performs better than random. This suggests that combining METAR observations with convective indices improves the performance of the model.

Table 2 summarises the importance of each feature for the model. The most preferred features are those that account for atmospheric instability as well as atmospheric dynamics. Among the surface-level features Pressure and T 2m are ranked the highest. This suggests the model successfully leverages atmospheric dynamics – such as the passage of weather fronts – to identify convective initiation. This conclusion is reinforced by the model ranking pressure at a two-hour lag (t-4) as its most valuable surface predictor, which effectively captures the characteristic pressure falls that typically precede frontal boundaries or the drop in pressure during the destabilisation of the atmosphere due to heating preceding intra-air-mass convection initiation. It is also evident that convective indices reflecting the vertical profile of the atmosphere are always more important for the model than the others. As for the indices, as noted in the previous section, the K Index discriminates well between TH and NTH samples in all months. This is why it holds the highest importance for the model. The second place is for TT, which

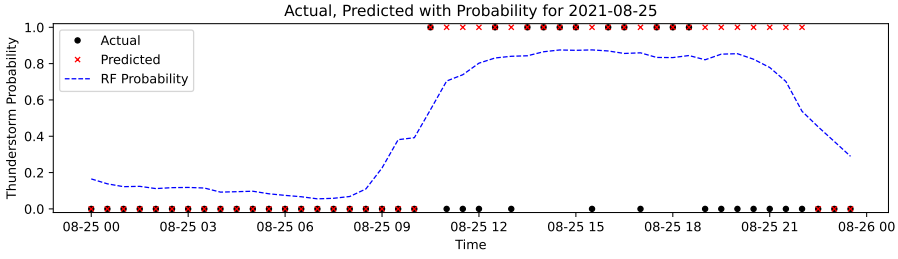
Table 2

Feature importance for the RF model in the different tests. Only the 20 most important features are displayed for Test 2 and Test 3

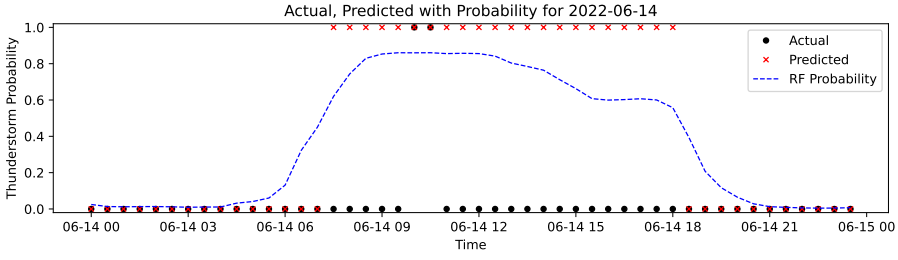
Test 1		Test 2		Test 3	
Feature	Importance	Feature	Importance	Feature	Importance
Pressure(t-4)	0.161	KI(t0)	0.145	KI(t0)	0.099
Pressure(t-3)	0.133	KI(t-1)	0.143	KI(t-2)	0.097
Pressure(t-2)	0.127	KI(t-2)	0.111	KI(t-3)	0.091
T 2m(t-4)	0.102	KI(t-3)	0.074	KI(t-1)	0.087
T 2m(t-3)	0.077	KI(t-4)	0.071	KI(t-4)	0.070
Pressure(t-1)	0.066	TT(t-1)	0.058	TT(t-1)	0.067
RH 2m(t0)	0.047	TT(t0)	0.049	TT(t-3)	0.050
T 2m(t-2)	0.044	TT(t-2)	0.049	TT(t0)	0.040
Wind Spd(t0)	0.043	TT(t-3)	0.044	TT(t-2)	0.039
Pressure(t0)	0.031	TT(t-4)	0.034	TT(t-4)	0.030
RH 2m(t-1)	0.029	SWEAT(t-1)	0.025	SWEAT(t-1)	0.027
T 2m(t-1)	0.024	SWEAT(t-2)	0.019	SWEAT(t-3)	0.026
T 2m(t0)	0.023	SWEAT(t0)	0.017	Pressure(t-4)	0.018
RH 2m(t-3)	0.022	MU CIN(t-4)	0.017	Pressure(t-3)	0.016
Wind Spd(t-1)	0.018	LI(t-1)	0.014	T 2m(t-4)	0.016
RH 2m(t-4)	0.017	MU CIN(t-3)	0.013	LI(t0)	0.015
RH 2m(t-2)	0.015	LI(t0)	0.012	SWEAT(t-2)	0.015
Wind Spd(t-2)	0.009	SWEAT(t-3)	0.011	SWEAT(t-4)	0.015
Wind Spd(t-3)	0.007	SWEAT(t-4)	0.011	LI(t-2)	0.014
Wind Spd(t-4)	0.005	MU CIN(t-2)	0.010	T 2m(t-3)	0.012

is most likely because there is no well-defined monthly trend during the selected study period from April to September. Finally, the model demonstrates a clear preference for temporal proximity regarding instability metrics: recent indices carry more importance for the model than those from older timesteps. This aligns perfectly with the physical understanding of atmospheric instability, which can evolve rapidly; consequently, the most current observation serves as the most robust predictor.

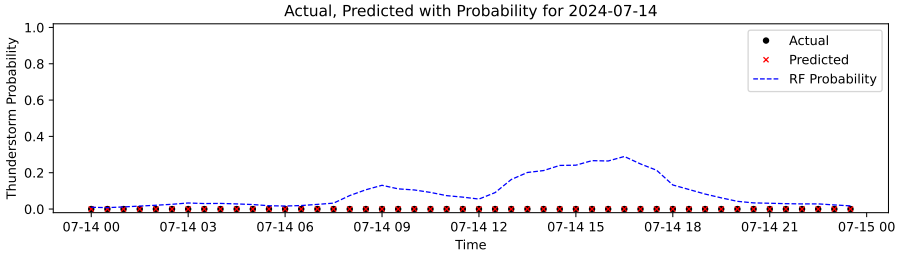
Figure 2 presents results from the Random Forest model Test 3 for three selected case studies. For 25 August 2021, the model accurately captures the onset of convection but fails to detect the cessation of storm activity, producing prolonged false alarms (Fig. 2a). This behaviour reflects a known limitation of reanalysis data: convection itself is not explicitly resolved, and model predictors describe the broader unstable environment rather than storm triggers. As a result, the model maintains high probabilities of thunderstorms despite the absence of observed events. A similar outcome is noted for 14 June 2022 (Fig. 2b), where



(a)



(b)



(c)

Fig. 2. Results from the Random Forest model (model prediction with black dot and actual observation with red cross) from Test 3 with higher temporal resolution of 30 min. The figure also includes the probability of thunderstorms (dashed line), estimated by the model. (a) Results from August 25, 2021, (b) results from June 14, 2022, (c) results from July 14, 2024

the model predicts an unrealistically early onset and prolonged duration of thunderstorms. In contrast, during non-convective conditions on 14 July 2024 (Fig. 2c), forecasted probabilities remain low throughout the day, demonstrating reliable performance under stable atmospheric regimes. In conclusion, these results show that convective indices computed from reanalysis data provide valuable skill in identifying favourable thunderstorm environments, particularly when combined with surface observations. However, both instability-based and dynamical indices do not accurately predict the timing of initial development and duration of thunderstorm activity, resulting in frequent false alarms.

Discussion and conclusions. This study demonstrates the applicability of data-driven machine learning methods for diagnosing thunderstorm occurrence at Sofia Airport using physically meaningful atmospheric predictors. A Random Forest classifier was trained on surface METAR observations and convective indices derived from ERA5 reanalysis for the period April–September 2014–2024. As a first step an analysis of the computed convective indices has been performed. It reveals a pronounced seasonal signal, with maximum instability occurring during the summer months, consistent with the regional convective climatology. Thermodynamic instability indices exhibit clear differences between thunderstorm and non-thunderstorm cases, whereas kinematic parameters such as 0–6 km bulk wind shear show weaker seasonal variability. This suggests that, in the Sofia region, thermodynamic conditions play a dominant role in thunderstorm occurrence, while dynamical forcing provides a secondary contribution.

Then three different RF models have been developed – the first using only ground-based observations as features, the second using only the computed convective indices as features, and the third using a combination of ground-based observations and convective indices. The most favoured RF features are those related to atmospheric instability, humidity, and dynamics. Convective instability indices reflecting the vertical profile are consistently more important. The K Index effectively distinguishes between TH and NTH samples for all months, making it the most important. The index with second place with respect to importance is TT, likely due to the lack of a clear monthly trend during the study period from April to September. Crucially, the integration of lagged variables allowed the model to capture the temporal evolution of the atmosphere, with recent instability metrics and lagged surface pressure proving highly effective at identifying dynamic triggers like frontal passages or the destabilisation of the atmosphere preceding intra-air-mass convection initiation.

The results confirm that ensemble-based machine learning approaches are well suited for convective-weather classification problems. The Random Forest model demonstrates robust performance in diagnosing non-convective and stable atmospheric conditions, maintaining low false-alarm rates in such environments. However, in some convective cases the model tends to diagnose earlier storm initiation and delayed dissipation compared to observations, resulting in an overestimation of thunderstorm duration. This highlights a limitation of machine learning models that rely solely on environmental predictors, without explicit information on storm-scale processes. Improving short-term thunderstorm prediction therefore likely requires additional predictors that capture triggering mechanisms and storm-scale processes unresolved in reanalysis products. Future improvements may be achieved through the inclusion of additional predictors, such as radar or satellite-derived observations, and by exploring deep learning approaches better suited for spatiotemporal data.

REFERENCES

[1] Haupt S. E., D. J. Gagne, W. W. Hsieh et al. (2022) The history and practice of AI in the environmental sciences, Bull. Am. Meteorol. Soc., **103**(5), E1351–E1370.

[2] Breiman L. (2001) Random forests, Mach. Learn., **45**(1), 5–32.

[3] Hill A. J., G. R. Herman, R. S. Schumacher (2020) Forecasting severe weather with random forests, Mon. Weather Rev., **148**(6), 2135–2161.

[4] McGovern A., K. L. Elmore, D. J. Gagne et al. (2017) Using artificial intelligence to improve real-time decision-making for high-impact weather, Bull. Am. Meteorol. Soc., **98**(10), 2073–2090.

[5] May R. M., K. H. Goebbert, J. E. Thielen et al. (2022) MetPy: A meteorological Python library for data analysis and visualization, Bull. Am. Meteorol. Soc., **103**(10), E2273–E2284.

[6] Hersbach H., B. Bell, P. Berrisford et al. (2020) The ERA5 global reanalysis, Q. J. R. Meteorol. Soc., **146**(730), 1999–2049.

[7] Wallace J. M., P. V. Hobbs (1977) Atmospheric Science: An Introductory Survey, New York, Academic Press, 397 pp.

[8] Galway J. G. (1956) The Lifted Index as a predictor of latent instability, Bull. Am. Meteorol. Soc., **37**(10), 528–529.

[9] George J. J. (1960) Weather Forecasting for Aeronautics, New York, U.S. Department of Commerce, National Weather Service, 162–163.

[10] Miller R. C. (1975) Notes on analysis and severe-storm forecasting procedures of the Air Force Global Weather Central (Tech. Rep. No. 200), AWS, 190 pp.

[11] Tsonevski I. (2025) Influence of environmental conditions on the development of thunderstorm clouds over Eastern Bulgaria (PhD dissertation), National Institute of Meteorology and Hydrology, Bulgaria.

[12] Pedregosa F., G. Varoquaux, A. Gramfort et al. (2011) Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., **12**, 2825–2830.

[13] Heidke P. (1926) Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst, Geogr. Ann., **8**(4), 301–349.

[1]*Department Meteorology and Geophysics, Faculty of Physics,*
Sofia University "St. Kliment Ohridski", 5 James Bourchier Blvd, 1164 Sofia, Bulgaria
e-mails: dibonchev@uni-sofia.bg
guerova@phys.uni-sofia.bg, https://orcid.org/0000-0001-5638-3387,
Scopus ID: 8558549200

[2]*Bulgarian Air Traffic Authority, 1 Brussels Blvd, 1540 Sofia, Bulgaria*
e-mail: nikolay.penov@bulatsa.com, Scopus ID: 57211242596