26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

A Random Forest analysis of variables influencing pyrocumulonimbus (pyroCb) occurrence over temperate southeast Australia

Wenyuan Ma [a,b], **Jason John Sharples** [a,b] **Zlatko Jovanoski** [a]

[a] School of Science, University of New South Wales, Canberra, ACT, Australia.
[b]NSW Bushfire and Natural Hazards Research Centre, NSW, Australia.
Email: wenyuan.ma@unsw.edu.au

Abstract: Pyrocumulonimbus clouds (pyroCbs) are extreme cloud formations that develop when intense wildfire heat and moisture generate deep convective columns that penetrate the upper troposphere and, in some cases, reach the lower stratosphere, thereby posing significant risks to the climate system, ecosystems, and fire behaviour. Climate projections indicate that the risk of extreme fire events will increase in the future, highlighting the importance of understanding potential occurrence patterns of pyroCbs and developing effective risk mitigation and management strategies. The temperate southeast Australia is a pyroCb prone region, with almost all recorded pyroCb events in southeast Australia occurred there (Figure 1). This study explores the potential of a machine learning technique—Random Forest—in assessing pyroCb risk across temperate southeastern Australia. This study compiled and integrated multi-source environmental data related to pyroCb events and standard wildfires (i.e., those that did not trigger pyroCb development) that occurred in temperate southeastern Australia between 1980 and 2020. Ten potential predictor variables were considered, covering atmospheric, surface fire weather, topographic, and vegetation conditions. The Random Forest algorithm was used to evaluate the relative importance of each variable. Results show that atmospheric variables, particularly the Continuous Haines (C-Haines) index, consistently emerged as the most important predictors, followed by surface fire weather and topographic variables, while vegetation variable was identified as the least important. The Random Forest model demonstrated strengths in capturing non-linear and threshold-based relationships; however, the rarity of pyroCb events and limited data availability presented challenges for predictive accuracy. This study highlights both the potential and limitations of Random Forest for pyroCb risk assessment and underscores the importance of selecting appropriate modelling approaches for effective risk evaluation of small-sample natural hazard.

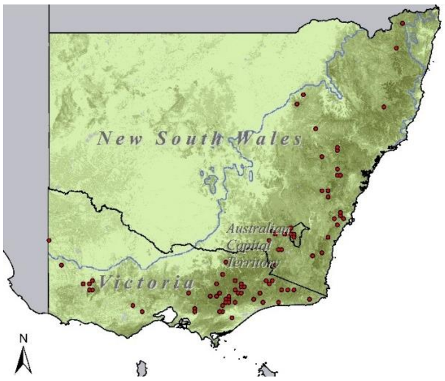

Figure 1. PyroCb distribution (red circles) in southeast Australia. The blue line represents the boundary of the temperate zone over southeast Australia. The background represents terrain ruggedness, with darker colours indicating rougher terrain.

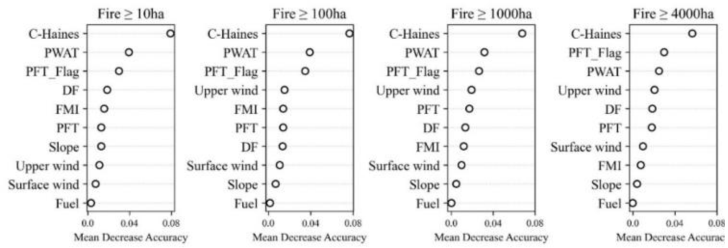

Figure 2 The importance ranking of explanatory variables in the whole study area. The subtitles indicate the wildfire size. The vertical axis indicates the variable, and the horizontal axis indicates the mean decrease accuracy when the given variable is removed.

Keywords: pyrocumulonimbus clouds (pyroCbs), Random Forest, extreme fire.

1. INTRODUCTION

Pyrocumulonimbus clouds (pyroCb) are extreme cloud formations that develop when intense wildfire heat and moisture generate deep convective columns, with potentially profound adverse impacts on the climate system, ecosystems, and wildfire behaviour itself. Climate projections indicate an increase in the risk of extreme fire events both in Australia (Di Virgilio *et al.* 2019; Dowdy *et al.* 2019; Ndalila *et al.* 2020) and globally (Abatzoglou *et al.* 2019). This underscores the importance of improving our understanding of the likely occurrence patterns of pyroCbs, including their spatial distribution, temporal variability, and associated environmental conditions, and informing effective risk mitigation and management strategies.

Previous studies have employed numerical simulation (Badlan *et al.* 2021a, 2021b) and statistical models (Ma *et al.* 2023; Ma *et al.* 2025) to investigate the key factors influencing pyroCb development. These efforts have contributed to the identification of critical factors, such as dynamic fire behaviour, atmospheric conditions, and surface fire weather, that influence pyroCb development. However, these methods face limitations in capturing complex nonlinear relationships and interactions among variables. Against this backdrop, machine learning, an approach that leverages algorithms to learn patterns from data and make predictions, has been increasingly applied in climate and environmental research. It also offers new tools for understanding the complex fire–atmosphere coupling processes involved in pyroCb formation. Tazi et al. (2022) developed a machine learning framework using Random Forests, Convolutional Neural Networks (CNNs), and CNNs pretrained with Auto-Encoders to predict the occurrence of pyroCbs from a given fire six hours in advance. Their findings showed that the Random Forest model exhibited the highest skill in forecasting pyroCb development. Overall, the application of machine learning in pyroCb research has demonstrated considerable potential, particularly the Random Forest algorithm offers distinct benefits in assessing variable importance, which is especially valuable for pyroCb risk assessment. However, such approaches have not been applied to pyroCb research in southeastern Australia.

To explore the potential of machine learning for pyroCb risk assessment, this study compiled and integrated multi-source environmental data related to pyroCb events and standard wildfires (i.e., those that did not trigger pyroCb development) that occurred in temperate southeastern Australia between 1980 and 2020. We employed the Random Forest algorithm to model the occurrence of pyroCb events, evaluating the relative importance of each variable and the model's predictive performance. Through this analytical framework, the study aims to achieve two main objectives: (1) to assess the applicability of Random Forest as a tool for pyroCb risk prediction, and (2) to enhance our understanding of the spatial patterns of pyroCb occurrence in Australia. The findings offer valuable insights for future extreme fire risk forecasting and provide a scientific basis for disaster early warning and management strategies.

2. MATERIALS AND METHODS

The temperate region of southeastern Australia is a pyroCb-prone area—almost all documented pyroCb events in southeastern Australia have occurred within this region (Figure 1). Accordingly, temperate southeastern Australia, including Australian Capital Territory (ACT) and parts of New South Wales (NSW) and Victoria (VIC), was selected as the study area.

Table 1. Description of explanatory variables

Types	Variables	Details	Data source
Atmosphere	C-Haines	Continuous Haines Index (Mills *et al.* 2010)	
	PWAT	Precipitable water	
	PFT	Pyrocumulonimbus Firepower Threshold (Tory and Kepert, 2021)	
	PFT_Flag	Pyrocumulonimbus Firepower Threshold Flag (Tory and Kepert, 2019)	ERA5 hourly reanalysis (Hersbach et al., 2018)
	Upper wind	Wind speed (km/h) at 500hPa	
Surface fire weather	DF	Drought factor	
	FMI	Fuel Moisture Index (Sharples *et al.* 2009)	
Topography	Slope	Slope Relief Classification ("0": smooth surface; "1": rugged surface)	Slope Ruggedness Classification 1" SRTM DEM-S (Gallant and Austin 2012), reclassify into binary (0/1) variable
Vegetation	Fuel	Vegetation type ("0": non-forest; "1": forest)	National Vegetation Information System (NVIS, Australian Government Department of Agriculture, Water and the Environment, 2020), reclassify into binary (0/1) variable

The dependent variable in the model is binary (pyroCb/non-pyroCb), indicating whether a pyroCb event occurred. Records of pyroCb occurrences were obtained from the Australian PyroCb Register (McRae, 2015). Non-pyroCb events were defined as standard wildfires that did not trigger pyroCb and were extracted from government wildfire history databases in New South Wales (NSW Rural Fire Service, 2021) and Victoria (Department of Environment, 2021). The preprocessing and filtering of standard wildfires are detailed by Ma et al. (2023). The explanatory variables were grouped into four categories: atmospheric, surface fire weather, topographic, and vegetation variables, comprising a total of ten variables. A detailed list of variables is provided in Table 1. Because standard wildfires and some pyroCbs only have date information without precise time of occurrence, this study focuses on mid-afternoon conditions (06:00 UTC, corresponding to 16:00 AEST), which typically coincides with the most hazardous fire weather conditions in Australia (Di Virgilio *et al.*, 2019). The multicollinearity diagnostics were conducted using the Variance Inflation Factor (VIF), and all variables had VIF values below 5, indicating no serious multicollinearity.

To assess whether standard wildfire size influences the conditions associated with pyroCb occurrence, and whether the importance of variables varies across regions, we divided the dataset into several subsets based on fire size and geographic location. Individual models were developed for each subset to compare the relative importance of variables across different fire sizes and regions. Table 2 shows the number of pyroCb events and standard wildfires included in each subset. Given the substantial imbalance between the number of standard wildfires and pyroCb events, a down-sampling strategy was implemented to ensure class balance in the dataset and to avoid potential biases in model training, following the approach adopted in previous studies (Ma et al., 2023; Ma et al., 2025). Specifically, for each subset in New South Wales and Victoria, 50 standard wildfires were randomly selected, while 100 were selected from the whole study area. These randomly sampled standard wildfires were then combined with all pyroCb events in the corresponding region to form a random modelling sample, ensuring an approximate 1:1 ratio between pyroCb and standard wildfires in each model. This down-sampling process was repeated 100 times per subset, and models were trained on each of these random samples. Summary statistics of model outputs across the 100 runs were then calculated for final analysis.

Table 2. Number of PyroCb events for NSW and VIC and the number of standard fires characterized according to size.

Class	Fire size	Number		
		Total	NSW (including ACT)	VIC
PyroCb		93	40	53
Standard wildfires	⩾10ha	7876	6652	1224
	⩾100ha	3908	3368	540
	⩾1000ha	1077	901	176
	⩾4000ha	370	292	78

Random Forest has been applied in studies of wildfire drivers and often demonstrates superior predictive performance compared to regression (Oliveira *et al.* 2012; Guo *et al.* 2016; Ma *et al.* 2020). Random Forest evaluates variable importance primarily through two methods: node split contributions and permutation importance (Breiman 1984). Node split contributions measure importance by calculating the cumulative reduction in Gini impurity. While permutation importance assesses it by randomly shuffling given variable's values and observing the decrease in model accuracy. The greater the drop in accuracy, the more important the variable is. This study employed permutation importance to quantify variable importance, using the Mean Decrease Accuracy as the evaluation metric. The final importance score for each variable was calculated as the average value of Mean Decrease Accuracy over 100 independently sampled model runs.

Partial dependence plots are a visualization tool for interpreting machine learning models, helping us intuitively understand the relationship between variables and the target outcome. For Random Forest, partial dependence plots show how the model's predictions change as the value of a specific variable varies. In this study, partial dependence plots are used to illustrate how the probability of pyroCb occurrence changes as the values of each explanatory variable. The modelling was performed with R 4.0.0 (R Core Team, 2020).

3. RESULTS

3.1. The importance and impact of variables

There are slight variations in variable importance across different fire size categories (Figure 2). Atmospheric variables demonstrate remarkable importance, with C-Haines consistently identified as the most influential variable across all samples, exhibiting an importance index (mean decrease accuracy) significantly higher than that of other variables. PWAT is ranked as the second most important variable, except in the subset of larger fires (Fire≥4000ha). The importance of DF, FMI, and PFT fluctuates across different fire size models. The importance of wind increases with fire size, with Upper wind having a higher importance than Surface wind. The categorical variables Slope and Fuel are ranked lowest. Overall, atmospheric variables outperform surface fire weather variables, followed by topography (slope), while vegetation variables (fuel) are consistently ranked as the least important.

Atmospheric and surface fire weather variables encompass several distinct variables. To further examine their relative importance, the models containing only atmospheric or surface fire weather variables were also created (Figure 3). In the model including only atmospheric variables, the ranking of variable importance remained largely consistent with that observed in the all-variable model. C-Haines continued to exhibit the highest importance, while upper-level wind and PFT were of lowest importance. However, in the model including only surface fire weather variables, the importance rankings differed from those in the full-variable model, particularly for small to medium fires (fire ≥10 ha, ≥100 ha, ≥1000 ha). In these subsets, FMI emerged as the most important surface variable, surpassing DF. In contrast, for large fires (fire ≥4000 ha), FMI was ranked the least important, with DF the most important surface fire weather variable, consistent with the full-variable model results.

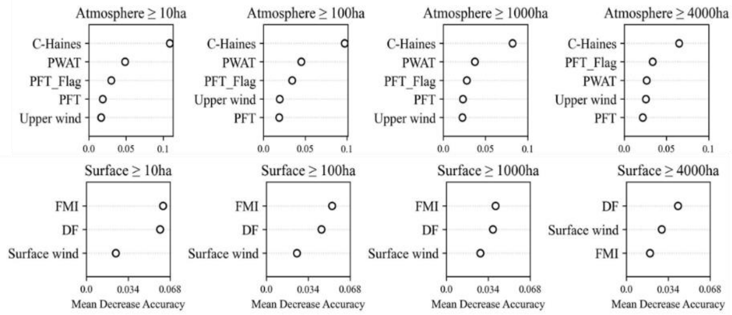

Figure 3 The importance ranking of atmospheric and surface weather variables in the whole study area.

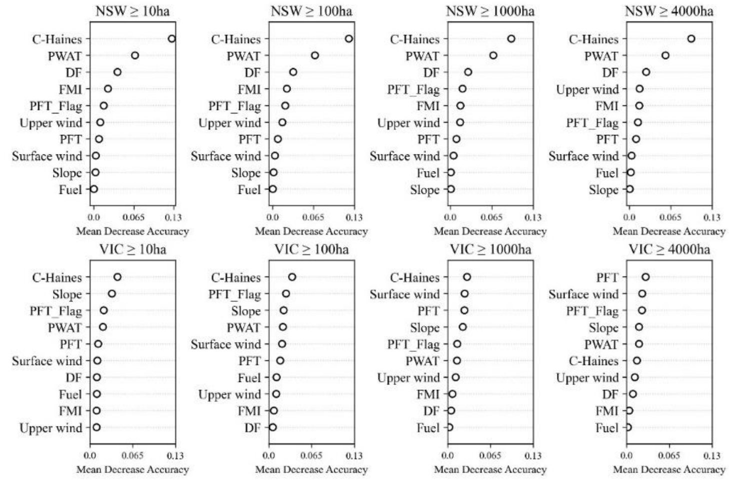

Figure 4 The importance ranking of all variables in different regions.

In the Random Forest models of different regions, the variable importance rankings for NSW and VIC exhibit notable differences (Figure 4). In the models with all variables, the importance of surface fire weather variables, particularly DF, is slightly higher in NSW compared to the entire study area. Meanwhile, categorical variables, particularly Slope, show significantly increased importance in VIC.

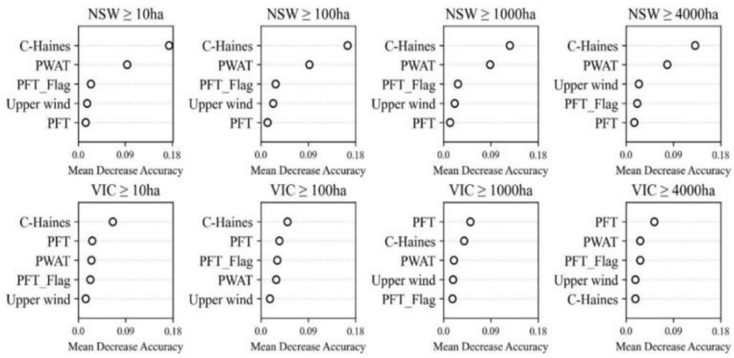

Figure 5 The importance ranking of atmospheric in different regions.

In the atmospheric-only models (Figure 5), C-Haines remains the most important variable across all subsets in NSW. However, its importance decreases as fire size increases in VIC. This spatial heterogeneity may be partly related to data quality issues in Victoria (Ma et al., 2023), while another possible reason is that extreme fire events in VIC are often associated with frontal passages, where synoptic-scale dynamics—such as temperature gradients, wind shifts, and uplift along the front—become the dominant drivers of convection, thereby reducing the relative explanatory power of C-Haines compared with smaller fires. In the surface-only models (Figure 6), the variable importance results differ sharply between NSW and VIC. In NSW, Surface wind is the least important variable, whereas in VIC, it serves as the most important variable.

The partial dependence plot (Figure 7) illustrates how the probability of pyroCb occurrence varies with the values of each variable. Both C-Haines and DF are positively correlated with the probability of pyroCb occurrence, with the probability increasing more rapidly once the variables exceed certain thresholds (C-Haines > 10 and DF > 7). PWAT initially shows a positive correlation with pyroCb occurrence probability, but this relationship becomes negative once PWAT exceeds 9 mm. The probability decreases with FMI, PFT, and PFT_Flag, each at a different rate. The decrease with FMI is the slowest, leveling off when FMI exceeds 13. The decrease with PFT is relatively faster but remains consistent. The decrease with PFT_Flag is the steepest but confined to the range where PFT_Flag is between 20 and 60; below 20, the probability shows an upward trend, and above 60, the probability becomes stable. In addition, the effects of Surface wind and Upper wind on the probability of pyroCb occurrence are relatively flat and exhibit minimal fluctuations.

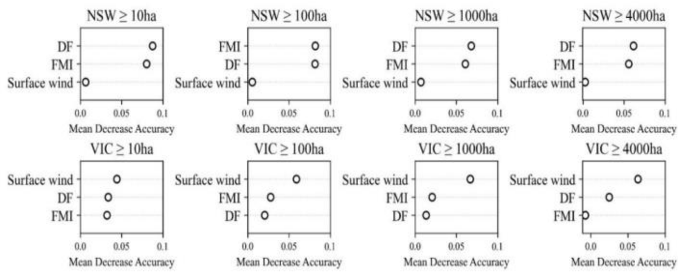

Figure 6 The importance ranking of surface fire weather variables in different regions.

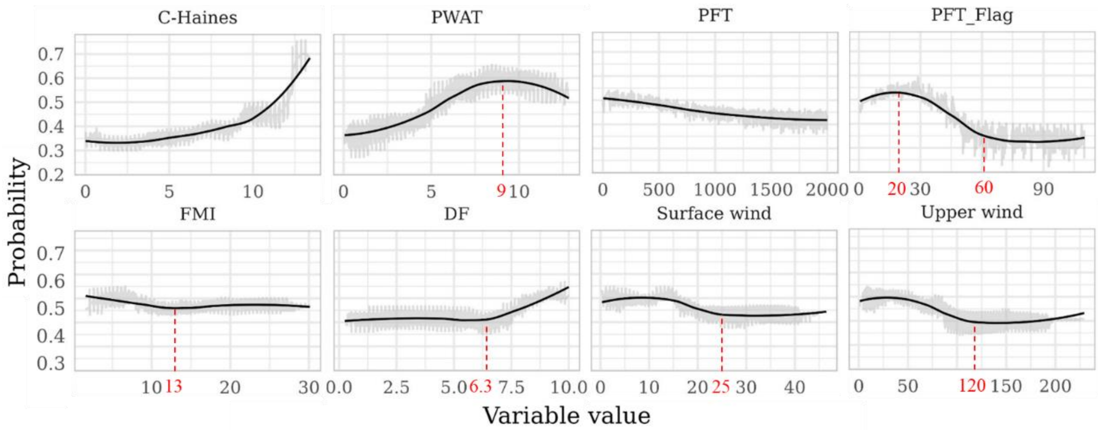

Figure 7. Partial dependence plots for different explanatory variables. The horizontal axis represents the values of given variable; the vertical axis represents the probability of pyroCb occurrence. Important thresholds are highlighted in red. The black curve represents the relationship between probability and variable values, obtained by smoothing the mean results across 100 samples. The grey shading represents the 95% confidence interval of the results from these 100 samples.

3.2. Model accuracy

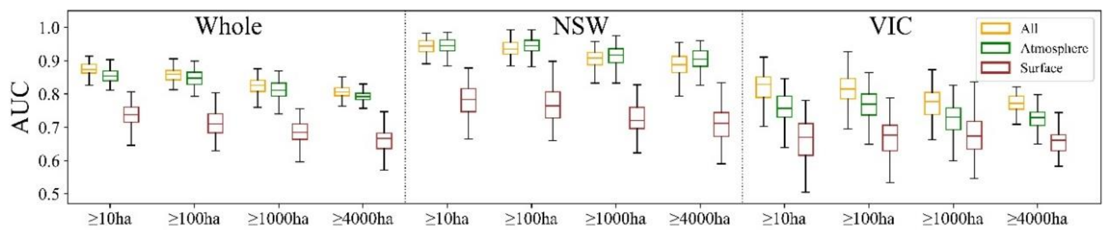

Figure 8. The AUC of Random Forest models over whole study area, NSW and VIC. The labels on the horizontal axis represent the area of standard wildfires. "All" in the legend indicates the all-variable model; "Atmosphere" and "Surface" refer to models using only atmospheric and surface fire weather variables, respectively.

The AUC (Area Under the Curve) from ROC (Receiver Operating Characteristics) curve (Hanley *et al.,* 1982) is used to evaluate model accuracy. The closer the AUC is to 1, the higher the model's accuracy. All all-variable models achieved AUC values above 0.8, indicating that the all-variable random forest models have strong explanatory power for pyroCb occurrence. The models including only atmospheric variables also performed well, with all AUC values exceeding 0.7—in both the whole study area and the VIC subset, their AUC was only slightly lower than that of the all-variable models, and in the NSW subset, the atmospheric-only model even outperformed the all-variable model. By contrast, models including only surface fire weather variables yielded notably lower AUC values, further highlighting the superior explanatory power of atmospheric variables compared to surface fire weather variables, particularly in NSW. Additionally, model performance was highest in the NSW subset and lowest in VIC (Figure 8).

4. DISCUSSION AND CONCLUSION

This study employed the Random Forest algorithm to quantify the importance of variables influencing pyroCb occurrence and to assess their predictive potential. The variable importance rankings obtained are broadly consistent with those reported in previous studies based on logistic regression (Ma et al., 2025), but some differences were observed, particularly in the evaluation of categorical variables (Slope and Fuel). In the logistic regression analyses, categorical variables, especially Slope, were assigned very high importance, whereas their importance in the Random Forest model was obviously lower. This discrepancy may be attributed to the way categorical variables are handled in logistic regression, where they are typically converted into 0/1 dummy variables. Changes in the state of categorical variables can result in abrupt shifts in predicted probability, leading to marginal effects and potential overestimation of their importance (Kuhn and Johnson 2013). Another notable difference is that atmospheric variables such as PWAT and Upper wind were assigned greater importance in the Random Forest model. This likely reflects the models' differing capacities to capture variable relationships. The relationships between PWAT, Upper wind, and pyroCb occurrence are not strictly linear (Figure 7) or their influence may only become significant beyond certain thresholds. Such non-linear or threshold-based relationships are difficult to capture with logistic regression but can be naturally modelled by Random Forest through its decision tree structure, which may explain the higher importance scores assigned to these variables.

The similarity between this study and previous research based on logistic regression models (Ma et al., 2025) is that C-Haines is consistently identified as the most important variable, while surface fire weather variables are generally ranked lower than atmospheric variables. This further highlights the crucial role of atmospheric variables, particularly C-Haines, in assessing pyroCb risk. It's also noteworthy that a simple index like C-Haines has better explanatory power than a more sophisticated and process-based metric like PFT. In addition, the variable importance rankings differ noticeably across the whole study area, NSW, and VIC, highlighting the importance of accounting for spatial variability in pyroCb risk assessments.

The Random Forest model offers notable advantages, such as the ability to effectively compute variable importance and flexibly capture non-linear relationships between predictors and outcomes, making it a valuable approach for assessing pyroCb risk. However, we found that the AUC of the Random Forest models (Figure 8) were lower than reported in previous studies using logistic regression (Ma et al., 2025). This discrepancy may be attributed to the relatively small sample size used in this study, which can lead to overlapping samples across decision trees within the Random Forest algorithm and, in turn, reduce overall model performance. In contrast, the traditional logistic regression model may provide more stable performance under such conditions. Although machine learning algorithms have been widely applied across various fields with impressive results, the rarity of pyroCb events and the limited availability of data pose ongoing challenges to enhancing the generalizability and predictive accuracy of these models. Their potential for assessing natural hazard risks, particularly in small-sample contexts, warrants further investigation. Moreover, selecting an appropriate algorithm is crucial. This study presents an initial exploration of the potential of machine learning in pyroCb risk assessment. Future work will examine other algorithms that may be better suited to small-sample data for more robust and detailed modelling.

ACKNOWLEDGMENTS

This project has been assisted by the New South Wales Government through the NSW Bushfire and Natural Hazards Research Centre.

REFERENCES

Australian Government Department of Agriculture, Water and the Environment. National Vegetation Information System (2020). Viewed October 2021. https://www.environment.gov.au/fed/catalog.

Abatzoglou JT, Williams AP & Barbero R (2019) 'Global Emergence of Anthropogenic Climate Change in Fire Weather Indices', Geophysical Research Letters, 46(1), 326.

Badlan RL, Sharples JJ, Evans JP & McRae RHD (2021a) Factors influencing the development of violent pyroconvection. Part I: fire size and stability, International Journal of Wildland Fire, 30(7), 484.

Badlan RL, Sharples JJ, Evans JP & McRae RHD (2021b) Factors influencing the development of violent pyroconvection. Part II: fire geometry and intensity, International Journal of Wildland Fire, 30(7), 498.

Breiman L (Ed.) (1984) Classification and Regression Trees, The Wadsworth statistics/probability series, Wadsworth International Group: Belmont, Calif.

Department of Environment, Land, Water & Planning. Victoria Fire History Records, Viewed 5 May 2022. https://data.aurin.org.au/dataset/vic-govt-delwp-datavic-fire-fire-history-na.

Di Virgilio G, Evans JP, Blake SAP, Armstrong M, Dowdy AJ, Sharples J & McRae R (2019) Climate Change Increases the Potential for Extreme Wildfires, Geophysical Research Letters, 46(14), 8517.

Dowdy AJ (2018) 'Climatological Variability of Fire Weather in Australia', Journal of Applied Meteorology and Climatology, 57(2), 221.

Dowdy AJ, Ye H, Pepler A, Thatcher M, Osbrough SL, Evans JP, Di Virgilio G, & McCarthy N (2019) Future changes in extreme weather and pyroconvection risk factors for Australian wildfires, Scientific Reports, 9(1), 10073.

Gallant J & Austin J (2012) 'Slope derived from 1" SRTM DEM-S', Viewed 10 March 2025, https://data.csiro.au/collections/#collection/CIcsiro:5142v4.

Guo F, Selvalakshmi S, Lin F, Wang G, Wang W, Su Z & Liu A (2016) 'Geospatial information on geographical and human factors improved anthropogenic fire occurrence modelling in the Chinese boreal forest', Canadian Journal of Forest Research, 46(4), 582.

Hersbach H, Bell B, Berrisford P, Biavati G, Horányi A, Muñoz Sabater J, Nicolas J, Peubey C, Radu R, Rozum I, Schepers D, Simmons A, Soci C, Dee D & Thépaut J-N (2018) ERA5 Hourly Data on Pressure Levels from 1940 to Present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

Kuhn M, Johnson K (2013) Applied Predictive Modeling, Springer New York: New York, NY, Viewed 5 August 2025, http://link.springer.com/10.1007/978-1-4614-6849-3.

Ma W, Feng Z, Cheng Z, Chen S & Wang F (2020) Identifying Forest Fire Driving Factors and Related Impacts in China Using Random Forest Algorithm, Forests, 11(5), 507.

Ma W, Wilson CS, Sharples JJ & Jovanoski Z (2023) Investigating the Effect of Fuel Moisture and Atmospheric Instability on PyroCb Occurrence over Southeast Australia, Atmosphere, 14(7), 1087.

Ma W, Sharples JJ, Wilson CS, Jovanoski Z (2025) Geographical drivers of pyrocumulonimbus and their regional differences over temperate southeast Australia, *Geomatics, Natural Hazards and Risk* (under review).

McRae R (2025) 'Australia PyroCb Register', Published on-line, Viewed 29 January 2022, https://www.highfirerisk.com.au/pyrocb/register.htm.

Mill GA & McCaw WL (2010) Atmospheric Stability Environments and Fire Weather in Australia: Extending the Haines Index, Centre for Australian Weather and Climate Research: Melbourne, Viewed 11 August 2022, https://nla.gov.au/nla.obj-2968722768.

Ndalila MN, Williamson GJ, Fox-Hughes P, Sharples J & Bowman DMJS (2020) 'Evolution of a pyrocumulonimbus event associated with an extreme wildfire in Tasmania, Australia', Natural Hazards and Earth System Sciences, 20(5), 1497.

NSW Rural Fire Service. 'NSW Fire History', viewed November 9, 2021. https://portal.spatial.nsw.gov.au.

Oliveira S, Oehler F, San-Miguel-Ayanz J, Camia A & Pereira JMC (2012) 'Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest', Forest Ecology and Management, 275, 117.

Sharples JJ, McRae RHD, Weber RO, Gill AM (2009) 'A simple index for assessing fire danger rating', Environmental Modelling & Software, 24(6),764.

Tazi K, Salas-Porras ED, Braude A, Okoh D, Lamb KD, Watson-Parris D, Harder P & Meinert N (2022) 'pyrocast: a Machine Learning Pipeline to Forecast Pyrocumulonimbus (PyroCb) Clouds', Viewed 23 December 2024, https://arxiv.org/abs/2211.13052.

Tory KJ (2019) Pyrocumulonimbus firepower threshold: A pyrocumulonimbus prediction tool. Australian Journal of Emergency Management Monograph, 5, 21.

Tory KJ and Kepert JD (2022) Pyrocumulonimbus Firepower Threshold: Assessing the Atmospheric Potential for pyroCb, Weather and Forecasting, 36(2),439.

