PYROCAST: a Machine Learning Pipeline to Forecast Pyrocumulonimbus (PyroCb) Clouds

Kenza Tazi
Department of Engineering
University of Cambridge
Cambridge, UK
kt484@cam.ac.uk

Emiliano Díaz Salas-Porras
Department of Electrical Engineering
University of Valencia
Valencia, Spain
emiliano.diaz@uv.es

Ashwin Braude
Laboratoire Atmosphères,
Milieux, Observations Spatiales
Institut Pierre-Simon Laplace
Guyancourt, France

Daniel Okoh
Centre for Atmospheric Research
National Space Research
and Development Agency
Abuja, Nigeria

Kara D. Lamb
Department of Earth and
Environmental Engineering
Columbia University
New York, US

Duncan Watson-Parris
Department of Atmospheric,
Oceanic and Planetary Physics
University of Oxford
Oxford, UK

Paula Harder
Fraunhofer Institute for
Industrial Mathematics
Kaiserslautern, Germany

Nis Meinert
Pasteur Labs & ISI
New York, US

Abstract

Pyrocumulonimbus (pyroCb) clouds are storm clouds generated by extreme wildfires. PyroCbs are associated with unpredictable, and therefore dangerous, wildfire spread. They can also inject smoke particles and trace gases into the upper troposphere and lower stratosphere, affecting the Earth's climate. As global temperatures increase, these previously rare events are becoming more common. Being able to predict which fires are likely to generate pyroCb is therefore key to climate adaptation in wildfire-prone areas. This paper introduces PYROCAST, a pipeline for pyroCb analysis and forecasting. The pipeline's first two components, a pyroCb database and a pyroCb forecast model, are presented. The database brings together geostationary imagery and environmental data for over 148 pyroCb events across North America, Australia, and Russia between 2018 and 2022. Random Forests, Convolutional Neural Networks (CNNs), and CNNs pretrained with Auto-Encoders were tested to predict the generation of pyroCb for a given fire six hours in advance. The best model predicted pyroCb with an AUC of 0.90 ± 0.04.

1 Introduction

More than 17 million people have been affected and USD 144 billion lost through major wildfire events over the last 30 years [6]. In addition, the degradation of air quality due to the creation of aerosols and ozone from fires results in between 260 000 and 600 000 premature deaths each year [8].

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

The risk posed by wildfires to people and the environment is increasing due to climate change. By the end of the century, the frequency of wildfires, compared to a 2000 – 2010 reference period, is predicted to increase by a factor of 1.31 to 1.57 with the number of extreme wildfires increasing even further [25, 2].

Pyrocumulonimbus (pyroCb) clouds are produced by large and intense wildfires [13, 19]. PyroCbs create their own weather fronts which can make wildfire behaviour unpredictable through strong winds and ignite new fires through lightning strikes [16]. PyroCbs also inject wildfire aerosols and trace gases into the stratosphere where they remain for several months [16, 28, 15]. These events, which can be on the scale of a volcanic eruption, have important impacts on the Earth's climate, e.g., through direct absorption of solar radiation or cloud formation [10, 24, 9]. PyroCbs could also hinder the recovery of the ozone layer [18, 20].

Despite the risk posed by pyroCbs, the conditions leading to their occurrence and evolution are still poorly understood. Previous pyroCb research has generally been limited to the study of occurrences linked to single wildfire events or over a limited study areas. Geostationary and local weather information has been used to characterise the properties of these clouds [23, 12] and to create an empirical thresholding algorithm to detect pyroCb from GOES17 satellite imagery [13]. This paper presents a machine learning pipeline named 'PYROCAST', with the aim of monitoring, forecasting, and understanding the drivers behind pyroCb events. This pipeline will aid scientists to systematically research pyroCb, as well as help policymakers and emergency services efficiently allocate resources and evacuate residents and emergency responders.

PYROCAST [1] is made up of three components:

1. A database (Section 2) containing labelled geostationary satellite images, meteorological, and fuel data associated with individual wildfire events. To the best of the authors' knowledge, the database is the most comprehensive pyroCb database published to date.
2. A forecast model (Section 3), a Random Forest (RF) model that can predict pyroCb occurrence from a wildfire over a six-hour horizon. This is the first model that attempts to forecast pyroCb and the first application of machine learning to this field. Convolutional Neural Networks (CNNs) and CNNs pretrained with Auto-Encoders (AE-CNNs) are also explored in this study.
3. A causal discovery framework, a set of tools to understand the characteristics and causes of pyroCb published in a companion paper [3]. This is the first attempt to create a causal model for pyroCb formation.

2 PYROCAST Database

To create the PYROCAST database, historical pyroCb events were manually collated from blogs and sparse databases including the Australian PyroCb registry [11], the CIMSS PyroCb Blog [17], Annastasia Sienko's master's thesis [23] and the PyroCb Online Forum [5]. The events were then matched to known historical wildfire events in the GlobFire database [1] to determine the start and end date of each fire. PyroCb that could not be associated with any wildfire were given arbitrary star and end dates, three days before and after the pyroCb sighting respectively.

Geostationary satellite imagery from Himawari-8, GOES16, and GOES17 was then matched to the wildfire locations and dates. The spatial coverage of these satellites overlaps with most of the recorded pyroCb located in North America, Australia and Russia. The satellite imaging instruments also have high temporal (10 min) and spatial resolutions ($0.5 - 2$ km) to study the evolution of the wildfires. Six wavelength channels (0.47 µm, 0.64 µm, 0.86 µm, 3.9 µm, 11.2 µm, and 13.3 µm) were chosen to detect and predict pyroCb. Images were downloaded on an hourly basis during local daytime hours at the wildfire locations from Amazon Web Services [21, 22] using a custom parallelisation pipeline. Each scene was interpolated to 1km resolution and cropped to 200 by 200-pixel image.

The cropped geostationary images were then fed into a pyroCb detection algorithm developed by Peterson et al. at the US Naval Research Laboratory (NRL) [13]. The algorithm was used to output pyroCb masks and flags for the forecast model. Finally, the geostationary images were matched with meteorological and fuel data from a climate reanalysis model. For this study, ERA5 reanalysis

[1] Code and data are available at `https://doi.org/10.56272/fpib2524`

[7] was used as it is global, up-to-date, accurate, has a high temporal resolution (hourly), and is easily accessible via the Copernicus Data Store API [4]. Nineteen variables were downloaded and interpolated to the geostationary grid. The full list of geostationary satellite and ERA5 variables, with reasons for selection, can be found in Table 2 of the Appendix. Overall, the PYROCAST database includes over 148 PyroCb events linked to 111 wildfires between 2018 and 2022, equivalent to over 18 thousand hourly observations. Most importantly, it is science and machine learning ready, allowing for the systematic study of the characteristics and causes of PyroCb.

3 PYROCAST Forecast Model

3.1 Setup

To create a pyroCb forecast model, multiple experiments with a combination of model architectures were performed: Random Forest (RF), Convolutional Neural Network (CNN), and Auto-Encoder pretrained Convolutional Neural Network (AE-CNN). `PyTorch` was used to implement CNNs and AE, and `Scikit Learn` for the RFs and metrics. Models were trained for three different learning tasks: 'detection' where labels and input correspond to the same timestamp, 'forecast with weather oracle' where output labels and meteorological inputs correspond to timestamps six hours after that of the geostationary inputs, and 'forecast' where output labels correspond to timestamps six hours after that of the geostationary and meteorological inputs.

The models were trained on five different types of input: geostationary channels (gs), three meteorological variables (w3) selected by importance using a RF for a detection task (convective available potential energy, boundary layer height, and relative humidity at $650\,$hPa), geostationary channels and important meteorological variables (gs+w3), all meteorological variables (w19), and geostationary and all meteorological variables available (gs+w19). Only data from North America and Australia were used for training. A cross-validation scheme with five folds was used given that multiple pyroCb observations can belong to the same wildfire event. The folds were defined by randomly assigning wildfire events to one of five clusters. A fold scheme defined by clustering by latitude and longitude is also presented in the Appendix. These results correspond to the performance that could be expected applying the forecast model to an unseen region such as Europe.

3.2 Training

The RF models consisted of 500 trees, where tree splits were chosen to minimise Gini impurity. For the RF models, 11 features were aggregated across the spatial dimensions: mean, standard-deviation, minimum, maximum, and the quantiles corresponding to the following percentiles: 0.01, 0.05, 0.25, 0.5, 0.75, 0.95, 0.99. Trees were grown to a maximum depth of 10 splits. Each tree was trained on a bootstrap subsample of the data. In evaluating the quality of splits, weighting was used to compensate for unbalanced labels.

The CNN model consisted of six convolutional layers, followed by two fully connected layers. Max pooling was used after each convolutional layer and drop-out (25% drop-out probability) was used after the convolutional layers. ReLU activation functions were used for each layer. A cross-entropy loss function was used to train the CNN classifier. ADAM optimization with a batch size of 64 and a learning rate of 0.001 was performed to train the model. CNN models were trained for five epochs only as overfitting was observed after a low number of epochs. This is due to the small amount of independent data (84 events, and $\sim$6k training observations).

AEs were used to obtain pretrained initializations for the weights of the CNN. The CNN described above corresponds to the encoder and a mirror decoder was used to project the final 16-dimensional hidden layer to outputs with the same dimension as the inputs. The mean squared error (MSE) was used to measure the reconstruction loss. We also used ADAM optimization with a batch size of 64 and learning rate of 0.001 to train the AEs. Given the richer labels, the AE-CNN models were trained for 40 epochs without overfitting.

3.3 Results

The average 'Area Under the Curve' (AUC) for each experiment is shown in Table 1. For the 'detection' task, the geostationary channels contributed significantly more to the performance. The AE-

Table 1: Average validation AUC across 5 folds with standard deviations.

		gs	w3	gs+w3	w19	gs+w19
Detection	RF	$\mathbf{0.96 \pm 0.00}$	0.84 ± 0.04	0.96 ± 0.01	0.90 ± 0.04	0.96 ± 0.01
	CNN	0.88 ± 0.19	0.69 ± 0.06	0.89 ± 0.14		
	AE-CNN	$\mathbf{0.97 \pm 0.01}$	0.72 ± 0.05	0.94 ± 0.01		
Forecast with oracle	RF	0.84 ± 0.03	0.83 ± 0.06	0.88 ± 0.04	0.89 ± 0.05	$\mathbf{0.90 \pm 0.04}$
	CNN	0.55 ± 0.06	0.67 ± 0.06	0.66 ± 0.10		
	AE-CNN	0.62 ± 0.06	0.72 ± 0.06	0.72 ± 0.04		
Forecast	RF	0.84 ± 0.03	0.82 ± 0.05	0.87 ± 0.03	0.88 ± 0.05	$\mathbf{0.90 \pm 0.04}$
	CNN	0.55 ± 0.06	0.73 ± 0.03	0.71 ± 0.05		
	AE-CNN	0.62 ± 0.06	0.72 ± 0.03	0.75 ± 0.05		

CNN and RF trained on the geostationary channels were the best detection models. The performance of the 'forecast with weather oracle' and 'forecast' models are similar. In practice, however, the performance for the 'forecast with weather oracle' will diminish when the oracle is replaced by a weather forecast.

For the forecast tasks, both the geostationary and meteorological data contributed to the predictive power of the models. This agrees with what the literature has suggested so far: it takes a combination of a very strong wildfire (which the geostationary imagery can detect) and the right meteorological conditions to generate a pyroCb event [14, 26, 27]. The RF achieved the best performance: 0.90 ± 0.04 AUC for both the 'forecast with weather oracle' and 'forecast' tasks using all the available variables. The quantity of data used to train the models may be too small to leverage the full potential of a CNN model and to exploit the spatially resolved information in the geostationary and meteorological data. The 11 aggregated features may be enough to extract most of the relevant information to predict pyroCb.

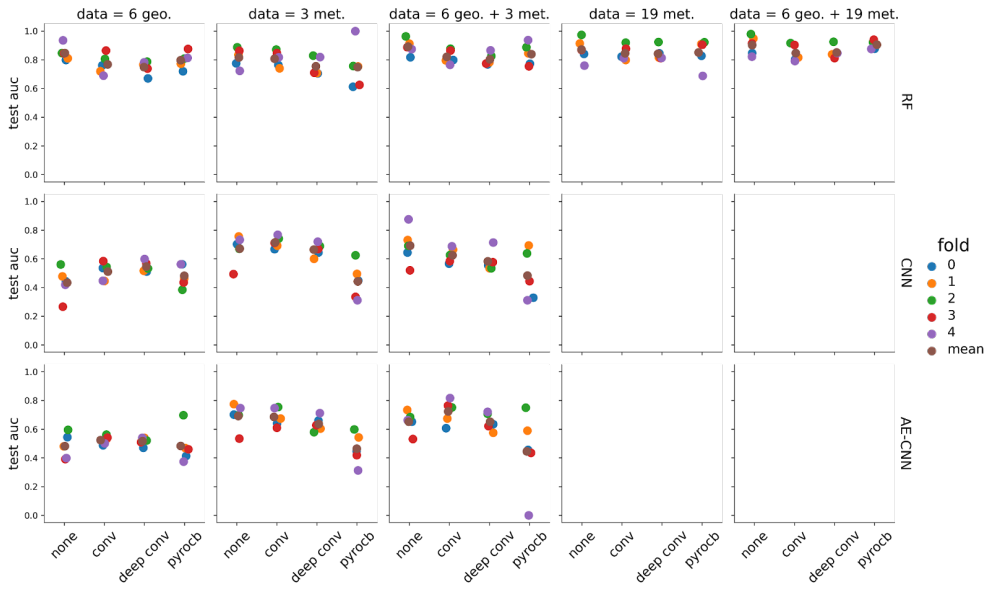

Figure 1: Validation AUC for different forecast models and input sets as a function of the NRL flag: 'None' for no convection or pyroCb observed, 'Convection' for when convection is observed but no deep convection or pyroCb, 'Deep-convection' for when deep convection is observed but no pyroCb, 'PyroCb' for when a pyroCb is already detected.

The performance of the 'forecast' model as a function of the wildfire state was also analysed. These states correspond to a grouping of the NRL algorithm flags (cf. Section 2). Figure 1 summarises the AUC of the forecast model as a function of the current state of the wildfire. In general, the RF model performed best when there was no sign of convection or pyroCb six hours before the label was

recorded. On the other hand, models struggled the most to predict the pyroCb when pyroconvection was already detected.

4 Conclusion

This paper presents PYROCAST, a pipeline to monitor, forecast, and understand the drivers behind pyroCbs clouds formed by extreme wildfires. The PYROCAST database provides the most comprehensive pyroCb dataset published to date with labelled geostationary satellite images and environmental data for 148 pyroCb events across North America, Australia and Russia between 2018 and 2022. The PYROCAST forecast model is trained using this data. A RF model showed the most skill in forecasting pyroCb occurrence from a wildfire six hours in advance with an AUC of 0.90 ± 0.04. This is the first attempt to forecast pyroCb and the first application of machine learning to this field. Further work could include expanding the database through the creation and application of a detection model for PyroCb and PyroCu clouds, the precursors of PyroCb. More data would in turn improve forecast performance. Finally, an evaluation of model uncertainty and confidence could also enable better decision making if the forecast model is deployed during a wildfire.

Acknowledgements

This work is the result of the 2022 Frontier Development Lab Europe Research Sprint. We are grateful for the support of the organizers, mentors, and sponsors. In particular, we would like to thank David Peterson, Michael Fromm, Annastasia Sienko, and Raul Ramos for their insight and help. Funding for this study was also provided by the European Research Council (ERC) Synergy Grant "Understanding and Modelling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant agreement No. 855187).

References

[1] Tomàs Artés, Duarte Oom, Daniele De Rigo, Tracy Houston Durrant, Pieralberto Maianti, Giorgio Libertà, and Jesús San-Miguel-Ayanz. A global wildfire dataset for the analysis of fire regimes and fire behaviour. *Scientific data*, 6(1):1–11, 2019.

[2] Andrew J Dowdy, Hua Ye, Acacia Pepler, Marcus Thatcher, Stacey L Osbrough, Jason P Evans, Giovanni Di Virgilio, and Nicholas McCarthy. Future changes in extreme weather and pyroconvection risk factors for australian wildfires. *Scientific reports*, 9(1):1–11, 2019.

[3] Emiliano Díaz Salas-Porras, Kenza Tazi, Ashwin Braude, Daniel Okoh, Kara Lamb, Duncan Watson-Parris, Paula Harder, and Nis Meinert. Identifying causes of pyrocumulonimbus (pyrocb). In *NeurIPS 2022 Workshop-Causality for Real-world Impact*, 2022.

[4] European Centre for Medium-Range Weather Forecasts. Copernicus data store api, 2019. URL `https://github.com/ecmwf/cdsapi`.

[5] PyroCb Online Forum. Worldwide pyrocb information exchange, 2022. URL `https://groups.io/g/pyrocb`.

[6] Debarati Guha-Sapir, R Below, and P Hoyois. Em-dat: The cred/ofda international disaster database. Université Catholique de Louvain, Brussels, Belgium, 2022. URL `www.emdat.be`.

[7] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[8] Fay H Johnston, Sarah B Henderson, Yang Chen, James T Randerson, Miriam Marlier, Ruth S DeFries, Patrick Kinney, David MJS Bowman, and Michael Brauer. Estimated global mortality attributable to smoke from landscape fires. *Environmental health perspectives*, 120(5):695–701, 2012.

[9] GP Kablick III, Douglas R Allen, Michael D Fromm, and Gerald E Nedoluha. Australian pyrocb smoke generates synoptic-scale stratospheric anticyclones. *Geophysical Research Letters*, 47 (13):e2020GL088101, 2020.

[10] Ziming Ke, Yuhang Wang, Yufei Zou, Yongjia Song, and Yongqiang Liu. Global wildfire plume-rise data set and parameterizations for climate model applications. *Journal of Geophysical Research: Atmospheres*, 126(6):e2020JD033085, 2021.

[11] Richard HD McRae. Australia pyrocb register, 2022. URL `https://www.highfirerisk.com.au/pyrocb/register.htm`.

[12] David Peterson, Edward Hyer, and Jun Wang. Quantifying the potential for high-altitude smoke injection in the north american boreal forest using the standard modis fire products and subpixel-based methods. *Journal of Geophysical Research: Atmospheres*, 119(6):3401–3419, 2014.

[13] David A Peterson, Michael D Fromm, Jeremy E Solbrig, Edward J Hyer, Melinda L Surratt, and James R Campbell. Detection and inventory of intense pyroconvection in western north america using goes-15 daytime infrared data. *Journal of Applied Meteorology and Climatology*, 56(2):471–493, 2017.

[14] David A Peterson, Edward J Hyer, James R Campbell, Jeremy E Solbrig, and Michael D Fromm. A conceptual model for development of intense pyrocumulonimbus in western north america. *Monthly Weather Review*, 145(6):2235–2255, 2017.

[15] David A Peterson, James R Campbell, Edward J Hyer, Michael D Fromm, George P Kablick, Joshua H Cossuth, and Matthew T DeLand. Wildfire-driven thunderstorms cause a volcano-like stratospheric injection of smoke. *NPJ climate and atmospheric science*, 1(1):1–8, 2018.

[16] David A Peterson, Michael D Fromm, Richard HD McRae, James R Campbell, Edward J Hyer, Ghassan Taha, Christopher P Camacho, George P Kablick, Chris C Schmidt, and Matthew T DeLand. Australia's black summer pyrocumulonimbus super outbreak reveals potential for increasingly extreme stratospheric smoke events. *NPJ climate and atmospheric science*, 4(1): 1–16, 2021.

[17] CIMSS PyroCb. Pyrocb blog, 2022. URL `http://pyrocb.ssec.wisc.edu/`.

[18] LA Rieger, WJ Randel, AE Bourassa, and S Solomon. Stratospheric temperature and ozone anomalies associated with the 2020 australian new year fires. *Geophysical Research Letters*, 48 (24):e2021GL095898, 2021.

[19] B Rodriguez, NP Lareau, DE Kingsmill, and CB Clements. Extreme pyroconvective updrafts during a megafire. *Geophysical Research Letters*, 47(18):e2020GL089001, 2020.

[20] Michael J Schwartz, Michelle L Santee, Hugh C Pumphrey, Gloria L Manney, Alyn Lambert, Nathaniel J Livesey, Luis Millán, Jessica L Neu, William G Read, and Frank Werner. Australian new year's pyrocb impact on stratospheric composition. *Geophysical Research Letters*, 47(24): e2020GL090831, 2020.

[21] Amazon Web Services. Noaa geostationary operational environmental satellites (goes) 16, 17 & 18, 2022. URL `https://registry.opendata.aws/noaa-goes/`.

[22] Amazon Web Services. Jma himawari-8, 2022. URL `https://registry.opendata.aws/noaa-himawari/`.

[23] Annastasia V Sienko. Investigation and documentation of pyrocumulonimbus clouds. Master's thesis, University of Wisconsin-Madison, 2017.

[24] Matthias Stocker, Florian Ladstädter, and Andrea K Steiner. Observing the climate impact of large wildfires on stratospheric temperature. *Scientific reports*, 11(1):1–11, 2021.

[25] A Sullivan, E Baker, and T Kurvits. Spreading like wildfire: The rising threat of extraordinary landscape fires, 2022. URL `https://www.unep.org/resources/report/spreading-wildfire-rising-threat-extraordinary-landscape-fires`.

[26] Kevin J Tory and Jeffrey D Kepert. Pyrocumulonimbus firepower threshold: Assessing the atmospheric potential for pyrocb. *Weather and Forecasting*, 36(2):439–456, 2021.

[27] Kevin J Tory, William Thurston, and Jeffrey D Kepert. Thermodynamics of pyrocumulus: A conceptual study. *Monthly Weather Review*, 146(8):2579–2598, 2018.

[28] Pengfei Yu, Owen B Toon, Charles G Bardeen, Yunqian Zhu, Karen H Rosenlof, Robert W Portmann, Troy D Thornberry, Ru-Shan Gao, Sean M Davis, Eric T Wolf, et al. Black carbon lofts wildfire smoke high into the stratosphere to form a persistent plume. *Science*, 365(6453): 587–590, 2019.

Appendix

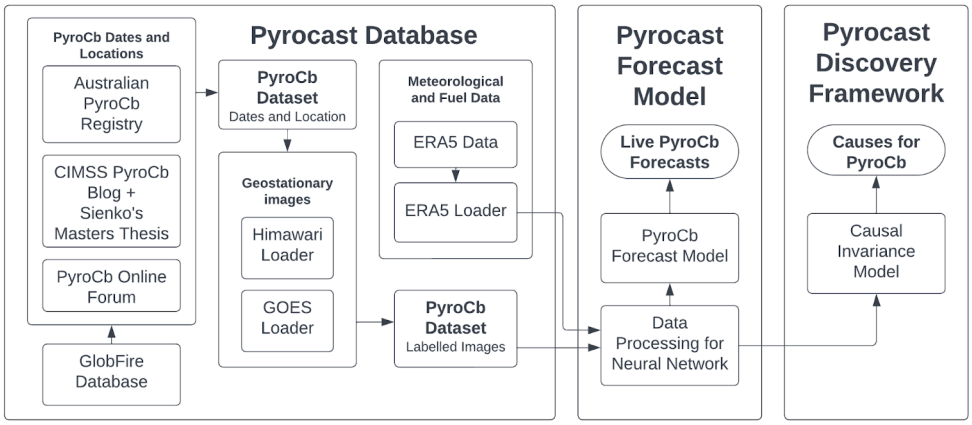

Figure 2: Diagram of PYROCAST pipeline.

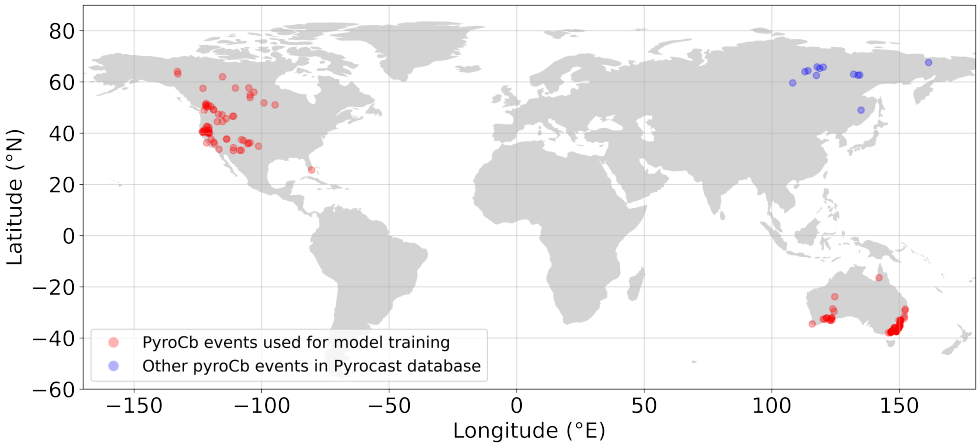

Figure 3: Spatial distribution of pyroCb events in the PYROCAST database.

Table 2: Geostationary satellite and ERA5 fuel and meteorological variables with description and motivation for selecting as pyroCb predictors.

Variable	Description	Sensitive to
*ch*1	0.47 µm	smoke, haze
*ch*2	0.64 µm	terrain type
*ch*3	0.86 µm	vegetation
*ch*4	3.9 µm	thermal emissions & cloud ice crystal size
ch{5,6}	{11.2, 13.3} µm	thermal emissions & cloud opacity
{*u*,*v*}	{*u*,*v*} comp. of wind at 250 hPa	upper-lvl dynamics influencing rising motion
{*u*,*v*}10	10 m {*u*,*v*} component of wind	change in fire intensity and spread
*fg*10	10 m gusts since prev. post-processing	(same as above)
blh	boundary layer height	height of turbulent air at the surface
cape	convective available potential energy	energy for air to ascend into atmosphere
cin	convective inhibition	energy that will prevent air from rising
z	geopotential	energy needed for air to ascend into atmosphere as a function of altitude
{*slhf*, *sshf*}	surface {latent, sensible} heat flux	heat released or absorbed {from, neglecting} phase changes
w	surface vertical velocity	ascent speed of the plume from the wildfire
cv{*h*,*l*}	fraction of {high, low} vegetation	available fuel for the wildfire
type{*H*,*L*}	type of {high, low} vegetation	(same as above)
r{650,750,850}	rel. humidity at {650,750,850} hPa	condensation of vapour into clouds

Table 3: Average validation AUC across five clustered folds with standard deviations. These results are indicative model performance when generalising to an unseen region such as Europe.

		gs	w3	gs+w3	w19	gs+w19
Detection	RF	0.94 ± 0.02	0.76 ± 0.03	0.93 ± 0.03	0.93 ± 0.04	0.93 ± 0.04
	CNN	0.95 ± 0.01	0.65 ± 0.07	0.87 ± 0.16		
	AE-CNN	$\mathbf{0.97 \pm 0.01}$	0.67 ± 0.14	0.91 ± 0.01		
Forecast with oracle	RF	0.70 ± 0.08	0.75 ± 0.05	0.76 ± 0.05	0.79 ± 0.07	$\mathbf{0.80 \pm 0.06}$
	CNN	0.52 ± 0.03	0.65 ± 0.06	0.63 ± 0.08		
	AE-CNN	0.58 ± 0.05	0.59 ± 0.05	0.68 ± 0.06		
Forecast	RF	0.70 ± 0.08	0.70 ± 0.06	0.73 ± 0.08	0.74 ± 0.05	$\mathbf{0.77 \pm 0.06}$
	CNN	0.52 ± 0.03	0.63 ± 0.08	0.64 ± 0.09		
	AE-CNN	0.58 ± 0.05	0.66 ± 0.06	0.68 ± 0.09		

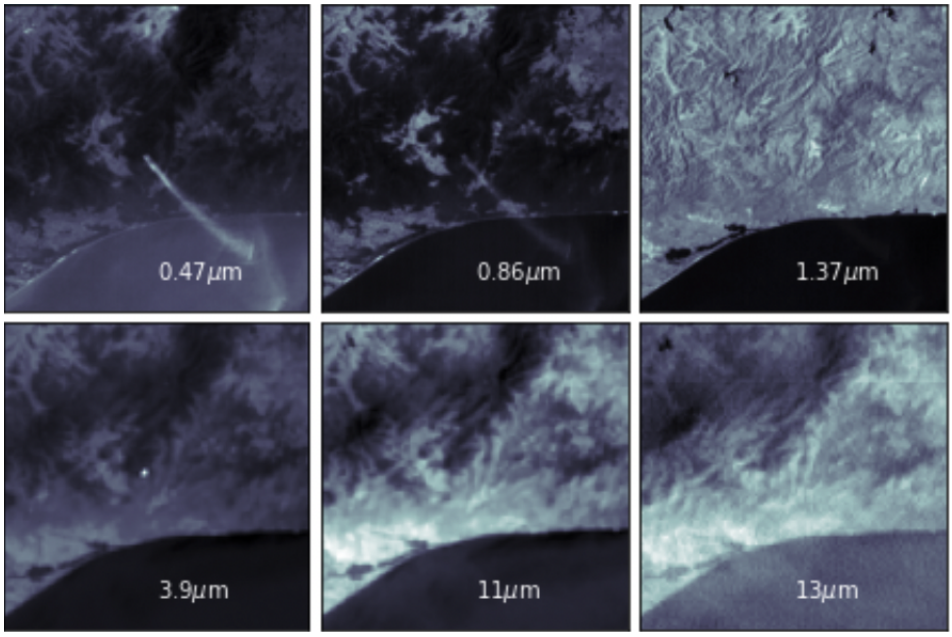

Figure 4: Example of geostationary images over the six wavelength channels a wildfire event in Timbarra, Australia (January 2019).

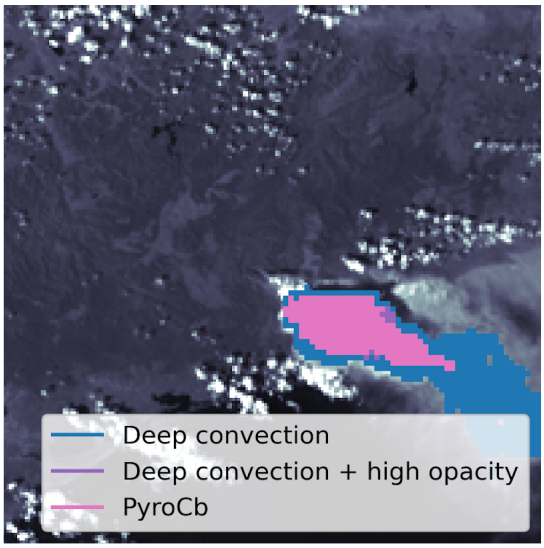

Figure 5: Example of NRL label masks overlaid onto geostationary satellite image (averaged over the 0.47 µm, 0.64 µm and 0.86 µm channels) of a wildfire event in Timbarra, Australia (January 2019).

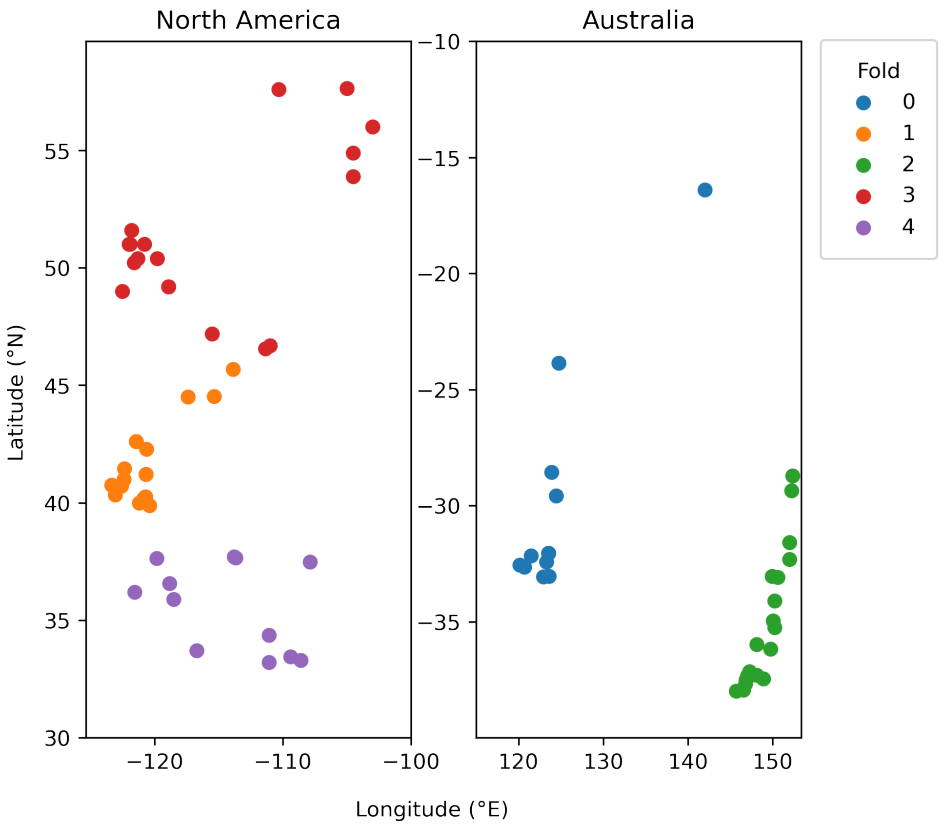

Figure 6: Clustered cross-validation folds.