Identifying causes of Pyrocumulonimbus (PyroCb)

Anonymous Author(s)
Affiliation
Address
email

Abstract

A first causal discovery analysis from observational data of pyroCb (storm clouds generated from extreme wildfires) is presented. Invariant Causal Prediction was used to develop tools to understand the causal drivers of pyroCb formation. This includes a conditional independence test for testing $Y \perp\!\!\!\perp E|X$ for binary variable Y and multivariate, continuous variables X and E and a greedy-ICP search algorithm that relies on fewer conditional independence tests to obtain a smaller more manageable set of causal predictors. With these tools we identified a subset of seven causal predictors which are plausible when contrasted with domain knowledge: surface sensible heat flux, relative humidity at 850hPa, a component of wind at 250 hPa, 13.3 µm thermal emissions, convective available potential energy and altitude.

1 Introduction

More than 17 million people have been negatively affected and USD144 billion lost during major wildfire events occurring during the last 30 years [8]. A changing climate is further increasing the risk that wildfires pose to people and the environment. By the end of the century, the frequency of wildfires is expected to increase by a factor of 1.3 to 1.6, compared to a 2000-2010 reference period [30]. The number of extreme wildfires increasing even further [3]. Furthermore, the degradation of air quality due to the creation of aerosols and ozone from fires resulted in between 260 000 and 600 000 premature deaths each year [12].

Pyrocumulonimbus (also known as PyroCbs) are storm clouds generated by large and intense wildfires [18, 24]. These storm clouds generate their own weather fronts which can make wildfire behaviour increasingly unpredictable: strong winds fan the flames of the fire and change its direction while lightning ignites new fires in the vicinity [21]. This unpredictability increases dangers for firefighters and other first responders, and reduces decision timescales for evacuating surrounding communities. PyroCbs also transport wildfire aerosols into the stratosphere, where they may remain for several months [21, 33, 20]. These events, which can transport amounts of aerosols on the scale of a volcanic eruptions, may have important impacts on the Earth's climate (e.g. through direct absorption of solar radiation or cloud formation [14, 29, 13]). PyroCbs may also disrupt the recovery of the ozone layer [23, 25].

Despite the risk posed by pyroCbs, the conditions leading to their occurrence and evolution are still poorly understood [19, 32, 31]. Although it is well-known that pyroCbs are generated only from large and intense wildfires, surface-based variables and fire weather indices (such as the Haines index [9]) have been shown to be poor predictors of the potential for pyroCb formation [17]. The meteorological environment in which intense fires develop is also an important precursor of pyroCb formation. As smoke plumes rise in the atmosphere, they cool and entrain surrounding colder, dryer air, potentially reducing condensation; it is still not clear how significant a role aerosol particles, moisture, and heat released by the fire itself is to this process [32, 31]. Several competing factors play a role, as the boundary layer must be dry enough to support intense fire activity and wildfire spread rates, while

Submitted to the NeurIPS Causal ML for Real-World Impact Workshop (CML4Impact 2022). Do not distribute.

also supporting conditions favorable for thunderstorm formation. Recent studies suggest that both a deep, dry, and unstable lower troposphere combined with a moist and unstable mid-troposphere are key ingredients for producing favorable conditions for pyroCb [19].

This study presents an analysis of pyroCb causes. This framework is part of the 'Pyrocast' machine learning pipeline that was developed with the aim of monitoring, forecasting and understanding the causes behind pyroCb events. This pipeline was developed by the authors and will be published in a companion publication. In this paper, the Pyrocast database is first summarised. Invariant Causal Prediction for PyroCb is defined in Section 3. Model implementation and results are then discussed in section 4 and 5.

2 Pyrocast database

The Pyrocast database contains geostationary satellite images, meteorological, and fuel data associated with individual wildfire events. It also contains pixel-wise labels of whether the wildfire produced a pyroCb at given space-time coordinates or not. The Pyrocast database contains four sources of information: 1/ images from geostationary satellites: 6 channels; 2/ meteorological information: 15 variables; 3/ fuel information: four variables on the type of vegetation; and lastly 4/ labels indicating the presence of pyroCb. Table 1 includes a description of these variables and their shorthand. All variables are available for a 200 by 200 grid of 1km^2 pixels. For this study we complemented the PyroCast database with the altitude from the world elevation model [4] (denoted alt). We also calculated the magnitude of the wind at 10m and at 250hPa (denoted uv10 and uv250 respectively) using the respective u and v components. The total 6 geostationary channels and 22 meteorological and fuel variables are defined in Table 1.

3 Invariant Causal Prediction

3.1 Problem statement

Consider the setting where we have an i.i.d.[1] sample of the random vectors $X = (X_1, X_2, ..., X_p)^\top \in \mathbb{R}^p$, $E = (E_1, E_2, ..., E_q)^\top \in \mathbb{R}^q$ and $Y \in \{0, 1\}$. For a variable of interest Y, E is a set of environment variables which may be causes of X but neither direct causes of Y, nor effects (direct or indirect) of Y. We assume that Y is generated causally from a subset $S^* \subseteq \{1, ..., p\}$ of the p variables considered, (i.e. there is causal sufficiency) by which we mean it is generated from a Structural Causal Model with equation:

$$Y = g(X_{S^*}, \epsilon), \epsilon \sim F, \epsilon \perp\!\!\!\perp X_{S^*} \tag{1}$$

where g and F are arbitrary function and distribution respectively. A set $S \subseteq \{1, ..., p\}$ is a generic subset of the full set of candidate causes. We refer to S and X_S interchangeably. The problem we deal with in this work is to infer the set S^* of direct causes of the variable of interest Y.

3.2 Linear and nonlinear ICP

The Invariance Causal Prediction (ICP) framework [16] is based on the observation that Y is independent of E given X_{S^*}, denoted $Y \perp\!\!\!\perp E|X_{S^*}$. Assuming we have a set of candidate causes X which includes S^*, the causal subset, and an environment variable E that we know does not directly cause Y or is an effect of Y, then we can search for S^* by applying a conditional independence test $Y \perp\!\!\!\perp E|X_S$ on Y, E and subsets $S \subseteq \{1, ..., p\}$ of X. ICP then selects as causal variables the intersection of all those subsets S where the corresponding conditional independence test is not rejected:

$$\hat{S}^* = \bigcap_{S: p_S > \alpha} S \tag{2}$$

where p_S is the p-value associated to the conditional independence test of $Y \perp\!\!\!\perp E|X_S$, with null-hypothesis corresponding to conditional independence, which is not rejected at significance level α if $p_S > \alpha$.

[1] i.i.d. = independently and identically distributed

Table 1: Geostationary and ERA5 fuel and meteorological variables.

Abbr.	Name	Reason for including
ch1	0.47 µm	sensitive to smoke, haze
ch2	0.64 µm	sensitive to terrain type
ch3	0.86 µm	sensitive to vegetation
ch4	3.9 µm	sensitive to thermal emissions from fire & cloud crystal size
ch5	11.2 µm	sensitive to thermal emissions and cloud opacity (with ch6)
ch6	13.3 µm	sensitive to thermal emissions and cloud opacity (with ch5)
u10	10m u component of wind	possible factor in change in fire intensity and spread
v10	10m v component of wind	
fg10	10m wind gust since previous post processing	
u	u comp. of wind at 250hPa	
v	v comp. of wind at 250hPa	
blh	boundary layer height	height of turbulent air at the Earth's surface
cape	convective available potential energy	energy for smoke plume to ascend into upper atmosphere
cin	convective inhibition	energy that will prevent an air parcel from rising
z	geopotential	energy not for smoke plume to ascend into upper atmospher virtue of a location's altitude
slhf	surface latent heat flux	heat released or absorbed from a phase change
sshf	surface sensible heat flux	heat released or absorbed neglecting a phase change
w	surface vertical velocity	speed of the ascent of the plume from the wildfire
cvh	fraction of high vegetation	available fuel for the wildfire
cvl	fraction of low vegetation	
typeH	type of high vegetation	
typeL	type of low vegetation	
r650	relative humidity at 650hPa	condensation of vapour into clouds
r750	relative humidity at 750hPa	
r850	relative humidity at 850hPa	

In [16] an algorithm to implement ICP is presented which assumes linear relationships between causes X and effects Y and a categorical univariate, variable E. In [10] a more general algorithm is introduced which allows for nonlinear relationships between cause and effects and which allows for continuous and multivariate environment variables E. This algorithm uses a conditional independence test based on regression models of Y on X and Y on (X, E).

However, there are two problems with implementing these algorithms:

- No available conditional independence test for a binary variable Y and continuous, multivariate X and E
- Combinatorial nature of ICP: not feasible to realize all tests needed to include all sets in the case of a large powerset (with cardinality over 250 million).

3.3 Greedy-ICP

Instead, as in [10] we use an *invariant target prediction* conditional independence test. This consists of predicting the target Y with two models. The first uses X_S only as predictors. The second uses X_S and E as predictors. If the null is true, and $Y \perp\!\!\!\perp E | X_S$, then the out-of-sample performance of both models is statistically indistinguishable. To deal with a the binary nature of our target variable Y we compare the ROC curves of both models by performing the non-parametric hypothesis test on the difference of the area under the curve (AUC) introduced by [2]. To deal with the large amount of candidate causal variables we use the following *greedy-ICP* search algorithm to obtain a causal importance ordering of the variables:

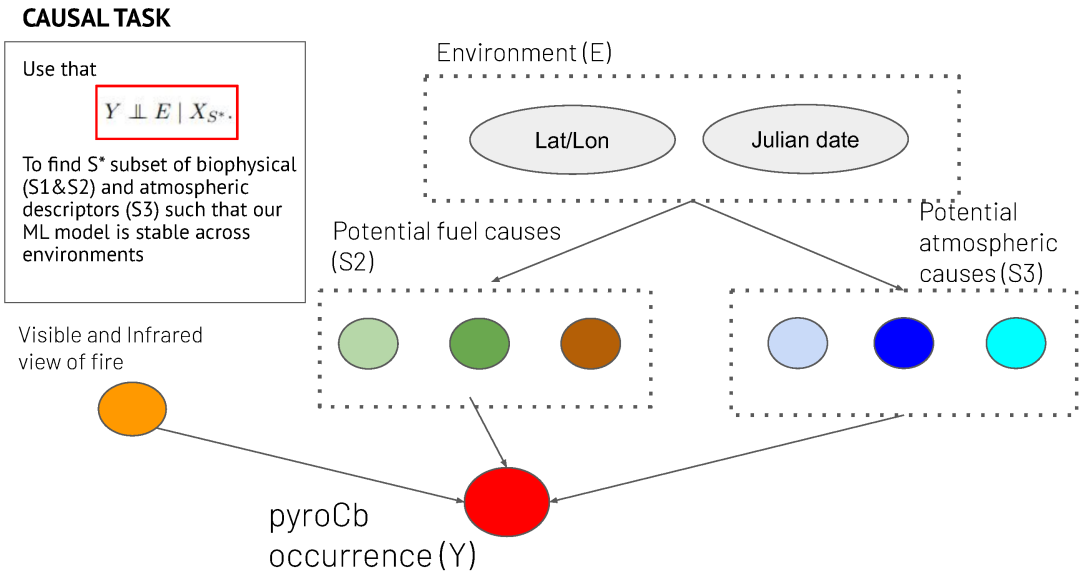

Figure 1: ICP for inferring causes of pyroCb.

The algorithm is initialised by setting $S := \{1, ..., p\}$ to be all candidate causal variables. While the cardinality of S is larger than one:

1. For each variable i in S perform conditional independence test for $S' = S \backslash \{i\}, Y \perp\!\!\!\perp E | X_{S'}$ obtaining corresponding p-value p_i
2. Remove variable j from S such that $j =_i p_i$ and redefine $S \leftarrow S \setminus \{j\}$.

The causal predictors, at a significance level α, are taken to be all those not eliminated when the first p-value below α is computed. It is not guaranteed that the set inferred by greedy-ICP is a superset of that inferred by the exhaustive ICP. Its use in this work is as an exploratory tool to obtain a causal importance ordering of the variables and as a tool to obtain a smaller set of candidate causes on which we can perform the full exhaustive ICP.

4 Setup for pyroCb

To determine the causes of pyroCb formation, the geostationary imagery and the 22 meteorological and fuel variables are used as our candidate causes X. Figure 1 summarizes the ICP framework for inferring the causes of pyroCb.

Each observation consists of a 200 x 200 pixel grid of observations. With the exemption of the discrete categorical variables, we aggregate these observations by taking 11 statistics: mean, standard deviation, minimum, maximum plus 7 quantiles corresponding to the 1, 5, 25, 50, 75, 95 and 99 percentiles. This means we end up with a vector of predictors $X \in \mathbb{R}^{296}$. However, when applying ICP we refer to the original 28 vector valued variables meaning we include or exclude all the different spatial aggregations of a given variable in a model jointly. As environment variables we used the latitude, longitude and start Julian date of the wildfire events.

The conditional independence test used involves fitting models of X_S and E on Y for different subsets S. We use Random Forest (RF) models consisting of 100 trees, where tree splits were chosen

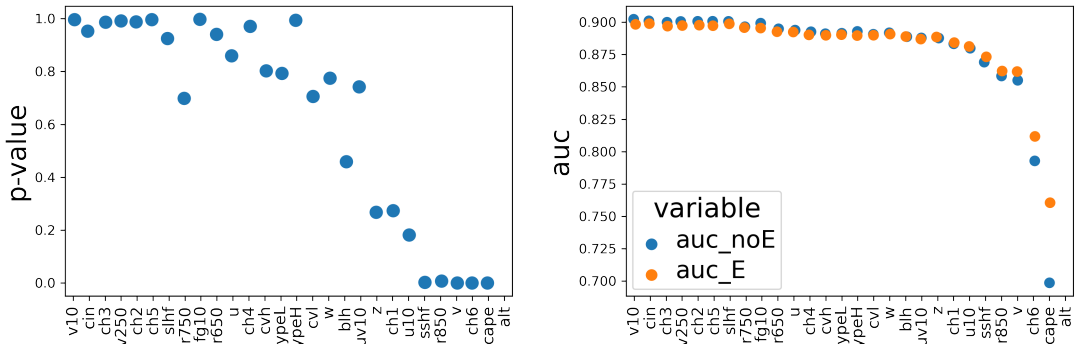

Figure 2: Sequence of p-values (left) and AUC values (right) for corresponding excluded variable produced by greedy ICP.

to minimise Gini impurity. 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 hypothesis test, introduced in [2], on the difference of the AUC between the model containing the environment and that excluding the environment is used as a conditional independence test. If E is conditionally independent of Y given a subset S then including E should not result in a statistically distinguishable AUC. The test is carried out on out-of-sample predictions of Y. For this a 5-fold cross validation (CV) scheme is implemented where each fold is constructed at a wildfire event level by randomly assigning each event to one of five clusters. The generalization performance obtained with this CV scheme is denoted *event generalization* performance.

To validate results from the greedy-ICP algorithm we consider *spatial generalization*. For this we used a 5-fold cross validation scheme where each fold is constructed at a wildfire event level by applying a k-means algorithm on the latitude and longitude of all wildfire events. Figure 8, in the Appendix, illustrates these clusters.

5 Results

Figure 2 shows the sequence of excluded variables, the corresponding maximum p-value of the conditional independence tests and the corresponding AUC of the model with and without the environment E. The order of exclusion of variables with greedy-ICP is the following: v10, cin, ch3, uv250, ch2, ch5, slhf, r750, fg10, r650, u, ch4, cvh, typeL, typeH, cvl, w, blh, uv10, z, ch1, u10, sshf, r850, v, ch6, cape and alt. Notice that the final variable, alt, has no p-value p_i as this refers to the the set $S \setminus \{i\}$ and excluding it would leave the set empty. We can see that with the greedy search algorithm we would exclude all variables except sshf, r850, v, ch6, cape and alt at a 0.05 significance level. This is largely in agreement with what the literature has suggested in the sense that a very large and intense fire (ch6, the 13.3 µm channel sensitive to thermal emission) mid-tropospheric moisture source (r850), conditions favorable for thunderstorms (cape), an unstable boundary layer (sshf) are conditions necessary for pyroCb generation [19, 32, 31]. Additionally, the v wind component at 250hPa and the altitude of the location of the wildfire are found to be important causal factors. This is plausible since wind at high altitudes contributes to an unstable atmosphere and a high surface altitude will diminish the amount of energy needed for the hot moist air to rise and condense at high altitudes. Next we obtain independent validation of this result by performing cross-validation of random forests regression models of Y on X_S which follow the same sequence of excluding variables one-at-a-time as the greedy ICP algorithm. We use the random (event generalization), and spatial (spatial generalization), cross-validation schemes described in section 4. Figure 8 illustrates what events are included in each fold. If greedy ICP has ordered variables correctly, we expect spatial generalization performance to be more robust to excluding non-causal variables: these variables are robust to environment.

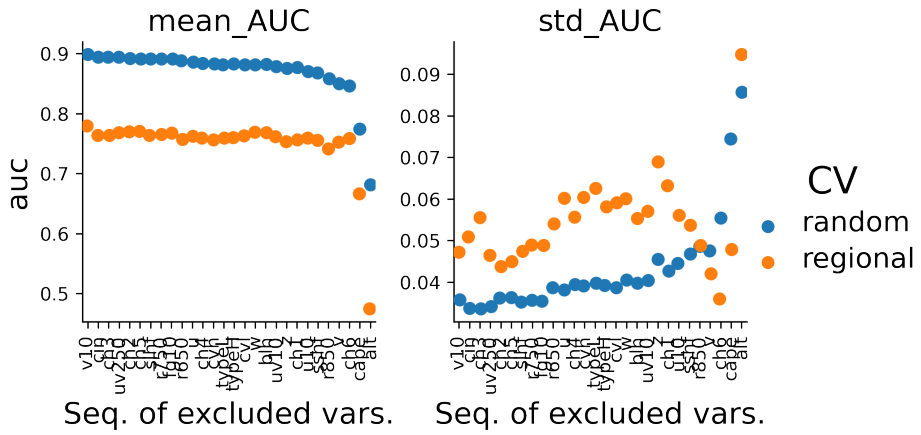

Figure 3: Mean (left panel) and standard deviation (right panel) of out-of-sample AUC for random-CV (blue) and spatial-CV (orange)

Figure 3 indeed shows that excluding non-causal variables leads to negligible loss in spatial generalization performance as measured by mean AUC and actually improves when considering the standard deviation of AUC across spatial folds. In contrast excluding variables deteriorates event generalization (both in mean and standard deviation) since these non-causal variables were detecting local associations between the non-causal variables and pyroCb occurrence.

We subsequently take one of the accepted subsets found with greedy-ICP and perform a more exhaustive ICP on this subset. Although the 7-variable subset $\{u10, sshf, r850, v, ch6, cape, alt\}$ was accepted, we take a larger 11-variable subset $\{blh, uv10, z, ch1, u10, sshf, r850, v, ch6, cape, alt\}$ since excluding blh from S was the first time the p-value dropped below 0.5. We consider all subsets of size 8 or larger of this set. This amounts to 232 subsets. Performing the conditional independence test $Y \perp\!\!\!\perp E|X_S$ on these sets resulted in 60 being accepted at significance level of 0.05. Unfortunately, the intersection of the 60 subsets is the empty set, meaning this more exhaustive version of ICP does not shed light on the causal drivers. As is pointed out in [10] this is often the case when the candidate causal predictors have high dependence among them. One solution pointed out in that work is to consider *defining sets* which are sets of sets where the interpretation is that one of the sets is the true set of causal drivers but we cannot say which. However computing the defining sets for the 60 accepted sets results in a defining set of 45 sets of potential causal drivers. This does not get us much closer to identifying the most plausible causal variables. In section 6.2 of the appendix we deal with dependence and redundancy in our set of 11 predictors by clustering variables together using a measure of the dependence between them and present results for exhaustive ICP on the clusters of variables.

6 Conclusions

Using a novel conditional independence test for $Y \perp\!\!\!\perp E|X_S$ and a greedy-ICP search algorithm we arrived at a set of candidate causes for pyroCb: surface sensible heat flux (sshf), relative humidity at 850hPa (r850), a component of wind at 250 hPa (v), 13.3 µm thermal emissions (ch6), convective available potential energy (cape) and altitude (alt). This set is plausible when contrasted with the existing literature. As more exhaustive implementations of ICP suggested, other sets S of predictors, with few intersecting variables with the one found, are also causally viable in the sense that, given these sets, pyroCb occurrence is independent of the environment. This is likely due to the fact that, as with most real world datasets, there is a lot of dependence and redundancy in the dataset. More exhaustive implementation of the original ICP algorithm either yielded the empty set, or a seemingly insufficient set of predictors to more satisfyingly describe the causal factors behind pyroCb generation. Future work involves learning invariant causal representations of the data such that pyroCb occurrence is independent of the environment given these representations, they are minimal in the sense that they exclude non causal information, and such that they allow us to circumvent the combinatorial nature of ICP (large number of hypothesis needed) and its problems dealing with dependent variables.

References

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] Elizabeth R. DeLong, David M. DeLong, and D L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44 3:837–45, 1988.

[3] 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.

[4] European environment agency. World digital elevation model (etopo5), 2003. URL `https://www.eea.europa.eu/data-and-maps/data/world-digital-elevation-model-etopo5`.

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

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

[7] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, ALT'05, page 63–77, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 354029242X. doi: 10.1007/11564089_7. URL `https://doi.org/10.1007/11564089_7`.

[8] 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`.

[9] Donald A Haines. A lower atmosphere severity index for wildlife fires. *National Weather Digest*, 13:23–27, 1989.

[10] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018. doi: doi:10.1515/jci-2017-0016. URL `https://doi.org/10.1515/jci-2017-0016`.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. 2016. URL `https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssb.12167`.

[17] David A Peterson, Edward J Hyer, James R Campbell, Michael D Fromm, Johnathan W Hair, Carolyn F Butler, and Marta A Fenn. The 2013 rim fire: Implications for predicting extreme fire spread, pyroconvection, and smoke emissions. *Bulletin of the American Meteorological Society*, 96(2):229–247, 2015.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

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

[25] 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.

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

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

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

[29] 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.

[30] 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`.

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

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

[33] 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Research is ongoing. To be published after concluding work.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Standard deviations for AUC results
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] Not significant, can run on any laptop in a few hours.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] Public information
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] To be published at a later date
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] Consent not necessary, data is public.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

6.1 Pyrocast database

The following description of the Pyrocast database is reproduced from a companion paper by the same authors to be published soon.

To create the Pyrocast database, historical pyroCb events were manually collated from blogs and sparse databases including the Australian PyroCb registry [15], the CIMSS PyroCb Blog [22], Annastasia Sienko's master's thesis [28] and the PyroCb Online Forum [6]. 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 dates three days before and after the pyroCb sighting.

Geostationary satellite imagery from Himawari-8, GOES16, and GOES17 was then matched to the wildfire locations and dates. The spatial coverage of satellites overlaps with most of the recorded pyroCb located in North America and Australia. The satellite imaging instruments also have a high temporal (10 min) and spatial resolution (0.5-2km) to study the evolution of the wildfires. Six wavelength channels (0.47µm, 0.64µm, 0.86µm, 3.9µm, 11.2µm, and 13.3µm) were chosen to best detect and predict pyroCb. Images were downloaded on an hourly basis during local daytime hours for the wildfire locations from Amazon Web Services [26, 27] using a custom parallelisation pipeline. All the images were interpolated to 1km resolution and cropped to 200 by 200-pixel images.

The cropped geostationary images were then fed into a pyroCb detection algorithm developed by Peterson et al. at the US Naval Research Laboratory (NRL) [18]. The algorithm is used to output pyroCb masks and flags for the forecasting model. Finally, the geostationary images were matched with meteorological and fuel data from a climate reanalysis model. For this study, ERA5 reanalysis [11] 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 [5]. Nineteen variables were downloaded and interpolated to the geostationary grid.

To the best of our knowledge, the Pyrocast database is the most comprehensive global PyroCb database. It 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.

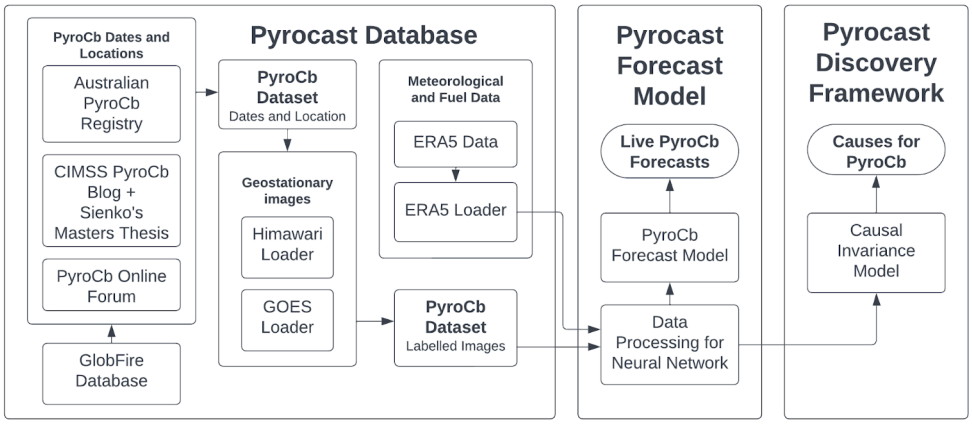

Figure 4: Diagram of Pyrocast pipeline.

6.2 ICP on clusters of variables

In order to deal with dependence and redundancy in the set of 11 predictors, identified in the greedy-ICP algorithm, we measure the dependence between the 11 candidate causal variables in order to cluster variables together. Figure 9 shows the Hilbert Schmidt Independence Criterion (HSIC) measure of dependence between the 11 variables.

HSIC, introduced in [7], is a good choice in this case because of the non-linear nature of the relationship between the predictor variables and because, recall, each variable is multivariate and includes statistical summaries such as the mean and standard deviation, of the variable across the 200 by 200km grid. We use the matrix of dependencies shown in figure 9 to cluster each variable with all other variables with which it has an HSIC measure of 0.25 or larger. This results in the following non-mutually exclusive clusters:{alt, z}, {cape}, {ch6, ch1}, {v}, {r850},{sshf, ch1, blh}, {u10, z, uv10},{ch1, ch6, sshf, blh}, {z, alt, u10, uv10}, {uv10, u10, z} and {blh, sshf, ch1}. We repeat the exhaustive ICP considering this time all subsets of size 8 or larger of these 11 clusters. Since the clusters are not mutually exclusive, the resulting subsets have duplicate variables. Once duplicates are removed this results in 28 unique subsets. We performed the conditional independence test on these sets to obtain 12 subsets for which the conditional independence test was not rejected at a significance level of 0.05. The intersection of these subsets is $\{z, uv10, alt, sshf\}$.

As before, altitude at the location of the wildfire (alt, z) and an unstable boundary layer (sshf) are found to be causal drivers of pyroCb. However the variables r850, v, ch6 and cape, suggested by results from greedy-ICP are not found. It is likely that dependence and redundancy between variables is still an issue causing the exhaustive ICP method to identify sets of variables with a small overlap.

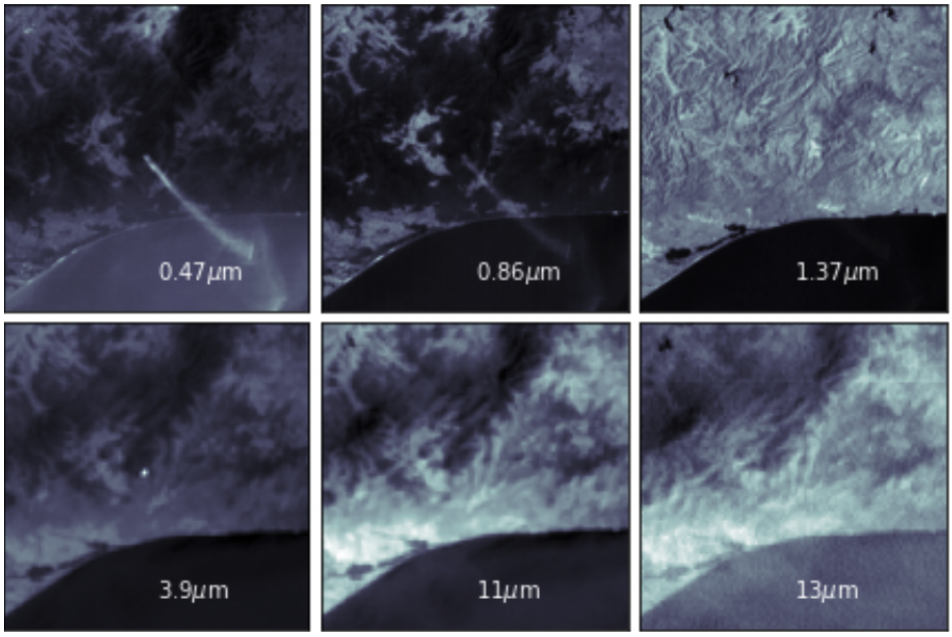

Figure 5: Example of geostationary images over the six wavelength channels of a wildfire event in Timbarra, Australia (January 2019).

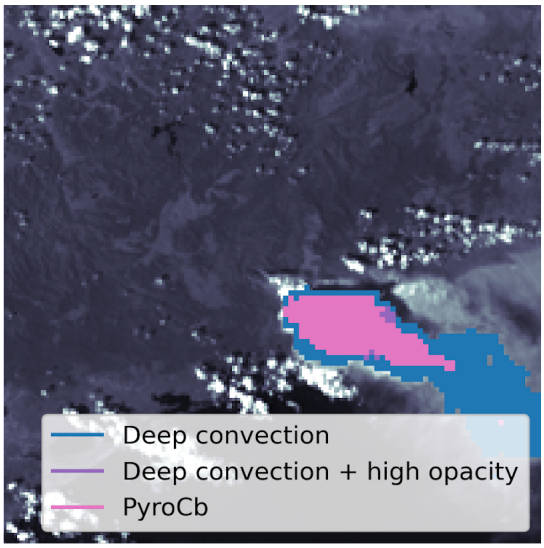

Figure 6: Example of NRL label masks overlaid onto geostationary satellite image (averaged over the 0.47 µm, 0.64 µm and 0.86 µm channels) of a wildfire event in Timbarra, Australia (January 2019).

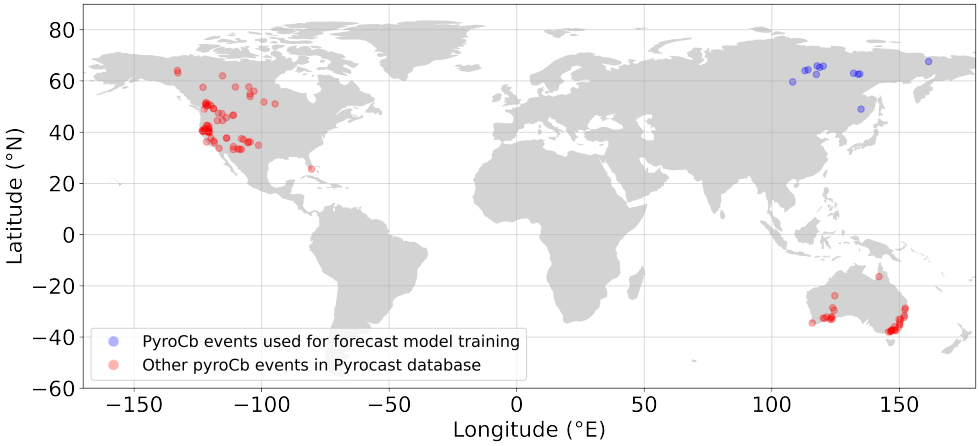

Figure 7: Spatial distribution of pyroCb events in the Pyrocast database.

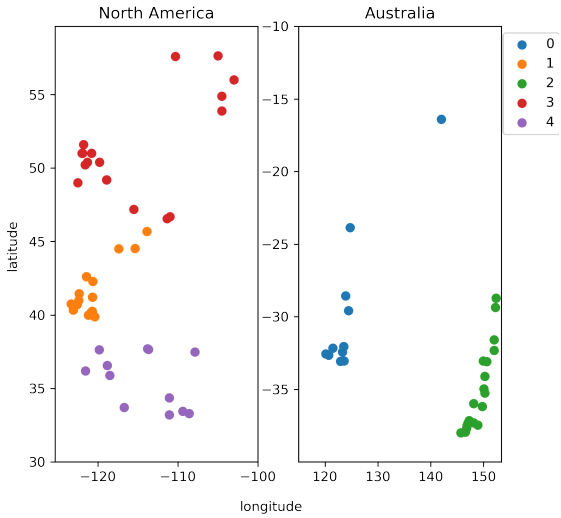

Figure 8: Clusters based on latitude and longitude for spatial cross-validation

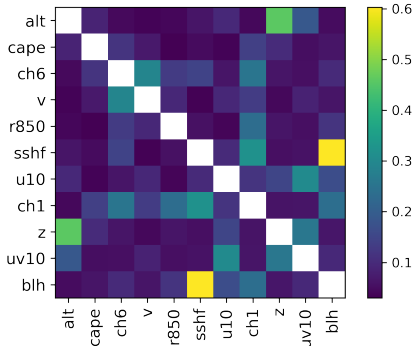

Figure 9: normalized HSIC between each of the 11 variables considered