Identifying the Causes of Pyrocumulonimbus (PyroCb)

Emiliano Díaz Salas-Porras
Image Processing Laboratory
University of Valencia
Valencia, Spain
emiliano.diaz@uv.es

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

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
New York, US

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 850 hPa, 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 $\$144$ billion USD lost during major wildfire events occurring during the last 30 years [8]. A changing climate further increases 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 is increasing to an even greater degree [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 (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 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 timescales for evacuating surrounding communities. PyroCbs also trans-

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

port wildfire aerosols into the stratosphere where they may remain for several months [21, 34, 20]. These events, which can transport amounts of aerosols on the scale of 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, 33, 32]. Although it is well-known that pyroCbs are generated solely 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 the role played by aerosol particles, moisture, and heat released by the fire itself is to pyroCb formation [33, 32]. Several competing factors play a role, as the boundary layer must be dry enough to support intense fire activity and wildfire spread rates, while 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 the causes of pyroCb and is part of the PYROCAST machine learning pipeline that was developed with the aim of detecting, forecasting and understanding the drivers of pyroCb events. This pipeline was developed by the authors of the study presented here and is published in a companion paper [31]. This study is structured as follows: in Section 2, the PYROCAST database is summarised; the Invariant Causal Prediction methodology is introduced in Section 3; implementation details are described in Section 4 and, finally, results are discussed in Section 5.

The PYROCAST database, all code used in its construction, and all code used to produce results in this work are publicly available at `https://spaceml.org/` (`https://doi.org/10.56272/fpib2524`).

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 indicating whether the wildfire produced a pyroCb at given space-time coordinates. The PYROCAST database contains four sources of information: (1) images from geostationary satellites: six 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 200×200 grid of $1\,\text{km}^2$ 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 $10\,\text{m}$ and at $250\,\text{hPa}$ (denoted *uv10* and *uv250*, respectively) using the relevant *u* (east-west) and *v* (north-south) components. The six geostationary channels and 22 meteorological and fuel variables are defined in Table 1.

3 Invariant Causal Prediction

3.1 Problem Statement

Consider the setting of independent and identically distributed samples 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. Let us assume that Y is generated *causally* from a subset $S^* \subseteq \{1, ..., p\}$ of the p variables considered, hence there is causal sufficiency, and Y is generated from a Structural Causal Model obeying:

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

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 for the sake of brevity. The problem we deal with in this work is to infer the set S^* of direct causes of the variable of interest Y.

Table 1: Geostationary and ERA5 fuel and meteorological variables. Description and motivation for selecting as a candidate cause of pyroCb.

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-level dynamics which influence 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

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 that 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:

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

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

In [16] an algorithm to implement ICP is presented that assumes linear relationships between causes X and effects Y, and a categorical, univariate variable E. In [10] a more general algorithm is introduced that allows for nonlinear relationships between cause and effects and 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:

- Unavailability of a conditional independence test for a binary variable Y and continuous, multivariate X and E.
- Combinatorial nature of ICP: it is not feasible to realize all tests needed to include all sets in the case of a large powerset (with cardinality over 250 million).

These issues are addressed with greedy-ICP.

3.3 Greedy-ICP

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, while the second uses both X_S and E. If the null-hypothesis is true, and $Y \perp\!\!\!\perp E|X_S$, then the out-of-sample performance of both models is statistically indistinguishable. To deal with the binary nature of our target variable Y we compare the receiver-operator curves (ROC) of both models by performing the non-parametric hypothesis test on the difference of the area under the curve (AUC) introduced by [2]. To address 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:

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 $Y \perp\!\!\!\perp E|X_{S'}$ where $S' = S \setminus \{i\}$, obtaining corresponding p-value p_i.
2. Remove variable j from S such that $j = \arg\max_i p_i$ 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.

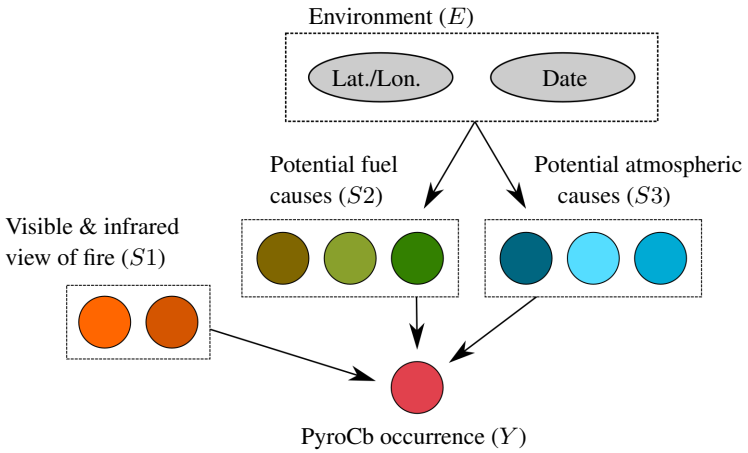

Figure 1: Schematic representation of assumptions made about causes of pyroCb. Boxes group variables (circles) into categories relevant to the pyroCb study. Arrows indicate assumed causal relationships between variables. Arrows that start/end at a given box indicate that at least one of its variables is assumed to be a cause/effect with respect to the end/start point of the arrow. The causal task of ICP is to use that $Y \perp\!\!\!\perp E|X_{S^*}$ (assuming S^* is the set of true causes), in order to find S^*, a subset of biophysical ($S1$ & $S2$) and atmospheric descriptors ($S3$) such the prediction model for Y based on X_{S^*} is stable across environments.

Each observation consists of a 200×200 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, and seven quantiles corresponding to the 1, 5, 25, 50, 75, 95 and 99 percentiles. We arrange these predictors in a new vector $X \in \mathbb{R}^{296}$. However, note that 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 to minimise the 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 in order 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.

We validate results from the greedy-ICP algorithm by considering *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 9 in the Appendix delineates these clusters.

5 Results

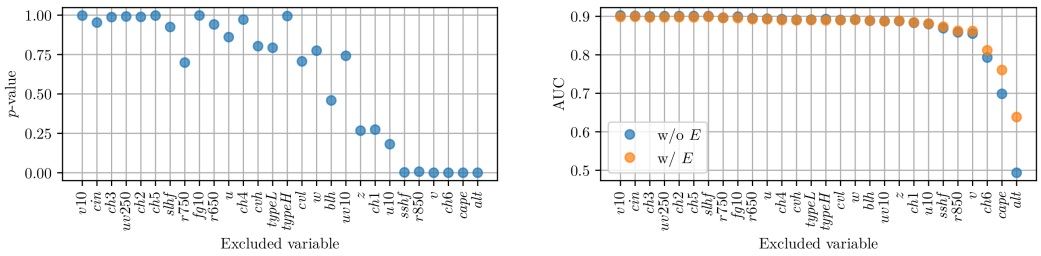

Figure 2: Sequence of p-values (left) and AUC values (right) for sequence of candidate cause sets produced by greedy ICP. In both cases, the abscissa, read from left to right, shows the order in which causal candidates are excluded from consideration by the greedy ICP algorithm. On the left, for every variable i, the p-value corresponding to the null-hypothesis that $Y \perp\!\!\!\perp E | X_S$, where S is composed of excluding all variables up to variable i, is shown. Generally, as more variables are excluded the causally necessary hypothesis of independence starts to be rejected. On the right we show the AUC, for prediction models including the environment variables (w/ E, in blue) and excluding these variables (w/o E in orange), for the same shrinking sequence of candidate causal predictors. As more variables are excluded, including environment information increasingly helps prediction models.)

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*, *ch*3, *uv*250, *ch*2, *ch*5, *slhf*, *r*750, *fg*10, *r*650, *u*, *ch*4, *cvh*, *typeL*, *typeH*, *cvl*, *w*, *blh*, *uv*10, *z*, *ch*1, *u*10, *sshf*, *r*850, *v*, *ch*6, *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 $sshf$, *r*850, v, *ch*6, *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 (*ch*6, the 13.3 µm channel sensitive to thermal emission) mid-tropospheric moisture source (*r*850), conditions favorable for thunderstorms (*cape*), an unstable boundary layer (*sshf*) are conditions necessary for pyroCb generation [19, 33, 32]. Additionally, the v wind component at 250 hPa 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 9 shows 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.

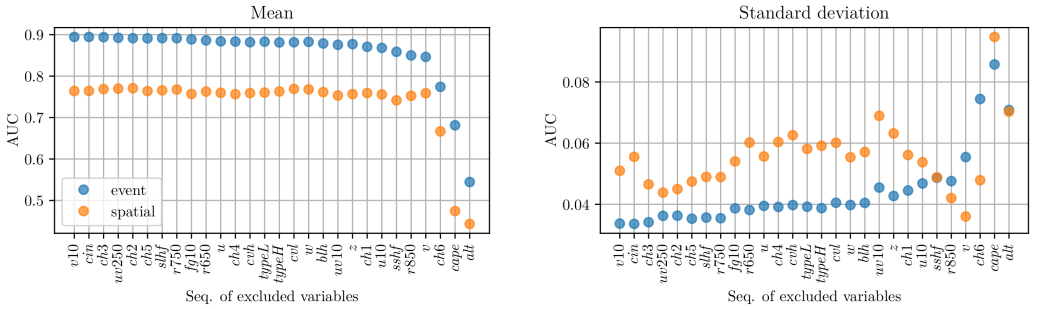

Figure 3: Validation of greedy ICP results. Since ICP should help find variables which have an invariant relationship to the target Y, across different environments, we use spatial-CV to verify that, as causal candidates are excluded from consideration, the spatial generalization performance remains stable, at least for those subsets of variables where the causally necessary conditional independence criterion $Y \perp\!\!\!\perp E|X_S$ is not rejected. We compare to event generalization which does not measure robustness across environments. Mean (left) and standard deviation (right) of out-of-sample AUC for event-CV (blue) and spatial-CV (orange) are shown.

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 seven-variable subset {*u*10, *sshf*, *r*850, *v*, *ch*6, *cape*, *alt*} was accepted, we take a larger 11-variable subset {*blh*, *v*10, *z*, *ch*1, *u*10, *sshf*, *r*850, *v*, *ch*6, *cape*, *alt*} since excluding *blh* from S was the first time the p-value dropped below 0.5. We consider all subsets of size eight 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 A.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 identified a set of candidate causes for pyroCb: surface sensible heat flux (*sshf*), relative humidity at 850 hPa (*r*850), a component of wind at 250 hPa (*v*), 13.3 µm thermal emissions (*ch*6), convective available potential energy (*cape*), and altitude (*alt*). This set is plausible when contrasted with the existing literature. There are other sets of predictors that are causally viable in the sense that they obey the causally necessary conditional independence condition and are similarly minimal (contain no additional non-causal information). As more exhaustive implementations of ICP suggest,

these sets have little overlap with the one identified with greedy-ICP. This is likely because, 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 causes of pyroCb generation. Future work will involve 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 tests needed) and its problems dealing with dependent variables.

Acknowledgements

This work is the result of the 2022 Frontier Development Lab Europe Aerosols challenge. 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

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. *Journal of the Royal Statistical Society Series B*, 78(5):947–1012, 2016. URL `https://EconPapers.repec.org/RePEc:bla:jorssb:v:78:y:2016:i:5:p:947-1012`.

[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] Kenza Tazi, Emiliano Díaz Salas-Porras, Ashwin Braude, Daniel Okoh, Kara Lamb, Duncan Watson-Parris, Paula Harder, and Nis Meinert. Pyrocast: a machine learning pipeline to forecast pyrocumulonimbus (pyrocb) clouds. In *NeurIPS 2022 Workshop-Tackling Climate Change with Machine Learning*, 2022.

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

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

[34] 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]

A Appendix

A.1 PYROCAST Database

The following description of the database is reproduced from the companion PYROCAST paper by the same authors [31].

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 events 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 the three 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 (0.5-2 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), based on their sensitivity to aerosol, vegetation type and temperature at different altitudes, 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 1 km resolution and cropped to 200×200 pixel images.

The cropped geostationary images were then fed into a pyroCb detection algorithm developed by Peterson et al. [18] at the US Naval Research Laboratory (NRL). This algorithm uses relative differences in brightness temperature between the 3.9 µm, 11.2 µm and 13.3 µm geostationary images to label individual spatial pixels as either containing or not containing a pyroCb cloud. 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 the authors 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 18 000 hourly observations. Most importantly, it is science and machine learning ready, allowing for the systematic study of the characteristics and causes of pyroCb.

A.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 4 shows the Hilbert Schmidt Independence Criterion (HSIC) measure of dependence between the 11 variables.

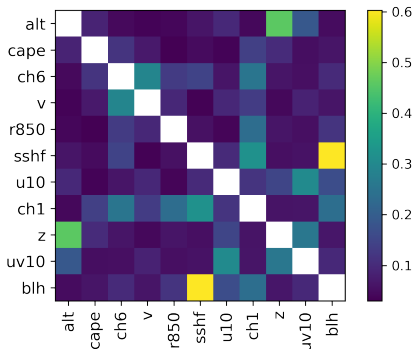

Figure 4: Normalized HSIC between each of the 11 variables considered.

HSIC, introduced in [7], is a good choice for this case because of (1) the non-linear nature of the relationship between the predictor variables and because, (2) each variable is multivariate and includes statistical summaries such as the mean and standard deviation, of the variable across the 200×200 km grid. We use the matrix of dependencies shown in Figure 4 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:$\{alt, z\}$, $\{cape\}$, $\{ch6, ch1\}$, $\{v\}$, $\{r850\}$,$\{sshf, ch1, blh\}$, $\{uv10, z, v10\}$,$\{ch1, ch6, sshf, blh\}$, $\{z, alt, u10, uv10\}$, $\{uv10, u10, z\}$, and $\{blh, sshf, ch1\}$. We repeat the exhaustive ICP considering this time all subsets of size eight 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.

A.3 Additional Figures

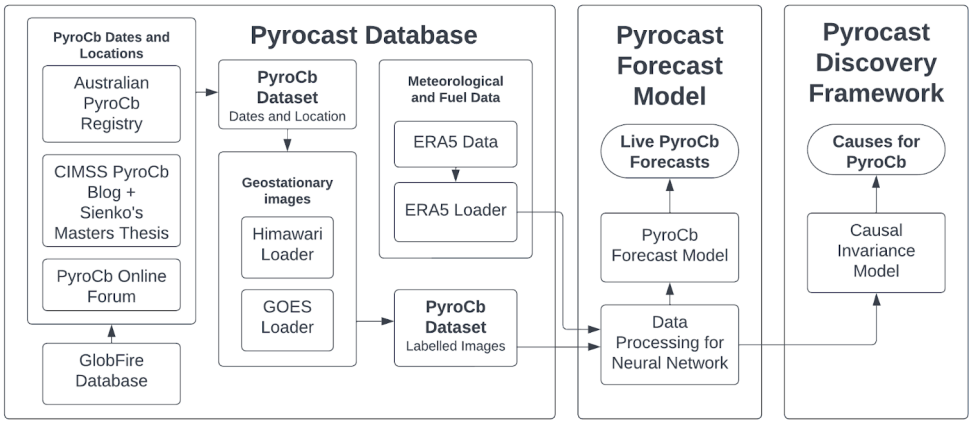

Figure 5: Diagram of PYROCAST pipeline. The pipeline has three components: (1) PYROCAST database, including pyroCb event labels, geostationary imagery in visible and infrared wavelengths, fuel descriptors and meteorological descriptors; (2) PYROCAST forecast model, a 6-hour ahead machine learning forecast model for predicting pyroCb occurrence from intense wildfires, and (3) PYROCAST discovery framework, with tools for understanding the causes of pyroCb events, among which, are those presented in this work. More information on the first two components can be found in [31] and at `https://doi.org/10.56272/fpib2524`.

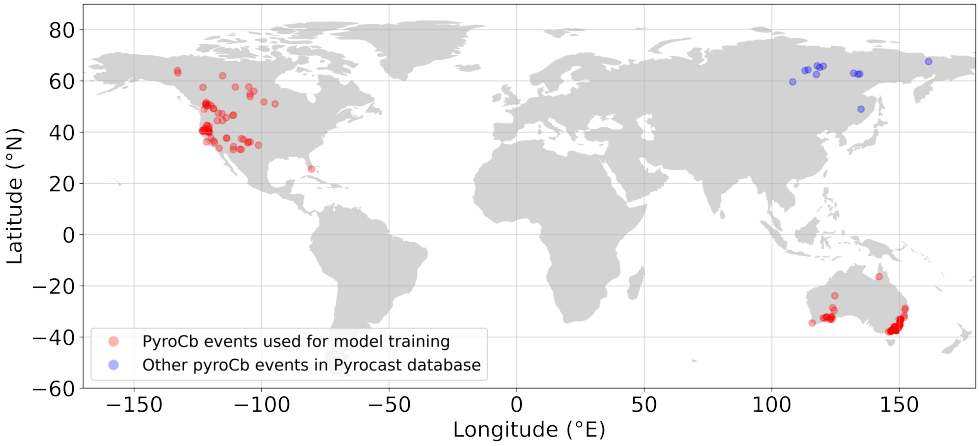

Figure 6: Spatial distribution of pyroCb events in the PYROCAST database.

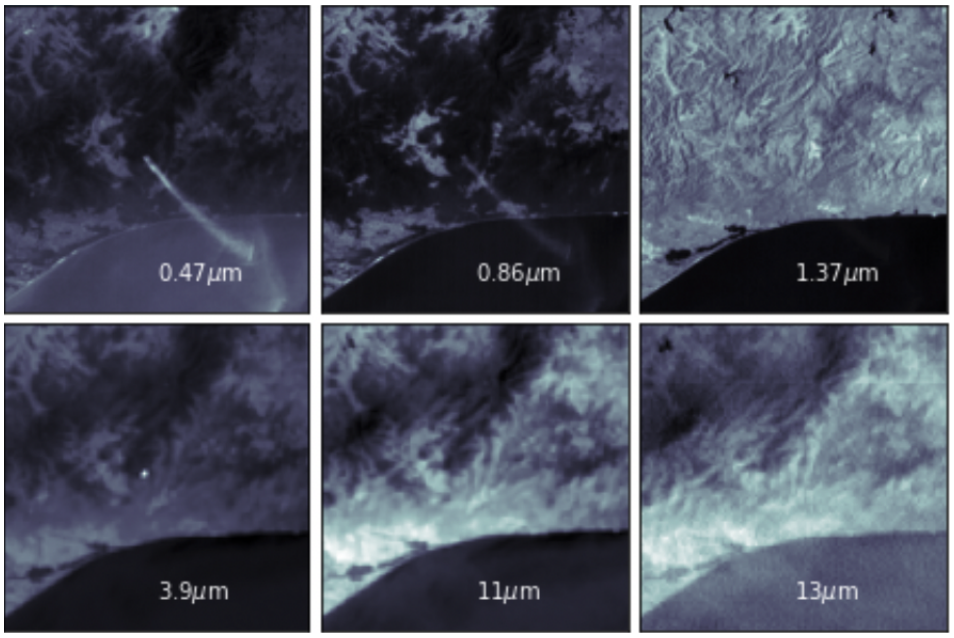

Figure 7: Example of geostationary images over the six wavelength channels of a wildfire event in Timbarra, Australia (January 2019). The centre of the wildfire can clearly be seen as a bright dot in the centre of the $3.9\,\mu$m image, which is most sensitive to high temperatures, while the smoke plume emanating from the fire can be seen most clearly in the $0.47\,\mu$m image, which is most sensitive to small-particle aerosols.

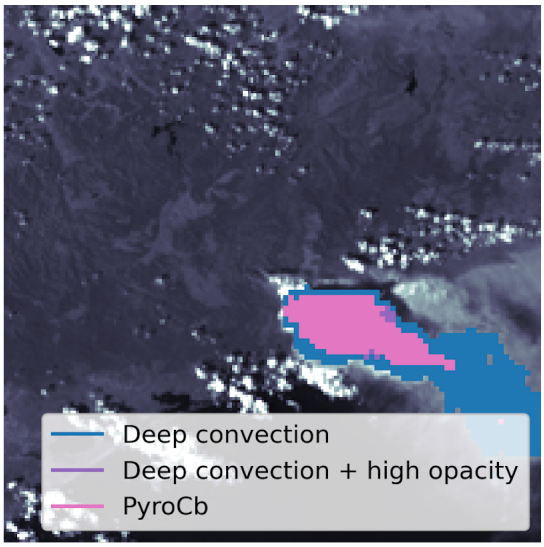

Figure 8: Example of NRL label masks overlaid onto geostationary satellite image (averaged over the 0.47 µm, 0.64 µm and 0.86 µm channels) of a wildfire event in Timbarra, Australia (January 2019).

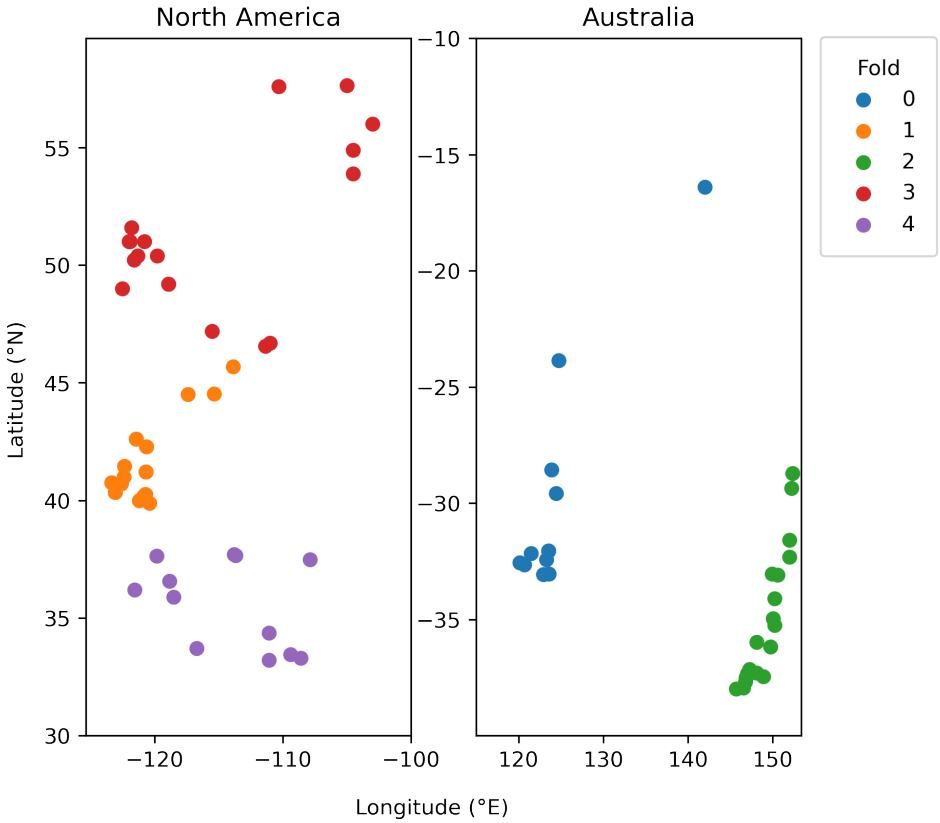

Figure 9: Clusters based on latitude and longitude for spatial cross-validation.