# Wildfire Dataset (1996-2021)

Provided by Marcin: <https://drive.google.com/file/d/10s4cq9_eYp3LtPHNH3NSofq2wvT6Y9uC/view>

Data is obtained from <https://www.alberta.ca/wildfire-maps-and-data> and aggregated into one csv. Some of the column names are changed. Columns of interest with changed names:

* ‘date’ = ‘reported\_date’
* ‘latitude’ = ‘fire\_location\_latitude’
* ‘general\_cause’ = ‘general\_cause\_description’ (not same as ‘true\_cause’)

Feature dictionary: 2006+: <https://open.alberta.ca/dataset/a221e7a0-4f46-4be7-9c5a-e29de9a3447e/resource/1b635b8b-a937-4be4-857e-8aeef77365d2/download/fp-historical-wildfire-data-dictionary-2006-2021.pdf>

1996:2005:

<https://open.alberta.ca/dataset/301bf91d-6db7-4004-8cc9-40ac7e7b42f7/resource/771776a4-33f3-45ae-822d-8e82bc1461c7/download/af-historic-wildfires-1996-2005-data-dictionary.pdf>

**‘discovered\_date’** has nan values which is why ‘reported\_date’ is used over this. Comparison between discovered\_date and reported\_date:

* Q1 = 0 min
* Median = 1 min
* Q3 = 4 min
* Outliers (1.5\*IQR) = 2358
* Min = 0 max = 233 days std = 41 hours

Majority of data falls within several minutes justifying use of ‘reported\_date’ over ‘discovered\_date’ for those entries. Proposed approaches for outliers:

* Leave them as is and use the reported\_date
* Remove them, they comprise of 8% of the dataset
* Look into the possibility of using discovered\_date for these entries
  + Should look closer at the outlier cases and see if this is reasonable
  + Large outliers may be typos

**‘fire\_start\_date’** is present in the Alberta datasets but removed from the cleaned dataset provided by Marcin. This feature represents the known (reported from responsible party) or estimated (using for example storm data) fire start time. Comparison of ‘fire\_start\_date’ to ‘reported\_date’:

* IQR ranges from 00:15:00 to 05:45:00 (hh:mm:ss)
* Number of outliers (based off 1.5\*IQR) is 4981

There are typos in the ‘fire\_start\_date’ that lead to nonsense dates. Perhaps significant manual effort is required to fix this dataset. Proposed approach for handling fire start date:

* Use reported\_date as the fire start date
  + Majority of data lies within 6 hours of estimated fire\_start\_date
  + fire\_start\_date is, itself, an estimate based off weather measurements or human reports
  + Outliers in the difference between reported\_date and fire\_start\_date may be influenced by typos or poor estimates

**‘fire\_type’** gives a class for the predominant fire type at assessment time.

* Ground fire
  + burns ground fuel layer
* Surface fire
  + A fire that burns in the surface fuels and involves less than 10% of the trees torching.
  + Campfires, brush piles and windrows that have not burned into the duff layer would be considered surface fires
* Crown fire
  + A fire that advances through the crown fuel layer and involves more than 10% of the tree crowns

**‘current\_size’** is the final burned area of the fire (unburned islands and bodies of water are removed).

We are interested in predicting the start of a fire so the following information is not needed:

* Assessment information
* Firefighting information

Furthermore, fire events are sparse and having weather information on a more consistent basis (both spatial and temporal) is useful. Therefore, any weather, vegetation, or topography will be obtained from the ERA5 dataset. The **final features** of interest from the **Alberta wildfire dataset** are:

* Latitude and longitude
* Date
* General cause
* Fire type (maybe depending on how the problem is posed)
* Discovered date (depending on how outliers are handled)

# ERA5 Dataset (2000 to 2023)

Available at:

<https://github.com/google-research/arco-era5?tab=readme-ov-file#data-description>

A historical record of global climate data. Information of interest includes weather, vegetation, and topography. Commonly used in literature.

Dataset can be interpolated to a custom grid resolution. Interpolation for continuous variables is bilinear and for discrete is the nearest neighbor. Specifying a coarser resolution is possible but care should be taken to avoid aliasing.

**Spatial resolution** is 0.25 degrees.

**Temporal resolution** is hourly.

Data is taken from full\_37-1h-0p25deg-chunk-1.zarr-v3/

### Features to include:

high\_vegetation\_cover <https://codes.ecmwf.int/grib/param-db/28>

low\_vegetation\_cover <https://codes.ecmwf.int/grib/param-db/27>

leaf\_area\_index\_high\_vegetation <https://codes.ecmwf.int/grib/param-db/67>

<https://codes.ecmwf.int/grib/param-db/66>

type\_of\_high\_vegetation <https://codes.ecmwf.int/grib/param-db/30>

type\_of\_low\_vegetation <https://codes.ecmwf.int/grib/param-db/29>

slope\_of\_sub\_gridscale\_orography <https://codes.ecmwf.int/grib/param-db/163>

### Possible additional features:

skin\_temperature <https://codes.ecmwf.int/grib/param-db/235>

* Modeled, probably don’t include

surface\_solar\_radiation\_downwards <https://codes.ecmwf.int/grib/param-db/169>

surface\_solar\_radiation\_downward\_clear\_sky

* <https://codes.ecmwf.int/grib/param-db/228129>
* This and the above are obtained from models but the description claims it to be quite accurate

standard\_deviation\_of\_orography <https://codes.ecmwf.int/grib/param-db/160>

# FWI ERA5 (1979-2021)

Available at:

<https://drive.google.com/drive/folders/1WWlDZ2lL941TzJAL7xubrJproQuskksA>

This dataset uses ERA5 to compute the FWI (fire weather index) system (<https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi>).

Includes an overwintering procedure (<https://doi.org/10.5194/essd-12-1823-2020>).

**Spatial resolution** is 0.25 degrees.

**Temporal resolution** is daily.

Daily aggregates of weather data are used as inputs, FWI variables are outputs. Both inputs and outputs will be considered features.

### Inputs (aggregated from ERA5):

24hr\_accumulated\_precipitation\_YYYY.nc [24 hour accumulated precipitation (noon to noon)]

24hr\_max\_temperature\_YYYY.nc [maximum daily temperature]

global\_noon\_LST\_2m\_temperature\_YYYY.nc [local standard time 2m temperature]

global\_noon\_LST\_relative\_humidity\_YYYY.nc [local standard time 2m relative humidity]

global\_noon\_LST\_wind\_speed\_YYYY.nc [local standard time 10 m wind speed]

### Outputs (FWI)

build\_up\_index\_YYYY.nc [Build Up Index, BUI]

daily\_severity\_rating\_YYYY.nc [Daily Severity Rating, DSR]

drought\_code\_YYYY.nc [Drought Code, DC]

duff\_moisture\_code\_YYYY.nc [Duff Moisture Code, DMC]

fine\_fuel\_moisture\_code\_YYYY.nc [Fine Fuel Moisture Code, FFMC]

fire\_weather\_index\_YYYY.nc [Fire Weather Index, FWI]

initial\_spread\_index\_YYYY.nc [Initial Spread Index, ISI]