# A Supplementary Material

We present a detailed description of the dataset based on "datasheet for datasets" [Gebru et al., 2018]. We only explain details pertaining to *WildfireDB* (as suggested in [Gebru et al., 2018]); for example, as the dataset does not pertain to people, topics such as re-identification, criminal history, gender composition, etc. are not relevant to the dataset in consideration.

## A.1 Motivation

*WildfireDB* is created to enable data-driven forecasting of wildfires which in turn can aid emergency response. Traditional models that forecast the spread of fires are *physics-based*, that model the effect of each covariate on the spread of fire in closed-form. Data-driven models can accommodate a diverse set of covariates and capture complex non-linear combinations of such features to predict the dynamics of how fires propagate in the real world. The lack of a comprehensive dataset in this regard was the primary motivator behind the creation of *WildfireDB*.

The dataset was created by a collaboration between Vanderbilt University, Stanford University, and University of California, Riverside. The dataset was initially curated for studying how principled decision making under partially observable state spaces can aid response to wildfires. Multiple agencies have helped fund the creation of the dataset (see the following author initials and funding sources). Author SS was funded by Agriculture and Food Research Initiative Competitive Grants no. 2019-67022-29696 and 2020-69012-31914 from the USDA National Institute of Food and Agriculture, AM was funded by the Center of Automotive Research at Stanford (CARS) and National Science Foundation Award Number IIS181495, MW was funded by National Science Foundation Award Number IIS181495, TD was funded by the Department of Management Science and Engineering at Stanford University, and VG was funded by Agriculture and Food Research Initiative Competitive Grant no. 2020-69012-31914 from the USDA National Institute of Food and Agriculture.

## A.2 Composition

Each instance in the dataset consists of information (vegetation type, canopy height, etc.) of a specific discretized spatial area (referred to as the reference cell) observed to be on fire at a given point in time. It also consists of information about one adjacent spatial area (referred to as a neighboring cell) and whether the neighboring cell was on fire at the subsequent time step or not. Each instance also consists of the local weather conditions (precipitation, humidity, etc.) and the relative strength of wind from the reference cell to the neighboring cell. The spatial and temporal granularity of the data are based on the maximum resolution at which fire occurrence data is available; each discrete spatial area is 375m x 375m and the time resolution is that of a day. The "label" of each instance is the FRP (fire radioactive power) of the neighboring cell. We envision that the data can be used to predict the spatial-temporal spread of fire. However, it is entirely possible to use the data to visualize historical fire occurrences, vegetation types, and weather; in such cases, the presence of a label is not required. The total number of data points is 17,820,835.

The data consists of all spatial areas detected to be on fire in continental United States between the years 2012-2017 through VIIRS sensors on the joint NASA/NOAA Suomi National Polar-orbiting Partnership (Suomi NPP) and NOAA-20 satellites. We point out that the direction of wind is missing for about $40\%$ of the data. However, the magnitude of the wind is present for all data points.

## A.3 Collection Process and Author Contributions

Each data source is listed in section 2 of the main paper. The data was collected by authors TD, SS, MW, and AM. TD and AM evaluated multiple sources pertaining to wildfire data and finalized the chosen data source (VIIRS). The data was collected by TD directly from Earth Data, an open source data portal managed by NASA. SS merged the vector and raster data (see section 3). MW collected the weather data from Meteostat [Meteostat, 2020] and matched each instance of observed fire with relevant weather information. MW generated the benchmark results. The visualization platform was created by VG. The manuscript was written by AM and SS. Feedback about the manuscript was provided by MK, AD, AE, and RS. The overall process was supervised by AM. All authors were compensated through their salaries/research stipend at their respective universities.

### A.4 Preprocessing

Preprocessing steps are mentioned in the main body of the paper.

### A.5 Uses

The primary purpose of the dataset is to forecast the spread of fires as a function of relevant covariates. The dataset has already been used to optimize response to wildfires [Diao et al., 2020]. We also tested the use of the data in Vanderbilt University's graduate level course on Big Data (Topics of Big Data, CS:4266/5266). The dataset can also be used for visualizing the spread of historical fires and studying the occurrence of wildfires themselves.

### A.6 Distribution

We have released the data as open-source, which means that the data can be distributed freely. We require that the paper accompanying this release be cited when the dataset is used. The visual interface associated with the dataset can be accessed through `raptor.cs.ucr.edu/wildfiredb`. We maintain an up-to-date description about the data at `https://wildfire-modeling.github.io/`. The data itself is hosted at `https://doi.org/10.5281/zenodo.5636429`.

### A.7 Maintenance

The dataset is currently being maintained by the authors SS, MW, and AM. The contact information of the authors are provided in the dataset webpage and on the main body of the paper. Any issues and/or inconsistencies found with the data should be reported to SS, MW, and AM. We welcome other contributors who want to augment the data and request them to contact author AM. An older version of the dataset was released at the Neurips AI for Earth Sciences Workshop 2020. We will currently continue to maintain the earlier release although the current version subsumes the earlier data.