MOTIVATION.

[question1] For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
[answer1] 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.

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] 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).

[question3] Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
[answer3] 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.

[question4] Any other comments?
[answer4] No

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
[answer5] 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.

[question6] How many instances are there in total (of each type, if appropriate)?
[answer6] The total number of data points is 17,820,835.

[question7] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

[answer7]

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] 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.

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] 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.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] 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

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11]

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12]

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13]

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions

of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
[answer14]
[question15] Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
[answer15]
[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
[answer16]
If the dataset does not relate to people, you may skip the remaining questions in this section.
[SkipA] NO
[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
[answer17]
[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.
[answer18]
[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
[answer19]
[question20] Any other comments?
[answer20] No.

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

[answer21] Each data source is listed in section 2 of the main paper.

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

[answer22] 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.

[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

[answer23]

[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24] 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. All authors were compensated through their salaries/research stipend at their respective universities

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer25]

[question26] Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

[answer26]

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipB] NO

[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

[answer27]

[question28] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

[answer28]

[question29] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
[answer29]
[question30] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
[answer30]
[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer31]
[question32] Any other comments?
[answer32] No.

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
[answer33] Preprocessing steps are mentioned in the main body of the paper.
[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
[answer34]
[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
[answer35]
[question36] Any other comments?
[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.
[answer37] 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).

[question38] Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

[answer38]

[question39] What (other) tasks could the dataset be used for?

[answer39] The primary purpose of the dataset is to forecast the spread of fires as a function of relevant covariates. The dataset can also be used for visualizing the spread of historical fires and studying the occurrence of wildfires themselves.

[question40] Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

[answer40]

[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.

[answer41]

[question42] Any other comments?

[answer42]

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

[answer43] We have released the data as open-source, which means that the data can be distributed freely.

[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

[answer44] 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.

[question45] When will the dataset be distributed?

[answer45]

[question46] Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

[answer46] We

require that the paper accompanying this release be cited when the dataset is used.

[question47] Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

[answer47]

[question48] Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

[answer48]

[question49] Any other comments?

[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50] The dataset is currently being maintained by the authors SS, MW, and AM.

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51] The contact information of the authors are provided in the dataset webpage and on the main body of the paper.

[question52] Is there an erratum? If so, please provide a link or other access point.

[answer52] Any issues and/or inconsistencies found with the data should be reported to SS, MW, and AM.

[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

[answer53] We will currently continue to maintain the earlier release although the current version subsumes the earlier data.

[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

[answer54]

[question55] Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

[answer55] An older version of the dataset was released at the Neurips AI for Earth Sciences Workshop 2020.

[question56] If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

[answer56] We welcome
other contributors who want to augment the data and request them to contact author AM.

[question57] Any other comments?

[answer57] No.