PNNL-34891

An Inventory of AI-ready Benchmark Data for US Fires, Heatwaves, and Droughts

September 2023

X Lin
Z Hou

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-34891

An Inventory of AI-ready Benchmark Data for US Fires, Heatwaves, and Droughts

September 2023

X Lin
Z Hou

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Extreme weather events, including fires, heatwaves, and droughts, have significant impacts on earth, environmental, and energy systems. Mechanistic and predictive understanding, as well as probabilistic risk assessment of these extreme weather events, are crucial for detecting, planning for, and responding to these extremes. Records of extreme weather events provide an important data source for understanding present and future extremes, but the existing data needs preprocessing before it can be used for analysis. Moreover, there are many nonstandard metrics defining the levels of severity or impacts of extremes. In this study, we have compiled a comprehensive benchmark data inventory of extreme weather events, including fires, heatwaves, and droughts. The dataset covers the period from 2001 to 2020 with a daily temporal resolution and a spatial resolution of 0.5°×0.5° (~55km×55km) over the continental United States (CONUS), and a spatial resolution of 1km × 1km over the Pacific Northwest (PNW) region, together with the co-located and relevant meteorological variables. By exploring and summarizing the spatial and temporal patterns of these extremes in various forms of marginal, conditional, and joint probability distributions, we gain a better understanding of the characteristics of climate extremes. The resulting AI/ML-ready data products can be readily applied to ML-based research, thereby fostering and encouraging AI/ML research in the field of extreme weather. This study can contribute significantly to the advancement of extreme weather research, aiding researchers, policymakers, and practitioners in developing improved preparedness and response strategies to protect communities and ecosystems from the adverse impacts of extreme weather events.

Acknowledgments

This research was supported by Earth and Biological Sciences Directorate (EBSD)'s Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). We'd like to acknowledge the valuable inputs from our colleagues: Aowabin Rahman, Sohom Datta, and Sing-Chun Wang and Jiwen Fan. PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

AI	Artificial Intelligence
AR	Arkansas
BA	Burned Area
BC+OC	Black Carbon and Organic Carbon Aerosols
CA	California
CONUS	The Continental United States
CPC	Climate Prediction Center
Daymet	Daily Surface Weather Data for North America
DOE	U.S. Department of Energy
FAIR	Finable, Accessible, Interoperable, Reusable
FHS	Fire Hot Spot
gridMET	Gridded Surface Meteorological
HI	Heat Index
HW	Heatwave
ID	Idaho
LA	Louisiana
LHF	Latent Heat Flux
maxFRP	Maximum Fire Radiative Power
MCD64A1	MODIS Thermal Anomalies and Fire Daily
MERRA-2	Modern-Era Retrospective Analysis for Research and Applications, Version 2
ML	Machine Learning
MODIS	Moderate Resolution Imaging Spectroradiometer
MOD14A1	MODIS Thermal Anomalies and Fire Daily
NARR	North American Regional Reanalysis
NLDAS-2	North American Land Data Assimilation System Phase 2
OR	Oregon
PDSI	Palmer Drought Severity Index
PNNL	Pacific Northwest National Laboratory
PNW	Pacific Northwest
PRISM	Parameter-elevation Regressions on Independent Slopes Model
Qv	Specific Humidity
WA	Washington
RH	Relative Humidity
SH	Sensible Heat Flux
SM	Soil Moisture
SPEI	Standardized Precipitation Evapotranspiration

SPI	Standardized Precipitation Index
T	Temperature
Tmax	Maximum Temperature
Tmean	Mean Temperature
Tmin	Minimum Temperature

Contents

Abstract....................ii
Acknowledgments....................iii
Acronyms and Abbreviations....................iv
1.0 Introduction....................1
2.0 Data Source Identification and Acquisition....................3
2.1 Heatwave data....................3
2.2 Fire data....................4
2.3 Drought data....................4
2.4 Relevant meteorological variables....................5
3.0 Data Compilation and Labeling of Extreme Events....................6
4.0 Extremes' Behavioral Characteristics Data through Exploratory Data Analysis....................7
4.1 Temporal and spatial distribution of HWs, fires and droughts....................7
4.2 The cross-dependence of compound extremes....................12
5.0 Data Availability and Usage Notes....................14
6.0 Summary and Next Steps....................15
7.0 References....................17

Figures

Figure 1. Schematic overview of the generation process for the AI-ready data inventory of extreme events....................2
Figure 2. The distribution of total HW days from 2001 to 2020 over CONUS for different HIs i.e., (a) HI02, (b) HI04, (c) HI05, (d) HI06, (e) HI09, (f) HI10. The HWs are defined using the 0.5°×0.5° CPC Global Unified daily temperature anomaly data....................7
Figure 3. The distribution of total HW days from 2001 to 2020 over PNW for HWs defined using different HIs i.e., (a) HI02, (b) HI04, (c) HI05, (d) HI06, (e) HI09, (f) HI10. The HWs are defined using the 1km × 1km Daymet daily temperature data....................8
Figure 4. The distribution of (a) the total number of FHS and (b) the maxFRP for each 0.5°×0.5° over CONUS from 2001 to 2020; (c) the total number of fire occurrence and (d)the maxFRP for each 1km×1 km over PNW from 2001 to 2020....................9
Figure 5. The time series of (a) gridded annual mean of HW days, (b)total number of FHS and Burned Area ,and (c) gridded annual mean drought days for extremes over CONUS(0.5°×0.5°); The time series of (d) gridded annual mean of HW days, (e)total number of FHS and Burned Area ,and (f) gridded annual mean drought days for extremes over PNW(1km×1 km)....................10
Figure 6. The total HW days for months from May to October; (b) stacked bar plots showing total FHS count in HW days and non-HW days(middle)....................11

Figure 7. (a)The distribution of the number of FHS for fires occurred on HW days(red) versus non-HW days(green); (b) the distribution of the number of FHS for fires occurred with(red) or without(green) long HWs (duration greater than 8 days). 11
Figure 8. The time series of daily total HW days, total number of FHS, and total drought days from May to October in 2018 over CONUS (a, b, c) and PNW (d, e, f). 12
Figure 9. (a)Marginal probability of HW labeled using index HI02;(b) Marginal probability for fires; (c) conditional probability of the occurrence of fires given that HWs (HI02) have occurred; (d) conditional probability of the occurrence of HWs (HI02) given that fires have occurred;(e) the probability of the co-occurrence of HWs (HI02) and fires. 13

Tables

Table 1. Summary of data sources used in this study. 3
Table 2. Heat indices (HIs) used to define HW events. 4

1.0 Introduction

Extreme weather events such as fires, heatwaves(HWs), and droughts cause significant socioeconomic and environmental damage around the worldwide. Understandings the mechanisms of these events and their potential drivers is crucial for detecting, planning, and responding to such challenges as well as mitigating their impacts. Historical records of extreme weather and their associated meteorological parameters offer invaluable data for identifying trends of the extremes, assessing, and managing risks associated with these events, deciphering their primary determinants, and projecting potential alterations under climate change. Previous studies show that machine learning (ML) algorithms are increasingly employed in weather prediction research, extreme event analysis, meteorological patterns extractions using historical weather data and simulations. However, one of the challenges facing scientists who conduct ML-based extreme weather analysis is that data from multiple sources are rarely in a form suitable for direction application. These datasets typically exhibit varying fidelity, spatiotemporal resolution and coverage, and therefore need preprocessing before they can be used for analysis. High-quality, AI-ready datasets enable scientists and researchers to apply data-driven ML/AI methods to extreme weather research without expending excessive time on data collection and compilation. Such datasets not only streamline the data collection and compilation process but also empower scientists and researchers to apply cutting-edge ML/AI techniques to their studies of extreme weather.

The objective of this study is to develop a benchmark data inventory of US extreme weather events (i.e., fires, HWs, and droughts) to support advanced ML/AI research in extreme weather. The data inventory is based on the intensive compilation of multi-fidelity data from various sources, resulting an AI-ready FAIR(i.e., finable, accessible, interoperable, reusable) dataset that users can easily query, search, and extract attributes of interest for advanced ML development. The data inventory includes two data products with a daily temporal resolution covering the period from 2001 to 2020: (1) fires, heatwaves, and droughts with a spatial resolution of 0.5°×0.5° (~55km×55km) over the continental United States (CONUS); (2) fires, heatwaves, and droughts with a spatial resolution of 1km × 1km over the Pacific Northwest (PNW) region. The choice of different spatial resolution for CONUS and PNW is because the trade-off between spatial resolution and coverage. Coarser spatial resolution allows for the efficient coverage of large geographic area (CONUS) and datasets are more manageable in terms of computing power and storage capacity. While for relatively small region like PNW, finer spatial resolution provides more detail and precision, and the datasets are also manageable. Both data products include the co-located and relevant meteorological variables. Evaluating extreme events across areas with significant differences on a daily scale can be challenging, especially when accounting for the variability of local extremes. Consequently, we incorporate multiple extreme event characteristics (e.g., coverage, timing, intensity, frequency) into the datasets and label these events using threshold-based methods. This approach help establish standards for defining extreme events from various perspectives. Furthermore, we use the compiled datasets to investigate spatial and temporal patterns of compound extremes, specifically fires, HWs, and droughts across PNW and CONUS. Exploratory data analysis is employed to identify co-occurrence and describe the cross-dependence conditional distribution of the compound extremes. The overview of the generation process for the AI-ready data inventory of extreme events is shown in Figure 1.

In summary, our work consists of the following specific steps:

(1) Data compilation - Integrating fires, HWs, and drought data and their co-located meteorological parameters from multiple sources with different fidelity, spatial and temporal resolution/coverage to produce more consistent, informative, and AI-ready datasets surpassing those offered by any individual data source.

(2) Labeling or characterizing extreme events at daily resolution and in a long-term (i.e., 20 years) dataset, as well as defining extreme events that cover different areas using data with different spatial resolutions (i.e., a spatial resolution of 1km × 1km for PNW, and a spatial resolution of 0.5°×0.5° for CONUS).

(3) Applying exploratory data analysis to visualize and summarize the spatiotemporal distributions of individual and compound extremes in PNW and CONUS.

(4) Gathering user feedbacks from potential stakeholders to enhance the quality and utility of data packages or to inform future improvements and maintenance.

(5) Ensuring accessibility and usability of the AI-Ready data inventory on the PNNL DataHub platform.

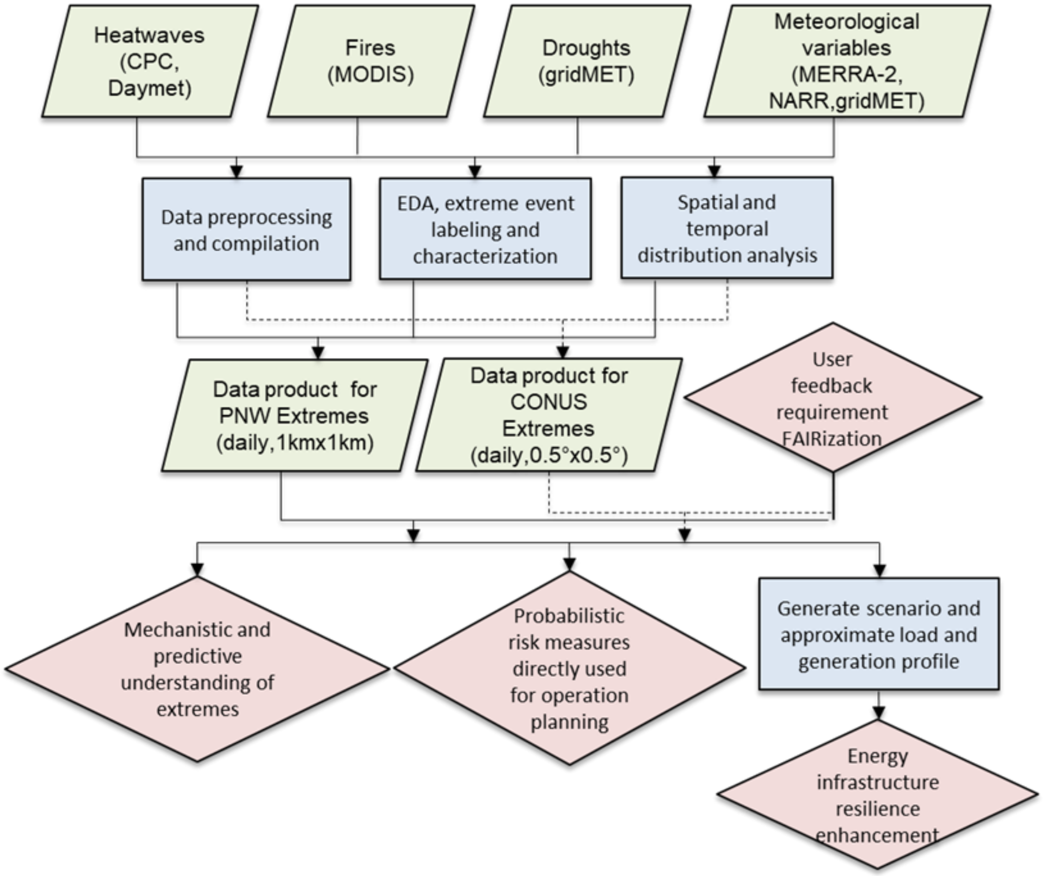
Heatwaves (CPC, Daymet)
Fires (MODIS)
Droughts (gridMET)
Meteorological variables (MERRA-2, NARR,gridMET)
Data preprocessing and compilation
EDA, extreme event labeling and characterization
Spatial and temporal distribution analysis
Data product for PNW Extremes (daily,1kmx1km)
Data product for CONUS Extremes (daily,0.5°x0.5°)
User feedback requirement FAIRization
Mechanistic and predictive understanding of extremes
Probabilistic risk measures directly used for operation planning
Generate scenario and approximate load and generation profile
Energy infrastructure resilience enhancement

Figure 1. Schematic overview of the generation process for the AI-ready data inventory of extreme events.

2.0 Data Source Identification and Acquisition

The raw data concerning extreme events such as fires, HWs, and droughts primarily originate from the sources outlined in Table 1, which include the temperature anomaly data from CPC and Dayment; fire related data from two MODIS products; drought indices from gridMET; and meteorological data from gridMET, NARR and MERRA-2. Detailed descriptions about the data sources are elaborated in the subsequent subsections.

Table 1. Summary of data sources used in this study.

○	Description	Variables	Study Region	Original resolution	Targeted resolution
CPC	Global Unified temperature data (daily)	Tmax, Tmin, Tmean	CONUS	0.5°×0.5°	0.5°×0.5°
Daymet	Daily Surface Weather Data(daily)	Tmax, Tmin, Tmean	PNW	1km x 1km	1km x1km
MODIS MOD14A1	Thermal Anomalies and Fire Detection data (daily)	FHS, MaxFRP	CONUS/PNW	500m x 500m	0.5°×0.5° 1km x1km
MODIS MCD64A1	Fire Burned Area Product (monthly)	Burned Area (BA)	CONUS	1km x 1km	0.5°×0.5°
gridMET	Gridded Surface Meteorological data (daily)	Drought indices (i.e., SPI, SPEI, PDSI), RH, QV, etc.	CONUS/PNW	4km x 4km	0.5°×0.5° 1km x1km
NARR	North American Regional Reanalysis data (3-hourly)	RH, SM, QV, U-wind,V-wind etc.	CONUS	~32km	0.5°×0.5°
MERRA-2	Modern-Era Retrospective Analysis for Research and Applications (3-hourly)	RH,BC+OC etc.	CONUS	0.5°×0.625°	0.5°×0.5°

2.1 Heatwave data

HWs over CONUS are defined using the temperature data sourced from the Climate Prediction Center (CPC) Global Unified daily gridded temperature data provided by NOAA Physical Sciences Laboratory. This daily data product is obtained from the global telecommunications system (GTS) data and is gridded using the Shepard Algorithm. The spatial resolution of this data is 0.5°×0.5°(approximately 55km×55km).

To define HWs in the PNW region, temperature data with a 1km × 1km spatial resolution from the Daily Surface Weather Data for North America (Daymet) are employed. Daymet is a research product of the Environmental Sciences Division at Oak Ridge National Laboratory, Oak Ridge. Daymet provides long-term, continuous, gridded estimates of daily weather and climatology variables by interpolating and extrapolating ground-based observations through statistical modeling techniques. The maximum temperature, duration, start and end date of all HWs are defined for each grid (i.e., 0.5°×0.5°, or 1km × 1km) using selective heat indices defined with daily mean, maximum and minimum of temperature, its historical percentiles, and duration of high temperature days, as shown in Table 2.

Table 2. Heat indices (HIs) used to define HW events.

Heat indices (HI)	*Temp Metric*	*Threshold*	*Duration*
HI02	Daily Mean	> 95th percentile	3+ consecutive days
HI04	Daily Mean	> 99th percentile	3+ consecutive days
HI05	Daily Max	> 95th percentile;	3+ consecutive days
HI06	Daily Max	T1 > 97.5th percentile T2 > 81st percentile	Everyday >T2; 3+ consecutive days >T1 Avg T max > T1 for the whole period
HI09	Daily Min	> 95th percentile	3+ consecutive days
HI10	Daily Min	T1> 97.5th percentile T2> 81st percentile	Everyday >T2; 3+ consecutive days >T1 Avg T max > T1 for the whole period

2.2 Fire data

In this study, two MODIS products are utilized: the MODIS Thermal Anomalies and Fire Daily (MOD14A1) Version 6, and MODIS Burned Area Product (MCD64A1) Version 6. MOD14A1 datasets encompass all fire-related thermal anomaly detection, including those caused by wildfires, agricultural field burning, prescribed fires, etc. These datasets are generated at approximately 1-kilometer (km) spatial resolution and daily temporal resolution. The variables within these datasets include the fire mask, pixel quality indicators, maximum fire radiative power (MaxFRP), and the position of the fire pixel within the scan.

Individual 1-km pixels (grids) are assigned to one of nine fire mask pixel classes, which indicate the different confidence levels of fire occurrence. A value of 7 indicates a low confidence detection, a middle value of 8, and a value of 9 signifies a high confidence detection. In this study, we only use the fire pixels (grids) with the two highest confidence levels of fire occurrence to summarize the daily fire features. The fire features over CONUS include the maximum MaxFRP, and the total number of fire hotspots (FHS) within each 0.5°×0.5° (~55km×55km) grid. Here, a FHS is a 1-km pixel with confidence levels of fire occurrence not less than 8. To estimate the daily fire burned area (BA) for CONUS, an event-delineation algorithm is utilized to derive fire events from the MODIS MCD64A1 burned area product by identifying the optimal spatial-temporal aggregation of burned pixels. For the PNW region, the summarized fire features include the maximum MaxFRP, and fire occurrence with confidence levels not less than 8 for each 1km×1 km grid.

2.3 Drought data

The drought indices used in this study are obtained from the Gridded Surface Meteorological (gridMET) dataset, which has a 4-km spatial resolution and 5-day temporal resolution. These drought indices include the standardized precipitation index (SPI) , the standardized precipitation evapotranspiration index(SPEI) , the Palmer Drought Severity Index (PDSI) , among others. The SPI and SPEI indices are provided at different time scales, corresponding to the time aggregation of precipitation, reference evapotranspiration, and precipitation minus reference evapotranspiration, respectively. The available time scales are 14-day, 30-day, 90-day, 180-day, 270-day, 1 year, 2 years and 5 years.

In this study, we include the SPI indices (i.e., SPI-14d, SPI-30d, SPI-90d,) and SPEI indices (i.e., SPEI-14d, SPEI-30d, SPEI-90d), aggregated at time scales of 14 days, 30 days, and 90 days. The sub-monthly PDSI is calculated using a modified version of the Palmer formula which uses reference evapotranspiration and precipitation from gridMET, and a static soil water holding capacity layer (top 1500mm) from STATSGO. Modifications to the coefficients of the original Palmer formula are applied to calculate PDSI. The baseline period for PDSI calculations is 1979-2018.

2.4 Relevant meteorological variables

An extensive compilation of meteorological variables that spatially and/or temporally coincide with HWs, fires, and droughts is created to enable and facilitate the development of AI-driven insights into the underlying physical mechanisms and possible predictive models of these extremes. These meteorological variables include accumulated precipitation, soil moisture (SM), latent heat flux (LHF), sensible heat flux (SH), and specific humidity (Qv) at 250 and 850 hPa, relative humidity (RH), U-wind and V-wind at 250, 500, and 850 hPa, and carbonaceous aerosols, namely black carbon and organic carbon aerosols (BC+OC), among others. The meteorological parameters mentioned above are sourced from the North American Regional Reanalysis (NARR) , the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) , and gridMET.

NARR data is available every 3 hours at a 32-km horizontal grid spacing and comprises 45 vertical layers. Data from MERRA-2 is available every 3 hours at an approximate spatial resolution of 0.5° × 0.625° and includes 72 hybrid-eta levels. We use precipitation, SM, LHF, SH from NARR, and RH, QV, BC+OC, U-wind, and V-wind at 250, 500, and 850 hPa from MERRA-2.

The daily meteorological variables from gridMET are available at a spatial resolution of 4 km. Variables from gridMET include RH, QV, precipitation, vapor pressure deficit, wind speed, and direction at 10 m. gridMET blends spatial attributes of gridded climate data from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) with desirable temporal attributes (and additional variables) from regional reanalysis - the North American Land Data Assimilation System Phase 2 (NLDAS-2) using climatically aided interpolation. The resulting gridMET is a spatially and temporally complete gridded dataset of surface meteorological variables.

3.0 Data Compilation and Labeling of Extreme Events

The data compilation steps for this study involve several key stages, including defining HW events, labeling temperature data, summarizing fire-related features, rescaling fire features, drought indices, as well as the co-located meteorological variables, to the same spatial and temporal resolution. The outputs include two data products, both with daily temporal resolution. The data product for CONUS has a spatial resolution of 0.5°×0.5°, while the data product for PNW has a spatial resolution of 1km×1km. The following paragraphs describe these steps in detail.

First, we utilize the CPC Global Unified daily temperature anomaly data and Daymet daily temperature anomaly data to summarize HW events over CONUS and PNW based on the Heat indices (HIs) defined in Table 2. For each HI, the duration, start dates and end dates, maximum temperature for each grid (i.e., 0.5°×0.5°, 1km × 1km) are summarized. According to the summarized HW characteristics, we then label the daily temperature data as HW days or not (0/1), with 1 indicating a HW day and 0 meaning a non-HW day for a specific grid. Since there are six HIs in Table 2, we have six different HW labels, one for each HI. This labeling process helps identify and characterize the occurrence of HWs across the specific study area (i.e., CONUS, PNW).

The fire-related features from the MODIS thermal anomaly data product, MOD14A1, have daily temporal and 1km×1km spatial resolution. For PNW, the fire-related features are the daily fire occurrence and fire intensity, i.e., MaxFRP (fire radiative power), for each 1km×1km spatial grid. Only fire pixels with the two highest confidence level(e.g., level 8 and level 9) of fire occurrence are considered. The nearest neighbor method is used to match fire features to 1km×1km spatial grid. For CONUS, the ~1km MODIS fire features are summarized by treating each ~1km fire pixel as an active FHS and calculate the number of FHS within each 0.5°×0.5° grid. When calculating the number of FHS, only the fire pixels with the two highest confidence level (e.g., level 8 and level 9) of fire occurrence are considered. For each 0.5°×0.5° grid, we also summarize the maximum of MaxFRP to represent the fire intensity. The daily fire burned area within each 0.5°×0.5° grid is estimated from the MODIS MCD64A1 burned area product by using an event-delineation algorithm to aggregate burned area pixels into distinct fire events. These steps help characterize fire activity in our study area (i.e., CONUS).

The drought indices from gridMET have 4-km spatial resolution and daily temporal resolution. For the seven drought indices (i.e., PDSI, SPI-14d, SPI-30d, SPI-90d, SPEI-14d, SPEI-30d, SPEI-90d), we rescale their spatial resolution from 4km×4km to 0.5°×0.5° (CONUS) and 1km×1km (PNW), using bilinear interpolation and nearest neighbor method, respectively. For the meteorological variables from the NARR and MERRA-2, the daily mean, maximum and minimum values within each 0.5°×0.5° grid are summarized to match the resolution of the CPC Global Unified temperature data for CONUS. Meteorological variables from gridMET are rescaled to 1km×1km spatial resolution using the nearest neighbor method for PNW.

Finally, the HW labels, temperature data, the summarized fire-related features (e.g., maximum of MaxFRP, total number of FHS), the drought indices, as well as the meteorological variables are combined to create comprehensive data products for both PNW and CONUS. These data products will facilitate a more in-depth analysis of the relationships between heatwaves, fires, and droughts.

4.0 Extremes' Behavioral Characteristics Data through Exploratory Data Analysis

To enhance the efficiency of AI/ML development and implementation, we have generated spatial and temporal joint distributions and behaviors of compound extremes over the study period and regions, which can be directly employed for risk estimation and other predictive models. The co-occurrence of heatwaves, wildfires, and droughts during 2001 to 2020 has also been closely examined. Additionally, the study delved into the differences in fire patterns and behaviors on days with heatwaves (HW) compared to non-heatwave (non-HW) days, using statistical tests to assess hypotheses regarding the average number of wildfire occurrences. To gain deeper insights, particular emphasis was placed on a historical summer period, specifically May to October in 2018.

4.1 Temporal and spatial distribution of HWs, fires and droughts

The spatial distributions of the total HW days over the period from 2001 and 2020 are shown in Figure 2 (CONUS) and Figure 3 (PNW). While there are fewer HW days for HI04 and more HW days for HI10, heat indices HI02, HI05,HI06 and IH09 have a comparable number of HW days over the 20 years period. The overall patterns of these HIs are quite similar, with more HW days in the western and southern regions of the US. The distribution of HW days in PNW defined using the Daymet temperature data (1km × 1km) (Fingure3) shows similar patterns with those defined using the CPC Global Unified temperature data (0.5°×0.5°) (Figure2), with more HWs in the central and southeast of PNW.

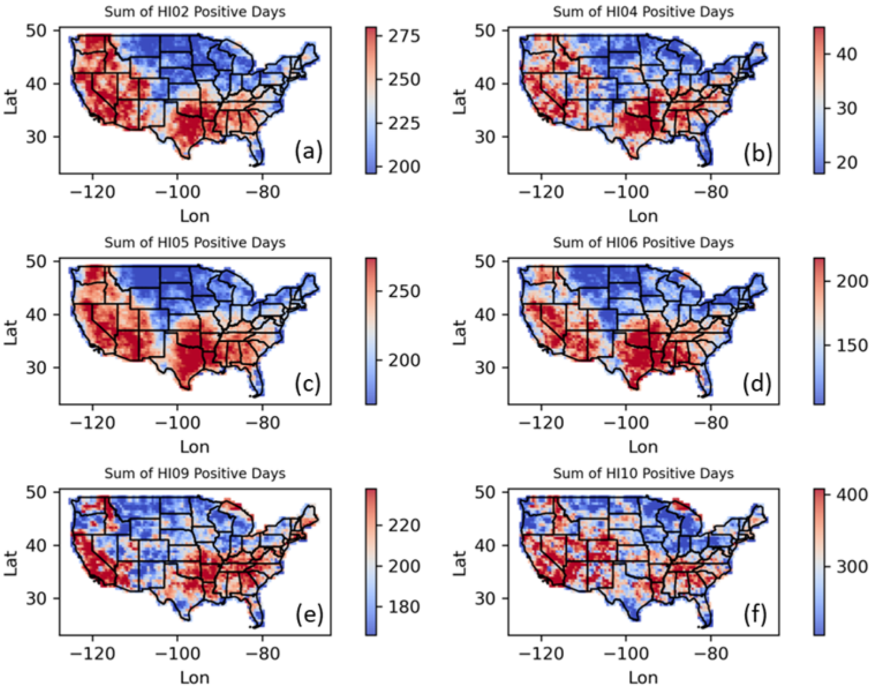
Sum of HI02 Positive Days
Sum of HI04 Positive Days
Sum of HI05 Positive Days
Sum of HI06 Positive Days
Sum of HI09 Positive Days
Sum of HI10 Positive Days
Lat
Lon
50
40
30
−120
−100
−80
275
250
225
200
40
30
20
250
200
200
150
220
200
180
400
300
(a)
(b)
(c)
(d)
(e)
(f)

Figure 2. The distribution of total HW days from 2001 to 2020 over CONUS for different HIs i.e., (a) HI02, (b) HI04, (c) HI05, (d) HI06, (e) HI09, (f) HI10. The HWs are defined using the 0.5°×0.5° CPC Global Unified daily temperature anomaly data.

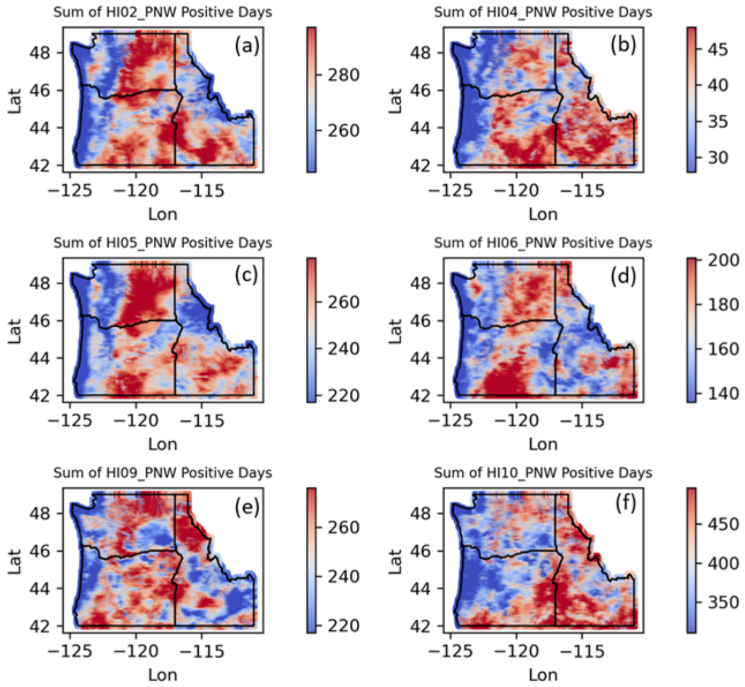
Sum of HI02_PNW Positive Days
(a)
Sum of HI04_PNW Positive Days
(b)
Sum of HI05_PNW Positive Days
(c)
Sum of HI06_PNW Positive Days
(d)
Sum of HI09_PNW Positive Days
(e)
Sum of HI10_PNW Positive Days
(f)
Lat
Lon
48
46
44
42
−125
−120
−115
280
260
45
40
35
30
240
220
200
180
160
140
450
400
350

Figure 3. The distribution of total HW days from 2001 to 2020 over PNW for HWs defined using different HIs i.e., (a) HI02, (b) HI04, (c) HI05, (d) HI06, (e) HI09, (f) HI10. The HWs are defined using the 1km × 1km Daymet daily temperature data.

The spatial distribution of fire characteristics, including the maximum of maxFRP, total number of FHS(i.e., 0.5°×0.5° for CONUS), and fire occurrence (1km×1 km for PNW) for each grid over the period from 2001 -2020, are summarized. Figure 4 display the distribution of total number of FHS in CONUS (Figure 4a) and fire occurrence in PNW (Figure 4c) . Overall, there are more fires occurred in the western US and southeast US(Figure 4a) in terms of frequency. Fires in the western US, especially in CA, OR, WA and ID, have higher intensity (maxFRP) (Figure 4b). Fires occurred in other regions of the US are notably smaller in spatial extent, frequency, and intensity. For PNW, the fire characteristics with spatial resolution of 0.5°×0.5°(Figure 4a) and 1km×1 km(Figure 4c) both show that more fires occur in central WA, southwest OR and north ID.

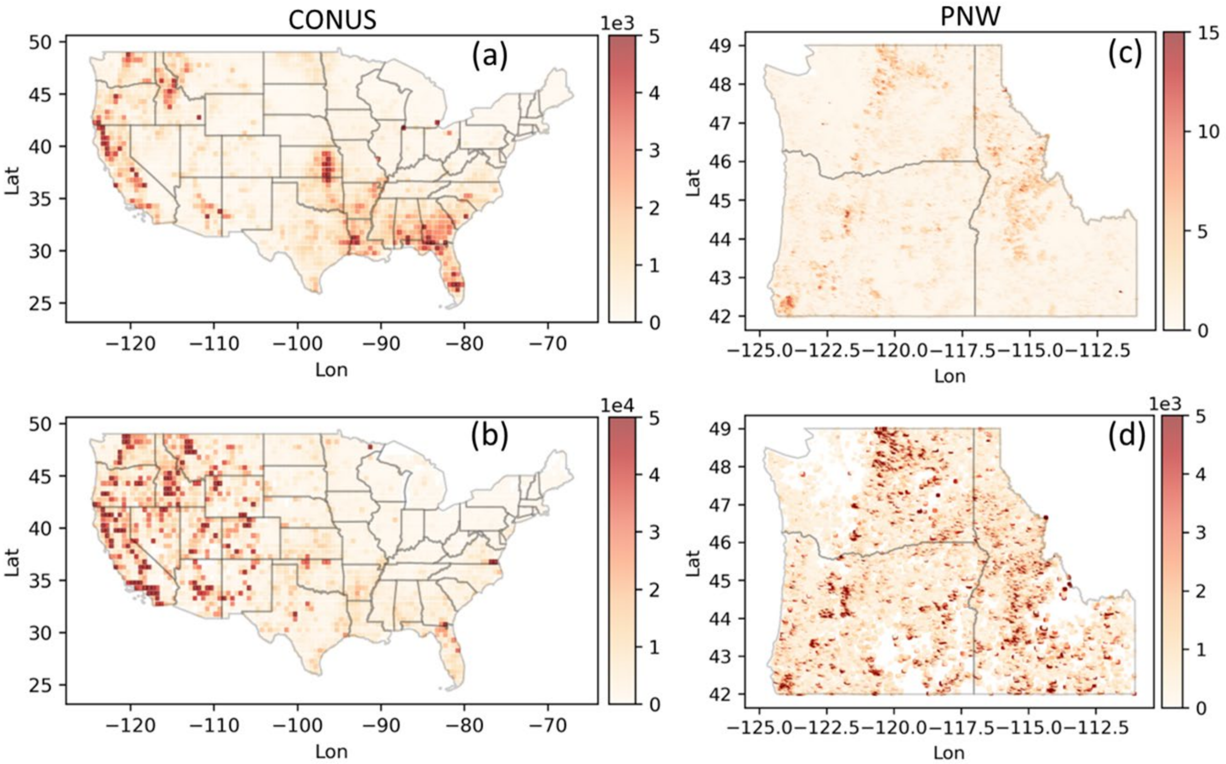

Figure 4. The distribution of (a) the total number of FHS and (b) the maxFRP for each 0.5°×0.5° over CONUS from 2001 to 2020; (c) the total number of fire occurrence and (d)the maxFRP for each 1km×1 km over PNW from 2001 to 2020.

In addition to checking the spatial distribution of HWs and Fires, we also explored their temporal variability and seasonal patterns over CONUS and PNW, and the co-occurrence of these events during 2001 to 2020. Figure 5a shows the time series of annual average HW days for different HIs. The gridded average HW days for a specific year are obtained by first calculating the annual total HW days for each grid, and then calculating their mean values over all grids. The gridded annual average values for fire features and drought indices are summarized in the same way.

For HWs, the six indices have very similar variabilities over the years(Figure 5a). But we can see that the average number of HW days defined using index HI04 is obviously less than HW defined using other indices. The occurrence and burned area of fires are summarized using MODIS data. FHS_c8c9 are the total number of FHS summarized using MODIS active fire data with confidence level not less than 8 (orange line in Figure 5b), while FHS_c9 are the total number FHS summarized using MODIS activate file data with confidence level 9 only (blue line in Figure 5b). For drought indices, when calculating the drouth days we only include the days labeled as severe drought and extreme drought, that's for days with PDSI values less than -4.0 and other drought indices less than -1.6. From Figure 5, we can see that the peaks of the time series for HWs, fires and drought align quite well, especially in 2002, 2007, 2012, and 2020.

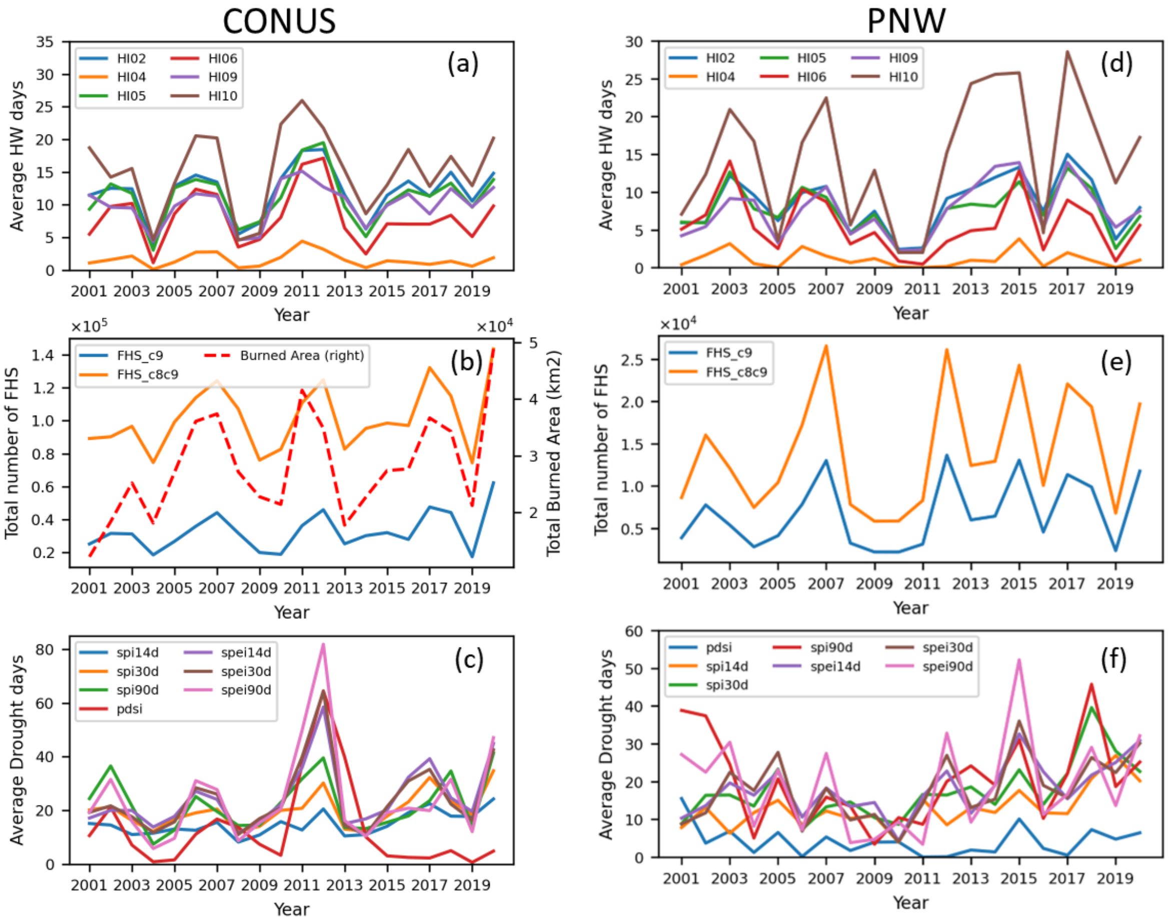

Figure 5. The time series of (a) gridded annual mean of HW days, (b)total number of FHS and Burned Area ,and (c) gridded annual mean drought days for extremes over CONUS(0.5°×0.5°); The time series of (d) gridded annual mean of HW days, (e)total number of FHS and Burned Area ,and (f) gridded annual mean drought days for extremes over PNW(1km×1 km).

We also examined the distribution of HW and fires across different summer months (Figure 6). More than 80% of the HWs occur in July and August, about 15% of them occur in June and September (Figure 6a). The fire season typically starts around May and reaches its peak in August. There are more HW-related Fires in July and August than in other months (Figure 6b).

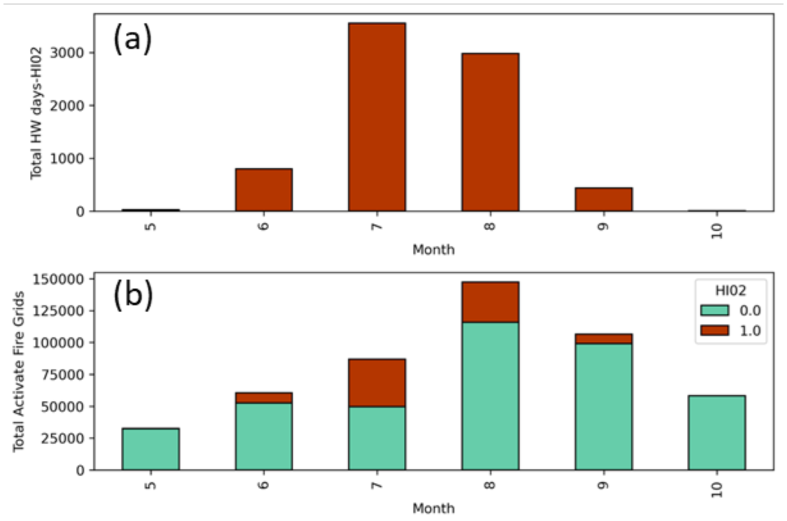

Figure 6. The total HW days for months from May to October; (b) stacked bar plots showing total FHS count in HW days and non-HW days(middle).

Statistical t-tests were conducted to check the hypothesis that the number of FHS is the same for fires occurred in HW and non-HW days, and the hypothesis that the average number of FHS is the same for fires occurred with or without long HWs (duration greater than 8 days). Both t-tests return p-values much smaller than 0.05, suggesting a statistically significant difference in the average number of FHS for fires occurred on different HW-related or non-HW days (Figure 7).

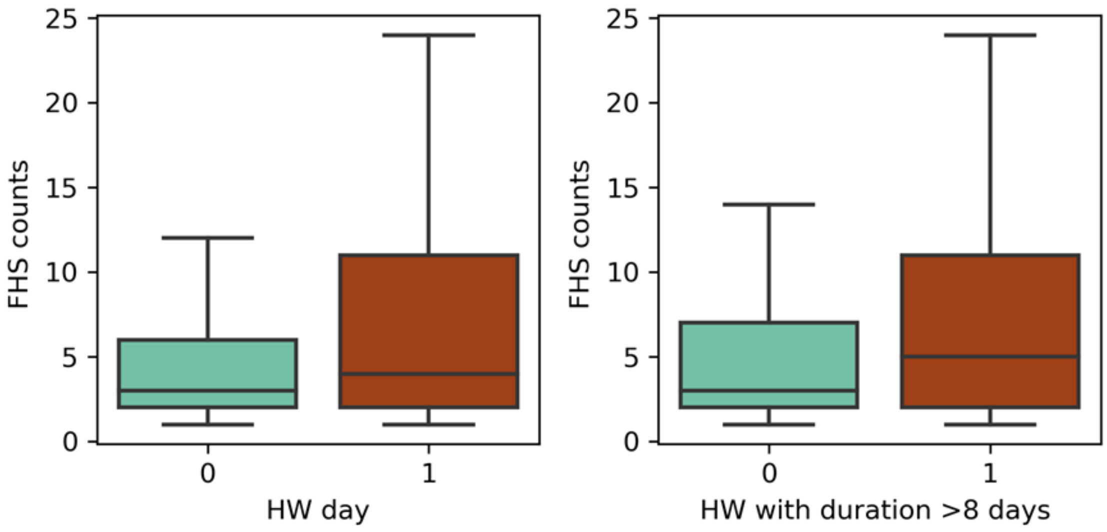

Figure 7. (a)The distribution of the number of FHS for fires occurred on HW days(red) versus non-HW days(green); (b) the distribution of the number of FHS for fires occurred with(red) or without(green) long HWs (duration greater than 8 days).

Next, we investigated one historical period, namely, May to October 2018, to check the co-occurrence of HWs, fires, and droughts across CONUS and the PNW region. The 2018 North American HW season started in late May and reached its peak in mid-July (Figure 8a). HW in the PNW region started a little bit late, around July (Figure 8d). We can see fire seasons started early and reached its peak in August for the CONUS (Figure 8b). Fire season for PNW started a bit late in mid-July and reached its peak in August too (Figure 8e). The fires that started around July coincided with the HW events during July and August. Regarding the drought index, there is an obvious peak in August for drought indices over PNW region (Figure 8f).

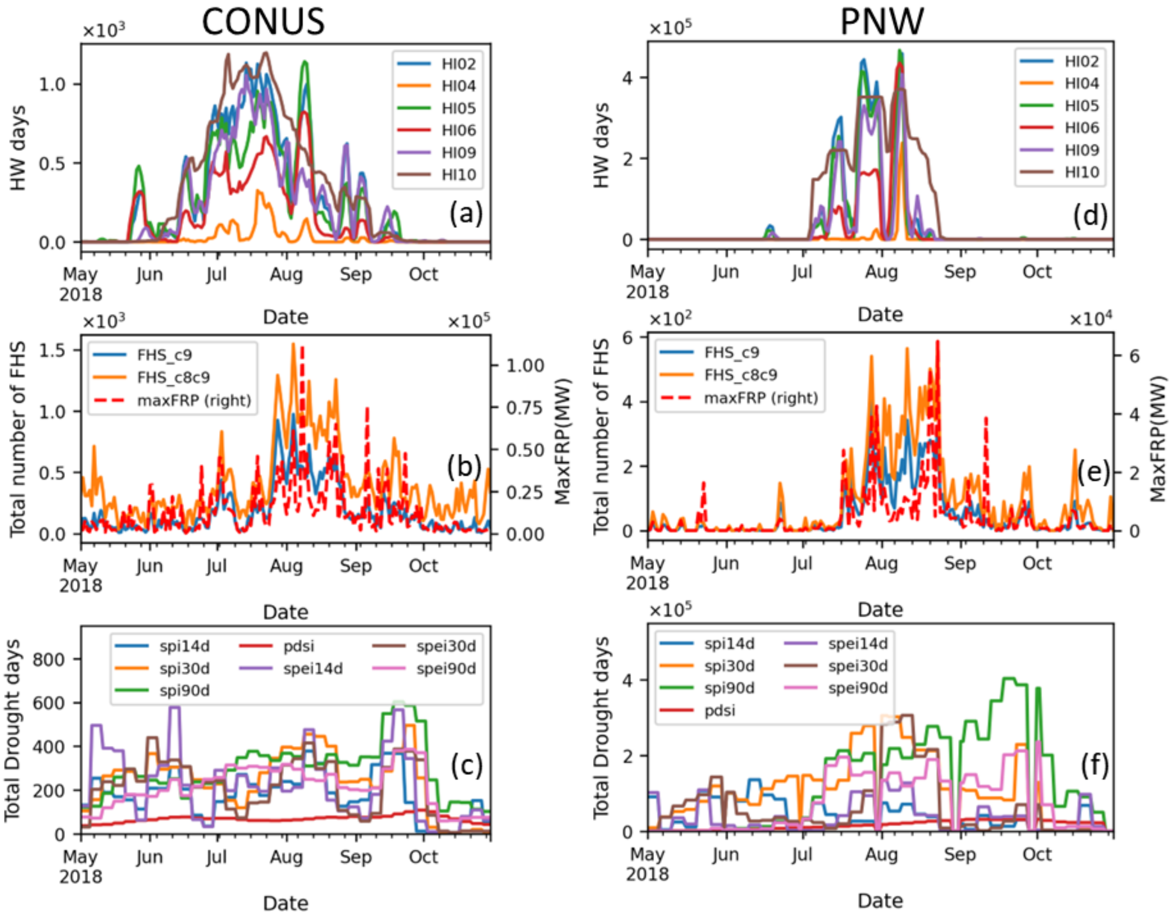

Figure 8. The time series of daily total HW days, total number of FHS, and total drought days from May to October in 2018 over CONUS (a, b, c) and PNW (d, e, f).

4.2 The cross-dependence of compound extremes

The cross-dependence between fires and HWs over CONUS are evaluated by examining the marginal and conditional probabilities of HWs and fires from 2000 to 2020. Overall, the probability maps show that the western(e.g., CA, OR,WA,ID) and south region(e.g., LA,AR) of the US have higher probability for the co-occurrence of compound extremes (i.e., fires and HWs) compared to other regions; the occurrence of HW will increase the chance of fire occurrence, especially in western US.

The marginal and conditional probabilities of HWs and fires are defined as follows:
P(H): the marginal probability of the occurrence of HWs.
P(F): the marginal probability of the occurrence of fires.
P(H∩F): the probability of the co-occurrence of HWs and fires

$P(F|H)$: the probability of the occurrence of fires given that HWs have occurred.

$$P(F|H\ )=(p(F\cap H))/(p(H))$$

$P(H|F)$: the probability of the occurrence of HWs given that fires have occurred.

$$P(H|F)=\ (p(F\cap H))/(p(F))$$

We can see that the probability for the occurrence of HW - P(H) over CONUS is around 0.02 to 0.06, with higher probability in the western and southern regions of the US than in other regions (Figure 9a). The probability of the occurrence of fires P(F) over CONUS is around 0 to 0.05. The west costal region and southeast region have notably higher probabilities of fire occurrence (Figure 9b). The probability of the co-occurrence of fires and HWs $P(H\cap F)$ is generally low, with probability values less than 0.005. However, if HWs have occurred, the probability of fire occurrence (i.e., P(F|H)) increases significantly compared to the marginal probability of fire occurrence P(F), especially in the western region of the US (Figure 9c). Similarly, in western US, if fires occur, it’s very likely that day is a HW day (Figure 9d) with P(H|F) around or greater than 0.5.

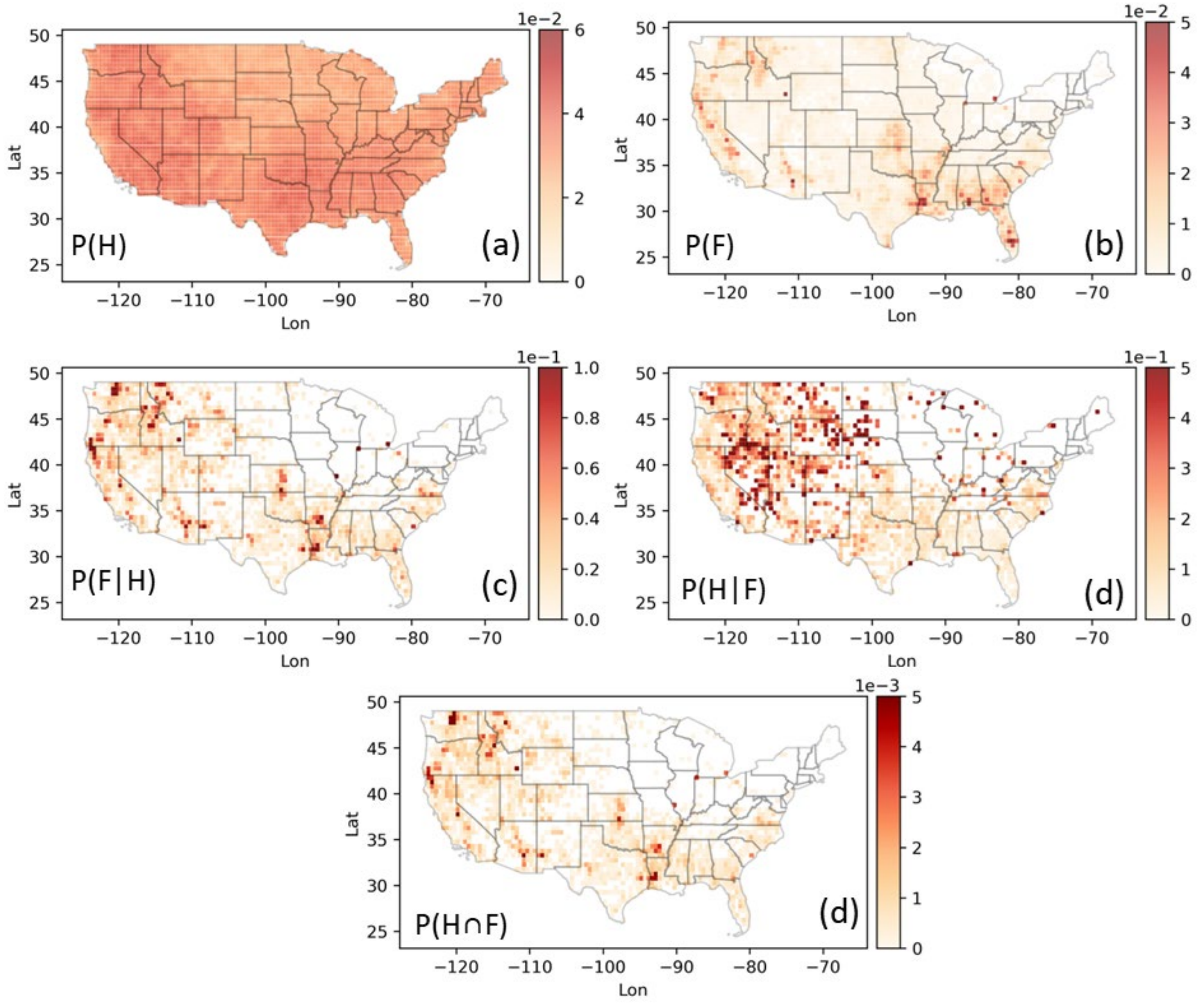

Figure 9. (a)Marginal probability of HW labeled using index HI02;(b) Marginal probability for fires; (c) conditional probability of the occurrence of fires given that HWs (HI02) have occurred; (d) conditional probability of the occurrence of HWs (HI02) given that fires have occurred;(e) the probability of the co-occurrence of HWs (HI02) and fires.

5.0 Data Availability and Usage Notes

The data inventory of US fires, HWs, droughts and the co-located meteorological data are available in CF compliant netCDF file format for the time period 2001–2020, covering the separate spatial extents of CONUS (0.5°×0.5°) and PNW(1km×1km). These data products have been published to the PNNL DataHub, a data management and sharing platform used by researchers and scientists to store, organize, and collaborate on various types of data. The complete data products can be accessed and downloaded at https://doi.org/10.25584/2004956.

The dataset with 0.5°×0.5° resolution for CONUS can be used to help build more accurate climate models for the entire CONUS, which can help in understanding long-term climate trends, including changes in the frequency and intensity of extreme events, predicting future extreme events as well as understanding the implications of extreme events on society and the environment. The data can also be applied for risk assessment of the extremes. For example, ML/AI models can be developed to predict wildfire risk or forecast HWs by analyzing historical weather data, and past fires or heatwaves, allowing for early warnings and risk mitigation strategies. Using this dataset, AI-driven risk assessment models can also be built to identify vulnerable energy and utilities infrastructure, improve grid resilience, and suggest adaptations to withstand extreme weather events.

The high-resolution 1km×1km dataset over PNW are advantageous for real-time, localized, and detailed applications. It can enhance the accuracy of early warning systems for extreme weather events, helping authorities and communities prepare for and respond to disasters more effectively. For example, ML models can be developed to provide localized HW predictions for specific neighborhoods or cities, enabling residents and local emergency services to take targeted actions; the assessment of drought severity in specific communities or watersheds within the PNW can help local authorities manage water resources more effectively.

6.0 Summary and Next Steps

In this study, we compiled a comprehensive AI-ready data inventory of historical extreme events with daily temporal resolution. This inventory covers the separate spatial extents of CONUS (0.5°×0.5°) and PNW(1km×1km) and is intended for various applications and studies. Exploratory data analysis was performed to examine the spatial and temporal distribution of these events over the study period (i.e., 2001-2020) in these regions using the developed data products. The co-occurrence of HWs, fires, and droughts from 2001 to 2020 has also been closely examined. Furthermore, we calculated the marginal, conditional, and joint probabilities of these HWs and fires to understand how the presence of heatwaves influences the likelihood of wildfires and the cross-dependencies among these compound extreme events.

The main outcome of this study contains two data products for CONUS and PNW, and a scientific data paper offering a comprehensive and structured description and exploratory data analysis of the data products. The data products have been published on the PNNL DataHub to enhance the accessibility and usability of the AI-Ready data inventory. We are currently working on the data manuscript, and plan to publish it within the next one or two months.

The AI-ready data inventory can serve the following purposes: (1) support the mechanistic and predictive understanding of extreme events and their impact on the earth's biological and environmental systems; (2) provide probabilistic maps for both individual and compound extreme events, enhancing the direct risk assessment of these extreme weather events for future detection, planning, and decision making; (3) generate various extreme event scenarios and approximate electricity load and generation profiles in energy systems to enhance infrastructure resilience in the face of extreme events.

With additional funding, we aim to further develop, consolidate, and utilize the data products to advance extreme event analysis for general applications in earth, environmental and energy systems, including but are not limited to the following:

(1) Publication of the data paper: An additional month is needed to refine the data paper.

(2) Expansion of exploratory data analysis: Our current exploratory data analysis focuses on the spatial and temporal distribution of single or compound extremes. Further analysis can delve into discovering correlations and causal inference of these extremes in different regions.

(3) AI/ML driven analysis for mechanistic and predictive understanding of extreme events: Extreme events are influenced by a combination of co-located environmental factors which are essential for mechanistic and predictive understanding of extremes. ML/AI models can be developed to predict wildfire risk or forecast HWs by analyzing historical weather data, and past fires or HWs, enabling early warnings and risk mitigation strategies.

(4) Incorporation of co-located environmental factors for AI/ML driven analysis: While our data products include co-located meteorological variables such as precipitation, soil moisture (SM), relative humidity, specific humidity (Qv), and wind data, feedback from energy and earth system experts suggested the need for more. Variables such as vegetation distributions, land cover types, population and housing density could be important controlling factors for the occurrence of wildfires and droughts. Inclusion of such data would enable a more comprehensive analysis of extreme events.

(5) Addition of data for future projections/trends: Our current data products span the historical period from 2001 and 2020. It’s sufficient for studies such as mechanistic understanding of extremes and probabilistic risk assessment. But it may not suffice for exploring future trends of these extremes. To scrutinize long-term climate shifts and forecast these extremes, data from climate model simulations and fire behavior model simulations that consider both natural variability and human-induced fluctuations are indispensable.

We anticipate these additional efforts will significantly amplify our comprehension of extreme events and provide a more comprehensive and inclusive data inventory for AI/ML-centric studies on these extremes.

7.0 References

Abatzoglou, John T. "Development of Gridded Surface Meteorological Data for Ecological Applications and Modelling." *International Journal of Climatology* 33, no. 1 (2013): 121-31. https://doi.org/10.1002/joc.3413.

Alley, William M. "The Palmer Drought Severity Index as a Measure of Hydrologic Drought 1." *JAWRA Journal of the American Water Resources Association* 21, no. 1 (1985): 105-14. https://doi.org/10.1111/j.1752-1688.1985.tb05357.x.

Anderson, G Brooke, and Michelle L Bell. "Heat Waves in the United States: Mortality Risk During Heat Waves and Effect Modification by Heat Wave Characteristics in 43 Us Communities." *Environmental health perspectives* 119, no. 2 (2011): 210-18. https://doi.org/10.1289/ehp.1002313.

Balch, Jennifer K, Lise A St. Denis, Adam L Mahood, Nathan P Mietkiewicz, Travis M Williams, Joe McGlinchy, and Maxwell C Cook. "Fired (Fire Events Delineation): An Open, Flexible Algorithm and Database of Us Fire Events Derived from the Modis Burned Area Product (2001–2019)." *Remote Sensing* 12, no. 21 (2020): 3498. https://doi.org/10.3390/rs12213498.

Beguería, Santiago, Sergio M Vicente-Serrano, Fergus Reig, and Borja Latorre. "Standardized Precipitation Evapotranspiration Index (Spei) Revisited: Parameter Fitting, Evapotranspiration Models, Tools, Datasets and Drought Monitoring." *International journal of climatology* 34, no. 10 (2014): 3001-23. https://doi.org/10.1002/joc.3887.

Bochenek, Bogdan, and Zbigniew Ustrnul. "Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives." *Atmosphere* 13, no. 2 (2022): 180. https://doi.org/10.3390/atmos13020180.

Daly, Christopher, Michael Halbleib, Joseph I Smith, Wayne P Gibson, Matthew K Doggett, George H Taylor, Jan Curtis, and Phillip P Pasteris. "Physiographically Sensitive Mapping of Climatological Temperature and Precipitation across the Conterminous United States." *International Journal of Climatology: a Journal of the Royal Meteorological Society* 28, no. 15 (2008): 2031-64. https://doi.org/10.1002/joc.1688.

Das, P., and K. Chanda. "Bayesian Network Based Modeling of Regional Rainfall from Multiple Local Meteorological Drivers." [In English]. *Journal of Hydrology* 591 (Dec 2020): 125563. https://doi.org/10.1016/j.jhydrol.2020.125563. <Go to ISI>://WOS:000599757800073.

Fang, Wei, Qiongying Xue, Liang Shen, and Victor S Sheng. "Survey on the Application of Deep Learning in Extreme Weather Prediction." *Atmosphere* 12, no. 6 (2021): 661. https://doi.org/10.3390/atmos12060661.

Gelaro, R., W. McCarty, M. J. Suarez, R. Todling, A. Molod, L. Takacs, C. Randles*, et al.* "The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (Merra-2)." *J Clim* Volume 30, no. Iss 13 (Jun 20 2017): 5419-54. https://doi.org/10.1175/JCLI-D-16-0758.1. https://www.ncbi.nlm.nih.gov/pubmed/32020988.

Giglio, Louis, and Christopher Justice. "Mod14a1 Modis/Terra Thermal Anomalies/Fire Daily L3 Global 1km Sin Grid V006." *NASA EOSDIS Land Processes DAAC* 10 (2015). https://doi.org/10.5067/MODIS/MOD14A1.006.

Liu, Yi, Ye Zhu, Liliang Ren, Vijay P Singh, Xiaoli Yang, and Fei Yuan. "A Multiscalar Palmer Drought Severity Index." *Geophysical Research Letters* 44, no. 13 (2017): 6850-58. https://doi.org/10.1002/2017GL073871.

Mesinger, F., G. DiMego, E. Kalnay, K. Mitchell, P. C. Shafran, W. Ebisuzaki, D. Jovic*, et al.* "North American Regional Reanalysis." [In English]. *Bulletin of the American Meteorological Society* 87, no. 3 (Mar 2006): 343-60. https://doi.org/10.1175/Bams-87-3-343. <Go to ISI>://WOS:000236534000016.

Mitchell, Kenneth E, Dag Lohmann, Paul R Houser, Eric F Wood, John C Schaake, Alan Robock, Brian A Cosgrove*, et al.* "The Multi-Institution North American Land Data Assimilation System (Nldas): Utilizing Multiple Gcip Products and Partners in a Continental Distributed Hydrological Modeling System." *Journal of Geophysical Research: Atmospheres* 109, no. D7 (2004). https://doi.org/10.1029/2003JD003823.

Reichstein, Markus, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, and Nuno Carvalhais. "Deep Learning and Process Understanding for Data-Driven Earth System Science." *Nature* 566, no. 7743 (2019): 195-204. https://doi.org/10.1038/s41586-019-0912-1.

Schneider, Tapio, Shiwei Lan, Andrew Stuart, and Joao Teixeira. "Earth System Modeling 2.0: A Blueprint for Models That Learn from Observations and Targeted High-Resolution Simulations." *Geophysical Research Letters* 44, no. 24 (2017): 12,396-12,417. https://doi.org/10.1002/2017GL076101.

Svoboda, Mark, Michael Hayes, and Deborah Wood. "Standardized Precipitation Index: User Guide." (2012).

Thornton, MM, R Shrestha, Y Wei, PE Thornton, SC Kao, BE Wilson, BW Mayer*, et al.* "Daymet: Daily Surface Weather Data on a 1-Km Grid for North America, Version 4 R1." *ORNL DAAC, Oak Ridge, Tennessee, USA. Single Pixel Extraction Tool| Daymet (ornl. gov)* (2022). https://doi.org/10.3334/ORNLDAAC/2129.

Vogel, Martha M, Jakob Zscheischler, Erich M Fischer, and Sonia I Seneviratne. "Development of Future Heatwaves for Different Hazard Thresholds." *Journal of Geophysical Research: Atmospheres* 125, no. 9 (2020): e2019JD032070. https://doi.org/10.1029/2019JD032070.

Wang, Sally S-C, and Yuxuan Wang. "Quantifying the Effects of Environmental Factors on Wildfire Burned Area in the South Central Us Using Integrated Machine Learning Techniques." *Atmospheric Chemistry and Physics* 20, no. 18 (2020): 11065-87. https://doi.org/10.5194/acp-20-11065-2020.

Zarch, Mohammad Amin Asadi, Bellie Sivakumar, and Ashish Sharma. "Droughts in a Warming Climate: A Global Assessment of Standardized Precipitation Index (Spi) and Reconnaissance Drought Index (Rdi)." *Journal of hydrology* 526 (2015): 183-95. https://doi.org/10.1016/j.jhydrol.2014.09.071.

Zscheischler, Jakob, Seth Westra, Bart JJM Van Den Hurk, Sonia I Seneviratne, Philip J Ward, Andy Pitman, Amir AghaKouchak*, et al.* "Future Climate Risk from Compound Events." *Nature Climate Change* 8, no. 6 (2018): 469-77.

Pacific Northwest
National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov