scientific data

Check for updates

OPEN

DATA DESCRIPTOR

SeasFire cube - a multivariate dataset for global wildfire modeling

Ilektra Karasante[1]✉, Lazaro Alonso[2]✉, Ioannis Prapas[1,3], Akanksha Ahuja[1,4], Nuno Carvalhais[2,5,6] & Ioannis Papoutsis[1,7]

Frequent, large-scale wildfires threaten ecosystems and human livelihoods globally. To effectively quantify and attribute the antecedent conditions for wildfires, a thorough understanding of Earth system dynamics is imperative. In response, we introduce the SeasFire datacube, a meticulously curated spatiotemporal dataset tailored for global sub-seasonal to seasonal wildfire modeling via Earth observation. The SeasFire datacube consists of 59 variables including climate, vegetation, oceanic indices, and human factors. It offers 8-day temporal resolution, 0.25° spatial resolution, and covers the period from 2001 to 2021. We showcase the versatility of SeasFire for exploring the variability and seasonality of wildfire drivers, modeling causal links between ocean-climate teleconnections and wildfires, and predicting sub-seasonal wildfire patterns across multiple timescales with a Deep Learning model. We have publicly released the SeasFire datacube and appeal to Earth system scientists and Machine Learning practitioners to use it for an improved understanding and anticipation of wildfires.

Background & Summary

Wildfires, as integral components of terrestrial ecosystems, play a significant role in shaping ecological development through disturbance and regeneration[1]. However, the increasing influence of both climate change and human activities has disrupted the natural fire cycle and modified ecosystems[2,3]. Considering the expected significant changes in climate over the next century, it is imperative to reevaluate wildfire adaptation and mitigation strategies[4]. In terms of impact, wildfires exert a significant ecological influence by enhancing nutrient cycling, initiating ecological succession, creating diverse habitats, and supporting fire-adapted species[5,6]. However, when wildfires exceed expected intensity or frequency, they can have devastating impacts on ecosystem services, infrastructure, communities, and public health[7–11].

To effectively mitigate wildfires, a robust characterization of the complex dynamics spanning atmospheric, oceanic, and terrestrial processes is imperative. Recognizing the urgency inherent in fire prediction and management, we synthesize a novel dataset, named SeasFire, which paves the way for the development of data-driven methods for forecasting wildfire patterns and impacts on a sub-seasonal scale, incorporating teleconnections. The dataset contains 59 global variables including climatic, meteorological, environmental, and human-related wildfire drivers. It also includes historical data on burned areas and carbon emissions, covering the period from 2001 to 2021. These variables are distributed within a 0.25° resolution striking a balance between offering detailed spatial information, ensuring computational efficiency, and facilitating compatibility with existing data sources. The SeasFire datacube[12] has an 8-day temporal resolution, effectively capturing both short-term fluctuations and maintaining a sufficient number of data points across a 21-year duration. This 8-day interval is particularly well-suited for capturing environmental phenomena characterized by seasonal or sub-seasonal cycles.

There have been significant efforts to consolidate wildfire datasets, as illustrated in Table 1, which delineates regions, temporal extents, and characteristics for each dataset. While these datasets provide valuable insights into specific aspects of wildfires, a notable gap exists in the availability of multivariate Earth Data Cubes dedicated to wildfires. Additionally, these datasets have predominantly focused on specific regions within the United States of America and Europe, leading to two consequential limitations. Firstly, the absence of globally scaled

[1]National Observatory of Athens, Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, Penteli, 15236, Greece. [2]Max Planck Institute for Biogeochemistry, Department of Biogeochemical Integration, Jena, 07745, Germany. [3]Universitat de València, Image Processing Laboratory, València, 46022, Spain. [4]University of Cambridge, Department of Engineering, Cambridge, CB2 1PZ, United Kingdom. [5]Universidade Nova de Lisboa, Departamento de Ciências e Engenharia do Ambiente, Faculdade de Ciências e Tecnologia, Caparica, 2829-516, Portugal. [6]ELLIS Unit Jena, Jena, Germany. [7]National Technical University of Athens, School of Rural, Surveying and Geoinformatics Engineering, Zografou, 15773, Greece. ✉e-mail: ile.karasante@noa.gr; lalonso@bgc-jena.mpg.de

Name	Description	Fire Drivers	Region	Years
MODIS[74]	The 1 km MODIS active fire product detects thermal anomalies using a contextual algorithm and provides location, brightness temperature and spatio-temporal attributes.	✗	GLB	2000 -Present
VIIRS[74]	The 375 m VIIRS active fire product detects fires with improved resolution and nighttime performance compared to MODIS and adds fire radiative power.	✗	GLB	2012 -Present
FireCCI[75]	The FireCCI burned area product provides a sum of the burned area, standard error, fraction of burnable area, fraction of observed area, number of patches, and a sum of the burned area for each land cover category.	✗	GLB	2001 -Present
Fire Atlas[76]	The Global Fire Atlas tracks the dynamics of individual fires to determine the location and timing of ignitions, duration and size of fires, daily expansion of fires, along with line length, speed, and direction of fire spread.	✗	GLB	2003–2016
NIFC[77]	The geospatial dataset includes data on the boundaries of wildfires and offers historical information in the United States, including fire boundaries, size, and other attributes using GeoMAC.	✗	US	2000–2018
GeoMAC[78]	Geospatial Multi-Agency Coordination Group (GeoMAC) is an internet-based mapping tool that was initially created for fire managers to obtain real- time fire perimeter data for active wildfires facilitating immediate monitoring and management.	✗	US	2000–2019
CNFDB[79]	Canadian National Fire DataBase serves as a comprehensive repository of wildfire information in Canada, including fire location, size, cause, and suppression details.	✗	CAN	1980 -Present
EFFIS[80]	Acts as a central hub for wildfire information and datasets in Europe, offering data on fire events, burned areas, and fire danger indices.	✗	EU	2015 -Present
Bushfire[81]	Offers vital information on wildfires in Australia, including fire history, severity, and vegetation data.	✓	AUS	1900 -Present
Wildfires Australia[82]	Amount, size and anomalies of surface temperature of wildfires on an interactive dashboard	✓	AUS	2011–2020
Kaggle[83]	Provides information on different fire sizes, frequencies, and causes, offering a comprehensive resource for studying 1.88 million wildfires in the US.	✗	US	1992–2015
Sentimental Wildfire[84]	Integrates geophysical satellite data from the Global Fire Atlas with Twitter's social data and applies sentiment analysis to social media for more accurate predictions of wildfire characteristics.	✓	US, AUS	2016
Incidents[85]	Captures various hazards, with fire-related incidents constituting the major-ity, highlighting the significance of fire-related events.	✗	US	1999–2014
WildfireDB[86]	Encompasses 17 million data points, allowing in-depth understanding of fire spread dynamics in the continental USA over the past decade.	✗	US	2012–2017
FIRE-ML[87]	Provides a daily wildfire forecasting dataset for the contiguous United States, including active fire detections, land cover, and more.	✓	US	2012–2020

Table 1. Summary of Wildfire Datasets (GLB (Global), US (United States), AUS (Australia), CAN (Canada), EU (Europe)).

variables impedes the ability to model Earth system processes and dynamics affecting wildfires on a global scale. This limitation hinders a comprehensive understanding of the interconnectedness of these processes across different regions. Secondly, the potential positive impacts of anticipating seasonal wildfires are constrained in resource-efficient countries and regions. SeasFire's global coverage not only enables a more comprehensive understanding of global wildfire dynamics but also contributes to the development of more robust and location-agnostic models. Such models have the potential to generalize well across diverse geographical areas, transcending the limitations posed by regional-focused datasets.

Consolidating a comprehensive, global, time-series dataset for wildfire modeling using Earth Observation variables poses a significant challenge, as navigating through the ever-expanding data landscape presents formidable hurdles. Indeed, the domain of Earth Observation is undergoing exponential growth, evidenced by the availability of over 500 publicly accessible datasets[13]. These datasets, hosted on various geospatial platforms, such as the Climate Data Store (CDS), Google Earth Engine, and the Copernicus Open Access Hub, collectively amass hundreds of terabytes of data[14]. They offer diverse spatiotemporal scales and modalities, forming a rich repository of information. This wealth of data is distributed across multiple services and repositories. Adding to the complexity, data is collected from a diverse array of sensors such as Landsat, MODIS, and Sentinel, each possessing unique characteristics and capabilities. To harness the potential of such extensive datasets for deriving valuable insights and developing accurate wildfire models, the scientific community must adeptly handle demanding tasks of data selection, access, and harmonization. These challenges can be effectively addressed through the utilization of datacubes[15–17]. Datacubes offer a streamlined approach to managing and analyzing large, multi-dimensional Earth Observation datasets, providing a unified framework for efficient data exploration and model development.

SeasFire cube design is a scientific asset based on a novel paradigm for data-driven wildfire research using analysis-ready data. The cube specifications and granularity allow the development of Earth system deep learning models for wildfire science and beyond, that capture the long spatiotemporal interactions of Earth system variables. Distinguished by its incorporation of ocean climate indices, SeasFire can be used to probe the Earth's spatiotemporal interactions, such as memory effects and teleconnections to capture the dynamic and non-linear interactions of the Earth system components, particularly in the context of seasonal wildfire forecasting. This innovative dataset empowers researchers in the Earth system sciences, facilitating rigorous analytical and predictive modeling on both regional and global scales. In addition, as a cloud-friendly dataset, it also tackles computational challenges and removes storage constraints for seamless data analysis. While SeasFire is particularly suited for machine learning applications in wildfire forecasting, its spatial resolution also makes it highly valuable for global wildfire model development. Researchers can use it to explore hypotheses related to wildfire drivers on multiple temporal scales (sub-seasonal, seasonal, annual), as well as to investigate potential drivers of

inter-annual variability due to its relatively long time frame. The SeasFire datacube[12] can also be used to explore the spatiotemporal distribution of wildfire carbon emissions, track changes, and identify sources and sinks, acknowledging variations based on fuel types. Beyond wildfires, SeasFire enables the study of environmental phenomena such as vegetation dynamics and drought monitoring.

The dataset has already been used for modeling global wildfire patterns with deep learning models[18,19]. Prapas *et al.*[18], use semantic segmentation on the SeasFire cube for burned area pattern forecasting. An extension of this work is TeleVit[19], a transformer model that captures teleconnection information to improve performance at larger forecasting horizons. TeleViT combines local views at higher resolution (0.25°), global views at lower resolution (1°), and time-series of ocean-climate indices, a setting that is allowed by the versatile structure of the SeasFire datacube[12]. Moreover, the Pi-school organization has employed the SeasFire datacube[12] to comprehend sub-seasonal to seasonal forecasts of global burned areas, harnessing explainable artificial intelligence (AI) techniques integrated with deep learning models. Their work can be found in the GitHub repository (https://github.com/PiSchool/noa-xai-for-wildfire-forecasting).

Methods

This section provides an overview of the dataset and all necessary curation steps undertaken for building an Earth system datacube for seasonal and sub-seasonal fire forecasting. Adhering to FAIR principles (Findable, Accessible, Interoperable, and Reusable)[14,20], we designed and constructed the SeasFire datacube[12], ensuring that it offers a user-friendly experience for researchers and analysts, to discover, access, and leverage these variables for various applications. The datacube's architecture is designed to be flexible, accommodating the addition of extra variables, as well as including many variables that grant users the freedom to extend them as necessary over time.

Development of SeasFire datacube. There are three foundational concepts essential to our work: datasets, datacubes (alternatively referred to as data cubes or simply cubes), and data arrays. A dataset functions as a container that accommodates multiple data variables, while a datacube represents a specialized form of dataset specifically designed for spatiotemporal data. Data arrays denote individual variables that may constitute a dataset or datacube. These concepts are frequently employed collectively, particularly in the context of managing complex, multi-dimensional data in scientific and geospatial applications.

Datacubes aim to tackle the challenges posed by Big Data through their cloud-optimized architectures[21–26]. Datacube initiatives have become pivotal in the realm of Earth observation and data analysis, exemplified by the Committee on Earth Observation Satellites (CEOS) as a founding partner of the Open Data Cube initiative[26] and a spectrum of diverse projects worldwide. Flagship projects in this domain include Digital Earth Australia[24,25], the Colombian Data Cube[23], the Swiss Data Cube[27], and several others. Furthermore, noteworthy initiatives include the Semantic Austrian EO Data Cube Infrastructure[22] and the Earth System Data Lab (ESDL)[16], purposefully designed to enhance the efficiency of analyzing analysis-ready data (ARD). CEOS defines ARD as "satellite data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis without additional user effort"[26,28]. Additionally, customized regional initiatives like the Regional Earth System Data Lab[29] further augment the datacube landscape. Furthermore, the existence of large-scale datacube initiatives on a global level, like the Earth System Data Cube[16], highlights the growing importance of this approach in the field of Earth observation.

Building a datacube can be achieved through various methods, with options ranging from dedicated software platforms such as the Open Data Cube (ODC)[26] to Python libraries like xarray[30] or its Julia equivalent YAXArray.jl, each fulfilling distinct roles. ODC primarily emphasizes data storage and retrieval, whereas xarray centers its focus on the manipulation of data within memory. Many research datacubes[16,31,32] tailored to specific research domains, have been developed by using xarray[30] as a key component, harnessing its capabilities for handling diverse data types, including climate data, environmental data, and scientific measurements.

Consequently, by leveraging xarray we developed the SeasFire datacube[12], a harmonized spatiotemporal Earth system datacube capable of accommodating multiple datasets. Datacubes have the ability to conduct concurrent analyses, by effectively converting vast data volumes into easily accessible and valuable insights. To address the challenges of managing multidimensional arrays, we provide the cube in.Zarr specification[33]. This is particularly well-suited for cloud-based environments, as it provides efficient chunk access mechanisms, facilitating seamless parallel processing of data. We exploit cloud data optimization to foster the exploration of the complex and interconnected dynamics of the multivariate Earth system, employing a standardized approach to generate customized datacubes instantly. This approach aims to help Earth system scientists and machine learning practitioners select the variables and dimensions essential for training their models.

SeasFire datacube variables. The SeasFire datacube[12] offers a comprehensive collection of variables that capture key environmental factors associated with fire drivers as well as fire targets such as burned area, fire radiative power, and wildfire-induced carbon dioxide emissions. In total, this datacube comprises 59 variables spanning various domains. These include climatic elements, featuring ocean-climate indices; atmospheric parameters, including temperature, pressure, and humidity-related variables; land-related aspects, like land cover and population density variables; and several masks as biomes and land-sea mask, all of which exert influence on fire behavior. These categories are organized as shown in Table 2. Furthermore, the datacube includes distinct data arrays that capture various ways of summarizing a variable. For instance, it provides temperature data at a 2-meter height in maximum, minimum, and average values. Additionally, the datacube contains arrays of data that originate from breaking down information initially found in a combined variable. An example of this is the comprehensive land cover category that encompasses multiple classes. A process of aggregation and regridding, results

Provider	Dataset	Variables	Time Span/Res.	Spatial Res.
Climate Data Store	ERA5 hourly data on single levels from 1940 to present[50]	Mean sea level pressure; Total precipitation; Relative humidity; Vapor Pressure Deficit; Sea Surface Temperature; Skin temperature; Wind speed at 10 meters; Temperature at 2 meters; Surface net solar radiation; Surface solar radiation downwards; Volumetric soil water levels 1,2,3 and 4	2001–2021 /hourly	0.25°
Climate Data Store	CEMS Global Fire Assimilation System Historical Data[88]	Drought Code; Fire Weather Index	2001–2021 /hourly	0.25°
Atmosphere Data Store	CAMS global biomass burning emissions based on fire radiative power (GFAS)[89]	Carbon dioxide emissions from wildfires; Fire radiative power	2003–2021 / daily averages	0.1°
NOAA, National Oceanic, and Atmospheric Administration	Climate Indices: Monthly Atmo- spheric and Ocean Time Series[90]	Western Pacific Index; Pacific North Amer- ican Index; North Atlantic Oscillation; Southern Oscillation Index; Global Mean Land/Ocean Temperature; Pacific Decadal Oscillation; Eastern Asia/Western Russia; East Pacific/North Pacific Oscillation; Nino 3.4 Anomaly; Bivariate ENSO Timeseries; Arctic Oscillation	2001–2021 /monthly	—
ESA CCI	Land cover classifi- cation gridded maps from 1992 to present derived from satellite observations[48]	No data; Agriculture; Forest; Grassland; Wetlands; Settlement; Shrubland; Sparse vegetation; bare areas, permanent snow and ice, Water Bodies	2001–2021 /yearly	300 m
NASA LP DAAC at the USGS EROS Center	MOD11C3 v006[51] / MCD15A2H v006[52]/ MOD13C1[47]	Land Surface Temperature / Leaf Area Index / Normalized Difference Vegetation Index	2001–2021 / 8 & 16 days averages	0.05° /500 m /0.05°
RESOLVE Bio-diversity and Wildlife Solu- tions	RESOLVE Ecore- gions 2017[53]	biomes	static / -	—
NASA SEDAC, Socioeconomic Data and Applica- tions Center	GPWv411: UN-Adjusted Popula- tion Density (Grid- ded Population of the World Version 4.11)[91]	Population density	2000; 2005; 2010; 2015; 2020 / 5 years	0.25°
Global Wildfire Information Sys- tem (GWIS)	GlobFire Fire Perimeters (2002–2023)[37]	Burned Areas	2002–2023 /monthly	500 m
ESA CCI	MODIS Fire_cci Burned Area pixel product version 5.1 (FireCCI51)[54]	Burned Areas; Fraction of burnable area; Number of patches; Fraction of observed area	2001–2020 /monthly	0.25°
ORNL DAAC	Global Fire Emis- sions Database (GFED.v4)[38]	Burned Areas (large fires only); basis re- gions (mask)	2001–2015 /monthly	0.25°

Table 2. Summary of curated datasets, encompassing information on providers, variables, and spatiotemporal resolution.

in the creation of individual variables, each focused on representing a specific land cover class. Incorporating a comprehensive set of variables within the datacube, enriches the information available to machine learning models, enabling a more customized and understandable modeling of seasonal fire patterns.

There are additional datasets that could further enhance its capabilities. Users looking to expand the scope of their analyses may consider incorporating additional data such as human land fragmentation data, topography data, and lightning data. As for the later they could help explore natural wildfire ignition patterns and better understand spatial-temporal variations in fire occurrences driven by lightning, complementing the existing atmospheric and environmental variables within the datacube. Integrating data—such as human land fragmentation, urban expansion, and agricultural intensification— have the potential to enhance our understanding of anthropogenic influences on wildfire dynamics, particularly in the wildland-urban interface[34], as well as the inclusion of Global Human Settlement Layer data[35] that could provide further insights into human-wildfire interactions and urban expansion trends. Finally, incorporating global topographic variables[36] -even in such coarse resolution- could possibly offer valuable insights into broader fire behavior patterns across large landscapes.

Burned areas variables. The datacube contains three different products related to burned areas: the GWIS[37], the GFED[38], and the FCCI[39] dataset. Global burned area products still rely on moderate resolution sensors, with 250–500 m pixel sizes and 1–2 days revisit time[40]. Despite Landsat and Copernicus Sentinel-2 satellites openly providing better resolutions (10–30 m), the obstacle of extensive processing effort required to produce comprehensive, long-term global datasets remains[40]. Therefore securing global, daily data on burned areas has been a notable challenge. Since 2001, the MODIS Aqua and Terra satellites have been delivering consistent and quality-checked data, inspiring various organizations to compile global burned areas datasets using a variety of data sources, methodologies, and spatiotemporal resolutions. A diverse range of burned area products has been developed using data from these satellite sensors. Some noteworthy examples include the MCD64A1 v6.0 dataset[41], which provides spectral reflectance data at 500 meters resolution, along with MODIS 1-kilometer active fire data[42–44], spanning from 2000 to the present day. Given the benefits of that product, numerous researchers and organizations have advanced their efforts to develop robust datasets for burned areas, each with its unique set of advantages and drawbacks. For instance, the GFED[38] dataset provides daily data up to 2015 up to 500 meters resolution, while the FCCI[39] (Fire Climate Change Initiative) dataset, developed by the European Space Agency (ESA), offers monthly data at 250 m resolution. Furthermore, the GWIS[37] dataset has produced global shapefiles on a yearly scale along with the ignition dates, spanning from 2002 to 2023 at a 500-meter resolution. Due to variations in data processing applied for each dataset, the burned area products and corresponding

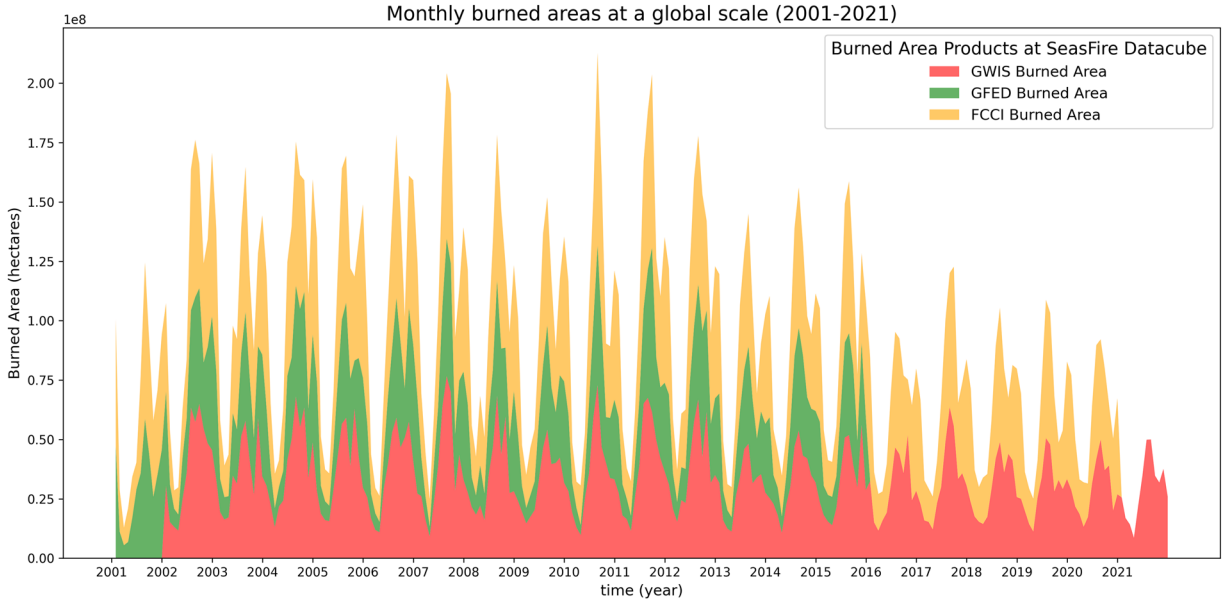

Fig. 1 Global monthly timeseries of the three burned areas products in the SeasFire datacube (GFED Burned Area: gfed_ba, GWIS Burned Area: gwis_ba, FCCI Burned Area: fcci_ba).

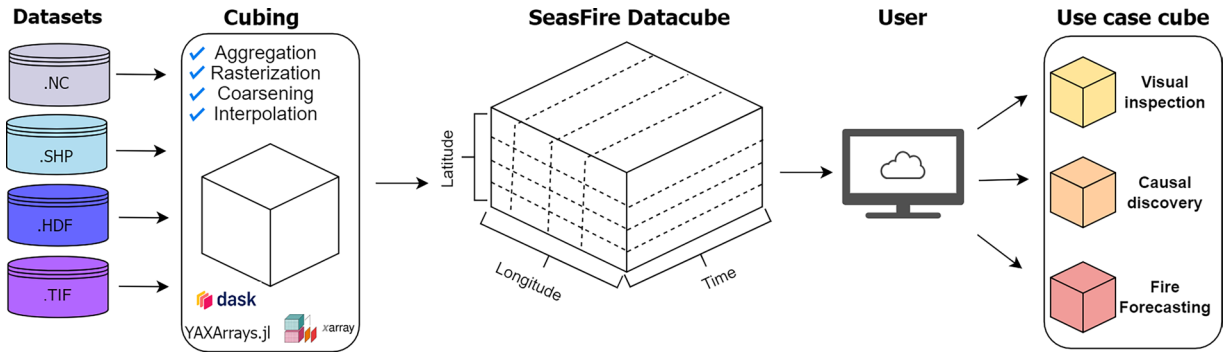

Fig. 2 Visualization of the datacube construction workflow. The data are collected from various sources and in different formats. The input data are preprocessed using interpolation, aggregation, coarsening, and rasterization and then appended in the datacube on the corresponding date. After the creation of the datacube, each user can extract task-related machine-learning datacubes.

values in hectares are different (Fig. 1). These differences are well-documented and explained in relevant literature[45,46]. By including those three datasets in the SeasFire datacube[12] we allow researchers to select the one that aligns most effectively with their specific requirements.

Assumptions & Specifications. The SeasFire datacube[12] offers a spatiotemporal perspective that facilitates the examination of global fire patterns. Table 2 summarizes the various dataset collections, including their spatiotemporal details. The data are organized in a three-dimensional grid, with each grid cell covering a spatial resolution of 0.25° in latitude and longitude. A 0.25° global grid aligns with various global datasets and standards, facilitating interoperability and compatibility with existing data sources (e.g. Climate data store, ESA CCI, NASA SEDAC) and at the same time has a fine resolution to capture variations in atmospheric and climatic data. Temporally, the datacube provides information aggregated over 8-day intervals, allowing for seasonal and sub-seasonal analysis and forecasting. The decision to use an 8-day resolution is deliberate, given the availability of the MODIS NDVI[47] product in a 16-day resolution, making the 8-day interval well-suited for resampling. The 8-day granulation commences on January 1st of each year, resulting in an annual total of 46 datetimes. With a span of 21 years, the datacube allows for long-term trend analysis and exploration of interannual variability.

Data acquisition and ARD generation. Figure 2 illustrates the practical implementation of the aforementioned concept. The flowchart delineates the process, commencing with the collection of relevant datasets. Subsequently, these datasets undergo reprocessing in order to be integrated to as a single datacube. Users can then

Provider	Dataset	Acquisition	Format
Climate Data Store	ERA5 hourly data on single levels from 1940 to present[50]	Web API: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview	Netcdf
Climate Data Store	CEMS Global Fire Assimilation System Historical Data[88]	Web API: https://cds.climate.copernicus.eu/cdsapp#!/dataset/cems-fire-historical?tab=overview	Netcdf
Atmosphere Data Store	CAMS global biomass burning emissions based on fire radiative power (GFAS)[89]	Web API: https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-fire-emissions-gfas?tab=form	Netcdf
NOAA, National Oceanic, and Atmospheric Administration	Climate Indices: Monthly Atmospheric and Ocean Time Series[90]	Web Scrapping: https://psl.noaa.gov/data/correlation/	text
ESA CCI	Land cover classification gridded maps from 1992 to present derived from satellite obser- vations[48]	Web Scrapping: https://data.ceda.ac.uk/neodc/esacci/fire/data/burned_area/MODIS/pixel/v5.1/compressed	Netcdf
NASA LP DAAC at the USGS EROS Center	MOD11C3 v006[51] / MCD15A2H v006[52] / MOD13C1[47]	Web API: https://lpdaac.usgs.gov/product_search/?status=Operational	Netcdf / Shapefile/Netcdf
RESOLVE Bio-diversity and Wildlife Solu- tions	RESOLVE Ecoregions 2017[53]	Web Scrapping: https://storage.googleapis.com/teow2016/Ecoregions20 17.zip	Shapefile
NASA SEDAC, Socioeconomic Data and Applica- tions Center	GPWv411: UN-Adjusted Population Density (Gridded Population of the World Ver- sion 4.11)[91]	Web Scrapping:https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-adjusted-to-2015-unwpp-country-totals-rev11/data-download	Netcdf
Global Wildfire Information Sys- tem (GWIS)	GlobFire Fire Perimeters (2001–2020)[37]	Web Scrapping: https://gwis.jrc.ec.europa.eu/apps/country.profile/downloads	Shapefile
ESA CCI	MODIS Fire_cci Burned Area pixel product version 5.1 (FireCCI51)[54]	Web API: https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-land-cover?tab=form	Netcdf
Oak Ridge Na- tional Laboratory (ORNL) Dis-tributed Active Archive Center (DAAC)	Global Fire Emissions Database (GFEDv4)[38]	Web Scrapping: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1293	hdf

Table 3. Dataset acquisition and formats.

LCCS Class	Long Name	Description
lccs_class_0	No data	LCCS code [0]
lccs_class_1	Agriculture	LCCS codes [10,11,12,20,30,40]
lccs_class_2	Forest	LCCS codes [50,60,61,62,70,71,72,80,81,82,90,100]
lccs_class_3	Grassland	LCCS codes [110,130]
lccs_class_4	Wetland	LCCS codes [160,170,180]
lccs_class_5	Settlement	LCCS code [190]
lccs_class_6	Shrubland	LCCS codes [120,121,122]
lccs_class_7	Sparse vegetation, bare areas, permanent snow and ice	LCCS codes [140,150,151,152,153,200,201,202,220]
lccs_class_8	Sparse vegetation, bare areas, permanent snow and ice	LCCS code [210]

Table 4. LCCS Classes and code grouping in SeasFire datacube (Source: Annex 6 of the Product User Guide. For detailed information, visit the guide: https://cds.climate.copernicus.eu/datasets/satellite-land-cover?tab=documentation).

load the variables essential for their specific use case, whether on a global scale spanning 21 years or for regional areas and specific time ranges.

We acquired our data from various sources documented in Table 3, such as the European Centre for Medium-Range Weather Forecasts (ECMWF), Copernicus Climate Data Store (CDS) (https://cds.climate.copernicus.eu/cdsapp#!/home). In the same table, we catalog the acquisition methods and formats of each dataset. We collected data covering different temporal resolutions: (a) hourly-daily records for meteorological variables, (b) monthly (e.g. ocean-climate indices), or (c) yearly and sub-yearly (e.g. land cover, and population density) for the study period.

We processed each data variable, by filling in missing values (e.g. population density is provided for every five years, so we forward filled the timestamps of the datacube for the next four years), transforming the data into a consistent format (e.g. shapefiles to netcdf, and missing_flag as np.nan for all variables), and applying the appropriate land-sea mask to each variable where needed. We then merged all the data, creating a single Zarr file. A detailed data acquisition, aggregation process, and assumptions made are described in detail for each variable in Table 5. All the acquisition-related information is also compacted and stored as a dictionary on its variable's metadata, so the user can have access directly to the data information.

The main processing techniques are described below and appear on each dataset in Table 5:

Aggregation. Aggregation is the process of creating a more generalized representation of data by either resampling timeseries data or reducing its resolution or granularity. Inherently aggregation results in information loss, which in the SeasFire cube dataset is not relevant for the scope of global wildfire modeling. Temporal aggregation occurs when we convert hourly or daily meteorological data into 8-day averages or cumulative totals,

Dataset	Variables	A	I	F	R
ERA5 hourly data on single levels from 1940 to present	Mean sea level pressure; Total precipitation; Relative humidity; Vapor Pressure Deficit; Sea Surface Temperature; Skin temperature; Wind speed at 10 meters; Temperature at 2 meters; Surface net solar radiation; Surface solar radiation downwards; Volumetric soil water levels 1,2,3 and 4	✓	✓	✗	✗
CEMS Global Fire Assimilation System Historical Data.	Drought Code; Fire Weather Index	✓	✓	✗	✗
CAMS global biomass burning emissions based on fire radiative power (GFAS)	Carbon dioxide emissions from wildfires; Fire radiative power	✓	✓	✗	✗
Climate Indices: Monthly Atmo- spheric and Ocean Time Series	Western Pacific Index; Pacific North Amer- ican Index; North Atlantic Oscillation; Southern Oscillation Index; Global Mean Land/Ocean Temperature; Pacific Decadal Oscillation; Eastern Asia/Western Russia; East Pacific/North Pacific Oscillation; Nino 3.4 Anomaly; Bivariate ENSO Timeseries; Arctic Oscillation	✓	✗	✗	✓
Land cover classification gridded maps from 1992 to present derived from satellite observations	No data; Agriculture; Forest; Grassland; Wetlands; Settlement; Shrubland; Sparse vegetation; bare areas, permanent snow and ice, Water Bodies	✓	✓	✗	✗
MOD11C3 v006 / MCD15A2H v006/ MOD13C1	Land Surface Temperature / Leaf Area Index / Normalized Difference Vegetation Index	✓	✓	✗	✓
RESOLVE Ecoregions 2017	biomes	✗	✗	✗	✓
GPWv411: UN-Adjusted Popula- tion Density (Gridded Population of the World Version 4.11)	Population density	✓	✓	✗	✗
GlobFire Fire Perimeters (2001–2020)	Burned Areas	✓	✗	✓	✓
MODIS Fire_cci Burned Area pixel product version 5.1 (FireCCI51)	Burned Areas; Fraction of burnable area; Number of patches; Fraction of observed area	✓	✓	✗	✗
Global Fire Emissions Database (GFEDv4)	Burned Areas (large fires only); basis re- gions (mask)	✓	✓	✗	✓

Table 5. Variables and processing techniques. A: Aggregation, I: Interpolation, F: Filtering, R: Rasterization.

based on the dataset's requirements. Similarly, GWIS daily wildfire data are compressed into 8-day intervals. As a result, all fire events that occur within this window are merged into a single summary, making it impossible to track the progression of individual fires across days.

Thematic aggregation involves merging certain categorical classes to simplify classification. This approach was applied to the ESA Land Cover Dataset[48] to enhance its usability for global wildfire modeling. We consolidated similar land cover classes into nine broader thematic categories, as shown in Table 4. The original dataset contains numerous land cover codes with fine distinctions between land types. However, thematic aggregation merges similar classes, representing only the primary thematic category. For instance, broad-leaved and coniferous forests are now grouped under a single "Forest" class in the SeasFire cube.

Spatial aggregation is discussed below in the rasterization paragraph.

Interpolation. Interpolation is a mathematical process used to estimate values for missing or unknown data points by leveraging the available data. It is often employed to fill gaps in a dataset or create a smoother representation of the data. In certain cases, interpolation serves as a technique for resampling or resizing data. We applied nearest-neighbor interpolation, where needed, to ensure compatibility between our datacube's grid and the spatiotemporal datasets we included.

Filtering. Filtering is the process of selectively extracting or retaining specific portions of the data while removing unwanted components. It can involve removing noise, outliers, or irrelevant data points based on specific criteria or filters. Filtering was performed in the GWIS dataset, where all the active fire data were removed.

Rasterization. Rasterization is the process of converting high resolution vector-based data (such as points, lines, or polygons) into a raster or grid format. It involves spatial aggregation by assigning values or attributes to each cell or pixel in the raster grid based on the characteristics of the original vector data. Rasterization is used in the ESA Land Cover and GWIS dataset for mapping each shapefile to the correct grid cell.

The aggregation of the original 300-meter ESA Land Cover dataset into a 0.25-degree resolution resulted in some loss of spatial detail. In the original dataset, numerous land cover codes are present, offering fine distinctions between various land types. However, through rasterization, the shift to a coarser resolution resulted in a loss of spatial details. While the datacube reflects the percentage of each land cover type within each 0.25° grid cell, it fails to indicate the precise spatial arrangement of these classes within the cell.

Similarly, the transformation of the GWIS dataset from vector to raster format involves a trade-off between precision and scalability, affecting spatial detail. In the vector format, fires are represented with precise boundaries, enabling detailed analysis of fire patterns and impacts. However, the raster format summarizes burned areas within 0.25° grid cells, losing the specific locations, shapes, and boundaries of fires. This shift results in more generalized fire data, obscuring fine-scale spatial details crucial for understanding wildfire behavior.

GWIS Burned areas dataset. The GWIS burned area data array, derived from the GlobFire vector dataset[37]. It consists of global fire perimeters provided as yearly ESRI shapefiles (.shp). Each shapefile includes essential attributes such as a unique fire identification code, initial date (initialdat), geometry data, and the burned areas in hectares. In order to transform the vector dataset into a rasterized product,we first intersected the polygons with our grid using the geopandas library[49], then quantified the burned areas in hectares for each grid cell, and

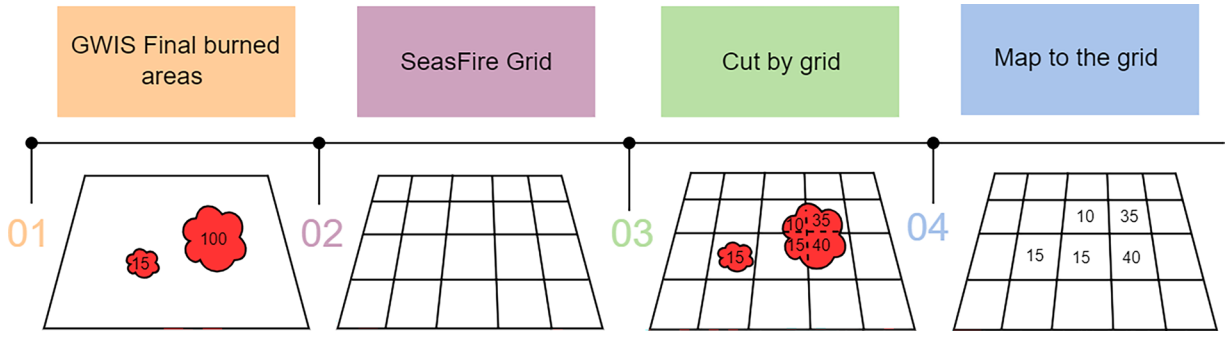

Fig. 3 Workflow toy example of rasterization technique.

gwis_ba	(time, latitude, longitude)	float32	dask.array<chunksize=(966, 180, 360), me...
aggregation :	Spatio-Temporal | sum		
creator_notes :	The missing data for the year 2001 has been filled with NaN values. The dataset is constructed using the ignition dates of the burned areas shapefiles. Each year is divided into 46 periods of 8 days, with the last period in December consisting of either 5 or 6 days, depending on whether it is a leap year.		
downloaded_from :	https://gwis.jrc.ec.europa.eu/apps/country.profile/downloads		
long_name :	Burned Areas from GWIS		
provider :	Global Wildfire Information System (GWIS)		
units :	hectares (ha)		

	Array	Chunk
Bytes	3.73 GiB	238.79 MiB
Shape	(966, 720, 1440)	(966, 180, 360)
Count	17 Tasks	16 Chunks
Type	float32	numpy.ndarray

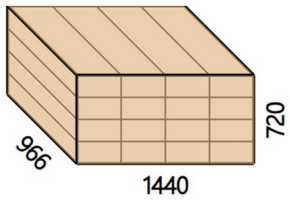

Fig. 4 Metadata example of gwis_ba variable in the datacube. Description of the variable, aggregation performed, units, download link, long name, dataset provider, and notes for the user.

finally converted the data into our datacube format, with the help of the xarray library[30]. The entire rasterization process involved several distinct steps, which are outlined as follows, and visually represented in Fig. 3:

1. **Data Processing:** We treat each yearly shapefile separately containing all the fire events that started during that period.
2. **Create a Grid:** We create a grid according to the spatial dimensions of the datacube [−90°, 90°, −180°, 180°] with the step of 0.25°.
3. **Split the geometry of fire across grid cells:** Intersect the grid with the geodataframe. For example, Fig. 3, demonstrates how fire events are allocated to each grid cell. In this step we change the coordinate reference system from geographic WGS84 (ESPG:4326) with degree units to an equal area projected WGS84 (ESPG:8857) with units in metres, to calculate the area in hectares.
4. **Assign burned areas to the datacube:** We assign the hectares of each geometry to the grid cell of the SeasFire cube it belongs to, using a rasterio.features.geometry_mask (https://rasterio.readthedocs.io/en/latest/api/rasterio.features.html) module.

Data Records

The SeasFire datacube version 0.4[12] is stored in a.zarr file and can be accessed via Zenodo https://zenodo.org/record/13834057. The dataset's acquisition process is comprehensively described in Table 2. The original data sources are cited within this paper (refs. [37,38,47,48,50–63]). Every variable includes inherent descriptions, details on aggregation, long names, dataset providers, and user notes, all stored as attributes (metadata) (Fig. 4) within the Zarr file.

The Zarr format allows optimized chunking for efficient access, storage, and time-based processing of multidimensional gridded data. Each variable is divided into 16 large chunks, structured with dimensions of time (966), latitude (180), and longitude (360), as illustrated in Fig. 4. Zarr integrates seamlessly with Dask and Xarray, facilitating advanced data analysis while allowing users to work with specific sections of the cube without

Dataset	Comparisons with other datasets	Uncertainty estimates	performance check across smaller areas
ERA5 hourly data on single levels from 1940 to present	✓	✗	✓
CEMS Global Fire Assimilation System Historical Data.	✓	✗	✓
CAMS global biomass burning emissions based on fire radiative power (GFAS)	✓	✗	✓
Climate Indices: Monthly Atmo- spheric and Ocean Time Series	N/A	N/A	N/A
Land cover classification gridded maps from 1992 to present derived from satellite observations	✓	✗	✓
MOD11C3 v006 / MCD15A2H v006/ MOD13C1	✓	✓	✓
RESOLVE Ecoregions 2017	✓	✗	✓
GPWv411: UN-Adjusted Popula- tion Density (Gridded Population of the World Version 4.11)	N/A	N/A	N/A
GlobFire Fire Perimeters (2001–2020)	✓	✗	✓
MODIS Fire_cci Burned Area pixel product version 5.1 (FireCCI51)	✓	✓	✓
Burned Areas from GFED (large fires only)	✓	✗	✓

Table 6. Overview of dataset's evaluation.

loading the entire dataset. Extending the datacube with new variables is straightforward, requiring minimal effort if the new data aligns with the existing spatial or temporal axes and compatible format. The SeasFire datacube can be easily updated with other satellite EO data collections (e.g., MODIS, Sentinel-5P), as well as in the time dimension. As new data will be organized and pre-processed following the protocols presented in this paper, new data streams can be readily included.

Technical Validation

The SeasFire datacube[12] gathers and harmonizes validated data records from the respective data providers as shown in Table 6. To enhance the reliability and accuracy of the SeasFire datacube[12], a comprehensive technical validation process was conducted, combining analysis and discovery techniques. This validation aimed to assess the quality, consistency, and usability of the datacube for effective wildfire monitoring and analysis. Although wildfire prediction is the primary aim, the data can also be used in the form of individual timeseries. Below we provide some examples to demonstrate the sanity check of some data in the form of data exploration (Visual Inspection) and causal links (Causality).

Moreover, throughout the development of this datacube, we have committed to continuous refinement and enhancement, resulting in three distinct versions. Each version marks a substantial step forward in terms of quality and utility, and comprehensive changelogs for all are available on Zenodo (https://zenodo.org/record/8055879) for reference. These versions are a testament to our ongoing commitment to delivering the most valuable and up-to-date resource for the research community.

Visual inspection. During data analysis, a preliminary visual inspection plays a pivotal role in unveiling the datacube's underlying characteristics. This process entails a survey of the datacube's variables, with a focus on those used for causality assessment and subsequent machine learning modeling, gaining a deeper understanding of the temporal and spatial patterns of wildfire dynamics around the globe. In Fig. 5, we included an example of how visual examination reveals trends, seasonal patterns, and relationships within the data, offering valuable clues for the formulation of hypotheses and the design of appropriate modeling strategies. This task here has been done using the plotting library Makie.jl[64].

Causality. Causal analysis can be a method for validating the quality of datasets by confirming theoretical Earth system science cause-and- effect relationships among variables, strengthening the overall reliability of the datacube. To demonstrate this, we conducted an experiment linking climate, meteorology, and burned areas in the European Mediterranean and Boreal region over two decades, using the SeasFire datacube[12]. The PCMCI method[65] for causal discovery, designed for timeseries data, employs the Peter-Clark (PC) algorithm that identifies all the causal graphs that are consistent with the available data, and Momentary Conditional Independence (MCI) test to further assess causal relationships. The MCI test takes into account factors like auto-correlation and erroneous edge detections, enhancing the accuracy of causal discovery. Tigramite (https://tocsy.pik-potsdam.de/tigramite.php), a versatile Python framework, supports PCMCI and other methods for causal discovery, even accommodating nonlinearities. However, geoscientific timeseries, like those involving oceanic and atmospheric processes, pose challenges due to non-Gaussian noise[66], making it difficult for statistical tests to capture complex nonlinear relationships between variables.

Our case study employed linear partial correlation tests (ParCorr) with PCMCI under specified assumptions[67,68], via preprocessing, parameter tuning, and causal network learning. The causal graphs are presented in Fig. 6, and illustrate the causal relationship (positive or negative), strength (0 to |1|), and time lag (0 to N months) of that relationship. The initial preprocessing involved transforming variables including Burned Areas, North Atlantic Oscillation, Arctic Oscillation, El-Niño in 3.4 region, precipitation, vapor pressure deficit, and temperature into monthly timeseries while ensuring their stationarity through anomaly calculation for all variables except oscillations and El-Niño in 3.4 region. For parameter tuning, a maximum lag of 6 months was chosen to account for seasonal changes, and an alpha significance threshold of 0.05 was applied to the independence test. To improve the outcome and minimize the presence of spurious links, we established a causal order as it appears in the grey circle in the middle of Fig. 6. In this order, the possible causal links are the following. The

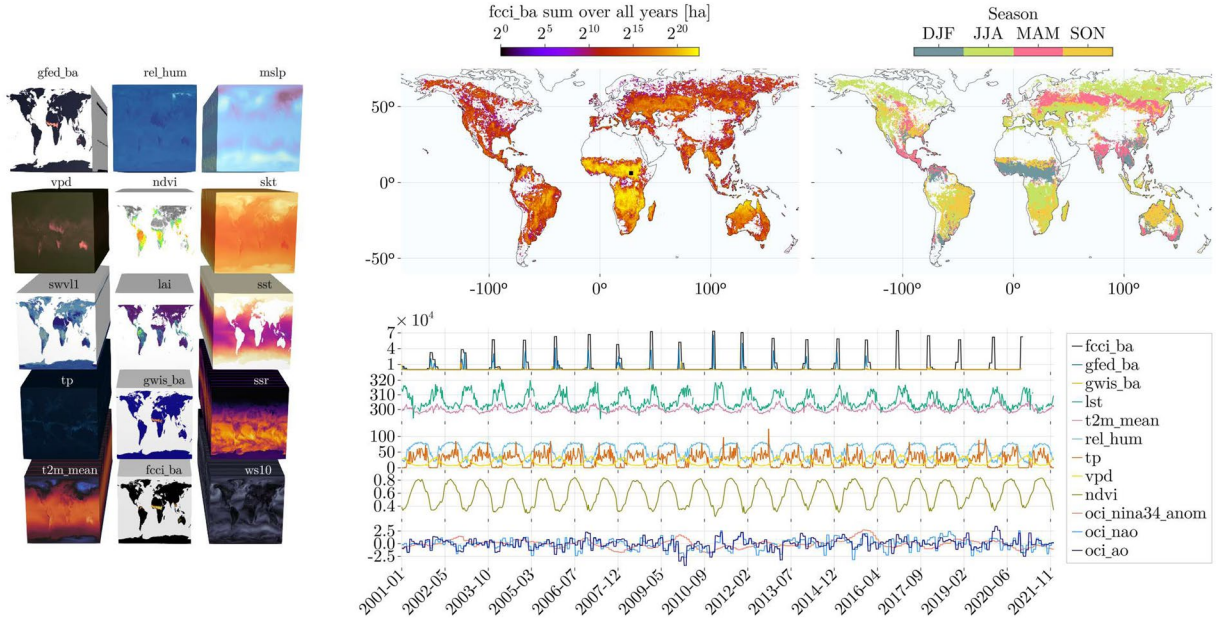

Fig. 5 Visual inspection of the datacube that primarily highlights key variables. The left panel displays the variables selected for the subsequent machine learning forecasting approach. The top/middle map illustrates the cumulative burned area across multiple years, with the black square denoting the location of maximum accumulation. This location serves as the basis for extracting the timeseries displayed at the bottom. Additionally, we also include data on El-Niño (oci_nina34_anom), the North Atlantic Oscillation (oci_nao), and the Arctic Oscillation (oci_ao), all of which are employed in the subsequent causality analysis. In the top-right map, we present the season with the highest burned area, represented by the acronyms DJF(Dec-Jan-Feb), JJA(Jun-Jul-Aug), MAM(Mar-Apr-May), and SON(Sep-Oct-Nov).

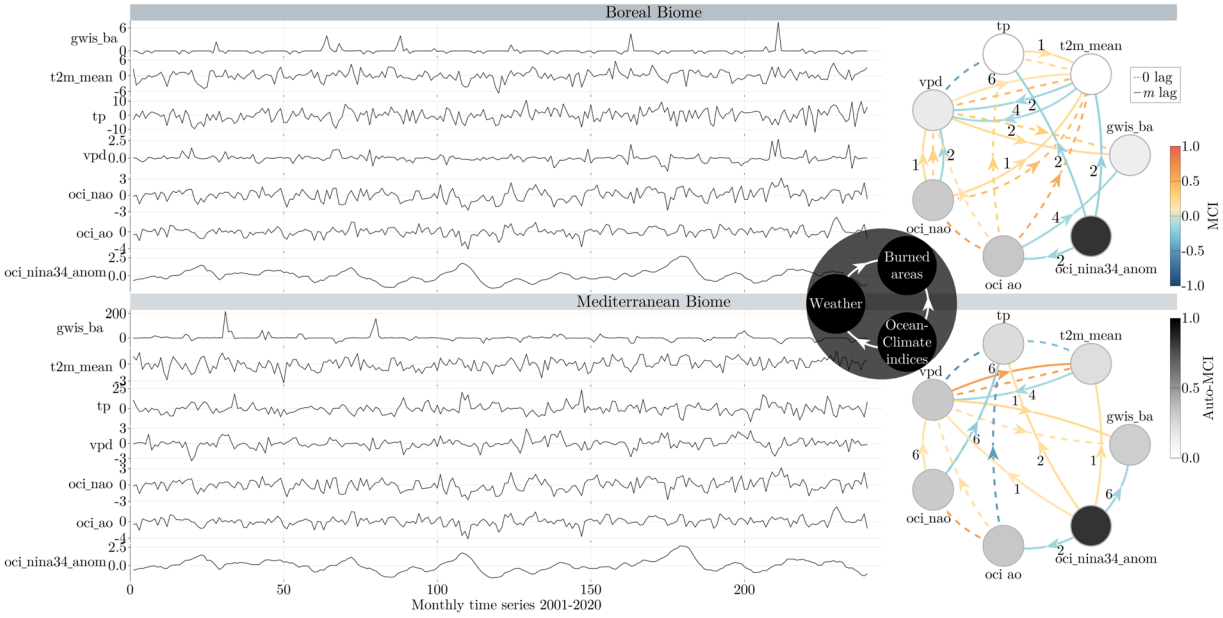

Fig. 6 Timeseries, causal order and PCMCI causal discovery graphs for the Mediterranean and Boreal forests. The color of each node indicates its self-correlation (the relationship with itself over time). In the causal networks, each node on the right side corresponds to a variable, as seen in the timeseries on the left. The color of the links indicates the partial correlation value between variables, which reveals the direction and strength of the inferred causal connection between them. For lagged connections, you can find the time delay (in months) indicated by small labels on the curved arrows, while dashed lines denote instantaneous causal connections occurring without any time delay. The illustrated key climatic and environmental variables include: oci_ao (Arctic oscillation), oci_nao (North Atlantic oscillation), oci_nina34_anom (El-Niño at 3.4 region), vpd (vapor pressure deficit), temp_mean (Mean air temperature at 2 meters), tp (total precipitation), and gwis_ba (GWIS burned areas).

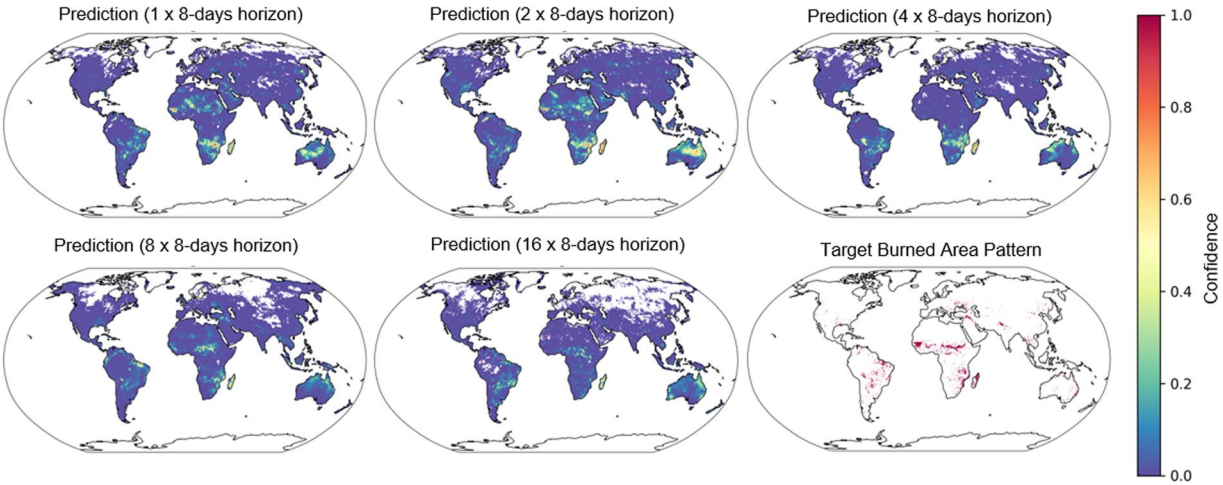

Fig. 7 Comparison between the prediction of the model and the target burned area pattern for the different forecasting horizons. The target for all the predictions comes from the datetime 2019-11-01 from the datacube, while the input for the different predictions is shifted backward by several 8-days indicated by the forecasting horizon. Confidence (softmax score of the positive prediction) bar indicates the model's level of certainty in predicting the occurence of a burned area.

ocean climate indices were considered causal for themselves, meteorology, and burned areas. Meteorology was deemed causal for itself and for burned areas only, and lastly, the burned areas are not considered causal for any of the variables.

The graphical analysis underscores the preeminent role of vapor pressure deficit timeseries, indicative of dryness[69], in influencing burned areas for both the Euro-Mediterranean and Euro-Boreal biomes. Notably, lags of 0 and 1-month exhibit significance for the Euro-Mediterranean, while lags of 0 and 2-months are influential for the Euro-Boreal biome. A closer look at the graphs reveals the following insights. The El-Niño (node Nino_3.4_anom) timeseries demonstrates a substantial self-memory (high Auto-MCI) of its past patterns, although this doesn't necessarily imply the predictability of the timeseries itself. As anticipated, a negative association between vapor pressure deficit (indicating dryness) and precipitation exists as contemporaneous links (lines without arrows), along with a positive correlation between vapor pressure deficit and temperature at 0 and 6-month lag time. Notably, the Arctic oscillation (oci_ao) and the North Atlantic oscillation (oci_nao) are strongly positively correlated in this context at 0-month lag.[70,71]. Lastly, concerning the lagged nonlinear causalities (lines with arrows- ParCorr test), across the Euro-Mediterranean region, Benassi *et al.*[72] confirm the outcomes of our analysis, wherein a positive El-Niño influence augments rainfall, while the same influence in boreal regions leads to negative temperature anomalies. As expected[71], Arctic Oscillation exhibits a leading role in negatively affecting precipitation at 0-month lag within Mediterranean forests, whereas it positively influences temperature contemporaneously in Boreal regions. These causal findings, in alignment with the existing literature, serve as compelling evidence that the timeseries data under examination are indeed reliable.

Usage Notes

The Seasfire datacube[12] is available for unlimited use under the Creative Commons License 4.0 International. We strongly recommend that users access the cube using multidimensional arrays, for instance, through Xarrays (Python) or YAXArrays.jl (Julia). Additionally, any customized Python or Julia code employed in the analyses of the SeasFire datacube[12] is accessible in the accompanying GitHub repository. Users are encouraged to utilize this code as a foundational resource for their individual analyses. Notebook examples are available for both programming languages (https://github.com/SeasFire/seasfire-datacube/tree/main). The datacube can be accessed both locally and in a cloud environment. The choice between these two options should be contingent upon the user's available computational resources. It is essential for users to verify that their local or cloud environment possesses adequate Random Access Memory (RAM) capacity to accommodate the processing requirements of the datacube. The specific RAM requirements may vary depending on the nature of the analyses performed, but as a general guideline, approximately 4 GB of RAM is needed for loading a single variable. The data within SeasFire has undergone preprocessing to a certain extent, which is well documented in the medatada of each variable. Nevertheless, users may find it necessary to conduct further data cleaning, normalization, or transformation procedures depending on their specific research questions.

Downstream Application: Machine Learning Modeling. When working with datasets containing historical data relevant to a particular phenomenon, machine learning can be leveraged to model connections between root causes and resulting outcomes.

In the process of translating data from a datacube into a machine learning task, several critical steps must be carefully navigated, which are elaborated upon as follows. Initially, one must clearly delineate the task's objectives and desired outcomes, while simultaneously devising a strategy for data sampling, mindful of the temporal and spatial aspects. A robust evaluation split is pivotal, as it determines how the dataset will be divided into

training, validation, and test subsets, steering clear of potential data leakage. Setting an appropriate baseline, often in the form of a simple model, provides a reference point for advanced model comparisons. Lastly, defining relevant evaluation metrics is crucial, ensuring that the chosen measures accurately reflect the task's real-world implications, thus paving the way for effective model development and evaluation. This approach is exemplified in Prapas *et al.*[18], where burned area pattern forecasting is framed as a segmentation task. Building on this foundation, we conducted a demo machine learning task using the same 8 variables from the datacube as input features and deriving the target from the GWIS burned area variable. To refine the analysis, the patch size was reduced to 80 × 80 pixels. Figure 7 illustrates the progression of the model's predictions across increasing forecast horizons (spanning 1 to 16 intervals of 8 days) and their alignment with the actual burned area patterns.

Other potential downstream applications. Machine learning holds immense potential across various fronts in advancing our understanding of wildfires. Firstly, by harnessing machine learning, we could improve the prediction of critical variables such as vegetation growth, drought conditions, and fire weather patterns on sub-seasonal to seasonal timescales, developing early warning systems and enabling more proactive wildfire mitigation efforts. Secondly, shifting from traditional segmentation to regression modeling provides an alternative avenue for comprehending the continuous relationships between variables, resulting in more precise predictions of wildfire occurrences. Moreover, machine learning can extend its reach to forecast wildfire emissions, offering a comprehensive assessment of their environmental impact, a vital aspect for evaluating their effects on air quality and the global carbon cycle. Additionally, by incorporating teleconnection indices to the modeling, we could gain deeper insights into the complex web of interactions between climatic factors and wildfires.

Limitations. The datacube's temporal and spatial resolution provides a detailed perspective that enhances its effectiveness in identifying intricate and detailed patterns and trends for global and regional analyses[19]. However, the inherent characteristics of the source data may limit its applicability in highly localised studies. The 0.25° spatial resolution means each grid cell covers an area of approximately 730 km2. While the 0.25° spatial resolution enables the identification of large-scale patterns and global trends, offering a broad view of environmental dynamics, it may fall short in representing localized variations, especially in areas where fine-scale topographic or land-use factors significantly influence wildfire behavior. The resolution may be too coarse to capture fine-scale variations in topography or land use that influence fire behavior at the level of individual communities or small watersheds. For instance, it might not distinguish between areas on opposite sides of a narrow mountain range that experience different microclimates. Microclimate modeling usually requires spatial resolutions of tens of meters or less and temporal resolutions of hours or less[73]. Similarly, the temporal resolution, characterized by an 8-day interval, may be insufficient for scenarios that require precise tracking of daily changes, which is often critical in fire spread dynamics.

Code availability

All code necessary for the technical validation and the machine learning demo is publicly available on GitHub (https://github.com/SeasFire/seasfire-datacube-paper). Code regarding working with the SeasFire datacube[12] in both Python and Julia, as well as for adding new variables, is available in the GitHub repository (https://github.com/SeasFire).

Received: 8 December 2023; Accepted: 28 January 2025;
Published online: 03 March 2025

References

1. McLauchlan, K. K. *et al.* Fire as a fundamental ecological process: Research advances and frontiers. *J. Ecol.* **108**, 2047–2069, https://doi.org/10.1111/1365-2745.13403 (2020).
2. United Nations Environment Programme. Number of wildfires to rise by 50% by 2100 and governments are not prepared, experts warn. https://www.unep.org/news-and-stories/press-release/number-wildfires-rise-50-2100-and-governments-are-not-prepared (2022).
3. Harper, A. R., Doerr, S. H., Santin, C., Froyd, C. A. & Sinnadurai, P. Prescribed fire and its impacts on ecosystem services in the uk. *Sci. The Total. Environ.* **624**, 691–703, https://doi.org/10.1016/j.scitotenv.2017.12.161 (2018).
4. Jolly, W. M. *et al.* Climate-induced variations in global wildfire danger from 1979 to 2013. *Nat. Commun.* **6**, https://doi.org/10.1038/ncomms8537 (2015).
5. Steel, Z. L., Campos, B. R., Frick, W. F., Burnett, R. D. & Safford, H. D. The effects of wildfire severity and pyrodiversity on bat occupancy and diversity in fire-suppressed forests. *Sci. Reports* **9**, https://doi.org/10.1038/s41598-019-52875-2 (2019).
6. Schoennagel, T. *et al.* Adapt to more wildfire in western north american forests as climate changes. *Proc. Natl. Acad. Sci.* **114**, 4582–4590, https://doi.org/10.1073/pnas.1617464114 (2017).
7. Cascio, W. E. Wildland fire smoke and human health. *Sci. Total. Environ.* **624**, 586–595, https://doi.org/10.1016/j.scitotenv.2017.12.086 (2018).
8. Liu, J. C., Pereira, G., Uhl, S. A., Bravo, M. A. & Bell, M. L. A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke. *Environ. Res.* **136**, 120–132, https://doi.org/10.1016/j.envres.2014.10.015 (2015).
9. Holm, S. M., Miller, M. D. & Balmes, J. R. Health effects of wildfire smoke in children and public health tools: a narrative review. *J. exposure science & environmental epidemiology* **31**, 1–20, https://doi.org/10.1038/s41370-020-00267-4 (2021).
10. Beranek, C. T. *et al.* Severe wildfires promoted by climate change negatively impact forest amphibian metacommunities. *Divers. Distributions* **29**, 785–800, https://doi.org/10.1111/ddi.13700 (2023).
11. Xu, Z., Zhang, Y., Blöschl, G. & Piao, S. Mega forest fires intensify flood magnitudes in southeast australia. *Geophys. Res. Lett.* **50**, https://doi.org/10.1029/2023gl103812 (2023).
12. Alonso, L. *et al.* SeasFire Cube: A Global Dataset for Seasonal Fire Modeling in the Earth System — zenodo.org. https://doi.org/10.5281/zenodo.13834057 [Accessed 27-09-2024] (2022).
13. Xiong, Z., Zhang, F., Wang, Y., Shi, Y. & Zhu, X. X. Earthnets: Empowering ai in earth observation, arXiv:2210.04936v3 (2022).

14. Giuliani, G., Masó, J., Mazzetti, P., Nativi, S. & Zabala, A. Paving the Way to Increased Interoperability of Earth Observations. *Data Cubes. Data* **4**, 113, https://doi.org/10.3390/data4030113 (2019).
15. Baumann, P., Misev, D., Merticariu, V. & Huu, B. P. Datacubes: Towards space/time analysis-ready data. In *Lecture Notes in Geoinformation and Cartography*, 269–299, https://doi.org/10.1007/978-3-319-72434-8_14 (Springer International Publishing, 2018).
16. Mahecha, M. D. *et al*. Earth system data cubes unravel global multivariate dynamics. *Earth Syst. Dyn.* **11**, 201–234, https://doi.org/10.5194/esd-11-201-2020 (2020).
17. Montero, D. *et al*. Earth System Datacubes: Avenues for advancing Earth System research. *Environmental Data Science*, **3**, e27, 1–31, https://doi.org/10.1017/eds.2024.22 (2024).
18. Prapas, I. *et al. Deep learning for global wildfire forecasting* https://doi.org/10.48550/ARXIV.2211.00534 (2022).
19. Prapas, I. *et al*. Televit: Teleconnection-driven transformers improve subseasonal to seasonal wildfire forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 3754–3759 (2023).
20. Wilkinson, M. D. *et al*. Addendum: The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **6** https://doi.org/10.1038/s41597-019-0009-6 (2019).
21. Kopp, S. *et al*. Achieving the full vision of earth observation data cubes. *Data* **4**, 94, https://doi.org/10.3390/data4030094 (2019).
22. Sudmanns, M., Augustin, H., van der Meer, L., Baraldi, A. & Tiede, D. The austrian semantic EO data cube infrastructure. *Remote. Sens.* **13**, 4807, https://doi.org/10.3390/rs13234807 (2021).
23. Ariza-Porras, C. *et al*. CDCol: A geoscience data cube that meets colombian needs. In *Communications in Computer and Information Science*, 87–99, https://doi.org/10.1007/978-3-319-66562-7_7 (Springer International Publishing, 2017).
24. Dhu, T. *et al*. Digital earth australia – unlocking new value from earth observation data. *Big Earth Data* **1**, 64–74, https://doi.org/10.1080/20964471.2017.1402490 (2017).
25. Lewis, A. *et al*. The australian geoscience data cube — foundations and lessons learned. *Remote. Sens. Environ.* **202**, 276–292, https://doi.org/10.1016/j.rse.2017.03.015 (2017).
26. Killough, B. Overview of the open data cube initiative. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, https://doi.org/10.1109/igarss.2018.8517694 (IEEE, 2018).
27. Giuliani, G. *et al*. Building an earth observations data cube: lessons learned from the swiss data cube (SDC) on generating analysis ready data (ARD). *Big Earth Data* **1**, 100–117, https://doi.org/10.1080/20964471.2017.1398903 (2017).
28. Lewis, A. *et al*. CEOS analysis ready data for land (CARD4l) overview. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, https://doi.org/10.1109/igarss.2018.8519255 (IEEE, 2018).
29. Estupinan-Suarez, L. M. *et al*. A regional earth system data lab for understanding ecosystem dynamics: An example from tropical south america. *Front. Earth Sci.* **9** https://doi.org/10.3389/feart.2021.613395 (2021).
30. Hoyer, S. & Hamman, J. xarray: N-D labeled arrays and datasets in Python. *prep, J. Open Res. Softw*. https://doi.org/10.5334/jors.148 (2016).
31. Kondylatos, S., Prapas, I., Camps-Valls, G. & Papoutsis, I. Mesogeos: A multi-purpose dataset for data-driven wildfire modeling in the mediterranean, https://doi.org/10.48550/ARXIV.2306.05144 (2023).
32. Kondylatos, S. *et al*. Wildfire Danger Prediction and Understanding With Deep Learn- ing. *Geophys. Res. Lett.* **49**, e2022GL099368 https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022GL099368 (2022).
33. Miles, A. *et al*. zarr-developers/zarr-python: v2.4.0, https://doi.org/10.5281/zenodo.3773450 (2020).
34. Driscoll, D. A. *et al*. How fire interacts with habitat loss and fragmentation. *Biol. Rev.* **96**, 976–998, https://doi.org/10.1111/brv.12687 (2021).
35. Schiavina, M., Melchiorri, M. & Pesaresi, M. GHS-SMOD R2023A - GHS settlement layers, application of the Degree of Urbanisation methodology (stage I) to GHS-POP R2023A and GHS-BUILT-S R2023A, multitemporal (1975–2030), https://doi.org/10.2905/A0DF7A6F-49DE-46EA-9BDE-563437A6E2BA (2023).
36. Amatulli, G. *et al*. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Sci. data* **5**, 1–15, https://doi.org/10.1594/PANGAEA.867115 (2018).
37. Artés, T. *et al*. A global wildfire dataset for the analysis of fire regimes and fire behaviour. *Sci. Data* **6**, 296, https://doi.org/10.1038/s41597-019-0312-2 (2019).
38. Randerson, J., Van Der Werf, G., Giglio, L., Collatz, G. & Kasibhatla, P. Global fire emissions database, version 4.1 (gfedv4). *ORNL DAAC* https://doi.org/10.3334/ORNLDAAC/1293 (2017).
39. Chuvieco, E., Pettinari, M. L., Lizundia-Loiola, J., Storm, T. & Padilla Parellada, M. Esa fire climate change initiative (fire cci): Modis fire cci burned area pixel product, version 5.1. *Centre for Environ. Data Analysis (CEDA)* https://doi.org/10.5285/58F00D8814064B79A0C49662AD3AF537 (2018).
40. Lizundia-Loiola, J., Franquesa, M., Khairoun, A. & Chuvieco, E. Global burned area mapping from sentinel-3 synergy and VIIRS active fires. *Remote. Sens. Environ.* **282**, 113298, https://doi.org/10.1016/j.rse.2022.113298 (2022).
41. Giglio, L., Justice, C., Boschetti, L. & Roy, D. Mcd64a1 modis/terra+aqua burned area monthly l3 global 500m sin grid v006, https://doi.org/10.5067/MODIS/MCD64A1.006 (2015).
42. Giglio, L., Loboda, T., Roy, D. P., Quayle, B. & Justice, C. O. An active-fire based burned area mapping algorithm for the MODIS sensor. *Remote. Sens. Environ.* **113**, 408–420, https://doi.org/10.1016/j.rse.2008.10.006 (2009).
43. Giglio, L., Schroeder, W. & Justice, C. O. The collection 6 MODIS active fire detection algorithm and fire products. *Remote. Sens. Environ.* **178**, 31–41, https://doi.org/10.1016/j.rse.2016.02.054 (2016).
44. Giglio, L., Boschetti, L., Roy, D. P., Humber, M. L. & Justice, C. O. The collection 6 MODIS burned area mapping algorithm and product. *Remote. Sens. Environ.* **217**, 72–85, https://doi.org/10.1016/j.rse.2018.08.005 (2018).
45. Turco, M., Herrera, S., Tourigny, E., Chuvieco, E. & Provenzale, A. A comparison of remotely-sensed and inventory datasets for burned area in Mediterranean europe. *International Journal of Applied Earth Observation and Geoinformation,***82**, 101887 https://doi.org/10.1016/j.jag.2019.05.020 (2019).
46. Valencia, G. M., Anaya, J. A., Velásquez, É. A., Ramo, R. & Caro-Lopera, F. J. About validation-comparison of burned area products. *Remote. Sens.* **12**, 3972, https://doi.org/10.3390/rs12233972 (2020).
47. Didan, K. MOD13C1 MODIS/Terra Vegetation Indices 16-Day L3 Global 0.05Deg CMG V006, https://doi.org/10.5067/MODIS/MOD13C1.006 Type: dataset (2015).
48. Copernicus Climate Change Service. Land cover classification gridded maps from 1992 to present derived from satellite observations, https://doi.org/10.24381/CDS.006F2C9A Type: dataset (2019).
49. Jordahl, K. *et al*. geopandas/geopandas: v0.8.1, https://doi.org/10.5281/zenodo.3946761 (2020).
50. Hersbach, H. *et al*. ERA5 hourly data on single levels from 1959 to present, https://doi.org/10.24381/cds.adbb2d47 Type: dataset (2018).
51. Wan, Z., Hook, S. & Hulley, G. MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V061, https://doi.org/10.5067/MODIS/MOD11A2.061 Type: dataset (2021).
52. Myneni, R., Knyazikhin, Y. & Park, T. MCD15A2H MODIS/Terra+Aqua Leaf Area Index/FPAR 8-day L4 Global 500m SIN Grid V006, https://doi.org/10.5067/MODIS/MCD15A2H.006 [Dataset] (2015).
53. Dinerstein, E. *et al*. An ecoregion-based approach to protecting half the terrestrial realm. *BioScience* **67**, 534–545, https://doi.org/10.1093/biosci/bix014 (2017).

54. Chuvieco, E., Pettinari, M. L., Lizundia Loiola, J., Storm, T. & Padilla Parellada, M. ESA Fire Climate Change Initiative (Fire_cci): MODIS Fire_cci Burned Area Grid product, version 5.1, https://doi.org/10.5285/3628CB2FDBA443588155E15DEE8E5352 Type: dataset (2019).
55. Arctic Oscillation (AO) [dataset]. Barnston, Anthony G. and Livezey, Robert E.: https://journals.ametsoc.org/view/journals/mwre/115/6/1520-0493_1987_115_1083_csapol_2_0_co_2.xml. Accessed 25.07.2023 from: https://psl.noaa.gov/data/timeseries/month/ (1987).
56. North Atlantic Oscillation Index (NAO) [dataset], Jones Phillip. D., Jonsson T, and Wheeler D.: https://rmets.onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0088(19971115)17:13%3C1433::AID-JOC203%3E3.0.CO;2-P. Accessed 25.07.2023 from: https://psl.noaa.gov/data/timeseries/month/ (1997).
57. East Atlantic (EA) [dataset]. Accessed 25.07.2023 from: https://psl.noaa.gov/data/timeseries/month/.
58. Global Mean Land/Ocean Temperature Index from NASA/GISS [dataset]. Smith Thomas M. *et al.*: https://journals.ametsoc.org/view/journals/clim/9/6/1520-0442_1996_00_1403_rohsst_2_0_co_2.xml. Accessed 25.07.2023 from: https://data.giss.nasa.gov/gistemp/tabledata_v3/GLB.Ts.txt (1996).
59. Niño 3.4 calculated from the HadISST1.1 dataset at NOAA/ESRL [dataset]. Rayner, N. A. et al.: https://www.metoffice.gov.uk/hadobs/hadisst/HadISST_paper.pdf. Accessed 25.07.2023 from: https://psl.noaa.gov/data/timeseries/month/ (2003).
60. West Pacific (WP) [dataset]. Barnston, Anthony G. and Livezey, Robert E.: https://journals.ametsoc.org/view/journals/mwre/115/6/1520-0493_1987_115_1083_csapol_2_0_co_2.xml. Accessed 25.07.2023 from: https://psl.noaa.gov/data/timeseries/month/ (1987).
61. East Pacific/North Pacific Oscillation (EP-NP) [dataset]. Bell, Gerald D. and Janowiak, John E.: https://journals.ametsoc.org/view/journals/bams/76/5/1520-0477_1995_076_0681_acawtm_2_0_co_2.xml. Accessed 25.07.2023 from: https://psl.noaa.gov/data/timeseries/month/ (1995).
62. Pacific Decadal Oscillation (PDO) [dataset]. Mantua, Nathan J. and Hare, Steven R. and Zhang, Yuan: https://journals.ametsoc.org/view/journals/bams/78/6/1520-0477_1997_078_1069_apicow_2_0_co_2.xml. Accessed 25.07.2023 from: https://psl.noaa.gov/data/timeseries/month/ (1997).
63. Pacific North American Index (PNA) [dataset]. Barnston, Anthony G. and Livezey, Robert E.: https://journals.ametsoc.org/view/journals/mwre/115/6/1520-0493_1987_115_1083_csapol_2_0_co_2.xml. Accessed 25.07.2023 from: https://psl.noaa.gov/data/timeseries/month/ (1987).
64. Danisch, S. & Krumbiegel, J. Makie.jl: Flexible high-performance data visualization for Julia. *J. Open Source Softw.* **6**, 3349, https://doi.org/10.21105/joss.03349 (2021).
65. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. & Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.* **5**, https://doi.org/10.1126/sciadv.aau4996 (2019).
66. Runge, J. *et al*. Inferring causation from time series in earth system sciences. *Nat. Commun*. **10**, https://doi.org/10.1038/s41467-019-10105-3 (2019).
67. Pearl, J. Causality, https://doi.org/10.1017/cbo9780511803161 (2009).
68. Spirtes, P., Glymour, C. & Scheines, R. Causation, prediction, and search, https://doi.org/10.7551/mitpress/1754.001.0001 (2001).
69. Li, S. *et al*. Increasing vapor pressure deficit accelerates land drying. *J. Hydrol.* **625**, 130062, https://doi.org/10.1016/j.jhydrol.2023.130062 (2023).
70. Almendra-Martín, L. *et al*. Influence of atmospheric patterns on soil moisture dynamics in Europe. *Sci. The Total. Environ.* **846**, 157537, https://doi.org/10.1016/j.scitotenv.2022.157537 (2022).
71. Climate.gov. Understanding Climate Variability: The North Atlantic Oscillation. https://www.climate.gov/news-features/understanding-climate/climate-variability-north-atlantic-oscillation Accessed 25.07.2023 (2009).
72. Benassi, M. *et al*. El niño teleconnection to the euro-mediterranean late-winter: the role of extratropical pacific modulation. *Clim. Dyn.* **58**, 2009–2029, https://doi.org/10.1007/s00382-021-05768-y (2021).
73. Wolf, K. D., Higuera, P. E., Davis, K. T. & Dobrowski, S. Z. Wildfire impacts on forest microclimate vary with biophysical context. *Ecosphere* **12**, e03467, https://doi.org/10.1002/ecs2.3467 (2021).
74. Earth Data, N. FIRMS Frequently Asked Questions | Earthdata — earthdata.nasa.gov. https://www.earthdata.nasa.gov/faq/firms-faq (2023).
75. Climate, E. Esa climate change initiative – fire CCI product user guide. https://climate.esa.int/media/documents/Fire_cciD4.2_PUG-MODIS_v1.0.pdf (2020).
76. Andela, N. *et al*. The global fire atlas of individual fire size, duration, speed and direction. *Earth Syst. Sci. Data* **11**, 529–552, https://doi.org/10.5194/essd-11-529-2019 (2019).
77. NIFC. Historic Perimeters Combined 2000-2018 GeoMAC — data-nifc.opendata.arcgis.com. https://data-nifc.opendata.arcgis.com/datasets/nifc::historic-perimeters-combined-2000-2018-geomac/about (2018).
78. National Interagency Fire Center. National Interagency Fire Center — data-nifc.opendata.arcgis.com. https://data-nifc.opendata.arcgis.com/search?q=geomac (2019).
79. CNFDB. Canadian Wildland Fire Information System | Canadian National Fire Database (CNFDB) — cwfis.cfs.nrcan.gc.ca. https://cwfis.cfs.nrcan.gc.ca/ha/nfdb (2022).
80. JRC-EFFIS. Welcome to EFFIS. https://effis.jrc.ec.europa.eu/ (2008).
81. Government of Australia. Product catalogue - Geoscience Australia — ecat.ga.gov.au. https://ecat.ga.gov.au/geonetwork/srv/eng/catalog.search#/metadata/147763 (2023).
82. Hagg, W. & Cankaya, K. ArcGIS Australia wildfire Dashboards — arcgis.com. https://www.arcgis.com/apps/dashboards/11e177420af74b1b8c5ecd59ae3a85a4 (2020).
83. Short, K. C. Spatial wildfire occurrence data for the united states, 1992-2015 [fpa_fod_20170508]. https://doi.org/10.2737/RDS-2013-0009.4 (2017).
84. Lever, J. & Arcucci, R. Sentimental wildfire: a social-physics machine learning model for wildfire nowcasting. *J. Comput. Soc. Sci.* **5**, 1427–1465, https://doi.org/10.1007/s42001-022-00174-8 (2022).
85. Denis, L. A. S. *et al*. All-hazards dataset mined from the US national incident management system 1999–2020. *Sci. Data* **10**, https://doi.org/10.1038/s41597-023-01955-0 (2023).
86. Singla, S. *et al*. Wildfiredb: An open-source dataset connecting wildfire occurrence with relevant determinants. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks* (2021).
87. Graff, C. A. Fire-ml: A remotely-sensed daily wildfire forecasting dataset for the contiguous united states. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning* (2021).
88. CMES. Fire danger indices historical data from the Copernicus Emergency Management Service, https://doi.org/10.24381/CDS.0E89C522 Type: dataset (2019).
89. Kaiser, J. W. *et al*. Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power. *Biogeosciences* **9**, 527–554, https://doi.org/10.5194/bg-9-527-2012 (2012).
90. National Oceanic and Atmospheric Administration. Climate indices. https://psl.noaa.gov/data/climateindices/. [Accessed 27-09-2024].
91. Center for International Earth Science Information Network (CIESIN), C. U. Gridded population of the world, version 4 (gpwv4): Population density, revision 11, https://doi.org/10.7927/H49C6VHW (2018).

Acknowledgements
This work has been funded by the European Space Agency (ESA) SeasFire project under the ESA Future EO-1 Science for Society Call, supported by ESA's Network of Resources Initiative, and by the Horizon Europe MeDiTwin project (Grant Agreement No. 101159723).

Author contributions
Conceptualization: I.Prapas, N.C., I.Papoutsis; Data curation: I.K., L.A.; Implementation: I.K., L.A., I.Prapas; Technical Validation: I.K., L.A., I.Prapas; Data visualization: I.K., L.A.; Writing - Original Draft: I.K., A.A.; Writing - Review, and Editing: All authors; Supervision: I.Prapas, N.C., I.Papoutsis; Funding acquisition: N.C., I.Papoutsis.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to I.K. or L.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.