scientific **data**

Check for updates

DeepExtremeCubes: Earth system spatio-temporal data for assessing compound heatwave and drought impacts

Chaonan Ji [1,2 ✉], Tonio Fincke[3], Vitus Benson [4,5], Gustau Camps-Valls [6], Miguel-Ángel Fernández-Torres [6], Fabian Gans[4], Guido Kraemer [1,2], Francesco Martinuzzi [1,2,7], David Montero [1,2,8], Karin Mora [1,2], Oscar J. Pellicer-Valero[6], Claire Robin[4,5], Maximilian Söchting [1,2,9], Mélanie Weynants[4] & Miguel D. Mahecha [1,2,8,10 ✉]

With climate extremes' rising frequency and intensity, robust analytical tools are crucial to predict their impacts on terrestrial ecosystems. Machine learning techniques show promise but require well-structured, high-quality, and curated analysis-ready datasets. Earth observation datasets comprehensively monitor ecosystem dynamics and responses to climatic extremes, yet the data complexity can challenge the effectiveness of machine learning models. Despite recent progress in deep learning to ecosystem monitoring, there is a need for datasets specifically designed to analyse compound heatwave and drought extreme impact. Here, we introduce the *DeepExtremeCubes* database, tailored to map around these extremes, focusing on persistent natural vegetation. It comprises over 40,000 globally sampled small data cubes (i.e. minicubes), with a spatial coverage of 2.5 by 2.5 km. Each minicube includes (i) Sentinel-2 L2A images, (ii) ERA5-Land variables and generated extreme event cube covering 2016 to 2022, and (iii) ancillary land cover and topography maps. The paper aims to (1) streamline data accessibility, structuring, pre-processing, and enhance scientific reproducibility, and (2) facilitate biosphere dynamics forecasting in response to compound extremes.

Background & Summary

There has been an unprecedented rise in the frequency and severity of climate extremes[1]. These rising extremes can have severe ecological[2] and socio-economic consequences[3], challenging our established paradigms of climate science[4]. For instance, in 2018, central and northern Europe experienced a record-breaking Compound Heatwave and Drought (CHD) event, which extensively impacted agriculture, forests, water supply, and the socio-economic sector[5]. Given the increasing intensity and adverse impacts of CHD events in the warming climate, it is critical to understand their intricate dynamics and interactions with climate drivers, spatial conditions, timing, and terrestrial ecosystems[6–9].

The exponential increase in Earth observation data represents a significant advancement but also introduces complex data management and analysis challenges[10,11]. In an era marked by rapid advances in remote sensing capabilities, including satellite observations, aerial imaging, and ground-based records, researchers have access to unprecedented amounts of information. These data are crucial for understanding the impacts of climate

[1]Remote Sensing Centre for Earth System Research (RSC4Earth), Leipzig University, Leipzig, 04103, Germany. [2]Institute for Earth System Science and Remote Sensing, Leipzig University, Leipzig, 04103, Germany. [3]Brockmann Consult GmbH, Hamburg, 21029, Germany. [4]Max Planck Institute for Biogeochemistry, Jena, 07745, Germany. [5]ELLIS Unit Jena, Jena, 07745, Germany. [6]Image Processing Laboratory (IPL), Universitat de València, València, 46980, Spain. [7]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Leipzig, 04105, Germany. [8]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, 04103, Germany. [9]Image and Signal Processing Group, Leipzig University, Leipzig, 04109, Germany. [10]Helmholtz Centre for Environmental Research – UFZ, Leipzig, 04318, Germany. ✉e-mail: chaonan.ji@uni-leipzig.de; miguel.mahecha@uni-leipzig.de

extremes[12–15]. Effective sampling strategies are required to harness this data deluge, ensuring relevance and manageability. Data cubes provide a flexible and efficient way to organise and analyse large volumes of multidimensional data, making such datasets manageable and streamlined across variables and spatio-temporal scales[10,16].

Machine Learning (ML) has been introduced into climate science as a valuable tool to understand and predict climate extremes and their impacts, as well as to decipher the interactions between climate and ecosystems[17–22]. Moreover, Deep Learning (DL) allows the identification of complex patterns and correlations that might elude traditional data science methods, thereby helping scientists to better understand the underlying mechanisms of climate variability and change. However, since ML generally performs best with large sample sizes, extreme impact prediction often has significantly smaller sample sizes compared to non-extreme conditions, which complicates the application of ML. In this dataset, we tackle this issue by oversampling extreme areas using the minicube strategy, which has a large distribution in space instead of time. This method introduces additional biases, as ML tends to amplify them. Nonetheless, this trade-off is necessary and must be considered when training models.

The sophisticated Earth observation databases that train ML models for analysing climate extremes are growing. These datasets primarily focus on addressing the scarcity of curated data concerning complex weather patterns and climate extremes' impacts on ecosystems. For instance, The ExtremeWeather dataset[23] provides labelled extreme weather events (i.e., tropical depression, tropical cyclone, extratropical cyclone, atmospheric river) as boxes, along with climatic and meteorological variables on a global grid of 768 by 1152. This dataset allows training ML models to leverage spatial and temporal information to predict the localisation of extreme weather events. ClimateNet[24] provides an expert-labelled dataset that enables pixel-level identification of extreme events using ML models. Additionally, cross-domain and high-resolution datasets are designed to include localised variables critical for analysing responses to climate extremes, incorporating data from diverse domains. For example, EarthNet2021[25] aims to bridge the data gap by integrating a variety of data variables such as precipitation, temperature, sea-level pressure, digital elevation models, and Sentinel-2 Multi-Spectral Instrument (MSI) images, offering a holistic view of Earth system. A model trained on EarthNet2021 can forecast optical satellite images of high perceptual quality. The newly enhanced version, GreenEarthNet[26], focuses more on predicting vegetation dynamics and includes an improved high-quality cloud mask[27]. The FluxnetEO data cubes[28] provide fully gap-filled Nadir BRDF Adjusted Reflectance (NBAR) data from MODIS, as well as Land Surface Temperature (LST) and several vegetation indices for the Fluxnet sites[29], aiming for modelling carbon and water fluxes. Moreover, DynamicEarthNet[30] tracks daily land use and land cover changes across 75 global regions from 2018 to 2019, focusing on detecting land cover changes. BigEarthNet[31] is a large-scale benchmark dataset consisting of Sentinel-2 satellite images with multi-label land use and land cover. Presto's Training Dataset[32] is a high-resolution dataset that provides detailed data for training ML models to significantly improve the prediction and understanding of climate extremes and their impacts.

However, these multi-purpose initiative datasets do not focus specifically on the impact of CHD extremes. We need harmonised datasets tailored for spatio-temporal ML methodologies, aiming to train ML methods to forecast and explain the impacts of extreme events such as droughts and heatwaves. Given the importance of CHD extremes and challenges arising from data biases and their repercussions, this paper is poised to propose a solution that encapsulates precision and reproducibility. Here, we present the *DeepExtremeCubes* dataset[33], a collection of minicubes that use a sampling methodology to focus on capturing the impact of CHD extremes globally. Specifically, we introduce 1) a globally stratified sampling procedure, 2) a reproducible data processing pipeline combining multi-modal data, and 3) a representative global dataset to train ML models on CHD extremes, which is analysis-ready and shared in cloud-native format.

Methods

The analysis of CHD extremes necessitates examining a broad range of Earth observation variables across climatic, meteorological, ecological, and topographical dimensions at various spatial and temporal scales[10]. Sampling these relevant datasets is crucial to focus on CHD impacts and to understand the complex interactions of different drivers, spatial conditions, and timing of these processes. The *DeepExtremeCubes* dataset[33] employs sampled minicubes targeted at regions experiencing extreme CHD events and their surroundings, facilitating a more detailed investigation. From a practical viewpoint, managing the vast, high-dimensional Earth system datasets requires significant computational resources. Segmenting these datasets into smaller, manageable subsets (i.e. minicubes) can enhance machine learning computations efficiency[11].

Input data sources. Two categories of input data sources are used to create *DeepExtremeCubes*[33]. One is the reference dataset used to determine the strata. This encompasses the Dry and hot extreme events database (Dheed dataset, a predefined global dataset of CHD extreme events)[34,35] and the European Space Agency (ESA) Climate Change Initiative (CCI) land cover map[36]. The other category of data sources consists of the comprehensive Earth system datasets from which the data included in individual minicubes is extracted. We first introduce the reference datasets and then provide details on the comprehensive datasets within the generated minicubes.

Dheed event detection dataset. Dheed[34,35] is a database of labelled CHD events utilising daily aggregated ERA5 reanalysis data[37]. A spatially and temporally example piece of a Dheed dataset is shown in Fig. 1. The maximum daily Temperature at 2 m (Tmax) is used to detect heatwaves, and the daily differences of Precipitation and Evapotranspiration (PE) averaged over 30, 90, and 180 days (PE30, PE90, PE180) are employed to detect droughts[34,35]. The CHD events in Dheed are defined as groups of spatio-temporally connected daily extremes. Extremely hot and dry days are defined as those when (1) the maximum temperature exceeds the 99th percentile for 1950-2021 at that location, and (2) at least one of three drought indicators (Precipitation- Evapotranspiration averaged over the last 30, 90, and 180 days) falls below the 1st percentile. The Dheed database covers a time
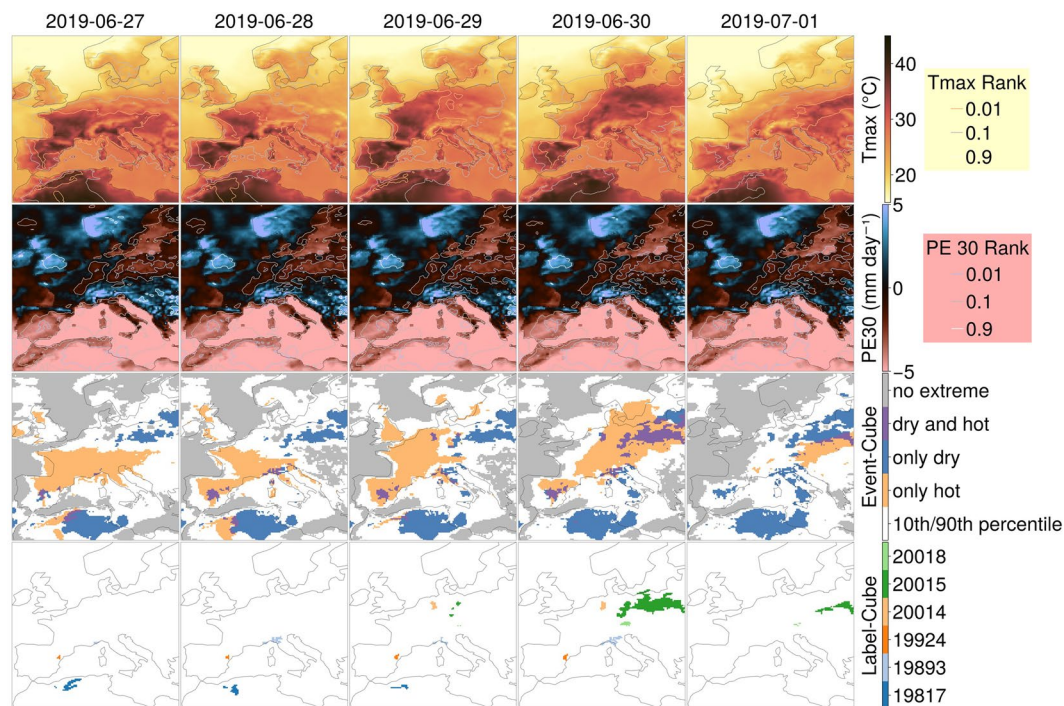
Fig. 1 An example of CHD detection in the Dheed database showing the evolution across Europe of the maximum daily temperature (Tmax, top row), the Precipitation - Evapotranspiration balance averaged over the previous 30 days (PE30, second row), along with the threshold of 0.01 on the ranked values used to detect extremes (Rank, first and second rows), the synthesis of the four indicators (Event-Cube, third row) and the labelled CHD events lasting at least three days (Label-Cube, fourth row) from 2019-06-27 to 2019-07-01.

range from 2016-01-01 to 2021-12-31, with a spatial resolution of 0.25°. Groups of spatio-temporal grid cells with extreme values connected across space and/or time are each assigned a unique event label. Dheed's labelled events have been benchmarked against extreme events documented in the literature or the media. The Dheed Label-Cube containing the labelled events (last row in Fig. 1) is used for sampling the locations of the minicubes.

Land cover map. The ESA CCI land cover dataset[36] employs the GlobCover unsupervised classification chain framework to generate global annual land use maps from 1992 to 2020[38]. It uses a combination of multi-year and multi-sensor strategies, incorporating data from various satellites such as ENVISAT-MERIS (2003-2012)[39], AVHRR (1992-1999)[40], SPOT-Vegetation (1999-2013)[41], and PROBA-Vegetation (2013-2020)[42,43]. The dataset categorises 37 land cover classes according to the United Nations Land Cover Classification System[44] and offers the data at a 300 m spatial resolution in GeoTIFF and NetCDF formats. The selection criteria for a reliable land cover map aim to best meet the requirements to analyse vegetation responses to CHD extremes. First, the map must cover most of the study period from 2016 to 2022, ensuring data continuity to reflect land cover changes. Second, data must be readily accessible for direct download and use. Third, a detailed classification of vegetation types is crucial, particularly focusing on persistent vegetation covers such as broad-leaved trees, needle-leaved trees, and grassland. While some vegetation classes were merged to simplify the initial sampling process, it was essential to retain the original and finer classifications in the minicubes for subsequent analyses. Various global land cover maps were evaluated[45–47], but none met these criteria as well as the ESA CCI WorldCover map. For instance, the Global Land Analysis and Discovery (GLAD) laboratory's Land Cover and Land Use Change (LCLUC) data[45], based on Landsat[48], offers high accuracy in certain non-forest regions and detailed classifications of open canopy forests in Africa. However, its infrequent updates (every five years) during 2000 and 2020 and limited vegetation classification hinder its suitability as the reference land cover map for the *DeepExtremeCubes* dataset[33]. The ESA CCI WorldCover map[47] offers high-resolution data for 2020 and 2021. However, it does not cover the entire study period and lacks comprehensive tree-type classifications. Therefore it is not suitable for the detailed vegetation analysis required for this study.

Data sources within each minicube. In addition to subsets of the Dheed dataset and the CCI land cover map, each minicube contains multiple data modalities: (1) Sentinel-2 MSI surface reflectance (L2A) time series data[49], (2) a corresponding deep-learning-based cloud mask[26,27], (3) ERA5-Land meteorological reanalysis variables[50], and (4) data from the Copernicus Digital Elevation Model (DEM)[51]. The ERA5-Land reanalysis data provides information on the historical weather conditions, represented by variables such as temperature, humidity, soil moisture, and others. Sentinel-2 MSI satellite images are a proxy observation for vegetation health. We include bands B02, B03, B04, B05, B06, B07, and B8A, which can be used to compute vegetation indices[52,53]. Together, the ERA5-Land and Sentinel-2 data allow us to study the impact of CHD extremes on vegetation
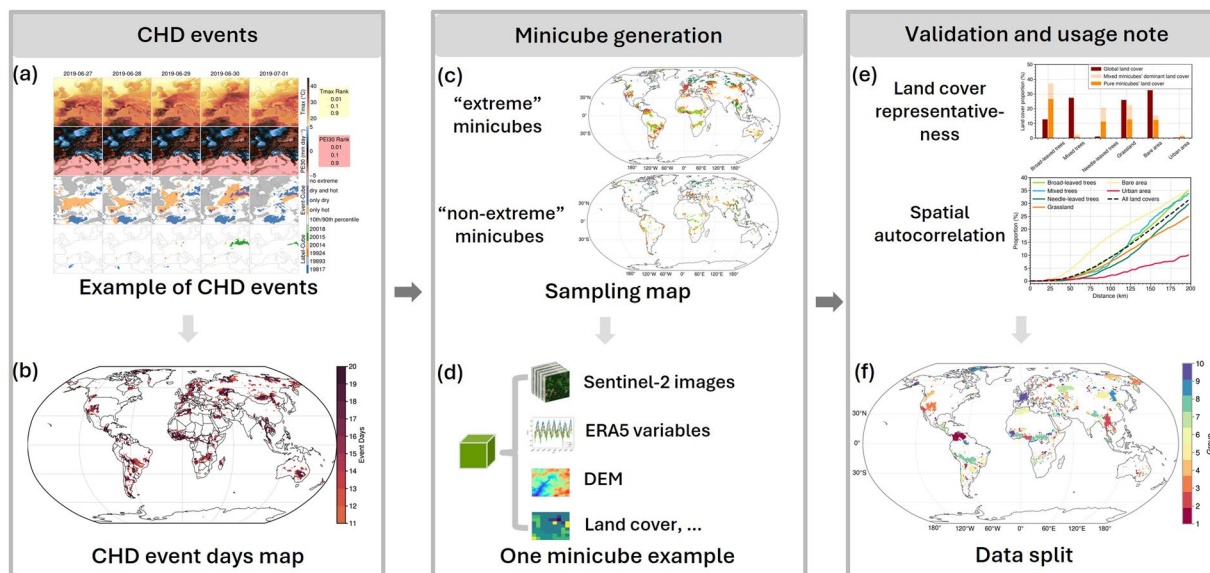
Fig. 2 A schematic diagram illustrating the *DeepExtremeCubes* database[33] development and validation workflow. It includes (**a**) The Dheed dataset presenting an example of a CHD event in this data. (**b**) The CHD event days map shows locations that experienced ten or more event days according to the Dheed event detection dataset. (**c**) Minicubes that experienced ten or more CHD days ("extreme" minicubes) and those that did not experience any CHD days in the Dheed dataset ("non-extreme" minicubes). (**d**) A demonstration of a minicube, which includes remote sensing images, climatic and meteorological variables, Digital Elevation Model (DEM), land cover, etc. (**e**) The representativeness and spatial autocorrelation of the *DeepExtremeCubes* dataset[33] for minicubes with different land cover types. (**f**) The spatial data split of *DeepExtremeCubes* for potential users to train their forecasting models.

in the *DeepExtremeCubes* dataset[33]. In Sentinel-2 images, pixels obscured by clouds and cloud shadows can be difficult to distinguish from actual changes in the underlying ecosystem, which may challenge subsequent analyses' accuracy. By incorporating the EarthNet Cloud Cover Mask[26], which is based on the CloudSEN12 dataset[27], obscured pixels can be filtered out, ensuring that vegetation biological dynamics are based on clear and reliable optical remote sensing data. In addition, the Copernicus DEM data is included as one of the key factors in climate-vegetation interactions. The Copernicus DEM provides topographical data at 30 m, enabling us to consider how elevation influences local climate conditions, subsurface hydrology and vegetation patterns. It is crucial in regions where elevation varies significantly and is also important on a global scale. For our minicubes, Sentinel-2 and the cloud mask are spatio-temporal arrays, ERA5-Land is included as a single-pixel time series (temporal array), and the DEM is a static image (spatial array).

Approach. We incorporated comprehensive data sources to develop the *DeepExtremeCubes* dataset[33], which comprises minicubes with a spatial size of 2.5 km by 2.5 km, covering the period from 2016 to 2022. The schematic approach is shown in Fig. 2. The Dheed dataset was used to generate a CHD event days map to determine the sampling locations for the minicubes. Subsequently, the *DeepExtremeCubes* minicubes were created with various variables. Additionally, we prepared a spatial data split strategy for subsequent users to train their ML forecasting models.

Sampling locations for minicubes. The sampling of minicube locations began with identifying areas frequently experiencing CHD extremes. We sampled from both areas affected by extremes and surrounding areas with similar land covers ("extreme" and "non-extreme" locations). This approach allows ML models to learn from both CHD-impacted and non-impacted instances, enhancing prediction accuracy and facilitating accurate estimation of carbon sequestration loss in CHD extremes. Additionally, we adjusted the locations based on land cover, with particular emphasis on persistent vegetation land covers.

The Dheed Label-Cube was compressed to a CHD event days map, marking all pixels that experienced ten or more CHD event days during 2016 and 2021. Due to the large size of the Dheed dataset, it could not be processed quickly in the following steps. Therefore, we aggregated the temporal dimension and generated the CHD event days map (see Fig. 2(b)). All pixels that experienced ten or more CHD event days were marked as potential central sampling locations. Around 80% of the "extreme" minicubes were located in heavily impacted areas (i.e., areas marked in the event days map), while roughly 20% "non-extreme" minicubes were situated in the vicinity of "extreme" areas and did not experience any CHD events (i.e., areas with zero event days). This step serves two main purposes: first, it enables the models to learn from CHD-impacted and non-impacted instances, enhancing predictive accuracy. Second, it allows for more accurate computation of carbon sequestration losses within regions covered by minicubes by comparing paired minicubes from both "extreme" and "non-extreme" conditions. To avoid spatial autocorrelation at very close distances in random sampling within "extreme" areas, where
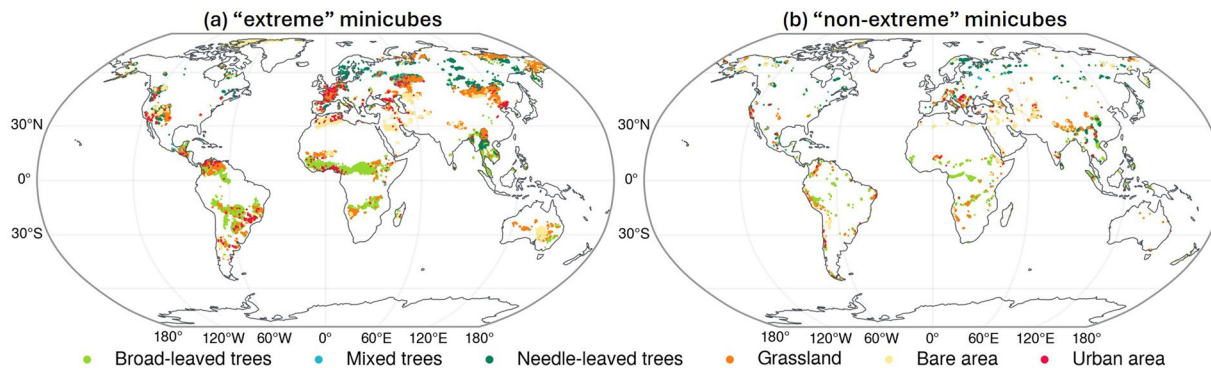
Fig. 3 The distribution of all sampled minicube locations. (**a**) Minicubes that experienced ten or more CHD days ("extreme" minicubes) and (**b**) minicubes that did not experience CHD events during 2016 and 2021 ("non-extreme" minicubes). The land cover types depicted represent either the pure land cover type for the pure land cover minicubes or the dominant land cover for the minicubes covered by mixed land covers.

sampled locations tend to cluster, we maintained a spatial grid of 0.125° (half of the Dheed dataset's resolution). We selected no more than one sample per grid cell for each land cover type.

We defined a set of target vegetation types that best summarise all the persistent land covers while keeping a focus on vegetation. In this study, we define "persistent land covers" as land covers that show minimal changes in their spatial extent or composition during the dataset period (2016-2022). This is crucial as stable land covers are less affected by anthropogenic activities, thereby minimising confounding effects from these changes. This stability is essential as the primary objective of the derived dataset is to assess vegetation responses to CHD events. In this sampling step, we merged most of the land covers to simplify the sampling categories (see details in Table s1). These merged vegetation classes include broad-leaved trees, needle-leaved trees, mixed trees, and grassland. To enhance the diversity of land covers and examine prediction accuracy through comparisons between vegetation land covers and other persistent land covers, we also included bare area and urban area. Thereby, we focused on six land cover types (i.e. broad-leaved trees, needle-leaved trees, mixed trees, grassland, bare area, and urban area) in total. In addition, we assessed the purity of land covers within defined minicubes to evaluate the varying behaviours of the ML prediction model concerning pure versus mixed land covers. Given the ESA CCI land cover map's resolution of 300 meters per pixel, a minicube (2.5 km by 2.5 km) encompasses approximately 9×9 pixels. We established a spatial window of 81 pixels (9×9), centred on the central pixel, to determine the purity of each minicube's land cover. If 65 or more pixels within this window exhibit the same land cover (equivalent to an 8×8 pixel area, about 80% of the spatial coverage of a minicube), the central pixel is classified under "pure land cover" and is eligible to be the central point of a minicube. Conversely, if fewer than 65 pixels share a single cover type, the central pixel is considered to have "mixed land cover." Given the importance of land cover purity in one minicube for ML prediction models, we set a lower threshold such that the central pixel must display at least 50% similarity (40.5 pixels in a 9×9 matrix). If a land cover meets this threshold range of 50%–79%, the central pixel is still considered a potential sampling location and is marked with this land cover as the dominant land cover. We also list the second dominant land cover by tallying the remaining land cover classes and selecting the most frequent. Figure S1 presents the detailed distribution of (a) the pure land cover map and the mixed land cover maps, including (b) the dominant land cover map and (c) the secondary land cover map.

In summary, the minicube location sampling is based on two factors: whether the area is impacted by CHD extremes, and is purely or mixed covered by the target land covers. First, the minicube location sampling primarily focuses on selecting areas affected by CHD extremes, specifically only those with more than ten event days. From this set 80% of all mincubes are chosen. The other 20% is selected from areas surrounding these extremes, which did not experience CHD extremes. We want to distinguish between minicubes covered by pure or mixed land covers, to facilitate exploration of land cover purity for prediction. This categorisation considers pure land cover (about 80%–100% covered by one land cover) and mixed land cover (about 50%–79% covered by one land cover). When a minicube has mixed coverage, its second land cover class is provided to offer additional contextual information. The results of the minicube sampling are shown in Fig. 3. The samples' latitude and longitude coordinates indicate the central location of each minicube.

Minicube Generation and variables. We generated minicubes given the previously established central locations. A minicube is a dataset covering an area of 2.56 km by 2.56 km around the central location, ranging from 2016 to 2022, incorporating elements from various data sources (as listed in Table 1). To minimise the distortion of the original data, we opted to maintain separate spatial and temporal resolutions, which can be harmonised during subsequent processing if necessary. We collected data from various sources during the generation process and documented details about processing steps in configuration files. We consolidated the variables into a single dataset[54] and stored them in the Zarr format. The generation process comprised two stages: initially, we created a "base-minicube" composed exclusively of Sentinel-2 data. In the second stage, this "base-minicube" was updated with the remaining components. By adopting this strategy, we reduce the frequency of Sentinel-2 data pooling, which is the most resource-intensive part of the process. Additionally, we generate corresponding configuration files for these two phases. The final collection consists of 42,477 minicubes.

Data	Variable	Spatial size/pixel	Spatial resolution/m	Temporal extent	Temporal resolution/day
CCI land cover map	Land cover class	9 * 9	300	2016	—
Copernicus DEM	DEM	128 * 128	20	—	—
Sentinel-2 L2A	Bands 2-8A, Scene Classification Layer	128 * 128	20	2016.01.01 - 2022.10.10	5
EarthNet cloud mask	Cloud mask	128 * 128	20	2016.01.01 - 2022.10.10	5
ERA5-Land data	Min, max, and mean of selected variables	1 * 1	—	2016.01.01 - 2022.10.10	5
Event Data	Event codes and labels	1 * 1	—	2016.01.01 - 2021.12.31	1

Table 1. The components of a minicube.

Attributes	Description
mc_id	Identifier of the minicube within the *DeepExtremeCubes* system with latitude, longitude, generation version, and generation date.
path	Path to the minicube file in the storage system.
location_source	Source file for the event location data.
location_id	Geographical identifier associated with the latitude and longitude of the minicube.
version	The generation version of the minicube
type	Type of minicube. Either "full", "backup", or "base", where "full" means that the minicube has all required properties, "backup" refers to cubes that were created not based on the determined events and "base" means they are missing properties.
geometry	Geometric boundary of the minicube.
creation_date	Date and time when the minicube was created.
modification_date	Date and time when the minicube was last modified.
events	IDs of detected compound events of heatwaves and droughts were recorded at the minicubes location, along with each respective event's start and end times.
class	The type of land cover in case the minicube occupies ≥80% of one land cover. None if the minicube has mixed land cover.
dominant_class	The most frequent land cover in case it covers between 50% and 80%. None for pure land cover cubes.
second_dominant_class	The second most frequent land cover in case of a mixed land cover minicube. None for pure land cover cubes.
s2_l2_bands	Version of the Sentinel-2 reflectance data in the minicube.
ERA5-Land	Version of the ERA5-Land climate reanalysis data in the minicube.
cci_landcover_map	Version of the CCI land cover data in the minicube.
copernicus_dem	Version of the Copernicus DEM data in the minicube.
de_africa_climatology	Version of climatology data specific to Africa in the minicube (applicable only to some cubes).
event_arrays	Version of data of recorded events in the minicube.
s2cloudless_cloudmask	Version of cloud masking data generated using the S2Cloudless model (where applicable).
sen2cor_cloudmask	Version of cloud masking data generated using the Sen2Cor tool (where applicable).
unetmobv2_cloudmask	Version of cloud masking data processed using the UnetMobV2 model (where applicable).
continent	Continent where the minicube is located.
remarks	Additional notes.

Table 2. Summary of minicube attributes in the *DeepExtremeCubes* dataset.

Data Records

The DeepExtremeCubes dataset[33,55] is accessible via OPARA, with maintained versions available at zenodo, with this section being the primary source of information on the availability and content of the data being described. The database includes the *DeepExtremeCubes* minicubes, a registry table providing all attributes within each minicube, and a demonstration Jupyter notebook to help explore the *DeepExtremeCubes* minicubes. Table 2 lists all attributes in the registry table. Users can use it to find minicubes based on specific criteria, such as spatial extent, components, land classes, labelled extreme events or located continent.

Technical Validation

Land cover representation. To validate the land cover representation in the *DeepExtremeCubes* dataset, we compared it with those of the global land cover (see Fig. 4). Our analysis reveals an overrepresentation of broad-leaved and needle-leaved trees in minicubes compared to the global distribution. Conversely, grassland and urban area align proportionally with the global distribution, while mixed trees samples are underrepresented. This underrepresentation is the result of our method, which defines the dominant land cover in each minicube as its primary land cover, and mixed trees often coincide with either broad-leaved or needle-leaved trees. From the ML perspective, having a larger number of minicubes covered by broad-leaved or needle-leaved trees, as opposed to mixed trees, enhances prediction accuracy.
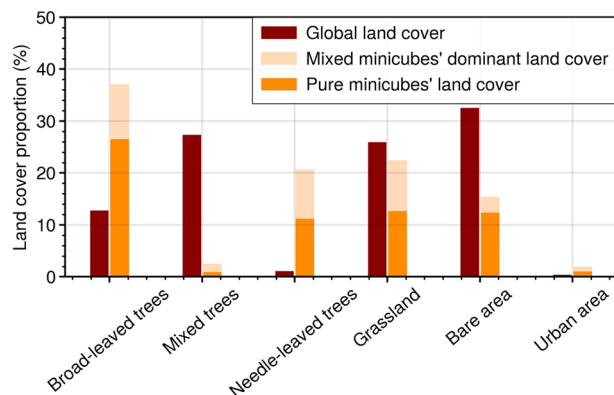
Fig. 4 Land cover representation in the *DeepExtremeCubes* dataset. The two plotted datasets use different reference points. The red bars consider the global areas of selected land cover as 100%, while the light and heavy orange bars treat the total sum of land cover types across all minicubes as 100%.
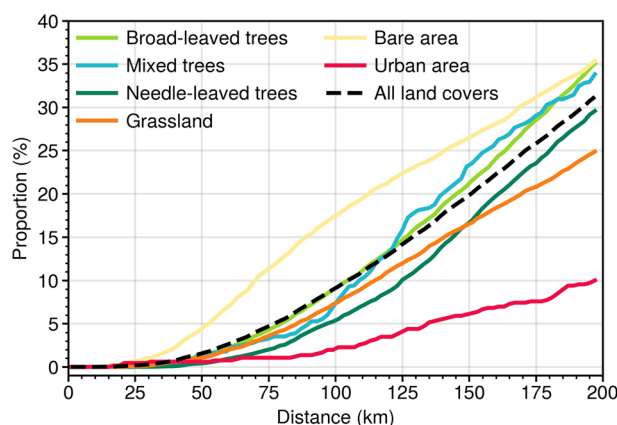

Fig. 5 The proportion within a certain distance where an "extreme" minicube can find a "non-extreme" minicube covered by the same land cover.

Spatial analysis of minicube distribution. To assess the impact of our stratified sampling on the spatial distribution of minicube locations, we computed the spatial autocorrelation of their extreme event occurrences. The analysis yielded a Moran's I value of 0.89, with a p-value of 0.001 and a z-score of 300.48. The high Moran's I value (0.89) reveals a significant and strong positive spatial autocorrelation among minicube locations. This suggests that locations with similar event statuses ("extreme" or "non-extreme") are clustered together. The statistical significance of this clustering is confirmed by the p-value of 0.001 and the high z-score (300.48), indicating that the observed spatial pattern is highly unlikely to be the result of random chance. Additionally, the proximity of "extreme" minicubes to "non-extreme" minicubes demonstrates a correlation in our dataset due to the sampling strategy. This effective sampling approach ensures that "non-extreme" samples are geographically close to "extreme" samples, confirming that our dataset's spatial correlation is preserved.

We analysed the shortest distance along which an "extreme" minicube can find a "non-extreme" minicube with the same land cover and computed the proportion of these minicubes in the total "extreme" minicubes along that distance. It maintains comparable environmental and vegetation characteristics while differing only in the impact of the CHD event on the "extreme" minicube. The results indicate that at a surrounding distance of 200 km, approximately 25% to 35% of extreme minicubes of each land cover type (excluding urban areas) can find a non-extreme minicube with the same land cover as a reference. Urban areas have a 10% probability at this distance, which is attributed to their low coverage in both the global map and minicube coverage (see Fig. 5). At a surrounding distance of 100 km, extreme minicubes covered by bare areas have a relatively high probability of finding a non-extreme minicube, as bare areas dominate global land coverage. A similar pattern applies to mixed trees. Although the likelihood of finding a paired minicube covered with mixed trees within a surrounding distance of 75 km is relatively low, the high global coverage increases the chances of finding a paired non-extreme minicube at larger distances. For the three main vegetation land covers (broad-leaved trees, needle-leaved trees, and grassland), there is a 5% to 10% chance of finding a paired non-extreme minicube with the same land cover type within a surrounding distance of 100 km. Needle-leaved trees slightly exceed grassland at a distance of 150 km, but overall, they exhibit a similar proportion.
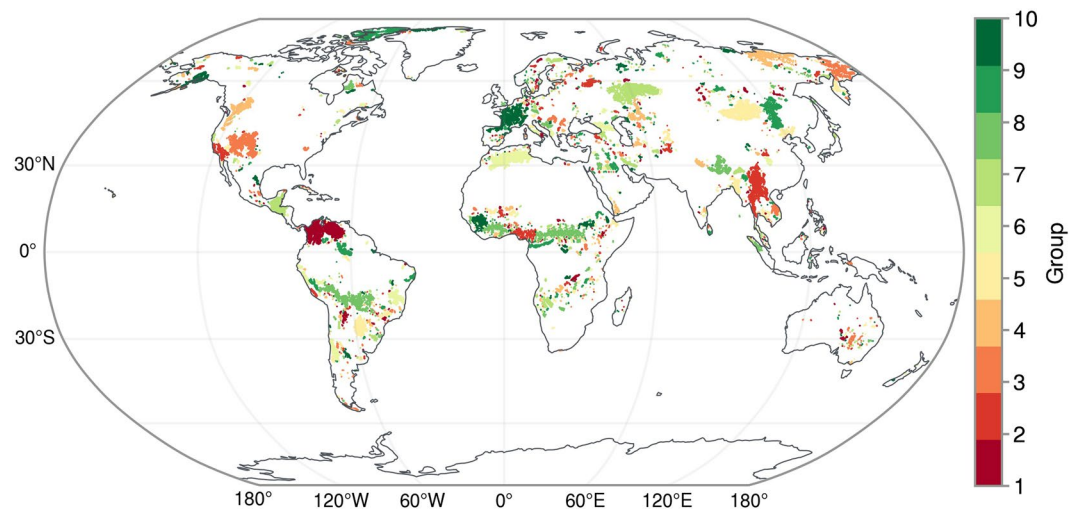
Fig. 6 The spatial division of minicube locations into ten folds using a spatially blocked design to reduce autocorrelation between folds.

Limitation of the data. The *DeepExtremeCubes* dataset focuses on forecasting and analysing the impact of CHD extremes on persistent vegetation types. Although we included bare area and urban area as additional land covers, the diversity of land cover in the real world is not fully encompassed. For example, land covers primarily impacted by anthropological effects were omitted (e.g. croplands). This is the trade-off we must make to narrow variables for better prediction and training results in ML models.

The created event days map includes areas experiencing ten or more event days in the Dheed dataset. This might not be the best approach for a base map for minicube sampling. As mentioned by Weynants *et al.*[35], the amount and volume of extreme events generally follow a power-law distribution, with a few extremely large events and many small ones. Additionally, events in the Dheed dataset with small spatial coverage and short duration could be false alarms. Therefore, selecting the event days map masking areas of more than ten event days might be simplistic and effective as it accurately reflects the real CHD detection results. Considering this limitation, we propose an alternative strategy, setting criteria that require a volume greater than 1000 units, an area exceeding 30 pixels (equivalent to a 0.25° increment), and a duration longer than five days. Using these conditions, we identified 114 significant events out of 26,935 events. The resulting event days map is shown in Fig. s2 in the supplementary information. This alternative map can directly replace the 10-day event map.

Usage Notes

The spatial distribution of the minicubes is strongly uneven. Climatic and meteorological data are correlated across large spatial ranges. This leads to minicubes being clustered in and around extreme events. When designing AI methods, one should circumvent this spatial autocorrelation, and samples may not be randomly selected from the minicubes collection to create the training and test sets. To limit the dependence between training, validation and test sets, we implemented a split of the collection into ten folds, ensuring that the distance between minicubes locations from different folds is larger than 50 km[56]. This split is created in three steps. First, we build a balltree from the haversine distance for all locations (lon, lat). A ball tree (also called metric tree) is a tree that is created from successively splitting points into surrounding hyperspheres whose radii are determined from the given metric[57,58]. The Haversine distance is the angular distance between two points on the surface of a sphere. Second, we create clusters of locations, ensuring that the distance between locations from different clusters is always larger than 50 km. Third, clusters are distributed into ten groups, in decreasing order of size, always adding the largest cluster to the smallest group. Figure 6 shows the spatial distribution of the resulting groups. One application case is that Pellicer *et al.*[59] used the DeepExtremeCubes dataset and demonstrated a convolutional LSTM-based approach to forecast Earth surface impacts during CHD events, achieving high accuracy while leveraging explainable AI to analyze predictive drivers and regime changes both before and during these events.

The criteria for defining "paired minicubes" are primarily determined by the user's study objectives and model training requirements. Users can typically adjust the definition of "paired minicubes" with criteria such as the distance between minicubes, the study period, land cover types (both primary and secondary), specific CHD extremes, and the model's training needs, such as the volume of input data required. If users wish to compute standard spectral indices[53,60], the minicubes allow for easy (and even lazy) computation of the required indices, tailored to their specific objectives.

Code availability

The code to create the minicubes is hosted on GitHub.

8

References

1. Fischer, E. M. *et al.* Storylines for unprecedented heatwaves based on ensemble boosting. *Nature Communications* **14**, 4643 (2023).
2. Flach, M. *et al.* Vegetation modulates the impact of climate extremes on gross primary production. *Biogeosciences* **18**, 39–53 (2021).
3. Lee, R. *et al.* Reclassifying historical disasters: From single to multi-hazards. *Science of the Total Environment* **912**, 169120 (2024).
4. Linkov, I. *et al.* Changing the resilience paradigm. *Nature Climate Change* **4**, 407–409 (2014).
5. Rousi, E. *et al.* The extremely hot and dry 2018 summer in central and northern europe from a multi-faceted weather and climate perspective. *Natural Hazards and Earth System Sciences* **23**, 1699–1718 (2023).
6. Olesen, J. E. & Bindi, M. Consequences of climate change for european agricultural productivity, land use and policy. *European journal of agronomy* **16**, 239–262 (2002).
7. Seneviratne, S. *et al. Changes In Climate Extremes And Their Impacts On The Natural Physical Environment* (Cambridge University Press, Cambridge, UK, and New York, NY, USA, 2012).
8. Zscheischler, J. *et al.* Future climate risk from compound events. *Nature Climate Change* **8**, 469–477 (2018).
9. Mahecha, M. D. *et al.* Biodiversity and climate extremes: known interactions and research gaps. *Earth's future* **12**, e2023EF003963 (2024).
10. Mahecha, M. D. *et al.* Earth system data cubes unravel global multivariate dynamics. *Earth System Dynamics* **11**, 201–234 (2020).
11. Montero, D. *et al.* Earth system data cubes: Avenues for advancing earth system research. *Environmental Data Science* **3**, e27 (2024).
12. Yang, J. *et al.* The role of satellite remote sensing in climate change studies. *Nature climate change* **3**, 875–883 (2013).
13. Smith, A. M. *et al.* Remote sensing the vulnerability of vegetation in natural terrestrial ecosystems. *Remote sensing of environment* **154**, 322–337 (2014).
14. Mahecha, M. D. *et al.* Detecting impacts of extreme events with ecological in situ monitoring networks. *Biogeosciences* **14**, 4255–4277 (2017).
15. Massaro, E. *et al.* Spatially-optimized urban greening for reduction of population exposure to land surface temperature extremes. *Nature Communications* **14**, 2903 (2023).
16. Montero, D. *et al.* On-demand earth system data cubes. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 7529–7532 (2024).
17. Reichstein, M. *et al.* Deep learning and process understanding for data-driven earth system science. *Nature* **566**, 195–204 (2019).
18. Camps-Valls, G., Tuia, D., Zhu, X. & Reichstein, M. *Deep Learning For The Earth Sciences: A Comprehensive Approach To Remote Sensing, Climate Science And Geosciences* (Wiley & Sons, 2021).
19. Martinuzzi, F. *et al.* Learning extreme vegetation response to climate drivers with recurrent neural networks. *Nonlinear Process. Geophys.* **31**, 535–557 (2024).
20. Montero, D. *et al.* Recurrent neural networks for modelling gross primary production. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 4214-4217 (IEEE, 2024).
21. Beigaitė, R. *et al.* Identifying climate thresholds for dominant natural vegetation types at the global scale using machine learning: Average climate versus extremes. *Global Change Biology* **28**, 3557–3579 (2022).
22. Gonzalez-Calabuig, M. *et al.* The aide toolbox: Artificial intelligence for disentangling extreme events [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine* **12**, 113–118 (2024).
23. Racah, E. *et al.* Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. *Advances in neural information processing systems* **30** (2017).
24. Prabhat. *et al.* Climatenet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development* **14**, 107–124 (2021).
25. Requena-Mesa, C., Benson, V., Reichstein, M., Runge, J. & Denzler, J. Earthnet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1132–1142 (2021).
26. Benson, V. *et al.* Multi-modal learning for geospatial vegetation forecasting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).
27. Aybar, C. *et al.* Cloudsen12, a global dataset for semantic understanding of cloud and cloud shadow in sentinel-2. *Scientific data* **9**, 782 (2022).
28. Walther, S. *et al.* Technical note: A view from space on global flux towers by modis and landsat: the fluxneteo data set. *Biogeosciences* **19**, 2805–2840 (2022).
29. Baldocchi, D. D. How eddy covariance flux measurements have contributed to our understanding of global change biology. *Global Change Biology* **26**, 242–260 (2019).
30. Toker, A. *et al.* Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21158–21167, https://doi.org/10.48550/arXiv.2203.12560 (2022).
31. Sumbul, G. *et al.* Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine* **9**, 174–180 (2021).
32. Tseng, G., Zvonkov, I., Purohit, M., Rolnick, D. & Kerner, H. Lightweight, pre-trained transformers for remote sensing timeseries. *Preprint at* https://doi.org/10.48550/arXiv.2304.14065 (2023).
33. Ji, C., Fincke, T. & Mahecha, M. D. Deepextremecubes dataset. *Zenodo* https://doi.org/10.5281/zenodo.13904460 (2024).
34. Weynants, M., Linscheid, N. & Gans, F. Dheed: a global database of dry and hot extreme events. *Zenodo* https://doi.org/10.5281/zenodo.11546130 (2024).
35. Weynants, M. *et al.* Dheed: an era5 based global database of dry and hot extreme events from 1950 to 2022. *Earth System Science Data Discussions* **2024**, 1–31, https://doi.org/10.5194/essd-2024-396 (2024).
36. Esa cci land cover dataset, https://maps.elie.ucl.ac.be/CCI/viewer/download.php (2024).
37. Hersbach, H. *et al.* ERA5 hourly data on single levels from 1940 to present, https://doi.org/10.24381/cds.adbb2d47 (2023).
38. Defourny, P. *et al.* Observed annual global land-use change from 1992 to 2020 three times more dynamic than reported by inventory-based statistics (2023).
39. RAST, M. Esa medium resolution imaging spectrometer (meris). *International journal of remote sensing (Print)* **20** (1999).
40. Cracknell, A. P. *Advanced very high resolution radiometer AVHRR* (Crc Press, 1997).
41. Saint, G. Spot 4 vegetation system: association with high resolution data for multiscale studies. *Advances in Space Research* **17**, 107–110 (1996).
42. Huby, G., Kleihorst, R. P., Mellab, K. & Grignard, L. Proba-v, a vegetation satellite. In *Sensors, Systems, and Next-Generation Satellites XIV*, vol. 7826, 205–210 (SPIE, 2010).
43. CCI, E. L. C. Product user guide version 2.0 (2017).
44. Rajib, A. *et al.* Human alterations of the global floodplains 1992–2019. *Scientific Data* **10**, 499 (2023).
45. Potapov, P. *et al.* The global 2000-2020 land cover and land use change dataset derived from the landsat archive: first results. *Frontiers in Remote Sensing* **3**, 856903 (2022).
46. Song, X.-P. *et al.* Global land change from 1982 to 2016. *Nature* **560**, 639–643 (2018).
47. Zanaga, D. *et al.* Esa worldcover 10 m 2021 v200. *Zenodo*, https://doi.org/10.5281/zenodo.7254221 (2022).
48. Loveland, T. R. & Dwyer, J. L. Landsat: Building a strong future. *Remote Sensing of Environment* **122**, 22–29 (2012).
49. Drusch, M. *et al.* Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment* **120**, 25–36 (2012).

50. Muñoz-Sabater, J. *et al.* ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth system science data* **13**, 4349–4383 (2021).
51. Copernicus DEM - global digital elevation model. https://doi.org/10.5270/ESA-c5d3d65 (2024).
52. Zeng, Y. *et al.* Optical vegetation indices for monitoring terrestrial ecosystems globally. *Nature Reviews Earth &; Environment* **3**, 477–493 (2022).
53. Montero, D. *et al.* A standardized catalogue of spectral indices to advance the use of remote sensing in earth system research. *Scientific Data* **10**, 1–20 (2023).
54. Consult, B. xcube, https://github.com/xcube-dev/xcube (2023).
55. Ji, C. *et al.* Deepextremecubes dataset. *OPARA* https://doi.org/10.25532/OPARA-703 (2024).
56. Weynants, M. & Gans, F. DeepExtremes/cv-groups-minicubes, https://github.com/DeepExtremes/cv-groups-minicubes (2023).
57. Omohundro, S. M. *Five Balltree Construction Algorithms* (International Computer Science Institute Berkeley, 1989).
58. Carlsson, K. *et al.* KristofferC/NearestNeighbors.jl: v0.4.18, https://doi.org/10.5281/zenodo.11941167 (2024).
59. Pellicer-Valero, O. J., Ángel Fernández-Torres, M., Ji, C., Mahecha, M. D. & Camps-Valls, G. Explainable Earth Surface Forecasting under Extreme Events. *Preprint at* https://doi.org/10.48550/arXiv.2410.01770 (2024).
60. Bannari, A., Morin, D., Bonn, F. & Huete, A. A review of vegetation indices. *Remote sensing reviews* **13**, 95–120 (1995).

Author contributions

C.J. contributed to conceptualization and methodology, conducted the experiment, software, validation, data curation, writing the original draft, and project administration. T. F. conducted the creation and storage of minicubes. M.W. and F.G. provided expertise in using the Dheed dataset and ran the spatially blocked dataset split. V.B.: conceptualization, methodology. M.A.F.T.: conceptualization, methodology. K.M.: conceptualization, methodology, and contributed critically to the drafts. F.M.: validation. D.M.: software, visualization. O.P.: methodology, validation. C.R. contributed to the spatially blocked dataset split. M.M.: conceptualization, supervision, and funding acquisition. All authors contributed to writing - review, and editing. Authors from the third to the second-to-last are ordered alphabetically.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04447-5.

Correspondence and requests for materials should be addressed to C.J. or M.D.M.

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.