Earth's Future

RESEARCH ARTICLE
10.1029/2024EF005446

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

Explainable Earth Surface Forecasting Under Extreme Events

Oscar J. Pellicer-Valero[1], Miguel-Ángel Fernández-Torres[1], Chaonan Ji[2,3], Miguel D. Mahecha[2,3,4,5], and Gustau Camps-Valls[1]

[1]Image Processing Laboratory (IPL), Universitat de València, València, Spain, [2]Remote Sensing Centre for Earth System Research (RSC4Earth), Leipzig University, Leipzig, Germany, [3]Institute for Earth System Science and Remote Sensing, Leipzig University, Leipzig, Germany, [4]Image and Signal Processing Group, Leipzig University, Leipzig, Germany, [5]Helmholtz Centre for Environmental Research—UFZ, Leipzig, Germany

Abstract With climate change-related extreme events on the rise, high-dimensional Earth observation data present a unique opportunity for forecasting and understanding impacts on ecosystems. This is, however, impeded by the complexity of processing, visualizing, modeling, and explaining this data. We train a convolutional long short-term memory-based architecture on the novel DeepExtremeCubes data set to showcase how this challenge can be met. DeepExtremeCubes includes around 40,000 long-term Sentinel-2 minicubes (January 2016–October 2022) worldwide, along with labeled extreme events, meteorological data, vegetation land cover, and a topography map, sampled from locations affected by extreme climate events and surrounding areas. When predicting future reflectances and vegetation impacts through the kernel normalized difference vegetation index, the model achieved an R^2 score of 0.9055 in the test set. Explainable artificial intelligence was used to analyze the model's predictions during the October 2020 Central South America compound heatwave and drought event. We chose the same area exactly 1 year before the event as a counterfactual, finding that the average temperature and surface pressure are generally the most important predictors. In contrast, minimum evaporation anomalies play a leading role during the event. We also found the anomalies of the reflectances in the timestep before the extreme event to be critical predictors of its impact on vegetation. The code to replicate all experiments and figures in this paper is publicly available at https://github.com/DeepExtremes/txyXAI.

Plain Language Summary As climate change intensifies, extreme droughts and heatwaves are becoming more frequent, making it increasingly important to understand and predict their effects on ecosystems. In this study, we used advanced machine learning to analyze large-scale satellite imagery from the Sentinel-2 mission, alongside environmental data, to predict how vegetation will be affected by these events. Our model accurately forecasted changes in vegetation cover, even during extreme events. We focused on a significant heatwave and drought event that occurred in Central South America in October 2020, and by comparing this event to data from the same area one year earlier, we found shifts in the most critical environmental factors, such as temperature and evaporation, that influence vegetation health. Our freely available data set and code open the door for future research in analyzing complex climate data and forecasting environmental impacts, helping scientists and policymakers better prepare for the effects of climate change.

Key Points:

- Comprehensive AI toolset for spatio-temporal data modeling and explainability
- Excellent results on vegetation impact forecasting on Sentinel-2
- Analysis of the October 2020 Central South America heatwave with explainable AI

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

O. J. Pellicer-Valero,
oscar.pellicer@uv.es

Citation:

Pellicer-Valero, O. J., Fernández-Torres, M.-Á., Ji, C., Mahecha, M. D., & Camps-Valls, G. (2025). Explainable Earth surface forecasting under extreme events. *Earth's Future*, *13*, e2024EF005446. https://doi.org/10.1029/2024EF005446

Received 11 OCT 2024
Accepted 26 AUG 2025

Author Contributions:

Conceptualization: Miguel-Ángel Fernández-Torres, Miguel D. Mahecha, Gustau Camps-Valls
Data curation: Oscar J. Pellicer-Valero, Chaonan Ji
Formal analysis: Miguel-Ángel Fernández-Torres
Funding acquisition: Miguel D. Mahecha, Gustau Camps-Valls
Investigation: Oscar J. Pellicer-Valero, Miguel-Ángel Fernández-Torres
Methodology: Oscar J. Pellicer-Valero, Miguel-Ángel Fernández-Torres
Project administration: Chaonan Ji

1. Introduction

Climate change is amplifying the frequency and intensity of extreme weather events, which in turn disrupts ecosystem functioning, diminishing carbon sequestration (Reichstein et al., 2013), water retention (Terrado et al., 2014), and biodiversity (Mahecha et al., 2024). All these factors contribute to harvest failures, directly affecting human well-being. Multi-hazard events often manifest as compound events (Zscheischler et al., 2020), posing a much greater impact on society and the environment than individual events. For instance, a Compound Heatwave and Drought (CHD) event can result in global food production issues (Gaupp et al., 2019). The ability to predict and understand the genesis and drivers of compound events could enable the effective implementation of mitigating measures by governmental bodies (Programme, 2020). Applying explainable Artificial Intelligence (XAI) methods to Deep Learning (DL) models trained on Remote Sensing (RS) data offers a theoretical possibility to predict and analyze event impacts.

Early examples of forecasting models for RS data (e.g., predicting satellite reflectance some timesteps ahead) and vegetation impact monitoring (e.g., predicting some vegetation impact index) include Hong et al. (2017), which

Software: Oscar J. Pellicer-Valero
Supervision: Miguel D. Mahecha, Gustau Camps-Valls
Validation: Oscar J. Pellicer-Valero
Visualization: Oscar J. Pellicer-Valero
Writing – original draft: Oscar J. Pellicer-Valero
Writing – review & editing: Miguel-Ángel Fernández-Torres, Chaonan Ji, Miguel D. Mahecha, Gustau Camps-Valls

used a convolutional Long Short-Term Memory (convLSTM) network (Shi et al., 2015) on weather satellite COMS-1 data; or Xu et al. (2019), in which Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) over LSTM networks were applied on FY-2E satellite cloud maps. These forecasting models are generally trained in a self-supervised manner, using future data as prediction targets, eliminating the need for manual labeling and allowing for the collection of arbitrarily large RS data sets.

Requena-Mesa et al. (2021) made a substantial contribution to the field by providing the EarthNet21 data set along with some baseline models. This data set was specifically tailored to the task of spatio-temporal high-resolution meteorology-guided Earth surface forecasting. It contained around 32,000 Sentinel-2 spatio-temporal arrays (also known as "minicubes") and a static Digital Elevation Model (DEM). In the same vein, a convLSTM was trained to predict future Normalized Difference Vegetation Index (NDVI) values, commonly used as a proxy for vegetation health states, conditioned on past reflectances and future ERA5 atmospheric reanalysis data (Hersbach et al., 2020). Various follow-up works have emerged since (Diaconu et al., 2022; Kladny et al., 2024; Martinuzzi et al., 2024). For example, Gao et al. (2022) developed a transformer-based architecture that improved the EarthNet21 metrics; Robin et al. (2022) focused on the African continent; and, very recently, Benson et al. (2024) proposed GreenEarthNet, a new data set with improved cloud mask, better model baselines (using a transformer), and evaluation. Overall, the main limitations of EarthNet21-like approaches are the short context window (50 days, with 5-day period, for a total of 10 samples), the locality of the data (both in the sense that only Europe was considered and that only past data from the same location is used as context), and the non-causality of the meteorological data used for the conditioned prediction (which is assumed to be available up to 100 days in the future at ERA5-like accuracy).

Parallel to these developments, a wave of foundation models Bommasani et al. (2021) has started to appear in the field of RS. In general, these models are characterized by using a wide variety of input modalities, such as Sentinel-2, Sentinel-1, ERA5 variables, DEM, and land-covers, and employing a self-supervised pretraining objective such as masked modeling, wherein pixels/patches, channels/bands, or even complete timesteps are removed from the inputs. As shown in Tseng et al. (2024), the model learns to reconstruct them. After training, the decoder part from the pretrained masked autoencoder is dropped, and the encoder part can then be applied (either by linear-probing the encoder's output or by fine-tuning the encoder's weights) to a variety of downstream tasks (Cong et al., 2022; Reed et al., 2023; Smith et al., 2024; Sun et al., 2022). In all the previous examples, the authors reported improved performance compared to training a model directly on the downstream task. However, smaller data sets seem to benefit the most from the pretraining. Unfortunately, all analyzed methods are limited to static satellite images (do not consider the time dimension) and the modalities seen during training (e.g., they cannot adapt to other sensors), showing visually degraded reconstructions for the masked areas.

One of the downsides of these high-dimensional models is their lack of transparency (i.e., knowing their weights does not help us understand how they work). The field of XAI was therefore born to make them interpretable (i.e., help us understand their decisions) and explainable (i.e., link these interpretations to domain knowledge) (Roscher et al., 2020). One of the most successful sub-fields in XAI is feature attribution, which attempts to assess the impact of an input feature (either positive or negative) on the prediction outcome (Molnar, 2023). For instance, Mateo-Sanchis et al. (2023) trained an LSTM model to predict crop yield, and then used the attribution method Shapley Additive Explanations (SHAP) (Lundberg & Lee, 2017) to conclude that high temperatures during the growth season have a negative impact on crop yield; the authors also discovered the existence of critical periods in the crop growth cycle for corn, soybean, and wheat. State-of-the-art (SOTA) convolutional Neural Networks (CNNs) for the BigEarthNet (Sumbul et al., 2019) and SEN12MS (Schmitt et al., 2019) data sets were explained in Kakogeorgiou and Karantzalos (2021) with 10 different XAI methods, selecting Grad-CAM (Selvaraju et al., 2016), Occlusion, and Lime (Ribeiro et al., 2016) as the best for their purposes. Almost no authors, however, have applied XAI to spatio-temporal models for RS data. A single example was found: Huang et al. (2023) trained a convLSTM model for soil moisture prediction in China from ERA5 data and then used Permutation Importance (Altmann et al., 2010) for global (i.e., model-wide) interpretations and Smooth Gradient (Smilkov et al., 2017) for local ones (i.e., sample-wise), the latter based both on temporal and spatial aggregations (medians). For a recent and exhaustive review on XAI for RS, refer to Höhl et al. (2024). As can be appreciated, the literature has a considerable gap regarding XAI application and visualization for highly dimensional spatio-temporal models. The challenge is furthered by the fact that very compute-intensive XAI methods, such as Occlusion or Lime, cannot be used in practice for such data.

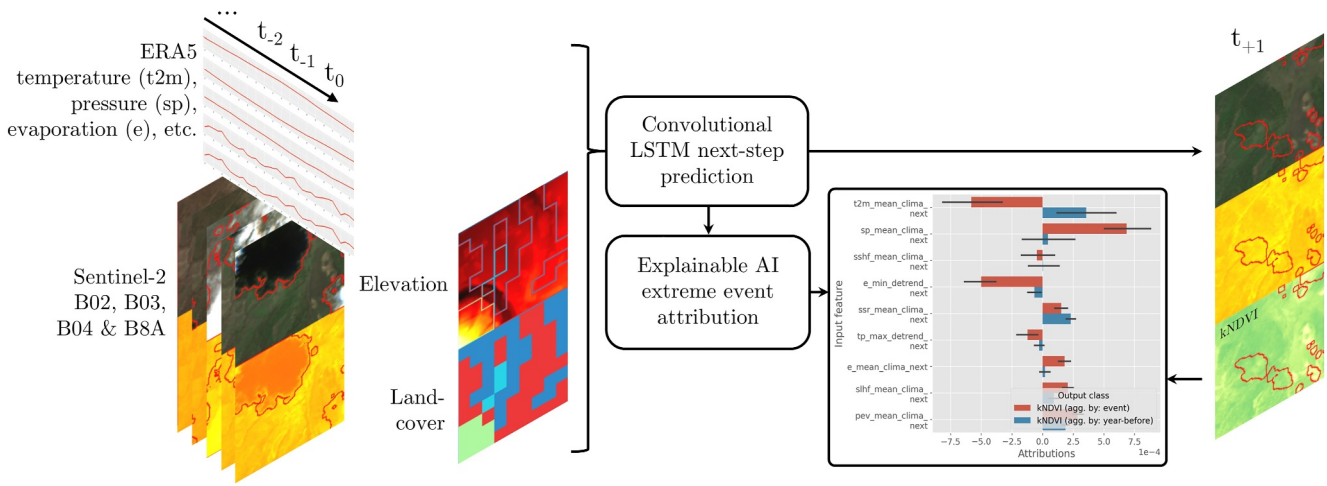

Figure 1. A convolutional LSTM model was trained to forecast future Sentinel-2 reflectances and vegetation impacts given previous timesteps (augmented with ERA5 meteorology, elevation, and land cover). Explainable AI was then used to gather insights into the effects of extreme events on vegetation by comparing the model's attributions during the events and comparable non-event situations. kNDVI: kernel normalized difference vegetation index (Camps-Valls et al., 2021).

The literature on using XAI to explain extreme events such as heatwaves and droughts is scarce. XGBoost was used in Mardian et al. (2023) to predict the Canadian Drought Monitor severity index given many manually crafted features and indices; then, the average SHAP values were used to obtain global feature attribution. XAI has also been applied to understand events such as rainfall extremes (Rampal et al., 2022), river flooding (Jiang et al., 2022), or wildfires (Kondylatos et al., 2022). Li et al. (2024) represented SHAP attributions for wildfires over spatial maps, indicating the most important features of each region and leading to interesting insights. No papers were found on using XAI for compound heatwaves and drought events.

Our main contributions, illustrated in Figure 1, are as follows: First, we provide a comprehensive toolset for processing and visualizing spatio-temporal data, as well as for training and evaluating DL models, and successfully apply it to the novel very long-context DeepExtremeCubes data set. In particular, regarding data processing, we introduce an efficient technique for causal climatology computation while, from a modeling standpoint, we ensure consistent prediction of both reflectances and the kernel NDVI vegetation index via a novel loss function. Second, we offer tools for applying XAI to high-dimensional data, with visualizations at both local (single input) and global levels, demonstrated through the analysis of the October 2020 central South America heatwave. To our knowledge, this is the first publication detailing DL training on global long-context, high-resolution RS data, as well as the first to apply XAI to such data and visualize the results. The code for replicating all experiments and figures is publicly available at https://github.com/DeepExtremes/txyXAI.

2. Materials and Methods

2.1. Data Set and Data Preprocessing

The DeepExtremeCubes data set was used for conducting all the experiments. It contains ~40,000 128 pixels × 128 pixels, globally sampled small data cubes (i.e., minicubes), with a spatial coverage of 2.5 by 2.5 km. Each minicube includes (a) Sentinel-2 L2A images, (b) ERA5-Land variables (Muñoz-Sabater et al., 2021) and generated CHD event cube covering 2016 to 2022 (Weynants et al., 2024), and (c) ancillary land cover and topography maps (Copernicus DEM—Global Digital Elevation Model, 2024). For more information on this data set, please refer to Ji et al. (2025). The variables in each minicube were preprocessed into three large tensors according to their dimensionality (see Table S1 in Supporting Information S1 for a summary).

2.1.1. Spatio-Temporal Data

They are represented by tensor $x_{st} \in \mathbb{R}^{C_{st} \times T \times W \times H}$ with dimensions $C_{st} \times T \times W \times H$, being $C_{st} = 9$ channels, $T = 495$ timesteps (5-day period), and $W = H = 128$ pixels (at 30 m per pixel). It contains a variety of variables (see example minicube in Figure 2):

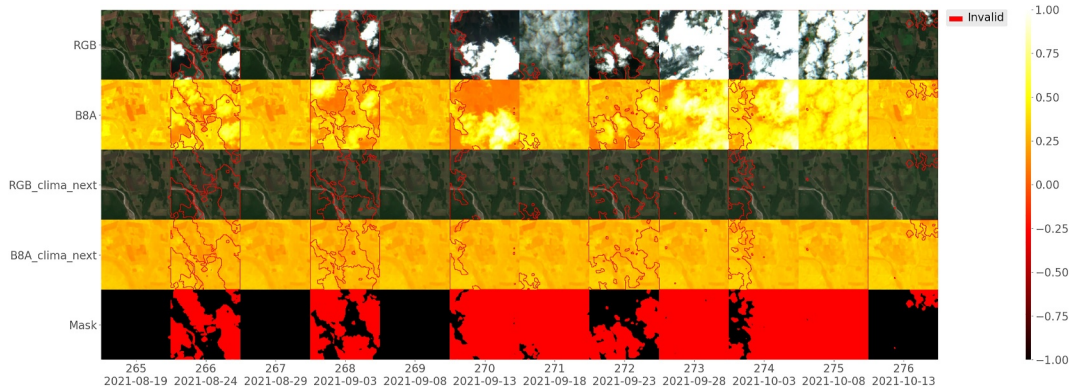

Figure 2. Overview of the spatio-temporal inputs (tensor x_{st}) for a single minicube sampled at location −99.47, 24.17 (China). Only a few timesteps are shown from 2021-08-19 to 2021-10-13.

- Spatio-temporal dynamic satellite reflectance data (*B02*, *B03*, *B04*, and *B8A* Sentinel-2 bands) at 30 m × 30 m × 5 days resolution.
- Their climatologies for the next timestep (*B02_clima_next*, *B03_clima_next*, *B04_clima_next*, and *B8A_clima_next*, to be explained hereafter).
- The *cloud_mask*: with segmentation masks for clouds, cloud shadows, snow, and cloud-free pixels. Instead of imputing the cloud-covered regions, the cloud mask was fed directly as an input so the model could learn to deal with the missing data.

The computation of the climatology requires further explanation. For each spatial location, climatology refers to the average value of a variable for a given day of the year based on previous years of observations. For our purposes, climatology was computed by splitting the year into monthly bins, averaging all observations within each bin, and then interpolating between bins to get the final values for a given time of the year. This process is efficient, as it only requires a linear interpolation between two values for a given timestep; causal, as, for a given timestep, it only uses data that has been observed up to that point, which is crucial for deployment in a real-time scenario; and dynamic, since it constantly gets updated as new samples are observed. In addition, the climatology was calculated for the next timestep (the one the model will attempt to predict), helping the model in its task. This is reflected in the name of these variables (**_clima_next*). Note that the first year of fully available data (2017) was used to kick-start the climatology computation and was excluded from training (see Section 2.2.4).

2.1.2. Spatial Data

They are represented by tensor $x_s \in \mathbb{R}^{C_s \times H \times W}$, with $C_s = 34$ channels. It contains:

- A one-hot encoding of the LCCS vegetation-focused land cover, with a total of 34 different classes (downscaled from 240 m × 240 m resolution to 30 m × 30 m using nearest-neighbor interpolation).
- The Copernicus DEM (*cop_dem*), which has the same 30 m native resolution as the spatio-temporal data.

2.1.3. Temporal Data

They are represented by tensor $x_t \in \mathbb{R}^{C_t \times T}$ with $C_t = 24$ channels, containing exclusively ERA5-Land-derived variables: *e* (evaporation, m of water equivalent), *pev* (potential evaporation, m), *slhf* (surface latent heat flux, J/m^2), *sp* (surface pressure, Pa), *sshf* (surface sensible heat flux, J/m^2), *ssr* (surface net shortwave (solar) radiation, J/m^2), *t2m* (temperature at 2m, K), *tp* (total precipitation, m). These variables, available at a 6-hourly rate, were aggregated to match the 5-day spatio-temporal data. Several aggregation strategies were applied (* stands for any of the above variables):

- **_min_detrend_next*: Minimum anomaly (with respect to climatology, i.e., data was detrended by subtracting the climatology) over the following 5 days.
- **_max_detrend_next*: Maximum anomaly (with respect to climatology) over the following 5 days.
- **_mean_clima_next*: The mean climatology value over the following 5 days.

Note that the variables include *_next* in their name, meaning that their values refer to the next timestep. Therefore, this is a non-causal ERA5 usage, assuming that in a real-time scenario, ERA5-quality atmospheric forecasts would be available up to 5 days in the future. For comparison, similar works (Benson et al., 2024; Requena-Mesa et al., 2021) follow the much stronger assumption that such forecasts would be available for up to 100 days.

2.1.4. Data Standardization

For the Sentinel-2 bands, data were left unmodified; for the *cop_dem*, data were divided by the maximum height in Earth (8,849 m); for the ERA5-variables, 0.01% and 99.99% percentiles were computed on the training set values, and variables (v) were standardized ($\hat{v}$) as follows:

$$\hat{v} = \frac{v - \text{perc}(v, 0.01\%)}{\text{perc}(v, 99.99\%) - \text{perc}(v, 0.01\%)},$$

where *perc* is the percentile operation. Finally, to get rid of any remaining large values that could destabilize training, all data were additionally clipped to the $[-5, 5]$ range.

2.2. Earth Surface and kNDVI Forecasting Model

2.2.1. Task

A convLSTM was trained on this data (x_{st}, x_s and x_t) for the task of vegetation impact forecasting through the prediction of kernel NDVI (kNDVI), a common proxy for vegetation health state,

$$\text{kNDVI} = \tanh\left[\left(\frac{\text{B8A} - \text{B04}}{\text{B8A} + \text{B04} + \epsilon}\right)^2\right]. \quad (1)$$

In particular, the model was trained to predict the same Sentinel-2 input bands but one timestep (five days) into the future (outputs: *B02_next*, *B03_next*, *B04_next*, and *B8A_next*). From these predicted output bands, kNDVI was computed by following Equation 1 (where $\epsilon = 1 \cdot 10^{-5}$ to avoid division by zero). This self-supervised task might seem simple, but in practice, the model must often effectively predict many more days into the future, as many minicubes are almost perpetually covered by clouds (in the training set, more than 50% of the minicubes have at least a 50% cloud coverage), making the problem much harder.

Following the idea proposed by Benson et al. (2024), model inputs *B02_clima_next*, *B03_clima_next*, *B04_clima_next*, and *B8A_clima_next* were added to the predicted outputs to obtain the final *B02_next*, *B03_next*, *B04_next*, and *B8A_next*. Hence, the model is forced to predict the difference with respect to the climatology (i.e., the anomalies) instead of wasting modeling capacity in reconstructing the full reflectances. Besides, the training becomes much more stable as anomalies are typically zero-centered and small in variance.

2.2.2. Architecture

The employed architecture (see Figure 4) is a 3-layer convLSTM (Shi et al., 2015) with a total of 768k parameters implemented using PyTorch Lightning (Paszke et al., 2019). Each convLSTM cell first applies a 2D convolution (kernel size 3×3, padding 2×2, output dimension 240) to its input, which consists of the concatenation of x_{st}, x_s, and x_t along the channel dimension for the first cell, or otherwise, the hidden state H_t from the previous cell; the convolutions allow the model to learn spatial patterns in the input data. Then, the LSTM part of the cell takes the H_{t-1} tensor (from the same cell, but the previous timestep) and splits into the input, input gate, output, and forget tensors to go through the standard LSTM pathways (Hochreiter & Schmidhuber, 1997) and create the new updated states H_t and C_t; the LSTM connections allow the architecture to detect and model temporal patterns in the input. At the output of the third layer, a pixel-wise multi-layer perceptron adapts the hidden dimensionality of the last H_t (60) to the desired output dimensionality (4 bands) using a single 10-unit hidden layer with ReLU activation.

Two more models were tested as ablations: *LSTM* (495k parameters), where 3×3 convolutions were replaced by 1×1 convolutions (so that the model behaved as a pixel-wise standard LSTM network with no spatial context), and the hidden dimension was increased to try to match the number of parameters of the convLSTM (it could not

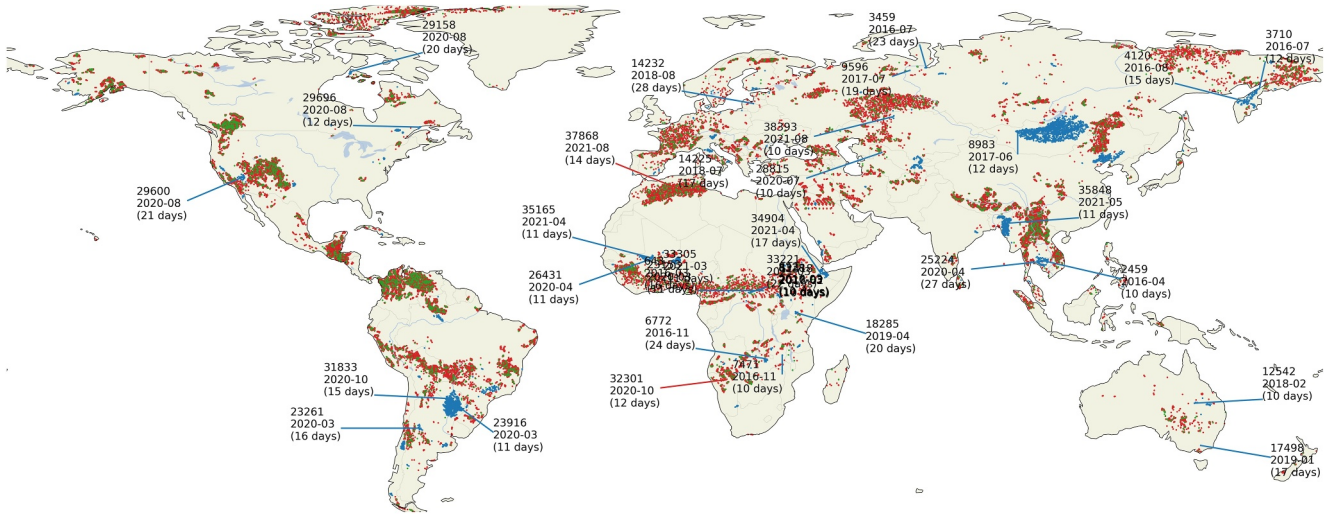

Figure 3. Map of the geographical distribution of the minicubes according to the subset to which they belong (red: train, green: validation, blue: test). Also, Compound Heatwave and Drought events appearing in the test set and lasting at least 10 days have been marked on the map, along with their label ID, occurrence time, and duration.

be matched precisely because of memory limitations); and *Conv* (744 k parameters), in which the three convLSTM cells were replaced by three consecutive 3 × 3 convolutions, with the number of intermediate channels chosen to approximately match the original model in parameter count.

2.2.3. Loss

The L1 loss was used for training, which generally preserves high-frequency details better than the more widely used L2 loss. However, this loss was not directly applied; instead, kNDVI was first computed from network outputs *B04_next* and *B8A_next*. Then the L1 loss was applied to the different outputs independently (B02_next, B03_next, B04_next, B8A_next, and kNDVI_next), and their respective losses were aggregated (by weighted sum) with weights 0.125, 0.125, 0.125, 0.125, and 0.5, respectively. Hence, the network is encouraged to accurately predict *kNDVI_next* (as the relative weight is much higher), not by directly predicting it, but rather by accurately predicting the actual reflectance values from which it is calculated. We argue that this is a better training objective than directly predicting derived maps such as kNDVI since the model does not need to waste capacity in learning the mapping between reflectances and kNDVI; yet, by propagating the loss through the kNDVI, we directly encourage good kNDVI predictions, which is of high interest for the analysis of vegetation impacts.

Unlike in previous impact assessment studies using NDVI, in this work, we chose kNDVI as our main study output (compared to NDVI) for two main reasons: Firstly, it is a better proxy to canopy structure, leaf pigment content (and, subsequently, plant photosynthetic potential), and correlates better with Gross Primary Production and Solar-Induced Chlorophyll Fluorescence than other indices (such as NDVI and Near-Infrared Reflectance of Vegetation -NIRv-) at many spatial and temporal scales (Camps-Valls et al., 2021). Secondly, it allows for the direct optimization of the model with respect to this index. In comparison, propagating the loss directly through the NDVI formula leads to unstable training due to the small denominators leading to arbitrarily large NDVI values. In contrast, the tanh in kNDVI likely provides a smoothing effect, stabilizing training as it bounds predictions. Finally, the loss function was masked using the cloud mask to exclude pixels affected by clouds, cloud shadows, missing data, or other errors. As a result, the model learned to ignore such pixels entirely and never predicts clouds since doing so offers no benefit during training.

2.2.4. Train, Validation and Test Subsets

For validation purposes, the minicubes were split into two sets: *train* + *val* and *test*, such that the locations for any given cube in *train* + *val* were at least 50 km away from all minicube locations in *test*. The data set was then split into the following subsets:

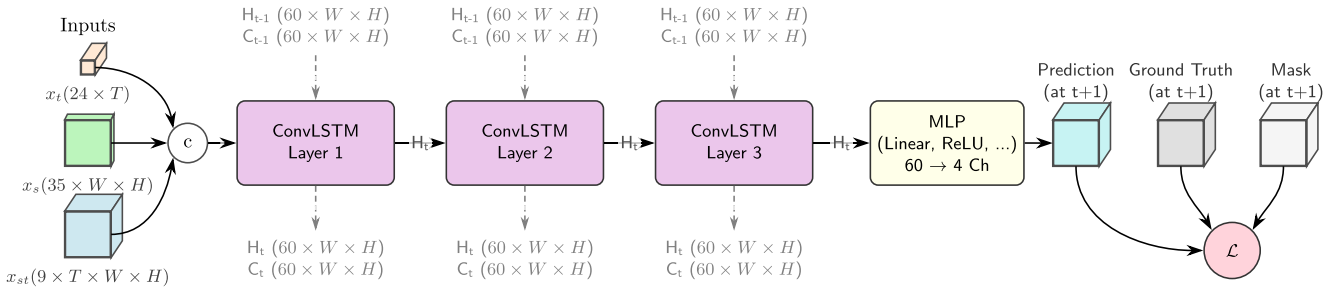

Figure 4. Architecture of the proposed convolutional LSTM (ConvLSTM) model: It consists of three sequential ConvLSTM cells, followed by a multi-layer perceptron (MLP). The three input tensors: temporal data x_t (24 × 110), spatial data x_s (35 × 128 × 128), and spatiotemporal data x_{st} (9 × 110 × 128 × 128) are concatenated and fed into the first ConvLSTM cell, producing after the MLP the output (4 × 128 × 128) as output, which represents the prediction at the next timestep. Each convLSTM cell receives the hidden states H_{t-1}, C_{t-1} from the same cell at the previous timestep (top), and generates the new hidden states H_t, C_t for the next timestep (bottom). The loss is computed from the predicted kNDVI and the ground truth kNDVI, masking out the cloud-covered regions.

- Training set: random 80% of the minicubes from *train* + *val*, from January 2018 until December 2021.
- Validation set: remaining 20% of the minicubes from *train* + *val*, from January 2022 until October 2022. Therefore, these minicubes are temporally uncorrelated from those in the training set but geographically close.
- Test set: all cubes from *test* from January 2022 until October 2022. The cubes in this set are spatially (>50 km away) and temporally uncorrelated with the training set.

Figure 3 shows a map of the geographical distribution of the minicubes according to their subset. Note that the first two years of data were not included in any of the subsets: 2016 was removed because only one of the two Sentinel-2 satellites was online during this first year of the mission, and 2017 was removed because at least a year of observations was needed to compute the climatology (as explained in Section 2.1).

2.2.5. Training

The model was trained for 12,500 training steps with a batch size of one (the largest one that would fit in GPU memory) using batch gradient accumulation to keep the effective batch size fixed at eight. The AdamW optimizer with a fixed learning rate of 0.001 was used for optimization. All the decisions regarding the specific values for the different hyperparameters (such as learning rate, number of layers, training schedule, etc.) were based on extensive experimentation on the validation set. Every epoch took around 16 hr on a high-end A100-SXM4-80 GB GPU.

2.2.6. Evaluation

Once trained, the predicted kNDVIs were evaluated over the training, validation, and test sets using several performance metrics on a per-cube basis. Only the last year of data (unseen by the model) was used for the validation and test sets, with the test set remaining completely unseen until the end to serve as a proxy for the model's actual performance. Both cloud-covered/unavailable pixels and pixels belonging to a non-vegetation land cover were ignored as we focused on vegetation impacts. The final value for the metric was computed as the grand mean of the per-cube metrics. The considered metrics were the following: L_1 (mean absolute error), L_2 (mean square error), R^2 score (fraction of kNDVI variance explained by the model), Normalized Nash-Sutcliffe Efficiency (NNSE, used to assess the predictive skill of hydrological models, and equivalent to $\frac{1}{2-R^2}$), and bias (average systematic absolute error committed by the model).

Additionally, the results from all the previous metrics were compared against the results of two *naïve* (parameter-less) models: the climatology at the prediction timestep and the last non-cloud-covered available value. As shown in Figure 5 and Table S2 in Supporting Information S1, these simple models constitute a strong baseline.

2.3. Explainable AI for Extreme Event Understanding

For obtaining both local and global explanations, we used Python's Captum library (Kokhlikyan et al., 2020), a well-known open-source library for XAI built on PyTorch implementing SOTA attribution methods. After some initial testing on a variety of methods, including Input × Gradient (Shrikumar et al., 2016) and Gradient SHAP

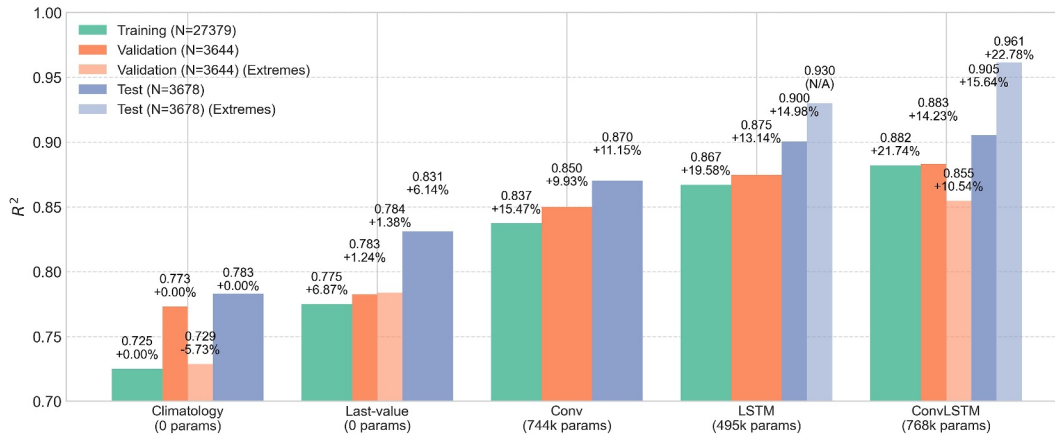

Figure 5. Comparison of R^2 scores for kNDVI prediction across different models and data splits (Training, Validation, Test). Bars show the R^2 value, with the percentage improvement relative to the Climatology baseline. For some models and splits, timestep results identified as extreme events are also shown in lighter colors. Model parameter counts and data split sample sizes (N) are indicated below each model's name and in the legend.

(Lundberg & Lee, 2017), finally, Integrated Gradients (IG) (Sundararajan et al., 2017) (computed over nine integration steps) was used for the experiments presented here. In IG, attributions are calculated as the integral of the gradients of the outputs with respect to the inputs along the path from a given baseline (e.g., 0) to the input, hence theoretically solving the issues of sensitivity and implementation invariance of other methods. Out of all the output classes (*B02_next*, *B03_next*, *B04_next*, and *B8A_next*), attributions were only calculated with respect to *kNDVI_next*, since the focus of the XAI stage is to analyze the impacts on vegetation. Also, the baselines were chosen to be zero for all input variables except for the input reflectances *B02*, *B03*, *B04*, and *B8A_next*, for which they were set to the climatology, so that the anomalies in the inputs are attributed (Mamalakis et al., 2023).

2.3.1. Local and Global Attributions

A naïve use of attribution methods for our problem would yield large attribution tensors since an attribution score is produced for every input and output value combination. Therefore, some local aggregations were performed. More precisely, the output's spatio-temporal dimensions were always aggregated by taking the mean over selected timesteps. While some local attribution tensors were represented as is, global attributions were obtained by taking the mean of the input's spatio-temporal dimensions over several minicubes. This procedure is further clarified in Section 2.3.2.

2.3.2. The October 2020 Central South America Heatwave

The XAI analysis in this paper was focused on the October 2020 Central South America heatwave, which affected a large geographical extension (from southern Peruvian Amazon to southeastern Brazil) for a long time (September 23rd to October 15th) and had a strong impact on the region (reaching record temperatures of 10°C above normal, some locations reporting maximum temperatures above 40°C for several days). A persistent atmospheric blocking caused the heatwave: a warm air mass lingered for several days, leading to significant temperature anomalies, likely exacerbated by very low soil moisture, as solar energy heated the atmosphere rather than evaporating the non-existent water, which, in turn, worsened drought conditions, escalating fires and impacting natural and human systems (Marengo et al., 2022).

This compound drought and heatwave event had been selected for building the DeepExtremeCubes data set (and labeled with event IDs 32379 and 31833), which allowed us to use it for our study. For the XAI analysis of this event, minicubes undergoing event ID 31833, for which the event lasted at least 10 days (two timesteps), were selected (and relegated to the test set, see Figure 3), for a total of 24 minicubes. Then, the average attributions of these minicubes were obtained for the model's outputs over the event and for the same period but exactly 1 year before, allowing us to compare feature importance for event and non-event predictions. Global attributions were obtained by taking the mean over the input's spatio-temporal dimensions and all 24 minicubes during the event,

and one year before the event. Finally, one minicube was selected, and the complete disaggregated local attributions were analyzed.

2.4. A General Toolset

The code for this Section 2 was developed to be as general as possible. For the data preprocessing, the Pytorch DataLoader applies different processing steps depending on the dimensionality of the data (x_{st},x_s, or x_t) and its type (real-valued or categorical), including code for dimension-checking, data visualization, etc. On the model side, plugging in custom models is straightforward. Similarly, the XAI code developed for this work bridges the gap between standard XAI libraries such as Captum (which is intended for single input images, and single output attributions for a given output class), and geospatial use cases, such as this one, where there can be multiple inputs of different sizes, and the output of the model is also a high-dimensional tensor. As such, it should be able to handle various kinds of tasks (regression or classification), data dimensionalities (x_{st},x_s, x_t), attribution methods (IG, SHAP, Input × Gradient, etc.), aggregation strategies (mean, absolute maximum value, PCA, or clustering), etc., while providing automatic yet meaningful summaries and visualizations. In summary, all these tools should be helpful to a much wider audience beyond any specific application such as DeepExtremeCubes.

3. Results and Discussion

3.1. Performance Evaluation

The Supplementary Material (Table S2 in Supporting Information S1) summarizes the forecasting model's quantitative results. A visual comparison focusing on the R^2 metric is provided in Figure 5. As can be seen, the convLSTM outperforms the baseline climatology by more than 15% over the training, validation, and test sets, meaning that the model has successfully learned to generalize. Although absolute scores seem better in the test set (Figure 5), relative improvements are similar across splits (see Table S2 in Supporting Information S1), possibly suggesting that this set might contain cubes that are inherently easier to predict.

Comparing against the ablated models (*LSTM* and *Conv*), the convLSTM consistently performs better across splits and metrics (Figure 5 and Table S2 in Supporting Information S1), suggesting that the model benefits both from the spatial context captured by convolutions and the temporal dependencies captured by the LSTM gates. Also, in the validation set, the performance of the convLSTM falls significantly when evaluated only on timesteps labeled as extreme events (lighter bars in Figure 5), which is expected, as they represent anomalous conditions and are hence more challenging to predict. This performance drop is also observed in the climatology baseline (since an extreme is, by definition, a deviation from climatology) but not as much in the last-value baseline. Curiously, the convLSTM performance in the test set is excellent even in the presence of extremes ($R^2 = 0.9614$ compared to $R^2 = 0.9054$ for all test data), which justifies using XAI over these events in Section 3.2. The detailed metrics in Table S2 of Supporting Information S1 also show that the performance on non-extreme timesteps is practically identical to the overall performance, reflecting the relatively small proportion of timesteps labeled as extremes.

For a qualitative assessment of the results, Figure 6 shows the model's predictions and the Ground Truth (GT) reflectance values for a few timesteps of a single test minicube. On the one hand, for the first few timesteps, the model accurately predicts the progressive thawing of the snow and the sudden snowfall present in the fourth timestep (despite the two previous timesteps being entirely covered by snow). On the other hand, the convLSTM correctly predicts the gradual greening of the scene over the latter timesteps. Furthermore, the forecasts seem robust to cloud coverage while doing a reasonable job at gap-filling, even if not explicitly trained for the task.

3.2. Explainable AI for the October 2020 Central South America Heatwave

In Section 3.1, the convLSTM was shown to have strong performance for vegetation impact forecasting (kNDVI) under extreme events in the test set, which is a prerequisite for any later XAI analysis to be useful (i.e., there are no insights to be gained from a model that performs poorly). Figure 7 shows the average attributions (over time, space, and minicubes) for variables in tensors x_t,x_s and x_{st} of minicubes affected by the October 2020 central South America heatwave, with red bars representing the attribution for the event, and the blue bars representing the attributions for the same period, but 1 year before.

10.1029/2024EF005446

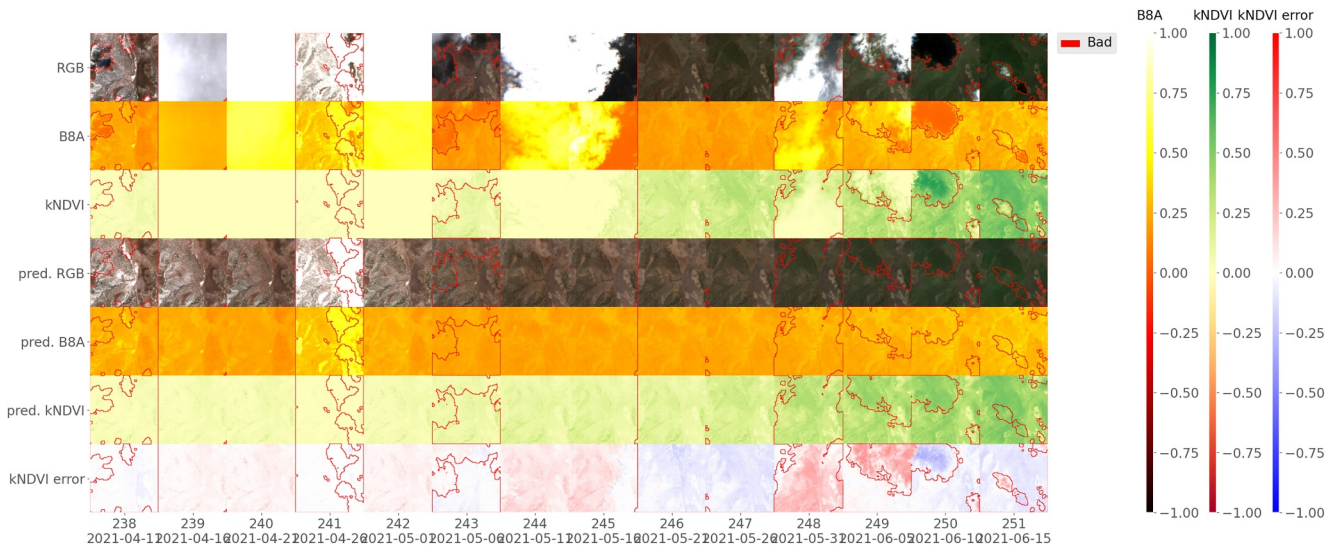

Figure 6. A Minicube at location 101.95°E, 46.97°N (Mongolia), from 2021-04-11 to 2021-06-15. Top three rows: ground truth *RGB_next*, *B8A_next*, and *kNDVI_next*. Next three rows: model predictions. Last row: Model error on predicted kNDVI computed as *kNDVI_predicted - kNDVI_next* (note that the error in cloud-covered regions is ignored in error calculations). Red outline: *cloud_mask* (labeled as "bad"). Minicube L_1: 0.0351, L_2: 0.0587, R^2: 0.9288, NNSE: 0.9335, bias: 0.0095.

Analyzing x_t (ERA5-derived) variables, *t2m_mean_clima* and *sp_mean_clima* (the average climatology of the temperature at 2 m, and of the surface pressure) seem to be the most critical general predictors for kNDVI, with higher average temperatures leading to higher kNDVI values and, conversely, higher pressures resulting in lower kNDVI values. During extremes, a few instantaneous variables stand out, such as *e_min_detrend_next* (minimum detrended evaporation at the prediction timestep). Here, the negative sign indicates that increased evaporation (i.e., low values of this variable) corresponds to an expected negative impact on the output kNDVI. For the non-event attribution, this effect seems to be much smaller. Regarding x_s variables (land cover and DEM), the Figure has been omitted because the error bars were large, and it was hence deemed uninformative. The land cover class with the largest average absolute attribution was *mosaic_natural_vegetation*.

For x_{st} (reflectance-derived and cloud mask variables), the most relevant input is, surprisingly, the cloud mask; however, there is a large error bar in the estimation of its mean, meaning that the sign of its contribution varies significantly. The signs of *B04* and *B8A*, as well as their respective climatologies (*B04_clima_next* and

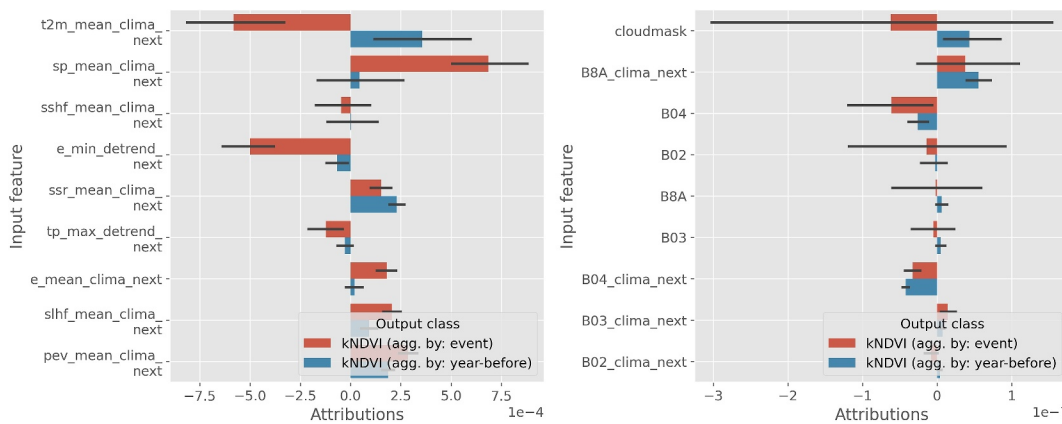

Figure 7. Bar plot of the average attributions (over time, space, and minicubes) for top nine variables (sorted by descending average of absolute value) in tensor x_t (left) and x_{st} (right) of 24 minicubes affected by event 31833 (October 2020 central South America heatwave). The red bars represent the attribution for the model's outputs coinciding with the event, while the blue bars represent the attributions for the same period but 1 year before. Error bars represent the 95% confidence intervals of the mean computed via bootstrapping. The bar plot for x_s has been omitted because the error bars were large and it was hence deemed uninformative.

10.1029/2024EF005446

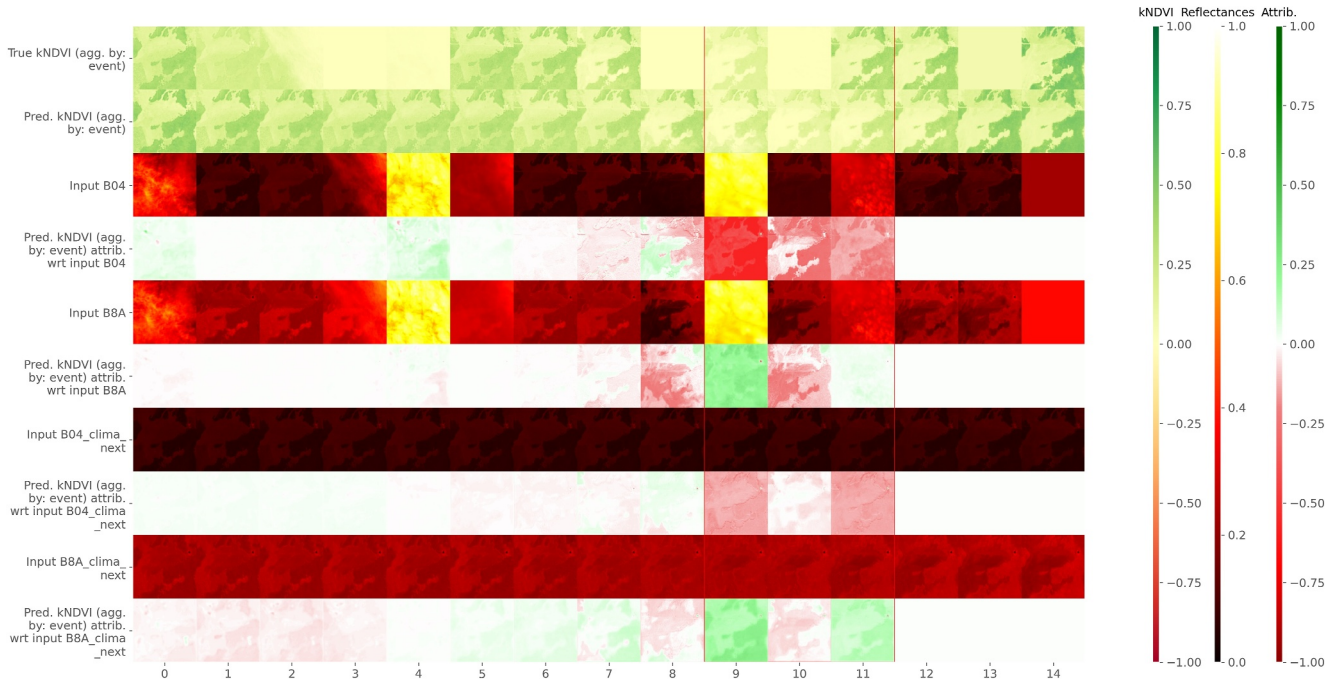

Figure 8. Full attributions for a few timesteps of input x_{st} from a minicube sampled at −58.17°E −23.98°N (Paraguay) and affected by event 31833 (October 2020 central South America heatwave). The first two rows represent the ground truth kernel Normalized Vegetation Index (kNDVI) and the model's prediction. In contrast, the rest of the rows represent an input feature (only the top four by importance) and its corresponding attribution map (green: positive, red: negative) in an alternating pattern. Thus, the first two rows are offset by one timestep into the future with respect to the rest. There is a red outline around timesteps 9–11, signaling where event 31833 occurred, aligned to the time of the true and predicted kNDVI (2020-10-03, 2020-10-08, and 2020-10-13). Attributions were rescaled to lie within the $[-1, 1]$ range while remaining zero-centered.

B8A_clima_next), correspond with the signs that these bands have in the numerator of the kNDVI Equation 1. Finally, the importance of *B04_clima_next* and *B8A_clima_next* is slightly lower during extremes, where conditions are, by definition, exceptional.

Overall, Figure 7 should be analyzed with caution, as the average attribution over many timesteps and locations can hide large positive and negative contributions and lead to results where a larger attribution value does not necessarily translate into more attention paid by the model. Still, Figure 7 is informative in at least two ways: Firstly, it provides a sorting by importance, as variables are sorted by descending average of the absolute value of the attributions (hence, values are all made positive before averaging). Secondly, it provides the average effect of a given variable in the model's output, which can be meaningful if the error bars are not too large.

Figures 8 and 9 show the full attribution tensors for inputs x_{st} and x_t, respectively, from a minicube affected by event 31833 (October 2020 central South America heatwave). The first two rows of Figure 8 represent the GT kNDVI and the model's prediction, whereas the rest of the rows represent an input feature and its corresponding attribution map in an alternating pattern. Thus, the first two rows are offset by one timestep into the future with respect to the rest. A red outline around timesteps 9–11 (2020-10-03, 2020-10-08, and 2020-10-13), signaling where event 31833 occurred, that is aligned to the timestep of the first two rows. We see that the model accurately predicted the reduction in greenness during the extreme event, which is desirable for the explanations to be meaningful.

A few interesting observations can be drawn from Figure 8: Firstly, the strongest attributions for bands *B04* and *B8A* seem to correspond to the timestep right before the start of the extreme event (step 9), meaning that this step was crucial for the prediction of the lower kNDVI values due to the extreme. Also, the signs of these attributions align with the signs of bands *B04* and *B8A* in the kNDVI equation (Equation 1). Secondly, the attributions become smaller as we move back in time from the attributed timestep (although they never quite reach zero if we could look at the full image), and they are precisely zero after the attribution timesteps (since the convLSTM model is sequential and cannot look into the future). Thirdly, cloud-occluded reflectances have a strong attribution

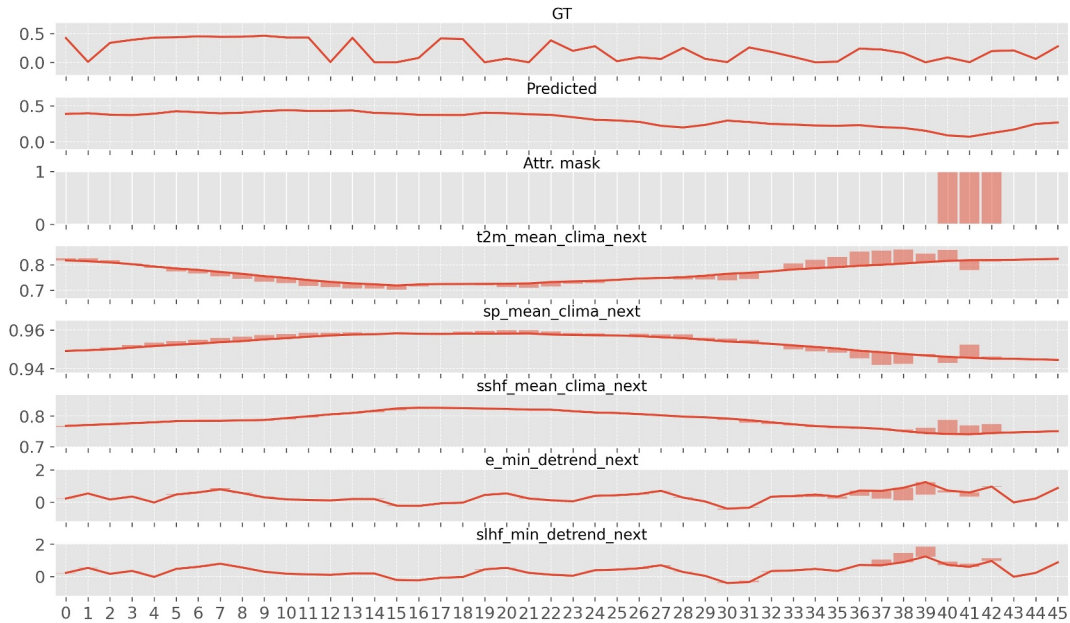

Figure 9. Full attributions for a few timesteps of input x_t from a minicube sampled at −58.17–23.98 (Paraguay) and affected by event 31833 (October 2020 Central South America heatwave). The first three plots represent, respectively, the average Ground Truth kNDVI, the predicted kNDVI, and a mask (aligned to the time of the true and predicted kNDVIs) indicating the timesteps (30–32, equivalent to 2020-10-03, 2020-10-08, and 2020-10-13) over which the attribution (and also the event) took place. In contrast, the rest of the plots show the top five most important input feature values (as a line) and their corresponding attributions (as a bar: positive when going upward and negative otherwise).

(opposite to the idea that they should not contribute), perhaps meaning that the presence of clouds is itself predictive of future kNDVIs. Fourthly, different parts of a single tile contribute differently to the attribution. In particular, B04 (Red) shows strong negative (red) attribution in areas of timestep eight where B04 reflectance is high (less vegetation), which aligns well with expectations since higher red reflectance means less chlorophyll, contributing to a lower predicted kNDVI. Also, B8A (NIR) shows positive (green) attribution where B8A reflectance is high (more vegetation biomass) since higher NIR reflectance means healthier/denser vegetation, contributing to a higher predicted kNDVI. Unfortunately, the complexity makes further analysis extremely hard, supporting the use of aggregations of the previous study despite the limitations.

Similarly, the first three plots in Figure 9 represent, respectively, the average GT kNDVI, the predicted kNDVI, and a mask indicating the timesteps (30–32, equivalent to 2020-10-03, 2020-10-08, and 2020-10-13) over which the attribution (and also the event) took place; whereas the rest of the plots show the input feature values as a line, along with their corresponding attributions as a bar (positive when going upwards, and negative otherwise). Compared to Figure 8, a longer context was chosen for this figure. Firstly, we see that the average predicted kNDVI (second plot) goes down progressively over time, reaching a minimum that coincides with the event; this cannot be as easily appreciated in the original GT signal due to the perpetual presence of clouds, which is alleviated in the cloudless prediction. Furthermore, climatology variables' attributions go much farther back than the instantaneous variables, which are only relevant for the last few timesteps before the prediction, helping explain why the climatology variables tend to have higher attributions when averaged over time. The climatology variables seem to undergo a change regime during the extreme, with temperature no longer being positively related to kNDVI and surface pressure starting to do so. Conversely, instantaneous variables are important right before the event but, surprisingly, lose their relevance during the event itself. Overall, Figure 9 highlights that the model's prediction of the kNDVI dip during the heatwave is strongly influenced by climate variables related to heat and water stress in the period leading up to and during the event.

4. Conclusion

The successful marriage of high-dimensional Earth Observation data modeling and XAI can help us predict and understand the occurrence of extreme events affecting ecosystems, particularly CHD events, which are

particularly interesting due to their vast impacts. This study provides the tools to make it possible through the accurate forecasting of kNDVI using Deep Learning models, the explanation of these models' behavior in the presence of such events, and their visualization.

From the data processing point of view, an efficient and causal climatology was employed as direct input to the model and for detrending other variables, keeping only the anomaly signal. Both variables were later proven helpful in the XAI analysis, with climatology being important during non-extremes and anomalies during extremes. From the modeling side, predicting only the anomalies in reflectance with respect to the climatology helped the model converge, while introducing the kNDVI computation in the loss (instead of predicting it directly) helped the model learn both reflectances and kNDVI efficiently. From the explainability point of view, Captum's attribution methods have been adapted to spatio-temporal data, and we have developed the tools to visualize the explanations, allowing us to gain interesting insights that substantiate the model's behavior during extremes.

However, some significant limitations remain: the model predicts only one timestep ahead, the provided explanations cannot always be linked satisfactorily to the domain knowledge, and the data handling is computationally costly. For future research, some promising avenues include using XAI storylines for extreme event understanding, exploring other XAI aggregation methods beyond simple averaging (e.g., PCA, clustering), or training more modern (e.g., attention-based) architectures. By making the code publicly available, together with the public DeepExtremeCubes data set, we hope to encourage the community to use these tools and baseline models to further improve Earth surface forecasting ability and extreme event understanding, broadening this relatively unexplored field.

Inclusion in Global Research

This research was conducted within the framework of the DeepExtremes project, funded by the European Space Agency (ESA). It addresses the increasing frequency and intensity of climate extremes, one of the most critical impacts of climate change, emphasizing the need for advanced methodologies that can detect, predict, and understand the impacts of such events. We acknowledge the successful collaboration of several public and private research institutions, including the Universitat de València, Leipzig University, Max Planck Institute, and Brockmann Consulting, whose contributions were invaluable. The project adhered to all necessary authorizations, permits, and formal agreements. The outcomes of this research are intended to support researchers and organizations dedicated to understanding and mitigating the effects of climate extremes on ecosystems and human societies.

Data Availability Statement

We employed the *DeepExtremeCubes* database, comprising over 40,000 spatially sampled small data cubes (i.e., minicubes) globally, with a spatial coverage of 2.5 by 2.5 km. Each minicube includes (a) Sentinel-2 L2A images, (b) ERA5-Land variables and generated extreme event cubes covering 2016 to 2022, and (c) ancillary land cover and topography maps. The *DeepExtremeCubes* dataset is being permanently stored in the OPARA Repository of TU Dresden (Ji et al., 2024). The code to replicate all experiments and figures in this paper is publicly available at https://github.com/DeepExtremes/txyXAI (Pellicer-Valero, 2025) under the MIT License.

Acknowledgments
This work was supported by the ESA AI4Science project "Multi-Hazards, Compounds and Cascade events: DeepExtremes", 2022–2024; the European Union's Horizon 2020 research and innovation project XAIDA: Extreme Events—Artificial Intelligence for Detection and Attribution, 2021–2024 (Grant agreement No 101003469); and the European Union's Horizon 2022 project ELIAS: European Lighthouse of AI for Sustainability, 2023–2027 (Grant agreement No. 101120237). The authors acknowledge the support from the computer resources provided by Artemisa (funded by the European Union ERDF and Comunitat Valenciana), as well as the technical support provided by the Instituto de Física Corpuscular, IFIC (CSIC-UV).

References

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/BIOINFORMATICS/BTQ134

Benson, V., Robin, C., Requena-Mesa, C., Alonso, L., Carvalhais, N., Cortés, J., et al. (2024). Multi-modal learning for geospatial vegetation forecasting. *arXiv*, 27788–27799. https://doi.org/10.1109/cvpr52733.2024.02625

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. *arXiv*. Retrieved from https://arxiv.org/abs/2108.07258v3

Camps-Valls, G., Campos-Taberner, M., Álvaro, M.-M., Walther, S., Duveiller, G., Cescatti, A., et al. (2021). A unified vegetation index for quantifying the terrestrial biosphere. *Science Advances*, *7*(9), 7447–7473. https://doi.org/10.1126/sciadv.abc7447

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., et al. (2022). Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Neurips 2022*. Retrieved from https://sustainlab-group.github.io/SatMAE/

Copernicus DEM—global digital elevation model. (2024). Copernicus DEM—Global digital elevation model. https://doi.org/10.5270/ESA-c5d3d65

Diaconu, C. A., Saha, S., Gunnemann, S., & Zhu, X. X. (2022). Understanding the role of weather data for Earth surface forecasting using a convLSTM-based model. In *IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 1361–1370). IEEE Computer Society. https://doi.org/10.1109/CVPRW56347.2022.00142

Gao, Z., Shi, X., Wang, H., Zhu, Y., Wang, Y., Li, M., & Yeung, D.-Y. (2022). Earthformer: Exploring space-time transformers for Earth system forecasting. *arXiv*. Retrieved from http://arxiv.org/abs/2207.05833

Gaupp, F., Hall, J., Mitchell, D., & Dadson, S. (2019). Increasing risks of multiple breadbasket failure under 1.5 and 2°C global warming. *Agricultural Systems*, *175*, 34–45. https://doi.org/10.1016/J.AGSY.2019.05.010

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. arXiv. Retrieved from http://www.github.com/goodfeli/adversarial

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The Era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/QJ.3803

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Höhl, A., Obadic, I., Fernández-Torres, M.-Á., Najjar, H., Oliveira, D., Akata, Z., et al. (2024). Opening the black-box: A systematic review on explainable ai in remote sensing. *IEEE Geoscience and Remote Sensing Magazine*.

Hong, S., Kim, S., Joh, M., & Song, S.-K. (2017). Psique: Next sequence prediction of satellite images using a convolutional sequence-to-sequence network. In *Workshop on deep learning for physical sciences*. nips.

Huang, F., Zhang, Y., Zhang, Y., Shangguan, W., Li, Q., Li, L., & Jiang, S. (2023). Interpreting conv-LSTM for spatio-temporal soil moisture prediction in China. *Agriculture*, *13*, 971. https://doi.org/10.3390/agriculture13050971

Ji, C., Fincke, T., Benson, V., Camps-Valls, G., Fernández-Torres, M.-Á., & Gans, F. (2024). Deepextremecubes [Dataset]. *Universität Leipzig*. https://doi.org/10.25532/OPARA-703

Ji, C., Fincke, T., Benson, V., Camps-Valls, G., Fernández-Torres, M.-Á., Gans, F., et al. (2025). Deepextremecubes: Earth system spatio-temporal data for assessing compound heatwave and drought impacts. *Scientific Data*, *12*(1), 149. https://doi.org/10.1038/s41597-025-04447-5

Jiang, S., Bevacqua, E., & Zscheischler, J. (2022). River flooding mechanisms and their changes in Europe revealed by explainable machine learning. *Hydrology and Earth System Sciences*, *26*(24), 6339–6359. https://doi.org/10.5194/hess-26-6339-2022

Kakogeorgiou, I., & Karantzalos, K. (2021). Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, *103*, 102520. https://doi.org/10.1016/J.JAG.2021.102520

Kladny, K. R., Milanta, M., Mraz, O., Hufkens, K., & Stocker, B. D. (2024). Enhanced prediction of vegetation responses to extreme drought using deep learning and Earth observation data. *Ecological Informatics*, *80*, 102474. https://doi.org/10.1016/j.ecoinf.2024.102474

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., et al. (2020). Captum: A unified and generic model interpretability library for pytorch. arXiv. Retrieved from https://arxiv.org/abs/2009.07896v1

Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., et al. (2022). Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters*, *49*(17), e2022GL099368. https://doi.org/10.1029/2022GL099368

Li, H., Vulova, S., Rocha, A. D., & Kleinschmit, B. (2024). Spatio-temporal feature attribution of European summer wildfires with explainable artificial intelligence (Xai). *Science of The Total Environment*, *916*, 170330. https://doi.org/10.1016/J.SCITOTENV.2024.170330

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4766–4775. Retrieved from https://arxiv.org/abs/1705.07874v2

Mahecha, M. D., Bastos, A., Bohn, F., Eisenhauer, N., Feilhauer, H., Hickler, T., et al. (2024). Biodiversity and climate extremes: Known interactions and research gaps. *Earth's Future*, *12*(6), 0–5. https://doi.org/10.1029/2023EF003963

Mamalakis, A., Barnes, E. A., & Ebert-Uphoff, I. (2023). Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience. *Artificial Intelligence for the Earth Systems*, *2*(1), e220058. https://doi.org/10.1175/aies-d-22-0058.1

Mardian, J., Champagne, C., Bonsal, B., & Berg, A. (2023). Understanding the drivers of drought onset and intensification in the Canadian prairies: Insights from explainable artificial intelligence (XAI). *Journal of Hydrometeorology*, *24*(11), 2035–2055. https://doi.org/10.1175/JHM-D-23-0036.1

Marengo, J. A., Ambrizzi, T., Barreto, N., Cunha, A. P., Ramos, A. M., Skansi, M., et al. (2022). The heat wave of October 2020 in central South America. *International Journal of Climatology*, *42*(4), 2281–2298. https://doi.org/10.1002/joc.7365

Martinuzzi, F., Mahecha, M. D., Camps-Valls, G., Montero, D., Williams, T., & Mora, K. (2024). Learning extreme vegetation response to climate drivers with recurrent neural networks. *Nonlinear Processes in Geophysics*, *31*(4), 535–557. https://doi.org/10.5194/npg-31-535-2024

Mateo-Sanchis, A., Adsuara, J. E., Piles, M., Munoz-Mari, J., Perez-Suay, A., & Camps-Valls, G. (2023). Interpretable long short-term memory networks for crop yield estimation. *IEEE Geoscience and Remote Sensing Letters*, *20*, 1–5. https://doi.org/10.1109/LGRS.2023.3244064

Molnar, C. (2023). Interpretable machine learning, a guide for making black box models explainable. Independently published. Retrieved from https://christophm.github.io/interpretable-ml-book/

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pellicer-Valero, O. J. (2025). *Deepextremes/txyxai: Initial release [code]*. Zenodo. https://doi.org/10.5281/zenodo.15242307

Programme, U. N. E. (2020). *Adaptation gap report 2020*. United Nations.

Rampal, N., Gibson, P. B., Sood, A., Stuart, S., Fauchereau, N. C., Brandolino, C., et al. (2022). High-resolution downscaling with interpretable deep learning: Rainfall extremes over New Zealand. *Weather and Climate Extremes*, *38*, 100525. https://doi.org/10.1016/J.WACE.2022.100525

Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., et al. (2023). Scale-Mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 4065–4076). arXiv. https://doi.org/10.1109/iccv51070.2023.00378

Reichstein, M., Bahn, M., Ciais, P., Frank, D., Mahecha, M. D., Seneviratne, S. I., et al. (2013). Climate extremes and the carbon cycle. *Nature 500*, *500*(7462), 287–295. https://doi.org/10.1038/nature12350

Requena-Mesa, C., Benson, V., Reichstein, M., Runge, J., & Denzler, J. (2021). Earthnet2021: A large-scale dataset and challenge for Earth surface forecasting as a guided video prediction task. *arXiv*, 1132–1142. https://doi.org/10.1109/cvprw53098.2021.00124

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *NAACL-HLT 2016 - 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies, proceedings of the demonstrations session* (pp. 97–101). https://doi.org/10.18653/v1/n16-3020

Robin, C., Requena-Mesa, C., Benson, V., Alonso, L., Poehls, J., Carvalhais, N., & Reichstein, M. (2022). Learning to forecast vegetation greenness at fine resolution over Africa with convlstms. *arXiv*. Retrieved from http://arxiv.org/abs/2210.13648

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, *8*, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

Schmitt, M., Hughes, L. H., Qiu, C., & Zhu, X. X. (2019). Sen12ms—A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. ISPRS annals of the photogrammetry. *Remote Sensing and Spatial Information Sciences*, *4*, 153–160. https://doi.org/10.5194/isprs-annals-IV-2-W7-153-2019

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C., & Observatory, H. K. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *arXiv*.

Shrikumar, A., Greenside, P., Shcherbina, A. Y., & Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv, 1, 0-5*. Retrieved from https://arxiv.org/abs/1605.01713v3

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: Removing noise by adding noise. *arXiv*. Retrieved from https://arxiv.org/abs/1706.03825v1

Smith, M. J., Fleming, L., & Geach, J. E. (2024). Earthpt: A time series foundation model for Earth observation. *arXiv*. Retrieved from http://arxiv.org/abs/2309.07207

Sumbul, G., Charfuelan, M., Demir, B., & Markl, V. (2019). Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *Igarss 2019 - 2019 ieee international geoscience and remote sensing symposium* (pp. 5901–5904). https://doi.org/10.1109/IGARSS.2019.8900532

Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., et al. (2022). Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–22. https://doi.org/10.1109/TGRS.2022.3194732

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *34th international conference on machine learning, ICML 2017* (Vol. 7, pp. 5109–5118). Retrieved from https://arxiv.org/abs/1703.01365v2

Terrado, M., Acuña, V., Acuña, A., Ennaanay, D., Tallis, H., & Sabater, S. (2014). Impact of climate extremes on hydrological ecosystem services in a heavily humanized Mediterranean Basin. *Ecological Indicators*, *37*, 199–209. https://doi.org/10.1016/j.ecolind.2013.01.016

Tseng, G., Cartuyvels, R., Zvonkov, I., Purohit, M., Rolnick, D., & Kerner, H. (2024). Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv*. Retrieved from http://arxiv.org/abs/2304.14065

Weynants, C., Melanie, J. I., Linscheid, N., Weber, U., Mahecha, M. D., & Gans, F. (2024). Dheed: An era5 based global database of dry and hot extreme events from 1950 to 2022. *Preprint at*. https://doi.org/10.5281/zenodo.13710040

Xu, Z., Du, J., Wang, J., Jiang, C., & Ren, Y. (2019). Satellite image prediction relying on GAN and LSTM neural networks. *IEEE International Conference on Communications, 2019-May*. https://doi.org/10.1109/ICC.2019.8761462

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., et al. (2020). A typology of compound weather and climate events. *Nature Reviews Earth and*, *1*(7), 333–347. https://doi.org/10.1038/s43017-020-0060-z