Identifying Spatio-Temporal Drivers of Extreme Events

Mohamad Hakam Shams Eddin **Juergen Gall**

Institute of Computer Science, University of Bonn
Lamarr Institute for Machine Learning and Artificial Intelligence
{shams, gall}@iai.uni-bonn.de

Abstract

The spatio-temporal relations of impacts of extreme events and their drivers in climate data are not fully understood and there is a need of machine learning approaches to identify such spatio-temporal relations from data. The task, however, is very challenging since there are time delays between extremes and their drivers, and the spatial response of such drivers is inhomogeneous. In this work, we propose a first approach and benchmarks to tackle this challenge. Our approach is trained end-to-end to predict spatio-temporally extremes and spatio-temporally drivers in the physical input variables jointly. By enforcing the network to predict extremes from spatio-temporal binary masks of identified drivers, the network successfully identifies drivers that are correlated with extremes. We evaluate our approach on three newly created synthetic benchmarks, where two of them are based on remote sensing or reanalysis climate data, and on two real-world reanalysis datasets. The source code and datasets are publicly available at the project page https://hakamshams.github.io/IDE.

1 Introduction

A frontier research challenge is to understand the affects of global change on the magnitude and probability of extreme weather events [1]. Overall, the evolution of extreme events such as agricultural droughts results from stochastic processes [2], conditions at ecosystem scales [3, 4], and the interaction between the Earth land and atmospheric variables as a part of a complex system of feedbacks [5]. However, the relative impacts of these factors differ depending on the event [2]. The time delays between extremes and their drivers vary seasonally [6–9], and the spatial response of these drivers is inhomogeneous [4]. A major challenge is therefore to model the spatio-temporal relations between extremes and their drivers during the development of these events [4]. The overarching goal of this modelling is to improve our understanding of the patterns and impacts of such events. This would improve our ability to project duration and intensity of extreme events and hence assisting in adaptation planning [10, 11].

In this work, we propose an approach that identifies spatio-temporal drivers in multivariate climate data that are correlated with the impact of extreme events. For the extreme events, we focus on agricultural droughts as an example, which can be measured by extremely low values of the vegetation health index (VHI). As drivers for such measurable extreme events, we consider anomalies in atmospheric and hydrological state variables like temperature or soil moisture, as well as land-atmosphere fluxes like evaporation. The task of identifying end-to-end spatio-temporal drivers for measurable impacts of extremes has not been addressed before, and it is very challenging since the drivers can occur earlier in time and at a different location than the measured extreme event as illustrated in Fig. 1.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

To address this challenging task, we propose a network that is trained to predict spatio-temporally extremes. Instead of simply predicting the extremes, the network quantizes the spatio-temporal input variables into binary states and predicts the extremes only from the spatio-temporal binary maps for each time series of input variables. In this way, the network is enforced to identify only drivers in the input variables that are spatio-temporally correlated with extreme events. While the network is trained using annotations of impacts of extreme events, which can be derived from remote sensing or reported data, we do not have any annotations of drivers or anomalies in the input variables.

Since drivers of extreme events are not fully understood, we cannot quantitatively measure the accuracy of the identified drivers on real-world data. We therefore propose a framework for generating synthetic data that can be used to assess the performance of our model as well as other baselines quantitatively. We evaluate our approach on three synthetic datasets where two of them are based on remote sensing or reanalysis climate data. Our evaluation shows that our approach outperforms approaches for interpretable forecasting, spatio-temporal anomaly detection, out-of-distribution detection, and multiple instance learning. Furthermore, we conduct empirical studies on two real-world reanalysis climate data. Our contributions can be summarized as follow:

- We introduce the new task of identifying spatio-temporal drivers of extreme events and three benchmarks for evaluating this highly important task.
- We propose a novel approach for identifying spatio-temporal drivers in climate data that are spatio-temporally correlated with the impacts of extreme events.
- We further verify our approach on two long-term real-world reanalysis datasets including various physical variables from five biogeographical diverse regions.

2 Related works

Anomalies and extremes detection in climate data. The identification of climatic changes and extreme weather has been a subject of many studies [12, 13]. Typical algorithms for extreme events detection are built upon domain knowledge in setting usually empirical thresholds for the physical variables through sensitivity experiments [6, 14]. Many works applied multivariate and statistical methods to detect extreme events such as droughts [6, 9, 15–17]. However, individual events are difficult to generalized across multiple events [2] and predefined indicators become less effective with changing climate [18]. Thus, machine and deep learning methods have been proposed as an alternative to classical methods, i.e., for supervised anomaly detection [19, 20] and for the detection of extremes in climate data [14, 21–24].

While methods for forecasting vegetation indices [25–28] do not focus on extremes, future impacts of extremes like agricultural droughts can be derived from forecast vegetation indices like the vegetation health index (VHI). For instance, the work [28] uses a climate simulation as input and forecasts the vegetation health index. Since predicting VHI directly is difficult, the approach predicts the normalized difference vegetation index and the brightness temperature instead. Both indices are then normalized and used to estimate VHI. Although we obtain the impacts of extreme events in our study from vegetation indices, our approach is not limited to such extremes. Since we use a binary representation of extremes, our approach can also be applied to other extremes that cannot be derived from satellite products, but that are stored in a binary format in databases.

While we aim to learn the relations between the impacts of extreme events and their relevant drivers from a data-driven perspective, spatio-temporal relations within the Earth system can also be inferred by causal inference and causal representation learning [29–36]. In contrast to statistical methods [37], data-driven methods do not require a prior hypothesis about drivers for extremes. Instead, they generate hypotheses that can be verified by statistical methods in a second step. We believe that this is an important direction since climate reanalysis provides huge amount of data and it is infeasible to test all combinations. This is also known as a curse of dimensionality in causal discovery problems [38] and data-driven approaches are therefore needed to generate potential candidates.

Anomaly detection algorithms. Since we focus on anomalies in land and meteorological data as drivers, we give an overview of approaches for anomaly detection and discuss their applicability to our task, which has not been previously addressed. *One-class:* The main stream in one-class anomaly detection is to model the distribution of the normal data during training and consider the deviation from the learned features as anomalies. This includes distance-based [39–41], patch-based

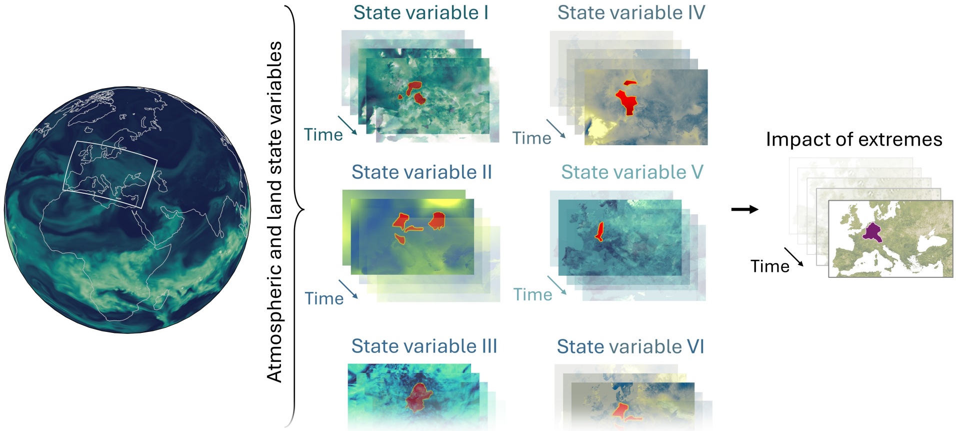


Figure 1: Overview of the objective of this work. We are interested in identifying spatio-temporal relations between the measurable impacts of extremes like the vegetation health index 🟪 and their drivers 🟥. As drivers, we focus on anomalies in state variables of the land-atmosphere and hydrological cycle. The task is very challenging since the drivers can occur at a different region than the extreme event and earlier in time.

[41–43], student-teacher [44–47], and embedding-based approaches [43, 48–51]. In general, the alignment between the anomaly type and the assumptions of the methods is the critical factor for their performance [52]. One of the limiting factors to apply these methods to climate data is that they assume priori knowledge about what is considered as normal. Furthermore, not all detected anomalies are drivers of an extreme event. *Reconstruction-based:* These methods assume that a trained model to reconstruct normal data will be unsuccessful in reconstructing anomalies, while it will reconstruct normal data well. Despite being widely applied for anomaly detection problems [53–64], these methods face the same problem as the one-class methods. In addition, many studies showed that anomalies can still be reconstructed by the trained model [65]. *Self-supervised learning:* These methods rely on the hypothesis that a model trained for a pretext task on normal data will be successful only on similar normal data during inference [66–69]. In addition to the above discussed limitations, finding a suitable pretext task for anomaly detection is challenging. For instance, common tasks such as solving a jigsaw puzzle [69] will fail in homogeneous regions. *Pseudo-anomaly:* The intuition in pseudo-based anomaly detection is to convert the problem of unsupervised learning into a supervised one by synthesizing abnormal data during training [70–77]. Since these methods depend partially on the degree to which the proxy anomalies correspond to the unknown true anomalies [44], applying these approaches to our task would require some knowledge about the coupling between the variables and extremes. *Multivariate anomaly detection:* Multivariate approaches detect anomalies simultaneously in multiple data streams [78–82]. The main difference to our task is that we aim to detect drivers across multiple data streams that do not necessary occur simultaneously. *Multiple instance learning:* Multiple instance learning (MIL) has been proposed for weakly supervised anomaly detection [83–91]. In MIL-based algorithms, the model is provided with labeled positive and negative bags where each bag includes a set of instances. The model is then trained to classify the instances inside the bags giving only the high level supervision, i.e., label of the bag [92]. A weakly supervised approach has been also applied for hyperspectral anomaly detection [93]. Most algorithms such as [94–97] choose the top-k potentially anomalous snippets within each video. This makes it challenging to apply them since the abnormality ratio varies in real-world applications [98]. Furthermore, the MIL detector can be biased toward a specific class depending on the context [99].

3 Method

Our aim is to design a model that is capable of identifying spatio-temporal drivers of extremes in multivariate climate data, i.e., Earth observations or climate reanalysis. In particular, we want to identify anomalies that are spatio-temporally related to extreme events like agricultural droughts. This is different to standard anomaly detection since we are not interested in all anomalies, but in spatio-temporal configurations of variables that potentially cause an extreme event with some time

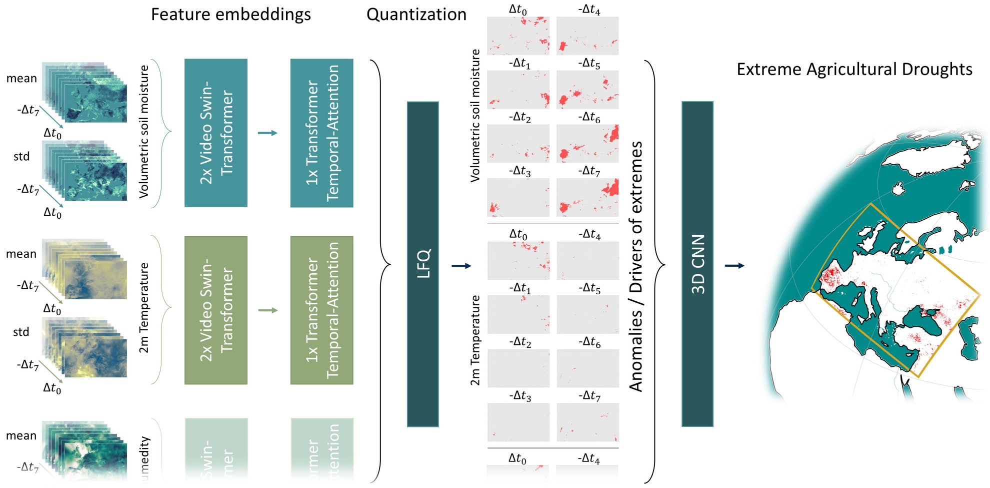


Figure 2: An overview of the proposed model to identify the spatio-temporal relations between extreme agricultural droughts and their drivers. The input variables are first encoded into features. In a subsequent step, a lockup free quantization layer (LFQ) takes the extracted features and classifies the variables into a binary representation of drivers, where we consider the drivers as anomalous events in the input variables. Finally, a classifier is used to predict impacts of extreme events from the identified drivers.

delay and at a potentially different location as illustrated in Fig. 1. To achieve this, we propose a network that is trained end-to-end on observed extremes where we focus on agricultural droughts. The network classifies the input variables before an extreme event occurs into spatio-temporal drivers without additional annotations besides the annotated extremes. The network then aims to predict future extremes based on the identified drivers. It needs to be noted that we are interested in input variables that do not define an extreme event, but we aim to find anomalies in input variables that are correlated with the occurrence of an extreme event. We will thus denote potential drivers as anomalies.

An overview of our approach is shown in Fig. 2. The input are weekly climate variables at sequential time steps $(\Delta t_{-7}, \dots, \Delta t_0)$ and the model is composed of three main parts. First, a feature extractor extracts relevant features from each input variable independently. The second component is a quantization layer that takes the extracted features as input and classifies the input variables into drivers. The role of the quantization layer is to transform the input variable into a binary representation (1 = drivers and 0 otherwise). This ensures that no additional information is encoded as an input to the subsequent classifier except that if the input variable at a specific location and time is a driver or not. The third component is a classifier that takes as input the variables, location, and time where drivers have been identified and predicts where extreme droughts occur at the time step Δt_0. All model components are trained jointly.

We denote the input data as $\mathbf{X} \in \mathbb{R}^{V \times C \times T \times Lat \times Lon}$, where V is the input variables, C is the channel dimension for each variable (i.e., mean and standard deviation of the week), T is the temporal resolution $(\Delta t_{-T+1}, \dots, \Delta t_0)$, and Lat and Lon are the spatial extensions. The model has two outputs, $\mathbf{Q} \in \mathbb{Z}_2^{V \times T \times Lat \times Lon}$ representing the binary classification of the input variables into potential drivers of the extreme events, and probabilities $\mathbf{E} \in \mathbb{R}^{Lat \times Lon}$ to predict extreme events at the time step Δt_0. In the following, we describe the model components:

Feature extraction. First, embedded features $f_\theta : \mathbf{X} \rightarrow \mathbf{Z} \in \mathbb{R}^{V \times K \times T \times Lat \times Lon}$ are extracted independently for each input variable $v \in V$ with K embedding dimensions and parameters θ. The rationale behind this independence is to prevent that drivers leak into other variables. We use the Video Swin Transformer model [100] as backbone to capture long-range interrelations across time and space. The input $\mathbf{X}$ is projected into a higher feature dimension K and followed by two Video Swin Transformer layers. The first layer has two consecutive 3D shifted window blocks for a spatio-temporal feature extraction. The second layer consists of one block for a temporal feature extraction. The later is useful to focus only on the temporal evolution of the variables. An ablation study regarding the backbone is provided in Sec. C.4.

Quantization layer. The role of this layer is to map $\mathbf{Z}$ from an embedded space into a compact binary representation $\mathbf{Q}$ suitable for detecting drivers. Using vector quantization (VQ) [101], each embedded feature vector $z \in \mathbf{Z}$ is assigned into a learnable codebook feature vector $z_q \in \mathbf{Z}_q$ based on the Euclidean distance:

$$VQ : z \rightarrow z_q, \text{where } q = \underset{q \in \{1,\ldots,Q\}}{\arg\min} \; \|z - z_q\|_2 \,, \tag{1}$$

where Q is the size of the codebook. Recently, lookup-free quantization (LFQ) [102] substitutes the learnable codebook with a set of integers $\mathbb{Q}$ with $|\mathbb{Q}| = Q$ and represents the embedding space as a Cartesian product of binary numbers. This omits the need for a distance metric to do the nearest vector assignment and simplifies the quantization. Based on experimental results, we built the vector quantizer on LFQ with two integers $Q = 2$ ($q = 1$ for drivers and $q = 0$ otherwise). Given a feature vector z, LFQ first maps z into a scalar value $z_l \in \mathbf{Z}_l \subset \mathbb{R}^{V \times 1 \times T \times Lat \times Lon}$. For multi-modality, we use two sequential 3D CNNs on each input variable followed by a shared linear layer that maps z to z_l and reduces the dimensions from K to 1. The quantization is then given by the sign of z_l:

$$z_q = \text{Linear}\big(\text{sign}(z_l)\big) = \text{Linear}(-\mathbb{1}_{\{z_l \leq 0\}} + \mathbb{1}_{\{z_l > 0\}}), \;\; \text{and} \;\; q = \mathbb{1}_{\{z_l > 0\}} \,, \tag{2}$$

where $q \in \mathbf{Q}$ represents the class ($q = 1$ or $q = 0$), and Linear is a linear layer that converts sign(z_l) back to the dimension K of the input after the quantization. Note that $\mathbf{Z}_q$ has only two unique vectors $z_{q=1}$ for a driver and $z_{q=0}$ otherwise.

Prediction of extreme events. We use a classifier that predicts the probably of extreme events $\mathbf{E}$ at the time step Δt_0 from the identified drivers $\mathbf{Z}_q$. We use a 3D CNN classifier instead of a transformer to reduce the computations. For training, we only know the ground truth of extremes at time step Δt_0 denoted by $\hat{\mathbf{E}} \in \mathbb{Z}_2^{Lat \times Lon}$. While we could compute the cross-entropy between $\mathbf{E}$ and $\hat{\mathbf{E}}$, we found that a single 3D CNN is insufficient to detect all drivers that are correlated with an extreme event. Instead, we use $V{+}1$ 3D CNNs where each predicts $\mathbf{E}_v$. While the first V 3D CNNs take the identified drivers for a single variable v as input, the last one takes the identified drivers of all variables as input. The multiple CNNs are only used for training. During inference, $\mathbf{E}$ is only predicted by the multivariate CNN where all variables are jointly used. The loss is thus given by

$$\mathcal{L}_{(extreme)} = -\sum_{v=1}^{V+1} \big(\hat{\mathbf{E}} \log(\mathbf{E}_v) + (1 - \hat{\mathbf{E}}) \log(1 - \mathbf{E}_v)\big)\mathbf{S} \,, \tag{3}$$

where $\mathbf{S} \in \mathbb{Z}_2^{T \times Lat \times Lon}$ is a mask for the valid regions. We actually utilize a weighted version of $\mathcal{L}_{(extreme)}$ to mitigate the class imbalance issue (Sec. C.3). While the loss $\mathcal{L}_{(extreme)}$ ensures that extreme events can be predicted from the identified drivers, we need to add standard loss terms for the quantization to ensure that the learned codes and thus drivers are compact:

$$\mathcal{L}_{(quantize)} = \lambda_{(commit)} \|\mathbf{Z}_l - \text{sg}(\text{sign}(\mathbf{Z}_l))\|_2^2 + \lambda_{(ent)} \mathbb{E}[H\big(\text{sign}(\mathbf{Z}_l)\big)] - \lambda_{(div)} H[\mathbb{E}\big(\text{sign}(\mathbf{Z}_l)\big)] \,. \tag{4}$$

The commitment loss $\|\mathbf{Z}_l - \text{sg}(\text{sign}(\mathbf{Z}_l))\|_2^2$ prevents the outputs of the encoder from growing and encourages $\mathbf{Z}_l$ to commit to the codes [101], where sg stands for the stopgradient operator with zero partial derivative. The term $\mathbb{E}[H\big(\text{sign}(\mathbf{Z}_l)\big)]$ encourages that the entropy per quantized code is low [102, 103], meaning that it provides more confident assignments. Whereas the term $H[\mathbb{E}\big(\text{sign}(\mathbf{Z}_l)\big)]$ increases the entropy inside the batch to encourage the utilization of all codes [102, 103]. The last important ingredient is a loss that ensures that only spatio-temporal regions are identified that correlate with an extreme event. To this end, we look at regions and intervals where no extreme event occurred and use these examples without drivers. Formally, we use $\hat{\mathbf{E}}_t \in \mathbb{Z}_2^{Lat \times Lon}$ as the union of extreme ground truth at all time steps $(\Delta t_{-T+1}, \ldots, \Delta t_0)$ and compute the loss by

$$\mathcal{L}_{(driver)} = \lambda_{(driver)} |\mathbf{Z}_q - \text{sg}(z_{q=0})| (1 - \hat{\mathbf{E}}_t)\mathbf{S} \,, \tag{5}$$

where $z_{q=0}$ is the quantization code for normal data without drivers. The model is trained end-to-end with the joint optimization of the loss function:

$$\min_{\theta,\phi,\psi} \; \underbrace{\mathcal{L}_{(extreme)}\big(\mathbf{E}, \hat{\mathbf{E}}, \mathbf{S}\big)}_{\text{predicts extremes from drivers}} + \underbrace{\mathcal{L}_{(quantize)}\big(\mathbf{Z}_l\big)}_{\text{encourages confident quantization}} + \underbrace{\mathcal{L}_{(driver)}\big(\mathbf{Z}_q, \hat{\mathbf{E}}_t, \mathbf{S}, \mathbf{Z}_{q=0}\big)}_{\text{assigns drivers to the same code in the codebook}} \;, \tag{6}$$

where θ, ϕ, ψ are the learnable parameters, and $\lambda_{(commit)}$, $\lambda_{(ent)}$, $\lambda_{(div)}$, and $\lambda_{(driver)}$ are weighting parameters. Ablation studies are provided in Sec. 5.1 and in Appendix Sec. C.

4 Dataset

4.1 Defining extreme agricultural droughts from remote sensing

We are interested in a specific impact of extreme events namely extreme agricultural drought. To define such extreme event, we rely on the observational satellite-based vegetation health index (VHI) obtained from NOAA [104]. This remote sensing product cannot be directly derived from the input reanalysis, which makes the task very challenging. VHI approximates the vegetation state based on a combination of the brightness temperature and normalized difference vegetation index (VHI $= 0$ for unfavorable condition and VHI $= 100$ for favorable condition). Extreme agricultural droughts are usually defined as VHI < 26 [104]. The dataset has a temporal coverage of 1981-onward and is provided globally on a weekly basis. We mapped this dataset into the same domains of the reanalysis data as described in Sec. 4.2 and used this dataset as ground truth for extreme events. Note that VHI is a general vegetation index and should be interpreted carefully. Details about this index, the dataset and pre-processing are provided in the Appendix Sec. I.

4.2 Climate reanalysis

Reanalysis data aim to provide a coherent and complete reconstruction of the historical Earth system state as close to reality as possible. During reanalysis, short-term forecasts from numerical climate models are refined with observations within the so called data assimilation framework [105]. We conducted the experiments on two real-world reanalysis datasets; CERRA reanalysis [106] and ERA5-Land [107]. ERA5-Land is widely used for global climate research and it is provided hourly at $0.1^\circ \times 0.1^\circ$ on the regular latitude longitude grid. CERRA is a regional reanalysis for Europe and is provided originally at 5.5km $\times$ 5.5km on its Lambert conformal conical grid with a 3-hourly temporal resolution. We aggregated these two datasets on a weekly basis and selected the years within the period overlapping with the remote sensing data. In addition, we mapped ERA5-Land into 6 CORDEX domains [108] over the globe and conduct experiments on each region separately. We do the experiments with 6 common variables from ERA5-Land and CERRA based on their connections to agricultural droughts. For each variable and week, we computed the mean and standard deviation separately. More details regarding the variables and the domains along with the training/validation/test splits are provided in the Appendix Sec. H and Tables 20 and 21.

4.3 Synthetic dataset

Although ground truth for extreme droughts can be obtained from remote sensing, an important methodological question remains as how to reliably have a meaningful quantitative evaluation of the identified drivers and their relations to the extreme events. To solve this critical issue, we introduce a new synthetic dataset that mimics the properties of Earth observations including drivers and anomalies. We are aware that the dynamic of the synthetic data are simplified compared to real Earth observations. However, we rely on this generated dataset to perform the quantitative evaluation of the proposed approach. In a first step, we generate the normal data. For instance to generate synthetic data of 2m temperature from CERRA reanalysis, the normal signal at a specific time and location is generated based on the typical value of 2m temperature at that time and location (i.e., the mean or median value from a long-term climatology). The second step is to generate anomalies conditioned on the occurrence of extremes. To achieve this, we assign binary spatio-temporally connected flags as extreme events randomly within the datacube and track their precise spatio-temporal locations. Then based on a predefined coupling matrix between the variables and the extreme event, we generate anomalous events only for the variables that are defined to be correlated with the extremes. We consider these anomalies as the drivers for the extreme events. Finally, we add additional random anomalous events for all variables. We synthesize overall 46 years of data; 34 years for training, 6 subsequent years for validation and the last 6 years for testing. The challenge is to identify the drivers, i.e., the anomalous events that are correlated with extreme events. Examples of the synthetic data are shown in Fig. 3 and in Appendix in Figs. 7-12. Technical details are explained in Appendix Sec. A.

5 Experimental results

First, we conducted experiments and ablation studies on the synthetic datasets (Sec. 4.3). We also empirically verified the effectiveness of the proposed design compared to baselines on this synthetic dataset. Then, we validate the model on two real-world datasets over five continents in Sec. 5.2.

Setup and implementation details. We set the hidden dimension K to 16 by default. The temporal resolution is $T = 6$ for the synthetic data and $T = 8$ for real-world data. Since, seasonal cycles are typical in climate data, we deseasonalize locally by subtracting the median seasonal cycle and normalizing by the seasonal variance for each pixel. Details regarding the model and implementation setup are given in Appendix Sec. G. For evaluation, we use the F1-score, intersection over union (IoU), and overall accuracy on both classes (OA).

5.1 Experiments on the synthetic datasets

We show the results on the Synthetic CERRA described in Sec. 4.3 and in Appendix Table 3. The generated dataset mimics a set of variables ($V = 6$) using statistics from the real-world CERRA reanalysis [106]. We artificially correlated four variables with extremes (2m temperature, total cloud cover, total precipitation, and volumetric soil moisture) and kept two variables uncorrelated (albedo and relative humidity).

Comparison to the baselines. We compare the new approach to interpretable forecasting approaches using integrated gradients [109] and to 8 baselines from 3 different categories of anomaly detection approaches; one-class unsupervised [39, 110, 51], reconstruction-based [59, 65], and multiple instance learning [94–96]. We also compare to a naive baseline which labels all variables as drivers for pixels where extreme events occur. The implementation details of these baselines are given in Appendix Sec. E.

The quantitative results are shown in Table 1. The naive baseline is impacted by two main issues; first by the time delay between drivers and extreme events, and second not all variables are correlated with the extremes. The second issue affects the one-class and reconstruction-based baselines where

Table 1: Driver detection results on the synthetic CERRA reanalysis. The best performance on each metric is highlighted in a bold text. (±) denotes the standard deviation for 3 runs.

		Validation			Testing		
	Algorithm	F1-score (↑)	IoU (↑)	OA (↑)	F1-score (↑)	IoU (↑)	OA (↑)
	Naive	47.93	31.52	98.61	51.24	34.45	98.37
	Integrated Gradients I [109]	24.15±9.94	14.12±6.71	92.18±2.94	23.14±7.05	13.27±4.65	91.58±2.25
	Integrated Gradients II [109]	31.23±4.40	18.58±3.05	95.26±1.03	30.34±4.27	17.95±2.94	94.19±1.22
One-Class	OCSVM [39]	28.21±2.49	16.44±1.67	95.64±0.16	29.98±2.26	17.66±1.54	94.91±0.19
	IF [110]	34.99±0.56	21.28±0.42	97.16±0.02	37.16±0.67	22.84±0.51	96.61±0.03
	SimpleNet [51]	75.31±0.07	60.39±0.10	99.20±0.01	73.50±0.24	58.11±0.30	98.91±0.02
Rec.	STEALNet [59]	55.98±0.90	38.87±0.86	98.47±0.03	57.74±0.95	40.60±0.93	98.22±0.03
	UniAD [65]	47.53±0.17	31.18±0.14	97.44±0.02	49.23±0.41	32.65±0.36	97.18±0.05
MIL	DeepMIL [94]	70.68±1.61	54.68±1.91	99.22±0.03	71.54±1.60	55.72±1.92	99.09±0.04
	ARNet [95]	72.92±0.85	57.39±1.06	99.26±0.01	73.68±0.86	58.34±1.08	99.13±0.02
	RTFM [96]	60.09±0.31	42.95±0.31	98.34±0.03	61.88±0.28	44.81±0.30	98.12±0.02
	Ours*	82.78±0.53	70.63±0.78	99.51±0.02	80.44±0.70	67.28±0.97	**99.29**±0.04
	🐍 Ours†	**83.45**±0.37	**71.60**±0.54	**99.52**±0.01	**80.65**±0.24	**67.58**±0.33	**99.29**±0.01

*Video Swin Transformer backbone [100] †Mamba backbone [111]

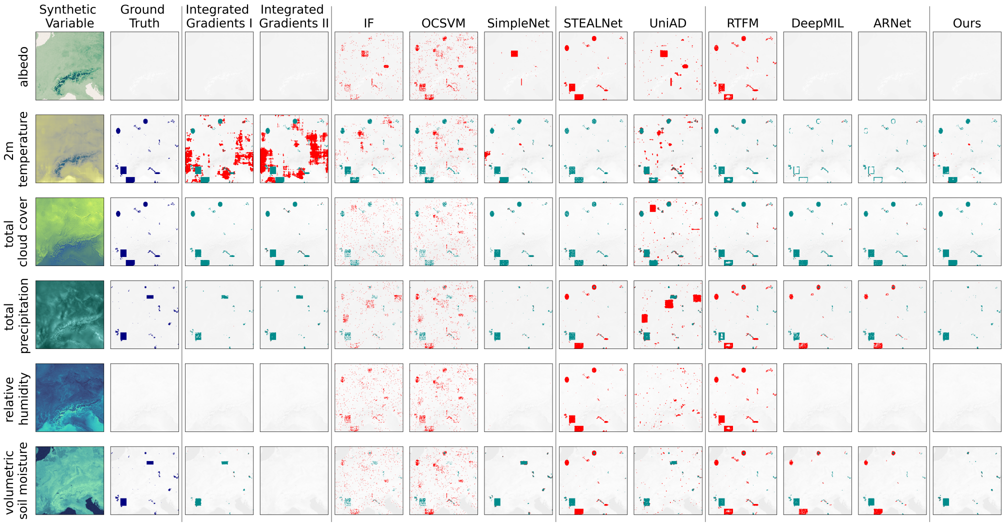


Figure 3: Qualitative results on the synthetic CERRA reanalysis from the test set at time step 2160. 🟩 is the prediction, 🟦 is the ground truth, and 🟥 is the false positive. Albedo and relative humidity are not correlated with extremes, meaning that they do not contain drivers, but only random anomalies.

they suffer mostly from false positives. In fact, both integrated gradients models achieve high F1-scores for detecting extremes (93.32 for Integrated Gradients I and 93.80 for Integrated Gradients II), but they have worse performance on identifying the drivers. SimpleNet is trained with our model as a feature extractor which explains its good performance. However, SimpleNet showed a drop of performance when it is tested on other datasets (see Appendix Sec. B for results on two more synthetic datasets). Among the reconstruction-based approach, STEALNet outperforms UniAD. This is probability because STEALNet exploits more weakly supervision information during training by maximizing the reconstruction loss for locations with extreme flags. MIL-based baselines are more suitable for the task. Finally, our model consistently outperforms the baselines on all metrics.

Qualitative samples in comparison with baselines are presented in Fig. 3. The qualitative examples indicate that our model and the MIL-based baselines except RTFM are capable of learning which variables are correlated with the extremes. The main weakness of RTFM is the reliance on feature magnitudes and the cross attention module (see Appendix Sec. E and Table 2), which make it more prone to produce false positives. Other baselines predict incorrect relations between the variables and extremes. Regarding the explainable AI methods, when we add more interactions between the variables (Integrated Gradients II), the gradients tend to omit some variables (soil moisture). Both integrated gradients models have also difficulties with the synthetic t2m, which includes red noise by design. These results demonstrate that networks that predict the extremes directly from the input variables utilize much more information even when it is not correlated with an extreme. It is thus beneficial to introduce a bottleneck into the network that enforces the network to explicitly identify drivers of extremes.

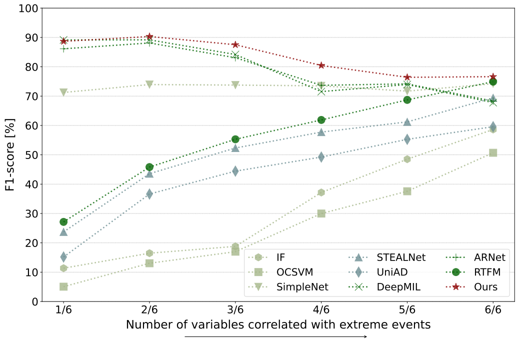


Figure 4: F1-score with different correlation settings between the input variables and extremes.

Performance on easy-to-hard correlation settings. We conduct an additional experiment to assess the model performance in relation to the correlation setup between the variables and extremes. We generate a synthetic CERRA dataset starting with only one correlated variable with the target extreme. We then generate different versions of the dataset by increasing the number of correlated variables with the extremes up to 100%. This analysis allows us to point out the strengths and weaknesses of

Table 2: Ablation studies from the validation set. The metric is F1 on the driver/extreme detection.

(a) Loss function

$\mathcal{L}_{(quantize)}$	$\mathcal{L}_{(driver)}$	$\mathcal{L}_{(extreme)}$ multi-head	F1-score (↑)
✓	✗	✗	29.99 / 48.94
✓	✗	✓	02.51 / 92.88
✓	✓	✗	31.23 / 66.40
✓	✓	✓	**82.78 / 92.45**

(b) Key model architecture

Temporal-attention	Cross-attention	Shared-(f_θ)	F1-score (↑)
✓	✗	✓	69.59 / 90.42
✗	✓	✗	67.18 / 93.78
✗	✗	✗	82.39 / 91.97
✓	✗	✗	**82.78 / 92.45**

(c) T dimension

T	F1-score (↑)
1	83.38 / 69.25
4	81.36 / 90.68
6	**82.78 / 92.45**
8	77.33 / 90.30

the comparative models for different scenarios and where our model becomes more effective, as well as where it could struggle most. The results are shown in Fig. 4. One-class, reconstruction-based and RTFM baselines benefit with increasing the number of correlated variables. In case of 6/6, the task reduces to an anomaly detection task. The performance of our model and MIL-based baselines generally decreases when the number of correlated variables increases as the task of finding all correlated anomalies becomes harder. Nevertheless, our approach performs best in all settings.

Ablation study. We conducted a set of ablation studies. This includes three main experiments:

Loss functions. As shown in Table 2 (a), $\mathcal{L}_{(quantize)}$ and $\mathcal{L}_{(driver)}$ are essential for training. As other quantization models, ours can not be trained without $\mathcal{L}_{(quanitze)}$, which ensures that the outputs of f_θ do not grow and commit to the binary embedding. $\mathcal{L}_{(driver)}$ unifies the representation of drivers for all variable as $q = 1$, which boosts the extreme detection. Moreover, the results demonstrate the impact of using $V+1$ 3D CNNs (multi-head) instead of one for $\mathcal{L}_{(extreme)}$. If a single 3D CNN is used, drivers are only identified in a small subset of variables. We discuss this more in detail in Sec. C.7.

Feature extractor. In Table 2 (b), we show the benefit of having independent feature extractors for driver detection. In a first experiment, we share the feature extractor f_θ among the variables. The performance is worst. Second, we replaced the temporal attention by a cross attention between the variables similar to [112] and [113] where each variable performs a cross attention with the other variables. We can see a drop of performance for the second experiment. We noticed that anomalies propagate between variables when adding connections in the feature extraction stage. This also explains the poor performance of RTFM compared to other MIL baselines. The best performance is shown for the proposed setup (last row), which also shows the benefit for the temporal attention.

Temporal resolution T. Table 2 (c) evaluates the impact of the temporal resolution on driver and extreme detection. T=6 provides a good balance between driver and extreme detection. More ablation studies on other aspects of the model design can be found in Appendix Sec. C.

5.2 Experiments on real-world datasets

We evaluate our model on two reanalysis data with diverse geographical and climate regions (Sec. 4.2). The input for the experiment is the normalized mean and standard deviation of each week. We exclude pixels over water surfaces, desert, and snow. In addition to the quantitative evaluation on the synthetic data, we aim to verify our method considering the following aspects:

Quantitative results. We expect that the developed model can identify drivers in real-world scenarios. We demonstrate this by measuring how well the model can predict extreme agricultural droughts from the identified drivers. The results verify that the model can predict the droughts across different regions and datasets (see Appendix Sec. D and Table 14). Note that compared to the synthetic dataset, the real-world drought prediction is much more difficult.

Extreme detection without anomaly detection. We trained the model without the quantization step, meaning without driver detection. This can be considered as an upper bound on the extreme detection accuracy since there is no information reduction by the quantization. We found that when we trained on the synthetic and real-world EUR-11 data, the F1-score for detecting extreme events increased only by $\sim 0.96\%$ and $\sim 1.93\%$, respectively, compared to the model with quantization (see Appendix Table 12). This verifies that the detected drivers are highly correlated with the extremes.

Qualitative results and spatial distribution. In Fig. 5 (a), we show the spatial distribution of the identified drivers at a specific time over EUR-11. Shown are the identified drivers up to 7 weeks

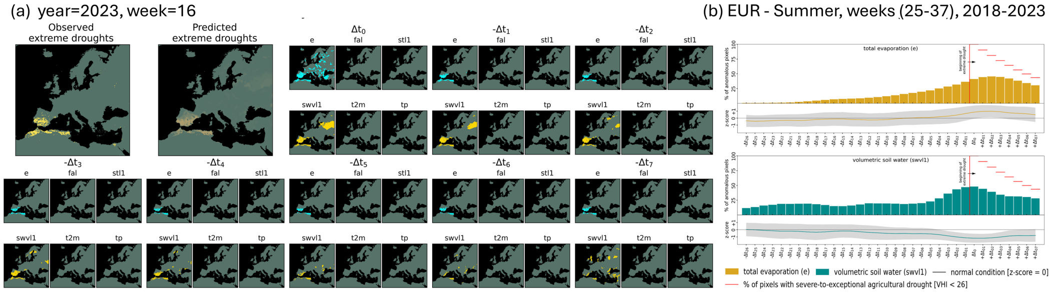


Figure 5: (a) Qualitative results on ERA5-Land over the EUR-11 domain. Shown are the identified drivers localized spatio-temporally 7 weeks before the extreme agricultural drought events. (b) Temporal evolution of drivers during the extremes.

(Δt_{-7}) before the extreme agricultural droughts at time Δt_0. We can see that the prediction of drivers and extremes are spatially correlated with the ground truth.

Physical consistency. In Fig. 5 (b), we show the relation between the input reanalysis data, extreme droughts, and identified drivers. For this experiment, we selected pixels with extreme events during summer (weeks 25-38) and visualize the average distribution of drivers with time. The red line at Δt_0 indicates the beginning of the extreme droughts. Z_{score} in the underneath curve represents the deviation from the mean computed from the climatology. It is expected that evaporation reduces soil moisture, which dries out the soil and vegetation [4]. Our model indicates that over Europe, the evaporation and soil moisture are the most informative variables to detect drivers related to extreme droughts. All pixels experienced a pronounced decline in soil moisture and an increase in evaporation as the events evolved. Please see Sec. D.3 for more discussion on the scientific validity.

6 Discussions and conclusions

We introduced a model that can identify the spatio-temporal relations between impacts of extreme events and their drivers. For this, we assumed that there exist precursor drivers, primarily as anomalies in assimilated land surface and atmospheric data, for every observable impact of extremes. We demonstrated the effectiveness of our approach by measuring to which degree the identified drivers can be used to predict extreme agricultural droughts. Apart from experiments on two real-world datasets, we also presented a new framework to generate synthetic datasets that can be used for spatio-temporal anomaly detection and climate research. The results on the synthetic datasets show that the approach is not limited to droughts and can be applied to other extremes. While we have shown that our approach outperforms other approaches, the study has some limitations. First, evaluating ability to handle a very large number of climate variables in a unified model needs further examination. Similarly, performing an additional pre-processing of specific variables like accumulating precipitation over many weeks might also improve the results. Second, modelling the temporal relations is limited by the time window T. Moreover, teleconnections of climatic anomalies can occur in distant regions on Earth, e.g., affects of El Niño and La Niña variability on drought and flood [12]. Modelling and disentangle such large spatio-temporal relations across the globe is an open research problem. Third, it would be appealing to provide scores for drivers instead of a binary classification. This could be achieved by measuring the distance to the nearest code in the LFQ. Forth, the prediction of the model depends on the capacity of reanalysis data to accurately represent the local environmental factors and land–atmospheric feedbacks. Most importantly, drawing conclusions on drivers from weak predictive models may lead to unreliable interpretations. Finally, our model does not identify causal relationships.

Despite these limitations, our approach demonstrates a clear capability in identifying drivers and anomalies in climate data which would allow a more timely event attributions during and right after extreme events. The identified spatio-temporal relations between extreme events and their drivers could support the understanding and forecasting of extremes.

7 Acknowledgments and Disclosure of Funding

We thank Petra Friederichs, Till Fohrmann, Sebastian Buschow, and Svenja Szemkus for the insightful discussions on detection and attribution of weather and climate extremes. We would also like to thank Olga Zatsarynna and Emad Bahrami for the technical discussion related to feature extraction. Finally, we thank the four anonymous reviewers for their comments and suggestions which improved the quality of this paper.

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1502/1–2022 – project no. 450058266 within the Collaborative Research Center (CRC) for the project Regional Climate Change: Disentangling the Role of Land Use and Water Management (DETECT) and by the Federal Ministry of Education and Research (BMBF) under grant no. 01IS24075C RAINA.

We acknowledge EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy, through EuroHPC Regular Access Call - proposal No. EHPC-REG-2024R01-076. The authors also gratefully acknowledge the granted access to the Marvin cluster hosted by the University of Bonn.

References

[1] Stephanie C. Herring, Nikolaos Christidis, Andrew Hoell, and Peter A. Stott. Explaining extreme events of 2020 from a climate perspective. *Bulletin of the American Meteorological Society*, 103(3):S1 – S129, 2022.

[2] Jana Sillmann, Thordis Thorarinsdottir, Noel Keenlyside, Nathalie Schaller, Lisa V. Alexander, Gabriele Hegerl, Sonia I. Seneviratne, Robert Vautard, Xuebin Zhang, and Francis W. Zwiers. Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather and Climate Extremes*, 18:65–74, 2017.

[3] M. B. Ek and A. A. M. Holtslag. Influence of soil moisture on boundary layer cloud development. *Journal of Hydrometeorology*, 5(1):86 – 99, 2004.

[4] Diego G. Miralles, Pierre Gentine, Sonia I. Seneviratne, and Adriaan J. Teuling. Land–atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges. *Annals of the New York Academy of Sciences*, 1436(1):19–35, 2019.

[5] JL Geirinhas, AC Russo, R Libonati, DG Miralles, AM Ramos, L Gimeno, and RM Trigo. Combined large-scale tropical and subtropical forcing on the severe 2019–2022 drought in south america. *npj Climate and Atmospheric Science*, 6(1):185, 2023.

[6] Shengzhi Huang, Qiang Huang, Jianxia Chang, Guoyong Leng, and Li Xing. The response of agricultural drought to meteorological drought and the influencing factors: A case study in the wei river basin, china. *Agricultural Water Management*, 159:45–54, 2015.

[7] Fei Wang, Hexin Lai, Yanbin Li, Kai Feng, Zezhong Zhang, Qingqing Tian, Xiaomeng Zhu, and Haibo Yang. Dynamic variation of meteorological drought and its relationships with agricultural drought across china. *Agricultural Water Management*, 261:107301, 2022.

[8] Meng Dai, Shengzhi Huang, Qiang Huang, Xudong Zheng, Xiaoling Su, Guoyong Leng, Ziyan Li, Yi Guo, Wei Fang, and Yongjia Liu. Propagation characteristics and mechanism from meteorological to agricultural drought in various seasons. *Journal of Hydrology*, 610:127897, 2022.

[9] Shengpeng Cao, Lifeng Zhang, Yi He, Yali Zhang, Yi Chen, Sheng Yao, Wang Yang, and Qiang Sun. Effects and contributions of meteorological drought on agricultural drought under different climatic zones and vegetation types in northwest china. *Science of The Total Environment*, 821:153270, 2022.

[10] Peter A. Stott, Nikolaos Christidis, Friederike E. L. Otto, Ying Sun, Jean-Paul Vanderlinden, Geert Jan van Oldenborgh, Robert Vautard, Hans von Storch, Peter Walton, Pascal Yiou, and Francis W. Zwiers. Attribution of extreme weather and climate-related events. *WIREs Climate Change*, 7(1):23–41, 2016.

[11] Zengchao Hao, Vijay P. Singh, and Youlong Xia. Seasonal drought prediction: Advances, challenges, and future prospects. *Reviews of Geophysics*, 56(1):108–141, 2018.

[12] Mahashweta Das and Srinivasan Parthasarathy. Anomaly detection and spatio-temporal analysis of global climate system. In *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*, SensorKDD '09, page 142–150, New York, NY, USA, 2009. Association for Computing Machinery.

[13] Sancho Salcedo-Sanz, Jorge Pérez-Aracil, Guido Ascenso, Javier Del Ser, David Casillas-Pérez, Christopher Kadow, Dušan Fister, David Barriopedro, Ricardo García-Herrera, Matteo Giuliani, et al. Analysis, characterization, prediction, and attribution of extreme atmospheric events with machine learning and deep learning techniques: a review. *Theoretical and Applied Climatology*, 155(1):1–44, 2024.

[14] Yunjie Liu, Evan Racah, Joaquin Correa, Amir Khosrowshahi, David Lavers, Kenneth Kunkel, Michael Wehner, William Collins, et al. Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv preprint arXiv:1605.01156*, 2016.

[15] G. Sepulcre-Canto, S. Horion, A. Singleton, H. Carrao, and J. Vogt. Development of a combined drought indicator to detect agricultural drought in europe. *Natural Hazards and Earth System Sciences*, 12(11):3519–3531, 2012.

[16] Yu Zhang, Zengchao Hao, Sifang Feng, Xuan Zhang, Yang Xu, and Fanghua Hao. Agricultural drought prediction in china based on drought propagation and large-scale drivers. *Agricultural Water Management*, 255:107028, 2021.

[17] Xiang Zhang, Nengcheng Chen, Jizhen Li, Zhihong Chen, and Dev Niyogi. Multi-sensor integrated framework and index for agricultural drought monitoring. *Remote Sensing of Environment*, 188:141–163, 2017.

[18] Yohannes Yihdego, Babak Vaheddoost, and Radwan A Al-Weshah. Drought indices and indicators revisited. *Arabian Journal of Geosciences*, 12:1–12, 2019.

[19] Mohamed Dahoui. Use of machine learning for the detection and classification of observation anomalies, 01/2023 2023.

[20] Peter Düben, Umberto Modigliani, Alan Geer, Stephan Siemen, Florian Pappenberger, Peter Bauer, Andy Brown, Martin Palkovic, Baudouin Raoult, Nils Wedi, and Vasileios Baousis. Machine learning at ecmwf: A roadmap for the next 10 years, 01/2021 2021.

[21] M. Flach, F. Gans, A. Brenning, J. Denzler, M. Reichstein, E. Rodner, S. Bathiany, P. Bodesheim, Y. Guanche, S. Sippel, and M. D. Mahecha. Multivariate anomaly detection for earth observations: a comparison of algorithms and feature extraction techniques. *Earth System Dynamics*, 8(3):677–696, 2017.

[22] Evan Racah, Christopher Beckham, Tegan Maharaj, Samira Ebrahimi Kahou, Mr Prabhat, and Chris Pal. 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.

[23] Prabhat, K. Kashinath, M. Mudigonda, S. Kim, L. Kapp-Schwoerer, A. Graubner, E. Karaismailoglu, L. von Kleist, T. Kurth, A. Greiner, A. Mahesh, K. Yang, C. Lewis, J. Chen, A. Lou, S. Chandran, B. Toms, W. Chapman, K. Dagon, C. A. Shields, T. O'Brien, M. Wehner, and W. Collins. Climatenet: an expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development*, 14(1):107–124, 2021.

[24] Elizabeth Weirich-Benet, Maria Pyrina, Bernat Jiménez-Esteve, Ernest Fraenkel, Judah Cohen, and Daniela I. V. Domeisen. Subseasonal prediction of central european summer heatwaves with linear and random forest machine learning models. *Artificial Intelligence for the Earth Systems*, 2(2):e220038, 2023.

[25] Adam B. Barrett, Steven Duivenvoorden, Edward E. Salakpi, James M. Muthoka, John Mwangi, Seb Oliver, and Pedram Rowhani. Forecasting vegetation condition for drought early warning systems in pastoral communities in kenya. *Remote Sensing of Environment*, 248:111886, 2020.

[26] Christian Requena-Mesa, Vitus Benson, Markus Reichstein, Jakob Runge, and Joachim Denzler. 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 (CVPR) Workshops*, pages 1132–1142, June 2021.

[27] Vitus Benson, Claire Robin, Christian Requena-Mesa, Lazaro Alonso, Nuno Carvalhais, José Cortés, Zhihan Gao, Nora Linscheid, Mélanie Weynants, and Markus Reichstein. Multi-modal learning for geospatial vegetation forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27788–27799, June 2024.

[28] M. H. Shams Eddin and J. Gall. Focal-tsmp: deep learning for vegetation health prediction and agricultural drought assessment from a regional climate simulation. *Geoscientific Model Development*, 17(7):2987–3023, 2024.

[29] Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

[30] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[31] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

[32] Xavier-Andoni Tibau, Christian Reimers, Andreas Gerhardus, Joachim Denzler, Veronika Eyring, and Jakob Runge. A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections. *Environmental Data Science*, 1:e12, 2022.

[33] Julien Boussard, Chandni Nagda, Julia Kaltenborn, Charlotte Emilie Elektra Lange, Philippe Brouillard, Yaniv Gurwicz, Peer Nowack, and David Rolnick. Towards causal representations of climate model data. *arXiv preprint arXiv:2312.02858*, 2023.

[34] Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

[35] Raanan Yehezkel Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. From temporal to contemporaneous iterative causal discovery in the presence of latent confounders. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39939–39950. PMLR, 23–29 Jul 2023.

[36] Jose María Tárraga, Eva Sevillano-Marco, Jordi Muñoz-Marí, María Piles, Vasileios Sitokonstantinou, Michele Ronco, María Teresa Miranda, Jordi Cerdà, and Gustau Camps-Valls. Causal discovery reveals complex patterns of drought-induced displacement. *iScience*, 27(9):110628, 2024.

[37] J. Vogel, P. Rivoire, C. Deidda, L. Rahimi, C. A. Sauter, E. Tschumi, K. van der Wiel, T. Zhang, and J. Zscheischler. Identifying meteorological drivers of extreme impacts: an application to simulated crop yields. *Earth System Dynamics*, 12(1):151–172, 2021.

[38] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019.

[39] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999.

[40] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. In *International Conference on Learning Representations*, 2021.

[41] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[42] Shenzhi Wang, Liwei Wu, Lei Cui, and Yujun Shen. Glancing at the patch: Anomaly localization with global and local feature comparison. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 254–263, 2021.

[43] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *Pattern Recognition. ICPR International Workshops and Challenges*, pages 475–489, Cham, 2021. Springer International Publishing.

[44] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9737–9746, June 2022.

[45] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021.

[46] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020.

[47] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. *arXiv preprint arXiv:2103.04257*, 2021.

[48] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2806–2814, June 2021.

[49] Oliver Rippel, Patrick Mertens, Eike König, and Dorit Merhof. Gaussian anomaly detection by modeling the distribution of normal data in pretrained deep features. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021.

[50] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14298–14308, 2022.

[51] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20402–20411, 2023.

[52] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. ADBench: Anomaly detection benchmark. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[53] Laurenz Strothmann, Uwe Rascher, and Ribana Roscher. Detection of anomalous grapevine berries using all-convolutional autoencoders. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3701–3704, 2019.

[54] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. In *30th IEEE/IES International Symposium on Industrial Electronics (ISIE)*, June 2021.

[55] Yunseung Lee and Pilsung Kang. Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder. *IEEE Access*, 10:46717–46724, 2022.

[56] Miro Miranda, Laura Zabawa, Anna Kicherer, Laurenz Strothmann, Uwe Rascher, and Ribana Roscher. Detection of anomalous grapevine berries using variational autoencoders. *Frontiers in Plant Science*, 13, 2022.

[57] Mohammad Sabokrou, Masoud Pourreza, Mohsen Fayyaz, Rahim Entezari, Mahmood Fathy, Jürgen Gall, and Ehsan Adeli. Avid: Adversarial visual irregularity detection. In C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 488–505, Cham, 2019. Springer International Publishing.

[58] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. Attention guided anomaly localization in images. In *European Conference on Computer Vision*, pages 485–503. Springer, 2020.

[59] Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 207–214, 2021.

[60] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12147–12156, 2023.

[61] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14592–14601, 2023.

[62] Neelu Madan, Nicolae-Cătălin Ristea, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah. Self-supervised masked convolutional transformer block for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):525–542, 2024.

[63] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 649–655, 2022.

[64] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2404.06564*, 2024.

[65] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022.

[66] Chaoqin Huang, Qinwei Xu, Yanfeng Wang, Yu Wang, and Ya Zhang. Self-supervised masking for unsupervised anomaly detection and localization. *IEEE Transactions on Multimedia*, 25:4426–4438, 2023.

[67] Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11701–11708, 2020.

[68] Hannah M. Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 474–489, Cham, 2022. Springer Nature Switzerland.

[69] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*, pages 494–511. Springer, 2022.

[70] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021.

[71] Anne-Sophie Collin and Christophe De Vleeschouwer. Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7915–7922. IEEE, 2021.

[72] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9659–9669, 2021.

[73] Masoud Pourreza, Bahram Mohammadi, Mostafa Khaki, Samir Bouindour, Hichem Snoussi, and Mohammad Sabokrou. G2d: Generate to detect anomaly. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2003–2012, 2021.

[74] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2020.

[75] Kamalakar Vijay Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 5541–5550, 2023.

[76] Jingtao Li, Xinyu Wang, Hengwei Zhao, Liangpei Zhang, and Yanfei Zhong. A unified remote sensing anomaly detector across modalities and scenes via deviation relationship learning. *arXiv preprint arXiv:2310.07511*, 2023.

[77] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. *arXiv preprint arXiv:2403.05897*, 2024.

[78] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.

[79] Dongha Lee, Sehun Yu, Hyunjun Ju, and Hwanjo Yu. Weakly supervised temporal anomaly segmentation with dynamic time warping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7355–7364, 2021.

[80] Zhijie Zhong, Zhiwen Yu, Yiyuan Yang, Weizheng Wang, and Kaixiang Yang. Patchad: Patch-based mlp-mixer for time series anomaly detection. *arXiv preprint arXiv:2401.09793*, 2024.

[81] Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. *Proceedings of VLDB*, 15(6):1201–1214, 2022.

[82] Yu Zheng, Huan Yee Koh, Ming Jin, Lianhua Chi, Khoa T. Phan, Shirui Pan, Yi-Ping Phoebe Chen, and Wei Xiang. Correlation-aware spatial–temporal graph learning for multivariate time-series anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):11802–11816, 2024.

[83] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[84] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 358–376. Springer, 2020.

[85] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE Transactions on Image Processing*, 30:4505–4515, 2021.

[86] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021.

[87] Seongheon Park, Hanjae Kim, Minsu Kim, Dahye Kim, and Kwanghoon Sohn. Normality guided multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2665–2674, January 2023.

[88] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3769–3777, 2023.

[89] Zhenting Qi, Ruike Zhu, Zheyu Fu, Wenhao Chai, and Volodymyr Kindratenko. Weakly supervised two-stage training scheme for deep video fight detection model. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 677–685. IEEE, 2022.

[90] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 729–745, Cham, 2022. Springer Nature Switzerland.

[91] Hitesh Sapkota and Qi Yu. Bayesian nonparametric submodular video partition for robust anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3212–3221, 2022.

[92] Minqi Jiang, Chaochuan Hou, Ao Zheng, Xiyang Hu, Songqiao Han, Hailiang Huang, Xiangnan He, Philip S Yu, and Yue Zhao. Weakly supervised anomaly detection: A survey. *arXiv preprint arXiv:2302.04549*, 2023.

[93] Tao Jiang, Weiying Xie, Yunsong Li, Jie Lei, and Qian Du. Weakly supervised discriminative learning with spectral constrained generative adversarial network for hyperspectral anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6504–6517, 2022.

[94] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[95] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020.

[96] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4955–4966, 2021.

[97] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 387–395, 2023.

[98] Yixuan Zhou, Yi Qu, Xing Xu, Fumin Shen, Jingkuan Song, and Hengtao Shen. Batchnorm-based weakly supervised video anomaly detection. *arXiv preprint arXiv:2311.15367*, 2023.

[99] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8022–8031, 2023.

[100] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[101] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[102] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation, 2023.

[103] Aren Jansen, Daniel PW Ellis, Shawn Hershey, R Channing Moore, Manoj Plakal, Ashok C Popat, and Rif A Saurous. Coincidence, categorization, and consolidation: Learning to recognize sounds with minimal supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–125. IEEE, 2020.

[104] Wenze Yang, Felix Kogan, and Wei Guo. An ongoing blended long-term vegetation health product for monitoring global food security. *Agronomy*, 10(12), 2020.

[105] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[106] S. Schimanke, M. Ridal, P. Le Moigne, L. Berggren, P. Undén, R. Randriamampianina, U. Andrea, E. Bazile, A. Bertelsen, P. Brousseau, P. Dahlgren, L. Edvinsson, A. El Said, M. Glinton, S. Hopsch, L. Isaksson, R. Mladek, E. Olsson, A. Verrelle, and Z.Q. Wang. Cerra sub-daily regional reanalysis data for europe on single levels from 1984 to present, 2021.

[107] J. Muñoz Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach, B. Martens, D. G. Miralles, M. Piles, N. J. Rodríguez-Fernández, E. Zsoter, C. Buontempo, and J.-N. Thépaut. Era5-land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383, 2021.

[108] F. Giorgi, C. Jones, and G. R. Asrar. Addressing climate information needs at the regional level: the cordex framework. *Bulletin - World Meteorological Organization*, 58(3):175–183, 2009.

[109] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

[110] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39, 2012.

[111] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[112] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[113] Christian Lessig, Ilaria Luise, Bing Gong, Michael Langguth, Scarlet Stadler, and Martin Schultz. Atmorep: A stochastic model of atmosphere dynamics using large scale representation learning. *arXiv preprint arXiv:2308.13280*, 2023.

[114] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2022.

[115] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *International Conference on Learning Representations*, 2022.

[116] Woncheol Shin, Gyubok Lee, Jiyoung Lee, Eunyi Lyou, Joonseok Lee, and Edward Choi. Exploration into translation-equivariant image quantization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[117] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *The Twelfth International Conference on Learning Representations*, 2024.

[118] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

[119] Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024.

[120] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024.

[121] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024.

[122] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024.

[123] Jinyoung Park, Hee-Seon Kim, Kangwook Ko, Minbeom Kim, and Changick Kim. Videomamba: Spatio-temporal selective state space model. *arXiv preprint arXiv:2407.08476*, 2024.

[124] Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Björn Ommer. Zigma: A dit-style zigzag mamba diffusion model. In *Arxiv*, 2024.

[125] Almudena García-García, Francisco José Cuesta-Valero, Diego G Miralles, Miguel D Mahecha, Johannes Quaas, Markus Reichstein, Jakob Zscheischler, and Jian Peng. Soil heat extremes can outpace air temperature extremes. *Nature Climate Change*, 13(11):1237–1241, 2023.

[126] Preet Lal, Gurjeet Singh, Narendra N. Das, Andreas Colliander, and Dara Entekhabi. Assessment of era5-land volumetric soil water layer product using in situ and smap soil moisture observations. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[127] Dara Entekhabi, Eni G. Njoku, Peggy E. O'Neill, Kent H. Kellogg, Wade T. Crow, Wendy N. Edelstein, Jared K. Entin, Shawn D. Goodman, Thomas J. Jackson, Joel Johnson, John Kimball, Jeffrey R. Piepmeier, Randal D. Koster, Neil Martin, Kyle C. McDonald, Mahta Moghaddam, Susan Moran, Rolf Reichle, J. C. Shi, Michael W. Spencer, Samuel W. Thurman, Leung Tsang, and Jakob Van Zyl. The soil moisture active passive (smap) mission. *Proceedings of the IEEE*, 98(5):704–716, 2010.

[128] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

[129] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[130] Xianfeng Liu, Xiufang Zhu, Yaozhong Pan, Shuangshuang Li, Yanxu Liu, and Yuqi Ma. Agricultural drought monitoring: Progress, challenges, and prospects. *Journal of Geographical Sciences*, 26:750–767, 2016.

[131] I. Meza, S. Siebert, P. Döll, J. Kusche, C. Herbert, E. Eyshi Rezaei, H. Nouri, H. Gerdener, E. Popat, J. Frischen, G. Naumann, J. V. Vogt, Y. Walz, Z. Sebesvari, and M. Hagenlocher. Global-scale drought risk assessment for agricultural systems. *Natural Hazards and Earth System Sciences*, 20(2):695–712, 2020.

[132] María Pedro-Monzonís, Abel Solera, Javier Ferrer, Teodoro Estrela, and Javier Paredes-Arquiola. A review of water scarcity and drought indexes in water resources planning and management. *Journal of Hydrology*, 527:482–493, 2015.

[133] Compton J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150, 1979.

[134] F.N. Kogan. Application of vegetation index and brightness temperature for drought detection. *Advances in Space Research*, 15(11):91–100, 1995. Natural Hazards: Monitoring and Assessment Using Remote Sensing Technique.

[135] Felix Kogan, Wei Guo, and Aleksandar Jelenak. Global vegetation health: Long-term data records. In Felix Kogan, Alfred Powell, and Oleg Fedorov, editors, *Use of Satellite and In-Situ Data to Improve Sustainability*, pages 247–255, Dordrecht, 2011. Springer Netherlands.

[136] F. N. Kogan. Remote sensing of weather impacts on vegetation in non-homogeneous areas. *International Journal of Remote Sensing*, 11(8):1405–1419, 1990.

[137] Anne F Van Loon, Tom Gleeson, Julian Clark, Albert IJM Van Dijk, Kerstin Stahl, Jamie Hannaford, Giuliano Di Baldassarre, Adriaan J Teuling, Lena M Tallaksen, Remko Uijlenhoet, et al. Drought in the anthropocene. *Nature Geoscience*, 9(2):89–91, 2016.

[138] Jiawei Zhuang, raphael dussin, André Jüling, and Stephan Rasp. JiaweiZhuang/xESMF: v0.3.0 Adding ESMF.LocStream capabilities, March 2020.

[139] Mohamad Hakam Shams Eddin and Juergen Gall. Identifying spatio-temporal drivers of extreme events [data set], 2024.

A Synthetic data

To the best of our knowledge, ground truth labels of drivers or anomalies which are correlated with extreme events impacts barely exist. This is especially the case when it comes to drivers or anomalous events which are related to a specific definition of an extreme event within the Earth system (i.e., drought or flood). For this reason, we designed a framework to generate artificial datasets which can be adopted to the task. Our framework is inspired by and related to Flach, et al. [21]. However, we differ in the following two aspects:
(1) First, we aim to generate multivariate anomalous that are correlated with a specific extreme event, while their aim is to generate anomalous that can occur simultaneously in multivariate data streams similar to [78–81]. Think of an increasing/decreasing of temperature, existing approaches are interested in temperature anomalies regardless of the subsequent extreme events they might cause, while we focus on temperature anomalies that can cause a particular extreme events in the near future.
(2) Second, we generate the synthetic data based on real-world data stream signals, while they use trigonometric functions (i.e., sine function) to mimic Earth observations across spacetime.

In the following, we explain our overall framework in more details:
(1) First, we generate the normal base signals $\mathbf{B} \in \mathbb{R}^{V \times T \times Lat \times Lon}$ for a set of different variables V, where T is the temporal extension, and Lat and Lon are the spatial extensions. For instance, to synthesize CERRA [106] $\mathbf{B}$ signals, we take the mean values from the CERRA climatologoy pixel-wise. This represents the typical value of $\mathbf{B}$ at specific time (week) and location (lat, lon). By definition, $\mathbf{B}$ inherits the intrinsic properties of the simulated variables including the existence of seasonality and correlations among variables.
(2) In the next step, we induce binary extreme events $\mathbf{E}^{ex} \in \mathbb{Z}_2^{T \times Lat \times Lon}$ within the datacube and track the precise spatio-temporal location of these events. Similar to Flach, et. al [21], the type and duration of extreme events vary within the datacube. For instance, we alter the duration of the events between long and short extreme events. The spatial distribution of the extreme event vary also between a local event at one pixel (LocalEvent), a rectangular event (CubeEvent), a Gaussian shape (GaussianEvent), an onset event that starts at specific time and lasts until the end of the series (OnsetEvent), and a random walk event that starts at a specific pixel and affecting neighboring pixels with time (RandomWalkEvent).

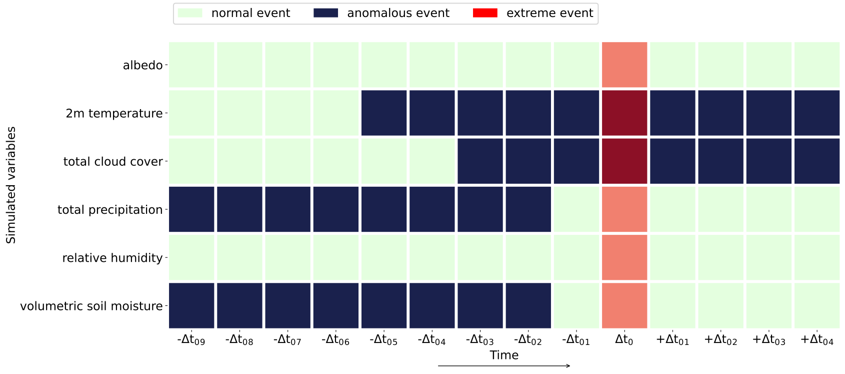


Figure 6: An example of the randomly generated coupling matrix between the synthetic variables and extremes for the synthetic CERRA dataset.

(3) We randomly define a coupling matrix $\mathbf{M}$ between the variables and the extreme event. An example is shown in Fig. 6, i.e., the anomalous events for 2m temperature start 5 time steps $(-\Delta T_a)$ before the extreme and last for 4 time steps afterwards $(+\Delta T_a)$. While albedo and relative humidity are not coupled with the extremes. Based on $\mathbf{M}$ and $\mathbf{E}^{ex}$, we generate the binary anomalous events matrix $\mathbf{E}^{a} \in \mathbb{Z}_2^{T \times Lat \times Lon}$.
(4) Similar to steps (2)-(3), we generate binary random anomalous events $\mathbf{E}^{r} \in \mathbb{Z}_2^{T \times Lat \times Lon}$. However, these events are uncorrelated with the extremes and generated randomlies for all variables.
(5) We sample noise signals $\mathbf{N} \in \mathbb{R}^{V \times T \times Lat \times Lon}$ for each variable. The noise signals could be sampled from a normal Gaussian distribution (GaussianNoise), a standard Cauchy distribution

(CauchyNoise), a double exponential distribution (LaplaceNoise), or from a spatiotemporal correlated noise across the datacube (RedNoise). (6) Using the indices $v \in \{1, \ldots, V\}$, $t \in \{1, \ldots, T\}$, $lat \in \{1, \ldots, Lat\}$, and $lon \in \{1, \ldots, Lon\}$, the synthetic signal $\mathbf{\Phi}_{(v,t,lat,lon)}$ is generated with the following formula:

$$\mathbf{\Phi}_{(v,t,lat,lon)} = \mathbf{B}_{(v,t,lat,lon)} + \mathbf{\Lambda}_{(v,t,lat,lon)} . \mathbf{\Theta}_{(v,t,lat,lon)} \,, \tag{7}$$

$$\begin{aligned} \mathbf{\Theta}_{(v,t,lat,lon)} &= \mathbf{B}_{(v,t,lat,lon)} . (2^{(kb.(\mathbf{E}^{r}_{(v,t,lat,lon)} \vee \mathbf{E}^{a}_{(v,t,lat,lon)}))} - 1) \\ &\quad + \mathbf{N}_{(v,t,lat,lon)} . 2^{(kn.(\mathbf{E}^{r}_{(v,t,lat,lon)} \vee \mathbf{E}^{a}_{(v,t,lat,lon)}))} \\ &\quad + ks.(\mathbf{E}^{r}_{(v,t,lat,lon)} \vee \mathbf{E}^{ex}_{(v,t,lat,lon)}) . \sigma_N \,, \end{aligned} \tag{8}$$

$$\mathbf{\Lambda}_{(v,t,lat,lon)} = \begin{cases} \dfrac{\mathbf{\Delta}_{(v,t,lat,lon)}}{\delta}, & \text{if } \mathbf{E}^{a}_{(v,t,lat,lon)} = 1, \\ +1, & \text{otherwise.} \end{cases} , \tag{9}$$

$$\mathbf{\Delta}_{(v,t,lat,lon)} = \begin{cases} -1, & \text{if } \mathbf{\Theta}_{(v,t,lat,lon)} \leq 0, \\ +1, & \text{otherwise,} \end{cases} , \tag{10}$$

where $\mathbf{\Theta}_{(v,t,lat,lon)}$ is the induced anomaly, kb, kn, and ks are control parameters for the events magnitudes, σ_N is the standard deviation of the noise signal, $\delta \in \{-1, 1\}$ is the predefined coupling sign with extremes from $\mathbf{M}$, and $\mathbf{\Lambda}_{(v,t,lat,lon)}$ controls the sign of the induced anomaly for extremes (i.e., a deficiency in soil moisture ($\delta = -1$) and an increased in temperature ($\delta = +1$) during an extreme drought event).

Using this framework, we generated 3 types of datasets; (1) synthetic CERRA reanalysis (Tables 3), (2) synthetic NOAA remote sensing (Tables 4), and (3) synthetic artificial data (Tables 5).

We generate the NOAA base signals from [104]. For the NOAA and artificial datasets, we also considered artificial linear (LinearCoupling) and non-linear (QuadraticCoupling) dependencies among the variables as additional data properties. For instance to generate a new dependent base signal from independent bases:

$$\mathbf{B}_{(t,lat,lon)} = \begin{cases} \sum_{v=1}^{V} w_{(v)} . B_{(v)}, & \text{if LinearCoupling,} \\ \sum_{v=1}^{V} \frac{1}{\sqrt{2}} w_{(v)} . (B^2_{(v)} - 1), & \text{otherwise,} \end{cases} , \tag{11}$$

$$\mathbf{E}^{r}_{(t,lat,lon)} = \vee_{v=1}^{V} \mathbf{E}^{r}_{(v,t,lat,lon)} \,, \tag{12}$$

where $w_{(v)} \in \mathbb{R}^V$ are weighted coefficients sampled either from a normal (NormWeight) or a Lablace (LaplaceWeight) distribution. In addition, we add an option to generate disturbed weights where weights vary spatio-temporally based on the locations of anomalies. Anomalous events that occur in one of the independent base signals propagate to the new generated dependent signals.

The third dataset (Artificial) does not depend on real-world data but rather consists of basis signals with trigonometric functions where we also add a linear latitudinal gradient (latGrad). Finally, we mask out some regions to exhibit no anomalies i.e., pixels over water surface. Each generated dataset consists of 52×46 time steps corresponding to 46 years of simulated data in 7-day intervals. Examples of the synthesized dataset can be found in Figs. 7-12. The results on the synthetic CERRA reanalysis are reported in Table 1 and the results on synthetic NOAA and the artificial data are shown Sec. B in Tables 6 and 7.

Table 3: Configurations of the synthetic CERRA reanalysis.

Synthetic CERRA reanalysis								
Ind. variables	Dep. variables	Coupled variables with extreme	Dimension	Extreme events	$-\triangle T_a$	$+\Delta T_a$	% Extreme	% Correlated anomalous
6	0	4	{lon=200, lat=200, time=52×46}	{CubeEvent(n=200, sx=35, sy=35, sz=25), RandomWalkEvent(n=1100, s=125), LocalEvent(n=2600, sz=17), GaussianEvent(n=340, sx=35, sy=35, sz=25), OnsetEvent(n=25, sx=17, sy=17, os=0.98)}	9	4	1.16	1.69 % Random anomalous 1.32
Base	Dependency	Weights	Noise	Random events	δ	kb	kn	ks
Albedo	-	-	WhiteNoise(meu=0, sigma=0.01)	{CubeEvent(n=320, sx=35, sy=35, sz=25), RandomWalkEvent(n=3000, s=125), LocalEvent(n=4000, sz=17), GaussianEvent(n=300, sx=35, sy=35, sz=25)}	-	0.30	0.20	0.50
2m Temperature	-	-	RedNoise(meu=0, sigma=0.90)	{OnsetEvent(n=18, sx=17, sy=17, os=0.98), RandomWalkEvent(n=1800, s=125), LocalEvent(n=160, sz=17), GaussianEvent(n=350, sx=35, sy=35, sz=25)}	+1	0.01	0.01	0.50
Total cloud cover	-	-	LaplaceNoise(meu=0, sigma=0.70, lambda=1)	{CubeEvent(n=300, sx=35, sy=35, sz=25), RandomWalkEvent(n=2000, s=125), LocalEvent(n=2800, sz=17), GaussianEvent(n=290, sx=35, sy=35, sz=25)}	-1	0.03	0.08	0.50
Total precipitation	-	-	WhiteNoise(meu=0, sigma=0.04)	{CubeEvent(n=320, sx=35, sy=35, sz=25), RandomWalkEvent(n=3000, s=125), LocalEvent(n=4000, sz=17), GaussianEvent(n=300, sx=35, sy=35, sz=25)}	-1	0.07	0.20	0.50
Relative humedity	-	-	CauchyNoise(meu=0, sigma=0.7)	{OnsetEvent(n=18, sx=17, sy=17, os=0.98), RandomWalkEvent(n=1800, s=125), LocalEvent(n=160, sz=17), GaussianEvent(n=350, sx=35, sy=35, sz=25)}	-	0.06	0.06	0.50
Volumetric soil moisture	-	-	WhiteNoise(meu=0., sigma=.017)	{CubeEvent(n=300, sx=35, sy=35, sz=25), RandomWalkEvent(n=2000, s=125), LocalEvent(n=2800, sz=17), GaussianEvent(n=290, sx=35, sy=35, sz=25)}	-1	0.10	0.10	0.50

Table 4: Configurations of the synthetic NOAA remote sensing.

Synthetic NOAA								
Ind. variables	Dep. variables	Coupled variables with extreme	Dimension	Extreme events	$-\triangle T_a$	$+\triangle T_a$	% Extreme	% Correlated anomalous
2	3	4	{lon=200, lat=200, time=52×46}	{CubeEvent(n=200, sx=35, sy=35, sz=25), RandomWalkEvent(n=1100, s=125), LocalEvent(n=2600, sz=17), GaussianEvent(n=340, sx=35, sy=35, sz=25), OnsetEvent(n=25, sx=17, sy=17, os=0.98)}	9	4	0.79	1.02 % Random anomalous 1.76
Base	Dependency	Weights	Noise	Random events	δ	kb	kn	ks
NDVI	-	-	WhiteNoise(meu=0, sigma=0.30)	{CubeEvent(n=320, sx=35, sy=35, sz=25), RandomWalkEvent(n=3000, s=125), LocalEvent(n=4000, sz=17), GaussianEvent(n=300, sx=35, sy=35, sz=25)}	-	0.25	0.10	0.50
BT	-	-	WhiteNoise(meu=0, sigma=0.5)	{CubeEvent(n=300, sx=35, sy=35, sz=25), RandomWalkEvent(n=2000, s=125), LocalEvent(n=2800, sz=17), GaussianEvent(n=290, sx=35, sy=35, sz=25)}	+1	0.01	0.01	0.50
-	Quadratic Coupling()	Norm Weight()	WhiteNoise(meu=0, sigma=0.65)	-	-1	0.01	0.01	0.50
-	Linear Coupling()	Laplace Weight()	WhiteNoise(meu=0, sigma=.065)	-	-	0.01	0.01	0.50
-	Linear Coupling()	Laplace Weight()	WhiteNoise(meu=0., sigma=.065)	-	+1	0.01	0.01	0.50

Table 5: Configurations of the synthetic artificial data.

Synthetic artificial								
Ind. variables	Dep. variables	Coupled variables with extreme	Dimension	Extreme events	$-\triangle T_a$	$+\triangle T_a$	% Extreme	% Correlated anomalous
3	3	4	{lon=200, lat=200, time=52×46}	{CubeEvent(n=200, sx=35, sy=35, sz=25), RandomWalkEvent(n=1100, s=125), LocalEvent(n=2600, sz=17), GaussianEvent(n=340, sx=35, sy=35, sz=25), OnsetEvent(n=25, sx=17, sy=17, os=0.98)}	9	4	1.24	1.81 % Random anomalous 2.93
Base	Dependency	Weights	Noise	Random events	δ	kb	kn	ks
SineBase(shift=0, amp=3, nOsc=46, latGrad=True)	-	-	RedNoise(meu=0, sigma=0.20)	{CubeEvent(n=320, sx=35, sy=35, sz=25), RandomWalkEvent(n=3000, s=125), LocalEvent(n=4000, sz=17), GaussianEvent(n=300, sx=35, sy=35, sz=25)}	-	0.35	0.35	0.35
CosineBase(shift=0, amp=3, nOsc=46, latGrad=True),	-	-	LaplaceNoise(meu=0, sigma=0.08, lambda=1)	{OnsetEvent(n=18, sx=17, sy=17, os=0.98), RandomWalkEvent(n=1800, s=125), LocalEvent(n=160, sz=17), GaussianEvent(n=350, sx=35, sy=35, sz=25)}	+1	0.35	0.35	0.35
ConstantBase(const=0, latGrad=True)	-	-	WhiteNoise(meu=0, sigma=0.07)	{CubeEvent(n=300, sx=35, sy=35, sz=25), RandomWalkEvent(n=2000, s=125), LocalEvent(n=2800, sz=17), GaussianEvent(n=290, sx=35, sy=35, sz=25)}	-1	0.90	0.90	0.90
-	Quadratic Coupling()	Norm Weight()	WhiteNoise(meu=0, sigma=0.65)	-	-1	0.35	0.35	0.35
-	Linear Coupling()	Laplace Weight()	WhiteNoise(meu=0, sigma=.065)	-	-	0.35	0.35	0.35
-	Linear Coupling()	Laplace Weight()	WhiteNoise(meu=0., sigma=.065)	-	-1	0.35	0.35	0.35

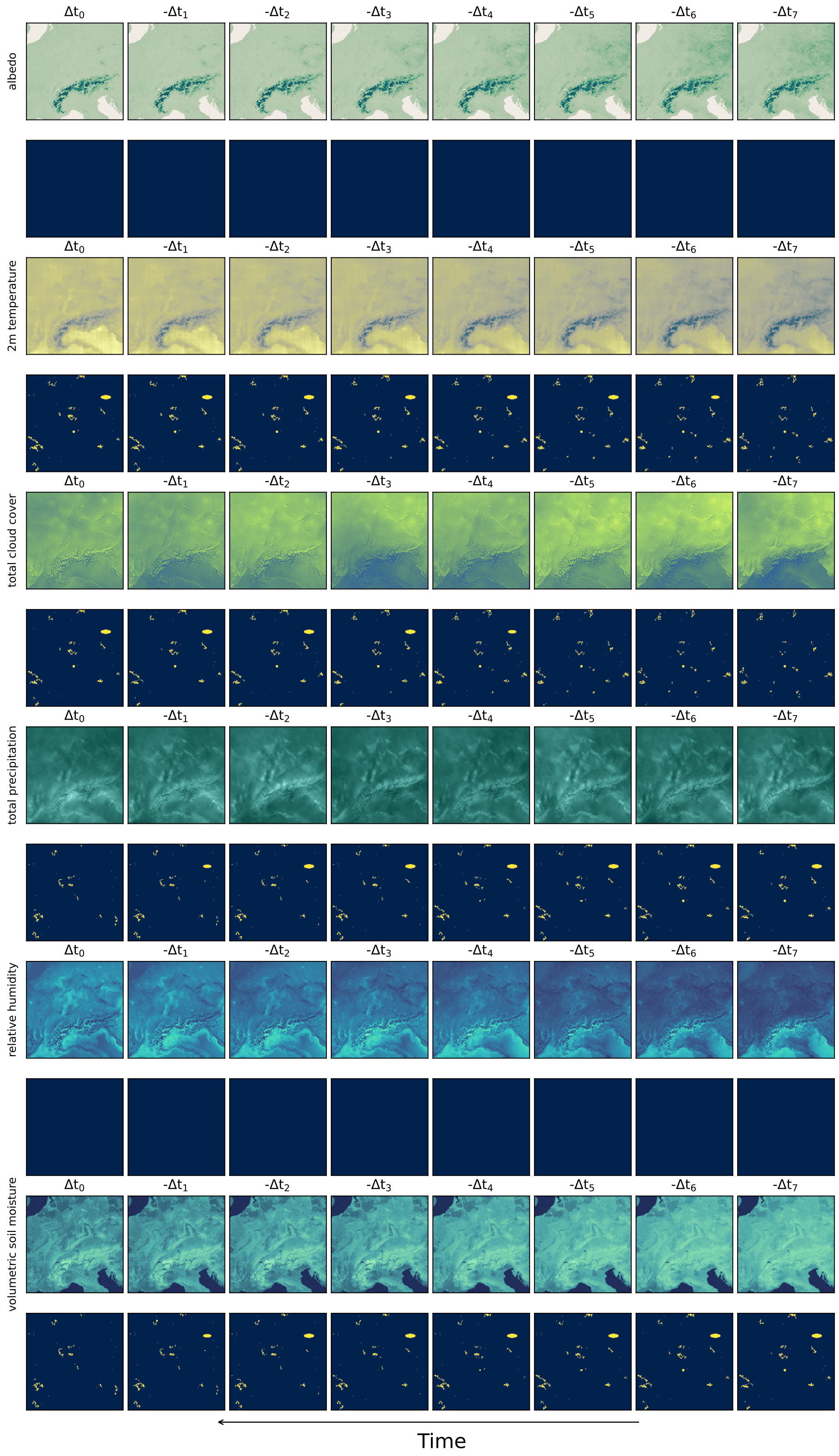


Figure 7: Examples of the synthetic CERRA reanalysis data described in Table 3. The drivers/anomalies 🟨 are visualized under each variable directly. Here, albedo and relative humidity are not correlated with the extremes.

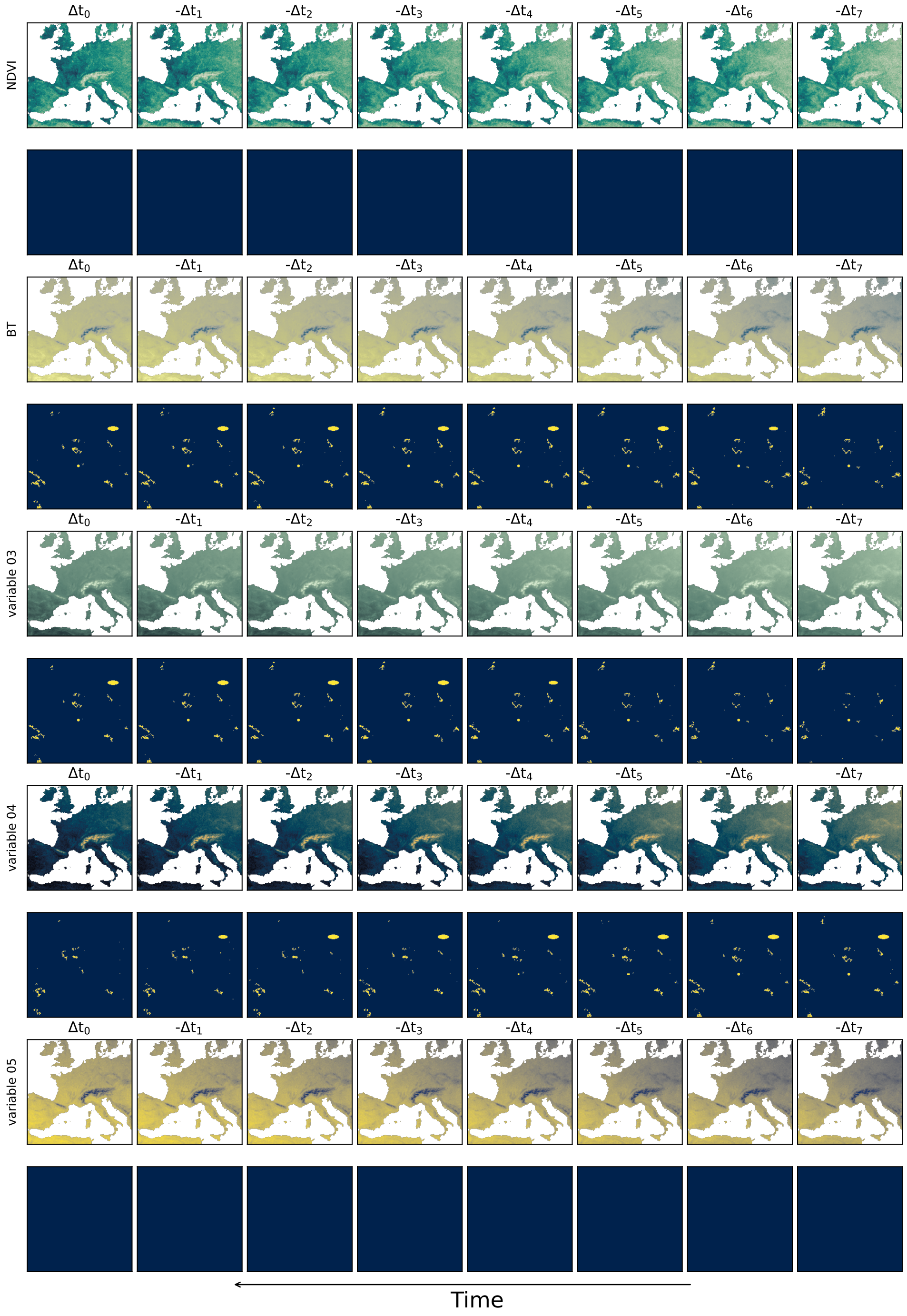


Figure 8: Examples of the synthetic NOAA remote sensing described in Table 4. The drivers/anomalies 🟨 are visualized under each variable directly. Here, NDVI and variables 05 are not correlated with the extremes.

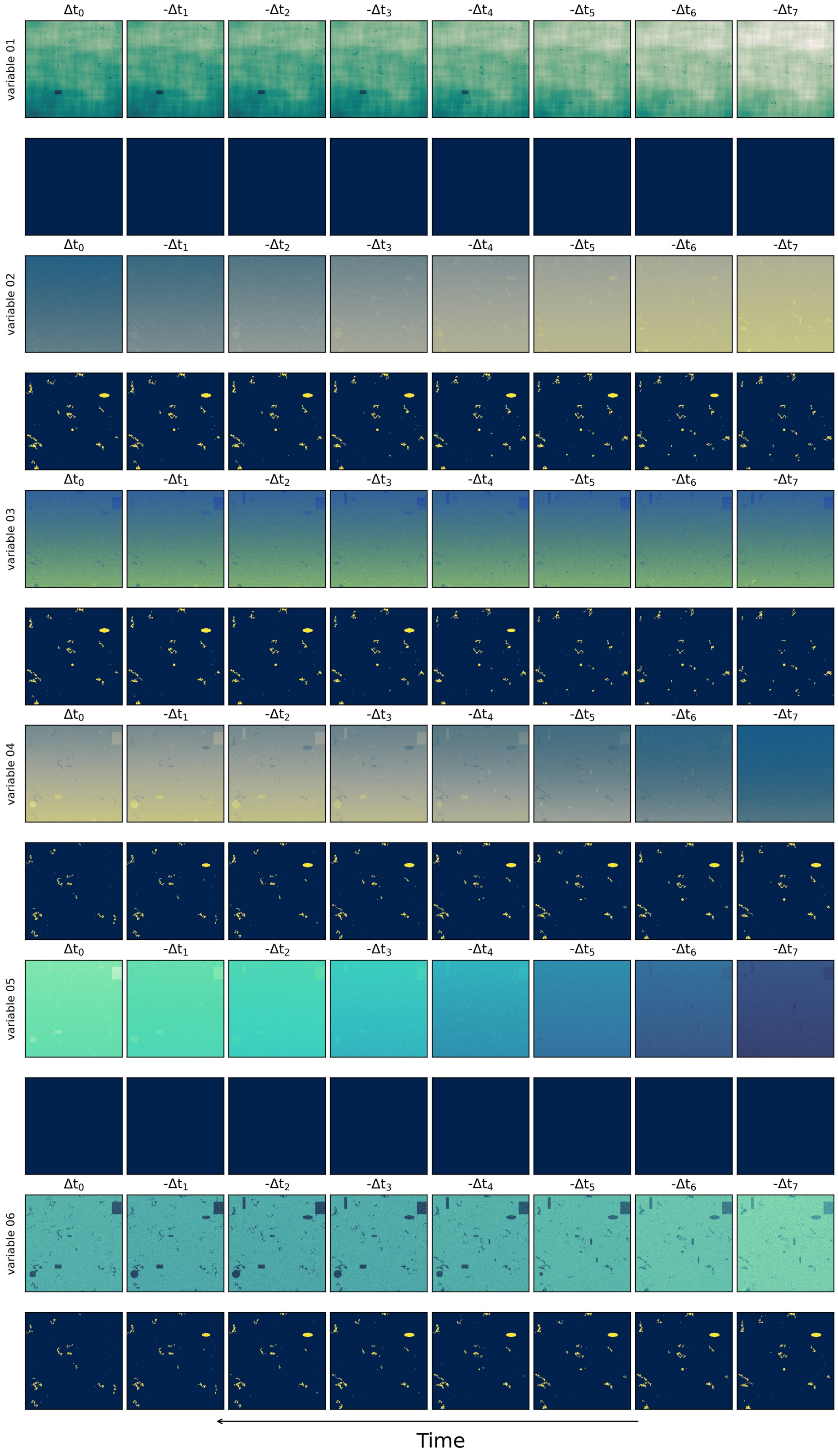


Figure 9: Examples of the synthetic artificial data described in Table 5. The drivers/anomalies 🟨 are visualized under each variable directly. Here, variables 01 and 05 are not correlated with the extremes.

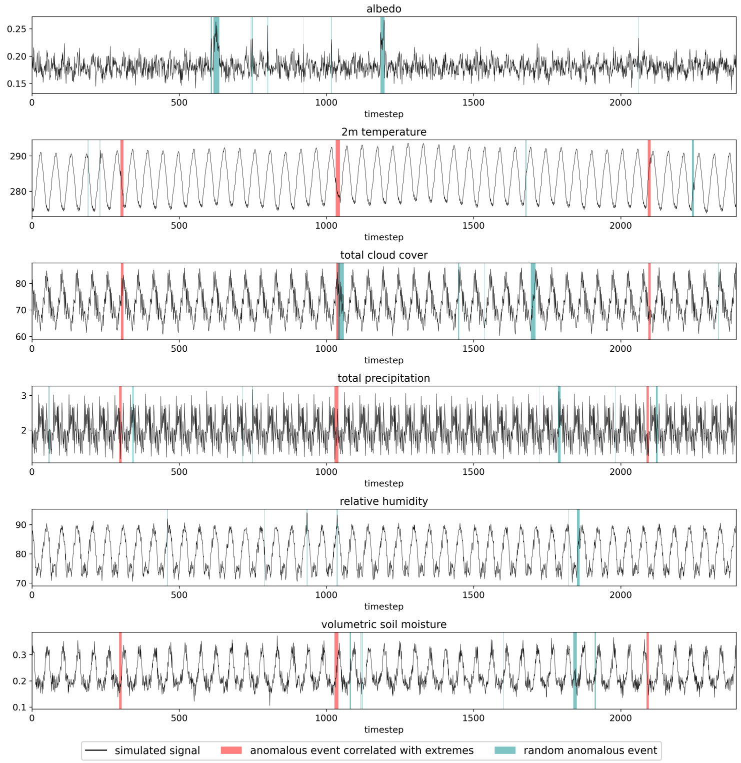


Figure 10: Visualization of the generated signals $\mathbf{\Phi}$ for 6 different variables from the synthetic CERRA reanalysis described in Table 3. The time series are shown for the location $(\text{lat} = 50, \text{lon} = 50)$. 🟥 are the drivers/anomalies which are correlated with extremes, and 🟩 are random anomalies.

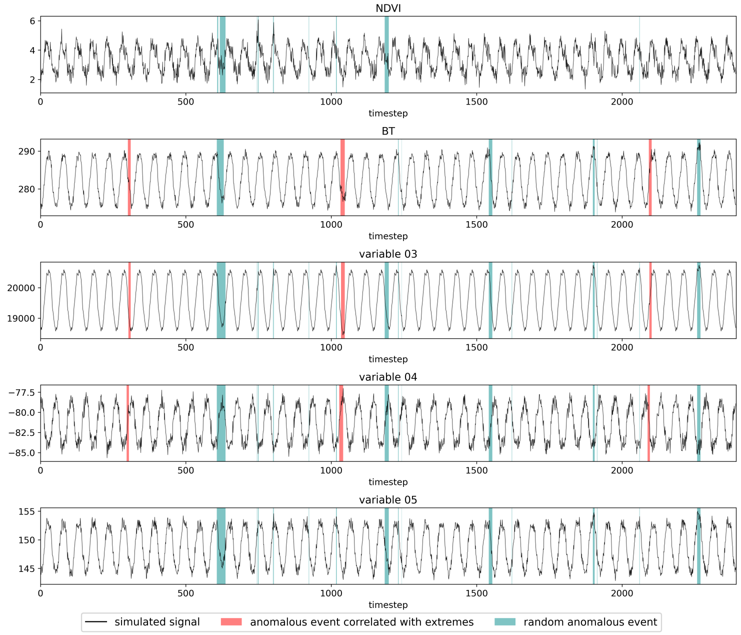


Figure 11: Visualization of the generated signals $\mathbf{\Phi}$ for 5 different variables from the synthetic NOAA data described in Table 4. The time series are shown for the location $(\text{lat} = 50, \text{lon} = 50)$. 🟥 are the drivers/anomalies which are correlated with extremes, and 🟩 are random anomalies.

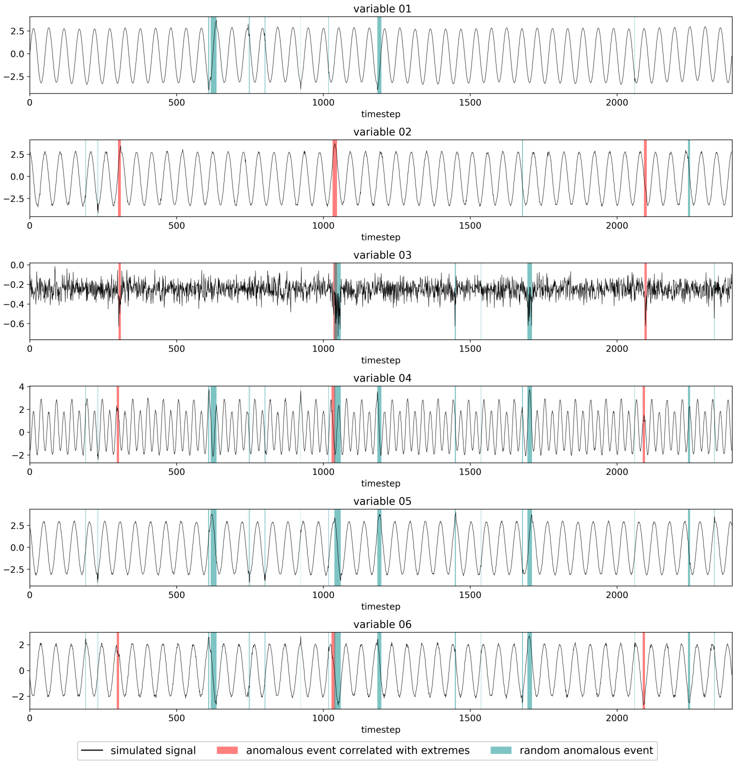


Figure 12: Visualization of the generated signals $\mathbf{\Phi}$ for 6 different variables from the synthetic artificial data described in Table 5. The time series are shown for the location (lat $= 50$, lon $= 50$). 🟥 are the drivers/anomalies which are correlated with extremes, and 🟩 are random anomalies.

B Additional results on the synthetic data

In Tables 6 and 7, we report the results on the synthetic NOAA remote sensing and artificial data described in Sec. A. We noticed a drop of performance for all models for the synthetic artificial data (Table 7). This explained as the ratio of anomalies is higher than the other two datasets. Second, we used for this artificial dataset red noise and quadratic coupling to generate dependent base signals. This makes this dataset harder for training. SimpleNet exhibits a dramatic dropped out of performance when it is tested on these two synthetic datasets. This illustrates that the model performance is highly dependent on the dataset and the backbone the model was trained on. Our model still outperforms all baselines on these datasets.

Table 6: Driver detection results on the synthetic NOAA remote sensing. The best performance on each metric is highlighted in a bold text. (±) denotes the standard deviation for 3 runs.

		Validation			Testing		
	Algorithm	F1-score (↑)	IoU (↑)	OA (↑)	F1-score (↑)	IoU (↑)	OA (↑)
	Naive	47.47	31.12	99.17	51.07	34.29	98.95
One-Class	OCSVM [39]	37.94±13.11	24.21±9.93	98.60±0.30	39.25±12.57	25.17±9.67	98.26±0.37
	IF [110]	44.14±1.18	28.34±0.97	98.68 ±0.01	45.13±1.30	29.15±1.09	98.49±0.01
	SimpleNet [51]	56.94±0.20	39.80±0.19	99.29±0.02	57.24±0.31	40.10±0.30	99.08±0.02
Rec.	STEALNet [59]	56.97±0.86	39.84±0.84	99.06±0.01	58.65±0.88	41.50±0.88	98.83±0.02
	UniAD [65]	45.24±3.58	29.30±2.94	98.65±0.21	46.99±4.21	30.81±3.54	98.36±0.29
MIL	DeepMIL [94]	71.77±0.38	55.97±0.46	99.55±0.01	72.02±0.31	56.28±0.38	99.42±0.01
	ARNet [95]	71.06±0.46	55.11±0.56	99.53±0.01	71.43±0.51	55.56±0.62	99.40±0.01
	RTFM [96]	60.30±0.26	43.16±0.27	98.92±0.02	61.91±0.21	44.83±0.23	98.70±0.02
	Ours*	**81.93**±0.23	**69.40**±0.36	**99.69**±0.01	**82.55**±0.25	**70.28**±0.36	**99.61**±0.01
	🐍 Ours†	81.44±0.47	68.70±0.66	**99.69**±0.01	82.07±0.45	69.59±0.65	99.60±0.01

*Video Swin Transformer backbone [100] †Mamba backbone [111]

Table 7: Driver detection results on the synthetic artificial data. The best performance on each metric is highlighted in a bold text. (±) denotes the standard deviation for 3 runs.

		Validation			Testing		
	Algorithm	F1-score (↑)	IoU (↑)	OA (↑)	F1-score (↑)	IoU (↑)	OA (↑)
	Naive	46.89	30.63	98.49	51.09	34.31	98.28
One-Class	OCSVM [39]	27.96±1.76	16.27±1.20	97.77±0.11	33.05±1.29	19.80±0.93	97.42±0.12
	IF [110]	28.55±1.21	16.66±0.82	97.63±0.02	34.15±1.25	20.60±0.91	97.19±0.03
	SimpleNet [51]	34.57±0.54	20.90±0.39	97.56±0.04	41.18±0.35	25.93±0.28	97.94±0.09
Rec.	STEALNet [59]	56.40±0.78	39.28±0.76	98.33±0.03	58.23±0.92	41.09±0.92	98.11±0.04
	UniAD [65]	49.48±1.60	32.88±1.41	97.66±0.11	52.49±1.25	35.60±1.16	97.57±0.09
MIL	DeepMIL [94]	20.18±23.67	13.38±16.45	71.33±20.53	18.75±21.34	12.04±14.39	71.08±20.48
	ARNet [95]	48.98±5.30	32.59±4.55	98.82±0.06	44.75±5.53	28.98±4.49	98.53±0.07
	RTFM [96]	59.90±0.31	42.75±0.32	98.27±0.04	61.41±0.46	44.31±0.48	98.07±0.04
	Ours*	**70.20**±0.43	**54.08**±0.51	**98.90**±0.03	**70.33**±0.71	**54.24**±0.84	**98.74**±0.08
	🐍 Ours†	66.63±5.41	50.19±5.91	98.81±0.12	67.64±5.95	51.39±6.60	98.71±0.18

*Video Swin Transformer backbone [100] †Mamba backbone [111]

C Ablation studies

In this section, we do ablation analyses on different aspects of the proposed model. All experiments are done on the validation set of synthetic CERRA reanalysis and evaluated with the F1-scores for drivers anomalies and extreme events detection.

C.1 Quantization layer

In Table 8, we study the performance of our model with different vector quantization algorithms. The first row **Threshold (Tanh)** represents a simple straight through estimator. For this quantization, we first map the input into a scalar value followed by a Tanh activation. Then we set positive values to be anomalies ($q = 1$). In **Random Quantization (RQ)** [114] the input is projected with a randomly initialized weights and then compared with a randomly initialized codes. **Vector Quantization (VQ)** is the standard quantizer which uses an Euclidean distance [101] or a cosine similarity [115]. We further add an orthogonality loss for VQ similar to [116]. **Finite Scalar Quantization (FSQ)** maps the input into a bounded scalar value. The code is then assigned based on the rounded value in the discrete space. As seen, **Lockup free quantization (LFQ)** [102] performs the best. We speculate that LFQ does not need to learn the code vectors which simplifies the task of quantization.

Table 8: Ablation studies on the quantization layer. The metric is F1-score on the driver/extreme detection.

Quantization Layer	F1-score (↑)
Threshold (Tanh)	70.07 / 84.98
RQ (Euclidean distance) [114]	71.52 / 87.23
VQ (Cosine similarity) [115]	78.83 / 88.98
VQ (Cosine similarity + Orthogonality) [115, 116]	78.54 / 86.83
VQ (Euclidean distance) [101]	77.42 / 87.27
VQ (Euclidean distance + Orthogonality) [101, 116]	79.26 / 88.28
FSQ [117]	76.57 / 88.65
LFQ [102]	**82.78 / 92.45**

C.2 Objective function for quantization

This part studies the loss function $\mathcal{L}_{(quantize)}$ used in Eq. (4). We denote the term $\|\mathbf{Z}_l - \text{sg}(\text{sign}(\mathbf{Z}_l))\|_2^2$ as $\mathcal{L}_{(commit)}$, $H[\mathbb{E}(\text{sign}(\mathbf{Z}_l))]$ as $\mathcal{L}_{(div)}$, and $\mathbb{E}[H(\text{sign}(\mathbf{Z}_l))]$ as $\mathcal{L}_{(ent)}$. As seen from Table 9, $\mathcal{L}_{(commit)}$ is essential for training to prevent the input for the quantization layer from growing. While $\mathcal{L}_{(ent)}$ and $\mathcal{L}_{(div)}$ (rows 4, 5, and 7) improve the results compared to the model with $\mathcal{L}_{(commit)}$ (second row).

Table 9: Ablation studies on the loss function $\mathcal{L}_{(quantize)}$ in Eq. (4). The metric is F1-score on the anomaly/extreme detection.

$\mathcal{L}_{(ent)}$	$\mathcal{L}_{(commit)}$	$\mathcal{L}_{(div)}$	F1-score (↑)
✓	✗	✗	00.00 / 00.00
✗	✓	✗	81.86 / 91.18
✗	✗	✓	00.00 / 00.00
✓	✓	✗	81.87 / 91.48
✗	✓	✓	82.26 / 92.30
✓	✗	✓	00.00 / 00.00
✓	✓	✓	**82.78 / 92.45**

C.3 Objective function for extreme events prediction.

Due to class imbalanced, we add class weighting applied to the binary cross entropy loss $\mathcal{L}_{(extreme)}$ to predict extremes. The weighting is based on the logarithm of the inverse square roots of the relative class frequencies in the batch. The weighted loss function achieves better results as shown in Table 10.

Table 10: Ablation studies on the weighted loss function $\mathcal{L}_{(extreme)}$ in Eq. (3). The metric is F1-score on the driver/extreme detection.

$\mathcal{L}_{(extreme)}$	F1-score (↑)
unweighted $\mathcal{L}_{(extreme)}$	81.14 / 90.91
weighted $\mathcal{L}_{(extreme)}$	**82.78 / 92.45**

C.4 Feature extractor (f_θ)

We evaluate three types of feature extractors, 3D CNN, Video Swin Transformer [100], and SSM Vision Mamba [111] with local scans. All models are trained from scratch. To build the 3D CNN model, we replaces the attention blocks in Swin Transformer with 3D convolutions and keep the overall architecture. For Mamba, we use local scans. Vision Mamba backbones replace the attention module with a linear selective state space model which provides an efficient alternative for Video Swin Transformer [118–124]. From Table 11 we notice that Swin Transformer achieves better results compared to 3D CNN. The parameters are also less compared to 3D CNN. The results also indicate that using Mamba instead of Swin Transformer achieves similar or better results when less parameters are used. In contrast to Swin Transformer, Mamba can be scaled to the global scale. However, we have used Swin Transformer in all experiments to have a fair comparison with the baselines due to its commonality as a backbone.

Table 11: Ablation study on the backbone f_θ used for feature extraction. The metric is F1-score on the driver/extreme detection.

Backbone f_θ	Hidden dimension (K)	Parameters	F1-score (↑)
3D CNN	8	63k	57.15 / 91.21
Video Swin Transformer	8	19k	81.22 / 91.16
Mamba	8	15k	82.15 / 90.18
3D CNN	16	250k	70.93 / 93.75
Video Swin Transformer	16	62k	82.78 / 92.45
Mamba	16	56k	83.45 / 93.12
3D CNN	32	998k	84.95 / 93.43
Video Swin Transformer	32	230k	84.14 / 93.12
Mamba	32	214k	84.00 / 93.43

Table 11 shows that increasing the model parameters still does not show a sign of overfitting.

C.5 Lossy vs lossless driver detection

We trained the model without the quantization layer. In other words, we remove the driver detection step and trained the model to predict extreme droughts directly from the extracted features with one classification head. The reason behind this experiment is to check the information loss through the quantization/driver detection step. Table 12 shows the improvement in performance on both the synthetic and real-world datasets. Adding the driver detection step identifies the related drivers to the events. However, at the cost of a slight decrease in accuracy on extreme events prediction.

Table 12: Ablation study on the driver detection step. The metric is F1-score on extreme detection where ($\Delta\text{F1} = \text{F1}_{(\text{without quantization})} - \text{F1}_{(\text{with quantization})}$).

Dataset	ΔF1-score (↓)
Synthetic CERRA Reanalysis	+0.96%
ERA5-Land (EUR-11)	+1.93%

C.6 Weighting parameters in the main objective function

Table 13: Sensitivity studies on the weighting parameters used in the main loss function in Eq. (6). The metric is mean F1-score on the driver/extreme detection for 3 random seeds.

(a)

$\lambda_{(driver)}$	F1-score (↑)
1	02.45 / 58.07
10	03.34 / 75.73
100	**82.78 / 92.45**
200	80.85 / 91.69
1000	79.94 / 86.44

(b)

$\lambda_{(commit)}$	F1-score (↑)
1.0	79.81 / 92.02
1.5	81.43 / 91.96
2.0	82.27 / 92.11
3.0	**82.78 / 92.45**
4.0	82.76 / 91.89

(c)

$\lambda_{(ent)}$	F1-score (↑)
0.01	82.52 / 92.61
0.1	**82.78 / 92.45**
1.0	82.72 / 92.46

(d)

$\lambda_{(div)}$	F1-score (↑)
0.01	77.51 / 91.71
0.1	**82.78 / 92.45**
0.5	82.48 / 92.21

We study the role of the weighting parameters in the main loss function in Eq. (6). We can observe from Table 13 that $\lambda_{(driver)}$ plays a crucial role in anomaly detection. Having $\lambda_{(driver)}$ small makes the model less constrained and reduces supervision on where extremes were reported in the training data. Bigger values of $\lambda_{(driver)}$ make the model more constrained to identify drivers near or in overlap with the extremes and thus reduce the overall performance. We can also observe that the model is less sensitive to $\lambda_{(ent)}$ compared to other parameters. We selected the final default weighting parameters in the experiments based on the average performance on both driver and extreme events prediction.

C.7 Objective functions

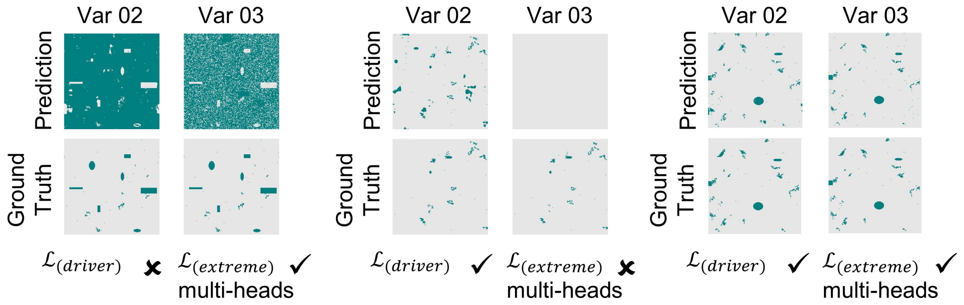


Figure 13: Supplementary to the ablation study in Table 2.

Without the loss $\mathcal{L}_{(driver)}$, the detection of drivers/anomalies is not reliable since pixels at regions and intervals where no extreme event occurred can be assigned to $z_{q=1}$ (driver) as well. In case of $\mathcal{L}_{(extreme)}$ multi-heads, we observe that anomalies are identified in a small subset of variables because the network omits some variables if there is a correlation with other variables. Please see Fig. 13. In case of a single head such flips occur less often, but they can occur. If the loss $\mathcal{L}_{(driver)}$ is used, such flips cannot occur and the multi-head improves both F1-scores by a large margin. When comparing rows 2 and 4 in Table 2 (a), there is a slight decrease in extreme prediction but a large improvement in anomaly detection. Note that there is always a trade-off between extreme and anomaly detection. The anomalies generate a bottleneck of information. When more information goes through the bottleneck the better the extreme prediction gets. Without any anomaly detection, the extreme prediction is best as shown in Table 12, but the increase in F1 score is only moderate. This is also visible in rows 2 and 3 in Table 2 (b). Cross-attention improves extreme prediction, but it hurts the detection of anomalies since the information is propagated between the variables.

D Results on the real-world reanalysis data

D.1 Quantitative results

It is difficult to quantify the quality of the predicted drivers and anomalies on real-world data. Therefore, we hypothesize that the model can predict extreme agricultural droughts from the identified drivers and anomalies, only if those drivers and anomalies are causally correlated with the extreme events. To verify this, we report in Table 14 the prediction accuracy on extreme agricultural droughts. The results show that using the identify drivers as input, the model can predict extreme agricultural droughts across different regions and datasets.

Table 14: Quantitative results for extreme droughts detection on real-world data. (±) denotes the standard deviation for 3 runs.

		Validation			Testing		
Dataset	Region	F1-score (↑)	IoU (↑)	OA (↑)	F1-score (↑)	IoU (↑)	OA (↑)
CERRA	Europe	22.31±0.74	12.56±0.47	90.45±1.05	28.63±1.45	16.71±0.98	89.13±1.44
ERA5-Land	Europe	31.87±0.39	18.96±0.28	95.45±0.25	21.52±0.86	12.06±0.54	95.84±0.21
	Africa	22.49±0.42	12.67±0.27	85.59±1.48	18.53±0.48	10.21±0.29	78.63±1.43
	North America	27.39±0.63	15.87±0.42	93.84±0.13	31.74±0.48	18.86±0.34	89.95±0.42
	South America	29.30±1.52	17.17±1.04	89.31±0.33	28.96±1.67	16.94±1.13	83.99±0.23
	Central Asia	20.99±0.43	11.73±0.27	95.38±0.04	25.01±0.11	14.29±0.07	94.51±0.04
	East Asia	18.82±0.72	10.39±0.44	93.10±0.42	25.35±0.20	12.58±0.12	93.73±0.38

The performance depends on the type and ratio of extremes, spatio-temporal resolution, the quality and consistency between the remote sensing and the reanalysis data. F1 scores substantially increase when the threshold on VHI is increased (see Table. 16 in Sec. D.3). Note that it is not required to predict all extremes in order to learn specific relations from the predicted events.

D.2 Robustness

To further check the model robustness, we train the model on the same EUR-11 region with 6 different combinations of physical variables. We anticipate that if a variable is relevant to extremes over specific region i.e., Europe it should appear in all identified sets of variables i.e., we expect soil moisture (swvl1) and evaporation (e) to be always presented as explanatory variables for extremes over Europe. Our experimental results in Table 15 confirms this assumption.

Table 15: Quantitative results from ERA5-Land EUR-11 data for different combination of physical variables. The metric is F1-score on the extreme droughts detection for the validation/test sets.

Input variables	Selected variables	F1-score(↑)
{t2m, fal, e, tp, stl1, swvl1}	{e, swvl1}	31.87 / 21.52
{d2m, t2m, sp, e, tp, stl1}	{e}	32.72 / 23.48
{d2m, fal, sp, e, skt, swvl1}	{e, swvl1}	32.92 / 22.41
{t2m, sp, e, skt, stl1, swvl1}	{e, swvl1}	32.33 / 21.61
{t2m, fal, sp, skt, stl1, swvl1}	{swvl1}	24.34 / 16.05
{t2m, e, tp, swvl1}	{e, swvl1}	30.38 / 21.59

D.3 Scientific validity

Surely, soil temperature is a key factor in the drought processing and soil moisture–temperature feedback [125]. State variable of the land surface such as albedo (fal/al) and soil temperature (stl1/sot) should be very related to reflectance on the ground and consequently to VHI from remote sensing. However, our approach indicates that these variables are not informative enough for the model to identify drivers and predict extremes. To investigate this issue, we conducted 4 more experiments on both CERRA and ERA5-Land where we trained models that take only one variable al/fal or stl as input and predict the extreme events directly without the driver/anomaly detection step (similar to Sec. C.5). In all of these experiments, the F1-score was very low. In the next experiment, we increased the threshold for VHI and trained new models to predict extremes directly. The results for the validation set are shown below in Table 16:

The first potential reason to consider is that some land surface variables might deviate from the reality. ERA5-Land does not use data assimilation directly. The evolution and the simulated land fields

Table 16: The relation between the definition of extremes from VHI and the model prediction. The metric is F1-score (↑) on the extreme droughts detection for the validation sets.

Input variables	Dataset	Domain	VHI < 26	VHI < 40	VHI < 50
{stl1}	ERA5-Land	EUR-11	05.67	31.53	58.36
{t2m, fal, e, tp, stl1, swvl1}	ERA5-Land	EUR-11	33.80	46.72	68.71

are controlled by the ERA5 atmospheric forcing. Another reason might be that when training only on extremes (VHI < 26), there are not enough samples to learn the relations. Note that VHI is a combination of both TCI and VCI. Most extremes (VHI < 26) might result from a deficiency in both stl/t2m and vsw. This might also explain why stl and albedo cannot be that informative to predict very extreme events. Last row in Table 15 shows the result when we discard albedo and soil temperature as input variables from ERA5-Land.

Moreover, state variables of the hydrological cycle in ERA5-land like volumetric soil water variable (swvl1/vsw) has biases [126]. One solution to improve the validity of investigation is to use satellite observations for the top layer [127]. However, the experiments showed that the model relates soil moisture anomalies with the extremes in VHI and provides reasonable predictions.

D.4 Spatial Resolution

Table 17: The impact of spatial resolution on the model prediction. The metric is F1-score on the extreme droughts detection for the validation sets.

Dataset	Domain	Spatial resolution	F1-score(↑)
ERA5-Land	EUR-11	0.1°	31.87
ERA5-Land	EUR-11	0.2°	30.09

E Baselines

We evaluate our approach with an interpretable forecasting approach using integrated gradients [109] and 8 baselines from 3 main related categories in anomaly detection; one-class unsupervised [39, 110, 51], reconstruction-based [59, 65], and multiple instance learning [94–96]. Note that these baselines are not directly applicable to the task we addressed in this paper. We modified and trained all baselines from scratch. For this, we relied on the officially released codes and started from the default hyperparameters.

Integrated Gradients [109] is an axiomatic attribution method which can be used as a post-hoc method to explain the model prediction. We trained two models that predict extreme events directly from the input variables and then we applied a post-hoc integrated gradients from Pytorch Captum [128]. Both models use the same Swin Transformer backbone as our model but without the anomaly detection step. For Integrated Gradients II, we added a cross attention. For these baselines, we compute the gradient only with respect to predicted extremes and computed a different threshold for each variable separately.

Isolated Forest (IF) [110] is an ensemble of random trees that isolate input instances by randomly selecting abounded splits for features. A shorter averaged path for a recursive partitioning implies an anomalous input. We empirically set the number of estimators to 100.

OCSVM [39] is a one-class support vector machine solved using stochastic gradient descent. We use a radial basis kernel with 100 components.

For IF and OCSVM, we extracted multivariate features as the distances between the input variables and then train a model on each input variable separately. We sampled 400k normal data points (locations without extreme flags) for training and defined the expected ratio of anomalous to be roughly equivalent to the ratio of extreme events in the data.

SimpleNet [51] is categorized as an embedding-based one-class algorithm for anomaly detection. In SimpleNet, a feature extractor is first used to extract local features from normal data followed by a feature adaptor. Then an anomaly feature generator induces anomalous features by adding a Gaussian noise in the feature space. Finally, a discriminator is trained to distinguish between the normal and anomalous generated features. During inference, the feature generator is discarded and the trained discriminator is expected to separate anomalous from normal input features. During inference, we compute the median anomaly score for both normal pixels and pixels with extreme flags. Then, we set a threshold for anomalous pixels based on the average of the two former median anomaly scores. We trained SimpleNet with our pretrained model as a feature extractor and set the feature dimension for feature adaptor to 512. The extracted features are scaled by 10^{-2} and the anomalous feature generator uses a Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma = 1.5)$. The discriminator composes of two linear layers with a hidden dimension of 96. We set th^{+} and th^{-} to 1.0 and trained with AdamW optimizer for 48 epochs with a batch size of 2 and a learning rate of 3×10^{-4} with 1×10^{-5} weight decay.

STEALNet [59] or synthetic temporal anomaly guided end-to-end video anomaly detection. This is a reconstruction-based algorithm. STEALNet was trained to maximize the reconstruction loss for locations with extreme flags and to minimize the reconstruction loss otherwise. We set the dimensions for the auto-encoder as 96, 128, and 256. We trained with a batch size of 8 for 100 epochs using Adam optimizer with 1×10^{-3} weight decay and a learning rate of 3×10^{-4}.

For STEALNet and UniAD, we compute during inference the mean reconstruction loss for both normal pixels and pixels with extreme flags. Then, we set a threshold for anomalous pixels based on the average of these two values.

DeepMIL [94] is a multiple instance learning for anomaly detection in surveillance videos. In MIL, each video is represented as a bag of snippets (instances). We define positive instances as pixels with extreme event flags where the exact information (which variable is anomalous) within the positive instances is unknown. In the original implementation, the ranking is enforced only on the top instance with the highest anomaly score in each bag. We modified the ranking loss to top-k and set $K = 100$.

ARNet [95] or anomaly regression net is another MIL-based approach. The ranking loss is based on a top-k binary cross entropy. In addition, there is a center loss to reduce the intra-class discriminative features of normal instance. We set $\alpha = 400$ and $\lambda_{center} = 20$.

RTFM [96] or robust temporal feature magnitude learning is a MIL-based algorithm which learns to distinguish between the normal and anomalous scores by selecting the top-k snippets with the highest feature magnitudes. We modified the multi-scale temporal network (MTN) to capture the local and global spatial dependencies between pixels. We set the dimension for MTN to 32, $K = 100$, the margin to separate features to $\times 10^{2}$ and $\alpha = 1 \times 10^{-4}$.

We use the same Video Swin Transformer backbone as our model to train DeepMIL, ARNet, and RTFM. During training, we noticed that the ranking loss becomes biases toward one variable when the ranking is computed among all variables, so we computed a loss for each variable and then average the losses. RTFM only worked when we added a cross attention between the variables. We suppose this to be related to the feature magnitude learning. We trained with a batch size of 2 using Adam optimizer and a learning rate of 6×10^{-4} for 100 epochs with a weight decay of 1×10^{-3}. We also added an instance dropout of 0.5.

F Computational efficiency

In Table 18, we report the inference time with a fixed input of 6 variables, 8 days and 200 × 200 spatial resolution. STEALNet is based on a 3D CNN auto-encoder without a self-attentions which explains its efficiency but on the cost of accuracy. In the last row, we show the estimated time when we discard the classification head to detect extreme events and only use the model for anomaly detection.

The training on the real-world data for EUR-11 took about $\sim$ 21 hours with a Swin Transformer model, $K = 16$, and 4 NVIDIA A100 GPUs. Table 19 gives a rough estimation for training on the synthetic CERRA for 1 epoch. SimpleNet was trained with a pretrained backbone. The training time includes some postprocessing to compute metrics on the training set. The time might also differ depending on the I/O during training and the number of available workers.

Table 18: Inference time in seconds for our model and other DL baselines.

Algorithm	GPU[1] (sec)	Parameters (M)
SimpleNet	0.156 ±0.003	0.203
STEALNet	0.003 ±0.000	6.005
UniAD	3.733 ±0.012	3.674
DeepMIL	0.193 ±0.008	0.285
ARNet	0.191 ±0.003	0.285
RTFM	0.257 ±0.001	0.319
Ours	0.132 ±0.000	0.479
Ours[2]	0.122 ±0.000	0.145

[1] NVIDIA GeForce RTX 3090 GPU

[2] Our model is used only for driver/anomaly detection

Table 19: Training time on the synthetic CERRA for 1 epoch.

Algorithm	Time (min)	GPU
SimpleNet	~ 2	A100
STEALNet	~ 1	A100
UniAD	~ 11	4 × A100
DeepMIL	~ 13	A40
ARNet	~ 13	A40
RTFM	~ 20	A40
Ours	~ 8	A40

G Implementation details

The training was done mainly on clusters with NVIDIA A100 80GB and NVIDIA A40 48GB GPUs. In the following, we highlight the main technical implementations and hyperparameters for training:

Synthetic data For the synthetic data we set the embedding dimension $K = 16$. We use one layer Video Swin Transformer with {depth=[2, 1], heads=[2, 2], window size=[[2, 4, 4], [6, 1, 1]]}.

For the quantization layer, we use two sequential 3D CNN layers on each variable with: {[kernel=(3, 3, 3), stride=(1, 1, 1), padding=(1, 1, 1)]} and followed by a shared linear layer.

The classifier g_ψ consists of the following layers: {[3D CNN, kernel=(3, 3, 3), stride=(3, 1, 1), padding=(0, 1, 1)], [3D CNN, kernel=(2, 3, 3), stride=(2, 1, 1), padding=(0, 1, 1)]}. We set $\lambda_{(ent)} = \lambda_{(div)} = 0.1$, $\lambda_{(anomaly)} = 100$, and $\lambda_{(commit)} = 3$. The models were trained with Adam optimizer [129] for 100 epochs with a batch size of 4. We use a linear warm up of 2 epochs and a cosine decay with an initial learning rate of 2×10^{-3} and a weight decay of 3×10^{-3}.

Reanalysis data We set the embedding dimension $K = 24$ for CERRA and CAS-11 datasets and $K = 16$ for the rest of ERA5-Land datasets. We use one layer Video Swin Transformer with {depth=[2, 1], heads=[2, 2], window size=[[2, 4, 4], [8, 1, 1]]}. The quantization layer is similar to the one for the synthetic data. The classifier g_ψ consists of the 3 layers each has {[D CNN, kernel=(2, 3, 3), stride=(2, 1, 1), padding=(0, 1, 1)}. We set $\lambda_{(ent)} = \lambda_{(div)} = 0.1$, and $\lambda_{(anomaly)} = 100$ by defaults and $\lambda_{(commit)} = 1.0$ for CAS-11. For CERRA, we set $\lambda_{(ent)} = \lambda_{(div)} = 0.01$. Due to the high resolution on the reanalysis data, we use gradient checkpoint during training. For CERRA reanalysis, we also cut the boundaries between low and high latitudes focusing on the central region. This results in a final grid with 512×832 cells.

To handle missing data and temporal gaps in the input reanalysis data, we first normalize the data using the pre-computed statistics and then replace the invalid pixels with zero values.

H Reanalysis data

The raw reanalysis data were provided by the Climate Data Store (CDS) [107, 106]. Technical details regarding the reanalysis datasets are provided in Tables 20 and 21. CORDEX regions [108] used in this study are shown in Fig. 14. For all regions on ERA5-Land, we used the following variables: {"t2m", "fal", "e", "tp", "swvl1", "stl1"}. For CERRA, we did the experiments with: {"t2m", "al", "tcc", "tp", "vsw", "r2"}.

Table 20: Datasets used in the experiments on real-world data. CORDEX domains are defined based on [108].

Dataset	Region	CORDEX	Resolution	Train	Val	Test
CERRA	Europe	-	1069×1069	1984-2015	2016-2018	2019-2021
ERA5-Land	Europe	EUR-11	412×424	1981-2017	2018-2020	2021-2024
ERA5-Land	Africa	AFR-11	804×776	1981-2017	2018-2020	2021-2024
ERA5-Land	North America	NAM-11	520×620	1981-2017	2018-2020	2021-2024
ERA5-Land	South America	SAM-11	668×584	1981-2017	2018-2020	2021-2024
ERA5-Land	Central Asia	CAS-11	400×612	1981-2017	2018-2020	2021-2024
ERA5-Land	East Asia	EAS-11	668×812	1981-2017	2018-2020	2021-2024

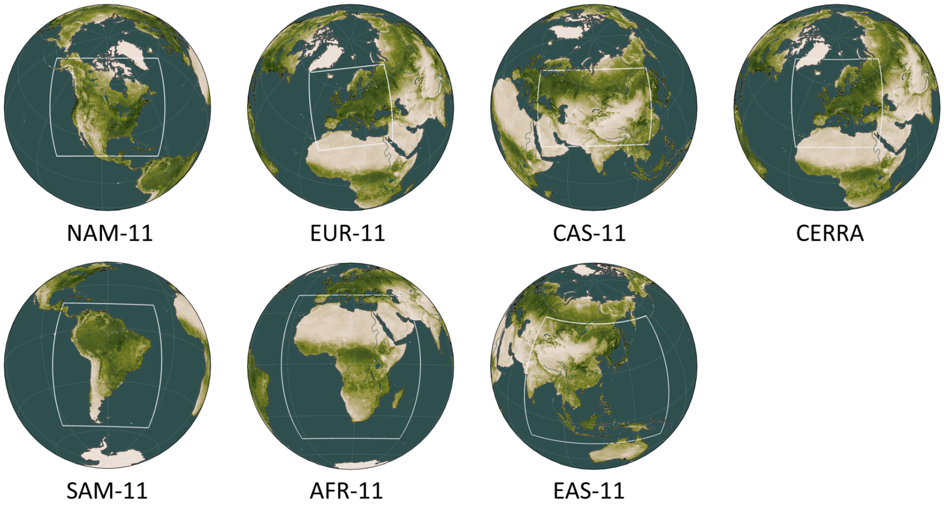


Figure 14: The definition of the domains used in the study. ERA5-Land reanalysis is mapped onto the CORDEX domains [108]. CERRA has its own domain definition [106].

I Agricultural drought definition and remote sensing data

I.1 Satellite-derived agricultural drought

It is generally challenging to define what exactly constitutes an extreme. Extremes can be categorized from the perspective of their impacts. For instance, extreme drought can be categorized into 4 types based on their impacts [130, 11]; meteorological or climatological drought, agricultural drought [131], hydrological drought, and socioeconomic drought [132]. Meteorological drought is mainly related to the dryness and can be defined based on a deficiency in temperature or precipitation. Agricultural drought measures the impact of stress on vegetation and usually defined as soil water deficits. It is also widely conceived that drought originates and progresses from meteorological into agricultural drought [6, 17]. Hydrological drought is related to the water storage. While socioeconomic drought can be measured based on supply and demand related to weather and deficit in water supply. In this paper, we are interested in extreme agricultural droughts.

Table 21: Details regarding the processed variables from ERA5-Land [107] and CERRA [106] reanalysis.

Dataset	Variable name	Long name	Unit	Height
CERRA	al	albedo	%	surface
	hcc	high cloud cover	%	above 5000m
	lcc	low cloud cover	%	surface-2500m
	mcc	medium cloud cover	%	2500m-5000m
	liqvsm	liquid volumetric soil moisture	m^3/m^3	top layer of soil
	msl	mean sea level pressure	Pa	surface
	r2	2 metre relative humidity	%	2m
	si10	10 metre wind speed	m/s	10m
	skt	skin temperature	K	surface
	sot	soil temperature	K	top layer of soil
	sp	surface pressure	Pa	surface
	sr	surface roughness	m	surface
	t2m	2 metre temperature	K	2m
	tcc	total Cloud Cover	%	above ground
	tciwv	total column integrated water vapour	kg/m^2	surface
	tp	total Precipitation	kg/m^2	surface
	vsw	volumetric soil moisture	m^3/m^3	top layer of soil
	wdir10	10 metre wind direction	∘	10m
ERA5-Land	d2m	2m dewpoint temperature	K	2m
	t2m	2m temperature	K	2m
	fal	forecast albedo	%	surface
	skt	skin temperature	K	surface
	stl1	soil temperature	K	soil layer (0 - 7 cm)
	sp	surface pressure	Pa	surface
	e	total evaporation	m of water equivalent	above ground
	tp	total precipitation	m	surface
	swvl1	volumetric soil water	m^3/m^3	soil layer (0 - 7 cm)

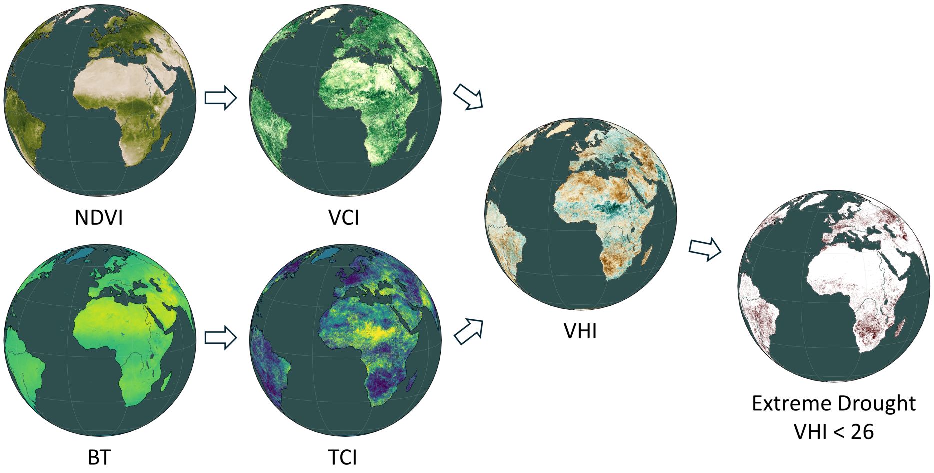


Figure 15: An overview of the extreme agricultural droughts definition from remote sensing.

Worldwide satellite observations allow for almost real-time monitoring of drought and vegetation conditions. In practice, vegetation states can be estimated from land surface reflectances acquired from satellites. As a result, the reflectances on the ground can be employed as agricultural drought indicators and as proxies for vegetation health. To define extreme agricultural drought events, we processed satellite-based vegetation health dataset [104] from the National Oceanic and Atmospheric Administration (NOAA) (https://www.star.nesdis.noaa.gov/smcd/emb/vci/VH/index.php [last access: 22 May 2024)]). This dataset consists of long-term remote sensing data acquired from a system of NOAA satellites: the Advanced Very-High-Resolution Radiometer (AVHRR) which starts from 1981 until 2012 and the new system the Visible Infrared Imaging Radiometer Suite (VIIRS) from 2013 onwards. The dataset has a global coverage with $\sim 0.05°$ ($\sim$4km) spatial resolution. The normalized difference vegetation index (NDVI) [133] and brightness temperature (BT) [134] are the two key products of the dataset. The BT is an infrared (IR)-based calibrated spectrum radiation. While the NDVI is a combination of the near-infrared (NIR) and red (R) bands. To remove the effects of clouds, atmospheric disturbance, and other error sources, the data were aggregated temporally into a smoothed product on a weekly basis. The weekly temporal coverage is needed for outliers and discontinuities removal and is suitable to study the phenological phases of vegetation and consequently to define agricultural drought [104, 135]. Based on the long-term upper and lower bounds of the ecosystem (maximum and minimum values of the NDVI and BT), agricultural drought indicators such as vegetation condition index (VCI), thermal condition index (TCI), and vegetation health index (VHI) can be derived [134, 136]. VHI is a combination of VCI and TCI (Fig. 15) and it fluctuates between 0 (unfavourable condition) and 100 (favourable condition). Values outside the range are clipped. Based on this definition of vegetation health, extreme agricultural drought can be defined when VHI < 26. Please note that vegetation stress detected by VHI could not be necessarily caused by a drought event i.e., a change in the land cover can change the signal as well [15, 137]. Thus, VHI should be interpreted carefully.

I.2 Pre-processing of the remote sensing data

The remote sensing dataset is provided on the Plate Carrée projection (geographic latitude and longitude). The target agricultural drought data and reanalysis data have to be aligned in the same coordinate systems and over the same regions. To realize this, we mapped the remote sensing data onto the Lambert conformal conical grid for CERRA and onto the rotated coordinate systems over the different CORDEX domains for ERA5-Land. For the mapping, we use the first-order conservative mapping using the software from Zhuang et. al [138]. To calculate the spatial averaged, we excluded coastal lines, invalid, and water body pixels. Furthermore, we combined the dataset with masks obtained from the quality assurance metadata for pixels over no vegetation and very cold areas. As mentioned in Sec. I.1, a temporal decomposition was conducted to remove some discontinuity and aggregate the data into a weekly product. However, some pixels will still be empty. To tackle this issue, we first checked if the pixel was covered by another satellite and averaged the measurements of the satellites. If it was not the case, we flagged the pixel as invalid and discard it from the training and evaluation. This remote sensing dataset serves as a reference of extreme agricultural drought events to train and evaluate the performance of the model. Table 22 shows the ratio of extreme events in the datasets. Please note that there is no ground truth for drivers or anomalies in our real-world dataset. We only report the ratio of extreme agricultural drought events, which can be detected using remote sensing data.

Table 22: Details regarding the ratio of extreme events in the pre-processed NOAA remote sensing data.

Region	Domain	Extremes (%)	
		Val	Test
Europe	CERRA	4.34	5.32
Europe	EUR-11	3.20	2.86
Africa	AFR-11	6.41	6.87
North America	NAM-11	3.68	6.61
South America	SAM-11	5.16	6.53
Central Asia	CAS-11	3.60	4.38
East Asia	EAS-11	3.16	3.05

J Additional results

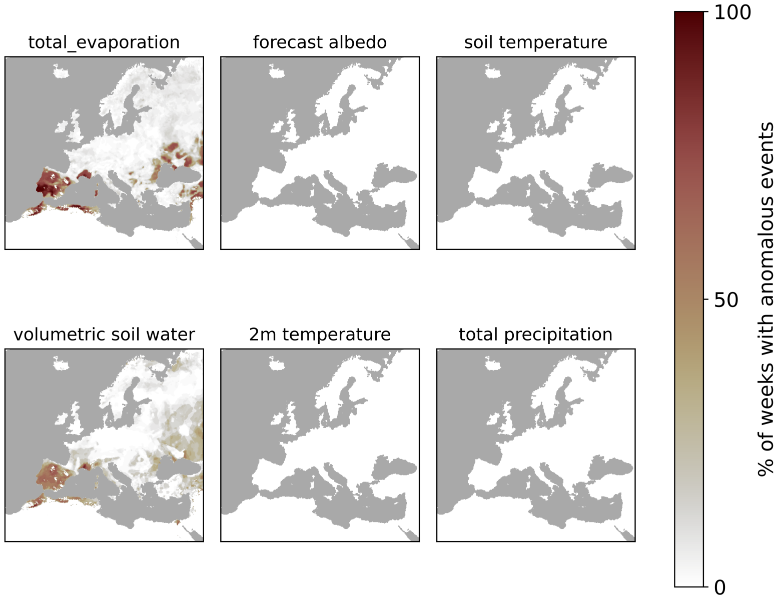


Figure 16: The averaged spatial distribution of drivers and anomalies related to Portugal in Europe. For this experiment, we use prediction on EUR-11 from ERA5-Land and select frames (times) within the period 2018-2024 where there were extreme drought of at least 25% of the pixels in the Portugal. Then we normalize the identified drivers and anomalies by the total number of frames to obtain the final map. As can be seen drivers are spatially centered around where extremes were reported.

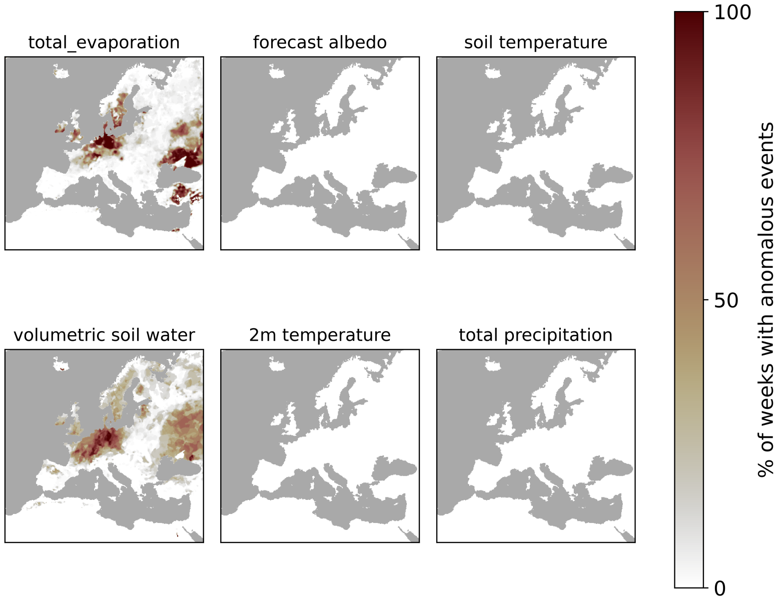


Figure 17: The averaged spatial distribution of drivers and anomalies related to a specific place in Europe (North Rhine-Westphalia). For this experiment, we use prediction on EUR-11 from ERA5-Land and select frames (times) within the period 2018-2024 where there were extreme drought of at least 25% of the pixels in the North Rhine-Westphalia. Then we normalize the identified drivers and anomalies by the total number of frames to obtain the final map. As can be seen drivers are spatially centered around where extremes were reported.

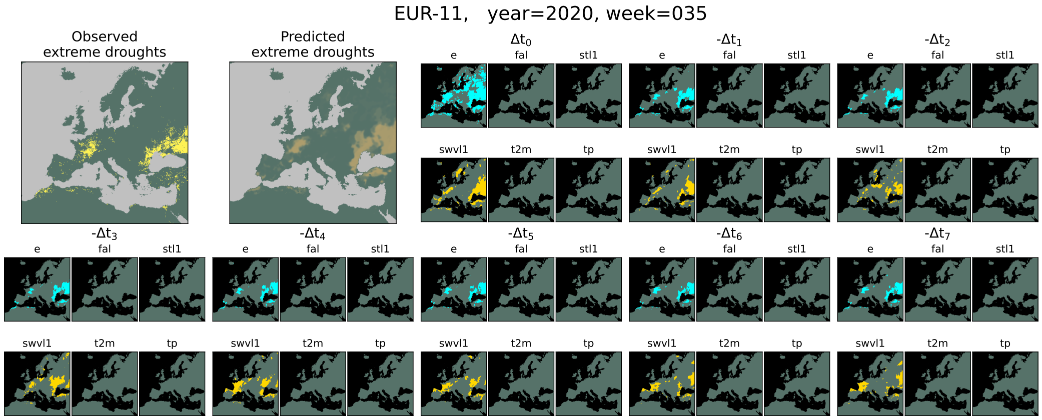

EUR-11, year=2020, week=035
Observed extreme droughts
Predicted extreme droughts
Δt0
-Δt1
-Δt2
-Δt3
-Δt4
-Δt5
-Δt6
-Δt7
e
fal
stl1
swvl1
t2m
tp

Figure 18: Qualitative results on ERA5-Land for Europe (EUR-11). Shown are the identified drivers and anomalies for each variable along with the prediction of extreme agricultural droughts on the top left.

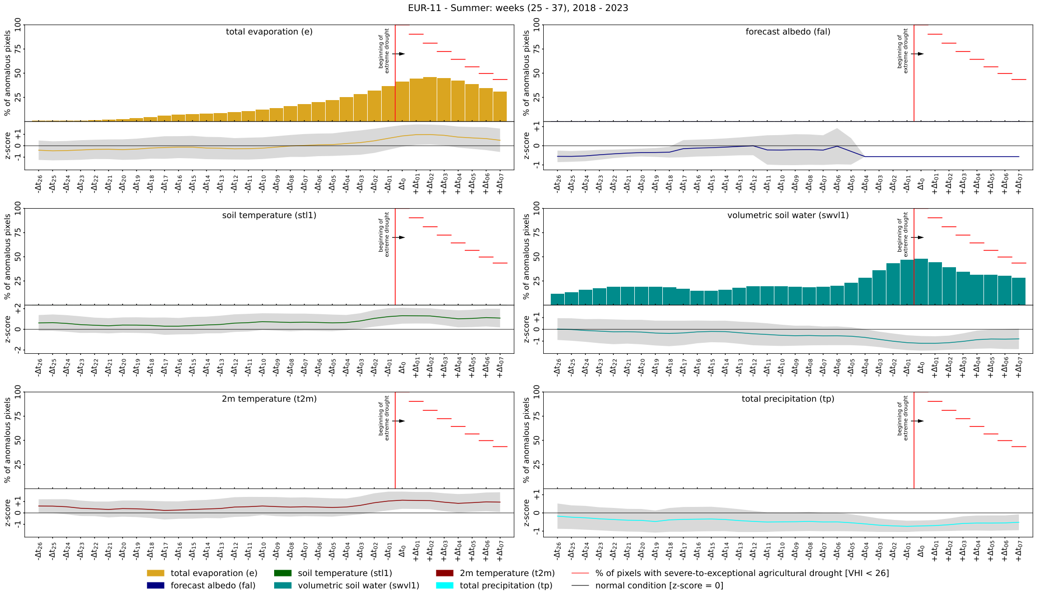


Figure 19: Temporal evolution of drivers and anomalies related to the extremes in ERA5-Land for Europe (EUR-11). For this experiments, we select pixels with extreme events during summer (weeks 25-38) for the years 2018-2023 and compute the average distribution of drivers and anomalies with time. The red line at δt_0 indicates the beginning of the extreme droughts. Z_{score} in the underneath curve represents the deviation from the mean computed from the ERA5-Land climatology.

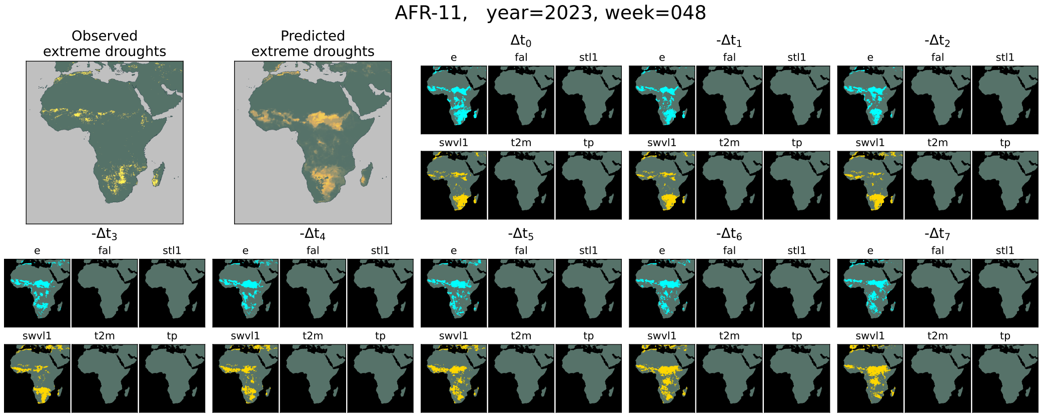

AFR-11, year=2023, week=048
Observed extreme droughts
Predicted extreme droughts
Δt0
-Δt1
-Δt2
-Δt3
-Δt4
-Δt5
-Δt6
-Δt7
e
fal
stl1
swvl1
t2m
tp

Figure 20: Qualitative results on ERA5-Land for Africa (AFR-11). Shown are the identified drivers and anomalies for each variable along with the prediction of extreme agricultural droughts on the top left.

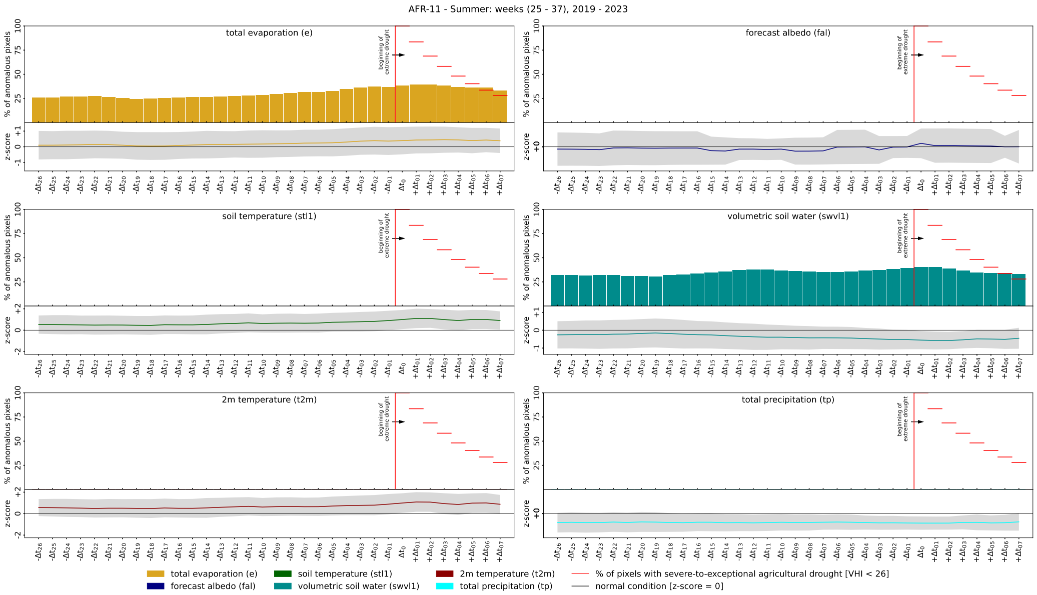
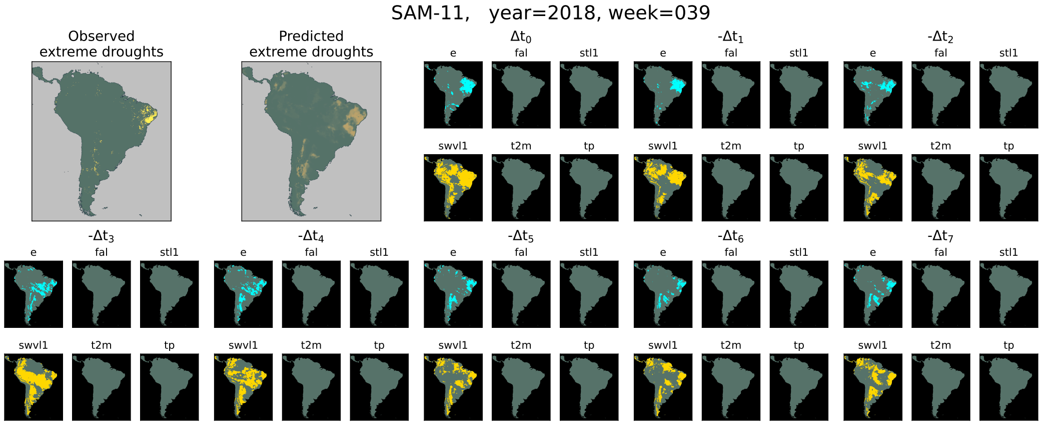


Figure 21: Temporal evolution of drivers and anomalies related to the extremes in ERA5-Land for Africa (AFR-11). For this experiments, we select pixels with extreme events during summer (weeks 25-38) for the years 2019-2023 and compute the average distribution of drivers and anomalies with time. The red line at δt_0 indicates the beginning of the extreme droughts. Z_{score} in the underneath curve represents the deviation from the mean computed from the ERA5-Land climatology.



SAM-11,   year=2018, week=039
Observed extreme droughts
Predicted extreme droughts
Δt0
-Δt1
-Δt2
-Δt3
-Δt4
-Δt5
-Δt6
-Δt7
e
fal
stl1
swvl1
t2m
tp

Figure 22: Qualitative results on ERA5-Land for South America (SAM-11). Shown are the identified drivers and anomalies for each variable along with the prediction of extreme agricultural droughts on the top left.

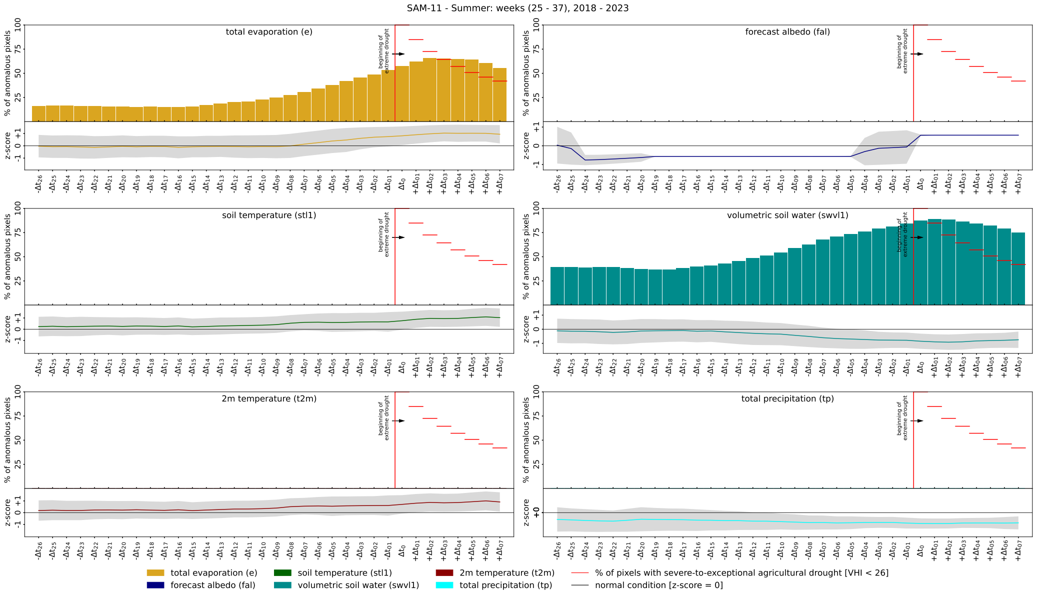


Figure 23: Temporal evolution of drivers and anomalies related to the extremes in ERA5-Land for South America (SAM-11). For this experiments, we select pixels with extreme events during summer (weeks 25-38) for the years 2018-2023 and compute the average distribution of drivers and anomalies with time. The red line at δt_0 indicates the beginning of the extreme droughts. Z_{score} in the underneath curve represents the deviation from the mean computed from the ERA5-Land climatology.

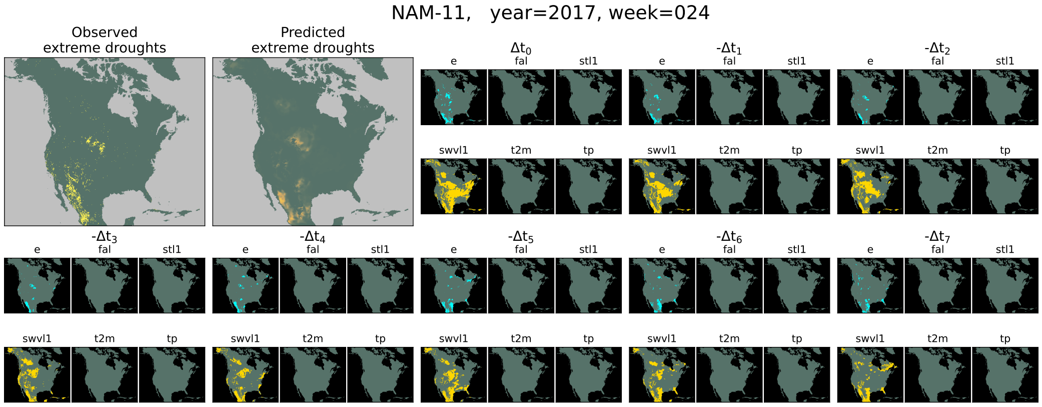

NAM-11, year=2017, week=024
Observed extreme droughts
Predicted extreme droughts
Δt₀
-Δt₁
-Δt₂
-Δt₃
-Δt₄
-Δt₅
-Δt₆
-Δt₇
e
fal
stl1
swvl1
t2m
tp

Figure 24: Qualitative results on ERA5-Land for North America (NAM-11). Shown are the identified drivers and anomalies for each variable along with the prediction of extreme agricultural droughts on the top left.

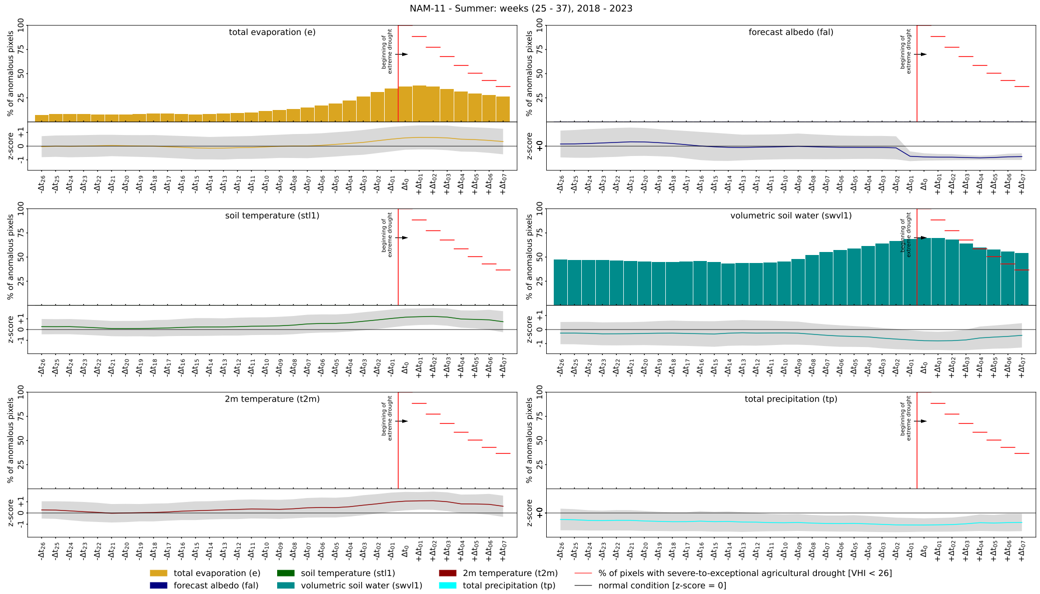


Figure 25: Temporal evolution of drivers and anomalies related to the extremes in ERA5-Land for North America (NAM-11). For this experiments, we select pixels with extreme events during summer (weeks 25-38) for the years 2018-2023 and compute the average distribution of drivers and anomalies with time. The red line at δt_0 indicates the beginning of the extreme droughts. Z_{score} in the underneath curve represents the deviation from the mean computed from the ERA5-Land climatology.

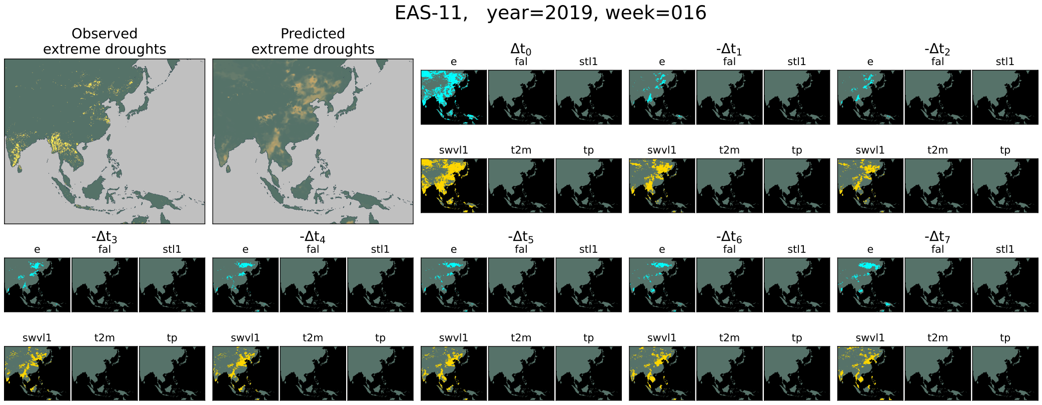

EAS-11, year=2019, week=016
Observed extreme droughts
Predicted extreme droughts
Δt0
-Δt1
-Δt2
-Δt3
-Δt4
-Δt5
-Δt6
-Δt7
e
fal
stl1
swvl1
t2m
tp

Figure 26: Qualitative results on ERA5-Land for East Asia (EAS-11). Shown are the identified drivers and anomalies for each variable along with the prediction of extreme agricultural droughts on the top left.

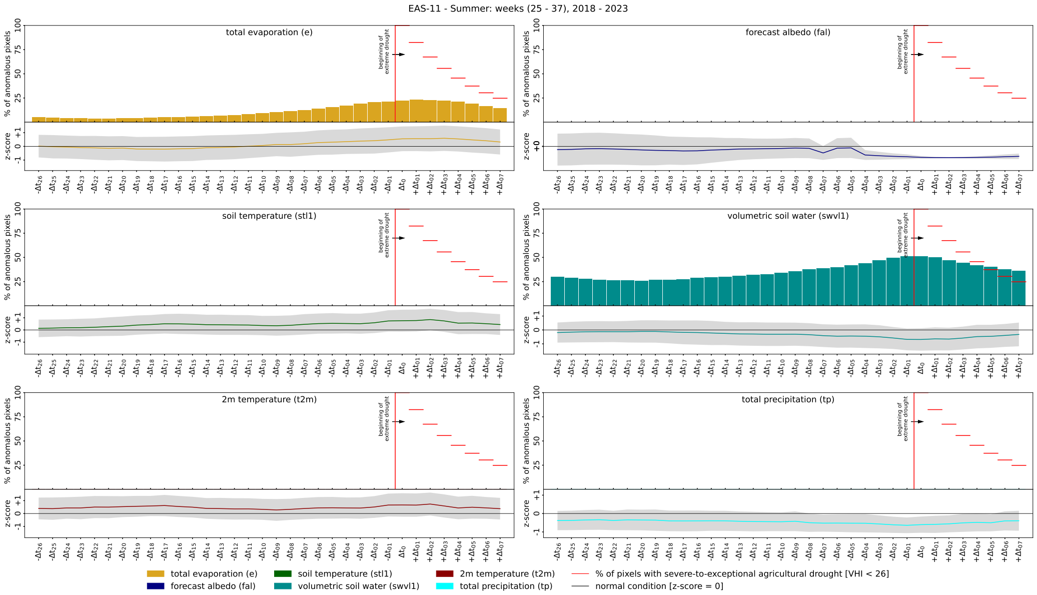


Figure 27: Temporal evolution of drivers and anomalies related to the extremes in ERA5-Land for East Asia (EAS-11). For this experiments, we select pixels with extreme events during summer (weeks 25-38) for the years 2018-2023 and compute the average distribution of drivers and anomalies with time. The red line at δt_0 indicates the beginning of the extreme droughts. Z_{score} in the underneath curve represents the deviation from the mean computed from the ERA5-Land climatology.

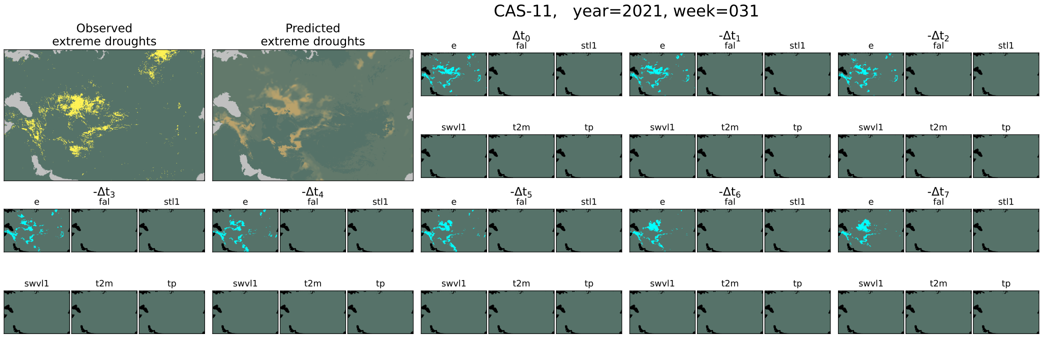

CAS-11, year=2021, week=031
Observed extreme droughts
Predicted extreme droughts
Δt$_0$
-Δt$_1$
-Δt$_2$
-Δt$_3$
-Δt$_4$
-Δt$_5$
-Δt$_6$
-Δt$_7$
e
fal
stl1
swvl1
t2m
tp

Figure 28: Qualitative results on ERA5-Land for Central Asia (CAS-11). Shown are the identified drivers and anomalies for each variable along with the prediction of extreme agricultural droughts on the top left.

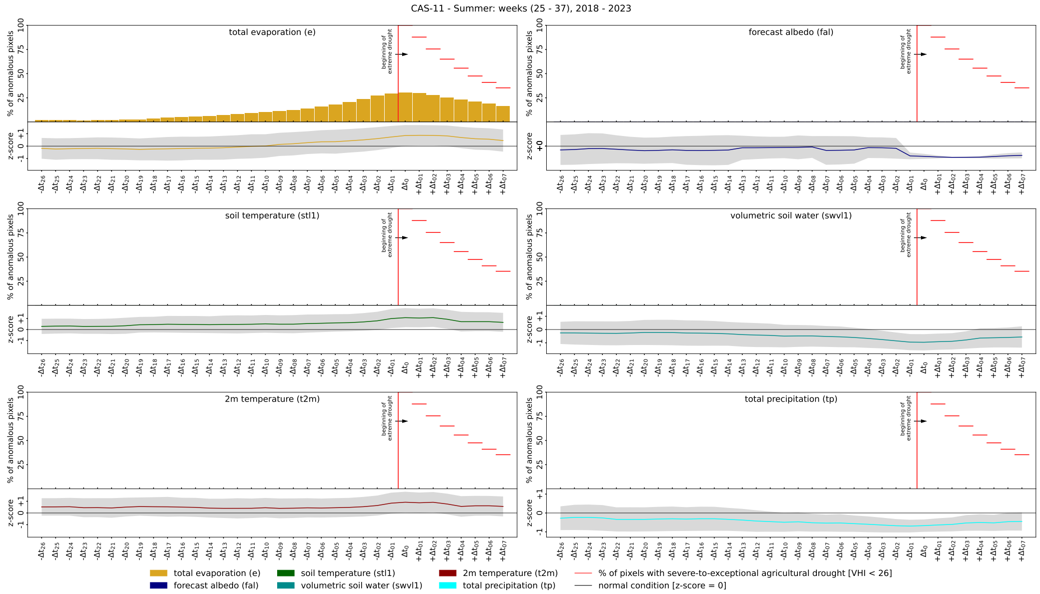


Figure 29: Temporal evolution of drivers and anomalies related to the extremes in ERA5-Land for Central Asia (CAS-11). For this experiments, we select pixels with extreme events during summer (weeks 25-38) for the years 2018-2023 nd compute the average distribution of drivers and anomalies with time. The red line at δt_0 indicates the beginning of the extreme droughts. Z_{score} in the underneath curve represents the deviation from the mean computed from the ERA5-Land climatology.

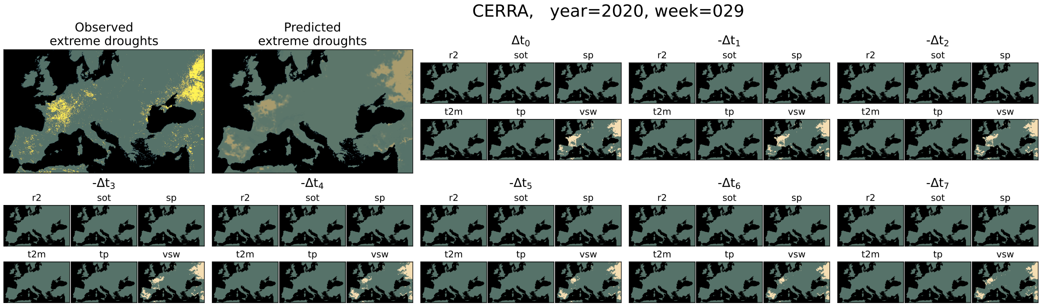

CERRA, year=2020, week=029
Observed extreme droughts
Predicted extreme droughts
Δt₀
-Δt₁
-Δt₂
-Δt₃
-Δt₄
-Δt₅
-Δt₆
-Δt₇
r2
sot
sp
t2m
tp
vsw

Figure 30: Qualitative results on CERRA reanalysis for Europe. Shown are the identified drivers and anomalies for each variable along with the prediction of extreme agricultural droughts on the top left.

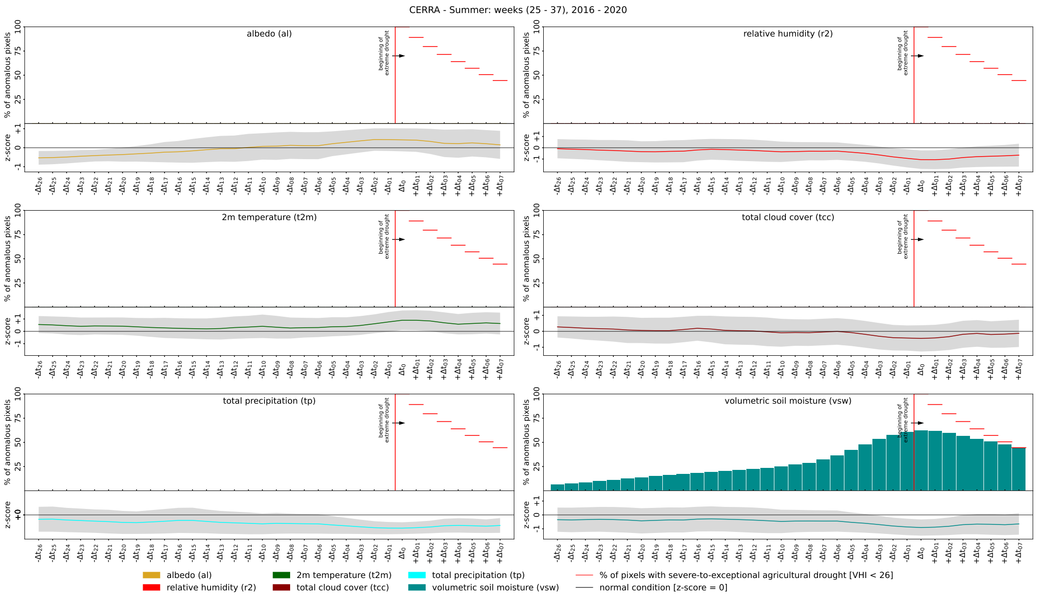


Figure 31: Temporal evolution of drivers and anomalies related to the extremes in CERRA reanalysis for Europe. For this experiments, we select pixels with extreme events during summer (weeks 25-38) for the years 2016-2020 and compute the average distribution of drivers and anomalies with time. The red line at δt_0 indicates the beginning of the extreme droughts. Z_{score} in the underneath curve represents the deviation from the mean computed from the CERRA climatology.

K Code and data availability

The source code to reproduce the results is available on GitHub at `https://github.com/HakamShams/IDEE`. The source code for the synthetic data generation is also available on GitHub at `https://github.com/HakamShams/Synthetic_Multivariate_Anomalies`. The pre-processed data used in this study are available at `https://doi.org/10.60507/FK2/RD9E33` [139].

L Broader impacts

There are generally no direct negative social impacts for conducting climate science researches. However, anomaly detection algorithms in general could be adapted for video surveillance and might infringe privacy considerations. Although this is the risk of developing anomaly detection algorithms.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Sec. 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Sec. 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper dose not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See implementation details in Sec. 5 and Appendix Sec. G.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and datasets are publicly available. See Sec. K.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See implementation details in Sec. 5 and Appendix Sec. G.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the mean and standard deviation of the results for 3 different random seed runs. See Tables 1, 6, 7, 14, and 18.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix Sec. F and Sec. G.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Sec. 6 and Appendix Sec. L.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We think the paper does not pose such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original publications for the raw data and provide URLs when applicable.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: See Sec. 4 and Appendix Sec. A, H, and I.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.