Multivariate Anomaly Detection for Earth Observations: A Comparison of Algorithms and Feature Extraction Techniques

Milan Flach[1], Fabian Gans[1], Alexander Brenning[2,4], Joachim Denzler[3,4,5], Markus Reichstein[1,4,5], Erik Rodner[3,4], Sebastian Bathiany[6], Paul Bodesheim[1], Yanira Guanche[3,4], Sebastian Sippel[1], and Miguel D. Mahecha[1,4,5]

[1]Max Planck Institute for Biogeochemistry, Department Biogeochemical Integration, P. O. Box 10 01 64, D-07701 Jena, Germany
[2]Friedrich Schiller University Jena, Department of Geography, Jena, Germany
[3]Friedrich Schiller University of Jena, Department of Mathematics and Computer Sciences, Computer Vision Group, Jena, Germany
[4]Michael Stifel Center Jena for Data-driven and Simulation Science, Jena, Germany
[5]German Centre for Integrative Biodiversity Research (iDiv), Leipzig, Germany
[6]Wageningen University, Department of Environmental Sciences, Wageningen, Netherlands

Correspondence to: Milan Flach (milan.flach@bgc-jena.mpg.de)

Abstract. Today, many processes at the Earth's surface are constantly monitored by multiple data streams. These observations have become central to advance our understanding of e.g. vegetation dynamics in response to climate or land use change. Another set of important applications is monitoring effects of climatic extreme events, other disturbances such as fires, or abrupt land transitions. One important methodological question is how to reliably detect anomalies in an automated and generic way within multivariate data streams, which typically vary seasonally and are interconnected across variables. Although many algorithms have been proposed for detecting anomalies in multivariate data, only few have been investigated in the context of Earth system science applications. In this study, we systematically combine and compare feature extraction and anomaly detection algorithms for detecting anomalous events. Our aim is to identify suitable workflows for automatically detecting anomalous patterns in multivariate Earth system data streams. We rely on artificial data that mimic typical properties and anomalies in multivariate spatiotemporal Earth observations like sudden changes of basic characteristics of time series such as the sample mean, the variance, changes in the cycle amplitude and trends. This artificial experiment is needed as there is no 'gold standard' for the identification of anomalies in real Earth observations. Our results show that a well chosen feature extraction step (e.g. subtracting seasonal cycles, or dimensionality reduction) is more important than the choice of a particular anomaly detection algorithm. Nevertheless, we identify 3 detection algorithms (k-nearest neighbours mean distance, kernel density estimation, a recurrence approach) and their combinations (ensembles) that outperform other multivariate approaches as well as univariate extreme event detection methods. Our results therefore provide an effective workflow to automatically detect anomalies in Earth system science data.

Keywords. statistical process control, process monitoring, Earth observations, artificial data, multivariate outlier detection, novelty detection, detection of extreme events, anomaly detection, event detection, k-nearest neighbours, kernel density estima-

tion, recurrences, support vector data description, kernel null ~~foley-sammon transform, mahalanobis~~ Foley-Sammon transform, Mahalanobis distance, Hotelling's T^2, multivariate exponential moving average.

1 Introduction

The Earth system can be conceptualized as a system of highly interconnected subsystems (e.g. atmosphere, biosphere, hydrosphere, lithosphere). Each of these subsystems can be monitored and characterized by multiple variables. Technological progress over the past decades has led to a boost in satellite technologies (Pfeifer et al., 2011; Nagendra et al., 2012) as well as ground station development and routine monitoring (Baldocchi et al., 2001; Dorigo et al., 2011; Ciais et al., 2014). Additionally, advanced computational methods efficiently integrate remote sensing and in-situ information to routinely derive novel data products (e.g., Beer et al., 2010; Jung et al., 2011; Tramontana et al., 2016). One key scientific challenge is co-interpreting these multiple views on the Earth system, in particular to address the impacts of changes in the climate system, the land use system, and other transformations.

Of particular importance is the analysis of extreme events like droughts, fires, heat waves or floods which are expected to change in a future climate (Kharin et al., 2013). One matter of concern are changes in hydrometeorological extremes that may translate into anomalies in vegetation dynamics, or extremes in vegetation dynamics that might result from slight changes in climatological conditions or human intervention, and that can have severe consequences for vegetation and the carbon cycle (Easterling et al., 2000; Meehl and Tebaldi, 2004; Seneviratne et al., 2012; Reichstein et al., 2013). Apart from natural events, one also aims at detecting events that are a direct consequence of human interference, e.g., detecting deforestation activities is required to assess the compliance with laws or agreements on forest conservation and climate change.

The flood of observational data is accompanied by a similar increase in data from Earth system models (Overpeck et al., 2011). As large amounts of data are difficult to handle and to translate to the quantities of human interest, it can be easy to overlook events of particular importance. For example, using a simple semi-automatic detection scheme to identify abrupt climate shifts in simulations of future climate, Drijfhout et al. (2015) found a number of abrupt events that have previously been overlooked in simulations.

In observations, anomalous events are often detected using extreme event detection methods suitable for univariate data streams (e.g., Alexander et al., 2006; Rahmstorf and Coumou, 2011; Zhou et al., 2011; Donat et al., 2013; Lehmann et al., 2015). Univariate extreme event detection can also be used to infer knowledge about underlying drivers of extremes (Zscheischler et al., 2014a); it is particularly valid when the variable of interest is either of specific importance or integrates a wide array of relevant processes. However, some information might only be inferred when taking the multivariate combination of several data streams into account (Vicente-Serrano et al., 2010; Seneviratne et al., 2012; Fischer, 2013; Zscheischler et al., 2015). For instance, a significant fraction of carbon extremes events in Europe is not associated with univariate climate extremes (Zscheischler et al., 2014b). Earth observations are multivariate and naturally characterized by strong dependencies and correlations in space, time, and across dimensions (Leonard et al., 2013). We assume that any suitable anomaly detection algorithm needs to consider these data properties. By considering multivariate constellations for anomaly detection, it might become possible to

gain further information, i.e. about anomalies which cannot be detected with univariate extreme event detection methods (for a review of approaches see, e.g., Ghil et al., 2011).

Multivariate approaches in geoscience make use of anomalies occurring simultaneously in multiple data streams, often referred to as coincidences or coexceedances (e.g., Donges et al., 2011b; Rammig et al., 2015; Zscheischler et al., 2015; Donges et al., 2016; Guanche et al., 2016) (Siegmund et al., 2016). An alternative is the copula approach introduced to the field e.g. by Schoelzel and Friedrichs (2008); Durante and Salvadori (2010). However, the copula approach so far is limited to 2 or 3 simultaneous data streams (Mikosch, 2006) which makes it unsuitable for high dimensional data as used in this paper.

Interestingly, there are multiple industrial applications that likewise require anomaly detection. In this context, anomaly detection has become a standard procedure in the wake of Harold Hotelling's publication of the T^2 control chart in 1947 (Hotelling, 1947; Lowry and Woodall, 1992). Consider, for instance several sensors observing some industrial production chain. These (potentially correlated) sensor data streams can be monitored with a Statistical Process Control (SPC) algorithm (Lim et al., 2014; Ge et al., 2013; Lowry and Montgomery, 1995). The basic idea is to raise an alarm as soon as an anomaly according to the SPC is detected, meaning that the production chain is 'out of control'. Despite the obvious analogy, the ideas of SPC are largely unknown in the geoscience community to the best of our knowledge. Conceptually, the industrial application is equivalent to the idea of monitoring environmental variables. However, data differ. Earth observations (EOs) exhibit strong (potentially non-linear) dependencies among the variables, seasonal cycles are typically present in both temporal mean and variance. The variables may also encode dynamic feedbacks and abrupt transitions. EOs are possibly more strongly corrupted by noise compared to industrial applications. Furthermore industrial applications are typically less affected by low-frequency variability than Earth observations. The most problematic aspect when considering SPC concepts in Earth system sciences is, however, defining states of normality.

The objective of this study is to provide an overview and comparison of anomaly detection algorithms and their combination with feature extraction techniques for identifying multivariate anomalies in EOs. ~~Therefore~~ Spatio-temporal EOs are therefore stored in the Earth system data cube (data cube), which is a 4 dimensional array of latitudes, longitudes, time and different measurement variables. To detect multivariate anomalies in EOs, we define an anomaly to be any consecutive spatiotemporal part of the data cube, which differs with respect to the mean, the variance, the amplitude of the seasonal cycle of trends from the 'normal' rest of the data cube. We adapt algorithms from SPC and novelty detection. The study is structured as follows: First, we create a series of artificial Earth system data cubes that try to mimic a series of real world features (in terms of multiple variables, seasonal cycles and correlation structure, etc). We are aware that these artificial data cubes are not 'real' simulations of Earth system data cubes. However, relying on artificial data in this paper is motivated by the fact that a meaningful quantitative evaluation of unsupervised anomaly detection algorithms and feature extraction techniques in 'real' Earth observation data is difficult due to the lack of ground-truth data (Zimek et al., 2012). Second, we use these artificial data to evaluate the capability of different algorithms to detect multivariate anomalous events~~.~~, including compound events (i.e. events where none of the single variables is extreme, but their joint distribution is anomalous and might lead to an extreme impact) (Seneviratne et al., 2012; Leonard et al., 2013). Specifically, we evaluate the performance of the algorithms to detect multivariate changes in the mean (comparable to an extreme event), the amplitude of the annual cycle, the variance and onset

of trends. Using the artificial dataset as testbed we apply various feature extraction schemes (Sect. 3.1), several detection algorithms (Sect. 3.2) as well as combinations of detection algorithms (ensembles, Sect. 3.4) to compare their performance in identifying anomalous events (Sect. 3.3). From this comparison we select suitable combinations of feature extraction (Sect. 4.1) and a few algorithms (Sect. 4.2) as well as ensembles of algorithms (Sect. 4.3) as the best ones applicable to EOs including suggestions for their specific usage (Sect. 5).

2 Experimental Setup

2.1 Generation Principle of the Artificial Data

Ground truth for detecting anomalies in multivariate data is rare, in particular for detecting anomalies in 'real' EOs. Thus, we generate artificial data that represent common properties of EOs, including anomalies. In particular, we focus on the existence of seasonality, correlations among variables, and non-Gaussian distributions. Data generation assumes that each subsystem of the Earth has uncorrelated intrinsic properties, i.e. it is dominated by a few independent components. Consequently, generating these independent components (which cannot directly be monitored) is the first step. We then derive variables that contain elements of all independent components and correspond to the 'observable' measurements as a set of correlated variables (Fig. 1).

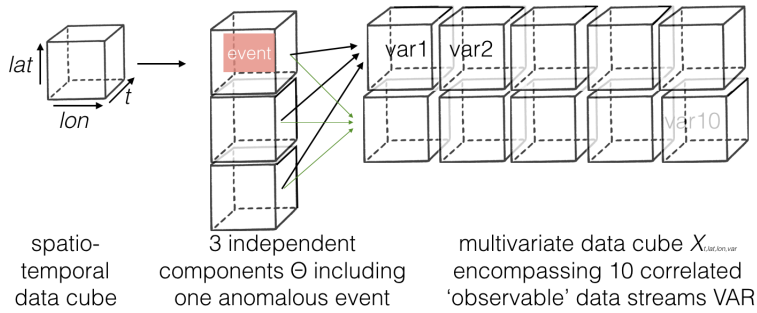

Figure 1. Combination of 3 independent component cubes to derive 10 correlated variables X as 'observable' measurements. The anomalous event is propagated into some variables of X.

More precisely, as basic version we create 3 independent components for the artificial data, each consisting of a signal (Gaussian, $sd = 1.0$) which includes seasonality in some cases (Sect. 2.3). Anomalous events are induced in one of the independent components for which we track the exact spatiotemporal location. These 3 independent components are then weighted with randomly generated linear (or non-linear, Sect. 2.3) weights to create a set of 10 correlated variables, which represent the artificial data cube, i.e. try to mimic 'observable measurements'. We add some additional measurement noise (Gaussian, $sd = 0.3$) to the data cube. For more technical details of this generation scheme we refer the reader to the Appendix A.

Our standard data cube $X_{t_{i,j},lat,lon,var}$ encompasses $t_{i,j} = 1,\ldots,T$ time steps ($T = 300$) corresponding e.g. to a 6.5-year time series of satellite images in 8-day intervals, $lat = 1,\ldots,LAT$ latitudes ($LAT = 50$), $lon = 1,\ldots,LON$ longitudes ($LON = 50$) and $var = 1,\ldots VAR$ data streams, or variables ($VAR = 10$).

2.2 Generating Anomalous Events

Anomalous events are introduced on the independent components only and then propagated from the independent component to some of the variables in the data cube with random weights. The anomalies are contiguous in space and time. The center of the anomaly is assigned randomly. The challenge is to detect the propagated anomaly through the unsupervised algorithms, i.e. without using the information about the spatiotemporal location of the anomaly. With this data cube generation scheme, we can generate anomalies by controlling the type of the anomalous event (event type), the magnitude of the anomalous event as well as the spatiotemporal location.

We create 4 data cubes using the following temporary event types:

a Shift in the baseline, i.e. shift of the running mean of a time series (*BaseShift*) (Fig. 2 ~~a)~~ (a)). This event type is closely related to "extremes" in real-world Earth observations.

b Onset of a trend in the time series (*TrendOnset*) (Fig. 2 ~~b)~~ (b)).

c Change in the amplitude of the mean seasonal cycle of a time series (*MSCChange*) (Fig. 2 ~~c)~~ (c)), which might happen in the real-world carbon cycle as response to combined drought-heatwaves (Ciais et al., 2005).

d Change in the variance of the time series (*VarianceChange*) (Fig. 2 ~~d)~~ (d)), e.g., in temperature (Huntingford et al., 2013).

2.3 Additional ~~Complications~~Data Properties

Apart from the basic data cubes we want to test the influence of a certain ~~complication~~ data properties on the anomaly detection algorithm. In order to do so, we create data cubes, each with one added ~~complication~~data property, i.e. we increase the number of independent components (*MoreIndepComponents*) or use a squared dependency among independent components (*NonLinearDep*) instead of a linear one. Furthermore typical EO variables are often driven by extrinsic forcings, i.e. the Earth's solar system orbit, rotation, and axis tilt, thus we add a seasonal cycle modifying the signal (*SeasonalCycle*). In a global context, the mean is rarely constant; we therefore introduce a linear latitudinal trend into the baseline (*LatitudinalGradient*). In the basic case, the signal of our independent components follows a Gaussian distribution. In the more complicated versions, we also implement alternative scenarios with Laplacian ('doubly exponentially') distributed signals (*LaplacianNoise*) and signals which exhibits spatiotemporal correlation with red noise (*CorrelatedNoise*). Signal-to-noise ratio is 0.3 in the basic version, one ~~complication~~ additional data property increases the signal-to-noise ratio to 1.0 (*NoiseIncrease*). Also the shape and duration of anomalous events differs. We double (*LongExtremes*) or reduce the temporal duration of the anomalous events (*ShortExtremes*) and change the spatial shape from rectangular to randomly affecting neighbouring grid cells (*RandomWalkExtreme*).

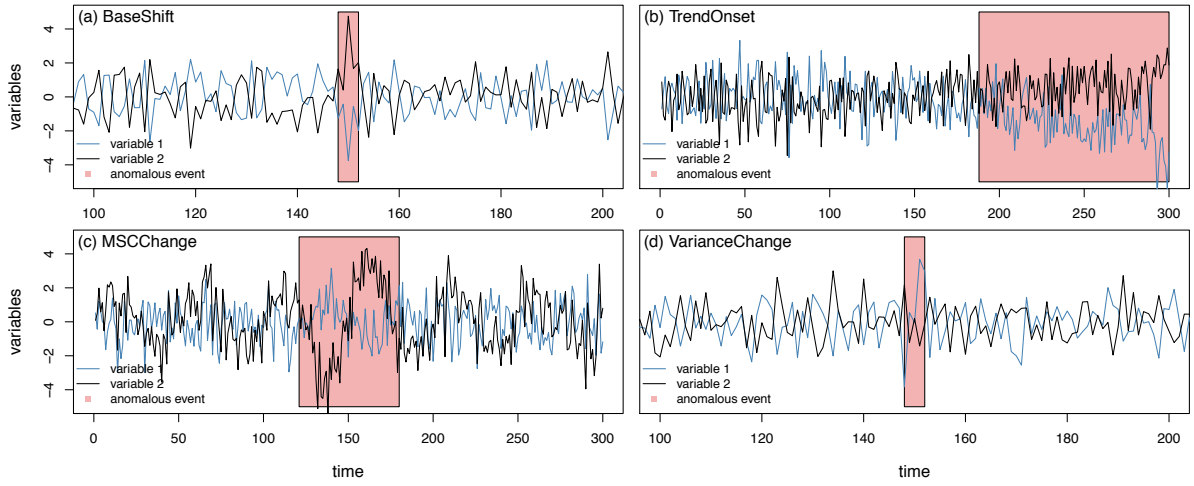

Figure 2. Visualization of the 4 different event types (a-d) with 2 variables along time $t_{i,j} = 1, \ldots, T$ ($T = 300$). The 2 variables contain an anomalous event (~~here: 60 time steps long~~red shape) which is propagated through the underlying independent components with randomly drawn weights within the generation process of the variables. For illustration purposes 2 variables are shown for one specific magnitude of the anomaly. The artificial data farm encompasses 10 variables and anomalous events of 20 different magnitudes ranging from very subtle to exceptionally high changes.

2.4 Experiment design

Each data cube with a specific type of the event is generated 20 times, each time with a different magnitude of the anomalous event (Appendix A). We introduce 10 spatially contiguous anomalous events into the independent components, with a spatial extent of 20 latitude and longitude steps each. Each event has a temporal extent of 5 time steps (which would be equivalent to 40 consecutive anomalous days in a 6.5 year record). Our total amount of anomalies equals about 3 % of the total data cube which we consider to be a realistic scenario (comparable to e.g., Zscheischler et al., 2014a). Some latitudes and longitudes do not exhibit any anomaly by design. The algorithms (Sect. 3.2) are expected to be able to deal with parts of the data cube that do not exhibit anomalies at all, as this is also very likely to happen for applications in real Earth observations.

Our experiment comprises 36 different event type combinations of ~~complications~~data properties, each repeated 20 times with varying event magnitudes (Appendix A). The entire set of artificial data cube consists of 720 data cubes, corresponding to ≈87 GB of data [1].

[1] Code to reproduce the data farm is provided in Appendix A.

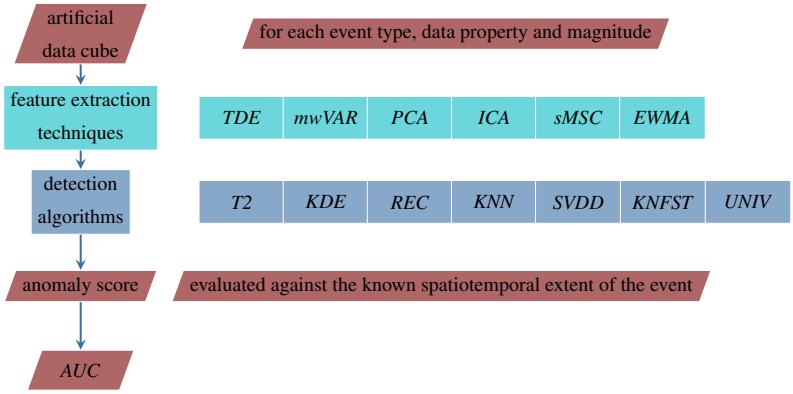

Figure 3. Data processing for detecting multivariate anomalies. We extract relevant features from each artificial data cube before applying the detection algorithms. The detection algorithms output some anomaly score which we evaluate against the known extent of the event using the Area Under the Curve (*AUC*). Feature extraction elements on the right hand side are understood as options and can be combined with each other.

3 Workflows to Detect Anomalies

The idea of this study is to elaborate workflows that contain both data preprocessing via feature extraction and algorithms for the detection of anomalous events (Fig. 3). In the following we introduce these 2 elements separately and explain the performance evaluation strategy afterwards.

3.1 Feature extraction

'Feature extraction' is a process to derive information from the data and condense it into non-redundant characteristic patterns. This may facilitate data interpretation (van der Maaten, 2009). In our study the aim is to maximize the ~~event detection rate~~ detection of anomalous events by providing relevant features. Feature extraction is often an element of data preprocessing. A very simple form of feature extraction could be to subtract the mean seasonal cycle~~(~~. We consider the anomaly time series ~~becomes the feature then)~~to be the extracted feature in this case. Here, we concentrated mainly on feature extraction methods that are used in the context of classical multivariate SPC (Lowry and Montgomery, 1995), data-based process monitoring in industry (Ge et al., 2013), and univariate extreme event detection. The following feature extraction methods are used in this study:

Subtracting the Median Seasonal Cycle (*sMSC*) is one way to deseasonalize time series. Deseasonalization may be instrumental in detecting anomalous events across different seasons. The remaining part of the time series is often referred to as anomalies and used here as input feature.

Computing the Moving Window Variance (*mwVAR*) is a popular technique for detecting trends in the variance in univariate time series (e.g., Huntingford et al., 2013). We ~~compute the with~~ choose a window size of 10 and ~~subtract the median~~ compute

the variance in the running window along the time series of each variable. We use the estimates of the ~~temporal moving window variance~~ *mwVAR* time series as feature to detect multivariate anomalies in the variance.

Time Delay Embedding (*TDE*) increases the feature vector Y_t with time delayed vectors ($Y_t = (X_t - 0\tau, X_t - 1\tau, X_t - (m-1)\tau)$) to include temporal context information. In the univariate case, this approach ideally creates an image of the attractor of a dynamical system (Takens, 1981). ~~The theoretical consideration does not hold true for~~ In high dimensional multivariate data ~~, but is nevertheless used in practical applications~~ applications it is used to include information of the dynamics in the feature vector (e.g., Koçak et al., 2004; Ge et al., 2013; Smets et al., 2009). Critical hyperparameters are the time delay τ and the number of dimensions m. We fix m to 3 (corresponding to the number of independent components within the data farm creation) and τ to 6 which is a compromise between the typical choice of the first zero crossing of the temporal autocorrelation function or the first local minimum of the mutual information (Webber and Marwan, 2015) (here: 11.5 corresponding to one quarter of the annual cycle with 46 time steps) and an accurate temporal detection (requires small τ).

Principal Component Analysis (*PCA*) is a data rotation, used to find an orthogonal (uncorrelated) subspace of the data of $n_{PC} \leq VAR$ variables (Von Storch and Zwiers, 2001). We choose n_{PC} such that at least 95 % of the variance in the original data cube are explained. By assuming a homogeneous covariance structure within the entire data cube, we perform the *PCA* globally, i.e. with the same rotation matrix for all grid cells. The combination of *TDE* and *PCA* is sometimes referred to as dynamic *PCA* when considering subsequent lags in the time series (Lee et al., 2004).

Independent Component Analysis (*ICA*) can be regarded as a nonlinear alternative to PCA; it has become a standard technique of data-based process monitoring~~, trying~~. We use one *ICA* variant which tries to separate different sources of data by maximizing the negentropy, a measure of non-Gaussianity of the data (Hyväringen and Oja, 2000)[2]. We apply *ICA* globally to each data cube. The hyperparameter is the number of independent components (sources). We choose the number of independent components to be equal to n_{PC} (see *PCA*) for consistency reasons (Majeed and Avison, 2014).

Exponentially Weighted Moving Average (*EWMA*) is one way of reducing the noise of the time series and taking temporal information into account. It is common in the context of classical multivariate SPC to detect only 'significant' outliers (Lowry and Woodall, 1992). The multivariate feature time series Y is computed recursively as

$$Y_{t_i} = \lambda X_{t_i} + (1-\lambda) Y_{t_{i-1}}. \quad (1)$$

The hyperparameter λ determines the degree of exponential weighting between 1 (no weighting) and zero (common choice $0.1 \leq \lambda \leq 0.3$, Santos-Fernández, 2012). We stay in this range with $\lambda = 0.15$.

There is of course a multitude of alternative approaches available in the literature, but we focus on the previously summarized ones as they are widely used and efficiently implemented. Furthermore, different feature extraction methods can also be combined (Fig. 3). As the number of possible combinations is considerably large, we focus here on dimensionality reduction techniques (*ICA*, *PCA*) combined with some *EWMA* to reduce the noise level afterwards. Depending on the event type and ~~complication~~data properties, additionally removing seasonality (*sMSC*) or including the variance *mwVAR* seems to be straightforward. Information about the dynamics (*TDE*) can be included before applying dimensionality reduction techniques, to keep

[2]We use the *fastICA* algorithm implemented in the julia package MultivariateStats.jl (https://github.com/JuliaStats/MultivariateStats.jl).

the dimensionality of the system as low as possible. In the following, combinations are noted in the order in which they were applied (e.g., *PCA_EWMA* means first applying *PCA*, then applying *EWMA* on the *PCA* features). In some cases this might lead to non-commutative combinations, especially for non-linear feature extraction techniques (*ICA*, *TDE*).

3.2 Anomaly Detection Algorithms

We use several detection algorithms which we implemented in the julia package MultivariateAnomalies.jl[3]. ~~We fix~~ Some anomaly detection algorithms require the estimation of parameters (Details are given below for each algorithm separately). In that case we fix the model parameters for the entire data cube. ~~Model~~ We estimate model parameters (σ, ε, Q, μ~~) and~~ , see below) and train the models themselves (Support Vector Data Description, Kernel Null Foley-Sammon Transform, see below) ~~are estimated~~ based on a random subsample of 5000 data points obtained from the entire data cube. To account for variability of the model parameter estimation, we resample 3 times. More resampling is not affordable due to high computational costs of processing the large number of data cubes. However, very little random variability is observed with this sample size for the best algorithms. Thus, we consider a resampling of 3 times to be sufficient for a first attempt accounting for variability in the parameterization. The following algorithms are investigated for anomaly detection.

Univariate Approach (*UNIV*). A simple approach to define extremes in univariate data is to identify all points above (or below) a certain quantile. This so-called 'peak-over threshold' approach can be transferred to deal with multiple univariate data streams. In this case, one would consider a data point to be extreme, if one or several of the univariate variables are ~~below or above~~ above (or below) a certain quantile threshold of the marginal distributions of each single variable. (here: globally) (e.g., Ledford and Tawn, 1996; Bae et al., 2003; Donges et al., 2016). Applications of the so-called cooccurrence or coincidence analysis can be found in Donges et al. (2011b); Rammig et al. (2015); Zscheischler et al. (2015); Guanche et al. (2016); Siegmund et al. (2016). For comparing the algorithms, we are interested in the information that at least one variable is above a certain threshold. We compute this information for ~~all possible thresholds to~~ different thresholds (in terms of quantiles of the marginal distributions between 0.0 to 1.0, accuracy 0.01) to get a score, i.e. a ranking of the extremeness of the data points.

Hotelling's T^2 (*T2*) computes the squared Mahalanobis distance of each data point X_t to its temporal mean μ weighted with the covariance matrix Q (Hotelling, 1947):

$$(X_t - \mu)' Q^{-1} (X_t - \mu) \tag{2}$$

A crucial prerequisite is the estimation of the covariance matrix Q, which is estimated from the random subsample of 5000 data points. Combining the feature extraction *EWMA* with *T2* equals the traditional multivariate exponential weighted moving average (Lowry and Woodall, 1992; Lowry and Montgomery, 1995).

Apart from computing weighted distances to the mean (like *T2*), it is also possible to compute pairwise Euclidean distances in variable space $d(X_{t_i}, X_{t_j})$ between vectors X_{t_i} and X_{t_j} of time step t_i and t_j for all possible timesteps $t_i, t_j = 1 \dots T$. The resulting matrix D with $D_{ij} = d(X_{t_i}, X_{t_j})$ is often referred to as distance matrix or dissimilarity matrix. For real-world data,

[3] https://github.com/milanflach/MultivariateAnomalies.jl

variables have to be standardized with care before computing the distance matrix (Sect. 5). However, in the used artificial data the variables are already comparable by construction, thus standardization is not needed. The following algorithms are based on pairwise distances.

k-nearest neighbours (*KNN*) can be used for anomaly detection by considering the mean distance to the k-nearest neighbors (k-nearest neighbours Gamma (*KNN-Gamma*)) and the ~~mean~~ length of the mean of the vectors pointing from X_{t_i} to its k-nearest neighbors (k-nearest neighbours Delta (*KNN-Delta*)) (Harmeling et al., 2006; Ramaswamy et al., 2000). With that approach KNN-Delta considers also the direction of the neighbors, i.e. has higher values in case its nearest neighbours are pointing in one direction, which is in contrast to the directionless distance of KNN-Gamma. We fix the hyperparameter k at 10 after carefully trying different choices for k without seeing major effects on preliminary results. Furthermore, we ~~take advantage of the temporal structure of anomalous events~~ exclude trivial temporal autocorrelations by excluding 5 neighbouring time steps ($abs(t_i - t_j) \geq 5$) to be also nearest neighbours.

Recurrences (*REC*). Within the framework of the theory of nonlinear dynamical systems, each state of a dynamical system will revisit a particular region in its phase space, if waiting for a sufficiently long time (Poincaré, 1890). These dynamics can be visualized in the recurrence plot and are quantified with several metrics usually referred to as recurrence quantification analysis (Marwan et al., 2007). It seems straightforward to use the concept of recurrence analysis to detect states in a dynamical system that are considered to be rare or unusual. Faranda and Vaienti (2013) used the concept of recurrences and combined it with extreme value theory. We want to use a more general approach without binning the time series. We count the number of observations ζ falling into a certain ε-ball in a system of multiple variables, condensed by their distance $d(X_{t_i}, X_{t_j})$:

$$\zeta(X_{t_i}) = \sum_{j=1}^{T} \Phi(\varepsilon - d(X_{t_i}, X_{t_j})) \tag{3}$$

$\Phi(z)$ is the Heaviside function, coding the distances to binary values ($\Phi(z) = 0$ if $z < 0$, $\Phi(z) = 1$ otherwise). ~~A~~ Am ε-hyperball containing only few recurrent observations is considered to be rare in comparison to the majority of ζ. We compute ~~$1 - \zeta cdot T^{-1}$~~ $1 - \zeta \cdot T^{-1}$ to get anomaly scores, which are more likely to be an anomaly for high score values. $\zeta \cdot T^{-1}$ known as local recurrence rate or degree ~~of centrality~~ density in recurrence analysis (Marwan et al., 2007; Donner et al., 2010) (Donges et al., 2012). ε is the crucial hyperparameter, defining the radius of the ball. Typical choices of ε in recurrence analysis are ~~using~~ quantiles of the distribution of elements of the distance matrix, e.g., 5 % or 10 % (Donges et al., 2011a; Flach et al., 2016). As we are not interested in small scale variations falling of *REC*, but more in major anomalies we estimate ε as median of the distance matrices on the random subsample. This choice turned out to be the optimal choice for ε in a small simulation, varying the thresholds between the 5 % to 95 % quantile of the element of the distance matrix (~~not shown~~Fig. S1). We exclude 5 neighbouring timesteps to be counted as recurrences (similar to *KNN*). *KNN* has similarities to *REC*, as one could also choose a data-adaptive k such that $\zeta = k$.

The distance matrix D can be transformed into a kernel matrix $K = exp(-0.5 \cdot D \cdot \sigma^{-2})$, i.e. by computing pairwise dissimilarities using Gaussian kernels centered on each data point.

Kernel Density Estimation (*KDE*) is a standard technique for estimating densities based on column means of the kernel matrix K (Parzen, 1962). The bandwidth σ of the kernel is a hyperparameter. We estimate σ by using the median of the

temporal distance matrix on the random subsample, which is a common choice (Schölkopf and Smola, 2001; Schölkopf et al., 2015).

Support Vector Data Description (*SVDD*) models the distribution of the training data with an enclosing hypersphere in a high-dimensional kernel feature space (Tax and Duin, 2004). As usual a kernel matrix of the random subsample is used for training. Although being a rather simple data description, a hypersphere in the kernel feature space can result in complex nonlinear decision boundaries in the original space of predictor variables if a nonlinear kernel function is used. Beside the σ hyperparameter of the kernel function (see *KDE*), the *SVDD* approach has a parameter called outlier ratio ν (fixed to 0.2). ν controls the amount of training samples that can be located outside of the hypersphere to prevent overfitting. As anomaly score for testing, its distance to the center of the hypersphere in the kernel feature space is computed. Testing requires pairwise similarities between test and training samples. For performance reasons in terms of computation time, we used the LIBSVM (Chang and Lin, 2013) implementation of the one-class ~~Support~~ support vector machine (Schölkopf et al., 2001), which is an alternative formulation that leads to identical data descriptions as *SVDD* in our setup.

Kernel Null Foley-Sammon Transform (*KNFST*) maps the training data into a so-called null space, in which the training samples have zero variance, i.e., all training samples are mapped to the same point called the target value (Bodesheim et al., 2013). Nonlinearity is incorporated by using a kernel matrix containing pairwise similarities of the training samples (training on the random subsample as for *SVDD*). Since all training samples are represented by a single target value in the one-dimensional null space, the anomaly score of a test sample is the absolute difference between its projection in the null space and this target value. The projection of the test sample requires pairwise similarities to the training samples. Compared to *SVDD* no parameters need to be tuned except for σ of the kernel function that ~~are~~ is fixed to the same values for all kernel methods.

3.3 Ranking of the Workflows

Given the large number of potential combinations of feature extraction and anomaly detection algorithms, we need an objective criterion to compare the performances of the numerous possible workflows. We use the Area Under the receiver operator characteristics Curve (*AUC*) as our measure of detection skill for a specific event type (Fawcett, 2006). The *AUC* is based on the fraction of events that are correctly detected (true positives) and the fraction of (false) detections among all non-events (false positives), for all possible decision thresholds that could be applied to scores produced by the algorithms. *AUC* values of 0.5 would be achieved by random detection, and values below 0.5 indicate that a lower score is more likely assigned to (true) anomalies than to non-anomalies.

For each data cube with a given event magnitude and event type we compute the *AUC* for each ~~complication~~data properties, feature extraction and algorithm combination. This leads to an entire catalogue of possible combinations, namely $1.27 \cdot 10^5$ (4 event types, 20 event magnitudes, 11 ~~complications~~data properties, 18 feature extraction combinations, 8 algorithms). The number of combinations strongly requires simplification to infer knowledge about which combination is advisable to use. Hence, we focus on events of magnitudes typically detected in real world data i.e. ~~changes in~~ deviations from the mean (extremes) larger than 2 sd (e.g., temperature extremes in Hansen et al., 2012), a relative increase or decrease in the mean annual cycle amplitude of 25 % (which might happen, e.g. in the carbon cycle after combined drought and heatwaves (Ciais

et al., 2005), or in the Arctic due to abrupt sea ice losses (Bintanja and van der Linden, 2013; Bathiany et al., 2016)) or an increase in the signal variance of 25 % (e.g. in temperature, Huntingford et al., 2013).

One way of summarizing the results of such a large number of combinations is treating the *AUC* values as the outcomes of an experiment in which the different design decisions (e.g., feature extraction techniques, anomaly detection algorithms) are the experimental factors. As a control treatment we introduce the simplest possible approach to detect the anomaly: *UNIV* approach on the selected event type, without any further ~~complications~~ data properties (e.g. short extremes or increase measurement noise) on the event type and without prior feature extraction. In order to assess the (averaged) effect of each experimental factor, we fit a linear mixed-effects model (Pinheiro et al., 2016) to the *AUC* data (fixed effects: ~~complications~~data properties, feature extraction, anomaly detection algorithms; random effect: magnitude of the event). This model's coefficients express the overall effect of a factor level with respect to the control while averaging over all other experimental factors. They are considered to be significant for $p < 0.01$.

Additionally, we compute the Resampling Variation of Parameter estimation of the anomaly detection algorithms (*RVP*) as mean difference of the maximum *AUC* and minimum *AUC* for each resampling $i = 1 \dots 3$ (Sect. 3.2).

$$RVP_{algorithm} = mean(max(AUC_{comp,feat,magn,event,i}) - min(AUC_{comp,feat,magn,event,i})) \quad (4)$$

3.4 Ensembles of Anomaly Detection Algorithms

Summarizing the output of several anomaly detection algorithms is one way to create more robust results (Thompson, 1977). For better comparability of the algorithms' outputs, we rank them by computing the percentiles of the algorithm scores. These are then aggregated into ensemble scores by computing the Minimum (min, 'Consensus voting'), the Mean ('Balanced voting') or the Maximum (max, 'Risky voting') of the scores of selected well performing algorithms (e.g., Aggarwal, 2012; Zimek et al., 2013).

4 Results & Discussion

In the following, we present the performance of the workflows in subsections corresponding to feature extraction techniques (Sect. 4.1), anomaly detection algorithms (Sect. 4.2), and ensembles of detection algorithms (Sect. 4.3). Specifically, we present the *AUC* difference to the *UNIV* control, i.e. the output of the linear mixed-effects model on the experimental factors 'feature extraction' and 'detection algorithm' (Fig. 4). The corresponding tables present the estimates as well as the *RVP* (Tab. 1, 2). Apart from the model the full range of *AUC* values with respect to different event magnitudes, ~~complications~~ data properties and event types is presented in Appendix B, Fig. B1.

4.1 Feature Extraction Techniques

Feature extraction techniques are often more important than the detection algorithm itself (Fig. 4). However, we find that choosing a suitable feature extraction technique largely depends on the event type of interest. Therefore, the feature extraction techniques are presented for different event types separately.

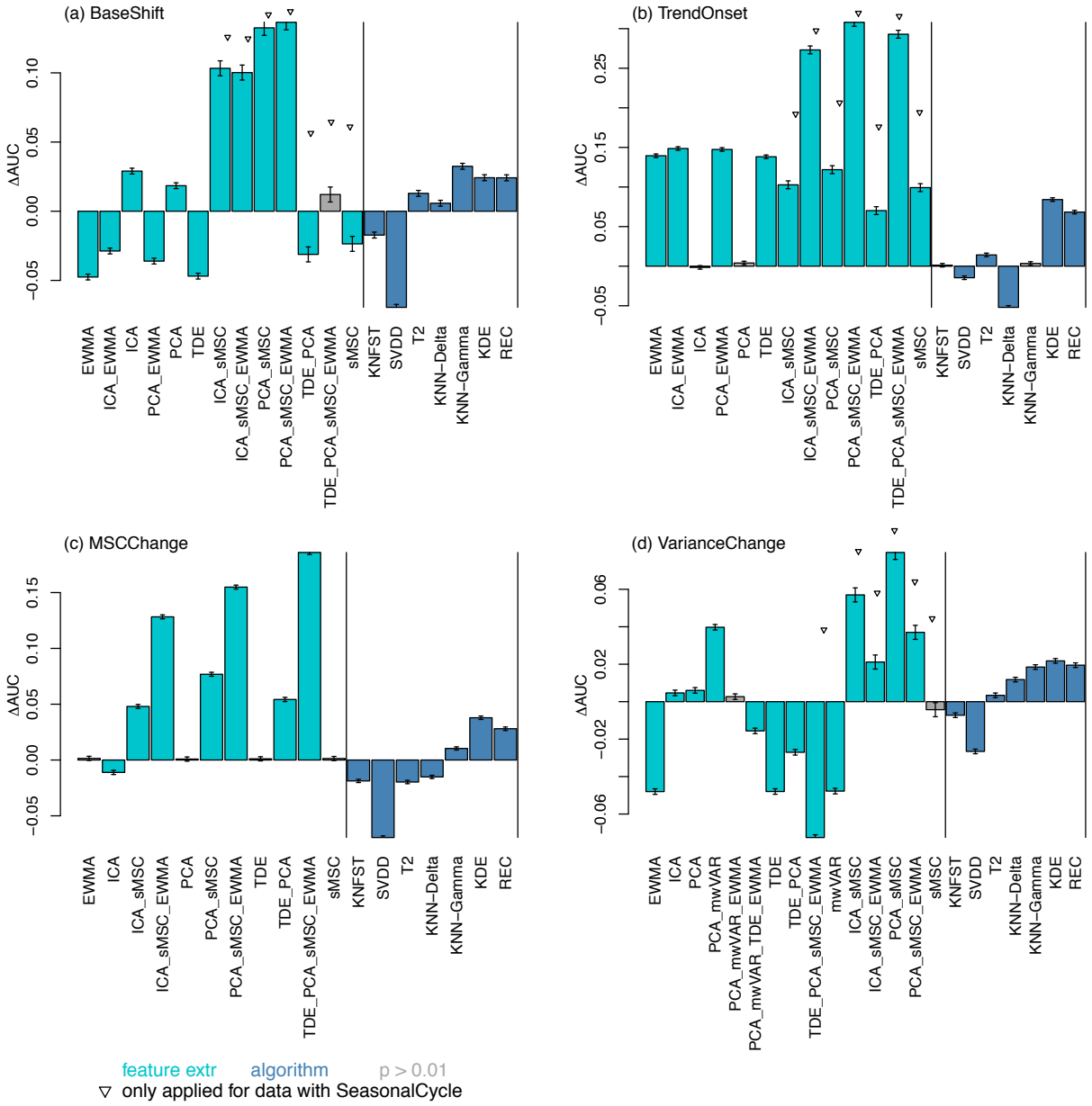

Figure 4. *AUC* difference with respect to the *UNIV* control in the experimental factors 'feature extraction' and 'detection algorithm' for the event types (a-d).

BaseShift. Shifts in the baseline are simulated to mimic extreme events. Increasing the magnitude (in terms of standard deviations) of a *BaseShift* makes it easier to detect the event (Fig. B1). Dimensionality reduction (via *PCA* or *ICA*) is a crucial feature extraction technique step as it derives meaningful uncorrelated subsets of the data (Fig. 4 a). The combination of dimensionality reduction with some temporal smoothing (*EWMA*) does not exhibit better overall performance (Fig. 4 a) as it fails for *ShortExtremes* due to oversmoothing. Nevertheless *EWMA* can improve the detection rate for special cases, i.e. long events (*LongExtremes*) and high signal to noise ratios (*NoiseIncrease*) (Fig. B1).

TrendOnset. Results look very similar to those of *BaseShift*, except that temporal smoothing with *EWMA* has a stronger positive effect than for *BaseShift*. This may be related to the fact that events for *TrendOnset* are longer than those for *BaseShift*. Since the algorithms used in this work are not specifically designed to detect the onset of linear trends, we speculate that their capability to detect such anomalies may be related to their ability to detect base shifts. While algorithms specifically designed to detect changes in trends (e.g., Forkel et al., 2013)) were not included in our work due to our focus on more generic types of anomalies, such specialized algorithms may perform better for this particular class of anomaly.

MSCChange. In the detection of *MSCChange* most feature extraction algorithms showed some skill in the detection of an amplitude increase, while only a subset of these succeeded also in detecting decreases in amplitude (Fig. B1). We focus on the the latter ones, which have one step in common: they subtract the median seasonal cycle before applying the detection algorithm (*sMSC*) (Fig. 4 c). In line with the results for *TrendOnset* and *BaseShift*, temporal smoothing in combination with dimensionality reduction improves detection by a large margin (*PCA_sMSC_EWMA*). Furthermore accounting for temporal dynamics with a time delay embedding *TDE* is even more suitable (*TDE_PCA_sMSC_EWMA*).

VarianceChange. The algorithms used are hardly able to detect any decrease in variance (Fig. B1). This may be due to an 'overwriting' of the decrease in signal variance with the independent noise since we are using a signal to noise ratio of 0.3. Thus, we exclude a decrease in the variance from the evaluation of the detection algorithms compared to the control. The detection of an increase in the variance can be improved by a combination of dimensionality reduction and variance in a moving window (*PCA_mwVAR*) (Fig. 4 d). Using the variance in a moving window is a popular approach (Huntingford et al., 2013) although it has to be applied with care when used in conjunction with normalization procedures (Sippel et al., 2015).

SeasonalCycle. Seasonality is occurring in most EOs. Not accounting for the seasonal cycle has a negative impact on the *AUC* (Appendix B, Fig. B2 a,b,d). However, if we subtract the median cycle within the feature extraction step (*PCA_sMSC_EWMA*, Fig. 4 a,b,d)), we can almost account for the negative *AUC* impact of the seasonal cycle, as in our experimental setting anomalous events do occur independently of seasonality. However, depending on the research question, independence of seasonality might not always be the case: some EOs may depend, e.g. on vegetation activity, which results in a strong dependence on seasonality.

4.2 Performance of Multivariate Anomaly Detection Algorithms

In contrast to the investigated combinations of feature extraction methods, we can identify 3 of the tested algorithms performing on average almost equally well for most event types given a suitable feature extraction as discussed before (Sect. 4.1).

Table 1. Average *AUC* difference of the Anomaly Detection Algorithms to the *UNIV* control for each event type.

	KNFST	*SVDD*	*T2*	*KNN-Delta*	*KNN-Gamma*	*KDE*	*REC*
BaseShift	-0.017	-0.069	0.013	0.006	0.032	0.024	0.024
TrendOnset	0.001	-0.015	0.014	-0.052	0.003	0.084	0.068
MSCChange	-0.023	-0.072	-0.023	-0.019	0.007	0.039	0.029
VarianceChange	-0.007	-0.027	0.003	0.012	0.018	0.022	0.019
Mean	-0.012	-0.046	0.002	-0.013	0.015	0.042	0.035
RVP	0.007	0.111	0.003			0.000	0.001

KDE and *REC* exhibit overall highest *AUC* and lowest *RVP* (Tab. 1). Their estimated mean differences are rather small, since *REC* can be considered as a binary form of the *KDE*. As *REC* uses a threshold ε for defining the hyperball of recurrences, the results can exhibit slightly higher *AUC* than *KDE* (~~not shown~~Fig. S1). However, with *REC* the caveat is that the parameter ε is not necessarily optimally chosen.

KNN. In most of the cases, *KNN-Gamma* performance is better than the *UNIV* control, but only as good as the *UNIV* control for detecting *TrendOnset*. This may be due to the fact that for *TrendOnset*, the mean distance to the *KNN* does not change, unless considering a very large number of *KNN* or excluding a large fraction of temporally near data points to be within the *KNN*. When excluding *TrendOnset* the mean performance increases to 0.019 which is comparable to *KDE* and *REC*. In contrast, *KNN-Delta* does not yield high *AUC*, probably because we do not construct anomalies in the data cube explicitly with a direction that is accounted for by *KNN-Delta* (mean length of the ~~mean~~ vectors to its *KNN*). The finding that simple algorithms like *KNN-Gamma* (or *KDE*, *T2*) are very competitive, if not favourable algorithms, goes in line with results of Harmeling et al. (2006); Killourhy and Maxion (2009); Ding et al. (2014) on various data sets.

On average, *KNFST* and *SVDD* perform worse than or equally well as the Univariate control algorithm (*UNIV*). Also the *RVP* is highest among the algorithms (Tab. 1). It has already been reported, that *SVDD* can exhibit remarkable fluctuations in the results for sample sizes smaller than 1000 data points (Ding et al., 2014). However, we use ~~5000~~ 5,000 points for training. Thus, we suggest that the fluctuations are due to the fact that *SVDD* and *KNFST* use a training set that is chosen at random and may itself contain anomalies. In the current setting the size of the training sample (5,000) is rather small compared to the spatiotemporal size of the data cube (750,000), and it does not seem to be sufficient to train these algorithms on the data cube. Increased sample sizes, however, would heavily increase memory demand and computing time, rendering kernel algorithms computationally inapplicable. Furthermore Ding et al. (2014) shows, that the sample size has a remarkable effect for *SVDD* (better performance for lager sample sizes). However, even with very large sample sizes *SVDD* is still performing worse than *KNN* in in the setting of Ding et al.. Training and testing *SVDD* on each pixel ~~did~~ does also not improve the results ~~(not shown)~~ ~~.~~ as the amount of anomalies differs between different pixels in our setting. Training and testing *SVDD* on each pixel assumes the same amount of anomalies in each pixel (constant outlier ratio assumed by the fixed ν parameter) which is contrary to the generation of the artificial data farm.

We explicitly do not want to state that ~~these 2 algorithms~~ *KNFST* and *SVDD* are generally worse algorithms, i.e. they are just not built for these massive amounts of data. *KNFST* and *SVDD* outperform others in very different setting (novelty detection in images) (Bodesheim et al., 2013).

T2 exhibits good performance for detecting starting trends and shifts in the mean. However, it also exhibits the third largest *RVP* (Tab. 1) indicating that the estimation of the covariance matrix may be sensitive to random variation in the data. Nevertheless, the *RVP* is still far better than for *SVDD*. The robust estimation of the mean and covariance matrix might be a difficult task (Smetek and Bauer, 2007; Rousseeuw and Hubert, 2011) for which rather complex algorithms like the (fast) minimum determinant covariance estimator have been proposed, which are closely related to *T2* (Rousseeuw and Van Driessen, 1990). Furthermore, *T2* assumes a multivariate Gaussian distribution and linear dependencies among the variables. Thus, it is not preferable for the ~~complications~~ data properties *NonLinearDep* and *CorrelatedNoise* unless combined with a nonlinear feature extraction technique like ICA (Fig. B1).

4.3 Ensembles

The selection of algorithms for computing the ensemble is a compromise between accurate detection of and diversity amongst the selected algorithms (Zimek et al., 2013). We select the 4 best algorithms (*4b*, *KDE*, *REC*, *KNN-Gamma*,*T2*) and the 3 best distance-based algorithms (*3d*, *KDE*, *REC*, *KNN-Gamma*) for computing their ensembles. We assume that this choice accounts for accuracy (best algorithms selected) as well as for diversity (different algorithms selected).

Overall, ensemble building improves the anomaly detection rate. The mean *AUC* of each of the ensemble members (*3d*: $+0.030$, *4b*: $+0.023$) is lower than the *AUC* of the ensemble, regardless of whether the maximum or the mean is used for ensemble aggregation. Minimum aggregation of ensemble members, however, performs worse than the individual ensemble members *REC* and *KDE*. Using the maximum or mean yields consistently higher *AUC* than using the minimum (Tab. 2). The superior performance of the maximum choice compared to the minimum indicates that single algorithms overlook more often anomalous events than raising false alarm. Nevertheless, the maximum has the caveat that even a single algorithm may cause a false alarm (Zimek et al., 2013), e.g. due to a poor parameterization or inadequate assumptions about properties of the data. Thus, a more 'balanced voting' procedure like the mean is the preferable choice and more stable with respect to possible error sources. Among the mean ensembles, the *3d* or *4b* ensembles perform equally well (0.041 vs. 0.039 ± 0.001 overall) (Tab. 2).

4.4 Limitations

High Dimensionality. The utility of distance-based outlier detection algorithms as used in this paper is often questioned in the context of high dimensional data (Zimek et al., 2012). The 'Curse of Dimensionality' states that the difference between near and far distances diminishes with increasing dimensionality. However, Zimek et al. (2012) showed in the case of *KNN* that the contrary is true for outliers with fixed magnitude in otherwise uncorrelated data. Dimensionality reduction as crucial feature extraction transforms the data into few (ideally) meaningful and uncorrelated variables. Thus, the findings of Zimek et al. (2012) provide strong arguments for applying dimensionality reduction on correlated data. We anticipate that ~~his~~ their findings are the reason of the superior performance of dimensionality reduction here.

Table 2. *AUC* difference of the ensembles of anomaly detection algorithms to the *UNIV* control. Ensembles are computed out of the 4 best algorithms (*4b*, *KDE*, *REC*, *KNN-Gamma*, *T2*) and the 3 best distance-based algorithms (*3d*, *KDE*, *REC*, *KNN-Gamma*).

	3d-max	*3d-mean*	*3d-min*	*4b-max*	*4b-mean*	*4b-min*
BaseShift	0.042	0.037	0.033	0.042	0.038	0.030
TrendOnset	0.059	0.058	0.033	0.060	0.056	0.020
MSCChange	0.033	0.040	0.032	0.033	0.037	0.017
VarianceChange	0.027	0.027	0.025	0.023	0.026	0.022
Mean	0.040	0.041	0.031	0.039	0.039	0.022
RVP	0.001	0.001	0.001	0.001	0.001	0.001

Heuristic Choices. Within the parameterization process, several heuristic choices are made. We exclude 5 time steps to be counted as recurrences or k-nearest neighbours. We fix several parameters, e.g. the number of nearest neighbours is fixed to 10. Also other parameter choices are rather heuristic (e.g. σ), although commonly used. ~~Within the data farmcreation, we assume that the number of~~ The artificial data farm's intrinsic dimension is 3 ~~, the~~ as it was created from three independent components. Therefore the embedding dimension m is fixed accordingly although it can be inferred based on the data by determining the number of false nearest neighbours (Kennel et al., 1992; Hegger et al., 1999). The signal to noise ratio of our artificial data farm is 0.3. Furthermore, the choice of the ~~complications~~ data properties might influence the results for each event type, as the standard deviation of *AUC* values over all ~~complications~~ data properties (0.05) is rather large, compared to the average *AUC* gain of the 3 best algorithms with respect to the control (+0.03). However, the ordering of the algorithms is also important to derive rankings of algorithms (Hornik and Meyer, 2007). By choosing different subsets of the ~~complications~~data properties, we observe that the 3 best algorithms (*KDE*, *REC*, *KNN-Gamma*) are on top, independently of the chosen ~~complication~~data property. Therefore, the ~~complications~~ data properties might have an influence on the *AUC* values themselves, but not on the choice of the 3 top candidates.

5 Remarks on Applications for Real Earth Observations

Our versions of the artificial data cubes were generated to test different algorithms for their capability to deal with typical properties of Earth observation data. The workflows were chosen to be as generic as possible, and therefore their application to 'real' data with slightly different properties should be made as easy as possible. Nevertheless, several points have to be considered, when applying the algorithms on real EOs.

A typical preprocessing of Earth observations is to center variables to zero-mean and standardize to unit variance (also known as z transformation). A standardization of this kind is of key importance in global EOs. Real multivariate observations often have different physical units or ranges, which have to be made comparable before analysing. However, standardization has to be applied with care. Differences of the mean and variance between geographically distinct or even adjacent grid cells as well as seasonal cycles might corrupt any further analysis. We recommend to subtract the median seasonal cycle before

standardization. The median is prefered over the mean as mean seasonal cycles are affected by changes in the amplitude of the cycle. Standardization can be applied globally (i.e. with global spatiotemporal mean and variance), regionally (i.e. with spatiotemporal mean and variance in subregions of the globe), or locally (i.e. with temporal mean and variance in each grid-cell). Global standardization might be more robust than local, but detects only anomalies in high-variance regions. Local standardization assumes that the number of extreme anomalies is equal in each grid cell, which is a rather strong assumption. Thus, a regional standardization is favourable in regions with similar mean and variance.

Especially variables presenting a signal from the biosphere are known to exhibit heteroscedasticity, e.g. the variance during growing season is substantially larger than during the rest of the year (Fig. 5). Atmospheric variables in high latitudes also show higher variability during the cold season, e.g. temperature variability might be higher over ice (cold season) than over open water (warm season) (Hansen et al., 2012). Specifically for global applications, using estimates of variance or standard deviation locally (in each grid-cell) leads to an underestimation of the variance during growing season and thus to an overestimation of anomalies due to standardization especially in the Northern latitudes (Guanche et al., 2016). Thus, we recommend to account for the heteroscedastic pattern by adjusting the variance during the growing season within similar regions. We also recommend this kind of adjustment for the covariance matrix used, e.g., in *T2* or *PCA* as well as for the parameterization of *KDE* or *REC*.

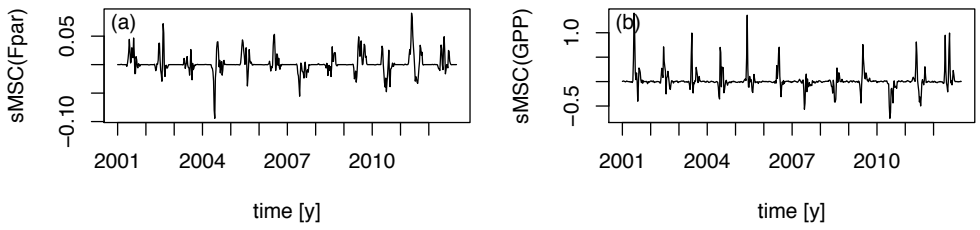

Figure 5. The residual time series obtained by subtracting the median seasonal cycle from (a) the fraction of absorbed photosynthetic radiation (fPAR) and (b) Gross Primary Productivity (GPP) at Northern latitudes exhibit heteroscedastic patterns.

Furthermore, anomalies are also overestimated when using a reference period for the estimation of the variance (Sippel et al., 2015). However, with 300 observations in 8-day intervals, as used in this study, this issue is expected to be less pronounced than for fewer observations as it scales with the length of the time series. Nevertheless, we rather recommend to use estimates of the variance of the entire time series or to correct for the overestimation in the out-of-reference period as shown in Sippel et al. (2015).

Regarding the parameterization process of the algorithms we use fixed parameters for σ, ε, k, ν, mean vector, and covariance matrix globally on the entire artificial data cube. Local parameterization assumes the same amount of anomalies in each region, which is neither suitable for the artificial data by construction nor for global real data. Thus we recommend to ~~parametrise~~ parameterize globally or within similar regions. Classification of the Earth into similar regions and applying multivariate extreme detection in each region will be the subject of future research.

6 Conclusions

Our aim is to identify suitable methods for detecting anomalies in highly multivariate, correlated, and seasonally varying data streams as they are common in Earth system science. In particular, we are interested in detecting shifts in mean (extremes), changes in the amplitude of the seasonal cycle, temporal changes in the variance and onsets of trends. We test a wide range of workflows (i.e. combining feature extraction techniques and anomaly detection algorithms). All experiments are based on artificial data, designed to mimic real world Earth observations.

We can show that, on average over different anomaly types and data ~~'complications'~~ properties, 3 multivariate anomaly detection algorithms (*KDE*, *REC*, *KNN-Gamma*) outperform univariate extreme event detection as well as other multivariate approaches (mean *AUC* compared to univariate control: $+0.030$). Additional slight improvement can be achieved by combining the best algorithms into ensembles using an aggregation by averaging score quantiles ($+0.041$). In contrast, the tested machine-learning algorithms (*SVDD* -0.05, *KNFST* -0.01) may fail due to overfitting to the training sample.

However, we also find for the considered type of events that including a suitable feature extraction technique in the detection workflow is often more important than the choice of the event detection algorithm itself. Yet, we find that the feature extraction has to be explicitly designed for the event type of interest, i.e. time delay embedding (for detecting changes in the cycle amplitude) and exponential weighted moving average (for detecting trends, long extremes and removing uncorrelated noise in the signal) increases the detection rate of the anomalous events. Including features of the variance within a moving window works partly for detecting increases in the variance, but fails to detect a decrease in the variance due to the relatively high observational noise level. In general, if the data comprises seasonality, subtracting it and using the remaining time series as input feature is essential. Furthermore, we improve the detection rate of multivariate anomalies in highly correlated data streams by adding a dimensionality reduction method to the workflow (in line with results of Zimek et al., 2012).

The proposed workflows are capable of dealing with common properties of Earth observations like seasonality, non-linear dependencies as well as (to a certain degree) non-Gaussian distributions and noise. Nevertheless, they have to be applied with care to Earth observations, i.e. standardization issues along with strong heteroscedastic patterns (e.g. in Biosphere variables of Northern latitudes) may lead to an overestimation of anomalies. Future work will explore the potential of the identified workflows on rediscovering known and potentially unknown extremes as well as other anomalies in a set of real Earth system science data streams. We anticipate that an automated application of our workflows might enable the establishment of automated Earth system process control in a very generic manner.

Appendix A: ~~Detailed Results~~

~~versus event magnitude for all combinations (grey) and the Univariate control (red). Columns of the matrix represent different , rows represent complications. Additional colored workflows represent the workflows with the 5 highest mean values for the magnitudes > 2 sd (> 0.6 respectively).~~

~~Effect of the complications on the 3 best detection algorithms (, ,) presented as difference of the control for the (a-d).~~

Appendix A: Technical Details on Generating the Artificial Data

Within the generation process, we assume that the signal S is additive to the baseline B. The baseline might represent reoccuring patterns like seasonality or a constant mean. In addition, binary event parameters $ev_{t,lat,lon}$ are introduced, which allow for switching the anomaly on ($ev_{t,lat,lon} \neq 0$~~and off~~) and off ($ev_{t,lat,lon} = 0$) ('normality'). The event type and magnitude of the event is controlled by a parameter separately for the baseline (k_b), the signal (k_s) and a mean-shift parameter (k_m) scaled with the standard deviation of the data sd.

$$\Theta_{t,lat,lon} = B_{t,lat,lon} \cdot 2^{(k_b \cdot ev_{t,lat,lon})} + S_{t,lat,lon} \cdot 2^{(k_s \cdot ev_{t,lat,lon})} + k_m \cdot ev_{t,lat,lon} \cdot sd \tag{A1}$$

For a basic version, 3 independent components $\Theta_{t,lat,lon,var}$ are created with the signal consisting of Gaussian noise ($sd = 1$). Each component represents intrinsic properties of the Earth system. Furthermore, we assume that properties of the Earth system $\Theta_{t,lat,lon}$ are not measured directly but indirectly via a set of correlated variables, i.e. representing patterns of these intrinsic properties. Hence, these variables propagate anomalous events that occur in one independent component. This set of correlated variables X_{var} is created by weighting the intrinsic properties Θ_{var} with randomly drawn linear (or non-linear) weights w_j plus additional measurement noise ϵ (Gaussian, $sd = 0.3$) added to each variable.

$$X_{var} = \sum_{j=1}^{j=3} w_j \cdot \Theta_j + \epsilon \tag{A2}$$

Using this data generation scheme, a standard data cube $X_{t_{i,j},lat,lon,var}$ is created, encompassing 300 time steps (T), 10 temporally correlated variables (VAR) and the total number of latitudes (LAT) and longitudes (LON) fixed to 50 each. We induce anomalous events with a spatial extent of 40 % of the latitude and longitude and 10 events, each with a temporal extent of 5 time steps. Our total amount of anomalies equals about 3 % of the total data cube.

In the basic version we create 4 data cubes each with a different temporary event type:

- Shift in the baseline, i.e. shift of the running mean of a time series (*BaseShift*) (Fig. 2 a)
- Change in the variance of the time series (*VarianceChange*) (Fig. 2 b)
- Change in the amplitude of the mean seasonal cycle of a time series (*MSCChange*)(Fig. 2 c)
- Onset of a trend in the time series (*TrendOnset*) (Fig. 2 d)

Regarding the ~~complications~~data properties, some of the event type ~~complication~~ data property combinations are excluded (Tab. A1). In detail, we do not expect a *TrendOnset* to 'infect' neighboured cells (*TrendOnset* plus *RandomWalkExtreme*) and a *TrendOnset* can hardly be called a *TrendOnset* if it encompasses only one time step (*ShortExtremes*).

The artificial data farm can be created after cloning into https://github.com/CAB-LAB/DataFarm. Generation is done with the following command within the programming language julia, version 0.4:

using SurrogateCube; SurrogateCube.DataFarm.makeDataFarm(300,50,50,PathToFolder)

Table A1. Parameter settings for the generation of the artificial data farm. Details are given for each event type and ~~complication~~ data property (in brackets).

		Basic	(~~Complication~~data property)	*BaseShift*	*VarianceChange*	*MSCChange*	*TrendOnset*
Independent comp.	Θ	3	(MoreIndep Components)	3 (6)	3 (6)	3 (6)	3 (6)
Dependency (Θ)		linear (w)	(*NonLinearDep* squ)	w (squ)	w (squ)	w (sq)	w (sq)
Baseline	B	const. = c	(*SeasonalCycle* s, *LatitudinalGradient* lg))	c (s, lg)	c (s, lg)	s (lg)	c (s, lg)
Signal	S	gaussian g	(*LaplacianNoise* l, *CorrelatedNoise* r)	g (l, r)	g (l, r)	g (l, r)	g (l, r)
Variables	VAR	10		10	10	10	10
Noise	ϵ	0.3	(*NoiseIncrease*)	0.3 (1)	0.3 (1)	0.3 (1)	0.3 (1)
Events							
Event number		10	(*ShortExtremes*, *LongExtremes*)	10 (50, 5)	10 (50, 5)	1	1
Spatial extent		1000		1000	1000	4	1000
Temporal extent		5	(*ShortExtremes*, *LongExtremes*)	5 (1,10)	5 (1,10)	92 (46, 184)	150
Magnitudes				k_m = 0.2-4	k_s = -2:2	k_b = -2:2	k_m = 0.2-4
Shape		rect.	(RandomWalk Extreme rw)	rect (rw)	rect (rw)	rect (rw)	rect

Appendix B: Detailed Results

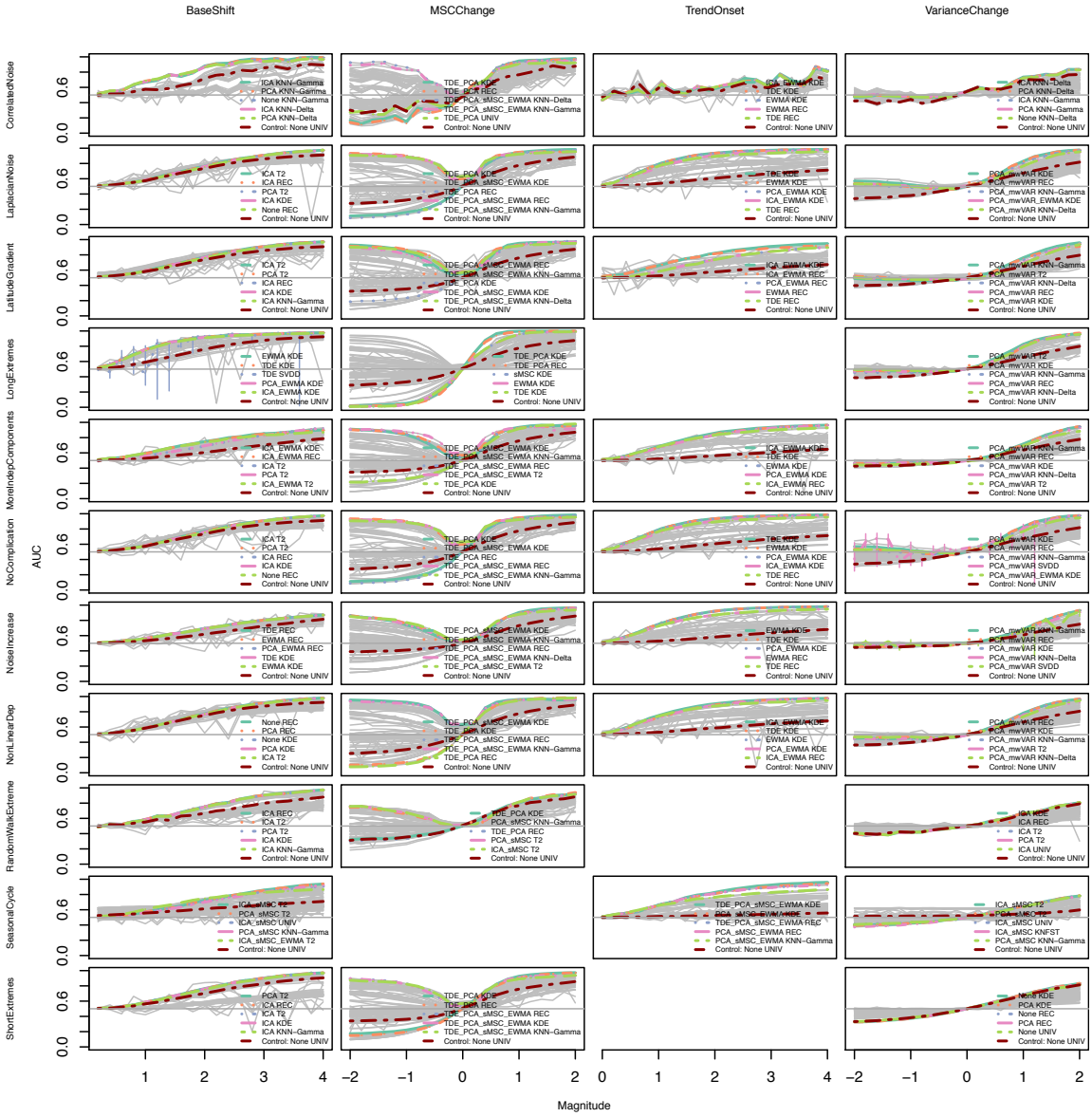

Figure B1. *AUC* versus event magnitude for all combinations (grey) and the Univariate control (red). Columns of the matrix represent different event types, rows represent data properties. Additional colored workflows represent the workflows with the 5 highest mean values for the magnitudes > 2 sd (> 0.6 respectively).

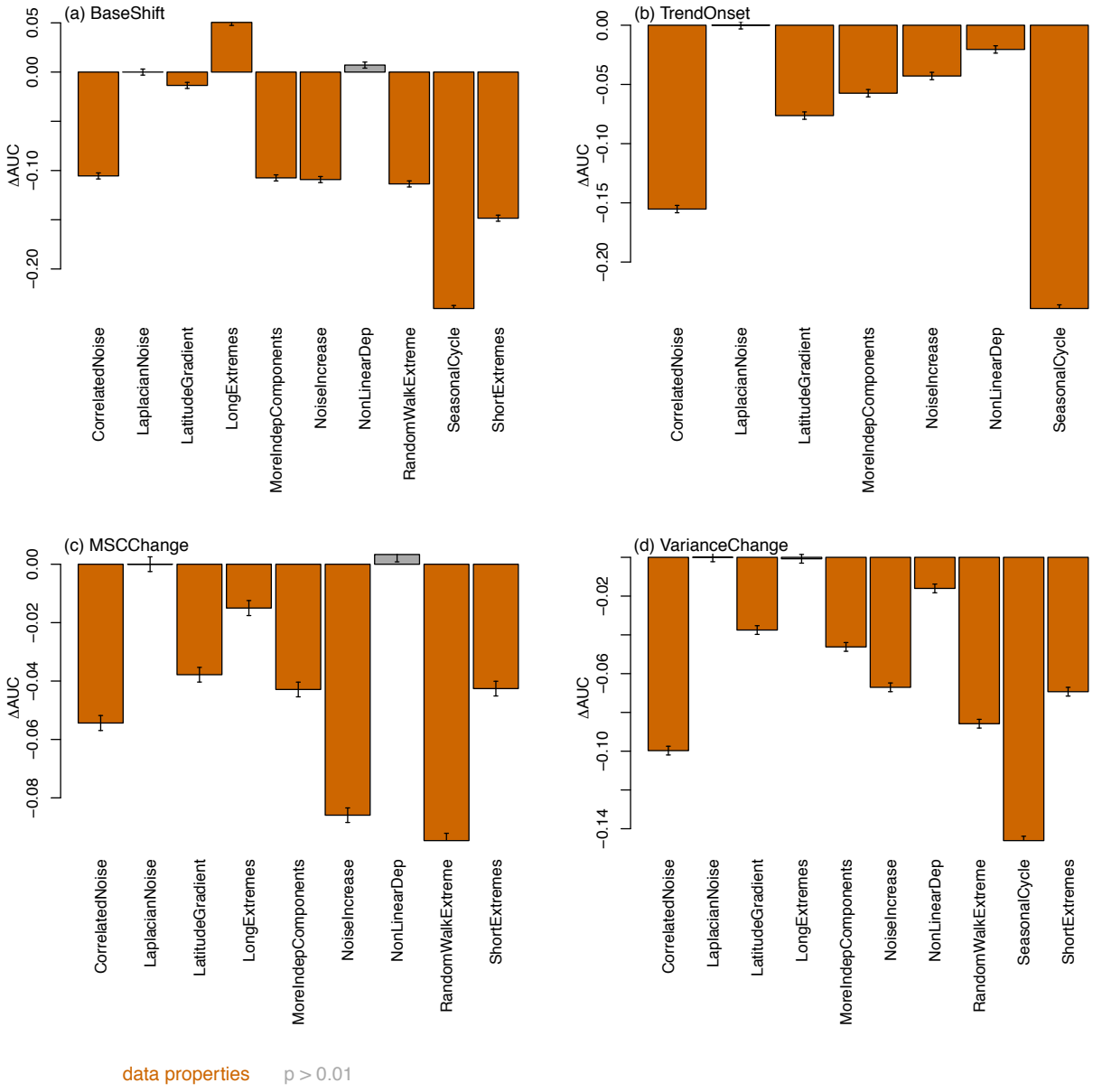

Figure B2. Effect of the data properties on the 3 best detection algorithms (*KDE*, *REC*, *KNN-Gamma*) presented as *AUC* difference of the *UNIV* control for the event types (a-d).

Author contributions. M.F. and M.D.M designed the study in collaboration with F.G., A.B., J.D., M.R. and E.R.; M.F. implemented the algorithms including contributions from F.G., P.B. and E.R.; M.F. wrote the manuscript with contributions from all co-authors.

Acknowledgements. This research has received funding by the International Max Planck Research School for Global Biogeochemical Cycles (IMPRS), the European Space Agency via the STSE project CAB-LAB and the BACI project, a European Union's Horizon 2020 research and innovation programme under grant agreement No 64176. We thank Simone Girst for her kind language check. Two reviewers provided valuable suggestions for improvement.

References

Aggarwal, C. C.: Outlier Ensembles, ACM SIGKDD Explorations Newsletter, 14, 49–58, 2012.

Alexander, L. V., Zhang, X., Peterson, T. C., Caesar, J., Gleason, B., Klein Tank, A. M. G., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., Tagipour, A., Rupa Kumar, K., Revadekar, J., Griffiths, G., Vincent, L., Stephenson, D. B., Burn, J., Aguilar, E., Brunet, M., Taylor, M., New, M., Zhai, P., Rusticucci, M., and Vazquez-Aguirre, J. L.: Global observed changes in daily climate extremes of temperature and precipitation, J. Geophys. Res, 111, D05 109, 2006.

Bae, K.-H., Karolyi, G. A., and Stulz, R. M.: A New Approach to Measuring Financial Contagion, Review of Financial Studies, 16, 717–763, 2003.

Baldocchi, D., Falge, E., and Wilson, K.: A spectral analysis of biosphere–atmosphere trace gas flux densities and meteorological variables across hour to multi-year time scales, Agricultural and Forest Meteorology, 107, 1–27, 2001.

Bathiany, S., Notz, D., Mauritsen, T., Raedel, G., and Brovkin, V.: On the Potential for Abrupt Arctic Winter Sea Ice Loss , Journal of Climate, 29, 2703–2719, 2016.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Altaf Arain, M., Baldocchi, D., Bonan, G. B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K. W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F. I., and Papale, D.: Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate , Science, 329, 843–838, 2010.

Bintanja, R. and van der Linden, E. C.: The changing seasonal climate in the Arctic, Sci Rep, 3, 2013.

Bodesheim, P., Freytag, A., Rodner, E., Kemmler, M., and Denzler, J.: Kernel Null Space Methods for Novelty Detection, CVPR, pp. 3374–3381, 2013.

Chang, C.-C. and Lin, C.-J.: LIBSVM: A Library for Support Vector Machines, ACM Transactions on Intelligent Systems and Technology, 2, 27:1–27:27, 2013.

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogée, J., Allard, V., Aubinet, M., Buchmann, N., Bernhofer, C., Carrara, A., Chevallier, F., De Noblet, N., Friend, A. D., Friedlingstein, P., Grünwald, T., Heinesch, B., Keronen, P., Knohl, A., Krinner, G., Loustau, D., Manca, G., Matteucci, G., Miglietta, F., Ourcival, J. M., Papale, D., Pilegaard, K., Rambal, S., Seufert, G., Soussana, J. F., Sanz, M. J., Schulze, E. D., Vesala, T., and Valentini, R.: Europe-wide reduction in primary productivity caused by the heat and drought in 2003, Nature, 437, 529–533, 2005.

Ciais, P., Dolman, A. J., Bombelli, A., Duren, R., Peregon, A., Rayner, P. J., Miller, C., Gobron, N., Kinderman, G., Marland, G., Gruber, N., Chevallier, F., Andres, R. J., Balsamo, G., Bopp, L., Bréon, F. M., Broquet, G., Dargaville, R., Battin, T. J., Borges, A., Bovensmann, H., Buchwitz, M., Butler, J., Canadell, J. G., Cook, R. B., DeFries, R., Engelen, R., Gurney, K. R., Heinze, C., Heimann, M., Held, A., Henry, M., Law, B., Luyssaert, S., Miller, J., Moriyama, T., Moulin, C., Myneni, R. B., Nussli, C., Obersteiner, M., Ojima, D., Pan, Y., Paris, J. D., Piao, S. L., Poulter, B., Plummer, S., Quegan, S., Raymond, P., Reichstein, M., Rivier, L., Sabine, C., Schimel, D., Tarasova, O., Valentini, R., Wang, R., van der Werf, G., Wickland, D., Williams, M., and Zehner, C.: Current systematic carbon-cycle observations and the need for implementing a policy-relevant carbon observing system, Biogeosciences, 11, 3547–3602, 2014.

Ding, X., Li, Y., Belatreche, A., and Maguire, L. P.: An experimental evaluation of novelty detection methods, Neurocomputing, 135, 313–327, 2014.

Donat, M. G., Alexander, L. V., Yang, H., Durre, I., Vose, R., Dunn, R. J. H., Willett, K. M., Aguilar, E., Brunet, M., Caesar, J., Hewitson, B., Jack, C., Klein Tank, A. M. G., Kruger, A. C., Marengo, J., Peterson, T. C., Renom, M., Oria Rojas, C., Rusticucci, M., Salinger, J.,

Elrayah, A. S., Sekele, S. S., Srivastava, A. K., Trewin, B., Villarroel, C., Vincent, L. A., Zhai, P., Zhang, X., and Kitching, S.: Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset, Journal of Geophysical Research: Atmospheres, 118, 2098–2118, 2013.

Donges, J. F., Donner, R. V., Rehfeld, K., Marwan, N., Trauth, M. H., and Kurths, J.: Identification of dynamical transitions in marine palaeoclimate records by recurrence network analysis, Nonlinear Processes in Geophysics, 18, 545–562, 2011a.

Donges, J. F., Donner, R. V., Trauth, M. H., Marwan, N., Schellnhuber, H.-J., and Kurths, J.: Nonlinear detection of paleoclimate-variability transitions possibly related to human evolution, PNAS, 108, 20 422–20 427, 2011b.

Donges, J. F., Heitzig, J., Donner, R. V., and Kurths, J.: Analytical framework for recurrence network analysis of time series, Phys. Rev. E, 85, 046 105, 2012.

Donges, J. F., Schleussner, C. F., Siegmund, J. F., and Donner, R. V.: Event coincidence analysis for quantifying statistical interrelationships between event time series, The European Physical Journal Special Topics, 225, 471–487, 2016.

Donner, R. V., Zou, Y., Donges, J. F., Marwan, N., and Kurths, J.: Recurrence networks – A novel paradigm for nonlinear time series analysis, New J. Phys., 12, 033 025, 2010.

Dorigo, W. A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A., and Jackson, T.: The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements, Hydrology and Earth System Sciences, 15, 1675–1698, 2011.

Drijfhout, S., Bathiany, S., Beaulieu, C., Brovkin, V., Claussen, M., Huntingford, C., Scheffer, M., Sgubin, G., and Swingedouw, D.: Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models, Proceedings of the National Academy of Sciences, 112, E5777–E5786, 2015.

Durante, F. and Salvadori, G.: On the construction of multivariate extreme value models via copulas, Environmetrics, 21, 143–161, 2010.

Easterling, D. R., Meehl, G. A., Parmesan, C., Changnon, T. R. K., and Mearns, L. O.: Climate Extremes: Observations, Modeling, and Impacts, Science, 289, 2068–2074, 2000.

Faranda, D. and Vaienti, S.: A new recurrences based technique for detecting robust extrema in long temperature records, Geophysical Research Letters, 40, 5782–5786, 2013.

Fawcett, T.: An introduction to ROC analysis, Pattern Recognition Letters, 27, 861–874, 2006.

Fischer, E. M.: Robust projections of combined humidity and temperature extremes, Nature Climate Change, 3, 126–130, 2013.

Flach, M., Lange, H., Foken, T., and Hauhs, M.: Recurrence Analysis of Eddy Covariance Fluxes, in: Recurrence Plots and Their Quantifications: Expanding Horizons, edited by Webber, Jr., C. L., Ioana, C., and Marwan, N., pp. 301–319, Springer Proceedings in Physics, Cham, 2016.

Forkel, M., Carvalhais, N., Verbesselt, J., Mahecha, M., Neigh, C., and Reichstein, M.: Trend Change Detection in NDVI Time Series: Effects of Inter-Annual Variability and Methodology, Remote Sensing, 5, 2113–2144, 2013.

Ge, Z., Song, Z., and Gao, F.: Review of Recent Research on Data-Based Process Monitoring, Industrial & Engineering Chemistry Research, 52, 3543–3562, 2013.

Ghil, M., Yiou, P., Hallegatte, S., Malamud, B. D., Naveau, P., Soloviev, A., Friederichs, P., Keilis-Borok, V., Kondrashov, D., Kossobokov, V., Mestre, O., Nicolis, C., Rust, H. W., Shebalin, P., Vrac, M., Witt, A., and Zaliapin, I.: Extreme events: dynamics, statistics and prediction, Nonlinear Processes in Geophysics, 18, 295–350, 2011.

Guanche, Y., Rodner, E., Flach, M., Sippel, S., Mahecha, M. D., and Denzler, J.: Detecting Multivariate Biosphere Extremes, 6th International Workshop on Climate Informatics, in Review, 1–4, 2016.

Hansen, J., Sato, M., and Ruedy, R.: Perception of climate change, PNAS, pp. E2415–E2423, 2012.

Harmeling, S., Dornhege, G., Tax, D., Meinecke, F., and Müller, K.-R.: From outliers to prototypes: Ordering data, Neurocomputing, 69, 1608–1618, 2006.

Hegger, R., Kantz, H., and Schreiber, T.: Practical implementation of nonlinear time series methods: The TISEAN package, Chaos: An Interdisciplinary Journal of Nonlinear Science, 9, 413, 1999.

Hornik, K. and Meyer, D.: Deriving Consensus Rankings from Benchmarking Experiments, in: Advances in Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization, edited by Decker, R. and Lenz, H.-J., pp. 163–170, Springer, Berlin Heidelberg, 2007.

Hotelling, H.: Multivariate Quality Control - Illustrated by the Air Testing of Sample Bombsights, in: Techniques of Statistical Analysis, edited by Eisenhart, C., Hastay, M. W., and Wallis, W. A., pp. 111–184, McGraw-Hill, New York, 1947.

Huntingford, C., Jones, P. D., Livina, V. N., Lenton, T. M., and Cox, P. M.: No increase in global temperature variability despite changing regional patterns, Nature, 500, 327–330, 2013.

Hyväringen, A. and Oja, E.: Independent component analysis: algorithms and applications, Neural Networks, 13, 411–430, 2000.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, J. Geophys. Res, 116, G00J07, 2011.

Kennel, M. B., Brown, R., and Abarbanel, H. D. I.: Determining embedding dimension for phase- space reconstruction using a geometrical construction, Physical Review A, 45, 3403–3411, 1992.

Kharin, V. V., Zwiers, F. W., Zhang, X., and Wehner, M.: Changes in temperature and precipitation extremes in the CMIP5 ensemble, Climatic Change, 119, 345–357, 2013.

Killourhy, K. S. and Maxion, R. A.: Comparing Anomaly-Detection Algorithms for Keystroke Dynamics, IEEE/IFIP International Conference on Dependable Systems & Networks, pp. 125–134, 2009.

Koçak, K., Şaylan, L., and Eitzinger, J.: Nonlinear prediction of near-surface temperature via univariate and multivariate time series embedding, Ecological Modelling, 173, 1–7, 2004.

Ledford, A. W. and Tawn, J. A.: Statistics for near independence in multivariate extreme values, Biometrika, 83, 169–187, 1996.

Lee, J.-M., Yoo, C., and Lee, I.-B.: Statistical monitoring of dynamic processes based on dynamic independent component analysis, Chemical Engineering Science, 59, 2995–3006, 2004.

Lehmann, J., Coumou, D., and Frieler, K.: Increased record-breaking precipitation events under global warming, Climatic Change, 132, 501–515, 2015.

Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., Risbey, J., Schuster, S., Jakob, D., and Stafford-Smith, M.: A compound event framework for understanding extreme impacts, Wiley Interdisciplinary Reviews: Climate Change, 5, 113–128, 2013.

Lim, S. A. H., Antony, J., and Albliwi, S.: StatisticalProcess Control (SPC) in the food industry - A systematic review and future research agenda, Trends in Food Science & Technology, 37, 137–151, 2014.

Lowry, C. A. and Montgomery, D. C.: A review of multivariate control charts , IIE Transactions, 27, 800–810, 1995.

Lowry, C. A. and Woodall, W. H.: A Multivariate Exponentially Weighted Moving Average Control Chart, Technometrics, 34, 46–53, 1992.

Majeed, W. and Avison, M. J.: Robust Data Driven Model Order Estimation for Independent Component Analysis of fMRI Data with Low Contrast to Noise, PLoS ONE, 9, e94 943, 2014.

Marwan, N., Carmen Romano, M., Thiel, M., and Kurths, J.: Recurrence plots for the analysis of complex systems, Physics Reports, 438, 237–329, 2007.

Meehl, G. A. and Tebaldi, C.: More Intense, More Frequent, and Longer Lasting Heat Waves in the 21st Century, Science, 305, 994–997, 2004.

Mikosch, T.: Copulas: Tales and facts, Extremes, 9, 3–20, 2006.

Nagendra, H., Lucas, R., Honrado, J. P., Jongman, R. H. G., Tarantino, C., Adamo, M., and Mairota, P.: Remote sensing for conservation monitoring: Assessing protected areas, habitat extent, habitat condition, species diversity, and threats, Ecological Indicators, pp. 1–15, 2012.

Overpeck, J. T., Meehl, G. A., Bony, S., and Easterling, D. R.: Climate Data Challenges in the 21st Century, Science, 331, 700–703, 2011.

Parzen, E.: On Estimation of a Probability Density Function and Mode, The Annals of Mathematical Statistics, 33, 1–1065–1076, 1962.

Pfeifer, M., Disney, M., Quaife, T., and Marchant, R.: Terrestrial ecosystems from space: a review of Earth observation products for macroecology applications, Global Ecology and Biogeography, 21, 603–624, 2011.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team: nlme: Linear and Nonlinear Mixed Effects Models, http://CRAN.R-project.org/package=nlme, R package version 3.1-128, 2016.

Poincaré, H.: Sur le probleme des trois corps et les équations de la dynamique, Acta Mathematica, 13, 5–7, 1890.

Rahmstorf, S. and Coumou, D.: Increase of extreme events in a warming world , PNAS, 108, 17 905–17 909, 2011.

Ramaswamy, S., Rastogi, R., and Shim, K.: Efficient Algorithms for Mining Outliers from Large Data Sets, SIGMOD record, 29, 427–438, 2000.

Rammig, A., Wiedermann, M., Donges, J. F., Babst, F., von Bloh, W., Frank, D., Thonicke, K., and Mahecha, M. D.: Coincidences of climate extremes and anomalous vegetation responses: comparing tree ring patterns to simulated productivity, Biogeosciences, 12, 373–385, 2015.

Reichstein, M., Bahn, M., Ciais, P., Frank, D., Mahecha, M. D., Seneviratne, S. I., Zscheischler, J., Beer, C., Buchmann, N., Frank, D. C., Papale, D., Rammig, A., Smith, P., Thonicke, K., van der Velde, M., Vicca, S., Walz, A., and Wattenbach, M.: Climate extremes and the carbon cycle , Nature, 500, 287–295, 2013.

Rousseeuw, P. J. and Hubert, M.: Robust statistics for outlier detection, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1, 73–79, 2011.

Rousseeuw, P. J. and Van Driessen, K.: A Fast Algorithm for the Minimum Covariance Determinant Estimator, Technometrics, 41, 212–223, 1990.

Santos-Fernández, E.: Multivariate Statistical Quality Control Using R, vol. 14 of *SpringerBriefs in Statistics*, Springer, New York Heidelberg Dordrecht London, 1 edn., 2012.

Schoelzel, C. and Friedrichs, P.: Multivariate non-normally distributed random variables in climate research – introduction to the copula approach, Nonlinear Processes in Geophysics, 15, 761–772, 2008.

Schölkopf, B. and Smola, A.: Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, USA, 2001.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C.: Estimating the Support of a High-Dimensional Distribution, Neural Computation, 13, 1443–1471, 2001.

Schölkopf, B., Muandet, K., Fukumizu, K., Harmeling, S., and Peters, J.: Computing functions of random variables via reproducing kernel Hilbert space representations, Statistics and Computing, 25, 755–766, 2015.

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., and Zhang, X.: Changes in climate extremes and their impacts on the natural physical environment, in: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (IPCC SREX Report), edited by Field, C., Barros, V., Stocker, T., Qin, D., Dokken, D., Ebi, K., Mastrandrea, M., Mach, K., Plattner, G.-K., Allen, S., Tignor, M., and Midgley, pp. 109–230, Cambridge University Press, 2012.

Siegmund, J. F., Sanders, T. G. M., Heinrich, I., van der Maaten, E., Simard, S., Helle, G., and Donner, R. V.: Meteorological Drivers of Extremes in Daily Stem Radius Variations of Beech, Oak, and Pine in Northeastern Germany: An Event Coincidence Analysis, Frontiers in Plant Science, 7, 220, 2016.

Sippel, S., Zscheischler, J., Heimann, M., Otto, F. E. L., Peters, J., and Mahecha, M. D.: Quantifying changes in climate variability and extremes: Pitfalls and their overcoming, Geophys. Res. Lett, 42, 9990–9998, 2015.

Smetek, T. E. and Bauer, K. W.: Finding Hyperspectral Anomalies Using Multivariate Outlier Detection, Proc. 2007 IEEE Aerosp. Conf., pp. 1–24, 2007.

Smets, K., Verdonk, B., and Jordaan, E. M.: Discovering Novelty in Spatio/Temporal Data Using One-Class Support Vector Machines, Proceeding of International Joint Conference on Neural Networks, pp. 2956–2963, 2009.

Takens, F.: Detecting Strange Attractors in Turbulence, Warwick 1980, in: Dynamical Systems and Turbulence, pp. 366–381, Springer, 1981.

Tax, D. M. and Duin, R. P. W.: Support Vector Data Description, Machine Learning, 54, 45–66, 2004.

Thompson, P. D.: How to Improve Accuracy by Combining Independent Forecasts, Monthly Weather Review, 105, 228–229, 1977.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, Biogeosciences, 13, 4291–4313, 2016.

van der Maaten, L. J. P.: Feature extraction from visual data, TiCC Dissertation Series, 2009.

Vicente-Serrano, S. M., Beguería, S., and López-Moreno, J. I.: A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index, Journal of Climate, 23, 1696–1718, 2010.

Von Storch, H. and Zwiers, F. W.: Statistical Analysis in Climate Research, Cambridge Univ. Press, Cambridge, U. K, 2001.

Webber, Jr., C. L. and Marwan, N.: Mathematical and Computational Foundations of Recurrence Quantifications , in: Recurrence Quantification Analysis, pp. 3–43, Springer, Cham Heidelberg New York Dordrecht London, 2015.

Zhou, B., Gu, L., Ding, Y., Shao, L., Wu, Z., Yang, X., Li, C., Li, Z., Wang, X., Cao, Y., Zeng, B., Yu, M., Wang, M., Wang, S., Sun, H., Duan, A., An, Y., Wang, X., and Kong, W.: The Great 2008 Chinese Ice Storm: Its Socioeconomic–Ecological Impact and Sustainability Lessons Learned, Bulletin of the American Meteorological Society, 92, 47–60, 2011.

Zimek, A., Schubert, E., and Kriegel, H.-P.: A survey on unsupervised outlier detection in high-dimensional numerical data, Statistical Analysis and Data Mining, 5, 363–387, 2012.

Zimek, A., Campello, R. J. G. B., and Sander, J.: Ensembles for Unsupervised Outlier Detection: Challenges and Research Questions, SIGKDD Explorations, 15, 11–22, 2013.

Zscheischler, J., Mahecha, M. D., von Buttlar, J., Harmeling, S., Jung, M., Rammig, A., Randerson, J. T., Schölkopf, B., Seneviratne, S. I., Tomelleri, E., Zaehle, S., and Reichstein, M.: A few extreme events dominate global interannual variability in gross primary production, Environ. Res. Lett., 9, 035 001, 2014a.

Zscheischler, J., Reichstein, M., Harmeling, S., Rammig, A., Tomelleri, E., and Mahecha, M. D.: Extreme events in gross primary production: a characterization across continents, Biogeosciences, 11, 2909–2924, 2014b.

Zscheischler, J., Orth, R., and Seneviratne, S. I.: A submonthly database for detecting changes in vegetation-atmosphere coupling, Geophys. Res. Lett, 42, 9816–9824, 2015.

Response on the

***Interactive comment on* "Multivariate Anomaly Detection for Earth Observations: A Comparison of Algorithms and Feature Extraction Techniques" *by* Milan Flach et al.**

Anonymous Referee #1

Received and published: 9 December 2016

*** General comments**

The manuscript describes a systematic and comprehensive study of methods for extraction of anomalies and features from artificially-generated multivariate datasets. The presentation is clear, the manuscript is well written, and the study is sound as a comparison of methods for multivariate data analysis, though its value for earth observations is in my opinion not convincing.

Response: We would like to thank the reviewer for the positive feedback. Regarding the reviewer's concern about the values for EOs, we consider our study as relevant in this context, because current anomaly or extreme detection in Earth observations is mostly done with peak-over-threshold techniques (p. 2, l. 22). These do not consider the multivariate and potentially non-linear correlation structure between multiple variables which we have in EOs. It is therefore an important motivation for our paper to provide a sound basis for alternative and more general approaches. This paper systematically analyses and proposes several algorithms and workflows which consider the structure among multiple variables and furthermore might also reveal novelties about so called compound extremes beyond known patterns, i.e. anomalies where none of the single variables is extreme itself, but their combination is anomalous and leading to an extreme impact. The consideration of compound events does play an increasing role within the community, but is typically confined to known compound events (e.g. heat and droughts) and not very generic. The comparison is performed on artificial data, which were explicitly built to mimic current EOs as ground truth is missing for 'real' EOs (p.3, l.25). Furthermore, available time periods as well as sample sizes are rather small for detecting anomalies in 'real' EOs, which empowers the use of an ensemble of artificial data for method comparison. The application of the proposed workflows to EOs will follow soon.

Although I understand the rationale for using artificial data, particularly when comparing the performance of different methodological approaches, the artificial events that are considered in the study seem to be unrealistically exaggerated, particularly the amplitude change in the seasonal cycle (Fig. 2 c) and and the change in variance (Fig.
C1 2 d). For example climate related changes in the seasonal cycle or in variance are far more subtle (in terms of magnitude) and much more difficult to identify in real data that the ones exemplified in Fig. 2.

Response: Please consider that Figure 2 is only an illustration. As described in the manuscript (p. 6, l. 9), we analyse each type of anomaly across 20 different magnitudes - from very minor perturbations to entirely exaggerated values. The generic formula B1 shows that we are effectively exploring the full range of

perturbations between very subtle changes (Appendix B: k = 0.2) to exceptionally high changes (k = 4.0) as if it was a model parameter sensitivity analysis. In the revised manuscript, we will add an additional sentence for clarification, explain it in the figure caption and show more realistic magnitudes in Figure 2.

I'm uncomfortable with the term "Complication" used throughout the manuscript to refer to specific characteristics of the artificial data. For example a seasonal cycle can hardly be seen as a complication, it's a feature of the data, not necessarily something complex as it is implicit from denoting it a "complication".

Response: We agree with the Reviewer, that the term 'complication' is far away from being optimal. However, 'data features' might be misunderstood with 'feature extraction techniques', which we wanted to prevent. Therefore, we suggest to rephrase 'complications' into 'data properties' in the revised version.

I think that the comprehensiveness of the study is a strength and paradoxically maybe the greatest weakness of the work, because the results need to be necessarily presented in a highly summarized way, here as difference in AUC values (which itself are already a reduction of a ROC curve to a single number) to a univariate approach without "complications" (UNIV). I don't doubt the technical correctness of the results, but in my opinion it's difficult to assess their relevance, particularly in the context of real earth observation data. I find the conclusions of the study quite obvious and realistic (the importance of deseasoning or dimensionality reduction), whether they would require such a wide statistical study on a artificial data farm is not obvious to me.

Response: We thank the reviewer for this comment. It has two components: (1) the presentation of the results, and (2) the overall relevance.
(1) Indeed, we decided to highly summarize the results in the main part of the study. However, more detailed results, for instance the effect of different magnitudes on specific event types and complications can still be inferred from the Appendix Fig. A1.
(2) The importance of dimensionality reduction as one way to enhance the performance of anomaly detection algorithms (p.18, l.12) has not been shown before to the best of our knowledge, in particular not for EOs. We are convinced that a highly multivariate system like the Earth with seasonality and potentially non-linear dependencies among the variables requires specific workflows like the ones we propose, i.e. the results of our study are relevant in this context. Currently we are working on applying the algorithms on 'real' EOs with very promising results, i.e., results that capture the major known events globally. Our overarching objective is developing workflows to open a path to a series of scientific studies exploring extreme compound events in depth.

*** Specific comments**

If I understood correctly the length of the generated time series is only of 300 time steps (Appendix B), which may be in itself a major factor influencing the performance of some of the methods.

Response: Indeed, the length of the time series is a factor influencing the performance of the multivariate anomaly detection algorithms. One crucial point of our study is, that Earth observations are typically short. We seek to understand the

performance characteristics of various algorithms and feature extraction methods on short time series. Furthermore, please note that Ding et al. (2014) studied this effect in detail, changing the data set size between 50-30000. The only algorithm on which the size of the data set had a remarkable effect was the Support Vector Data Description (SVDD). SVDD performance increased with the size of the data set. However, even the best performance of SVDD was worse than the other algorithms. Therefore, we conclude that the size of the data set is not influencing the results of the top algorithms (KDE, KNN, REC, T2). We include this aspect in a second version of the paper

Although I'm keen on the transference of methodological approaches across different areas, and in this case the use of statistical process control (SPC) methods typically used in other contexts (e.g. industry), the restriction of feature extraction methods to the ones used in classical multivariate SPC seems to me an unnecessary restriction. Many feature extraction methods, e.g. wavelets, are routinely used with earth observations precisely because they perform very well in that kind of data.

Response: We are aware that the list of feature extraction algorithms as well as the list of anomaly detection algorithms can hardly be complete. We did not restrict the feature extraction methods only to the ones used in classical statistical process control. We also included non-standard ones from process monitoring in industry (e.g., Independent Component Analysis) and of course from univariate extreme extreme event detection (e.g., subtracting the mean seasonal cycle) (p.6, l. 28). We agree with the reviewer that wavelets perform very well on EOs, e.g., for extracting information about dominant frequencies in the data. However, event detection is another task. We are not aware that wavelets improve the detection rate of multivariate anomalous events, but we will consider this as an interesting aspect for future research.

*** Technical corrections**

Page 9, line 31: cdot notation

Response: We changed it.

References:
Ding, X., Li, Y., Belatreche, A., & Maguire, L. P. (2014). An experimental evaluation of novelty detection methods. Neurocomputing, 135(C), 313–327. http://doi.org/10.1016/j.neucom.2013.12.002

Response on the

Interactive comment on "Multivariate Anomaly Detection for Earth Observations: A Comparison of Algorithms and Feature Extraction Techniques" by Milan Flach et al.

Reviewer: R. V. Donner (Referee #1)

Reviewer:
Flach et al. present a detailed inter-comparison between a selection of recently applied methodological approaches for detecting multivariate anomalies in Earth observation (EO) data, including a performance assessment based on artificially generated time series data that capture some of the essential features (and complications) of real-world observation. The topic is timely and important, since with the fastly growing amount of big data from remote sensing, the automated identification of key features and particularly unexpected behaviors becomes a crucial task. In this regard, I warmly welcome this study and believe that it can be an important milestone in its field, even though it necessarily presents just a case study and can thus not be complete by definition.

Prior to accepting this very interesting work for final publication in Earth System Dynamics, I would like to ask the authors to address a couple of questions I came up with when working through their material.

Response:
We would like to thank the Reviewer for his positive feedback and the numerous suggestions for improvement of the paper.

Reviewer:
1. It would be good if the authors could clarify already in the abstract which kind of anomalies they aim to address. From Figure 2, it is evident that the four considered types of "events" (or better, episodic behaviors) – base shift, trend onset, change in mean seasonal cycle amplitude, change in variance – affect predominantly the basic statistical features of the data, while their dynamical characteristics (respectively, those of the residuals after removing the seasonal variability component) are largely unaffected. Since this is in contrast to some recent works (including papers by the reviewer's group) which have particularly focused on "dynamical anomalies", it might be worth clarifying this from the beginning. In this context, it is interesting that the authors also consider recurrence characteristics, which are commonly used for detecting changes in the dynamical patterns. However, what they consider here is just a variant of the recurrence rate, which is essentially a statistical characteristic again, as opposed to more sophisticated complexity measures that can be defined within this framework as well.

Response:
We agree with the reviewer that our artificial detection experiment does mostly not affect dynamical characteristics of time series which might be revealed by numerous more complex measures derived from recurrence quantification techniques or recurrence networks. However, the experiment was also not meant to do so. We focus on basic time series characteristics, which are often perceived as "extremes" in the public. As proposed by the reviewer we extend the sentence in the abstract for clarification as follows (p.1, l.9): We rely on artificial data that mimic typical properties and anomalies in multivariate spatiotemporal Earth observations *like sudden changes of basic characteristics of time series such as the sample mean, the variance, changes in the cycle amplitude and trends.*

Reviewer:
2. The motivation for choosing the specific settings in the artificial data could be further clarified, especially regarding explicit statements on typical features of real-world EO data. In this context, I was wondering why the authors study only short-term correlated noises, whereas much of the stochastic background signals in common geophysical variables exhibits long-term memory, which might strongly complicate the anomaly detection. Do the authors consider their white/red noises mostly reflecting additive measurement uncertainties or "true" dynamical components, e.g., due to variables and/or scales not resolved by the measurement process.

Response:
We thank the reviewer for this comment. According to the reviewer's suggestion, we will motivate the settings of the artificial data with typical real-world EO data features at p.5, ll. 12-18:

1. *Shift in the baseline, i.e. shift of the running mean of a time series (BaseShift) (Fig. 2 (a)). This event type is closely related to "extremes" in real-world Earth observations.*
2. *Onset of a trend in the time series (TrendOnset) (Fig. 2 (b)).*
3. *Change in the amplitude of the mean seasonal cycle of a time series (MSCChange) (Fig. 2 (c)), which might happen in the real-world carbon cycle as response to combined drought-heatwaves (Ciais et al., 2005).*
4. *Change in the variance of the time series (VarianceChange) (Fig. 2 (d)), e.g., in temperature (Huntingford et al., 2013).*

In real-world EO data, we are typically dealing with rather short time series (e.g., less than 10-15 years). Long-term memory processes cannot reliably diagnosed with such short time series (Ghil et al., 2011, e.g.). However,

our main focus in this paper is distinguishing anomalies from 'normal' short term noise and not to detect dynamical changes in the processes of a system or to infer anything about the (dynamical) reason behind the anomaly. In case an extreme anomalies occurs, it will certainly impact people, ecosystems, etc., regardless whether it was a random natural event or due to a change in the system's dynamics. In this paper, we want to detect such kind of events. Therefore, we consider the red/white noise in our artificial data farm to reflect both 'true' dynamical components (esp. In the 'signal' of the independent components) as well as measurement uncertainty (which is also explicitly added as additional white noise on the top of each variable). It would indeed be a very different but nevertheless very interesting question how multivariate anomaly detection algorithms perform in the presence of long-term memory and how to distinguish anomalies which occur due to long-term memory from anomalies in short term noise. This question is beyond the scope of this paper, but we thank the reviewer for this interesting aspect for future research.

Reviewer:
3. To me, the idea behind mwVAR is not fully clear. Subtracting the median mwVAR just removes a constant factor from the time series as it is described now. Maybe it should be better explained here how this specific "preprocessing" step works.

Response:
We thank the reviewer to point to this formulation and understanding issue. Subtracting a constant (the median of moving window variance) is indeed not influencing the results of algorithms based on pairwise distances. We rephrase the paragraph (p.7, l. 17 - p.8, l. 2) to: *Computing the moving window Variance (mwVAR) is a popular technique for detecting trends in the variance in univariate time series (e.g., Huntingford et al., 2013). We choose a window size of 10 and compute the variance in the running window along the time series of each variable. We use the estimates of the mwVAR time series as feature to detect multivariate anomalies in the variance.*

Reviewer:
4. On p.8, l.22, the authors address the model parameters. However, these parameters have not been introduced before, so it is hard to grasp their meaning at this point.

Response:
For better understandability, it seems to be more logical to us to describe the parameter estimation procedure once, before introducing the anomaly detection algorithms. To clarify, we changed p.9, l. 6 to: *S*ome anomaly detection algorithms require the estimation of parameters (Details are given

below for each algorithm separately). In that case we fix the model parameters for the entire data cube. We estimate model parameters (σ, ε, Q, μ, see below) and train the models themselves (Support Vector Data Description, Kernel Null Foley-Sammon Transform, see below) based on a random subsample of 5000 data points obtained from the entire data cube.

Reviewer:
5. Quite a bit of potentially interesting material is "not shown" by the authors. I understand and agree the need to focus on the most important aspects, but maybe the authors could consider preparing some supplementary material containing these additional results.

Response:
We are pleased that the reviewer is interested in additional aspects which we did not show. We are only aware of two aspects, which are referred to but not shown in the paper:

1. "Training and testing SVDD on each pixel did also not improve the results" (previously p.15, l.2). We thank the reviewer for pointing to this aspect as this is not only an experimental result, but even a theoretical finding. Training and testing SVDD with the same parameterisation (ν) on each pixel assumes the same number of anomalous events in each pixel. Therefore, it cannot improve the detection rate in datasets with varying number of anomalous events. We propose to rephrase the sentence (now p.15, ll.24-27) to:

   Training and testing SVDD on each pixel does also not improve the results as the amount of anomalies differs between different pixels in our setting. This contrasts the assumption when training SVDD on each pixel with constant outlier ratio (ν parameter).

2. AUC values of different σ (KDE) or ε (REC) choices (previously p.10, l.4 and p.14, l.10). We will prepare supplementary material including one figure (S1, attached), which shows a small simulation (500 repetitions) in which we are trying to detect one anomalous event (BaseShift) with different σ (KDE) or ε (REC) choices. We change σ (or ε, respectively) between the 0.05 and 0.95 quantile of the distance matrix. Results exhibit, that REC has slightly higher AUC values for optimal ε choices, whereas KDE is largely insensitive to different σ choices in the given range.

Reviewer:
6. In general, the parameter selection in the different methods is not well motivated (e.g., embedding delay and dimension, number of nearest neighbours, outlier ratio). Some more words on these aspects would be helpful.

The authors shortly address the subjectivity of parameter selection on p.16, ll.3-6, but do not mention that there are established ways to make (some of) these parameter selections at least a bit more objective. I do not request a detailed discussion on this aspect, but it would be worth mentioning it at least.

Response:
We explicitly wanted to point to the heuristic parameters choices, as we are aware that this is a crucial aspect for our results. Therefore, we selected the parameters very carefully. However, we assume that the reviewer's concerns especially about the time delay embedding for which much more objective criteria exist and apologize for not mentioning these criteria in the manuscript. To address this issue we will extend the following sentences:

1. p.17, ll.6-8: The artificial data farm's intrinsic dimension is 3 as it was created from three independent components. Therefore the embedding dimension m is fixed accordingly although it can be inferred based on the data by determining the number of false nearest neighbours (Kennel et al., 1992; Hegger et al., 1999).

2. p.8, ll.8-12: We fix m to 3 (*corresponding to the number of independent components within the data farm creation*) and τ to 6 which is a compromise between the typical choice of the first zero crossing of the temporal autocorrelation function *or the first local minimum of the mutual information* (Fraser and Swinney, 1986; Webber and Marwan, 2015).

Technical comments:

Reviewer:
- p.1, keywords: please capitalize the names Mahalanobis, Foley and Sammon

Response:
Done

Reviewer:
- p.3, ll.1-3: The papers by Donges, Rammig, Zscheischler et al. use only a bivariate form of event coincidence analysis. Since the authors refer her to the "truly multivariate" case, a better reference would be Siegmund et al., Front. Plant Sci., 7, 733, 2016, who introduced a multivariate version of event coincidence analysis.

Response:
We now refer additionally to Siegmund et al. (2016) at the p.3, l. 5 as well as on p.9, l.20.

Reviewer:
- p.3, l.21: The term "data cube" should be explicitly defined here – it is intuitively clear (especially in connection with Fig. 1), but especially the spatial component (2d vs. 3d) could differ from what is considered in this paper.

Response:
We defined it now as follows (p.3, l.21): *S*patio-temporal EOs are therefore stored in the Earth system data cube, which is a 4 dimensional array of latitudes, longitudes, time and different measurement variables. To detect multivariate anomalies in EOs, we define an anomaly to be any consecutive spatiotemporal part of the data cube ...

Reviewer:
- p.3, ll.28-30: It is not clear if the authors wish to consider "multivariate events" or "compound events" (i.e., such that are anomalous with respect to the marginal feature distribution of a single variable or the joint feature distribution of a (sub)set of variables.

Response:
We thank the reviewer for this question. It is definitely within our scope to consider also "multivariate events", i.e. anomalous events where none of the single variables is extreme itself, but their joint feature distribution is anomalous. However, "compound events" are usually defined as multivariate events which are additionally leading to an extreme impact (Seneviratne et al., 2012; Leonard et al., 2013), which is not possible to evaluate with the artificial data farm. Nevertheless, we consider our study an important scoping study also in the context of compound events, as the proposed algorithms and workflows are in general capable to detect multivariate anomalous events, which might include compound events (with impact) in real EOs. For clarification we add (p.3, ll.32-34): Second, we use these artificial data to evaluate the capability of different algorithms to detect multivariate anomalous events, *including compound events (i.e. events where none of the single variables is extreme, but their joint distribution is anomalous and might lead to an extreme impact) (Seneviratne et al., 2012; Leonard et al., 2013).*

Reviewer:
- p.4, l.15: Why is Appendix B referenced in the paper before Appendix A. I think that changing the order of both Appendices would be more logical.

Response:
We changed the order according to the reviewer's suggestions.

Reviewer:
- p.5, ll.6-9: replace a, b, c, d by (a), (b), (c), (d)

Response:
Done

Reviewer:
- p.5, l.14: I think that it is not the Earth observations (EOs) that are driven by extrinsic forcings, but the EO variables.

Response:
We changed EOs to EO variables.

Reviewer:
- p.6, l.24: In fact, what you study is the maximization of the rate of correct detections at simultaneous minimization of false detections (this is essentially what the ROC analysis does).

Response:
We thank the reviewer for this comment. However, we are not convinced that mentioning details on ROC analysis facilitates understanding of the essential point here, which deals with the term feature extraction and its justification. To clarify we change the term "event detection rate" to "detection of anomalous events" (p.6, l.24). The exact definition of ROC/AUC characteristics is given later, p. 10, ll.31-33.

Reviewer:
- p.6, l.26: "the anomaly time series becomes the feature then" – maybe the authors should explicitly state here what they "define" (consider) to be meant by a feature.

Response:
The definition of feature extraction is already given a few sentences before (p.6, l. 23). Therefore, we changed the sentence to: *A very simple form of feature extraction could be to subtract the mean seasonal cycle. We consider the anomaly time series to be the extracted feature in this case.*

Reviewer:
- Figure 3: Since the authors allow for combining different feature extraction techniques, they should emphasize here that their application might be non-commutative in some cases. For example, TDE must be performed after sMSC, otherwise, the signal would be dominated by seasonality and potentially reflect different features than those one is actually interested in.

Response:
We thank the reviewer for this comment and add a sentence on that (p.9, ll. 2-3): *In some cases this might lead to non-commutative combinations, especially for non-linear feature extraction techniques.*

Reviewer:
- p.7, l.9: "This theoretical consideration does not hold true for high dimensional multivariate data." Do the authors have a reference for this? I am not convinced that this statement is correct in general. In particular, one may refer to multi-channel SSA (mSSA), which essentially combines TDE for multivariate data with PCA. What is the difference between mSSA and "dynamic PCA" mentioned in p.7, ll.18-19?

Response:
We thank the reviewer for his comment on multi-channel SSA. Dynamic PCA and mSSA are not different in technique, although their purpose differs (extracting main frequencies, versus smoothing for subsequent process monitoring). We removed the statement about the theoretical consideration of high dimensional multivariate data.

Reviewer:
- p.8, l.2: To my knowledge, there are various variants of ICA, and the one maximizing the negentropy is just one version among several others.

Response:
The currently used formulation was indeed not ideal. We specify the sentence as follows: *We use one ICA variant which tries to separate different sources of data by maximizing the negentropy*

Reviewer:
- p.8, l.12: "in the literature"

Response:

Done

Reviewer:
- p.8, l.21: "we fix the model parameters"

Response:
Done

Reviewer:
- p.8, ll.22-23: "model parameters. . . and the models themselves. . . are estimated" – better use the terms model selection and parameter estimation separately

Response:
Done

Reviewer:
- p.8, ll.24- 25: Do the authors mean "more resampling is NOT affordable. . ."?

Response:
Yes, indeed. We changed it.

Reviewer:
- p.8, l.26: "a resampling of 3" – 3 what?

Response:
... 3 times. We changed it.

Reviewer:
- p.8, l.30: "if one or several of the univariate variables are below or above a certain quantile threshold" – again: do the authors mean marginal quantiles or multivariate quantiles (i.e., multivariate or compound extremes)? Page 9, ll.2-3 suggests that they refer to extremes in the marginals.

Response:
For the "univariate approach" we refer to quantiles in the distribution of each single variable separately, i.e. to extremes in the marginals. To clarify we changed p.9, l.17 to: In this case, one would consider a data point to be extreme, if one or several of the univariate variables are above (or below) a certain quantile threshold *of the marginal distributions of each single variable.*

Reviewer:
- p.9, ll.1-2: The event coincidence analysis the authors refer to here is a bivariate (or, in its extension, multivariate) statistical method. Its relevance in the context of the present work is not clear, since I do not find information that statistical interrelationships between anomalies in different variables are considered here.

Response:
We totally agree with the reviewer that the mentioned coincidence analysis do not consider interrelationships between different variables. Therefore, we also write p.9, l.15, that the technique is "multiple univariate". We would not consider it, to be a "real" multivariate technique as the following ones. However, it is the simplest technique for detecting anomalies in multiple data streams. Thus, we use the technique as benchmarking for the other algorithms.

Reviewer:
- p.9, l.3: Details on the definition of the threshold exccedance score should be given.

Response:
We will add the details on that: different thresholds in terms of quantiles of the marginal distributions between 0.0 to 1.0 (accuracy 0.01) are used.

Reviewer:
- p.9, l.12: I suppose that the authors are using standardized variables; otherwise, defining distances across different variables might not make much sense in the real- world data case. I recommend elaborating a bit more on this aspect.

Response:
In our artificial data, the variables are already comparable by construction, so standardisation is not needed. However, for the real-world data standardisation is important. Furthermore it might even be an additional error source, if not applied with care. We elaborate already on that on p. 17, l.21-24, but nevertheless add an additional sentence for clarification: For real-world data, variables have to be standardized with care before computing the distance matrix (Sect. 5). However, in our artificial data farm the variables are already comparable by construction, thus standardization is not needed.

Reviewer:
- p.9, ll.15-16: This formulation should be checked again; for me, the difference between the two measures does not become obvious from the given description.

Response:
We change it to: K-nearest neighbours (KNN) can be used for anomaly detection by considering the mean distance to the k-nearest neighbors (KNN-Gamma) and the length of the mean of the vectors pointing from X_t to its k-nearest neighbors (KNN-Delta). With that approach KNN-Delta considers also the direction of the neighbors, i.e. has higher values in case its nearest neighbours are pointing in one direction, which is in contrast to the directionless distance of KNN-Gamma.

Reviewer:
- p.9, ll.18-19: In how far do the authors really "take advantage" here? Isn't it rather that you wish to exclude trivial information due to autocorrelation in your variables?

Response:
Indeed, "take advantage" was rather meant in the sense of improving the algorithm's capability to deal with autocorrelated data. Thus, we reformulate according to the reviewers suggestion to: *We exclude trivial temporal*

autocorrelations by excluding 5 neighbouring time steps to be also nearest neighbours.

Reviewer:
- p.9, l.29: "An epsilon-hyperball"

Response:
Done

Reviewer:
- p.9, l.31: $\zeta \cdot T^{-1}$

Response:
Done

Reviewer:
- p.9, l.32: "degree of centrality" is not the proper network theoretic term (it would be "degree centrality" or just "degree"); however, what the authors consider here is not the degree, but the "degree density" (cf. Donges et al., Phys. Rev. E, 85, 046105, 2012).

Response:
We thank the reviewer pointing this out and changed the term in degree density, as well as adding the correct citation.

Reviewer:
- p.9, l.33: "quantiles of the distribution of elements of the distance matrix" (also on p.10, ll.3-4)

Response:
Done

Reviewer:
- p.10, l.20: "of the one-class support vector machine"

Response:
Done

Reviewer:
- p.10, l.28: "that is fixed"

Response:
Done

Reviewer:
- p.11, ll.7-9: Temperature extremes represent strong deviations from the mean rather than "changes in the mean".

Response:
We changed "changes in the mean" to "deviations from the mean".

Reviewer:
- p.14, l.17: Do the authors mean "mean length of the vectors"?

Response:
Indeed, we changed it.

Reviewer:
- p.16, l.1: "that these findings" or "that their findings"

Response:
Changed to "their findings".

Reviewer:
- p.17, l.22: "parameterise"

Response:
Done

Reviewer:
- p.18: It is a bit unusual to write the Conclusions completely in present tense. Maybe you wish to consider using present perfect here?

Response:
We rewrote the first paragraph of the Conclusions in past tense.

Reviewer:
- p.18, ll.12-13: Maybe it is worth clarifying here again that the results apply for the considered types of anomalies?

Response:
We added "for the considered event types" for clarification.

Reviewer:
- Figure A2: It would be interesting to see these charts detailed for the different detection algorithms (e.g. using different colors for the respective bars). Maybe the authors could add some corresponding figure as supplementary material?

Response:
We prepared an additional figure as suggested by the reviewer in the supplementary material.

Reviewer:
- p.21, l.4: I suggest putting the two equations in brackets.

Response:
Done

Reviewer:
- The authors should check/revise/complete the following citations: Bintanja and van der Linden (2013), Faranda and Vaienti (2013) [remove publisher], Pfeifer et al. (2011) [capitalization of "Earth"], Pinheiro et al. (2016) [capitalization of "R"], Poincare (1890) [incomplete reference], Smetek and Bauer (2007), van der Maaten (2009), Webber and Marwan (2015) [page numbers].

Response:
Done

References

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogee, J., Allard, V., Aubi- net, M., Buchmann, N., Bernhofer, C., Carrara, A., Chevallier, F., De No- blet, N., Friend, A. D., Friedlingstein, P., Gruenwald, T., Heinesch, B., Keronen, P., Knohl, A., Krinner, G., Loustau, D., Manca, G., Matteucci, G., Miglietta, F., Ourcival, J. M., Papale, D., Pilegaard, K., Rambal, S., Seufert, G., Soussana, J. F., Sanz, M. J., Schulze, E. D., Vesala, T., and Valentini, R. (2005). Europe-wide reduction in primary productivity caused by the heat and drought in 2003. Nature, 437(7058):52–533.

Fraser, A. M. and Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. Physical Review A, 33:1–7.

Ghil, M., Yiou, P., Hallegatte, S., Malamud, B. D., Naveau, P., Soloviev, A., Friederichs, P., Keilis-Borok, V., Kondrashov, D., Kossobokov, V., Mestre, O., Nicolis, C., Rust, H. W., Shebalin, P., Vrac, M., Witt, A., and Zaliapin, I. (2011). Extreme events: dynamics, statistics and prediction. Nonlinear Processes in Geophysics, 18(3):295–350.

Hegger, R., Kantz, H., and Schreiber, T. (1999). Practical implementation of nonlinear time series methods: The TISEAN package. Chaos: An Interdisciplinary Journal of Nonlinear Science, 9(2):413.

Huntingford, C., Jones, P. D., Livina, V. N., Lenton, T. M., and Cox, P. M. (2013). No increase in global temperature variability despite changing regional patterns. Nature, 500(7462):327–330.

Kennel, M. B., Brown, R., and Abarbanel, H. D. I. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. Physical Review A, 45:3403–3411.

Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., Risbey, J., Schuster, S., Jakob, D., and Stafford-Smith, M. (2013). A compound event framework for understanding extreme impacts. Wiley Interdisciplinary Reviews: Climate Change, 5(1):113–128.

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., and Zhang, X. (2012). Changes in climate extremes and their impacts on the natural physical environment. In Field, C., Barros, V., Stocker, T., Qin, D., Dokken, D., Ebi, K., Mastrandrea, M., Mach, K., Plattner, G.-K., Allen, S., Tignor, M., and Midgley, editors, Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (IPCC SREX Report), pages 109–230. Cambridge University Press.

Siegmund, J. F., Sanders, T. G. M., Heinrich, I., van der Maaten, E., Simard, S., Helle, G., and Donner, R. V. (2016). Meteorological Drivers of Extremes in Daily Stem Radius Variations of Beech, Oak, and Pine in North-eastern Germany: An Event Coincidence Analysis. Frontiers in Plant Science, 7:220.

Webber, Jr., C. L. and Marwan, N. (2015). Mathematical and Computational Foundations of Recurrence Quantifications. In Recurrence Quantification Analysis, pages 3–43. Springer, Cham Heidelberg New York Dordrecht London.