Detecting changes in essential ecosystem and biodiversity properties- towards a Biosphere Atmosphere Change Index: BACI

Deliverable 5.1: First Methods for Novelty Detection for Synthetic Change Index

Project title:	Detecting changes in essential ecosystem and biodiversity properties- towards a Biosphere Atmosphere Change Index
Project Acronym	BACI
Grant Agreement Number:	640176
Main pillar:	Industrial Leadership
Topic:	EO-1-2014: New ideas for Earth-relevant space applications
Start date of the project:	1st April 2015
Duration of the project:	48 months
Dissemination level:	Public
Responsible of the deliverable:	Joachim Denzler Phone: +49 3641 9 46420 Email: joachim.denzler@uni-jena.de
Contributors:	Erik Rodner, Milan Flach, Yanira Guanche, Miguel Mahecha, Paul Bodesheim
Date of submission:	March 29, 2016

Contents

Summary

This deliverable is dedicated to the works done within the *WP5 - Synthetic Index and Attribution Scheme: the BACIndex*. WP5 is divided into 4 main tasks and this first report refers to the first one: *Task 5.1 - Time-series extension null space and GP methods and combination with spatial clustering*.

The main focus of our work in this task is the analysis, comparison and development of methods and techniques that allow for the automatic detection and location of abnormal events in multivariate time series which is a key element towards addressing the issue of developing a "Biosphere Atmosphere Change Index". Several existent methods and some newly developed ones have been intensively tested and their results have been compared.

This deliverable is divided into three main sections:

1. **Evaluation of Novelty Detection Algorithms.** In this section, a comparison between different techniques to detect extreme events is presented. This comparison has been done by using an artificial dataset created in such a way that it represents the same characteristics as real earth observation data.

2. **Kernel Functions for Mixed Discrete-Continuous and Partially Observed Time Series.** Within this section, we present a set of kernel functions with potential application to earth observation data.

3. **Max-Divergent Regions - A new Machine Learning Algorithm for Extreme Event Detection.** In this last section, we present a newly developed method to detect abnormal regions or events in multivariate time series. The method is based on maximizing the Kullback-Leibler divergence and it has been applied to a set of real data of marine climate.

The work summarized in this deliverable represents the accomplishment of the duties encompassed in Task 5.1.

1 Evaluation of Novelty Detection Algorithms

This section summarizes the work done comparing different methods and techniques to detect multivariate abnormal events. A more detailed description of the methods as well as the results obtained is intended to be submitted to a peer-reviewed journal as: *'Detecting Multivariate Events in Artificial Earth Observartion Data'- Milan Flach et al.*.

1.1 Artificial dataset

In order to test different methods for event detection, we have used an artificial dataset that allowed us to have total control of the events before applying the methods to real data. This artificial test data set was initially developed for applications of this kind by the ESA STSE project "Coupled Atmosphere Biosphere virtual LABoratory, CAB-LAB" [1]. The final objective is to select the best methods with respect to their performance in detecting events in real data. Therefore, the artificial multivariate data cube needs to fulfill some requirements such as seasonality, correlation, and non-linearity.

Three independent components $\Theta_{t,lat,lon,var}$ are created with a normal variability of Gaussian noise, $sd = 1.$, each representing intrinsic properties of the earth system [6]:

$$\Theta_{t,lat,lon} = B_{t,lat,lon} \cdot 2^{(k_b \cdot ev_{t,lat,lon})} + N_{t,lat,lon} \cdot 2^{(k_n \cdot ev_{t,lat,lon})} + k_s \cdot ev_{t,lat,lon} \cdot sd \qquad (1)$$

where N represents a common variation or noise, added to a baseline B and $ev_{t,lat,lon}$ are the events manually introduced. The magnitude of the event is multiplied by a parameter separately for the baseline (k_b), the noise (k_n) and a mean-shift parameter k_s scaled with the standard deviation of the data sd.

Earth system properties, $\Theta_{t,lat,lon}$, are not directly measured but indirectly monitored through correlated variables, $\boldsymbol{X}(t)$.The artificial set of correlated variables ($\boldsymbol{X}$) is created from the earth system properties (Θ_{var}) by weighting them with linear random weights w. This process is illustrated in Figure 1.

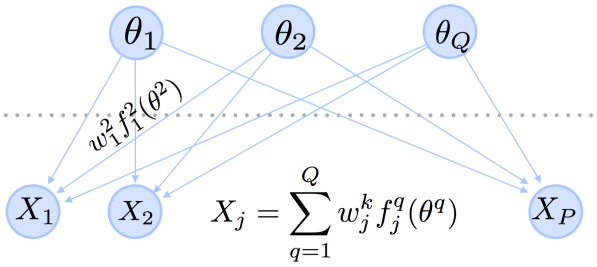

Figure 1: Combination of independent components Θ_{var} to a set of correlated variables $\boldsymbol{X}$.

Thus, we finally create an artificial datacube $\boldsymbol{X}_{t,lat,lon,var}$ consisting of 10 variables var, with 300 observations or timesteps t and a spatial coverage of 100 points of lat and lon yielding a grid of 10000 points.

[1] see: *http://earthsystemdatacube.net/*

1.2 Events definition

Within the artificial datacube, we introduced 5 different types of events resulting in 5 different datacubes. These events together with their abbreviations are:

- Shift in the mean of stationary data (*BaseShift*)
- Change in the variance (*VarianceShift*)
- Change in the amplitude of the mean annual cycle (*MACShift*)
- Trend in the time series (*TrendOnset*)
- Trend with an abrupt change to the initial state (*TrendJump*)

To complement these events and to increase the difficulty for event detection, we also experimented with the following cases:

- Increase of the noise signal (*NoiseIncrease*)
- Long tailed Cauchy distributed noise (*NonGaussianNoise*)
- Red noise with temporal long term correlation and spatial correlation of the intrinsic components (*CorrelatedNoise*)
- Use of different spatial distribution shapes of the noise to an event, randomly affecting its neighbor in the next timestep (*RandomWalkExtreme*)
- Increase of the number of intrinsically independent components from 3 to 6 (*MoreIndepComponents*)
- Different temporal length of the extremes (*LongExtremes*,*ShortExtremes*)

1.3 Methods

The methodology proposed to achieve a Change Index can be divided in different steps: *i)* Preprocessing, *ii)* Feature Extraction, *iii)* Event Detection and *iv)* calculating the Change Index (Figure 2). The preprocessing step includes any pre-treatment done to the variables like normalizing them or transforming them by its logarithms for example.

1.3.1 Feature Extraction Techniques

The following techniques, or combinations of them, have been tested:

- **Spatial filter (SF)**: consists of a spatial moving average of surrounding cells. We use a 3x3 filter matrix, weighting the value of each pixel itself with 50%, the value of its neighbours with 7.5% or 5% at the edges.

- **Time Delay Embedding (TDE)**: increases the feature vector with time delayed vectors to include temporal context information [5] [17]. Critical hyperparameters are the time delay τ and the number of dimensions. We fix τ to 6 and the dimension to 3, which is a compromise between the typical choice of the first zero crossing of the temporal auto correlation function (here: 12) and an accurate temporal detection (requires small τ).

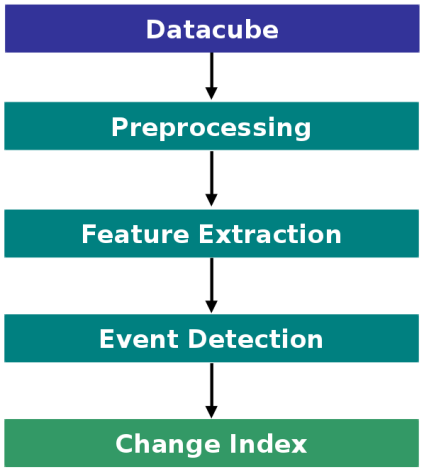

Figure 2: Diagram of the methodology.

- **Principal Component Analysis (PCA)**: transforms a set of correlated variables into a set of linearly uncorrelated variables, called principal components. We choose the number of principal components such that at least 95% of the variance in the original datacube are explained.

- **Independent Component Analysis (ICA)**: is one standard technique of data-based process monitoring that separates a multivariate signal in additive non-gaussian subcomponents which are assumed to be statistically independent [5].

- **Substraction of the Mean Annual Cycle (sMAC)**: it is a common technique in extreme detection in environmental variables that present a seasonal behavior. The remainder part of the time series is often referred to as anomalies [21] [20].

- **Exponential 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 statistical process control to detect only 'significant' outliers [12].

1.3.2 Event Detection Methods

- **Hotelling's** T^2: this method computes the mahalanobis distance from each data point x to its mean [8].
- **Multivariate Exponential Weighted Moving Average (MEWMA)**: it is based in Hotelling's T^2, but with an exponential weighting of the data in the temporal dimension [12] [11] .
- **Kernel Density Estimation (KDE)**: is a standard technique of estimating densities using gaussian kernels centered on each datapoint [15].
- **K-nearest neighbors (KNN)**: can be used for outlier detection by considering the mean distance (γ) and the length of the mean vectors (δ) of the K nearest neighbors [7].

- **Support Vector Data Description (SVDD)**: tries to fit a hypersphere around the given data by allowing some slack variables [19].
- **Recurrences**: a concept based on the theory of nonlinear dynamical systems assuming that each state of a dynamical system will revisit a certain region in its phase space, if waiting for a sufficiently long time [16]. These dynamics can be visualized in the recurrence plot and are quantified with a bunch of measures. We use the local recurrence rate to quantify rare events [13]. A similar technique has been already applied in computer vision applications,[9].
- **Kullback-Leibler Divergence**: This new developed method will be explained in detail in Section 3.
- **Kernel Null Foley-Sammon Transform (KNFST)**: maps the training data into a 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 [2]. As for SVDD, a kernel matrix is used as input and learning the null space model is performed based on a random sample of 5000 points. In the testing phase, we compute pairwise similarities between test data and training samples with the kernel function, which are used to project the test data into the null space in order to finally compute the absolute distance to the training target value as a measure of discrepancy.
- **Univariate**: Extremes in an univariate sense are any data above (or below, depending on being maximum or minimum extremes) a certain threshold. This is the simplest way to define an extreme and therefore is included here for comparison reasons.

1.4 Preliminary Results

To compare the results, the Area Under the Receiver Operator Characteristic Curve (AUC) was calculated. This metric allows us to evaluate the performance of each method.

In Figure 3, results are shown for the different types of events applied to the entire artificial datacube created (10 variables, 300 timesteps and a spatial grid of 10.000 nodes, 100x100). For each event type, the x-axis shows the difficulty parameter used to create the event (we refer to this as complication), the colors represent the method used and the symbols the feature extraction.

From the results obtained it can be seen that for a given event type and complication, the value of AUC (and therefore the ability to correctly detect the events) can vary largely depending on the method and the feature extraction technique used. In case of detecting changes in the baseline (BaseShift), fast recovering trends (TrendOnset) and a shift in the variance (VarianceShift), Gamma (K-nearest neighbors) is the best algorithm. However, it is closely followed by Parzen and Recurrences, which yield highest results for a change in the mean annual cycle (MACShift) and a starting trend in the time series (TrendOnset). Regarding the feature extraction techniques, PCA-EWMA is the one to choose for BaseShift, TrendOnset and TrendJump. For MACShift TDE-PCA-sMAC-EWMA exhibits considerably better results. In the case of VarianceShift,

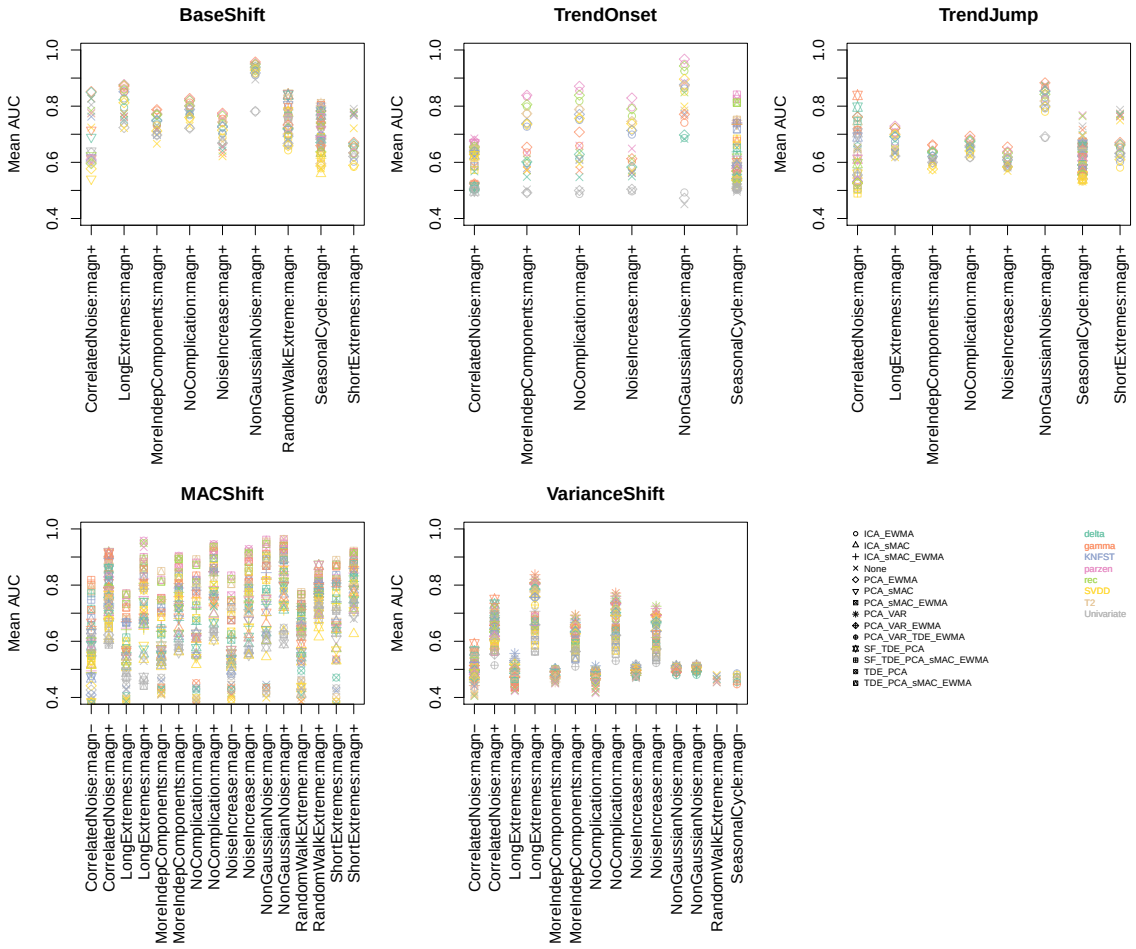

Figure 3: Comparison results in terms of AUC for the different types of events.

PCA-VAR is the best.

In this figure the Kullback-Leibler divergence method was not included. This method, in its current state, is computationally demanding when dealing with large spatial grids. Therefore a smaller grid of 5 latitudes and longitudes (5x5) has been tested. The results of the tests done to this smaller grid are presented in Figure 4.

In this smaller grid tested, for the case of studying the BaseShift and TrendJump the Kullback-Leibler divergence and the K-nearest neighbors methods get the promising results in combination with the PCA-VAR feature extraction technique. Further work will be done in terms of improving the Kullback-Leibler divergence method and its applicability to large-scale datasets. Please also note that only a single early prototypical version of the KL-method has been tested so far.

1.5 Conclusions

Several methods and techniques were analyzed and compared in their ability to detect different types of events in a multivariate dataset. To perform these experiments and comparisons, an artificial dataset representing the same characteristics of Earth Ob-

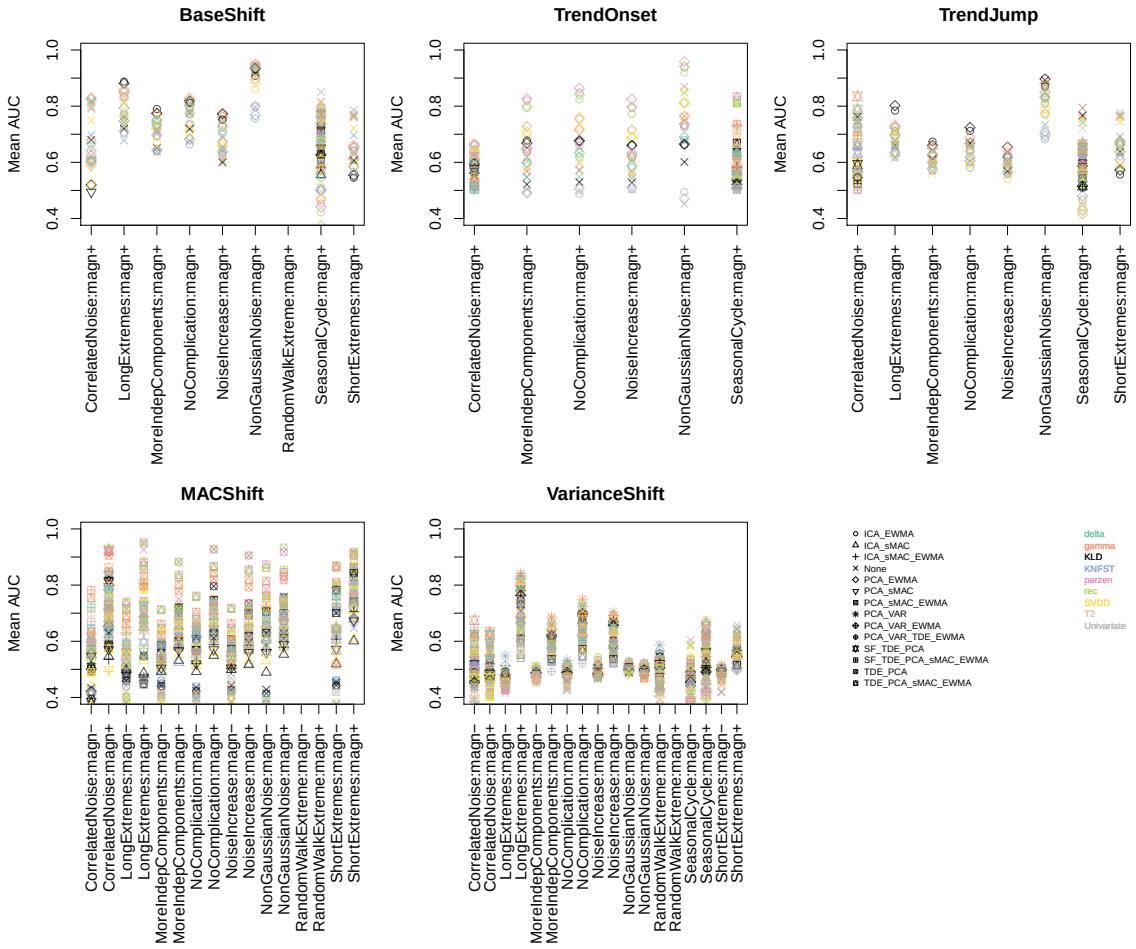

Figure 4: Comparison results in terms of AUC, for a smaller grid and including an early prototypical version of the Kullback-Leibler divergence method, for the different types of events.

servation data was previously created.

From the results obtained, we are able to provide some guidance in terms of which method performs better regarding the kind of events to be detected: the Gamma, Parzen and Recurrences methods are good choices for detecting changes in the baseline, the variance, fast recovering trends and changes in the mean annual cycle. In addition, in terms of the feature extraction techniques tested: the combination PCA-EWMA gives better results for trends detection and changes in the baseline; PCA-VAR is the best option for shifts in the Variance and TDE-PCA-sMAC-EWMA outperforms the other techniques when detecting changes in the mean annual cycle.

With this analysis we are able to go one step further and apply the selected methods to real data.

2 Kernel Functions for Mixed Discrete-Continuous and Partially Observed Time Series

As can be seen in our report about existing novelty detection algorithms, a large number of the techniques depends on kernel functions that measure the similarity between data points. The behavior of the algorithm and the model learning highly depends on the kernel function used.

The multivariate time series we have to deal with in BACI make choosing the right kernel function even more challenging. This is mainly due to: (1) different scales of the variables, (2) a mixture between discrete and continuous variables and (3) missing measurements for certain spatio-temporal positions. Whereas (1) can be tackled by proper normalization of the ranges, issue (2) requires new types of kernel functions and issue (3) asks for statistically reasonable strategies of imputation and marginalization.

2.1 Kernel functions for mixed discrete continuous time series

A standard kernel function often used as a default solution in machine learning is the *squared exponential kernel* (often: *radial basis function* or *Gaussian kernel*):

$$K_{SE}(\mathbf{x}, \mathbf{x}_*) = \sigma^2 \exp\left(-\frac{(\mathbf{x} - \mathbf{x}_*)^2}{2\ell}\right) \quad , \tag{2}$$

where σ^2 and ℓ are kernel-specific hyperparameters. The kernel K_{SE} is often a good starting point because it can be applied to a large set of problems and corresponds to an infinite dimensional Hilbert space. However, since it directly uses the Euclidean distance between two data points it theoretically requires variables coming from a continuous input space. For a given data domain within BACI this is only true for some of the variables. Other variables come from a discrete space related to a given categorization. For these cases, the Euclidean distance is not reasonable, since a discrete value of 1 is not closer to 2 than to every other positive discrete value within the categorization classes.

Therefore, it is necessary to define kernel functions for non-continuous attributes. In [3], the authors compare 14 different similarity measures for discrete data. In the following, we consider additive kernel functions, where the similarity between two points is defined by:

$$K(\mathbf{x}, \mathbf{x}_*) = \frac{1}{D} \sum_{d=1}^{D} S_d(\mathbf{x}, \mathbf{x}_*) \ . \tag{3}$$

with S_d being kernel functions specific for each dimension d of the input variables.

We chose two rather simple but good performing measures from [3] as a basis for discrete kernel function. The *Overlap* measure returns 1 if the value for attribute d is equal and 0 otherwise. This kernel function is equivalent with a Gaussian kernel with zero variance parameter and corresponds to a delta impulse. Using it for discrete and especially categorical attributes is very intuitive, since no relation on the values is defined.

The second measure is called *Goodall4* [3]:

$$S_d(\mathbf{x}_d, \mathbf{x}_{*d}) = \begin{cases} \frac{f_d(\mathbf{x}_d)(f_d(\mathbf{x}_d)-1)}{N(N-1)} & \text{if } \mathbf{x}_d = \mathbf{x}_{*d} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $f_d(x)$ is the frequency of how often the attribute d takes value x. In case the values of two examples are equal, the similarity highly depends on the frequency of the values. More unlikely values also yield a lower similarity of the examples. This kernel functions is especially useful when discrete and categorical attributes have non-uniform probability distributions. Unusual values that have been rarely observed in the training set might be related to data noise and are down-weighted in the kernel function.

We test these kernels for an upscaling regression task related to WP 4.1. For the GP regression in our experiments, we use different kernel functions for different variables of our data. We build a base kernel $\mathbf{K}_c$ on continuous variables using the SE kernel function and a base kernel $\mathbf{K}_d$ on discrete dimensions using either *Overlap* or *Goodall4*. Both $\mathbf{K}_c$ and $\mathbf{K}_d$ are combined in order to get the final kernel $\mathbf{K}$. We choose a combination by adding both base kernels since it produces a final kernel which has high values if either of the two base kernels has a high value.

3 Max-Divergent Regions – A New Machine Learning Algorithm for Extreme Event Detection

We show how to find extreme regions in Earth Observation data. Our approach is based on maximizing the Kullback-Leibler divergence between the data distribution within a region and the distribution outside of the region. Modelling the data distribution can be done either by kernel density estimation [15] or Gaussian assumptions, depending on the type of input data and the runtime requirements of the algorithm.

3.1 Definitions and problem description

We first focus on finding extreme regions in time series $(\boldsymbol{x}_t)_{t=1}^n$, where $\boldsymbol{x}_t \in \mathbb{R}^D$ is a multivariate observation at time step t. Furthermore, we assume that all the data is given in advance and we want to detect extreme regions a posteriori in an offline batch fashion.

A very intuitive way of characterizing an extreme region is that its data distribution differs significantly from the data distribution of the remaining time series.

In the following, we will denote a region with $I = \{t \mid t_1 \leq t < t_2\}$ and its data distribution with p_I. The remaining set of data points is denoted by $\Omega = \{1, \ldots, n\} \setminus I$ with data distribution p_Ω. To develop a concrete algorithm based on above definition, the following questions have to be answered:

1. How can we calculate a difference between data distributions p_I and p_Ω?
2. How can the data distributions be modelled and estimated?
3. How is it possible to find the region with maximum difference in the data distributions?

In the following, we propose to use the empirical Kullback-Leibler divergence to measure the difference between distributions. Furthermore, we show that by modelling the data distributions either by kernel density estimation or simple Gaussian assumptions, we can compute the empirical KL-divergence in an efficient manner, which also allows for greedy optimization of the region later on.

3.2 Maximizing the Kullback-Leibler divergence

The empirical Kullback-Leibler divergence of two distributions p_Ω and p_I is defined as follows:

$$\mathsf{KL}(p_I, p_\Omega) = -\int p_I(\boldsymbol{x}) \log \frac{p_I(\boldsymbol{x})}{p_\Omega(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \ . \tag{5}$$

The KL divergence is zero for identical distributions and large for "significantly different" data distributions. We approximate it using an empirical expectation over the set of

extreme points leading to:

$$\mathsf{KL}(p_I, p_\Omega) \approx KL_{I,\Omega} = \frac{1}{|I|} \sum_{t \in I} \log \frac{p_I(\mathbf{x}_t)}{p_\Omega(\mathbf{x}_t)} \tag{6}$$

$$= \frac{1}{|I|} \sum_{t \in I} \left(\log p_I(\boldsymbol{x}_t) - \log p_\Omega(\boldsymbol{x}_t) \right) \tag{7}$$

This resulting criterion is very intuitive since it is calculating the differences of log-likelihoods for p_I and p_Ω. To find the region belonging to an extreme event, we maximize the KL divergence with respect to the region I:

$$\hat{I} = \mathrm{argmax}_x I \in \mathcal{I} \; \mathsf{KL}(p_\Omega, p_I) \ . \tag{8}$$

The set $\mathcal{I}$ contains suitable regions and is important to integrate prior expectations about extreme regions, such as a range of possible region sizes. Naive brute-force optimization of the KL divergence requires $\mathcal{O}(|\mathcal{I}| \cdot T)$ operations, where T is the time needed to evaluate the KL divergence and $\mathcal{I}$ is usually $\mathcal{O}(n \cdot n')$ with n' being possible lengths of an extreme region. A property of the KL-divergence is its asymmetry, *i.e.* $\mathsf{KL}_{I,\Omega} \neq \mathsf{KL}_{\Omega,I}$. The question is therefore which version fits best to our task. Other work [10] often relied on a symmetric version of it. In our experiments, we evaluate different choices and provide further insights on their differences.

Parameterizing the KL-divergence To allow for tuning our algorithm, we use a parameterized version of the KL divergence with a hyperparameter $\alpha > 0$:

$$\mathsf{KL}^\alpha(p_\Omega, p_I) = \frac{1}{n} \sum_{t=1}^{n} p_\Omega(\mathbf{x}_t) \log \frac{p_I^\alpha(\mathbf{x}_t)}{p_\Omega(\mathbf{x}_t)} \tag{9}$$

However, one should use something like the power divergence [18] or the density power divergence [1].

3.3 Maximally divergent intervals for arbitrary smooth distributions

A very flexible way to model and estimate distributions is kernel density estimation. For a given kernel function K, the estimate for p_I is defined by:

$$p_I(\mathbf{x}) = \frac{1}{|I|} \sum_{t_1 \leq t < t_2} K(\mathbf{x}, \mathbf{x}_t) \tag{10}$$

for an arbitrary multivariate observation $\mathbf{x}$. We use a similar estimate for p_Ω. As a kernel function, we use the common Gaussian kernel:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{\sqrt{2\pi}^D \sigma^D} \exp\left(\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2} \right) \tag{11}$$

with hyperparameter σ^2.

3.4 Relationship to the density ratio method of Liu et al

The most similar paper to ours is the method of Liu [10], which also uses a divergence criterion to detect changes in the data. However, there are major differences between their approaches and ours.

First, [10] considers always the last k data points in a time series $\boldsymbol{X}(t) = \{\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \ldots, \boldsymbol{x}_{t-k+1}\}$ and compares the distributions of $\boldsymbol{X}(t+k)$ and $\boldsymbol{X}(t)$, i.e. chunks of the time series with k data points. Second, they use the f-divergence defined by:

$$\text{f-div}(p_\Omega, p_I) = \int p_\Omega(\mathbf{x}) \cdot f\left(\frac{p_I(\mathbf{x})}{p_\Omega(\mathbf{x})}\right) \mathrm{d}\mathbf{x} \tag{12}$$

with the KL-divergence being a special case for $f = \log$ and with our notations in place. Furthermore, they present multiple methods where the density ratio in the above formula is not computed explicitly like in our case but estimated indirectly with different kernel methods:

$$\frac{p_I(\mathbf{x})}{p_\Omega(\mathbf{x})} = \sum_{t=1}^{n} \alpha_t \cdot K(\mathbf{x}_t, \mathbf{x}) \ , \tag{13}$$

where α_t are coefficients to be estimated. Whereas, directly estimating the ratio might have a benefit especially for small-sample cases, it would be not trivial to develop efficient algorithms for optimization of the interval I.

In summary, the method of Liu applied to our maximum-divergent region scenario could indeed lead to a performance gain but likely at the cost of a dramatically increased inference time.

3.5 Experiments

3.5.1 Data

Wave data (significant wave height, H_s and wave period, T) at a location near the North-Western Spanish coast were obtained from the hindcast DOW 1.1 (downscaled ocean waves, [4]), developed by IH Cantabria. Although these data have an hourly temporal resolution, for computational reasons we have used them 3-hourly aggregated.

3.5.2 One-dimensional application

Initially, despite the data used present a longer coverage we have tested the method only in one winter: from the beginning of November 2007 until the end of February 2008, therefore we are working with a dataset of about 1200 observations. From the time series presented in Figure 5 we extracted the Peaks Over Threshold (POT). Setting the threshold in 5 meters (corresponding to the percentile of 90% of the time series) and imposing a minimum lag of 3 days between peaks we extracted up to 8 peaks. The minimum lag of 3 days between peaks is related to the location of the data: in the North-Atlantic the typical duration of a storm is of 3 days, therefore for peaks separated

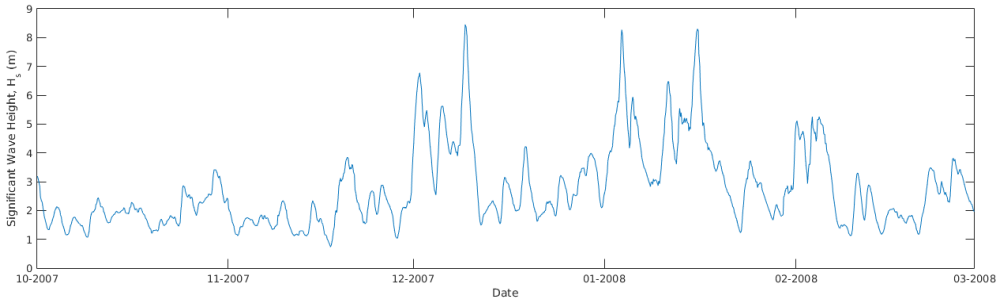

Figure 5: Wave data time series.

in time more than that can be assumed to be caused by different storms [14].

We have applied our KL method to the same time series, setting the size of the regions to be find equal to 24 timesteps (3 days with 8 observations per day) to be consistent with the 3 days independence lag imposed in the POT. In Figure 6 the results obtained with both methods are shown: the black dots represent the 8 peaks extracted with the POT. Shaded areas represent the regions obtained with the KL-divergence method. The areas colored in blue were obtained assuming that the points in the region follow a Gaussian distribution while the areas colored in red were obtained by modeling the regions through a Kernel Density Estimation (Parzen method, [15]). Purple areas are regions where both method overlap. Only the 8 most divergent regions extracted with each approximation are depicted. Notice that with the case of the Parzen approximation, 7 from the 8 peaks occur within the extreme regions defined by the method.

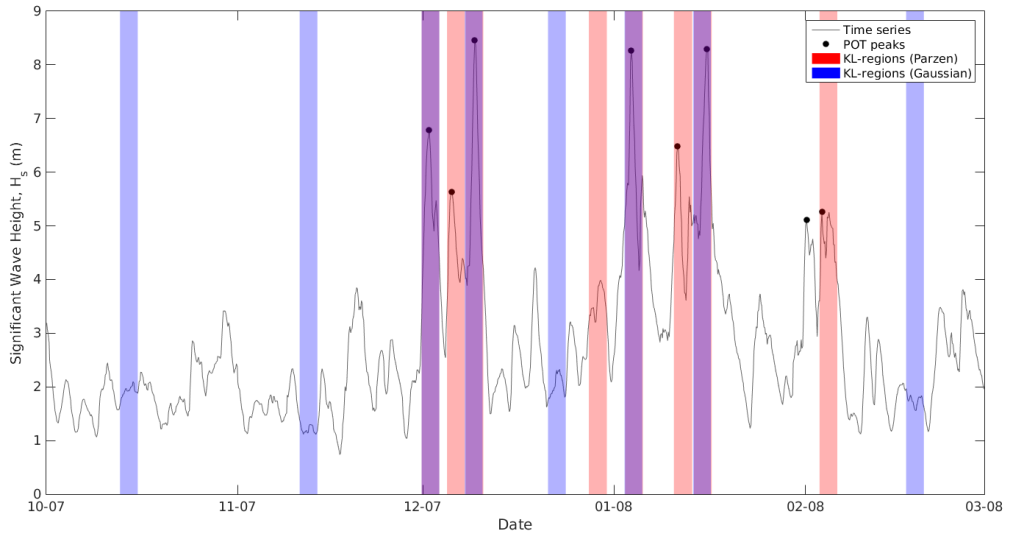

Figure 6: Results obtained with the KL-divergence method compared to a POT.

Initially, the Parzen approximation seems to be much more appropriated to accurately represent the data. But the Gaussian approximation might be useful when dealing with very long time series because is less computationally demanding. This is the

case of the data used in these tests, the temporal coverage of the wave record expands from 1957 until 2007, 51 years which makes a dataset of about 150.000 observations (once they have been 3-hourly aggregated). Applying the Parzen method to this temporal series was unfeasible, but it could be done with the Gaussian approximation, Figure 7.

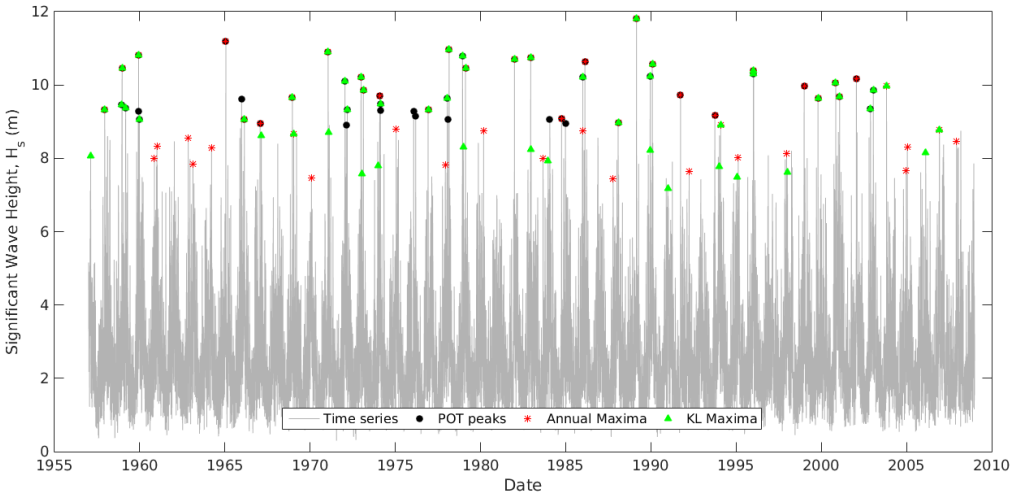

Figure 7: Results obtained in a long term time series.

The pale grey line represents the time series of wave height. From this time series, the 51 annual maximum were extracted (red asterisks). In addition, a POT was applied in such a way that the same number of peaks (black dots) were extracted: i.e. 3 days of independence lag and a threshold of 8.91 meters. Finally, the Kullback-Leibler method was used to extract the 51 most divergent regions with the Gaussian approximation. Regions of 24 observations were detected and the maximum within these regions was extracted (green triangles).

The KL method developed performs satisfactory and is able to detect a representative number of extremes. It performs better than the Annual Maximum compared with the POT, detecting 33 of the 51 same extremes when the Annual Maximum and the POT only coincides in 29 of the 51 events.

Compared with the Annual Maxima the Kullback-Leibler method presents the advantage of not being restricted to the annual constraint, so if there are two extreme events within the same year they can be detected. Additionally, when compared to the POT it has the advantage of no need for a threshold definition: the most divergent regions are detected and decreasingly sorted so instead of defining a threshold we only need to select the number of regions we want to extract from the time series.

In addition, the Kullback-Leibler method proposed here provides more information than the other two. Not only peaks are detected but abnormal regions were the distribution is significantly different from the rest of the time series. In this example it has not been explored but the method allows searching and optimizing the size of the regions detected. This could be useful in many cases were the variables involved are not so

related to physical processes like waves. Or even working with wave data, in locations were the combination of locally and regionally generated waves (sea and swell components) is important.

3.5.3 Two-dimensional application

Despite being the most important variable to define marine climate, the wave height alone may not be sufficient to fully characterize the wave conditions. As a minimum, the mean period associated to the wave height is required in most of the cases.

We have tested the Kullback-Leibler divergence method to the bivariate scenario (H_s and T) of the winter of 2007-2008. We have applied the method with the Parzen approach and extracted the 10 most divergent regions. The size of the regions, as in the 1-D case, was fixed to 24 timesteps.

In Figure 8 the results obtained are depicted. The left plot shows the two time series (H_s and T) with the 10 regions while the left graph represents the bivariate scatter plot and the colored circles mark the 10 regions. As it can be seen from both plots, the method is able to detect groups of data that define an abnormal event. These events might be either in the upper right tail of the joint distribution (maximum events: i.e. big waves and/or long periods) or in the lower left tail (minimum events: i.e. calm sea conditions with almost no waves).

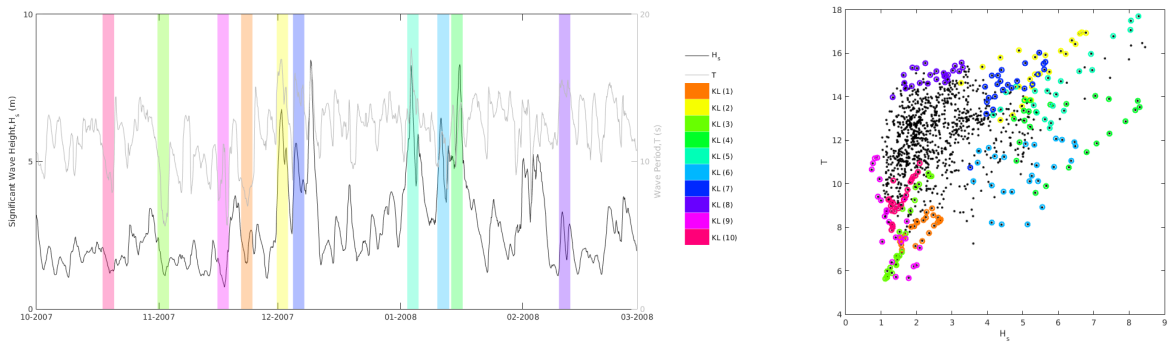

Figure 8: Results in a bivariate case (H_s and T).

3.5.4 Higher dimensions

The way the method has been exposed and developed allows its application to high-dimensional problems. For the sake of a simpler graphical representation and intuitive understanding we have not included in this report examples with more variables. In the methods comparison developed within this Task 5.1 (see Section 1), applied to an artificial dataset (to be later extrapolated to real data) this method was successfully used with 10 variables.

3.6 Conclusions

Within this last part of the report a newly developed method to detect extreme events in multivariate time series has been presented and explained in detail. The methodology proposed is based on the use of Kullback-Leibler divergence to find regions of the time series where the difference between this region and the rest of the time series is maximum. These events where the divergence is relatively large are assumed to be extreme events.

Applications of the method to real data in both univariate and bivariate cases were showed. The method has been also used and compared with other methods by using artificial data as it was explained in the Section 1 of this report.

This promising method will be more tested and explored in more real environmental variables databases from the project.

Conclusions

This report refers to the works done related to the *Task 5.1 - Time-series extension null space and GP methods and combination with spatial clustering* within the *Work Package 5 - Synthetic Index and Attribution Scheme: the BACIndex*.

The works done can be divided in two main parts:

In a first part a comparison between different methods and techniques to detect extreme events was done. This comparison was done using a dataset of artificially created variables were the extremes were known. By doing so, the comparison was fully controlled and gave the opportunity to determine which methods performed better depending on the kind of extreme events analyzed.

Within the methods compared, a newly developed method was introduced. This method is based in the maximization of the Kullback-Leibler divergence to detect regions of the time series were the distribution is notably different to the rest of the time series. This new method performs successfully in different kind of real and artificial dataset.

The work presented in this deliverable represents the successful accomplishment of the duties enclosed in Task 5.1.

References

[1] A. Basu, I:R: Harris, N.L. Hjort, and M.C. Jones. Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, 85:549–559, 1998.

[2] P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler. Kernel null space methods for novelty detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3374–3381, 2013.

[3] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the eighth SIAM International Conference on Data Mining*, pages 243–254, 2008.

[4] P. Camus, F.J Mendez, and R. Medina. A hybrid efficient method to downscale wave climate to coastal areas. *Coastal Engineering*, 53:1–25, 2011.

[5] Zhiqiang Ge and Zhihuan Song. *Multivariate Statistical Process Control*. Advances in Industrial Control. Springer, Springer London Dordrecht Heidelberg New York, 1 edition, November 2013.

[6] Ruey-Shiang Guh and Yeou-Ren Shiue. An effective application of decision tree learning for on-line detection of mean shifts in multivariate control charts. *Computers & Industrial Engineering*, 55(2):475–493, September 2008.

[7] Stefan Harmeling, Guido Dornhege, David Tax, Frank Meinecke, and Klaus-Robert Müller. From outliers to prototypes: Ordering data. *Neurocomputing*, 69(13-15):1608–1618, August 2006.

[8] H. Hotelling. *Multivariate Quality Control*. Techniques of Statistical Analysis. McGraw-Hill, New York, 1947.

[9] M. Körner and J. Denzler. Temporal self-similarity for appearance-based action recognition in multiview setups. *Computer Analysis of Images and Patterns*, pages 163–171, 2013.

[10] Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

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

[12] C A Lowry and W H Woodall. A Multivariate Exponentially Weighted Moving Average Control Chart. *Technometrics*, pages 1–9, May 1992.

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

[14] F.J. Mendez, M. Menendez, A. Luceño, and I. J. Losada. Estimation of the long-term variability of extreme significant wave height using a timer-dependent Peak Over Threshold (POT) model. *Journal of Geophysical Research*, 111, 2006.

[15] Emanuel Parzen. On Estimation of a Porbability Density Function and Mode. *The Annals of Mathematical Statistics*, 33:1–1065–1076, September 1962.

[16] H. Poincare. Sur le probleme des trois corps et les equations de la dynamique. *Acta Mathematica*, 13, 1890.

[17] Koen Smets, Brigitte Verdonk, and Elsa M Jordaan. Discovering Novelty in Spatio/Temporal Data Using One-Class Support Vector Machines. *Proceeding of International Joint Conference on Neural Networks*, pages 2956–2963, July 2009.

[18] S.Patra, Y. Maji, A. Basu, and L. Pardo. The power divergence and the density power diergence families: the mathematical connection. *The Indian Journal of Statistics*, 75:16–28, 2013.

[19] David M J Tax and Robert P W Duin. Support vector domain description. *Pattern Recognition Letters*, 20:1191–1199, November 1999.

[20] J Zscheischler, Markus Reichstein, S Harmeling, A Rammig, E Tomelleri, and Miguel D Mahecha. Extreme events in gross primary production: a characterization across continents. *Biogeosciences*, 11(11):2909–2924, 2014.

[21] Jakob Zscheischler, Miguel D Mahecha, Jannis von Buttlar, Stefan Harmeling, Martin Jung, Anja Rammig, James T Randerson, Bernhard Schölkopf, Sonia I Seneviratne, Enrico Tomelleri, Sönke Zaehle, and Markus Reichstein. A few extreme events dominate global interannual variability in gross primary production. *Environ. Res. Lett.*, 9(3):035001, March 2014.