



**HAL**  
open science

# On the ranking of variable length discords through a hybrid outlier detection approach

Hussein El Khansa, Carmen Gervet, Audrey Brouillet

► **To cite this version:**

Hussein El Khansa, Carmen Gervet, Audrey Brouillet. On the ranking of variable length discords through a hybrid outlier detection approach. 2022. hal-03838246

**HAL Id: hal-03838246**

**<https://hal.science/hal-03838246>**

Submitted on 3 Nov 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# On the ranking of variable length discords through a hybrid outlier detection approach

Hussein El Khansa<sup>1</sup>[0000-0001-8477-4887], Carmen Gervet<sup>1</sup>[0000-0002-8062-2808],  
and Audrey Brouillet<sup>1</sup>[0000-0001-9466-0197]

Espace-Dev, Univ. Montpellier, IRD, U.Guyane, U.Reunion, Montpellier, France

**Abstract.** In this paper we are interested in identifying insightful changes in climate observations series, through outlier detection techniques. Discords are outliers that cover a certain length instead of being a single point in the time series. The choice of the length can be critical, leading to works on computing variable length discords. This increases the number of discords, with potential overlapping, subsumption and reduced insightful results. In this work we introduce a hybrid approach to rank variable length discords and extract the most prominent ones, that can yield more impactful results. We propose a ranking function over extracted variable length discords that accounts for contained point anomalies. We investigate the combination of pattern wise anomaly detection, through the Matrix Profile paradigm, with two different point wise anomaly detectors. We experimented with MAD and PROPHET algorithms based on different concepts to extract point anomalies. We tested our approach on climate observations, representing monthly runoff time series between 1902 and 2005 over the West African region. Experimental results indicate that PROPHET combined with the Matrix Profile method, yields more qualitative rankings, through an extraction of higher values of extreme events within the variable length discords.

**Keywords:** Matrix profile · Prominent discord discovery · point and patten outlier detection · time series · Climate data

## 1 Introduction

Outlier detection is a field of machine learning to analyze data, covering numerous application domains including fraud detection, insurance [9], medical [30], internet of things[22], cyber security and hydrology [23]. Overall, an outlier can be generally defined as an observation that is significantly dissimilar to other data observation or an observation that does not behave like the expected typical behavior of the other observations. An outlier detection method can be applied to multiple data types such as images, transaction data, sequence data including genomics and time series. There are two main types of anomaly detection, point wise and pattern also called collective anomalies or discords[3]. Discords denote the most unusual time series sub sequences, and are detected using similarity measures that compare sub sequences with each other for a

given length. Existing approaches can be either approximate (e.g. HOT-SAX [11], QUICK MOTIF [15], Rare Rule Anomaly[21]) or exact like the matrix profile (MP) paradigm [27], used to detect motifs or discord. The choice of the sub sequence length is critical, set as input parameter and often specified with experts' knowledge which clearly influences the outcome. This observation lead more recently to works on computing variable length discords, over a given length interval (e.g. GraphAn [2], PanMatrix [17]). The increased usage of these approaches has also raised their weaknesses. Indeed, the number of extracted discords is larger, with potential overlapping and subsumption among them, leading to the question of "how can we select the more relevant or actionable ones?". We are not aware of approaches seeking a ranking of variable lengths discords to this date.

In this work, we address this issue in the field of climate data analysis [12]. Climate impact models provide time series based both on historical data and projected data generated from complex simulations of physical processes. The impact models concern many fields, including yields of agricultural crops, biodiversity, runoff of water infiltration. Such models are commonly analyzed using multi-model ensembles means to mitigate uncertainties. With a primary goal of estimating future trends, the use of outlier detection methods has seen limited interest in this field to this date, with some works on hydrological impact models [28,26,28]. Another particularity of these models is the very large scale of the spatio-temporal data series, covering daily or monthly data over 150 years for each spatial point in a given region.

Our objective is to investigate a ranking approach over variable length discords, by integrating two unsupervised anomaly detection frameworks, point anomaly detectors and pattern based anomaly detection using MP. The intuition behind our work, is that a discord will dominate another one, if proportionally to its length, it also contains anomalous points that are greater in numbers and extreme values within the discord sequence. In our work, we compare two types of point outlier detectors, model and predictor based and draw conclusions on their insights to rank variable length discords. We consider respectively the MAD and PROPHET methods. We also define and specify a ranking function over variable length discords. MAD is a general point anomaly detection method, based of how far a point is from the median, while PROPHET is more tailored to time series and detects anomaly points that diverge from the expected values with respect to the projected seasonality and trends derived in the learning phase of PROPHET. Our approach is evaluated on total (surface + subsurface) runoff observations, that reflect the soil water levels.

Our main contributions are: 1) a hybrid approach to rank variable length discords, through the combination of exact discord method with point detectors and a ranking function; and 2) the evaluation through an experimental study over large scale climate impact data, showing the differences among point anomalous detectors and their added value as an effective means to rank variable discords.

This paper is organized as follows. Section 2 gives a background on the concepts and methods we use, section 3 presents our approach. We conclude in section 4.

## 2 Background

In this section we review existing concepts and approaches we make use of, relative to outlier detection in time series, and review the use of point outliers to extract pattern outliers.

Point outliers over time series can be specified as global outliers when compared to all values in the time series, and local outliers when compared to its neighboring points. Subsequence or pattern outliers in time series, extract collective consecutive points that behave unusually compared to other subsequences of similar length. In such cases a single point in the sequence might or not be also an anomaly on its own.

Methods for point outlier detection can be classified as model outlier detectors or predictor based outlier detectors. Model or statistical outlier detectors derive a probabilistic model which captures the distribution of the time series. If an instance in the time series has a low probability of belonging to the estimated model, it is flagged as an outlier[4]. Examples of model based point outlier detectors are Median Absolute Deviation (MAD) [10] and Minimum Covariance Determinant (MCD) [19].

Predictor based outlier detectors proceed differently. They construct a predictor (like forecast model or a regressor) to learn the normal behavior of the time series, and then exploits the learned information to predict future values. Prediction errors which are predictions that significantly deviate from the true value are labeled as point outliers [14]. Many predictor methods have been developed including ARIMA[20], and SVM[16] and PROPHET [24].

In this work, we will apply one point outlier from each type to extract point outliers within computed variable length discords, as a means to rank those discords. Note that point outliers have also been used to directly segment data series and thus extract discords. We briefly survey these approaches and their suitability for our goal.

### 2.1 Point outliers detectors and series segmentation

In [6] the authors proposed IForest-based anomaly detection to detect drift in streaming data, where a fixed sliding window is applied, and in each window the IForest detector is applied to generate the anomaly rate by averaging the instance depth of the forest, if the anomaly rate is smaller than the threshold then there is a drift. Also, in [5] the authors proposed sliding-window convolutional variational autoencoder (SWCVAE) an algorithm to detect point anomalies in robot, by applying sliding windows. Each window is fed to a convolutional variational autoencoder, a trained artificial neural network(ANN), to detect point wise anomaly. The ANN will give each instance in the sliding window an anomaly score, then these scores are averaged to calculate the anomaly score of the discord. If an anomaly score is above the predefined threshold then it is classified as a discord. In [29] the author combines both the result of SAX as a pattern outlier detection and combined with point outlier detection second-order-difference (SOD) and Chebyshev Inequality (CI)-based methods, where discord that does

not contain point anomaly detected by SOD and CI are discarded. Those approaches contribute both ranking mechanisms of discords through a common concept, that is the segmentation of the time series and the use of a point outlier detector on each subsequence to give it an anomaly score.

With respect to our goal of comparing variable length discords over large time series, these approaches offer scoring mechanisms for subsequences, but are applied to fixed size windows, and go through the whole data set. This requires the application of outlier detectors over very large sets of sliding windows, without considering variable length windows.

In our approach, we purposely focus on variable length discords, and separate the issue of computing variable length discords, through an exact dedicated approach, with that of exploiting point anomaly detectors for ranking purpose. Part of our goal is to analyse the contribution and added value of point anomaly detectors to rank the discords of variable lengths.

## 2.2 Matrix Profile, MAD and PROPHET

In this subsection we provide the necessary background on the three methods used in our work, namely the STOMP algorithm to compute Matrix Profiles and extract discords, and the point anomaly detectors MAD and PROPHET.

*Matrix Profile methods* Matrix Profile is a powerful data structure that stores the z-normalized Euclidean distances between each subsequence and its closest neighbor sequence. Two main algorithms have been defined to compute the exact solution of the matrix profile by performing all-pairs-similarity-search on time series, namely STAMP[27] and STOMP [31]. We adopted STOMP for its runtime and scalability over large time series. STOMP computes the matrix profile by calculating the distance profile of every subsequent in time series  $T$  and then selecting the minimum value in each distance profile. It uses a similarity search algorithm to compute the matrix profile faster in time complexity of  $O(n^2)$ . The produced Matrix can be used to detect discords, where the highest  $K$  subsequences in the Matrix Profile profile are the top  $K$  discords, specifying the highest distance of these subsequences with its closest neighbor. The concepts used in this work are recalled here after, relative to discords and the Matrix Profile data structure.

**Definition 1.** *A time series  $T$  is a sequence of real-valued numbers  $t_i : T = [t_1, t_2, \dots, t_n]$  where  $n$  is the length of  $T$ .*

**Definition 2 (Matrix Profile).** *A matrix profile  $P_m$  of time series  $T$  and given length  $m$  is a meta series of the Euclidean distances vector between each subsequence  $T_{i,m}$  of given length  $m$  where  $i$  varies, and its nearest neighbor (closest match) in time series  $T$ , together with the corresponding position vector for each closest neighbor associated with  $\min(D_{i,m})$ . We denote it  $P_m = [\min(D_{1,m}), \dots, \min(D_{n-m+1,m})]$ , where  $D_{i,m} (1 \leq i \leq n - m + 1)$  is the distance profile  $D_{i,m}$  of time series  $T$  for subsequences of length  $m$ .*

**Definition 3 (Discord).** *The discord denoted  $\Delta_{j,m}$  is a subsequence  $T_{j,m}$  of length  $m$  starting at the position  $j$ , that has the largest distance to its nearest neighbor. In matrix profile the largest distance corresponds to the maximum distance value in  $P_m$ .*

*MAD* The Median Absolute Deviation outlier detection method is a point outlier detector that uses the Absolute deviation around the median to measure the distance between a data point and the median[10]. The median is considered a tendency measurement of the data similar to the mean, but unlike the mean it is more robust to detect relevant point outliers in the data, being immune to noise, and to the sample size [13].

Given a time series  $T = x_1, x_1, \dots, x_i, \dots, x_n$  such that  $i$  is the time unit, MAD first calculates the median of  $T$ , this median is denoted by  $\tilde{x}$ , and its absolute deviation, by subtracting  $\tilde{x}$  from each  $x$  in  $T$ :

$$MAD = median(|x_i - \tilde{x}|) \quad (1)$$

For each  $x_i$  we calculate their modified z-score using the predefined formula [25]:

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD} \quad (2)$$

In statistics, the z-score for each  $x_i$  measures the difference between the value and the mean how many standard deviations away a value is from the mean in terms of standard deviation . MAD uses the following formula to calculate a z-score:

$$Z - Score = \frac{x_i - \mu}{\sigma} \quad (3)$$

where:  $\mu$  is the mean value of the data set,  $\sigma$  is the standard deviation of the data set where if the Z-score is 0 then it means the value of  $x_i$  is equal to the mean, if z-score is 2 then it is 2 standard deviation ( $2 * \sigma$ ) away from the mean. If the absolute value of the modified z-core of  $x_i$  is larger than a threshold then is considered as an outlier. [10] proposed the value 3.5 as a threshold, also the default one in MAD.

*PROPHET* It is an open source time series forecasting model, that detects outliers based on the data learned through the forecasting component. Its purpose is to forecast future data, and detect outliers, with an uncertainty estimation. Once PROPHET fits the given time series, it generates a data set containing the following variables with an estimation of the uncertainty through a confidence interval for each value. Then for each value in the Time series, it calculates the error and uncertainty based on the confidence interval of each value. If the error of a value is bigger than its uncertainty it will be flagged as an anomaly. PROPHET is based on generalized additive model (GAM), which means that the model is the sum of several components summarized below:

$$y_t = g(t) + s(t) + h(t) + \varepsilon(t) \quad (4)$$

Where:

- $g(t)$  denotes the growth function, a piece-wise linear function. The number of linear growths (straight lines), is determined by the number of change points in the time series, that can be entered manually or detected automatically by PROPHET.

$$g(t) = (k + \mathbf{a}(t)^\top \boldsymbol{\delta}) t + (m + \mathbf{a}(t)^\top \boldsymbol{\gamma}) \quad (5)$$

where  $k$  is the growth rate scalar value,  $m$  the offset value,  $a(t)$  a binary vector  $\in \{1, 0\}$  indicating if there is a growth rate adjustments at time  $t$ .  $\boldsymbol{\delta}$  is a vector that contains the growth rate value. For instance if there is 10 change point then the vector of  $\boldsymbol{\delta}$  will have 10 values.

- $\boldsymbol{\gamma}$  is vector similar too that contains the offset parameter adjustment
- $s(t)$  the seasonality, modeled using Fourier Series as a function of time.
- $h(t)$  holiday seasons parameter, that we will not consider
- $\varepsilon(t)$  the error estimation using a maximum a posteriori estimation.

Not all components are mandatory, and in our case since we will be experimenting on climate impact data.  $h(t)$  not considered for instance.

### 3 Our approach

Our work is first motivated by the insightful ranking of variable length discords, to extract those that are combined with more eventful changes and potential impact on forecasted data trends. We do so by combining an exact, robust pattern outlier detection approach, with two different point outlier detectors. The exact approach ensures reliable extraction of discords, while the point outlier detectors will contribute to their ranking. We use the Matrix Profile method, STOMP as an exact approach to extract discords, that also offers the qualities of being parameter-free and that does not utilize a similarity or distance threshold. This is important since the point outlier detectors do not have these advantages, and an increase in thresholds tuning would reduce the robustness of the overall results. With respect to the point outlier detectors, we experimented with the generic MAD model approach, and the predictor PROPHET method that also handles trends and seasonality.

Our Algorithm 1 is based on four 4 main phases :

1. (Lines 8-12) Extract variable length discords (Matrix Profile method)
2. (Line 13) Detect point outliers (using PROPHET, MAD) over the time series
3. (Lines 14-20) Score the discords based on the anomalous points they contain
4. (Lines 21-25) Filter and sort the discords with a ranking function

*Extract variable lengths discords.* We compute the Matrix Profile for variable window sizes  $j$  over  $T$ , where  $4 < j < \text{length}(T)/2$ . For each window size, we extract the top discord. The output is a list of top discords, one per length.

*Detect point outliers in the time series.* This step seeks point wise outliers (through PROPHET and MAD), over the whole time series  $T$ . The outcome is a Boolean list denoting for each data in the time series if it is an anomaly.

*Scoring the discords.* This step integrates the variable length discords and Boolean anomaly list for scoring purpose. For each discord we count the anomalous points that are included in each discord  $\Delta_{i,j}$ . Clearly a simple count is not the only element in the scoring, and the discords' length needs to be taken into account. Note that some discords can overlap, thus share anomalous points. To address this issue we define a scoring function as the ratio of the count over each length.

*Filtering and sorting.* This procedure filters the discords based on their count, as well as their starting dates, and length. Based on our past work and analysis [7] of subsumption among variable length discords, it is important to compare variable length discords with identical starting dates. The filtering stage groups discords sharing a starting date  $i$  and for each group selects the discord with highest ratio, meaning it subsumes the others. Then the discords (of variable length) are sorted in descending order of their ratio value.

---

**Algorithm 1:** matrix prophet ensemble

---

```

input : Time Series  $T$ 
output: list <Discords> SortedList
1 initialization
2 int  $m = \text{length}(T)/2$ 
3 list  $\langle \Delta \rangle$  top_discord_list=[ ] // List to store top discords
4 list  $\langle \Delta \rangle$  group_list=[ ] // List to group discords with similar date
5 list  $\langle \text{bool} \rangle$  anomaly_list=[ ] // Boolean list of anomaly tag per value
6 list  $\langle \Delta \rangle$  sorted_list=[ ] // List of ranked discords
7 list  $\langle \text{int} \rangle$  Windows = [4, 5, 6, 7, 8, ..., m]; // List of Window sizes
8 foreach  $j$  in Windows do
9    $P_j \leftarrow \text{STOMP}(T, j)$  // Matrix profile for window size  $j$ 
10   $\Delta_{i,j} \leftarrow \max(P_j)$  // Discord of length  $j$ 
11  top_discord_list  $\leftarrow$  top_discord_list.add( $\Delta_{i,j}$ )
12 end
13 anomaly_list  $\leftarrow$  PROPHET or MAD( $T$ ) // Store output of detector
14 foreach  $\Delta_{i,j}$  in top_discord_list do
15   start= $\Delta_{i,j}.i$  // Store the start date of discord
16   end= $\Delta_{i,j}.i + \Delta_{i,j}.j$  // Store the end date of the discord
17   length= $\Delta_{i,j}.j$  //Store the discord length
18   int count $\leftarrow$  count_Anomaly( $T$ ,start,end)//Anomaly count per  $\Delta_{i,j}$ 
19    $\Delta_{i,j}.\text{ratio} \leftarrow$  count/length
20 end
21 group_list  $\leftarrow$  groupBy(top_discord_list) // Group discord list per similar
   starting date
22 foreach group in group_list do
23   sorted_list  $\leftarrow$  add(max(group) ) // Get discord with highest ratio in
   each group
24 end
25 sorted_list  $\leftarrow$  Quicksort(sorted_list, $\Delta.\text{ratio}$ )
26 return sorted_list

```

---



### 3.1 Experimental comparison of MAD and PROPHET

*Climate data.* In our experimental study we used observed monthly runoff data obtained from Global Runoff Reconstruction dataset (GRUN) [8]. In climate and corresponding impact data science, runoff is an impact variable that can be used to quantify flood and drought risks at regional and global scales [1]. It is usually provided in  $\text{kg}/\text{m}^2/\text{s}$  or in  $\text{mm}/\text{day}$ . The source of the data has a  $0.5^\circ \times 0.5^\circ$  spatial grid resolution with a focus on the Sahel region. This data gives indicators on the soil water content over the time period between 1902 and 2005 (i.e. 104 years, 1248 months). This is a standard observation period in historical climate analysis. We spatially average the monthly runoff data over the grid box  $[5^\circ\text{W}-25^\circ\text{E} ; 10^\circ\text{N}-18^\circ\text{N}]$ .

*Point anomaly detectors.* A first comparative study between MAD and PROPHET on the runoff data set is shown in Figure 1. Those point anomaly detectors use different methods to extract anomalies, and we can see that some anomalous points are shared and many differ. Over the whole data series, MAD extracts a large set of anomalies (green), essentially in the upper part of the data, thus detecting high water levels. On the other hand, PROPHET extracts fewer anomalies, some shared with MAD for the highest values of anomalous water levels, and most importantly detected the drought periods in terms of anomalous points (purple crosses), in 1900-1910 and 1970-1990. This can be explained by the fact that PROPHET through the learning phase was able to identify a normality within seasons, whereas MAD focuses on extremes in the overall data series.

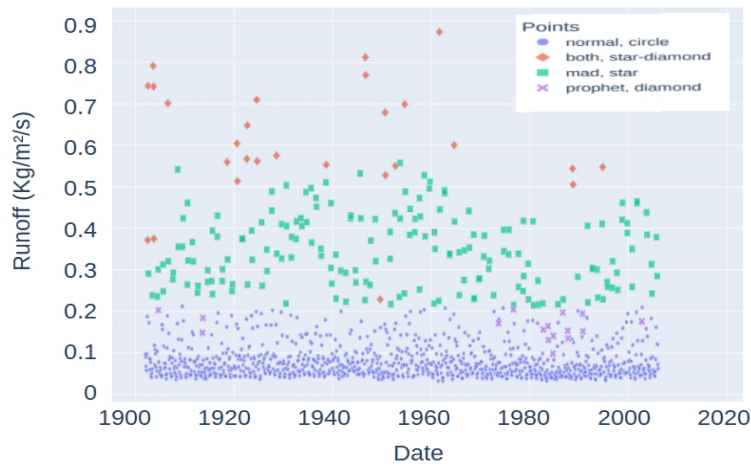


Fig. 1: Anomaly points extracted by MAD and PROPHET: similarities and differences

If we zoom on the period 1981-1984 where 1983 is recorded as one of the driest years in the West African region, this aspect is striking. MAD does not extract any anomalous point, whereas PROPHET detects three anomalies displayed in Figure 2. By not accounting for, or detecting, seasonality MAD misses important point outliers in such time series. On the other hand, PROPHET takes into account seasonality through the learning phase, and we can see that it extracts anomalies with respect to months which are learnt to be dry and those that should not be. We note that the anomalous points do not necessarily reflect extreme values overall, but anomalous points with respect to the anticipated values (learnt during the forecasting phase).

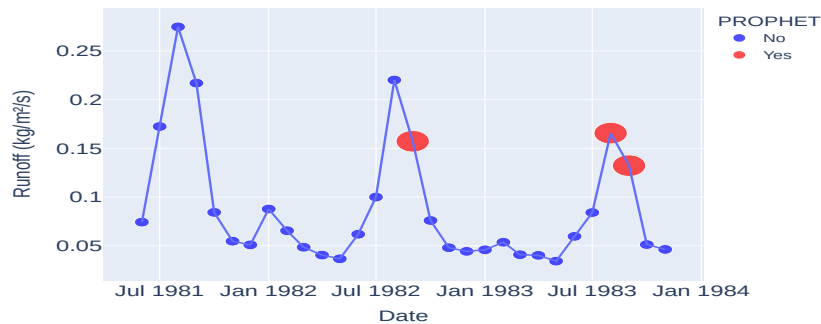


Fig. 2: Anomalous points detected by PROPHET in the subsequence over 1981-1984

We also studied the early period 1902-1905, a period of higher runoff values in average. We can see that both MAD and PROPHET detect anomalies during this period, with some common points, showcased in Figures 3 and 4. Those figures are interesting as they also show how MAD does detect some anomalous points not tagged as such by PROPHET. This reflects the role played by the median function and the z-score defined by MAD.

All experiments were run on an Intel(R) Xeon(R) Bronze 3106 CPU processor at 1.70GHz with 8 core with 64 GB of RAM. The main computational cost are the  $N/2$  runs of STOMP to compute the variable length discords, taking around 9 minutes. MAD is in constant time, and PROPHET took 1.5 minutes.

### 3.2 Integrated approach: ranking and outcomes

Having compared the commonalities and differences between two point anomaly detectors, we saw that even though they are generic, they can have a great impact on the outliers' detection. This is related to their models and potentially the nature of the data. We now present the results of our methodology. Both

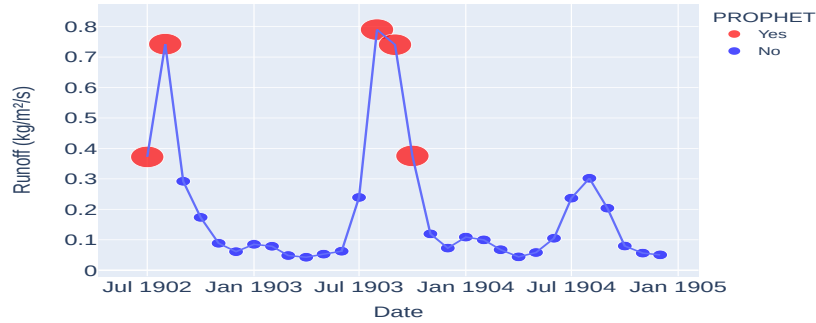


Fig. 3: Anomalous points from PROPHEET in the subsequence 07/1902-01/1905

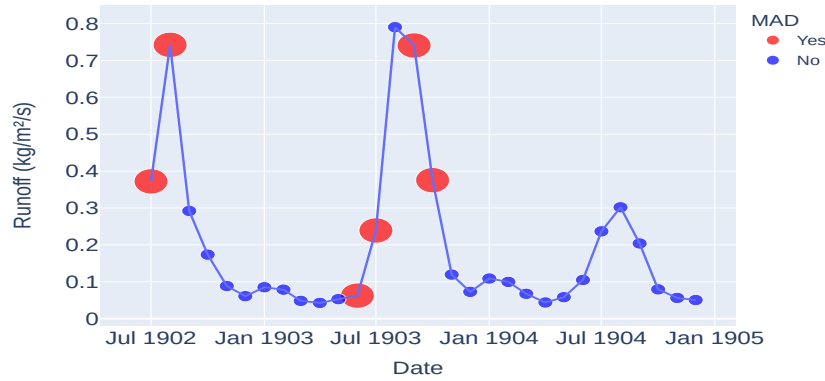


Fig. 4: Anomalous points detected by MAD in the subsequence 07/1902-01/1905

detectors are applied and compared as a means to rank the variable lengths discords produced.

Figure 5 represents respectively for PROPHEET and MAD, the top discords ranking with respect to our ratio function. Note that the discords of variable lengths are specified by their starting date, window length and the value of the ratio function. Note that since we are comparing discords of variable lengths, at this stage we only extract the top discord for each given length  $m$ , to reduce the combinatorics of our approach. Extracting top  $k$  discords per length is part of future work. We recall that the ratio values depend on the length and the number of anomalous points within each discord. In this figure we picked the top 20 discords relative to each detector output ranking. The discords highlighted in yellow, correspond to a filtering, that extracts among those discords the top ones outside a proximity temporal interval of  $\pm$  three months. For instance for the year 1982 we have 3 ranked discords starting from July till September with

overlapping periods. We kept the one with highest ranking to reduce redundant discord periods. Thus the colored row in the Table 5 represent the top 5 discords resulting from the filtering procedure.

PROPHET				MAD			
date	window	ratio	points	date	window	ratio	points
1983-08-01	97	0.12	12	1904-06-01	14	0.29	4
1982-09-01	108	0.12	13	1904-05-01	15	0.27	4
1982-08-01	121	0.11	13	1902-03-01	43	0.26	11
1982-07-01	122	0.11	13	1904-04-01	16	0.25	4
1902-07-01	62	0.1	6	1904-03-01	32	0.25	8
1902-06-01	63	0.1	6	1904-08-01	24	0.25	6
1902-05-01	64	0.09	6	1902-02-01	93	0.25	23
1902-04-01	65	0.09	6	1902-01-01	107	0.23	25
1902-03-01	66	0.09	6	1903-09-01	35	0.23	8
1902-02-01	67	0.09	6	1927-04-01	623	0.22	140
1984-08-01	123	0.09	11	1927-09-01	617	0.22	138
1902-01-01	68	0.09	6	1912-07-01	591	0.22	132
1984-07-01	125	0.09	11	1927-12-01	614	0.22	137
1984-06-01	129	0.09	11	1912-05-01	592	0.22	132
1984-05-01	130	0.08	11	1927-11-01	615	0.22	137
1984-04-01	131	0.08	11	1912-08-01	589	0.22	131
1983-09-01	143	0.08	12	1927-10-01	616	0.22	137
1903-10-01	12	0.08	1	1904-02-01	18	0.22	4
1983-07-01	158	0.08	13	1903-10-01	36	0.22	8

Fig. 5: Ranking table of top 20 discords

*Analysis of the results:*

1. *Point anomalous detectors do matter.* While STOMP produces exact matrix profiles to derive variable length discords, the rankings differ between PROPHET and MAD. By detecting the anomalies in the driest period, where droughts occurred (1982-1984), PROPHET manages to rank top discords in this period. We can see that the variable length are not negligible ranging in this period from 97 months to 123 months. MAD detects mainly the periods with high water levels (1902-1904), yielding high ranking for the corresponding discords. It corresponds to the periods with higher soil water levels and larger inter-annual variability, that comes before the later continuous long-term drying trends observed within Sahel [7]. PROPHET extracted and ranked the 1<sup>st</sup>, 2<sup>nd</sup> and 11<sup>th</sup> discords, that represent the intense droughts

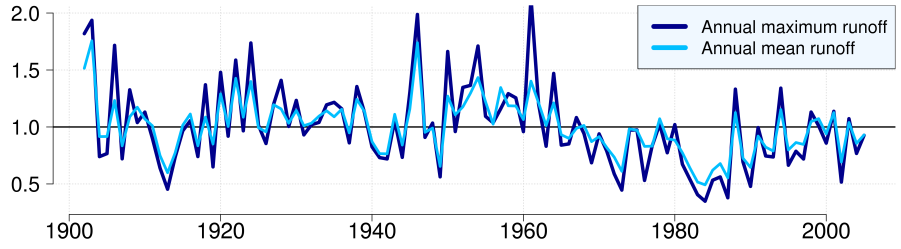
that occurred in Sahel, while the 3<sup>rd</sup> and 19<sup>th</sup> discords represent the the period where the high soil water levels and larger inter-annual variability.

2. *Rankings are insightful.* We can see in the table that the discords vary greatly in length and also that the number of anomalous points extracted in those are not proportional to the length. Thus the integration of exact variable length discords with different anomalous point detectors does bring complementary information and the scoring function is adequate. This information was able to extract the prominent discords over the runoff data that reflected both the driest periods, and periods with highest soil water levels.
3. *Climate-related impact insights.* Those results are compared to basic statistical analysis of runoff data to demonstrate the thematic insights brought by combining anomalous point detection and discord methods (Figure 6). The first rank of combined PROPHET and discord methods detects an anomalous event/period from August 1983 for 97 months with 12 anomalous points (Table 5, Figure 2). Following discords 2, 3 and 4 exhibit larger time windows but include this first top discord. This is consistent with well known intense droughts that occurred in Sahel between 1982 and 1987 (Figure 6a, 6b), with the most severe drought ever recorded within the African continent in 1983-1984 (e.g. [18]). Moreover, following ranks of combined PROPHET and discords emphasize top discords starting in 1902 for at least 60 months (Table 5). The first ranked of combined MAD and discord methods also detects "anomalies" in 1904 for 14 months, with following ranked ones including this first top ranked (Table 5). These results coincide with large positive anomalies both in annual maximum and annual mean in 1902 and 1903, and with very large year-to-year variability from 1902 to 1907 (Figure 6a, 6c). These early 1900s result also illustrate the well known soil drying trend resulting from a rainfall deficit observed between 1900 and 2013 within the region ([8]). PROPHET detects both 1980s dry events and 1904 abrupt year-to-year changes, and it is important to note that MAD does not detect any 1980s dry signal in the first top 20 ranked discords (Table 5).

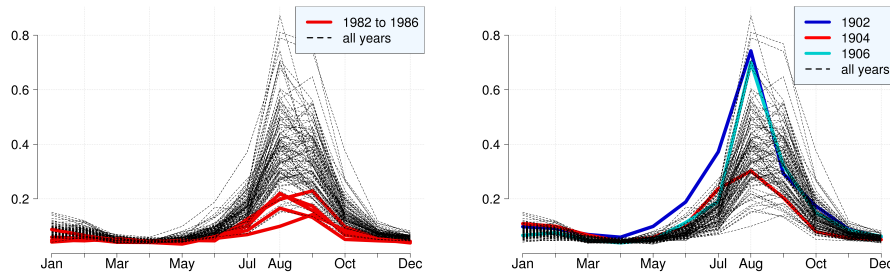
## 4 Conclusion and future work

In this paper, we proposed a hybrid approach to rank variable length discords that combine an exact pattern wise outlier approach with anomalous point detectors. Our methodology is generic. We chose the Matrix Profile as the pattern outlier detector because it is exact and parameter free, generated variable length discords, and used point wise anomaly detectors to give each discord a score. This score together with the discord length were used to define a ranking function.

We tested two different point anomaly detectors, a model based detector (MAD) and predictor based one (PROPHET). We run experiments on large scale Monthly runoff data of the Sahel region in West Africa from 1902 till 2005. Our results show that PROPHET is more adequate than MAD, as its top 5 ranked discords did capture the actual discords. The top 5 discords extracted with MAD were only able to reflect the periods with highest water level. They also show that



(a) Time series of normalized annual runoff



(b) Seasonal cycles

(c) Seasonal cycles

Fig. 6: **(a)** Time series of normalized annual maximum and mean runoff (relative to their respective mean climatologies over 1902-2005). **(b)**-**(c)** Seasonal cycles of absolute runoff values for each year between 1902 and 2005 (in mm/day).

PROPHET is suitable both for detecting specific events (i.e. droughts in 1982, 1983, 1984, etc) and larger timescale anomalies (i.e. dry decade of 1980s). MAD is suitable to detect abrupt inter-annual change (i.e. dry 1904/1905 relative to wet 1902/1903 and 1906 rainy seasons).

In summary, this work contributed a novel approach to exploit variable length discords and be able to rank them with respect to a complementary data analysis, point anomalous detectors. The ranking function confirmed pattern outliers for the climate impact data at hand.

Future work includes running more experimental studies on different impact climate data and other data sets. We also intend to extend our integration with alternative point anomaly detectors such as ARIMA, Random Forest and LSTM, towards a comprehensive approach, given the specificity and impact they have on the extracted point outliers.

*Acknowledgements* The authors would like to thank the Occitanie Region, who partially funded this research, and the reviewers for their comments and suggestions.

## References

1. Nigel W. Arnell and B. Lloyd-Hughes. The global-scale impacts of climate change on water resources and flooding under new climate and socio-economic scenarios. *Climatic Change*, 122(1-2):127–140, 2014.
2. Paul Boniol, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. Graphan: Graph-based subsequence anomaly detection. *Proceedings of the VLDB Endowment*, 13(12):2941–2944, 2020.
3. Heraldo Borges, Reza Akbarinia, and Florent Masseglia. Anomaly detection in time series. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems L*, pages 46–62. Springer, 2021.
4. Varun Chandola, Arindam Banerjee, and Vipin Kumar. Outlier detection: A survey. *ACM Computing Surveys*, 14:15, 2007.
5. Tingting Chen, Xueping Liu, Bizhong Xia, Wei Wang, and Yongzhi Lai. Unsupervised anomaly detection of industrial robots using sliding-window convolutional variational autoencoder. *IEEE Access*, 8:47072–47081, 2020.
6. Zhiguo Ding and Minrui Fei. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proceedings Volumes*, 46(20):12–17, 2013.
7. Hussein El Khansa, Carmen Gervet, and Audrey Brouillet. Prominent discord discovery with matrix profile : Application to climate data insights. In *Computer Science & Technology Trends*. Academy and Industry Research Collaboration Center (AIRCC), May 2022.
8. Gionata Ghiggi, Vincent Humphrey, Sonia I Seneviratne, and Lukas Gudmundsson. Grun: an observation-based global gridded runoff dataset from 1902 to 2014. *Earth System Science Data*, 11(4):1655–1674, 2019.
9. Anton Hansson and Hugo Cedervall. Insurance fraud detection using unsupervised sequential anomaly detection, 2022.
10. Boris Iglewicz and David Caster Hoaglin. *How to detect and handle outliers*, volume 16. Asq Press, 1993.
11. Eamonn Keogh, Jessica Lin, and Ada Fu. Hot sax: Efficiently finding the most unusual time series subsequence. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. Ieee, 2005.
12. Philomène Le Gall, Anne-Catherine Favre, Philippe Naveau, and Clémentine Prieur. Improved regional frequency analysis of rainfall data. *Weather and Climate Extremes*, 36:100456, 2022.
13. Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766, 2013.
14. Tianyu Li, Mary L Comer, Edward J Delp, Sundip R Desai, James L Mathieson, Richard H Foster, and Moses W Chan. Anomaly scoring for prediction-based anomaly detection in time series. In *2020 IEEE Aerospace Conference*, pages 1–7. IEEE, 2020.
15. Yuhong Li, Hou U Leong, Man Lung Yiu, and Zhiguo Gong. Quick-motif: An efficient and scalable framework for exact motif discovery. In *2015 IEEE 31st International Conference on Data Engineering*, pages 579–590. IEEE, 2015.
16. Junshui Ma and Simon Perkins. Time-series novelty detection using one-class support vector machines. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 3, pages 1741–1745. IEEE, 2003.

17. Frank Madrid, Shima Imani, Ryan Mercer, Zachary Zimmerman, Nader Shakibay, and Eamonn Keogh. Matrix profile xx: Finding and visualizing time series motifs of all lengths using the matrix profile. In *2019 IEEE International Conference on Big Knowledge (ICBK)*, pages 175–182. IEEE, 2019.
18. I. Masih, S. Maskey, F. E. F. Mussá, and P. Trambauer. A review of droughts on the african continent: a geospatial and long-term perspective. *Hydrology and Earth System Sciences*, 18(9):3635–3649, 2014.
19. Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.
20. Manassakan Sanayha and Peerapon Vateekul. Fault detection for circulating water pump using time series forecasting and outlier detection. In *2017 9th International Conference on Knowledge and Smart Technology (KST)*, pages 193–198. IEEE, 2017.
21. Pavel Senin, Jessica Lin, Xing Wang, Tim Oates, Sunil Gandhi, Arnold P Boedihardjo, Crystal Chen, and Susan Frankenstein. Time series anomaly discovery with grammar-based compression. In *Edbt*, pages 481–492, 2015.
22. Arnaldo Sgueglia, Andrea Di Sorbo, Corrado Aaron Visaggio, and Gerardo Canfora. A systematic literature review of iot time series anomaly detection solutions. *Future Generation Computer Systems*, 2022.
23. Ilias Siniosoglou, Panagiotis Radoglou-Grammatikis, Georgios Efstathopoulos, Panagiotis Fouliras, and Panagiotis Sarigiannidis. A unified deep learning anomaly detection and classification approach for smart grid environments. *IEEE Transactions on Network and Service Management*, 18(2):1137–1151, 2021.
24. Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.
25. Rand R Wilcox. *Fundamentals of modern statistical methods: Substantially improving power and accuracy*, volume 249. Springer, 2001.
26. Feng Ye, Zihao Liu, Qinghua Liu, and Zhijian Wang. Hydrologic time series anomaly detection based on flink. *Mathematical Problems in Engineering*, 2020, 2020.
27. Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 1317–1322. Ieee, 2016.
28. Yufeng Yu, Yuelong Zhu, Shijin Li, and Dingsheng Wan. Time series outlier detection based on sliding window prediction. *Mathematical problems in Engineering*, 2014, 2014.
29. Meng Yue. An integrated anomaly detection method for load forecasting data under cyberattacks. In *2017 IEEE Power & Energy Society General Meeting*, pages 1–5. IEEE, 2017.
30. Haibo Zhang, Wenping Guo, Shiqing Zhang, Hongsheng Lu, and Xiaoming Zhao. Unsupervised deep anomaly detection for medical images using an improved adversarial autoencoder. *Journal of Digital Imaging*, pages 1–9, 2022.
31. Yan Zhu, Zachary Zimmerman, Nader Shakibay Senobari, Chin-Chia Michael Yeh, Gareth Funning, Abdullah Mueen, Philip Brisk, and Eamonn Keogh. Matrix profile ii: Exploiting a novel algorithm and gpus to break the one hundred million barrier for time series motifs and joins. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 739–748. IEEE, 2016.