PROMINENT DISCORD DISCOVERY WITH MATRIX PROFILE: APPLICATION TO CLIMATE DATA INSIGHTS

Hussein El Khansa, Carmen Gervet, Audrey Brouillet

PROMINENT DISCORD DISCOVERY WITH MATRIX PROFILE : APPLICATION TO CLIMATE DATA INSIGHTS

Hussein El Khansa[1] [0000−0001−8477−4887], Carmen Gervet[1] [0000−0002−8062−2808], and Audrey Brouillet[1] [0000−0001−9466−0197]

[1]Espace-Dev, Univ. Montpellier, IRD, Univ. Guyane, Univ. La Réunion, Montpellier, France

Hussein.El-khansa@ird.fr, carmen.gervet@umontpellier.fr, audrey.brouillet@ird.fr

ABSTRACT

The definition and extraction of actionable anomalous discords, i.e. pattern outliers, is a challenging problem in data analysis. It raises the crucial issue of identifying criteria that would render a discord more insightful than another one. In this paper, we propose an approach to address this by introducing the concept of prominent discord. The core idea behind this new concept is to identify dependencies among discords of varying lengths. How can we identify a discord that would be prominent? We propose an ordering relation, that ranks discords and we seek a set of prominent discords with respect to this ordering. Our contributions are 1) a formal definition, ordering relation and methods to derive prominent discords based on Matrix Profile techniques, and 2) their evaluation over large contextual climate data, covering 110 years of monthly data. The approach is generic and its pertinence shown over historical climate data.

KEYWORDS

Prominent discord discovery, Large time series, Matrix profile, Climate data.

1. INTRODUCTION

The analysis of climate data towards the extraction of global climate trends using ensemble mean approaches is receiving a wide interest. To this date, the search for pattern anomaly or outlier patterns has received less attention in climate data studies. Such data is contextual due to the geo-localization and timestamps of the data series relative for instance to soil humidity, temperature or rainfall. It comes from historical sources or complex simulation impact models (physical processes of atmosphere and ocean) such as Earth System Models ([1], [2]). Existing approaches mainly focus on seeking long-term trends, and the study of abnormal behaviour tend to focus on the search of extreme values.

An important element in climate data analysis is the observation window. For instance, a thirty years window has been commonly accepted for climate studies, and is now being reconsidered given the impact it can have on change detection (e.g. [3]). Thus to our knowledge, in climate data analysis, we note 1) a lack of robust outlier pattern detection, and 2) a need to consider very large data sets to minimize the bias induced by the limited observation window. We propose to address those issues in this paper, by introducing a novel concept of pattern outlier, and evaluating our approach to the field of climate data.

In the realm of data mining, outlier detection is receiving much interest, and has shown its benefits in a wide range of applications including fraud detection [4], cybersecurity [5] and the health sector [6]. Identifying outliers through data sets contributes towards decision support, risk and impact studies. Various definitions of what constitutes an outlier have been proposed, along with associated detection methods. A survey [7] specifies twelve different interpretations of outliers from the perspective of different studies. Overall, an outlier can be commonly defined as "an observation that is significantly dissimilar to other data observations or an observation that does not behave like the expected typical behavior of the other observations*"* [8]. The observations can specify a single point outlier, or a shape/pattern denoting an abnormal sub-sequence over a time series data. The latter form our outliers are also called discords.

In this article, we investigate discords over contextual time series. Furthermore, we are interested in very large data series to mitigate the potential impact of the length of the data set at hand upon the results. We apply our approach to large data series coming from monthly data between 1902 and 2005. A scalable and exact approach that has proven its computational and space efficiency to detect discords is the Matrix Profile method [9]. It requires the length of the sub-sequence (window) to be set as a parameter. The chosen window has a strong impact to detect meaningful discords. In existing studies, the detection algorithm is run with different window [10], leading potentially to multiple discords. The ranking of such discords is challenging since it requires meaningful criteria to prefer a discord to another. If the data is labelled such ranking is possible since it can be related to an event, but in case of unlabelled data it usually requires expert knowledge.

We propose a novel concept and approach to identify relevant discords over different windows automatically. To do so, we introduce the concept of *prominent discord* that specifies the most significant discord as an anomalous pattern over the longest continuous period, from a shared starting position. A core benefit of the prominent discord is to gain insight onto discords that coincide over their start date, while searching for the ones with longer and subsuming anomalous pattern. For instance if a drought is found through a window of four months, we seek whether it belongs to a longer dry period that may last six months, one year, ten years. By doing so, we search for the longest span of a discord that would subsume other discords and relates to similar occurrences. In contrast to point outliers that are likely to indicate extreme events, resulting from potential long term changes, prominent discords would reveal the anomalous patterns that cause such events.

We formally define the concept of prominent discord, propose a detection algorithm and present an application to large date sets of so called climate impact runoff data. This means that they are historical data relative to surface and subsurface runoff observations, a variable that provides information about flood and drought risks depending on values being high or low.

Figure 1 gives an intuition of what a prominent discord is. We show three discords of respective lengths 13, 37 and 58; all starting at the same position in the series. The reading of the plots corresponds to daily runoff data over five years in millimetres. The Xs covers the daily timestamp while the Ys the height of the runoff in the Sahelian region. Peaks clearly indicate the rainy season. For each window length the prominent discord is highlighted in a continuous line over the time spam. We notice that when the window size changes we have three prominent discords, all starting at the same position date and the one corresponding to the window size of 58 months covers the other two. It is the most prominent one.

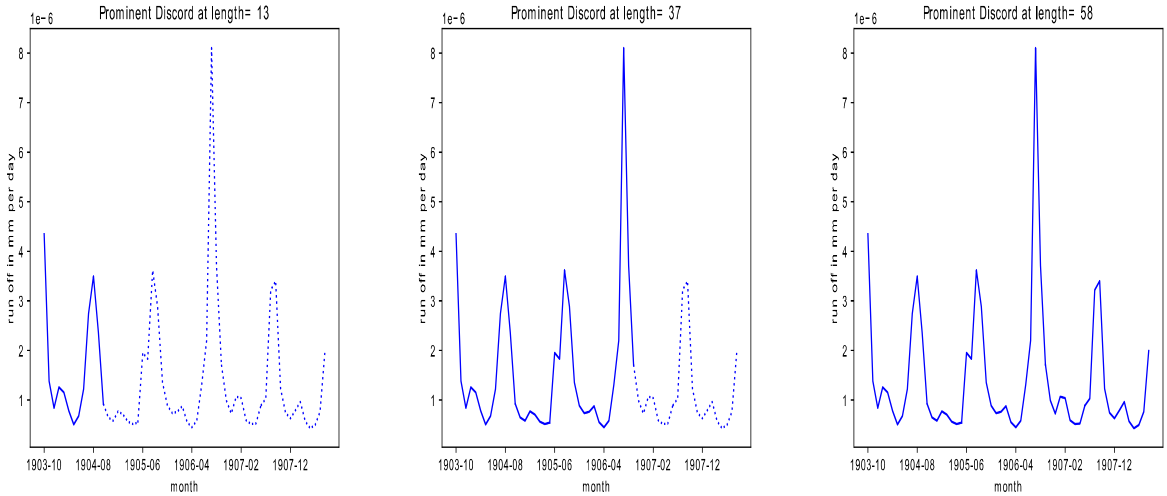


Figure 1. Prominent discord and subsequent discords

The longest one discovered is the most prominent discord: it subsumes the other two. It actually gives insight into lasting changes and anomalous behaviour that could lead towards a change of system state altogether when studying global change and climate data.

The main contributions of this paper are: 1) the novel concept, formalization and method to compute prominent discords and extract the most prominent ones applied to large-scale time series, 2) its application to climate-related data to detect and evaluate the relevance and insights of such discords from a climate point of view.

The paper is organized as follows: section 2 gives a background and related work in the field, section 3 presents our conceptual and methodological contributions, and section 4 its evaluation before concluding in section 5.

2. BACKGROUND AND RELATED WORK

In this section we review previous work relative to time series discord discovery, more specifically with variable length discords and very large time series such as climate models data.

2.1. Climate data analysis

In climate data studies, long-term trends, defined as a tendency towards a climate change pattern, are often characterized by basic statistical measures, such as the average rate of variables increase/decrease over a given time period [11]. In the field of weather extreme events (e.g. droughts, floods, heatwaves, storms), a common approach consists in quantifying how a given climate indicator jumps out, against the background of former climate records (in intensity, frequency or duration; ; [12]). These approaches are conducted under an arbitrary choice of a base time window since a climate reference is inherently defined to assess whether a long-term change and extreme event occurrences can be considered as emergent and/or anomalous. This choice greatly affects the meaning and the robustness of climate studies outcome when this reference is shifted [3].

2.2. Time series discord discovery

Time series discord detection is receiving an increasing interest in data mining since its formalization ([13]–[15]). Efficient and exact methods have been proposed to discover discords in data series [9], [16]. These approaches require some parameter settings including the size of the observation window. The window size is fixed and needs to be specified as an input parameter (HOT SAX [17], QUICK MOTIF [18]). As a result, recent works have drawn on the challenges and insight limitations of a fixed set window size leading to research towards computing all possible discords within a size range, using different methods such as quadratic regression [19], dynamic time warping (DTW) [20], or a graph-based approach [21].

PanMatrix [10] and VALMOD [22] compute variable-length top k discords, using the Matrix Profile method for different window sizes, given a value k. VALMOD considers an interval of possible subsequence lengths as initial parameter; whereas PanMatrix computes exact distances for subsequences of all lengths.

These approaches address the issue of multiple discord computations, but there has been no attempt to order the discords of variable length, and seek those that would have more impact in terms of revealing a potential change of state. Also, the role played by the data series time span has received little attention with respect to its link to discord discovery. Both issues are important for climate model data towards insightful impact studies. A key element is to investigate coincident variable length discords to be able to extract actionable insight through the identification of prominent discords, longest ones sharing a starting position with smallest discords, and thus subsuming them. This is the goal of our approach that can be considered as a meta discord discovery problem over very large data series, with no a priori interpretation of outlier patterns. In other words, we seek *discords for which all the subsequences within a resulting length interval are also discords sharing their starting position*.

We use the Matrix profile approach, because it is an exact method to compute discords for a given window length. It is also computational and space efficiency. Let us now the key notions at hand that will be used to formalize our concept of prominent discord.

2.3 Matrix profile and discords

The matrix profile is a data structure computed to discover discords and motifs using similarity search algorithms [9]. Many algorithms have been proposed with different space and time performance (e.g. STOMP and GPU-STOMP [23], SCRIMP and SCRIMP++ [24]). The data structure builds on the following notions [23], recalled hereafter for further usage and completeness :

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

Definition 2. *A subsequence $T_{i,m}$ of a time series T is a continuous subset of values in T, of length m and starting at position i. Formally, $T_{i,m} = [t_i, t_{i+1}, \ldots, t_{i+m-1}]$, where $1 \leq i \leq n - m + 1$.*

Definition 3 (Distance Profile). *A distance profile $D_{i,m}$ of time series T and length m is a vector of the z-Euclidean distances between a given query subsequence $T_{i,m}$ and all subsequences of length m in the time series T. Formally, $D_{i,m} = \left[d_{i,1}, d_{i,2}, \ldots, d_{i,n-m+1}\right]$, where $d_{i,j}$ $(1 \leq i, j \leq n - m + 1)$ is the distance between $T_{i,m}$ and $T_{j,m}$ with $i \neq j$.*

Definition 4 (z-Euclidian distance). The z-normalized Euclidean distance $d_{i,j}$ between two subsequences $T_{i,m}$ and $T_{j,m}$ of length m, is defined by:

$$d_{i,j} = \sqrt{2m\left(1 - \frac{T_{i,m}.T_{i,m} - m\mu_i\mu_j}{m\sigma_i\sigma_j}\right)} \tag{1}$$

where μ_i and σ_i are respectively the mean and standard deviation of $T_{i,m}$, μ_j and σ_j the mean and standard deviation of $T_{j,m}$.

Definition 4 (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 $D_{i,m} (1 \le i \le n-m+1)$ is the distance profile $D_{i,m}$ of time series T for subsequences of length m.P_m = [$min(D_{1,m})$,...,$min(D_{n-m+1,m})$], where $D_{i,m}(1 \le i \le n-m+1)$ is the distance profile $D_{i,m}$ of time series T for subsequences of length m.*

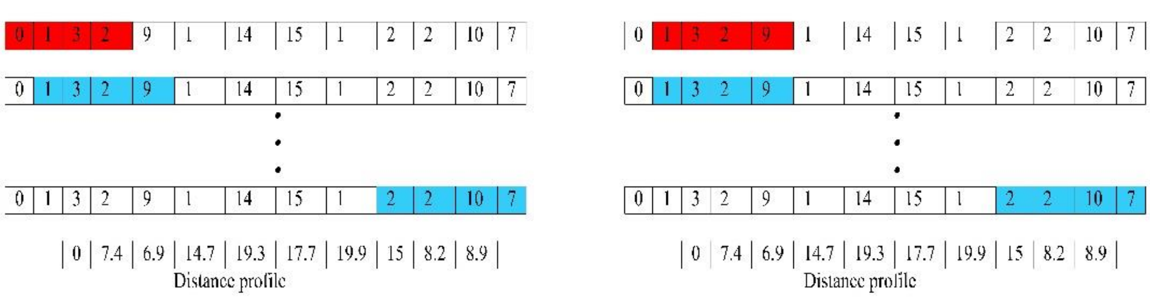
Definition 5 (Index vector). *A matrix profile index vector V_m, associated with a matrix profile P_m denotes the vector of starting position j of the subsequence corresponding to the minimal distance. It is specified by the vector: $V_m = [V_1, V_2, \dots, V_{n-m+1}]$, such that $V_i = j$ if $d_{i,j} = min(D_{i,m})$.*

Definition 6 (Discord). *A discord denoted $\Delta_{j,m}$ is a subsequence $T_{j,m}$ of length m starting at the position j in V_m, that corresponds to the maximum distance value in P_m.*

In other words, the discord of length *m*, is the subsequence in the data series (specified by its starting position *j*), such that among the shortest Euclidean distance, it is the one with the maximal value, ie. with largest anomalous pattern among all.

The following example gives the Distance profiles for two susbsequences (in red) of window size 4 over a time series of length 13. The distance profile vector gives the z-Euclidian distance between the chosen subsequence and all the other ones (the next one is highlighted in blue). There are 10 of them, one per subsequence of size 4, sliding over the time series.

The resulting matrix profile extracts for each subsequence the smallest distance in each distance profile (yellow). Finally, we extract the discord from the matrix profile that is the subsequence corresponding to the greatest distance value in the profile. In this example, the biggest distance value is 14.1,distance between subsequence 7 and subsequence 3 ($V_7 = 3$). Thus the discord is the seventh subsequence [14,15,1,2] in the time series.



										Matrix Profile	Index vector
0	7.4	6.9	14.7	19.3	17.7	19.9	15	8.2	8.9	6.9	9
7.4	0	10.9	7.9	15.7	18.8	19.1	158.8	1.4	8.4	1.4	8
6.9	10.9	0	16.8	16.1	13.6	18.8	14	11.6	6.2	6.2	9
14.7	7.9	16.8	0	16.8	19.8	18	19.4	8.2	13.4	7.9	9
19.3	15.7	16.1	16.8	0	20.7	23.6	18.7	15.3	14.4	11.4	8
17.7	18.8	13.6	19.8	20.7	0	19.2	23.1	19.8	14.4	13.6	2
19.9	19.1	18.8	18	23.6	19.2	0	14.1	20.1	20.5	14.1	3
15	15.8	14	19.4	18.7	23.1	14.1	0	16.2	16.1	14	4
8.2	1.4	11.6	8.2	15.3	19.8	20.1	16.2	0	8.6	1.4	1
8.9	8.4	6.2	13.4	11.4	14.4	20.5	16.1	8.6	0	6.2	0

Figure 2. Distance and matrix profile, sequences of length 4

3. PROMINENT DISCORDS: DEFINITION AND ALGORITHMS

In this section we present our approach and contributions, that include the concept of prominent discord and the method we developed to extract a set of prominent discords. The overall problem we address is the following:

Problem addressed: Given a data series T, a large range of possible lengths and a unit step, compute all the variable length discords, and find the discords, such that all of their subsequences within a length interval and shared starting position in T are discords. From these prominent discords, identify a relevant ordering that relates to their length, and number of subsumed discords.

3.1. Definitions

We work at a meta level with respect to the matrix profile and the discord since we compute matrix profiles over sequences of variable lengths, and seek the longest discords that contain discords with a given ordering relation. To formalize our problem, we introduce three concepts: **discord profile**, **discord subsumption ordering** and **prominent discord**. Note that to allow a reliable reasoning over a large number of subsequences, the interval for the variable lengths is set to [4,..,*n/*2].

Definition 7 (Discord profile). *The discord profile ΔP of a time series T is a set of discords $\Delta_{j,l}$ of variable lengths l and starting positions j, such that $\Delta P = \{\Delta_{j,l} \mid j \in 1..n/2-1, 0 < l \leq n/2\}$.*

Definition 8 (Discord subsumption ordering). *A discord $\Delta_{j,m}$ subsumes a discord $\Delta_{i,l}$, specified as $\Delta_{i,l} \preccurlyeq_{\Delta} \Delta_{j,m}$ if and only if, $i = j$ and $l < m$..*

Note that this ordering relation is a partial order since we assume that two discords with different starting position in the time series are not comparable. The motivation behind this ordering is that such discords might relate to very different events, whereas discords that co-occur in their starting position are more likely to share the root event for the outlier pattern. With identical start position, two discords can relate to the same anomalous pattern. This is not necessarily true for different starting positions.

Definition 9 (Prominent discord). *A Prominent discord $\bar{\Delta}_{j,l_{start},l_{end}}$ of a time series T is the top discord $\Delta_{j,l_{end}}$ of a lattice of all discords $\Delta_{j,l}$ in ΔP such that $\Delta_{j,l_{start}} \preccurlyeq_{\Delta} \Delta_{j,l} \preccurlyeq_{\Delta} \Delta_{j,l_{end}}$.*

Our approach will compute a set of prominent discords for a given time series. We propose an ordering that accounts for 1) the number of subsumed discords (of different lengths of course), and 2) the relative length of the shortest subsumed discord. The idea behind this ordering is to exploit discords for insight studies on the anomalous patterns that can have a lasting impact, and pertained change of behavior in the time series. Intuitively, a point outlier can be the *consequence* of an existing change of behavior (e.g. rising number of extreme weather events), whereas a discord of droughts of length 4, also found in subsequences of lengths 8 and 15 for example, can indicate a first anomalous pattern, that pertains as an anomalous pattern in longer discords. The longest discord subsuming a much shorter one, can indicate a potential important weather change, and impact on soils, agriculture etc.

The ranking function sorts the prominent discords in decreasing order of $sort((l_{end} - l_{start})/l_{start})$, where the top value corresponds to the longest set of subsumed discords ($l_{end} - l_{start}$), and the lower one the length of the first subsumed discord (l_{start}). As illustrated in Figure 3 the prominent discord A will outrank prominent discord B even though they subsume the same number of discords, because A builds upon a shorter outlier (l_{start}).

Prominent discord A

Prominent discord B

Figure 3. Prominent discord ordering: A is preferred to B with higher ratio function value

3.2.2 Algorithms

To derive the set of prominent discords, we first derive the discord profile (Algorithm 1) over variable length discords and extract the prominent ones (Algorithm 2) through a counting method based on shared starting positions. Note that Algorithm 1 makes use of an efficient matrix profile computation algorithm (line 5), the **STOMP** algorithm [23] omitted for space reasons. This algorithm, like other matrix profile computation methods (eg, STAMP) derives the z-normalized Euclidean distance to measure efficiently the distance between subsequences.

Algorithm 1 takes as input the whole time series, a maximum subsequence length and list of variable lengths (line 2--3), called windows (from the terminology of the matrix profile approach) that specifies the subsequence lengths considered. For each window size (lines 4--7), we compute the matrix profile, extract the discord $\Delta_{j,l}$ to be stored in the Discord profile list ΔP.

Algorithm 1: Discord Profile: Compute the list of variable length discords

```
  input  : Time Series T
  output: Discord Profile ΔP
1 initialization
2 int m = length(T)/2
3 list(int) Windows = [4, 5, 6, 7, 8, ..., m]
4 foreach l in Windows do
5 |   P_l ← STOMP (T, l) // Matrix profile for window size l
6 |   Δ_{j,l} ← max(P_l) // Discord of size l
7 |   Discord Profile ΔP ← ΔP.add(Δ_{j,l})
8 end
9 return ΔP
```

The main algorithm, Algorithm 2, computes the list of prominent discords ΔC and returns the sorted list of prominent discords (in decreasing value of respective $(count/l_{start})$ value. It takes as input the time series and returns the sorted list of prominent discords $\bar{\Delta}_{j,l_{start},l_{end}}$ including its starting position, first discord length and longest one. Line 3 initializes a counter of discords having identical starting positions. In line 4 the discord profile ΔP is computed from Algorithm 1 and contains all the discords Δ_{jl} one per length l considered in Algorithm 1. Line 5 and 6 define the variables used to extract the length l and starting position j of a discord in ΔP. Line 7 extracts the length of the first discord. Lines 9--11 increment the count as long as the next discord has the same starting position as the current one denoted by i. Lines 12--14 create a new prominent discord with starting position j, starting length s, last length $s + count$. It is added to the prominent discord list. Lines 15--16 re-initialize the count, and new starting s position to the one of the next discord in ΔP. Line 19 sorts the prominent discord list in decreasing order according to the proposed function $(l_{end} - l_{start})/l_{start}$; and finally line 20 returns the final sorted list ΔC.

Algorithm 2: Sorted list of Prominent Discords

```
   input  : Time Series T,,number of prominent discord K
   output:  List of sorted Prominent Discords ΔC
1  initialization
2  ΔC ← [ ] /List of prominent discords
3  int count = 1 /Increment counting of subsumed discords
4  ΔP ← Discord_Profile(T)
5  Var j / variable that extracts the starting position of Δ_jl in ΔP
6  Var l / variable that extracts the length of Δ_jl in ΔP
7  int s = ΔP[0].l
8  for i ← 0 to i ≤ length(ΔP)-1 do
9      if ΔP[i].j == ΔP[i + 1].j then
10         count + +
11     end
12     else
13         Δ̄_{j,s,s+count} = newΔ̄(ΔP[i].j, s, s + count)
14         ΔC ← ΔC.add(Δ̄)
15         count = 1
16         s = ΔP[i + 1].l
17     end
18 end
19 ΔC ← Quicksort(ΔC, (l_end − l_start)/ l_start)
20 return ΔC // sorted list of prominent discords
```

Worst case time complexity. Algorithm 2 (calling algorithm 1) overall runs in the worst case in $O(n^3)$ where n is the length of the time series. To decompose, we have: Algorithm 1 calls $n/2$ times STOMP, thus runs in the worst case in $O(n^2 \times n) = O(n^3)$. The For loop in Algorithm 2 runs in the length of the discord profile list thus in the worst case $O(n)$ since there is one discord per length ($n/2$ variable length discords), and the list of prominent discords is sorted in the worst case in $O(nlogn)$.

4. EXPERIMENTAL EVALUATION AND COMPARISONS

We now present an application of our method to the analysis of large datasets relative to runoff historical climate data. Runoff data correspond to measures of waters in terms of distance (in mm) above the land surface to reach a stream but also to infiltrate the soil surface. 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. We also compare our proposed approach to other discord discovery methods.

The climate data. We consider observed monthly runoff data, defined in climate data science as an impact variable analyzed to quantify flood and drought risks at regional and global scales (e.g. [12], [25], [26]). These monthly runoff observations are obtained from the Global Runoff Reconstruction dataset (GRUN) that covers the 1902-2014 time period (113 years), with a 0.5° × 0.5° spatial grid resolution [27]. We focus our analysis on the Sahel region, a particularly soil water vulnerable area, and we spatially average monthly runoff over the corresponding grid box [5°W-25°E ; 10°N-18°N]. Our prominent discord approach is then applied to the obtained Sahel time series between 1902 and 2005 (i.e. 104 years, 1248 months), a period commonly considered in historical climate analysis.

4.1. Prominent discords discovery

For the dataset of monthly runoff observation data over 1248 months, we applied our approach and derived the list of all prominent discords, including their ranking based on our proposed ratio function, after calculating the discords for all variable length windows in the interval [4,..,1248/2], with a monthly step increment. The set of prominent discords was derived in 3.5 minutes.

Figure 4 shows for five prominent discords, including the most prominent one, their respective subsumed discords. The X-axis indicates window lengths up to 130 months, and the Y-axis the discords starting date. Each blue dot represents a discord with its window length (Xs) and its starting date (Ys). The arrows show the prominent discords with all the subsumed discords of coinciding starting dates. The length of the arrows illustrates how many discords the prominent one subsumes. Here we have four prominent discords. The most prominent discord starts at the position date 1903-10-15, with a lower window size of 13 months for its first subsumed discord, and upper length of 58 months.

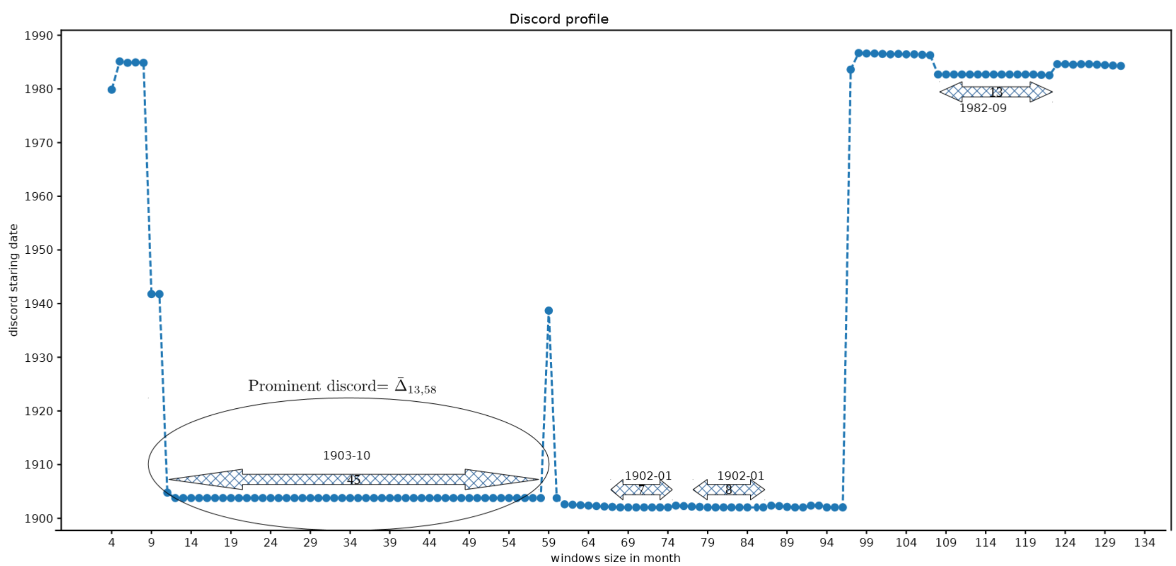


Fig.4. Prominent discords derived from observed runoff monthly data.

Table 1 shows the top five prominent discords in ΔC according to our proposed ratio ordering. The first and second ones are found in Figure 4.

date	starting window size	ending window size	ratio
1903-10-15	13	58	3.46
1982-09-15	109	120	0.10
1902-09-15	242	264	0.09
1903-09-15	193	209	0.08
1902-01-15	80	86	0.075

Table 1. Top 5 prominent discords sorted by the ratio function

4.2. Comparison with alternative approaches

We compared our approach with existing discord discovery methods, illustrating mainly the importance of considering an exact approach for the prominent discord extraction, and the need of variable length discord computation without parameterized length settings.

We considered HOT-SAX, an extension of the SAX algorithm [28]. SAX discretizes time series into words and detects motifs in time series but not discords. HOT-SAX algorithm was developed to detect discords. It builds a suffix tree that stores the words generated by SAX. A word with the least number of occurrences is a discord. A requirement is that the number of extracted discords is predefined.

Rare Rule Anomaly (RRA) [29] is an algorithm that uses grammar-based compression able to detect motifs and discords in time series. Similar to HOT-SAX, RRA uses SAX algorithm to discretize the time series. A grammatical induction algorithm (ex: Sequitur [30]) is used to generate the grammar. These generated grammars are used to detect the discords.

One of the main parameters for HOT-SAX and RRA is the window size. To be able to compare them with our approach, we use the value l_{end} of each prominent discord as the window parameter.

HOT SAX		
window size	start	end
58	1902-01-01	1906-11-01
120	1943-08-01	1953-08-01
264	1972-02-01	1994-02-01
209	1973-06-01	1990-11-01
86	1902-07-01	1909-09-01

Table 2. HOT SAX results using windows extracted from the l_{end} of each prominent discord

RRA		
window size	start	end
58	1902-02-01	1907-04-01
120	1944-08-01	1954-11-01
264	1953-01-01	1977-08-01
209	1953-12-01	1971-12-01
86	1943-04-01	1951-01-01

Table 3. RRA results using windows extracted from the l_{end} of each prominent discord

We can see that both HOT SAX and RRA lead to different results in terms of starting and end date of the prominent discord for a given window size. This comes from the fact that they are not exact methods and given an input window length, lead to a different discord.

We also compared with the PAN MATRIX algorithm. It calculates variable length discords, using the SKIMP algorithm to compute the matrix profile for all motif lengths.

Table 4 shows the top five discords with the PAN MATRIX. Compared with our results in Table 1, the top 5 discords are different, and the top 5 discord of PAN MATRIX are not relative to subsequences since they are not based on an ordering among variable length discords. PAN MATRIX is efficient to calculate variable length discord, but is not designed to order discords.

Pan Matrix	
Date	Window size
1952-07-01	619
1952-07-01	618
1952-07-01	617
1952-07-01	616
1952-07-01	615

Table 4. PAN MATRIX top 5 discords

4.3. Analysis and insights for climate data analysis

To analyze those results, as well as the relevance and insights of our prominent discord and proposed orderings, we compared with statistical analysis of those historical climate data (e.g. long-term trends, seasonal cycles, standard-deviation). It is worth noting that the prominent discord approach is unsupervised, and does not consider any climate behaviour nor known physical processes.

Figure 6 shows the annual average runoff data over the Sahel region illustrating the maximum, minimum and average annual values, providing a general idea of the global fluctuations and extremes. Figure 7 relates the monthly runoff values together with the top five prominent discords we discovered, cf Figure 5. For each color, the dotted lines indicate the starting position of a prominent discord and the vertical lines the respective l_{start} and l_{end}.

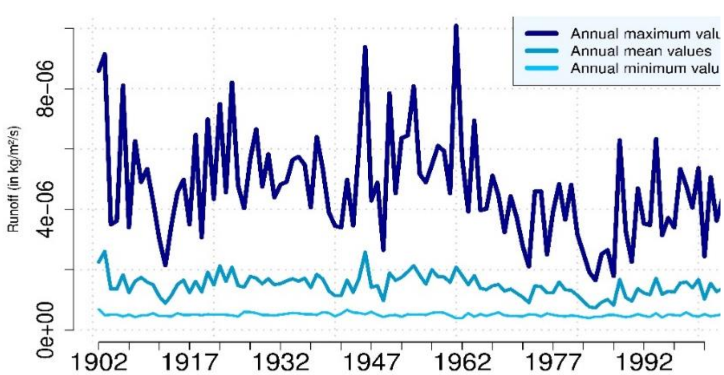


Figure 6. Annual mean runoff values

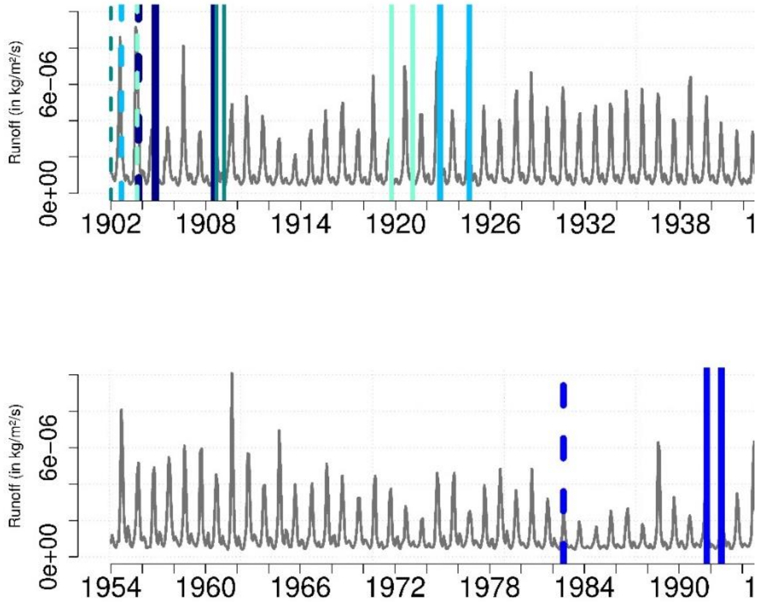


Figure 7. Monthly runoff with prominent discords between 1900 and 2005

According to climate studies, a well known soil drying trend mostly resulting from a rainfall decrease is observed between 1900 and 2013 in the Sahel ((e.g. [27])). We find consistent trends using runoff data, particularly in the annual maximum runoff time evolution between 1902 and 2005 ([26]). In our work, four of the five first prominent discords coincide with starting dates and length intervals during the first 22 years of the time series (Table 5 and Figure 7). The corresponding averaged monthly runoff over 1902-1924 also shows a mean of $15e^{-07}$ kg/m^2/s and a standard-deviation (i.e. an inter-annual variability indicator; ([31]) of 0.14 kg/m^2/s, whereas the entire 1902-2005 time series is characterized by a mean of $12e^{-07}$ kg/m^2/s and a standard deviation of 0.12 kg/m^2/s. These four prominent discords may thus illustrate the specific time pattern between 1902 and 1924 with higher soil water amount and larger inter-annual variability, before the upcoming continuous long-term drying trend observed within Sahel.

The second prominent discord is detected for a starting date at 1982-09-15 and window sizes in [109 ; 120 months]. This period corresponds to the time period with the smallest mean runoff values of the entire time series (Fig. 6 and Fig. 7). The associated 1982-1992 mean runoff is $9e^{-07}$ kg/m^2/s compared to a mean runoff of $12e^{-07}$ kg/m^2/s over 1900-2005. This prominent discord illustrates the temporal pattern resulting from intense droughts that occurred in Sahel in the 1980s. During that decade, the most severe drought ever recorded over the African continent occurred during 1983-1984 (Figure 2 in [32]).

In this study, we demonstrated two major usefulness of our proposed prominent discord approach and ordering relations, in climate data analysis in terms of real insights towards the emergence of a long-term change, and the detection of recurring anomalous events.

First, we showed that the prominent discord concept does capture a 20-30 years pattern illustrating a different (former) climate regime compared to the rest of the considered time series (*emergence*).

Second, we showed that the subsumption ordering and prominent discords ranking capture the time pattern at a decade scale of the driest recorded yearly event (*recurring anomaly detection*).

5. Conclusion and future work

In this paper we proposed a new concept in the realm of variable length discords, the prominent discord. It focuses on identifying anomalous patterns that last through time and subsume sets of discords. The main contributions are the formalization and method to compute prominent discords and extract the most prominent ones applied to large scale time series, and the application to climate-related data.

Our ordering and results show their relevance in the field of climate data series. In particular, they show that through such ordering we gain insights on the anomalous patterns that have a lasting impact, and pertained change of behavior in the time series. This is new to our knowledge.

Future works include further experimental studies on different impact climate data and other data sets, to evaluate the thematic insights of our approach, as well as some optimization of the algorithm to ensure scalability.

Acknowledgements

The authors would like to thank the Occitanie Region, who partially funded this research, and the readers of previous versions for the tremendous guidance.

References

[1] Randall, D. A.; Wood, R. A.; Bony, S.; Colman, R.; Fichefet, T.; Fyfe, J.; Kattsov, V.; Pitman, A.; Shukla, J.; Srinivasan, J.; others. (2007). Climate models and their evaluation, *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC (FAR)*, Cambridge University Press, 589–662

[2] Knutti, R.; Masson, D.; Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there: CLIMATE MODEL GENEALOGY, *Geophysical Research Letters*, Vol. 40, No. 6, 1194–1199. doi:10.1002/grl.50256

[3] Sippel, S.; Zscheischler, J.; Heimann, M.; Otto, F. E. L.; Peters, J.; Mahecha, M. D. (2015). Quantifying changes in climate variability and extremes: Pitfalls and their overcoming, *Geophysical Research Letters*, Vol. 42, No. 22, 9990–9998. doi:10.1002/2015GL066307

[4] Hilal, W.; Gadsden, S. A.; Yawney, J. (2022). Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances, *Expert Systems with Applications*, Vol. 193, 116429. doi:10.1016/J.ESWA.2021.116429

[5] Hopkins, S.; Kalaimannan, E.; John, C. S. (2020). Sub-Erroneous Outlier Detection of Cyber Attacks in a Smart Grid State Estimation System, *2020 11th IEEE Annual Ubiquitous Computing, Electronics \& Mobile Communication Conference (UEMCON)*, 447–454

[6] Xia, H.; An, W.; Li, J.; Zhang, Z. J. (2020). Outlier knowledge management for extreme public health

events: understanding public opinions about COVID-19 based on microblog data, *Socio-Economic Planning Sciences*, 100941

[7] Ayadi, A.; Ghorbel, O.; Obeid, A. M.; Abid, M. (2017). Outlier detection approaches for wireless sensor networks: A survey, *Computer Networks*, Vol. 129, 319–333

[8] Hodge, V.; Austin, J. (2004). A survey of outlier detection methodologies, *Artificial Intelligence Review*, Vol. 22, No. 2, 85–126

[9] Yeh, C.-C. M.; Zhu, Y.; Ulanova, L.; Begum, N.; Ding, Y.; Dau, H. A.; Silva, D. F.; Mueen, A.; Keogh, E. (2016). Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets, *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 1317–1322

[10] Madrid, F.; Imani, S.; Mercer, R.; Zimmerman, Z.; Shakibay, N.; Keogh, E. (2019). Matrix profile xx: Finding and visualizing time series motifs of all lengths using the matrix profile, *2019 IEEE International Conference on Big Knowledge (ICBK)*, 175–182

[11] Hartmann, D. L.; Klein Tank, A. M. G.; Rusticucci, M.; Alexander, L. V; Brö\"nnimann, S.; Charabi, Y.; Dentener, F. J.; Dlugokencky, E. J.; Easterling, D. R.; Kaplan, A.; Soden, B. J.; Thorne, P. W.; Wild, M.; Zhai, P. M. (2013). Observations: Atmosphere and Surface, T. F. Stocker; D. Qin; G.-K. Plattner; M. Tignor; S. K. Allen; J. Boschung; A. Nauels; Y. Xia; V. Bex; P. M. Midgley (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 159–254. doi:10.1017/CBO9781107415324.008

[12] Sillmann, J.; Kharin, V. V; Zwiers, F. W.; Zhang, X.; Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections: CMIP5 PROJECTIONS OF EXTREMES INDICES, *Journal of Geophysical Research: Atmospheres*, Vol. 118, No. 6, 2473–2493. doi:10.1002/jgrd.50188

[13] Fu, A. W.-C.; Leung, O. T.-W.; Keogh, E.; Lin, J. (2006). Finding time series discords based on haar transform, *International Conference on Advanced Data Mining and Applications*, 31–41

[14] Chiu, B.; Keogh, E.; Lonardi, S. (2003). Probabilistic discovery of time series motifs, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 493–498. doi:10.1145/956750.956808

[15] Kleijnen, J. P. C. (2008). Response surface methodology for constrained simulation optimization: An overview, *Simulation Modelling Practice and Theory*, Vol. 16, No. 1, 50–64. doi:10.1016/j.simpat.2007.10.001

[16] Yankov, D.; Keogh, E.; Rebbapragada, U. (2008). Disk aware discord discovery: Finding unusual time series in terabyte sized datasets, *Knowledge and Information Systems*, Vol. 17, No. 2, 241–262

[17] Keogh, E.; Lin, J.; Fu, A. (2005). Hot sax: Efficiently finding the most unusual time series subsequence, *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 8--pp

[18] Li, Y.; Leong, H. U.; Yiu, M. L.; Gong, Z. (2015). Quick-motif: An efficient and scalable framework for exact motif discovery, *2015 IEEE 31st International Conference on Data Engineering*, 579–590

[19] Leng, M.; Chen, X.; Li, L. (2008). Variable length methods for detecting anomaly patterns in time series, *2008 International Symposium on Computational Intelligence and Design* (Vol. 2), 52–56

[20] Vy, N. D. K.; Anh, D. T. (2016). Detecting variable length anomaly patterns in time series data, *International Conference on Data Mining and Big Data*, 279–287

[21] Boniol, P.; Palpanas, T. (2020). Series2graph: Graph-based subsequence anomaly detection for time

series, *Proceedings of the VLDB Endowment*, Vol. 13, No. 12, 1821–1834

[22] Linardi, M.; Zhu, Y.; Palpanas, T.; Keogh, E. (2020). Matrix profile goes MAD: variable-length motif and discord discovery in data series, *Data Mining and Knowledge Discovery*, Vol. 34, No. 4, 1022–1071

[23] Zhu, Y.; Zimmerman, Z.; Senobari, N. S.; Yeh, C.-C. M.; Funning, G.; Mueen, A.; Brisk, P.; Keogh, E. (2016). Matrix profile ii: Exploiting a novel algorithm and gpus to break the one hundred million barrier for time series motifs and joins, *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 739–748

[24] Zhu, Y.; Yeh, C.-C. M.; Zimmerman, Z.; Kamgar, K.; Keogh, E. (2018). Matrix profile XI: SCRIMP++: time series motif discovery at interactive speeds, *2018 IEEE International Conference on Data Mining (ICDM)*, 837–846

[25] Arnell, N. W.; Lloyd-Hughes, B. (2014). The global-scale impacts of climate change on water resources and flooding under new climate and socio-economic scenarios, *Climatic Change*, Vol. 122, Nos. 1–2, 127–140. doi:10.1007/s10584-013-0948-4

[26] Gosling, S. N.; Zaherpour, J.; Mount, N. J.; Hattermann, F. F.; Dankers, R.; Arheimer, B.; Breuer, L.; Ding, J.; Haddeland, I.; Kumar, R.; Kundu, D.; Liu, J.; van Griensven, A.; Veldkamp, T. I. E.; Vetter, T.; Wang, X.; Zhang, X. (2017). A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C, *Climatic Change*, Vol. 141, No. 3, 577–595. doi:10.1007/s10584-016-1773-3

[27] Ghiggi, G.; Humphrey, V.; Seneviratne, S. I.; Gudmundsson, L. (2019). GRUN: an observation-based global gridded runoff dataset from 1902 to 2014, *Earth System Science Data*, Vol. 11, No. 4, 1655–1674. doi:10.5194/essd-11-1655-2019

[28] Lin, J.; Keogh, E.; Lonardi, S.; Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms, *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD '03*, Vol. 13, 2–11. doi:10.1145/882082.882086

[29] Senin, P.; Lin, J.; Wang, X.; Oates, T.; Gandhi, S.; Boedihardjo, A. P.; Chen, C.; Frankenstein, S. (n.d.). Time series anomaly discovery with grammar-based compression., *Researchgate.Net*

[30] Nevill-Manning, C.; Research, I. W.-J. of A. I.; 1997, undefined. (1997). Identifying hierarchical structure in sequences: A linear-time algorithm, *Jair.Org*, Vol. 7, 67–82

[31] Hansen, J.; Sato, M.; Ruedy, R. (2012). Perception of climate change, *Proceedings of the National Academy of Sciences*, Vol. 109, No. 37, E2415--E2423. doi:10.1073/pnas.1205276109

[32] Masih, I.; Maskey, S.; Mussá, F. E. F.; Trambauer, P. (2014). A review of droughts on the African continent: a geospatial and long-term perspective, *Hydrology and Earth System Sciences*, Vol. 18, No. 9, 3635–3649. doi:10.5194/hess-18-3635-2014

Authors

Hussein Al Khansa is a computer science PhD student from the university of Montpellier, within the Espace Dev laboratory. He carried out his Masters degree at the Lebanese International University, in Lebanon.

Carmen Gervet is professor of computer science at the university of Montpellier and director of the Espace-Dev research unit. She is specialized in Artificial Intelligence and more specifically decision support systems and constraint-based reasoning.

Audrey Brouillet is a post-doctoral researcher in the laboratory Espace-Dev. She obtained her PhD in climatology from the university Paris Saclay in 2020. She specializes in climate and impact data analysis, in the context of climate change.