A Visual Analytics Approach to Multiscale Exploration of Environmental Time Series

Mike Sips, *Member, IEEE*, Patrick Köthur, Andrea Unger, Hans-Christian Hege, and Doris Dransch

Abstract—We present a Visual Analytics approach that addresses the detection of interesting patterns in numerical time series, specifically from environmental sciences. Crucial for the detection of interesting temporal patterns are the time scale and the starting points one is looking at. Our approach makes no assumption about time scale and starting position of temporal patterns and consists of three main steps: an algorithm to compute statistical values for all possible time scales and starting positions of intervals, visual identification of potentially interesting patterns in a matrix visualization, and interactive exploration of detected patterns. We demonstrate the utility of this approach in two scientific scenarios and explain how it allowed scientists to gain new insight into the dynamics of environmental systems.

Index Terms—Time series analysis, multiscale visualization, visual analytics.

◆

1 INTRODUCTION

Advancement of recent technology allows geoscientists to measure and to simulate a wide range of variables of environmental systems. The resulting environmental time series encompass long time periods at high (and sometimes varying) sampling rates. To study the temporal behavior of observed environmental systems, scientists need to detect interesting patterns in these time series.

Since the dynamics of environmental systems are typically not completely understood, the detection of interesting patterns in environmental time series comprises two major challenges. First, researchers often have difficulties to specify in advance what constitutes an 'interesting' pattern. The significance of a pattern depends on the specific application, the analysis question, the occurrence of the pattern in time, and its concordance with domain knowledge. Second, the time scales on which interesting temporal patterns occur are often difficult to determine. Environmental systems show interesting patterns at very different time scales. Here, a time scale denotes the length of intervals of a logical division of the time series.

In this paper, we present a Visual Analytics approach that addresses the detection of interesting patterns in environmental time series. To address the two major challenges mentioned above, we identified in close collaboration with geoscientists four important design requirements. Our approach allows users to visually detect potentially interesting patterns in a comprehensive visual overview, and to visually inspect and evaluate patterns simultaneously at many time scales. This enables users to quickly assess the interestingness of detected patterns. The proposed system is characterized by its conceptual simplicity and effectiveness in finding hidden temporal patterns, though only basic statistical parameters are calculated. Its effectiveness is based on the combination of (a) an algorithm to compute statistical values for all possible time scales and starting positions of intervals, and (b) the human visual system to detect potentially interesting patterns in matrix visualizations.

- *Mike Sips is with German Research Center for Geosciences GFZ, e-mail: sips@gfz-potsdam.de.*
- *Patrick Köthur is with German Research Center for Geosciences GFZ, e-mail: koethurp@gfz-potsdam.de.*
- *Andrea Unger is with German Research Center for Geosciences GFZ, e-mail: unger@gfz-potsdam.de.*
- *Hans-Christian Hege is with Zuse Institute Berlin, e-mail: hege@zib.de*
- *Doris Dransch is with German Research Center for Geosciences GFZ, e-mail: dransch@gfz-potsdam.de.*

In particular, the contributions of the paper are the following:

- We propose a straightforward, yet significant, approach that captures characteristics of the temporal behavior of environmental systems across all possible time scales and starting positions of intervals based on a user-chosen statistical measure.

- We utilize the tremendous pattern recognition apparatus of humans to detect hidden patterns of environmental time series in a matrix visualization called *Pinus* view which presents the variation of a user-chosen statistical measure across all possible time scales and starting positions of intervals.

- We demonstrate, how patterns, detected at different time scales, can be efficiently explored with visual queries and tightly coupled visual components.

- We explain, how our prototype enabled scientists to gain new insight into the dynamics of environmental systems.

2 RELATED WORK

2.1 Automated Mining of Temporal Patterns

In mining algorithms, the detection of potentially interesting patterns is an algorithmic process which is based on pre-defined assumptions about the patterns. Many mining algorithms detect specific patterns in time series data, e.g., periodic patterns [9, 10, 32], surprising patterns [18], sequential patterns [1], or subsequence motifs [24]. Environmental systems, however, may show behavior that does not meet pre-defined features, e.g., type of periodicity. An example are regime changes which are difficult to detect by mining algorithms alone (see Section 5.1 for further details). Mining algorithms typically require users to specify important mining parameters in advance, e.g., lengths of periods or lengths of motifs. Users often have difficulties to specify these parameters. This is especially true for environmental time series where non-trivial processes exhibit temporal patterns with different and typically unknown durations. Our approach makes no assumptions about temporal patterns. It combines fundamental statistical quantities and the tremendous pattern recognition abilities of the human user to support effective detection of potentially interesting temporal patterns at different time scales.

2.2 Time Series Visualization

Graphical representations of time series support users in the discovery of a wide range of potentially interesting patterns by taking advantage of the human ability to visually detect such patterns and to assess them using expert knowledge [5]. Many time series visualizations have been proposed over the last decades. Here, we discuss only methods that are relevant for this paper, and refer to [2] for a comprehensive overview.
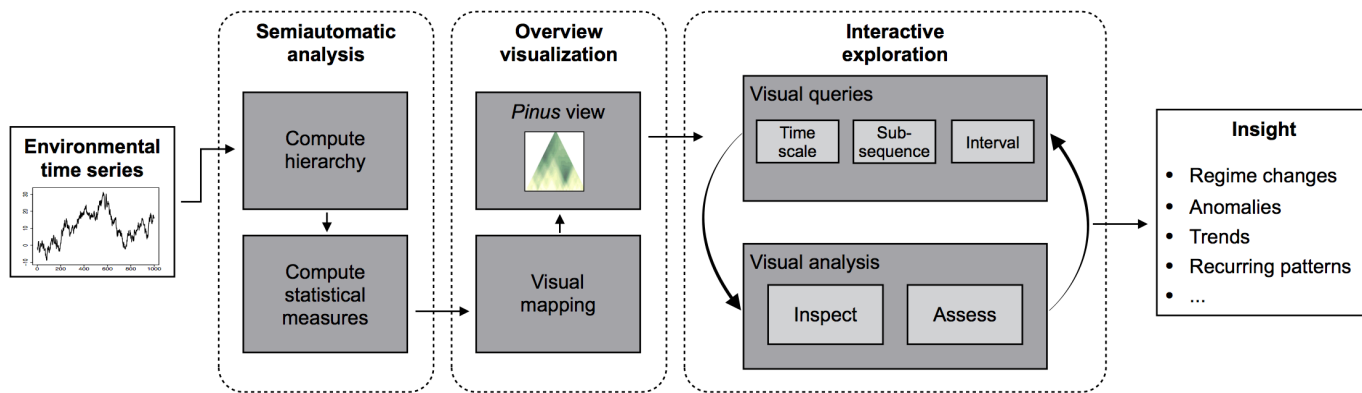
Fig. 1. Our approach to the analysis of environmental time series; the combination of (a) an algorithm to compute statistical values for all possible time scales and starting positions of intervals, and (b) the human visual system to detect potentially interesting patterns in matrix visualizations. Interactive exploration allows users to quickly decide on the interestingness of a detected pattern. The achieved results show the significance of our approach to the analysis of environmental systems.

Spiral displays [27, 29] facilitate the detection of different kinds of periodicity in time series data. They map a time series to a spiral-shaped time axis (see also [8]) and allow users to search for periodic patterns by adjusting the spirals cycle length. The utility of spiral approaches is limited to periodic data, e.g., seasonal cycles.

Time series visualization tools often tightly couple interactive visualization with efficient mining algorithms. Hao and collaborators [12] summarize and visualize time series by their recurring patterns. Recurrences, however, are only one feature of interest in the analysis of environmental time series. TimeSearcher [4, 13] allows users to specify various regions of interest within a time series, and to find time series in time series databases that exhibit similar temporal patterns. Specifying regions of interest requires users to have a general idea of what constitutes interesting patterns. VizTree [21] takes a different approach; it computes a symbolic approximation of a time series, and then visualizes the sequence of symbols in a suffix/subsequence tree. The subsequence extraction and discretization in VizTree provides an elegant approach to compute compact visual representations of large time series by presenting the frequency of recurring subsequences to the user. The shape of the subsequence tree, and thus the temporal features that can be explored, depend on the symbolic approximation parameters, i.e, length of the sliding window and number of segments per window. Our approach does not rely on such mining parameters. In contrast to VizTree, the pixel space of a *Pinus* view grows quadratically with the length of the time series. However, as the examples show, this visualization can be powerful.

2.3 Multiscale Approaches

Many multiscale visualization approaches support users in the exploration of large time series data by presenting time series at several levels of abstraction. Multiscale visualization methods facilitate the information seeking process "Overview first, zoom and filter, details on demand" [26]. Information mural [16] computes a miniature version of large time series data; the density of overplotting is visualized using grayscale shading. TraXplorer [15], KronoMiner [35], Line Graph Explorer [19] and LiveRAC [22] start with an overview of the time series data. As users zoom, more detail is provided by changing the visual representation of the time series data. Hao and collaborators [11] presents a space-efficient multiresolution matrix representation of time series; the abstraction level for each cell is determined based on a numerical degree of interest scale. BinX [3] uses binning along the time axis to present user-defined levels of aggregation of the time series. Curve density estimates [20] visually represent large line graphs as a smooth curve for a specific level of detail.

These multiscale approaches adapt the visual representation of time series to the available screen space. The detection of interesting patterns relies on browsing through different levels of abstraction. In contrast, we decompose a time series into a hierarchical data structure

which describes the temporal behavior of the time series at all possible time scales (see also [33, 34]). A *Pinus* view is a visual representation of this hierarchy. It serves as a starting point for interactive exploration of interesting patterns.

2.4 Matrix visualization

A *Pinus* view is a matrix visualization of the hierarchical decomposition of a time series. Matrix visualizations can display massive data sets in a compact visual overview (see also [17]). They have a substantial history in helping numerical analysts and algorithm designers to gain insight into the behavior of algorithms [28, 30]. They are also a standard technique in many software packages such as MATLAB and R. We refer to [6, 31] for a comprehensive overview.

3　DESIGN REQUIREMENTS

To better understand the requirements for a tool that supports scientists in detecting interesting patterns in environmental time series, we conducted informal interviews with expert users from different domains: climate scientists trying to understand the complexity of the Earth's climate system and ocean modelers trying to relate temporal patterns in time series of coupled subsystems. Based on their feedback, we identified the following requirements:

R1 Semiautomatic Analysis. The tool to be developed should enable researchers to conduct analyses that characterize the temporal behavior of environmental systems across all time scales and starting positions of intervals. Thus, a descriptive statistical measure should be computed across all possible time scales and starting positions of intervals. The measure should be selectable by the user.

R2 Overview Visualization. Domain experts need to see the computational results of the analysis in an overview visualization. The overview visualization should allow users to study the variation of the temporal behavior across all time scales and starting positions of intervals, and to detect potentially interesting patterns.

R3 Visual Queries and Interaction. To query the time series data directly in the overview visualization is crucial to facilitate insight into complex temporal behavior. Users should be able to select the specific time scales and subsequences of interest in the overview visualization, and to display the selected data in other linked visualizations.

R4 Multiple and Coordinated Views. To support visual exploration of potentially interesting patterns, the tool should provide different perspectives to the time series data. It is important that

the tool provides a common time axis in all views to support side-by-side comparisons.

4 OUR APPROACH

We propose an approach for detection of interesting patterns in environmental time series: an algorithm to characterize the temporal behavior, visual detection of potentially interesting patterns in a matrix visualization, and interactive exploration of detected patterns. Figure 1 presents our approach in detail.

4.1 Definitions

We begin with some basic notations and definitions. Given a time series $\{Y_i\}$, we define an interval $s(p,w)$ as w contiguous time steps $p,\ldots,p+w-1$ starting at time step p. Note, the interval $s(p,w)$ can contain a varying number of data items $y_j \in \{Y_i\}$ with $p \leq j \leq p+w-1$ for time series with unequal sampling rates.

To analyze characteristics of the temporal behavior of the time series $\{Y_i\}$ at scale w, analysts divide $\{Y_i\}$ into intervals $s_j(p_j,w)$ over all starting positions $p_1,p_2,\cdots$ with $p_i < p_j, 1 \leq i < j \leq N-w$ where N denotes the time step of the last data item of the time series $\{Y_i\}$. Statistical measures then quantitatively characterize the temporal behavior within the intervals s_j.

4.2 Basic Concept

We compute statistical measures for all possible time scales w and starting positions p. A specific feature of our approach is its flexibility due to the wide range of statistical measures provided, e.g., mean, variance, entropy. This allows for capturing a large set of temporal patterns to address different analysis tasks and different types of data (design requirement **R1**).

An overview visualization, called *Pinus*, presents the results of the semiautomatic analysis as a function of time scale w and starting position p. This compact and easy-to-interpret matrix visualization allows users to identify potentially interesting patterns at multiple time scales and to locate these patterns in the original time series (design requirement **R2**).

With our tool, users can retrieve and visualize patterns they have identified in the *Pinus* view. They can easily formulate queries through interactive selections (design requirement **R3**), and study the results of these visual queries in tightly coupled visual components (design requirement **R4**).

4.3 Semiautomatic Analysis

4.3.1 Algorithm

The algorithm computes the result of a statistical measure over all time scales and intervals as a half $N \times N$ matrix. As statistical measures, our prototype implements mean $\bar{y} = \frac{1}{N}\sum_{i=1}^{N}y_i$, variance $\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$, and discrete entropy $\sum_{i=1}^{K} p_i \log p_i$ with a quantization of the domain of the observed values into K bins (see Section 5.1.4 for further details).

A brute-force algorithm would compute the mean, variance or entropy for each interval $s(p,w)$ one at the time, e.g., taking the sum of $p+w-1$ data items to compute the mean of the interval $s(p,w)$.

We speed-up the computation of the result matrix C as follows. To compute the result matrix C for the statistical quantity mean, the algorithm utilizes the hierarchical structure of the result set to compute the sum of w consecutive time steps on the fly. The algorithm initializes the N cells $c(p,1)$ of the first column with the data item at time step p. Note, the cell $c(p,1)$ is zero if a time step p does not contain a valid data item. The algorithm computes the sum of $w > 1$ consecutive time steps starting at time step p and ending at time step $p+w-1$ in a bottom-up manner using $c(p,w) = c(p,w-1) + c(p+w-1,1)$. The mean of $c(p,w)$ is the normalized value $c(p,w)/w$ for $1 \leq p \leq N-w+1$.

The computation of the variance requires the matrix C and an additional matrix $C2$. The algorithms initializes the first column of $C2$ with the square of the data items. Here, the cell $c2(p,w)$ stores the sum of the squares of the w consecutive time steps starting at time step p and

ending at time step $p+w-1$. The computation of the sum of squares for each cell is similar to the schema of matrix C.

To compute entropy of subsequences, the algorithm determines the bin for each data item at time step p the bin in which it falls; the bin label is stored in the vector $ybin[p]$. Next, the algorithm stores the bin counts of K bins for each subsequence starting at time step p in a matrix B. The algorithm initializes the cells b of B with $b(p,ybin[p]) = 1$ and zero elsewhere. To compute the entropy of the subsequences starting at time step p and ending at time step $p+w-1$, the algorithm iterates over all possible window lengths $w > 1$ and updates the histogram for each starting position p on the fly using $b(p,ybin[p+w-1]) = b(p,ybin[p+w-1]) + 1$. The entropy of the subsequence is the discrete entropy of the updated histogram $b(p,l)$ for $1 \leq l \leq K$ and is stored in the cell $c(p,w)$. Note, the entropy of the cells $c(p,1)$ are zero.

4.3.2 Time Complexity

To determine the computational complexity of the algorithm, we analyze the computational cost for the outer and the inner loop. Let us start with the inner loop; to iterate over all possible starting positions p. The computational cost to determine the statistical quantities for all intervals $s(p,w)$ of a given w is linear in the length of the time series. Since we iterate over N different window lengths w in the outer loop, the total computational effort to compute the result matrix C is $O(N^2)$.

Although N typically ranges from 10^3 to 10^5 in environmental sciences, and thus, N^2 is large, the computation is feasible. Furthermore, the result set can be restricted in practice, as follows: First, since micro-scales often do not provide interesting features in large environmental time series, users can choose a reasonably large initial window size w. Second, while the algorithm considers all N time scales of a time series as a default, users can control the sampling of the window lengths, e.g., considering only every second time scale (interval length). Practical experiments showed that the characteristic temporal behavior of environmental time series can still be captured; what is a reasonable sampling of the time scales depends on the application scenario and the original sampling rate. A third complementary option is to sample the starting positions. The current prototype considers all time steps of the original time series, but this condition could be relaxed, if necessary. One should also have in mind that this concerns only a pre-computation step that does not affect interactive work.

Table 1 summarizes the computational performance in the scientific application scenarios (see Section 5).

Table 1. Summary of the computational performance in the application scenarios.

Data	N	Sampling	Timing
Ocean	876	Regular	0.01 s
Antarctica	4601	Varying	1.6 s

4.4 *Pinus* View

In principle, a *Pinus* view is a matrix visualization. that presents each item of the resulting hierarchy.

4.4.1 Visual Mapping

To display the values represented by the $N \times N$ matrix C, we need a mapping $f : I \to D$ from the index space I of C to the screen space $D = \{1,\ldots,W\} \times \{1,\ldots,H\}$, where W and H are the dimensions of the screen window.

We use two different mappings $(i,j) \mapsto f(i,j)$. The **basic visual layout**, defined by $f(j,k) = (k,j)$ and the **symmetrical layout**, defined by $f(j,k) = (k + ((N/2) - (|k_{max}(j)|/2)), j)$, where $|k_{max}(j)|$ denotes the number of intervals at time scale j. Both versions are shown in Figure 2. The visual appearance of these layouts differs significantly. In our collaborations, scientists preferred the symmetrical layout because they considered it more intuitive and easier to read. More specifically, scientists found it much easier to relate a pixel in
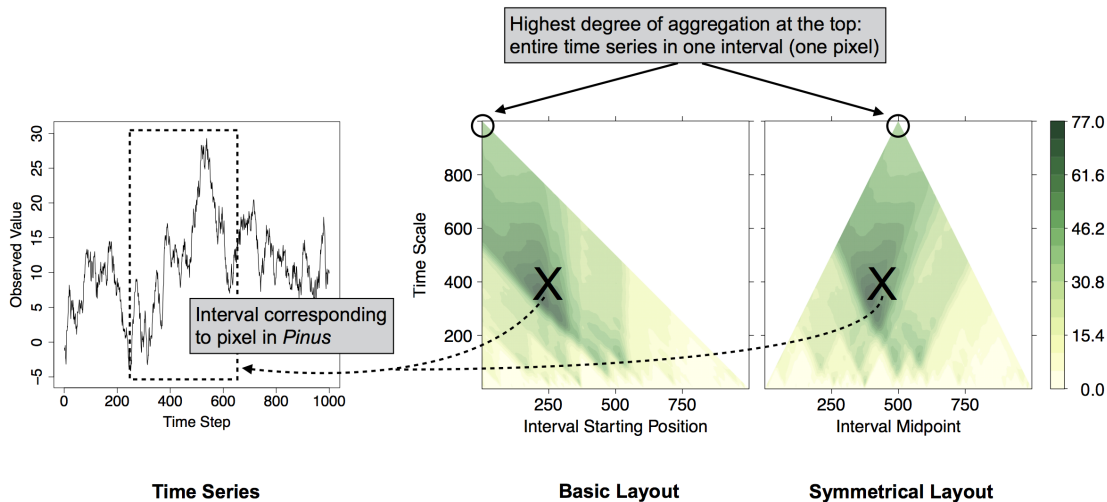
Fig. 2. Principle visual layouts of a *Pinus* view. This figure illustrates the relation of *Pinus* pixels to the temporal behavior of the original time series. The *Pinus* view depicts the variation of the time series across all scales and starting points. Color indicates variance values of subsequences.

the symmetrical *Pinus* view to the corresponding interval of the original time series.

The pixel space for time series with N time steps is N^2. If this exceeds the available screen space, similar options as described in Section 4.3.2 can be applied to render a *Pinus* view of large time series. For the Antarctica time series in Figure 4, the sampling rate for starting positions and interval length is 300. The initial window size, i.e., the initial time scale, is 1000. To further support users in the exploration of large time series, we provide a 'time zoom' that allows them to zoom into regions of interest. The 'time zoom' is similar to a magnifying glass and allows users to inspect finer starting positions or scales. Please note, depicting a *Pinus* view for large time series also becomes feasible with the increasing availability of giga-pixel displays.

4.4.2 Detection and Interpretation of Interesting Patterns

Figure 2 shows the visual layout of a *Pinus* view. A *Pinus* view presents a triangular-shaped multiscale graphical representation of a given time series. Each pixel in a *Pinus* view represents a statistical quantity attributed to some specific interval of the given time series. In the example in Figure 2 it depicts the statistical quantity variance.

Users can interpret a *Pinus* view in two ways. For specificity let us assume that we use the basic layout. The first strategy is to read a *Pinus* view from left to right to locate potentially interesting patterns. Changes in the statistical quantity become visible by color that varies along the horizontal direction. The second strategy is to read a *Pinus* view along the vertical direction. The pixel at the very top represents the variance of the entire time series. Continuing from there in downward direction allows users to study a time series from a global to a micro scale. The pixels in each vertical line represent an interval with a certain starting position. In the symmetrical layout, where intervals are mapped to their midpoints on the time axis, vertical lines represent intervals with the same midpoint. Albeit the two layouts are different, they both support users to locate detected patterns in the original time series.

A typical task is to identify subsections with high fluctuations, their time scale and their starting positions. Figure 3 illustrates how the answers to these questions can be easily read off. The length of the time series is 1000; the statistical measure depicted in the *Pinus* view (Figure 3, left) is variance. The *Pinus* view reveals two regions, A and B, with particularly high variances. Reading off the changes in color along the vertical direction, it becomes obvious that region B represents a much stronger trend than region A. The *Pinus* view also shows that the high variance values in region A occur on a much longer time horizon than in region B. Looking at the original time series (Figure 3, right), we can easily relate region B to region b in the original time se-

ries, where the observable shows a strong negative trend between time steps 600 and 800. The explanation of region A is not that obvious at a first glance. The large variances indicated by A are induced by a steady rise in the observed values between the time steps 1 and 600 (region a in Figure 3, right).

Please note, we have chosen a rather small time series as a didactic example. The temporal characteristics can easily be detected in the line plot – down to the time scales that are visually resolved. This, however, will not be the case for long environmental time series that contain many thousands of time steps. Of course, the *Pinus* view can be utilized more generally for depicting all kinds of variables that can be associated to intervals of time series.

In addition to identifying minima and maxima in a *Pinus* view, one can also distinguish isolines and enclosed regions. We gathered qualitative feedback to understand the interpretation of these features. It became quite clear that it is highly dependent on the specific application and task, and, therefore, requires domain knowledge. For example, in climatological time series, rather large regions of relatively intense color, e.g., representing entropy, may hint at significant climatological processes.

4.4.3 Selection of Interesting Time Scales

We received ambiguous user feedback regarding strategies for selecting appropriate time scales from the *Pinus* view. Again, it seems to be dependent on the application and task. Thus, we cannot provide clear-cut guidelines. One exemplary strategy that was applied in some cases is the following. Scientists start at the bottom of the *Pinus* view and scan the x-axis for comparatively rapid changes in the statistical measure – i.e., changes in color – that are followed by a rather long subsequence of either relatively high or low values. Once they have visually detected such a region in the *Pinus*, they scan upward along the y-axis to select the largest time scale that still exhibits this feature. The rationale behind this strategy is that sometimes the same important environmental processes become visible at different time scales. In such a case, scientists prefer a larger time scale because it reduces the amount of data and ignores insignificant information.

4.5 Interactive Exploration

Figure 4 shows a typical analysis scenario. The *Pinus* view is the starting point for detecting potentially interesting patterns (see A). Our approach supports users to visually inspect and evaluate detected patterns (see B). An important feature of our analysis tool is that users can easily formulate queries against the time series data to derive different visual views of detected patterns (see C). The tool provides an intuitive visual mechanism to formulate queries through interactive se-
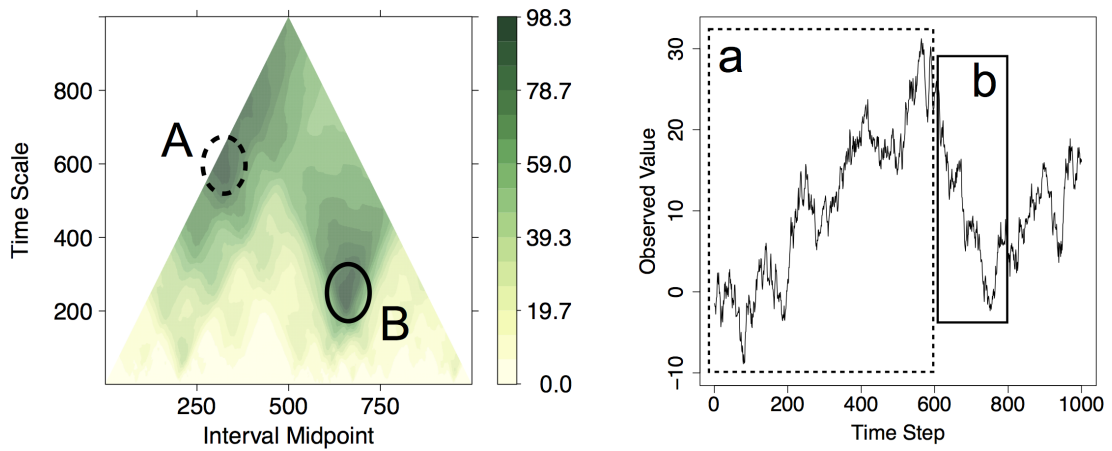
Fig. 3. An exemplary time series and its *Pinus* view. Users can identify regions *A* and *B* with variance values in the *Pinus* view and their corresponding temporal patterns *a* and *b* in the time series. Color indicates variance values of subsequences.

lections. Users can derive versions of the time series at different time scales by defining a horizontal line brush in *Pinus*. The selected time range depends on starting and end point of the line brush, which can span a single interval, any subsequence, or the complete time series. To support effective visual analysis, our tool provides immediate visual feedback to users by presenting the results of the visual queries in additional visual components. In the current prototype, we provide dot and line plots, as well as heat-maps. Adapting time scale, starting position, and size of the inspected subsequence in these additional visual components in *C* in turn affects the *Pinus* view, which visually highlights the currently inspected time scale and subsequence. This provides a flexible mechanism to inspect individual time scales at various levels of detail.

4.6 Data Quality

Data quality is an important issue in the analysis of environmental systems. Data is often subject to noise due to errors at various stages of measurement. Our approach is robust against noisy temporal data. Since noise usually increases with frequency, it becomes visible through large variations between adjacent cells along the horizontal direction, e.g., when using mean as statistical quantity. Another data quality aspect are missing values. At finer scales, missing values at a particular time step lead to missing values of the statistical quantity which are then visually encoded as empty cells. Thus, the resulting *Pinus* view shows visual gaps at fine scales. At coarser scales, sometimes only a few values contribute to the statistical quantity; this increases the uncertainty of the computed quantities. Please note, computed quantities are not wrong but more uncertain in comparison with other subsequences.

5 USE CASES

In this section, we discuss two use cases from geoscientific domains: analysis of the Earth's climate system and ocean modeling. These use cases represent different stages of collaboration with geoscientists. In the first use case, we explain how our analysis tool enabled them to gain new insight into the dynamics of the Earth's climate system. The results of the second use case represent an early stage of collaboration.

5.1 Regime Changes of the Earth's Climate System

An important scientific aim of climate research is to gain insight into the dynamics of the Earth's climate system. Our task here is to support the collaborators in identifying regime changes in the Antarctic temperature of the past. In the following, we explain how we achieved this and how we exceeded expectations of our collaborators.

5.1.1 Antarctic Temperature of the Past

Time series of the global temperature of the past are a valuable source of information on the dynamics of the Earth's climate system. The climate is continuously changing on very different time scales. Climate scientists assume a significant fluctuation in temperature values between ice ages and greenhouse periods. These fluctuations in temperature values could be a robust indicator for regime changes of the climate system. A major scientific challenge in climate research projects is to identify fluctuations in temperature values as regime changes. The time scales on which regime changes occur constitute important scientific insights. Many external factors such as the solar cycle drive the Earth's climate system. Well-understood time scales of regime changes will allow scientists to relate them to such external factors. A successfully established correlation between the time scale of regime changes and external factors will facilitate a better understanding of the complex dynamics of the Earth's climate system.

5.1.2 Data: Glacial Climate Record

We use glacial climate record data derived from an ice core from Dronning Maud Land, Antarctica [23]. The ice core represents South Atlantic temperature of the past 150k years. Scientists determine the age of the different layers of the ice core by measuring the occurrence of the isotope ^{18}O in the ice. A common measure of the temperature is the so-called $\delta^{18}O$ value, representing a normalized ratio of ^{18}O and ^{16}O isotopes. This analysis method usually leads to uneven temporal sampling rates. In our case the time series data had a dense sampling of temperatures for the past 20k years, and a sparse sampling rate for the years 100k to 150k before present. With our approach we could easily handle the uneven sampling rate, which is a challenge for many other approaches. The sample size of the Antarctica time series is 4601 data points. We will refer to this glacial climate record as the 'Antarctica time series' throughout this section.

5.1.3 Task: Detect Robust Indicators of Regime Changes

The task is to support scientists in detection of regime changes as well as their time scale and timing. In the following, we report how scientists utilized our tool to detect regime changes.

5.1.4 Characterize Temporal Fluctuations

Scientists used variance and entropy as statistical measures. In the initial analysis steps, it became apparent that entropy is superior for solving the task. Here the term entropy refers to the discrete entropy of the continuous temperature domain.

We compute the entropy of the Antarctica time series according to the following algorithm.

1. **Global quantization of the temperature domain.** The temperature domain is divided into a number of bins with width Δ.
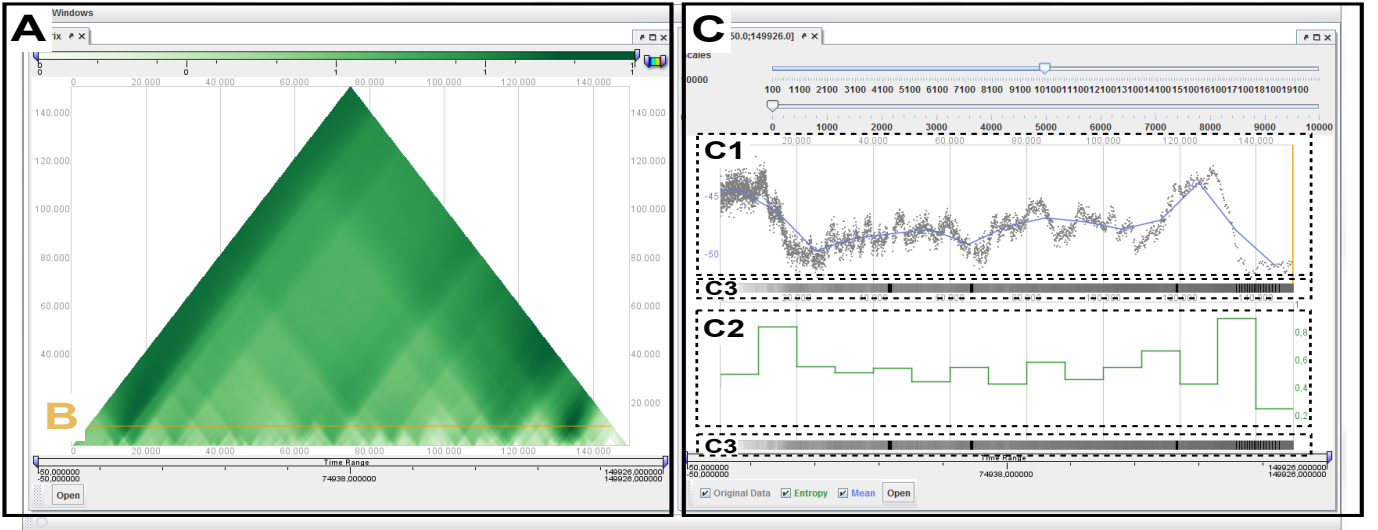
Fig. 4. The visual interface of the prototype. The figure shows the Antarctica time series. The panel *A* (left) is the *Pinus* view of the Antarctica. The *Pinus* view shows the variation of the entropy for many time scales and time steps. The color palette goes from white (zero entropy) to dark green (max. entropy). The slider at the bottom supports zooming in time. Scientists can select a time scale directly in the *Pinus* view (*B*). Panel *C* (right) shows the result of the query. The sliders in *C* are the starting position and time scale of subsequences. Panel *C*1 displays the data points of Antarctica time series (dots), and the Antarctica time series at 10k year time scale (line). Panel *C*2 shows the distribution of the entropy at 10k year time scale. Since the 10k year time scale has been chosen, the chart displays 15 entropy values. *C*1 and *C*2 are aligned along a common time axis; each entropy value covers a 10k years horizon. The heatmap (C3) shows the density of data points at a particular time step. Black color represents missing values.

2. **Probability density estimation.** For a given interval $s(p, w)$, the probability density $f(x_i)$ is computed, based on the global quantized temperature domain, as bin count, that is, the number of temperature values y_i falling into the same bin. Let v_k the bin count of the k-th bin, then $f(x_i) = v_k/N$. Note that $\Delta f(x_i)$ are the building blocks of a density histogram. This density histogram is completely determined by the subsequence $s(p, w)$ and the choice of the global Δ (see [25] for further readings).

3. **Compute entropy.** The discrete entropy is calculated for each interval $s(p, w)$ by computing the discrete entropy of the corresponding histogram from step 2.

Glacial climate in the circum-Antarctic regions has a millennial variability in temperature with amplitudes of $1 - 3°C$. We decided to compute the entropy using a bin width $\Delta = 2°C$.

5.1.5 *Pinus*: Detect Strong Temperature Fluctuations

The *Pinus* view (Figure 4) reveals a significant difference in the entropy values around the 10k year time scale in comparison to other time scales and intervals. The high entropy values around the 10k year time scale indicates high disorder in the distribution of the temperature values. The discrepancy in the entropy values across all possible time scales is interesting, since it suggests that these fluctuations are regime changes and that they occur at the 10k years scale.

5.1.6 Interactive Exploration

Scientists conducted several visual queries against the Antarctica time series to understand under which conditions the disorder in the temperature values at the 10k years scale occurred. Figure 4 (*B*) shows a query that revealed interesting temporal features. Panel *C* (right) shows the result of this query. Panel *C*1 displays the data points of Antarctica time series (dots), and the Antarctica time series at 10k year time scale (line). Panel *C*2 shows the distribution of the entropy at 10k year time scale. Note, the chart displays 15 entropy values which correspond to 10k year intervals. Since both views are aligned along a common time axis, it becomes obvious that high entropy values are correlated to temperature shifts from cool to a warm condition. Panel *C*2 shows a clear bi-polar pattern where the entropy values between the

two high-value poles fluctuate around 0.5. Scientists identified these low entropy intervals as an interglacial state that shows stable climate condition.

5.1.7 Insights

Our tool allowed our collaborators to establish scientific evidence for their initial assumption. First, the two high entropy values at the 10k years scale seem to be regime changes. Important indicators are the compact shape of high entropy regions along the horizontal direction, e.g., abrupt temperature fluctuations, and the persistence of the high entropy values across higher time scales. Second, our tool supported scientists to relate the high entropy values to the 10k-20k years before present interval and 130k-140k years before present interval, respectively. From the occurrence at these time intervals and from their temporal distance of about 100k years, our collaborators derived the hypothesis that the detected regime changes might be related to the 100k years cycle of the Milankovitch cycles. The 100k years cycle of the Milankovitch cycles describes the transition from a circular to an ellipsoidal orbit of the Earth around the Sun.

5.1.8 Significance of our Approach

Our collaborators use our analysis tool in their daily scientific work. The reported results are just the first findings; the next step in this collaboration is to extent the tool to compare time series of well-known factors that impact the Earth's climate system. Our collaborators plan to report interesting findings in their community.

5.2 Ocean Modeling

The aim of ocean modeling is to study or predict the behavior of the ocean because a direct observation or manipulation is difficult or even impossible. Ocean modeling allows scientists to (a) evaluate existing theories about different processes in the ocean by comparing the model output to measured data, and to (b) perform experiments which cannot be conducted in the real world.

Our collaboration with ocean modelers is at an early stage. As a first step, we decided to utilize our tool to explore well-understood observational data. In the following, we report the results of this test. Since with our tool anticipated patterns were successfully detected at

(a) Sea-level anomalies time series
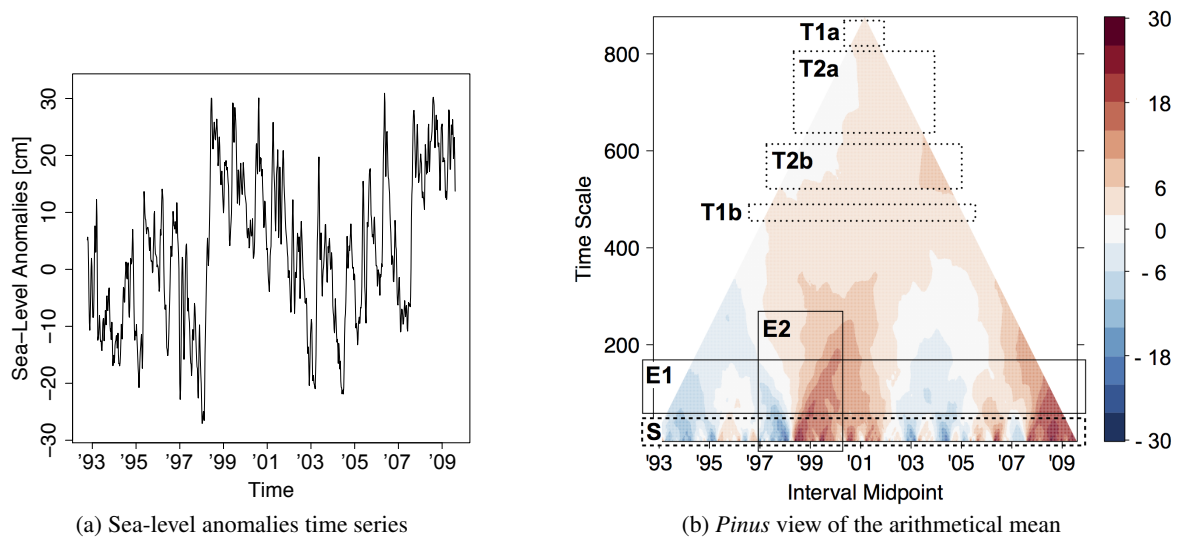
(b) *Pinus* view of the arithmetical mean

Fig. 5. Analysis of sea-level height anomalies from October 1992 through July 2009. Color indicates mean values of subsequences. A positive global trend becomes apparent at time scales in regions *T2a* and *T2b*. On the contrary, choosing the time scales in regions *T1a* or *T1b* for the analysis would not reveal this trend. Time scales in region *S* show a pronounced seasonal cycle. Further, the *Pinus* depicts El Niño and La Niña phenomena (region *E1*) and extreme sea surface heights (region *E2*).

multiple time scales, the next step in this collaboration is to utilize it for detection of patterns in the output of an ocean model.

5.2.1 Data: Sea-level Anomalies

The data represent sea-level anomalies of the Philippine Sea, obtained by a combination of satellite altimeters[1]. Sea-level anomalies are the deviation of the sea surface from a long-term mean. The time series comprises weekly sea-level records from October 1992 to July 2009 with 876 time steps in total.

5.2.2 Detect Anticipated Patterns at Multiple Time Scales

The Philippine Sea time series contains several characteristic patterns at varying time scales that have been extensively studied by geoscientists. The aim of this study was to evaluate, whether scientists are able to detect with the *Pinus* all of the following patterns:

- **A positive global trend.** In the considered time period, a clear sea-level rise has been observed.

- **A pronounced seasonal cycle**. The height of the ocean's surface is influenced by the temperature of the atmosphere. Warm air temperature induces a warming of the upper layer of the ocean. Since the density of water is inversely related to its temperature, a warm ocean has an increased volume. Therefore, sea surface heights generally increase with air temperature. Because the ocean reacts with some delay, the seasonal cycle can be split into a summer/autumn state and a winter/spring state.

- **Inter-annual variations.** The sea level of the Philippine Sea is strongly influenced by the El Niño Southern Oscillation (ENSO). ENSO is an oceanographic-meteorologic phenomenon which influences the sea level in the form of two distinct events. An El Niño event induces low sea surface heights while a La Niña event causes increased sea surface heights. These two events recur approximately every two to seven years with varying duration and intensity (for further reading, see [7]).

- **Extreme sea surface heights from 1997 to 2000.** These extremes were caused by a very strong El Niño event in 1997/98 followed by a rather long La Niña condition.

[1]http://www.aviso.oceanobs.com/duacs/ [access date 2010-04-15].

5.2.3 Results

Scientists selected the arithmetical mean as the statistical measure for the analysis. Figure 5 shows the resulting *Pinus* view along with the original time series. The *Pinus* exhibits distinct patterns on very different time scales. One can easily identify a positive global trend by studying the upper part of the overview visualization. Here, the colors progress from white to light orange. The trend is also an illustrative example of how the choice of time scale can influence the analysis outcome. The time scales in regions *T1a* and *T1b* in Figure 5 do not reveal a global trend while regions *T2a* and *T2b* clearly show it.

The small time scales in region *S* depict the seasonal cycle in the sea level data. Periods of rather low sea-level anomalies (winter/spring state) and periods of rather high sea-level anomalies alternate. The fluctuations between these two states appear to be of varying intensity. This is caused by a strong influence of El Niño/La Niña events which induce high or low sea-levels, respectively. This impacts the arithmetical mean. Therefore, El Niño and La Niña patterns can be detected at larger time scales (see region *E1*). The extreme sea-level heights from 1997 to 2000 can also be identified in the *Pinus* view (Figure 5, region *E2*). A short but very intense period of relatively low sea level (1997/98 El Niño) is followed by a long period of rather high sea level (La Niña condition). We can also observe the tremendous influence of these extreme events on larger time scales.

In summary, our approach enabled ocean modelers to easily detect a wide range of anticipated patterns in the sea-level anomalies time series: strictly periodic recurrences, sporadic recurrences, trends, and outliers. These results encourage the application of our approach to actual ocean model output.

6 SUMMARY AND CONCLUSION

We proposed a multiscale analysis approach that captures interesting patterns in environmental time series. To design and implement a versatile, application-oriented and field-ready tool, we conducted informal interviews with geoscientists. Based on their feedback, we derived four important design criteria. The tool should offer (i) computation of (user-selectable) statistical quantities that characterize the temporal behavior across all time scales and starting positions of patterns, (ii) visual detection of potentially interesting patterns, (iii) specification of visual queries, and (iv) immediate visual feedback in multiple linked views to facilitate an efficient interactive exploration of patterns at different time scales via linked views.

To meet these requirements, we utilize a matrix-like visualization to

display the statistical quantities of the hierarchical decomposition of a environmental time series. Therefore, we decided to compute rather basic statistical measures for (up to) all intervals of the time series and display them in an orderly way. For visualization of the resulting matrix we proposed a compact and easy-to-interpret visualization, called *Pinus*. A *Pinus* view presents the variation of the user-chosen statistical measure across time scales and starting positions of intervals.

We demonstrated the versatility of the *Pinus* view in pattern mining (detection of patterns, determination of the time scale and position of potentially interesting patterns). We discussed two possible layouts of a *Pinus* view: basic layout and symmetrical layout. An evaluation of the differences between the two layouts regarding perception and interaction, as well as the implications for effective visualization is left for future work. We also plan to identify additional effective mappings, e.g., for different analysis tasks or symbolic time series, utilizing known matrix visualization techniques. It is obvious that the *Pinus* view can be used for display of every statistical and non-statistical signature associated to intervals of time series. A drawback of our approach, in comparison to other visualization methods, is the quadratic pixel space needed to render a *Pinus* view.

We explained how our approach enabled scientists to identify subsections of time series as regime changes of the Earth's climate system. Our tool allowed scientists to establish sound scientific grounding that the regime changes are correlated with Milankovitch cycles. These results demonstrate the significance of the proposed approach to the analysis of environmental systems. However, a more extensive evaluation needs to be done to derive general guidelines on how to detect patterns in a Pinus view. The next step is a longitudinal user study to gather empirical evidence on the effectiveness and efficiency of the method.

Overall, the research presents a first step to a Visual Analytics approach that supports users in detecting hidden patterns in large environmental time series. However, our approach is not limited to environmental science. Since it makes no specific assumptions and only very fundamental statistical quantities are depicted, the proposed approach is generic. The method can be applied to all kinds of time series. Whether this reveals interesting features depends on the application. The next step of this research project aims at algorithms and visualization techniques that allow our collaborators to explore correlations between time series, as well as analysis of truly massive time series.

7 ACKNOWLEDGMENTS

REFERENCES

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering ICDM 1995*, pages 3–14, 1995.

[2] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of time-oriented data*. Springer, London, 2011.

[3] L. Berry and T. Munzner. Binx: Dynamic exploration of time series datasets across aggregation levels. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pages 5–6. IEEE Computer Society, 2004.

[4] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman. Interactive pattern search in time series. In *Proceedings of Conference on Visualization and Data Analysis, VDA 2005*, pages 175–186. SPIE, 2005.

[5] S. K. Card, J. D. Mackinlay, and B. Shneiderman, editors. *Readings in information visualization: Using vision to think*. Morgan Kaufmann, San Francisco and CA, 1999.

[6] C. H. Chen, H. G. Hwu, W. J. Jang, C. H. Kao, Y. J. Tien, S. Tzeng, and H. M. Wu. Matrix visualization and information mining. In *Proceedings in Computational Statistics 2004 (Compstat 2004)*, pages 85–100. Physika Verlag, Heidelberg, 2004.

[7] H. F. Diaz and V. Markgraf. *El Niño and the southern oscillation: Multiscale variability and global and regional impacts*. Cambridge University Press, Cambridge, 2000.

[8] G. M. Draper, Y. Livnat, and R. F. Riesenfeld. A survey of radial methods for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 15(5):759–776, 2009.

[9] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid. Periodicity detection in time series databases. *IEEE Transactions on Knowledge and Data Engineering*, 17(7):875–887, 2005.

[10] J. Han, G. Dong, and Y. Yin. Efficient mining of partial periodic patterns in time series database. In *Proceedings of the 15th International Conference on Data Engineering*, pages 106–115, 1999.

[11] M. C. Hao, U. Dayal, D. A. Keim, and T. Schreck. Multi-resolution techniques for visual exploration of large time-series data. In *EuroVis07: Joint Eurographics - IEEE VGTC Symposium on Visualization*, pages 27–34. Eurographics Association, 2007.

[12] M. C. Hao, M. Marwah, H. Janetzko, U. Dayal, D. A. Keim, D. Patnaik, N. Ramakrishnan, and R. K. Sharma. Visual exploration of frequent patterns in multivariate time series. *Information Visualization*, 11(1):71–83, 2012.

[13] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004.

[14] O. Hoeber, G. Wilson, S. Harding, R. Enguehard, and R. Devillers. Exploring geo-temporal differences using gtdiff. In *Proceedings of the 2011 IEEE Pacific Visualization Symposium*, PACIFICVIS '11, pages 139–146. IEEE Computer Society, 2011.

[15] W. Javed and N. Elmqvist. Stack zooming for multi-focus interaction in time-series data visualization. In *IEEE Pacific Visualization Symposium PacificVis 2010*, pages 33–40. IEEE Computer Society, 2010.

[16] D. F. Jerding and J. T. Stasko. The information mural: A technique for displaying and navigating large information spaces. *IEEE Transactions on Visualization and Computer Graphics*, 4(3):257–271, 1998.

[17] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2000.

[18] E. Keogh, S. Lonardi, and B. Chiu. Finding surprising patterns in a time series database in linear time and space. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 550–556. ACM, 2002.

[19] R. Kincaid and H. Lam. Line graph explorer: scalable display of line graphs using focus+context. In *Proceedings of the working conference on Advanced visual interfaces*, AVI '06, pages 404–411, New York, NY, USA, 2006. ACM.

[20] O. D. Lampe and H. Hauser. Curve Density Estimates. In *EuroVis11: Eurographics/ IEEE Symposium on Visualization*, pages 633–642. Eurographics Association, 2011.

[21] J. Lin, E. Keogh, and S. Lonardi. Visualizing and discovering nontrivial patterns in large time series databases. *Information Visualization*, 4(2):61–82, July 2005.

[22] P. McLachlan, T. Munzner, E. Koutsofios, and S. North. Liverac: interactive visual exploration of system management time-series data. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 1483–1492, New York, NY, USA, 2008. ACM.

[23] E. C. Members. One-to-one coupling of glacial climate variability in greenland and antarctica. *Nature*, 444(7116):195–198, 2006.

[24] A. Mueen, E. J. Keogh, Q. Zhu, S. Cash, and M. B. Westover. Exact discovery of time series motifs. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009*, pages 473–484, 2009.

[25] D. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1992.

[26] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages VL '96*, pages 336–343. IEEE Computer Society Press, 1996.

[27] C. Tominski and H. Schumann. Enhanced interactive spiral display. In *Proceedings of the Annual SIGRAD Conference, Special Theme: Interactivity*, pages 53–56. Linköping University Electronic Press, 2008.

[28] A. Tuchman and M. W. Berry. Matrix visualization in the design of numerical algorithms. *INFORMS Journal on Computing*, 2(1):84–92, 1990.

[29] M. Weber, M. Alexa, and W. Müller. Visualizing time-series on spirals. In

K. Andrews, editor, *Proceedings of the IEEE Symposium on Information Visualization 2001*, pages 7–13. IEEE Computer Society, 2001.

[30] L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[31] H.-M. Wu, S. Tzeng, and C. houh Chen. *Matrix Visualization in Handbook of Data Visualization*, pages 681–708. Springer Handbooks of Computational Statistics. Springer, 2008.

[32] J. Yang, W. Wang, and P. S. Yu. Mining asynchronous periodic patterns in time series data. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):613–628, 2003.

[33] I. Zaliapin, A. Gabrielov, and V. Keilis-Borok. Multiscale trend analysis. *Fractals*, 12(3):275–292, 2004.

[34] X. Zhang, D. E. Shasha, Y. Song, and J. T. L. Wang. Fast elastic peak detection for mass spectrometry data mining. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):634–648, 2012.

[35] J. Zhao, F. Chevalier, and R. Balakrishnan. Kronominer: using multi-foci navigation for the visual exploration of time-series data. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 1737–1746. ACM, 2011.