Novel Algorithm for Mining ENSO-Oriented Marine Spatial Association Patterns from Raster-Formatted Datasets

Xue Cunjin [1,2] and Liao Xiaohan [3,*]

[1] Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth,
    Chinese Academy of Sciences, Beijing 100094, China; xuecj@radi.ac.cn
[2] Key Laboratory of the Earth Observation, Beijing 100094, China
[3] Institute of Geographical Science and Natural Resources Research, Chinese Academy of Sciences,
    Beijing 100101, China
[*] Correspondence: liaoxh@igsnrr.ac.cn; Tel.: +86-10-64888956

Abstract: The ENSO (El Niño Southern Oscillation) is the dominant inter-annual climate signal on Earth, and its relationships with marine environments constitute a complex interrelated system. As traditional methods face great challenges in analyzing which, how and where marine parameters change when ENSO events occur, we propose an ENSO-oriented marine spatial association pattern (EOMSAP) mining algorithm for dealing with multiple long-term raster-formatted datasets. EOMSAP consists of four key steps. The first quantifies the abnormal variations of marine parameters into three levels using the mean-standard deviation criteria of time series; the second categorizes La Niña events, neutral conditions, or El Niño events using an ENSO index; then, the EOMSAP designs a linking–pruning–generating recursive loop to generate ($m + 1$)-candidate association patterns from an m-dimensional one by combining a user-specified support with a conditional support; and the fourth generates strong association patterns according to the user-specified evaluation indicators. To demonstrate the feasibility and efficiency of EOMSAP, we present two case studies with real remote sensing datasets from January 1998 to December 2012: one considers performance analysis relative to the ENSO-Apriori and Apriori methods; and the other identifies marine spatial association patterns within the Pacific Ocean.

1. Introduction

The ENSO (El Niño Southern Oscillation) is the dominant year-to-year climate signal on Earth, a cycle of alternating warm El Niño and cold La Niña events. Its relationships with marine environments constitute a complex interrelated system [1], e.g., during a warm phase of the ENSO, a positive sea surface temperature (SST) and a negative sea level anomaly co-occur in the eastern tropical Pacific Ocean [2], and in a cool phase, an abnormal increase in dry conditions occurs over the Pacific Ocean, SSTs drop in the central Pacific Ocean, and SST gradients increase in the east-west Pacific Ocean, which indirectly dominate primary productivity and sea surface chlorophyll-a concentrations in the Pacific Ocean [3]. In recent decades, raster-formatted datasets derived from remote sensing images, reanalysis products and numerical simulations have provided an important source of data and offer new opportunities to improve our understanding of these relationships on a large scale [4,5].

To obtain such association patterns against the ENSO from raster-formatted datasets, conventional methods of spatiotemporal analysis first extract the spatial and temporal characteristics by applying

empirical orthogonal functions [6], canonical analysis [7], or singular value decomposition [8], and then, their correlations with the ENSO are investigated by statistical analysis or wavelet spectrum analysis [2,3]. Although these methods can obtain the spatial distribution and temporal characteristics of the marine parameters when ENSO events occur, they have two bottlenecks, i.e., one is that few quantitative relationships among geographical parameters have been obtained, and the other is that great challenges exist in exploring the spatial association patterns among multiple geographical parameters. Compared with these conventional methods, an inductive spatiotemporal data mining technique shows more promise for discovering association patterns among multiple geographic parameters [9]. In recent decades, this method has gained attention as a means of understanding these relationships using the raster-formatted datasets [10–12].

When using data mining techniques to address the marine association characteristics against the ENSO from raster-formatted datasets, we need to resolve two issues. One is improving the efficiency of mining algorithms, and the other is retaining the most spatial information during the mining process.

Regarding association pattern mining algorithms, regardless of whether they are Apriori-type or non-Apriori, a core idea is to find frequent itemsets from transactions by applying a support threshold, and the computational complexity depends on the number of times the database is scanned. To reduce the number of database scans, many algorithms have been developed in recent decades. Considering that database scans mostly depend on the numbers of frequent 1-itemsets, mutual information is used to pre-extract the pair-wise related items and to then find all frequent itemsets [13,14]. By first filtering the unrelated 1-itemsets, the mutual-information-based algorithms greatly reduce the number of database scans and thus improve the implementation efficiency [15]. Some other documents also reduce the number of database scans by predefining an efficient data structure, e.g., Tsay and Chiang created cluster-based tables to find frequent itemsets, and for k-frequent itemsets, the number of database scans is less than k [16]; Wu and Huang (2011) defined the frequent closed enumeration table to store maximal itemsets to reduce database scans [17]; and Liu et al. (2012) used the directed itemsets graph to store the information of frequent itemsets, which realizes scanning a database only once [18]. In addition, for dealing with raster-formatted datasets, the above mining algorithms are extended with spatial regions, e.g., the spatial clusters [19], object-oriented technologies [20–23], and event-coverage domains [24]. Since large numbers of grid pixels are replaced by typical regions and thereby simplify the mining process, however, these techniques result in a loss of large amounts of spatial information.

To reduce the loss of spatial information during a mining process, an approach considers each grid pixel as an independent time series and mines an association pattern one-by-one for each grid pixel. For example, Julea et al. (2011) proposed a grouped frequent sequence pattern mining algorithm for agricultural monitoring, which was aimed at extracting an evolution of each grid pixel with time series images [25]; Romani et al. (2013) developed a RemoteAgri system to discover the Plateau–Valley–Mountain (P–V–M) association patterns for monitoring sugar cane fields with time series of remote sensing images and found that the P–V–M pattern mainly analyzed the association patterns between two geographical parameters [26]; and Saulquin et al. (2014) designed an event-based mining algorithm for dealing with SST anomalies relative to ENSO events, which considers each one-dimensional time series as a series of significant time-scale events for each grid pixel [12]. Generally, each grid pixel may have several patterns, and each pattern may evolve several geographical parameters; therefore, the complicated association patterns from remote sensing images make it impossible for a user to analyze an entire set and find the most interesting ones [27].

Generally, these mining algorithms are able to obtain the marine association patterns against the ENSO. However, as they treat the items equivalently, i.e., do not consider a core idea of one given item, these algorithms have great potential for improving the mining efficiency. In addition, taking one given item as a core, the mining algorithm may easily visualize and find the interesting spatial association patterns from raster-formatted datasets. Thus, the motivations behind this manuscript lie in two aspects: the spatial association patterns among multiple marine environmental parameters against ENSO events are more complicated, and the ENSO is taken as a core item to simplify the

mining process and to then improve the mining efficiency. The proposed novel algorithm aims at exploring marine spatial association patterns using long-term time series of raster-formatted datasets, which we called an ENSO-Oriented Marine Spatial Association Pattern mining algorithm (EOMSAP). The remainder of this manuscript is organized as follows. Section 2 describes the workflow for the EOMSAP. In Section 3, two experiments using real image datasets are described. One tests the advantages and disadvantages of our proposed algorithm, and the other proves the soundness of our work with the discovered marine spatial association patterns against ENSO events in the Pacific Ocean. Section 4 presents our discussions and conclusions.

2. EOMSAP with Raster-Formatted Datasets

2.1. Algorithm Workflow

In this manuscript, marine spatial association patterns refer to an abnormal variation of one to several marine parameters against ENSO events in a specified spatial region. The design of EOMSAP with raster-formatted datasets includes the extraction and representation of abnormal variations of marine parameters from long-term time series, the identification of El Niño and La Niña events from ENSO indices, the recursive algorithm with linking and pruning functions, and the generation of marine spatial association patterns. Figure 1 shows the detailed workflow of our proposed algorithm.
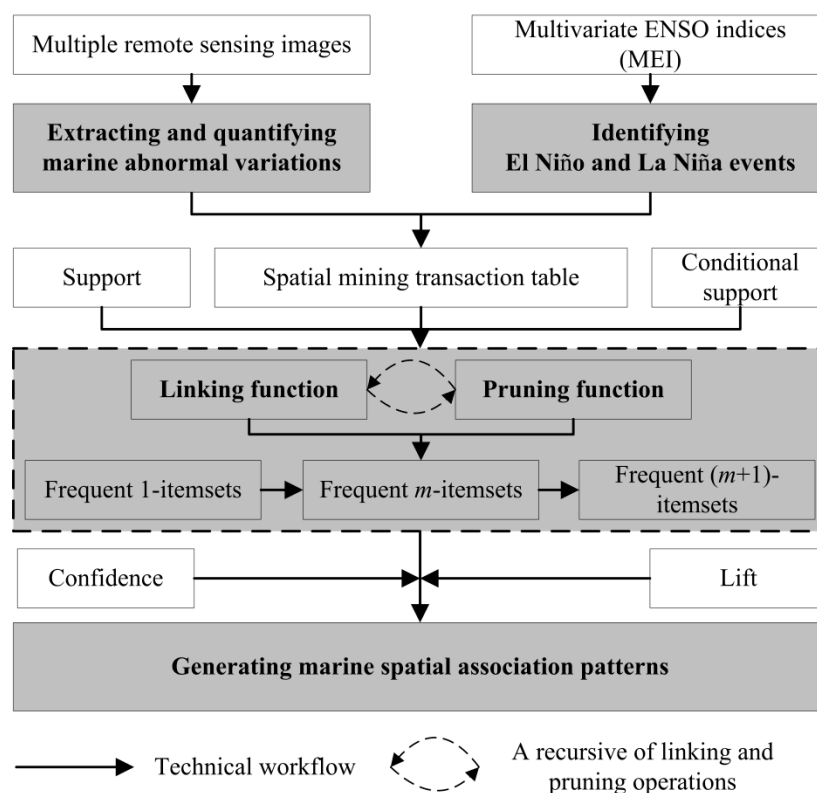

Figure 1. Workflow of the proposed algorithm. The four key steps are indicated by gray shading.

2.2. Quantization of Marine Abnormal Variations from Raster-Formatted Datasets

An abnormal variation is a deviation from an averaged status obtained from a long-term series (e.g., daily, monthly, seasonally or yearly). Obviously, long-term marine parameters have seasonal variations that are mainly dominated by solar radiances. However, against the background of global climate change, spatiotemporal patterns that deviate from normal seasonal cycles are of particular interest in anomalous climate event analysis. With little prior knowledge, the *z-score* algorithm is more suitable for removing seasonal fluctuations [28].

Both quantitative and Boolean mining are unable to address continuous values. Therefore, before carrying out the rule mining, the abnormal variations need to be quantified into continuous intervals, which are used to represent the intensities of variations [15,28]. Many quantization strategies are available (e.g., cluster-based, equal-density, equal-area, or equal-depth methods), but very often, they are closely related to the specific domain [14]. In this manuscript, our goal is to discover the abnormal association relationships among marine parameters against global climate change, and the quantization algorithm should describe the intensities of abnormal variations. The mean-standard deviation of the time series was a criterion used to quantify the marine parameters into three ranks, −1, 0 and +1, indicating negative changes, no changes and positive changes, respectively. For a specified grid pixel, i.e., the ith row and jth column in a raster-formatted dataset, the formula is shown as Equation (1).

$$f_{Rank}(V) = \begin{cases} +1; & V \geq \mu + \delta \\ 0; & -(\mu + \delta) < V < \mu + \delta \\ -1; & V \leq -(\mu + \delta) \end{cases} \tag{1}$$

where μ and δ are the mean and standard deviation values of the time series in the specified grid pixel (i^{th} row and j^{th} column), respectively, and V is the abnormal variation at a given time in the specified grid pixel.

From long-term raster-formatted datasets, the quantification of abnormal marine parameter variations consists of the following steps:

1. Step 1: Calculate the mean and standard deviation of the time series' real values of marine parameters from long-term raster-formatted datasets.
2. Step 2: Extract the abnormal variations of marine parameters using the *z-score* algorithm.
3. Step 3: Calculate the mean and standard deviation values on the basis of long-term abnormal variations of marine parameters.
4. Step 4: Quantify the abnormal variations into continuous intervals (i.e., −1, 0 and +1), using Equation (1) for each time and each grid pixel.

2.3. Identification of ENSO Events

There are many indices that describe ENSO events, including the Southern Oscillation Index; anomalies of SST in El Niño region 12 (90° W~80° W, 10° S~0°), region 3 (150° W~90° W, 5° S~5° N), and region 4 (160° E~150° W, 5° S~5° N) [29]; the Multivariate ENSO Index (MEI); the Oceanic Niño Indices [30]; and the precipitation-based ENSO index [31]. In this study, we used the MEI (http://www.esrl.noaa.gov/psd/enso/mei/), provided by the U.S. National Oceanic and Atmospheric Administration's Earth System Research Laboratory Physical Sciences Division. It is based on six observed variables over the tropical Pacific: sea-level pressure, the zonal and meridional components of the surface wind, SST, surface air temperature, and the total cloudiness fraction of the sky [32].

Different percentile definitions are used to rank ENSO events as strong, moderate or weak [32]. However, using too many types would make it difficult to identify abnormal variations of marine parameters related to the ENSO. Considering the consistency of abnormal variations in marine parameters and ENSO events, the mean-standard deviation algorithm was used to catalog ENSO events into three ranks: −1, 0 and +1, which indicate a La Niña event, a neutral condition, and an El Niño event, respectively. The criteria are similar to Equation (1).

2.4. A Recursive Algorithm

Apriori is a seminal algorithm for finding frequent itemsets using candidate generation and is based on the three steps referred to as link–prune–generation [33]. Since its introduction and subsequent widespread application, the core idea of Apriori has been shared and improved in the development of quantitative relationship mining [34]. This manuscript uses the core idea of link–prune–generation to design the EOMSAP for exploring abnormal association patterns among

marine parameters against ENSO events. The key implementations consist of two steps. The database in these steps is composed of mining transaction tables.

Step 1: Generate the frequent 1-itemset related to the ENSO by scanning the database one time for each item (i.e., marine parameter) and each quantification type (i.e., -1, 0 and $+1$). Next, use Equation (2) to calculate support (S), denoted as $S(\mathrm{A}[k])$, and use Equation (3) to calculate conditional support (CS) against ENSO events, denoted as $CS(\mathrm{A}[k]||\mathrm{ENSO}[l])$. If and only if the inequalities in Equation (4) are true, the frequent 1-itemset related to ENSO is generated, denoted as $CS(\mathrm{A}_1[k_1]||\mathrm{ENSO}[l])$.

$$S(\mathrm{A}_1[k_1]\mathrm{A}_2[k_2]\cdots\mathrm{A_m}[k_m]) = \frac{n(\mathrm{A}_1[k_1]\mathrm{A}_2[k_2]\cdots\mathrm{A_m}[k_m])}{N} \times 100\% \tag{2}$$

$$CS(\mathrm{A}_1[k_1]\mathrm{A}_2[k_2]\cdots\mathrm{A_m}[k_m]||\mathrm{ENSO}[l]) = \frac{n(\mathrm{A}_1[k_1]\mathrm{A}_2[k_2]\cdots\mathrm{A_m}[k_m]\mathrm{ENSO}[l])}{n(\mathrm{ENSO}[l])} \times 100\% \tag{3}$$

$$\left.\begin{array}{l} S(\mathrm{A}_1[k_1]\mathrm{A}_2[k_2]\cdots\mathrm{A_m}[k_m]\mathrm{ENSO}[l]) \geq \tau_s \\ CS(\mathrm{A}_1[k_1]\mathrm{A}_2[k_2]\cdots\mathrm{A_m}[k_m]||\mathrm{ENSO}[l]) \geq S(\mathrm{A}_1[k_1]\mathrm{A}_2[k_2]\cdots\mathrm{A_m}[k_m]) \end{array}\right\} \tag{4}$$

where m is the number of items involved in the mining model, which goes from 1 to the total number of marine parameters (M). For one item, m is equal to 1, while for M items, m is equal to M; $n(\mathrm{A}_1[k_1]\mathrm{A}_2[k_2]\ldots\mathrm{A_m}[k_m])$ is the number of co-occurrences of items $\mathrm{A}_1, \mathrm{A}_2\ldots\mathrm{A_m}$ at level $k_1, k_2, \ldots, k_m$; $n(\mathrm{ENSO}[l])$ is the number of occurrences of an ENSO[l] event; $n(\mathrm{A}_1[k_1]\mathrm{A}_2[k_2]\ldots\mathrm{A_m}[k_m]\mathrm{ENSO}[l])$ is the number of co-occurrences of items $\mathrm{A}_1, \mathrm{A}_2\ldots\mathrm{A_m}$ at level $k_1, k_2, \ldots, k_m$ and the ENSO[l] event; $k_1, k_2, \ldots, k_m$ are one of the quantification types (i.e., -1, 0 and $+1$); l is the ENSO type (i.e., $+1$, El Niño and -1, La Niña); and τ_s is the user-specified threshold of marine parameters. The first inequality in Equation (4) means that only the variation type $k_1, k_2, \ldots, k_m$ of marine parameters $\mathrm{A}_1, \mathrm{A}_2\ldots\mathrm{A_m}$ and the ENSO[l] event satisfying the user-specified minimum support are meaningful. The second means that only when the supports of marine parameters $\mathrm{A}_1, \mathrm{A}_2\ldots\mathrm{A_m}$ at variation type $k_1, k_2, \ldots, k_m$ against an ENSO[l] event are not less than their support in the database are their co-variations of marine parameters regarded as association patterns against ENSO[l].

Step 2: Generate frequent ($m + 1$)-itemsets from candidate m-itemsets using a recursive algorithm with linking–pruning, where m is not less than 2. Within this step, the linking and pruning functions are run recursively until no more frequent itemsets are generated. The *Linking Function* generates the candidate ($m + 1$)-itemsets from the m-itemsets by step-by-step linking without scanning the database, while the *Pruning Function* removes the false ($m + 1$)-itemsets according to Equation (4).

For a clear description of the workflow finding frequent itemsets against ENSO events, we give an example with simulated data in Table 1.

Table 1. Quantitative data in a database for example 1.

	$\mathbf{A_1}$	$\mathbf{A_2}$	$\mathbf{A_3}$	$\mathbf{A_4}$	$\mathbf{A_5}$	**ENSO**
0	+1	+1	+1	0	+1	+1
1	+1	0	+1	0	+1	+1
2	+1	+1	0	+1	0	0
3	+1	+1	+1	+1	+1	0
4	+1	+1	+1	0	0	+1
5	+1	+1	+1	+1	+1	−1
6	+1	0	+1	0	+1	−1
7	0	+1	+1	+1	0	0
8	+1	+1	0	0	+1	0
9	0	+1	+1	+1	0	−1

Example 1: Table 1 shows quantitative change for five marine parameters ($\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_5$) and an ENSO event. The $+1$, 0 and -1 of the marine parameters mean positive changes, no changes and

negative change, respectively. The ±1 of the ENSO means an El Niño or La Niña event, respectively. In this case, the support threshold is set to 20.0%.

The supports of A_1, A_2, A_3, A_4 and A_5 with El Niño events are 30.0%, 20.0%, 30.0%, 0.0% and 20.0%, respectively. Their independent supports are 80.0%, 80.0%, 80.0%, 50.0% and 60.0%, and their conditional supports against El Niño are 100%, 66.7%, 100%, 0.0% and 66.7%, respectively. According to Equation (4), A_4 fails to meet the first inequality and A_2 fails the second inequality; thus, the frequent 1-itemsets are A_1, A_3 and A_5, denoted as (A_1[+1]|ENSO[+1], A_3[+1]|ENSO[+1] and A_5[+1]|ENSO[+1]). The *LinkingFunction* generates three candidate 2-itemsets, which are (A_1[+1]A_3[+1]|ENSO[+1], A_1[+1]A_5[+1]|ENSO[+1] and A_3[+1]A_5[+1]|ENSO[+1]). The *PruningFunction* verifies that they are all frequent 2-itemsets. Repeating the *LinkingFunction* and *PruningFunction* generates one frequent 3-itemset, which is A_1[+1]A_3[+1], A_5[+1]|ENSO[+1].

With similar processing, the frequent itemsets against a La Niña event include three 1-itemsets and two 2-itemsets, which are (A_3[+1]|ENSO[−1], A_4[+1]|ENSO[−1], and A_5[+1]|ENSO[−1], A_3[+1]A_4[+1]|ENSO[−1], and A_3[+1]A_5[+1]|ENSO[−1]).

2.5. Generating Meaningful Marine Spatial Association Patterns

In this step, the key issue is to determine which frequent itemsets are meaningful according to the minimum thresholds of the evaluation indicators. Generally, the specified thresholds are defined by users according to their research domains. For each frequent itemset, its evaluation indicators (e.g., confidence and lift) are calculated by scanning the database once. If the evaluation indicators satisfy the user-specified thresholds, a frequent itemset is meaningful.

In this manuscript, we use confidence and lift as evaluation indicators for generating meaningful marine spatial association patterns. *Confidence* describes the occurrence probability of marine abnormal variations ($A_1[k_1]A_2[k_2]\ldots A_m[k_m]$) assuming that an ENSO event occurs, which has the same formula as Equation (3).

Lift describes the impact on marine abnormal variations of the occurring ENSO event; that is, once an ENSO event has occurred, how much does the occurrence probability of marine abnormal variations change? *Lift* is defined as:

$$Lift(A_1[k_1]A_2[k_2]\cdots A_m[k_m]ENSO[l]) = \frac{n(A_1[k_1]A_2[k_2]\cdots A_m[k_m]ENSO[l]) \times N}{n(ENSO[l]) \times n(A_1[k_1]A_2[k_2]\cdots A_m[k_m]}$$ (5)

where $n(A_1[k_1]A_2[k_2]\cdots A_m[k_m]ENSO[l])$, $n(A_1[k_1]A_2[k_2]\cdots A_m[k_m])$, $n(ENSO[l])$ and N have similar meanings as in Equations (2) and (3).

3. Experiments

In this section, we present two case studies for marine environments using long-term time series of remote sensing products. One experiment illustrates the feasibility and effectiveness of EOMSAP compared with the Apriori algorithm. The other explores marine spatial association patterns against ENSO events in the Pacific Ocean. The performance of our proposed algorithm was evaluated against the quantitative Apriori algorithm [34], which extends the candidate generation procedure by using the interest measure to prune and uses a different data structure (i.e., hash-table and R-tree) to count candidates.

To better illustrate our proposed algorithm, we also developed the quantitative Apriori. The main revision is that only the frequent itemsets related to the ENSO, not all frequent ones, are obtained during the process of linking and pruning. The revised algorithm is denoted as ENSO-Apriori. All the algorithms described in this manuscript have been developed and integrated into the Marine Spatiotemporal Association Patterns Mining System (MarineSTAPMining) software, which is registered by the National Copyright Administration of P.R. China (No. 2014SR013444). The MarineSTAPMining software was developed by the authors and integrated with several association pattern-mining

algorithms, including EOMSAP, Apriori, quantitative Apriori, ENSO-Apriori, MIQarma [15] and FP-Tree [35]. To smooth out any variations, each experiment was run five times, and the average result was taken. The experimental hardware environment includes an Intel core i7-3520M CPU at 2.90 GHz with a 1-level cache memory of 0.5 MB and 2-level of 4 MB, a 500 GB hard disk, and 4.0 GB of memory.

3.1. Research Area and Datasets

Our study was conducted on long-term marine remote sensing products, including SST, sea surface chlorophyll-a, sea surface precipitation, and sea level anomaly. The MEI was used to identify the ENSO events. The Pacific Ocean from 100° E to 60° W and 50° S to 50° N was the research area, as shown in Figure 2. Areas CLSObj1 to CLSObj6 were used to test the EOMSAP's performance, and the entire research area was used to obtain the marine spatial association patterns. Table 2 summarizes the datasets used.
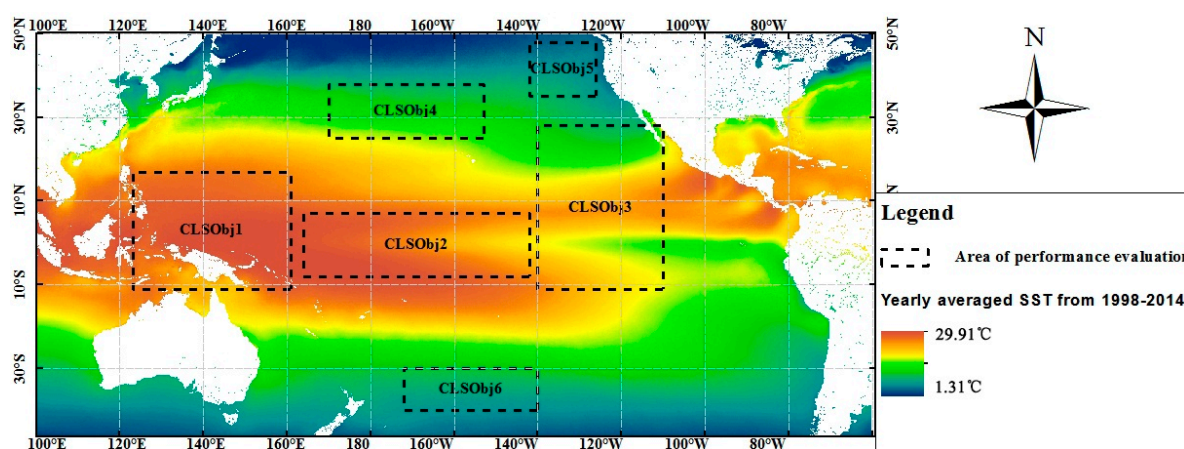

Figure 2. The research area. The background colors show the yearly averaged sea surface temperature (SST) from 1998 to 2014.

Table 2. Sources and resolutions of the remote sensing products and MEI used in this manuscript.

	Product	Source	Timespan (DD-MM-YY)	Temporal Resolution	Spatial Coverage	Spatial Resolution
1	SST [1]	NOAA/PSD	01-12-1981–30-04-2014	Monthly	Global	1° × 1°
2	Chl-*a* [2]	SeaWifs	01-09-1997–30-11-2010	Monthly	Global	9 × 9 km
		MODIS	01-07-2002–31-05-2014	Monthly	Global	9 × 9 km
3	Sea surface precipitation [3]	TRMM	01-01-1998–28-02-2014	Monthly	Global	0.25° × 0.25°
4	SLA [4]	AVISO	01-01-1993–31-12-2012	Monthly	Global	0.25° × 0.25°
5	ENSO [5]	MEI	01-01-1950–30-05-2014	Monthly	-	-

[1] Obtained from (http://www.esrl.noaa.gov/psd/) and provided by the National Oceanic and Atmospheric Administration (NOAA)/the Office of Oceanic and Atmospheric Research (OAR)/Earth System Research Laboratory (ESRL) Physical Sciences Division (PSD) [36]. [2] Obtained from the Sea-Viewing Wide Field-of-View Sensor (SeaWiFS) and Moderate Resolution Imaging Spectroradiometer (MODIS) projects and their level-three standard mapped images (SMI) [37]. [3] Obtained from Version 7 of the Tropical Rainfall Measuring Mission (TRMM Product 3B43), provided by the Goddard Distributed Active Archive Center (GES DISC DAAC), [4] Produced by Ssalto/Duacs and distributed by Archiving, Validation and Interpretation of Satellites Oceanographic Data (AVISO) with support from the National Centre for Space Studies. [5] Provided by NOAA-ESRL Physical Sciences Division [32]. SST = sea surface temperature; Chl-*a* = monthly anomaly of sea surface chlorophyll-a; SLA = sea level anomaly; ENSO = El Niño Southern Oscillation.

In both cases, to obtain uniform datasets from remote sensing products with the same spatial and temporal resolution, an analysis period of January 1998 to December 2012 was selected. Monthly anomalies of the research area elements with a spatial resolution of 1° in grid projection and with a temporal resolution of one month were calculated to remove seasonal effects. The resulting anomalies were denoted as SSTA (monthly anomaly of SST), CHLA (monthly anomaly of sea surface

chlorophyll-a), SLAA (monthly anomaly of sea level anomaly), and SSPA (monthly anomaly of sea surface precipitation).

3.2. Performance Evaluation and Analysis

Monthly image datasets for the six different regions, denoted as CLSObj1, CLSObj2, CLSObj3, CLSObj4, CLSObj5 and CLSObj6 in Figure 2, and the MEI are used to test the algorithms. Each region contains four parameters, i.e., SSTA, CHLA, SSPA and SLAA. Using Equation (1), each marine environmental parameter was quantified (1) at each region and (2) in each interval within the time period. The parameters were quantified as −1, 0 or +1, indicating negative change, no change or positive change, respectively. The MEI in each interval was quantified in the same manner. As noted previously, −1, 0 and +1 indicate a La Niña event, neutral condition, or El Niño event, respectively.

Apart from the data pretreatment, there are two factors responsible for the performance of the mining algorithms. One is the number of database scans, and the other is the computation time of each scan. The former is jointly determined by the minimum support threshold and the number of evolved items, i.e., the marine parameters, and the latter is determined by the database record sizes.

3.2.1. Computational Complexity

The computational complexity of EOMSAP is classified into two categories—the number of database scans and the intensive computing. Given the quantitative rank, R, and the total number of marine environmental parameters, M, R and M are used to calculate the computational complexity of mining algorithms.

EOMSAP scans a database in three stages. The first stage builds the frequent 1-items by scanning the database for each quantitative level for each item, i.e., $R \times M$. The computational complexity is $O(R \times M)$. The second stage generates all candidates of frequent items by a recursive loop of linking and pruning functions. According to the recursive algorithm, the number of database scans to generate the frequent 1-items related to ENSO is $R^2 \times C_{M-1}^1$, the frequent 2-items related to ENSO is $R^2 \times C_{M-1}^2$, and so on. Thus, the total number of database scans is $R^2 \times \left(C_{M-1}^1 + C_{M-1}^2 + \ldots + C_{M-1}^{M-2} \right)$, and the computational complexity is $O(R^2 \times M^{M-2})$. The third stage finds a meaningful cascading pattern by scanning the database once for each candidate pattern, that is, the total number of candidate patterns determines the computational complexity, which is $O(R \times M + R^2 \times M^2 + \ldots + R^{M-1} \times M^{M-1})$. According to the mining process of both ENSO-Apriori and Apriori, this manuscript also analyzes their computational complexities. Comparisons with EOMSAP show that during the first stage, these three algorithms have similar numbers of database scans; thus, they have the similar computational complexity, $O(R \times M)$. In the second stage, the ENSO-Apriori and Apriori have similar numbers of database scans, $R^2 \times \left(C_M^2 + C_M^3 + \ldots + C_M^{M-1} \right)$; thus, their computational complexity is $O(R^2 \times M^{M-1})$. In the third stage, EOMSAP and the ENSO-Apriori have similar computational complexity, and the computational complexity of Apriori is the largest one, being $O(R \times M + R^2 \times M^2 + \ldots + R^{M-1} \times M^{M-1} + R^M \times M^M)$.

The intensive computing mainly involves computing to find the frequent items from all the candidates and to generate the meaningful patterns from all the frequent items. The first depends on the total number of the candidates of frequent items, and its computational complexity is similar to the second stage of generating all the candidates of frequent items. The latter depends on the total number of frequent items, and its computational complexity is similar to the third stage of generating all the meaningful patterns. Table 3 shows the computational complexity of EOMSAP and gives its comparisons with ENSO-Apriori and Apriori.

Table 3. Comparisons of the computational complexities among ENSO-oriented marine spatial association pattern (EOMSAP), ENSO-Apriori and Apriori.

		EOMSAP	ENSO-Apriori	Apriori
Number of Database Scans	Build frequent 1-items		$O(R \times M)$	
	Find all frequent candidates	$O(R^2 \times M^{M-2})$	$O(R^2 \times M^{M-1})$	
	Generate all meaningful patterns	$O(R \times M + R^2 \times M^2 + \ldots + R^{M-1} \times M^{M-1})$	$O(R \times M + R^2 \times M^2 + \ldots + R^{M-1} \times M^{M-1} + R^M \times M^M)$	
Intensive Computing	Find all frequent items	$O(R^2 \times M^{M-2})$	$O(R^2 \times M^{M-1})$	
	Generate all meaningful patterns	$O(R \times M + R^2 \times M^2 + \ldots + R^{M-1} \times M^{M-1})$	$O(R \times M + R^2 \times M^2 + \ldots + R^{M-1} \times M^{M-1} + R^M \times M^M)$	

3.2.2. Numbers of Database Scans

Database scans are one of the most important factors affecting the efficiency of finding frequent items. Generally, the more evolved the marine parameters, the greater the number of database scans, and the smaller the support threshold, the greater the number of database scans. Unlike the Apriori and ENSO-Apriori algorithms, the EOMSAP embeds conditional support, instead of only user-specified support, to find the frequent itemsets from candidate ones. As the conditional support only considers the items on the preconditions of ENSO occurrence during the process of linking and pruning, the number of database scans of EOMSAP is greatly reduced. Figure 3 compares their numbers of database scans. The database record size used was 180, and the number of evolved marine parameters was 5, 9, 13, 17, 21 and 25. The 5 items are the four marine parameters in the CLSObj1 region and the ENSO index, the 9 items are the four marine parameters in the CLSObj1 and CLSObj2 regions and the ENSO index, the 13 items are the four marine parameters in the CLSObj1, CLSObj2 and CLSObj3 regions and the ENSO index, and so on. The minimum support threshold was set to 5.0%, 7.5%, 10.0% and 15.0%.
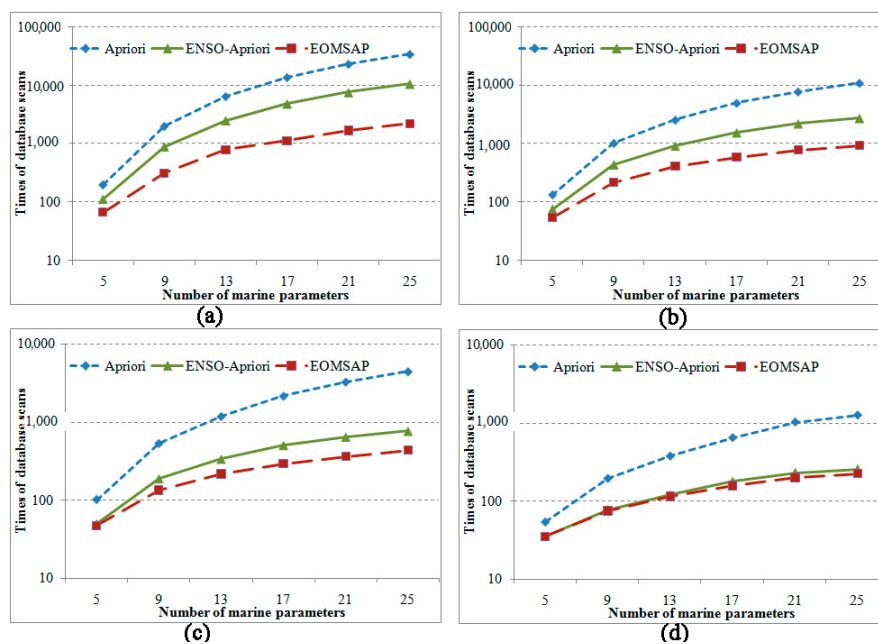
Figure 3. Comparisons of numbers of database scans. The support threshold in (**a–d**) is 5.0%, 7.5%, 10.0% and 15.0%, respectively.

With minimum support thresholds of 5.0% and 7.5%, the number of database scans for EOMSAP is much less than that for ENSO-Apriori, particularly with larger numbers of evolved items (Figure 3a,b), so the computational performance of EOMSAP is much better than that of ENSO-Apriori. When the support thresholds were set to 10.0% and 15.0%, the number of database scans of EOMSAP was less than that of ENSO-Apriori (Figure 3c,d). Actually, when the number of the evolved items is greater than 13, the computational performance of EOMSAP is little better than that of ENSO-Apriori, i.e., the EOMSAP's computation times are 1.20 s with 17 items, 1.65 s with 21 items and 2.15 s with 25

items, while the times for ENSO-Apriori are 1.29, 1.79 and 2.42 s, correspondingly. When the number of the evolved items is not greater than 13, EOMSAP and ENSO-Apriori have similar computational performance. With the exception of database scanning, there are two other issues responsible for the performance of EOMSAP. One is identifying El Niño and La Niña types when finding frequent 1-itemsets, and the other is calculating the conditional support against ENSO events for each candidate frequent itemset during the process of linking and pruning. Thus, when the support threshold increases, the EOMSAP's advantage in terms of the number of database scans will reduce and even disappear (Figure 3d).

3.2.3. Database Record Sizes

The database size represents the number of samples determining each occurrence of database scanning. In this case, 1, 10, 100, 1000 and 10,000 copies of 180 records (January 1998 to December 2012) with 25 items (six regions with four parameters and the ENSO index) from remote sensing products are produced, and the support was set to 7.5%. As with mining with duplications, we obtain similar results with 100,834 database scans by Apriori, 2707 scans by ENSO-Apriori and 913 scans by EOMSAP. Figure 4 shows the performance of EOMSAP, ENSO-Apriori and Apriori algorithms using different numbers of records.
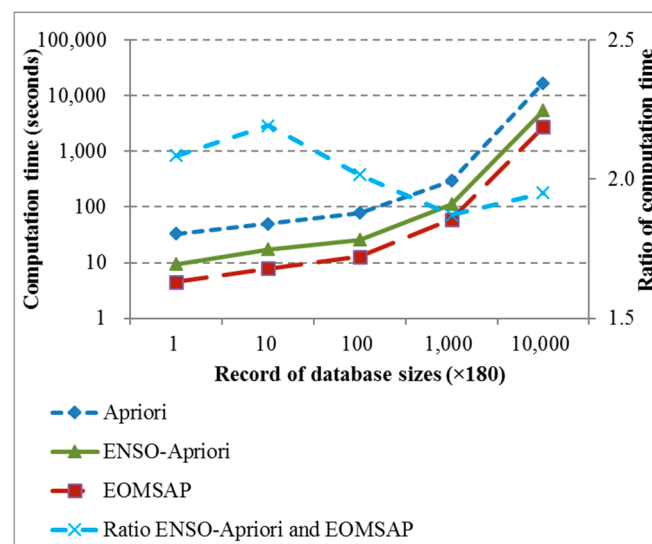

Figure 4. Performance analysis with different numbers of records.

Figure 4 shows that regardless of the number of database records, the computation efficiency of EOMSAP is much better than that of ENSO-Apriori and then Apriori, and that the computation time of EOMSAP, ENSO-Apriori and Apriori have similar increasing characteristics with an increase in the number of database records. For better analysis of their computational performance, the ratio of computation time between ENSO-Apriori and EOMSAP is calculated and shown on the right vertical axis. The computational performance of EOMSAP is always approximately two times better than that of ENSO-Apriori. That is, the number of database records has little effect on the computational performance of EOMSAP. However, it should be noted that with an increase in the number of database records from 1000 to 10,000 copies, its computation time increases exponentially. The reason may be due to the storage capacity of the storage units and the reading capacity of the cache memory CPU. By calculation, 1 to 100 copies of database records can be read by the CPU memory when scanning the database. However, 1000 and 10,000 copies are too large to be read by the CPU memory, and the records need to be read from a hard disk. The computation time of reading from a hard disk is exponentially larger than that from CPU memory.

3.3. Spatial Abnormal Association Patterns among Marine Environmental Parameters

In this case, monthly anomalies of marine parameters, i.e., SSTA, CHLA, SLAA and SSPA, within the Pacific Ocean, as shown in Figure 2, and the ENSO index were used. Marine spatial association patterns against ENSO were extracted by EOMSAP pixel by pixel with a support threshold of 5.0% and a confidence threshold of 75.0% and then mapped onto two-dimensional thematic maps. The selection of support and confidence thresholds is based on our many experiments and statistical analyses.

As El Niño and La Niña events have similar processes of mining marine spatial association patterns, this manuscript takes La Niña as an example to illustrate the feasibility of our proposed method. Figure 5a–c give examples of the abnormal variations of SSTA, SSPA and SLAA, respectively, against a La Niña event, and Figure 5d shows association patterns both in SSTA and SSPA against a La Niña event.
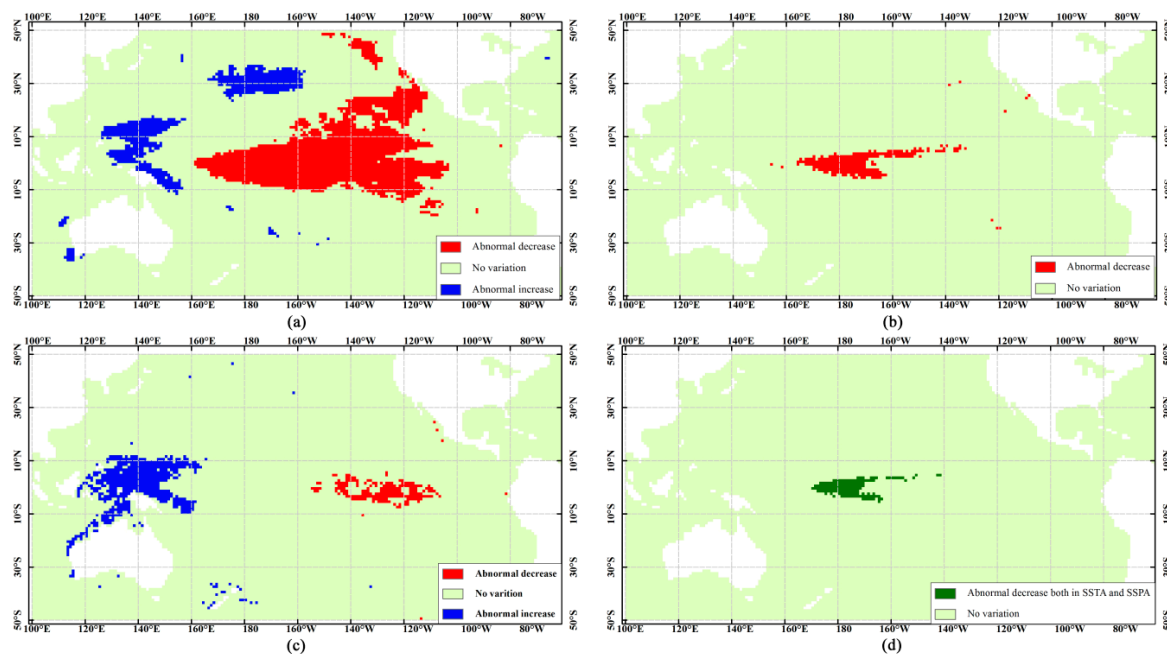

Figure 5. Association patterns against La Niña events. (**a**) SSTA abnormal variations; (**b**) SSPA abnormal variations; (**c**) SLAA abnormal variations; (**d**) Association patterns both in SSTA and SSPA. SSTA = monthly anomaly of SST; SSPA = monthly anomaly of sea surface precipitation; SLAA = monthly anomaly of sea level anomaly.

In Figure 5, the marine spatial association pattern in each grid pixel means that when a La Niña event occurs, the abnormal variation of the marine parameter will rise or drop abnormally, with a support not less than 5.0% and a confidence not less than 75.0%. That is, during the period from January 1998 to December 2012, the abnormal variation of marine parameters occurs not less than 9.0 times, and once a La Niña event occurs, the probability of a marine parameter changing abnormally is not less than 75.0%. In view of the number of analyzed geographical parameters, we not only obtain the spatial distribution of one marine parameter against La Niña events (Figure 5a–c), we also obtain the spatial association patterns among several parameters against La Niña events (Figure 5d). When La Niña events occur, the spatial variations of one marine parameter, e.g., SST, SSP or SLA, have been often documented in the literature [12,38,39]. Although previous studies have also analyzed marine environmental parameters alongside ENSO events using statistical analysis and empirical orthogonal decomposition with multiple remote sensing products [2,40–42], few studies have examined the associated relationships among several elements within a uniform framework [20]. That is, there have been few studies that discuss the spatial association patterns obtained such as in Figure 5d.

4. Discussion and Conclusions

In this manuscript, we proposed an original approach for exploring marine spatial association patterns against ENSO events with multiple long-term raster-formatted datasets. In this study, the process of quantifying abnormal variations and defining the ENSO used the mean-standard deviation of its time series, which is novel and different from the traditional static threshold. The results for January 1981 to December 2012 show that, except during a weak La Niña event from October 1995 to March 1996, the proposed method reached the same conclusions about ENSO events as previously reported [29,43]. In addition, the threshold to identify El Niño events is in the 29.49 percentile, and the threshold to identify La Niña events is in the 30.13 percentile. These almost agree with the 30.00 percentile [32].

The two datasets in this experiment came from real remote sensing products. One dataset covered six regions in the Pacific Ocean, chosen for testing the efficiency and feasibility of our proposed algorithm against the quantitative Apriori and ENSO-Apriori algorithms. As only the items related to ENSO are considered during the process of linking and pruning, the mined patterns and computational efficiency of EOMSAP and ENSO-Apriori are much better than that of Apriori. Comparisons of EOMSAP and ENSO-Apriori show that the greater the number of database scans, i.e., the more evolved the marine parameters, and the smaller the minimum support threshold, the more superior the performance of EOMSAP. With a decrease in the number of database scans, the superiority will decrease and even disappear. The reason is that, as well as the number of database scans, the identification and the calculation of conditional support during the process of linking and pruning also affects the performance of EOMSAP. In fact, the abnormal variation of a marine parameter can be considered as a low probability phenomenon, so a small minimum support threshold is more suitable for finding association patterns. In addition, in global climate changes, the evolved marine parameters are not only limited to the six regions, as shown in Figure 2; thus, the EOMSAP has great potential in real applications.

The other datasets considered the Pacific Ocean for exploring marine association patterns against ENSO events because it is sensitive to global climate change and regional sea–air interactions and is responsible for several marine abnormal variations. Compared with traditional spatiotemporal analysis, the information obtained from EOMSAP not only includes some that are well known to earth scientists, but also some that are new to earth scientists. For example, when La Niña events occur, the westward North Equatorial Current and South Equatorial Current and the eastward Equatorial Counter-Current result in the sea level anomaly increasing in the western Pacific Ocean and decreasing in the central Pacific Ocean, as shown in Figure 5c; thus, the SST decreases in the central and eastern Pacific Ocean and increases in the western Pacific Ocean (Figure 5a). Under the force of trade winds and the Walker circulation, the rainfall shifts westward, and SSPA in the middle of the tropical Pacific Ocean abnormally decreases [38] (Figure 5b). The more detailed and informative knowledge can help to improve our understanding of how and where the marine environmental parameters in different zones respond to ENSO events. However, further study is needed to determine the physical mechanisms behind the abnormal decrease in SSTA off the California coast, the abnormal increase in SSTA in the northern subtropical Pacific Ocean (Figure 5a), and the co-variations in the decrease of SSTA and SSPA (Figure 5d).

In summary, the main contributions of our algorithm and study are the following:

1. EOMSAP includes a process of quantification that ranks abnormal variations of marine parameters using long-term raster-formatted datasets and identification that defines ENSO events using the MEI. The quantification process has similar results with the prevalent algorithms.
2. EOMSAP reduces the number of database scans and improves the efficiency of finding frequent association patterns against ENSO by embedding the conditional support. The greater the number of evolved marine parameters considered, the greater the superiority of EOMSAP over

ENSO-Apriori and quantitative Apriori. Additionally, the lower the support threshold, the greater the superiority of EOMSAP over ENSO-Apriori and Apriori.

3. EOMSAP explores marine spatial association patterns within the Pacific Ocean against ENSO events using multiple long-term raster-formatted datasets. Among these spatial association patterns, some are well known to earth scientists, and some are new.

4. EOMSAP improves the abilities to address multiple remote sensing products and helps marine experts identify new phenomena or knowledge.

Although our proposed approach takes ENSO as a means to explore association patterns, marine parameters and ENSO events are equivalent during the mining process. That is, an ENSO event could be replaced by any marine parameter, and thus, the specified marine parameter-oriented spatiotemporal association pattern can be acquired. Such a mining model can improve our understanding of how and where the marine environmental parameters in different zones help to drive and respond to the variations of other parameters.

Author Contributions: Cunjin Xue led the research work, proposed the idea and structure of this manuscript, wrote the Introduction and Conclusion sections, and integrated the contributions from all authors as a whole. Xiaohan Liao proposed the idea of this manuscript and conceived and designed the experiments, and collaborated with Cunjin Xue in the writing of the Results and Discussion sections.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McPhaden, M.J.; Zebiak, S.E.; Glantz, M.H. ENSO as an integrating concept in earth science. *Science* **2006**, *314*, 1740–1745. [CrossRef] [PubMed]

2. Wang, C.; Fiedler, P.C. ENSO variability and the eastern tropical Pacific: A review. *Prog. Oceanogr.* **2006**, *69*, 239–266. [CrossRef]

3. Messié, M.; Chavez, F.P. Physical-biological synchrony in the global ocean associated with recent variability in the central and western equatorial Pacific. *J. Geophys. Res. Oceans* **2013**, *118*, 3782–3794. [CrossRef]

4. Korting, T.S.; Fonseca, L.M.G.; Camara, G. GeoDMA—Geographic Data Mining Analyst. *Comput. Geosci.* **2013**, *57*, 133–145. [CrossRef]

5. Yang, J.; Gong, P.; Fu, R.; Zhang, M.H.; Chen, J.M.; Liang, S.L.; Xu, B.; Shi, J.C.; Dickinson, R. The role of satellite remote sensing in climate change studies. *Nat. Clim. Chang.* **2013**, *3*, 875–883. [CrossRef]

6. Hannachi, A.; Jolliffe, I.T.; Stephenson, D.B. Empirical orthogonal functions and related techniques in atmospheric science: A review. *Int. J. Climatol.* **2007**, *27*, 1119–1152. [CrossRef]

7. Smith, T.M.; Arkin, P.A.; Sapiano, M.R.P. Reconstruction of Near-global Annual Precipitation using Correlations with Sea Surface Temperature and Sea Level Pressure. *J. Geophy. Res.* **2009**, *114*, D12107. [CrossRef]

8. Cherry, S. Some Comments on Singular Value Decomposition Analysis. *J. Clim.* **1997**, *10*, 1759–1761. [CrossRef]

9. Liao, S.H.; Chu, P.H.; Hsiao, P.Y. Data mining techniques and applications—A decade review from 2000 to 2011. *Expert Syst. Appl.* **2012**, *39*, 11303–11311. [CrossRef]

10. Hoffman, F.M.; Larson, J.W.; Mills, R.T.; Brooks, B.G.J.; Ganguly, A.R.; Hargrove, W.W.; Huang, J.; Kumar, J.; Vatsavai, R.R. Data Mining in Earth System Science (DMESS 2011). *Procedia Comput. Sci.* **2011**, *4*, 1450–1455. [CrossRef]

11. Su, F.Z.; Zhou, C.H.; Lyne, V.; Du, Y.Y.; Shi, W.Z. A data mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. *Ecol. Model.* **2004**, *174*, 421–431. [CrossRef]

12. Saulquin, B.; Fablet, R.; Mercier, G.; Demarcq, H.; Mangin, A.; Fantond'Andon, O.H. Multiscale Event-Based Mining in Geophysical Time Series: Characterization and Distribution of Significant Time-Scales in the Sea Surface Temperature Anomalies Relatively to ENSO Periods from 1985 to 2009. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 3543–3552. [CrossRef]

13. Ke, Y.P.; Cheng, J.; Ng, W. An information-theoretic approach to quantitative association rule mining. *Knowl. Inf. Syst.* **2008**, *16*, 213–244. [CrossRef]

14. Ke, Y.P.; Cheng, J.; Ng, W. Correlated pattern mining in quantitative databases. *ACM Trans. Database Syst.* **2008**, *33*, 1–44. [CrossRef]

15. Xue, C.J.; Song, W.J.; Qin, L.J.; Dong, Q.; Wen, X. A mutual-information-based mining method for marine abnormal association rules. *Comput. Geosci.* **2015**, *76*, 121–129.

16. Tsay, Y.J.; Chiang, J.Y. CBAR: An efficient method for mining association rules. *Knowl. Based Syst.* **2005**, *18*, 99–105. [CrossRef]

17. Wu, C.M.; Huang, Y.F. Generalized association rule mining using an efficient data structure. *Expert Syst. Appl.* **2011**, *38*, 7277–7290. [CrossRef]

18. Liu, X.B.; Zhai, K.; Pedrycz, W. An improved association rules mining method. *Expert Syst. Appl.* **2012**, *39*, 1362–1374. [CrossRef]

19. Wu, T.S.; Song, G.J.; Ma, X.J.; Xie, K.Q.; Gao, X.P.; Jin, X.X. Mining geographic episode association patterns of abnormal events in global earth science data. *Sci. China Ser. E Technol. Sci.* **2008**, *51*, 155–164. [CrossRef]

20. Xue, C.J.; Dong, Q.; Fan, X. Spatiotemporal association patterns of multiple parameters in the northwes tern Pacific Ocean and their relationships with ENSO. *Int. J. Remote Sens.* **2014**, *35*, 467–4483. [CrossRef]

21. Huang, P.Y.; Kao, L.J.; Sandnes, F.E. Efficient mining of salinity and temperature association rules from ARGO data. *Expert Syst. Appl.* **2008**, *35*, 59–68. [CrossRef]

22. Satheesh, A.; Patel, R. Use of object-oriented concepts in databases for effective mining. *Int. J. Comput. Sci. Eng.* **2009**, *1*, 206–216.

23. Rao, K.V.; Govardhan, A.; Rao, K.V.C. An object-oriented modeling and implementation of spatio-temporal knowledge discovery system. *Int. J. Comput. Sci. Inf. Technol.* **2011**, *3*, 61–76.

24. Li, G.Q.; Deng, M.; Zhang, W.L.; Chen, Y. Events-coverage based spatio-temporal association rules mining method. *J. Remote Sens.* **2010**, *14*, 468–481.

25. Julea, A.; Meger, N.; Bolon, P.; Rigotti, C.; Doin, M.P.; Lasserre, C.; Trouve, E.; Lazarescu, V.N. Unsupervised spatiotemporal mining of satellite image time series using grouped frequent sequential patterns. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 1417–1430. [CrossRef]

26. Romani, L.A.S.; de Avila, A.M.H.; Chino, D.Y.T.; Zullo, J.; Chbeir, R.; Traina, C.; Traina, A.J.M. A New Time Series Mining Approach Applied to Multitemporal Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 140–150. [CrossRef]

27. Blanchard, J.; Pinaud, B.; Kuntz, P.; Guillet, F. A 2D-3D visualization support for human-centered rule mining. *Comput. Graph.* **2007**, *31*, 350–360. [CrossRef]

28. Kumar, V. Discovery of Patterns in Global Earth Science Data Using Data Mining. *Lecture Notes Comput. Sci.* **2010**, *6118*, 2. [CrossRef]

29. Trenberth, K.E. The Definition of El Niño. *Bull. Am. Met. Soc.* **1997**, *78*, 2771–2777. [CrossRef]

30. Smith, T.M.; Reynolds, R.W. Improved extended reconstruction of SST (1854–1997). *J. Clim.* **2004**, *17*, 2466–2477. [CrossRef]

31. Curtis, S.; Adler, R. ENSO Indices Based on Patterns of Satellite-Derived Precipitation. *J. Clim.* **2000**, *13*, 786–2793. [CrossRef]

32. Wolter, K.; Timlin, M.S. El Nino/Southern Oscillation behavior since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext). *Int. J. Climatol.* **2011**, *31*, 1074–1087. [CrossRef]

33. Agrawal, R.; Srikant, R. Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile Bocca, 12–15 September 1994; Bocca, J.B., Jarke, M., Zaniolo, C., Eds.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA; pp. 407–419.

34. Srikant, R.; Agrawal, R. Mining quantitative association rules in large relational tables. In Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal, QC, Canada, 4–6 June 1996.

35. Han, J.W.; Pei, J. Mining Frequent Patterns by Pattern-Growth: Methodology and Implications. *SIGKDD Explor.* **2000**, *2*, 14–20. [CrossRef]

36. Reynolds, R.W.; Rayner, N.A.; Smith, T.M.; Stokes, D.C.; Wang, W. An improved in situ and satellite SST analysis for climate. *J. Clim.* **2002**, *15*, 1609–1625. [CrossRef]

37. Hooker, S.B.; McClain, C.R. The Calibration and Validation of SeaWiFS Data. *Prog. Oceanogr.* **2000**, *45*, 427–465. [CrossRef]

38. Curtis, S.; Salahuddin, A.; Adler, R.F.; Huffman, G.J.; Gu, G.; Hong, Y. Precipitation Extremes Estimated by GPCP and TRMM: ENSO Relationships. *J. Hydrometeorol.* **2007**, *8*, 678–689. [CrossRef]

39. Chen, G.; Wang, Z.; Qian, C.C.; Lv, C.; Han, Y. Seasonal-to-decadal modes of global sea level variability derived from merged altimeter data. *Remote Sens. Environ.* **2010**, *114*, 2524–2535. [CrossRef]

40. Wu, B.; Zhou, T.J.; Li, T. Contrast of Rainfall-SST Relationships in the Western North Pacific between the ENSO-Developing and ENSO-Decaying Summers. *J. Clim.* **2009**, *22*, 4398–4405. [CrossRef]

41. Murtugudde, R.; Wang, L.P.; Hackert, E.; Beauchamp, J.; Christian, J.; Busalacchi, A.J. Remote sensing of the Indo-Pacific region: Ocean colour, sea level, winds and sea surface temperatures. *Int. J. Remote Sens.* **2004**, *25*, 1423–1435. [CrossRef]

42. Casey, K.S.; Adamec, D. Sea surface temperature and sea surface height variability in the North Pacific Ocean from 1993 to 1999. *J. Geophys. Res. Oceans* **2002**, *107*, 3099. [CrossRef]

43. Li, X.Y.; Zhai, P.M. On indices and indictors of ENSO episodes. *Acta Meteorol. Sin.* **2000**, *58*, 102–119.