Canonical correlation and visual analytics for water resources analysis

Bybordi Arezoo[1], Thampan Terri[1], Linhares Claudio D. G.[2], Ponciano Jean R.[3], Travençolo Bruno A. N.[3], Paiva Jose Gustavo S.[3], and Etemadpour Ronak[4]

[1]The Graduate Center, CUNY
[2]University of São Paulo
[3]Federal University of Uberlândia
[4]The City College of New York, City University of New York

November 16, 2022

Abstract

In the last decades, the urbanization process and population growth resulted in a substantial increase of water consumption for agricultural, industrial, and residential purposes. The characterization of the interplay between environmental variables and water resources plays a critical role for establishing effective water management policies. In this paper, we apply the Canonical Correlation Analysis (CCA) in a set of climate and hydrological indicators to investigate the behavior of these environmental variables over time in different geographical regions of California, as well as the relationship among these regions. CCA served as base to establish a temporal graph that models the relation between the stations over time, and advanced graph visualization techniques are used to produce patterns that aids in the comprehension of the underlying phenomena. Our results identified important temporal patterns, such as heterogeneous behavior in the dry season and lower correlation between the stations in La Niña years. We show that the combination of CCA and visual analytics can assist water experts in the identification of important climate and hydrological events in different scenarios.

Canonical correlation and visual analytics for water resources analysis

Arezoo Bybordi[1], Terri Thampan[1], Claudio D. G. Linhares[3], Jean R. Ponciano[2], Bruno A. N. Travençolo[2], Jose Gustavo S. Paiva[2], Ronak Etemadpour[1]

[1]City University of New York, The Graduate Center, New York, United States
[2]Federal University of Uberlândia, Uberlândia, Brazil
[3]Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, Brazil

Key Points:

- Canonical correlation analysis of climate and hydrological indicators are used to find relationships between gauging stations.
- Temporal graph modeling and Visual Analytics are employed as a tool for temporal pattern identification between the stations.
- A case study shows the use of the proposed methodology to analyze stations located in different geographical regions of California.

Corresponding author: Ronak Etemadpour, `retemadpour@ccny.cuny.edu`

Abstract

In the last decades, the urbanization process and population growth resulted in a substantial increase of water consumption for agricultural, industrial, and residential purposes. The characterization of the interplay between environmental variables and water resources plays a critical role for establishing effective water management policies. In this paper, we apply the Canonical Correlation Analysis (CCA) in a set of climate and hydrological indicators to investigate the behavior of these environmental variables over time in different geographical regions of California, as well as the relationship among these regions. CCA served as base to establish a temporal graph that models the relation between the stations over time, and advanced graph visualization techniques are used to produce patterns that aids in the comprehension of the underlying phenomena. Our results identified important temporal patterns, such as heterogeneous behavior in the dry season and lower correlation between the stations in *La Niña* years. We show that the combination of CCA and visual analytics can assist water experts in the identification of important climate and hydrological events in different scenarios.

1 Introduction

Water is the one of the most important natural resources, with a significant irregular availability distribution around the planet. Several regions experienced in the last decades a urbanization process and population growth which resulted in an substantial increase of water consumption for agricultural, industrial and residential purposes (Bekchanov et al., 2017). The consequent social and economic impacts brought by this consumption increasing lead to the urgency in developing effective and sustainable water management policies and strategies. That is the case of the California state, located in a mountain region in arid and semi-arid climates, in which drought periods frequently occurs. Since 2012, and specially in 2014-2016, the state witnessed the worst drought in more than a century, with low precipitation and extreme high temperatures values (AghaKouchak et al., 2015, 2014). The consumed water in this state is mostly from surface sources, and comes from Colorado River and lake Oroville via a system of rivers and aqueducts. As the largest agricultural producer in the United States by value (Qin & Horvath, 2020), and at the same time having the country larger population area (Bureau, 2021), it is crucial to maintain an effective water management planning in California, as well as to comprehend one of the most extensive and complex water infrastructure system in the world (Stewart et al., 2020), in order to grant water distribution in all economic sectors.

A variety of water related data is collected by several institutions and governmental agencies in California, and publicly available in portals such as California Open Data Portal[1], California Water Data Consortium[2], California Water Boards[3], among others. These data provide information about several climatic and hydrological indicators, such as streamflow, precipitation, runoff, mountain snowpack, evapotranspiration and soil moisture (Chang et al., 2015; Kasiviswanathan & Sudheer, 2016), and are applied in scenarios such as climate changes analysis (Oo et al., 2020), streamflow (Alipour & Kibler, 2019), flood and drought prediction (Ouarda et al., 2001; Forootan et al., 2019), among other tasks.

Canonical correlation analysis (CCA) is a statistical tool for multivariate data analysis which investigates relationships among multiple sets of variables. The multivariate distributions and analyses of sets of hydrological random variables represent the best approach in deriving hydrological relationships of a probabilistic type (RICE, 1972). Sev-

[1] https://data.ca.gov/
[2] https://cawaterdata.org/
[3] https://www.waterboards.ca.gov/

eral works can be found applying Canonical correlation for the analysis of climate change patterns (Zhang et al., 2020), droughts evolution (Forootan et al., 2019), as well as for droughts/flood frequency analysis and estimation (Ouarda et al., 2001), specially in regional frequency analysis (RFA), to delineate hydrological neighborhoods at ungauged sites (Ouali et al., 2016; Desai & Ouarda, 2021). However, most of the works apply CCA in a single water related indicator, using or climate or hydrological ones, thus capturing only one perspective of the scenario. These works also do not consider the temporal aspect associated with these indicators, in the sense that they often employ summarization strategies that may hide important evolution patterns associated with these measures.

In this sense, this paper proposes to apply the Canonical Correlation Analysis (CCA) in a set of climate and hydrological indicators to investigate the relationship among these indicators and among different geographical regions of California, as well as how these relationships behave over time. Our study employs 17 hydrological and climate variables, daily collected over 40 years from over 130 stream gauging stations in California, measuring CCA in each day to verify similar behavior among them. We believe that employing a set of distinct indicators combines information from complementary aspects related to water availability and distribution, which may capture particular aspects from different geographic regions and thus improve the scenario comprehension by water management experts when making decisions.

However, the amount of data generated from daily measurements over 40 years is huge. So, in this paper we propose to model the relationship between gauge stations using temporal graphs. Each station is represented by a graph node, and the graph edges connecting two stations represent statistical significant CCA values regarding their behavior. Then, we employ a set of temporal graph visualization strategies proposed in (Linhares et al., 2017) to support an interactive visual analysis. By exploring the number and distribution of edges in each timestamp, this visualization strategy is able to reveal strategic temporal patterns which may represent seasonal and abnormal events, as well as structural patterns associated with geographical locations, such as similar behavior associated with geographical distances. We believe that CCA coupled with a temporal graph visual analytics strategy is a potential tool for providing a simple yet effective analysis of geographical locations behavior regarding water related indicators. This analysis can provide a better comprehension of the water scenario in California, and foster the creation and application of private and/or government policies which grant an efficient water management and help to better forecast extreme periods, mitigating their social and economic effects.

Our contributions are listed as follows:

- Application of Canonical Correlation Analysis in climate and hydrological indicators to identify similarities among geographic locations in California;
- A visual analysis strategy to: (i) Support and enhance the CCA analysis over different geographic California regions; (ii) Explore the evolution of CCA analysis results over time;
- A detailed discussion of the results obtained with the proposed approach, focusing on guiding water management experts in their decision making process.

The following sections describe related work, our approach to calculate the CCA and visually explore the results, as well as the discussion of the results of applying the proposed strategy to California scenario.

2 Related Work

In this section we discuss the related works that uses CCA or graph modeling in the study of water resources. We also present a brief description of visual analysis of temporal graphs.

2.1 Canonical correlation analysis

The analysis and comprehension of how water related indicators behave in different scenarios is important to guide the creation of policies aiming to grant water availability for citizens, as well as in predicting the occurrence of natural disasters, mitigating its negative effects, among other tasks. Several water related analysis strategies can be found in the literature, focused in a variety of research topics which include climate changes (AghaKouchak et al., 2014; Stewart et al., 2020), droughts/flooding analysis (AghaKouchak et al., 2015; Papaioannou et al., 2015; Vicente-Serrano et al., 2018), water management (Lund et al., 2018; Kamienski et al., 2019) or streamflow prediction (Li et al., 2018; Alipour & Kibler, 2019; Meng et al., 2019). These works use different water indicators as input for different computational strategies, in order to reveal interesting patterns to water experts and assist their decision making. In this sense, Canonical correlation analysis (CCA) in multivariate statistics can be useful to highlight the interrelations that may exist between two groups of variables by providing the general theoretical framework for the techniques of factorial discriminant analysis, multivariate regression and correspondence analysis (Ouarda et al., 2001).

A variety of works employ CCA in water related research, such as the analysis of multi predictor-rainfall relationships (Tukimat et al., 2019). They showed that the CCA is sufficient to show the predictors' capability and reliability based on the percentages of variance. In (Zhang et al., 2020), CCA is applied to link all the hydrological variables to El Niño Southern Oscillation (ENSO) Index through SST to identify the implicit relationship between the hydrological cycle on land and ENSO, based on the fact that precipitation' changes on land and ocean is related to Sea Surface Temperature (SST).

To study the correlation structure between two sets of variables represented by watershed characteristics and flood peak in a regional flood frequency analysis, CCA is used to determine the homogeneous hydrologic neighborhoods (Ouarda et al., 2001). Also for flood-risk management, Aguilar et al. (Schanze, 2006) used CCA to find the correlation between the prioritized variables that have an important role in dams operation with the precipitation intensities and flow rates during hurricanes in the Mexican coast. Desai et al. (Desai & Ouarda, 2021), used CCA to select the optimal hydrological neighborhoods for each hydrometric site located in the southern part of the province of Quebec, Canada. After selecting the optimal sites, multiple regression models, including non-linear/non-parametric methods and Artificial Neural Network (ANN) were used for regional flood estimation. Finally, a comparison between different methods for regional flood estimation, based on multiple regression models on data from the Balsas, Lerma and The Pánuco River Basins located in Mexico, showed that CCA-based estimations outperformed other techniques in identification of the exploratory tropical climate variables (Ouarda et al., 2008).

We use CCA to delineate homogeneous California watershed stations taking into account different sets of hydrological and climatic variables that are explained in Section 3.1.

2.2 Graph modeling

In recent years, several studies concentrated on the use of graphs modeling for water related research, specially for analysis of precipitation and streamflow dynamics. In this section, we briefly review some of these proposals, focusing on how the graphs are built and how the data were analyzed. Some of the works described in this section use the terminology from the complex network theory in their text. However, here we will employ the term graphs instead of networks for all papers.

(Sivakumar & Woldemeskel, 2014) used graph modeling to examine the connections in streamflow dynamics. Monthly streamflow data were collected over a period of 52 years from a large network of 639 monitoring stations in the United States (US). Each station is a node in the graph and the connections between the nodes are defined using a threshold on the linear cross-correlation streamflow values between stations.

In (Sivakumar & Woldemeskel, 2015), the authors analyzed monthly rainfall data recorded over a period of 68 years (1940-2007) at 230 rain gauge stations across Australia. Each station was represented by a node in the graph and the connections were established by analyzing the correlation between the nodes based on rainfall data. A correlation threshold is considered to identify the neighbors and clustering coefficient and degree distribution are used to analyze the network.

In the work by (Halverson & Fleming, 2015), a total of 127 hydrometric stations on the Canadian west coast from 2000–2009 was selected. Each station is considered a node and the links are generated based on a threshold on the correlation coefficient computed from the stations' streamflow data. They analyzed the community structure and betweenness of the graph.

Graph modeling was used by (Xu et al., 2020) to investigate the spatial connections and architecture of precipitation networks in the Yellow River Basin in China. The graph is built considering 379 stations as nodes and links are defined by correlation coefficients computed from rainfall data during a period of 56 years (1956-2012).

(Braga et al., 2016) investigate the dynamics of river flows by mapping daily time series from 141 different measuring stations of 53 Brazilian rivers from the period of 1931–2012. For each year, a graph is constructed using horizontal visibility graph approach. They analyzed the degree distribution and clustering coefficient of the 81 networks to study the evolution of flow fluctuation. Horizontal visibility graphs were also employed by (Serinaldi & Kilsby, 2016) in the analysis of the dynamics of streamflow fluctuations in the continental US. They used a data set consisting of 699 daily time series from 743 gage stations spanning up to 114 years.

In (Fang et al., 2017), the authors use the concept of community structure in graphs to classify catchments of the Mississippi river basin in the US. A community in this context is a group of individuals that connect more among themselves than to other groups. To construct the network, they use daily streamflow data from a network of 1663 gauging stations from 2008 to 2013. Six community structure methods were evaluated and have shown a high degree of consistency between them. They have shown that the correlation threshold influences the size and number of communities found. A similar approach using community structure was used in the analysis of US and Australian basins in the work by (Tumiran & Sivakumar, 2021).

In (Han et al., 2018), the temporal dynamics of streamflow are analyzed using graphs measures as degree centrality, clustering coefficient, and degree distribution. Each node here represents an year, and consists of a time series of (365 daily) streamflow values over a period of 151 years (1862-2013) from the Mississippi River basin.

(Yasmin & Sivakumar, 2018) proposed a different approach to construct the network that employs coupled phase space reconstruction for examining the temporal connections in streamflow data from each of 639 stations across the contiguous US. In another study, (Yasmin & Sivakumar, 2020) also examined clustering properties of the temporal dynamics of streamflow using the same coupled phase space reconstruction–network.

In (Agarwal et al., 2020), rainfall event series in 1229 stations across Germany are compared with each other using event synchronization. The employed data covers 110 years at a daily resolution from the period of 1901-2010. If two stations are significantly synchronized, a link between them is established. The authors proceed the analysis using several network measurements.

In summary, the proposals discussed in this section use different approaches to build the graphs and analyze the water related phenomena. Although temporal features were considered in the definitions of the graphs (especially to decide the links between the nodes), none of these works has modeled the underlying phenomena as a temporal graph. In addition, only few hydrological and climate variables are considered in each analysis, and most of the works concentrate in the analysis of only one feature (e.g. streamflow data, rainfall events). Finally, although several figures and maps are shown in these papers, most of these graphical elements are used only to map the features to the geographical location of the events. There is no deep visual analysis of the derived graphs, restricting the analysis to traditional graph measurements. The area of Information Visualization provides several methods and tools for graph visualization and analysis, and it can provide, by means of visual analysis, new insights for the phenomenon under study. In the next section we discuss about some of these methods.

2.3 Visual analysis of temporal graphs

Visualization techniques allow users to gain insights, generate knowledge, find patterns, trends, and anomalies in the data that were usually not expected. Moreover, graph visualization allows finding different structural, topological and temporal behaviors in the data, such as group formation and graph evolution, while preserving the user's mental map. We follow the visualization mantra of *"Overview first, zoom, details on demand"*, which lead us to discoveries in global and local perspectives (Shneiderman, 1996). To gain insights and ideas that lead to global discoveries, we analyze the data in overview representations. With the use of interactive tools, we can identify more local patterns and investigate in external sources how to interpret them, speeding up the analysis process and generating more reliability in the results.

There are several visualization strategies to visualize temporal graphs. We use the Dynamic Network Visualization (DyNetVis) software in our analysis, which is freely available interactive software that contains several state-of-the-art techniques for temporal visualization (Linhares et al., 2017, 2020). DyNetVis provides four types of visualization techniques: structural (node-link diagram), temporal (TAM – Temporal Activity Map), matrix, and community layouts. It offers several state-of-the-art methods to interact with each of these layouts.

3 Methodology

In this paper, we propose a novel methodology for the analysis of water resources based on CCA and visual analytics. The methodology consists of four major steps, illustrated in Figure 1: (i) collect a set of climate and hydrological indicators over time from stations located in different geographical regions of California; (ii) compute the Canonical Correlation Analysis (CCA) between the collected indicators; (iii) based on CCA results, establish a temporal graph to model the relation between the stations; and (iv) perform a visual analysis using graph visualization strategies to identify temporal patterns between the stations. The following sections describe the details of each step.

3.1 Data Collection

For the analysis, we employ 17 hydrologic and climatic features; streamflow, which is one of the essential hydrological features, a group of primary climate variables: max-

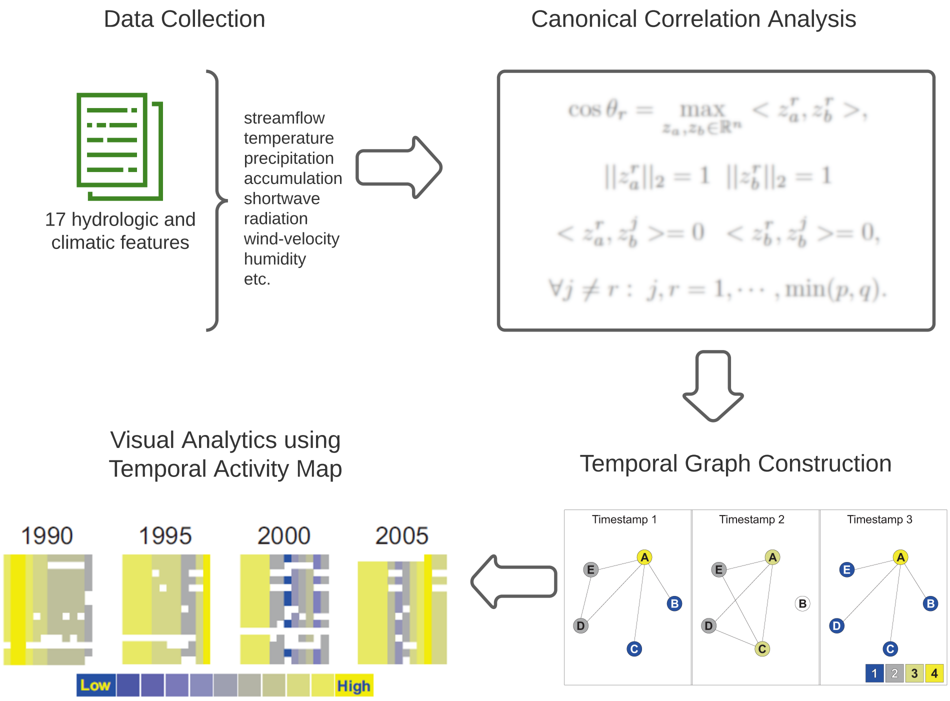

Figure 1. Major steps of the methodology used to analyse the relationship between water resource stations.

imum temperature, minimum temperature, precipitation accumulation, downward surface shortwave radiation, wind-velocity, humidity (maximum and minimum relative humidity and specific humidity; a group of derived variables: Reference evapotranspiration (ASCE Penman-Montieth), Energy Release Component, Burning Index, 100-hour and 1000-hour dead fuel moisture and mean vapor pressure deficit. Daily streamflow data is extracted from USGS using the dataRetrieval R package (DeCicco & Hirsch, 2016). The daily data is the average of streamflow values for the same day that has been calculated by USGS (California Water Data Maintainer, 2021-05-13). Other climatic features have been derived from The climatology lab from the University of Idaho (Abatzoglou J. et al., 2017).

First, the daily streamflow dataset for all the stations was retrieved. The ones with at least 80% daily streamflow data available from Oct 1979 to Sept 2019 were chosen to be considered for further analysis. At this step, we had over 300 streamflow stations, divided in **reference stations**, which are non-human-impacted stations, and **non-reference stations**, which are the ones that have been human-impacted, either by building a dam, agriculture field, or urban area. The locations information, such as coordinates and whether they are reference or non-reference, were extracted from the GAGES-II dataset given in USGS (Falcone, 2011).

The other variables extracted from the climatology lab were chosen from a gridMET dataset with average daily values for 4km square areas. For our analysis, we divided the experiments in two parts. First, we picked a sample of 15 stations containing both reference and non-reference stations from different California hidrological regions. For our second analysis, we picked all the 131 reference stations.

For each of the hydrologic and climatic features, we have standardized the daily values throughout the 40 years as it is the form required for CCA. As explained in (Uurtio et al., 2017), the variables, which in this case is a vector of length 17 for each station, are assumed to be jointly sampled from a multivariate normal distribution (each station has its own multivariate normal distribution) – we used z-score normalization for this task.

3.2 Canonical Correlation Analysis

Calculation details: use of time series comprising all years, sampling strategy, grouping strategy, etc.

We follow the notation of (Uurtio et al., 2017) to describe this two-view multivariate statistical method, and divide the features of the variable into two sections which we call **views**. We denote two views a and b with two matrices X_a and X_b having dimensions $n \times p$ and $n \times q$ respectively. The rows of those matrices show measurements for multivariate observations, which are assumed to be jointly sampled from a multivariate normal distribution. In this method, using some linear transformations, we try to find linear relations between variables of X_a and X_b. We use $X_a \in \mathbb{R}^{n\times p}$ and $X_b \in \mathbb{R}^{n\times q}$ as the linear transformations of positions $w_a \in \mathbb{R}^p$ and $w_b \in \mathbb{R}^q$. As a result, we get their images z_a and z_b in $\mathbb{R}^n$. In order to compute CCA, the angle $\theta \in [0, \frac{\pi}{2}]$ between z_a and z_b must be minimized. This is equivalent to maximizing the cosine of the angle. In this case, the cosine of θ is equal to dot product of z_a and z_b, since they are already mapped into an n-dimensional unit ball and $||z_a|| = ||z_b|| = 1$. The cosine of the angle is called the first canonical correlation, which is the only canonical correlation we consider in our study.

Let $r = 1, \cdots, q$ where $p > q$. The first r canonical correlation values can be found by recursively finding the next minimum enclosing angle. The general formula for canonical correlation values is

$$\cos\theta_r = \max_{z_a, z_b \in \mathbb{R}^n} < z_a^r, z_b^r >,$$

$$||z_a^r||_2 = 1 \quad ||z_b^r||_2 = 1$$
$$< z_a^r, z_b^j >= 0 \quad < z_b^r, z_b^j >= 0,$$
$$\forall j \neq r : \ j, r = 1, \cdots, \min(p, q).$$

For solving CCA in $X_a w_a = z_a$ and $X_b w_b = z_b$, the standard method of Singular Value Decomposition (SVD) is used. CCA can also be considered as a dimensionality reduction technique. One of the ways to determine the relevant number of positions or canonical correlation values is by applying statistical significance tests. In our case, we used the Bartlett-Lawley test to determine if the first canonical correlation is relevant or not. The statistical significance conveys the importance of the detected pattern. In fact, the statistical significance tests of canonical correlation values evaluate whether the obtained pattern can be considered to occur non-randomly (Uurtio et al., 2017). The sequential test procedure of Bartlett (Bartlett, 1938) determines the number of statistically significant canonical correlation values in the data. In fact, the hypothesis tested is

$$H_0 : \min(p, q) = k \text{ against } H_1 : \min(p, q) > k,$$

in which $k = 0, 1, \cdots, p$ and $p < q$. The number of statistically significant canonical correlation values will be considered k if the null hypothesis $H_0 : \min(p, q) = j$ is rejected for $j = 0, 1, \cdots, k-1$ but accepted for $H_1 : \min(p, q) > k - 1$. Bartlett-Lawley statistic is used for the test

$$L_k = -(n - k - \frac{1}{2}(p + q + 1) + \sum_{j=1}^{k} r_j^{-2}) \ln(\prod_{j=k+1}^{\min(p,q)} (1 - r_j^2)).$$

Here r_j is the j'th canonical correlation. The asymptotic null distribution of L_k is the χ^2 with $(p - k)(q - k)$ degrees of freedom. First, we check to see if no canonical relations exist between two views and then continue with more. In our case, we only checked the first step of the sequential test to see if the pattern detected by CCA was statistically significant or not, and we did not continue to check the rest of the sequential process. We follow this procedure because we only want to see if the pattern detected from CCA is non-random to consider those stations connected in the corresponding graph. Thus, if there is at least one statistically significant canonical correlation, we would accept that the pattern detected is not random, and consider those two stations to have similar behavior.

3.3 Temporal Graph Construction

We have made two groups of graph experiments. First, a small experiment using a 15-node graph corresponding to 15 reference stations (see Fig. 3), and a second experiment with 131 node graph corresponding to all the reference stations (131 reference stations in total). Here are how the nodes are connected. We denote the two stations (nodes) with X and Y in this way

$$X = \& \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_{18} \end{bmatrix}, Y = \& \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_{18} \end{bmatrix}$$

where x_i shows the i'th hydrological/climatic feature for station X and y_i shows the same for station Y. Canonical correlation values using all the values for each month of each year has been calculated. Two nodes are connected if they pass the hypothesis test of Bartlett-Lawley, as explained in Section 3.2. In this case, if the p-value for the χ^2 distribution is less than 0.05, we consider that it has passed the hypothesis test. If they pass the hypothesis test, they have a statistically significant canonical correlation and, therefore, have similar hydrological and climatic patterns. So we connect those two stations (nodes) by an edge. For each month of each year (479 months in total), canonical correlation values between each pair of 15 stations have been calculated. The same method is used for a larger graph, including all 131 reference stations.

3.4 Visual Analytics

Using Canonical Correlation Analysis, we produce a graphical structure with stations as nodes and edges showing similarities between hydrological/climatic features of the stations. In order to better comprehend, analyze and find the underlying patterns in the graphical structure, we employ recent visualization techniques available in DyNetVis (Linhares et al., 2017).

Among the available techniques, we used the node-link diagram (also called structural layout) to represent the structure and relation of the canonical network (Fig. 2(a)). This representation maps the nodes (stations) as circular shapes and edges (related stations) as straight lines. To represent the temporal evolution and the node correlation (activity), we use the Temporal Activity Map (TAM), which is a matrix-based layout that represents the nodes in the rows and timestamps in the columns (Linhares et al., 2017). As illustrated in Fig. 2(b), each square of the TAM layout is mapped using a different color, that represents the low (blue) or high (yellow) correlation between stations of the watersheds. White squares represents no activity during that time period, i.e. no significant correlation between the respective stations. To organize the nodes in TAM, we use the Community-based Node Ordering (CNO), wherein the nodes are approximated and grouped according to the community relationships among the elements (Linhares et al., 2019). This layout is used to identify the stations state changes over time (represented by a color scale), and also group then into communities with similar correlation pattern.

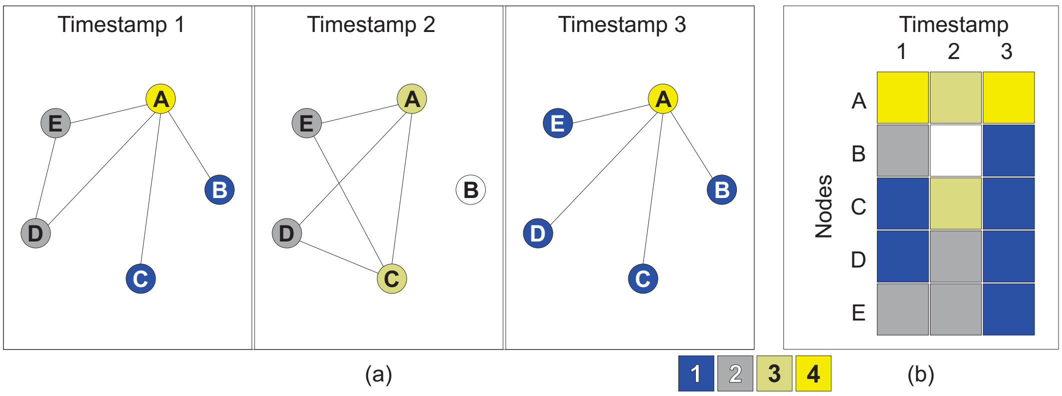

Figure 2. Visual representations used to represent the canonical correlation of the watershed stations: (a) node-link diagram and (b) Temporal Activity Map (TAM). The node-link diagram was divided into three timestamps and the respective timestamps are also shown in the TAM layout. Each node represent a station and the colors represent the correlation level of the station, which varies from blue (low) to yellow (high). In this example we can verify that station A has the highest correlation with other stations over time. Also, in *Timestamp 3*, nodes B–E show low correlation.

4 Results and Discussion

We applied the proposed methodology for two groups: (i) 15-node graph corresponding to 15 reference stations, and (ii) 131-node graph, considering all stations.

4.1 15-stations graph

Station ID	Mapped ID
10257600	1
10258000	2
10258500	3
10259000	4
10259200	5
10336645	6
10336660	7
10336676	8
10343500	9
11015000	10
11098000	11
11111500	12
11120500	13
11124500	14
11141280	15

Figure 3. Reference stations that compose our 15-stations graph. To improve the readability of some figures in this section, we will refer to these stations using a small incremental ID instead of the original station ID.

Fig. 3 shows the reference stations considered in our graph. To improve the readability of some figures in this section, we refer to these stations using a small incremental ID instead of the original station ID, as depicted in the figure. Different patterns regarding the correlation between stations may be observed when analyzing its evolution throughout the almost 40 years of collected data (1980 to 2018). We can see, for instance, a great discrepancy when analyzing different months of the year. There is a high correlation among almost all stations in the first four or five months in most of the years, as shown in Fig. 4, in a period that is often associated with high precipitation and streamflow values. On the other hand, there is low or no correlation in particular months, especially in the dry season, e.g. middle months of 1985 and 2010 in Fig. 4. Although low or no correlation in middle months can be perceived in most of the years, in some of them this period is extended or transferred to the fourth trimester, as shown in Fig. 4 for 1990 and 2000.

As mentioned, the degree of correlation involving different stations varies a lot over time, from moments with high correlation to moments with no correlation at all. Figure 5 shows a set of consecutive years (from 1997-2002), highlighting a station (id 12) that correlates with no other station in this entire period.

Despite no correlations involving station 12 are noticed from middle/1997 to middle/2002, station 14 is the one that presents the lowest correlation overall (Figure 6). When analyzing all years' data, one notice that this station ends up eventually correlating with

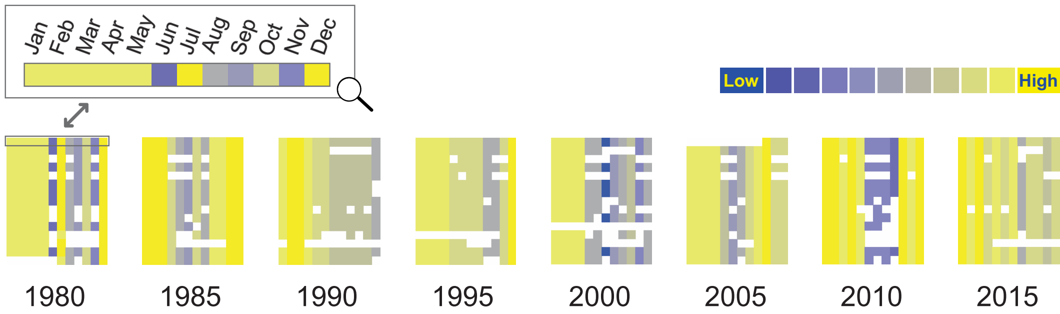

Figure 4. TAM layout for the 15-stations graph. Blue and yellow nodes indicate low and high correlation, respectively. Stations' ids omitted. As illustrated, each column (vertical line) in a year (matrix) corresponds to a month.

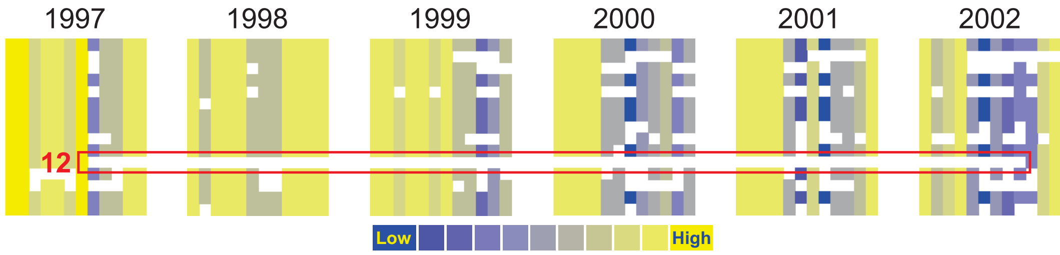

Figure 5. TAM layout for the 15-stations graph highlighting the absence of station 12 between 1997 and 2002, which indicates no correlation involving this station in the represented period. Blue and yellow nodes indicate low and high correlations, respectively.

all the others (Figure 6(a)), even though it has no correlation at all in almost half of the time (notice the absence of this station in several months of the years in Figure 6(b)).

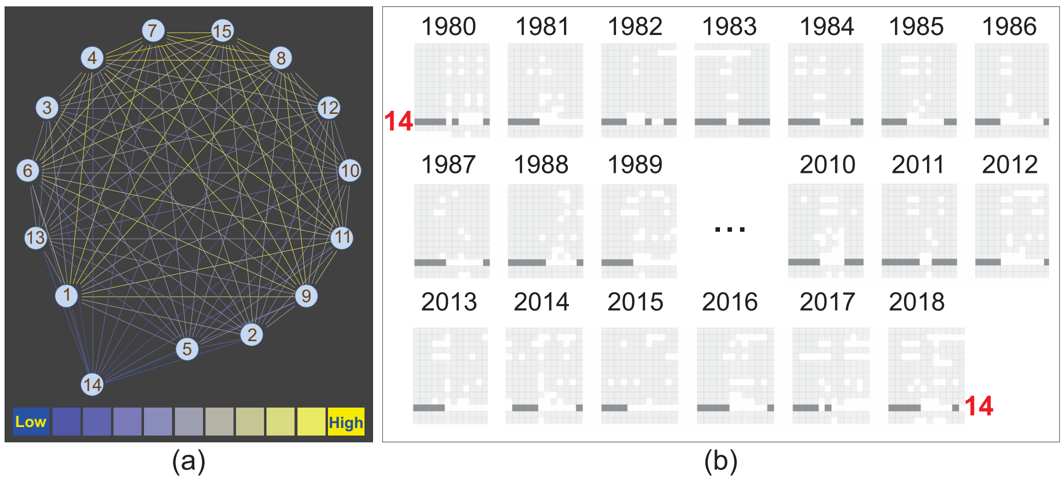

Figure 6. Node-link diagram and TAM layout for the 15-stations graph highlighting the node activity from node 14, which represents the node with the lowest activity among others.

The obtained results can also be related to climatic trends, as in the case of *La Niña*, an event that affects water levels. It is caused by the ocean surface cooling in the central and eastern tropical Pacific Ocean, leading to dry winters in Southern California. according to (Null, 2018), strong *La Niña* events occurred in 1988-1989, 1999-2000, 2007-2008, 2010-2011, and a moderate event occurred in 2011-2012. In Fig. 7, we notice the same years having more blue colors in the corresponding graph, showing fewer similarities between stations. The graph confirms *La Niña* events causing dry winters only for southern California. Different patterns in northern and southern California are apparent in the graphs by more blue pixels in the corresponding years. According to

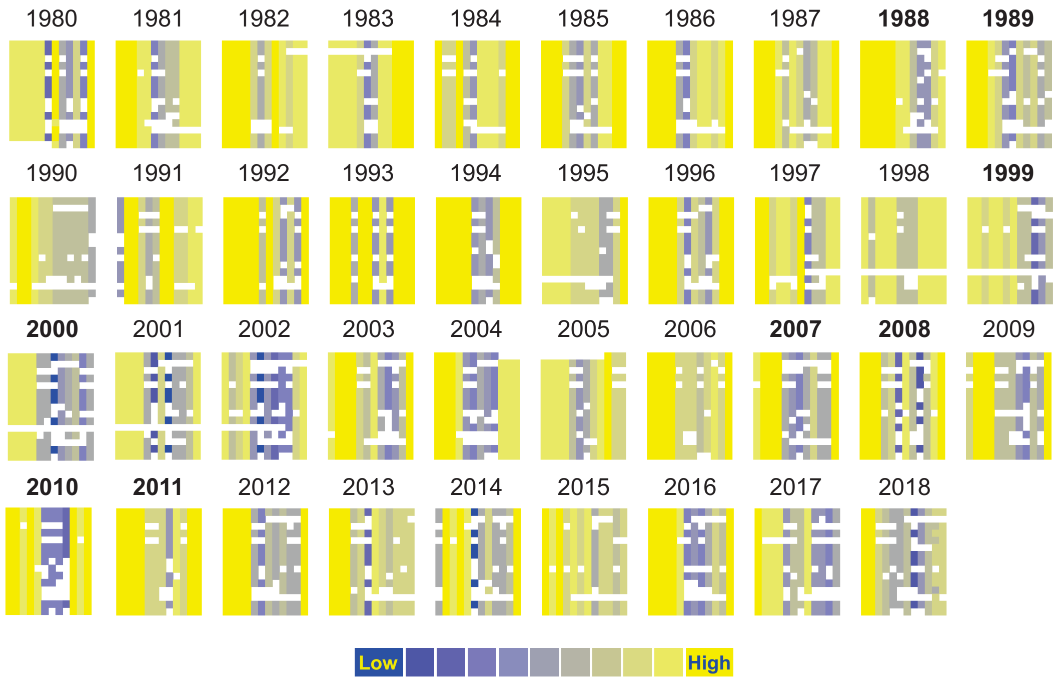

Figure 7. TAM layout for 15-stations graph. *La Niña* years (1988, 1989, 1999, 2000, 2007, 2008) have more blue pixels indicating less similarities (correlation) between different stations. This is caused by northern and southern regions having different patterns. Strong *La Niña* years are highlighted in bold.

According to Allan and Komar (Allan & Komar, 2002), in 1999 the Pacific Northwest witnessed successive *El Niño* and *La Niña* winters. As a result, between 1999 and 2000 the storms offshore from the Pacific Northwest generated waves that exceeded 33 feet, which increased the water levels and caused substantial coastal erosion. As can be seen in Figure 8, the average streamflow difference between 1999-2000 and previous years between January to April is higher in the north of California (e.g. stations no.6-9), approx. 0.50–0.90) than in the south of California (e.g. stations no.1-5 & 10), less than 0.10. The number of stations whose streamflow difference from previous years is less than 0.10 is greater in January when compared to July. Our Dynetvis visualization method shows there are higher strength of association between canonical variates (linear combination of variables) of stations in January to April based on the Canonical Correlation Coefficient measures. It means that climate and hydrological variables show higher correlations in January compared to July in 2000, and demonstrates how employing only streamflow values in the analysis of this scenario may hide interesting patterns produced by other variables. Our layout applied to a combination of variables is able to show patterns related to a broader range of aspects, which may better guide the analysts into a more criterious investigation.

According to NOAA results, the La Niña Winter of 1999-2000 was the warmest winter on record since 1900 in the southeastern United States and a colder winter from the Pacific Northwest to the Great Lakes. Therefore, the sea surface temperatures in the tropical Pacific fell below normal. La Niña caused an increasing in the precipitation in the Pacific Northwest and below normal precipitation in the Southern California.

Because of the lack of research addressing the effects of La Niña on winter temperatures, drawing a conclusion based only in these observations is rather complex. However, some studies show (Pierce, 2005) that a strong and intense La Niña can cause a below normal temperatures across most of the California state. Also this phase is characterized by above-average precipitation in the pacific Northwest and far Northern California and below-average precipitation in Southern California.

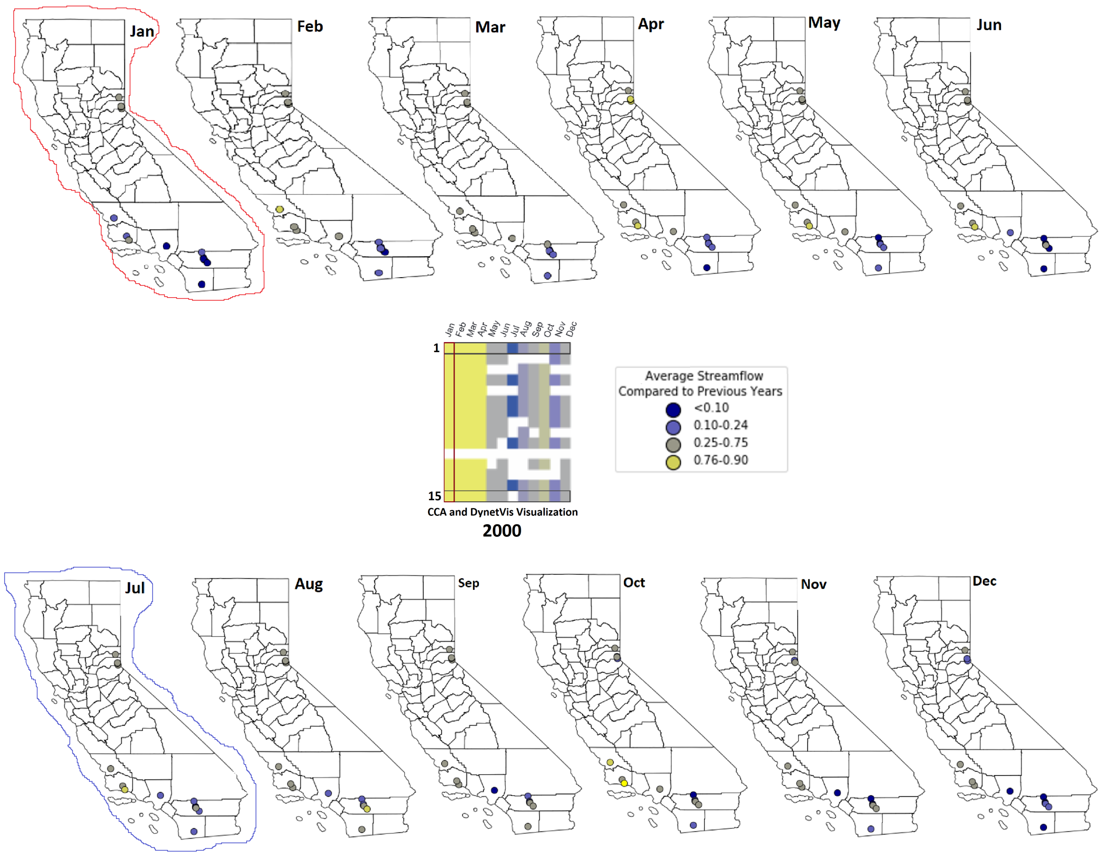

Figure 8. Monthly changes of streamflow in 2000 La Niña

4.2 131-stations graph

We have focused so far on local behaviors involving a small sample of stations. Now, we aim at identifying global patterns that can be noticed when analyzing different watersheds and geographical regions. For this purpose, we consider a set of 131 stations belonging to different watersheds.

As discussed in Section 2.2, community detection is an important and widely adopted strategy that allows the identification of meaningful groups of nodes on a given graph. In the context of this paper, this strategy is used to group watersheds and stations with similar overall behaviors. As illustrated in Fig. 9(a), three communities are obtained after applying Louvain (Blondel et al., 2008) community detection algorithm on the aggregated graph, thus considering the entire observation period at once to focus on overall patterns. By comparing the detected communities (Fig. 9(a)) with the actual watersheds (Fig. 9(b)), we can extract two relevant information. First, each community spreads over different geographical regions and is composed of stations belonging to various watersheds. As a consequence, we can notice stations with overall similar behavior geographically far from each other. Second, particular watersheds may contain stations categorized into different communities, so we can also observe stations with different overall behavior geographically near to each other.

Even though characteristics of specific periods of time may greatly affect temporal behaviors, as we will discuss later, geographical proximity is not mandatory for having similar behavior when considering the entire observation period.

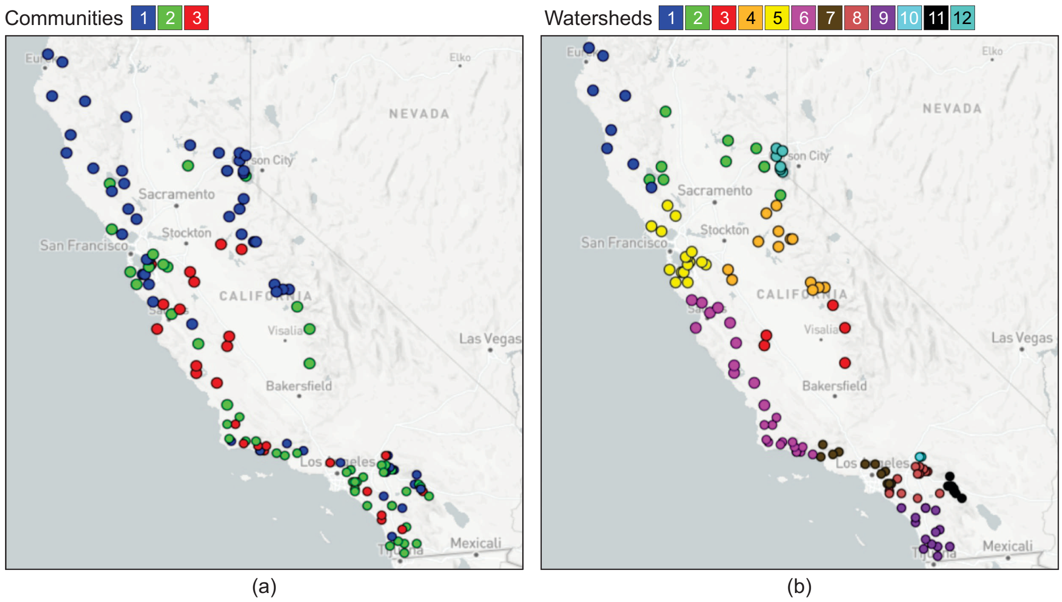

Figure 9. Geographical positioning of the 131 stations. Stations colored according to (a) communities, (b) watersheds. There are 56 stations in community C1 (blue), 49 in C2 (green), and 26 in C3 (red).

Figure 10 shows a TAM layout reorganized such that (i) the same months of different years are grouped, and (ii) the stations are grouped according to the underlying graph community structure detected by Louvain. Communities C1, C2, and C3 refer to the blue, green, and red communities from Fig. 9, respectively. By analyzing this TAM layout, we can now generalize the perception that Jan-Apr and Nov-Dec are the months with the highest correlation degrees, while Jun-Sept are those with the lowest degrees. In this latter period, it is also possible to identify several stations with low correlation – mainly concentrated in C1 –, and several stations with no correlation at all – mainly spread over C2 and C3.

Fig. 11 presents an analogous TAM layout, but now grouping the stations according to the associated watersheds. For almost all watersheds, variations over time in the degree of correlation among their stations follow the expected behavior, varying according to the characteristics of the season (see, e.g., watersheds W1 and W2). One exception, however, occurs with watershed W3: if we consider July and August of all years, we notice a correlation in only 13 months. Despite belonging to the same watershed and

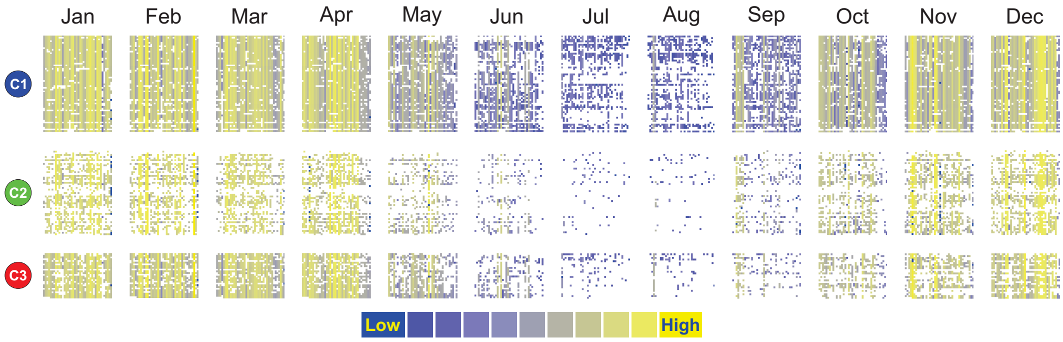

Figure 10. TAM layout for the 131-stations graph, considering all years. Communities C1, C2, and C3 are the same as those from Fig. 9(a). Blue and yellow nodes indicate low and high correlation, respectively.

being relatively close to each other, these locations are poorly correlated during this period of the year. Additionally, Fig. 11 shows that locations in watersheds W1 and W2 are more correlated then locations in W3. When comparing the two maps from Fig. 9, one sees that all stations from W1 are in C1, as well as the majority of stations from W2, which may justify this behavior. Stations from W3 are however split into C2 and C3, representing a heterogeneous watershed, and that is reflected in the poor correlation observed among its stations.

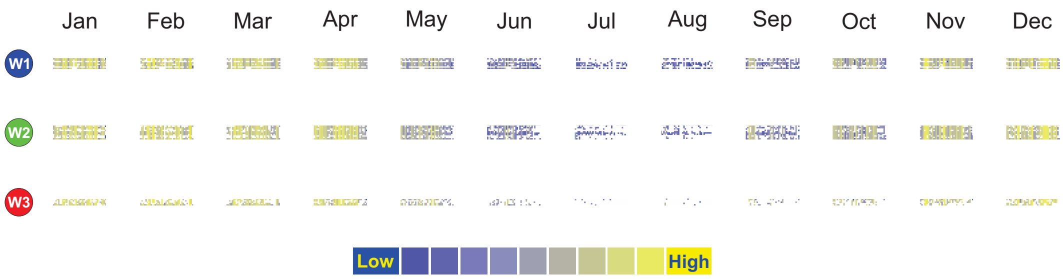

Figure 11. TAM layout for the 131-stations graph. Watersheds W1, W2, and W3 are the same as those from Fig. 9(b). Blue and yellow nodes indicate low and high correlation, respectively.

5 Conclusion

In this paper we applied Canonical Correlation Analysis (CCA) in a combination of climate and hydrological indicators to investigate behavior relationships among California water stations over time. Our analysis was supported by a set of temporal graph visualization approaches that provided means to explore temporal and structural aspects of these relationships, including the concentration of the analysis in specific time periods, or considering specific geographical regions.

Most of the related works employ a single variable, or a single category of variables to perform the analysis, and our experiments showed that such strategy may hide specific patterns and limit the comprehension of specific phenomena and their impacts. In this sense, our decision to apply CCA in this novel combination of climate and hydrological variables provided a broader view of multiple natural aspects and a better comprehension of the scenario. In addition, the layouts produced by the employed visual strategies provided an effective view of the CCA distribution over the regions and over time. We were able to notice general patterns, such as a discrepant behavior in specific months/years, including a significant heterogeneous behavior in the dry season, represented by low correlation values among the stations in these period. We could also identify stations with few or no correlation, representing regions with peculiar behavior. In this sense, our strategy is capable of guiding experts in make a deeper investigation in these stations, as it can represent abnormal natural events related or not to human intervention, or even represent stations in which the collection process was wrongly performed. We have found no significant correlation between the stations geographical location and their behavior, but we could notice that in general the stations located in a specific watershed present homogeneous behavior, and each watershed present a particular behavior over time, although they present similarities among each other. Although we were not able to justify all the produced patterns, we believe they may guide water experts into a more careful investigation.

We were able to identify some limitations in the data collection, as well as in our proposed analysis strategy. The data collected from the stations is deficient for some locations, due to stations technical issues and/or malfunctions, which resulted in missing data for some of them, and for some time periods. Although we considered only a subset of reference stations, this deficiency may influence the results of this analysis, specially if the aim is to investigate human intervention effects. We also considered the entire time period to compute the communities, which may hide behavior evolution patterns related to specific time periods and influence the communities identification. We intend to employ time windows to capture these localized behavior and enhance the analysis.

For future work, besides addressing the aforementioned limitations, it would be interesting to apply this analysis to other geographical regions, in order to identify specific behaviors. It would be also good to perform user experiments with water experts and managers, in order to collect feedback regarding other interesting behavior patterns, as well as possible modifications in the analysis process to enhance these and other findings.

6 Open Research

The network data used for all visualization in the study are available at `https://www.dynetvis.com/datasets`. The software used for analysis and visualization of the networks is freely available at `https://www.dynetvis.com`. DyNetVis website started in 2017 and it is a constant update to include new features (Linhares et al., 2020). The software is developed openly at `https://github.com/travencolo/DyNetVis`

Acknowledgments

This research was supported by grants #2020/10049-0, #2020/07200-9 and #2016/17078-0 from São Paulo Research Foundation (FAPESP), Conselho Nacional de Desenvolvimento Cientfico e Tecnologico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (CAPES).

References

Abatzoglou J. et al. (2017). *Climatology Lab.* `http://www.climatologylab.org/gridmet.html`. ([Online; accessed 13-May-2021])

Agarwal, A., Marwan, N., Maheswaran, R., Ozturk, U., Kurths, J., & Merz, B. (2020, may). Optimal design of hydrometric station networks based on complex network analysis. *Hydrology and Earth System Sciences*, *24*(5), 2235–

2251. doi: 10.5194/hess-24-2235-2020

AghaKouchak, A., Cheng, L., Mazdiyasni, O., & Farahmand, A. (2014). Global warming and changes in risk of concurrent climate extremes: Insights from the 2014 california drought. *Geophysical Research Letters*, *41*(24), 8847–8852.

AghaKouchak, A., Feldman, D., Hoerling, M., Huxman, T., & Lund, J. (2015). Water and climate: Recognize anthropogenic drought. *Nature News*, *524*(7566), 409.

Alipour, M., & Kibler, K. M. (2019). Streamflow prediction under extreme data scarcity: a step toward hydrologic process understanding within severely data-limited regions. *Hydrological Sciences Journal*, *64*(9), 1038–1055.

Allan, J. C., & Komar, P. D. (2002). Extreme storms on the pacific northwest coast during the 1997-98 el niño and 1998-99 la niña. *Journal of coastal Research*, 175–193.

Bartlett, M. S. (1938). Further aspects of the theory of multiple regression. *Mathematical Proceedings of the Cambridge Philosophical Society*, *34*(1), 33–40. doi: 10.1017/S0305004100019897

Bekchanov, M., Sood, A., Pinto, A., & Jeuland, M. (2017). Systematic review of water-economy modeling applications. *Journal of Water Resources Planning and Management*, *143*(8), 04017037.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*, P10008.

Braga, A., Alves, L., Costa, L., Ribeiro, A., de Jesus, M., Tateishi, A., & Ribeiro, H. (2016, feb). Characterization of river flow fluctuations via horizontal visibility graphs. *Physica A: Statistical Mechanics and its Applications*, *444*, 1003–1011. doi: 10.1016/j.physa.2015.10.102

Bureau, U. C. (2021, May). *2020 census apportionment results.* `https://www.census.gov/data/tables/2020/dec/2020-apportionment-data.html`.

California Water Data Maintainer. (2021-05-13). *USGS Current Water Data for California.* `https://waterdata.usgs.gov/ca/nwis/rt`. ([Online; accessed 13-May-2021])

Chang, F.-J., Tsai, Y.-H., Chen, P.-A., Coynel, A., & Vachaud, G. (2015). Modeling water quality in an urban river using hydrological factors–data driven approaches. *Journal of environmental management*, *151*, 87–96.

DeCicco, L. A., & Hirsch, R. M. (2016). *Data Retrieval R package USGS.* `https://cran.r-project.org/web/packages/dataRetrieval/vignettes/dataRetrieval.html`. ([Online; accessed 13-May-2021])

Desai, S., & Ouarda, T. B. (2021). Regional hydrological frequency analysis at ungauged sites with random forest regression. *Journal of Hydrology*, *594*, 125861.

Falcone, J. (2011). *GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow.* `https://water.usgs.gov/lookup/getspatial?gagesII_Sept2011`. U.S. Geological Survey, Reston, Virginia. ([Online; accessed 13-May-2021])

Fang, K., Sivakumar, B., & Woldemeskel, F. M. (2017, feb). Complex networks, community structure, and catchment classification in a large-scale river basin. *Journal of Hydrology*, *545*, 478–493. doi: 10.1016/j.jhydrol.2016.11.056

Forootan, E., Khaki, M., Schumacher, M., Wulfmeyer, V., Mehrnegar, N., van Dijk, A. I., … others (2019). Understanding the global hydrological droughts of 2003–2016 and their relationships with teleconnections. *Science of the Total Environment*, *650*, 2587–2604.

Halverson, M. J., & Fleming, S. W. (2015, jul). Complex network theory, streamflow, and hydrometric monitoring system design. *Hydrology and Earth System Sciences*, *19*(7), 3301–3318. doi: 10.5194/hess-19-3301-2015

Han, X., Sivakumar, B., Woldmeskel, F. M., & de Aguilar, M. G. (2018, mar). Temporal dynamics of streamflow: application of complex networks. *Geoscience Letters*, *5*(1). doi: 10.1186/s40562-018-0109-8

Kamienski, C., Soininen, J.-P., Taumberger, M., Dantas, R., Toscano, A., Salmon Cinotti, T., ... Torre Neto, A. (2019). Smart water management platform: Iot-based precision irrigation for agriculture. *Sensors*, *19*(2), 276.

Kasiviswanathan, K., & Sudheer, K. (2016). Comparison of methods used for quantifying prediction interval in artificial neural network hydrologic models. *Modeling Earth Systems and Environment*, *2*(1), 22.

Li, X., Sha, J., Li, Y.-m., & Wang, Z.-L. (2018). Comparison of hybrid models for daily streamflow prediction in a forested basin. *Journal of Hydroinformatics*, *20*(1), 191–205.

Linhares, C. D. G., Ponciano, J. R., Paiva, J. G. S., Rocha, L. E. C., & Travençolo, B. A. N. (2020, dec). DyNetVis - an interactive software to visualize structure and epidemics on temporal networks. In *2020 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM).* IEEE. doi: 10.1109/asonam49781.2020.9381304

Linhares, C. D. G., Ponciano, J. R., Pereira, F. S. F., Travençolo, B. A. N., Paiva, J. G. S., & Rocha, L. E. C. (2019). A scalable node ordering strategy based on community structure for enhanced temporal network visualization. *Computers & Graphics*, *84*, 185 - 198. Retrieved from `https://doi.org/10.1016/j.cag.2019.08.006` doi: 10.1016/j.cag.2019.08.006

Linhares, C. D. G., Travençolo, B. A. N., Paiva, J. G. S., & Rocha, L. E. C. (2017). DyNetVis: a system for visualization of dynamic networks. In *Proceedings of the symposium on applied computing* (pp. 187–194). Marrakech, Morocco: ACM. Retrieved from `http://doi.acm.org/10.1145/3019612.3019686` doi: 10.1145/3019612.3019686

Lund, N. S. V., Falk, A. K. V., Borup, M., Madsen, H., & Steen Mikkelsen, P. (2018). Model predictive control of urban drainage systems: A review and perspective towards smart real-time water management. *Critical Reviews in Environmental Science and Technology*, *48*(3), 279–339.

Meng, E., Huang, S., Huang, Q., Fang, W., Wu, L., & Wang, L. (2019). A robust method for non-stationary streamflow prediction based on improved emd-svm model. *Journal of hydrology*, *568*, 462–478.

Null, J. (2018). El niño and la niña years and intensities. golden gate weather services. *El Niño and La Niña Years and Intensities: Golden Gate Weather Services*.

Oo, H. T., Zin, W. W., & Kyi, C. C. T. (2020). Analysis of streamflow response to changing climate conditions using swat model. *Civil Engineering Journal*, *6*(2), 194–209.

Ouali, D., Chebana, F., & Ouarda, T. B. (2016). Non-linear canonical correlation analysis in regional frequency analysis. *Stochastic environmental research and risk assessment*, *30*(2), 449–462.

Ouarda, T. B., Bâ, K. M., Diaz-Delgado, C., Cârsteanu, A., Chokmani, K., Gingras, H., ... Bobée, B. (2008). Intercomparison of regional flood frequency estimation methods at ungauged sites for a mexican case study. *Journal of Hydrology*, *348*(1-2), 40–58.

Ouarda, T. B., Girard, C., Cavadias, G. S., & Bobée, B. (2001). Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology*, *254*(1), 157-173. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0022169401004887` doi: https://doi.org/10.1016/S0022-1694(01)00488-7

Papaioannou, G., Vasiliades, L., & Loukas, A. (2015). Multi-criteria analysis framework for potential flood prone areas mapping. *Water resources management*, *29*(2), 399–418.

Pierce, D. W. (2005). Effects of the north pacific oscillation and enso on seasonally averaged temperatures in california. *California Energy Commission CEC-500-2005-002, PIER Project Report*.

Qin, Y., & Horvath, A. (2020). Use of alternative water sources in irrigation: potential scales, costs, and environmental impacts in california. *Environmental Research Communications*, *2*(5), 055003.

RICE, R. M. (1972). Using canonical correlation for hydrological predictions. *Hydrological Sciences Journal*, *17*(3), 315–321.

Schanze, J. (2006). Flood risk management–a basic framework. In *Flood risk management: Hazards, vulnerability and mitigation measures* (pp. 1–20). Springer.

Serinaldi, F., & Kilsby, C. G. (2016, may). Irreversibility and complex network behavior of stream flow fluctuations. *Physica A: Statistical Mechanics and its Applications*, *450*, 585–600. doi: 10.1016/j.physa.2016.01.043

Shneiderman, B. (1996, September). The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 ieee symposium on visual languages* (p. 336-343).

Sivakumar, B., & Woldemeskel, F. M. (2014, nov). Complex networks for streamflow dynamics. *Hydrology and Earth System Sciences*, *18*(11), 4565–4578. doi: 10.5194/hess-18-4565-2014

Sivakumar, B., & Woldemeskel, F. M. (2015, jul). A network-based analysis of spatial rainfall connections. *Environmental Modelling & Software*, *69*, 55–62. doi: 10.1016/j.envsoft.2015.02.020

Stewart, I. T., Rogers, J., & Graham, A. (2020). Water security under severe drought and climate change: Disparate impacts of the recent severe drought on environmental flows and water supplies in central california. *Journal of Hydrology X*, *7*, 100054. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2589915520300055` doi: https://doi.org/10.1016/j.hydroa.2020.100054

Tukimat, N., Harun, S., & Tadza, M. (2019). The potential of canonical correlation analysis in multivariable screening of climate model. In *Iop conference series: Earth and environmental science* (Vol. 365, p. 012025).

Tumiran, S. A., & Sivakumar, B. (2021, jan). Catchment classification using community structure concept: application to two large regions. *Stochastic Environmental Research and Risk Assessment*, *35*(3), 561–578. doi: 10.1007/s00477-020-01936-4

Uurtio, V., Monteiro, J. a. M., Kandola, J., Shawe-Taylor, J., Fernandez-Reyes, D., & Rousu, J. (2017, November). A tutorial on canonical correlation methods. *ACM Comput. Surv.*, *50*(6). Retrieved from `https://doi.org/10.1145/3136624` doi: 10.1145/3136624

Vicente-Serrano, S. M., Miralles, D. G., Domínguez-Castro, F., Azorin-Molina, C., El Kenawy, A., McVicar, T. R., ... Peña-Gallardo, M. (2018). Global assessment of the standardized evapotranspiration deficit index (sedi) for drought analysis and monitoring. *Journal of Climate*, *31*(14), 5371–5393.

Xu, Y., Lu, F., Zhu, K., Song, X., & Dai, Y. (2020, jun). Exploring the clustering property and network structure of a large-scale basin's precipitation network: A complex network approach. *Water*, *12*(6), 1739. doi: 10.3390/w12061739

Yasmin, N., & Sivakumar, B. (2018, sep). Temporal streamflow analysis: Coupling nonlinear dynamics with complex networks. *Journal of Hydrology*, *564*, 59–67. doi: 10.1016/j.jhydrol.2018.06.072

Yasmin, N., & Sivakumar, B. (2020, nov). Study of temporal streamflow dynamics with complex networks: network construction and clustering. *Stochastic Environmental Research and Risk Assessment*, *35*(3), 579–595. doi: 10.1007/s00477-020-01931-9

Zhang, J., Tourian, M. J., & Sneeuw, N. (2020). Identification of enso signature in the boreal hydrological cycle through canonical correlation with sea surface temperature anomalies. *International Journal of Climatology*, *40*(15), 6219–6241.