Journal of Geochemical Exploration 259 (2024) 107438

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/gexplo

Robust CoDA balances and the role of the variance in complex riverine geochemical systems

Caterina Gozzi [a,*], Matthias Templ [b], Antonella Buccianti [a]

[a] University of Florence, Department of Earth Sciences, Via G. La Pira 4, 50121 Firenze, Italy
[b] University of Applied Sciences and Arts Northwestern Switzerland, Riggenbachstrasse 16, 4600 Olten, Switzerland

ARTICLE INFO

Keywords:
Robust principal balances
River basins
Water chemistry
Variance
Complex systems

ABSTRACT

This study introduces a robust method for analyzing the geochemical behavior of chemical species in river catchment water. It focuses on isometric log-ratio coordinates obtained from a sequential partition method that successively maximizes the explained variance in the data set. Robust orthonormal coordinates are created based on hierarchical clustering and robust estimation of the variation matrix. Applying this to the water chemistry of Italy's Arno and Tiber basins, the research reveals the associations of variables in data structure and processes across varying geological and climatic conditions. The method uncovers key contrasting geochemical processes and suggests that the behavior of simple balances characterized by lower variances (i.e., Ca^{2+}/HCO_3^- and Na^+/Cl^-) are mainly influenced by random fluctuations with no differences between classical or robust methods. However, when balances describe more complex geochemical processes resulting in frequency distributions affected by the presence of bimodality or outliers, significant differences among the two approaches emerge, compromising the data interpretation. The proposed metodology offers more insights into the investigation of catchment geochemistry's resilience to hydroclimatic changes, marking a significant step in understanding large-scale environmental dynamics.

1. Introduction

Rivers are a fundamental component of the hydrological cycle and the chemical composition of their water can be used to monitor the ecological status of the environment as well as the response to climatic perturbations. Climate changes induce substantial modifications in the watersheds, which are then reflected in the water chemical composition. Increased aridity makes forests more prone to fires and soils more subjected to erosion, compromising the quality and quantity of water. Despite the importance of several geological and geomorphological factors, the chemical weathering on the Earth's surface is mainly controlled by climate (Dinis et al., 2020; Schmeller et al., 2022). In this framework, riverine waters connect atmosphere, lithosphere, and biosphere through complex dynamics governed by thermodynamical laws, thus generating biogeochemical cycles of elements and modifying the status of the Earth Critical Zone (Kleidon et al., 2012). Furthermore, rivers have been fundamental in the development of human society and are still important today, representing a crucial component for water supply, irrigation, food and energy production, recreation and transportation (Patil et al., 2018). As a consequence, watersheds are the most extensively and deeply altered ecosystem on Earth today subjected to wide variability (Dai et al., 2023; Dede et al., 2023). This is particularly true for the Mediterranean catchments, which are expected to experience substantial climate changes during the 21st century (Lutz et al., 2016). To obtain generalizable insights into river catchment geochemistry, it is necessary to explore the organizing principles that could underlie their heterogeneity and complexity moving beyond the current status of explicitly characterizing local watersheds with very complex models (McDonnell et al., 2007). In this regard, it is crucial to conduct comparative studies of different catchment areas and implement robust methods capable of exploring the sources of geochemical variability on different scales. Analyzing the chemical composition of water allows us to understand how watersheds evolve and retrace the impact of different stressors, as well as identify potential drivers and feedback mechanisms.

Numerous studies suggest that variability, as measured by the statistics of variance, can indicate whether a system is more or less stable in time and space, whether flickering or transitions to alternative states

* Corresponding author.
E-mail address: caterina.gozzi@unifi.it (C. Gozzi).

https://doi.org/10.1016/j.gexplo.2024.107438
Received 24 July 2023; Received in revised form 10 February 2024; Accepted 12 February 2024
Available online 27 February 2024

occur (Dakos et al., 2014; Belle et al., 2017; Chen et al., 2018; Gozzi et al., 2021; Grziwotz et al., 2023). However, investigation of the variance for single variables, as given by the chemical concentration of elements/species dissolved in river water, is not significant for compositional data due to their relative nature (Aitchison, 1982, 1986; Gozzi et al., 2020). Therefore, in this work we use a compositional approach based on the analysis of variance of isometric coordinates obtained with the sequential partition method as defined by Egozcue and Pawlowsky-Glahn (2005) and then modified by Martn-Fernández et al. (2017). Nevertheless, in the presented approach, the sequence of orthonormal coordinates (also called balances) that successively maximize the explained variance in the data set is further implemented by applying a new robust methodology.

Applying the proposed methods, we compare the hydrochemical variability of rivers pertaining to two of the widest catchments in central Italy (Arno and Tiber river basins). The shape of the frequency distribution of the obtained balances is then analyzed to understand which group of variables is better associated with resilient behavior to hydroclimatic changes (Mitzenmacher, 2004; Gozzi and Buccianti, 2022). The newly developed method was implemented in the free and open-source R software environment (R Development Core Team, 2023), specifically in the robCompositions package (Templ et al., 2021; Filzmoser et al., 2018).

2. Materials and methods

2.1. Compositional data analysis

A deeper understanding of the nature of these types of data begins with Aitchison (1982) indicating that log-ratios provide a one-to-one mapping from the constrained sample space of compositional data (called the simplex) and the real space.

We present our data as equivalence classes instead of vector of constant sums (ie. sum to 100 %), taking into account that we are dealing with molar fractions or molarities that do not add up to a single constant (Buccianti and Pawlowsky-Glahn, 2005). The composition $\mathbf{x} = [x_1, \ldots, x_D]'$ with all its parts being strictly positive holds the same relative information in both x_i/x_j and $(ax_i)/(ax_j)$ for any non-zero scalar a. The composition can be expressed as a proportion, $\mathbf{x}^* = a\mathbf{x}$, by setting $a = 1/\sum x_i$. This composition $\mathbf{x}^*$ is part of the standard simplex, which is defined as

$$S_D = \left\{ \mathbf{x}^* = \left[x_1^* \ \ldots \ x_D^*\right]' \middle| x_i^* > 0, \sum_{i=1}^{D} x_i^* = 1 \right\} \quad .$$

The choice of 1 as the closing is arbitrary and does not affect the ratios between variables (van den Boogaart and Tolosana-Delgado, 2013; Filzmoser et al., 2018). Simplex geometry is distinct from Euclidean geometry, and the standard statistical tools have typically been designed for the latter. In fact, around 2000, the algebraic-geometric structure of the simplex was recognized (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001) leading to the proposal of the principle of working in coordinates (Mateu-Figueras et al., 2011). This trend is reflected in compositions being represented by orthonormal coordinates in a real Euclidean space, which is the same as the sample space of compositions.

There are several ways to define orthonormal bases in the simplex, and one of the most used is the identification of a sequential binary partition (SBP) of a compositional vector (Egozcue and Pawlowsky-Glahn, 2005). The motive is that such bases can easily lead to a better interpretation of the meaning of the constructed factors of the grouped parts of the composition. The definition of an orthonormal basis of a sequential partition of a composition is discussed in Egozcue and Pawlowsky-Glahn (2005), which links the idea of balance between groups of components to the sub-composition of a set of parts. Thus, when a SBP process is used, the corresponding coordinates are the balances between the groups of parts separated in each step of a binary partition, and they allow us both subcompositional analysis, i.e. intra-group ratios, and grouped parts analysis, i.e. inter-group ratios.

From a practical point of view, an SBP is a hierarchy of the parts of a composition. In the first order of the hierarchy, all parts are divided into two groups. In the following steps, each group is again divided into two groups. The process continues until all groups have a single part. For the kth order partition, it is possible to define the two subgroups formed at that level: if $i_1, i_2, \ldots, i_r$ are the r parts of the first subgroup (coded by $+1$) and $j_1, j_2, \ldots j_s$ the s parts of the second (coded by -1), the coordinate is defined as the normalized log-ratio of the geometric mean of each group of parts:

$$z_r = \sqrt{\frac{rs}{r+s}} ln \frac{\left(x_{i1}, x_{i2} \ldots x_{ir}\right)^{1/r}}{\left(x_{j1}, x_{j2} \ldots x_{js}\right)^{1/s}} = ln \frac{\left(x_{i1}, x_{i2} \ldots x_{ir}\right)^{a_+}}{\left(x_{j1}, x_{j2} \ldots x_{js}\right)^{a_-}}, \tag{1}$$

where

$$a_+ = +\frac{1}{r}\sqrt{\frac{rs}{r+s}}, \quad a_- = -\frac{1}{s}\sqrt{\frac{rs}{r+s}}. \tag{2}$$

The term a_+ refers to parts in the numerator, a_- to parts in the denominator, and the values of r and s correspond to the k-th order partition. Depending of the number of variables D, following the SBP process will lead to $D-1$ coordinates for which the shape of the frequency distribution can be analyzed to search for variance and resilience indications. However, the original proposal of Egozcue and Pawlowsky-Glahn (2005) then modified by Martn-Fernández et al. (2017) is sensible to the presence of outliers cases or skewness. In our approach, the sequence of orthonormal coordinates that successively maximize the explained variance in the data set is obtained by applying a new robust methodology. The balances obtained in the SBP process governed by the decreasing variance are computed using a hierarchical clustering applied on the robust variation matrix, as explained in the following section. The aim is to discover associations of variables more resilient to flickering processes and/or characterized by input from different sources in a robust framework.

2.2. Robust Orthonormal coordinates

Robust orthonormal coordinates are created on the basis of hierarchical clustering and robust adaptations. The robust methods that are subsequently used are presented first.

2.2.1. Robust covariance estimation

Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ of n coordinates, that is, compositions expressed in coordinates (cf. Eq. 1), of the $n \times D$ matrix $\mathbf{Z}$. The classical estimators for the covariance

$$\mathbf{S}_\mathbf{z} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{x}_i - \overline{\mathbf{z}})',$$

(so as the classical estimation of the (column-wise) arithmetic means: $\overline{\mathbf{z}}$) are sensitive to outliers with a finite sample breakdown point $\frac{1}{n}$ (Maronna et al., 2006), because an outlier observation can make these estimators arbitrarily large just by changing any of the values of a coordinate.

Nowadays, several robust counterparts of location and covariance are available, including the property of affine equivariance, that is, the covariance estimate is appropriately adjusted under shifted, rotated, or rescaled versions of $\mathbf{Z}$. The Minimum Covariance Determinant (MCD) estimator is a robust statistical method used for estimating the center and scatter of multivariate data. It is designed to be resistant to the influence of outliers in the data, making it particularly useful when dealing with datasets that may contain aberrant observations. The MCD estimator (Rousseeuw and Van Driessen, 1999) works by finding a subset of

h observations that has the smallest determinant of the covariance matrix. The MCD estimator then computes the mean and covariance matrix of this subset as an estimate for the location and covariance. It is worth mentioning that the MCD algorithm also includes a re-weighting step that also involves correction factors due to finite sample correction. After an initial estimate of the MCD subset is obtained, the re-weighting step is applied to adjust the influence of each data point on the MCD estimator. This process involves evaluating each data point's distance from these robust estimates to determine their weights in the final model. The goal is to refine these estimates by potentially including additional observations that align well with the already established robust pattern, but might not have been part of the initial MCD-selected subset.

The weights are then used to better represent data points that are closer to the MCD subset and less weight to those that are farther away. Estimates of the column-wise arithmetic means and the (Pearson) covariance matrix of the reweighted subset of data form the solution of the MCD estimator. An alternative is to use the MM estimator, which has higher efficiency but also higher computational costs. As it provided very similar results with our data sets, as is the case with some other studies (to give some examples: Todorov et al., 2011; Templ et al., 2019), we have not reported them. The version that uses the MM estimator can be selected in the procedures that we provide together with this article and that are available in the R package `robCompositions` (Templ et al., 2011).

2.2.2. Robust estimation of the variation matrix

We focus on estimating with sample data and for the definitions of theoretical concepts related to random variables, we refer to Filzmoser et al. (2018).

The estimated variation matrix is a $D \times D$ matrix of sample variances for all pairwise log-ratios of a compositional data sample. For a compositional data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_D)' \in \mathbb{R}^D$, the element (j,k) of the variation matrix, denoted as $\widehat{\mathbf{T}}$, is estimated by $var_s\left(ln(\mathbf{x}_j/\mathbf{x}_k)\right)$, where var_s denotes the sample variance, for $j,k \in \{1,\ldots,D\}$.

The variation matrix can be expressed by using centered log-ratio coordinates $\mathbf{Y}$ of $\mathbf{X}$ (more details and derivation in Filzmoser et al., 2018).

For an $n \times D$ matrix $\mathbf{X}$ of compositional sample data, with the compositions $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})$ in the rows of $\mathbf{X}$, for $i = 1, \ldots, n$, the matrix of clr coefficients $\mathbf{Y}$ is formed by the rows

$$\mathbf{y}_i' = (\mathrm{clr}(\mathbf{x}_i))' = \left(ln \frac{x_{i1}}{\sqrt[D]{\prod_{k=1}^{D} x_{ik}}}, \ldots, ln \frac{x_{iD}}{\sqrt[D]{\prod_{k=1}^{D} x_{ik}}} \right). \quad (3)$$

For $\mathbf{Z}$ the principal balance coordinate representation of $\mathbf{X}$ (see Eq. (1), the centered log-ratio coefficients $\mathbf{y}$ of this composition can be obtained by

$$\mathbf{Y} = \mathbf{V}\mathbf{Z} \quad \text{and} \quad \mathbf{Z} = \mathbf{V}'\mathbf{Y}, \quad (4)$$

see Egozcue et al. (2003) and Tolosana-Delgado et al. (2019). The $D \times (D-1)$ matrix $\mathbf{V}$ represents orthonormal basis vectors that express the relationship of the isometric logarithmic ratio coordinates in Eq. (1) and centered log-ratio coefficients. In our case, these orthonormal basis vectors are obtained in the SBP process governed by the decreasing variance using hierarchical clustering on the robust variation matrix. It represents the sign matrix of our obtained sequential binary partition. It is a matrix similar in shape to the principal balance coordinates in Eq. 1, but contains only values of $+1$, 0, or -1. These values indicate whether a particular part of the composition is involved in the numerator $(+1)$, the denominator (-1), or not involved at all (0) in each coordinate of the principal balance coordinate transformation of Eq. 1.

Thus the robust covariances $\widehat{\mathbf{C}}_{\mathrm{MCD}}$ (or alternatively with the MM estimator, $\widehat{\mathbf{C}}_{\mathrm{MM}}$) are estimated. It is irrelevant whether we work with pivot coordinates, pca scores or even alr coordinates; any coordinate representation (hence full-rank) would suffice. A robust estimate of the variation matrix $\mathbf{T}$ is then obtained by

$$\widehat{\mathbf{T}}_{\mathrm{MCD}} = \mathbf{J}diag(\mathbf{V}\widehat{\mathbf{C}}_{\mathrm{MCD}}\mathbf{V}') + diag(\mathbf{V}\widehat{\mathbf{C}}_{\mathrm{MCD}}\mathbf{V}')\mathbf{J} - 2\mathbf{V}\widehat{\mathbf{C}}_{\mathrm{MCD}}\mathbf{V}'. \quad (5)$$

Filzmoser et al. (2018) called this approach to estimate the robust variation matrix as variation based on robust pivot coordinates.

It is essential to employ Eq. (5) since the centered log-ratio coefficients are singular, and the MCD estimator would not work on singular matrices. Therefore, it is necessary to apply the robust covariance fit in isometric logarithmic coordinates and to work with the orthonormal basis vectors $\mathbf{V}$ to project the result to a centered log-ratio representation.

2.2.3. Obtaining principal coordinates from the robust variation matrix by hierarchical clustering

The initial start of a (agglomerative) hierarchical clustering is that each object forms a separate cluster, thus having n different clusters, with n the number of observations in a data set. Clusters are then merged by a stepwise procedure. In each step, the number of clusters is reduced by 1, merging the two most similar clusters. Thus, it is important to measure the dissimilarity between clusters. In the end, only one cluster is left, all objects lie in this cluster. Note that a "height" (dissimilarity when merging) is assigned to the newly obtained cluster and this height is used to represent these steps in a dendrogram. The dissimilarity between clusters can be measured in different ways. One example is single linkage, where the minimum distance of all members in one cluster to members of another cluster is measured. This tends to have clusters of different sizes, as large clusters tend to merge quickly.

The Ward method (Murtagh and Legendre, 2014) in hierarchical clustering is a criterion that minimizes the total within-cluster variance. At each step, the pair of clusters with minimum between-cluster distance is merged. To prevent any misunderstandings due to the numerous implementations of the Ward method, we will refer to it as *ward.D2* in the following, to emphasize that squared distances are employed as outlined in Murtagh and Legendre (2014). The Ward clustering technique has the advantageous characteristic that the distance between two clusters is proportional to the variance of the balance between the two clusters that are to be combined.

Instead of clustering observations, it is often useful to cluster the variables. The dissimilarity matrix, as input for a hierarchical clustering procedure, is then usually built on correlations between variables. More precisely, the dissimilarities between variables with index i and j can then be expressed by correlations r_{ij} after its transformation into dissimilarities by $d_{ij} = 1 - |r_{ij}|$. For elements of the dissimilarity matrix of compositional data, the variation matrix can be used to form the dissimilarity matrix as input to hierarchical clustering. Interpreting correlations based on symmetric pivot coordinates is more challenging than analyzing variations, so variations are typically used as the standard approach. Our software also includes the version with symmetric pivot coordinates (Kynčlová et al., 2017).

In Section 2.2.2 we have already introduced the sign matrix. More precisely in our problem, a sign matrix of dimension $(D \times (D-1))$ expresses the sequential binary partition based on a merged structure of a hierarchical clustering algorithm that contains values of $+1$, 0 or -1. For example, if variables 3 and 5 are first merged using the hierarchical clustering procedure in a compositional data set of 5 variables, then the first balance is built with a basis $e_1 = (0, 0, -1, 0, 1)$, that is, z_r in Eq. 1 is built with the logarithmic ratio of x_{i3} and x_{i5}. Subsequently, all balances are formed by the corresponding sequential binary partition based on the merged structure resulting from the hierarchical clustering algorithm. Thus, the first resulting coordinate contains the two parts with most similar variance of log-ratios considering all D clusters (each part of a composition is a cluster when starting the hierarchical clustering algorithm). The second coordinate expresses the second most similar

log-ratios but based on the remaining $(D-1)$ clusters, thus members of one cluster joined (noted by $+1$) to other members of another cluster (noted by -1).

In other words, to create compositional (principal) balances, a clustering algorithm is employed. The first step is to identify the smallest entry in the variation matrix **T**. The parts corresponding to this entry are then merged into a group. Next, the variation matrix is updated accordingly based on the previous merge of two clusters. This process is repeated iteratively, with groups of parts merged according to the smallest variance of the corresponding balance. Eventually, in the final phase of the algorithm, the last two remaining groups are merged to obtain the balance with the largest variance (Martn-Fernández et al., 2017). Our contribution is to robustify the procedure of Egozcue and Pawlowsky-Glahn (2005) and Martn-Fernández et al. (2017) by a robust variation matrix based on robust pivot coordinates and to show the consequences of such a robustification in practical applications.

Additional features of the procedure would be visualizing the role of each balance in a dendrogram explaining the decomposition of the total variance (Egozcue and Pawlowsky-Glahn, 2005). More precisely, the compositional dendrogram is the graphical representation of a sequential binary partition where the vertical scale is proportional to the total variance, as described in Pawlowsky-Glahn and Egozcue (2011).

Although the compositional dendrogram has indeed been implemented showing vertical quantile boxplots in addition, yet, in the spirit of simplicity similar to Martn-Fernández et al. (2017), our primary attention is dedicated to the facet of robust estimation.

2.2.4. *Comparing two hierarchical structures*

The cophenetic correlation between two hierarchical clustering results is a measure of how similarly the two clusterings preserve the pairwise distances between the original data points. It involves comparing the cophenetic distances from each hierarchical clustering's dendrogram to each other. To receive the cophenetic correlation between two clustering structures, the cophenetic distance between two points in a cluster is first calculated, which is the height at which these points are first joined together in the dendrogram. The cophenetic correlation coefficient is then calculated using the Pearson correlation from the cophenetic distances from both clustering results. Essentially, this correlation assesses whether the two hierarchical clustering results produce similar structures and groupings in the data. A high cophenetic correlation would indicate that both clusterings agree on the data's structure, while a low correlation would suggest that they are interpreting the data structure differently.

The variation matrix, either robustly estimated or classical, must be estimated before clustering it. This leads to a clustering structure, say the reference clustering structure. In a leave-on-out sensitivity analysis, the clustering is applied to the (robust or classical) variation matrix obtained with one observation left out. This leads to another clustering structure that can then be compared with the reference clustering structure using the cophenetic correlation. This is repeated for each

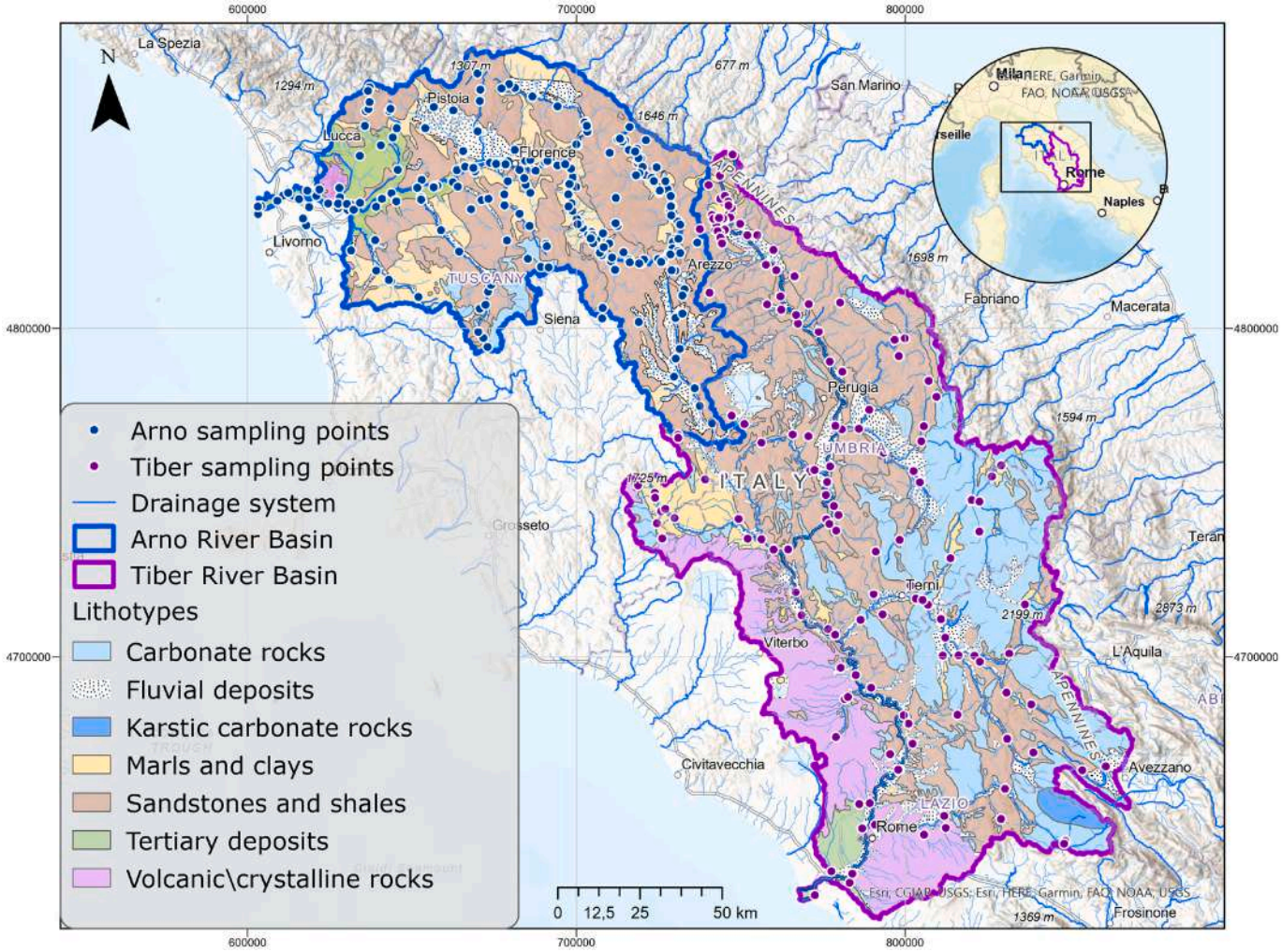

Fig. 1. Map illustrating the geographical positioning of the Arno and Tiber River basins in central Italy, their respective sampling points and primary lithotypes. The latter were derived and modified from ISPRA Ambiente (2017).

observation once left out. For the simulation procedure presented subsequently, the variation matrix is estimated once based on the entire data set prior to clustering. This leads again to a reference clustering structure. Another clustering structure can be obtained by introducing outliers into the data set before clustering. Again, these two clustering structures can then be compared using the cophenetic correlation.

2.3. Geochemical datasets and catchments settings

The proposed methodology was applied to two geochemical data sets associated with different Mediterranean catchments. The testing areas were identified by the two largest river basins in central Italy, the Arno and Tiber river basins, draining an area of 8228 km^2 and 17,156 km^2, respectively (Fig. 1). Regarding the Arno catchment, a sampling field trip, encompassing seasonal variability, was conducted in 2002–2003 (Nisi et al., 2008), followed by an additional upgrade in 2019 for monitoring purposes, resulting in a total of 521 samples. The data set for the Tiber River basin includes 160 water samples collected in 2017 and 62 samples taken in 2018 from selected locations during supplementary monitoring campaigns. The sampling strategy was uniform for both basins, covering the entire extent of their respective watersheds, from the major rivers to the smallest branches in the upper sub-catchments. More details on sampling and analytical methods can be found in Gozzi et al. (2019) and Nisi et al. (2008). All analyzes considered nine main dissolved species, defining the primary water composition as HCO_3^-, F^-, Cl^-, NO_3^-, SO_4^{2-}, Ca^{2+}, Mg^{2+}, Na^+, K^+, with concentrations measured in *mg/L*. These elements serve to trace the most relevant weathering processes that take place in each catchment.

The geological features of the Tiber and Arno river basins are intimately connected to the recent evolution of the Apenninic chain (e.g. Bonini, 1998). Situated in the south-eastern sector of the Northern Apennines, the Tiber basin is shaped by a NE to *E*-verging arcuate fold and thrust belt. It predominantly exhibits terrigenous deposits in its upper section, the carbonatic Apennine ridge to the south-east, and potassic and ultra-potassic volcanic complexes in the south-western area (Fig. 1). On the other hand, the Arno basin is located in the central sector of the Northern Apennines, featuring predominantly sedimentary folded and faulted Mesozoic and Tertiary units resulting from the formation of the Apennine chain (e.g. Carmignani et al., 1994). The Tiber river basin has a mean altitude of 520 m and only 6 % of the total area of the basin exceeds 1200 m (Panichi et al., 2005). The Arno river basin has a slightly lower average elevation (353 m) with 68 % of its surface characterized by a hilly landscape (Cencetti and Tacconi, 2005). Both the Arno and Tiber basins fall within the temperate climate zone. Their annual rainfall pattern exhibits a minimum in summer and two maxima, one in November–December and another at the end of winter.

3. Results

3.1. Classical and robust balances

The results of the SBP process for the river chemistry of the Arno and Tiber basins, derived from both classical and robust procedures, are presented in Figs. 2. The boxes provide a detailed description of the steps of the SBP process allowing to follow the fate of each chemical species participating in the balance construction. The eight balances are labeled

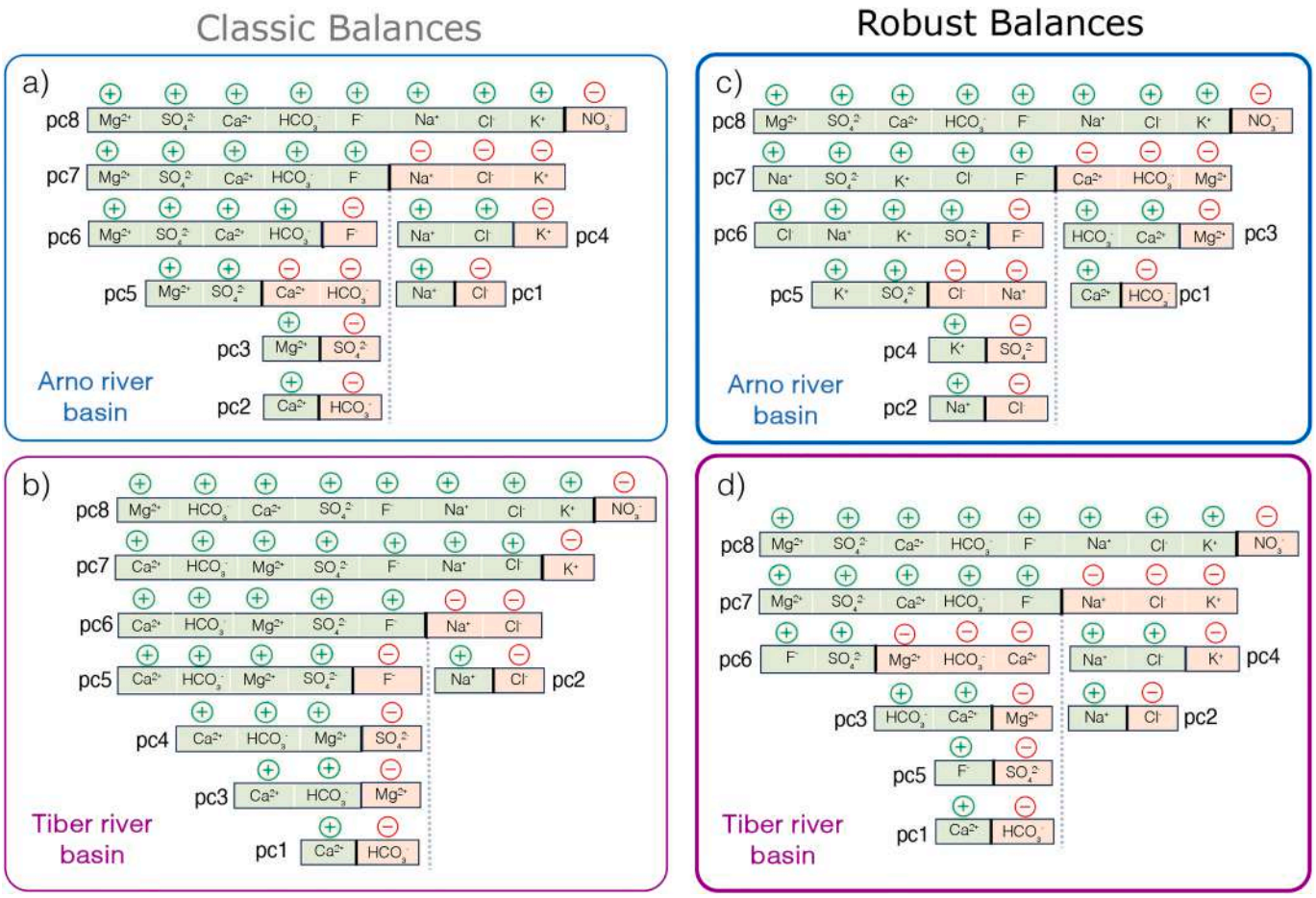

Fig. 2. SPB process for the waters of the Arno and Tiber river basins obtained using classical (a,b) and robust approaches (c,d) based on the principle of decreasing variance from *pc*8 to *pc*1.

*pc*1, *pc*2, …, *pc*8 and the explained variance decreases from *pc*8 to *pc*1. The results unveiled that the balance having the highest variance is *pc*8 (rest of the variables vs. NO_3^-) for both basins, notwithstanding the method applied. From *pc*7 to *pc*3, the balances derived from the classical approach deviate from the robust ones. In the case of the Arno basin, *pc*7 (robust) is distinguished by two sets of variables, namely Na^+, SO_4^{2-}, K^+, Cl^- and F^- versus Ca^{2+}, HCO_3^- and Mg^{2+}. Similarly, for the Tiber basin, *pc*7 (robust) identifies two groups but with different variables: Mg^{2+}, SO_4^{2-}, Ca^{2+}, HCO_3^- and F^-, versus Na^+, Cl^-, and K^+. The primary distinction between the two basins appears to be related to the assignment of F^- and SO_4^{2-} in *pc*7 (robust), which are associated with carbonate species (Mg^{2+}, HCO_3^-, and Ca^{2+}) in the Tiber basin and with silicate species in Arno. The robust balances associated with the lower variability (*pc*1 and *pc*2) are instead common to both cases and are represented by the simple Ca^{2+}/HCO_3^- and Na^+/Cl^- log-ratios. By comparing the two approaches, it is clear that when the variance of the balance is the highest or the lowest, similar results are obtained while in the "middle-earth" important differences are observed.

The behavior of the eight balances is illustrated for the two study cases in Fig. 3 using box-plots and frequency distribution diagrams. In the search for indices capable of monitoring the state of the environment, investigating the behavior of the balances associated with the lowest and the highest variance could represent a promising opportunity (Scheffer et al., 2015). When data appear unimodal and clearly tend to concentrate around a central value, variability could become a less evident feature. In general, the histograms of the balances that are sufficiently symmetrical, such as those obtained for Ca^{2+}/HCO_3^- and Na^+/Cl^-, indicate that the original molar ratios are log-normal distributed. Molar ratios that include elements as Ca^{2+} and HCO_3^- or Na^+ and Cl^- have been widely used in river chemistry investigation (Gaillardet et al., 1999). Moreover, when examining the average log-ratios calculated for unpolluted and polluted rivers in Europe, as reported in Berner and Berner (1996), the values stand at $ln(Ca^{2+}/HCO_3^-)$ equal to −1.2 and 0.38, and $ln(Na^+/Cl^-)$ equal to −0.99 and 0.19, respectively. For $ln(Ca^{2+}/HCO_3^-)$, median values of −1.22 were obtained for Tiber and Arno data, likely indicating unpolluted conditions. Conversely, the values of $ln(Na^+/Cl^-)$ appear to be closer to those reported for polluted conditions, especially for the Tiber river (0.17).

On the other hand, the balance *pc*8 associated with the highest variance, involving NO_3^- versus all the other variables in both catchments is characterized by presence of plurimodality, skewness and outliers. When the frequency distribution deviates from a symmetrical shape due to multimodality, high data fragmentation and the possible presence of alternative stable states emerges. These factors favor flickering processes, leading to increased instability and resilience loss (Dakos et al., 2014; Scheffer et al., 2015).

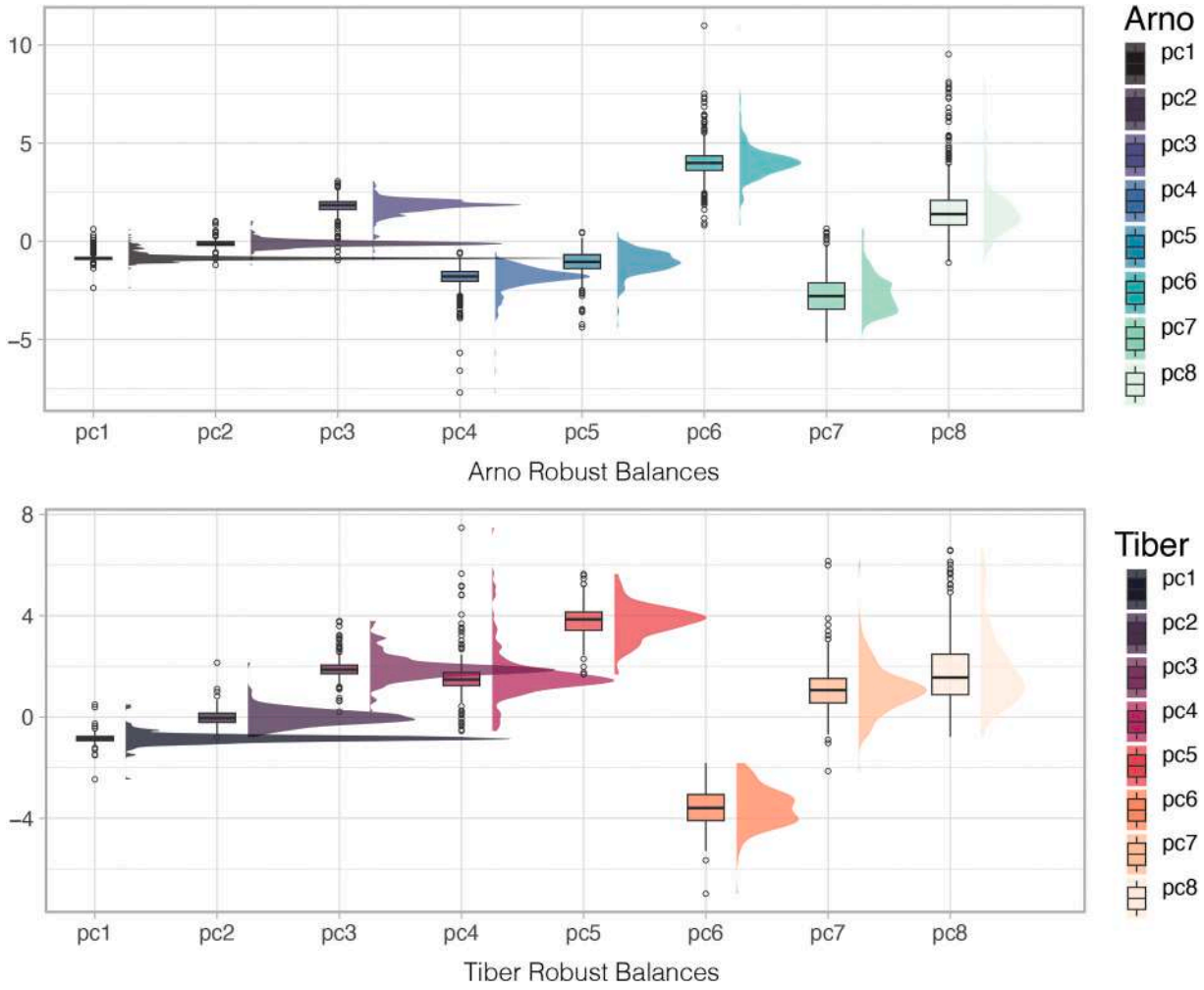

Fig. 3. Box-plots showing the eight principal balances obtained for Arno and Tiber river waters using the robust procedure. Variables involved in each balance are reported in Figs. 2(c-d).

3.2. Which observations have the greatest influence on principal balances? Simulation results

In the Tiber river basin, observations with a significant impact on both the variation matrix and the resulting SBP process comprise two saline springs located in the upper and central areas of the catchment, namely Acqua Cetra and S. Susanna. These are depicted in purple and dark pink, respectively, in Fig. 4. Concerning the Arno River, the most impactful observations are situated in the lower stretches of the Elsa and Era rivers, which drain recent marine deposits and scattered outcrops of marine evaporites (Dinelli et al., 1999).

In order to deeply investigate the influence of outliers values on our approach, a simulation study is carried out by using geochemical data with potential anomalous cases. The simulation of the data is based on the hypothesis of the multivariate normal distribution with parameters that have been estimated from the data of the Tiber river basin considering a specified number of outliers $n_{(out)}$. The approach is a balance between the complexity of the models and the taking in charge of a few multivariate characteristics of realistic data, such as the different covariance structures due to the presence of outliers.

The data generation proceeds as follows:

1. Identify potential outliers using robust Mahalanobis distances by flagging observations as outliers if they exceed a specified quantile threshold (0.975 in our case). The outlier detection procedure is based on (robust) Mahalanobis distances in isometric logratio coordinates (Filzmoser and Hron, 2008) and the MCD estimator (Rousseeuw and Van Driessen, 1999).
2. Represent the geochemical measurements of the Tiber River basin data in centered log-ratio coordinates.
3. Estimate the covariance matrix, say $\widehat{\mathbf{C}}$, for the non-outlier clr-transformed data using the Orthogonalized Gnanadesikan-Kettenring (OGK) method with the benefits of using pairwise estimation, avoiding singularity issues of centered log-ratio coefficients and the simpler meaning of them (*average* of one part to the whole composition) compared to, e.g., isometric coordinates (dominance of parts to other parts).
4. Similarly, compute the covariance matrix, say $\widehat{\mathbf{C}}_{(out)}$, for the identified outlier data.
5. Generate a new set of data points with the same dimension as the original Tiber River basin data, $\mathbf{X}^*$, from a multivariate normal distribution with the mean and covariance obtained in step 3.
6. If $n_{(out)}$ is greater than 0, replace the last $n_{(out)}$ points in $\mathbf{X}^*$ with outliers generated similarly, but using parameters from $\widehat{\mathbf{C}}_{(out)}$ and scaling the mean by a factor of 10. This scaling factor is used to generate larger outliers, since otherwise the outliers overlap with non-outliers. This data with some of its observations replaced by outliers is referred to as $\mathbf{X}^*_{(out)}$.
7. Convert the simulated back to the original space using the inverse centered log-ratio transformation.

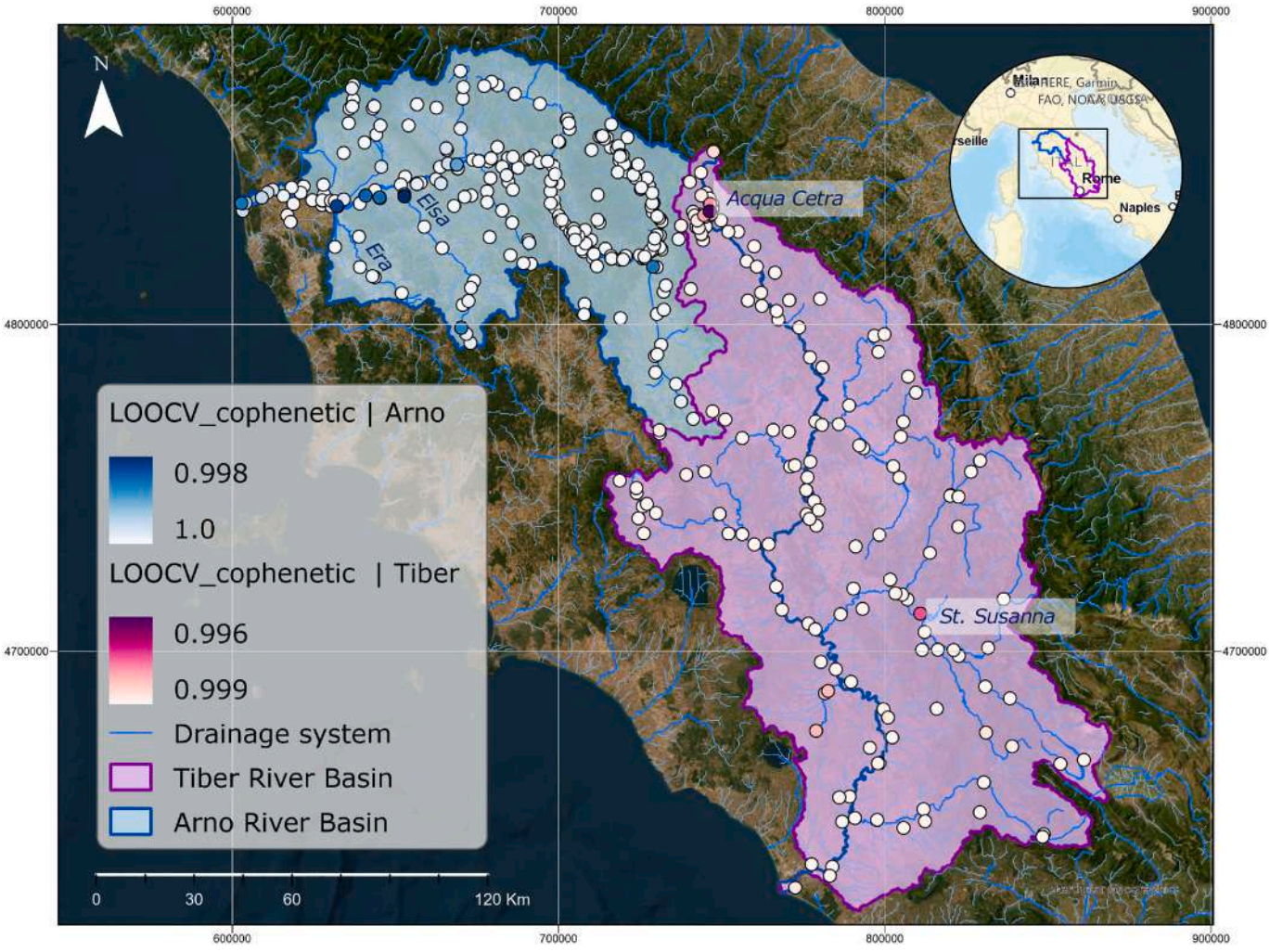

Fig. 4. Maps illustrating the observations with significant impact on both the variation matrix and the resulting SBP process.

The resulting data set comprises both regular observations and artificially introduced outliers, suitable for testing outlier detection and robust statistical methods in geochemical analysis. Fig. 5 shows the average results from 2000 simulations for each different number of outliers. In this way, the cophenetic correlation between the hierarchical clustering structure obtained from the variation matrix of the full data set without outliers $\mathbf{X}^*$ and those obtained from $\mathbf{X}^*_{(out)}$ is computed. For comparison reasons, the average linkage, single linkage, and a hierarchical clustering method called genieclust (Gagolewski, 2021) are compared. Average linkage and *ward.D2* gives comparable results. If one does not use a robust estimation of the variation matrix, the effect of outliers is quite noticeable (as seen in the left plot of Fig. 5). The cophenetic correlation between the hierarchical clustering structures is large as soon as few outliers are introduced. This means that a few outliers will change the hierarchical clustering structure obtained. Consequently, only a few outliers may have a large potential effect on the choice of principal coordinates. The graphic to the right of Fig. 5 shows that the results do not change much until 111 outliers of the 222 observations. Naturally, with more than 50 % outliers, the results are driven by the outliers that now represent the majority of the data points.

4. Discussion

Rapid and extensive global changes affect hydrological cycle so that the concept of invariance of climate, land use, morphology and dynamics for different spatio-temporal ranges cannot be sustained any more. In this framework, catchments can change from stationary systems to transient ones, a condition governed by greater variability. From a thermodynamic perspective, catchments are open and dissipative systems operating far from equilibrium. They have the capacity to undergo substantial structural changes and can spontaneously develop self-organization (Kleidon et al., 2009). In fact, the state far from thermodynamic equilibrium is maintained by a range of processes that continuously perform work, dissipate energy, and thereby produce entropy. Following Prigogine and Stengers (1984) all dissipative systems contain subsystems with permanent fluctuation. However, at a certain moment of the time, the fluctuations become so strong that break the original system to generate a new state. In this context, the irreversible processes governing dissipative structures are able to produce order rather than chaos. The nonequilibrium state is a source of ordering and progressive development, where the components of the systems interact with each other, and the system works as a whole. Inside the hydrological cycle, the water-rock system operates in equilibrium-non-equilibrium states everywhere in natural conditions, acting like a dissipative structure (Shvartsev, 2009).

Taking into account the previous considerations, the variation patterns in the concentration of the chemical components within natural waters are fundamental since they inherit the organization and history of the catchment (Rinaldo et al., 2014; Allen, 2017). The hypothesis is that the degree of variability reflects the sensitivity and adaptability of the system and that the investigation of this property can be related to the probability models that generate the data. Due to the interconnections inherent in compositional data, this task cannot be addressed investigating the role of single variables but rather that of more complex indices, such as balances obtained using a SBP process governed by the decreasing of variance.

The robust approach developed in this paper has yielded additional insights compared to prior investigations (Gozzi et al., 2019) and interesting results have been found that are valid regardless of the catchment investigated. In fact, for the two cases of study, *pc*8 balance (i.e., all the other variables vs. NO_3^-) explain the highest variance. From a geochemical point of view, this index could represent the contrast between anthropic sources and geogenic input (i.e. the whole water/rock interaction, either with silicate- or carbonate-dominated landscapes)

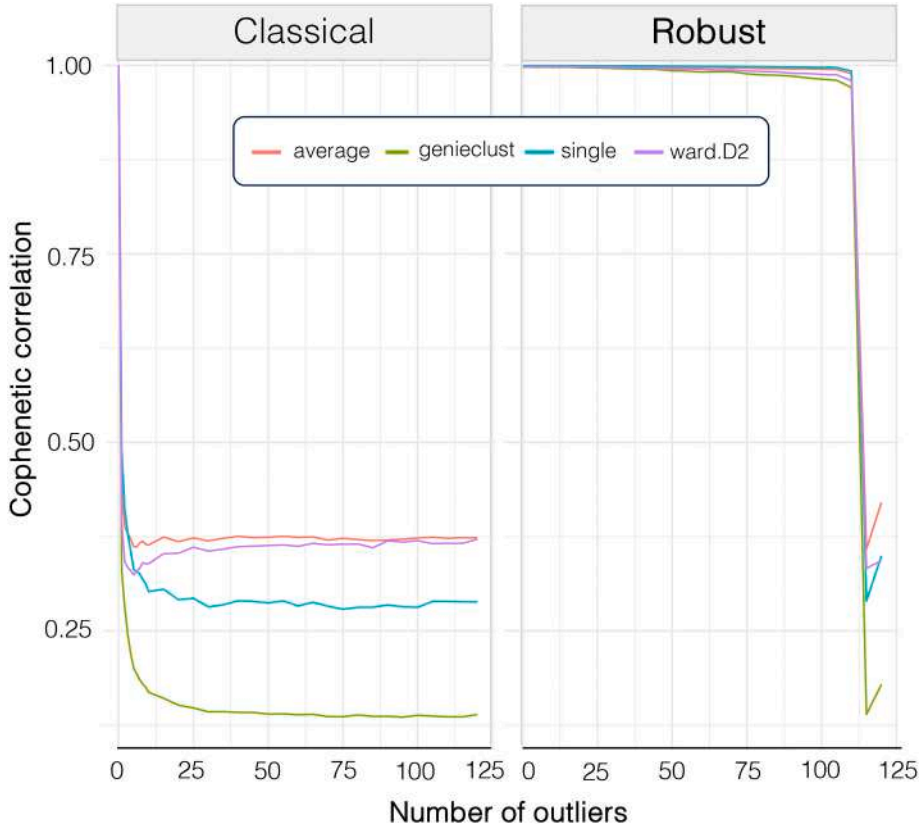

Fig. 5. The cophenetic correlation was used to compare the clustering structure of the variation matrix of the outlier-free data with that of the outlier-induced data, based on 2000 replications, while changing the number of outliers.

(Schlesinger et al., 2020). The persistence of the identical *pc*8 balance in both robust and classical methods indicates that its high variability is an inherent characteristic of the system. There are no outliers in this log-ratio; rather, there are extreme values that may still be part of the main distribution.

The distinct variable associations observed in *pc*7 for the two basins likely mirror the geological characteristics of their respective catchments. Notably, a more pronounced role of silicate weathering is evident in the Tiber river, while the Arno river exhibits a stronger influence of carbonate weathering. These results are likely attributed to the higher prevalence of volcanic rocks in the Tiber basin, as opposed to their almost negligible presence in the Arno basin (refer to Fig. 1). These distinctions would be less evident if solely the classical method had been employed.

To further explore potential sources of variability, also related to the spatial behavior of the variables, *pc*8 values are represented in Fig. 6 using graduated colors for two available seasons (summer and winter). In addition to the clear effects on the index variability of the lithological composition of the drainage areas, it should be noted that seasonality also affects the distribution of *pc*8. This could indicate that additional seasonal-induced processes or anthropic inputs could also play a role in increasing *pc*8 values, thus creating multi-modalities. For the Tiber river waters, lower values of *pc*8 (higher relative content of NO_3^-) are found along the main course especially in summer (yellow points; Fig. 6a). These are likely related to anthropogenic inputs from urban centers (e.g. city of Rome) and diffuse agricultural activities and farming areas (Gozzi

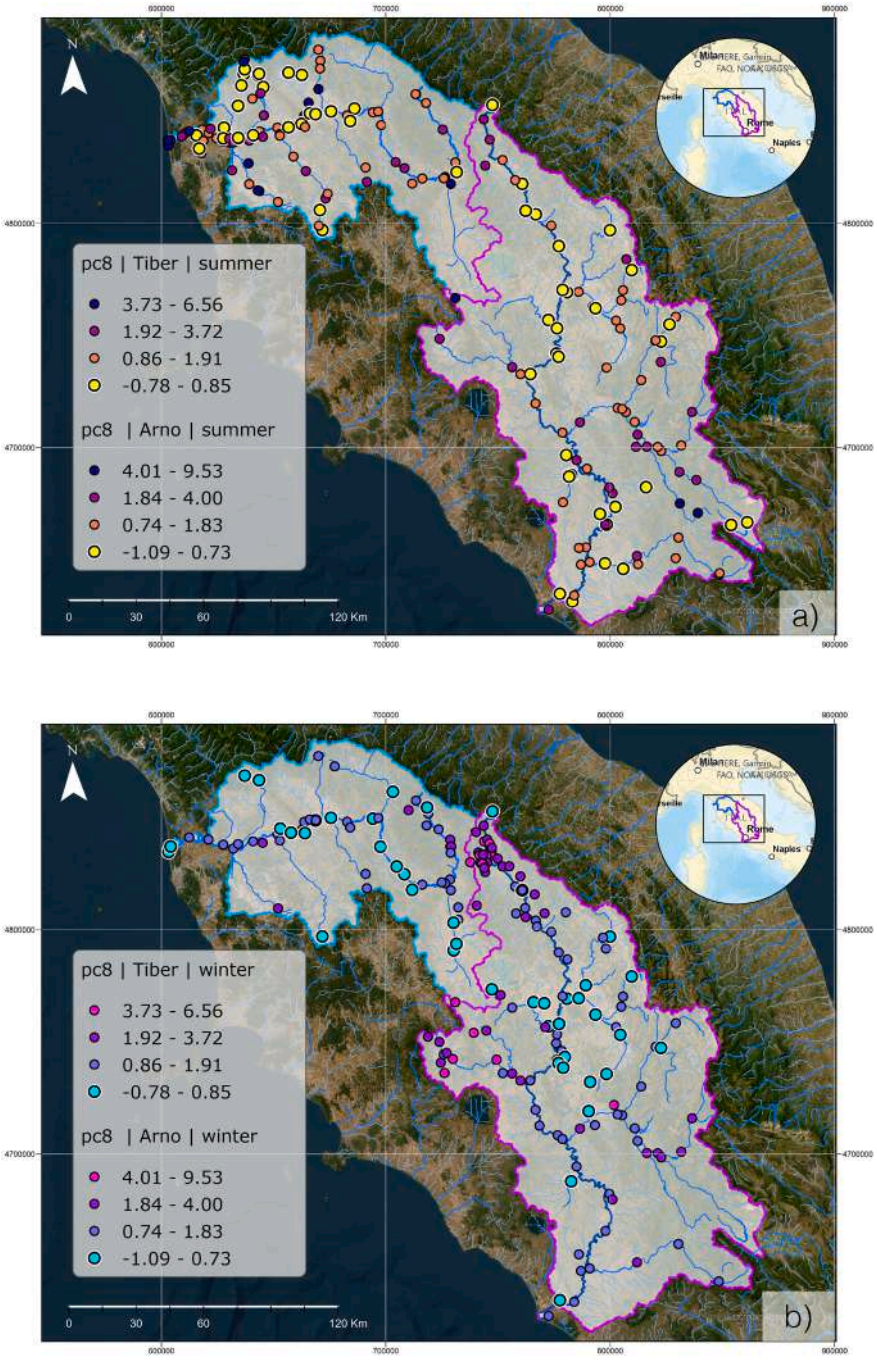

Fig. 6. Maps of *pc*8 balance (Mg^{2+}, SO_4^{2-}, Ca^{2+}, HCO_3^-, F^-, Na^+, Cl^-, K^+ versus NO_3^-) for the Tiber and Arno basins in a) summer and b) winter seasons. Values of *pc*8 represented using graduated symbols and class breaks determined with the natural breaks (Jenks) classification in ArcGIS Pro.

et al., 2021, 2020; Taussi et al., 2022). Similarly, the Arno River basin suffers from anthropic inputs such as direct discharges from urban areas (e.g., Florence), agricultural activities and waste water from industrial settlements (for example, nurseries, tanneries, and paper mills) (Cortecci et al., 2002; Dinelli et al., 2005; Nisi et al., 2008). This lead to higher contents of NO_3^-, especially downstream from Florence in the summer season (Fig. 6a). Conversely, during winter NO_3^--related effects appear more local in both river basins (see cyan points in Fig. 6b), due to dilution processes.

Following Scheffer et al. (2015) and Dakos et al. (2014) a histogram can be mirrored upside down, and the area related to the higher frequency can be transformed into a more or less deep hole. Generally speaking, the deeper the hole, the more difficult it is for a ball to escape, indicating a more stable system, within certain margins. The presence of multimodality or skewness in this context generates meta-stable valleys increasing the possibility of flickering between different states, thus creating instability (Seely and Macklem, 2012). Considering the results for the balance characterized by the lowest variance (i.e., Ca^{2+}/HCO_3^-), we can say that processes affecting water chemistry for the investigate catchment are simpler and mainly affected by random fluctuations around a well-defined barycenter. They are characterized by a scarce occurrence of outliers observations, thus representing the conditions of stability described by Scheffer et al. (2015) and Dakos et al. (2014). It can also be concluded that the positive values of the index could be due to Ca^{2+}'s excess over HCO_3^- owing to the possibility of Ca^{2+} being bound to SO_4^{2-}. This could indicate the presence of evaporite rocks such as

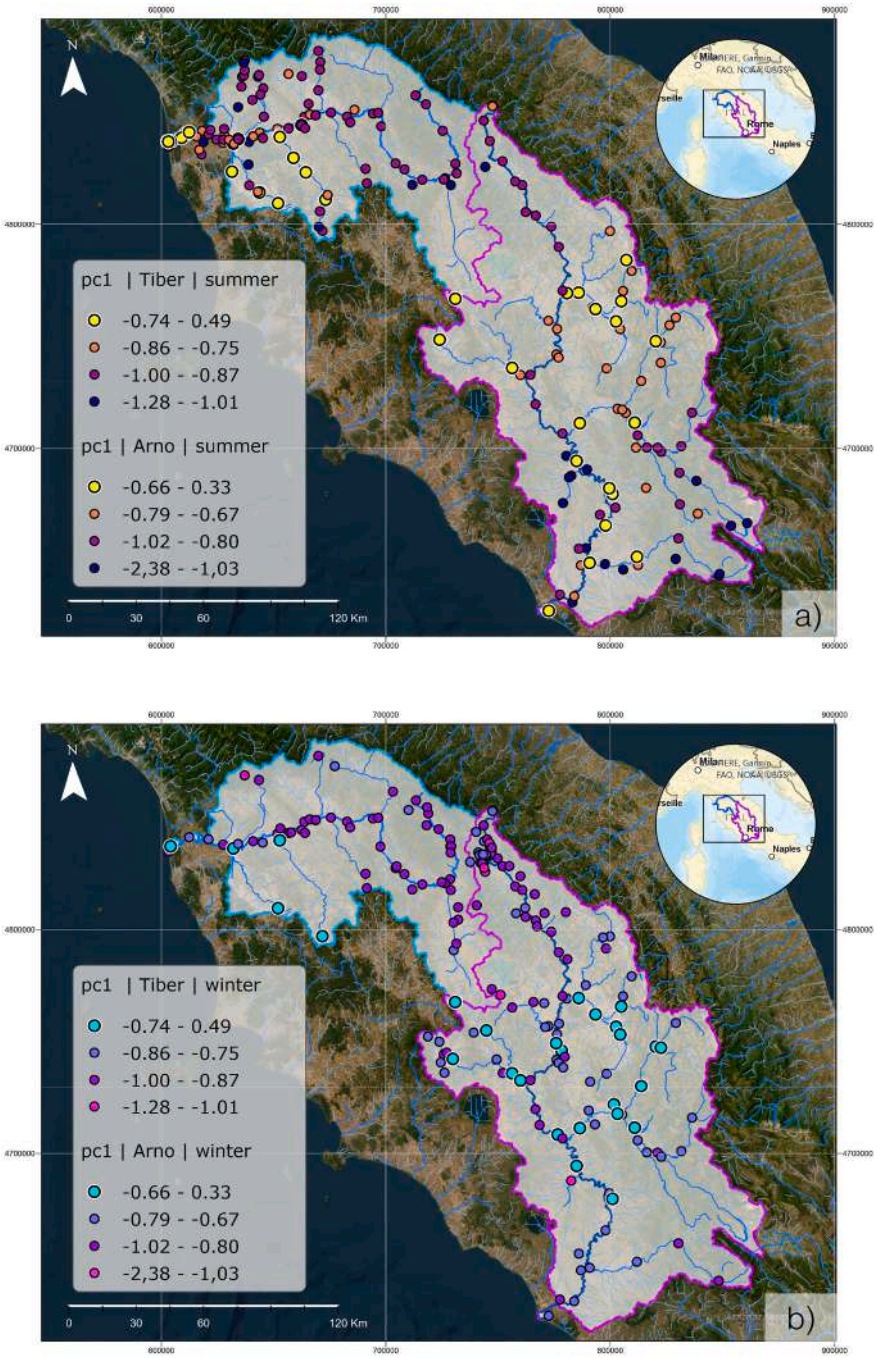

Fig. 7. Maps of *pc*1 balance (Ca^{2+}/HCO_3^-) for the Tiber and Arno basins in summer a) and winter b) seasons. Values of *pc*1 are represented using graduated colors and class breaks determined with the natural breaks (Jenks) classification in ArcGIS Pro.

anhydrite or gypsum within the catchment. This idea is supported by the maps of *pc*1 reported in Figs. 7a-b. For the Tiber basin high values of *pc*1 are reported for the Nera sub-basin most likely related to the effects of groundwater circulation in Triassic evaporites of the Narni-Amelia aquifer system (Dinelli et al., 1999; Frondini et al., 2012; Gozzi et al., 2021). Similarly, in the Arno basin, high values are found for Era and Elsa rivers that drain the recent marine deposits and the scattered outcrops of marine evaporites, both Messinian and Triassic (Dinelli et al., 1999). Additional sources of Ca^{2+} stemming from the weathering of Anorthite and Ca-bearing silicates could potentially contribute to the excess of Ca^{2+}. However, this does not appear to be the case, as elevated values of *pc*1 are predominantly observed in areas where carbonate rocks are prevalent (see Fig. 1). Overall, seasonal variability does not appear to have a substantial effect on *pc*1, confirming the role of this balance for monitoring weathering processes on a global scale and, potentially the effect of climate changes.

5. Conclusions

The objective of this study was to demonstrate the utilization of robust techniques to investigate the origins of geochemical variability and to conduct comparative analyses in different watersheds to find similar laws regulating the behavior of chemical species. This is particularly true for Mediterranean catchments, which are predicted to experience substantial climate change in the 21st century. Using the open-source R software environment, we implement a robust methodology for the analysis of variance of isometric coordinates (balances) obtained by sequential partitioning that successively maximizes the explained variance. Based on hierarchical clustering and robust estimation of the variation matrix, robust orthonormal coordinates are generated. When applying this compositional method to the Tiber and Arno river basins, the results indicate that they have similar balances, with the highest and lowest geochemical variability that are described exactly by the same indices in both basins. On the other hand, in the "middle-earth" situated between the lower and the higher variances interesting results emerge shedding light on the behavior of the variables participating in different water/rock interaction processes, mainly dominated by silicate or carbonate weathering. Furthermore, the frequency distributions of the balances were analyzed as resilience indicators along with their spatial distribution as support for interpretation. When a higher number of variables is considered, the balance between silicate and carbonate weathering processes appear to be counterpoised to anthropic inputs and seasonal variations, thus dominating variability and generating an index able to monitor instability in time or space. Contrary to this, the log-ratio Ca^{2+}/HCO_3^-, which shows the lowest variance, suggests that pure carbonate weathering processes are more stable and mainly influenced by random fluctuations, even though local variability induced by evaporitic outcrops could be detected. This work will be further improved by applying the proposed approach to a broader range of catchments experiencing significantly different geological and morpho-climatic conditions, for a deeper understanding of river geochemistry response to global hydroclimatic change, its variability and repeatability (Gozzi and Buccianti, 2024). The calculations were done using the recently added capabilities of the `robCompositions` R package (Templ et al., 2011) and the existing features of the `compositions` R package (van den Boogaart and Tolosana-Delgado, 2008). The `robustbase` R package (Mächler et al., 2018) is employed to perform robust estimation of covariances.

Funding

The University of Florence is acknowledged for financial assistance through the funds ["Fondi Ateneo 2023 and 2024 (A.B.; C.G.)"]. This work was supported by the National Biodiversity Future Centre (NBFC) and the National Centre for HPC, Data and Quantum Computing to University of Florence, Department of Earth Sciences, funded by the Italian Ministry of University and Research, PNRR, Missione 4 Componente 2, "Dalla ricerca all'impresa", Investimento 1.4, [Projects CN00000033 and CN00000013].

CRediT authorship contribution statement

Caterina Gozzi: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Conceptualization. **Matthias Templ:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Antonella Buccianti:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Conceptualization.

Declaration of competing interest

Co-Guest Editor of the Special Issue "VSI:CoDA 40 years after 1982" of Journal of Geochemical Exploration - C.G.

Data availability

Data will be made available on request.

Acknowledgments

Orlando Vaselli and Barbara Nisi are thanked for their scientific and technical support during the sampling campaigns of the Tiber river basin. The authors are grateful to two anonymous reviewers for their valuable comments on an earlier version of the manuscript.

References

Aitchison, J., 1982. The Statistical Analysis of Compositional Data (with discussion). Journal of the Royal Statistical Society Series B 44 (2), 139–177.
Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd, London.
Allen, P., 2017. Sediment Routing Systems: The Fate of Sediment from Source to Sink. Cambridge University Press, p. 442. https://doi.org/10.1017/9781316135754.
Belle, S., Baudrot, V., Lami, A., Musazzi, S., Dakos, V., 2017. Rising variance and abrupt shifts of subfossil chironomids due to eutrophication in a deep sub-alpine lake. Aquat. Ecol. 51, 307–319.
Berner, E.K., Berner, R.A., 1996. Global Environment: Water. Air and Geochemical cycles, New Jersey, Prentice Hall, p. 376.
Billheimer, D., Guttorp, P., Fagan, W., 2001. Statistical interpretation of species composition. J. Am. Stat. Assoc. 96 (456), 1205–1214. https://doi.org/10.1198/016214501753381850.
Bonini, M., 1998. Chronology of deformation and analogue modelling of the Plio-Pleistocene Tiber Basin: implication for the evolution of the Northern Apennines (Italy). Tectonophysics 285 (1–2), 147–165.
van den Boogaart, K., Tolosana-Delgado, R., 2013. Analyzing Compositional Data with R. Springer, Heidelberg.
van den Boogaart, K.G., Tolosana-Delgado, R., 2008. Compositions: a unified R package to analyze compositional data. Comput. Geosci. 34 (4), 320–338. https://doi.org/10.1016/j.cageo.2006.11.017.
Buccianti, A., Pawlowsky-Glahn, V., 2005. New perspectives on water chemistry and compositional data analysis. Math. Geol. 44(2) (37(7)), 703–727.
Carmignani, L., Decandia, F., Fantozzi, P., Lazzarotto, A., Liotta, D., Meccheri, M., 1994. Tertiary extensional tectonics in tuscany (northern apennines, Italy). Tectonophysics 238 (1), 295–315. https://doi.org/10.1016/0040-1951(94)90061-2 late Orogenic Extension.
Cencetti, C., Tacconi, P., 2005. The fluvial dynamics of the arno river. Giornale di Geologia Applicata 1 (01), 193–202.
Chen, N., Jayaprakash, C., Yu, K., Guttal, V., 2018. Rising variability, not slowing down, as a leading indicator of a stochastically driven abrupt transition in a dryland ecosystem. Am. Nat. 191 (1), E1–E14.
Cortecci, G., Dinelli, E., Bencini, A., Adorni-Braccesi, A., La Ruffa, G., 2002. Natural and anthropogenic so4 sources in the arno river catchment, northern tuscany, Italy: a chemical and isotopic reconnaissance. Appl. Geochem. 17 (2), 79–92.
Dai, Y., Wu, J., Yang, Q., Cheng, S., Liang, W., Hein, T., 2023. The ecosystem services concept in freshwater conservationand restoration. Aquat. Conserv. Mar. Freshwat. Ecosyst. 33 (12) https://doi.org/10.1002/aqc.3913.
Dakos, V., Carpenter, S., van Nes, E., Scheffer, M., 2014. Resilience indicators: prospects and limitations for early warnings of regime shifts. Philos. Trans. R. Soc. B 370 (20130263), 1–10.

Dede, M., Lam, K., Whitaningsih, S., 2023. Relationship between landscape and river ecosystem services. Global Journal of Environmental Science and Management 9 (3), 12. https://doi.org/10.22035/gjesm.2023.03.18.
Dinelli, E., Testa, G., Cortecci, G., Barbieri, M., 1999. Stratigraphic and petrographic constraints to trace element and isotope geochemistry of Messinian sulfates of Tuscany. Mem. Soc. Geol. Ital. 54 (01), 61–74.
Dinelli, E., Cortecci, G., Lucchini, F., Zantedeschi, E., 2005. Sources of major and trace elements in the stream sediments of the arno river catchment (northern Tuscany, Italy). Geochemical journal GJ 39 (01), 531–545. https://doi.org/10.2343/geochemj.39.531.
Dinis, P.A., Garzanti, E., Hahn, A., Vermeesch, P., Cabral-Pinto, M., 2020. Weathering indices as climate proxies. A step forward based on Congo and sw african river muds. Earth Sci. Rev. 201 (02) https://doi.org/10.1016/j.earscirev.2019.103039.
Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geosci. 37 (7), 795–828. https://doi.org/10.1007/s11004-005-73811-9.
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logration transfomations for compositional data analysis. Math. Geol. 35 (3), 270–300.
Filzmoser, P., Hron, K., 2008. Outlier detection for compositional data using robust methods. Math. Geosci. 40 (3), 233–248. https://doi.org/10.1007/s11004-007-9141-5.
Filzmoser, P., Hron, K., Templ, M., 2018. Applied Compositional Data Analysis: With Worked Examples in R. Springer International Publishing, Cham, p. 280. https://doi.org/10.1007/978-3-319-96422-5.
Frondini, F., Cardellini, C., Caliro, S., Chiodini, G., Morgantini, N., 2012. Regional groundwater flow and interactions with deep fluids in western Apennine: the case of Narni-Amelia chain (Central Italy). Geofluids 12, 182–196.
Gagolewski, M., 2021. Genieclust: Fast and robust hierarchical clustering. SoftwareX 15, 100722. https://doi.org/10.1016/j.softx.2021.100722.
Gaillardet, J., Dupré, B., Louvat, P., Allègre, C., 1999. Global silicate weathering and co_2 consumption rates deduced from the chemistry of large rivers. Chem. Geol. 159, 3–30.
Gozzi, C., Buccianti, A., 2022. Assessing indices tracking changes in river geochemistry and implications for monitoring. Nat. Resour. Res. 31 (2) https://doi.org/10.1007/s11053-022-10014-1.
Gozzi, C., Buccianti, A., 2024. Resilience and high compositional variability reflect the complex responsebof river waters to global drivers: the Eastern Siberian River Chemistry database. Sci. Total Environ. 908 (168120), 1–14.
Gozzi, C., Filzmoser, P., Buccianti, A., Vaselli, O., Nisi, B., 2019. Statistical methods for the geochemical characterisation of surface waters: the case study of the Tiber River basin (Central Italy). Comput. Geosci. 131, 80–88.
Gozzi, C., Sauro Graziano, R., Buccianti, A., 2020. Part–Whole Relations: New Insights about the Dynamics of Complex Geochemical Riverine Systems. Minerals 10 (501). https://doi.org/10.3390/min10060501.
Gozzi, C., Dakos, V., Buccianti, A., Vaselli, O., 2021. Are geochemical regime shifts identifiable in river waters? Exploring the compositional dynamics of the Tiber River (Italy). Sci. Total Environ. 785, 147268 https://doi.org/10.1016/j.scitotenv.2021.147268.
Grziwotz, F., Chang, C., Dakos, V., van Nes, E., Schwarzländer, M., Kamps, O., Heßlera, M., Tokuda, I., Telschow, A., Hsieh, C., 2023. Anticipating the occurrence and type of critical transitions. Sci. Adv. 9, 1–12.
ISPRA Ambiente, March 2017. Geoportale Ispra Ambiente. URL. http://geoportale.isprambiente.it/sfoglia-il-catalogo/?lang=en.
Kleidon, A., Schymanski, S., Stieglitz, M., 2009. Thermodynmaics, irreversibility, and optimality in land surface hydrology. Bioclimatology and Natural Hazard 107–118.
Kleidon, A., Zehe, E., Ehret, U., Scherer, U., 2012. Thermodynamics, maximum power, and the dynamics of preferential river flow structures on continents. Hydrol. Earth Syst. Sci. Discuss. 9, 7317–7378. https://doi.org/10.5194/hessd-9-7317-2012.
Kynčlová, P., Hron, K., Filzmoser, P., 2017. Correlation between compositional parts based on symmetric balances. Math. Geosci. 49, 777–796.
Lutz, S., Mallucci, S., Diamantini, E., Majone, B., Bellin, A., Merz, R., 2016. Hydroclimatic and water quality trends across three Mediterranean river basins. Sciences of the Total Environment 571, 1392–1406.
Mächler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E.L.T., Anna di Palma, M., 2018. robustbase: Basic Robust Statistics. URL. `http://robustbase.r-forge.r-project.org/`.
Maronna, R., Martin, R., Yohai, V., 2006. Robust Statistics: Theory and Methods. John Wiley & Sons, New York. ISBN 978-0-470-01092-1.
Martn-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosona-Delgado, R., 2017. Advances in principal balances for compositional data. Math. Geosci. 50, 273–298.
Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J., 2011. The principle fo working in coordinates. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd, Chichester, pp. 34–42. https://doi.org/10.1002/9781119976462.ch3.
McDonnell, J.J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M.L., Selker, J., Weiler, M., 2007. Moving beyond heterogeneity and process complexity: a new vision for watershed hydrology. Water Resour. Res. 43 (7) https://doi.org/10.1029/2006WR005467.
Mitzenmacher, M., 2004. A brief history of generative models for power law and lognormal distributions data analysis. Internet Math. 1 (2), 226–251.
Murtagh, F., Legendre, P., 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J. Classif. 31 (3), 274–295. https://doi.org/10.1007/s00357-014-9161-z.
Nisi, B., Buccianti, A., Vaselli, O., Perini, G., Tassi, F., Minissale, A., Montegrossi, G., 2008. Hydrogeochemistry and strontium isotopes in the Arno River Basin (Tuscany, Italy): Constraints on natural controls by statistical modeling. J. Hydrol. 360 (1–4), 166–183.
Panichi, C., Giuliano, G., Preziosi, E., Gherardi, F., Droghieri, E., 2005. Hydrochemical and isotopic characterisation of the base flow in the Tiber Basin. In: Relations between Surface Waters and Groundwaters, 124. Romana Editrice, p. 113.
Patil, R., Wei, Y., Shelmeister, J., 2018. Understanding hydro-ecological surprises for riverine ecosystem management. Curr. Opin. Environ. Sustain. 33 (08) https://doi.org/10.1016/j.cosust.2018.05.021.
Pawlowsky-Glahn, V., Egozcue, J., 2001. Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment (SERRA) 15 (5), 384–398. https://doi.org/10.1007/s004770100077.
Pawlowsky-Glahn, V., Egozcue, J., 2011. Exploring compositional data with the coda-dendrogram. Austrian Journal of Statistics 406, 103–113.
Prigogine, I., Stengers, I., 1984. Order out of Chaos: Man's New Dialogue with Nature. Bantam New Age Books, London, p. 349.
R Development Core Team, 2023. R: A Language and Environment for Statistical Computing. Version 4.2.0.. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.
Rinaldo, A., Rigon, R., Banavar, J., Maritan, A., Rodriguez-Iturbe, I., 2014. Evolution and selection of river networks: Statics, dynamics, and complexity. Proceedings of the National Academy of Sciences (PNAS) 111 (7), 2417–2424.
Rousseeuw, P., Van Driessen, K., 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics 41 (3), 212–223. https://doi.org/10.1080/00401706.1999.10485670.
Scheffer, M., Carpenter, S.R., Dakos, V., van Nes, E.H., 2015. Generic indicators of ecological resilience: Inferring the chance of a critical transition. Annu. Rev. Ecol. Evol. Syst. 46 (1), 145–167. https://doi.org/10.1146/annurev-ecolsys-112414-054242.
Schlesinger, W., Klein, E., Vengosh, A., 2020. Global biogeochemical cycle of fluorine. Glob. Biogeochem. Cycles 1–18. https://doi.org/10.1029/2020GB006722.
Schmeller, D.S., Urbach, D., Bates, K., Catalan, J., Cogalniceanu, D., Fisher, M.C., Friesen, J., Fuureder, L., Gaubek, V., Haver, M., Jacobsen, D., Roux, G.L., Linm, Y., Loyau, A., Machatei, O., Mayer, A., Palomo, I., Plutzar, C., Sentenac, H., Sommaruga, R., Tiberti, R., Ripple, W., 2022. Scientists' warning of threats to mountains. Sci. Total Environ. 853, 1–12. https://doi.org/10.1016/j.scitotenv.2022.158611.
Seely, A., Macklem, P., 2012. Fractal variability: an emergent property of complex dissipative systems. Chaos 22, 13108-1–013108-7.
Shvartsev, S., 2009. Self-organizing abiogenic dissipative structures in the geologic history of the earth. Earth Science Frontiers 16 (6), 257–275.
Taussi, M., Gozzi, C., Vaselli, O., Cabassi, J., Menichini, M., Doveri, M., Romei, M., Ferretti, A., Gambioli, A., Nisi, B., 2022. Contamination Assessment and Temporal Evolution of Nitrates in the Shallow Aquifer of the Metauro River Plain (Adriatic Sea, Italy) after Remediation Actions. Int. J. Environ. Res. Public Health 19 (19), 12231.
Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: an R-package for Robust Statistical Analysis of Compositional Data, Ch. 25. John Wiley & Sons, Ltd, pp. 341–355. https://doi.org/10.1002/9781119976462.ch25.
Templ, M., Gussenbauer, J., Filzmoser, P., 2019. Evaluation of robust outlier detection methods for zero-inflated complex data. J. Appl. Stat. 0 (0), 1–24. https://doi.org/10.1080/02664763.2019.1671961.
Templ, M., Hron, K., Filzmoser, P., 2021. robCompositions: an R-package for robust statistical analysis of compositional data, Ch. 25. Wiley Online Library, pp. 341–355. https://doi.org/10.1002/9781119976462.ch25.
Todorov, V., Templ, M., Filzmoser, P., 2011. Detection of multivariate outliers in business survey data with incomplete information. ADAC 5 (1), 37–56.
Tolosana-Delgado, R., Mueller, U., van den Boogaart, G., 2019. Geostatistics for compositonal data: an overview. Math. Geosci. 51, 485–526.