Science of the Total Environment

Evaluation of natural background levels of high mountain karst aquifers in complex geological settings. A Gaussian mixture model approach in the Port del Comte (SE, Pyrenees) case study

--Manuscript Draft--

Manuscript Number:	STOTEN-D-20-22426
Article Type:	Research Paper
Keywords:	High-mountain karst system; Alpine hydrologeology; Natural background levels; Compositional data; Model-based clustering; Gaussian mixing model
Corresponding Author:	Jorge Jódar, Ph.D. (a)Geological and Mining Institute of Spain Zaragoza, SPAIN
First Author:	Ignasi Herms, MSc
Order of Authors:	Ignasi Herms, MSc
	Jorge Jódar, Ph.D.
	Albert Soler, Ph.D.
	Luis Javier Lambán, Ph.D.
	Emilio Custodio, Ph.D.
	Joan Agustí Núñez, BSc
	Georgina Arnó, BSc
	Maribel Ortego, Ph.D.
	David Parcerisa, Ph.D.
	Joan Jorge, Ph.D.
Abstract:	The hydrogeological processes driving the hydrochemical composition of groundwater in the alpine pristine aquifer system of the Port del Comte Massif (PCM) are characterized through the multivariate statistical techniques Principal Component Analysis (PCA) and Gaussian Mixture Models (GMM) in the framework of Compositional Data (CoDa) analysis. Also, the groundwater Natural Background Levels (NBLs) for NO3 and SO4 and Cl are evaluated, which are specially important for indicating occurrence of groundwater contamination derived from the anthropic activities conducted in the PCM. The hydrogeochemical facies in the aquifer system of the PCM comprises low mineralized Ca-HCO 3 water for the main Eocene karst aquifer, but also Ca-SO 4 and highly mineralized Na–Cl water types for the minor aquifers discharging from the PCM. The NBL values of SO 4 , Cl and NO 3 obtained for the main karst aquifer are 14.33, 4.06 and 6.55 mg/L, respectively. These values are 35, 3 and 1.2 times lower than the respective official NBLs values that were determined by the water administration to be compared with in the case of conducting a polluting assessment characterization in the main karst aquifer. Official overestimation of NBLs can put important groundwater resources in the PCM at risk.
Suggested Reviewers:	Tibor Yvan Stigter, Ph.D. IHE Delft Institute for Water Education Department of Water Science & Engineering t.stigter@un-ihe.org Specialist in groundwater resources management
	Miguel Angel García-Vera, Ph.D. Head of water resources planning, Head of water resources planning in Confederación Hidrográfica del Ebro mgarciave@chebro.es Specialist in water resources planning from the Water resources Administration point of view

	Miguel Rodríguez-Rodríguez, Ph.D. Researcher, Universidad Pablo Olavide Departamento de Sistemas Físicos Químicos y Naturales: Universidad Pablo Olavide Departamento de Sistemas Fisicos Quimicos y Naturales mrodrod@upo.es Specialist in Karst hydrology
	Josep Mas-Pla, Ph.D. Researcher, Universitat de Girona josep.mas@udg.edu; Specialist in groundwater resources management
	Juan Antonio Barberá, Ph.D. Researcher, Universidad de Malaga jabarbera@uma.es Specialist in Karst hydrology
	Javier Heredia, Ph.D. Researcher, Instituto Geológico y Minero de España: Instituto Geologico y Minero de Espana j.heredia@igme.es Specialist in Karst hydrology
	Caoimhe Hickey, Ph.D. Researcher, Geological Survey of Ireland Caoimhe.Hickey@gsi.ie Specialist in Karst hydrology
	Vera Pawlowsky-Glahn, Ph.D. Researcher, Universitat de Girona vera.pawlowsky@udg.edu Specialist in Compositional data Analysis
	Juan José Egozcue, Ph.D. Reseach Profesor, Universitat Politecnica de Catalunya juan.jose.egozcue@upc.edu Specialist in Compositional data Analysis
	Mark Engle, Ph.D. Researcher, University of Texas maengle@utep.edu Specialist in Compositional data Analysis
Opposed Reviewers:	

To whom concern:

Knowledge on natural background levels of concentrations in groundwater is important to identify contamination of groundwater by anthropogenic activities. This is ever more important in high mountain aquifers where water resources are mostly generated for the downgradient depending areas. This is the case of the Port del Comte karst aquifer system discharges into the Cardener River, the most important tributary of the populated Llobregat River basin that provides drinking water to the city of Barcelona (NE Spain), where the population is larger than 2 million people, and where more than 70% of the water supply derives from surface waters. These waters also feed one of the major groundwater reservoirs for the Barcelona area, the aquifers of the Llobregat Delta (Otero et al., 2008).

High mountain aquifers are very sensitive to polluting activities. Management and preservation of these systems requires a well defined hydrochemical picture to compare with and warning in case of aquifer polluting threats affecting the water resources depending systems.

Processes such as aquifer recharge and runoff generation which force and describe the hydrological systems response also belong to the main knowledge areas considered by Science of the Total Environment, by including hydrosphere, biosphere, lithosphere, and anthroposphere.

Barcelona, 12/09/2020

The authors: Ignasi Herms. Jorge Jódar, Albert Soler, Luis Javier Lambán, Emilio Custodio, Joan Agustí Núñez, Georgina Arnó, Maribel Ortego, David Parcerisa and Joan Jorge

Evaluation of natural background levels of high mountain karst aquifers in complex geological settings. A Gaussian mixture model approach in the Port del Comte (SE, Pyrenees) case study

Herms, I.[a], Jódar, J.[b*], Soler, A.[c], Lambán, L.J.[b], Custodio, E.[d], Núñez, J.A.[a], Arnó, G.[a], Ortego, M.I.[e], Parcerisa, D.[f], Jorge, J.[f]

(a) Àrea de Recursos Geològics. Institut Cartogràfic i Geològic de Catalunya (ICGC), Barcelona, Spain

(b) Instituto Geológico Minero de España (IGME), Zaragoza, Spain

(c) Grup MAiMA, SGR Mineralogia Aplicada, Geoquímica i Geomicrobiologia, Departament de Mineralogia, Petrologia i Geologia Aplicada, Facultat de Ciències de la Terra, Universitat de Barcelona (UB), Barcelona, Spain

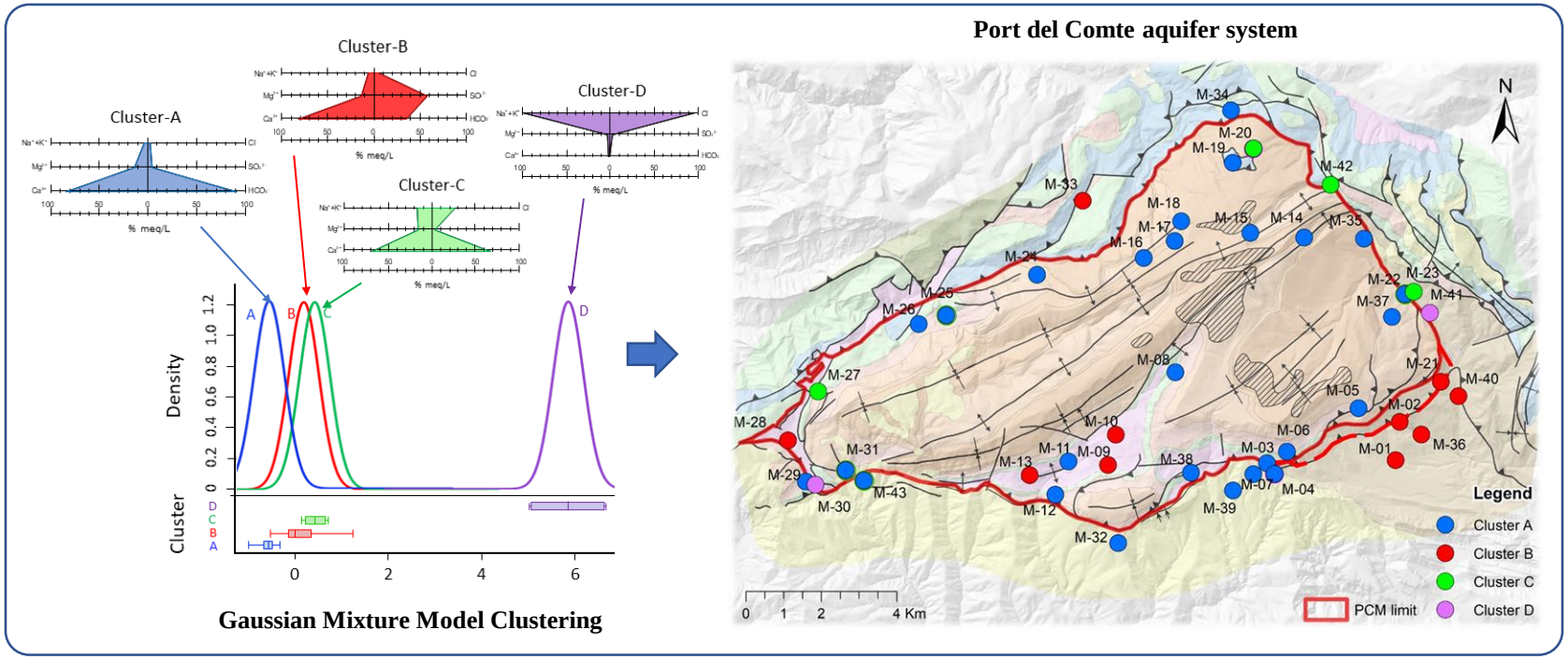
(d) Real Academia de Ciencias. Grupo de Hidrología Subterránea, Dept. de Ingeniería Civil y Ambiental, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

(e) Compositional and Spatial Data Analysis (COSDA) Research Group. Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya BarcelonaTech, Spain

(f) Departament d'Enginyeria Minera, Industrial i TIC. Universitat Politècnica de Catalunya (UPC), Manresa, Spain

* Corresponding author: j.jodar@igme.es

Graphical Abstract



Highlights (for review : 3 to 5 bullet points (maximum 85 characters including spaces per bullet point)

- The Gaussian Mixture Models are used to cluster springs by hydrochemical response
- The clustering analysis groups the springs of the aquifer system into four families
- The compositional analysis approach is required for hydrochemical data analysis
- The hydrogeological characterization of Port del Compte Aquifer system is presented
- Natural base levels of nitrate and sulphate in the Port del Compte Aquifers are given

Evaluation of natural background levels of high mountain karst aquifers in complex geological settings. A Gaussian mixture model approach in the Port del Comte (SE, Pyrenees) case study

Herms, I.[a], Jódar, J.[b*], Soler, A.[c], Lambán, L.J.[b], Custodio, E.[d], Núñez, J.A.[a], Arnó, G.[a], Ortego, M.I.[e], Parcerisa, D.[f], Jorge, J.[f]

(a) Àrea de Recursos Geològics. Institut Cartogràfic i Geològic de Catalunya (ICGC), Barcelona, Spain
(b) Instituto Geológico Minero de España (IGME), Zaragoza, Spain
(c) Grup MAiMA, SGR Mineralogia Aplicada, Geoquímica i Geomicrobiologia, Departament de Mineralogia, Petrologia i Geologia Aplicada, Facultat de Ciències de la Terra, Universitat de Barcelona (UB), Barcelona, Spain
(d) Real Academia de Ciencias. Grupo de Hidrología Subterránea, Dept. de Ingeniería Civil y Ambiental, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
(e) Compositional and Spatial Data Analysis (COSDA) Research Group. Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya BarcelonaTech, Spain
(f) Departament d'Enginyeria Minera, Industrial i TIC. Universitat Politècnica de Catalunya (UPC), Manresa, Spain
* Corresponding author: j.jodar@igme.es (J.Jódar)

Abstract

The hydrogeological processes driving the hydrochemical composition of groundwater in the alpine pristine aquifer system of the Port del Comte Massif (PCM) are characterized through the multivariate statistical techniques Principal Component Analysis (PCA) and Gaussian Mixture Models (GMM) in the framework of Compositional Data (CoDa) analysis. Also, the groundwater Natural Background Levels (NBLs) for NO3 and SO4 and Cl are evaluated, which are specially important for indicating occurrence of groundwater contamination derived from the anthropic activities conducted in the PCM.

The hydrogeochemical facies in the aquifer system of the PCM comprises low mineralized Ca-HCO_3 water for the main Eocene karst aquifer, but also Ca-SO_4 and highly mineralized Na–Cl water types for the minor aquifers discharging from the PCM. The NBL values of SO_4, Cl and NO_3 obtained for the main karst aquifer are 14.33, 4.06 and 6.55 mg/L, respectively. These values are 35, 3 and 1.2 times lower than the respective official NBLs values that were determined by the water administration to be compared with in the case of conducting a polluting assessment characterization in the main karst aquifer. Official overestimation of NBLs can put important groundwater resources in the PCM at risk.

Keywords: High-mountain karst system; Natural background levels; Compositional data; Model-based clustering; Gaussian mixing model.

1. Introduction

High mountain zones produce globally essential water resources that feed with fresh water to the lowland depending ecosystems and a large portion of the world's population (Viviroli et al., 2020). Mountain aquifers, specially those developed in karstifiable carbonate rocks, storage the infiltrated precipitation, thus maintaining important groundwater resources. These resources are typically released as groundwater through large springs that regulates the hydro-ecological regime of the downstream rivers (Kresic and Stevanović, 2010), and provide water resources during the dry season conspicuously in semi-arid regions, where they are often the primary source of drinking water (Stevanović, 2019).

Karst aquifers are much more vulnerable to pollution than other aquifers. Contaminants may easily enter the subsurface into the karst system and rapidly spread in the conduit system without any substantial attenuation (Marín and Andreo, 2015), threatening the water resources of a region at large scale. These aquifers need special protection (Drew and Hötzl 1999, Zwahlen, 2004). In this line, the European Union enacted the Water Framework Directive (2000/60/EC) (WFD, 2000) as an integrated approach to water management to minimize the likelihood of pollution to enter the aquifers. The WFD (2000) defines the rules that apply to the identification of the different groundwater bodies

(GWB), but also the criteria for chemical status assessment through defining pollutants threshold values (TVs) and groundwater natural background values (NBLs). The former are quality standards for pollutants in groundwater representative of those groundwater bodies considered to be at risk. The latter provides the information regarding the concentration of a given element, species or chemical substance present in solution which is derived by natural processes from geological, chemical, biological and atmospheric sources (Müller et al., 2006). In other words, NBLs are the corner stone to quantitatively evaluate whether groundwater is significantly affected or modified by anthropogenic influences (Nieto et al., 2005; Custodio et al., 2007).

It is not easy to define NBLs in high mountain karst aquifer systems (HMKS). For a given aquifer and a certain element, the corresponding NBL value is obtained by averaging the dissolved content of that element in groundwater discharge for the different springs draining the aquifer. HMKS are usually embedded in complex geological structures as a result of tectonic processes (e.g. faults, fold-and-thrust belts, wedge pinch out layers). This often causes a strong compartmentalization (Ballesteros et al, 2014) that may involve different lithologies (i.e. from carbonates to evaporites), thus generating a complex aquifer system. The geological variability of such aquifer system influences the hydrogeochemical signature of groundwater along the different flowlines, which typically converge while mixing around springs. As a result, a different hydrochemical composition than the expected may be obtained in the discharge of a spring given its geological setting (Lambán et al., 2015), thus complicating a consistent NBLs characterization for the different aquifers conforming the hydrogeological system.

To correctly define NBLs in HMKS it is necessary to have both a good hydrogeological characterization of the aquifers at the local scale, and a good characterization of the relevant hydrogeochemical finger prints describing the whole picture of the aquifer system. In the framework of compositional data analysis (CoDA) (Aitchison, 1986), in which the concentration of a given element dissolved in water is actually expressing a part of a whole, regardless of the dimensions in which the component concentration is expressed, either as weight per cent ratio [-] (e.g., %, mg/kg), or given as element mass per unit of dissolution volume [ML^{-3}] (e.g., mg/L) (Reimann et al., 2012; Buccianti and Grunsky, 2014; Filzmoser et al., 2018; Egozcue and Pawlowsky-Glahn, 2005, 2006; Pawlowsky-Glahn, et al. 2015). For CoDa multivariate statistical analysis (MSA)

techniques/tools have shown a proven track record in characterizing complex hydrogeological systems through the analysis of spatial variations in hydrochemical data. The combination of MSA tools (e.g. principal component analysis and clustering analysis) allow to investigate the factors controlling the processes taking place in aquifers driving the hydrogeochemical content of groundwater (Otero et al., 2005; Puig et al., 2011; Blake et al., 2016; Owen et al., 2016; Piña et al., 2018; Shelton et al., 2018). Clustering analysis (CA) methods have been largely used to separate groundwater samples, especially for large and/or complicated datasets, into homogeneous groups to reflect different source contributions to groundwater in the sampled springs (see Suk and Lee, 1999; Cloutier et al., 2008; Yidana, 2010; Kim et al., 2014; Yolcubal et al 2019, among others). This faculty makes CA methods a promising tool to correctly define NBLs in HMKS.

There are two mainstreams in CA, (1) the "hard clustering" methods like hierarchical clustering and partitioning methods (k-means, k-medoids: PAM, CLARA) where each data point (i.e. the sample) is assigned to one and only one cluster (hard assignment) and (2) the "soft clustering" methods like model-based clustering (e.g. the Gaussian Mixture Models – GMM) and fuzzy clustering where instead of assigning each data point into a unique specific cluster, it is assigned to all the clusters with different probabilities or weights (soft assignment) (Güler and Thyne 2004).

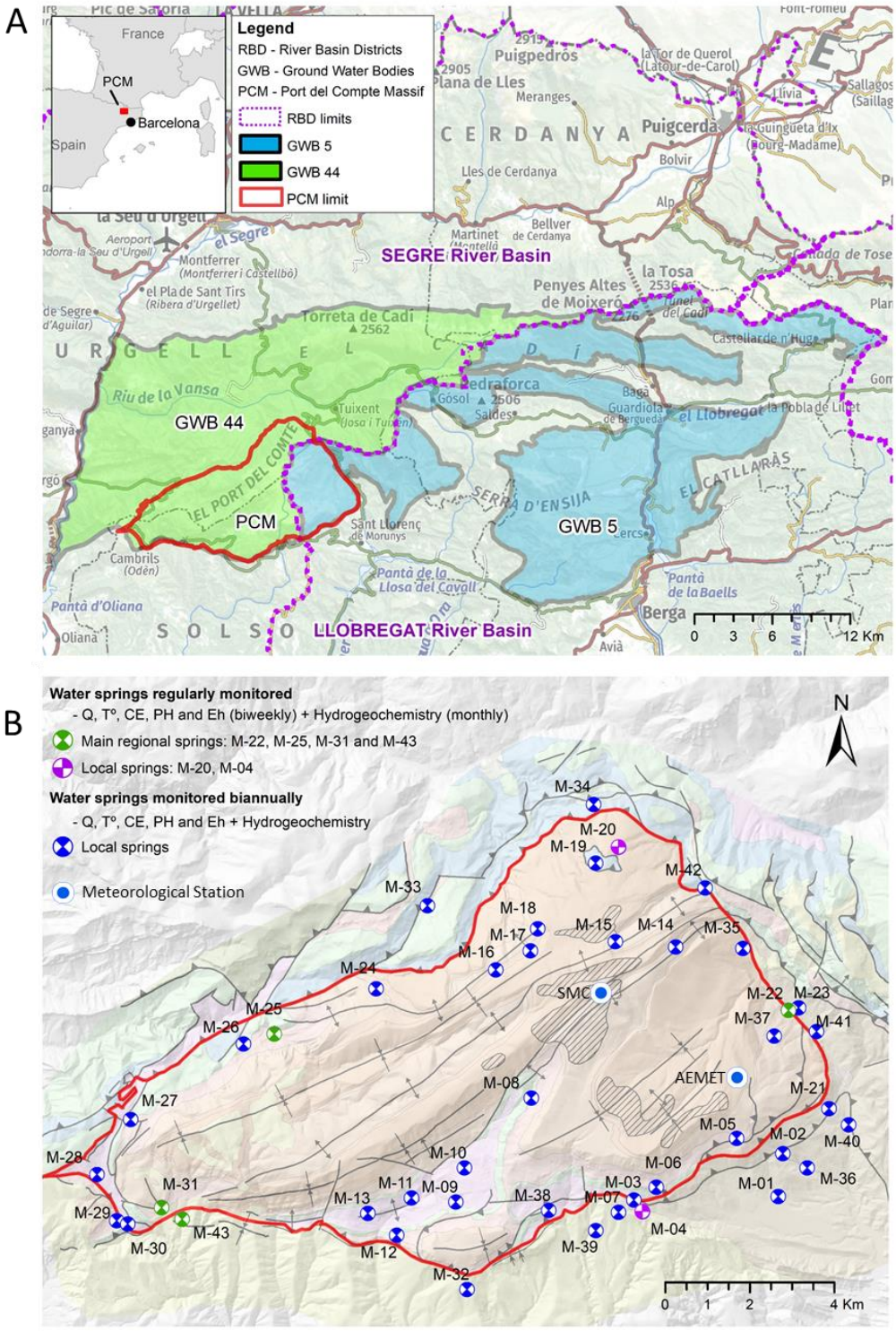
Soft clustering methods are getting more popular since they provide degrees of membership at different hydrogeochemical clusters, rather than clear-cut distinctions. As a result, they can better reflect the spatial continuity of a hydrological system while providing a more rigorous framework to validate the clustering results (Kim et al., 2014, 2015; Wu, et al., 2017, Bondu et al., 2020). Moreover, in the framework of HMKA where the limited number of observations often is a challenge, GMM clustering algorithms are shown able to provide valuable insights into hydrochemical processes, delineating the different groundwater sources imprinting the hydrochemical signature of the aquifer system despite a sparse hydrochemical dataset (Wu et al., 2017). GMM are specially well suited to provide a solid basement for NBLs determination in HMKS. Altough GMM have been used for some authors to evaluate NBLs (Kim, et al. 2015), surprisingly, there are no references in the scientific literature that use GMM in the framework of CoDA to evaluate NBLs in HMKS.

This work aims at filling this gap. To that end, we characterize the hydrochemical composition of the different aquifers associated to the alpine karst aquifer system of the Port del Comte Massif (PCM) to evaluate in a consistent way the NBLs for he different aquifers integrated in this HMKS. This is conducted through a MSA approach that combines in a CoDA framework both PCA and GMM clustering analysis.

2. Study area

The PCM is located in the South-Central Catalan Pyrenees (north-east of Spain), which constitute an orogenic system that runs along the boundary between the Iberian and European plates. It was formed from the Late Cretaceous to Miocene times (Muñoz, et al 2018). The elevation of the mountainous massif ranges from 900 m a.s.l. to 2390 m a.s.l. The massif constitutes an independent structural and hydrogeological system with a surface area of 110 km^2. The highest peaks of the massif conform a water divide between the upper Segre River basin to the NW and SW and the upper Cardener River basin (a tributary of the LLobregat River) to the SE (Fig. 1).

According to the Köppen-Geiger classification (Peel et al., 2007), the study area is characterized by a cold climate without a dry season and with a temperate summer. For the period 2005-2019, the average annual precipitation (P), temperature (T) and potential evapotranspiration (Hargreaves' method) at the SMC meteorological station located at 2315m a.s.l. (Fig. 1) are 1055 mm, 3.2° C and 525 mm, respectively. At elevations > 1800 m a.s.l. the snow covers the massif from December to March.



158

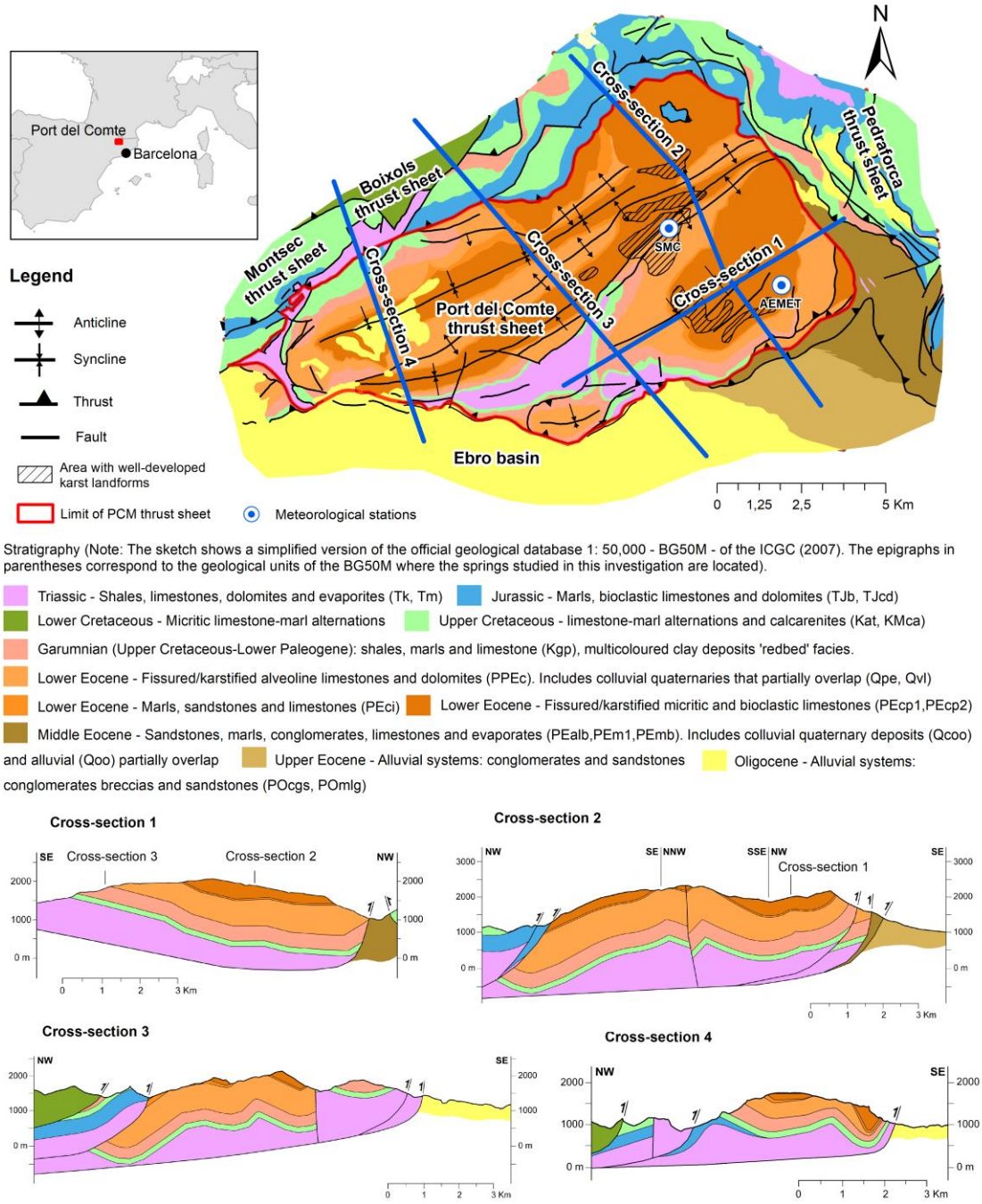
159

160 **Fig.1.** (A) Location map of the study area. (A). Delimintation of the groundwater bodies
161 affecting the PCM; GWB-44 belongs to the Segre river basin, and GWB-5 belongs to the
162 Llobregat river basin. (B) Location of the 43 monitored springs in the PCM

163

Geologically, the PCM constitutes an independent thrust sheet which presents complex structural shapes in its boundaries (Fig. 2), with different thrust sheets individualizing the whole domain in one independent structural system. The internal structure of the PCM is formed by a set of folds and thrusts. These folds have a constant direction NE-SW parallel to the NW limit (Vergés, 1999). The stratigraphic series contains limestones and evaporites mainly from the Triassic, Cretaceous limestones, calcarenites, and shales and Paleogene: Eocene-Oligocene limestones, sandstones and marls. The Jurassic marls, limestones and dolomites only outcrops in the NW part of the sheet. The limestones present a total thickness greater than 1300 m. From the geomorphological perspective, the PCM presents a rounded-soft landscape in the highest domains with no vegetation cover and almost no soil horizon development. The rest of the massif is covered by mountain meadows and forest with a shallow soil depth up to medium development ground cover. Many different karst forms appear progressively from 1950 m.a.s.l. upwards, being well developed at 2050m a.s.l. (see Fig 2, indicated as 'Area with well-developed karst landforms') with sinkholes, dolines and karren fields. They underline the heterogeneity of the karst system.

From the hydrogeological point of view, the PCM can be considered an independent unit and a multi-aquifer system. The main aquifer is formed by Lower Eocene – fissured and karstified limestones and dolomites and constitute one of the most important karst aquifers of the Catalan Pyrenees. The other existing aquifers and aquitards in the system are related to the levels of Cretaceous limestones, Triassic limestone and evaporites, other Paleogene conglomerates and sandstones, and also to small quaternary aquifers (draining small areas), which can be recharged locally at low or medium elevations. The lower Upper Cretaceous/Paleocene (Garumnian facies) substrate materials composed by siltstone and shales constitute an impervious layer for the overlaying Lower Eocene karst aquifer. The complex geologic structure of the system strongly influences the location of the existing karst springs, their groundwater geochemistry and their long-term hydrologic behaviour.



194

195 **Fig. 2.** Geological map and geological cross-sections of the PCM (modified from ICGC,
196 2007)

197 The hydrogeological conceptual model of the PCM aquifer system, as presented by
198 Herms, et al. (2019), considers that recharge is produced by infiltration of precipitation
199 as rainfall and snowmelt and it occurs both in a concentrated way through the local karst
200 conductive features, mostly situated at the top of the massif, and diffuse through the whole

domain. The infiltrated water percolates through the thick unsaturated zone (with more than 1000 m at the top of the massif) towards the saturated zone, and discharges through a large number of springs.

More that 100 springs were inventoried in the study zone. Nevertheless, only 43 of them discharge throughout the year (Fig. 1). These springs were monitored during the period September 2013 – October 2015. Most of them discharge small-scale local sub-surface water flows with flow rates ranging between 0.1 L/s to 1 L/s. Nevertheless, there are four 'regional' springs (M-22, M-25, M-31 and M-43) with flow rates comprised between 1 L/s and 900 L/s during the monitored period. These regional springs are recharged at medium to high elevations, and drain the system discharging through the limestones outcrops (M-31), quaternary deposits overlying the limestones (M-25, M-22), and also through well-developed karst conduits in the conglomeratic materials of the Ebro Basin (M-43). These conglomerates conform the southern foreland basin of the Pyrenees, which is located just at the southern border of the PCM. There is also a diffuse groundwater discharge through the 'Riu Fred' sub-basin to the North. With the exception of two singular groundwater wells on the SW and E edges of the PCM, there are no other water wells within the perimeter of the PCM that exploit the main karst aquifer. It is estimated that the regional water table of the karst system is between 1000 and 1100 m a.s.l. (Herms et al., 2019).

Although the whole PCM massif belongs to the same geomorphological structure, the SE sector has been assigned to GWB-5 ('Conca Alta del Cardener i Llobregat'), whereas the rest of the PCM was assigned to GWB-44 ('Cadí Port del Comte'). Table 1 summarizes the natural background levels at the 90th percentile values (NBL90) determined through the Pre-selection (PS) method described by the EU research project "BRIDGE" (2007) (Müller et al., 2006) using different control points for each GWB. It is worth noting the high values determined for SO_4 contents in both GWBs. Considering an aquifer of pristine waters related to the main Eocene karst aquifer of PCM composed only of limestone and dolomites, these values are surprisingly high.

Table 1. NBL90 values for Cl, NO_3 and SO_4 in the GWB-5 and GWB-44.

	NBL90		
	Cl [mg/L]	SO_4 [mg/L]	NO_3 [mg/L]

GWB-5	12 [a]	485 [a]	-
GWB-44	36[b]	609[b]	8[b]

(a) Data source: Agència Catalana de l'Aigua

(b) Data source: Confederación Hidrográfica del Ebro

3. Materials and methods

3.1. Sampling and analysis

In this work 43 springs were sampled twice per year (i.e. before snowfall and after snowmelt seasons) between September 2013 and October 2015. Nevertheless, in six of them (M-04, M-20, M-22, M-25, M-31 and M-43; Fig. 1) the groundwater sampling frequency was higher, every three to four weeks, to study hydrogeochemical evolution of the groundwater discharge. The springs M-22, M-25, M-31 and M-43 correspond to regional discharge points of the karst system, whereas M-04 and M-20 are considered representative of local small aquifers of the area (Herms et al., 2019).

A total of 288 groundwater samples were collected. Additionally, 10 snow samples (7 from natural snow and 3 from artificial snow produced in the existing sky resort in the NE zone of the PCM) and 2 water samples from water ponds destinated to artificial snow production were collected. In all cases, the in situ physico-chemical parameters Temperature (T), electrical conductivity (EC), pH, Eh and the total dissolved solid (TDS) were measured. The geochemical analysis considered major cations and anions.

All samples were filtered using a 0.45µm membrane filter and stored in new 200-500 mL polyethylene bottles washed with diluted nitric acid and rinsed with samples prior to sampling. Samples for cation analysis were acidified with ultrapure HNO_3, to pH<2 to prevent precipitation. Samples for anion analysis were not acidified. All water samples were preserved at 4 ºC before laboratory measurement. T, CE, pH, Eh and TDS were measured by a portable Hanna meter (Multiparameter Water Quality Meter HI9829). Total alkalinity was also determined in-situ by using the titration method in the first four campaigns as well as with an Alkalinity Test Checker de (HI755) of Hanna Instruments. The major cations (Ca, Mg, Na, K, NH_4) and anions (Cl, NO_3, HCO_3,CO_3, SO_4, and F) were determined in the Laboratori Ambiental d'Aigües de Terrassa: the cations were

analysed by inductively coupled plasma atomic emission spectrometry – ICP-OES (Agilent 5100 DV), and the anions by ion chromatography (Dionex, DX-120). Ionic balance errors were calculated using the USGS software PHREEQC (Parkhurst and Appelo, 2013) within the version PhreeqC Interactive (version 3.3.3 10424), and with the phreeqc.dat database, except for the most salinized natural waters (M-30 and M-41) related to deep flow through Keuper evaporates. The majority of samples had ionic balance errors below the recommended standard of ±5% (Appelo and Postma, 2005).

3.2 Data transformation using the CoDa approach

Geochemical datasets are mostly composed by compositional variables . That is, multivariate variables where the individual parts are parts of a whole (Buccianti and Grunsky, 2014). Classical examples refer to constant sum variables, but recent definitions of compositional data include all types of data representing parts of some whole. Ignoring the compositional character of these geochemical variables may lead to misleading results (Pawlowsky-Glahn et al., 2015). In this context, the CoDa methodology is used in this work. In order to avoid the problems derived from the compositional data character, three transformations, all based on log-ratios has been historically proposed, named as: additive log-ratio (alr) transformation and centered log-ratio (clr) transformation (Aitchison, 1986) and isometric log-ratio (ilr) transformation (Egozcue et al., 2003).

In this study, the hydrochemical dataset was transformed using, firstly clr and secondly ilr. The former transformation is described by

$$\mathrm{clr}(\bar{x}_i) = ln\left(\frac{x_i}{g(x_i)}\right);\ i = 1 \div D\ , \qquad (1)$$

where $g(x) = \sqrt[D]{\prod_{i=1}^{D} x_i}$ is the geometric mean of all the considered parts (ions), and D is the matrix dimension.

The ilr transformation allows to express hydrochemical compositions with respect to an orthonormal basis. Their coordinates, called balances, may be easily obtained using a Sequential binary partition (SBP) (Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2005, 2006; Pawlowsky-Glahn et al. 2015). The SBP has been widely used for many

293 authors on water chemistry studies (Engle and Rowan, 2013; Owen et al., 2016; Hee Kim
294 et al., 2019; Bondu et al., 2020). For a D column matrix, i.e. a D-part composition, D-1
295 balances are calculated from the SBP as

$$ilr(\bar{x}_i) = \sqrt{\frac{r_{i_+} \cdot r_{i_-}}{r_{i_+} + r_{i_-}}} \ln \frac{g(c_{i_+})}{g(c_{i_-})}; \;\; i = 1 \div D - 1 \,, \qquad (2)$$

296

297 where c_{i_+} and c_{i_-} the groups of parts separated in the i^{th} step of the SBP; r_{i_+} and r_{i_-} are the
298 numbers of parts included in c_{i_+} and c_{i_-}, respectively.

299 There are different software tools that allow to perform these transformations. The called
300 CoDaPack v.2.0. program (Comas-Cufi and Thió-Henestrosa, 2011) is a software
301 developed by the Research Group in Statistics and Compositional Data Analysis at
302 University of Girona (UdG). This software can be freely downloaded from
303 http://ima.udg.edu/codapack. It allows performing the log-ratio transformations and to
304 prepare different kind of plots to show the results. In this research, all statistical analyses
305 were done using the statistics program R version 3.6.1 (2019-07-05) (R Development
306 Core Team 2004), which is available for free under the GNU-public License and for all
307 platforms from http://www.cran.r-project.org, through the software RStudio, a graphical
308 user interface for R . For multivariate statistical analysis (MSA) using the CoDa approach,
309 the following packages for R software were used: {stats} version 3.6.1. (R-core R-
310 core@R-project.org); {compositions} version 1.40-5 (Van den Boogaart and Tolosana-
311 Delgado, 2008) {zCompositions} version 1.3.4 (Palarea-Albaladejo and Martín-
312 Fernández, 2015).

313 Water samples with solute dissolved concentrations lower than the detection limit (the
314 so-called 'left-censored values') put an extra challenge when addressing MSA
315 techniques. The censored data can be either removed, or replaced or imputed (e.g. values
316 below detection limit are rounded as zeros) (Carranza, 2011). Following the criteria used
317 for several authors (Reimann and Filzmoser, 2000; Farnham et al., 2002), in this work,
318 left-censored values were excluded from the MSA when they represented > 25% of the
319 total number of samples (i.e. when the variable had a 'medium–high' level of nondetects
320 according to Palarea-Albaladejo and Martín-Fernández, 2014). Different algorithms can
321 be applied within the {zCompositions} package for R for imputing these values (like
322 multRepl, multLN, lrEM and lrDA methods).

3.3 Principal Component Analysis (PCA) and Model-based clustering

The first step to apply any MSA, is to check the presence of left-censored data and the imputation of values. The function 'zPatterns' {zCompostions} is used to find and display patterns of zeros/missing values in the whole dataset (see pattern diagrams at Fig.SM.2.1 of Supp. Mat.). In this work, the left-censored detected values were imputed using the 'lrDA' (log ratio Data Argumentation) function. It is based on the log ratio Markov Chain Monte Carlo Data Argumentation algorithm (Palarea-Albaladejo and Martín-Fernández, 2015), and it has been already used by different authors to delineate water types (e.g. Owen et al 2016; Hee Kim et al., 2019). Following the commented procedure two data matrices were prepared:

- Dataset Matrix (300x8), corresponding to 300 water samples (288 groundwater samples and 12 snow and water ponds samples) and 8 variables (HCO_3, Cl, SO_4, NO_3, Ca, Mg, Na, K).

- Dataset Matrix (43x8), corresponding to 43 springs and 8 variables (HCO_3, Cl, SO_4, NO_3, Ca, Mg, Na, K), which represent the median hydrochemical composition of groundwater evaluated for each spring (Table SM.1. Supp. Mat.) The consideration of "median composition" of time series follows the requirements to estimate NBL's using the PS method (see section 3.4).

Table SM.3.1 (Supp. Mat.) shows the list of parameters 'included' and 'excluded' for the MSA and their justification.

PCA is a very common method that is based on dimensionality reduction of datasets. It helps deciphering hydrogeochemical patterns and to infer the controlling variables of the water chemistry (Merchán et al., 2015; Moya et al., 2015). In order to perform the PCA it is necessary to calculate the '*variation matrix*' of the dataset (Aitchison, 1986) as a first step to obtain a measure of the dependence of the different variables, that is, the parts of the composition. Each component of the variation matrix, τ_{ij}, describes the log-relationship between two of the composition x_i and x_j (in this case chemical species), and it's defined as

$$\tau_{ij} = \mathrm{var}\left(\ln\frac{x_i}{x_j}\right) = \frac{1}{N-1}\sum_{n=1}^{N}\ln^2\left(\frac{x_{ni}}{x_{nj}}\right) - \ln^2\left(\frac{g_i}{g_j}\right), \quad (3)$$

where N is the number of observations and g_i, g_j are the geometric mean values for the two variables in question. A small value of τ_{ij} (which is equivalent to τ_{ji}) implies a good proportionality between the two variables. The variation matrix, τ_{ij}, is obtained using the R function 'summary.acomp' of the package {compositions}.

Once the variation matrix is obtained, then the correlation between the variables x_i and x_j is estimated through the '*index of proportionality*' function, ρ_{ij} (Eq. 4) (Aitchison, 1986). The stronger the correlation between x_i and x_j the closer to 1 is the value of ρ_{ij}.

$$\rho_{ij} = \exp\left(\frac{-\tau_{ij}^2}{2}\right) \quad (4)$$

Data transformation following the CoDa approach is applied before using any MSA tool. In this case, PCA is applied using clr-transformed data (Eq. 1). with the function 'clr' of the {compositions} R package. The method provides a new matrix of standardized coordinates for each sample called 'the scores', and also a new matrix of variable 'loadings' with columns representing the principal components of the (clr-transformed) data.

The graphical representations of the PCA results of clr-transformed data were done using the well-known biplot graphic (Gabriel, 1971) (Fig. 3), where the individuals are expressed as dots and the variables as rays. However, the interpretation of the clr-biplot differs from the interpretation of the classical biplot. Therefore, the clr-biplot interpretation is conducted by following the criteria proposed by Aitchison and Greenacre (2002), which is well suited for analyzing compositional data (Otero et al., 2005; Engle and Rowan, 2013; Blake et al., 2016; Piña et al., 2018). The criteria can be summarized as:

- The length of a link (i.e . black shaded line) between the rays (red arrows) defining clr(x_i) and clr(x_j) is proportional to the variance of ln(x_i/x_j).

- If two rays lay near each other, their quotient might be almost constant, and they might be proportional.
- If two links between four different clr-variables are orthogonal, then the corresponding pairwise quotients may be independent.
- If three or more vectors lie on the same link, the corresponding sub-composition might have one single degree of freedom.
- If two links between four separate clr-variables are orthogonal then the corresponding pairs of variables may vary independently of each other.

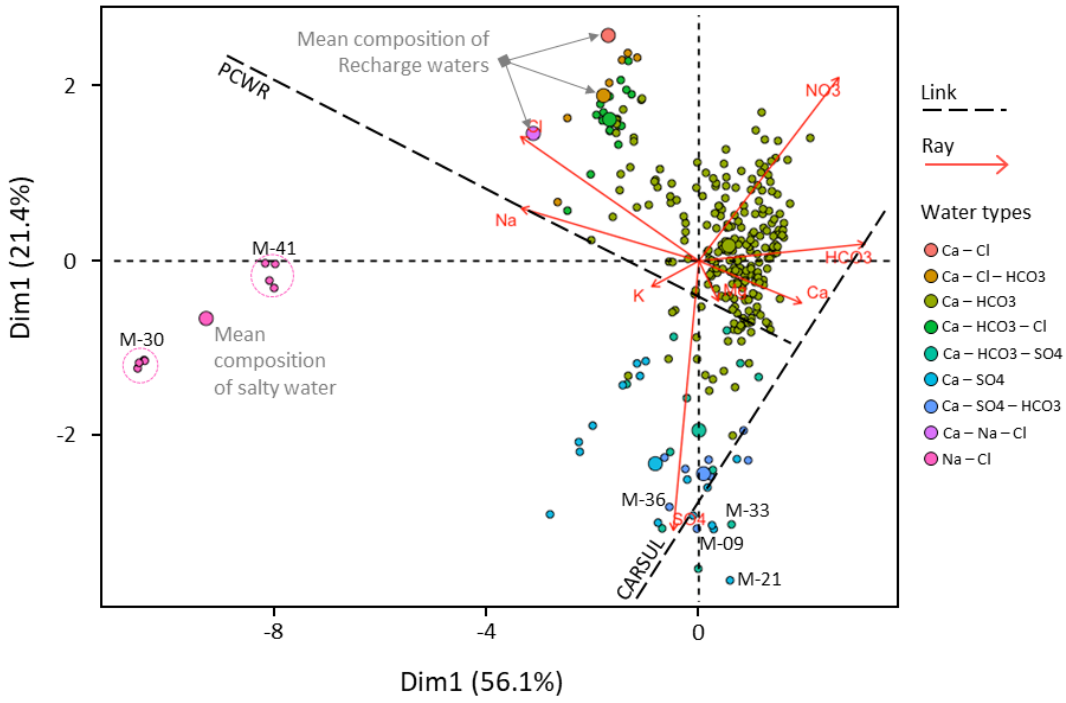


Fig. 3. clr-Biplot of the Principal Components PC1 and PC2 for the dataset Matrix (300x8). The label of the axes indicates the percentage of the variance explained by PC1 and PC2, respectively. The PCWR dashed line indicates the link between Pristine waters and groundwater with Water-Rock interaction. The CARSUL dashed line indicates the link between CARbonate and SULphate waters. The smaller circles correspond to the different water samples and their color indicates their corresponding water type, whereas the larger circles represent the average composition of the different water types. To illustrate this the groundwater samples from springs M-30 and a M-41 are indicated as well as the corresponding mean composition.

The cluster analysis is applied to group observations into several homogeneous clusters. It is based upon similarities between the observations and provides insights regarding the multivariate geochemistry characteristics (Bondu et al., 2020; Templ et al., 2008). In this work it is used the 'soft' model-based clustering method. One of the main advantages is that it uses a probability-based approach. Therefore the obtained partition can be interpreted from a statistical point of view unlike the classical 'hard' - or heuristic-based - algorithms (k-means, hierarchical clustering, etc.) (Bouveyron and Brunet-Saumard, 2014). The model-based clustering approach used was the finite mixtures of multivariate-normal or Gaussian distributions known as Gaussian Mixture Model (GMM), which is included as {Mclust} (Fraley and Raftery, 2002; Fraley et al. 2012; Scrucca et al., 2016), in R version: 5.4.6 (Raferty et al., 2020). It assumes that observed data come from a mixture of underlying probability distributions representative of two or more clusters.

The GMM assumes that following probability distribution function (PDF)

$$f(x) = \sum_{k=1}^{K} \omega_k f_k(x|\mu_k, D_k) \ , \tag{5}$$

where ω_k represents the weight or mixing proportion ($0 \leq \omega_k \leq 1$; $\sum_{k=1}^{K} \omega_k = 1$) or probability that an observation comes from the k_{th} mixture component, K is total number of components (i.e., groups or clusters), and f_k is the PDF of the observations for the k^{th} variable. Each component is usually modeled by a normal distribution (Eq. 6) with mean μ_k and covariance matrix D_k.

$$f_k(x|\mu_k, D_k) = \frac{1}{(2\pi)^{\frac{p}{2}} \cdot |D_k|^{\frac{1}{2}}} \exp\left[-\frac{(x-\mu_k)^2}{2 \cdot D_k}\right] \tag{6}$$

Taking into account Eq. 5 the conditional probability of assigning one observation to a given cluster is given by

$$P(cluster\ k|x) = \frac{\omega_k f_k(x|\mu_k, D_k)}{f(x)} \tag{7}$$

The greater the value of P the closer the association of sample x with the PDF corresponding to the cluster k. By definition, those samples for which P > 0.5 for PDF k constitute a ''cluster''.

For the different components K, the model parameters ω_k, μ_k, and D_k are estimated using the expectation–maximization (EM) algorithm (Dempster et al., 1977). The covariance matrix D_k describes the geometry of the clusters with its volume, shape and orientation The different combinantions of these parameters allows to define 14 multivariate mixture models grouped in three main families: spherical, diagonal, and, ellipsoidal which are included in the version used of {Mclust} package. In the other hand, this package uses the Bayesian Information Criterion (BIC) to find the optimum number of clusters and identify from those 14 multivariate mixing models, the one that best characterizes the data while maximizing BIC. More details of the GMM, BIC and EM mathematical approach, can be found on Biernacki and Govaert (1999); Fraley and Raftery (2002, 2012) and Raferty et al. (2020). In this study, model-based clustering has been applied to the dataset Matrix (43x8) of major ion data (HCO_3, Cl, SO_4, NO_3, Ca, Mg, Na, K), represented in this case using ilr-coordinates (Eq. 2).

3.4. Determination of Natural Background Levels (NBLs) and Threshold Values (TV)

After identifying the number of underlying clusters in the data set in hand based on MSA tools, the NBL and TV values for Cl, SO_4 and NO_3 are determined, which are the most common solutes causing specific groundwater pollution issues in HMKS, are determined. In this work the PS-method developed in the framework of the EU "BRIDGE" (2007) project (Müller et al., 2006) is applied since it has been successfully proven in many studies (Coetsiers et al., 2009; Ducci and Sellerino, 2012; Hinsby et al., 2008; Marandi and Karro, 2008; Parrone et al., 2019; Preziosi et al., 2010; Wendland et al., 2008; Zabala et al., 2016). The PS-method considers the following criteria for the data preparation before estimating the NBL's:

- Time series should be replaced by median averaging (i.e. all sampling sites contribute equally to the NBL estimation).
- Samples with incorrect ion balance (exceeding 10%) and samples with median NO_3 contents >10 mg/L must be rejected.
- Salt waters (i.e. NaCl) exceeding 1 g/L must not be considered.
- If samples are anaerobic (O_2 < 1 mg/L) or denitrification occurs, the dataset needs to be evaluated for the aerobic and anaerobic samples separately.

To obtain the NBL, the 90th percentile is advisable for small datasets (N ≤ 60 sampling points) or when human impact cannot be excluded from the data, which is the case of the case study in this research. For n > 60 the 97.7th percentile is preferred. Once the NBLs are defined then the TVs are obtained following the final methodology suggested by the EU "BRIDGE" project:

$$TV = \begin{cases} \frac{1}{2} \cdot (\mathrm{NBL} + \mathrm{Ref}); & NBL \leq Ref \\ NBL; & NBL > Ref \end{cases}, \quad (9)$$

where *Ref* is the reference value. In case of the Spanish Royal Decree 140/2003 of 7 February, laying down the health criteria for the quality of water intended for human consumption, the values of *Ref* for SO_4, NO_3 and Cl are 250 mg/L, 50 mg/L and 200 mg/L, respectively.

4. Results and discussion

4.1. Exploratory analysis of data and general water chemistry

In order to explore the internal structure of the dataset Matrix (43x8) of major ions (HCO_3, Cl, SO_4, NO_3, Ca, Mg, Na, K), different EDA plots (a combination of histogram, density trace, and one-dimensional scatterplot in one view) (Reimann et al., 2008), and boxplot in just one plot, were prepared (Fig. 4). Having this in mind, the ilr coordinates are adapted to the univariate case with the package {StatDa} (Filzmoser et al, 2009, 2009b). The variable of interest *x* (i.e. in this case Cl, NO_3 and SO_4) is single ilr-transformed (Eq. 11):

$$z = \frac{1}{\sqrt{2}} \cdot ln\left(\frac{x}{1-x}\right) \quad (11)$$

The resulted histograms show multi-model shapes in all the cases (i.e. major ions) suggesting that different populations are superimposed. In order to explain the dataset, and considering the geological setting of the area, a hypothetical mixture model with multiple components of different natural geogenic origin (possibly affected with local

anthropogenic sources) must be considered. Thus, a simply bi-modal distribution composed of natural vs anthropogenic contamination cannot be considered to establish the NBLs without taking into account the multivariate character of the data. Thus, the first step is to address separating the chemical groups or clusters.

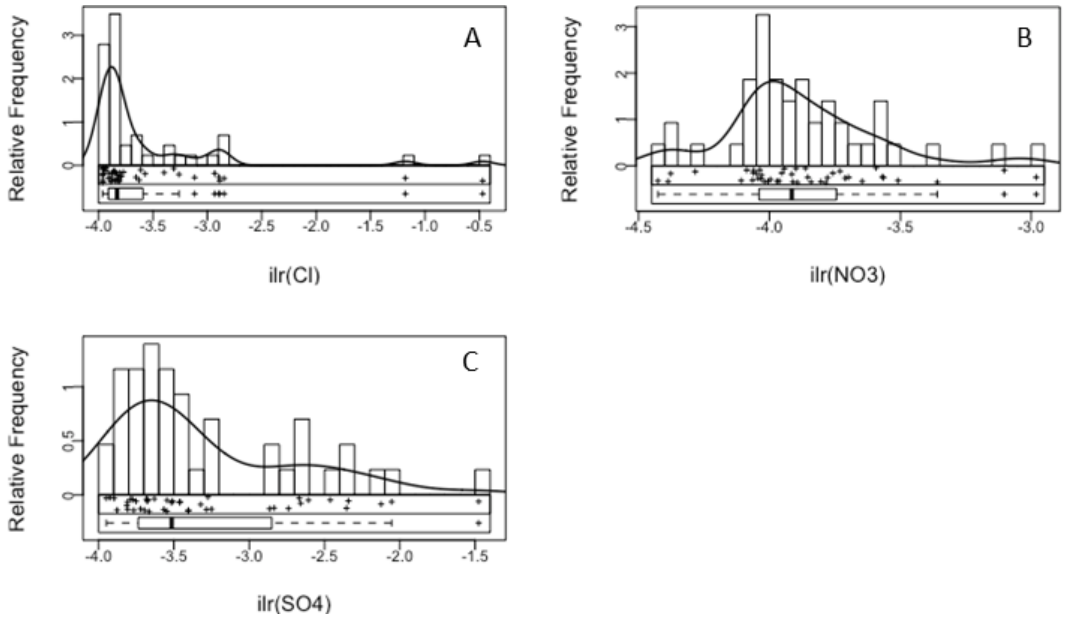
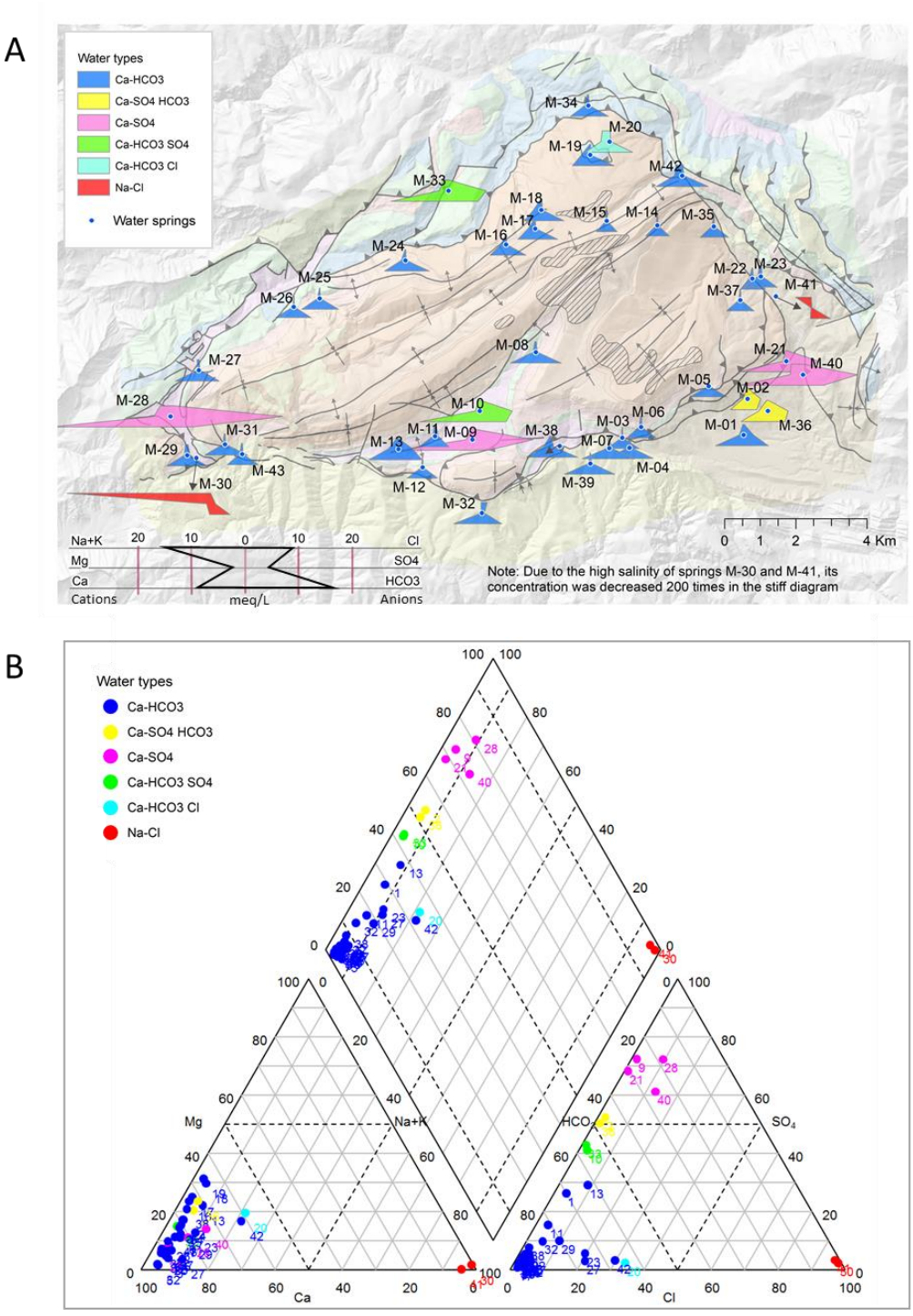


Fig. 4 EDA plots for Cl (A), NO_3 (B), and SO_4 (C).

Classical graphical methods for the classification of water chemistry data, such as Piper and modified Stiff diagrams were used as a first step to analyse the whole dataset (except those water samples from pluviometers (i.e. in total 288 samples). The Fig. 5 shows a map with the Stiff diagrams distribution over the PCM and also the corresponding modified Piper diagram. Based on that information it is possible to initially aggregate the groundwater discharge from the 43 springs into 6 types of hydrogeochemical facies (Table 2):



512

513 **Fig. 5.** Hydrochemical diagrams. (A) Stiff diagram map and (B) Piper diagram associated
514 to the selected springs in the PCM. In both cases, for every spring the ion dissolved
515 content values correspond to the median value associated to all samples taken from that
516 spring. The springs are classified by their hydrochemical facies.

517

518 **Table 2.** Identified water types

Water type	Num. Springs	Geological units[a]
$Ca\text{-}HCO_3$	32	Cretaceous (KMca, Kgp, Kat) Paleogene-Eocene (PEab, PEci, PEcp1, PEm1) Paleogene-Oligocene (POcgs, POmlg, PPEc) Quaternary (Qpe, Qt0, Qvl) Triassic-Jurassic (TJb, TJcd) Triassic Muschelkalk (Tm)
$Ca\text{-}HCO_3\text{-}Cl$	1	Paleogene-Eocene (PEcp2)
$Ca\text{-}SO_4$_	4	Quaternary (Qcoo) Triassic-Keuper (Tk)
$Ca\text{-}HCO_3\text{-}SO_4$	2	Triassic-Keuper (Tk)
Na-Cl	2	Triassic-Keuper (Tk)
$Ca\text{-}SO_4\text{-}HCO_3$	2	Paleogene-Eocene (Pemb)

(a) For a given Water type, the geological units based on ICGC, (2007) ordered by number of springs

519

520 At the first glance, the results show that diverse springs outcropping from different
521 geological units (ICGC, 2007) show similar groundwater facies, or also the same facies
522 can be shown from different points located at different geological units. In this context,
523 these graphical techniques should not be considered determinant alone to discriminate
524 between hydrochemical groups and therefore, their results should be considered only
525 preliminary. Table SM.1 (Supp. Mat.) shows the summary of the major ions content of
526 the 43 monitored springs (expressed as median values of time series for the period
527 September 2013 – October 2015) and also the water facies associated to them.

528 **4.2. PCA and dataset matrix size**

529 The variation matrix for the dataset Matrix (300x8) (Table 3) shows strong correlations
530 between different pairs of variables such as Ca and HCO_3, Na and Cl, and Mg and HCO_3.
531 Besides, NO_3 shows a high correlation with Ca and HCO_3, whereas almost no correlation
532 with SO_4. This result indicates that the most groundwater samples affected by nitrate
533 pollution are those from the Eocene karst aquifer with a Ca–HCO_3 hydrochemical
534 composition.

535

536 **Table 3.** The upper triangle over the main diagonal shows the '*index of proportionality*'
537 (Eq. 4) of the dataset Matrix (300x8). The lower triangle over the main diagonal shows

in italic the *'index of proportionality'* of the dataset Matrix (43x8). In both cases, the correlation values larger than 0.5 are shaded in blue.

	Ca	Mg	Na	K	HCO_3	Cl	NO_3	SO_4
Ca	--	0.88	0.06	0.51	0.98	0.04	0.57	0.44
Mg	*0.76*	--	0.34	0.69	0.67	0.26	0.27	0.52
Na	*0.01*	*0.10*	--	0.54	0	0.96	0	0.15
K	*0.49*	*0.75*	*0.58*	--	0.15	0.36	0.13	0.43
HCO_3	*0.94*	*0.41*	*0*	*0.07*	--	0	0.56	0.06
Cl	*0.01*	*0.07*	*0.99*	*0.40*	*0*	--	0	0.05
NO_3	*0.55*	*0.07*	*0*	*0.02*	*0.59*	*0*	--	0.01
SO_4	*0.17*	*0.24*	*0.07*	*0.25*	*0*	*0.02*	*0*	--

Initially the PCA is conducted with the whole dataset (N=300), including the hydrochemical composition of natural and artificial snow, water from ponds and groundwater samples. The PCA with clr transformed data shows that only with three principal components, the 87.4 % of total variance can be explained (Fig. 6). The PCA is affected by the presence of natural outliers, in our case from the Na-Cl hydro-facies, that completely distorts the shape of the biplots (Fig. 6B, 6C and 6D). The scores are classified according to the distinguished nine water types when considering the complete dataset.

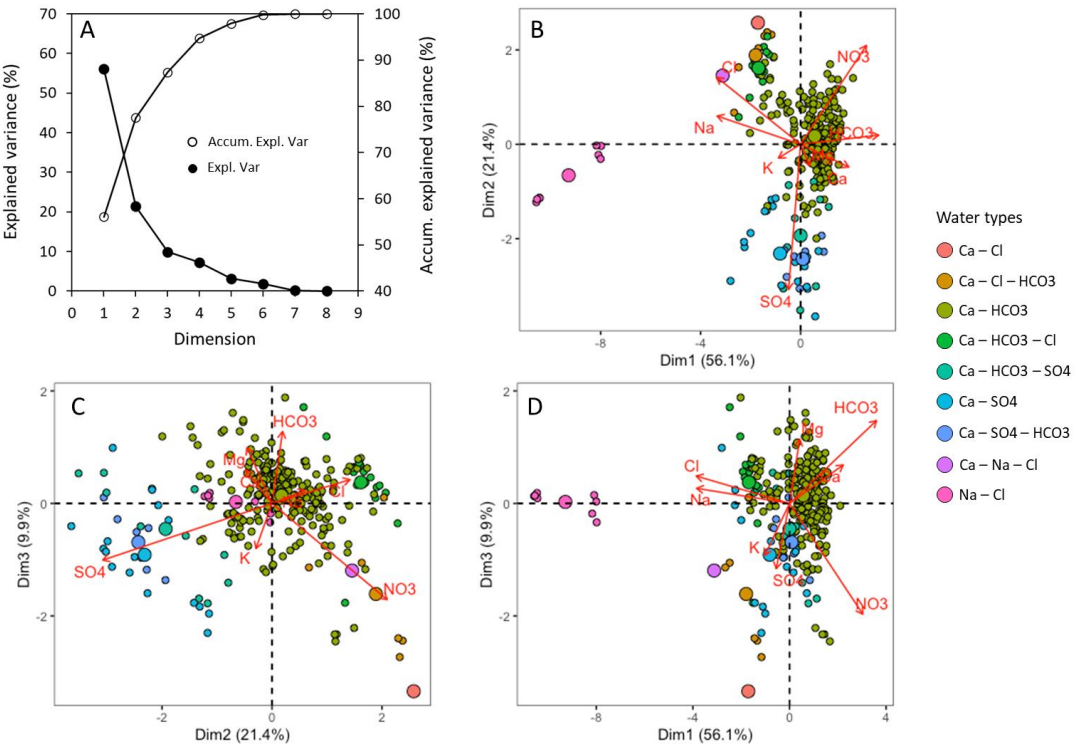


Fig. 6 (A) Scree-plot of dataset Matrix (300x8) showing the explained (solid circles) variance associated to every PC of the PCA, and the accumulated explained variance (empty circles) as the different PCs are accounted in the PCA. (B) Compositional biplot PC1 vs PC2 (C) Compositional biplot PC2 vs PC3 and (D) Compositional biplot PC1 vs PC3 showing scores (circles) and loadings (arrows) for clr transformed data. In the biplots, the bigger points represent the mean clr-value for each water type.

From the distribution of the water samples in the clr-biplots several subgroups of waters with clear similarities can be read. The biplot between PC2 and PC3 clearly separates sulfate waters. Moreover, looking at the biplot between PC1 and PC2 closely (Fig. 3), different hydrochemical spatial trends, likely associated with changes in terms of bedrock lithology, can be observed. In fact, it can be inferred that: (1) The highest clr-variances are shown for SO_4, Cl and NO_3, followed by Na and HCO_3. The lowest clr-variances are shown for Ca, K, and Mg; (2) The PCA has situated the saltiest waters (M-30 and M-41) separately in the western quadrant of the biplot. As can be shown, using clr-transformed data allows to correctly separate characteristic points of the domain, which correspond to the deepest drainage from the Keuper materials; (3) The groundwater samples from the remaining springs are located in the eastern north and south quadrants: the freshest waters that are more related to the upper Eocene karst aquifer are situated at the eastern-north quadrant and present some correlation with NO_3. The samples related to Cretaceous and Triassic materials appear to be more disperse, being most of them at the eastern-south part of the biplot with extreme values in springs M-21, M-9, M-33, M-36, among others.

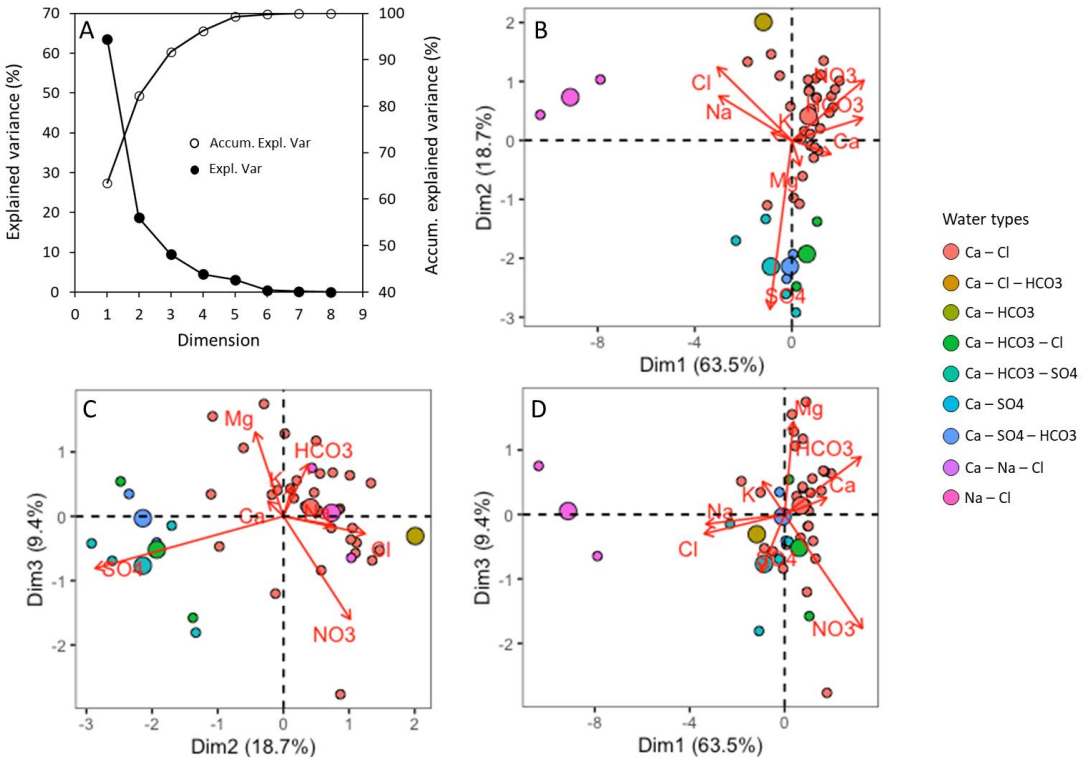
Taking into account the specific rules for interpreting clr-biplots, the following aspects can be highlighted:

- It is possible to draw a link between the vertices of Na, K and Mg indicating that these variables may form a sub-composition with a single degree of freedom.
- The vertices of SO_4, Ca and HCO_3 lie almost on a common link. This link is also almost orthogonal to the link drawn between Na, K and Mg suggesting that these two sub-compositions may vary independently of each other.
- The two indicated links can be interpreted as two independent set of processes on the hydrochemistry of the springs: (1) The "Pristine Character/Water-Rock interaction" link PCWR [Na, K, Mg] which represents as one end-member, the groundwaters influenced NaCl contributions derived from Keuper materials but

also to recharged waters (Ca–Cl–HCO_3; Ca–Cl, Ca–NaCl) at the upper part of the PCM, and representing the other end member the waters that have interacted longer with the Tertiary karst system materials. (2) The "CARbonate/SULfate dissolution" link 'CARSUL' [SO_4, Ca, HCO_3] representing the dissolution of different types of carbonate and sulfate rocks (HCO_3 as one end member of the link and SO_4 as the other one).

- Samples in the eastern-south quadrant of the biplot are more disperse having a stronger association with the SO_4 vertices.

In the case of the dataset Matrix (43x8) the variation matrix (Table 4) is consistent with that of the dataset Matrix (300x8), showing strong correlations between the same pairs of variables, and even with similar correlation values. The PCA with clr transformed data shows that when considering two or three PCs, it can be explained 87.4% and 91.7% of total variance, respectively (Fig. 7). Besides, the resulting clr-biplots are similar in shape to those of Matrix (300x8). As it can be shown, the reduction of the dataset matrices from (300x8) to (43x8) in the PCA does not introduce any relevant change in the final inference regarding the geochemical characteristics of groundwater. This is convenient from the perspective of dimensionality issues.



603 **Fig. 7** (A) Scree-plot of dataset Matrix (43x8) showing the explained variance (solid
604 circles) associated to every PC of the PCA, and the accumulated explained variance
605 (empty circles) as the different PCs are accounted for in the PCA. (B) Compositional
606 biplot PC1 vs PC2 (C) Compositional biplot PC2 vs PC3 and (D) Compositional biplot
607 PC1 vs PC3 showing scores (circles) and loadings (arrows) for clr transformed data. In
608 the biplots, the bigger points represent the mean value for each water type.
609

610 **4.3 Clustering analysis**

611 The principal aim of cluster analysis is to split a number of observations into groups that
612 are similar in their characteristics or behaviour (Reimann et al. 2008). To this end, the
613 dataset Matrix (43x8) is used. Before conducting the ilr transformation, an intuitive
614 sequential binary partition (SBP) is used to characterize the hydrochemical variability
615 within the domain wherein non-overlapping groups of parts, known as balances, are
616 defined. According to Egozcue and Pawlowsky-Glahn (2005; 2006), two methods for
617 performing SBP can be applied: (1) directly a PCA, and (2) by experienced judgment. In
618 this work SBP is based on knowledge of the groundwater chemistry in the study area and
619 on the resulting compositional biplot (Fig 7). As a result, seven groundwater partitions
620 are considered (Table 4): the ilr_1 balance separates the Ca-HCO_3 waters (mostly affected
621 by NO_3) from the rest; the ilr_2 separates those waters affected/non-affected by NO_3
622 pollution; the ilr_3 separates the contribution of calcite and dolomite to groundwater; the
623 ilr_4 separates Ca from HCO_3; the ilr_5 separates SO_4 waters from most salty waters; the
624 ilr_6 separates K from Na/Cl; and finally the ilr_7 separates Na and Cl.

625 **Table 4.** SBP of a 7-part composition (ilr_1, ilr_2, …, ilr_7) for describing isometric log
626 ratio (ilr) coordinates based on the separation of anions and cations related to the
627 hydrochemical composition of natural groundwaters for the clustering analysis.

ilr	Ca^{2+}	Mg^{2+}	Na^{+}	K^{+}	HCO_3^{-}	Cl^{-}	NO_3^{-}	SO_4^{2-}
ilr_1	+1	+1	-1	-1	+1	-1	+1	-1
ilr_2	+1	+1	0	0	+1	0	-1	0
ilr_3	+1	-1	0	0	+1	0	0	0
ilr_4	+1	0	0	0	-1	0	0	0
ilr_5	0	0	+1	+1	0	+1	0	-1
ilr_6	0	0	+1	-1	0	+1	0	0
ilr_7	0	0	+1	0	0	-1	0	0

628

In order to inspect the sensibility of considering 'hard' clustering methods to determine the optimal number of clusters (k), a first and preliminary analysis was performed using the {clValid} (Brock et al. 2008) and {NbClust} (Charrad et al. 2014) R packages and using clr and ilr coordinates. These packages contain a set of functions for both validating the results of different heuristic-based clustering algorithms (e.g., '*hierarchical*', '*k-means*', '*diana'*, '*pam'* and 'clara (*k-medoids*)'), and determining the relevant number of clusters in a dataset. The clustering models may account for different linkage methods (i.e., '*complete*', '*average*', 'single' and '*ward*') and dissimilarity metrics ('*Euclidean*' and '*Manhattan*', among others). In the case of the dataset Matrix (43x8), the best results have been obtained for the *hierarchical* and *k-medoids* models, and for k clusters between 2 and 4. Nevertheless, given the large number of methods and their corresponding options for setting the models it is actually difficult to select the most suitable one.

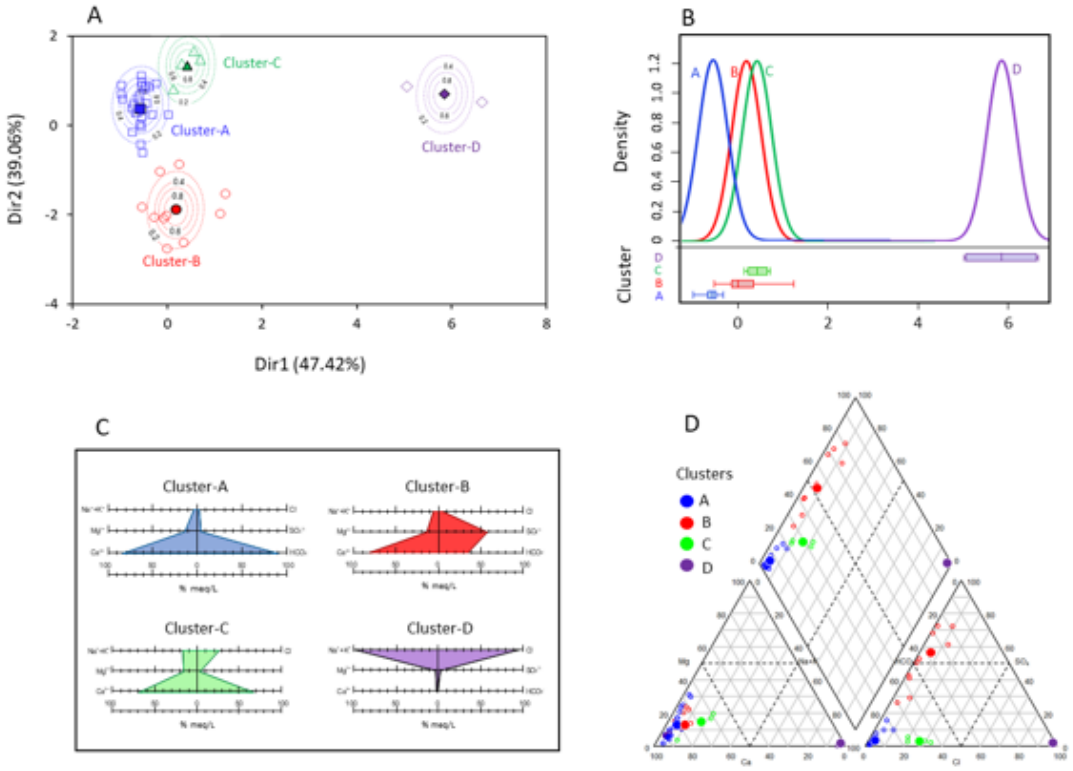
To avoid the clustering method selection issues, the cluster analysis was performed by solving a GMM using the {Mclust} R package (Fraley and Raftery, 2002; Fraley et al., 2012; Scrucca et al., 2016) and considering the ilr-transformed data. The obtained results indicates that in the case of the dataset Matrix (43x8), the best multivariate clustering option is obtained applying the 'EEI' model while considering a total of 4 clusters (see Fig. SM.4.1 in Suppl. Mat.).

The scatterplot matrix obtained with the model-based clustering process using the seven ilr coordinates. Being D the previous dimension of the original dataset matrix (43x8), *D*-1 coordinates can be shown in Fig. SM.4.2 (Suppl. Mat.). In order to visualize the clusters in a most suitable way, the dimension reduction function 'MclustDR' (Scrucca, L. 2010) for visualizing the classification structure obtained from the finite mixture of Gaussian densities of the {Mclust} package is used to reduce the dimensionality of the ilr matrix and estimate the principal components. Table. SM.4.1 and Fig. SM.4.3 (Suppl. Mat.) provide the scores of the reduced ilr-matrix and their representation in a scatterplot, respectively. As can be shown, the two main principal components explain 86,42% of the total variance. As a result, and only a glance at the scatterplot of PC1 and PC2 (Fig. 8A) the cluster division for the different springs reveals in a clear way, and each cluster can be described by the corresponding PDFs (Fig. 8B). Worth to point out the similarity between the distributions of samples in the 2D space (albeit in a symmetric plane). The Fig. 8C presents the mean hydrochemical composition of each cluster (Table SM.4.2 in

Suppl. Mat.) in terms of modified Stiff diagrams, and Fig. 8D shows in a Piper diagram how the mean hydrochemical composition of the clusters is representative of the composition of the corresponding springs.

The probabilistic GMM framework estimates the optimal number of clusters and provides for every spring the probability of belonging to these clusters (soft assignment). This approach is more interesting that the classical clustering approaches, in which the number of clusters is assumed fixed, and every spring is assigned to one and only one of the previously assumed clusters (hard assignment) (Kim et al., 2014). From an hydrochemical point of view, the soft assignment often provides the more interesting interpretation because the method reveals if one observation is influenced by several factors (Templ et al., 2008). Moreover, Wu et al., (2017) show how the probabilistic GMM clustering provides insights into hydrochemical processes affecting groundwater, even though with a limited number of observations, which is a common situation in high mountain kart aquifers such as PCM.

The conditional probabilities (P) of assigning one observation to a given cluster (Eq. 7) are given in Table SM.4.3 (Suppl. Mat.). In all cases, springs are assigned to one cluster with a probability > 0.95, and specifically, more than 83% of the springs reach the probability of '1'. The smaller probabilities occur in M-01 (P = 0.911 cluster A) and M-13 (P = 0.969 cluster B). Spring M-01 discharges from the Eocene karstic limestones. Nevertheless, this discharge might be affected by weak contributions of Tertiary sulfates (which are related to the locally known as 'Beuda gypsum Formation'). The discharge in M-13 shows a Ca-HCO_3 hydrogeochemical composition despite discharging from the Triassic (Muschelkalk) limestone aquifer. In this case the groundwater discharge is weakly affected by the underlying Keuper materials.



686 **Fig. 8.** (A) clr-Biplot PC1 vs PC2 with the distribution of each cluster after dimension
687 reduction for dataset Matrix (43x8) ilr-transformed data. The dashed lines correspond to
688 the probability zones of belonging a certain cluster in the subspace PC1-PC2. Solid
689 symbols correspond to the mean hydrochemical composition of the clusters (B) PDF's of
690 the resulting 4 clusters. (C) Modified Stiff diagram associated to the mean hydrochemical
691 composition of the clusters. (D) Piper diagram associated to the selected springs in the
692 PCM classified by their corresponding cluster to which they belong. Solid symbols
693 correspond to the mean hydrochemical composition of the clusters.

694 The hydrogeochemical description of each groundwater cluster can be summarized as:

695 • **Cluster A** is characterized by low mineralization and dominated by Ca–HCO_3
696 water type and slightly alkaline pH. In total 27 springs are grouped in this cluster
697 which correspond to 203 groundwater samples collected in the study from the
698 total of 288. All the springs drain directly or indirectly (i.e. covered by local
699 quaternary deposits) the Tertiary Eocene upper karst aquifer of the PCM (Fig. 9)
700 and from the higher parts of the mountain (944 - 2144 m a.s.l.). They are almost
701 mainly found inside the structural limits of the PCM sheet and at its boundaries
702 except some of them localized in Quaternary deposits or discharging karstic

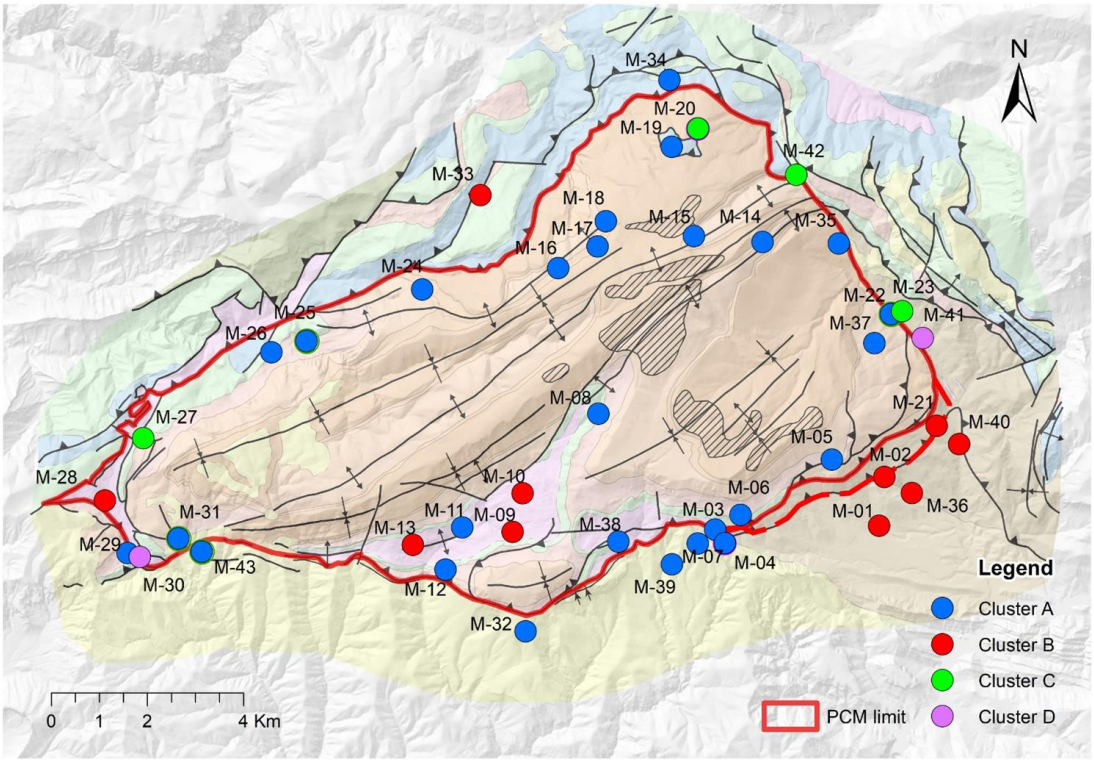
conduits trough the Oligocene carbonate karstic conglomerates situated just in the front of the thrust sheet (e.g. M-03, M-04, M-07, M-39, M-32 and M-43, which is one of the most important karst springs of the system). Another special case is the spring M-06 which lies over garumnian shales, marls and limestones (Kgp) outcropping materials. In this zone, a fault affecting the stratigraphy might allow the hydrological connection between the lower Eocene limestones (PPEc) and Kgp formations. This connection would explain the CA–HCO_3 water type associated to spring M-06, and also its classification in the cluster A, thus pointing the groundwater discharge origin as the Eocene Tertiary aquifer. Finally, the spring M-29 actually drains a Eocene limestone level situated at the west of the PCM boundary.

Cluster A presents the lower values of EC, which ranges between 186 and 486 μS/cm and the minimum values of groundwater temperatures. The concentrations of Cl and SO_4 are very low ranging between 2.5 and 15 mg/L and between 2.6 and 25.3 mg/L respectively. In 13 samples, the concentration of NO_3 is above 10 mg/L, and in one specific spring (M-32) exceeds in all samples the legal limit for potable water (50 mg/L). The average Saturation Indices (SI) estimated with the Phreeqc program (Parkhurst and Appelo, 2013) for calcite, gypsum and halite are 0.23, -2.67 and -9.68, respectively. The groundwaters are representative of the recharge of the karst system in the highest altitudes of the massif, where the dissolution of carbonates is the dominant geochemical process controlling groundwater chemistry.

- **Cluster B** encompasses water types from Ca–HCO_3 to Ca–HCO_3–SO_4, Ca–SO_4-HCO_3 and Ca–SO_4, which are characterized by moderate mineralization and slightly alkaline pH. This group includes 10 springs. A total of 40 groundwater samples collected in the study would correspond to this cluster. The springs related to Cluster B are situated either inside or outside the internal structural limits of the PCM thrust sheet. The springs situated inside (M-9, M-10 and M-13) occur mostly in (1) Cretaceous and Triassic (Keuper) materials outcropping in the area. These materials underly the principal aquifer of the massif (the Eocene carbonate karstic system), and (2) local shallow granular aquifers. The springs M-01, M-02, M-21 and M-36 are related to sediments with high content of Tertiary

gypsum from the Beuda Formation, which outcrops in small pinched belts located in front of the southeastern part of the PCM thrust sheet. Springs are located at the lowest parts of massif (altitudes ranging between 867 - 1456 m a.s.l.). The EC varies between 493 and 2102 μS/cm. The SO_4 concentration is quite high and it ranges between 88 and 989 mg/L exceeding in most cases the legal limit for potable water (250 mg/L). The concentration of Cl ranges between 3.8 and 94.5 mg/L. The average SI for calcite, gypsum and halite are 0.32, -0.99 and - 8.61 respectively.

- **Cluster C** includes water types from Ca–HCO3 and Ca–HCO3-Cl water types. This group includes 4 springs and a total of 37 groundwater samples from which 26 of them correspond to the spring M-20 (located at 1858m a.s.l.). Except the spring M-20, the rest (M-23, M-27, M-42) are located at the boundaries of the PCM sheet. The EC varies between 332 and 747 μS/cm. Although they have SO_4 concentration similar to cluster A with 9.7- 15.3 mg/L, the content of Cl is much higher ranging between 24 and 82 mg/L. These higher values compared to cluster A are interpreted as related with groundwater flow through areas with the presence of relict halite or salty water in closed pores in the Keuper materials. In the case of M-20 (which is located inside the PCM sheet) the salt is related to a klippe of Jurassic delineated into the geological map. Besides, in the catchment area of this spring, there are small outcrops of Keuper materials detected during the fieldwork. The average SI for calcite, gypsum and halite are 0.24, -2.32 and -7.42 respectively.
- **Group D** is the most evident and special. corresponding to Na–Cl type waters (Fig. 8). This group is composed of 2 salt springs (M-41 and M-30) located at the 993 and 1023 m a.s.l. at the East and West boundaries of the PCM sheet respectively. They are characterized by very high mineralization discharging from Keuper confined bedrocks and interpreted as the contribution of deep groundwater flow with elevated transit times that allows a significant solute difussion. The waters are slightly acidic to near-neutral pH. The M-41 and M-30 samples presents EC values of 57.2 and 247.1 mS/cm, Cl concentrations of 21 and 178.2 g/L, and SO_4 concentrations of 1.2 and 8.1 g/L respectively. The M-30 spring can currently be considered the saltiest spring of natural origin in Catalonia.



767
768
769 **Fig. 9.** Spatial distributions of the 43 clustered springs over the geological map of the
770 PCM based on the GMM. The description of the different geological materials is the same
771 presented in Fig. 2.

772

773 In the framework of multivariate statistics data analysis (e.g. PCA and data clustering),
774 specially dealing with compositional data (i.e. data that carry only information about the
775 relative abundance of each component on the whole, such as the hydrogeochemical data
776 sets), it is important to suitably transform the dataset using the CoDa approach (e.g., Eq.1
777 or Eq.2) before conducting any analysis, otherwise it is very likely to obtain wrong results
778 (Otero et al., 2005). Moreover, uninterpretable results are also obtained when applying
779 the classical standardization methodology known as "z-score" on compositional data,
780 which considers logarithms and then subtracts the mean and divides it by the standard
781 deviation to scale (Blake et al., 2016). To illustrate the importance of using correct CoDA
782 data transformations, the dataset Matrix (43x8) is used to apply the same MSA analysis
783 (PCA and the model-based clustering GMM) but using the classical standardization
784 approach (or Z-score normalization) instead of using the CoDa transformation approach.

As can be shown (see at the Annex.2) neither the PCA variable loadings nor the soft clustering results make any clear hydrogeological sense.

4.4 NBLs and TVs values.

Once the groundwater clusters are defined for the PCM, the NBL and TV's for NO_3, SO_4 and Cl have been obtained applying the PS-method (Müller et al., 2006). Taking into account the criteria required for data to be accounted when estimating the NBL's with this method (section 3.4), the following observations apply:

- The groundwater samples from M-30 and M-41 (the whole cluster D) present Cl concentrations of geogenic origin above the drinking water limit (>200 mg/L). Therefore, these samples are not considered in the NBL determination.
- NO_3 concentrations above the drinking water limit (>50 mg/L) are mostly observed in M-32 spring (cluster A). Besides, the springs M-32, M-10, M-11 and M-28 present NO_3 concentrations > 10 mg/L. Following the PS-method criteria, these springs have been excluded of the NBL determination.

The NBLs for the remaining groundwater samples belonging to the clusters A, B and C are obtained taking into account the 90^{th} percentiles (P90) of the corresponding cluster ECDF plots (Fig. 10B). The obtained NBL90 and TVs are presented in Table 5. The results indicate that Tertiary Eocene karst aquifer (cluster A), which is the principal aquifer inside the PCM presents the lowest NBL90 values for Cl, SO_4 and NO_3. The cluster B, which is related to the aquifers on the Cretaceous and specially the Triassic Keuper materials presents the highest NBL90 value for SO_4, and the Cluster C, which is generally related to local small aquifers located at the boundaries of the PCM presents the highest values of NBL90 for both Cl and NO_3.

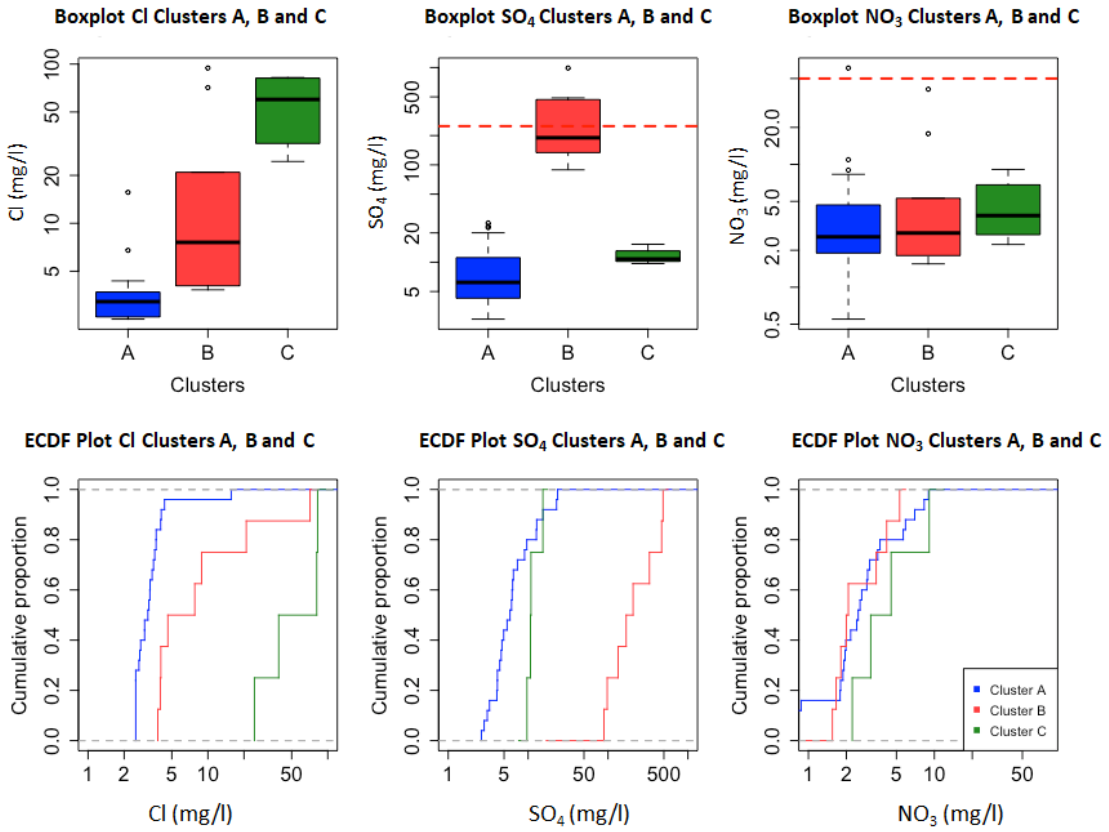


Fig. 10. (**A**) boxplots of the clusters A, B and C for SO_4, Cl and NO_3. The dashed red lines indicate the reference limits established in the Spanish Royal Decree 140/2003 (**B**) ECDF plots.

Comparing the obtained NBL90 values with those officially assigned to GWB-5 and GWB-44 (Table 1), it looks that the NBL official values of SO_4 assigned to both GWBs (485 and 609 mg/L, respectively) are likely conditioned by the interaction between fresh groundwater and most probably evaporites of the Upper Triassic (Keuper facies). These evaporites appear very often at the boundaries of many thrust sheets throughout the Southern Pyrenean zone. Additionally, the official NBL90 value of Cl assigned to GWB-44 is similar to that obtained for cluster B, which is related to the Keuper deposits. Likewise, the obtained NBL90 value of NO_3 for the Cluster C is similar to the official one for GWB-44. As can be shown, none of the official NBL90 values defined for GWB-5 and GWB-44 correspond to those values obtained for the Lower Eocene limestones and dolomites, which constitute by large the main aquifer of the PCM.

Table 5. Summary results of the NBL and TV's values derived from the PS-method (BRIDGE, 2007) for clusters A, B and C for the solutes Cl, SO_4 and NO_3.

	Cl [mg/L]		SO_4 [mg/L]		NO_3 [mg/L]	
Clusters	**$NBL_{90\%}$**	**TVs**	**$NBL_{90\%}$**	**TVs**	**$NBL_{90\%}$**	**TVs**
A	4.06 ± 2	8.12 ± 2	14.33 ± 2	29.66 ± 2	6.55 ± 2	13.1 ± 2
B	35.98 ± 2	71.96 ± 2	471.71 ± 2	471.71 ± 2	4.51 ± 2	9.02 ± 2
C	81.92 ± 2	140.96 ± 2	13.96 ± 2	27.92 ± 2	7.73 ± 2	15.46 ± 2

It is well known that high mountain karst aquifers generate highly valuable water resources for the downstream water depending ecosystems. Their protection and rational management is of utmost importance to sustain such ecosystems and satisfying their water demands (Kazakis et al., 2018). In this framework, NBLs provide an objective scale to compare with when the quality status of the aquifer is assessed. Nevertheless, these aquifers are often immersed in complex geological settings as happens in the axial zone of the Central Pyrenees (Lambán et al., 2015), in the Picos de Europa massif (Ballesteros et al., 2015), in the Jura Mountains (Luetscher and Perrin, 2005) and the Hochifen–Gottesacker Alps (Goldscheider, 2005), among others. The NBLs are obtained as a function of the hydrochemical content measured in the different springs discharging the system. Nevertheless, in geological complex zones it is difficult to assert if one certain spring is discharging groundwater from the aquifer of interest or not, because the geographical location of the spring may suggest an origin for the sampled groundwater while hiding mixing relations between groundwater flow lines from other local aquifers with different hydrogeochemical fingerprint (Lambán et al., 2015; Barbieri et al., 2017; Sánchez et al., 2017).

The European Union Water Framework Directive (WFD, 2000) defines a general framework for integrated river basin management in Europe to ensure their "good water status". Nevertheless, the river basin is often a entity hard to manage because the larger the size of the basin the larger is (1) the number of water bodies enclosed and (2) the likelihood of political-administrative boundaries issues to appear. To avoid such problems, instead of looking at river basins the WFD refocussed on the smaller scale "river basin districts", for which administrative structures were defined to correctly manage the corresponding bodies, thus ensuring -hopefully- the right management of whole river basin (Boeuf and Fritsch, 2016). In this line, the WFD includes the guidelines that apply to define the groundwater bodies (GWB). Even in this case, some scale issues

may arise when considering the definition of the GWB in mountain zones. By definition, the GWB are assumed to belong to a certain river basin. Despite of that, it is well known that groundwater basins, specially in mountain zones, may extend throughout several river basins (Struckmeier et al., 2006; Serianz et al., 2020). As a result, GWBs may include several aquifers or even only parts of them as it happens in the PCM, whose discharge contributes to both the Ebro and the Llobregat rivers through GWB-44 and GWB-5, respectively. This is the reason why there are two different sets of NBL applying for the same aquifer (Table 1).

The WFD recognises the importance of having well defined NBLs. Their characterization must be based on a consistent and rigorous hydrochemical criteria, given that they will be used to quantitatively assess whether or not anthropogenic pollution is taking place in the corresponding aquifer (Nieto et al., 2005). The hydrogeological fingerprint of each aquifer belonging to the same GWB may be different. Therefore, the criterion of defining a single set of NBLs for the whole GWB may have no sense. Moreover, such criterium may be counterproductive from a safety perspective, given that one may assume for the GWB some quantities of species or chemical substances present in solution as normal, when actually those concentrations may be already indicating the existence of a polluting issue in some aquifers of the GWB. This is even worst when only one of these aquifers play a relevant role from a water resources perspective, as happens in the PCM. Here, the Lower Eocene karst aquifer generates an overall mean groundwater discharge that represents 15% of the mean annual water consumption in the city of Barcelona (Herms et al., 2019). Therefore, from a water resources management perspective, it might worth defining NBLs at the local scale for each aquifer. In this line, the methodology presented in this work to "complement" the sample pre-selection method is a useful tool to objectively reel off the NBL of the different high mountain aquifers belonging to a given GWB. Besides, the proposed methodology provides the GWBs managing authorities a full-sense hydrochemical criteria to better protect the high mountain pristine and strategic aquifers, while ensuring the good status of the associated high mountain river basins.

6. Conclusions

The PCM is a complex hydrogeological system composed by a main Eocene karst aquifer that drives the hydrodynamical discharge response of the massif. The PCM also includes small aquifers whose discharge present a different hydrochemical composition. The discrepancies between the official NBLs of the GWBs associated to the PCM reveal the disparities in the hydrochemical composition of groundwater from the different sampled springs belonging the GWBs. To estimate correctly the NBLs associated to one aquifer it is necessary to consider only samples from springs discharging groundwater from the aquifer of interest. In high complex hydrogeological settings this selection is not easy and must be guided by a consistent and objective clustering method.

In the case of the PCM, four compositional groups have been identified by means of GMM clustering analysis. Most of the analysed springs are dominated by Ca–HCO_3 water type coming from the main aquifer of the area. There are some springs dominated by Ca–HCO_3, Ca–HCO_3–SO_4, Ca–SO_4-HCO_3, Ca–SO_4, Ca–HCO_3-Cl, Na–Cl water types derived from other small/local aquifers. Determination of NBLs values in the area must take into account the four groups defined in this study.

In complex aquifer systems the proposed soft clustering approach, which is based on probabilistic Gaussian mixture models, provides the optimal number of clusters for the sampled springs only based upon the observed compositional data while estimating the probability of belonging to everyone of these clusters for each spring. The presented clustering approach relies on multivariate statistics methods. In this framework it is essential to transform the dataset using the CoDa rules, specially when dealing with hydrochemical compositions. Otherwise, uninterpretable results will be likely obtained.

In the case of different existing aquifers with discrepant hydrochemical fingerprints in the same GWBs, it would be reasonable to evaluate the NBLs in all of them rather than having a single set of NBLs for the whole GWB. Otherwise, errors may appear when estimating the quality status of some of these aquifers, even if the overall assessed quality status of the GWB appears to be correct.

Acknowledgements:

This research has been supported by Agencia Estatal de Investigación (AEI) from the Spanish Government and the European Regional Development Fund (FEDER) from EU through PACE-ISOTEC (CGL2017-87216-C4-1-R) projects, the EFA210/16/ PIRAGUA project which is co-founded by the European Regional Development Fund (ERDF) through the Interreg V Spain-France-Andorre Programme (POCTEFA 2014-2020) of the European Union, the Catalan Government projects to support consolidated research groups MAG (Mineralogia Aplicada, Geoquimica i Geomicrobiologia, 2017SGR-1733) from Universitat de Barcelona (UB) and GREM (Grup de Recerca de Minería Sostenible) from the Universitat Politècnica de Catalunya (UPC), the Ministerio de Ciencia, Innovación y Universidades through the METhods for COmpositional analysis of DAta (CODAMET) project (Ref: RTI2018-095518-B-C22, 2019-2021). We thank the Hydrogeology and Geothermics Unit Team of the Institut Cartogràfic i Geològic de Catalunya (ICGC) by their helpful collaboration. We acknowledge the Confederacion Hidrográfica del Ebro (CHE) and Agència Catalana de l'Aigua (ACA) for providing the official NBLs for WGB44 and WGB5, respectively. Meteorological data have been kindly provided by the Spanish Meteorological Agency (AEMET) and the Meteorological Service of Catalonia (SMC).

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability Chapman and Hall, London, New York (416 pp.).

Aitchison, J., and Greenacre, M. 2002. Biplots of compositional data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 51(4), 375–392. doi:10.1111/1467-9876.00275

Appelo, C., Postma, D. 2005. Geochemistry, Groundwater and Pollution. 2nd edition. London: CRC Press, 683 pp. https://doi.org/10.1201/9781439833544

Ballesteros, D., Malard, A., Jeannin, P. Y., Jiménez-Sánchez, M., García-Sansegundo, J., Meléndez-Asensio, M., & Sendra, G., 2015. KARSYS hydrogeological 3D modeling of alpine karst aquifers developed in geologically complex areas: Picos de Europa National Park (Spain). Environmental Earth Sciences, 74(12), 7699-7714. https://doi.org/10.1007/s12665-015-4712-0

Barbieri, M., Nigro, A., Petitta, M., 2017. Groundwater mixing in the discharge area of San Vittorino Plain (Central Italy): geochemical characterization and implication for drinking uses. Environmental Earth Sciences, 76(11), 393. https://doi.org/10.1007/s12665-017-6719-1

Biernacki, C., Govaert, Gérard., 1999. Choosing models in model-based clustering and discriminant analysis. J. Stat. Comput. Simul. 64, 49–71. https://doi.org/10.1080/00949659908811966

Blake, S., Henry, T., Murray, J., Flood, R., Muller, M.R., Jones, A.G., Rath, V., 2016. Compositional multivariate statistical analysis of thermal groundwater provenance: A hydrogeochemical case study from Ireland. Appl. Geochem. 75, 171–188. https://doi.org/10.1016/j.apgeochem.2016.05.008

Boeuf, B., Fritsch, O., 2016. Studying the implementation of the Water Framework Directive in Europe: a meta-analysis of 89 journal articles. Ecology and Society, 21(2):19. http://dx.doi.org/10.5751/ES-08411-210219

Bondu, R., Cloutier, V., Rosa, E., Roy, M., 2020. An exploratory data analysis approach for assessing the sources and distribution of naturally occurring contaminants (F, Ba, Mn, As) in groundwater from southern Quebec (Canada). Appl. Geochem. 114, 104500. https://doi.org/10.1016/j.apgeochem.2019.104500

Bouveyron, C., Brunet-Saumard, C., 2014. Model-based clustering of high-dimensional data: A review. Comput. Stat. Data Anal. 71, 52–78. https://doi.org/10.1016/j.csda.2012.12.008

BRIDGE, 2007. Background cRiteria for the IDentification of Groundwater Thresholds. https://cordis.europa.eu/project/id/6538.

Brock, G., Pihur, V., Datta, S., Datta, S. 2008. clValid: An R Package for Cluster Validation. Journal of Statistical Software 25(4). doi: 10.18637/jss.v025.i04

Buccianti and Grunsky, 2014. Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes? J. Geochem. Explor. 141, 1–5. https://doi.org/10.1016/j.gexplo.2014.03.022

Carranza, E.J.M., 2011. Analysis and mapping of geochemical anomalies using logratio-transformed stream sediment data with censored values. J. Geochem. Explor. 110, 167–185. https://doi.org/10.1016/j.gexplo.2011.05.007

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A. 2014. NbClust: an R package for determining the relevant number of clusters in data set. J Stat Soft.61:1–36.

991 Cloutier, V., Lefebvre, R., Therrien, R., Savard, M.M., 2008. Multivariate statistical
992 analysis of geochemical data as indicative of the hydrogeochemical evolution of
993 groundwater in a sedimentary rock aquifer system. J. Hydrol. 353, 294–313.
994 https://doi.org/10.1016/j.jhydrol.2008.02.015
995 Coetsiers, M., Blaser, P., Martens, K., Walraevens, K., 2009. Natural background levels
996 and threshold values for groundwater in fluvial Pleistocene and Tertiary marine
997 aquifers in Flanders, Belgium. Environ. Geol. 57, 1155–1168.
998 https://doi.org/10.1007/s00254-008-1412-z
999 Comas-Cufí M, Thió-Henestrosa S. CoDaPack 2.0: a stand-alone, multi-platform
1000 compositional software. In: Egozcue JJ, Tolosana-Delgado R, Ortego MI, eds.
1001 CoDaWork'11: 4th International Workshop on Compositional Data Analysis. Sant
1002 Feliu de Guíxols; 2011.
1003 Custodio, E.; Nieto, P.; Manzano, M. 2007. Natural groundwater quality: policy
1004 considerations and European opinion.The Natural Baseline Quality of Groundwater
1005 (eds. W.M. Edmunds & P. Shand). Blackwell Publ., Oxford. Chap. 8: 178–194.
1006 ISBN: 978–14051–5675–2.Dempster, A. P. , Laird, N. M.; Rubin, D. B. 1977.
1007 Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the
1008 Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. pp. 1-38. DOI:
1009 10.2307/2984875
1010 Drew, D., Hötzl, H. (eds.), 1999. Karst Hydrogeology and Human Activities. Impacts,
1011 Consequences and Implications. – International Contributions to hydrogeology
1012 (IAH) 20, 322 p.
1013 Ducci, D., Sellerino, M., 2012. Natural background levels for some ions in groundwater
1014 of the Campania region (southern Italy). Environ. Earth Sci. 67, 683–693.
1015 https://doi.org/10.1007/s12665-011-1516-8
1016 Egozcue J, Pawlowsky-Glahn V, Mateu-Figueras F, Barceló-Vidal C. 2003. Isometric
1017 logratio transformations for compositional data analysis. Math Geol;35:279–300.
1018 Egozcue, J.J. and Pawlowsky-Glahn, V. 2005. Groups of parts and their balances in
1019 compositional data analysis. Mathematical Geology, 37(7), 795-828.
1020 Egozcue J, Pawlowsky-Glahn V. 2006. Simplicial geometry for compositional data. In:
1021 Buccianti A, Mateu-Figueros G, Pawlowsky-Glahn V, editors. Compositional data
1022 analysis in the geosciences: from theory to practice. Bath, UK: Geological Society
1023 Publishing House; p. 67–77.

Engle, M.A., Rowan, E.L., 2013. Interpretation of Na–Cl–Br Systematics in Sedimentary Basin Brines: Comparison of Concentration, Element Ratio, and Isometric Log-ratio Approaches, Math Geosci (2013) 45:87-101 DOI 10.1007/s11004-012-9436-z

Farnham, I.M., Sinh, A.k., Stetzenbach, K.J., Johannesson, K.H., 2002. Treatment of nondetects in multivariate analysis of groundwater geochemistry data. Chemometrics and Intelligent Laboratory Systems. Volume 60, Issues 1–2, 28 pp 265-281. https://doi.org/10.1016/S0169-7439(01)00201-5

Filzmoser P, Steiger B. 2009. StatDA: statistical analysis for environmantel data. R package version 1.1. http://cran.at.r-project.org/web/packages/StatDA/index.html.

Filzmoser P, Hron K, Reimann C. 2009b Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci Total Environ 407:6100–8. https://doi.org/10.1016/j.scitotenv.2009.08.008

Filzmoser, P., Hron, K., & Templ, M. 2018. Applied Compositional Data Analysis. With Worked Examples in R. Springer Series in Statistics. doi:10.1007/978-3-319-96422-5

Fraley C. and Raftery A. E. 2002. Model-based clustering, discriminant analysis and density estimation, Journal of the American Statistical Association, 97/458, pp. 611-631

Fraley C., Raftery A. E., Murphy T. B. and Scrucca L. 2012. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report No. 597, Department of Statistics, University of Washington

Gabriel, K.R., 1971. The biplot-graphic display of matrices with application to principal component analysis. Biometrika 58, 453e467.

Goldscheider, N. 2005. Fold structure and underground drainage pattern in the alpine karst system Hochifen-Gottesacker. Eclogae geol. Helv. 98, 1–17. https://doi.org/10.1007/s00015-005-1143-z

Güller, C., Thyne, G.D. 2004. Delineation of hydrochemical facies distribution in a regional groundwater system by means of fuzzy c- means clustering. Water Resour Res 40:W12503. https://doi.org/10.1029/2004WR003299

He Kim, S.H., Choi, B., Lee, G., Yun, S., Kim S. 2019. Compositional data analysis and geochemical modeling of CO2–water–rock interactions in three provinces of

Korea. Environ Geochem Health 41, 357–380. https://doi.org/10.1007/s10653-017-0057-9

Herms, I., Jódar, J., Soler, A., Vadillo, I., Lambán, L. J., Martos-Rosillo, S., Núñez, J.A., Arnó, G., Jorge, J., 2019. Contribution of isotopic research techniques to characterize high-mountain-Mediterranean karst aquifers: The Port del Comte (Eastern Pyrenees) aquifer. Science of The Total Environment, 656, 209-230. https://doi.org/10.1016/j.scitotenv.2018.11.188

Hinsby, K., Condesso de Melo, M.T., Dahl, M., 2008. European case studies supporting the derivation of natural background levels and groundwater threshold values for the protection of dependent ecosystems and human health. Sci. Total Environ. 401, 1–20. https://doi.org/10.1016/j.scitotenv.2008.03.018

ICGC, 2007. Mapa Geològic Comarcal de Catalunya 1:50,000. Full Alt Urgell (BDGC50M).http://www.icgc.cat/ca/Administracio-i-empresa/Descarregues/Cartografia-geologica-i-geotematica/Cartografia-geologica/Mapa-geologic-comarcal-de-Catalunya-1-50.000/Mapa-geologic-comarcal-de-Catalunya-1-50.000.

Kazakis, N., Chalikakis, K., Mazzilli, N., Ollivier, C., Manakos, A., Voudouris, K. 2018. Management and research strategies of karst aquifers in Greece: Literature overview and exemplification based on hydrodynamic modelling and vulnerability assessment of a strategic karst aquifer, Sci. Total Environ. 643, 592-609, https://doi.org/10.1016/j.scitotenv.2018.06.184.

Kim, K.-H., Yun, S.-T., Park, S.-S., Joo, Y., Kim, T.-S., 2014. Model-based clustering of hydrochemical data to demarcate natural versus human impacts on bedrock groundwater quality in rural areas, South Korea. J. Hydrol. 519, 626–636. https://doi.org/10.1016/j.jhydrol.2014.07.055

Kim, K.-H., Yun, S.-T., Kim, H.-K., Kim, J.-W., 2015. Determination of natural backgrounds and thresholds of nitrate in South Korean groundwater using model-based statistical approaches. J. Geochem. Explor. 148, 196–205. https://doi.org/10.1016/j.gexplo.2014.10.001

Kresic, N., Stevanović, Z., 2010. Groundwater hydrology of springs: engineering, theory, management, and sustainability. Butterworth-Heinemann, Oxford

Lambán, L. J., Jódar, J., Custodio, E., Soler, A., Sapriza, G., & Soto, R. (2015). Isotopic and hydrogeochemical characterization of high-altitude karst aquifers in complex geological settings. The Ordesa and Monte Perdido National Park (Northern Spain)

case study. Science of the Total Environment, 506, 466-479. http://dx.doi.org/10.1016/j.scitotenv.2014.11.030

Luetscher, M., Perrin, J. 2005. The Aubonne karst aquifer (Swiss Jura). Eclogae Geologicae Helvetiae, 98(2), 237-248. https://doi.org/10.1007/s00015-005-1156-7

Marandi, A., Karro, E., 2008. Natural background levels and threshold values of monitored parameters in the Cambrian-Vendian groundwater body, Estonia. Environ. Geol. 54, 1217–1225. https://doi.org/10.1007/s00254-007-0904-6

Marín, A.I., Andreo, B., 2015. Vulnerability to Contamination of Karst Aquifers. In: Stevanović Z. (eds) Karst Aquifers—Characterization and Engineering. Professional Practice in Earth Sciences. Springer, Cham. https://doi.org/10.1007/978-3-319-12850-4_8

Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V., 2003. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Mathematical Geology 35 (3), 253–278.

Merchán, D., Auqué, L.F., Acero, P., Gimeno, M.J., Causapé, J., 2015. Geochemical processes controlling water salinization in an irrigated basin in Spain: Identification of natural and anthropogenic influence. Sci. Total Environ. 502, 330–343. https://doi.org/10.1016/j.scitotenv.2014.09.041

Moya, C.E., Raiber, M., Taulis, M., Cox, M.E., 2015. Hydrochemical evolution and groundwater flow processes in the Galilee and Eromanga basins, Great Artesian Basin, Australia: A multivariate statistical approach. Sci. Total Environ. 508, 411–426. https://doi.org/10.1016/j.scitotenv.2014.11.099

Müller, D., Blum, A., Hart, A., Hookey, J., Kunkel, R., Scheidleder, A., Tomlin, C., Wendland, F., 2006. Final proposal for a methodology to set up groundwater threshold values in Europe. In: Report to the EU project ‘‘BRIDGE’’, Deliverable D18.

Muñoz, J.A., Mencos, J., Roca, E., Carrera, N., Gratacós, A., Ferrer, O. Fernández, O. 2018. The structure of the South-Central-Pyrenean fold and thrust belt as constrained by subsurface data. Geologica Acta, Vol.16, No 4, 439-460 DOI: 10.1344/GeologicaActa2018.16.4.7

Nieto, P., Custodio, E., Manzano, M., 2005. Baseline groundwater quality: a European approach. Environmental Science & Policy, 8(4), 399-409. https://doi.org/10.1016/j.envsci.2005.04.004

Otero, N., Tolosana-Delgado, R., Soler, A., Pawlowsky-Glahn, V., Canals, A., 2005. Relative vs. absolute statistical analysis of compositions: A comparative study of surface waters of a Mediterranean river. Water Res. 39, 1404–1414. https://doi.org/10.1016/j.watres.2005.01.012

Owen, D.Des.R., Pawlowsky-Glahn, V., Egozcue, J.J., Buccianti, A., Bradd, J.M., 2016. Compositional data analysis as a robust tool to delineate hydrochemical facies within and between gas-bearing aquifers: COMPOSITIONAL DATA ANALYSIS TO DELINEATE WATER TYPES. Water Resour. Res. 52, 5771–5793. https://doi.org/10.1002/2015WR018386

Palarea-Albaladejo J, Martin-Fernandez JA, Olea, RA. A bootstrap estimation scheme for chemical compositional data with nondetects. Journal of Chemometrics 2014; 28: 585-599. https://doi.org/10.1002/cem.2621

Palarea-Albaladejo and Martín-Fernández, 2015. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems, Volume 143, pp 85-96, ISSN 0169-7439, https://doi.org/10.1016/j.chemolab.2015.02.019.

Parkhurst, D.L., Appelo, C.A.J. 2013. Description of Input and Examples for PHREEQC Version 3—A Computer Program for Speciation, Batch-Reaction, One-Dimensional Transport, and Inverse Geochemical Calculations: U.S. Geological Survey Techniques and Methods, Book 6, Chap. A43: 1–497. (Available only at https://pubs.usgs.gov/tm/06/a43. Last access 28 August 2020).

Parrone, D., Ghergo, S., Preziosi, E., 2019. A multi-method approach for the assessment of natural background levels in groundwater. Sci. Total Environ. 659, 884–894. https://doi.org/10.1016/j.scitotenv.2018.12.350

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R. 2015. Modeling and Analysis of Compositional Data. ed. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom. 272 pages. ISBN: 9781118443064

Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen–Geiger climate classification. Hydrol. Earth Syst. Sci. 11, 1633–1644. https://doi.org/ 10.5194/hess-11-1633-2007.

Piña, A., Donado, L.D., Blake, S., Cramer, T., 2018. Compositional multivariate statistical analysis of the hydrogeochemical processes in a fractured massif: La

Línea tunnel project, Colombia. Appl. Geochem. 95, 1–18. https://doi.org/10.1016/j.apgeochem.2018.05.012

Preziosi, E., Giuliano, G., Vivona, R., 2010. Natural background levels and threshold values derivation for naturally As, V and F rich groundwater bodies: a methodological case study in Central Italy. Environ. Earth Sci. 61, 885–897. https://doi.org/10.1007/s12665-009-0404-y

Puig, R., Tolosana-Delgado, R., Otero, N., Folch, A., 2011. Combining isotopic and compositional data: a discrimination of regions prone to nitrate pollution. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), Compositional Data Analysis: Theory and Applications 390.

Raftery, AE., Scrucca, L., Brendan, T., Fop, M. 2020. Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation. Package 'mclust'. Version 5.4.6. https://mclust-org.github.io/mclust/

Reimann, C., Filzmoser, P. 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environmental Geology 39, 1001–1014. https://doi.org/10.1007/s002549900081

Reimann C, Filzmoser P, Garrett RG, Dutter R. 2008. Statistical data analysis explained. Applied environmental statistics with R. Wiley, Chichester, UK, 362 pp. ISBN: 978-0-470-98581-6

Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., 2012. The concept of compositional data analysis in practice — Total major element concentrations in agricultural and grazing land soils of Europe. Sci. Total Environ. 426, 196–210. https://doi.org/10.1016/j.scitotenv.2012.02.032

Sánchez, D., Antonio Barberá, J., Mudarra, M., Andreo, B., & Martín, J. F. 2017. Hydrochemical and isotopic characterization of carbonate aquifers under natural flow conditions, Sierra Grazalema Natural Park, southern Spain. Geological Society, London, Special Publications, 466(1), 275–293. doi:10.1144/sp466.16

Scrucca, L. 2010. Dimension reduction for model-based clustering. Statistics and Computing, 20(4), pp. 471-484.

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2016. Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models, The R Journal, 8/1, pp. 289-317.

Serianz, L., Cerar, S. & Šraj, M. 2020. Hydrogeochemical characterization and determination of natural background levels (NBL) in groundwater within the main lithological units in Slovenia. Environ Earth Sci 79, 373. https://doi.org/10.1007/s12665-020-09112-1

Shelton, J.L., Engle, M.A., Buccianti, A., Blondes, M.S., 2018. The isometric log-ratio (ilr)-ion plot: A proposed alternative to the Piper diagram. J. Geochem. Explor. 190, 130–141. https://doi.org/10.1016/j.gexplo.2018.03.003

Stevanović, Z., 2019. Karst waters in potable water supply: a global scale overview. Environ Earth Sci 78, 662. https://doi.org/10.1007/s12665-019-8670-9

Struckmeier WF, Gilbrich WH, Gun Jvd, Maurer S, Puri S, Richts A, Winter P, Zaepke M. 2006. WHYMAP and the groundwater resources map of the world at the scale of 1:50 000 000. Special edition for the 4th world water forum, Mexico City, March 2006. BGR Hannover/UNESCO, Paris.

Suk, H., and K.K. Lee. 1999. Characterization of a groundwater hydrochemical system through multivariate analysis: clustering into groundwater zones. Groundwater v. 37, no. 3pp. 358-366.

Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data: Problems and possibilities. Appl. Geochem. 23, 2198–2213. https://doi.org/10.1016/j.apgeochem.2008.03.004

Tolosana-Delgado, R., Otero, N., Soler, A. 2005. A compositional approach to stable isotope data analysis. Conference: CODAWORK'05

Van den Boogaart, K.G. and Tolosana-Delgado, R. (2008) "compositions": a unified R package to analyze Compositional Data, Computers & Geosciences. 34 (4), 320-338. https://doi.org/10.1016/j.cageo.2006.11.017

Vergés, J. 1999. Estudi geològic del vessant sud del Pirineu oriental i central. Evolució cinemàtica en 3D. PhD Thesis. University of Barcelona (UB), Faculty of Geology, 180 pp.

Viviroli, D., Kummu, M., Meybeck, M., Kallio, M., Wada, Y. 2020. Increasing dependence of lowland populations on mountain water resources. Nature Sustainability, 1-12. https://doi.org/10.1038/s41893-020-0559-9

Wendland, F., Blum, A., Coetsiers, M., Gorova, R., Griffioen, J., Grima, J., Hinsby, K., Kunkel, R., Marandi, A., Melo, T., Panagopoulos, A., Pauwels, H., Ruisi, M., Traversa, P., Vermooten, J.S.A., Walraevens, K., 2008. European aquifer typology: a practical framework for an overview of major groundwater composition at

European scale. Environ. Geol. 55, 77–85. https://doi.org/10.1007/s00254-007-0966-5

WFD 2000, Water Framework Directive, 2000. Directive 2000/60/CE of the European Parliament (ECOJ 22 December 2000). http://www.bygg.ntnu.no/borsanyi/eamn_web/documents/wfd-es.pdf.

Wu, X., Zheng, Y., Zhang, J., Wu, B., Wang, S., Tian, Y., Li, J., Meng, X., 2017. Investigating Hydrochemical Groundwater Processes in an Inland Agricultural Area with Limited Data: A Clustering Approach. Water 9, 723. https://doi.org/10.3390/w9090723

Yidana, S.M. 2010. Groundwater classification using multivariate statistical methods: Southern Ghana. Journal of African Earth Sciences 57(5):455-469 doi: 10.1016/j.jafrearsci.2009.12.002

Yolcubal, İ., Gündüz, Ö.C.A., Kurtuluş, N., 2019. Origin of salinization and pollution sources and geochemical processes in urban coastal aquifer (Kocaeli, NW Turkey). Environmental Earth Sciences, 78(6), 181. https://doi.org/10.1007/s12665-019-8181-8

Zabala, M.E., Martínez, S., Manzano, M., Vives, L., 2016. Groundwater chemical baseline values to assess the Recovery Plan in the Matanza-Riachuelo River basin, Argentina. Sci. Total Environ. 541, 1516–1530. https://doi.org/10.1016/j.scitotenv.2015.10.006

Zwahlen, F. (ed.), 2004. Vulnerability and risk mapping for the protection of carbonate (karst) aquifers, final report COST Action 620. EUR 20912, European Commission, Brussels, 297 p.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

The authors: Ignasi Herms. Jorge Jódar, Albert Soler, Luis Javier Lambán, Emilio Custodio, Joan Agustí Núñez, Georgina Arnó, Maribel Ortego, David Parcerisa and Joan Jorge