Vol. 22, No. 2(2023) IA235
Revista Mexicana de Ingeniería Química

Comprehensive assessment of groundwater quality in Mexico and application of new water classification scheme based on machine learning

Evaluación integral de la calidad de las aguas subterráneas en México y aplicación de un nuevo esquema de clasificación del agua basado en el aprendizaje automático

L. Díaz-González[1]*, M. Rosales-Rivera[2], L.A. Chávez-Almazán[3]
[1] *Centro de Investigación en Ciencias, Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos 62209, México.*
[2] *Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, 62209, México.*
[3] *Unidad de Innovación Clínica y Epidemiológica del Estado de Guerrero, Acapulco, Guerrero, 39715, México.*
Received: April 18, 2023; Accepted: June 26, 2023

Abstract
This study conducted a comprehensive evaluation of groundwater quality at 1,068 monitoring sites across all hydrologic-administrative regions in Mexico. Based on the analysis of 14 physicochemical and microbiological parameters, which include fluorides, fecal coliforms, nitrate-nitrogen, arsenic, cadmium, chromium, mercury, lead, manganese, iron, alkalinity, conductivity, water hardness, and total dissolved solids, it was found that 41% of the sites exhibited good water quality. Additionally, 23% of the sites presented regular water quality, while 36% of the sites showed poor water quality. Sites with good water quality exhibited lower concentrations of major ions (Ca, Mg, Na, K, SO_4, Cl, and HCO_3) compared to sites with regular and poor water quality. Water nomenclature was also estimated using the VL model based on Support Vector Machines with linear kernel, statistical techniques, and Monte Carlo simulation. This model classified 87% of the monitoring sites into four basic water classes: Na HCO_3 (47%); Na Cl (18%); Ca HCO_3 (17%); and Na SO_4 (5%). Furthermore, the t-SNE computational algorithm was applied to reduce the dimensionality of the data and visualize it in a 2D plot; in this context, the data corresponds to the chemical concentrations of major ions and contaminants. This algorithm obtained a clustering consistent with the water nomenclature estimated by the VL model. The contaminant study results revealed that all hydrologic-administrative regions presented at least one physicochemical-microbiological parameter that exceeded the acceptable levels defined by regulations of Mexico. Therefore, the implementation of environmental sanitation strategies is crucial to ensure the availability of high-quality water resources that are safe for human health.
Keywords: Support Vector Machine, Gradient Boosting, Log-ratio transform, Hill-Piper diagram, Visualization 2D t-SNE.

Resumen
Este estudio realizó una evaluación integral de la calidad del agua subterránea en 1,068 sitios de monitoreo en todas las regiones hidrológico-administrativas de México. Según el análisis de 14 parámetros fisicoquímicos y microbiológicos, que incluyen fluoruros, coliformes fecales, nitrato-nitrógeno, arsénico, cadmio, cromo, mercurio, plomo, manganeso, hierro, alcalinidad, conductividad, dureza del agua y sólidos disueltos totales, se encontró que el 41% de los sitios exhibieron agua de buena calidad. Adicionalmente, el 23% de los sitios presentaron agua de calidad regular, mientras que el 36% de los sitios mostraron agua de mala calidad. Los sitios con buena calidad de agua presentaron menores concentraciones de los iones mayores (Ca, Mg, Na, K, SO_4, Cl y HCO_3) en comparación con los sitios con calidad de agua regular y mala. También se estimó la nomenclatura del agua utilizando el modelo VL basado en Máquinas de Vectores de Soporte con kernel lineal, técnicas estadísticas y simulación Monte Carlo. Este modelo clasificó el 87% de los sitios de monitoreo en cuatro clases básicas de agua: Na HCO_3 (47%); Na Cl (18%); Ca HCO_3 (17%); y Na SO_4 (5%). Además, se aplicó el algoritmo computacional t-SNE para reducir la dimensión de los datos y visualizarlos en un gráfico 2D; en este contexto, los datos corresponden a concentraciones químicas de los iones mayoritarios y contaminantes. Este algoritmo obtuvo un agrupamiento coherente con la nomenclatura del agua estimada por el modelo VL. Los resultados del estudio de contaminantes revelaron que todas las regiones hidrológico-administrativas presentaron al menos un parámetro fisicoquímico-microbiológico que excedió los niveles aceptables definidos por la normatividad de México. Por lo tanto, la implementación de estrategias de saneamiento ambiental es crucial para garantizar la disponibilidad de recursos hídricos de alta calidad que sean seguros para la salud humana.
Palabras clave: Support Vector Machine, Gradient Boosting, Log-ratio transform, Hill-Piper diagram, Visualización 2D t-SNE.

* *Corresponding author. E-mail*: ldg@uaem.mx
https://doi.org/10.24275/rmiq/IA235
ISSN:1665-2738, *issn-e*: 2395-8472

1 Introduction

Water is an essential natural resource for the survival of organisms and ecosystems. However, it is susceptible to changes produced by anthropogenic activities and natural factors, making water quality management crucial. Monitoring physicochemical and microbiological parameters of water is essential to ensure its adequacy. Hydrogeochemical assessment and groundwater classification can contribute to sustainable development and mitigate potential adverse health consequences in the future (Amiri and Nakagawa, 2021).

Water classification plays a valuable role in offering initial insights into the complex hydrochemical mechanisms occurring beneath the surface (Kumar, 2013). However, many of these approaches (e.g., Hill, 1940; Piper, 1944; Durov, 1948; Handa, 1965; Chadha, 1999; Güler *et al*., 2002; Ahmad *et al*., 2003; Ray and Mukherjee, 2008; Giménez-Forcada, 2010; Al-Bassam and Khalil, 2012; Teng *et al*., 2016; Elhag, 2017; Shelton *et al*., 2018; Pérez-Espinosa, 2019) rely on the traditional Hill-Piper diagram (Hill, 1940; Piper, 1944), which utilizes two ternary diagrams based on the normalized concentrations (mM) of 4 cations (Ca, Mg, (Na + K)) and 4 anions (SO_4, Cl, (HCO_3+CO_3)) to identify five water types (Díaz-González *et al*., 2021).

For more than a century, geochemistry has made use of ternary diagrams to illustrate compositional variability among components of a geochemical sample set and to draw information from the spatial positioning of the points on the diagram. However, it is now widely documented that ternary diagrams suffer from issues of distortion and errors amplification-reduction, produced by closure and constant sum problems (Butler, 1979; Aitchison, 1986; Verma, 2015). Further, mixing trends in ternary diagrams have been wrongly considered to be straight lines, but show variations in thickness along the mixing curve when accounting for analytical errors (Verma, 2015; Verma, *et al*., 2021).

In response to these limitations, Verma *et al*. (2021) proposed a multidimensional model (7-hlr) for water classification, which utilizes linear discriminant analysis, canonical analysis, and hybrid log-ratio transformations. Subsequently, Díaz-González *et al*. (2021) introduced four new machine learning models (CB, VL, VP, VR) and compared them with the 7-hlr model proposed by Verma *et al*. (2021). These new models, based on CatBoost and Support Vector Machines (SVM), outperformed the 7-hlr model, demonstrating higher classification accuracy (Díaz-González *et al*., 2021). Recently, to facilitate the application and comparative analysis of these models, Díaz-González *et al*. (2022) developed in Python a freely available computer program called WCSystem (Water Classification System).

In this study, 1,068 groundwater samples collected from monitoring sites across 13 Mexican hydrologic-administrative regions reported by the National Water Commission of Mexico (CONAGUA) were processed using WCSystem software, revealing the prevalence of basic and hybrid water types. Overcoming the limitations of traditional ternary diagrams and introducing new models and computational tools, this study contributes to a more accurate water classification and assessment of groundwater quality in Mexico, providing valuable insights for groundwater resource management and sustainable development.

2 Water classification models encapsulated in the WCSystem program

2.1 *Training and validation databases of water classification models*

Training (50,000 samples) and validation (8,000 samples) databases were generated through Monte Carlo simulations of ionic charge-balanced concentrations of 8 ions (Ca, Mg, Na, K, SO_4, Cl, HCO_3, and CO_3; mM). The methodology for generating these databases was extensively described by Verma *et al*. (2021) and Díaz-González *et al*. (2021). The simulation procedure can be summarized as follows:

1. **Monte Carlo simulation procedure**: for each ion (Ca, Mg, Na, K, SO_4, Cl, HCO_3, and CO_3), uniformly distributed values IID U(0,1) were simulated to generate the training and validation datasets, using the Mersenne Twister pseudo-random number generator algorithm (Matsumoto and Nishimura, 1996; Law and Kelton, 2000; Verma and Quiroz-Ruiz, 2006). These values were scaled to range of 0-100 (Fig. 1a) to ensure the representability of the ternary diagrams. Fig. 1 provides schematic representations of the training database, including ternary diagrams (Fig. 1a-b), a 2D plot (Fig. 1c-d), and a 3D plot (Fig. 1e-f) for cation and anion data. Furthermore, Fig. 2a presents a histogram of 8 variables for a training database, which are uniformly distributed values U(0,1) and scaled to the range of 0-100.

2. **Ionic charge-balance (ICB) validation procedure**: for each simulated water sample, the ionic charge-balance (ICB; Nicholson, 1993) was calculated as follows: $ICB = \frac{|\sum_i^n cations + \sum_i^n anions|}{|\sum_i^n cations - \sum_i^n anions|}$, where i = 1 to 4; cations and anions are expressed in mM units. An exacting threshold of 0.00005% was defined as the permissible maximum unbalance, therefore when the unbalance was greater than this value, an unbalance factor ($F = \frac{|\sum_i^n cations|}{|\sum_i^n anions|}$) was calculated and multiplied by a pseudo-random increment ranging from 0 to 10%, which was applied for each ion (Ca, Mg, Na, K, SO_4, Cl, HCO_3, and CO_3). This procedure was applied iteratively until the sample was balanced, enabling the generation of samples with $ICB < \pm 0.00005\%$. Figure 2b shows a histogram of the training database after the ICB validation process.

3. **Initial assignment for 16 balanced classes**: the initial assignment for 16 balanced classes (with a minimum of 3021 samples and a maximum of 3247 samples) involved utilizing the concept of Greater Molar Concentration (GMC) for each cation and anion. These classes represent the cross-

combinations of four cations (Ca, Mg, Na,K) and four anions (SO_4, Cl, HCO_3,CO_3) classes.

4. **Hybrid log-ratio (hlr) transformation**: The *hlr* of the molar concentrations of 8 ions (Ca, Mg, Na, K, SO_4, Cl, HCO_3, and CO_3) was calculated as follows: $hlr_{(i+1)} = \ln\left(\frac{g(x_i,\ldots,x_n)^{1/n}}{x_{(i+1)}}\right)$, $i = 1,2,\ldots,(n-1)$, where $g()$ represents the geometric mean, x_i represents the concentration of each ion in the same order ($Ca\times Mg\times Na\times K\times SO_4\times Cl\times HCO_3\times CO_3)^{1/8}$, $x_{(i+1)}$ denotes one ion at a time from second (Mg) to last (CO_3), and n is the total number of ions ($n = 8$). Based on this equation, seven hlr variables (hlr_2 to hlr_8) were calculated as follows: $hlr_2=\ln(gm/Mg)$, $hlr_3=\ln(gm/Na)$, $hlr_4=\ln(gm/K)$, $hlr_5=\ln(gm/SO_4)$, $hlr_6=\ln(gm/Cl)$, $hlr_7=\ln(gm/HCO_3)$, and $hlr_8=\ln(gm/CO_3)$; where $gm=(Ca\times Mg\times Na\times K\times SO_4\times Cl\times HCO_3\times CO_3)^{1/8}$ and represents the geometric mean of 8 ions. According to Aitchison and Egozcue (2005), the utilization of the geometric mean treats the components symmetrically and provides a reasonable approach to quantify the interdependence among the parts. Finally, a histogram of the *hlr* transformations is presented in Fig. 2c.

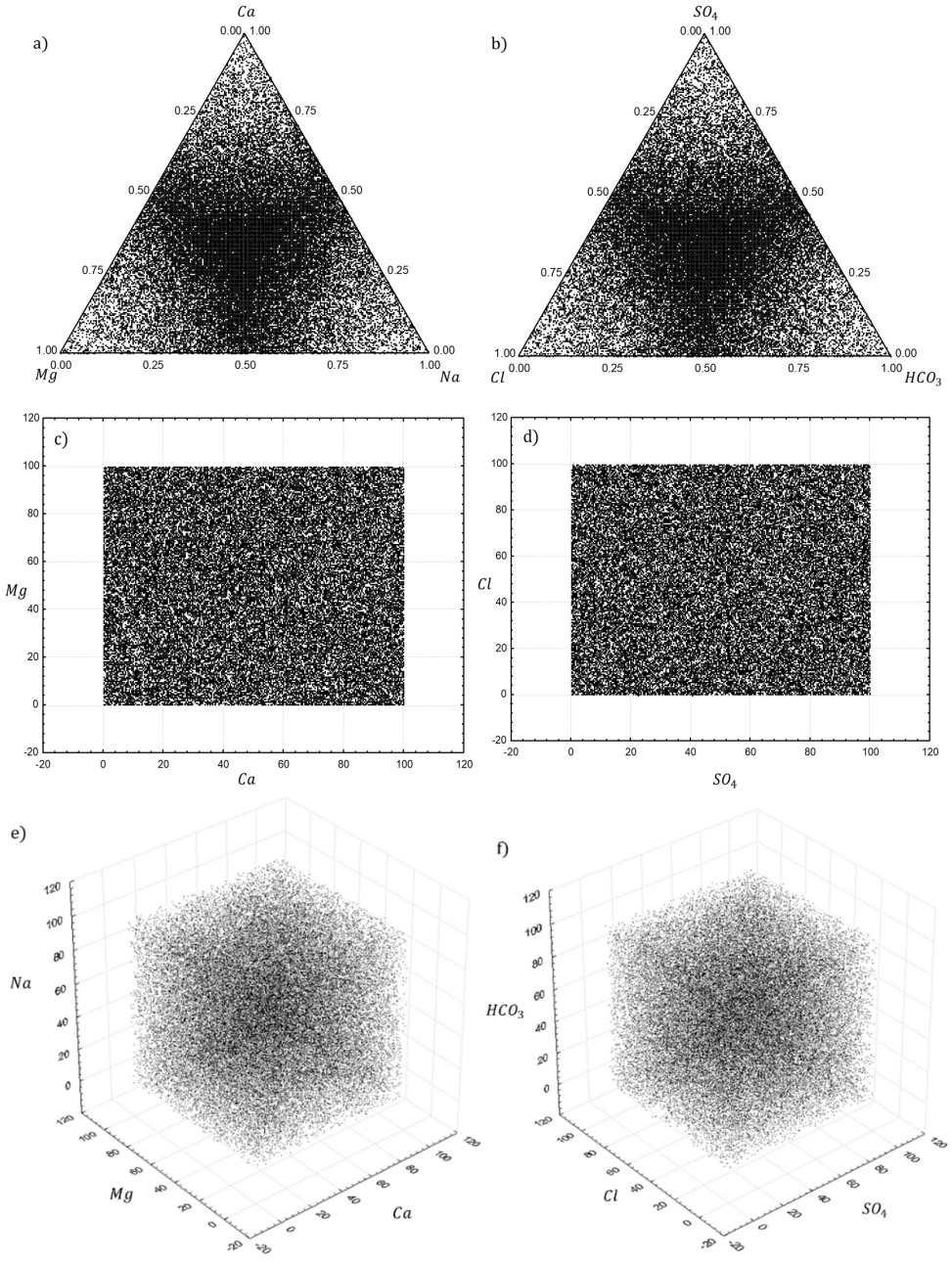

Fig. 1. Schematic representation (modified of Díaz-González *et al.* 2021) of the training database (50,000 simulated samples): (a-b) cation (Ca, Mg, and Na) and anion (SO_4, Cl, and HCO_3) data ternary diagrams; (c-d) cation (Ca and Mg) and anion (SO_4, and Cl) data 2D plot; (e-f) cation (Ca, Mg, and Na) and anion (SO_4, Cl, and HCO_3) data 3D plot.

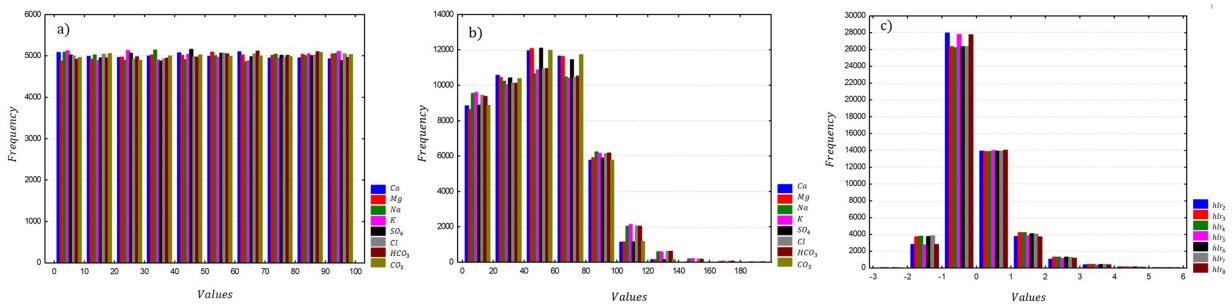

Fig. 2. Histograms (modified of Díaz-González *et al*. 2021) of simulated variables (50,000 samples); (a) data generated of the distribution U(0,1) multiplied by a scalar 100; (b) major ions data after ionic charge-balance procedure; (c) hlr_2-hlr_8 transformations of the 8 major ions.

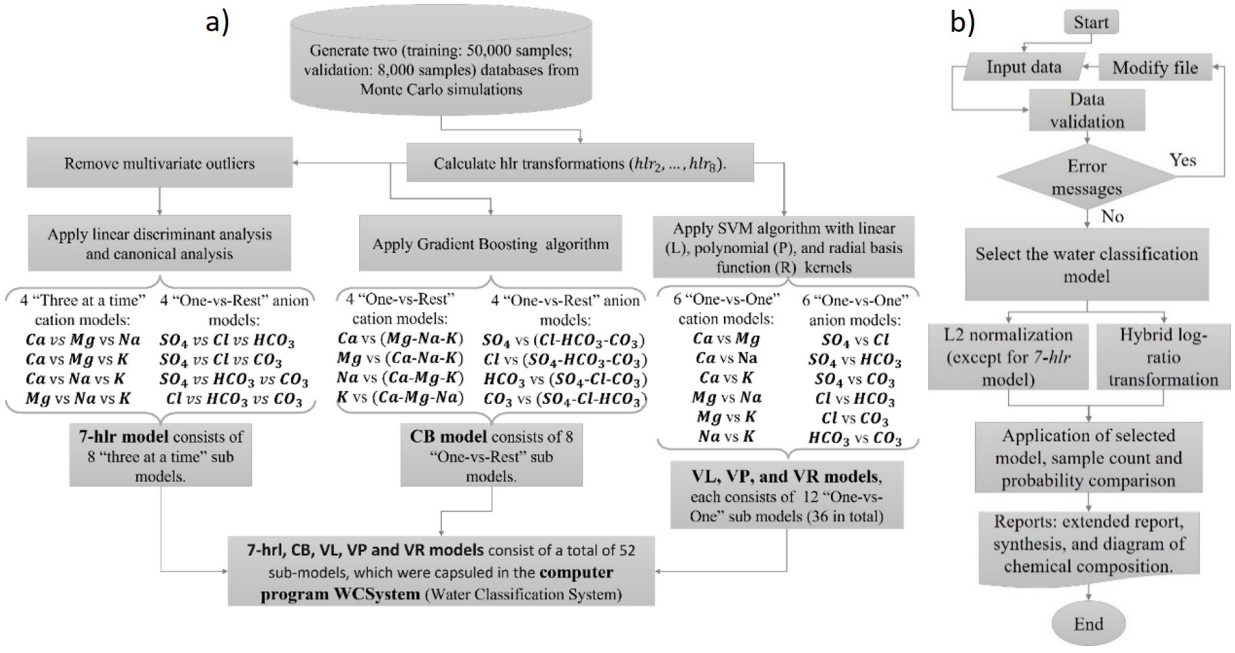

Fig. 3. Schematic diagrams (modified after Díaz-González *et al*., 2021, 2022) about the methodology and application of water classification models (7-hlr: Verma *et al*., 2021; CB, VL, VP, and VP: Díaz-González *et al*., 2021): a) flowchart of the methodology of five water classification models; b) flowchart of computer program WCSystem procedure.

2.2 *Water classification models based on machine learning*

These models were proposed by Verma *et al*. (2021) and Díaz-González *et al*. (2021) using 7 features (hlr_2 to hlr_8) for their training, and are briefly described below (see Fig. 3a).

1. **7-hlr model**: This model is based on linear discriminant and canonical analysis (LDCA) classification parametric technique, that transforms input data into a lower-dimensional space to maximize the ratio of between-class variance to within-class variance. This process makes it easier to distinguish between different classes, which in turn makes classification more effective (Tharwat *et al*., 2019). Since LDCA is parametric, it requires that the features of each class be multi-normally distributed. For this purpose, each class was depurated of multivariate discordant outliers using DOMuDaF (Discordant Outlier from Multivariate Data) program (Verma *et al*., 2016), which involved a transformation of the Wilks statistic W to F-test (Rencher, 2002). After applying this program, the final training database consisted of 46,292 outlier-free samples suitable for training LDCA.

   The 7-hlr model (Fig. 3a) consists of 8 "three at a time" LDCA classifiers (Ca-Mg-Na, Ca-Mg-K, Ca-Na-K, Mg-Na-K, SO_4-Cl-HCO_3, SO_4-Cl-CO_3, SO_4-CO_3-HCO_3, and Cl-HCO_3-CO_3). These classifiers enable the identification of 16 water types through cross combinations.

2. **CB model**: This model is based on the initial dataset (50,000 samples) and CatBoost machine learning algorithm available on CatBoost library (Prokhorenkova *et al*., 2018) for Python. CatBoost constructs decision trees sequentially, with each subsequent tree having decreased loss (Géron, 2019). The CB model consists of 8 "One-versus-Rest" binary classifiers as follows (Fig. 3a): (1) Ca vs (Mg-Na-K); (2) Mg vs (Ca-Na-K); (3) Na vs (Ca-Mg-K);

(4) K vs (Ca-Mg-Na); (5) SO_4, vs (Cl, HCO_3, and CO_3); (6) Cl, vs (SO_4, HCO_3, and CO_3); (7) HCO_3, vs (SO_4, Cl, and CO_3); and (8) CO_3, vs (SO_4, Cl, and HCO_3).

3. **VL, VP, and VR models**: These models are based on Support Vector Machines (SVM) using different kernels: linear, polynomial and the radial basis function, respectively. These models utilized an open-source SVC library for Python (Pedregosa *et al*., 2011). Each model consists of an ensemble of 12 "One-versus-One" binary classifiers as follows (Fig. 3a): (1) Ca vs Mg, (2) Ca vs Na, (3) Ca vs K, (4) Mg vs Na, (5) Mg vs K, and, (6) Na vs K, (7) SO_4, Cl, (8) SO_4 vs HCO_3, (9) SO_4 vs CO_3, (10) Cl vs HCO_3, (11) Cl vs CO_3, (12) HCO_3 vs CO_3.

The water type determined by each model is obtained by considering the probabilities associated with the competing fields in all sub-models. These models generate probability values for each of the cations or anions, enabling the determination of basic and hybrid water types. Of these probabilities, let us suppose that Pm is the highest probability and Pn is the second-highest probability. The conditions that define if the water type is basic are as follows: if ($P_m \geq 0.5$) and (($P_m - P_n$) ≥ 0.25) and ($P_n \leq 0.25$). Otherwise, a hybrid nomenclature is assigned, that is the highest probability cation or anion followed by the next highest cation or anion. Thus, for the 4 cations and anions separately, 4 basic and 12 hybrid classes can be achieved.

It should be noted that, the classification accuracy values obtained by these five models in an external validation set (8,000 simulated samples), were as follows: (i) VL model (99.8%), VP model (99.0%), VR model (98.9%), CB model (98.7%), and 7-hlr model (92.0%). It is crucial to consider this ranking to ensure the most accurate interpretation of the results of models. Thus, we only applied the VL model to simplify the presentation of the results of 1,068 sites of the groundwater network of Mexico.

2.3 *Availability, structure, and use of WCSystem software*

The WCSystem software, developed by Díaz-González *et al*. (2022), can be freely downloaded from the web portal http://tlaloc.ier.unam.mx/WCSystem. This program enables users to utilize and evaluate these new water classification models. The overall structure of WCSystem is presented in Fig. 3b. To utilize the software, the user is required to provide an input file containing the concentrations of 8 ions (Ca, Mg, Na, K, SO_4, Cl, HCO_3, and CO_3; mg/L) for each sample. If the file is error-free, the user can select one model from the "*Classification*" menu. Otherwise, the user must correct the corresponding error(s). After, the program proceeds to calculate the hybrid log-ratio transformations, applies L2 normalization (except for 7-hlr model; Pedregosa *et al*., 2011), calculate the discriminant functions associated with the chosen model, and calculates the probability of assigning each sample to a specific type of basic or hybrid water. WCSystem program generates three output files for each model: (1) The first file is an extended report that includes concentrations (mg/L and mM), hybrid log-ratio transformations (*hlr*2 to *hlr*8), probability values, and basic and hybrid water types for each sample; (2) The second file is a brief report that provides a count of samples assigned to certain water types (basic and hybrid), thus offering insights into the most probable water nomenclature within dataset; and (3) The third file is a diagram representing the chemical composition of processed samples, using a logarithmic scale on the Y-axis and mM units.

3 Groundwater monitoring sites of Mexico

Groundwater can be found in the saturated zone of the subsoil and moves slowly from places with high elevation and pressure to places with low elevation and pressure, such as rivers and lakes.

3.1 *Groundwater quality assessment*

Adequate monitoring is crucial for groundwater quality assessment and requires a representative monitoring network. In Mexico, the groundwater bodies are monitored by the National Water Commission of Mexico (CONAGUA). In 2020, a total of 1,068 monitoring sites were analyzed, with 98% of them being wells and the remaining 2% comprising cenotes, springs, and discharges. The analysis considered 14 physicochemical and microbiological parameters, including fluorides (F), fecal coliforms (F_C), nitrate-nitrogen (NO_3-N), arsenic (As), cadmium (Cd), chromium (Cr), mercury (Hg), lead (Pb), manganese (Mn), iron (Fe), alkalinity (Alk), conductivity (Cond), water hardness (W_H), and total dissolved solids (TDS). Based on these parameters, three types (as a three-color traffic light) of groundwater quality were established (Fig. 4a), considering the official Mexican regulations (DOF, 2022). Detailed information about the groundwater quality parameters of Mexico is available at: https://www.gob.mx/conagua/articulos/calidad-del-agua. The groundwater quality assessment of monitoring sites reveals the following findings:

1. **Good quality**: 41% of the sites were classified as having good quality (green color). These sites showed compliance with the acceptable limits of water quality for all 14 parameters analyzed.
2. **Regular quality**: 23% of the sites were categorized as having regular quality (yellow color). These sites showed non-compliance in one or more of the following parameters: alkalinity, conductivity, hardness, total dissolved solids, manganese, or iron.
3. **Poor quality**: 36% of the sites were classified as having poor quality (red color). These sites exhibited non-compliance in one or more of the following parameters: fluorides, fecal coliforms, nitrate-nitrogen, arsenic, cadmium, chromium, mercury, and lead.

Fig. 4a illustrates a map of the 1,068 groundwater quality monitoring sites across 13 Mexican hydrological-administrative regions, which are referred to as "regions"

for simplicity. The percentage distribution of monitoring sites in each region is as follows: *Centrales* (9%), *Península de Baja California* (8%), *Río Bravo* (6%), *Balsas* (6%), *Pacífico Norte* (6%), *Golfo Norte* (5%), *Aguas del Valle de México* (4%), *Frontera Sur* (3%), *Golfo Centro* (2%), *Pacífico Sur* (1%). Notably, half of the monitoring sites are concentrated in *Centrales del Norte*, *Lerma Santiago Pacífico*, and *Península de Yucatán* regions.

Fig. 4b presents a histogram of groundwater quality for these monitoring sites from Mexico. More than 40% of the sites in the following regions exhibited poor quality: *Centrales del Norte* (57%, 132 of 232), *Río Bravo* (55%, 36 of 65), *Pacífico Norte* (48%, 30 of 62), *Pacífico Sur* (44%, 7 of 16), and *Lerma Santiago Pacífico* (42%, 71 of 170). Meanwhile, *Península de Yucatán* (59%, 74 of 125) and *Aguas del Valle de México* (45%, 17 of 38) regions presented regular quality. Finally, *Balsas* (64%, 44 of 69), *Frontera Sur* (59%, 20 of 34), *Noroeste* (57%, 54 of 94), *Lerma Santiago Pacífico* (45%, 77 of 170), and *Golfo Centro* (43%, 9 of 21) regions demonstrated good quality.

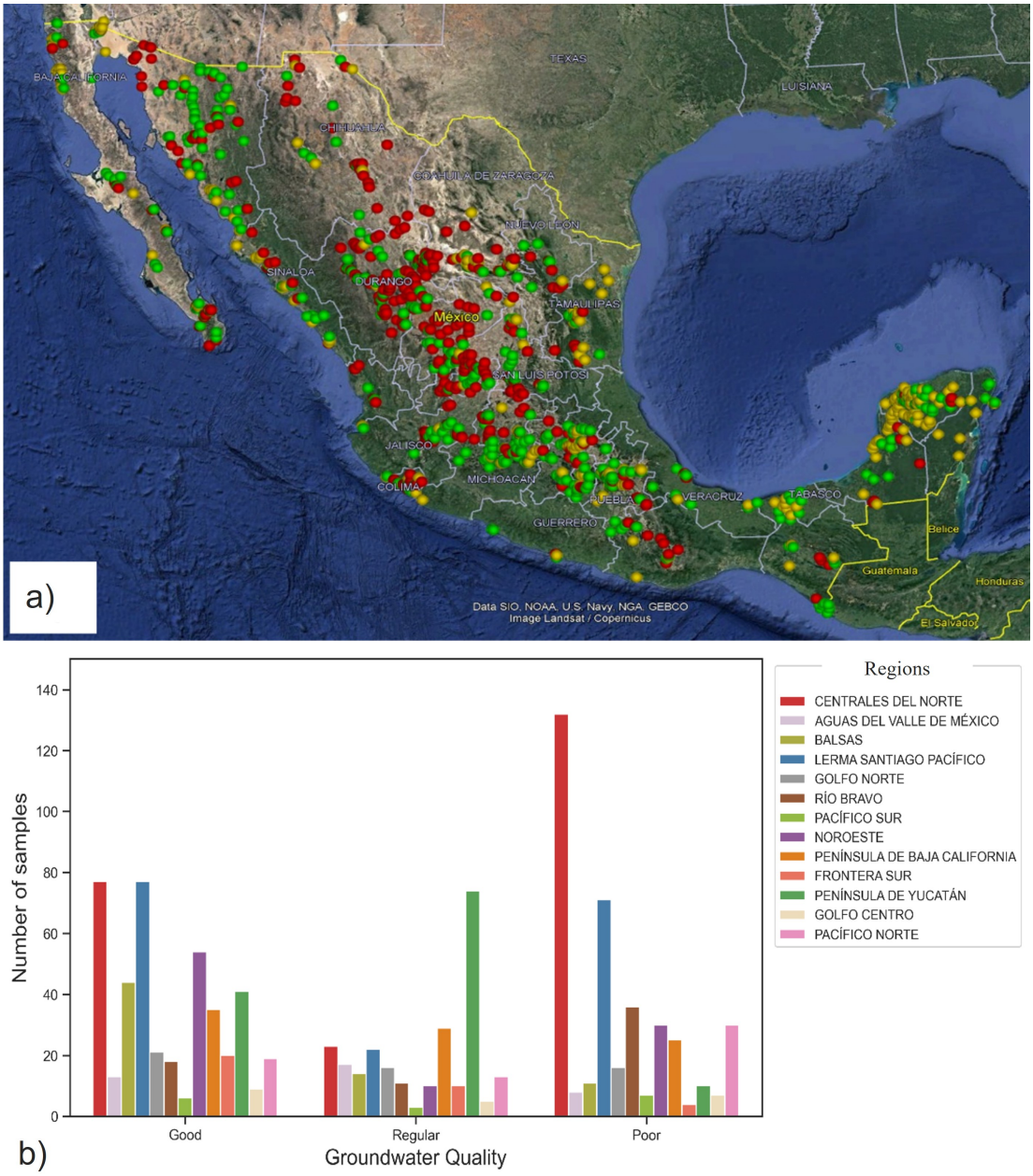

Fig. 4. Distribution of groundwater quality of 1,068 monitoring sites of Mexico sampled in 2020 by National Water Commission of Mexico (CONAGUA, 2022). a) Mexico map; the color indicates the quality of the groundwater: (i) green color indicates good quality, (ii) yellow color indicates regular quality, and (iii) red color indicates poor quality; b) Histogram of 1,068 sites grouped by groundwater quality level and categorized by hydrological-administrative regions of Mexico.

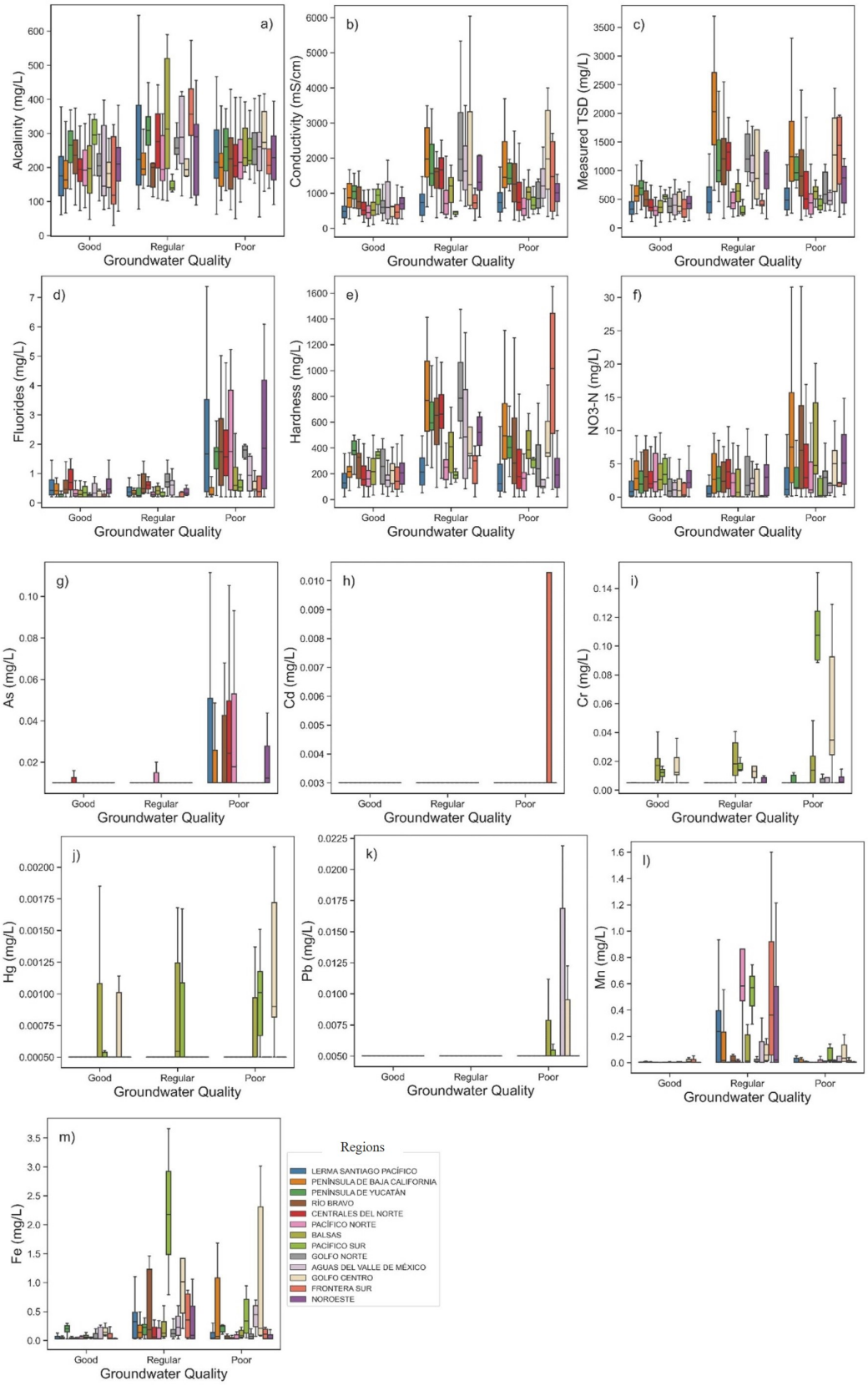

Fig. 5. Histogram of concentrations of physicochemical-microbiological parameters of 1,068 sites grouped by groundwater quality level and categorized by hydrological-administrative regions of Mexico: alkalinity (Alk), conductivity (Cond), total dissolved solids (TDS), fluorides (F), water hardness (W_H), nitrate-nitrogen (NO_3-N), arsenic (As), cadmium (Cd), chromium (Cr), mercury (Hg), lead (Pb), manganese (Mn), and iron (Fe).

3.1.1 Statistical analysis of physicochemical-microbiological parameters and major ions

Fig. 5 shows a box plot set of concentrations of alkalinity (Alk), conductivity (Cond), total dissolved solids (TDS), fluorides (F), water hardness (W_H), fecal coliforms (F_C), nitrate-nitrogen (NO_3-N), arsenic (As), cadmium (Cd), chromium (Cr), mercury (Hg), lead (Pb), manganese (Mn), and iron (Fe). The concentrations are grouped by groundwater quality level and regions. Overall, sites classified as good quality presented lower concentration of alkalinity, conductivity, total dissolved solids, water hardness and iron compared to sites categorized as regular and poor quality. Interestingly, sites with regular quality presented higher Mn concentrations compared to sites with good and poor quality. However, sites with poor quality demonstrated higher concentration of NO_3-N, As, Cr and Pb compared to sites with good and regular quality, except Pb, that was only reported in sites with poor quality. Particularly, four (*Lerma Santiago Pacífico, Río Bravo, Centrales del Norte*, and *Pacífico Norte*) showed high concentrations of arsenic, while two (*Pacífico Sur* and *Golfo Centro*) regions showed high concentrations of Cr. The concentration of Cd was not frequently reported. Finally, only *Balsas*, *Pacífico Sur* and *Golfo Centro* regions reported high concentrations of Hg.

3.1.2 Monitoring sites that exceed the permissible limits of physicochemical parameters

Fig. 6 shows for each region, the percentage of monitoring sites that exceeded the acceptable levels of Mexican regulations (DOF, 2022) of all physicochemical parameters of groundwater. The results demonstrated that in most of the Mexican regions, more than 10% of the monitoring sites exceeded the established water quality limits. In the case of conductivity, more than 20% of the monitoring sites in the *Baja California Peninsula*, *Golfo Centro*, and *Yucatán Peninsula* regions exceeded the regulations. However, the *Yucatan Peninsula* region had the highest water hardness (as $CaCO_3$) in the country, with 52% of monitoring sites having concentrations above the acceptable limit of 500 mg/L. Regarding heavy metals, Fe was detected in all regions. The *Pacífico Sur* (44%), *Aguas del Valle de México* (34%), and *Golfo Centro* (33%) regions presented the highest percentage of monitoring sites that exceeded the permissible limit of 0.3 mg/L for Fe. Likewise, approximately 30-40% of the sites in *Pacífico Sur* and *Centrales del Norte* regions showed Cr and As concentrations above acceptable limit of 0.05 mg/L. It is important to note that arsenic is a highly toxic, carcinogenic trace metal that can potentially contaminate groundwater sources, particularly in volcanic regions (Apostol, 2022). Other parameters that showed significant presence were total dissolved solids, fluorides, and nitrate-nitrogen. The *Río Bravo*, *Península de Baja California*, *Golfo Central*, *Golfo Norte* and *Aguas del Valle de México* regions presented the largest number of parameters with non-standard values. The pollution of heavy metals and other physicochemical parameters poses hazards to human health, biodiversity loss, disturbance in food chain

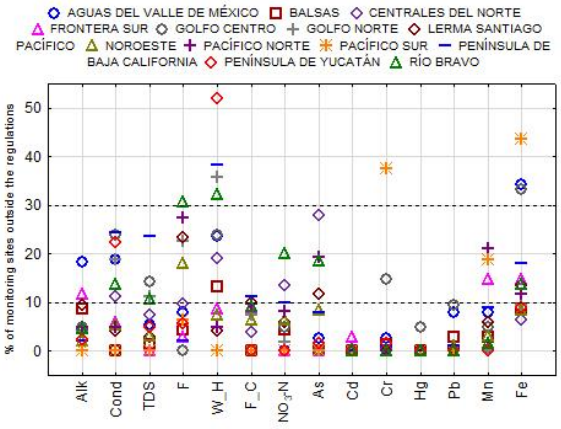

Fig. 6. Percentage of monitoring sites that exceeded the permissible limits (DOF, 2022) of the following physicochemical parameters of water: alkalinity, conductivity, total dissolved solids (TDS), fluorides (F), hardness (W_H), fecal coliforms, nitrate-nitrogen (NO_3-N), arsenic (As), cadmium (Cd), chromium (Cr), mercury (Hg), lead (Pb), manganese (Mn), iron (Fe). The monitoring sites were grouped by region.

and impacts on quality of environment. The presence of these parameters in the groundwater can be attributed to anthropogenic (e.g., fertilizer and industrial pollutants) as well as natural (e.g., volcanic action) sources (Krishan *et al*., 2021).

3.1.3 Multivariate data visualization

In Fig. 7, a set of radar plots shows multivariate data of normalized concentrations of 14 physicochemical-microbiological parameters and 8 major ions. For comparative purposes, all variables were normalized to a range of [0,1], and the same scale from 0 to 0.5 was maintained in all the subplots of Fig. 7. These radar diagrams showed some variations in the chemical composition of groundwater in all regions, which can be attributed to both geogenic and anthropogenic factors. In all the regions, except for *Aguas del Valle de México* and *Balsas* regions, high concentrations of HCO_3 and alkalinity were observed. Similarly, elevated concentration of nitrate-nitrogen was present in all regions, except for *Pacífico Norte* and *Lerma Santiago Pacífico* regions. Amiri and Nakagawa (2021) suggest that agricultural activities, fertilizer use, effluent leakage, and natural processes such as ammonia oxidation, can contribute to increased NO_3 concentration in groundwater. Three regions (Aguas del Valle de México, Balsas, and Frontera Sur) exhibited high normalized concentrations of K and moderate normalized concentrations of Na. The correlation between these components may be due to water-rock interaction processes, favoring the dissolution of alkaline minerals such as potassium feldspar and albite. Overall, the radar plots highlight the variations in each region of groundwater composition of 14 physicochemical-microbiological parameters and 8 major ions.

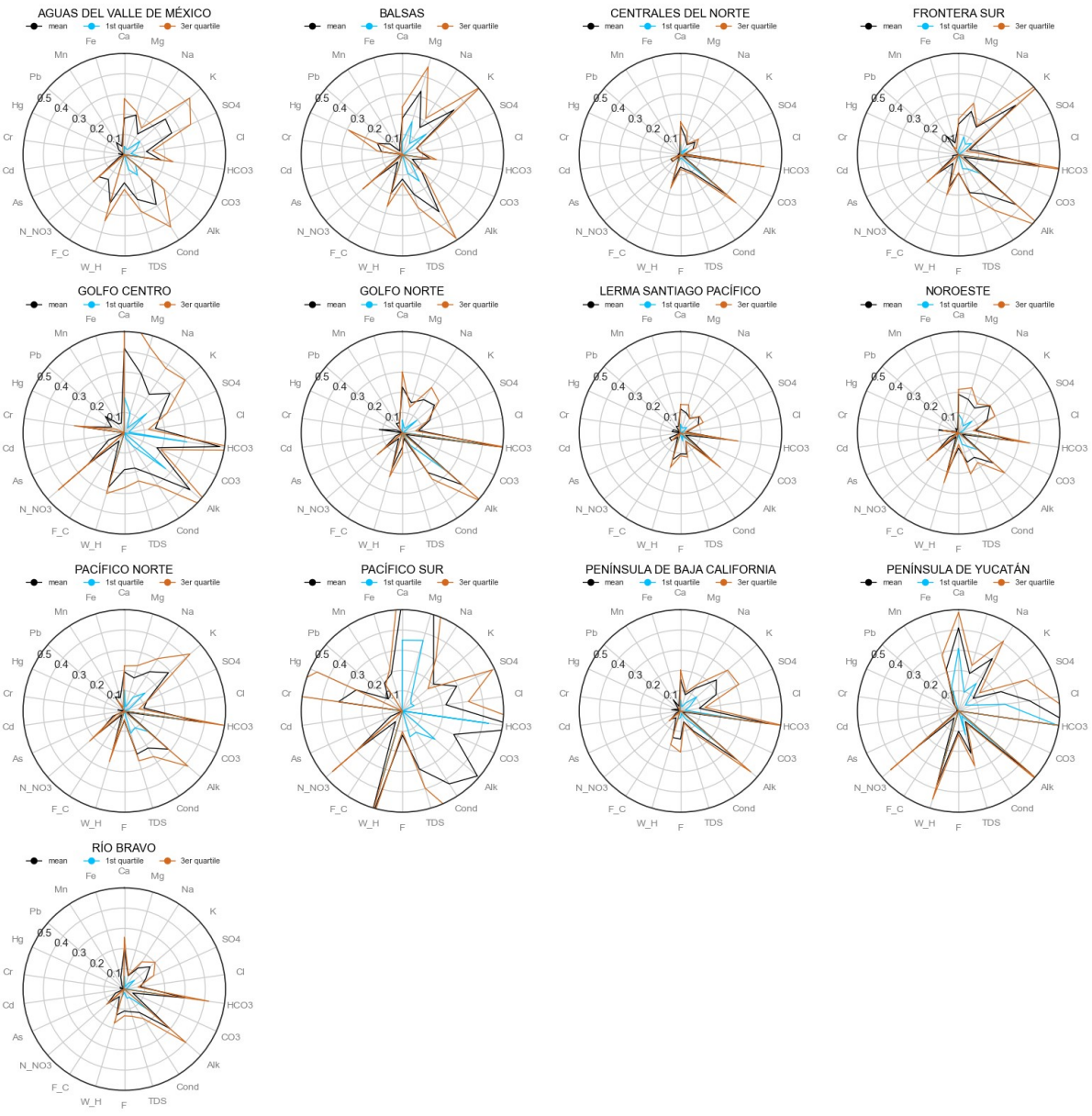

Fig. 7. Radar diagrams of normalized concentrations of 14 physicochemical-microbiological parameters and 8 major ions, namely: alkalinity (Alk), conductivity (Cond), total dissolved solids (TDS), fluorides (F), water hardness (W_H), fecal coliforms (F_C), nitrate-nitrogen (NO_3-N), arsenic (As), cadmium (Cd), chromium (Cr), mercury (Hg), lead (Pb), manganese (Mn), iron (Fe), Ca, Mg, Na, K, SO_4, Cl, HCO_3, and CO_3. The 1,068 monitoring sites were grouped by region. All variables were normalized within the range of 0 to 1, ensuring consistent scaling and comparability.

3.2 Water nomenclature from WCSystem program

Chemical compositions of 1,068 sites from Mexico were analyzed using the WCSystem program to infer the water nomenclature. Each of the five models implemented in this program provides the nomenclature as both basic and hybrid water classification types. However, before showing the results obtained from the water classification, a brief analysis of the variables used by the models was conducted. It is important to note that the CO_3 concentrations were imputed using a probable lower limit of detection (0.0005 mM), since it was not reported by the National Water Commission of Mexico (CONAGUA, 2022). Fig. 8 shows a box plot of concentrations (in mM) for Ca, Mg, Na, K, SO_4, Cl, and HCO_3 (mM) grouped by groundwater quality level and categorized by regions in Mexico. For all regions, sites with good water quality exhibited lower concentration of all the parameters compared to sites with regular and poor quality. This analysis provides insights into the variations in the major ions concentrations among different groundwater quality levels and regions in Mexico.

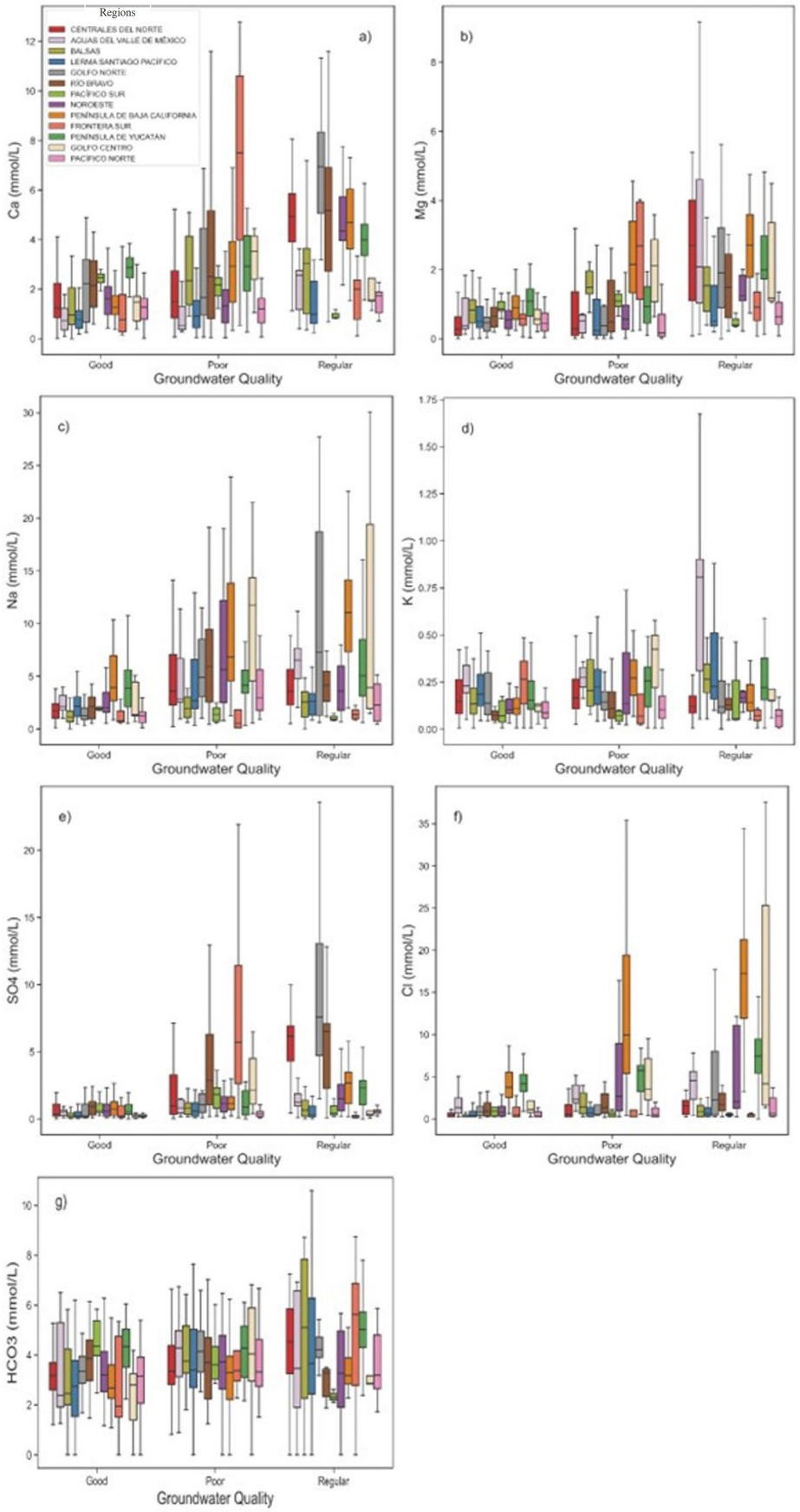

Fig. 8. Box plots of concentrations of Ca, Mg, Na, K, SO_4, Cl, and HCO_3 (mM) grouped by groundwater quality level and categorized by hydrological-administrative regions of Mexico.

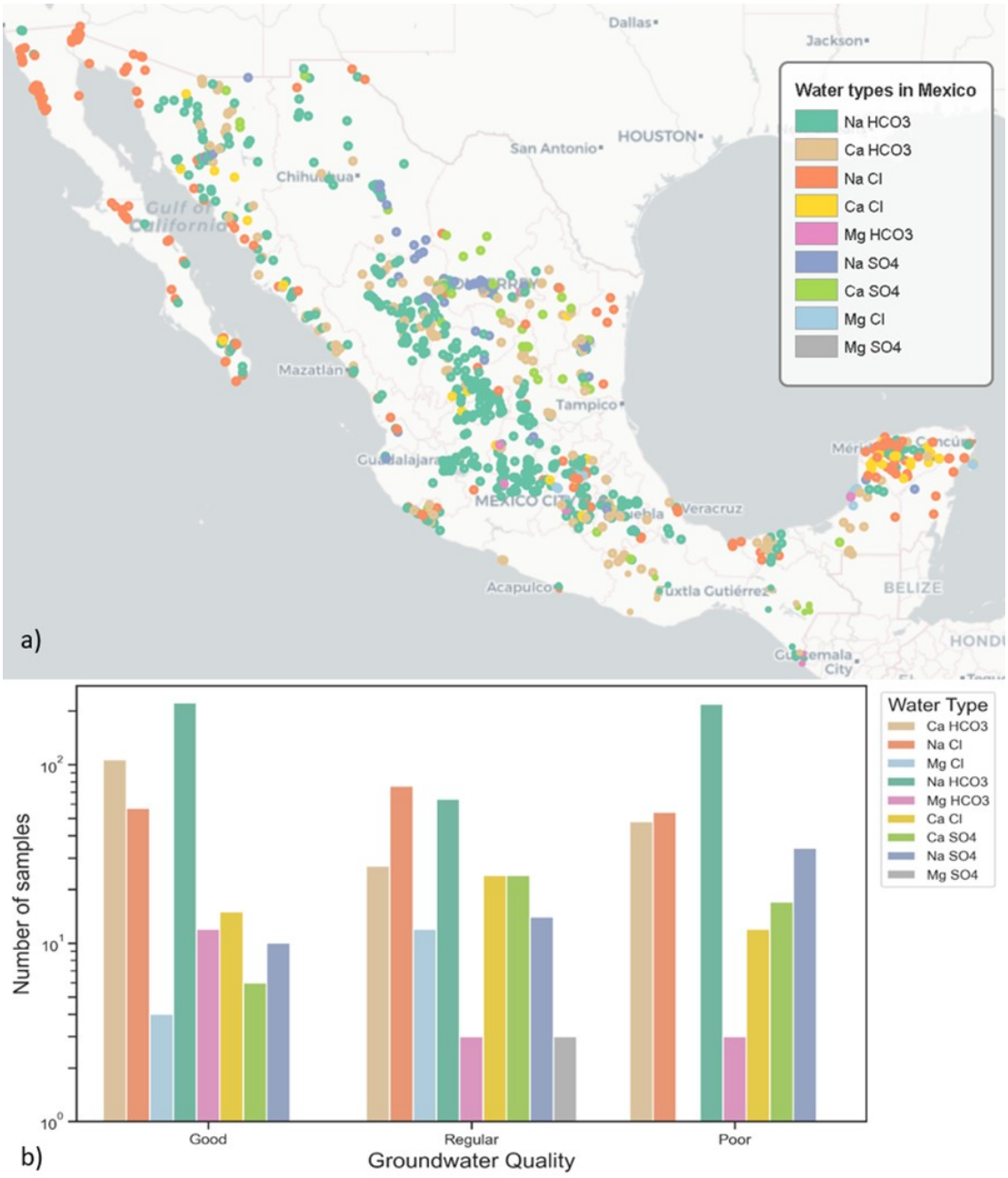

Fig. 9. Water types of 1,068 monitoring sites of Mexico sampled in 2020 by CONAGUA (2022). a) Map of basic nomenclature obtained by the VL model; b) Histogram of water types obtained by the VL model grouped by the water quality of the sites.

3.2.1 Basic water types

Fig. 9a shows a map of the basic water types provided by the VL model (as described in the "Development methodology of water classification models" section). The majority of samples (82%) were classified into three predominant classes (Table 1): Na HCO_3 (47%), Na Cl (18%), and Ca HCO_3 (17%). Additionally, six minority classes were identified (Table 1): Na SO_4 (5%), Ca Cl (5%), Ca SO_4 (4%), Mg HCO_3 (1.7%), Mg Cl (1.5%), Mg SO_4 (0.3%).

Fig. 9b presents a histogram of the basic nomenclature obtained by the VL model, grouped by water quality. Regarding the water quality of these 1,068 groundwater sites, 41, 23, and 36% were classified as good, regular, and poor quality, respectively. In general, these water types are distributed in all types of water quality. The predominant class, Na HCO_3 has fewer sites classified as regular quality compared to good and poor quality. Meanwhile, the third-class Ca HCO_3 has slightly more sites classified as good quality compared to the other types of quality. The minority class, Mg SO_4, was only observed in the regular water quality set. This analysis provides an overview of the basic water types and their distribution across different water quality levels.

3.2.2 Basic+hybrid water types

Each of the five models encapsulated into the WCSystem program offers “basic+hybrid" water types that can provide valuable insights into important environmental processes. However, we focus only on the VL model to simplify the “basic + hybrid" nomenclature results. In summary, the VL model classified 94% of the sites into the following 10 basic and hybrid types, each comprising at least 10 sites: Na HCO_3 (44%), Na Cl (16%), Ca HCO_3 (15%), Ca Cl (4%), Na SO_4 (4%), Ca SO_4 (4%), Ca-Na HCO_3 or Na-Ca HCO_3 (3%), Na Cl-HCO_3 or Na HCO_3-Cl (2%), Mg Cl (1%), Mg HCO_3 (1%). Additionally, 17 minority classes were identified (Table 1). This analysis highlights the distribution of water types “basic+hybrid", providing an understanding of the dominant types as well as the presence of minority classes in Mexican groundwater.

3.3 Multivariate data visualization using t-SNE algorithm

The multivariate data visualization utilized the t-distributed stochastic neighbor embedding (t-SNE) algorithm. This statistical technique, proposed by Van der Maaten and Hinton (2008), is designed to visualize high-dimensional data sets by reducing the number of dimensions while preserving the similarity relationships between observations.

Table 1. Summary frequency table of basic and hybrid water types obtained by the VL model grouped by hydrological-administrative regions.

Water nomenclature	Hydrological-administrative regions													
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	Total
Basic														
Na -HCO_3	17	43	40	24	130	128	18	23	33	9	4	13	24	506
Na -Cl	64	19	6	4	7	10	6	9	1	7	1	4	49	187
Ca -HCO_3	3	20	15	14	45	12	14	1	21	3	10	11	13	182
Na -SO_4	-	3	-	12	29	4	5	-	2	1	-	-	2	58
Ca -Cl	5	6	1	1	0	5	1	1	3	1	-	-	27	51
Ca -SO_4	-	3	-	9	17	1	9	-	2	-	1	4	1	47
Mg -HCO_3	-	-	-	1	1	5	-	-	7	-	-	2	2	18
Mg - Cl	-	-	-	-	-	5	-	4	-	-	-	-	7	16
Mg - SO_4	-	-	-	-	3	-	-	-	-	-	-	-	-	3
Hybrid+basic														
Na HCO_3	14	41	35	22	128	124	18	20	29	9	4	13	18	475
Na Cl	63	17	5	4	6	9	5	7	1	7	1	3	42	170
Ca HCO_3	1	19	13	12	39	10	13	1	18	3	9	10	12	160
Ca Cl	5	6	1	1	-	4	1	1	3	1	-	-	24	47
Na SO_4	-	3	-	8	24	3	5	-	1	1	-	-	1	46
Ca SO_4	-	2	-	8	13	-	9	-	2	-	1	3	1	39
Ca-Na HCO_3 or Na-Ca HCO_3	2	3	6	1	4	6	1	-	3	-	1	1	-	28
Na Cl-HCO_3 or Na HCO_3-Cl	4	-	1	-	-	1	1	4	-	-	-	1	9	21
Mg Cl	-	-	-	-	-	3	-	1	1	-	-	-	6	11
Mg HCO_3	-	-	-	-	1	5	-	-	2	-	-	2	1	11
Na SO_4-HCO_3 or Na HCO_3-SO_4	-	-	-	2	5	-	-	-	1	-	-	-	-	8
Ca-Na Cl or Na-Ca Cl	-	1	-	-	-	1	-	-	-	-	-	-	5	7
Ca-Na SO_4 or Na-Ca SO_4	-	-	-	3	3	-	-	-	-	-	-	-	-	6
Ca HCO_3-SO_4 or Ca SO_4-HCO_3	-	1	-	1	3	1	-	-	-	-	-	-	-	6
Mg-Na Cl	-	-	-	-	-	2	-	3	-	-	-	-	1	6
Mg-Na HCO_3 or Na-Mg HCO_3	-	-	-	-	-	-	-	-	4	-	-	-	1	5
Na Cl-SO_4 or Na SO_4-Cl	-	1	-	1	1	1	-	-	-	-	-	-	1	5
Ca-Mg HCO_3 or Mg-Ca HCO_3	-	-	1	-	-	-	-	-	2	-	-	-	-	3
Ca Cl-HCO_3 or Ca HCO_3-Cl	-	-	-	-	-	-	-	-	2	-	-	-	1	3
Ca-Na HCO_3-Cl or Na-Ca HCO_3-Cl	-	-	-	-	-	-	-	-	-	-	-	-	2	2
Ca-Na SO_4-Cl or Na-Ca SO_4-Cl	-	-	-	1	1	-	-	-	-	-	-	-	-	2
Mg SO_4	-	-	-	-	2	-	-	-	-	-	-	-	-	2
Ca SO_4-Cl	-	-	-	-	-	-	-	-	-	-	-	1	-	1
Mg-Na SO_4	-	-	-	-	1	-	-	-	-	-	-	-	-	1
Mg HCO_3-SO_4	-	-	-	1	-	-	-	-	-	-	-	-	-	1
Na-Ca HCO_3-SO_4	-	-	-	-	1	-	-	-	-	-	-	-	-	1
Na HCO_3-Cl	-	-	-	-	-	-	-	1	-	-	-	-	-	1

I: Península de Baja California; II: Noroeste; III: Pacífico Norte; IV: Río Bravo; V: Cuencas Centrales del Norte; VI: Lerma Santiago Pacífico; VII: Golfo Norte; VIII: Aguas del Valle De México; IX: Balsas; X: Golfo Centro; XI: Pacífico Sur; XII: Frontera Sur; XIII: Península de Yucatán.

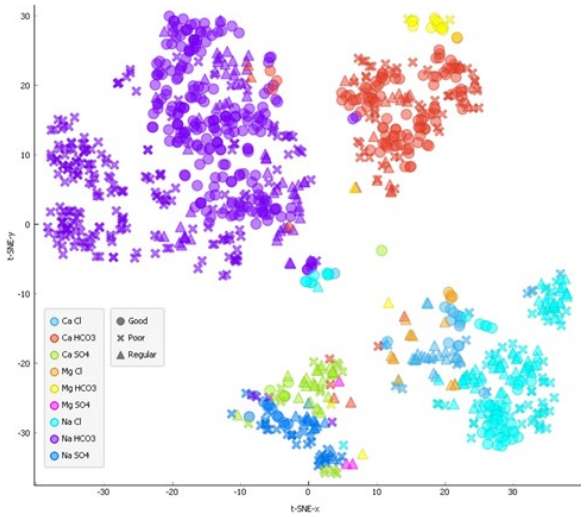

Figure 10. 2D visualization of monitoring sites using the t-SNE method (Van der Maaten and Hinton, 2008). The color of the symbol represents the basic water type obtained by the VL model and the symbol type presents the water quality determined.

The t-SNE algorithm aims to keep similar observations close to each other and dissimilar observations apart in the reduced-dimensional space. In this study, t-SNE was applied to reduce the dimensionality of the data (chemical concentrations of major elements and contaminants), and visualize it in a 2D plot. The result of this algorithm is presented in Fig. 10, where the color of the symbols represents the type of basic water determined by the VL model, and the type of symbol presents the water quality. As seen in this 2D visualization obtained by t-SNE algorithm, the sites were grouped according to their water nomenclature. The predominant class observed in the plot is Na HCO_3, followed by the Na Cl and Ca HCO_3 classes. This visualization provides insights into the clustering patterns and relationships between different water types and water quality.

3.4 Detailed analysis of Pacífico Sur hydrological-administrative region

In this section, we present the case study of the *Pacífico Sur* region to analyze in detail the water types provided by both the VL model and the Hill-Piper diagram. This region consists of groundwater 5 sites located in Guerrero state and 11 located in Oaxaca state. The chemical composition (mM) and results of water classification are presented in Table 2. To provide a graphical representation of the chemical composition of these 16 groundwater samples, Figure 11 is presented. Furthermore, the Hill-Piper diagram generated using the CCWater program (Pérez-Espinosa *et al*., 2019) is used to analyze the distribution of the sites within four distinct zones (Fig. 12 and Table 2): (i) zone 5: this zone includes 9 samples with carbonate hardness > 50%; (ii) zone 7: this zone includes 2 samples with Non-Carbonate alkali > 50%; (iii) zone 6: this zone includes 1 sample with non-carbonate hardness > 50%; and, (iv) ambiguous zone 9: this zone includes 1 sample that does not exhibit a cation-anion pair exceeding 50% dominance.

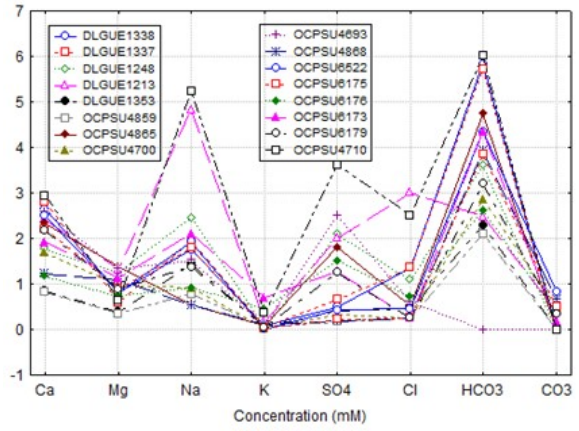

Figure 11. Concentration (mM) diagram of 16 groundwater sites from Pacífico Sur hydrological-administrative region.

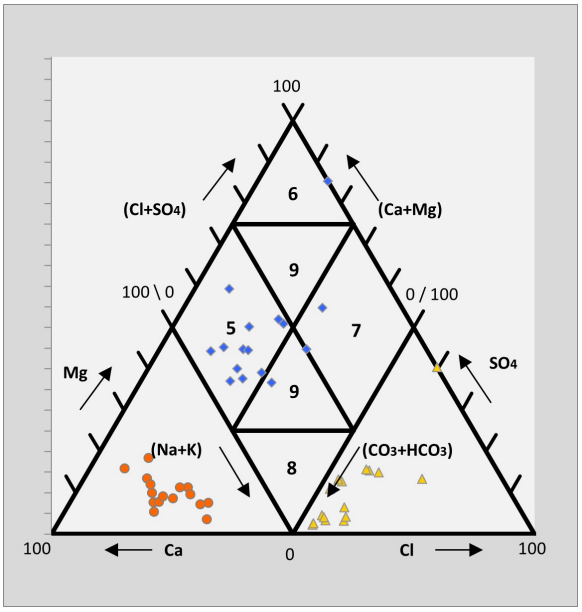

Figure 12. Hill-Piper ternary diagram of 16 groundwater sites from *Pacífico Sur* hydrological-administrative region.

The water nomenclature determined by the VL model is distributed across four water types: Ca HCO_3 (9 samples), Na HCO_3 (4 sites), Ca SO_4 (1 sample); Ca-Na HCO_3 (1 sample); and Na Cl (1 sample).

Finally, to address groundwater contamination in Mexico, here are some strategies that could be applied in collaboration among government entities, local communities, industries, and farmers: (i) apply strict regulations on industrial and agricultural activities near groundwater aquifers to prevent contamination; (ii) regulate and closely monitor the use of chemicals and toxic substances in industrial, agricultural, and domestic activities, as well as implement quality standards to regulate discharges; (iii) build and maintain effective wastewater treatment plants in urban areas, and promote responsible management practices at the domestic and community levels; (iv) educate the population about proper water management, and appropriate waste disposal methods; (v) develop monitoring programs that regularly assess groundwater quality and address contamination promptly;

Table 2. Application of the VL model and Hill-Piper diagram for the nomenclature of the groundwater sites from *Pacífico Sur* hydrological-administrative region (CONAGUA, 2022).

Sample		Chemical composition (mM)								Hill-Piper diagram					Water nomenclature
		Ca	Mg	Na	K	SO_4	Cl	HCO_3	CO_3	%Cl+SO_4	%HCO_3	%Na+K	%Ca+Mg	Class*	Basic+Hybrid
1	DLGUE1338	**2.65**	0.81	1.75	0.08	0.50	1.36	**5.83**	0.00	24	76	35	65	Zone 5	Ca HCO_3
2	DLGUE1337	**2.81**	0.59	1.94	0.06	0.65	1.36	**5.72**	0.00	26	74	37	63	Zone 5	Ca HCO_3
3	DLGUE1248	1.77	1.25	**2.45**	0.04	2.09	1.09	**3.61**	0.005	47	53	45	55	Zone 5	Na HCO_3
4	DLGUE1213	2.41	1.32	**4.83**	0.17	2.01	**3.00**	2.48	0.41	67	33	57	43	Zone 7	Na Cl
5	DLGUE1353	0.85	0.39	**1.44**	0.05	0.41	0.47	**2.29**	0.005	28	72	54	46	Zone 9	Na HCO_3
6	OCPSU4859	**0.82**	0.34	**0.77**	0.06	0.17	0.25	**2.10**	0.005	17	83	42	58	Zone 5	Ca-Na HCO_3
7	OCPSU4865	**2.34**	1.38	0.55	0.07	1.81	0.52	**4.74**	0.005	33	67	14	86	Zone 5	Ca HCO_3
8	OCPSU4700	**1.68**	0.97	0.91	0.03	0.31	0.25	**2.86**	0.47	17	83	26	74	Zone 5	Ca HCO_3
9	OCPSU4693	**2.59**	1.33	1.52	0.09	**2.51**	0.60	0.005	0.005	100	0	29	71	Zone 6	Ca SO_4
10	OCPSU4868	**1.22**	1.09	0.54	0.10	0.19	0.25	**3.93**	0.67	10	90	22	78	Zone 5	Ca HCO_3
11	OCPSU6522	**2.50**	0.81	1.92	0.01	0.42	0.45	**4.34**	0.83	17	83	37	63	Zone 5	Ca HCO_3
12	OCPSU6175	**2.21**	0.90	1.80	0.03	0.22	0.25	**3.85**	0.50	11	89	37	63	Zone 5	Ca HCO_3
13	OCPSU6176	**1.17**	0.74	0.92	0.46	1.51	0.71	**2.62**	0.33	46	54	42	58	Zone 5	Ca HCO_3
14	OCPSU6173	1.92	1.13	**2.10**	0.70	1.29	0.25	**4.34**	0.17	26	74	48	52	Zone 5	Na HCO_3
15	OCPSU6179	**2.17**	0.89	1.38	0.05	1.25	0.25	**3.20**	0.33	32	68	32	68	Zone 5	Ca HCO_3
16	OCPSU4710	2.95	0.65	**5.24**	0.35	3.62	2.51	**6.01**	0.005	50	50	61	39	Zone 7	Na HCO_3

*Zone 5: carbonate hardness > 50% (alkaline earths and weak acids dominate); zone 6: non-carbonate hardness > 50% (alkaline earths and dominant weak acids); zone 7: Non-carbonate alkali > 50% (alkalis and strong acids dominate); zone 9: ambiguous, no cation-anion pair > 50%. Note that the letter of the ions with the highest concentration in mM is presented in bold and underlined, which is consistent with the water type identified by the VL model.

annual sampling is insufficient, therefore, the sampling frequency should be increased especially in areas that exceeded the permissible limits of contamination defined by Mexican regulations.

In this context, Aguilar-Vilchis *et al*. (2023) demonstrated the biochemical use of Lerma River sediments for methane production and elimination organic pollutants in wastewater, which contributes to environmental remediation. Additionally, Canul-Chan *et al*. (2023) conducted a study on the biodegradation of crude oil and demonstrated its potential for effectively removing organic pollutants in wastewater.

Conclusions

Five models (7-hlr, Castboost, VL, VP, VR) for water classification are available for free through the WCSystem program. All these new models offer the possibility of detecting various basic and hybrid water types based on probabilities. These models outperform conventional approaches, such as the Hill-Piper diagram, that only provides four types (5, 6, 7 and 8 zones) and an ambiguous area (zone 9). Thus, we recommend that this new multidimensional scheme should replace the use of Hill-Piper ternary diagrams.

WCSystem accurately estimated the basic and hybrid water nomenclature for 1,068 groundwater sites in México. A half of the sites are concentrated in *Centrales del Norte*, *Lerma Santiago Pacífico*, and *Península de Yucatán* regions.

According to water quality categorization by CONAGUA, 41%, 23%, and 36% of the monitoring sites were classified as good, regular, and poor quality, respectively. The sites with good water quality exhibited lower concentrations of Ca, Mg, Na, K, SO_4, Cl, and HCO_3 compared to sites with regular and poor water quality.

In general, all regions showed high concentrations of HCO_3, Ca and Na, while high concentrations of Cl were observed in only three regions (*Aguas del Valle de México*, *Golfo Centro*, *Península de Baja California* and *Península de Yucatán*).

The recommended VL model classified 82% of the sites into three predominant basic classes (47% Na HCO_3; 18% Na Cl; and 17% Ca HCO_3). Furthermore, for basic+hybrid water types, the VL model classified 94% of the sites into 10 basic and combined hybrid classes, each with at least 10 sites. This demonstrates that groundwater in Mexico has high concentrations of Na and Ca cations, and HCO_3, SO_4, and Cl anions.

Additionally, t-SNE algorithm effectively grouped the monitoring sites according to their water nomenclature, where the predominant class was Na HCO_3, followed by the Na Cl and Ca HCO_3 classes.

Additionally, all regions exhibited at least one physicochemical-microbiological parameter that did not comply with the current regulations in Mexico. Therefore, it is necessary to apply environmental sanitation strategies to achieve a quality and safe water resource for human health. In future work we will carry out this analysis at the aquifer level to develop better environmental improvement strategies.

Acknowledgment

We express our gratitude to the two anonymous referees and the Editor-in-Chief, PhD. Angélica Román Guerrero, for their valuable feedback and suggestions, which greatly contributed to the improvement of our manuscript. In addition, we are grateful for the computer resources provided by “Laboratorio Nacional de Cómputo de Alto Desempeño” through project LANCAD 6-2023 "Aprendizaje máquina y profundo".

References

Aguilar-Vilchis, R., Hernández-Rodríguez, I.A., González-Blanco, G., Hernández-Soto, L.M., Aguirre-Garrido, J.F. and Beristain-Cardoso R. (2023). Characterization of sediments from the upper basin of the Lerma River, Mexico: Microbiome and biomethane potential. *Revista Mexicana de*

Ingeniería Química 22, IA2330. https://doi.org/10.24275/rmiq/IA2330

Ahmad, N., Sen, Z. and Ahmad, M. (2003). Ground water quality assessment using multirectangular diagrams. *Groundwater 41*, 828-832. https://doi.org/10.1111/j.1745-6584.2003.tb02423.x

Aitchison, J., (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London, UK.

Aitchison, J. and Egozcue, J.J. (2005). Compositional data analysis: where are we and where should we be heading? *Mathematical Geology 37*, 829-850. https://doi.org/10.1007/s11004-005-7383-7.

Al-Bassam, A.M. and Khalil, A.R. (2012). DurovPwin: a new version to plot the expanded Durov diagram for hydro-chemical data analysis. *Computers & Geosciences 42*, 1-6. https://doi.org/10.1016/j.cageo.2012.02.005

Amiri, V. and Nakagawa, K. (2021). Using a linear discriminant analysis (LDA)-based nomenclature system and self-organizing maps (SOM) for spatiotemporal assessment of groundwater quality in a coastal aquifer. *Hydrogeology Journal 603*, 127082. https://doi.org/10.1016/j.jhydrol.2021.127082

Apostol, G.L.C., Valenzuela, S. and Seposo, X. (2022). Arsenic in groundwater sources from selected communities surrounding Taal Volcano, Philippines: An exploratory study. *Earth 3*, 448-459. https://doi.org/10.3390/earth3010027

Butler, J.C. (1979). Trends in ternary petrologic variation diagrams; fact or fantasy? *American Mineralogist 64*, 1115-1121.

Canul-Chan, M., Rodas-Junco, B.A., Uribe-Riestra, E. and Houbron E. (2023). Biodegradation of crude oil present in wastewaters: evaluation of biosurfactant production and catechol 2,3 dioxygenase activity. *Revista Mexicana de Ingeniería Química 22*, Bio2932. https://doi.org/10.24275/rmiq/Bio2932

Chadha, D.K. (1999). A proposed new diagram for geochemical classification of natural waters and interpretation of chemical data. *Hydrogeology Journal 7*, 431-439. https://doi.org/10.1007/s100400050216.

CONAGUA (2022). Comisión Nacional del Agua de México. Informe técnico de calidad del agua en México. Available at: https://www.gob.mx/conagua/articulos/calidad-del-agua accessed: December 9, 2022.

DOF (2022). Diario Oficial de la Federación. NOM-127-SSA1-1994, Salud Ambiental. Agua para uso y consumo humano. Límites permisibles de calidad y tratamientos a que debe someterse el agua para su potabilización. Available at: https://www.dof.gob.mx/nota_detalle.php?codigo=2063863&fecha=22/11/2000 accessed: January 30, 2022.

Díaz-González, L., Uscanga-Junco, O.A. and Rosales-Rivera, M. (2021). Development and comparison of machine learning models for water multidimensional classification. *Journal of Hydrology 598*, 126234. https://doi.org/10.1016/j.jhydrol.2021.126234

Díaz-González, L., Uscanga-Junco, O.A., and Rosales-Rivera, M. (2022). WCSystem - A new computer program for water classification through five new multidimensional models and its application to geosciences. In: *Geochemical Treasures and Petrogenetic Processes*, (Armstrong-Altrin, J., Kailasa, P., and Verma S. eds) Springer.

Durov, S.A. (1948). Natural waters and graphic representation of their composition. Doklady Akademii Nauk SSSR. 59, 87-90.

Elhag, A.B. (2017). New diagram useful for classification of groundwater quality. *Journal of Geology & Geophysics 6*, 279. https://doi.org/10.4172/2381-8719.1000279

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems*. Second ed. Canada: O'Reilly Media.

Giménez-Forcada, E. (2010). Dynamic of sea water interface using hydrochemical facies evolution diagram. *Groundwater 48*, 212-216. https://doi.org/10.1111/j.1745-6584.2009.00649.x

Güler, C., Thyne, G.D., McCray, J.E., and Turner, A.K. (2002). Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeology Journal 10, 455-474. https://doi.org/10.1007/s10040-002-0196-6.

Handa, B.K. (1965). Modified Hill-Piper diagram for classification of groundwater in arid and semi-arid regions. *Geochemical Society of India Bulletin 1*, 20-24.

Hill, R.A. (1940). Geochemical patterns in Coachella Valley. *American Geophysical Union Trans. Union Part I. 21*, 46-49.

Krishan, G., Taloor, A.K., Sudarsan, N., Bhattacharya, P., Kumar, S., Ghosh, N.C, Singh, S., Sharma, A., Rao, M. S., Mittal, S., Sidhu, B. S., Vasisht, R., and Kour, R. (2021). Occurrences of potentially toxic trace metals in groundwater of the state of Punjab in northern India. *Groundwater for Sustainable Development 15*, 100655. https://doi.org/10.1016/j.gsd.2021.100655

Law, A.M., and Kelton, W.D. (2000). *Simulation Modeling and Analysis*. McGraw Hill, Boston.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vnderplas, J., Passos, A., Cournapeau, D., Brucher, M., Pérez-Espinosa, R., Pandarinath, K., and Hernández-Campos, F.J. (2019). CCWater-A computer program for chemical classification of geothermal waters. *Geosciences Journal 23*, 621-635. https://doi.org/10.1007/s12303-018-0064-6

Piper, A.M. (1944). A graphic procedure in the geochemical interpretation of water analyses. *Transactions of the American Geophysical Union 25*, 914-923. https://doi.org/10.1029/TR025i006p00914

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.

Ray, R.K, and Mukherjee, R. (2008). Reproducing the Piper trilinear diagram in rectangular coordinates. *Groundwater 46*, 893-896. https://doi.org/10.1111/j.1745-6584.2008.00471.x

Rencher, A.C. (2002). *Methods of Multivariate Analysis*. Wiley-Interscience, New York.

Shelton, J.L., Englea, M.A., Buccianti, A., and Blondes, M.S. (2018). The isometric log-ratio (ilr)-ion plot: a proposed alternative to the Piper diagram. *Journal of Geochemical Exploration 190*, 130-141. https://doi.org/10.1016/j.gexplo.2018.03.003.

Teng, W.C., Fong, K.L., Shenkar, D., Wilson, J.A., and Foo, D.C.Y. (2016). Piper diagram - a novel visualization tool for process design. *Chemical Engineering Research and Design 112*, 132-145. https://doi.org/10.1016/j.cherd.2016.06.002

Tharwat, A., Gaber, T., Ibrahim, A., and Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications 30*, 169-190. https://doi.org/10.3233/AIC-170729

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research 9*, 2579-2605.

Verma, S.P. (2015). Monte Carlo comparison of conventional ternary diagrams with new log-ratio bivariate diagrams and an example of tectonic discrimination. *Geochemical Journal 49*, 393-412. https://doi.org/10.2343/geochemj.2.0364

Verma, S.P., Rivera-Gomez, M.A., Díaz-González, L., and Quiroz-Ruiz, A. (2016). Log-ratio transformed major-element based multidimensional classification for altered high-Mg igneous rocks. *Geochemistry, Geophysics, Geosystems 17*, 4955-4972. https://doi.org/10.1002/2016GC006652

Verma, S.P., Uscanga-Junco, O.A., and Díaz-González, L. (2021). A statistically coherent robust multidimensional classification scheme for water. *Science of the Total Environment 750*, 141704. https://doi.org/10.1016/j.scitotenv.2020.141704

Verma, S.P., and Quiroz-Ruiz A. (2006). Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering. *Revista Mexicana de Ciencias Geológicas 23*, 133-161.