Stochastic Environmental Research and Risk Assessment (2024) 38:4527–4553
https://doi.org/10.1007/s00477-024-02817-w

ORIGINAL PAPER

Risk assessment for the surface water quality evaluation of a hydrological basin

Irene Brito[1] · **A. Manuela Gonçalves**[1] · **Ana Pedra**[1]

Accepted: 6 September 2024 / Published online: 15 October 2024

Abstract
This paper proposes a risk assessment methodology for evaluating the surface water quality of hydrological basins based on physico-chemical parameter concentrations. Considering the Douro River basin in Portugal and monthly recorded dissolved oxygen and conductivity parameter measurements in 18 water sampling stations from January 2002 to December 2013, the work intends to answer the research question of identifying the riskiest periods for water pollution in the year and classifying the water sampling stations in terms of risk for water pollution. The methodology consists first in determining the pollution risk implied by the physico-chemical parameters, based on the monthly water station measurements, using six different risk measures, namely mean, variance, loss probability, entropy, mean excess loss and value at risk. The risk values are ordered according to each risk measure and a final ranking is established through a ranking aggregation method. The final ranking permitted identifying the high risk period as ranging from May to October and the low risk period from November to April. Furthermore, July was classified as riskiest month concerning the dissolved oxygen concentration, and August as riskiest month regarding the conductivity levels. On the other hand, the ranking allowed classifying the water sampling stations, previously grouped in clusters, in terms of similar risk for water pollution: six sampling stations in the west of the basin formed the riskiest cluster in the dry period considering the dissolved oxygen concentrations, and four of those stations formed also the riskiest cluster concerning the conductivity levels.

Keywords Risk assessment · Risk measures · Risk ranking · Clustering · Water quality

1 Introduction

Water quality plays an important role in water resources management. Environmental problems, such as water pollution, threaten increasingly ecological water characteristics. Since rivers are the main water sources for population, agriculture and industry, it is fundamental and important to assess and monitor the river surface water quality. In 2000, the European Union established the Water Framework Directive (Directive 2000/60/EC of the European Parliament and of the Council - WFD), which outlines a framework for community action to protect inland surface waters, transitional waters, coastal waters, and groundwater. Currently, in Portugal, the entity responsible for the water quality management is the Portuguese Environment Agency (APA), which, among other objectives, is tasked with establishing a planning system adapted to the characteristics of water in different hydrographic regions. For this purpose, hydrometeorological and water quality data are collected at various monitoring stations, which are later disclosed and compiled in the repository of the National Water Resources Information System, SNIRH (2023).

Water quality is assessed by considering the physico-chemical and microbiological conditions that are necessary to sustain an healthy ecosystem. Thereby, risk analysis and risk assessment tools are essential to evaluate the water quality status, so that preventive and remedial actions can be considered and implemented for the protection of the water resources. The methodologies and techniques that have been developed and proposed in the literature for

✉ Irene Brito
ireneb@math.uminho.pt

A. Manuela Gonçalves
mneves@math.uminho.pt

Ana Pedra
pedracris2010@outlook.pt

[1] Centre of Mathematics, Department of Mathematics, University of Minho, 4800-045 Guimarães, Portugal

the risk assessment of water pollution in rivers consist of probabilistic approaches and approaches based on fuzzy set theory (see Ganoulis 2009 and references therein). Fuzzy stochastic methods were e.g. applied in Li et al. (2019) to analyse the risk of sudden water pollution in a river network system and in Rai et al. (2014) to assess the risk in transboundary rivers. Considering the probabilistic approaches, in several works, risk for water pollution is assessed through risk indices. For example, Zeleňáková et al. (2021) analysed the water quality of the Laborec river, Kumar et al. (2023) assessed the risk of heavy metal contamination in the Gomti river in India, Custodio et al. (2023) evaluated the microbial and toxic risks in the Andean river in Peru, Fang et al. (2023) used risk indices to assess pollution in the soil of the Huangshui river basin in China. López et al. (2020) analysed the water quality of the San Marcos river in Mexico through the quantification of physico-chemical and microbiological pollutants with risk quotients. Li et al. (2021) analyzed risks based on pollution indices in the surface water of the Yellow river in China.

Other water risk assessment frameworks, the water risk filter methodology WWF (2023), used e.g. in the work by Opperman et al. (2022), and the interdisciplinary water risk assessment framework proposed in Sandhu et al. (2023), take into account more risk categories and depend on different qualitative and quantitative water risk indicators.

In these approaches, the quantitative risk indices or indicators represent essentially mean estimates for pollution levels, that depend on the risk occurrence probability and the consequence, or they are based on average values of a given water quality parameter. However, one can find only few applications of other risk measures than the mean: the loss probability (under the designation of exceedance probability or probability of low water quality occurrence) was applied to river water quality analysis in Rehana et al. (2020), Gibbons (2003), Borsuk et al. (2002); applications of entropy were presented in the work by Singh et al. (2019), where the risk of heavy metal concentrations in water was assessed. The former works were additionally concerned with the integration of uncertainties, quantified essentially by variance, in the water quality analysis. Under this respect, Gronewold and Borsuk (2009) and Reckhow (2003) emphasized the importance and need of also estimating uncertainties for the modeling process.

In effect, the limitations of the existing risk assessment methods are that, in most approaches, the risk is quantitatively assessed only by a mean value or by a different single risk measure (in the few mentioned works) and that other quantitative risk characteristics are neglected.

The present work intends to address this limitation by proposing a method to assess the risk for water pollution quantitatively using different risk measures, that are determined for a certain water quality parameter through its representation by a random variable (taking into account the monthly concentrations). Beyond the mean, the variance and the loss probability, the entropy is used to measure the risk due to the uncertainty implied by the outcome probabilities, the mean excess loss considers the loss size over a certain threshold, and the value at risk looks up an adverse outcome associated with a specific probability. These risk measures are widely used in the actuarial and financial context (Kaas et al. 2008; Klugman et al. 2019) and have been applied in a pure or combined form to many problems in economy and finance (see e.g. Brachinger and Weber 1997; Brito 2022 and references therein) and in industrial settings (see e.g. Brito et al. 2022). The here proposed approach based on the different risk measures broadens, therefore, the quantitative risk assessment and has the advantage of taking into account different characteristics of the underlying parameter's distribution. As for the evaluation of the consequences of risk of water pollution due to the water quality parameter, i.e. the negative impact on the aquatic environment in the river and on environmental and human health, will be carried out elsewhere, since the main objective of the present study is to propose the risk assessment methodology for evaluating the surface water quality based on the physico-chemical parameters. However, the consequences could be analysed in a similar fashion (for that purpose numerical data are required that are directly related to the consequences of the risk factors, e.g. concerning the conductivity, the higher the conductivity levels, the higher is the risk for water pollution and the consequence is e.g. the higher amount of impurities in water, the harm to aquatic organisms).

The methodology consists in calculating the risk measures for a water quality parameter associated with different alternatives and ranking these alternatives according to their risk degree from the lowest to the highest risk. A final classification of the alternatives is obtained by applying a ranking aggregation method to the different individual rankings. In this work, the water quality parameters dissolved oxygen (DO) and conductivity will be considered using monthly measurements of 18 different water sampling stations of the Douro River hydrographic basin in Portugal (from January 2002 to December 2013) and the risk assessment method will be used to study the monthly risk for water pollution, in order to identify the most critical months and periods for water contamination in the year, and to classify water stations, previously grouped into clusters, in terms of their water pollution risk.

Cluster analysis is frequently applied in water quality assessment (see e.g. Barrie et al. 2023; Wu et al. 2023; Huang et al. 2011). Usually, this approach is applied to sets of water quality variables, composed of quantitative analytical data in order to group water quality monitoring sites

with similar water characteristics. Previous studies, such as those conducted by Gonçalves and Costa (2011), and Gonçalves and Alpuim (2011), for example, used this technique to classify water quality monitoring sites into homogeneous groups. In this study, cluster analysis will serve the same purpose. Specifically, hierarchical agglomerative clustering will be applied to a single water quality variable, the DO concentration and the conductivity level. DO and conductivity are physico-chemical parameters for surface water quality. Monitoring DO levels is crucial for assessing the health of aquatic ecosystems and the suitability of water for various uses (see e.g. Rudolph et al. 2002). The conductivity of water is another important parameter that can reveal changes in water quality (see e.g. Tejaswini et al. 2023). The aim is to identify homogeneous regions based on similarities in the temporal dynamics of water quality variables observed across time and space within the River Douro basin with the purpose to evaluate the surface water quality. In this analysis, seasonal variations will be considered by distinguishing between dry and wet periods (He et al. 2015; Du et al. 2013). The risk assessment method will then enable the analysis of water contamination risk in different seasonal periods and clusters.

This paper is organized as follows. In Sect. 2, the risk assessment methodology is presented, where in Sect. 2.1 the risk measures are defined, in Sect. 2.2 the risk ranking aggregation is explained, and in Sect. 2.3. the methodological framework is described. In Sect. 3, the methodology is applied to water quality data of the Douro River basin, where Sect. 3.1 contains the descriptions of the study area and of the data set, Sect. 3.2 is dedicated to the risk assessment considering the monthly DO and conductivity concentrations and Sect. 3.3 provides the risk assessment of water station clusters. The conclusions are presented in Sect. 4.

2 Risk assessment

In the context of water quality assessment, risk for water pollution is related with the loss of water quality characteristics, implied by deviations of certain physico-chemical and microbiological parameter concentrations from predefined quality standards. The parameter measurements can be described by a random variable and the risk associated with the parameter deviations can be described quantitatively using risk measures. Naive risk measures (see Brachinger and Weber 1997), such as the expected value or the variance, give an overall indication of the general mean concentration and about deviations from that mean. Other risk measures, such as the loss probability or the value at risk (Kaas et al. 2008; Klugman et al. 2019) have applications in the actuarial context, however suit well to the present context, since they take into account a threshold or target value that can be identified with the quality standard limits. The comparison of the risk measures, obtained for example for different monthly periods or for different clusters of water measurement stations, permits then ranking these alternatives according to each risk measure and identifying for example the most critical monthly periods in the year for water contamination or the most critical water station clusters. Since each risk measure can lead to a different ranking order, a ranking aggregation method will be applied to obtain a final, overall ranking.

In the following Section, six risk measures are defined that will be used in the risk assessment process.

2.1 Risk measures

Let X be the random variable representing a physico-chemical parameter concentration and consider a data-dependent frequency distribution for a given data sample of parameter measurements.

The following risk measures (in the list denoted by R_1 to R_6) will be used to assess the risk for water contamination:

R_1 Expected value, $\mathbb{E}[X]$:

Depending on the influence of the chemical parameter on water contamination, a higher mean can indicate a lower risk or a higher risk.

R_2 Variance, $\mathrm{Var}[X]$:

A higher variance implies a higher risk.

R_3 Loss probability:

Depending on the influence of the chemical parameter on water contamination, the loss probability can be defined by

$$P(X \leq r), \tag{1}$$

where r is a certain target level and outcomes lower than r represent a loss, or by

$$P(X \geq r), \tag{2}$$

where r is a certain target level and outcomes higher than r represent a loss. A loss probability can be interpreted as a probability of undesirable outcomes. A higher loss probability translates into a higher risk.

R_4 Entropy:

The entropy is given by

$$H(X) = -\sum_{k=1}^{K} p_k \ln(p_k), \tag{3}$$

where p_k, $k = 1, \ldots, K$, are the probabilities of the outcomes of X. The entropy measures the uncertainty and can

therefore be used to measure risk, where a higher entropy indicates a higher risk. R_5 Mean excess loss function and mean shortfall loss function:
The mean excess loss function is defined by

$$\mathbb{E}\big[X - r | X > r\big], \tag{4}$$

where *r* is a given value and outcomes higher than *r* represent a loss. The mean shortfall loss function is given by

$$\mathbb{E}\big[|X - r||X < r\big], \tag{5}$$

where *r* is a given value and outcomes lower than *r* represent a loss. The higher the mean excess loss or the mean shortfall loss, the higher is the risk for water contamination through high or, respectively, low parameter concentrations.
R_6 Value at risk (VaR):
The VaR at the confidence level *p*, $0 < p < 1$, is the quantity π_p defined by:

$$P(X \leq \pi_p) \geq p. \tag{6}$$

Here, the 95%-quantile, $\pi_{0.95}$, satisfying $P(X \leq \pi_{0.95}) \geq 0.95$, will be used. Depending on the interpretation of the chemical parameter, a lower VaR can indicate a higher or lower risk for water contamination.

2.2 Ranking of risks and ranking aggregation

Consider a set of alternatives $A = \{A_1, \ldots, A_I\}$ that are ranked according to different risk measures (or risk attributes) $R_1, \ldots, R_J$. Let the *I* rankings of the alternatives be given by the ranking matrix $RM = [r_{ij}]_{I \times J}$ as follows

$$RM = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_I \end{array} \begin{array}{c} \begin{array}{cccc} R_1 & R_2 & \cdots & R_J \end{array} \\ \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1J} \\ r_{21} & r_{22} & \cdots & r_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ r_{I1} & r_{I2} & \cdots & r_{IJ} \end{pmatrix} \end{array},$$

where r_{ij} is the rank order of alternative A_i for the risk attribute R_j. The row *i* of the ranking matrix, represented in vector form by $RM_{I=i} = (r_{i1}, r_{i2}, \ldots, r_{iJ})$, contains the ranking orders of the alternative A_i obtained with the *J* risk measures, whereas the column *j*, $RM_{J=j} = (r_{1j}, r_{2j}, \ldots, r_{Ij})$ contains the ranking of the alternatives for the attribute R_j. The risk ranking orders are determined from the lowest to the highest risk, where the lowest risk corresponds to the risk position number 1 and the highest risk to the risk position number *I* in the ordering. Given two alternatives A_{i_1} and A_{i_2} and considering the risk attribute R_j, A_{i_1} has a higher risk than A_{i_2}, which will be represented by $A_{i_1} \succ A_{i_2}$, if the risk attribute $r_{i_1 j}$ represents a higher risk than $r_{i_2 j}$. If $r_{i_1 j} = r_{i_2 j}$, then A_{i_1} and A_{i_2} are equally risky (tied in terms of risk), which is denoted by $A_{i_1} \sim A_{i_2}$. If $r_{i_1 j}$ represents a lower risk than $r_{i_2 j}$, then A_{i_1} is less risky than A_{i_2}, $A_{i_1} \prec A_{i_2}$.

The ranking of alternatives for a given attribute R_j can also be described through a pairwise comparison (Cook 2006; Kemeny and Snell 1962) and represented using an individual risk attribute comparison matrix $D^j = [d^j_{ik}]_{I \times I}$ with entries defined by

$$d^j_{ik} = \begin{cases} 1, & \text{if } A_i \succ A_k, \\ 0, & \text{if } A_i \prec A_k, \\ 0.5, & \text{if } A_i \sim A_k. \end{cases}, i, k = 1, \ldots I. \tag{7}$$

To obtain a final ranking (overall ranking) from the different rankings one can employ a ranking aggregation method, such as the Borda method or the Copeland method. These methods belong to the batch mode methods, where the aggregated ranking is obtained in one run (Ding et al. 2018).The Borda method (Cook 2006; Zahid and de Swart 2015) is based on the ranking positions given by the ranking vectors $RM_{I=i} = (r_{i1}, r_{i2}, \ldots, r_{iJ})$ for each alternative A_i. The Borda score (or Borda count) for the alternative A_i gives the final aggregated ranking position of A_i and is defined by

$$BS(A_i) = \sum_{j=1}^{J} r_{ij}. \tag{8}$$

The final ranking of the alternatives A_i, $i = 1, \ldots, I$, is then obtained by ranking the alternatives using the following Borda ranking rule, here adapted to the risk order relation:

$$BS(A_{i_1}) > BS(A_{i_2}) \Rightarrow A_{i_1} \succ A_{i_2}, \tag{9}$$

where $i_1, i_2 = 1, \ldots, I$. This means that A_{i_1} is riskier than A_{i_2} if A_{i_1} has a higher Borda score than A_{i_2}. If there is a tie between two alternatives, which happens if there exist two alternatives A_{i_1} and A_{i_2} such that $BS(A_{i_1}) = BS(A_{i_2})$, then the tie can be broken by leaving out the poorest grade of each alternative and recalculating the Borda score. This process may be repeated successively until a strict ordering of both alternatives is achieved.

The Copeland method (Ding et al. 2018; Klamler 2005), applied to the present risk ordering context, ranks the alternatives taking into account the difference between the number of alternatives for which a given alternative is riskier and the number of alternatives for which a given alternative has a lower risk. The Copeland score (or Copeland count) of A_i is defined by

$$CS(A_i) = |\{A_n \in A : A_i \succ A_n\}| - |\{A_m \in A : A_m \succ A_i\}|, \quad (10)$$

where $|\cdot|$ stands for the cardinality of a set. The final ranking of the alternatives A_i, $i = 1, \ldots, I$, by the Copeland ranking, here adapted to the risk order relation, is obtained as follows:

$$CS(A_{i_1}) > CS(A_{i_2}) \Rightarrow A_{i_1} \succ A_{i_2}. \quad (11)$$

If $CS(A_{i_1}) = CS(A_{i_2})$, then there is a tie between the alternatives A_{i_1} and A_{i_2}. In this case, the tie breaking rule described for the Borda ranking can be used, by leaving out the poorest grade of each alternative and recalculating the Copeland score. This process may be repeated successively until a strict ordering of both alternatives is achieved.

2.3 Methodological framework description

The methodological risk assessment framework for the water quality evaluation, applied in the next section, consists of the following steps, summarized in the diagram in Fig. 1 (more details about the assumptions and the methods are provided in the Sects. 2.1, 2.2 and in the next section).

Step 1–Data sampling: Identify the water quality parameter (e.g. dissolved oxygen) and select the water stations based on the criterion of having less than 5% missing water quality parameter data in a given time interval of regularly, equally spaced (monthly) collected data in the official database. Extract the corresponding water quality parameter data from the official database.

Step 2–Data preprocessing: Perform data imputation for missing values using a linear interpolation model. Go to step 4, or else, apply a data normalization method (e.g. standardization through the range) for the clustering analysis.

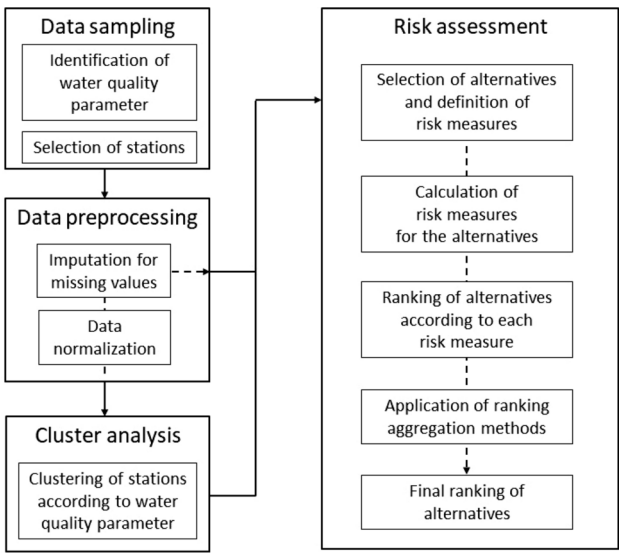

Fig. 1 Methodological risk assessment framework

Step 3–Cluster analysis: Cluster water sampling stations according to the water quality parameter using Ward's hierarchical agglomerative clustering method.

Step 4–Risk assessment:

(a) Select the alternatives for the risk assessment (e.g. months or clusters).
(b) Define the risk measures: mean, variance, loss probability, mean excess loss or mean shortfall loss function, and value at risk by specifying water quality parameter thresholds taking into account the guideline values.
(c) Calculate the risk measures using a frequency-based distribution corresponding to the alternatives' water quality parameter data samples.
(d) Rank the alternatives according to each risk measure from the lowest to the highest risk.
(e) Apply ranking aggregation methods (Borda method and Copeland method) to obtain an aggregated ranking.
(f) Determine the final risk assessment for the alternatives based on the aggregated ranking.

Considering the risk assessment framework in Step 4, it is tailored to the specific context of the study by taking into account the parameters' related risk perception and the parameters' quality standard guideline values as thresholds for defining the risk measures and by determining a frequency distribution from the parameter data for the calculation of the risk measures. Furthermore, the ranking methods take into account the adequate interpretation and perception of risk related to each risk measure, depending on the specific water quality parameter (as described in Sect. 2.1). The main advantage of this risk assessment framework is the integration of different risk measures that evaluate risk more deeply than the usual evaluation based on the mean. The risk measures consider different risk-related characteristics (e.g. uncertainty, excess, and mean excess). It is a pure quantitative approach and one limitation is that an appropriate data sample must be available for the assessment (e.g. monthly regularly collected data). The problem is that for certain water sampling stations there exists a high number of missing data. In the present study this limited, on the one hand, the number of water station clusters to be included in the analysis and, on the other hand, also the determination of the time range, as will be explained in the next section.

3 Application to water quality data of a hydrological basin

The methodology presented in the previous section will now be applied to water quality data in order to assess the risk for water pollution in the hydrological Douro River basin.

3.1 Study area and data set

The Douro River Hydrographic Region is located in the Iberian Peninsula (Portugal and Spain) and has a total area of 97500 km^2, of which 18600 km^2 are in Portuguese territory and 78900 km^2 are in Spanish territory. This work considers the Douro River Hydrographic Region RH3 in Northern Portugal which encompasses an area of 18600 km^2 (see Fig. 2). Given the crucial role that the hydrographic basin plays in the local ecosystem and the lives of the communities that depend on it, monitoring its quality to prevent and identify environmental problems is essential.

The data, measurements of several water quality variables, are collected in the water resources monitoring network (of the Ministry of the Environment in Portugal), in a set of water monitoring stations located along the Douro River hydrographic basin. The water resources monitoring network is made up of automatic and conventional stations. The data are available in the Portuguese National Information System for Water Resources, SNIRH (2023). Unfortunately, some of these stations have missing values or have even been deactivated at some time. The existence of missing data may represent a limitation for applying the assessment method since missing values can introduce bias in the results and reduce the statistical power of the study. For this reason, only the following 18 stations were selected for this work (see Fig. 2), based on the criterion of having less than 5% missing data: Alb. Alvão - V. Real (ALV), Alb. Bastelo (BAS), Alb. Camba (CAM), Alb. Fonte Longa (LON), Alb. Ranhados (RAN), Alb. Serra Serrada (SER), Alb. Sordo (SOR), Foz Corgo (COR), Foz Tâmega (Alb. Crestuma) (TAM), Melres (MEL), Moledo (MOL), Penude (PEN), Praia Aurora (AUR), Ponte Bateira (BAT), Ponte Vale Telhas (TEL), Quinta Maravilha (MAR), Semealho - Alb. Torrão (SEM), and Vau (VAU). Although, in general, it is acceptable if less than 5% of values in a data set are missing, in the present study an equal size of monthly observations for each station is needed, so that the missing values were replaced based on the available information. To deal with this problem, data imputation for missing values was performed. The imputation procedure consists in establishing a linear interpolation model by using the function *na.interp()* of the package *forecast* of the *R* software (see Moritz et al. 2015). In the case of data with seasonal behaviour, this function applies a STL (Seasonal-Trend Decomposition Procedure Based on Loess) decomposition, before the linear interpolation.

The present study is based on data from January 2002 to December 2013, because this observation period corresponds to a time interval, where the data were regularly and monthly collected. From 2013 onwards, the periodicity of data collection changed from 1 month to 3 months. Also, the irregularity of data collection after that date is significant due to management and restructuring aspects of water networks done by SNIRH (2023). Therefore, data from December 2013 onwards were not considered in the present study. Although the data set contains several water quality variables, this study focuses on the water quality variables: DO and conductivity, as mentioned above and justified in more detail in the next section.

3.2 Risk assessment: monthly DO and conductivity concentrations

Water quality parameters help to measure the quality of surface water in hydrological basins and to assess the risk for water pollution. Here, the physico-chemical parameters DO and conductivity are considered. The concentrations of these parameters change naturally throughout the year. For this reason the different risk measures, presented in Sect. 2.1, were used to assess the risk of water pollution taking into account the monthly nature of the data, where the aim is to identify the most critical months or periods for water pollution in the year influenced by the DO concentration and by the conductivity.

3.2.1 Monthly dissolved oxygen

Dissolved oxygen is one of the most important physico-chemical parameters for measuring the water quality in a river. DO involves both physical and chemical aspects: it measures the amount of oxygen (O_2) physically dissolved in water, which is influenced by chemical processes such as respiration, photosynthesis, and the decomposition of organic matter. A low amount of DO indicates poor water quality, that the water is highly polluted and that contaminants are consuming the dissolved oxygen. Therefore, risk is related to a decrease of DO, increasing the possibility for water contamination, where according to Portuguese standards (APA 2021), the DO quality situation in rivers is described by the following interval quality levels (in mg/l):

$$\text{excellent} - [8;12], \quad \text{good} - [6;8[, \quad \text{fair} -]-;6[.$$

Table 13 in Appendix 1 contains descriptive statistics of the DO concentration for each month in the observation period. Those results can be visualized in Fig. 3 through the boxplots determined for the monthly DO concentrations. One can observe a seasonal influence on the DO concentration, where, in general, in the hot or dry period the values tend to be lower than in the cold or wet period.

A more detailed analysis of the monthly DO concentration, in particular the monthly risk for water pollution, will be performed next using the methodology based on the risk

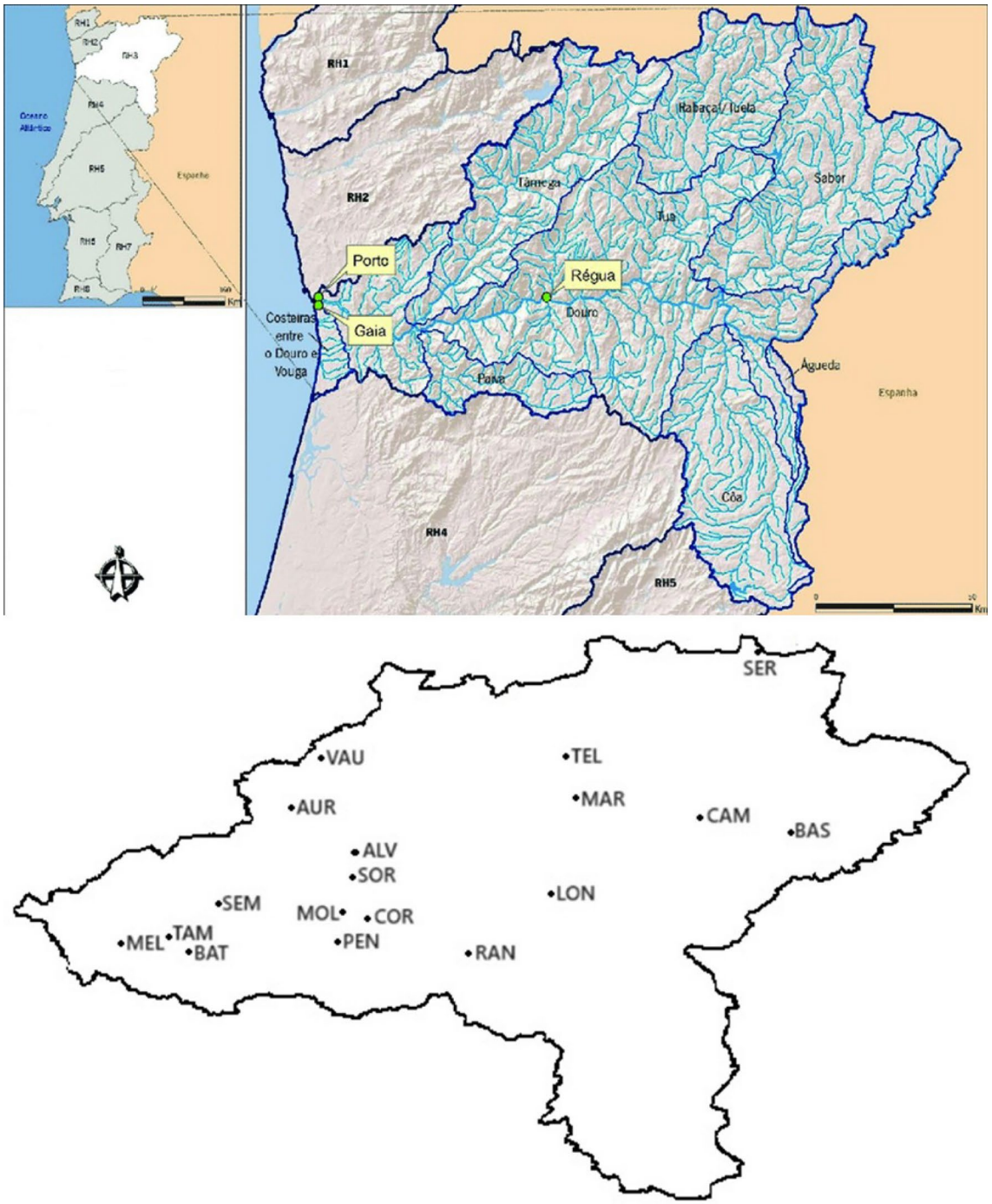

Fig. 2 The Douro River hydrological basin (left) and spatial distribution of the selected water quality monitoring stations (right)

measures presented in Sect. 2.1. Estimates for the risk measures R_1-R_6, were calculated for each month based on a frequency distribution determined from the data sample of month specific DO measurements (in an analogous way as done e.g in Yang et al. (2017), Brito (2023)).

As for the interpretation of the risk measures, considering R_1, a lower mean implies a higher risk for water pollution. Considering R_3, the loss probability $P(X \leq r)$ with $r = 8$ was determined, so that a loss occurs if the DO concentration falls below 8, meaning that water loses the excellent DO quality level. A higher loss probability translates into a higher risk. In R_5, the mean shortfall loss function, $\mathbb{E}\big[|X - r||X < r\big]$, r was set to $r = 8$, so that only DO values that do not belong to the highest quality level were considered, and the higher this risk value, the higher is the risk for water contamination through low DO concentrations.

Fig. 3 Boxplots for DO

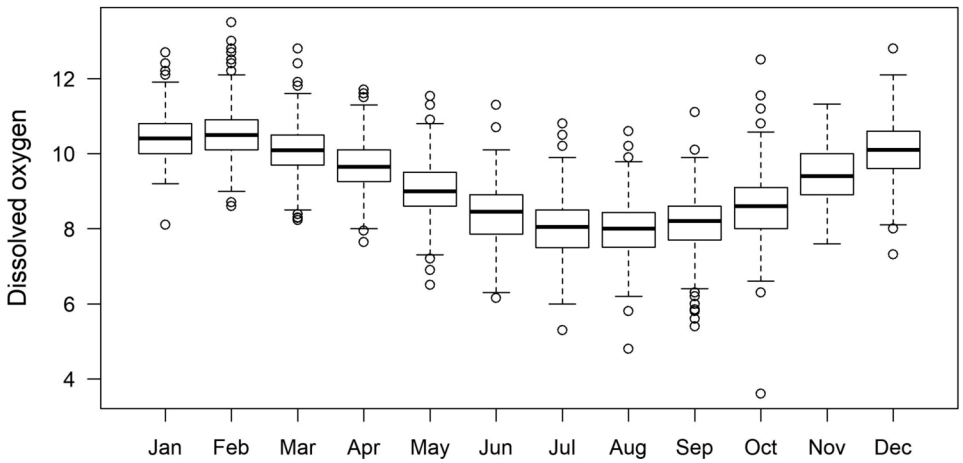

Table 1 Risk measures per month for DO

	R_1	R_2	R_3	R_4	R_5	R_6
January	10.4548	0.3690	0.0000	1.1802	0.0000	11.3223
February	10.5967	0.5084	0.0000	1.2909	0.0000	12.1636
March	10.0946	0.4366	0.0000	1.2451	0.0000	11.3319
April	9.6761	0.3906	0.0046	1.2304	0.0017	**10.4867**
May	**9.0214**	0.5260	**0.0417**	1.3423	**0.0301**	10.5111
June	**8.3785**	**0.6365**	**0.2130**	**1.4189**	**0.1686**	**9.7000**
July	**8.0184**	**0.6084**	**0.3611**	**1.4281**	**0.2915**	**8.8145**
August	**8.0018**	**0.5706**	**0.3426**	**1.3919**	**0.2772**	**8.8521**
September	**8.1314**	**0.6679**	**0.2731**	**1.4544**	**0.2486**	**9.6618**
October	**8.5514**	**0.9386**	**0.1898**	**1.6045**	**0.1630**	**9.6486**
November	9.4462	0.5467	0.0046	1.3544	0.0019	10.5759
December	10.1344	**0.5892**	0.0046	**1.3801**	0.0032	11.4026

R_1–mean, R_2–variance, R_3–loss probability, R_4–entropy, R_5–mean shortfall loss, R_6–value at risk)

The interpretation of R_6 is that a lower VaR indicates a higher risk for water contamination.

Table 1 contains the results of the risk measures obtained for each month, where the values for the six riskiest months in terms of each risk measure are highlighted in bold. The results are represented graphically in Figs. 11, 12, 13 in Appendix 2. In fact, one can see a seasonal influence on the DO concentration and on the risk for water pollution, where, in general, the riskiest months belong to the dry period, from May to September, except December, characterized by a high uncertainty, and April, attaining a lower VaR than May. Additionally, one can observe that October is also a critical month according to all risk measures and it is characterized by having the highest variance and entropy (see Figs. 11 and 12 in Appendix 2). Although all mean values belong to the excellent quality level, the loss probabilities and mean shortfall losses indicate that from April to December water looses the excellent quality. One can observe in Figs. 12 and 13 in Appendix 2 that these two risk measures (loss probability and mean shortfall loss) have a very similar graphical pattern. The values of the loss probability reveal that the probability of occurring DO concentrations lower than 8 is approximately 0.36 in July and 0.34 in August, only from November to April it is almost zero, where from January to March it is exactly zero. This means that assessing risk only by the mean can be misleading and therefore it is important to consider further risk measures.

Considering now the ranking of risks and the ranking aggregation, the set of alternatives consists in this case of the 12 months, $A = \{A_1, \ldots, A_{12}\} = \{\text{Jan}, \ldots, \text{Dec}\}$. Ranking the months according to the risk measures $R_1, \ldots, R_6$ from the lowest to the highest risk, where the rank order 1 corresponds to the lowest risk and the rank order 12 to the highest risk, one obtains the following risk ranking matrix (where an arithmetic mean was determined as the rank for months with equal risk values):

Table 2 Borda score, final Borda ranking (BR), Copeland score and final Copeland ranking (CR) for DO

	A_i											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
BS	12	14	17	25	37	54	66	60	61	58	33	31
BR	1	2	3	4	7	8	12	10	11	9	6	5
CS	−54	−50	−44	−28	−4	30	54	42	44	38	−12	−16
CR	1	2	3	4	7	8	12	10	11	9	6	5

Table 3 Borda score, final Borda ranking (BR), Copeland score and final Copeland ranking (CR) for DO after elimination of the riskiest month

	A_i										
	Jan	Feb	Mar	Apr	May	Jun	Aug	Sep	Oct	Nov	Dec
BS	12	14	17	25	37	53	59	59	56	33	31
BR	1	2	3	4	7	8	10/11	10/11	9	6	5
CS	−48	−44	−38	−22	2	34	46	46	40	−6	−10
CR	1	2	3	4	7	8	10/11	10/11	9	6	5

$$RM = \begin{array}{c} \\ \text{Jan} \\ \text{Feb} \\ \text{Mar} \\ \text{Apr} \\ \text{May} \\ \text{Jun} \\ \text{Jul} \\ \text{Aug} \\ \text{Sep} \\ \text{Oct} \\ \text{Nov} \\ \text{Dec} \end{array} \begin{array}{c} \begin{array}{cccccc} R_1 & R_2 & R_3 & R_4 & R_5 & R_6 \end{array} \\ \begin{pmatrix} 2 & 1 & 2 & 1 & 2 & 4 \\ 1 & 4 & 2 & 4 & 2 & 1 \\ 4 & 3 & 2 & 3 & 2 & 3 \\ 5 & 2 & 5 & 2 & 4 & 7 \\ 7 & 5 & 7 & 5 & 7 & 6 \\ 9 & 10 & 9 & 9 & 9 & 8 \\ 11 & 9 & 12 & 10 & 12 & 12 \\ 12 & 7 & 11 & 8 & 11 & 11 \\ 10 & 11 & 10 & 11 & 10 & 9 \\ 8 & 12 & 8 & 12 & 8 & 10 \\ 6 & 6 & 5 & 6 & 5 & 5 \\ 3 & 8 & 5 & 7 & 6 & 2 \end{pmatrix} \end{array}.$$

All risk measures rank differently the 12 months, confirming that each risk measure evaluates differently the specific risk related characteristics of the quality variable. A closer look reveals that January and February seem to get the lowest risk orders and, oppositely, July, August and September attained the highest risk orders. Calculating the Borda score using (8) and the Copeland score using (10) for each month, one obtains the results displayed in Table 2.

According to the final aggregated Borda ranking and Copeland ranking, the six riskiest months can be classified from the highest to the lowest risk as follows:

$$\text{BR,CR}: \ \text{July} \succ \text{September} \succ \text{August} \succ \text{October} \succ \text{June} \succ \text{May},$$

so that July is the riskiest month and May is the less riskiest one (among the six months). Considering the six months with the lowest risk, one obtains the following ranking with both methods:

$$\text{BR,CR}: \ \text{November} \succ \text{December} \succ \text{April} \succ \text{March} \succ \text{February} \succ \text{January}.$$

Both Borda and Copeland rankings classify January as the less riskiest month and February as second less riskiest month. November is the sixth less riskiest month.In sensitivity analysis related to ranking studies, one issue is the rank reversal problem. It consists in investigating the effect of eliminating an alternative on the ranking result in order to test the robustness of the model. A rank reversal occurs if the orders of the remaining alternatives change. Here, the occurrence of rank reversals is analysed by eliminating the worst alternative and studying the impact of the change on the ranking. A scenario of the worst alternative removal is tested as follows. The Borda and Copeland methods established the month of July as the riskiest alternative (see Table 2). Removing July results in the new Borda and Copeland rankings presented in Table 3. One can observe that no rank reversals occur, the only difference is that a tie appears between August and September, which are now classified as the riskiest months in the new ranking. However, the removal of the worst case does not affect the remaining ranking positions.

In general, the results obtained in this subsection confirm that it is important to take into account the different seasonal conditions and the distinction between dry and wet periods in the water quality assessment, as indicated by the monthly ranking of DO concentrations. The lower amount of DO in the dry period can be explained by the higher water temperature in that period since the temperature is one factor that influences the DO concentration in water (see e.g. Rajwa-Kuligiewicz et al. 2015; Rajesh and Rehana 2022). More attention should therefore be paid to the dry period, especially to July, in order to improve the water quality conditions. Since October is classified as belonging to the most critical months and it is even riskier than June and May, the results suggest that the critical and risky period for water pollution (the dry period) should range from May to October (note that usually the time horizon from May to September is considered as the dry period).

3.2.2 Monthly conductivity

Conductivity is considered a physico-chemical parameter that can, to some extent, reveal the quality of surface water. It measures the water's ability to conduct electrical current, which is influenced by the presence of dissolved ions such as salts and minerals. Therefore, it encompasses both physical aspects (electricity conduction) and chemical aspects (concentration and types of dissolved ions). As water quality changes, the ability of water to pass current through it changes. Conductivity measurements permit the detection of pollution events in rivers. A high conductivity indicates that the water contains a high quantity of contaminants and it has therefore a poor quality. For example, the occurrence of an agricultural runoff increases the conductivity due to the higher amount of impurities in the water. According to Portuguese standards APA (2021), the recommended values for conductivity (in μ S/cm) should be lower than 250 and the following quality levels are distinguished:

$$\text{excellent/good}: \ [-, 250[, \quad \text{fair}: \ [250; -[.$$

Descriptive statistics of the monthly conductivity measurements are listed in Table 14 in Appendix 1. The corresponding boxplots are depicted in Fig. 4. One can observe the presence of a maximum outlier (see Fig. 4 on the left), detected in June of 2011 in MEL. Since the extreme conductivity value may have been caused by an agricultural or industrial runoff or by vessel discharges, the outlier is considered in the present analysis. Only for a better graphical comprehension and comparisons of the boxplots, the outlier, recorded in June of 2011, was replaced by the average of the observations of that month (see Fig. 4 on the right).
One can observe in Fig. 4, that there are several occurrences throughout the year that led to conductivity values exceeding the excellent/good level of 250. The upper interquartile range tends to be larger from June to November, when compared with other months, indicating that in the hot or dry period the conductivity values may be in general higher than in the period from December to April. The following analysis based on the risk measures permits a more detailed and better risk assessment of water contamination due to the conductivity levels in the different months and periods of the year.

As in the analysis of the DO concentration, the estimates for the risk measures R_1-R_6 (see Sect. 2.1), were calculated for each month based on a frequency distribution determined from the data sample of month specific conductivity measurements.

The interpretation of certain risk measures is now opposite to the interpretation established for DO. Considering R_1, here, a higher mean implies a higher risk for water pollution. For R_3, the loss probability $P(X \geq 250)$ was determined, meaning that a loss occurs if the conductivity exceeds 250, when water looses the excellent/good conductivity quality level. A higher loss probability translates into a higher risk. Regarding R_5, the mean excess loss $\mathbb{E}\big[X - 250 | X > 250\big]$ was determined, so that only conductivity values of poorer quality (not belonging to the excellent/good quality level) were considered. The higher the mean excess loss, the higher is the risk for water contamination through high conductivity values. A higher VaR (R_6) indicates a higher risk for water contamination.

Table 4 contains the results of the risk measures obtained for each month, where the values for the six riskiest months in terms of each risk measure are highlighted in bold. The results can be visualized graphically in Figs. 14, 15, 16 in Appendix 2.

Considering now the ranking of risks and the ranking aggregation for the set of alternatives consisting of the 12 months, $A = \{A_1, \ldots, A_{12}\} = \{\text{Jan}, \ldots, \text{Dec}\}$, first, the risk ranking matrix is determined, yielding

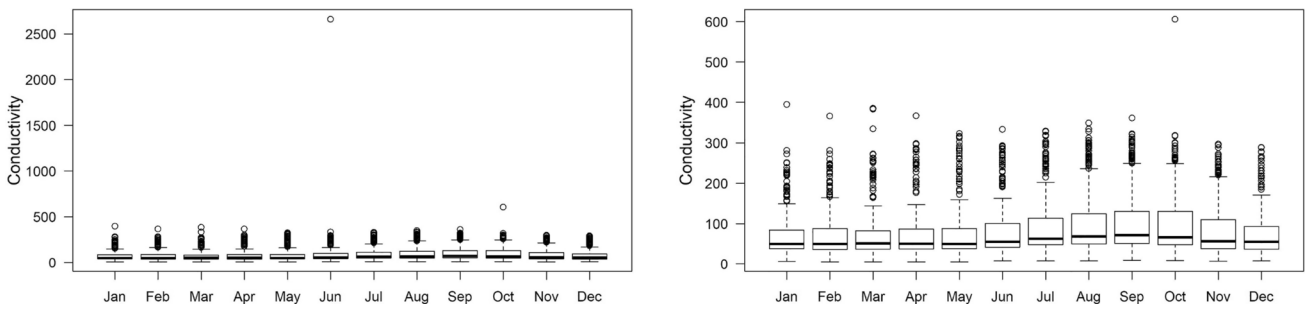

Fig. 4 Boxplots for conductivity with outlier (left panel) and excluding the outlier (right panel)

Table 4 Risk measures per month for conductivity

	R_1	R_2	R_3	R_4	R_5	R_6
January	73.4644	1710.177	0.0463	0.1875	0.5167	64.3529
February	75.0617	2114.430	0.0602	0.2275	0.4074	256.7692
March	75.8617	2512.134	0.0556	0.2146	1.8065	**282.5167**
April	78.1406	3636.882	0.0926	0.3085	1.5676	**266.9300**
May	80.3202	**3882.773**	**0.0972**	**0.3189**	**1.9639**	**270.2000**
June	**100.4379**	**34604.730**	**0.1250**	**0.3966**	**12.6347**	262.2731
July	**98.2710**	**5321.859**	**0.1481**	**0.4195**	**3.4373**	**273.2016**
August	**106.2050**	**5950.654**	**0.1713**	**0.4579**	**4.4325**	**275.8761**
September	**109.6451**	**5779.563**	**0.1806**	**0.4722**	**3.9005**	**271.6028**
October	**105.4852**	**6160.870**	**0.1713**	**0.4792**	**3.5030**	261.1293
November	**84.5579**	2737.653	0.0880	0.2978	0.2671	253.0368
December	76.7430	1921.232	0.0556	0.2146	0.4148	257.4664

R_1–mean, R_2–variance, R_3–loss probability, R_4–entropy, R_5–mean excess loss, R_6–Value at Risk

Table 5 Borda score, final Borda ranking (BR), Copeland score and final Copeland ranking (CR) for conductivity

	A_i											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
BS	9	18	30	35	42	55	52	63.5	63	57.5	25	18
BR	1	2/3	5	6	7	9	8	12	11	10	4	2/3
CS	−60	−42	−18	−8	6	32	26	49	48	37	−28	−42
CR	1	2/3	5	6	7	9	8	12	11	10	4	2/3

$$R = \begin{array}{c} \\ \text{Jan} \\ \text{Feb} \\ \text{Mar} \\ \text{Apr} \\ \text{May} \\ \text{Jun} \\ \text{Jul} \\ \text{Aug} \\ \text{Sep} \\ \text{Oct} \\ \text{Nov} \\ \text{Dec} \end{array} \begin{array}{c} \begin{array}{cccccc} R_1 & R_2 & R_3 & R_4 & R_5 & R_6 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 4 & 1 \\ 2 & 3 & 4 & 4 & 2 & 3 \\ 3 & 4 & 2.5 & 2.5 & 6 & 12 \\ 5 & 6 & 6 & 6 & 5 & 7 \\ 6 & 7 & 7 & 7 & 7 & 8 \\ 9 & 12 & 8 & 8 & 12 & 6 \\ 8 & 8 & 9 & 9 & 8 & 10 \\ 11 & 10 & 10.5 & 10 & 11 & 11 \\ 12 & 9 & 12 & 11 & 10 & 9 \\ 10 & 11 & 10.5 & 12 & 9 & 5 \\ 7 & 5 & 5 & 5 & 1 & 2 \\ 4 & 2 & 2.5 & 2.5 & 3 & 4 \end{pmatrix} \end{array},$$

where the rank order 1 corresponds to the lowest risk and the rank order 12 to the highest risk and an arithmetic mean was used as the rank for months with equal risk values.The results show that the risk for water contamination, as indicated by the risk measures and the corresponding orders, is higher in the period from June to October, especially in August, September and October. One can see the influence of the outlier in June on the variance and on the mean excess loss, which classify June as the worst month. The influence of the outlier can also be well observed in the graphical plots of the variance and the mean excess loss in Figs. 14 and 16, respectively, in Appendix 2. The mean and the loss probability rank September as the riskiest month. On the other hand, all risk measures, except entropy, classify January as the less riskiest month. Note that the mean values belong all to the excellent quality level (see Table 4), however according to the loss probabilities, exceedances of the conductivity level 250 have in fact occurred, although with low probabilities (lower than 0.2).

Determining the Borda score using (8) and the Copeland score using (10) for each month, one obtains the results in Table 5.

The final aggregated Borda ranking and the final aggregated Copeland ranking lead to the following classification of the six riskiest months (from the highest to the lowest risk):

$$\text{BR,CR}: \text{ August} \succ \text{September} \succ \text{October} \succ \text{June} \succ \text{July} \succ \text{May}.$$

August is the riskiest month and May is the less riskiest one (among the six months). As for the six months with the lowest risk, both methods lead to the following ranking:

$$\text{BR,CR}: \text{ April} \succ \text{March} \succ \text{November} \succ \text{December} \sim \text{February} \succ \text{January}.$$

January is the less riskiest month and April is the sixth less riskiest month. February and December are tied. Using the tie breaking rule described in Sect. 2.2, February is classified as having a lower risk than December, so that the final ranking for both methods read

$$\text{BR,CR}: \text{ April} \succ \text{March} \succ \text{November} \succ \text{December} \succ \text{February} \succ \text{January}.$$

To test the effect of the worst alternative removal on the orders of the remaining alternatives, the riskiest month August is eliminated and the Borda and Copeland scores are recalculated for the remaining months. From the results in Table 6, one can observe that the elimination of the riskiest month August does not imply the reversal of the alternative orders. Thus, the results remain consistent.The results of the risk assessment for water pollution through the monthly conductivity also show a clear distinction between the dry and wet period or between the hot and cold period. Since conductivity is also temperature dependent: as the temperature of the water increases, conductivity levels increase (see e.g. Hayashi 2004; Talbot et al. 1990), this can explain the higher conductivity levels observed in the dry period. In the present study, again, the most critical period in the year susceptible to pollution ranges from May to October. In that period, August followed by September and October are the riskiest months for the occurrence of pollution events, so that more attention should be paid to these months. As in the risk assessment based on the DO concentrations, also the conductivity measurements suggest that October should be included in the critical and risky hot or dry period.

3.3 Risk assessment of water station clusters based on DO and conductivity concentrations

Having determined the most susceptible months for the risk of water pollution and classified the months in terms of risk due to the DO concentration and conductivity levels, the aim is now to apply the risk assessment methodology to classify water sampling station clusters in terms of risk for water pollution, that is, to classify stations that are grouped in clusters based on similarity of parameter concentrations and thus on similar risk. According to the previous risk assessment methodology, the DO concentrations and conductivity values suggest a division of the months into a higher risk period for water contamination ranging from May to October and a lower risk period from November to April. Therefore, in the following analysis, the higher risk period from May to October will be considered as the dry period and the period from November to April, as the wet period.

The water sampling stations will first be grouped into similar locations using hierarchical agglomerative clustering with standardized data, by division through the range. Given a set of observed values $Z = \{z_1, \ldots, z_J\}$, the standardized values using the range are computed as follows

$$x_j = \frac{z_j}{\max(Z) - \min(Z)}, \quad j = 1, \ldots, J. \tag{12}$$

According to Milligan and Cooper (1988), in hierarchical agglomerative clustering analysis, using the standardization through the range and Ward's method (see Ward 1963) is more effective than using other standardization procedures combined with other linkage methods, such as e.g. the complete linkage method. For that reason, Ward's method of linkage with Euclidean distance will be used to measure the dissimilarity. Thus, starting with the singleton clusters (containing only one observation), the Euclidean distance between clusters are calculated and the most similar clusters are merged during the agglomeration process. Consider the data matrix $X = [x_{ikt}]$, $i = 1, \ldots, N$, $k = 1, \ldots, p$, $t = 1, \ldots, n$, containing the already standardized observed values, where x_{ikt} represents the value of the quality variable *k*, measured at the monitoring station *i* at month *t*. Let P_t be the set of all quality variables measured in the same month *t* in both water monitoring stations *i* and *j*. The average of the Euclidean distances between water monitoring stations *i* and *j* over all months *t* is given by the expression

$$d_{ij} = \frac{1}{n}\sum_{t=1}^{n}\left[\sum_{k\in P_t}(x_{ikt} - x_{jkt})^2\right]^{1/2}, i, j = 1, \ldots, N, \tag{13}$$

and this is the dissimilarity measure employed in the present study for the agglomeration of clusters. This methodology was based on the previous work of Gonçalves and Alpuim (2011).The clustering technique will be applied separately to the DO water quality variable and to the conductivity variable ($p = 1$), considering the monthly measurements from January 2002 to December 2013 ($n = 144$ months), in 18 monitoring sites of the Douro River hydrological basin ($N = 18$).

For each variable (DO and conductivity), the cluster analysis will be applied considering all the data (the data in the total period), the data from the dry period and the data from the wet period. The risk assessment methodology, described in Sect. 2, will then be followed to evaluate

Table 6 Borda score, final Borda ranking (BR), Copeland score and final Copeland ranking (CR) for conductivity after elimination of the riskiest month

	A_i										
	Jan	Feb	Mar	Apr	May	Jun	Jul	Sep	Oct	Nov	Dec
BS	9	18	29	35	42	53	62	59	55	25	18
BR	1	2/3	5	6	7	8	11	10	9	4	2/3
CS	−54	−36	−13	−2	12	34	32	46	38	−22	−36
CR	1	2/3	5	6	7	9	8	11	10	4	2/3

the risk for water pollution in the different water sampling station clusters. In more detail, the risk measures are calculated for each cluster, using the corresponding water quality parameter measurement data, and the clusters are then ranked according to their final aggregated ranking.

3.3.1 Cluster analysis and risk assessment: dissolved oxygen

Applying the cluster analysis to the DO data of the 18 stations (listed in Sect. 3.1), considering the total period, the wet and the dry period, leads to the cluster configurations presented by the dendrograms in Fig. 5. The 18 water sampling stations were grouped as follows:

$$C_{1D} = \{\text{COR, MEL, MOL, PEN, SEM, TAM}\} \tag{14}$$

$$C_{1T} = \{\text{COR, MEL, MOL, PEN, SEM, TAM, ALV, RAN, SOR}\} \tag{15}$$

$$C_{1W} = \{\text{COR, MEL, MOL, PEN, SEM, TAM, ALV, RAN, SOR, CAM, BAS, LON}\} \tag{16}$$

$$C_{2D} = \{\text{AUR, BAT, MAR, SER, TEL, VAU, BAS, CAM, LON, ALV, RAN, SOR}\} \tag{17}$$

$$C_{2T} = \{\text{AUR, BAT, MAR, SER, TEL, VAU, BAS, CAM, LON}\} \tag{18}$$

$$C_{2W} = \{\text{AUR, BAT, MAR, SER, TEL, VAU}\}, \tag{19}$$

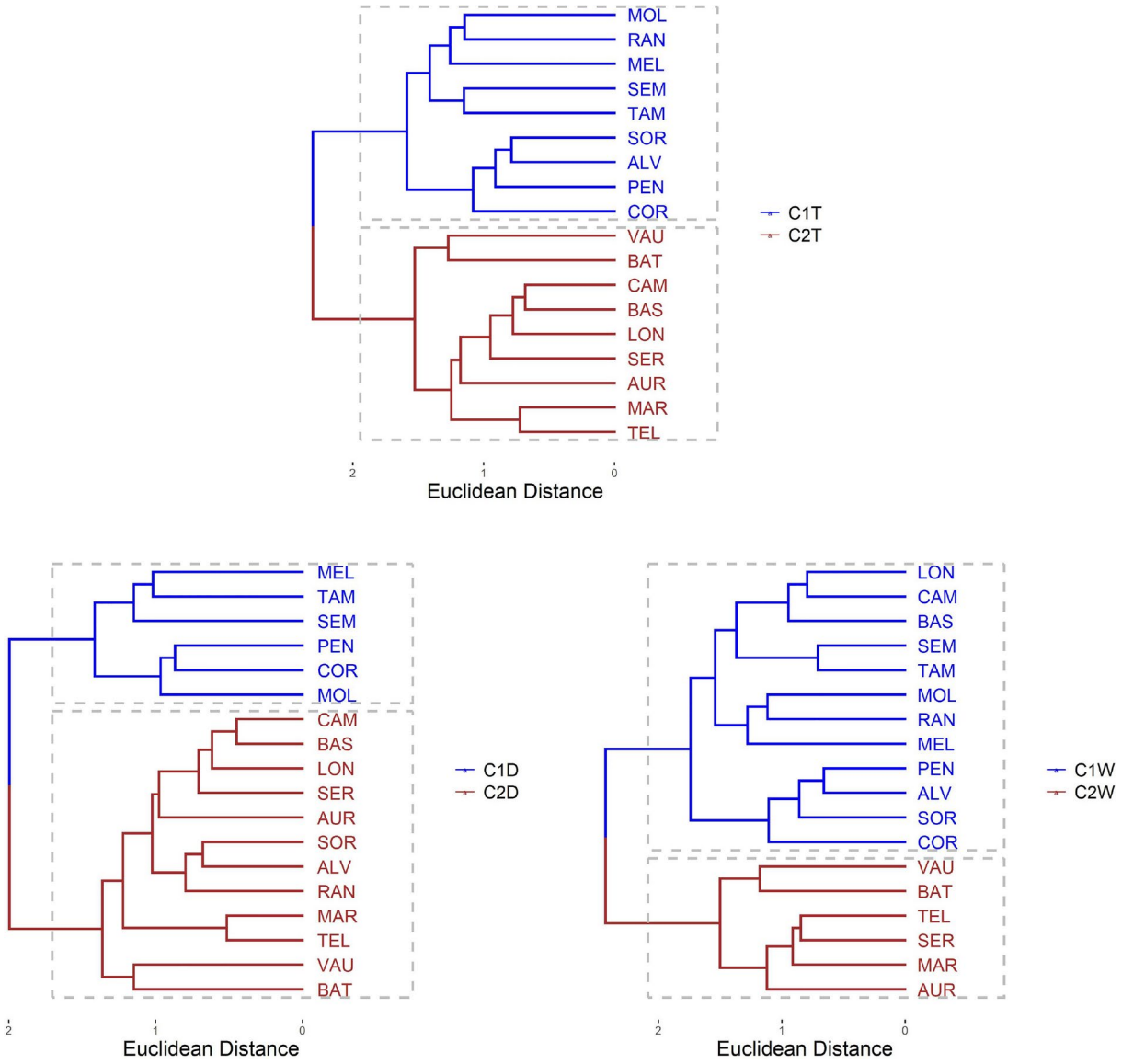

Fig. 5 Dendrograms of the DO variable, considering all the data (top), the data from the dry period (left) and the data from the wet period (right)

where C_{1D} and C_{2D} denote the clusters in the dry period, C_{1W} and C_{2W}, the clusters in the wet period, and C_{1T} and C_{2T} are the clusters in the total period. As can be seen, the obtained clusters are different according to the considered data period (total observed period, dry period and wet period), which highlights that the DO concentration changes over time with respect to the seasonal period (impact of precipitation, water temperature and other hydrometeorological factors) and influences differently the individual stations.
In order to explore the differences between the clusters, the descriptive statistics of DO concentrations are presented in Table 15. Those results can be observed in Fig. 6 through the boxplots of DO concentrations for each cluster. Its analysis shows that the clusters C_{1T}, C_{1D} and C_{1W} have slightly lower values and slightly higher variances when compared to the clusters C_{2T}, C_{2D}, and C_{2W}, respectively. At the same time, comparing the clusters of the dry and wet periods, it becomes evident that during the dry period, the DO concentrations are lower.

The risk assessment of the clusters and a more detailed analysis of the clusters' DO concentrations will be performed using the risk measures presented in Sect. 2.1. Estimates for the risk measures R_1-R_6 were calculated for each cluster based on a frequency distribution determined from the data sample of clusters specific DO measurements. The obtained risk measures are presented in Table 7 and the corresponding graphical representations can be found in Figs. 17, 18, 19 in Appendix 2. One can observe that the risk measures loss probability and mean shortfall loss exhibit again a very similar graphical pattern (see Figs. 18 and 19 in Appendix 2).

Considering the set of alternatives $A = \{C_{1T}, C_{2T}, C_{1D}, C_{2D}, C_{1W}, C_{2W}\}$, the corresponding risk ranking matrix, containing the rank orders for the six clusters, with respect to the risk measures R_1-R_6, where the rank order 1 corresponds to the lowest risk and 6 to the highest risk, is given by:

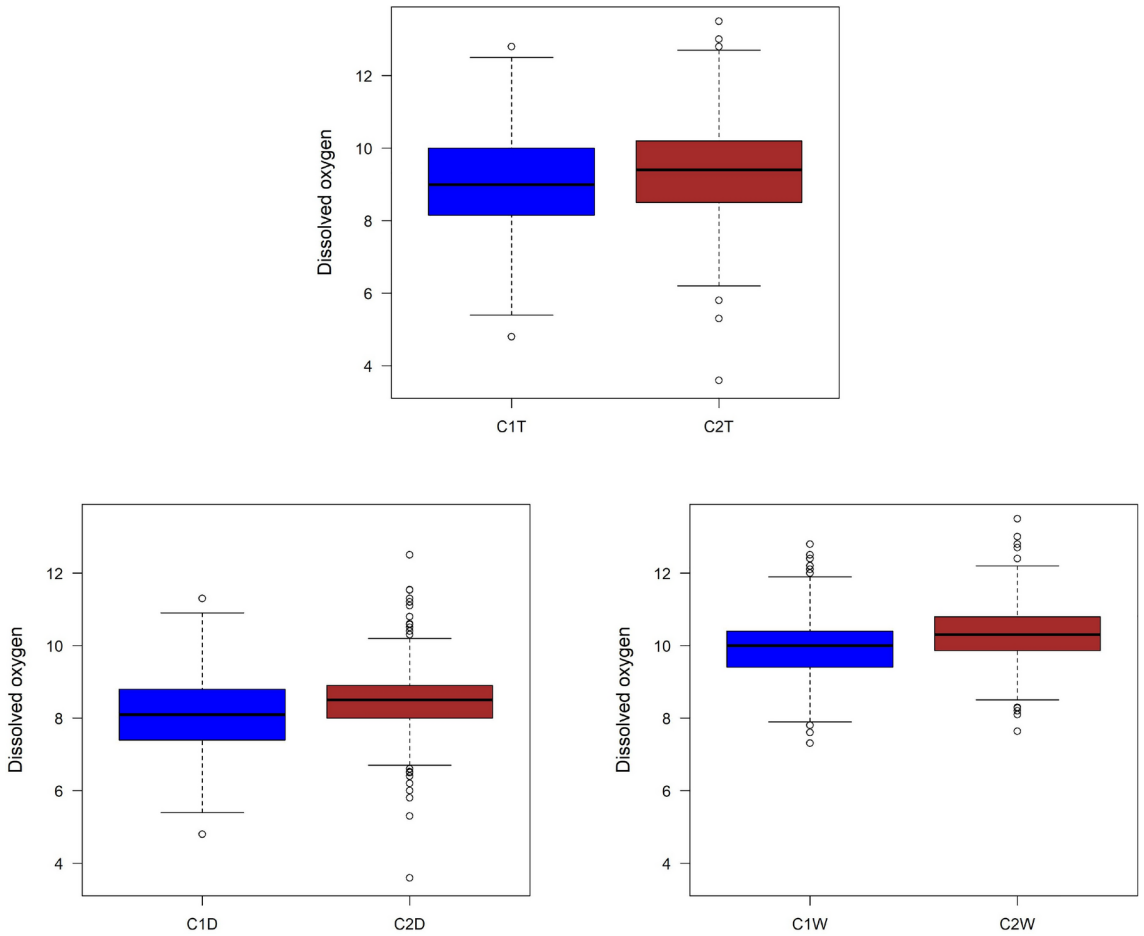

Fig. 6 Boxplots for DO, considering all the data (top), the data from the dry period (left) and the data from the wet period (right)

Table 7 Risk measures per cluster for DO

Period	Cluster	R_1	R_2	R_3	R_4	R_5	R_6
Total	C_{1T}	9.0423	1.5028	0.2253	2.4113	0.1424	10.9412
	C_{2T}	9.3753	1.4051	0.1389	2.3773	0.0650	10.9663
Dry	C_{1D}	8.1138	0.9779	0.4653	2.2084	0.3418	9.5444
	C_{2D}	8.4688	0.6973	0.2951	2.0367	0.1380	10.0270
Wet	C_{1W}	9.9244	0.6395	0.0081	2.3372	0.0021	11.1976
	C_{2W}	10.3526	0.6398	0.0023	2.3061	0.0008	11.7878

R_1–mean, R_2–variance, R_3–loss probability, R_4–entropy, R_5–mean shortfall, R_6–value at Risk

Table 8 Borda score, final Borda ranking (BR), Copeland score and final Copeland ranking (CR) for DO clusters

	A_i					
	C_{1T}	C_{2T}	C_{1D}	C_{2D}	C_{1W}	C_{2W}
BS	29	22	30	23	13	9
BR	5	3	6	4	2	1
CS	16	2	18	4	-16	-24
CR	5	3	6	4	2	1

$$RM = \begin{array}{c} \\ C_{1T} \\ C_{2T} \\ C_{1D} \\ C_{2D} \\ C_{1W} \\ C_{2W} \end{array} \begin{array}{c} \begin{array}{cccccc} R_1 & R_2 & R_3 & R_4 & R_5 & R_6 \end{array} \\ \begin{pmatrix} 4 & 6 & 4 & 6 & 5 & 4 \\ 3 & 5 & 3 & 5 & 3 & 3 \\ 6 & 4 & 6 & 2 & 6 & 6 \\ 5 & 3 & 5 & 1 & 4 & 5 \\ 2 & 1 & 2 & 4 & 2 & 2 \\ 1 & 2 & 1 & 3 & 1 & 1 \end{pmatrix} \end{array}.$$

The results indicate that the clusters of the wet period exhibit in fact lower risks with respect to almost all risk measures than the clusters of the dry period and of the total period, where the cluster of the wet period C_{2W} achieves the lowest risk with four risk measures. On the other hand, the cluster of the dry period C_{1D} seems to be the riskiest one, since it attains the highest risk with four risk measures. One can observe that the risk measures R_1 (mean), R_3 (loss probability) and R_6 (VaR) classify the six clusters in the same way, since $RM_{J=1} = RM_{J=3} = RM_{J=6}$. The risk measures R_2 (variance) and R_4 (entropy) assign both the same and the highest rank orders to C_{1T} and to C_{2T}, these having thus the highest uncertainty, however the remaining clusters are classified differently with these risk measures. Calculating the Borda scores and Copeland scores, using (8) and (10), for all clusters, permits aggregating the rankings as follows (see Table 8).According to the final aggregated Borda ranking and Copeland ranking, the clusters are classified from the highest to the lowest risk as follows:

$$\text{BR,CR}: \; C_{1D} \succ C_{1T} \succ C_{2D} \succ C_{2T} \succ C_{1W} \succ C_{2W},$$

so that C_{1D} is the riskiest cluster and C_{2W} is the less riskiest one. One can observe that $C_{1D} \subset C_{1T}$ (see (14) and (15)). The six stations contained in C_{1D}: COR, MEL, MOL, PEN, SEM, TAM, contribute therefore also to the high risk identified in C_{1T}, which is the second riskiest cluster. This means that the six stations in C_{1D} are among the 18 stations the most riskiest stations for water pollution due to lower DO concentrations in the dry period, however also in the total period (see Fig. 7, where the stations of the riskiest cluster are marked with a red box). The six stations of C_{2W} : AUR, BAT, MAR, SER, TEL, VAU, are the less riskiest stations for water pollution in terms of the DO concentration, especially in the wet period. These stations belong in the wet and in the total period to clusters classified with an intermediate risk, $C_{2W} \subset C_{2D}$ and $C_{2W} \subset C_{2T}$.

Now, it will be analysed if the elimination of the riskiest alternative will change the order of the remaining alternatives, i.e. if it results in the reversal of orders. Removing the riskiest cluster C_{1D} and recalculating the Borda and Copeland scores for the remaining clusters, one obtains the results in Table 9. One concludes that no rank reversals occur and the ordering of the alternatives remains consistent after eliminating the riskiest cluster.

In general, the results highlight in fact how important it is to distinguish the dry and wet periods in the water quality assessment. The analysis allows to classify the stations in terms of risk for water contamination depending on the DO concentrations and permits to identify station clusters with a poorer or better water quality, i.e. with respective higher or lower pollution risk, and this for each considered period (total observed period, dry period and wet period).

3.3.2 Cluster analysis and risk assessment: conductivity

Applying the clustering technique to the conductivity data of the 18 sampling stations, considering the total period, the wet and the dry period, one obtains the following station clusters:

Fig. 7 Spatial distribution of the riskiest stations according to DO

Table 9 Borda score, final Borda ranking (BR), Copeland score and final Copeland ranking (CR) for DO clusters after elimination of the riskiest cluster

	A_i				
	C_{1T}	C_{2T}	C_{2D}	C_{1W}	C_{2W}
BS	27	20	23	12	8
BR	5	3	4	2	1
CS	18	4	10	−12	−20
CR	5	3	4	2	1

$$C_{2D} = \{\text{COR}, \text{MEL}, \text{MOL}, \text{TAM}\} \tag{20}$$

$$C_{2T} = \{\text{COR}, \text{MEL}, \text{MOL}, \text{TAM}\} \tag{21}$$

$$C_{2W} = \{\text{COR}, \text{MEL}, \text{MOL}, \text{TAM}, \text{BAS}, \text{LON}\} \tag{22}$$

$$C_{1D} = \{\text{BAS}, \text{LON}, \text{AUR}, \text{ALV}, \text{BAT}, \text{CAM}, \text{MAR}, \text{PEN}, \text{RAN}, \text{SEM}, \text{SER}, \text{SOR}, \text{TEL}, \text{VAU}\} \tag{23}$$

$$C_{1T} = \{\text{BAS}, \text{LON}, \text{AUR}, \text{ALV}, \text{BAT}, \text{CAM}, \text{MAR}, \text{PEN}, \text{RAN}, \text{SEM}, \text{SER}, \text{SOR}, \text{TEL}, \text{VAU}\} \tag{24}$$

$$C_{1W} = \{\text{AUR}, \text{ALV}, \text{BAT}, \text{CAM}, \text{MAR}, \text{PEN}, \text{RAN}, \text{SEM}, \text{SER}, \text{SOR}, \text{TEL}, \text{VAU}\}, \tag{25}$$

where C_{1D} and C_{2D} denote the clusters in the dry period, C_{1W} and C_{2W}, the clusters in the wet period, and C_{1T} and C_{2T} represent the clusters in the total period. The dendrograms are presented in Fig. 8. As can be seen, the obtained clusters are identical when considering the total observed period or the dry period: $C_{1T} = C_{1D}$, $C_{2T} = C_{2D}$, but they are different in the wet period.

The descriptive statistics of the conductivity values are presented in Table 16. The analysis of the results, represented through the boxplots for each cluster in Fig. 9, shows that the clusters C_{2T}, C_{2D} and C_{2W} have slightly higher values and higher interquartile values when compared to C_{1T}, C_{1D}, and C_{1W}, respectively. In addition, it can be seen that the conductivity values are lower in the wet season.

Now, the aim is to evaluate the risk of water pollution, considering the conductivity values, for the clusters formed in the different time periods. The estimates for the risk measures R_1-R_6 (see Sect. 2.1), were calculated for each cluster based on a frequency distribution determined from the data sample of clusters specific conductivity measurements. The results are displayed in Table 10 and the corresponding graphical representations can be found in Figs. 20–22 in Appendix 2. One can notice a similar graphical pattern in the plots of the mean, variance, loss probability, mean excess loss and value at risk, in the sense that the clusters C_{2T}, C_{2D} and C_{2W} are characterized by higher risk values, compared to the complementary clusters.

Considering the set of alternatives $A = \{C_{1T}, C_{2T}, C_{1D}, C_{2D}, C_{1W}, C_{2W}\}$, the risk ranking matrix, containing the rank orders for the six clusters with respect to the risk measures R_1-R_6, where the rank order 1 corresponds to the lowest risk and 6 to the highest risk, is given by:

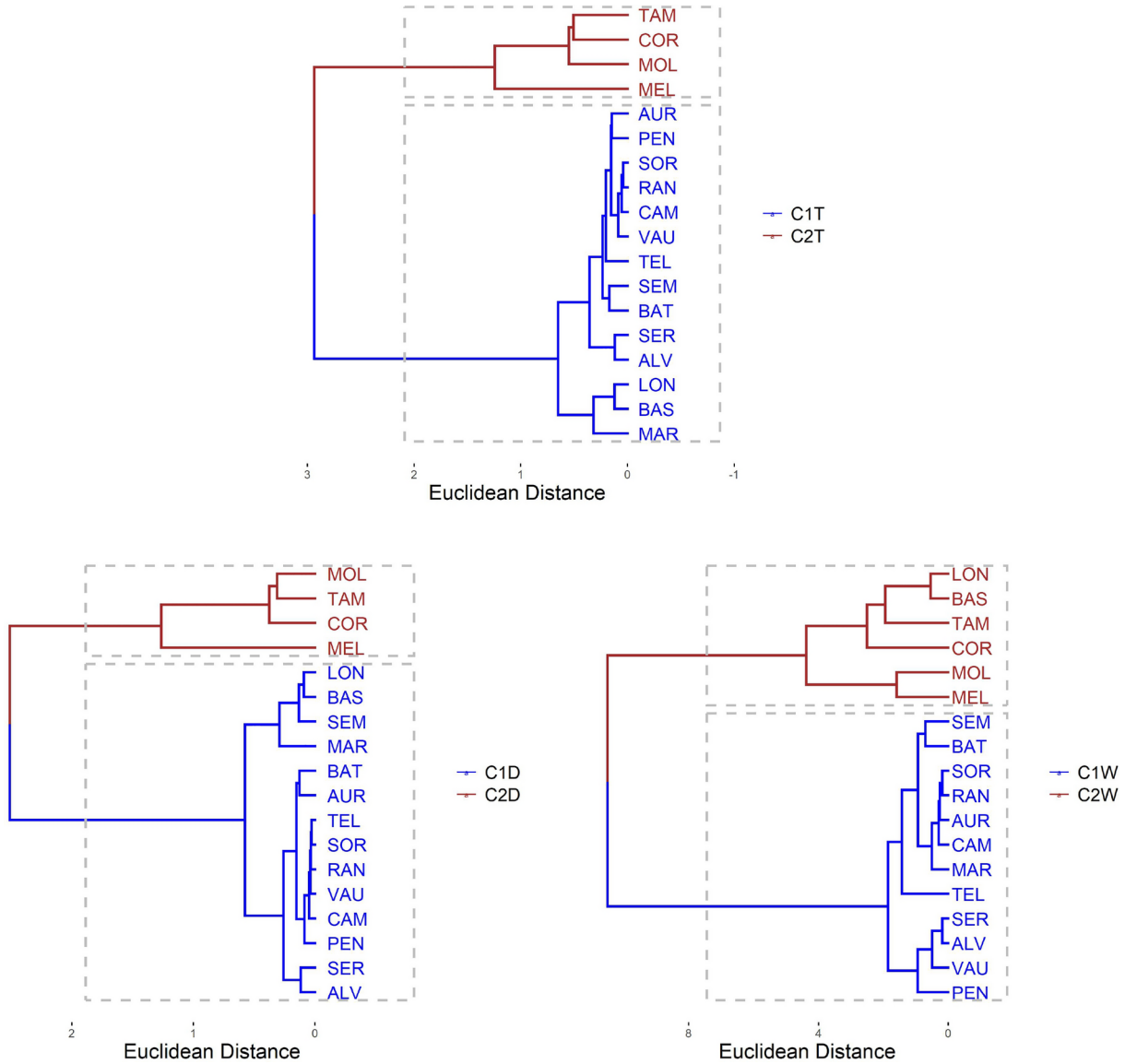

Fig. 8 Dendrograms of the conductivity variable, considering all the data (top), the data from the dry period (left) and the data from the wet period (right)

$$RM = \begin{array}{c} \\ C_{1T} \\ C_{2T} \\ C_{1D} \\ C_{2D} \\ C_{1W} \\ C_{2W} \end{array} \begin{array}{c} \begin{array}{cccccc} R_1 & R_2 & R_3 & R_4 & R_5 & R_6 \end{array} \\ \begin{pmatrix} 2 & 3 & 1 & 2 & 2 & 3 \\ 5 & 5 & 6 & 4 & 5 & 4 \\ 3 & 2 & 2 & 1 & 3 & 1 \\ 6 & 6 & 5 & 3 & 6 & 6 \\ 1 & 1 & 3 & 5 & 1 & 2 \\ 4 & 4 & 4 & 6 & 4 & 5 \end{pmatrix} \end{array}.$$

According to the orderings in the risk ranking matrix, one can affirm that C_{2D} attains the highest risk orders with almost all risk measures. However, one can not definitively conclude that the clusters of the wet period have in overall lower risks than those of the total and of the dry periods. Note that C_{2W} exhibits higher risks than C_{1T} and C_{1D} and, although C_{1W} achieves the lowest risk orders with R_1, R_2 and R_5, the cluster C_{1D} of the dry period outperforms C_{1W} according to R_3, R_4 and R_6. One can observe that the rankings corresponding to the risk measures R_1 (mean) and R_5 (mean excess loss) are the same, since $RM_{J=1} = RM_{J=5}$. The other risk measures lead to distinct rankings. In order to determine the final aggregated ranking of the clusters, the next step consists in calculating the Borda scores and Copeland scores for all clusters (see Table 11). The aggregated Borda ranking and Copeland ranking lead to the following classification of clusters from the highest to the lowest risk (see Table 10):

$$\text{BR,CR}: \; C_{2D} \succ C_{2T} \succ C_{2W} \succ C_{1T} \sim C_{1W} \succ C_{1D}.$$

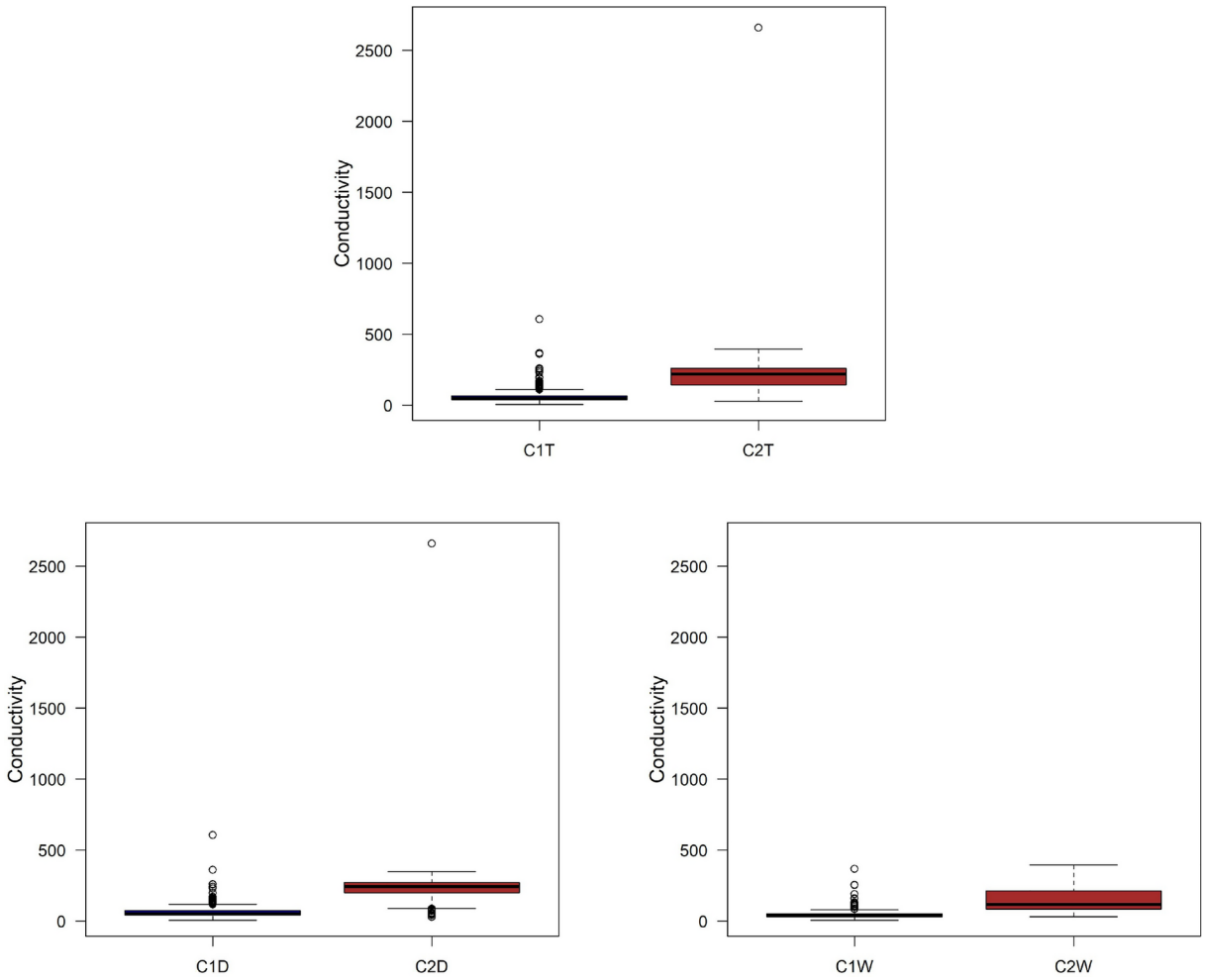

Fig. 9 Boxplots for Conductivity, considering all the data (top), the data from the dry period (left) and the data from the wet period (right)

Table 10 Risk measures per cluster for conductivity

Period	Cluster	R_1	R_2	R_3	R_4	R_5	R_6
total	C_{1T}	55.8027	737.5175	0.0015	0.2982	0.2898	123.1281
	C_{2T}	203.7632	16263.2320	0.5538	1.1571	9.3486	253.0465
dry	C_{1D}	60.5958	649.7241	0.0030	0.1234	0.4745	57.3406
	C_{2D}	238.1878	24145.3485	0.3958	0.9689	22.8556	286.9240
wet	C_{1W}	43.2271	596.7730	0.0035	1.3571	0.1472	68.7257
	C_{2W}	145.4605	5555.632	0.0810	2.5162	3.2942	267.6561

R_1–mean, R_2–variance, R_3–loss probability, R_4–entropy, R_5–mean excess loss, R_6–value at risk)

Table 11 Borda score, final Borda ranking (BR), Copeland score and final Copeland ranking (CR) for conductivity clusters

	A_i					
	C_{1T}	C_{2T}	C_{1D}	C_{2D}	C_{1W}	C_{2W}
BS	13	29	12	32	13	27
BR	2/3	5	1	6	2/3	4
CS	-16	16	-17	22	-16	12
CR	2/3	5	1	6	2/3	4

Fig. 10 Spatial distribution of the riskiest stations according to conductivity

Table 12 Borda score, final Borda ranking (BR), Copeland score and final Copeland ranking (CR) for conductivity clusters after elimination of the riskiest cluster

	A_i				
	C_{1T}	C_{2T}	C_{1D}	C_{1W}	C_{2W}
BS	13	27	12	12	26
BR	3	5	1/2	1/2	4
CS	-10	18	-22	-12	16
CR	3	5	1	2	4

The cluster C_{2D} is classified as the riskiest cluster and C_{1D} as the less riskiest one. One can observe that there is a tie between C_{1W} and C_{1T}. Applying the tie breaking rule, one gets $C_{1T} \succ C_{1W}$, so that the final ordering reads:

$$\text{BR,CR}: \; C_{2D} \succ C_{2T} \succ C_{2W} \succ C_{1T} \succ C_{1W} \succ C_{1D}.$$

Considering the three riskiest clusters, one has $C_{2D} = C_{2T} \subset C_{2W}$ (cf.(20)-(22)). One can therefore conclude that the four stations contained in C_{2D}, namely COR, MEL, MOL, TAM, present a high risk for water pollution taking into account the conductivity values (see Fig. 10, where the stations of the riskiest cluster are marked with a red box). On the contrary, the stations of $C_{1D} = C_{1T}$ (the complementary cluster of C_{2D}) form the less riskiest set for water pollution.

For studying the stability of the ranking considering the scenario of the worst alternative removal, the riskiest cluster C_{2D} is now eliminated and the Borda and Copeland scores are recalculated for the remaining clusters. The results in Table 12 reveal that now the tie between the second and third less riskiest clusters C_{1T} and C_{1W}, observed in the previous rankings (see Table 11), disappears. The new Copeland ranking solves the tie, C_{1W} being classified as the second less riskiest and C_{1T} as the third less riskiest cluster. The Borda ranking, instead, introduces a tie between the less riskiest cluster C_{1D} and C_{1W} (that was previously classified with the tied order 2/3). However, rank reversals do not occur, merely a tie is solved or introduced without reversing the general ordering and leaving the orders of the remaining clusters unchanged so that the method remains in general sufficiently stable.

4 Conclusions

In this paper, a risk assessment method based on different risk measures has been proposed for evaluating the surface water quality of hydrological basins. It has been shown that this method has allowed for the determination of the months with the highest risk of water pollution in the year and the identification of the riskiest water station clusters. The approach consists of calculating the risk measures: mean, variance, loss probability, entropy, mean excess loss, and value at risk, for a water quality parameter and ranking then the alternatives (months or water station clusters) according to each risk measure. A ranking aggregation method establishes then a final risk order ranking.

Since the risk of water pollution is related to the possibility of a loss of water quality characteristics, that are described by concentrations of specific parameters, such as DO or conductivity, and that can lead to environmental damages, the quantification of risk in different ways through the

risk measures has proved to be efficient for the risk assessment purpose and enhances usual approaches, where risk is judged using only the mean. Some risk measures have the advantage of considering deviations from predefined parameter quality standards (the mean excess loss can capture the risk implied by extreme events, e.g. a very high conductivity level caused by a discharge) and the entropy and variance are adequate to describe the risk related uncertainties.

The method has been applied to monthly DO and conductivity measurement data from 18 sampling stations of the Douro River basin, recorded from January 2002 to December 2013. The risk assessment has allowed to identify the riskiest months and periods for water pollution in the year: July is the riskiest month concerning the DO concentration; August is the riskiest month regarding the conductivity levels; October has been classified as the third and fourth riskiest month in the year, suggesting that the risky dry period should range from May to October (since usually October is excluded from the dry period). The risk assessment permitted also to classify water station clusters in terms of risk and to identify the most critical water stations. A cluster analysis has been conducted to group the 18 stations for the dry, the wet, and the total period, considering the DO and conductivity measurements, into different clusters. Regarding the DO concentrations, the cluster consisting of the stations: COR, MEL, MOL, PEN, SEM, TAM, in the dry period was classified as the riskiest one. As for the conductivity levels, the stations: COR, MEL, MOL, TAM, formed the riskiest cluster in the dry and in the total period. Appropriate environmental practices should therefore be implemented to reduce pollution events and to improve and protect the aquatic environment of these stations.

In the future, time series forecasting models will be applied to the data to examine the ability of the risk measures to predict the risk of water pollution. Another aim is to apply the proposed methodology to more recent data and to study more water quality parameters.

Appendix 1: Statistical summaries

See Tables 13, 14, 15 and 16.

Table 13 Descriptive measures of the DO variable per month

	R	1Q	ME	M	3Q	SD	CV
January	8.100 — 12.700	10.000	10.400	10.455	10.800	0.650	6.221
February	8.600 — 13.500	10.100	10.500	10.597	10.900	0.747	7.045
March	8.230 — 12.800	9.700	10.093	10.095	10.500	0.701	6.940
April	7.640 — 11.700	9.280	9.650	9.676	10.100	0.667	6.897
May	6.500 — 11.530	8.600	9.000	9.021	9.500	0.761	8.435
June	6.150 — 11.300	7.877	8.450	8.379	8.900	0.833	9.946
July	5.300 — 10.800	7.500	8.050	8.018	8.500	0.811	10.117
August	4.800 — 10.600	7.516	8.000	8.002	8.414	0.788	9.849
September	5.400 — 11.100	7.700	8.204	8.131	8.600	0.846	10.410
October	3.600 — 12.500	8.000	8.600	8.551	9.100	0.993	11.617
November	7.600 — 11.320	8.900	9.400	9.446	10.000	0.777	8.222
December	7.310 — 12.800	9.600	10.100	10.134	10.600	0.806	7.957

R–range, 1Q–1st quartile, ME–median, M–mean, 3Q–3rd quartile, SD–standard deviation, CV–coefficient of variation

Table 14 Descriptive measures of the conductivity variable per month

	R	1Q	ME	M	3Q	SD	CV
January	6.500 — 395.000	38.000	50.000	73.464	84.000	64.700	88.070
February	5.000 — 366.000	36.000	49.500	75.062	87.850	65.758	87.606
March	5.000 — 385.000	36.817	51.000	75.862	81.100	70.167	92.494
April	5.000 — 367.000	37.750	50.633	78.141	86.400	72.205	92.405
May	5.000 — 323.000	38.516	49.750	80.320	88.200	75.405	93.880
June	8.000 — 2660.000	41.782	55.000	100.438	100.250	191.422	190.588
July	7.800 — 329.000	48.000	63.000	98.271	113.250	83.275	84.740
August	7.900 — 349.000	50.000	68.000	106.205	123.250	87.818	82.688
September	9.200 — 361.311	51.300	71.550	109.645	129.825	87.011	79.357
October	9.000 — 606.000	47.500	66.000	105.485	129.750	89.101	84.467
November	7.500 — 297.000	38.000	56.500	84.558	109.775	70.237	83.063
December	8.000 — 289.000	37.000	55.550	76.743	93.025	63.759	83.081

R–range, 1Q–1st quartile, ME–median, M–mean, 3Q–3rd quartile, SD–standard deviation, CV–coefficient of variation

Table 15 Descriptive measures of the DO variable, for the obtained clusters

Period	Cluster	R	1Q	ME	M	3Q	SD	CV
total	C_{1T}	4.800 — 12.800	8.150	9.000	9.042	10.000	1.233	13.638
	C_{2T}	3.600 — 13.500	8.500	9.400	9.375	0.200	1.193	12.722
dry	C_{1D}	4.800 — 11.300	7.400	8.100	8.114	8.800	0.998	12.303
	C_{2D}	3.600 — 12.500	8.000	8.500	8.469	8.900	0.845	9.975
wet	C_{1W}	7.310 — 12.800	9.400	10.000	9.924	10.400	0.805	8.113
	C_{2W}	7.640 — 13.500	9.870	10.300	10.350	10.800	0.806	7.782

R–range, 1Q–1st quartile, ME–median, M–mean, 3Q–3rd quartile, SD–standard deviation, CV–coefficient of variation

Table 16 Descriptive measures of the conductivity variable, for the obtained clusters

Period	Cluster	R	1Q	ME	M	3Q	SD	CV
total	C_{1T}	5.000 — 606.000	37.000	50.000	55.800	66.750	34.927	62.591
	C_{2T}	29.300 — 2660.000	144.300	220.000	203.800	259.600	130.106	63.852
dry	C_{1D}	5.000 — 606.000	42.000	53.600	60.600	72.030	38.073	62.830
	C_{2D}	29.300 — 2660.000	199.000	244.800	238.200	272.200	158.160	66.401
wet	C_{1W}	5.000 — 367.000	32.000	42.000	43.230	53.000	24.965	57.753
	C_{2W}	31.000 — 395.000	85.920	116.500	145.460	214.000	74.812	51.431

R–range, 1Q–1st quartile, ME-median, M-mean, 3Q–3rd quartile, SD–standard deviation, CV–coefficient of variation

Appendix 2: Graphical plots of risk measures

Figs. 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 and 22

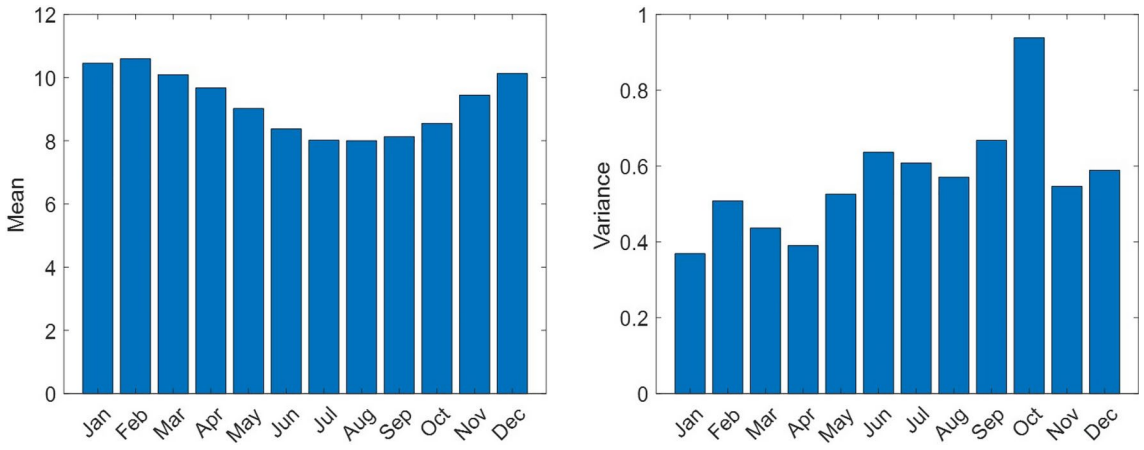

Fig. 11 Graphical plots of risk measures for DO: R_1–mean and R_2–variance

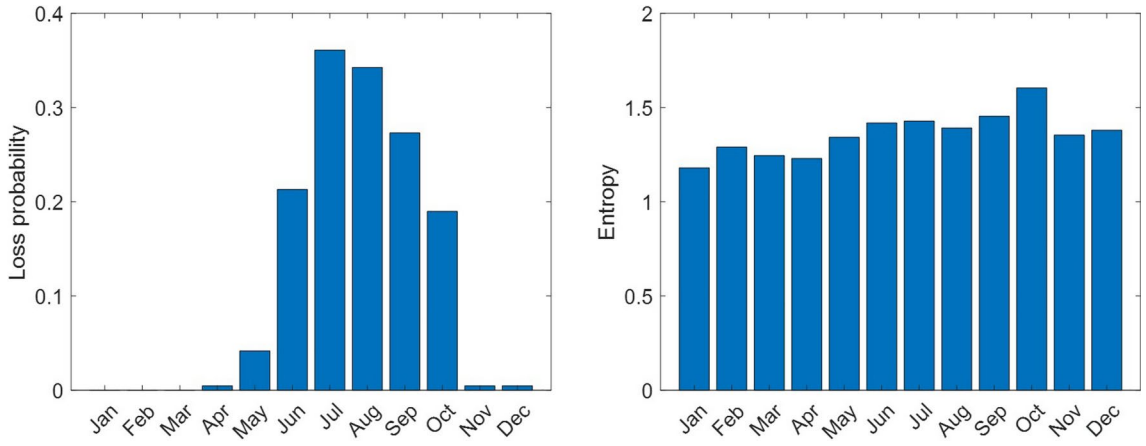

Fig. 12 Graphical plots of risk measures for DO: R_3–loss probability and R_4–entropy

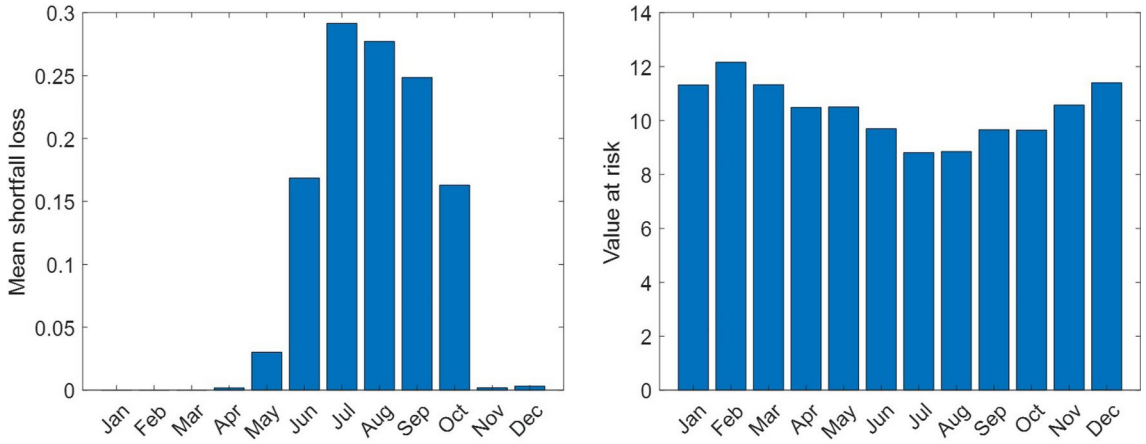

Fig. 13 Graphical plots of risk measures for DO: R_5–mean shortfall loss and R_6–value at risk

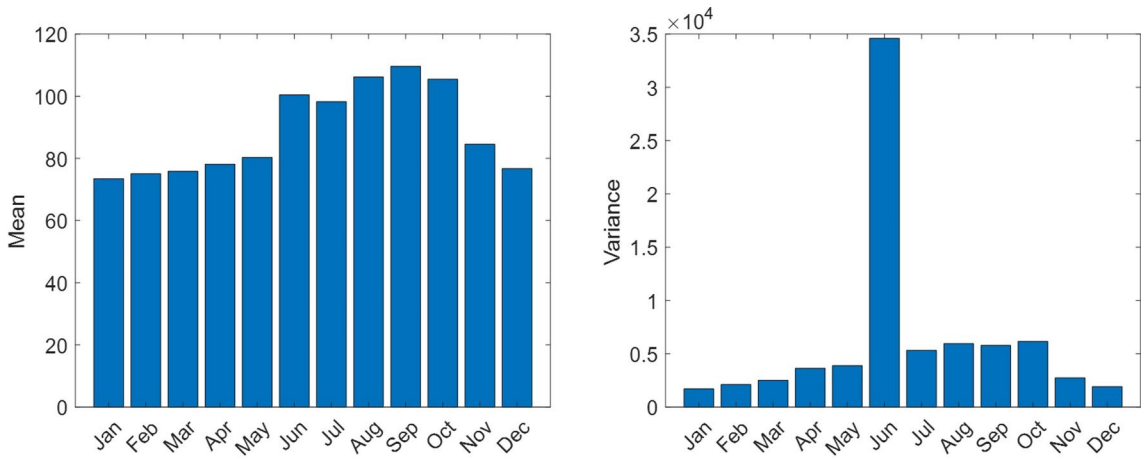

Fig. 14 Graphical plots of risk measures for conductivity: R_1–mean and R_2–variance

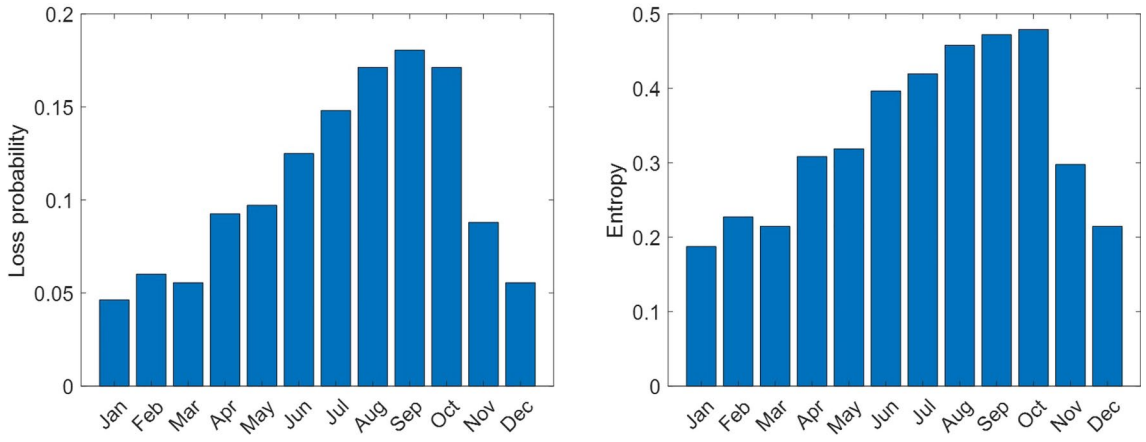

Fig. 15 Graphical plots of risk measures for conductivity: R_3–loss probability and R_4–entropy

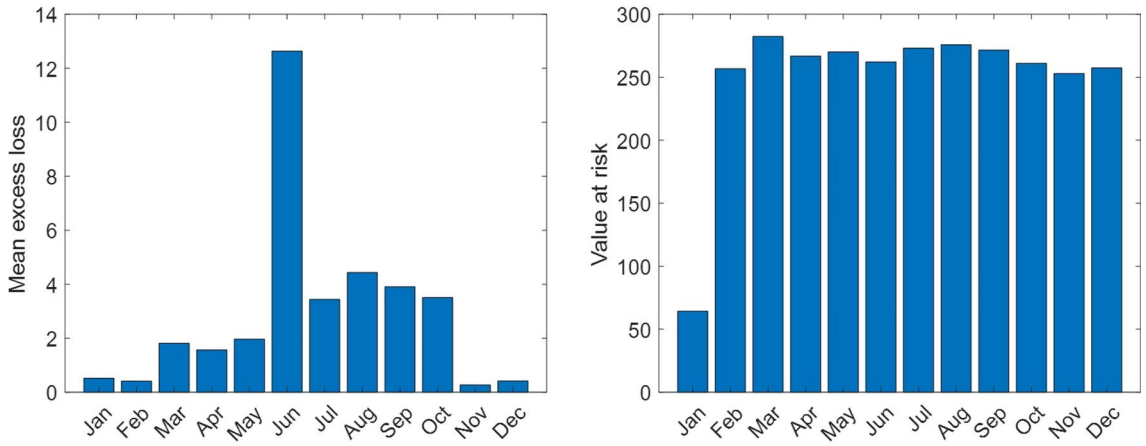

Fig. 16 Graphical plots of risk measures for conductivity: R_5–mean excess loss and R_6–value at risk

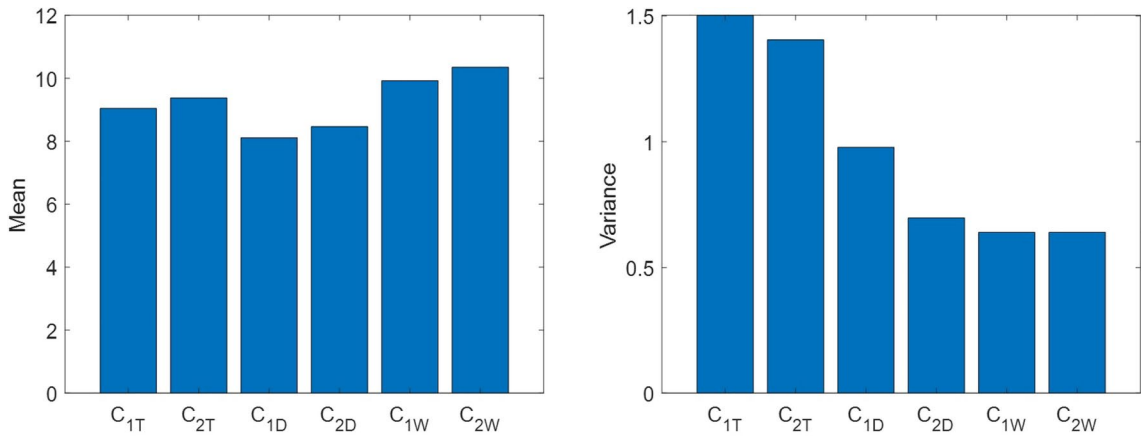

Fig. 17 Graphical plots of risk measures per cluster for DO: R_1–mean and R_2–variance

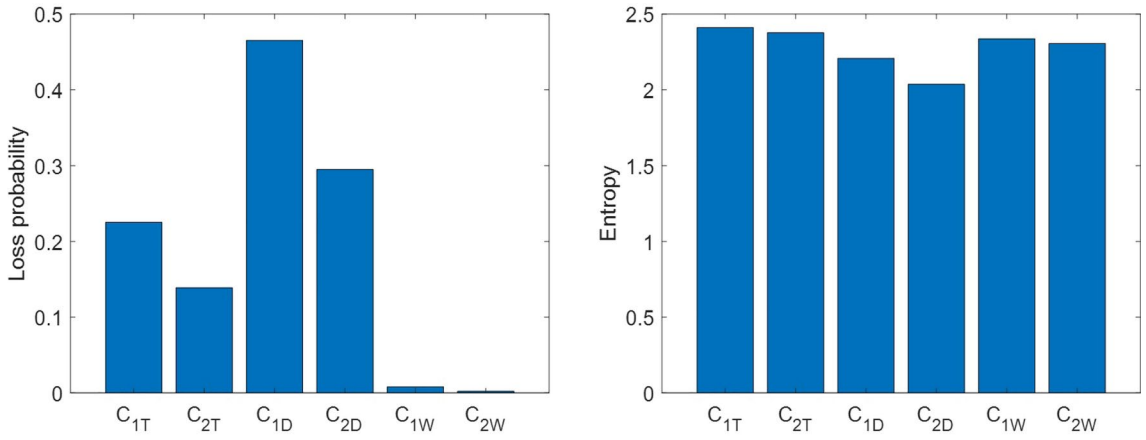

Fig. 18 Graphical plots of risk measures per cluster for DO: R_3–loss probability and R_4–entropy

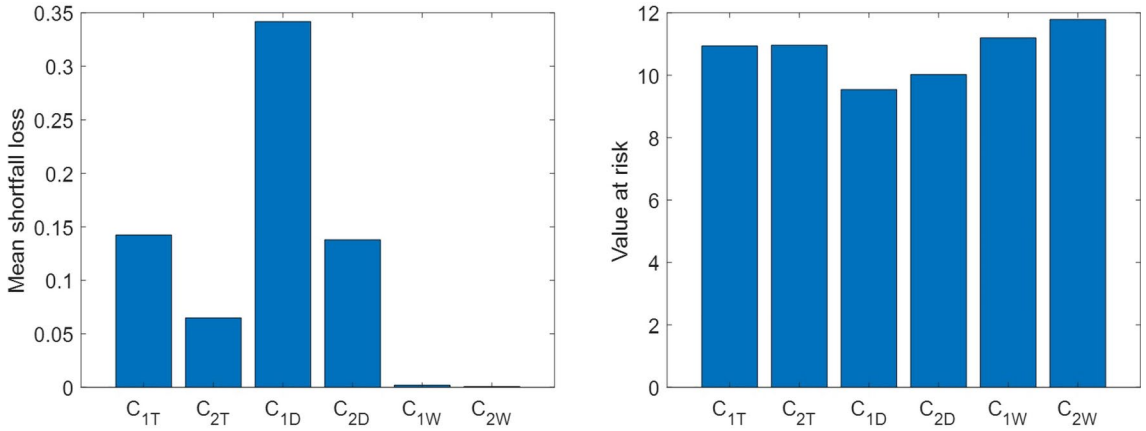

Fig. 19 Graphical plots of risk measures per cluster for DO: R_5–mean shortfall loss and R_6–value at risk

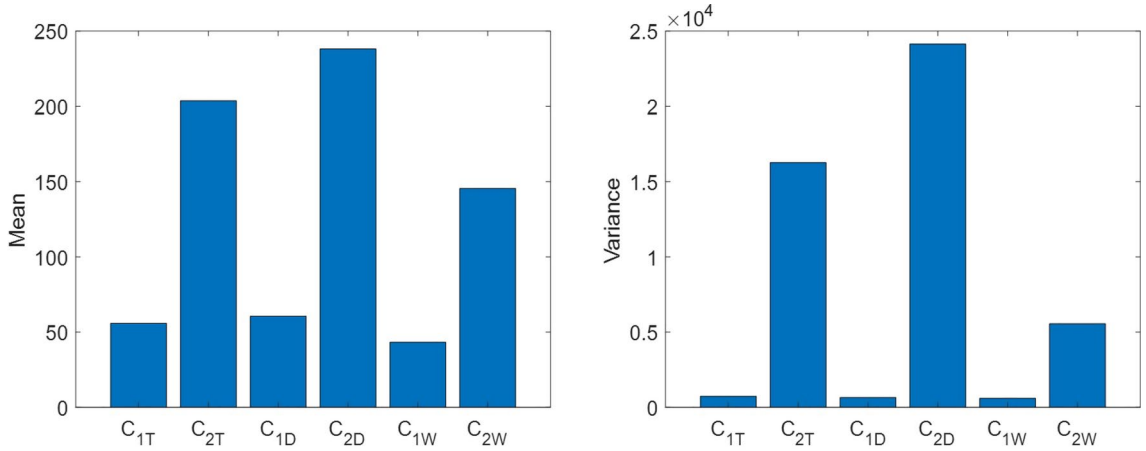

Fig. 20 Graphical plots of risk measures per cluster for conductivity: R_1–mean and R_2–variance

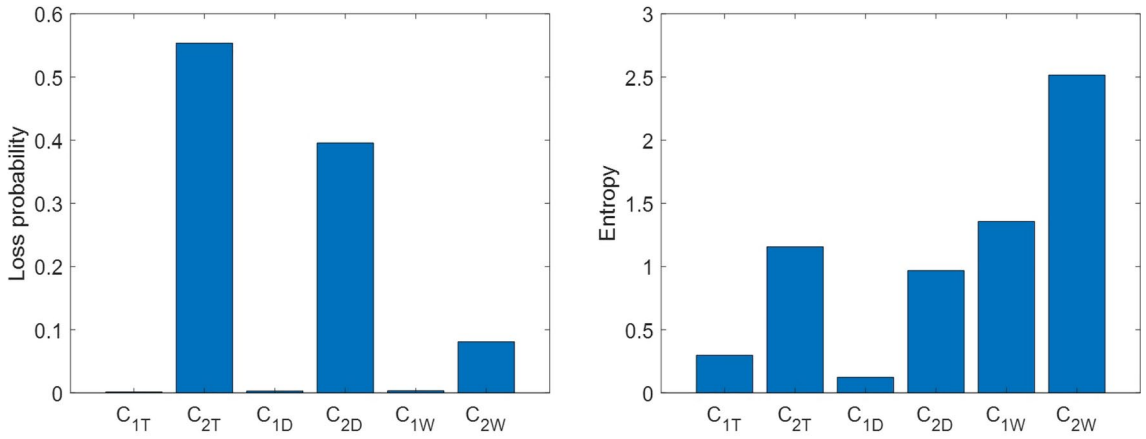

Fig. 21 Graphical plots of risk measures per cluster for conductivity: R_3–loss probability and R_4–entropy

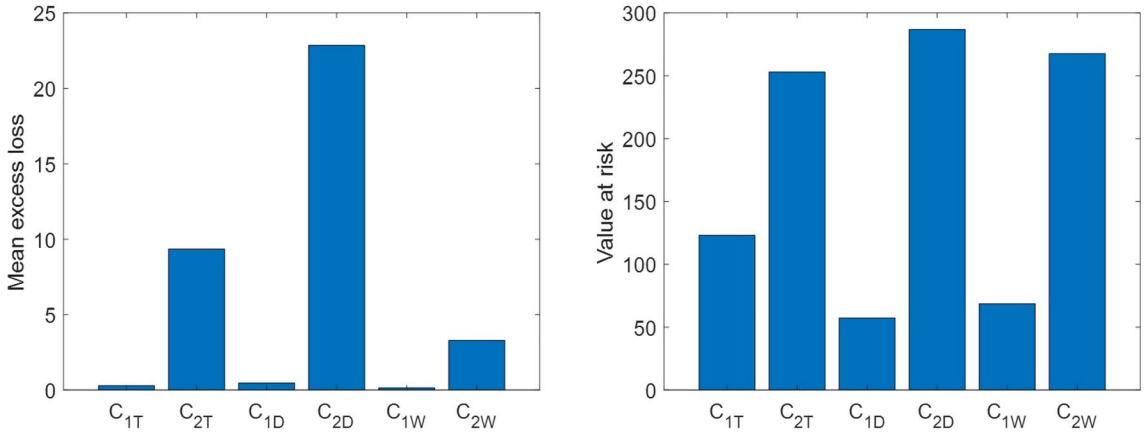

Fig. 22 Graphical plots of risk measures per cluster for conductivity: R_5–mean shortfall loss and R_6–value at risk

Acknowledgements Irene Brito and A. Manuela Gonçalves thank support from FCT through the projects UIDP/00013/2020 (https://doi.org/10.54499/UIDP/00013/2020) and UIDB/00013/2020 (https://doi.org/10.54499/UIDB/00013/2020). Ana Pedra thanks CMAT for the research fellowship (BI) UMINHO/BIM/2022/100. The authors thank Engineer José Vitorino, from the Portuguese Environment Agency (APA), for his availability and sharing of knowledge.

Author Contributions All authors elaborated and reviewed the manuscript.

Funding Open access funding provided by FCT|FCCN (b-on). This work was supported by FCT through the projects UIDP/00013/2020 https://doi.org/10.54499/UIDP/00013/2020andUIDB/00013/2020 https://doi.org/10.54499/UIDB/00013/2020.

Data availibility The data used in this research are available at https://snirh.apambiente.pt/ (Sistema Nacional de Informação de Recursos Hídricos) and further informations are provided in the manuscript.

Declarations

Conflict of interest The authors have no conflict of interest to declare.

References

APA (2021) Ficha técnica drh/deqa 2021 - critérios para a classificação das massas de Água. technical report, APA

Barrie A, Agodzo S, Frazer-Williams R et al (2023) A multivariate statistical approach and water quality index for water quality assessment for the rokel river in sierra leone. Heliyon 9(6):e16196. https://doi.org/10.1016/j.heliyon.2023.e16196

Borsuk ME, Stow CA, Reckhow KH (2002) Predicting the frequency of water quality standard violations: A probabilistic approach for tmdl development. Environ Sci Technol 36(10):2109–2115. https://doi.org/10.1021/es011246m

Brachinger HW, Weber M (1997) Risk as a primitive: a survey of measures of perceived risk. Oper Res Spektrum 19:235–250. https://doi.org/10.1007/s00477-012-0640-7

Brito I (2022) The normalized expected utility—entropy and variance model for decisions under risk. Int J Approx Reason 148:174–201. https://doi.org/10.1016/j.ijar.2022.06.005

Brito I (2023) A portfolio stock selection model based on expected utility, entropy and variance. Expert Syst Appl 213:118896. https://doi.org/10.1016/j.eswa.2022.118896

Brito I, Leão CP, Rodrigues MA (2022) Risk analysis and risk measures applied to the furniture industry. In: Machado J, Soares F, Trojanowska J et al (eds) Innovations in Mechanical Engineering. Springer International Publishing, Cham, pp 113–121

Cook WD (2006) Distance-based and ad hoc consensus models in ordinal preference ranking. Eur J Oper Res 172(2):369–385. https://doi.org/10.1016/j.ejor.2005.03.048

Custodio M, Penaloza R, Ochoa S et al (2023) Microbial and potentially toxic elements risk assessment in high Andean river water based on monte carlo simulation. Sci Rep 13:21473. https://doi.org/10.1038/s41598-023-48853-4

Ding J, Han D, Dezert J et al (2018) A new hierarchical ranking aggregation method. Inf Sci 453:168–185. https://doi.org/10.1016/j.ins.2018.04.041

Du J, Fang J, Xu W et al (2013) Analysis of dry/wet conditions using the standardized precipitation index and its potential usefulness for drought/flood monitoring in hunan province, china. Stoch Environ Res Risk Assess 27:377–387. https://doi.org/10.1007/s00477-012-0589-6

Fang Z, Hua C, He J et al (2023) Pollution assessment and source apportionment of heavy metal(loid)s in soil of Huangshui river basin, Ginghai province, china. Stoch Environ Res Risk Assess 37:4843–4855. https://doi.org/10.1007/s00477-023-02544-8

Ganoulis J (2009) Risk Analysis of Water Pollution. Wiley, Weinheim

Gibbons RD (2003) A statistical approach for performing water quality impairment assessments1. JAWRA J Am Water Resour Assoc 39(4):841–849. https://doi.org/10.1111/j.1752-1688.2003.tb04409.x

Gonçalves AM, Alpuim T (2011) Water quality monitoring using cluster analysis and linear models. Environmetrics 22:933–945. https://doi.org/10.1002/env.1112

Gonçalves AM, Costa M (2011) Clustering and forecasting of dissolved oxygen concentration on a river basin. Stoch Environ Res Risk Assess 25:151–163. https://doi.org/10.1007/s00477-010-0429-5

Gronewold AD, Borsuk ME (2009) A software tool for translating deterministic model results into probabilistic assessments of water quality standard compliance. Environ Model Softw 24(10):1257–1262. https://doi.org/10.1016/j.envsoft.2009.04.004

Hayashi M (2004) Temperature-electrical conductivity relation of water for environmental monitoring and geophysical data inversion. Environ Monitor Assess 96:119–128. https://doi.org/10.1023/B:EMAS.0000031719.83065.68

He Y, Ye J, Yang X (2015) Analysis of the spatio-temporal patterns of dry and wet conditions in the Huai river basin using the standardized precipitation index. Atmos Res 166:120–128. https://doi.org/10.1016/j.atmosres.2015.06.022

Huang J, Ho M, Du P (2011) Assessment of temporal and spatial variation of coastal water quality and source identification along Macau peninsula. Stoch Environ Res Risk Assess 25:353–361. https://doi.org/10.1007/s00477-010-0373-4

Kaas R, Goovaerts M, Dhaene J et al (2008) Modern Actuarial Risk Theory. Springer, Heidelberg

Kemeny J, Snell L (1962) Preference ranking: an axiomatic approach. Mathematical models in the social sciences pp 9–23

Klamler C (2005) The Copeland rule and Condorcet's principle. Econ Theory 25:745–749. https://doi.org/10.1007/s00199-004-0467-7

Klugman SA, Panjer HH, Willmot GE (2019) Loss Models: From Data to Decisions. Wiley, New York

Kumar A, Saxena P, Kisku G (2023) Heavy metal contamination of surface water and bed-sediment quality for ecological risk assessment of Gomti river, india. Stoch Environ Res Risk Assess 37:3243–3260. https://doi.org/10.1007/s00477-023-02447-8

Li D, Shi L, Dong Z et al (2019) Risk analysis of sudden water pollution in a plain river network system based on fuzzy-stochastic methods. Stoch Environ Res Risk Assess 33:359–374. https://doi.org/10.1007/s00477-018-01645-z

Li Z, Li Z, Tang X et al (2021) Distribution and risk assessment of toxic pollutants in surface water of the lower yellow river. Water, China. https://doi.org/10.3390/w13111582

López E, Patiño R, Vázquez-Sauceda ML et al (2020) Water quality and ecological risk assessment of intermittent streamflow through mining and urban areas of San marcos river sub-basin, mexico. Environ Nanotechnol Monitor Manag 14:100369. https://doi.org/10.1016/j.enmm.2020.100369

Milligan GW, Cooper MC (1988) A study of standardization of variables in cluster analysis. J Classif 5:181–204. https://doi.org/10.1007/BF01897163

Moritz S, SradÃ¡ A, Bartz-Beielstein T, et al (2015) Comparison of different methods for univariate time series imputation in R. Research Paper pp 1–20

Opperman JJ, Camargo RR, Laporte-Bisquit A et al (2022) Using the WWF water risk filter to screen existing and projected hydropower projects for climate and biodiversity risks. Water. https://doi.org/10.3390/w14050721

Rai SP, Sharma N, Lohani A (2014) Risk assessment for transboundary rivers using fuzzy synthetic evaluation technique. J Hydrol 519:1551–1559. https://doi.org/10.1016/j.jhydrol.2014.08.060

Rajesh M, Rehana S (2022) Impact of climate change on river water temperature and dissolved oxygen: Indian riverine thermal regimes. Sci Rep 12:9222. https://doi.org/10.1038/s41598-022-12996-7

Rajwa-Kuligiewicz A, Bialik RJ, Rowiński PM (2015) Dissolved oxygen and water temperature dynamics in lowland rivers over various timescales. J Hydrol Hydromech 63(4):353–363. https://doi.org/10.1515/johh-2015-0041

Reckhow KH (2003) On the need for uncertainty assessment in tmdl modeling and implementation. J Water Resour Plan Manag 129(4):245–246. https://doi.org/10.1061/(ASCE)0733-9496(2003)129:4(245)

Rehana S, Rajulapati CR, Ghosh S et al (2020) Uncertainty quantification in water resource systems modeling: case studies from India. Water. https://doi.org/10.3390/w12061793

Rudolph A, Ahumada R, Pérez C (2002) Dissolved oxygen content as an index of water quality in San vicente bay, Chile. Environ Monitor Assess 78:89–100. https://doi.org/10.1023/A:1016140819487

Sandhu G, Weber O, Wood MO et al (2023) An interdisciplinary water risk assessment framework for sustainable water management in Ontario, Canada. Water Resour Res 59(5):2022032959. https://doi.org/10.1029/2022WR032959

Singh K, Dutta R, A.S. Kalamdhad A (2019) Information entropy as a tool in surface water quality assessment. Environ Earth Sci DOI-urlhttps://doi.org/10.1007/s12665-018-7998-x

SNIRH (2023) Sistema nacional de informação de recursos hídricos. https://snirh.apambiente.pt/, Accessed 15 March 2023

Talbot JD, House WA, Pethybridge AD (1990) Prediction of the temperature dependence of electrical conductance for river waters. Water Res 24(10):1295–1304. https://doi.org/10.1016/0043-1354(90)90055-B

Tejaswini K, George B, Mukhopadhyay S, et al (2023) Conductivity sensors for water quality monitoring: a brief review. In: Technological Solutions for Water Sustainability: Challenges and Prospects: Towards a Water-secure India. IWA Publishing, https://doi.org/10.2166/9781789063714_0213

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244. https://doi.org/10.1080/01621459.1963.10500845

Wu J, Cheng SP, He LY et al (2023) Assessing water quality in the pearl river for the last decade based on clustering: characteristic, evolution and policy implications. Water Res 244:120492. https://doi.org/10.1016/j.watres.2023.120492

WWF (2023) WWF water risk filter. https://riskfilter.org/water/explore/data-and-methods, Accessed 2 July 2024

Yang J, Feng Y, Qiu W (2017) Stock selection for portfolios using expected utility-entropy decision model. Entropy. https://doi.org/10.3390/e19100508

Zahid MA, de Swart H (2015) The Borda majority count. Inf Sci 295:429–440. https://doi.org/10.1016/j.ins.2014.10.044

Zeleňáková M, Kubiak-Wojcicka K, Weiss R et al (2021) Environmental risk assessment focused on water quality in the Laborec river watershed. Ecohydrol Hydrobiol 21(4):641–654. https://doi.org/10.1016/j.ecohyd.2021.06.002