A Stage-Based Approach to Allocating Water Quality Monitoring Stations Based on the WorldQual Model: The Jubba River as a Case Study

Miguel A. Vega-Rodríguez[a,*], Carlos J. Pérez[b], Klara Reder[c], Martina Flörke[d]

[a] *Escuela Politécnica, Universidad de Extremadura, Avda de la Universidad s/n, 10003 Cáceres, Spain*
[b] *Facultad de Veterinaria, Universidad de Extremadura, Avda de la Universidad s/n, 10003 Cáceres, Spain*
[c] *Fraunhofer Institute for Energy Economics and Energy System Technology, Königstor 59, 34119 Kassel, Germany*
[d] *Hydrological Engineering and Water Resources Management, Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany*

Abstract

Ensuring adequate freshwater quality is an important aspect of integrated environmental management and sustainable development. One contribution towards this end is to monitor the water quality of river basins. An important issue in constructing a water quality monitoring network is how to allocate the stations. This is usually done by using in situ measurements of pollutants together with other information. A stage-based optimization approach has been developed to find the optimal sites to allocate the monitoring stations. The proposed approach constructs a network in a sequence of stages without the need for in situ pollution measurements. Instead, it uses pollutant estimates from the WorldQual model together with other social and hydrological criteria. The approach is computationally efficient and provides an ordered list of stations that can be used to initialize or augment a water quality network. This is especially relevant for consideration by developing countries since, with this approach, they can get an overview of their river basins, and

*Corresponding author

Email addresses: mavega@unex.es (Miguel A. Vega-Rodríguez), carper@unex.es (Carlos J. Pérez), klara.reder@iee.fraunhofer.de (Klara Reder), Martina.Floerke@hydrology.ruhr-uni-bochum.de (Martina Flörke)

then prioritize the initial distributions of the networks. The approach was applied successfully to the 741 751 km^2 of the Jubba River basin, but it is applicable to river basins of any size.

Keywords: Monitoring network; Multi-objective optimization; River basin; Stage-based approach; Water quality; WorldQual model.

1. Introduction

While water covers more than two-thirds of the Earth's surface, only 2.5% of it is fresh. In developing countries, 80% of people have no access to potable water, and cholera is still present in more than 50 countries (Adu-Manu et al., 2017). Fresh water is therefore an invaluable resource which needs to be first monitored, and then properly maintained.

According to the United Nations Environment Programme (UNEP) report (UNEP, 2016), water pollution has worsened since the 1990s in many rivers in Latin America, Africa, and Asia. However, it is still possible to cut short further pollution and restore the quality of polluted rivers. In this context, there is a need to design and implement effective intervention programs to control pollution in river basins, and, when necessary, to restore them.

Water quality control is a complex issue that requires the use of support techniques to provide relevant information on water resource management. The need to monitor the water quality and budgetary constraints make it necessary to develop assessment tools that can allow efficient control of water quality. Adu-Manu et al. (2017) reviewed methods for water quality monitoring which ranged from traditional manual methods to more technologically advanced ones. More recently, Nguyen et al. (2019) presented a review of design methods for river water quality monitoring networks.

A relevant part of the design of water quality monitoring networks is to establish the location of the stations. A well-designed network is one that provides the greatest possible amount of information from the fewest number of stations, so that the overall cost of monitoring is less. Several authors have addressed this issue from different viewpoints and based on different information.

Park et al. (2006) considered a genetic algorithm in a geographic information system framework to allocate monitoring stations in river basins. An optimization problem was formulated by considering different planning objectives such as compliance with water quality standards through biochemical oxygen demand measurements in the dry season, observation of water use, tracking of sources of pollution, and examination of water quality changes. A different approach, also based on a genetic algorithm, allowed Telci et al. (2009) to address this allocation problem based on the minimization of the average detection time of contamination events and the maximization of the reliability of the monitoring system at the same time as keeping the number of stations to a minimum. Remote sensing images were used by Chang & Makkeasorn (2010) who proposed an approach based on "grey integer programming" to select the locations considering a number of biophysical parameters and a budgetary constraint. Liyanage et al. (2016) considered a genetic algorithm to optimize avoidance of water quality standards violations, the population affected, distance of the site from the nearest downstream water intake, and coverage of upstream area. A fitness function was defined as the weighted sum (with weights defined by experts) of the partial objective functions with the number of stations included. Puri et al. (2017) proposed a genetic algorithm to select monitoring stations using mean annual *E. coli* flux data from the Spatially Referenced Regression Model on Watershed Attributes. The objectives were to minimize the number of monitoring stations, to cover large mean annual *E. coli* fluxes, and to minimize the uncertainty in the flux estimates. Constraints related to monitoring critical locations were also included in a multi-objective optimization problem. Zhu et al. (2019) implemented a discrete particle swarm optimization procedure for the allocation of water quality monitoring stations based on minimum pollution detection time, maximum pollution detection probability, and maximum centrality of monitoring locations while allowing reservation of some particular locations. This was done by considering the reduction of redundant monitoring locations.

The water quality parameters most commonly monitored correspond to general physicochemical characteristics (Nguyen et al., 2019). Villas-Boas et al. (2017) considered thirteen water quality parameters based on in situ measurements that were analysed in the laboratory. They identified the most relevant parameters, and showed how some of them were redundant and could be removed without significant information loss. However, obtaining many in situ measurements is both very costly and difficult to implement in many river basins, so some approaches have been developed based on pollutant simulations given by hydrological models. Al-Khafaji & Abdulraheem (2017) proposed a deterministic algorithm for the determination of optimal water quality monitoring stations based on two-dimensional hydrodynamic and water quality simulation models that had been used to estimate the distribution of total dissolved solids. Pérez et al. (2017) considered estimates of pollutant parameters such as biochemical oxygen demand, faecal coliform bacteria, and total dissolved solids from the WorldQual model (Voß et al., 2012) to feed a multi-objective evolutionary algorithm implemented in the framework of a geographic information system.

Assessing global water quality issues requires a multi-pollutant modelling approach. Strokal et al. (2019) focused on the need to integrate information on sources of pollutants such as plastic debris, nutrients, chemicals, and pathogens, among others. Vital signs for water quality usually cover some water quality parameters such as dissolved oxygen, temperature, salinity (total dissolved solids or electric conductivity), pH, turbidity, and faecal coliforms, among others. Impairments of these can produce an impact to the flora and/or fauna for a given water body. McCaffrey (2012) described some water quality parameters and provided their acceptable ranges. They also focused on their effects, for example, the temperature of the water is relevant because the amount of oxygen that will dissolve in water increases as the temperature decreases. Strokal et al. (2019) illustrated the potential of multi-pollutant modelling for hotspot analyses, and discussed scientific challenges and future directions for multi-pollutant modelling.

Sparse-data scenarios are especially interesting, and some approaches have been developed for such situations. Bastidas et al. (2017) designed water quality monitoring networks via optimization techniques, geographic information system technology, and a "matter-element" analysis of 5-day biological oxygen demand and total suspended solids. Scenarios with and without historical water quality data were addressed. Alilou et al. (2019) proposed a multi-criteria evaluation method including the analytic network process and fuzzy logic to identify locations of sampling points based on a "total potential pollution score" calculated without water quality data as input. This approach prioritizes the best candidate sampling points, and can be applied in settings where water quality data are scarce.

Redesigning water quality monitoring networks has also been addressed in the scientific literature. Sabzipour et al. (2019) applied two geostatistical methods, ordinary kriging and sequential Gaussian simulation, to electrical conductivity and dissolved oxygen concentration data, showing that the stations involved could be relocated to achieve an optimized network. Bastidas et al. (2017) also proposed a methodological approach which they used to redesign water quality networks.

In the present paper, a stage-based approach to allocating water quality monitoring stations is described and applied. This approach is based on the WorldQual model, a continental-scale water quality model that has been developed to obtain simulations of some pollutant loadings and in-stream concentrations in river basins. With the WorldQual model, such pollutants as biochemical oxygen demand, faecal coliform bacteria, and total dissolved solids can be estimated for any river worldwide. Besides detecting areas with pollutants, other planning objectives can be optimized, such as population, hydrological categorization, and the number of stations. The algorithm was developed to allow progressive selection of the stations, and is based on multi-objective artificial bee colony (MOABC) optimization. It is able to find optimal solutions from a very large number of alternatives in a computationally efficient way. It can be used to establish an initial network for river basins with no pre-existing network or in situ measurements, and then also to subsequently increase the number of stations in the network. The proposed approach was applied to the Jubba River basin (eastern Africa) whose drainage area is 741 751 km^2.

The main advantages of this proposal are:

- It uses pollutant estimates from the WorldQual model to detect non-compliance areas.
- Its approach is one of stage-based multi-objective optimization (MOABC).
- It is able to efficiently reduce the number of possible solutions to allow just one optimal solution to be chosen.
- It can be used both to establish a new network and then to progressively augment it.
- It is valid for river basins worldwide of any size.

The remainder of this paper is organized as follows. Section 2 presents the motivation for the proposed approach. Section 3 describes the study area, the WorldQual model, the planning objectives, and the optimization method. Section 4 presents and discusses the experimental results, and finally, Section 5 presents the conclusions.

2. The Motivating Problem

The decision of where to allocate the monitoring stations is an important part of the process of designing a water quality network. Many river basins around the world have no water quality monitoring networks, and the water quality data available in many parts of the world is inadequate.

The Global Environment Monitoring System for freshwater (GEMS/ Water) programme[1] was established more than 40 years ago to collect global water quality data to assess the status and trends of global inland water quality. Under the auspices of the UNEP, the GEMS/Water programme involves the World Health Organization (WHO), the World Meteorological Organization (WMO), and the United Nations Educational, Scientific and Cultural Organization (UNESCO). This programme uses GEMStat as a database of surface and ground water quality data. According to this GEMStat database, 71 of the 110 river basins with data have a density of at most 0.5 stations/10 000 km^2. The average densities for Latin America, Asia, and Africa are, respectively, 0.3, 0.08, and 0.02 stations/10 000 km^2 for the time period 1990 to 2010. Therefore, the GEMStat database station density is far below the typical minimum densities of around 1.5 to 4 stations/10 000 km^2 of the river basins of Europe and the USA (UNEP, 2016).

There is no internationally accepted scientific standardized process for the design of water monitoring networks. WMO (1994) provided a recommended minimum network density for different physiographic units (coastal, mountains, interior plains, hilly, small islands, and polar/arid). Borden & Roy (2015), through the International Institute for Sustainable Development (IISD), summarized the WMO recommendations for the minimum number of stations in three physiographic units: flat (1 station/1000-2000 km^2), mountainous (1 station/300-1000 km^2), and arid (1 station/5000-20 000 km^2). However, it was made clear that the recommendations are general and the final network density should be based on particular criteria such as network objectives or availability of finance and other resources.

To analyse the impact of the size of rivers on how best to locate the stations, it would be useful to have a categorization of their basins in terms of size. Again, there is no consensus on a standard categorization in this sense. Nguyen et al. (2019) presented a categorization by river basin area with the following classes: small (<100 km^2), medium (100-1000 km^2), large

[1] http://www.unep.org/gemswater/

(1000-10 000 km^2), and very large (>10 000 km^2), with reference to Higgins et al. (2005), Noble & Cowx (2002), and European Parliament and Council (2000). Those authors (Nguyen et al., 2019) reported that most studies in the literature concerned very large rivers in high- to middle-income countries.

In Africa, Asia, and South America there are, respectively, 8104, 45 804, and 9926 river basins, with mean areas in km^2 ($\pm$ standard deviation) of 3712 $\pm$ 64 044, 1025 $\pm$ 28 178, and 1815 $\pm$ 66 295, respectively. Figure 1 displays the comparative box plots for the area distributions of the river basins. All three distributions are strongly positively skewed, with many small river basins and few large ones.

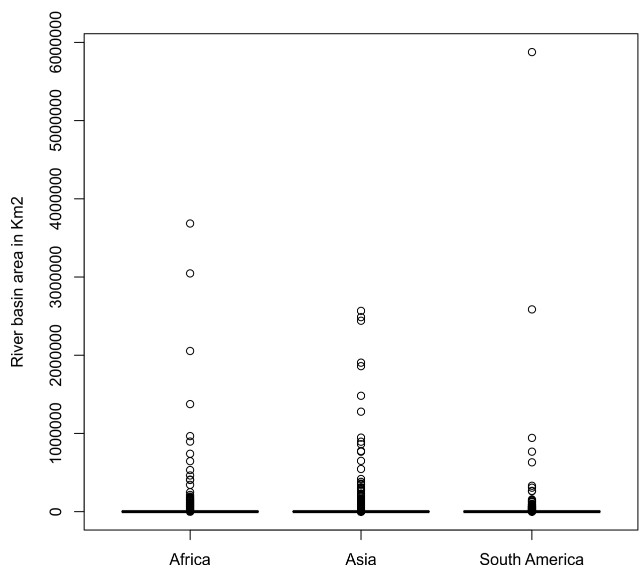


Figure 1: Box plots for the size distribution of the river basin areas in Africa, Asia, and South America.

According to UNEP (2016), in Africa, the average density of stations per 10 000 km^2 is 0.02. The Jubba River basin comprises parts of Somalia, Kenya, and Ethiopia, covering an area of 741 751 km^2. This value corresponds to the 99th percentile, which means that it is one of the largest rivers in Africa. The water of the Jubba River basin is very important for food production in this area. Water resources are strongly influenced by seasonal floods. A hydrometric network was operating in southern Somalia before the civil war in 1991, when data collection was abandoned until 2001 with the rebuilding of a much reduced network (Houghton-Carr et al., 2011). The objective of this rebuilt network was to measure rainfall, river flow, groundwater resources, land characteristics, degradation, and land suitability as well as to improve flood warning and flood management. To the best of the authors' knowledge, the network's stations do not provide accessible regular water quality information. Indeed, the GEMStat database does not report any water quality monitoring network in the Jubba River basin.

As an example, this situation motivates the need to develop tools that can help the authorities allocate water quality monitoring stations based on objective criteria that can be easily obtained without requiring actual in situ pollutant measurements. Using low-cost criteria such as pollutant estimates given by the WorldQual model to detect non-compliance areas may help in this task, and are applicable to river basins of any size anywhere in the world. If this is done in a progressive way, the approach could be useful both for constructing a new water quality network and for augmenting an existing network. While this is especially relevant for rivers where no water quality networks have been established, it can also be applied to progressively reconstruct old ones. Such progressive increase of a network also facilitates accommodation to different budgets.

3. Methods

3.1. Study area

The Jubba River is an important east African river located in southern Somalia. Its source is at the border with Ethiopia, and it flows directly south to empty into the Indian Ocean at Goobweyn. The two main tributaries, the Shebelle and Lagh Dera rivers, join the Jubba River close to its mouth. Balint et al. (2010) defined three catchment areas for this zone (Jubba, Shebelle, and Lagh Dera), although only the Jubba River has access to the sea. Together, these three catchments constitute the present work's study area, which we shall term the Jubba River basin. This basin drains a total of 741 751 km^2 and extends across parts of Somalia, Kenya, and Ethiopia (see Figure 2 for the location and distribution of the tributaries).

3.2. WorldQual model

The WorldQual model is a global scale water quality model (Voß et al., 2012). It is a module of WaterGAP3 (Water Global Assessment and Progno-

Figure 2: The Jubba-Shebelle-Lagh basin and its location in eastern Africa.

sis), which is a global water assessment model consisting of a global water use model and a global hydrology model (Alcamo et al., 2003; Verzano, 2009).

Voß et al. (2012) presented WorldQual, and illustrated its performance in applying it to model biological oxygen demand (BOD) and total dissolved solids (TDS) across Europe. The model was extended to model faecal coliform (FC) loadings for large European rivers and the resulting in-stream concentrations (Reder et al., 2015). It has also been applied to model FC and BOD in African rivers (Reder et al., 2014). Indeed, it is the only large-scale concentration model which has been applied to faecal coliform bacteria for various continents (Vermeulen et al., 2019). It has been validated and tested with a global sensitivity and uncertainty analysis (Reder et al., 2017).

The WorldQual model operates on monthly time steps and a 5-by-5 arc-minute grid spatial resolution (approximately 9-by-9 km at the equator). This resolution divides the Jubba River basin into 8710 cells of about 85 km^2 each in accordance with the drainage direction map of Lehner et al. (2008). Estimated concentrations of BOD, FC, and TDS were obtained from January 1990 to December 2010 on a monthly basis.

Figure 3 represents the entire Jubba River basin divided into main and secondary stretches based on Strahler stream order (Strahler, 1957). The streams of the river basin are represented as a mathematical tree, and the Strahler number is a numerical measure of their branching complexity. A stream with no children is a leaf, and its Strahler number is one. Such streams are regarded as secondary, and the remaining ones as main. Main channels (streams with Strahler number greater than one) are candidates for the allocation of stations.

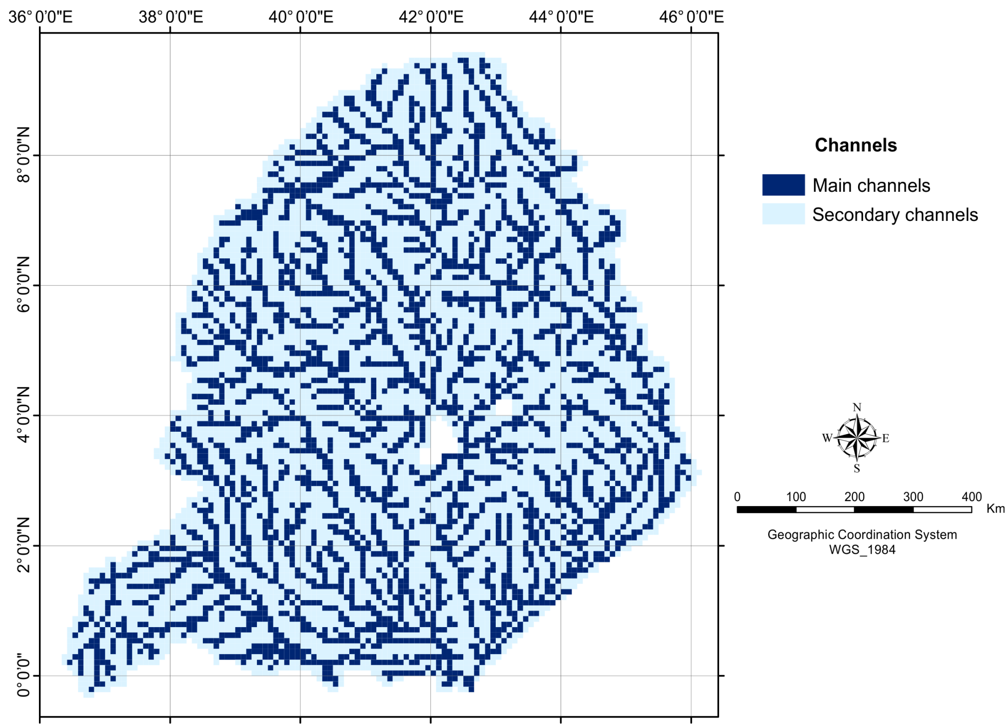


Figure 3: Grid representation of the Jubba River basin with main and secondary channels.

3.3. Planning objectives

In order to find macro-locations at which to optimally allocate water quality monitoring stations, planning objectives based on economic, environmental, social, and hydrological aspects are defined. Examples of the use of such planning objectives are found in Park et al. (2006), Liyanage et al. (2016), and Pérez et al. (2017).

Let X_{ij} denote a binary variable for the cell (i, j) in the river basin grid, with i being the index representing the row and j the index representing the column. This variable determines whether the cell (i, j) is $(X_{ij} = 1)$ or is not $(X_{ij} = 0)$ assigned to allocate a station. For the Jubba River basin, there are a total of 3097 candidate stations, which are labeled as main channels in Figure 3. The vector of binary components denoted by $\boldsymbol{x}$ represents the solution of the allocation problem.

A first criterion is to reduce the economic cost of building the water quality monitoring network. This is achieved by keeping the number of monitoring stations to a minimum. Mathematically, this is defined as minimizing the following objective function:

$$\phi_1(\boldsymbol{x}) = \sum_{i,j} X_{ij}. \tag{1}$$

Detecting lower compliance areas is the second criterion considered, which is related to an environmental aspect. This criterion is established so that the network shows the greatest potential capability to detect polluted areas. Mathematically, this is translated into maximizing the probabilities of detecting threshold violations. With the WorldQual model, the in-stream concentrations of BOD, FC, and TDS were estimated for each grid cell of the Jubba River basin on a monthly basis for a period of 21 years. These quantities are denoted for BOD, FC, and TDS, respectively, by $U_{ij}^{(t)}, V_{ij}^{(t)}$, and $W_{ij}^{(t)}$, for $t = 1, 2, \ldots T$, where $T = 252 = 12 \cdot 21$ is the total number of simulated measurements over the entire period for each pollutant. Values of BOD, FC, and TDS under 4 mg/l, 200 cfu/100 ml (colony-forming units per 100 ml), and 450 mg/l, respectively, are acceptable (UNEP, 2016). The probabilities of detection of threshold violations of BOD (U_{ij}), FC (V_{ij}), and TDS (W_{ij}) for each grid cell over the whole period of time are defined as:

$$P(U_{ij} > 4) = \frac{1}{T} \sum_{t=1}^{T} I[U_{ij}^{(t)} > 4], \tag{2}$$

$$P(V_{ij} > 200) = \frac{1}{T} \sum_{t=1}^{T} I[V_{ij}^{(t)} > 200], \tag{3}$$

$$P(W_{ij} > 450) = \frac{1}{T} \sum_{t=1}^{T} I[W_{ij}^{(t)} > 450], \tag{4}$$

where $I[\cdot]$ represents the indicator function.

The above probabilities are combined into a single value representing each cell, denoted Water Pollution Detection (WPD). This measure is defined for each cell as $D_{ij} = P(U_{ij} > 4) + P(V_{ij} > 200) + P(W_{ij} > 450)$. The greater the value of D_{ij}, the more pollution that cell contains. Therefore, cells with large values should be preferred for allocating stations in order to detect non-compliance areas. Figure 4 shows the WPD values in the Jubba River basin.

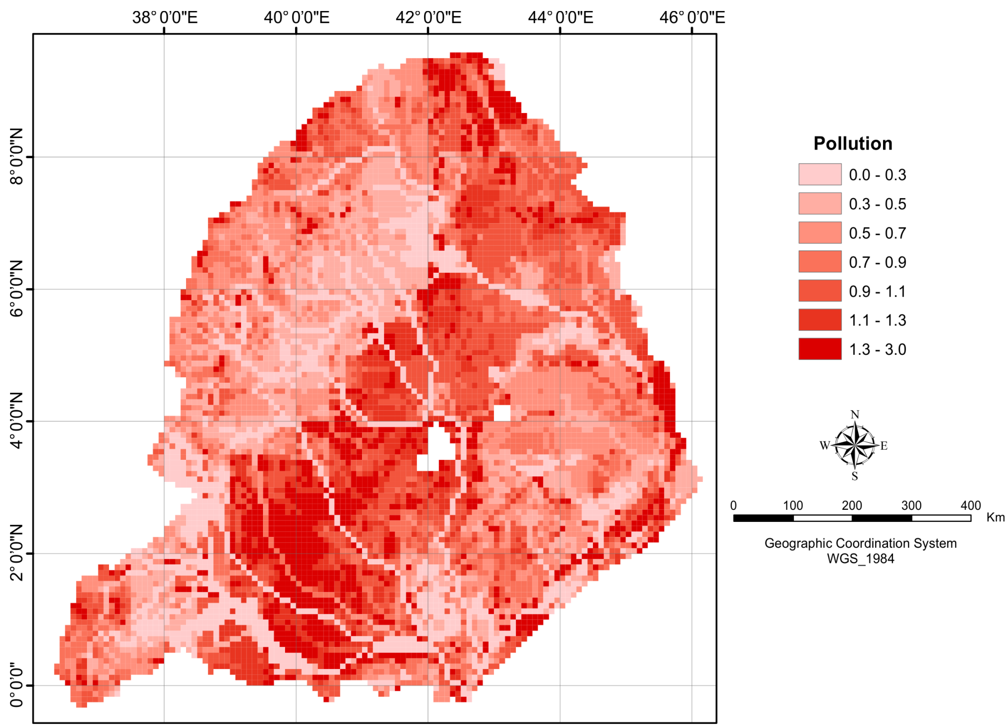


Figure 4: WPD for the cells in the Jubba River basin.

The objective function is defined as:

$$\phi_2(\boldsymbol{x}) = \sum_{i,j} D_{ij} \cdot X_{ij}, \tag{5}$$

which should be maximized.

The third criterion considers the population involved. The degree of protection of areas with large populations takes into account a social dimension. The areas with greater populations should be preferred for allocating monitoring stations over those with fewer people. This information is available in the WorldQual model, and the population count in each cell is denoted by C_{ij}. Figure 5 represents the population in each cell of the Jubba River basin.

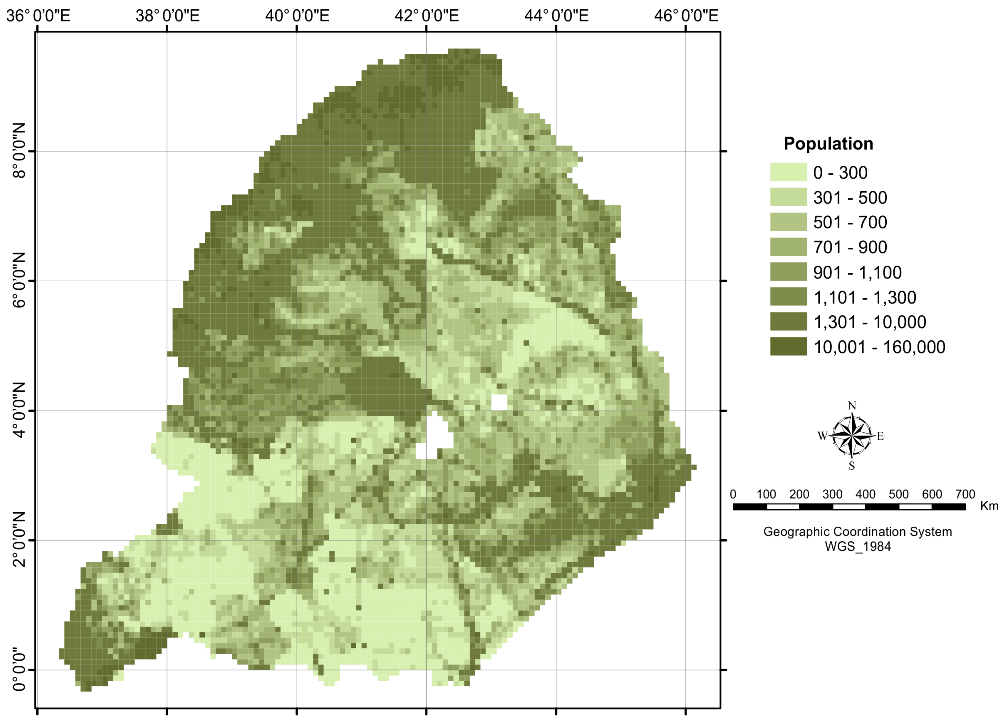


Figure 5: Population for the cells in the Jubba River basin.

Therefore, the objective function to maximize is:

$$\phi_3(\boldsymbol{x}) = \sum_{i,j} C_{ij} \cdot X_{ij}. \tag{6}$$

The last criterion is related to a hydrological aspect, involving a classification of the river basin stretches. The Strahler (or Horton–Strahler) number of the mathematical tree representing the river basin will be considered for this task (Horton, 1945; Strahler, 1957). The Strahler number for a node in that tree is just its number of children. It is a numerical measure of branching complexity, and is used to define stream size based on a hierarchy of tributaries. Let S_{ij} denote the variable representing the Strahler number of the cell (i, j). The greater the Strahler number, the more importance the cell has hydrologically. These numbers may be of great help to avoid the concentration of stations in areas of low hydrological importance. Figure 6 represents the Strahler number for the cells in the Jubba River basin.

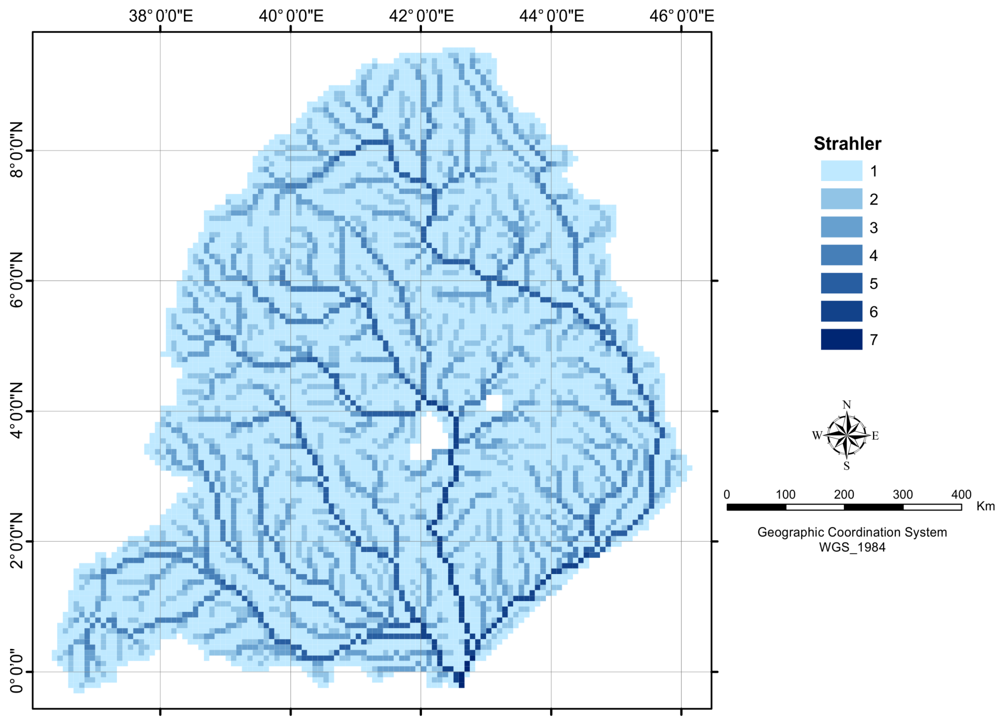


Figure 6: Strahler number for the cells in the Jubba River basin.

The objective function to maximize is:

$$\phi_4(\boldsymbol{x}) = \sum_{i,j} S_{ij} \cdot X_{ij}. \tag{7}$$

Besides the objective functions, two constraints are considered. The first is related to the maximum budget allowed. The total number of stations should be between a minimum m ($m \geq 1$) and a maximum M. The second is related to neighbouring cells. This constraint is aimed at avoiding touching cells from being selected.

Finally, the problem of how to allocate monitoring stations is defined as a multi-objective optimization problem in which the number of stations is minimized (Equation (1)) over a certain range of values while avoiding touching cells at the same time as maximizing the detection of non-compliance areas (Equation (5)), the population in the area (Equation (6)), and the hydrological importance (Equation (7)).

3.4. Stage-based multi-objective optimization algorithm

The optimization problem defined in the previous subsection contains multiple conflicting objectives. Multi-objective optimization is the natural way to solve it.

The Multi-Objective Artificial Bee Colony (MOABC) algorithm is an extension to multi-objective settings of the Artificial Bee Colony (ABC) algorithm proposed by Karaboga & Basturk (2007). MOABC is an evolutionary algorithm based on the intelligent foraging behaviour of honey bee swarms with a population defined as a colony with three groups of bees: (1) employed bees maintain the currently known solutions of the problem; (2) onlooker bees allow the exploitation of the best solutions found so far; and (3) scout bees allow the exploration of new solutions when some of the current solutions are exhausted (they cannot be further improved). This algorithm allows escape from local optima, and provides good approximations to the overall optimum. MOABC has been successfully applied in several contexts (see, e.g., Karaboga et al. (2014); Pérez et al. (2017); Huo & Liu (2018)).

In the present work, the MOABC algorithm was adapted to find optimal solutions in different stages by means of an efficient implementation that allows the allocation problem to be solved for river basins of any size. In spite of the fact that multi-objective optimization returns a set of non-dominated solutions, the proposed approach considers at each stage consensus stations which properly approach the overall optima. Algorithm 1 presents the pseudocode for the stage-based MOABC approach. A detailed explanation of the MOABC algorithm (lines 3-15) can be found in Pérez et al. (2017). In the present work, a stage-based version of MOABC is proposed. In this new version, the process is repeated for a maximum number of stages (line 2). Here, three stages will be considered. Each stage increases the number of stations used, taking into account the stations fixed in the previous stage. The first stage has no station fixed (line 1). Once MOABC has obtained the non-dominated solutions for that stage, they are saved (line 16) and are used to generate the consensus stations (line 17), i.e., the stations most used by all the non-dominated solutions. These consensus stations are saved as the exact solution for that stage (line 18). The consensus stations will then be used as fixed stations (line 19), adding more stations during the processing of the next stage.

The stage-based MOABC algorithm was configured with parameters that had been tested to work appropriately, i.e., a colony size of 50, a maximum

Algorithm 1 Stage-based MOABC pseudocode.

```
fixed_stations ← ∅
for stage = 1 to max_stages do
    non_dominated_solutions ← ∅
    #Problem-aware and random generation of the initial colony
    initial(fixed_stations)
    #Main steps of MOABC are repeated max. cycles or generations
    for cycle = 1 to max_cycles do
        send_employed_bees(fixed_stations)
        rank_and_crowding(colony_size)
        calculate_probabilities()
        send_onlooker_bees(fixed_stations)
        send_scout_bees(fixed_stations, cycle)
        rank_and_crowding(2 * colony_size)
        non_dominated_solutions ← export_colony()
    end for
    save_non_dominated_solutions(non_dominated_solutions, stage)
    consensus_stations ← get_consensus(non_dominated_solutions)
    save_consensus_stations(consensus_stations, stage)
    fixed_stations ← fixed_stations ∪ consensus_stations
end for
```

number of cycles of 3000, and a number of tries of 100 before taking a solution to be exhausted. Similar parameter values were proposed by the ABC authors[2]. Other values were tested for each of these parameters, but the best results were obtained with this configuration.

In order to ensure the statistical reliability of the results, the execution of the algorithm was repeated 31 times because of the stochastic nature of the proposed algorithm. The median results obtained from these 31 independent runs will be presented in the Results and Discussion section. This is a usual procedure in the optimization field since it avoids only the best execution being chosen and provides fairer information about the algorithm's performance (Birattari & Dorigo, 2007).

Finally, scatter plots will be used to represent the non-dominated solu-

[2]http://mf.erciyes.edu.tr/abc/

tions for each stage, and graphs and descriptive statistics to compare the objective values by stages.

4. Results and Discussion

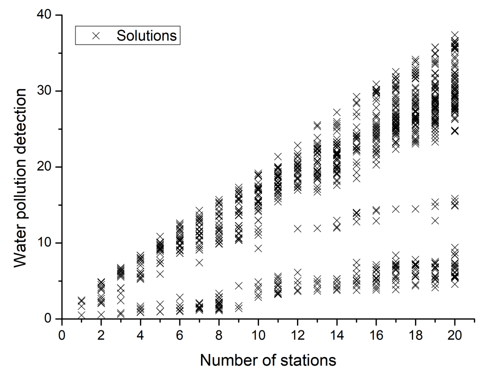
This section will present a detailed analysis of the solutions obtained in each stage. Then a comparative analysis of the results of the stages will be made, followed by a discussion of the stage-based selection of water monitoring stations.

4.1. Defining the stages

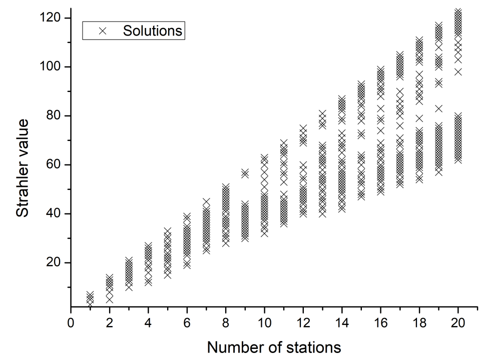
The Jubba River basin has a drainage area of 741 751 km^2 in a region with an arid climate. According to the minimum number of stations defined by WMO (WMO, 1994) and summarized by Borden & Roy (2015) of the IISD, arid regions should have at least 1 station/5000-20 000 km^2. This would mean between 37 and 148 water quality monitoring stations in the network. But, according to the GEMStat database, the average density for African river basins is 0.02 stations/10 000 km^2, which is very low compared to typical minimum densities of around 1.5 to 4 stations/10 000 km^2 of river basins in Europe and the USA (UNEP, 2016). These latter values for Europe and the USA would mean that there should be from 112 to 296 stations allocated to the Jubba River basin. Combining the two criteria, it would seem reasonable to establish a network with 112 stations for this large drainage area. Moreover, to be affordable, this should be performed in stages. A three-stage approach is proposed here, with the first comprising 20 stations, the second incorporating additional 30 stations, and the third adding the remaining 62 stations to the 50 stations already allocated in the two previous stages.

4.2. First stage: from 1 to 20 monitoring stations

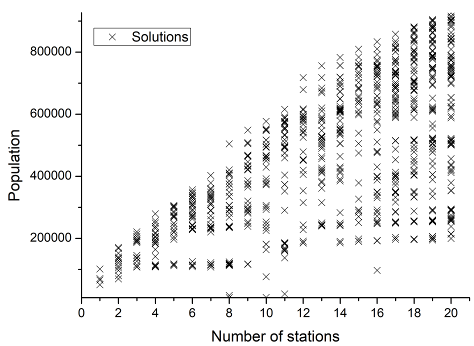
In the execution of this stage-based MOABC algorithm, a set of non-dominated solutions is generated. In the first stage, each solution corresponds to from 1 to 20 locations where the monitoring stations can be placed. Specifically, 880 different non-dominated solutions are obtained in the first stage for the median execution. Note that the four objectives are being optimized simultaneously, so that the graphical representation of the solutions involves four dimensions. For better understanding, the graphical representation will be decomposed into six 2D plots that include all the possible pairwise combinations of the four objectives. The objective values of these non-dominated solutions are shown in Figure 7.



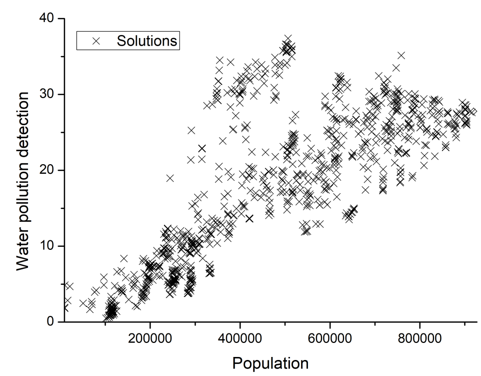
(a) WPD vs number of stations.



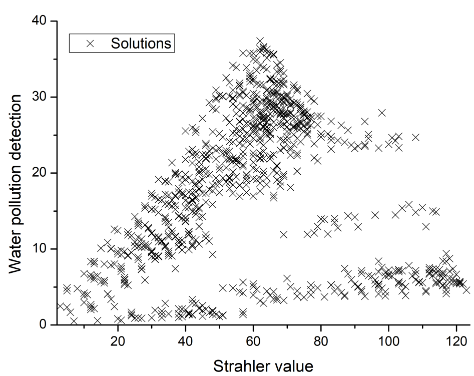
(b) Strahler number vs number of stations.



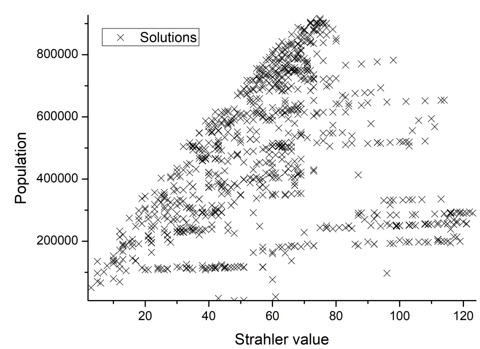
(c) Population vs number of stations.



(d) WPD vs population.



(e) WPD vs Strahler number.



(f) Population vs Strahler number.

Figure 7: Objective values of the non-dominated solutions obtained when the number of monitoring stations is limited to between 1 and 20.

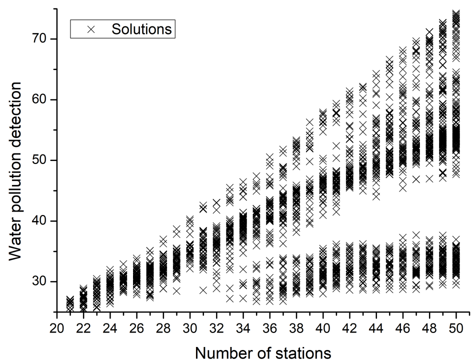
The following deductions can be made from the plots in Figure 7. Figure

7(a) shows that non-dominated solutions are obtained for all numbers of stations between 1 and 20, and that the dispersion increases with the number of stations. The more stations there are, the greater is WPD. This was to be expected since the greater the number of stations, the more threshold violations will be detectable. The relationship between the number of stations and the Strahler number is shown in Figure 7(b). Again as expected, when the number of stations increases, the Strahler number also increases since it is possible to cover more river stretches with high hydrological importance. The increase in the dispersion stands out more in Figure 7(c) with more population being covered as more stations are considered. This would result in more people being protected. Figure 7(d) shows that when the population increases, the WPD also increases. Areas with high population may produce more pollution in general, and hence greater water contamination. Figure 7(e) shows two types of behaviour for the relationship between the Strahler number and WPD. First, for most of the solutions, as the Strahler number increases, so does the WPD with a steep slope. This means that these solutions represent stations at which the pollution increases strongly with their growing hydrological importance. Second, there is another group of solutions for which this increase is weaker. Finally, Figure 7(f) shows that, as the Strahler number increases, so does the population. This means that people settle in areas close to important streams.

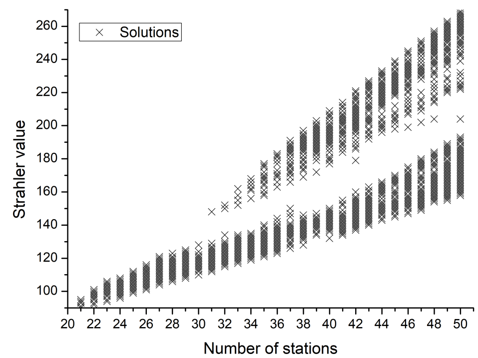
This procedure allowed the $2.56 \cdot 10^{51}$ possible solutions for from 1 to 20 monitoring stations to be reduced to the best 880 solutions (i.e., greater than 99.99% reduction). Now the best 20 locations based on these 880 non-dominated solutions have to be selected. To this end, a procedure based on the consensus solution is applied. The frequency of each cell (i.e., the number of times that the cell is present in these 880 non-dominated solutions) is calculated, and all the cells are ranked according to those frequencies. The 20 most repeated cells are then taken to be the locations of the monitoring stations given by this first stage.

4.3. Second stage: from 21 to 50 monitoring stations

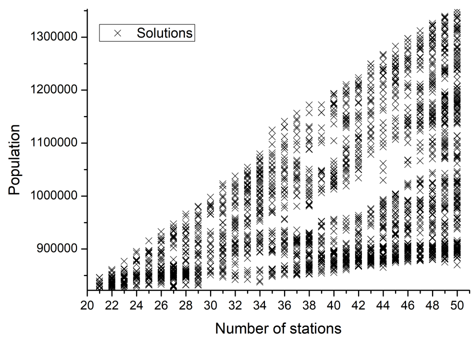
Now the focus turns to the next 30 monitoring stations, given that the first 20 have been settled. In this case, a total of 2352 different non-dominated solutions were found in the median execution. The objective values of these non-dominated solutions are shown in Figure 8.



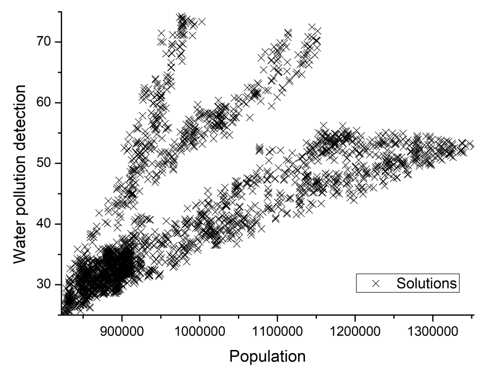
(a) WPD vs number of stations.



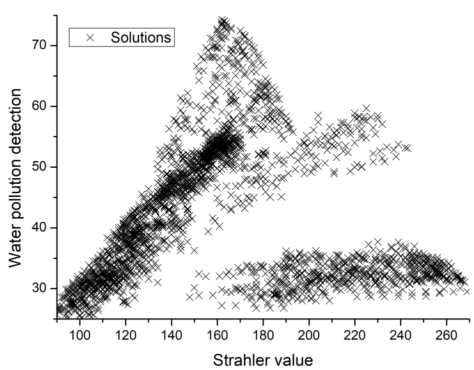
(b) Strahler number vs number of stations.



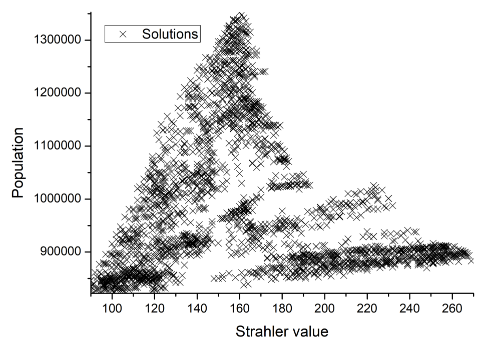
(c) Population vs number of stations.



(d) WPD vs population.



(e) WPD vs Strahler number.



(f) Population vs Strahler number.

Figure 8: All the non-dominated solutions obtained when the number of monitoring stations is limited to between 21 and 50.

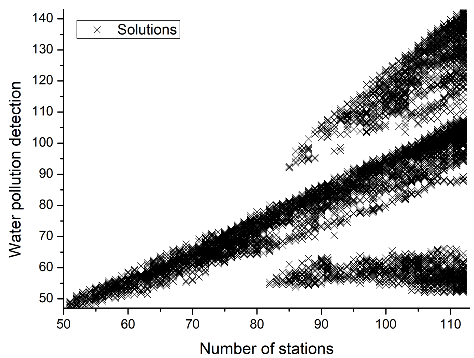
442 The comments made above for the first stage also apply to this stage,

but with some additional remarks. The increase in the number of solutions leads to greater dispersion than in the first stage. This can be seen in the relationship between the number of stations and each of the three objectives (Figures 8 (a,b,c)). While again WPD increases as the population grows, there now appear three positive slopes which correspond to three different areas related to the Strahler numbers (see Figures 8 (d, e)). In Figure 8 (f), one observes that larger Strahler numbers are related to lower populations.

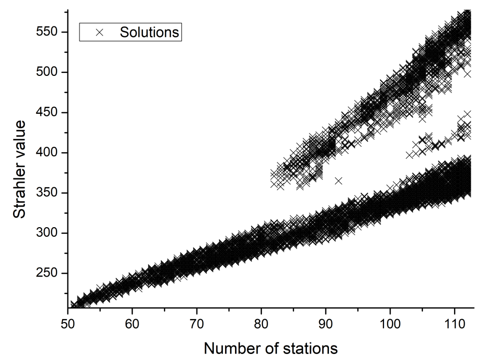
Since the process is progressive, in this stage there are 3077 (=3097-20) possible locations. The number of possible solutions with between 21 and 50 stations is then $5.71 \cdot 10^{109}$, which the algorithm reduces to 2352 solutions (again greater than 99.99% reduction). The consensus solution approach was applied to get the 30 best new locations for this second stage.

4.4. Third stage: from 51 to 112 monitoring stations

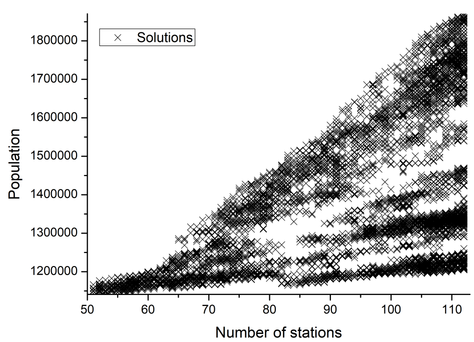
Now the focus is on allocating the last 62 monitoring stations given that the first 50 have been settled in the previous stages. A total of 6283 non-dominated solutions were found in the median execution. The objective values of these non-dominated solutions are shown in Figure 9.



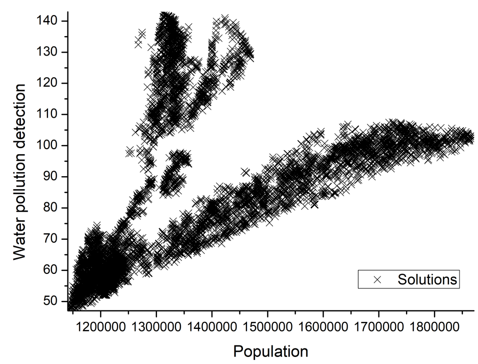
(a) WPD vs number of stations.



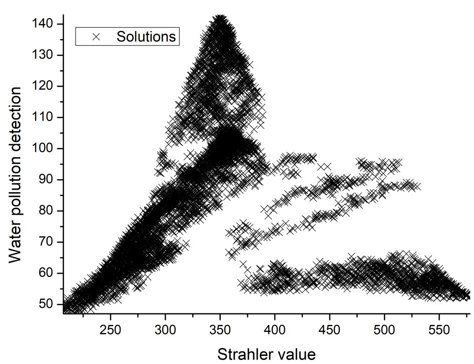
(b) Strahler number vs number of stations.



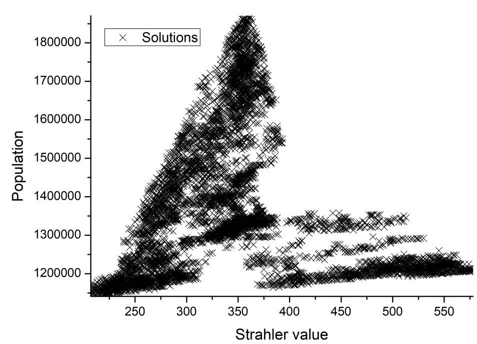
(c) Population vs number of stations.



(d) WPD vs population.



(e) WPD vs Strahler number.



(f) Population vs Strahler number.

Figure 9: All the non-dominated solutions obtained when the number of monitoring stations is limited to between 51 and 112.

The same trends for the objectives observed in the second stage are even clearer now due to the greater number of solutions found.

In this stage, there are 3047 (3097-50) possible locations for the stations 51 to 112. In this stage, there are $1.04 \cdot 10^{207}$ possible solutions, which the algorithm reduces to 6283, a computationally manageable number. Application of the consensus solution approach allowed the best 62 new locations to be determined for this stage. Combination with the solutions found in the previous stages yielded an ordered list of 112 locations for water quality monitoring stations.

4.5. Description of the monitoring station locations

Figure 10 shows the locations of the water quality monitoring stations by stage.

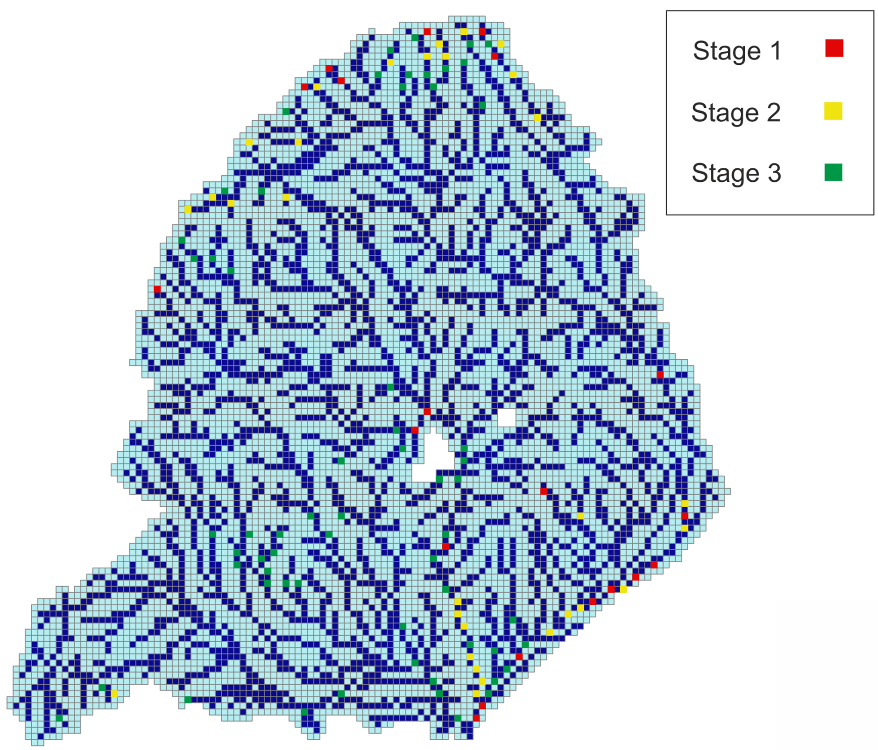


Figure 10: Progressive allocation of 112 monitoring stations in three stages.

The first stage assigned stations in different important areas. Specifically, stretches with low Strahler number and high pollution were assigned as stations in the north and northwest of the Jubba River basin (in Ethiopia). In the centre of the river basin (in Somalia), the stations assigned corresponded to sites with very high Strahler number and population but low pollution. Some stations were allocated in the east of the river basin, mainly corresponding to the Shebelle tributary. These stretches have very high Strahler numbers and high pollution. Finally, some stations were allocated close to the mouth, along the Jubba River itself. They combine high population with the maximum Strahler numbers, but not high pollution.

The second stage reinforces the coverage of some of the vast areas described for the previous stage, and adds stations in the southwest of the Jubba River basin (Kenya) with intermediate pollution, high population, and low Strahler numbers.

Finally, the third stage increases the coverage of all of the previously defined areas, and adds focuses on three new areas. One is in the west of the river basin (Ethiopia) with locations having low Strahler numbers and intermediate population and pollution. Another is around the Kutulo and Lak Bor tributaries (Kenya and Somalia) where the locations assigned have low population and Strahler number, but high pollution. The last area with assigned stations corresponds to the Lagh Dera tributary, especially close to where it flows into the Jubba River. These locations have low pollution and population, but very high Strahler number.

The construction of this network based on objective criteria has thus been able to allocate in a progressive way the water quality monitoring stations in the most significant parts of the river basin. The following subsection will compare the performance of the objective values in the three stages.

4.6. The performance of the stage-based approach

As was described above, the procedure was defined so that the number of stations increases up to 112 in three stages. This subsection will present a comparison of the evolution of the objective values for the locations selected in the three stages. Since the stations were ranked within each stage, it is possible to conform a water quality network with any number of stations from 1 to 112. Figure 11 presents the cumulative WPD for any number of stations together with its disaggregation by FC, BOD, and TDS.

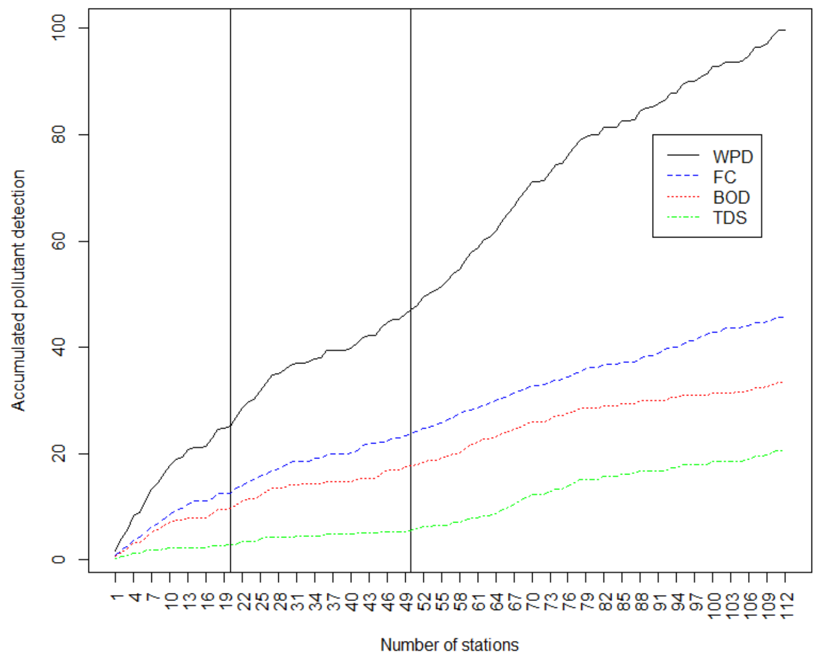


Figure 11: Cumulative amounts of WPD, FC, BOD, and TDS for from 1 to 112 stations.

Figure 11 shows a major increase of cumulative WPD during the first stage, with a more moderate rise during the other stages. Specifically, the slopes are 1.23, 0.70, and 0.85 for the first, second, and third stages, respectively. This behaviour shows that the first stations are very well allocated, covering the locations with the highest WPD. Then, as the number of monitoring stations increases, it becomes harder to proportionally detect this pollution because the best locations have already been assigned. The information in this figure would allow an exact number of stations for the network to be found for any threshold established for cumulative WPD. The disaggregated information shows that FC contributes the most, followed by BOD, and finally by TDS, and that this is the case for whatever number of stations. Again, a threshold of cumulative FC, BOD, or TDS could be established in order to find the exact number of stations required for the network.

Figure 12 is the analogous graph for the population. The slope of the cumulative population curve is steep to begin with, and then gradually starts to flatten out as the number of stations increases. In particular, the slope decreases from 49 267 in the first stage to 12 748 in the second and 5098 in the third. Again, this indicates that the first stations are very well allocated, and then the next stations become harder to match such a high benefit. This caveat notwithstanding, the graph allows a population threshold to be established so as to find the exact number of stations that need to be set up in the network.

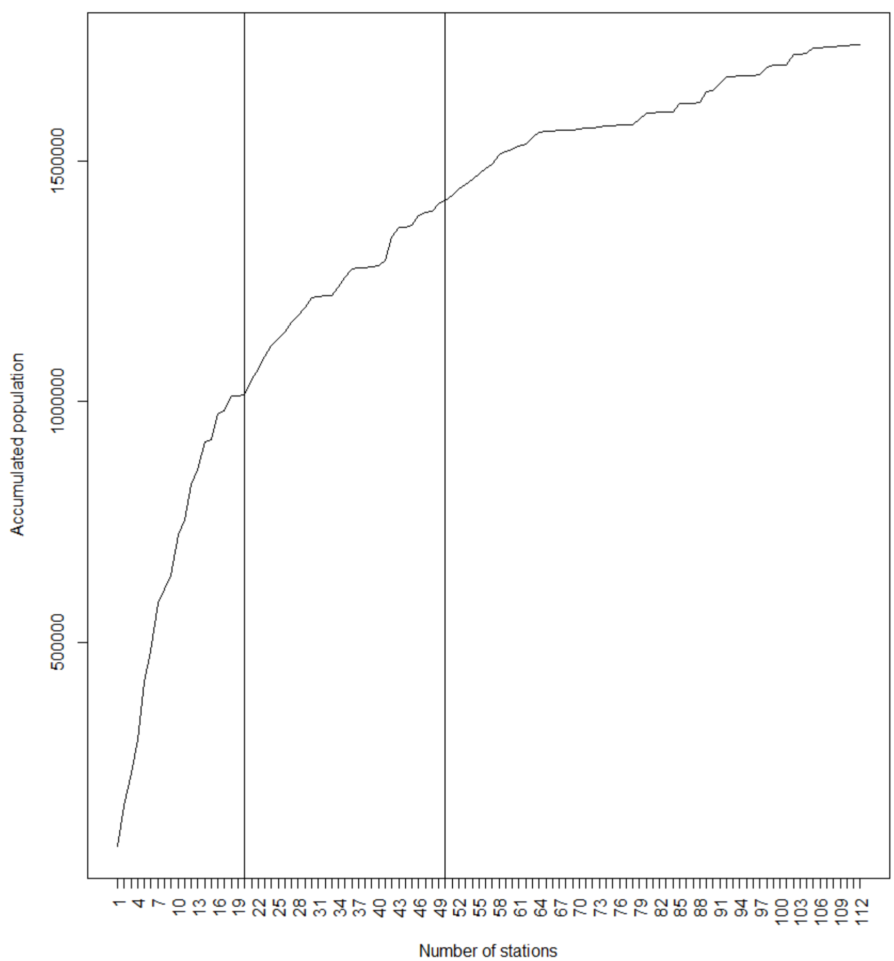


Figure 12: Cumulative population for from 1 to 112 stations.

Finally, the corresponding graph for the Strahler numbers is shown in Figure 13. In this case, the cumulative Strahler number curve has a fairly constant upward slope across all three stages, reflecting that a steadily growing number of streams are being monitored by the network. Again, the graph would allow a Strahler number threshold to be established to find the exact number of stations needed in the network.

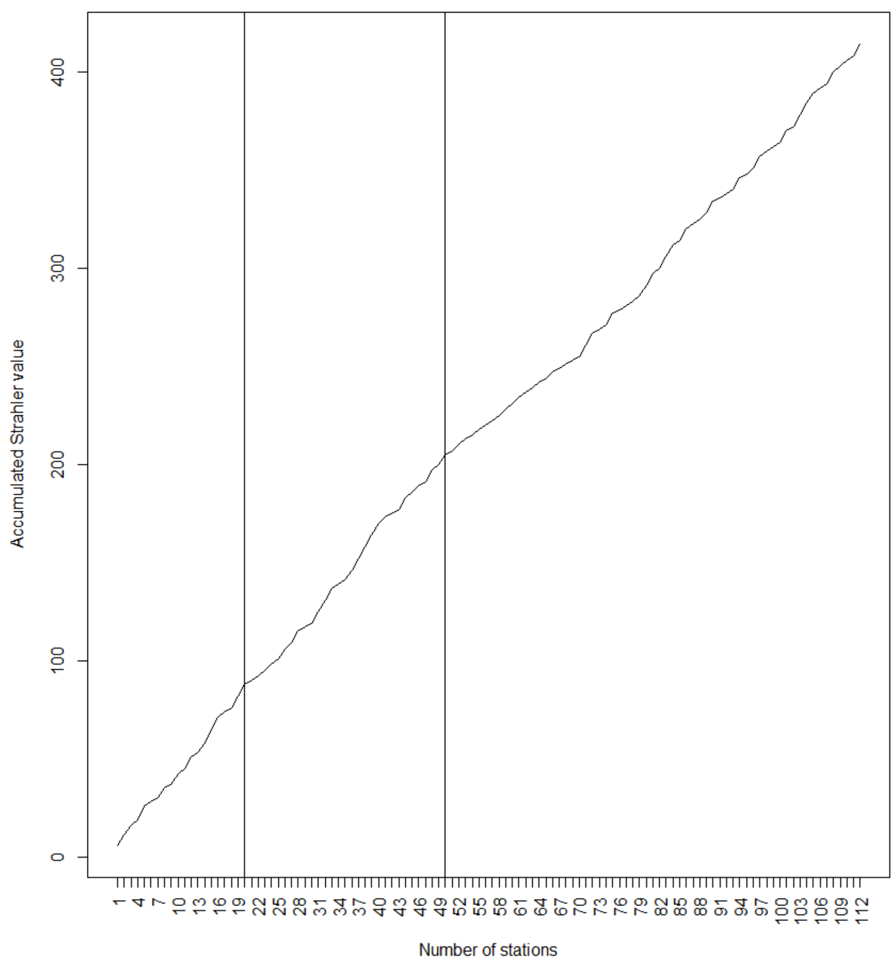


Figure 13: Cumulative Strahler numbers for from 1 to 112 stations.

By way of synthesis, Table 1 lists the objective values (WPD, population, and Strahler number) for each stage, as well as the percentage improvement of each stage over the preceding one.

Stage	WPD	Population	Strahler
1	25.08 -	1 013 221 -	88 -
2	47.10 87.80%	1 417 275 39.88%	205 132.95%
3	99.72 111.72%	1 740 415 22.80%	414 101.95%

Table 1: Objective values and percentage improvements for WPD, population, and Strahler number by stages.

For WPD, the percentage improvement from the first to the second stage is 87.80%, increasing to 111.72% from the second to the third stage. For the population objective, the respective improvements are 39.88% and 22.80%, and for the Strahler numbers they are both greater than 100%. Observe that percentage improvements higher than 100% are possible because the increment of monitoring stations from stage 1 to stage 2 is 150% (from 20 to 50 stations) and the augmentation of stations from stage 2 to stage 3 is 124% (from 50 to 112 stations). It can be concluded that the data from Table 1 lend support to the applicability of the proposed stage-based approach to substantially improving the objectives being considered with increases in the number of stations.

5. Conclusions

A stage-based optimization approach has been developed to construct river basin water quality monitoring networks. The approach allows a network to be created from scratch without the need for in situ pollution measurements. Instead, it uses, together with other social and hydrological criteria, pollutant estimates from the WorldQual model. At each stage, the proposed stage-based MOABC approach is able to efficiently reduce the huge number of possible solutions, and choose just one optimal solution. Moreover, the method followed provides a list of candidate stations ranked by importance, so that the network can be built up progressively. This approach is especially interesting when in situ measurements are absent since it allows a water quality network to be initialized and augmented. It can be applied to river basins of any size.

The results show that the approach finds the best places to allocate monitoring stations in the Jubba River basin by seeking a compromise between detection of pollutants, number of people affected, and the location's hydrological importance. At the same time, the number of stations required is reduced as much as possible – an aspect that is especially relevant for application by developing countries who may consider this approach as a way to obtain an overview of their river basins and then prioritize the initial distributions of the networks. The approach would help policy makers to take informed decisions based on environmental and sustainability assessments of their river basins.

As a future research line, this approach could be adapted to include more pollutants and water information, such as microplastics, nutrients, chemicals, pathogens, temperature, among others, with a previous correlation study that clarifies the possible relationships among all this information.

Acknowledgements

This research was supported by *Ministry of Science, Innovation and Universities* - Spain and *State Research Agency* - Spain (projects PID2019-107299GB-I00/AEI/10.13039/501100011033 and MTM2017-86875-C3-2-R), *Junta de Extremadura* - Spain (projects GR18090 and GR18108), *Centre for Environmental Systems Research, University of Kassel* - Germany (WaterGAP3 project), and *European Union* (European Regional Development Fund). There are no conflicts of interest.

References

Adu-Manu, K., Tapparello, C., Heinzelman, W., Katsriku, F., & Abdulai, J.-D. (2017). Water quality monitoring using wireless sensor networks: Current trends and future research directions. *ACM Transactions on Sensor Networks*, *13*, 1–41.

Al-Khafaji, M. S., & Abdulraheem, Z. A. (2017). A deterministic algorithm for determination of optimal water quality monitoring stations. *Water Resource Management*, *31*, 3575–3592.

Alcamo, J., Döll, P., Heinrichs, T., Kaspar, F., Lehner, B., Rösch, T., & Siebert, S. (2003). Development and testing of the WaterGAP 2 global model of water use and availability. *Hydrological Sciences Journal*, *48*, 317–337.

Alilou, H., Nia, A. M., Saravi, M. M., Salajegheh, A., Han, D., & Enayat, B. B. (2019). A novel approach for selecting sampling points locations to river water quality monitoring in data-scarce regions. *Journal of Hydrology*, *573*, 109–122.

Balint, Z., Njeru, L., Pappas, C., & Pron, P. (2010). *Atlas of the Juba and Shabelle Rivers in Somalia*. FAO-SWLIM project (1st ed.).

Bastidas, J. C., Vélez, J. J., Zambrano, J., & Londoño, A. (2017). Design of water quality monitoring networks with two information scenarios in tropical Andean basins. *Environmental Science and Pollution Research*, *24*, 20134–20148.

Birattari, M., & Dorigo, M. (2007). How to assess and report the performance of a stochastic algorithm on a benchmark problem: mean or best result on a number of runs? *Optimization Letters*, *1*, 309–311.

Borden, C., & Roy, D. (2015). *Water Quality Monitoring System Design*. Technical Report - International Institute for Sustainable Development.

Chang, N. B., & Makkeasorn, A. (2010). Optimal site selection of watershed hydrological monitoring stations using remote sensing and grey integer programming. *Environmental Modeling and Assessment*, *15*, 469–486.

European Parliament and Council (2000). Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy. *Official Journal L of the European Parliament and the Council*, *327*, 1–73.

Higgins, J. V., Bryer, M. T., Khoury, M. L., & Fitzhugh, T. W. (2005). A freshwater classification approach for biodiversity conservation planning. *Conservation Biology*, *19*, 432–445.

Horton, R. E. (1945). Erosional development of streams and their drainage basins: hydro-physical approach to quantitative morphology. *Bulletin of the Geological Society of America*, *56*, 275–370.

Houghton-Carr, H. A., Print, C. R., Fry, M. J., Gadain, H., & Muchiri, P. (2011). An assessment of the surface water resources of the Juba-Shabelle basin in southern somalia. *Hydrological Sciences Journal*, *56*, 759–774.

Huo, J., & Liu, L. (2018). An optimization framework of multiobjective artificial bee colony algorithm based on the MOEA framework. *Computational Intelligence and Neuroscience*, *2018*, 5865168.

Karaboga, D., & Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *Journal of Global Optimization*, *39*, 459–471.

Karaboga, D., Gorkemli, B., Ozturk, C., & Karaboga, N. (2014). A comprehensive survey: Artificial Bee Colony (ABC) algorithm and applications. *Artificial Intelligence Review*, *42*, 21–57.

Lehner, B., Verdin, K., & Jarvis, A. (2008). New global hydrography derived from spaceborne elevation data. *Eos Transactions*, *89*, 93–94.

Liyanage, C. P., Marasinghe, A., & Yamada, K. (2016). Comparison of optimized selection methods of sampling sites network for water quality monitoring in a river. *International Journal of Affective Engineering*, *15*, 195–204.

McCaffrey, S. (2012). *Water Quality Parameters & Indicators*. Report - Waterwatch Coordinator, Namoi Catchment Management Authority. New South Wales, Australia.

Nguyen, T. H., Helm, B., Hettiarachchi, H., Caucci, S., & Krebs, P. (2019). The selection of design methods for river water quality monitoring networks: A review. *Environmental Earth Sciences*, *78*, 1–17.

Noble, R., & Cowx, I. (2002). *Development of a River-Type Classification System (D1). Compilation and Harmonisation of Fish Species Classification (D2)*. Technical Report - University of Hull, UK.

Park, S. Y., Choi, J. H., Wang, S., & Park, S. S. (2006). Design of a water quality monitoring network in a large river system using the genetic algorithm. *Ecological modelling*, *199*, 289–297.

Pérez, C. J., Vega-Rodríguez, M. A., Reder, K., & Flörke, M. (2017). A multi-objective artificial bee colony-based optimization approach to design water quality monitoring networks in river basins. *Journal of Cleaner Production*, *166*, 579–589.

Puri, D., Borel, K., Vance, C., & Karthikeyan, R. (2017). Optimization of a water quality monitoring network using a spatially referenced water quality model and a genetic algorithm. *Water*, *9*, 704(1–11).

Reder, K., Alcamo, J., & Flörke, M. (2017). A sensitivity and uncertainty analysis of a continental-scale water quality model of pathogen pollution in African rivers. *Ecological Modelling*, *351*, 129–139.

Reder, K., Flörke, M., & Alcamo, J. (2015). Modeling historical fecal coliform loadings to large European rivers and resulting in-stream concentrations. *Environmental Modelling and Software*, *63*, 251–263.

Reder, K., Malsy, M., Bärlund, I., & Flörke, M. (2014). Modelling water pollution posing a risk to human and ecosystem health in Africa using fc and bod as indicator substances. In C. A. Brebbia (Ed.), *Water Pollution XII* (pp. 297–308). WIT Press volume 182.

Sabzipour, B., Asghari, O., & Sarang, A. (2019). Evaluation and optimal redesigning of river water-quality monitoring networks (RWQMN) using geostatistics approach (case study: Karun, Iran). *Sustainable Water Resources Management*, *5*, 439–455.

Strahler, A. N. (1957). Quantitative analysis of watershed geomorphology. *Transactions of the American Geophysical Union*, *38*, 913–920.

Strokal, M., Spanier, J. E., Kroeze, C., Koelmans, A. A., Flörke, M., Franssen, W., Hofstra, N., Langan, S., Tang, T., van Vliet, M. T., Wada, Y., Wang, M., van Wijnen, J., & Williams, R. (2019). Global multi-pollutant modelling of water quality: scientific challenges and future directions. *Current Opinion in Environmental Sustainability*, *36*, 116–125.

Telci, I. T., Nam, K., Guan, J., & Aral, M. M. (2009). Optimal water quality monitoring network design for river systems. *Journal of Environmental Management*, *90*, 2987–2998.

UNEP (2016). *A Snapshot of the World's Water Quality: Towards a global assessment*. Technical Report - United Nations Environment Programme. Nairobi, Kenya.

Vermeulen, L. C., van Hengel, M., Kroeze, C., Medema, G., Spanier, J. E., van Vliet, M. T., & Hofstra, N. (2019). Cryptosporidium concentrations in rivers worldwide. *Water Research*, *149*, 202–214.

Verzano, K. (2009). *Climate change impacts on flood related hydrological processes: Further development and application of a global scale hydrological model*. Ph.D. thesis - University of Kassel, Germany.

Villas-Boas, M. D., Olivera, F., & de Azevedo, J. P. S. (2017). Assessment of the water quality monitoring network of the Piabanha river experimental watersheds in Rio de Janeiro, Brazil, using autoassociative neural networks. *Environmental Monitoring and Assessment*, *189*, 1–15.

Voß, A., Alcamo, J., Bärlund, I., Voß, F., Kynast, E., Williams, R., & Malve, O. (2012). Continental scale modeling of in-stream river water quality: A report on methodology, test runs, and scenario application. *Hydrological Processes*, *26*, 2370–2384.

WMO (1994). *Guide to Hydrological Practices: Hydrology - From Measurement to Hydrological Information*. Technical Report - World Meteorological Organization.

Zhu, X., Yue, Y., Wong, P., Zhang, Y., & Ding, H. (2019). Designing an optimized water quality monitoring network with reserved monitoring locations. *Water*, *11*, 713 (1–25).