AN ONLINE PLATFORM FOR ENABLING OPEN DATA COLLECTION, STORAGE, ANALYSIS AND CLASSIFICATION OF WATER MANAGEMENT DATA FOR THE ATTICA REGION IN GREECE

DOI: 10.26341/issn.2241-4010-2025-13a-10-K02045

Antoniadis Ioannis
Student, Harokopio University of Athens, School of Digital Technology
it22008@hua.gr

Drakou G. Evangelia
Assistant Professor, Harokopio University of Athens, School of Environment Geography and Applied Economics, El. Venizelou 70 Ave, 176 76, Kallithea – Attiki, Greece
e.drakou@hua.gr

Tsadimas Anargyros
Teaching Associate, Harokopio University of Athens, School of Digital Technology, Omirou 9, Tavros 17778, Athens, Greece
tsadimas@hua.gr

Kousiouris George
Associate Professor, Harokopio University of Athens, School of Digital Technology, Omirou 9, Tavros 17778, Athens, Greece
gkousiou@hua.gr

Abstract

Water resource management is considered critical in an era where the specific resource is becoming increasingly scarce due to population growth, over-usage, changing weather patterns and climate conditions. The existence of detailed and concentrated data regarding its usage is predominantly needed in order to perform research, guide relevant policies and necessary management and planning.

The present paper aims at designing and implementing a backbone platform that may be used to gather, store, analyze, categorize and display relevant data to be used in understanding water consumption and availability patterns. The data refer to water availability in reservoirs, water production statistics, temperature and precipitation as well as supply network failures in the Attica region of Greece. To this end, it makes use of open data sources (eydap.gr, the main authority responsible for water supply in the Greek capital and surrounding areas, open-meteo.com for weather and precipitation conditions, as well as 3 services for geocoding and mapping- namely geodata.gov.gr, openstreetmap and geoapify).

The developed platform is able to retrieve historical data, as well as implement complex automated workflows for the daily update of these data from their sources. It is also able to parameterize the views and display of data based on user-needed configurations, in order to support a more fine grained and tailored analysis. Clustering is performed in order to create relevant categories in data ranges as well as classify observed values. Data acquisition is performed both in a typical web-based user friendly interface as well as through offering the same functionalities through a backend REST API that may be used to programmatically interact with it and submit filtering queries. Therefore it can support researchers in finding concentrated data for their studies through a single access point, as well as export that data for further analysis.

Key words: Water management, digital platform, open data, policy making, REST API, data analysis, clustering

1. Introduction

Water resources are becoming scarce around the world, either due to overutilization or change in climate patterns and conditions. The drive to invest further in relevant infrastructures and/or suitable water management schemes needs to be complemented by the appropriate baseline data that can inform decision-makers to prioritize actions and policies. One of the benefits of the modern era is the abundance of sensing devices, monitoring systems as well as data sources that can provide such a baseline. However in most cases these systems are fragmented and diverse, thus making it difficult to aggregate, analyze and distribute them to all relevant stakeholders.

Different data models, different interfaces, as well as different levels of information, create challenges to the data collection process, or the mapping of similar information based on different granularities and features of the data. For example, geographic location may be given in one case as GPS coordinates, in another as street names and in another at municipality level. Furthermore, filtering and exposure of that data from remote locations, based on researcher or policy maker needs is required in order to exploit their full potential. Thus exposing them through a set of RESTful services i.e., a web-based way to filter and retrieve data subsets, is necessary.

The aim of this research is to present the development of such a platform that collects, analyzes and stores all relevant data for the city of Athens. To this end it collects open data from the main water authority of Athens (EYDAP), regarding consumption of water and status of reservoirs, weather data from an extended area around these locations, as well as information on water distribution disruptions due to technical reasons (e.g. broken water pipes). The latter is needed in an effort to understand and prioritize areas with frequent such disruptions, in order to save water waste. Further analysis is conducted at reservoir levels, applying a classification scheme based on the available historical data, in order to understand the comparative status of them. The implementation of the platform is an open source project available through its github repository[1]. It has also been deployed and is accessible for end users[2].

The paper proceeds as follows. In Chapter 2, related work is presented, while in Chapter 3 the methodology of data collection and transformation is given. Chapter 4 describes the inner system architecture as well as the data serving mechanisms, while Chapter 5 presents the front-end (data presentation) as well as the analysis mechanisms for the classification of water levels and municipality disruptions. Finally Chapter 6 gives the final conclusions.

2. Related Work

Existing research highlights the critical role of digital platforms in advancing water resource management through open data collection, storage, and analysis (Horsburgh et al 2025). Cosgrove and Loucks (2015) discuss the pressing challenges in water management, particularly issues related to data accessibility, integration, and the need for robust decision-support systems. They emphasize the necessity of technological advancements to enhance data-sharing mechanisms and support evidence-based decision-making. Complementing this perspective, Delipetrev et al. (2014) present the development of a web-based application utilizing open-source software for water resource management, demonstrating the feasibility of real-time data access, modeling, and decision support in hydrological systems. Drost et al.

[1] https://github.com/giannisclipper/water-reserves

[2] https://water.ditapps.hua.gr/

(2022) present an automated event driven monitoring system exploiting satellite data, based on open source software. Only recently, Horsburgh et al (2025) recognizing the significance of such an infrastructure for decision-makers, made an attempt to develop an application for data acquisition and curation in a US application, highlighting the potential of such applications to minimize technological challenges decision-makers might face, given that several of the data repositories require data management expertise.

Several web-based platforms highlight the role of digital technologies in water management. WaterApp[3] enables real-time monitoring through IoT, while Utilios[4] provides insights into top water management software. WebbyLab[5] explores IoT applications for smart water monitoring, and FREEWAT[6], based on AkvaGIS from Criollo et al. (2019), offers an open-source GIS-based tool for groundwater management and classification in Belgium, while HydroShare (Horsburgh et al 2025) develops a similar application for the US. These platforms demonstrate the importance of digital innovation and data-driven decision-making, aligning with the need for an advanced online system for water data collection, storage, and analysis.

3. Methodology

The designed platform collects, stores and analyzes data concerning the water reserves and the water supply of Athens. As Figure 3.1 shows, the data of interest are about the remaining quantities in the water reservoirs, the quantities of the drinking water production, precipitation and temperature measurements, the water supply interruptions as well as their distribution over the municipalities. Moreover, the platform provides the collected data to other systems through a REST API or to end-users through a graphical UI.

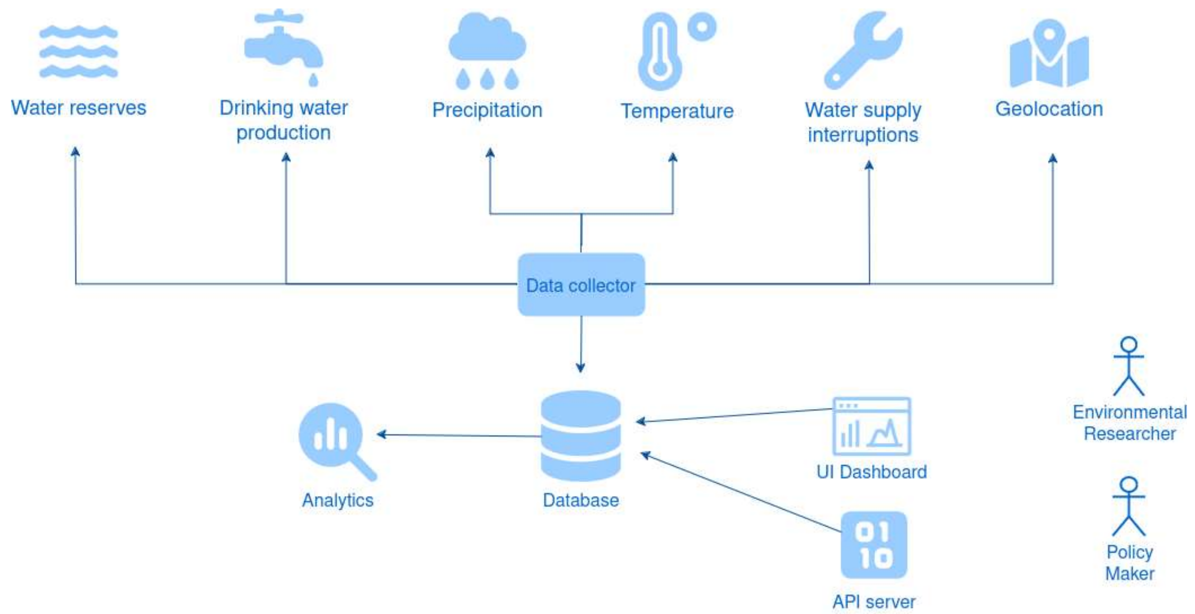

Figure 3.1. An overview of the platform functionality

[3] https://waterapp.in/

[4] https://utilios.io/best-water-management-software/

[5] https://webbylab.com/blog/iot-in-water-management/

[6] http://www.freewat.eu/

Data sources

Different datasets are created and maintained by the platform facilitating various data views and a better insight of the water reserves and the water supply of Athens. While the data is acquired from external data sources, transformation processes are required to make them coherent and usable. A DB schema has been designed ensuring data integrity and processing efficiency. The collected data is checked for missing, inconsistent or inaccurate values, and they are cleaned and transformed to follow the DB schema, the table relations and the data types. The transformed data is finally stored in the DB and becomes available for analysis, filtering and illustrations.

Water data acquisition

EYDAP[7] is the main authority responsible for water supply in the Greek capital and surrounding areas. The company uses raw water mainly from four surface water resources, the lakes of Marathonas, Yliki, Mornos and Evinos, while some more underground water resources are used for backup purposes (EYDAP, 2024a). Furthermore, the company operates four water treatment plants located in Galatsi, Acharnes, Polydendri and Aspropyrgos (EYDAP, 2024b) in order to make the raw water drinkable.

On an almost daily basis, the water quantities remaining in the four main reservoirs as well as the drinking water quantities produced from the four water treatment plants, are reported through the official website of EYDAP. All the daily values for the water reserves since 1985 and for the drinking water production since 1996, are available on this website. The users can reach them as HTML outputs by specifying a time range of interest (a single day, a week, a month or a year). The outputs are shaped as tables of values, containing the date, the apart quantities (per reservoir or per treatment plant accordingly) and the sum of the quantities.

After a calendar month is completed, EYDAP makes the data available through its open data hub[8] as well. The users can specify a time range of interest (a sequence of months). The result is an HTML output shaped as a list of values but can also be downloaded as a CSV file. However there is no offering of a programmatic way (e.g. REST API) to download these files on demand based on an easy programmable interface.

Additionally the company reports the water supply interruption events through this open data hub, within a few months delay since they were encountered. There are available interruption events from 2021. In a way similar to the water reserves or drinking water production, the users can specify a time range of interest (a sequence of months) and get the result as an HTML output shaped as a list of interruption events. An interruption event is described by the date when it was encountered and the address (a crossroad and a neighborhood) where the water supply network was disrupted and repaired. The output can also be downloaded as a CSV file.

The website and the open data hub of EYDAP, constitute the most essential data source for the platform, which utilizes all the historical data and automatically incorporates the new updates concerning the water reserves, the drinking water production and the water supply interruptions.

Weather data acquisition

A second important data source is the open-meteo.com, an open-source weather API offering free access for non-commercial use. It provides weather measurements for almost any place

[7] https://www.eydap.gr

[8] https://opendata.eydap.gr/

around the world and for a long time in the past. It follows the style of a typical REST API, it receives requests in the form of URLs with differentiated endpoints and query parameters, and it responds in JSON format.

The created platform utilizes open-meteo to collect daily precipitation and temperature measurements for eight cities located in Central Greece (Athens, Amfissa, Chalkida, Karpenisi, Lamia, Livadeia, Thiva, Agrinio), the wider area where the water reservoirs of EYDAP are located too. The platform has collected all the daily precipitation and temperature measurements for these cities since 1985 and goes on checking and getting new updates daily.

Data Extract-Transform-Load Approach

One of the most critical processes in the envisioned system is the Extract-Transform-Load (ETL) process responsible for acquiring and ingesting the diverse data in the unified data model of our system. The variables already mentioned (water reserves, drinking water production, precipitation and temperature measurements, water supply interruptions) are tracked on a daily basis. The datasets are updated in time order without duplicated or missing dates. The values are incorporated respecting the DB schema, the table relations and the data types. Although it is very infrequent, some values may be missing or invalid regarding the water reserves and the drinking water production. In such cases, the values are considered the same as the ones of the previous date, due to the slight daily change of these variables. As an example, Figure 3.2 illustrates the transformation of the water reserves data, originating from HTML documents, stored in SQL DB tables and provided in JSON format.

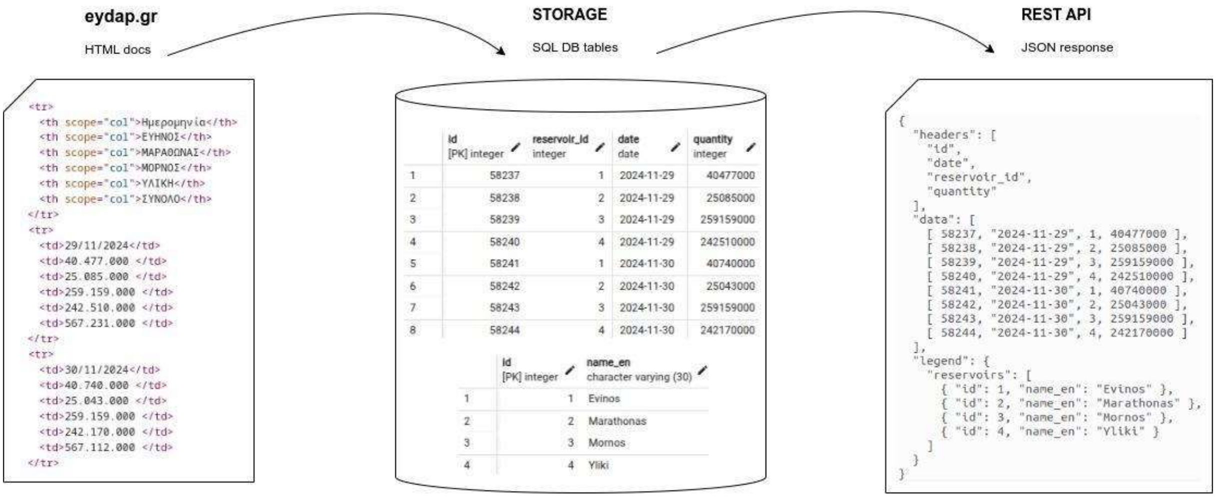

Figure 3.2. The transformation of the water reserves data, originating from HTML documents, stored in SQL DB tables and provided in JSON format.

Additional services and resources are used for the geospatial analysis of the water supply interruptions, which are very common in the region of Attica and for which any kind of organized information service would significantly improve local and regional management. At first, the platform utilizes openstreetmap.com, the well known online map, which is free to use under an open license. It is used for map representations of the municipalities of greater Athens. Then, geoadata.gov.gr which hosts open geospatial datasets from public sector organizations. It has provided the boundaries of the municipalities as bounding boxes of geographical coordinates.

Moreover, two geolocation services, nominatim.org and geoapify.com, are used for the classification of raw text addresses describing crossroads and neighborhoods into different municipalities. Nominatim is an open source project using OpenStreetMap data to search and find locations on earth and for the same purpose Geoapify is a commercial API providing a free tier. In most cases the use of Nominatim is enough to identify the municipality from a given address, but due to no strict format or notation of the addresses, the use of the second geolocation service was adopted to maximize the accuracy of matching the addresses to the municipalities. This process is needed since we considered that an analysis at the level of municipalities would be a desired level for policy makers to understand the status of the supply network and target investments. However, other means of grouping are not excluded by the organization of the data.

Figure 3.3. The steps of converting between crossroads and municipalities.

Figure 3.3 illustrates the steps of converting between crossroads and municipalities. It follows a detailed description of this process:

- Each one of the water supply interruptions given by EYDAP, is described including the address that it took place. e.g. "Bulgari and Tsamadou" KASTELLA

- The program needs to find a pair of geographical coordinates (latitude, longitude) representing the address of the water supply interruption. The two geolocation APIs (Nominatim, Geoapify) that the platform is powered by, do not respond well to queries describing crossroads. So the crossroads have to be split.

- A typical form ("road A & road B" AREA) is followed in the very most of the cases, but it is not a strict pattern. Sometimes the program is obliged to deal with differences in syntax or the presence of explanatory remarks. Nevertheless, the program tries to recognize and distinguish two roads and an area. e.g. road A: Bulgari, road B: Tsamadou, area: KASTELLA

- The program composes up to 3 alternative queries trying to identify the corresponding municipality:
  a) based on road A and area, e.g. Bulgari, KASTELLA, ATTICA, GREECE
  b) based on road B and area, e.g. Tsamadou, KASTELLA, ATTICA, GREECE
  c) based on area only, e.g. KASTELLA, ATTICA, GREECE

- Up to 6 requests (2 APIs x 3 queries) can be performed until some response fulfills all the query address terms. The response also includes the geographical coordinates of interest, e.g.: lat:37.9406, lon:23.6558

- That pair of the geographical coordinates is matched with the geographical polygon that encloses it, identifying consequently the corresponding municipality. For this purpose, a number of geographical polygons defining the boundaries of the municipalities are used. These polygons originate from geodata.gov.gr and they are stored locally in a geojson file for the needs of the platform. For example, the previous pair of coordinates is enclosed in the geographical polygon of Piraeus municipality identified by id: 9205

Data Analysis

A data analysis is performed to evaluate the new measurements of the water reserves, the drinking water production and the weather variables (precipitation, temperatures). It is a temporal analysis comparing the recent measurements with the historical ones. All the measurements over time are grouped into different clusters through a K-means clustering (Likas et al., 2003). K-means is a widely used clustering method, appropriate to distinguish a number of instances into different sets. This approach was chosen in order to summarize the data in meaningful and easy to interpret information. The clusters are ordered indicating lower to higher measurements. Thus the current value of each metric is classified in one of the available levels (e.g. low, medium, high etc) that are discovered from the historical data in a non-supervised manner.

A geospatial analysis is performed to classify the municipalities regarding how much they are disrupted from interruption events in a similar fashion. The events encountered within municipality areas are taken into account either over sq. kilometers or over thousands of inhabitants. Based on one of these two alternative indexes, the municipalities are grouped into clusters through a K-means clustering. The clusters are ordered indicating lower to higher levels of disruption.

Moreover, a correlation hypothesis was investigated between the precipitation measurements and the water reserves. For this purpose the Pearson and the Spearman correlation coefficients were used to check that hypothesis. The results of all the above appear in Chapter 5.

4. System architecture

Overview

The platform is a distributed system, based on a Three-Tier Client-Server Architecture. As Figure 4.1 represents, the platform consists of three distinct subsystems: a database, a backend application and a frontend application. A multi-tiered architecture encourages the separation of concerns and the modularity of the systems, and offers advantages regarding the maintainability of the implementation, the security of the database, the scalability of the deployment, as well as the performance of the applications (Nyabuto et al., 2024).

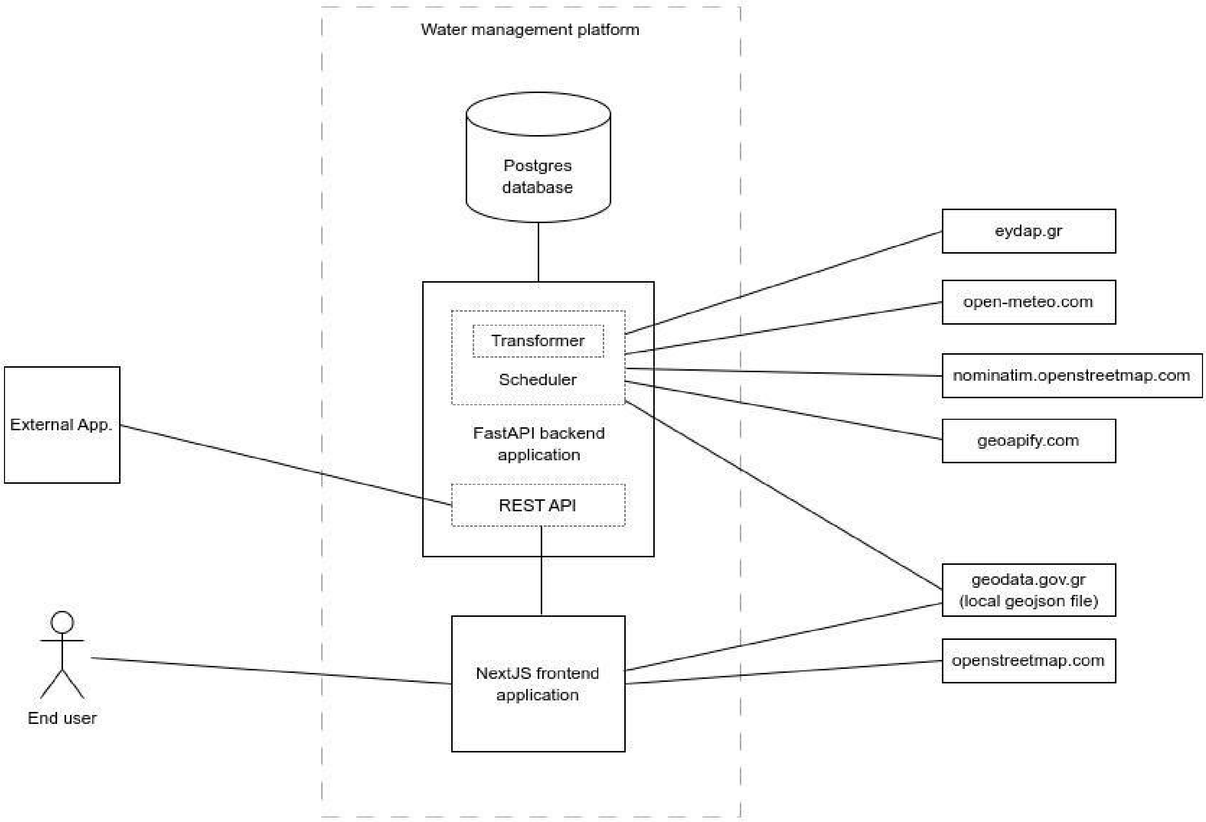

Figure 4.1. The platform as a Three-Tier Client-Server system

The first tier is an SQL database, based on the open source Postgres system. The DB schema is designed to incorporate and organize the data originated from the external sources, in such a way that ensures the integrity as well as the effective and efficient processing of them. A number of main tables are used to store the instances of the different datasets (water reserves, drinking water production, weather measurements, interruption events), while some auxiliary tables are used to codify different entities (water reservoirs, water treatment plants, locations of weather measurements, municipalities) which are included in the data and are useful for filtering and classification purposes.

The second tier of the system is a backend application written in FastAPI, a Python framework. It is designed to collect data from different data sources, as well as to store and analyze them. Additionally to provide them to other applications through a REST API service. The data is collected in various formats (HTML, JSON or CSV) and is transformed while it is incorporated in the platform. This tier implements the various transformation mechanisms that were described in the Methodology section and represented in Figures 3.2 and 3.3.

The third tier of the system, is a frontend application that carries out the interaction with the end users through a graphical interface. It is written in NextJS (a ReactJS framework). Typescript was also adopted in application development, to ensure the cohesion and the integrity of the code while it was being extended. Tailwind CSS (a CSS framework) was used for styling purposes, including responsive design to adjust layouts into different screen sizes. The Recharts library was used to create graphical charts and the Leaflet library to create map representations over OpenStreetMap.

Acquisition, Transformation and Data Serving Processes

At the very first time of its execution, the backend application initializes the database with the main body of the historical data (water reserves since 1985, drinking water production since 1996, water supply interruptions since 1985, precipitation and temperatures since 1985, water supply interruptions since 2021). This content has already been collected from the external data sources and stored in local HTML, JSON or CSV files. By retrieving these bulk updates from local files, the backend avoids overloading the data sources with massive requests. Additionally, the installation process is performed much faster.

Since the historical data has been loaded, the backend undertakes to communicate with the data sources on a daily basis, to check for possible new updates and to retrieve the new data whenever it exists. This operation is performed by five different schedulers running periodically. The four of them are used to check and retrieve updates, each one concerning a different dataset (water reserves, drinking water production, weather measurements, water supply interruptions). Figure 4.2 represents the typical flowchart of these procedures. The fifth one is used for geolocation purposes, to match the new water supply interruption events with the municipalities. Figure 4.3 represents the flowchart of that procedure. From a technical aspect, the schedulers have been developed to run in different threads, for not disturbing other backend operations with excessive delays during the updating processes.

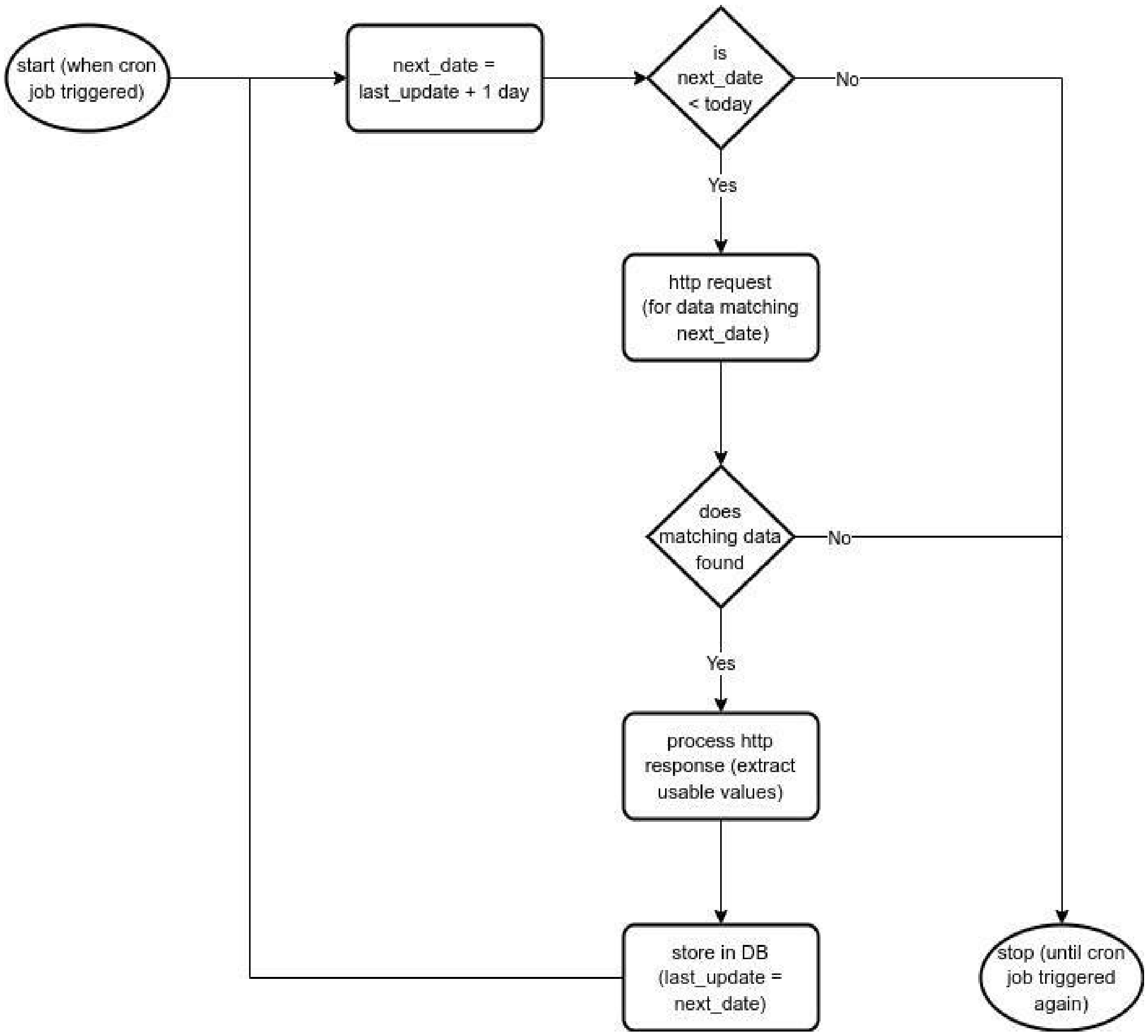

Figure 4.2. Retrieving new updates from the data sources

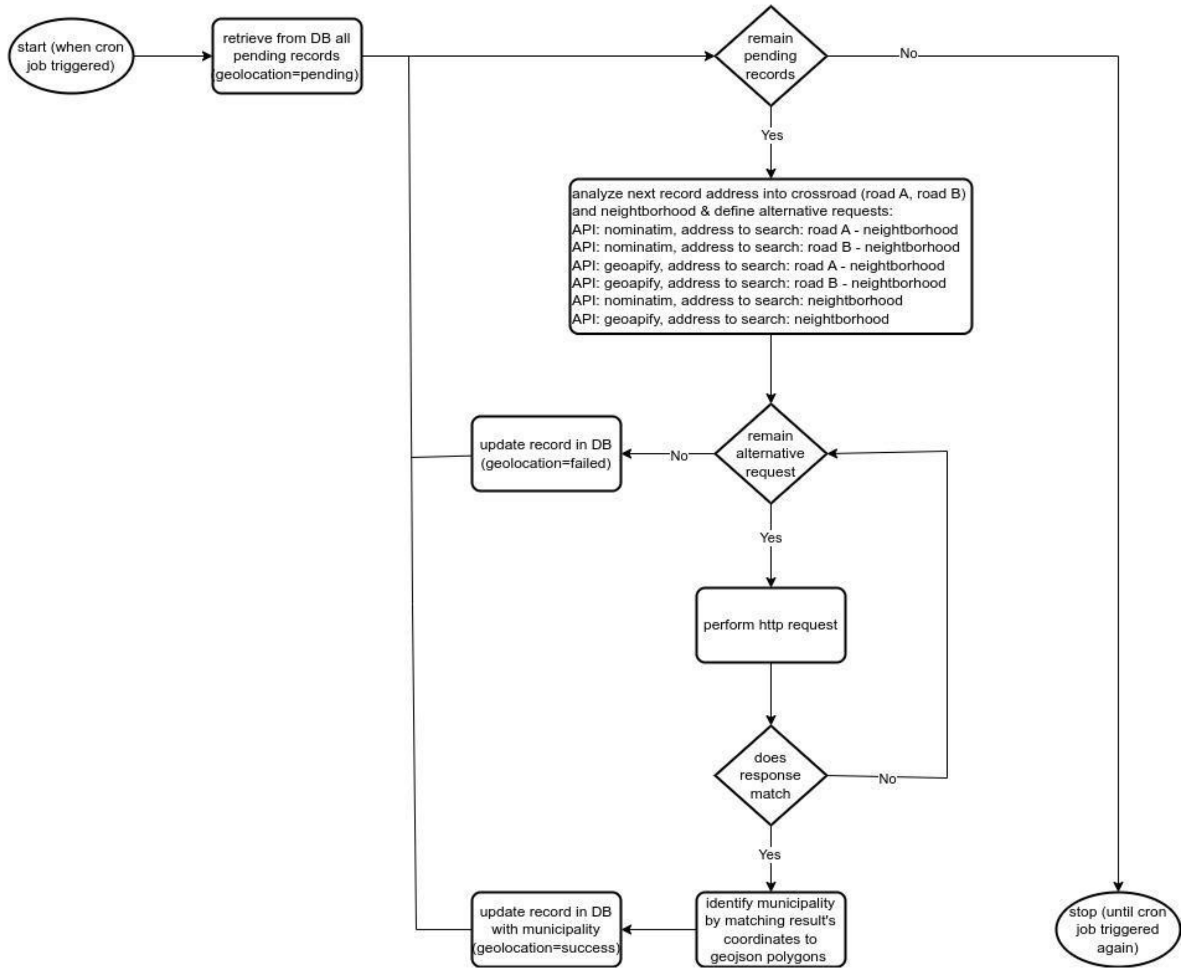

Figure 4.3. Matching water supply interruptions to municipalities

Furthermore, the backend operates as an HTTP server responding to client requests and providing them with data and analysis results. The communication with the clients is based on the principles and the architectural style of the REST API, which is qualified for network-based systems emphasizing in generality of interfaces, in independent deployment of components, in scalability, caching and reduced interaction latency (Fielding, 2000).

Different URL endpoints are used to serve different data categories and a number of query parameters are used to filter the data and specify the result. Validation checks are also used to ensure the integrity of the requests and the parameter values. The processing of the requests is carried out in asynchronous mode due not to disturb simultaneous requests or other backend operations with excessive delays. The responses are provided in JSON format. Table 4.1 illustrates all the available API endpoints.

5. Results

Dashboard presentation

The frontend application that carries out the interaction with the end-users, makes use of the data and the analysis that are available through the backend's REST API. It is also equipped with more features to address the needs of the end users, such as the graphical interface making the interaction between the users and the system easier, the charts illustrations and the map representations making the outputs more comprehensible, the options of downloading the outputs as PNG or CSV files for further processing, the options of getting unique URLs to reproduce or share specific data views.

Table 4.1. The available API endpoints

Method	Endpoint	Description	Response
GET	/api/v1/docs	API documentation (swagger-ui)	HTML
GET	/api/v1/status	Gets the current values of the values	JSON
GET	/api/v1/reservoirs?id=...	Gets the water reservoirs	JSON
GET	/api/v1/savings?time_range=2000,2023&time_aggregation=...	Selects water reserves data	JSON
GET	/api/v1/factories?id=...	Gets the water treatment plants	JSON
GET	/api/v1/production?time_range=2000,2023&time_aggregation=...	Selects drinking water production data	JSON
GET	/api/v1/locations?id=...	Gets the weather measurement locations	JSON
GET	/api/v1/weather?time_range=2000,2023&time_aggregation=...	Selects precipitation & temperature data	JSON
GET	/api/v1/municipalities?id=...	Gets the municipalities of greater Athens	JSON
GET	/api/v1/interruptions?time_range=2000,2023&time_aggregation=...	Selects water supply interruptions data	JSON

For this reason of reproducing and sharing specific views, the frontend application does not follow the single-page-application logic. Page URLs are composed of diverse endpoints and query parameters leading to different pages with content specified by the parameter values.

One of the pages is designed to illustrate the current status of the system, informing about the most recent updates of the variables and the evaluation of them compared to the historical data (Figure 5.1). It is actually a dashboard illustrating four variables -the water reserves, the drinking water production, the sum of the precipitation measurements and the temperature measurements (min, mean, max) in Athens. All these four layouts have the same structure, at top the users are informed when was the last update of the variables and what was the exact value. In the middle is shown the trend of the 30 most recent values. And at the bottom is indicated the evaluation score of the recent values compared to the historical data.

The other pages correspond to the different datasets (water reserves, drinking water production, precipitation, temperatures, water supply interruptions). In these pages the users can set parameters in order to filter the data and specify the exact output. Through these parameters, various preferences can be specified, such as time periods (years, months, dates), entities (water reservoirs, water treatment plants, locations, municipalities), aggregations (time, entity), computations (mean, sum).

The requested data are represented both as graphical charts (Figure 5.2) and as tables of raw values (Figure 5.3). A number of icon tools accompany these representations and users can change the type of chart output (line, area or bar charts are available), can download the outputs as image file (PNG) or as text file (CSV) or can get URLs to reproduce and share the whole page output or just a single chart or the table of values.

Figure 5.1. The dashboard with the most recent values of the variables and their evaluation.

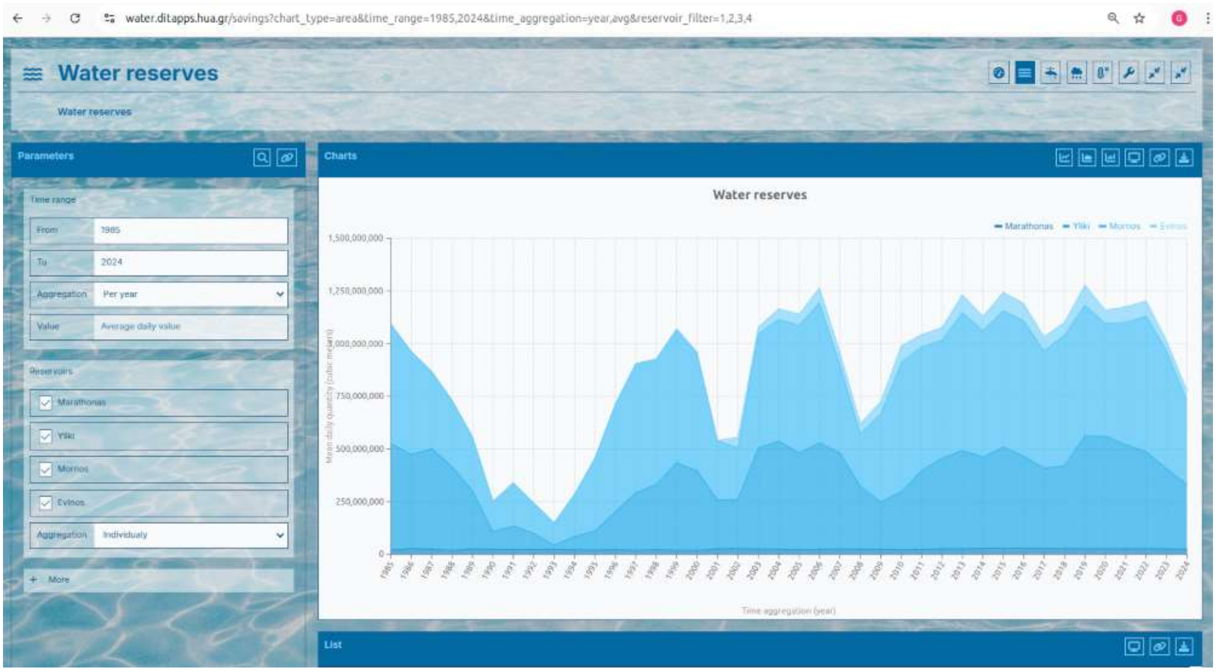

Figure 5.2. The water reserves per year are illustrated as an area chart.

Furthermore the page of the water supply interruptions includes the geospatial analysis of the interruption events. It is illustrated as an interactive map of the greater Athens where the municipalities are drawn from cold to warm colors regarding how much they were disrupted from the interruptions, based on the output of the k-means clustering classification. There are available two analysis alternatives. The one is based on the municipality area taking into account how many events are encountered per sq. km (Figure 5.4). The other is based on the municipality population taking into account how many events are encountered per 1000 inhabitants (Figure 5.5).

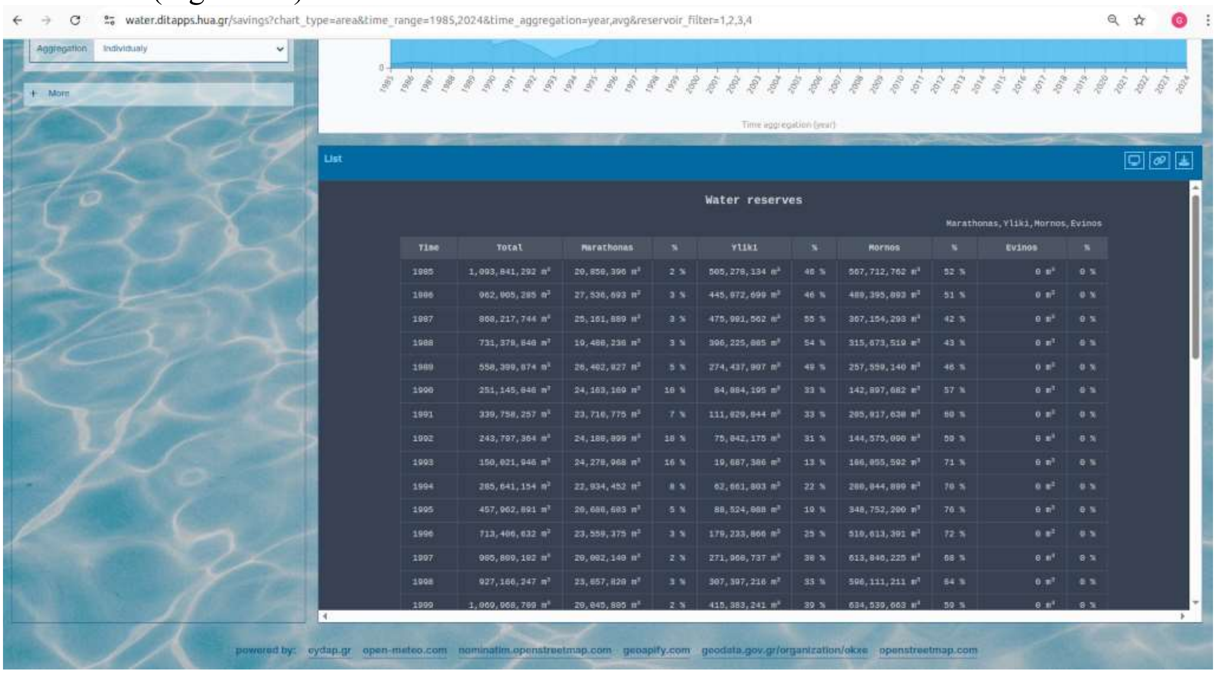

Figure 5.3. The water reserves per year are provided as a table of values .

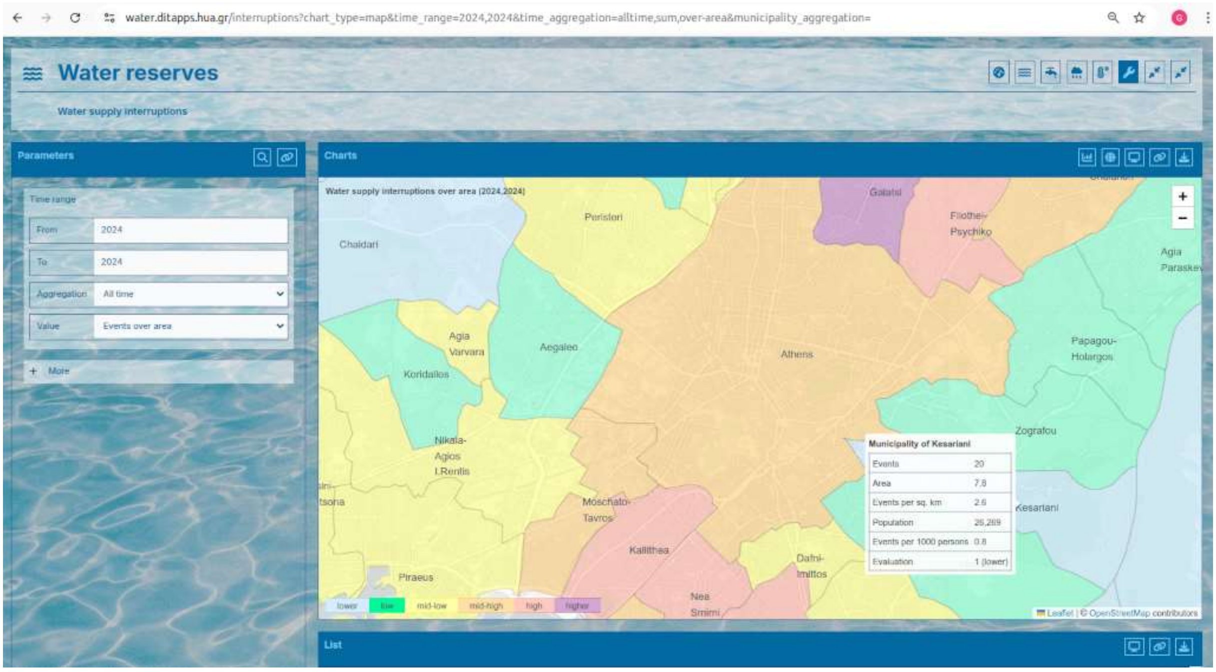

Figure 5.4. The distribution of the water supply interruptions among the municipalities (events per sq. km)

Data analysis on the acquired data

Clustering water reserve levels, water production and weather indexes

The recent values of the variables are evaluated compared to the historical data. Actually it is about 30-days-period comparisons in order to avoid possible single-day outliers. So, the average of the 30 most recent values of each variable is compared with the averages of the same 30-days-periods of the previous years. The K-means clustering algorithm is used, in order for these comparisons to be carried out.

Every single cluster is defined by a centroid value automatically estimated by the algorithm in an unsupervised learning manner. The instances are assigned to the set with the nearest centroid to them (Oti et al., 2021). For our purposes here, a K-means clustering is applied to group and classify all those averages across the years. The sets, or so-called clusters, correspond to different levels of values. Afterwards they are represented as a rating scale from lower to higher points.

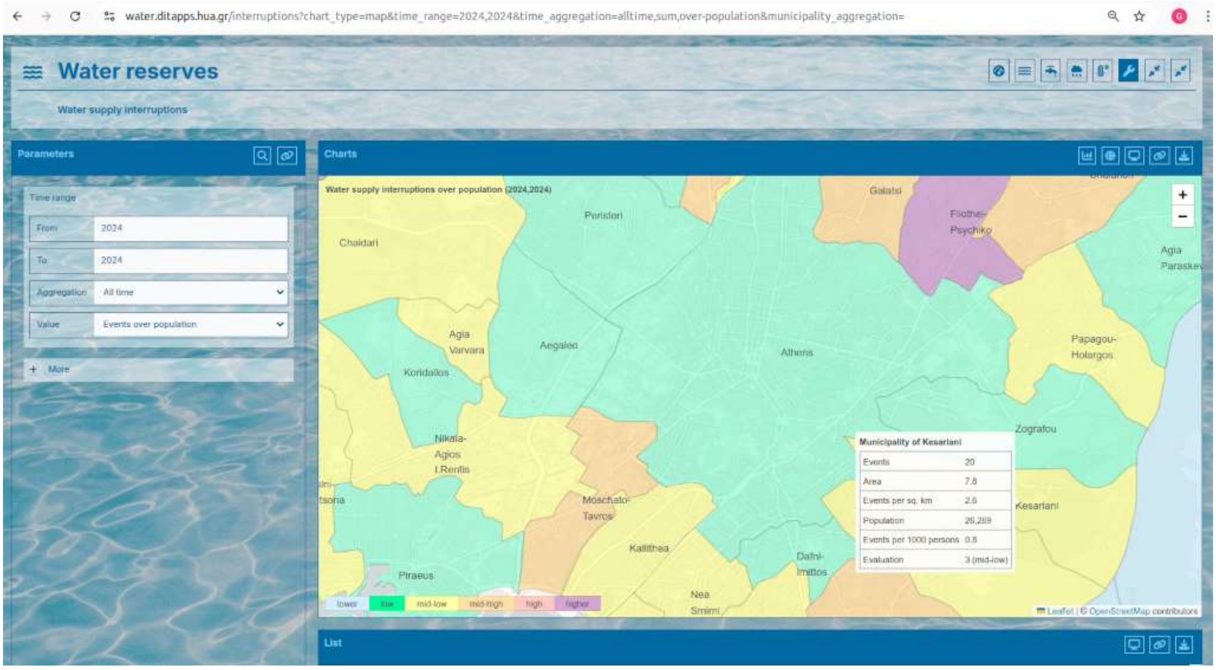

Figure 5.5. The distribution of the water supply interruptions among the municipalities (events per 1000 inhabitants)

The number of the clusters and consequently the length of the rating scale was a question. The representation of a very short scale could be naive and meaningless, on the other hand a quite long scale could seem too confusing. Intuitively, a five-point rating scale seemed to be appropriate.

Additionally, the Elbow and Silhouette methods were applied to check how well different numbers of clusters fit to the different datasets. The Elbow method measures the cluster cohesion, how closely related are the instances within their cluster, while the Silhouette takes into account not only the cluster cohesion but also how separated the clusters are between each other (Saputra et al., 2020). Finally, a four, a five or a six-point rating scale was adopted considering each specific dataset of the four.

For example, Figure 5.6 shows the scores of Elbow and Silhouette methods applied to the water reserves data. A number of six clusters was adopted for that dataset, although a number of two or three clusters gets an even better score regarding the Silhouette method. However the number of six clusters hits a good score regarding both methods, as well as it is closer to that intuitively supposed as more appropriate. Moreover, Table 5.1 shows the centroid values of these six clusters and Table 5.2 lists the averages of the daily water reserves per year and the evaluation score of them.

In a similar way, a four-point rating scale was adopted for the drinking water production evaluation, a five-point rating scale for the precipitation evaluation, and a six-point rating scale for the temperature evaluation.

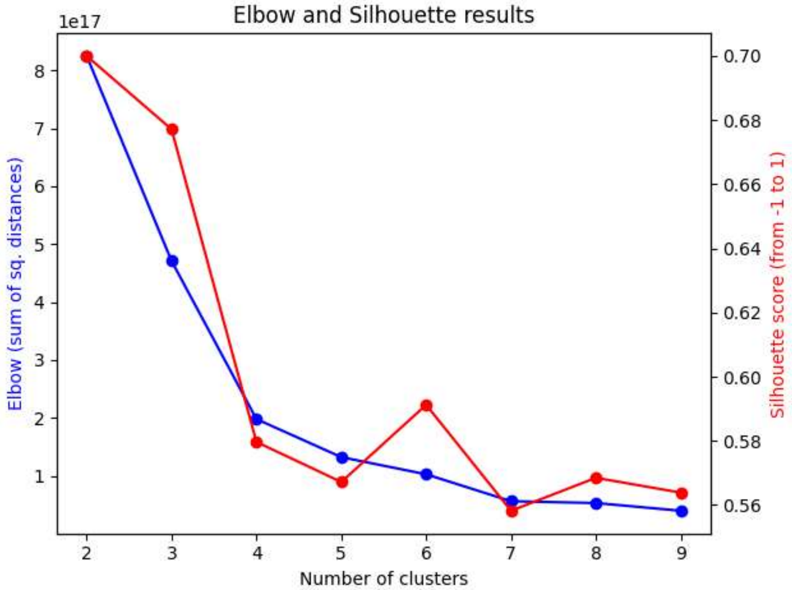

Figure 5.6. The results of Elbow and Silhouette methods applied to water reserves data

Table 5.1. The centroids and the labels regarding the water reserves evaluation

Centroids (m^3)	254,072,753	612,945,583	908,215,661	1,055,842,615	1,166,232,292	1,255,070,428
Labels	1 lower	2 low	3 mid-low	4 mid-high	5 high	6 higher

Table 5.2.The averages of the daily water reserves per year and the evaluation scores of them

Year	Daily average (m^3)	Difference (m^3)	Growth (%)	Evaluation score
1985	1,093,841,292	0	0	4 mid-high
1986	962,905,285	-130,936,007	-12	3 mid-low
1987	868,217,744	-94,687,541	-9.8	3 mid-low
1988	731,378,840	-136,838,904	-15.8	2 low
1989	558,399,873	-172,978,967	-23.7	2 low
1990	251,145,046	-307,254,827	-55	1 lower
1991	339,758,258	88,613,212	35.3	1 lower

1992	243,797,364	-95,960,894	-28.2	1 lower
1993	150,021,946	-93,775,418	-38.5	1 lower
1994	285,641,153	135,619,207	90.4	1 lower
1995	457,962,890	172,321,737	60.3	2 low
1996	713,406,632	255,443,742	55.8	2 low
1997	905,809,101	192,402,469	27	3 mid-low
1998	927,166,247	21,357,146	2.4	3 mid-low
1999	1,069,968,709	142,802,462	15.4	4 mid-high
2000	959,013,397	-110,955,312	-10.4	3 mid-low
2001	540,976,936	-418,036,461	-43.6	2 low
2002	555,437,096	14,460,160	2.7	2 low
2003	1,078,911,761	523,474,665	94.2	4 mid-high
2004	1,165,431,448	86,519,687	8	5 high
2005	1,140,740,940	-24,690,508	-2.1	5 high
2006	1,265,312,537	124,571,597	10.9	6 higher
2007	958,084,141	-307,228,396	-24.3	3 mid-low
2008	620,835,709	-337,248,432	-35.2	2 low
2009	725,166,695	104,330,986	16.8	2 low
2010	989,674,997	264,508,302	36.5	4 mid-high
2011	1,044,216,071	54,541,074	5.5	4 mid-high
2012	1,076,791,238	32,575,167	3.1	4 mid-high
2013	1,232,389,148	155,597,910	14.5	6 higher
2014	1,129,315,303	-103,073,845	-8.4	5 high
2015	1,243,746,000	114,430,697	10.1	6 higher
2016	1,190,968,396	-52,777,604	-4.2	5 high
2017	1,033,002,684	-157,965,712	-13.3	4 mid-high
2018	1,102,115,255	69,112,571	6.7	4 mid-high
2019	1,278,834,027	176,718,772	16	6 higher
2020	1,158,609,413	-120,224,614	-9.4	5 high
2021	1,175,615,283	17,005,870	1.5	5 high
2022	1,202,945,266	27,329,983	2.3	5 high
2023	1,014,061,532	-188,883,734	-15.7	4 mid-high
2024	776,313,716	-237,747,816	-23.4	3 mid-low

Classification of municipalities regarding the water supply interruptions

The municipalities are classified regarding how many water supply interruptions were encountered inside their boundaries. That classification can be carried out based on two alternative indexes, either the interruption events per sq. kilometer or the interruption events per one thousand inhabitants.

Regardless of the selected index, the municipalities are grouped into different clusters through a K-means clustering. Afterwards the clusters are represented as a six-point rating scale from lower to higher points. The number of the clusters and consequently the length of the scale was selected in the same way as it was discussed in the preceding section about the evaluation of the recent values compared to the historical data. Table 5.3 shows some municipality examples, with the results of both alternatives: classification based on area and classification based on population.

Table 5.3. Examples of municipalities classifications

Municipality	Events	Area (km^2)	Events /km^2	Classification based on area	Population	Events /1000 pers.	Classification based on population
Athens	2608	39.0	66.9	5 high	643.452	4.1	2 low
Kallithea	340	4.8	70.8	5 high	97.616	3.5	2 low
Elefsina	381	36.6	10.4	2 low	30.147	14.2	4 mid-high
Galatsi	392	4.0	98.0	6 higher	57.909	6.8	3 mid-low
Peristeri	314	10.1	31.1	3 mid-Low	133.630	2.4	1 lower
Piraeus	688	10.9	63.1	5 high	168.151	4.1	2 low
Vrilissia	11	3.9	2.8	1 lower	32.417	0.3	1 lower

Correlation hypothesis

A correlation hypothesis has been checked between the water reserves and the precipitation. It is more discussed as an indicative example of the possible relations between the different variables. Platform's precipitation measurements originate from eight cities located in Central Greece, in the wider area where the fourth water reservoirs of EYDAP are located too. Potentially other areas could be included based on a more complete hydrological justification.

At first, a similarity was visually detected in line charts between the growth of a single year precipitation and the trend of the following year water reserves (Figure 5.7). In order to check if an association exists between the two variables, Pearson and Spearman correlation coefficients were applied on the data. Such coefficients explore if a change in the one variable is associated with a change in the other, but no causation can be inferred from the correlation results (Schober et al., 2018).

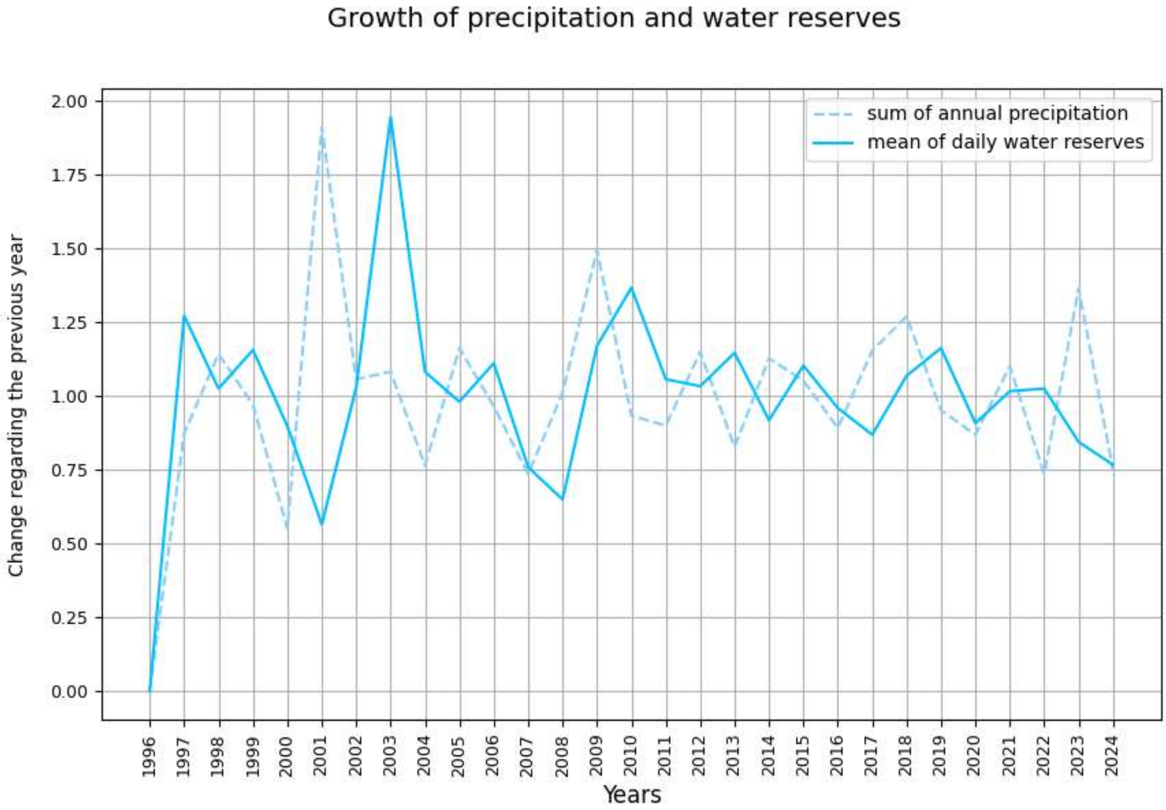

Figure 5.7. The change of the precipitation fall and the water reserves regarding the previous year quantities

Table 5.4 shows the water reserves and the precipitation measurements that the correlation coefficients were applied on. The precipitation of a single year was compared to the next year's water reserves. The time period that was taken into consideration was from 1996 assuming similar conditions regarding the water reserves infrastructure. In 1996 the construction of Evinos-Mornos tunnel was already completed and the water from the Evinos river was transported to boost the Mornos reservoir quantities (EYDAP, 2024a).

A first computation was carried out from 1996 to 2023 and both coefficients indicated a good correlation, the Pearson's value was 0.68 and the Spearman's 0.62. A second computation was carried out extending the time period by one year, from 1996 to 2024. The coefficients still indicated the existence of correlation but their values decreased, the Pearson's value became 0.58 and the Spearman's 0.53.

Table 5.4. The sum of the precipitation measurements and the mean of the daily water reserves per year

Year	Precipitation (mm)	Water reserves (m^3)	Year	Precipitation (mm)	Water reserves (m^3)
1996	5,647	713,406,632	2011	5,139	1,044,216,071
1997	4,923	905,809,101	2012	5,897	1,076,791,237
1998	5,615	927,166,247	2013	4,869	1,232,389,147
1999	5,440	1,069,968,708	2014	5,484	1,129,315,303

2000	2,991	959,013,397	2015	5,760	1,243,746,000
2001	5,711	540,976,936	2016	5,126	1,190,968,396
2002	6,025	555,437,095	2017	5,905	1,033,002,684
2003	6,511	1,078,911,760	2018	7,486	1,102,115,254
2004	4,969	1,165,431,447	2019	7,113	1,278,834,027
2005	5,763	1,140,740,940	2020	6,166	1,158,609,412
2006	5,552	1,265,312,537	2021	6,777	1,175,615,282
2007	4,080	958,084,140	2022	4,968	1,202,945,265
2008	4,116	620,835,709	2023	6,766	1,014,061,531
2009	6,144	725,166,694	2024	5,067	776,313,715
2010	5,723	989,674,997			

6. Conclusion

The 6 data sources (water reserves, drinking water production, precipitation, temperatures, water supply interruptions, municipalities' geographical boundaries) and the 3 web services (the two geolocation services and the openstreetmap) we have already seen, are effectively combined with software development technologies. The proposed system is featured by automated data acquisition through the REST API and by flexible data filtering through the UI. The created system managed to address efficiently the data collection and curation challenges, presenting diverse means of exploiting the available data, from a visual representation to a programmatically oriented usage through the provided API.

The researchers and practitioners might benefit from a rather convenient way to search and get data relevant to their investigation or management questions. The policy makers might prioritize the monitoring and the upgrading of the more disruptive areas of the water supply network according to the results of the geospatial analysis. Indicative data analysis has been performed in order to indicate the usefulness of the acquired data as well as their comparative classification in time.

Finally, due to the correlation results, more research may be appropriate to explore possible relations between the water reserves and more specific precipitation measurements or even other weather variables.

Acknowledgements
Part of the research leading to the results presented in this paper has received funding from the European Union's funded Project HUMAINE under grant agreement No. 101120218.

References

Cosgrove, W. J., & Loucks, D. P. (2015). Water management: Current and future challenges and research directions. *Water Resources Research*, 51(6), 4823-4839.

Criollo, R., Velasco, V., Nardi, A., De Vries, L. M., Riera, C., Scheiber, L., ... & Vázquez-Suñé, E. (2019). AkvaGIS: An open source tool for water quantity and quality management. *Computers & Geosciences*, 127, 123-132.

Delipetrev, B., Jonoski, A., & Solomatine, D. P. (2014). Development of a web application for water resources based on open source software. *Computers & Geosciences*, 62, 35-42.

Drost, S., Vogt, A., Danowski-Buhren, C., Jirka, S., Kirstein, V., Pakzad, K., & Rieke, M. (2022). WaCoDiS: Automated Earth Observation data processing within an event-driven architecture for water monitoring. *Computers & Geosciences*, 159, 105003.

EYDAP (2024a) Water supply resources, Available at: https://www.eydap.gr/en/thecompany/water/watersources (accessed: 30 Dec. 2024).

EYDAP (2024b) Water treatment plants, Available at: https://www.eydap.gr/en/thecompany/water/waterplants (accessed: 30 Dec. 2024).

Fielding, R. (2000). Architectural styles and the design of network-based software architectures. Ph.D. Dissertation. University of California, Irvine.

Horsburgh, J. S., Black, S., Castronova, A., & Dash, P. K. (2025). Advancing open and reproducible water data science by integrating data analytics with an online data repository, *Environmental Modelling & Software*, Volume 188, 2025, 106422, ISSN 1364-8152

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2), 451-461.

Nyabuto, G. M., Mony, V., & Mbuguah, S. (2024). Architectural Review of Client-Server Models. *International Journal of Scientific Research and Engineering Trends*, *10*(1), 139–143.

Oti, E. U., Olusola, M. O., Eze, F. C., & Enogwe, S. U. (2021). Comprehensive review of K-Means Clustering Algorithms. *International Journal of Advances in Scientific Research and Engineering*, 07(08), 64–69.

Saputra, D. M., Saputra, D., & Oswari, L. D. (2020). Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method. *Advances in Intelligent Systems Research*, *172*, 341–346.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768.