Dr. Bayomock Linwa André Claude, 2025, 13:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Efficient Water Data Consumption and Prediction Model: SODECI Case

[1]Dr. Bayomock Linwa André Claude, [2]Mrs. Dosso Nofogon Grace Marienne

[1]Chief of Computer Science Department, International University of Grand-Bassam
[2]Bachelor of Computer Science, International University of Grand-Bassam

Abstract - Water is an essential yet limited resource whose management is increasingly challenging due to urban growth, climate variability, and limited infrastructure, especially in developing countries. In Côte d'Ivoire, the national water utility SODECI oversees production, distribution, and consumption monitoring, but faces significant obstacles. These include reliance on manual meter readings, delayed or inaccurate data from smart meters, and the absence of real-time alerts. As a result, both utilities and consumers struggle with unreliable consumption data, leading to billing inaccuracies, undetected leaks, mistrust, and inefficient planning. Many water utilities in Sub-Saharan Africa operate with fragmented, low-digital systems and lack structured, continuous data needed for effective analysis. They have limited tools for understanding historical trends, identifying anomalies, or forecasting demand. Existing systems focus primarily on billing rather than data-driven monitoring or prediction. To address these gaps, the paper proposes a new system aimed at improving water consumption monitoring and understanding for both consumers and utilities in Côte d'Ivoire. The system integrates data collection, analysis, and visualization to enhance transparency and decision-making. It provides insights tailored to different stakeholders and supports proactive management of water use. The paper proposes a suitable and predictable water consumption data model that captures abnormal events and alerts consumers and producer. Algorithms that cleanse and retrieve the abnormal water consumption from a given data set are proposed too.

Index Terms – Water, Consumption, Prediction, Data Model, Algorithm, Analysis, Architecture, Transparency.

I. INTRODUCTION

Water is a vital and limited resource essential to public health, economic development, and environmental sustainability. As urban populations grow and climate conditions being more unpredictable, the water use management becomes a central concern for utilities, governments, and households alike.

For a country perspective, the water life cycle can be classified 5 main phases (with their associated metrics): water possession, clean water production, clean water transportation, water consumption. To capture and meet the population needs in water, the water consumption is the main metric frequently used to control and alert billed clients (individuals, households, or organizations).

Understanding and managing this metric is critical in development countries, where infrastructure and service delivery systems often face structural and budget limitations. In Côte d'Ivoire, for instance, water management and consumption are part of the objectives of National Water Distributor Company commonly called SODECI (Société de Distribution d'Eau de la Côte d'Ivoire).

Meanwhile, achieving these objectives is challenging because of factors correlations as real water counter measure, involvement of the client to provide real measure, intelligent counters that may delay the real time counter values information that are essentials to the account owners and/or the water supplier to take actions in case of an abnormal change water consumption events.

In many regions of Sub-Saharan Africa, Côte d'Ivoire included, remote access to get accurate and timely lecture data on water counter is limited; and alert distribution signals of water consumption change do not exist. Utility providers often rely on manual meter readings, incomplete records, or outdated systems, making it difficult to monitor consumption effectively.

This lack of data records in client water consumption changes can lead to inaccurate billing, undetected leaks, and inefficient resource planning. On the consumer side, clients often have no clear view of their usage patterns or how their bills are calculated, which can generate mistrust and limit opportunities for conservation. Furthermore, the absence of historical usage data hinders efforts to detect abnormal consumption or anticipate future needs.

These challenges are amplified by the rapid urbanization of cities, the variability of water availability, and the absence of modern systems for consumption monitoring in many communities. These observations point to a deeper issue: while the operational difficulties are well known, they are rarely addressed through tools designed specifically to fill the informational and analytical gaps.

The effective management of water consumption depends heavily on the availability of structured, reliable, and continuous data. However, in practice, there is often a disconnection between water consumption in the field and the information systems intended to track it. Most utilities operate with limited digital infrastructure, and their data collection processes are generally manual, fragmented, or error-prone.

There is often no systematic way to analyze past consumption trends, identify irregular usage patterns, or forecast future demand. This lack of analytical insight poses serious risks for both utilities and clients, including resource mismanagement, undetected anomalies, and reduced trust in service delivery.

Despite the critical need for more advanced monitoring and analysis, few tools exist that are tailored to the operational realities of water utilities in developing regions. In particular, there is a noticeable absence of systems that are accessible, user-friendly, and capable of supporting both agents and consumers in understanding and managing water consumption more effectively.

In SODECI for example, the existing water management tools mainly focus on billing and infrastructure, with low capabilities in data analysis or prediction tailored to challenges. There is also a lack of accessible platforms that serve both utility agents and consumers effectively.

In this paper, a system designed to improve the monitoring and understanding of water consumption for both clients and utility providers in Côte d'Ivoire is proposed. By leveraging available data, the system aims to enhance transparency and support better decision-making regarding water usage.

The system integrates data collection, analysis, and visualization components to provide users with insights into their water consumption patterns. It serves multiple stakeholders by offering relevant information tailored to their roles, thereby facilitating improved management and awareness of water use.

The organization of this paper is as follows: after the introduction, the motivations behind this work are explained, followed by a summary of related work. Then, the water consumption design model is presented, including the data collection approach, algorithms to improve data quality, data granularity, and the description of the data model. Afterwards, the application requirements are detailed, followed by the architectural solution. The implementation process is then described, and the results are presented and analyzed. And finally, the conclusion is given.

Motivations

The motivations behind this research are classified into four main categories: economic, social, environmental, and technical/operational. Each plays a significant role in justifying the implementation of

a water consumption monitoring and prediction system.

Economic motivation is one of the strongest driving forces. Water is a crucial economic good, and its management has a direct financial impact on both service providers and consumers. In many developing regions, including Côte d'Ivoire, water utility companies face significant challenges in collecting accurate and timely consumption data. These inefficiencies often lead to revenue loss, whether through under billing caused by faulty meters or the inability to detect unauthorized usage. For clients, the absence of detailed consumption data makes it difficult to plan budgets or understand unexpected charges on their invoices. By implementing a system capable of continuously monitoring and analyzing water usage, companies can improve the precision of billing, recover losses due to errors or fraud, and optimize their resource allocation. This, in turn, supports financial sustainability and enables long-term investment in infrastructure improvements.

From a social perspective, the lack of visibility into personal water consumption frequently creates frustration among clients. Many households and businesses receive bills that they cannot verify or understand, leading to mistrust toward the utility provider. By giving users access to historical and real-time data about their consumption, the system can strengthen communication between clients and the company, promote awareness, and reduce disputes. It also empowers individuals to take control of their usage habits, making service delivery more customer-focused.

The environmental motivation stems from the global imperative to preserve water resources. Water is increasingly scarce due to climate change, population growth, and pollution. Wasteful consumption—whether intentional or due to leaks and faulty infrastructure—poses a threat to sustainability. A smart monitoring and prediction system can identify unusual usage early, allowing rapid intervention and preventing further loss. It also encourages behaviors that are more responsible by making consumption patterns visible to users and stakeholders.

Lastly, the technical and operational motivation relates to the need for better decision-making tools in utility management. Traditional systems rely heavily on manual processes and delayed reporting, which limit responsiveness and scalability. Integrating data science techniques into the monitoring system allows for the detection of subtle trends, forecasting of future needs, and identification of areas at risk. This supports proactive decision-making, optimizes maintenance scheduling, and enhances service delivery overall. It also aligns with broader efforts to digitize public utilities and modernize infrastructure through technology.

II. RELATED WORK

Recent years have seen a growing interest in digital systems that monitor, analyze, and predict water consumption to enhance utility management and promote sustainability. This literature review presents an integrated view of prior research organized around key conceptual streams: data collection and sensor integration, forecasting algorithms for water demand, anomaly detection and alert mechanisms and data granularity and representation.

To begin, data collection and sensor integration refers to the use of smart devices, such as IoT-enabled sensors and smart water meters, to capture real-time or high frequency water consumption data. These technologies play a crucial role in digitizing water infrastructure and providing structured datasets essential for subsequent analysis. For example, Alvisi and Franchini (2018) proposed a monitoring framework based on district-metered areas, which allowed better segmentation and understanding of consumption behavior at a granular level. Similarly, Cominola et al. (2015) introduced a sensor-based architecture for disaggregating residential water use. Their system generated high-resolution data capable of distinguishing between different types of household usage, thereby supporting more precise and user-

specific water demand assessments. These contributions collectively highlight the foundational role of data acquisition in enabling intelligent water management systems.

Following data collection, the next step typically involves forecasting algorithms for water demand, which are designed to predict short-term or long-term consumption patterns. This component is critical for both operational planning and anomaly anticipation. Forecasting methods range from traditional statistical models to modern machine learning techniques. Herrera et al. (2010) applied artificial neural networks (ANNs) to estimate urban water demand, demonstrating superior performance in short-term prediction. In a comparative analysis, Adamowski et al. (2012) evaluated models such as support vector regression (SVR) and autoregressive integrated moving average (ARIMA), showing that machine-learning models often outperform classical approaches in capturing nonlinear demand patterns. Another noteworthy study by Zhou et al. (2000) introduced a hybrid model that integrated wavelet transforms with neural networks to capture both trends and anomalies in consumption time series. These diverse forecasting approaches underscore the adaptability of predictive tools across varying data environments and support the rationale for embedding such techniques into smart utility systems.

Complementing prediction is the task of identifying anomalies in water consumption, which encompasses both technical faults (e.g., leaks or meter malfunctions) and behavioral irregularities (e.g., fraud or unusual user activity). Anomaly detection and alert mechanisms refer to algorithmic processes that automatically flag such deviations from expected consumption patterns. Carrasco-Jiménez et al. (2021) proposed an unsupervised clustering approach to detect both point and pattern anomalies in household data, offering a tool to distinguish between different sources of irregularity without relying on labeled datasets. Labura et al. (2025) extended this work by combining Isolation Forest with FFT-based filtering in a robust framework tailored to smart meter data. Their model operated across both time and frequency domains, effectively detecting a wide range of consumption anomalies in data-scarce and infrastructure-constrained settings. These innovations reduce the burden of manual inspection and allow utilities to respond quickly to potential issues.

Closely tied to these functionalities is the issue of data granularity and representation, which concerns how consumption data is structured, segmented, and processed for downstream analysis. Pavlou et al. (2024) conducted a comparative study of model-based and learning-based approaches for non-intrusive household water end-use disaggregation. Using 1-minute resolution inlet flow data, both methods achieved similar classification performance, but the learning-based model offered real-time, privacy-preserving processing capabilities suitable for deployment on edge devices. This finding is particularly relevant to systems designed for transparency and responsiveness in regions with limited digital infrastructure, such as parts of Sub-Saharan Africa.

While these studies offer substantial advancements, there remain critical gaps and opportunities. Most existing platforms address specific tasks—such as forecasting or anomaly detection—in isolation. Few systems integrate prediction, real-time alerts, and billing estimation within a single, user-facing interface tailored to the operational realities of utilities in developing regions. Moreover, much of the literature is rooted in Western or Asian contexts, where data availability and infrastructure differ significantly from Sub-Saharan Africa. Issues such as multilingual user interfaces, context-specific anomaly types (e.g., billing errors or unregistered users), and inclusive system design are often overlooked.

Building upon these insights, the current paper proposes a comprehensive platform that not only forecasts water consumption but also integrates anomaly detection and multilingual user alerting, AlfredoECN (2018). This unified approach aims to bridge the gap between academic research and practical implementation by providing an adaptable solution for utilities like SODECI, ultimately contributing to improved service delivery,

transparency, and customer satisfaction in under-resourced environments.

Requirements

Functional requirements

User Management

FR1: The system must allow users (clients, admins) to create an account.

FR2: The system must enable users to log in and log out securely.

FR3: The system must allow users to change their passwords and see their profile information

FR4: Admins must be able to modify, and delete user accounts.

Data Input and Collection

FR5: The system must collect historical water consumption data directly from SODECI's database.

FR6: The system must automatically retrieve and integrate real-time water consumption data from SODECI's database.

FR7: The system must preprocess the collected data by validating and formatting it for analysis.

Data Analysis, Predictions and Visualization

FR8: The system must analyze water consumption patterns for individual clients.

FR10: The system must generate predictive models for future water consumption based on historical data.

FR11: The system must display results via user-friendly charts.

FR12: Users can save the different charts as screenshots.

Threshold and Alert Management

FR13: The system must allow users to set consumption thresholds

FR14: The system must send alerts via email o when consumption tends to be exceeded or exceeds the defined thresholds.

Performance and Security

FR17: The system must process and return analysis results within a reasonable time

FR18: The system must encrypt sensitive data, such as user passwords

FR19: The system must restrict access to admin and agent features to authorized personnel only.

Security requirements

The implementation of all solutions described above requires providing data safety to preserve private or confidential information. Therefore, security is an important aspect of the platform.

Sensitive Data

All information on the platform is considered sensitive. In order to access the platform and its functionalities, each user must provide validated credentials. The most sensitive data are:

User Data: User account data are crucial as they determine the level of access.

Roles: The role of a user cannot be determined by the user themselves to prevent unauthorized access.

Identifier and Personal Data: The identifier and personal data will be used for account verifications.

Password: To ensure security, passwords will be encrypted

Access Control

Client Accounts: Access to platform features related to clients is restricted to admins only. Clients cannot access the administration features.

Login Credentials:

Each user must login with their password and their username or identifier or email

Method Restrictions

Create or Modify User Accounts: All users can create accounts except agents. Only users with the role being set to admin can modify or remove a user account.

Sensitive Deletions: Deleting clients or their associated data is irreversible. Only admins are allowed to remove client accounts.

Network requirements

The system requires a stable and reliable internet connection for efficient operation. This internet connection is necessary for the system to communicate with external sources such as SODECI's database, access data, and perform real-time processing of water consumption and predictions.

Software requirements

The application requires specific software components to function correctly and efficiently. These requirements are divided into backend, frontend, libraries, and development tools.

Backend Environment

Programming Language: Python 3.x or higher, to support data processing, predictive modeling, and backend logic.

Database Management System: MySQL or a compatible relational database to store users, water consumption records, predictions, alerts, and billing data.

Frontend Environment

Web Application Framework: A framework capable of building interactive web interfaces and dashboards, such as Streamlit. Streamlit allows rapid development of dashboards for data visualization and reporting, Avikumar Talaviya (2024),.

Libraries and Dependencies

The system requires libraries that provide the following functionalities:

Data processing and manipulation: For handling numerical computations and structured data.

Visualization: For creating interactive charts and dashboards to monitor water consumption and predictions.

Machine Learning: For developing predictive models based on historical water consumption.

Security: For managing sensitive information, including encrypted passwords.

Email Integration: For sending automated alerts and notifications.

Development and Management Tools

Code Editor/IDE: Visual Studio Code or similar software to write, debug, and maintain Python code efficiently.

Version Control: A system like Git to track changes in the codebase and collaborate efficiently.

Database Management Tools: Optional graphical interfaces such as PhpMyAdmin or DBeaver to facilitate database administration.

III. WATER CONSUMPTION DATA MODEL

Data Collection Approach

The foundation of any predictive system lies in the quality and quantity of its dataset. Since access to real client consumption data from SODECI was restricted due to privacy and security concerns, an alternative strategy was required to simulate realistic data for model development and testing.

Initially, 300 dummy records were created using SQL queries. These records served as a proof of concept, allowing the system design to be validated and tested in terms of database integration, user registration, and preliminary data handling. However, the limited dataset was not sufficient for building robust predictive models, as machine learning algorithms typically require a larger and more diverse sample to capture consumption patterns and trends.

It is important to note the distinction between dummy records and synthetic records. Dummy data is generally minimal, static, and artificial, created mainly to test technical processes such as database connectivity or schema correctness, without aiming to reflect realistic behaviors. By contrast, synthetic data is generated to statistically resemble real-world data distributions, often introducing realistic variability across features like timestamps, meter readings, or consumption levels. In this way, dummy data ensures functionality, while synthetic data enables meaningful experimentation and predictive analysis.

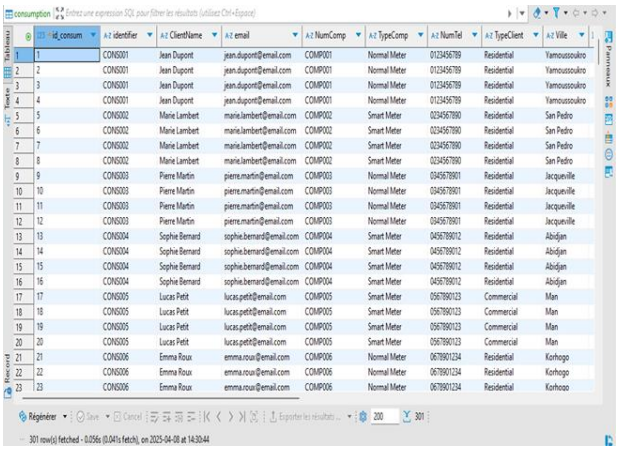
To overcome the limitations of the dummy dataset, the Faker Python library was employed to generate 10,000 synthetic data records. Faker is a widely used tool for creating artificial but realistic datasets by simulating attributes such as dates, numerical ranges, and categorical variables. In this project, Faker was customized to mimic the characteristics of water consumption data, including variables such as client identifiers, meter readings, timestamps, and consumption values.

The expansion from 300 to 10,000 records offered two main benefits:

- Statistical Reliability – Larger sample size enabled more accurate representation of possible consumption behaviors.
- Model Training and Validation – With more diverse records, predictive models could generalize better and avoid overfitting to a small dataset.

Although synthetic, this data collection approach provided a reliable foundation for analysis and prediction, enabling the development and testing of the system in the absence of real-world data. In future work, integration with actual SODECI

consumption data could further enhance accuracy and practical relevance.



id_consum	identifier	ClientName	email	NumComp	TypeComp	NumTel	TypeClient	Ville
1	CONS001	Jean Dupont	jean.dupont@email.com	COMP001	Normal Meter	0123456789	Residential	Yamoussoukro
2	CONS001	Jean Dupont	jean.dupont@email.com	COMP001	Normal Meter	0123456789	Residential	Yamoussoukro
3	CONS001	Jean Dupont	jean.dupont@email.com	COMP001	Normal Meter	0123456789	Residential	Yamoussoukro
4	CONS001	Jean Dupont	jean.dupont@email.com	COMP001	Normal Meter	0123456789	Residential	Yamoussoukro
5	CONS002	Marie Lambert	marie.lambert@email.com	COMP002	Smart Meter	0234567890	Residential	San Pedro
6	CONS002	Marie Lambert	marie.lambert@email.com	COMP002	Smart Meter	0234567890	Residential	San Pedro
7	CONS002	Marie Lambert	marie.lambert@email.com	COMP002	Smart Meter	0234567890	Residential	San Pedro
8	CONS002	Marie Lambert	marie.lambert@email.com	COMP002	Smart Meter	0234567890	Residential	San Pedro
9	CONS003	Pierre Martin	pierre.martin@email.com	COMP003	Normal Meter	0345678901	Residential	Jacqueville
10	CONS003	Pierre Martin	pierre.martin@email.com	COMP003	Normal Meter	0345678901	Residential	Jacqueville
11	CONS003	Pierre Martin	pierre.martin@email.com	COMP003	Normal Meter	0345678901	Residential	Jacqueville
12	CONS003	Pierre Martin	pierre.martin@email.com	COMP003	Normal Meter	0345678901	Residential	Jacqueville
13	CONS004	Sophie Bernard	sophie.bernard@email.com	COMP004	Smart Meter	0456789012	Residential	Abidjan
14	CONS004	Sophie Bernard	sophie.bernard@email.com	COMP004	Smart Meter	0456789012	Residential	Abidjan
15	CONS004	Sophie Bernard	sophie.bernard@email.com	COMP004	Smart Meter	0456789012	Residential	Abidjan
16	CONS004	Sophie Bernard	sophie.bernard@email.com	COMP004	Smart Meter	0456789012	Residential	Abidjan
17	CONS005	Lucas Petit	lucas.petit@email.com	COMP005	Smart Meter	0567890123	Commercial	Man
18	CONS005	Lucas Petit	lucas.petit@email.com	COMP005	Smart Meter	0567890123	Commercial	Man
19	CONS005	Lucas Petit	lucas.petit@email.com	COMP005	Smart Meter	0567890123	Commercial	Man
20	CONS005	Lucas Petit	lucas.petit@email.com	COMP005	Smart Meter	0567890123	Commercial	Man
21	CONS006	Emma Roux	emma.roux@email.com	COMP006	Normal Meter	0678901234	Residential	Korhogo
22	CONS006	Emma Roux	emma.roux@email.com	COMP006	Normal Meter	0678901234	Residential	Korhogo
23	CONS006	Emma Roux	emma.roux@email.com	COMP006	Normal Meter	0678901234	Residential	Korhogo

Table 1: Data sample

Algorithms to Improve the Quality of Collected Data

One important step applied to improve the quality of the collected dataset was capping, a data preprocessing technique used to manage extreme values or outliers. Outliers can occur naturally in datasets, but in the case of synthetic data generated with the Faker library, extremely high or unrealistic values sometimes appeared that did not reflect actual patterns of water consumption. If left untreated, such extreme values could distort statistical measures such as the mean and standard deviation, and negatively affect the accuracy of predictive models.

Capping addresses this issue by setting upper and lower thresholds for a variable, beyond which values are replaced by the threshold itself. For example, if typical household water consumption values fell within a defined realistic range, any unusually high consumption values beyond this range were capped at the maximum threshold. This ensured that the dataset remained representative of realistic conditions while still preserving variability within acceptable limits.

By applying capping, the dataset became more stable and less sensitive to extreme values, which improved the robustness of the prediction models. This preprocessing step was particularly important in ensuring that the synthetic data more closely approximated real-world consumption behaviors, thereby strengthening the reliability of the system's analysis and forecasts.

Data Granularity

In addition to applying capping techniques to handle extreme values, the project carefully considered data granularity, which refers to the level of detail at which data is collected and stored. For water consumption, the dataset was organized across multiple time scales: yearly, semester, quarterly, and monthly. This multi-level granularity enabled the system to capture both long-term trends and short-term variations in consumption patterns.

- Yearly granularity provides a broad overview, useful for strategic planning, detecting overall trends, and benchmarking total consumption.
- Semester-level granularity allows the identification of medium-term changes, such as shifts between wet and dry seasons or the effects of half-year policy interventions.
- Quarterly granularity is particularly effective for uncovering seasonal fluctuations and understanding patterns within a year.
- Monthly granularity captures the finest short-term variations, enabling the detection of anomalies, unusual spikes, or sudden drops in consumption.

Selecting these granularities ensured that the data was neither too coarse to overlook important patterns nor too fine to introduce unnecessary noise. For instance, monthly analysis highlights seasonal changes that would be invisible in yearly summaries, while annual aggregation provides a clear perspective for long-term forecasting. By aligning data granularity with the objectives of monitoring and prediction, the project ensured that the predictive models were trained on data reflecting meaningful consumption behaviors at multiple time scales.

Overall, the combination of carefully chosen granularity levels and capping of extreme values enhanced the accuracy, reliability, and interpretability of the dataset. This approach allowed the predictive models to deliver insights and forecasts that are both statistically robust and

practically useful for water consumption monitoring and management.

Description and Representation of the Data Model

The data model for this project is designed to support the analysis, prediction, and alerting of water consumption for clients. The model is implemented in MySQL and captures detailed information about users, consumption records, predictions, alerts, and geographic areas. It ensures data consistency, enforces relationships between entities, and supports predictive analytics.

Entities and Their Roles

The data model includes the following key entities:

- User
- A user is either an Agent or an Admin or a Client. Users can receive alerts and interact with consumption data.
- Neighborhood
- A Neighborhood is a geographic area within a city and contains multiple meters. Each neighborhood belongs to one zone and is associated with multiple users, such as clients and agents.
- City
- A City is the broader geographic area that includes various neighborhoods.
- Consumption
- A Consumption represents the amount of water used by a client as measured by a meter. A single consumption can generate predictions, analysis, alerts, invoices, and apply rates. It is linked to a specific client.
- Alert
- An Alert is generated based on a consumption or a prediction. Alerts can be preventive (warning of potential overuse) or informative (information on exceeding the threshold).

Prediction

- A Prediction forecasts future water consumption based on historical data and trends. Predictions are generated from a consumption record and can trigger alerts related to excessive usage or other significant factors.

Type of Alert

- Defines the two types of alerts: informative (provides general information) and preventive (warns of potential overuse).

Role

- This table defines the roles of users in the system. It helps in distinguishing between agents, clients and admin.

Predictive Alert

- This table stores alerts that are generated based on predictions. These alerts typically warn about potential future overuses based on consumption trends or predictions.

Informative Alert

- This table stores alerts that provide general information, generally not related to predictive or potential problems, but rather for information purposes, such as when a threshold has been exceeded.

Billing threshold

- This table store the different thresholds set by the client.

Relationships and Management Rules

The data model enforces several business rules and relationships to ensure data integrity:

- A user is either an admin, an agent or a client (role).
- A neighborhood is located in a city, and a city contains multiple neighborhoods.
- A client can consume water in one of three contexts: domestic, commercial, or industrial (type of client).
- A user can receive at least one alert, and an alert is sent to one user.
- A consumption can generate at least one prediction, and a prediction is generated by a consumption.
- A consumption can generate at least one alert, and an alert is generated by a consumption.
- A prediction can generate at least one alert, and an alert is generated by a prediction.
- An alert can be either preventive or informative (type of alert).
- An alert can be generated based on a threshold and a threshold can generate many alerts.

Attributes and Constraints

Each entity includes attributes with specific data types and constraints to ensure data quality and consistency.

Entity	Attibutes	Type	Constraints
User	User_id	Integer	unique
	Email	Chain of Characters	unique, xxxx@xxx.xx
	phone_number	Chain of Characters	Avoid special characters except +. Control length. avoid typing letters.
	Adress	Chain of Characters	None
	Password_hash	Chain of Characters	Encrypted
	Registration _date	Date time	YYYY-MM-DD HH-mmss
	Is_active	Boolean	True/False
	role_id	Chain of Characters	unique
	Username	Chain of Characters	unique
	Identifier	Chain of Characters	unique
	ClientName	Chain of Characters	None
	Assigned_city	Chain of Characters	None
Role	Role id	Chain of Characters	unique
	Role_name	Chain of Characters	No special characters
	Description	Chain of Characters	None
	Permissions	Chain of Characters	None
Consumption	ClientName	Chain of Characters	None
	Identifier	Chain of Characters	unique
	Email	Chain of Characters	unique, xxxx@xxx.xx
	Id_consum	Chain of Characters	unique
	Invoice id	Chain of Characters	unique
	Conso	Integer	None
	Start date	Date time	YYYY-MM-DD HH-mmss
	End date	Date time	YYYY-MM-DD HH-mmss
	Unit price	Integer	None
	Invoice price	Float	None
	Total price	Float	None
	Meter id	Chain of Characters	unique
	Phone number	Chain of Characters	Avoid special characters except +. Control length. avoid typing letters
	Penalties	Float	None
	Paiement Status	Chain of Characters	None
	Rate	Float	None
	Client type	Chain of Characters	None
	Meter type	Chain of Characters	None
	City	Chain of Characters	None
Prediction	Prediction_id	Chain of Characters	unique
	Client_id	Chain of Characters	unique

	Volume	Float	None
	Prediction_date	Date time	YYYY-MM-DD HH-mmss
	Confidence_level	Float	None
	Model_version	Chain of Characters	None
Billing_threshold	Created_at	Date time	YYYY-MM-DD HH-mmss
	Identifier	Chain of Characters	unique
	Threshold_value	Float	none
	Updated_at	Date time	YYYY-MM-DD HH-mmss
Alert	Alert_id	Chain of Characters	unique
	User_id	Chain of Characters	unique
	Alert_type	Chain of Characters	None
	Message	Chain of Characters	None
	Created_at	Date time	YYYY-MM-DD HH-mmss
	Prediction_id	Chain of Characters	unique
Alert_type	Type_alert_id	Chain of Characters	unique
	Threshold	Float	None
Predictice _Alert	Predictive_alert_id	Chain of Characters	unique
	Alert_type_id	Chain of Characters	unique
	Predicted_date	Date time	YYYY-MM-DD HH-mmss
Informative_Alert	Informative_alert_id	Chain of Characters	unique
	Alert_type	Chain of Characters	unique
	Current_value	Float	None
	Detected_at	Date time	YYYY-MM-DD HH-mmss
City	City_id	Chain of Characters	unique
	City_name	Chain of Characters	unique
Neighborhood	Neighborhood_id	Chain of Characters	unique
	City_id	Chain of Characters	unique
	Name	Chain of Characters	None

Table 2: Entities with Associated Attributes

Representation

The data model is be represented using an Class diagram, illustrating entities, their attributes, and the relationships between them.

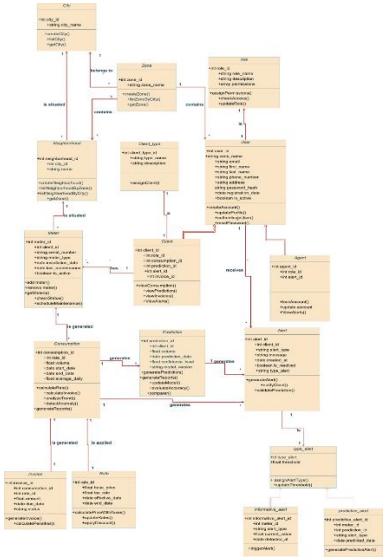


Figure 1: Class diagram

This structure ensures that the system can efficiently store, retrieve, and process data for monitoring,

analyzing, predicting, and alerting water consumption patterns while maintaining data integrity and consistency.

Architecture Solution

The proposed architecture is shown in Figure 2. It has 3 layers: Web, Application, and Data. The Web Layer contains presentation files and handles interactions with the Application Layer to support dynamic web features.

This layer typically uses event listeners to capture user actions during navigation. The Application Layer processes captured events and contains business logic. It interacts with the Data Layer using entity classes, where each entity class is mapped one-to-one with a database table. The Data Layer manages interactions with the physical database using database connectivity protocols like ODBC or similar mechanisms.

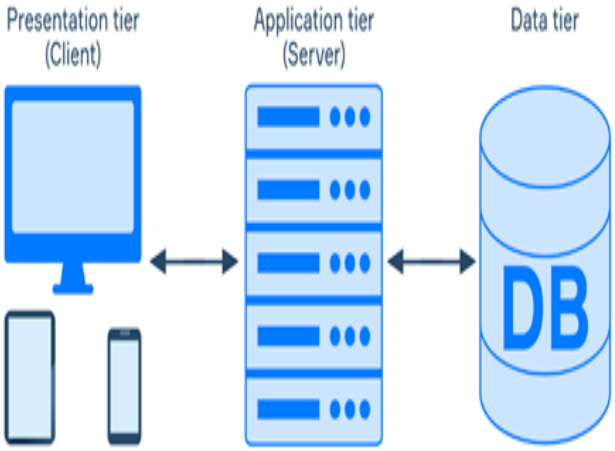


Figure 2: 3-Tiers Architecture for water consumption

The 3-Tiers Architecture was selected due to its functional approach, which separates the presentation, business logic, and data management layers. This design is appropriate for implementing the system for monitoring and predicting water consumption it ensures maintainability, scalability, and clarity in the development process.

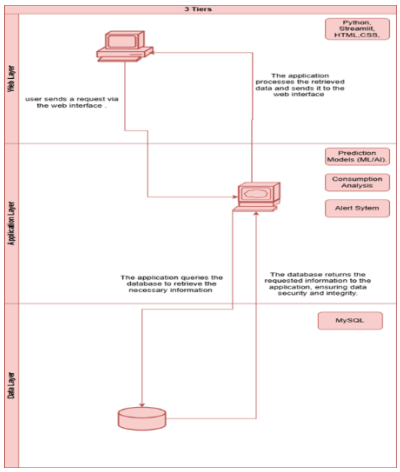


Figure 3: Information Flow Architecture

Algorithms and Implementation

For this research, a Linear Regression model was initially implemented to predict future water consumption based on features like city, time, and client type. However, the results were not satisfactory. The performance metrics, especially the R^2 score, were extremely low (e.g., R^2 = 0.00002), indicating that the model was nearly failing to explain the variance in the target variable. This poor performance led to consider alternative models.

Ridge Regression was then considered, a regularized version of linear regression that introduces a penalty term (controlled by the alpha parameter) to reduce the impact of extreme coefficient values. This approach is especially helpful in cases where the dataset is small or contains multi-collinearity. With Ridge Regression, the evaluation metrics improved significantly, indicating better and more consistent predictive performance.

However, despite using Ridge Regression, the limited number of real observations (only 300 data points) caused high variability in the predictions, which made the results unstable.

To initially address this issue, I applied a technique called capping, which consists of limiting the extreme values (outliers) in a dataset to reduce their disproportionate impact on the model. Capping helped reduce noise and improve the model's behavior on the small dataset.

Still, to better demonstrate the model's effectiveness without being hindered by small sample size and high variation, I used the Faker Python library to generate 10,000 synthetic water consumption entries. This larger dataset allowed the model to learn from a more representative and consistent distribution, leading to more reliable predictions and stronger performance metrics.

Ridge Regression Model

The main idea behind Ridge Regression is to add a penalty (regularization term) to the linear regression loss function to constrain the magnitude of the model coefficients. This regularization helps the model perform better when dealing with multicollinearity (high correlation between features) or when there is not enough data to build a reliable model. Ridge regression works by modifying the linear regression loss function as follows:

Formula 1: Ridge Expression

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

- yi : the actual value of the target variable for observation iii
- Xi: the explanatory (or feature) variables for observation iii
- β: the coefficients (parameters) to be estimated for each feature
- λ : the regularization parameter (also called alpha in some implementations) that controls the importance of regularization
- p: the number of explanatory variables

Number of Features

The ratio between the number of samples and the number of features is critical.

As a general rule, it is recommended to have at least 10 times more observations than features.
For example, my dataset contains 29 features, so a minimum of 290 observations is advisable.

Data Quality

Well-cleaned, low-noise, and representative data can reduce the need for large datasets.

If the features are highly correlated or redundant, more data may be needed to maintain model performance.

Regularization Strength (alpha parameter)

Ridge regression is more stable than standard linear regression, especially in cases of multi-collinearity or when the sample size is small.

Choosing an appropriate regularization parameter (alpha), typically through cross-validation, which is a model evaluation technique that involves splitting the dataset into multiple subsets (folds) to train and test the model on different portions, can improve model robustness even when data is limited.

Metrics for Model Evaluation

To evaluate the performance of the Ridge Regression model, I used some evaluation metrics. These metrics help to understand how well the model is performing and how accurate the predictions are compared to actual values.

Mean Absolute Error (MAE)

MAE measures the average magnitude of the errors between the predicted and actual values. It provides a clear indication of the model's accuracy.

Formula 2: MAE Expression

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Where:

- n is the number of observations
- y_i is the actual value
- $\hat{y}_i$ is the predicted value
- $|y_i - \hat{y}_i|$ is the absolute error for each observation

Range of Values

MAE is always non-negative. Lower values are better. An MAE of 0 indicates perfect predictions, while a larger value indicates worse performance.

R-Squared (R^2)

R^2 measures how well the model explains the variability in the target variable. It provides an indication of how well the independent variables predict the target variable.

Formula 3: R-Squared Expression

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where:

- y_i is the actual value
- $\hat{y}_i$ is the predicted value
- $\bar{y}$ is the mean of all actual values
- The numerator is the residual sum of squares (RSS)
- The denominator is the total sum of squares (TSS)

Range of Values

R^2 ranges from 0 to 1. A value of 1 indicates perfect predictions, while a value closer to 0 indicates poor performance. A negative R^2 indicates that the model performs worse than simply predicting the mean value.

These metrics were used to assess the overall performance of the Ridge Regression model. By evaluating the model's MAE and R^2. I was able to determine how well the model was performing in predicting future water consumption. Each metric provided a different perspective on the model's performance, allowing me to understand its strengths and areas for improvement.

Results

Login page

This screenshot shows the login interface of the application. It allows users to securely access their account using their unique credentials. Depending on the role (admin or regular user), different access rights and features are enabled upon login.

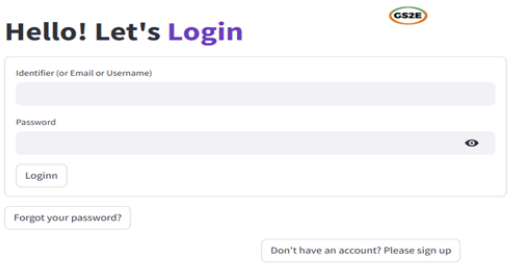


Figure 4: Login page

Water Consumption Analysis pages (By Period, City, and Client Type)

This section of the application provides comprehensive tools for analyzing water consumption. All users can access the Analysis by Period feature, which allows them to visualize their water usage over a selected timeframe, helping them understand their consumption patterns and identify irregularities. For administrators, additional advanced analytics are available: the Analysis by City page displays consumption trends across different cities, enabling regional comparisons and data-driven decisions, while the Analysis by Client Type presents segmented usage data based on user categories (domestic, commercial, industrial), offering insights to support targeted resource management strategies.

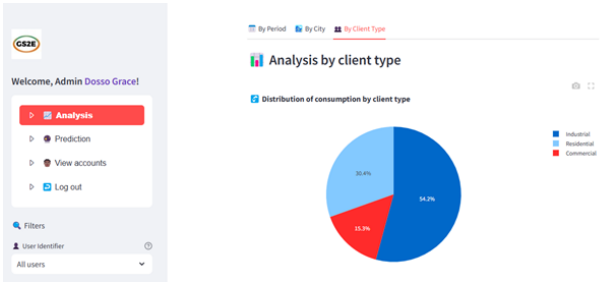


Figure 5: Analysis by client type page

Figure 6: Analysis by city page

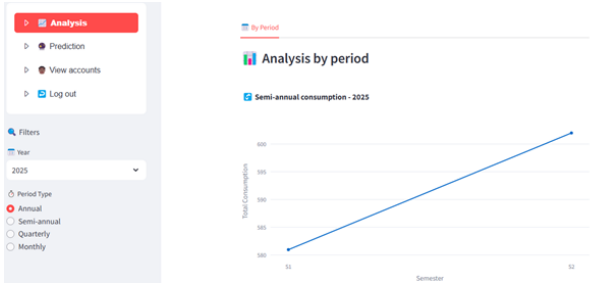


Figure 7: Analysis by period page

Prediction page

This interface displays predicted water consumption using the implemented machine learning model. It helps users and administrators anticipate future usage and prepare accordingly. The predictions are supported by data science techniques such as Ridge Regression.

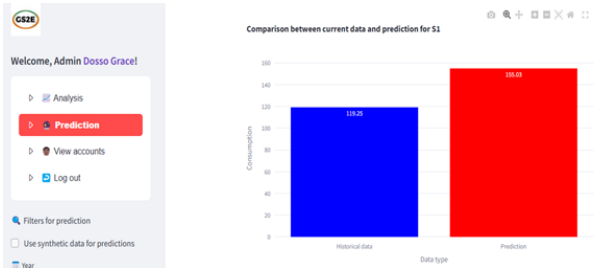


Figure 8: Prediction page

Profile page and Account management

This section allows each user to view their own profile details. Clients and agents can only access and view their personal information. However, administrators have extended privileges: they can view, update, and delete user accounts. They can also create agent and other administrators account. This functionality ensures that the system remains organized, secure, and easy to maintain.

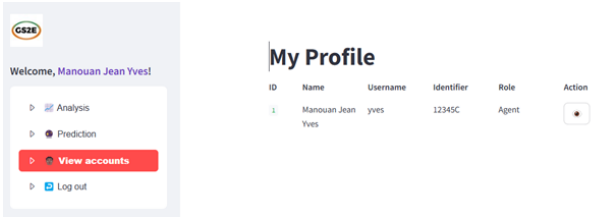


Figure 9: Profile page

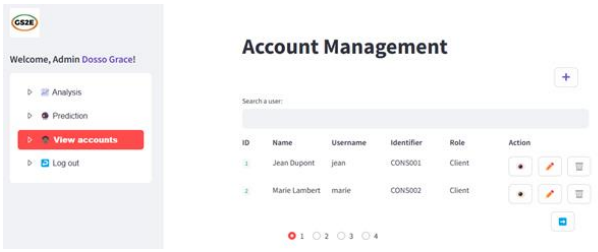


Figure 10: Account management page

Performance analysis of the model

This page presents the evaluation results of the predictive model. It includes key performance metrics such as Mean Absolute Error (MAE) and R^2 score. These metrics provide insight into the model's accuracy and reliability in predicting water consumption.

Figure 11: Performance analysis page

IV. CONCLUSION

This work presented the design and implementation of a comprehensive platform for analyzing and predicting water consumption patterns. The system integrates data collection and predictive modeling to provide actionable insights in a centralized, user-friendly interface. By combining efficient data management with accurate forecasting techniques, the platform enables timely detection of anomalies and supports informed decision-making for both clients and administrators.

The implemented solution demonstrates practical utility through its visualization tools, reporting features, and alert mechanisms, allowing stakeholders to monitor usage trends, detect irregularities, and optimize water consumption strategies. The predictive capabilities offer foresight into potential consumption spikes, helping to anticipate demand and improve resource allocation.

Overall, this system provides a scalable and reliable approach to understanding water consumption behaviors, emphasizing usability, performance, and practical relevance. Its architecture and features ensure that both individual users and organizational administrators can leverage data-driven insights for more effective water management, making it a valuable tool for planning, monitoring, and sustainability efforts.

REFERENCES

1. Adamowski, J. F., Chan, H. F., Prasher, S. O., Ozga-Zielinski, B., & Sliusarieva, A. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, and artificial neural network models for urban water demand forecasting in Montreal, Canada. Water Resources Management, 26(2), 433–454. https://doi.org/10.1007/s11269-011-9925-8
2. AlfredECN (2018), Maijet Review : Enhanced Tour Email Campaigns Ease, Mailjet: We Analyze This Email and SMS Marketing Tool
3. Alvisi, S., & Franchini, M. (2018). District-metered areas and advanced monitoring techniques for urban water networks. Urban Water Journal, 15(1), 12–22. https://doi.org/10.1080/1573062X.2017.1323741
4. Avikumar Talaviya (2024), Streamlit Tutorial: Building Web Apps with Code Examples, 12 October 2024, Streamlit Tutorial: Building Web Apps with Code Examples -
5. Carrasco-Jiménez, J., Muñoz, P., & Herrera, L. (2021). Unsupervised clustering for anomaly detection in household water consumption. Journal of Hydroinformatics, 23(4), 789–804. https://doi.org/10.2166/hydro.2021.055
6. Cominola, A., Giuliani, M., & Castelletti, A. (2015). Sensor-based architectures for disaggregated household water use monitoring. Environmental Modelling & Software, 69, 107–123. https://doi.org/10.1016/j.envsoft.2015.03.0030
7. Herrera, M., Torgo, L., Izquierdo, J., & Pérez-García, R. (2010). Predicting short-term urban water demand using artificial neural networks. Journal of Hydroinformatics, 12(3), 305–318. https://doi.org/10.2166/hydro.2010.059
8. Labura, D., Okoye, C., & Mensah, K. (2025). Robust anomaly detection in smart meter data using Isolation Forest and FFT-based filtering. Sustainable Cities and Society, 87, 104192. https://doi.org/10.1016/j.scs.2025.104192
9. Pavlou, A., Georgiou, A., & Tsiropoulos, I. (2024). Real-time non-intrusive household water end-use disaggregation: Model-based vs learning-based approaches. Water, 16(5), 987. https://doi.org/10.3390/w16050987
10. Zhou, H., Yang, Z., & Li, Y. (2000). Wavelet-neural network hybrid model for water consumption forecasting. Journal of Water Resources Planning and Management, 126(3), 152–160.