A predictive model for summer groundwater levels

Authors :
M. Ayoub Boufous
M. Abderrazak Derder
M. Arthur De Rouck
M. Ismail El Moufakir
M. Amine Maazizi
M. Dustin Pfundstein

December 1, 2024

Contents

1 Overview

1.1 Context

"More rapid evaporation and drying of soils worsen drought conditions. And yet, far too little is known about the true state of the world's freshwater resources. We cannot manage what we do not measure." This strong declaration from Celeste Saulo, Secretary-General of the World Meteorological Organization, underlines the critical challenge: we fail to measure or manage water resources upon which we depend.

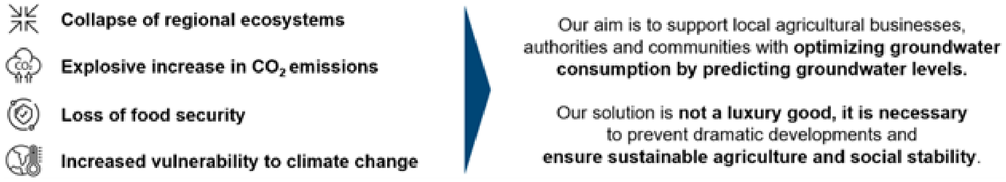
In the context of a world increasingly defined by climate change, water scarcity and insecurity have emerged among the most pressing issues of our time, driven by inefficient water use, overextraction, and intensifying droughts. At the heart of these crises is groundwater, a crucial resource that supplies 65% of drinking water and 25% of agricultural irrigation across the EU. Urbanization, population growth, and climate change are accelerating its depletion, creating a cascade of challenges for communities, ecosystems, and economies.

Groundwater mismanagement has far-reaching consequences that go beyond immediate water shortages. First, the local agriculture collapses and communities become dependent on imported food, implying CO2-intensive supply chains. Therefore, such imports increase carbon emissions and environmental costs exponentially. The reduction in local production further results in a rise in food prices and higher vulnerability to global market volatility, thus hitting the low-income households harder. Without effective groundwater management, regions experiencing droughts will have longer and more intense impacts, leaving their recovery uncertain. Furthermore, groundwater mismanagement results in a loss of resilience to environmental change.

Precise information about groundwater is not a luxury, it is an imperative for averting cascading crises and stabilizing agriculture in the interest of social and economic stability.

1.2 Our Project

To combat all these issues, we have designed an advanced groundwater level prediction model to meet the pressing need for groundwater level predictions. This model allows accurate forecasts that will give communities, farmers, and industries unparalleled opportunities to handle water efficiently and sustainably.



Collapse of regional ecosystems

Explosive increase in CO_2 emissions

Loss of food security

Increased vulnerability to climate change

Our aim is to support local agricultural businesses, authorities and communities with **optimizing groundwater consumption by predicting groundwater levels.**

Our solution is **not a luxury good, it is necessary** to prevent dramatic developments and **ensure sustainable agriculture and social stability**.

By equipping society with these tools, overextraction of the resource, destruction of sensitive ecosystems, and reduced availability of groundwater in the future will be averted. Our team of six students has devoted one weekend to design this innovative model. Please see the picture below for a detailed line-up of our team members with diverse backgrounds.

2 Business approach

Our solution offers significant economic and ecological benefits. It promotes sustainable water use, supports local food security, and reduces reliance on imports by stabilizing agriculture. It also helps users adapt to climate challenges, protects ecosystems by preventing groundwater overuse, and minimizes environmental impact. From a business perspective, it reduces costs by optimizing water usage and improving efficiency.

By linking water management and regional production to the real world, we raise our model to an even stronger social level. It becomes a tool to protect the regional economy, environment and social structures.

Imagine a regional farmer with no background in AI or coding, trying to navigate the complexities of water management amidst unpredictable climate conditions. Our solution is a simple, user-friendly Software-as-a-Service platform that can be easily installed on a smartphone. This tool provides real-time access to precise groundwater predictions, offering tailored guidance for every month. It tells the farmer exactly how much water they can use in their operations to maximize profits while maintaining sustainable practices. Beyond water usage, the platform also advises on the best crops to plant based on current and future water availability and which crops to avoid to minimize risk. Essentially, it functions like an online banking app—notifying the farmer how much "water budget" they have left to spend that month, empowering them to make informed, sustainable decisions with ease.

3 Scientific approach

The approach adopted in this work follows a rigorous and progressive methodology, structured into steps to ensure reliable and reproducible results. The main aspects of this approach are detailed below:

3.1 Problem formulation

The first step was to clearly define the problem: improving the prediction of a target variable using a partially incomplete dataset containing both numerical and categorical variables. The quality of the predictions was evaluated using appropriate metrics, with particular attention paid to model explainability.

3.2 Data preparation

A scientific strategy was implemented for data processing and preparation, following these steps:

3.2.1 Column Analysis

Variables were classified into three categories:

- **Important**: Variables strongly correlated with the target.
- **Less Important**: Complementary variables with moderate influence.
- **Not Important**: Variables deemed irrelevant and potentially excluded.

This classification was based on a thorough exploratory analysis and visualizations of relationships between columns. Continuous data was visualized, and it was noticed that a lot of data was skewed, so this data was isolated and transformed using a log scale.

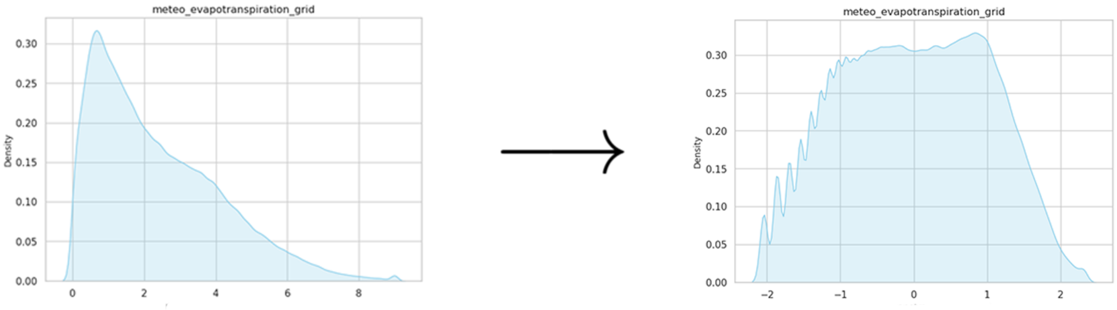


Figure 1: Log-transformation of a skewed feature

3.2.2 Feature Selection

Highly correlated features were identified and removed to reduce redundancy and mitigate multicollinearity, ensuring that models focused on the most informative variables. In addition, not important features and some less important features were removed to refine the dataset further.

3.2.3 Handling Missing Data

Three techniques were applied based on the importance of the variables:

- Predictive Regression: For critical columns where relationships needed to be preserved.
- Mean Imputation: For secondary columns to simplify processing without distorting global relationships.
- Missing data was categorized to better understand their patterns and address them accordingly.

3.2.4 Categorical Variable Encoding

Encoding was adapted according to the nature of the data:

- One-Hot Encoding for nominal variables.
- Label Encoding for ordinal variables.
- Dates were encoded starting from 1970 using Unix time encoding and transformed into seasonal features for better model interpretability.

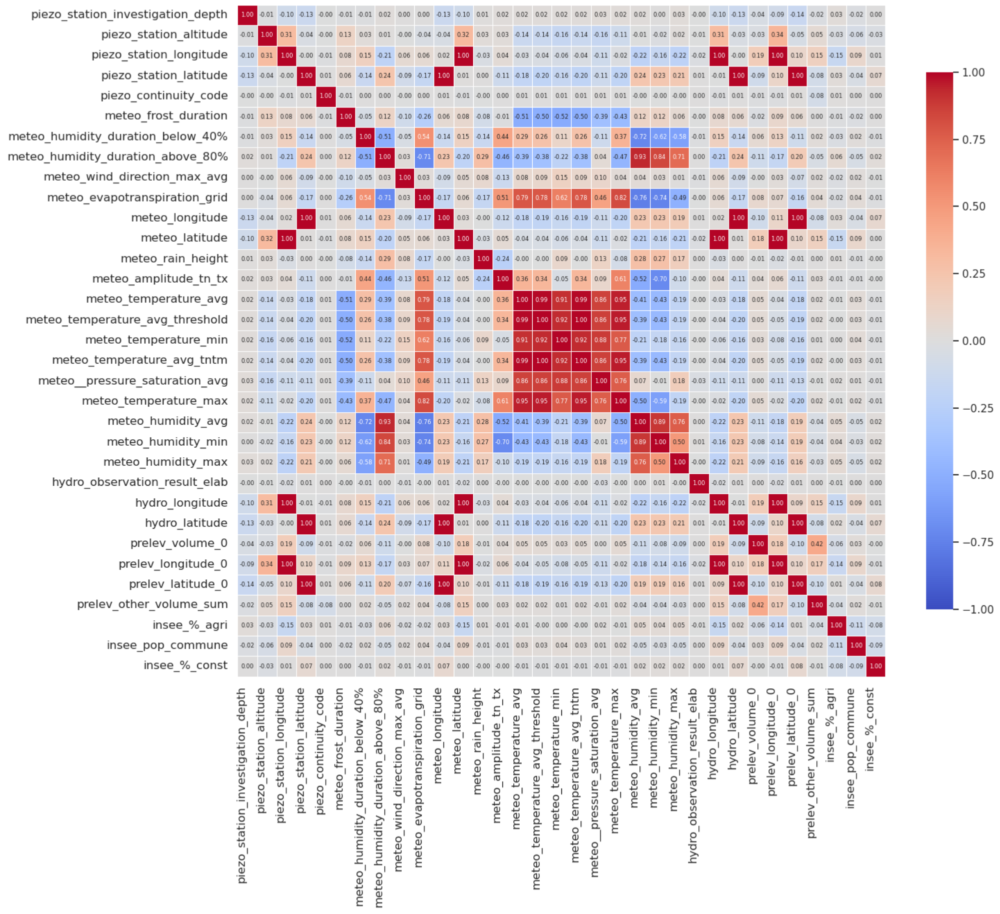


Figure 2: Correlation matrix

3.3 Modeling

Classification models were chosen as the primary approach. Three models were tested to maximize performance and address specific objectives. The details of the models used and the results obtained will be presented in the following section.

4 Results and Future Potential

Three different models were used :

- Random Forest Classifier: An ensemble method was chosen for its robustness.
  Result: A score of 50%, with good interpretability of decisions.

- Neural Networks: This was included for experimentation, but ultimately dropped due to suboptimal performance.
  Result: A score of 36%, with limited optimization.

- XGBoost: Selected as a comparison to Random Forest, it proved to be more performant.
  Result: A score of 41%, representing a strong balance between performance and flexibility.

The model could be more accurate, but it can still give an idea of the groundwater level in summer.