**I. Overview**

**Problem Statement**

**Groundwater** is a vital resource, providing essential support for agriculture, industry, and domestic use. Globally, it serves over 2.5 billion people and accounts for 40% of irrigation needs. As a **start-up** **technology company** specializing in **water shortage risks and insurance risk assessment service**, we recognize the critical importance of groundwater resources in sustaining agriculture, industry, and domestic needs. In France, groundwater is a critical resource, supplying approximately 65% of drinking water and 25% of agricultural irrigation needs. However, the country faces significant challenges with declining groundwater levels due to excessive withdrawals and the impacts of climate change. The summer of 2022 was France's hottest and driest on record, leading to depleted groundwater reserves and necessitating water restrictions in various regions[[1]](#footnote-0). By 2050, France is projected to have 30–40% less water available, intensifying the urgency for sustainable water management practices[[2]](#footnote-1). This situation poses substantial risks, including agricultural stress affecting crop production, economic challenges for water-dependent industries, ecosystem degradation impacting biodiversity, and increased vulnerabilities for communities reliant on consistent water access.

In France, the agricultural insurance market is predominantly composed of standardised products, such as fire and multi-risk insurance, with premiums reaching €936 million in 2023. These standard policies typically cover general risks like livestock mortality, fire, and hail. However, they often fall short in addressing the specific needs arising from regional climatic variations and diverse crop types. For instance, France's diverse climatic conditions lead to region-specific agricultural risks. Southern regions are more susceptible to droughts, while northern areas may experience flooding. Standard insurance products may not adequately cover these unique regional risks, leaving farmers vulnerable to losses not encompassed by general policies. Also, although the current French agricultural insurance system does cover natural disaster risks such as droughts and floods, the participation rate is still low, mainly due to the high premium costs and difficulty in widespread acceptance. Our machine learning-based pricing model can effectively solve this problem. By using accurate groundwater level predictions and regional risk assessments, our model can dynamically adjust insurance prices according to risk levels. Thus, we want to fill the gap in the market by providing financial institutions with **risk assessments with water shortage risk factors included**, helping clients effectively mitigate the financial impacts of water scarcity.

**Presentation of Project**

As a **start-up technology company focusing on insurance risk assessment service**, we recognize our responsibility to assist local companies, particularly agricultural enterprises, in mitigating downside risks associated with water shortages. By integrating advanced machine learning models into our insurance pricing strategies, we can provide more accurate risk assessments using **the pricing model embedded with machine learning techniques**, enabling the financial institution to **offer tailored insurance solutions that reflect the specific challenges faced by each client**. This enables the development of insurance premiums that more closely align with the actual risk profiles of different agricultural clients, ensuring fair pricing and adequate coverage. Through these efforts, we aim to support the long-term viability and success of local enterprises in the face of evolving environmental risks.

**Team Structure**

We are a team of five dedicated insurance analysts specializing in assessing regional water shortage risks and delivering tailored insurance solutions to agricultural enterprises. We are Wilen, Sid Ahmed and Yanis from University Sorbonne Paris Nord, and Siyuan and Yiying from HEC Paris. Collectively, we bring a diverse skill set encompassing data science, machine learning, environmental science, and insurance pricing modelling. This interdisciplinary expertise ensures that our solutions are both technically robust and practically applicable. We are committed to mitigating summer water shortages in France by equipping the government with precise, region-specific insights to optimize water resource management.

**II. Business Approach**

**Value of Project**

Our project aims to provide risk assessments by **developing a machine learning-based groundwater level prediction system for dynamically adjusting insurance prices according to risk levels**. Using an extensive dataset that combines piezometric data, meteorological variables, hydrological and abstraction metrics, and socio-economic indicators, our solution predicts groundwater levels during summer months. This solution focuses on predicting groundwater levels during summer with high accuracy.

These predictive insights enable us to **refine our risk assessment processes**, allowing for the calibration of insurance premiums that accurately reflect the specific water shortage risks faced by agricultural clients in different regions. Integrating this predictive capability into our pricing models ensures that premiums are commensurate with actual risk levels, promoting fairness and financial sustainability. Moreover, this data-driven approach enhances our ability to offer tailored insurance solutions, thereby strengthening our competitive position in the market and providing greater value to our clients. By leveraging advanced machine learning techniques, our solution transcends the limitations of traditional methods, offering a forward-looking perspective that aligns with the complex realities of climate change and water resource management.

This dynamic, risk-driven pricing mechanism not only helps to increase insurance participation rates, but also brings a more stable profit model to insurance companies while meeting the actual needs of farmers. With greater precision and flexibility, our model can provide farmers in different regions and crop types with tailored insurance solutions, injecting new vitality into the French agricultural insurance market.

**Real-World Integration**

The predictive model can be seamlessly incorporated into existing underwriting and pricing systems. By regularly updating the model with current data, insurers can dynamically adjust premiums and policy terms to reflect the latest risk assessments.

where P is the premium of the insurance;

E(L) is the expected loss;

Water\_Shortage\_Risk factor is the prediction result from the machine learning model;

Pt-1 is the premium of the insurance for the previous term;

L\_expenses is the Loss Adjustment Expenses;

R\_expenses is the Underwriting and Administrative Expenses;

M\_profit is the Margin for Profit and Contingencies

By regularly updating the model with current data, insurers can identify regions predicted to experience low groundwater levels during summer months, allowing for the development of tailored insurance products that address specific water shortage risks. For instance, in areas where the primary crop is grapes, insurers can offer policies that cover potential yield reductions due to drought conditions, with payouts linked to factors such as crop quantity and market prices. Additionally, insurers can design products that provide coverage for extreme climate events, including droughts and floods, with parameters based on rainfall and temperature data. This integration facilitates real-time decision-making, enhances the insurer's ability to respond promptly to changing environmental conditions, and offers clients insurance solutions that are closely aligned with their specific risk profiles and needs.

**Novelty**

Traditional insurance pricing often relies on historical data and generalized risk assessments, which may not account for the dynamic nature of groundwater levels and climate variability. Our solution leverages **advanced machine learning techniques** to provide forward-looking risk evaluations, capturing complex, non-linear relationships between environmental factors and groundwater levels. This proactive approach enables insurers to offer more tailored and responsive products, setting them apart in a competitive market.

By integrating our predictive model into existing underwriting and pricing systems, insurers can dynamically adjust premiums and policy terms based on the latest risk assessments. Regular updates with current data allow for real-time decision-making, enhancing the insurer's ability to respond promptly to changing environmental conditions. This integration facilitates the development of insurance products that are closely aligned with clients' specific risk profiles and needs.

**III. Scientific Approach**

**Data Pre-processing**

To ensure robust predictions, the data pre-processing involved several critical steps. We identified and excluded rows with invalid or extreme values in key features such as piezo\_station\_altitude, where values below -500 were considered invalid and removed to maintain data quality. Missing values in crucial features such as piezo\_station\_investigation\_depth were imputed using the median, to maintain integrity and avoid bias. We converted columns containing date information, such as piezo\_station\_update\_date and piezo\_measurement\_date, into standard datetime formats and removed invalid dates to ensure consistency in temporal data. Furthermore, we retained features, like meteo\_rain\_height, meteo\_temperature\_avg, and socio-economic variables based on their relevance to the domain. Additionally, categorical target labels such as piezo\_groundwater\_level\_category were encoded using hierarchical mappings to accurately represent their relative severity.

**Machine Learning Model**

We chose the Random Forest Classifier for our project due to its exceptional ability to handle high-dimensional and complex datasets characteristic of environmental and socio-economic variables influencing groundwater levels. Random Forests are robust to outliers and noise, thanks to their ensemble approach of aggregating multiple decision trees, which dilutes the impact of anomalies and prevents overfitting through methods like bootstrapping and feature randomness. They require no assumptions about data distribution, making them ideal for modeling the nonlinear and intricate relationships inherent in our data. The algorithm’s feature importance metrics enhance interpretability by highlighting key factors affecting predictions, which is crucial for understanding and decision-making. Additionally, Random Forests scale efficiently with large datasets and support parallel processing, ensuring computational efficiency. They also handle class imbalance effectively through class weighting and balanced subsampling. Compared to other models like Support Vector Machines or neural networks, Random Forests offer a superior balance of accuracy, interpretability, and computational efficiency, while being less sensitive to parameter tuning and overfitting. Their adaptability to new data allows for incremental learning, ensuring the model remains relevant as environmental patterns evolve. Lastly, strong community support and mature implementations in libraries like scikit-learn facilitate ease of use and reliable performance, making Random Forests the optimal choice for our predictive modeling needs.

**Scalability of the Model**

The Random Forest model is designed with scalability in mind to handle both increasing data volume and evolving environmental patterns. By leveraging parallel processing capabilities, the model efficiently trains and predicts on large datasets, making it suitable for real-time applications. Additionally, the use of cloud-based infrastructures such as AWS or Google Cloud enables the seamless integration of new data sources, including piezometric measurements and meteorological updates, ensuring that the model remains up-to-date. The modular architecture of the model allows for easy adaptation to new regions or industries, such as municipal water management or industrial insurance. Regular re-training pipelines ensure the model evolves alongside changing environmental conditions, while automated hyperparameter tuning optimizes performance across diverse datasets.

**Social Impact**

Our approach prioritizes the use of machine learning to analyze extensive datasets on groundwater levels, weather patterns, and agricultural demands. By focusing on data, we minimize the need for extensive physical resources that are traditionally used in risk assessment processes, such as on-site evaluations and manual data collection efforts. The predictive nature of our model means we can anticipate water shortage issues before they occur, thereby reducing the reactive need for intensive resource use in crisis situations.

By incorporating groundwater level predictions into our insurance pricing model, we inherently encourage farmers to adopt more sustainable water usage practices. Our tailored insurance products make it economically viable for farmers to invest in water-efficient technologies and crop varieties that are more resistant to drought conditions. Farmers who implement such practices may benefit from lower insurance premiums, as their risk levels decrease with reduced groundwater dependency. This not only helps in conserving water but also in maintaining ecological balance.

In addition, the digital nature of our services means that we heavily rely on cloud-based technologies and remote operations, reducing the need for physical office spaces, travel, and paper-based processes. This significantly lowers our carbon footprint and sets a standard for how technology can be leveraged to deliver essential services without the heavy environmental cost typically associated with traditional insurance and risk assessment practices.

**IV. Results and Future Potential**

**Results**

The Random Forest model achieved an accuracy of **30.76%**, reflecting the complexity of predicting groundwater levels using the available dataset. This score highlights both the challenges inherent in modeling such multi-class classification problems and the potential limitations in feature representation. The model’s effectiveness is constrained by the quality and variety of the input features, as some critical environmental or socio-economic factors may not be fully captured in the dataset. Additionally, the relatively high noise in certain variables and imbalances in the target categories may have affected the overall performance.

Despite these challenges, the Random Forest model remains robust due to its ability to manage heterogeneous data and its inherent interpretability. The feature importance analysis revealed that variables like piezo\_station\_altitude and meteo\_rain\_height contributed significantly to predictions, indicating that these factors are strongly correlated with groundwater levels. However, improvements in feature engineering—such as incorporating lagged variables, advanced climatic indices, or domain-specific transformations—could enhance the model’s predictive power. This score serves as a baseline for iterative refinement of the model and feature set to better capture the underlying complexities of groundwater prediction.

**Future Potentials**

**Benefits for Potential Clients**

The insurance products corporated with our unique pricing model will bring a wide range of benefits to buyers (the agricultural producers) and sellers (the insurance companies).

1. **Lower the premium threshold for high-risk areas**: For areas that have been at medium or low risk for a long time, the premium can be reduced, which will enhance the attractiveness of insurance and encourage more farmers to participate.
2. **Accurately cover high-risk periods and areas**: In specific time periods (such as the peak of summer drought) or specific high-risk areas, the premium will reflect the actual risk level to ensure a balance of interests between insurance companies and policyholders.
3. **Enhance fairness and flexibility**: Compared with the traditional fixed premium structure, our model is priced according to actual risks, which increases farmers' trust in insurance and attracts more farmers to join the insurance plan.

**Transferrable for Other Industries**

The integration of machine learning-based groundwater level prediction models into insurance pricing holds significant potential beyond the agricultural sector. Industries such as manufacturing, energy production, and municipal water supply, which heavily depend on consistent groundwater availability, can benefit from this advanced risk assessment approach. By accurately forecasting groundwater levels, insurers can develop tailored policies that reflect the specific water-related risks of these industries, leading to more precise premium pricing and enhanced financial resilience against water scarcity.

1. <https://www.france24.com/en/tv-shows/down-to-earth/20230615-france-s-growing-water-crisis> [↑](#footnote-ref-0)
2. <https://www.dw.com/en/water-crisis-in-france-drought-hit-regions-rely-on-tankers/a-65649291> [↑](#footnote-ref-1)