Predictive Modelling of Water Body Dynamics

Abstract: *This study aims to design and implement a robust predictive model for forecasting water levels in various types of water bodies, including springs, lakes, rivers, and aquifers. The model integrates data from multiple independent datasets, accounting for seasonal variations and environmental factors such as rainfall and temperature. By using machine learning techniques, including Linear Regression and Random Forest Regression, the study evaluates model performance based on standard metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The model's flexibility allows for adaptation to different waterbody types, and scenario analysis is used to simulate extreme events such as droughts and excessive rainfall. Results show that while both models perform well, Random Forest Regression consistently outperforms Linear Regression, providing better accuracy and robustness for water level prediction. This adaptable framework offers valuable insights for managing water resources across diverse waterbodies.*

INTRODUCTION AND BACKGROUND:

The topic of this investigation centers around predicting the water levels in various types of water bodies, including lakes, rivers, aquifers, and springs. The goal is to develop a machine learning model capable of accurately forecasting the water levels for each time interval (daily, monthly, etc.) throughout the year, under different seasonal and environmental conditions.

The Acea Group is one of the largest Italian multi-utility operators, managing water, electricity, and environmental services across the country. One of their core responsibilities is the sustainable management of water resources, especially the prediction and control of water levels in diverse water bodies, such as springs, lakes, rivers, and aquifers. These water bodies are essential sources for both daily consumption and the broader ecosystem, making it critical to maintain their health and balance. However, water bodies behave differently depending on their type, geographic location, and the environmental conditions surrounding them.

Goal of the study: The goal of this study is to design and implement a robust predictive model capable of forecasting the water levels in various types of water bodies across different seasons. This model will be informed by data from multiple independent datasets corresponding to different water body types, including springs, lakes, rivers, and aquifers. This study aims to:

1. Predict Water Levels: Develop a model that can accurately forecast the water levels for each type of water body, with consideration for seasonal and environmental factors.
2. Understand Environmental Impact: Investigate how factors such as rainfall and temperature influence water body levels and incorporate temporal delays into the model to account for these effects.
3. Evaluate Model Performance: Develop Linear Regression and Random Forest Regression for this data analysis. Assess the model's accuracy using standard evaluation metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2 score, Kurtosis and Skewness.
4. Simulate Extreme Scenarios: Use scenario analysis techniques to test the model's ability to handle extreme events such as droughts or excessive rainfall, and explore the model's performance under these challenging conditions.

Dataset Dictionary: The dataset for this competition[1] consists of nine independent datasets representing different types of waterbodies, including water springs, lakes, rivers, and aquifers. Each dataset contains unique features reflecting the characteristics of its corresponding waterbody, such as rainfall, hydrometry, depth, and temperature. The Acea Group manages four types of waterbodies: three datasets for water springs, one for lakes, one for rivers, and four for aquifers. Source[1] shows the clear description for the outcome variables in dataset, which features need to be predicted (I have also uploaded a data description xlsx file for this).

For this analysis, I'll show results for aquifers (because of page constraint), and the model pipeline can be applied to all waterbody types. Additional data analysis and results for other waterbodies are presented in the accompanying Jupyter notebook.

METHODS:

Data Exploration: Below is the correlation plot for one of the Aquifer named Auser. From the correlation plots for each waterbody it can be justified that rainfall and temperature are one of the important features to predict depth of waterbody.

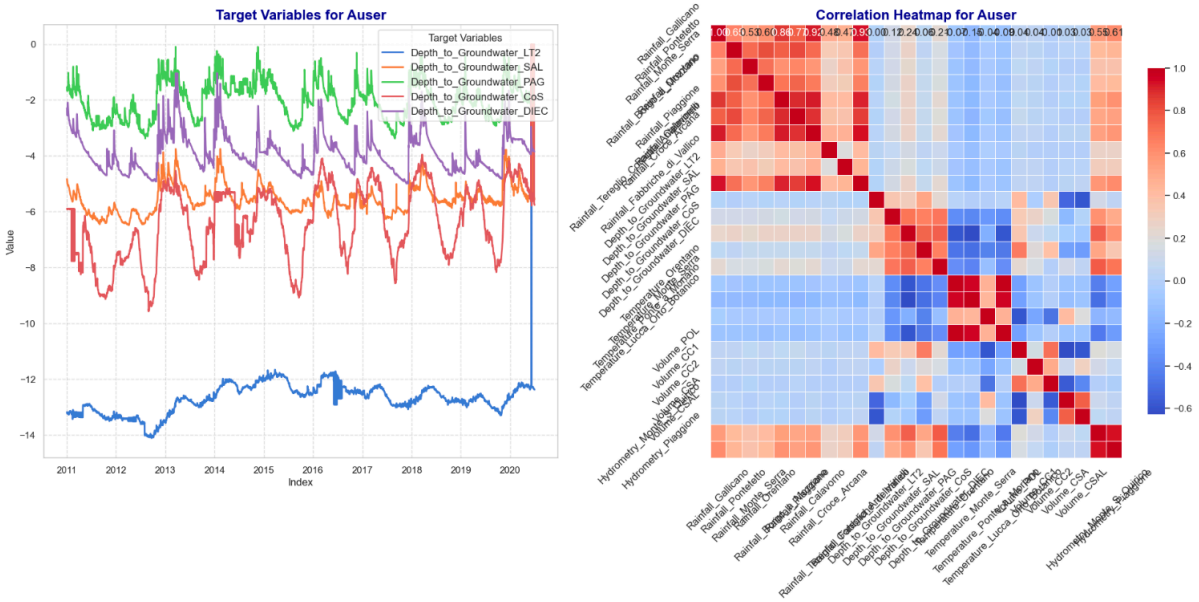

Figure 1. Target Variables and Correlation Heatmap for Auser

Data Preprocessing and Exploratory Data Analysis (EDA): In this study, we implement a structured pipeline for preparing and transforming the raw datasets into a suitable format for machine learning models. The pipeline consists of several key classes designed to handle missing data, engineer new features, perform exploratory data analysis, and split the data.

Data Preparation: The **DataPreparation** class is responsible for cleaning and preprocessing the raw data before feature engineering begins. It handles several crucial tasks, such as removing null values and setting up the data for further analysis. For imputing missing values, **KNNImputer** is employed. The KNN (K-Nearest Neighbors) algorithm imputes missing values based on the values of nearby data points, ensuring that the imputation reflects the structure of the data. The **distance** parameter is used as the weighting mechanism, meaning that data points closer to the missing values have more influence on the imputation than those further away.

Feature Engineering: Once the data is preprocessed, the **FeatureEngineering** class is used to create new features, resample the data, and apply transformations like shifting the target variable for time-based forecasting. The class takes the target variable, a shift period (default value of 1 day), and a resampling frequency (default of 'W' for weekly) as input parameters. As resampled frequency I used 'W' (weekly frequency), ‘M’ (month end frequency) and ‘SM' (semi-month end frequency (15th and end of month)). As shift period values 1, 2 and 3 are used. For understanding suppose if parameters were ‘M’ and ‘3’ the forecast is monthly with a predicted forecast of three months and so on.

Train-Test Split: The **TrainTestSplit** class is used to split the dataset into training and testing sets, ensuring that the model is trained on a portion of the data and evaluated on unseen data. The class takes a split ratio as input, which determines the proportion of data allocated to training and testing. The ratio I used is 0.8, meaning that 80% of the data is used for training, and 20% for testing.

Model Training: The **ModelTraining** class supports two types of models for water level prediction: **Linear Regression** and **RandomForestRegressor**. These models are chosen for their simplicity (Linear Regression) and flexibility in capturing non-linear relationships (RandomForestRegressor). Before feeding the data into the models, **RobustScaler** is applied as a preprocessing step. The RobustScaler is particularly useful for datasets with outliers, as it scales the features based on the interquartile range (IQR), making it more robust to extreme values than standard scaling methods. To ensure that the models are optimised for the best performance, **RandomizedSearchCV** is employed to tune the hyperparameters of the models. RandomizedSearchCV performs random sampling over the hyperparameter space and selects the best combination of parameters based on cross-validation performance.

For the selected best estimator, the following outputs are computed:

1. Predictions: The model makes predictions on the test set, which represent the predicted water levels for the given test data. These predictions are stored in a dataframe for later analysis.
2. Residuals: The residuals are calculated by subtracting the predicted values from the actual values. The residuals provide insight into how well the model fits the data, as larger residuals can indicate model misfit or unaccounted-for patterns in the data.
3. Feature Importances: For models like **RandomForestRegressor**, feature importances are extracted to understand which input variables are most influential in predicting the target variable.

RESULTS AND TESTS:

Model results (Linear Regression): Below are the results for the Linear Regression model applied to the four different aquifer waterbodies. Each waterbody was evaluated with different configurations for frequency and shift period, resulting in varying model performances measured by **Mean Squared Error (MSE)** and **Mean Absolute Error (MAE)**. The **Auser** aquifer achieved the best performance, with both low MSE and MAE, suggesting that a **semi-monthly frequency** with a **1-day shift period** was optimal for this dataset. The **Doganella** aquifer model, on the other hand, exhibited significantly higher error values, which may suggest that monthly data with a 1-day shift period is not sufficient to capture the complex dynamics of this waterbody (one of the limitation).

DATA	FREQUENCY	SHIFT PERIOD	MEAN SQUARED ERROR	MEAN ABSOLUTE ERROR
Auser	SM	1	0.205833	0.323666
Donganella	M	1	14.548504	2.736526
Luco	SM	1	0.976106	0.709982
Petrignano	W	3	4.756940	1.913417

Table 1. Best results (shift period and frequency) for each of the waterbody (for aquifer)

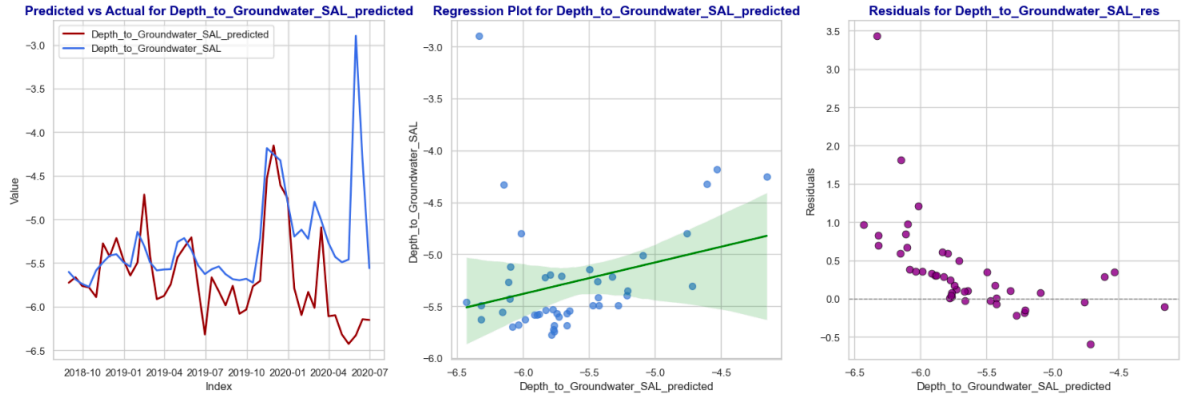

Figure 2. Prediction plot, Regression plot and Residual plot for results predicted (one of the predicted variable, others are in notebook)

Statistics (Linear Regression): The model evaluation includes key statistical metrics, including **Skewness**, **R² Score**, and **Kurtosis**, which help assess the distribution and performance of the predicted water levels. A value of R^2 score > 6.0 signifies a good fit. Also probability plot (for one of the predicted feature) shows a good fit.

Target	Skewness	R2_Score	Kurtosis
Depth_to_Groundwater_LT2_res	2.072372	0.747627	4.579202
Depth_to_Groundwater_SAL_res	2.816784	0.776099	10.829888
Depth_to_Groundwater_PAG_res	0.404571	0.879276	-0.722000
Depth_to_Groundwater_CoS_res	0.222787	0.711879	-0.989430
Depth_to_Groundwater_DIEC_res	0.312102	0.843491	-0.804389

Table 2. Model Evaluation using Statistical metrics (for Aquifers)

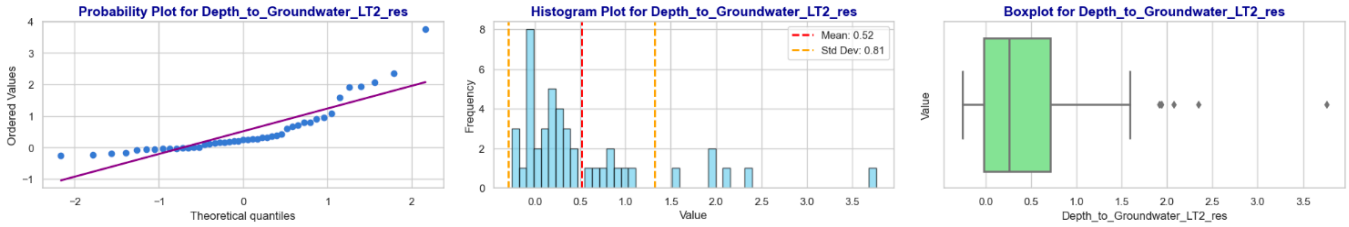

Figure 3. Probability plot, Histogram plot and Box plot for results predicted

Model Results (Random Forest Regression): Below are the results for the Random forest Regression model applied to the four different aquifer waterbodies. Each waterbody was evaluated with different configurations for frequency and shift period, resulting in varying model performances measured by **Mean Squared Error (MSE)** and **Mean Absolute Error (MAE)**.

DATA	FREQUENCY	SHIFT PERIOD	MEAN SQUARED ERROR	MEAN ABSOLUTE ERROR
Auser	SM	1	0.205833	0.323666
Donganella	M	1	14.548504	2.736526
Luco	SM	1	0.976106	0.709982
Petrignano	W	3	4.756940	1.913417

Table 3. Best results (shift period and frequency) for each of the waterbody (for aquifer)

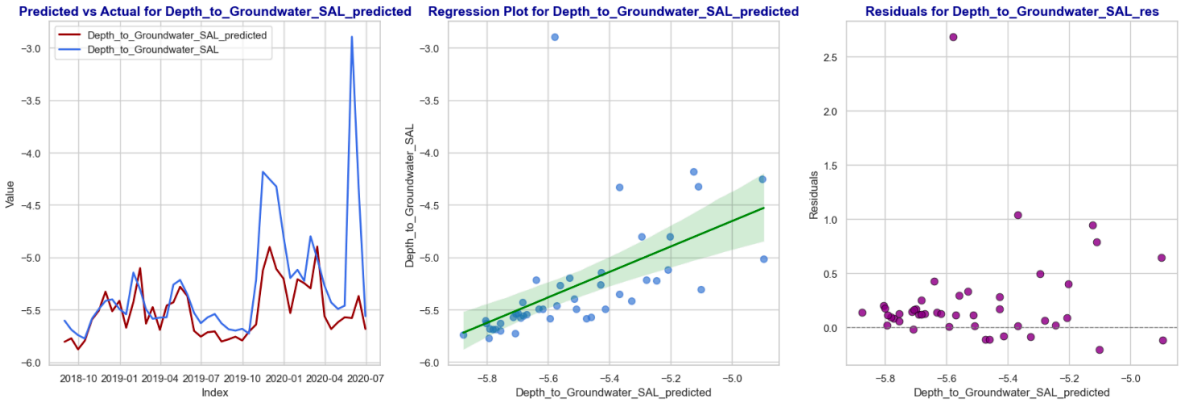

Figure 4. Prediction plot, Regression plot and Residual plot for results predicted (one of the predicted variable, others are in notebook)

Statistics (Random Forest Regression):

The model evaluation includes key statistical metrics, including **Skewness**, **R² Score**, and **Kurtosis**, which help assess the distribution and performance of the predicted water levels. . A value of R^2 score > 6.0 signifies a good fit. Also probability plot (for one of the predicted feature) shows a good fit.

Target	Skewness	R2_Score	Kurtosis
Depth_to_Groundwater_LT2_res	3.218844	0.663920	14.714665
Depth_to_Groundwater_SAL_res	3.773744	0.646540	17.041820
Depth_to_Groundwater_PAG_res	0.488829	0.611994	-0.277046
Depth_to_Groundwater_CoS_res	-0.154948	0.607474	-0.477583
Depth_to_Groundwater_DIEC_res	0.400145	0.602280	-0.289838

Table 4. Model Evaluation using Statistical metrics (for Aquifers)

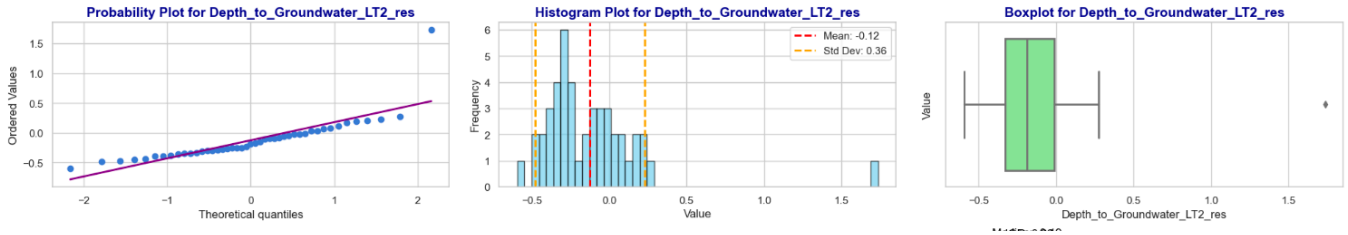

Figure 5. Probability plot, Histogram plot and Box plot for results predicted

CONCLUSION:

The model developed in this study offers a highly flexible and adaptable framework for predicting water availability across various types of waterbodies. Its modular structure allows for easy customization, making it applicable to different waterbody types by simply adjusting one or more components of the pipeline. This ensures that the model can be effectively applied to water springs, lakes, rivers, and aquifers. Throughout the evaluation, it was observed that both **Linear Regression** and **Random Forest Regression** models performed well across most waterbody types. However, **Random Forest Regression** consistently outperformed **Linear Regression**, delivering superior results with higher **R² scores** and better **kurtosis values**. These metrics indicate that Random Forest Regression provides a better fit to the data, with a reduced number of outliers, making it more robust and reliable for water level forecasting.

REFERENCES:

[1] Antimo Musone, Aredhel Bergström, Federico, Luisa Marotta, Maggie, and Maurizio Lucchesi, Acea Smart Water Analytics. https://kaggle.com/competitions/acea-water-prediction, 2020. Kaggle.

[2] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178

[3] Biau Gérard, Analysis of random forests model, *J. Mach. Learn. Res. 13*, null (3/1/2012), 1063–1095.

[4] The Acea Group website Link: https://www.gruppo.acea.it/en