Machine learning Prediction of River Flows in Southern Israel using Meteorological Data

Daniel Azenkot
[1]*Ben Gurion University, faculty on software and information engineering*

ABSTRACT

Southern Israel faces a significant challenge in managing water resources due to the absence of historical hydraulic flow data beyond the past 20 years. Although meteorological records spanning several decades are available, the lack of information on past river flows directly affects the region's aquifer systems, which are essential for water sustainability. A previous study conducted a year ago attempted to develop a machine learning model to predict flooding in certain rivers. However, it faced several issues that are addressed in the current study. This study continues to use 10-minute meteorological data from the stations at Beer Sheva, Lahav, Dorot, and Sede Boqer, along with hydrological data from stations along the Besor River. We aimed to create an improved model using proper methodologies. Specifically, we employed tree-based models such as XGBoost, Random Forest, and LightGBM, and compared their performance across multiple modeling approaches.

Key words: Machine Learning - Trees Models - Imbalanced Data - Flood Prediction

1 INTRODUCTION

Water resource management is a critical concern in arid regions, and Southern Israel is no exception. While there is extensive meteorological data spanning several decades that documents rainfall, temperature, pressure, and humidity patterns, there is a significant gap in historical hydraulic flow data. This data, which characterizes river flows over time, is only available and organized for the past two decades. This limitation hampers our understanding of the long-term dynamics of river flows, which are essential for effective water resource management in the region.

The lack of historical hydraulic data presents a pressing challenge for water resource managers, scientists, and policymakers. This urgency arises from the direct connection between river flows and aquifer water levels, which is central to understanding and maintaining water resources in the region. Typically, river flows significantly influence aquifer water levels, thereby affecting the overall water resource scenario. Consequently, the absence of historical flow data hinders our ability to gain a comprehensive understanding of the changes these aquifers have undergone since humans began extracting water from them for irrigation and daily use.

This research project addresses this critical knowledge gap by developing a machine learning model capable of predicting past river flows using high-resolution meteorological data with a 10-minute temporal resolution, building on previous work. In the earlier research, a 30-minute window was used to create new features for relative humidity, temperature, and rainfall at the rain station. Expanding on this, we incorporated additional windows of 60, 120, 300, 600, and 1200 minutes to create a more comprehensive dataset, aiming to better understand the predictors of flooding.

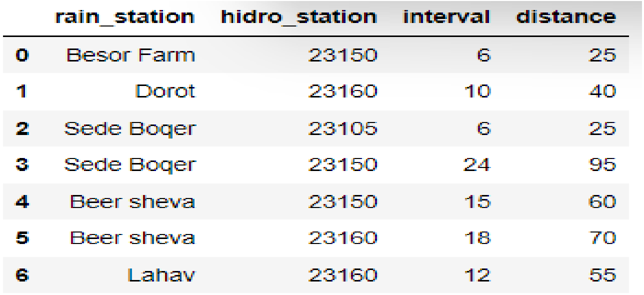
We utilize a range of tree-based machine learning techniques: XGBoost, Random Forest, and LightGBM, to develop models for simulating river flows. Subsequently, we employ SHAP analysis to identify the most influential factors within these models, revealing the key drivers behind historical flow patterns.

This research represents a significant advancement in understanding past river flow dynamics in Southern Israel and, consequently, historical changes in aquifer levels. The developed machine learning models have the potential to inform water resource management and planning in the region, aiding in the sustainable utilization of this vital natural resource.

Our research centers on seven distinct linkages between meteorological and hydraulic stations, with each link being customized to one of the three types of tree-based machine learning models mentioned.

2 METHODS

Machine Learning (ML) has emerged as a pivotal field in computer science and artificial intelligence, fundamentally transforming our capacity to glean insights from data. Ensemble methods such as **XGboost** (XGB), **LightGBM** (LGBM) and **Random For- est** (RF) have been proven in practice. RF is a bagging algorithm, consisted of decision trees that offers robustness to overfitting and excels in handling high dimensional data. On the other hand, XGB and LGBM are types of boosting models. XGB is an efficient and scalable gradient boosting framework, to construct predictive models for our dataset. XGB iteratively refines multiple weak learners to optimize predictive accuracy, utilizing advanced regularization techniques to prevent overfitting. LGBM, a high-performance gradient boosting framework, to develop predictive models for our dataset. LGBM employs a novel tree-based algorithm, focusing on leaf-wise growth and histogram-based techniques for faster training and improved accuracy. To interpret and comprehend the intricate models



	rain_station	hidro_station	interval	distance
0	Besor Farm	23150	6	25
1	Dorot	23160	10	40
2	Sede Boqer	23105	6	25
3	Sede Boqer	23150	24	95
4	Beer sheva	23150	15	60
5	Beer sheva	23160	18	70
6	Lahav	23160	12	55

Figure 1. This is the Link dataset. The interval column represents the assumed time, in hours, for rainfall at the meteorological station to reach the hydraulic station. The distance column indicates the distance between the stations in kilometers.

produced by these algorithms, the **Shapley Additive Explanations** (SHAP) framework has become indispensable. SHAP values offer a holistic method for model explainability, enabling practitioners and researchers to discern the impacts of individual features on model predictions. This enhances transparency and instills trust in ML models.

3 DATASETS

- **IMS** data pertains to information sourced from the Israeli Meteorological Service (IMS). Using an API connection, we obtained 130 CSV files from meteorological stations. For each station, we downloaded 21/22 CSV files, covering each year from 2000 to 2021. The IMS dataset comprises relative humidity, temperature, and rainfall measurements recorded at 10-minute intervals, at each meteorological station.
- **The Flows** data originates from the water authority in Israel. We acquired three hydrographs: the first spans from 2000 to 2009, the second from 2010 to 2019, and the third from 2020 to 2022. There were two classes: flood, with is denoted by 1, and normal flow which is denoted by 0. The dataset had missing values, we decided to remove them from the dataset. Before the removal of the missing values, there were 3971453 samples, and afterwards there were 3725716 across all the stations.
- **The Link** data pertains to a dataset that we pre processed from the Dead Sea and Arava Science Center. It aids us in establishing a time window and creating connections between meteorological stations and hydraulic stations. For each meteorological station, there exists a designated time frame during which the rainfall is presumed to reach the hydraulic station. With seven unique connections established, we developed machine learning models for each link to classify the occurrence of floods at the hydraulic station. See Figure 1 for the data set.

Using these three datasets, we compile a comprehensive dataset by linking the IMS data with the Flows dataset. This linkage is established for each instance where a meteorological station and a hydraulic station are matched in the Link dataset.

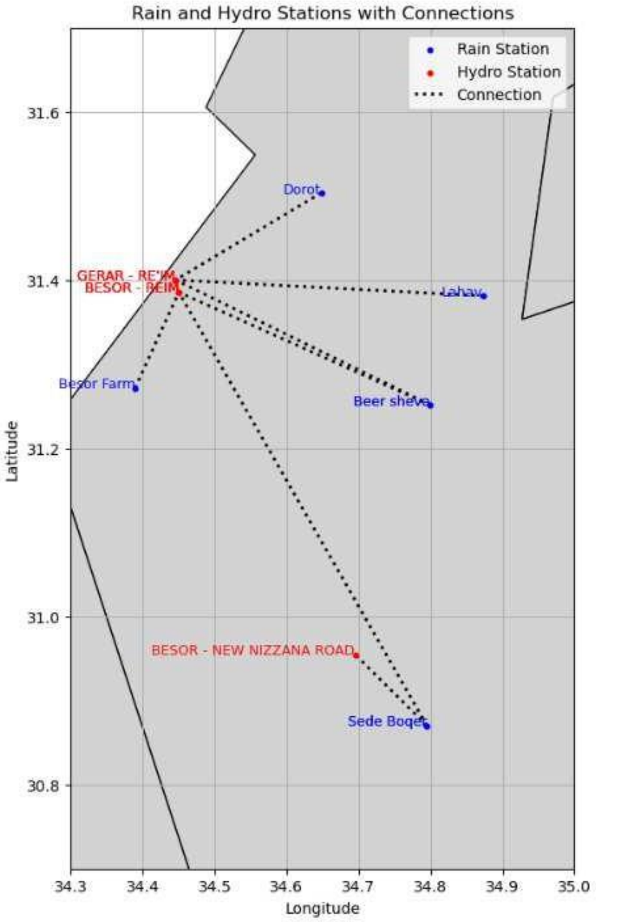
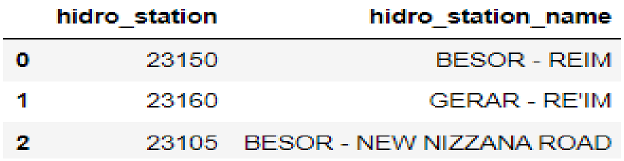


Figure 2. A map with the connections of the different meteorological and hydraulic stations



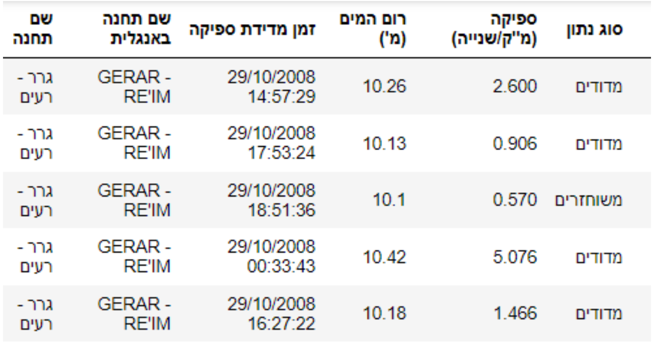
	hidro_station	hidro_station_name
0	23150	BESOR - REIM
1	23160	GERAR - RE'IM
2	23105	BESOR - NEW NIZZANA ROAD

Figure 3. The different hydraulic stations codes and names.

4 EXPLORATORY DATA ANALYSIS AND VISUALIZATIONS

The Flows dataset consists of recorded documents detailing flow and normal flow measurements at various hydraulic stations. Since the dataset lacks time intervals, as shown in the table below, we had to establish a minimum time span to consider two consecutive rows as part of the same flood event. After consulting with specialists from the Dead Sea and Arava region, we decided on a minimum interval of 3 hours.

For each flood occurrence, we defined start and end intervals based on this logic. The integration of the IMS dataset and the Flows dataset was then calculated using the interval time from the Link dataset. If a flood started or ended within the specified interval hours after an



שם תחנה	שם תחנה באנגלית	זמן מדידת ספיקה	רום המים (מ')	ספיקה (מ"ק/שנייה)	סוג נתון
גרר - רעים	GERAR - RE'IM	29/10/2008 14:57:29	10.26	2.600	מדודים
גרר - רעים	GERAR - RE'IM	29/10/2008 17:53:24	10.13	0.906	מדודים
גרר - רעים	GERAR - RE'IM	29/10/2008 18:51:36	10.1	0.570	משוחזרים
גרר - רעים	GERAR - RE'IM	29/10/2008 00:33:43	10.42	5.076	מדודים
גרר - רעים	GERAR - RE'IM	29/10/2008 16:27:22	10.18	1.466	מדודים

Figure 4. Here we see and example of five rows from the Flows data set. Here, we can see that the first and second row are referring to the same flood, while the 4th row is not connected to the others, since it was written long before them.

IMS measurement at the corresponding stations, we marked the IMS measurement as 1 (indicating a flood), and 0 otherwise.

5 EXPERIMENTS AND FEATURE ENGINEERING

Utilizing the comprehensive dataset, our objective was to enhance the feature set for the classification task beyond that of the previous project. Drawing on advice from the Dead Sea and Arava Science Center, we expanded the time windows to thoroughly investigate the impact of rainfall preceding both flood events and normal flow conditions. For each potential sample (with retrospective analysis extending to earlier instances), we generated the following features for every variable (Relative Humidity, Temperature, and Rainfall in mm) across time windows of 30, 60, 120, 180 300, 600, and 1200 minutes (half an hour, one, two, three, five, ten and twenty hours):

- Average
- Median
- Minimum
- Maximum
- Standard Deviation
- EMA (exponentially weighted mean)

following the feature Engineering, we also decided to remove samples which had no rainfall in the twenty hours, since there are no real value in these samples. The final data set is composed of 700280 samples.

The biggest reason for the added time windows is because empirically it has been shown that the biggest causes for floods may not always be strong rain in short periods of time, because the soil may will not be able to soak all the water in a long time. So, the time windows are created to see if we can predict whether the rain cause the flood or not, and what are the most telling causes.

Also, before applying any machine learning models, we aimed to identify the most significant features for predicting floods using statistical methods. We employed two statistical techniques, mutual information and F-value, to compare the relevance of the features we created. Consequently, for each relevant pair of meteorological and hydraulic stations, we calculated and plotted the twenty most significant features.

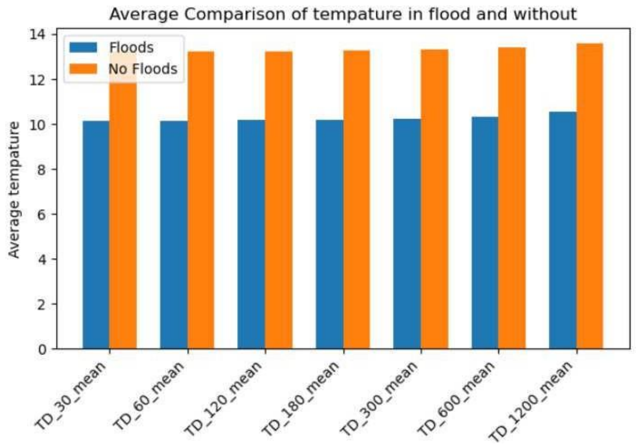


Figure 5. Each bin represents the average temperature in the X minutes preceding the IMS sample timestamp, across all the stations. For example, the bin TD30mean represents the average temperature in the 30 minutes prior. The graph clearly shows that temperatures tend to be lower during flood occurrences.

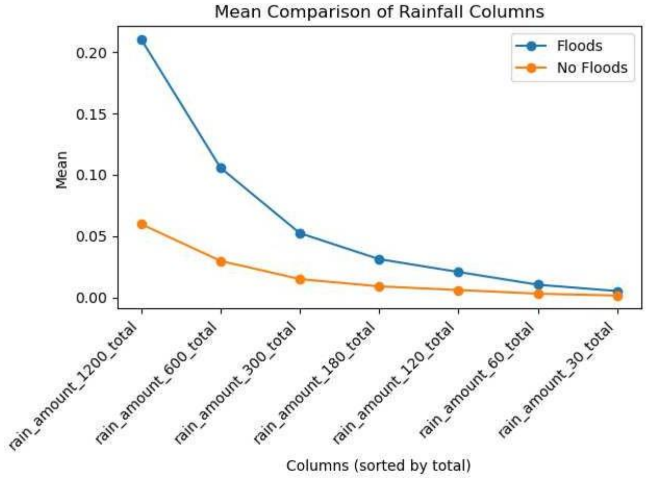
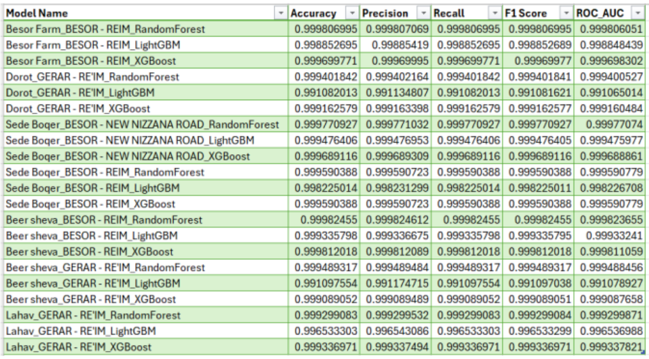


Figure 6. Each bin represents the average rainfall in the X minutes preceding the IMS sample timestamp across all stations. For example, the bin rainamount60total indicates the rainfall in mm that was messured in the station in the 60 minutes prior the IMS sample timestamp. Specialists at the Dead Sea and Arava Center anticipated a hyperbolic pattern for flood events, which was indeed observed in the graph.

We aimed to compare our results with those of the previous research. However, we encountered issues with the process of training and testing the models. Firstly, due to the data being unbalanced, with approximately a 1-10 ratio of floods to normal flows in the hydraulic stations (62842 against 637438), floods samples were up-sampled to match the number of normal flow samples. While this approach seemed appropriate, a major problem arose when the data was randomly split into training and testing sets. This led to significant data leakage from the training set into the testing set. Consequently, when we followed the same procedure, the results appeared unusually favorable, see figure 10 for the results.

Therefore, we attempted three methods to address the issue:

1. We divided the data into training and testing sets based on a temporal split: samples before 2016 were allocated to the training set, while later samples were assigned to the test set. Additionally, similar to the previous approach, we up-sampled the flood samples in the training set. This will be denoted as



Model Name	Accuracy	Precision	Recall	F1 Score	ROC_AUC
Besor Farm_BESOR - REIM_RandomForest	0.999806995	0.999807069	0.999806995	0.999806995	0.999806051
Besor Farm_BESOR - REIM_LightGBM	0.998852695	0.99885419	0.998852695	0.998852689	0.998848439
Besor Farm_BESOR - REIM_XGBoost	0.999699771	0.99969995	0.999699771	0.99969977	0.999698302
Dorot_GERAR - RE'IM_RandomForest	0.999401842	0.999402164	0.999401842	0.999401841	0.999400527
Dorot_GERAR - RE'IM_LightGBM	0.991082013	0.991134807	0.991082013	0.991081621	0.991065014
Dorot_GERAR - RE'IM_XGBoost	0.999162579	0.999163398	0.999162579	0.999162577	0.999160484
Sede Boqer_BESOR - NEW NIZZANA ROAD_RandomForest	0.999770927	0.999771032	0.999770927	0.999770927	0.99977074
Sede Boqer_BESOR - NEW NIZZANA ROAD_LightGBM	0.999476406	0.999476953	0.999476406	0.999476405	0.999475977
Sede Boqer_BESOR - NEW NIZZANA ROAD_XGBoost	0.999689116	0.999689309	0.999689116	0.999689116	0.999688861
Sede Boqer_BESOR - REIM_RandomForest	0.999590388	0.999590723	0.999590388	0.999590388	0.999590779
Sede Boqer_BESOR - REIM_LightGBM	0.998225014	0.998231299	0.998225014	0.998225011	0.998226708
Sede Boqer_BESOR - REIM_XGBoost	0.999590388	0.999590723	0.999590388	0.999590388	0.999590779
Beer sheva_BESOR - REIM_RandomForest	0.99982455	0.999824612	0.99982455	0.99982455	0.999823655
Beer sheva_BESOR - REIM_LightGBM	0.999335798	0.999336675	0.999335798	0.999335795	0.99933241
Beer sheva_BESOR - REIM_XGBoost	0.999812018	0.999812089	0.999812018	0.999812018	0.999811059
Beer sheva_GERAR - RE'IM_RandomForest	0.999489317	0.999489484	0.999489317	0.999489317	0.999488456
Beer sheva_GERAR - RE'IM_LightGBM	0.991097554	0.991174715	0.991097554	0.991097038	0.991078927
Beer sheva_GERAR - RE'IM_XGBoost	0.999089052	0.999089489	0.999089052	0.999089051	0.999087658
Lahav_GERAR - RE'IM_RandomForest	0.999299083	0.999299532	0.999299083	0.999299084	0.999299871
Lahav_GERAR - RE'IM_LightGBM	0.996533303	0.996543086	0.996533303	0.996533299	0.996536988
Lahav_GERAR - RE'IM_XGBoost	0.999336971	0.999337494	0.999336971	0.999336971	0.999337821

Figure 7. The results obtained are highly unrealistic across all models, with almost perfect scores observed in all aspects. This highlights the potential issues stemming from the procedures employed in the previous project.

2. We adopted the same temporal split method as the first approach but incorporated time series cross-validation with five folds. Our objective was to maximize the F1-score to obtain a more reliable evaluation metric for addressing the challenges posed by the imbalanced dataset. We will denote this method as 2).

3. For the same train and test split as before, every pair of stations, we took the top twenty features via mutual info, based on the training data set, and also tried to regularize the models in an attempt to avoid over fitting. We will denote this method as 3).

4. For the same train and test split as before, every pair of stations, we took the top twenty features via F-Value. based on the training data set, and also tried to regularize the models in an attempt to avoid over fitting. We will denote this method as 4).

6 RESULTS

Due to the unbalanced nature of the dataset described earlier, we decided to measure the results using the ROC-AUC score and F1-score. The ROC-AUC score evaluates the model's ability to distinguish between positive samples (floods) and negative samples (normal flows). Its values range from 0 to 1, where 0.5 indicates random guessing (similar to the results of the dummy models we tried), and 1 represents a perfect model. The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both. The harmonic mean gives more weight to lower values, ensuring the F1-score is high only when both precision and recall are high. Like the ROC-AUC score, the F1-score ranges from 0 to 1, with higher scores indicating better model performance.

As can be seen from the graphs below, we could not definitely decide which model or frame work were more effective.

However, as can be seen from Figures 11 and 12, below, they are better then a dummy model which guesses only normal flows, and never guesses any floods. For example. the F1 score and AUC- ROC score for the meteorological station LAHAV and hydraulic station 23160 is at 0.74 for the dummy model, as oppose to 0.81 for the xgboost model for the same pair of stations. Also, there are no differences between the the performances of each model. The two graphs illustrate that there is no significant difference between the models, where the procedure was without cross validation. It was the same for all the procedures. Worth note, the time series cross validation produced the exact same results as without it!!

A big problem that was addressed without major success is the tendency of the models to over fit. Trying to address the issue, we selected the top twenty features like Figure 9 and method 3, and limited the depth of the trees. However, the models continued to show a great over fit, which needs to be solved in future work.

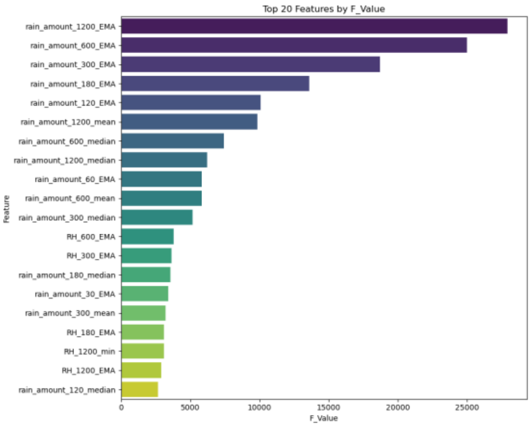


Figure 8. These are the top twenty features with the highest F-values for predicting flood occurrences at the Beer Sheva meteorological station and the Besor Reim hydraulic station. Note, scaling was performed before selection the features.

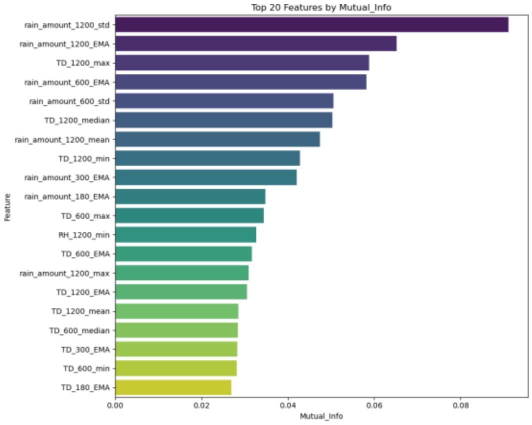


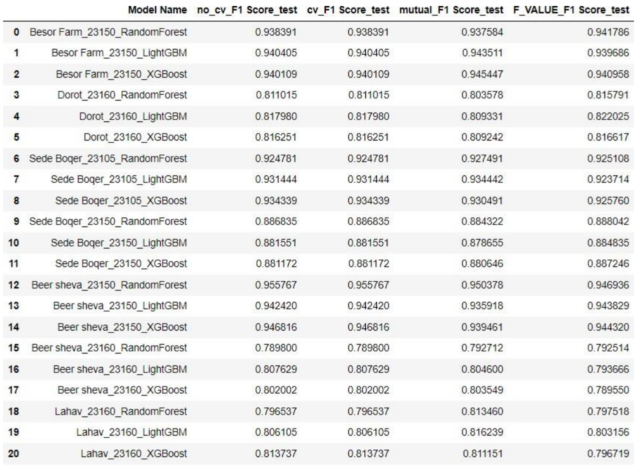
Figure 9. These are the top twenty features with the highest mutual info for predicting flood occurrences at the Beer Sheva meteorological station and the Besor Reim hydraulic station.

FUTURE WORK AND DISCUSSION

As demonstrated, the results were mediocre, with an approximately and average ROC-AUC score of around 0.6. While this score is better than random guessing, it falls short of significant improvement. Two potential factors could explain these issues:

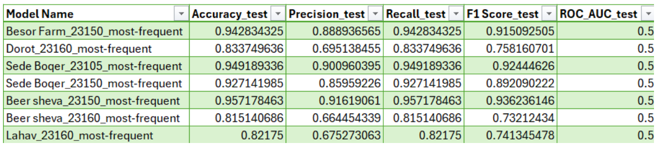
- Rainfall originates from various locations, implying that restricting models to consider only a pair of stations may not be ideal.
- The assumption regarding the time it takes for rainfall to reach the river, along with the documentation referencing the same floods, may not be optimal, necessitating further experimentation.
- Other regularization methods should be considered to mitigate over fitting.

On the other hand, there may have been optimistic results. The methods used to develop and compare the models are not infallible, and as seen from Figure 9 and Figure 16, the most impactful features were those that looked back 20 hours, which could also be a step in the right direction.



	Model Name	no_cv_F1 Score_test	cv_F1 Score_test	mutual_F1 Score_test	F_VALUE_F1 Score_test
0	Besor Farm_23150_RandomForest	0.938391	0.938391	0.937584	0.941786
1	Besor Farm_23150_LightGBM	0.940405	0.940405	0.943511	0.939686
2	Besor Farm_23150_XGBoost	0.940109	0.940109	0.945447	0.940958
3	Dorot_23160_RandomForest	0.811015	0.811015	0.803578	0.815791
4	Dorot_23160_LightGBM	0.817980	0.817980	0.809331	0.822025
5	Dorot_23160_XGBoost	0.816251	0.816251	0.809242	0.816617
6	Sede Boqer_23105_RandomForest	0.924781	0.924781	0.927491	0.925108
7	Sede Boqer_23105_LightGBM	0.931444	0.931444	0.934442	0.923714
8	Sede Boqer_23105_XGBoost	0.934339	0.934339	0.930491	0.925760
9	Sede Boqer_23150_RandomForest	0.886835	0.886835	0.884322	0.888042
10	Sede Boqer_23150_LightGBM	0.881551	0.881551	0.878655	0.884835
11	Sede Boqer_23150_XGBoost	0.881172	0.881172	0.880646	0.887246
12	Beer sheva_23150_RandomForest	0.955767	0.955767	0.950378	0.946936
13	Beer sheva_23150_LightGBM	0.942420	0.942420	0.935918	0.943829
14	Beer sheva_23150_XGBoost	0.946816	0.946816	0.939461	0.944320
15	Beer sheva_23160_RandomForest	0.789800	0.789800	0.792712	0.792514
16	Beer sheva_23160_LightGBM	0.807629	0.807629	0.804600	0.793666
17	Beer sheva_23160_XGBoost	0.802002	0.802002	0.803549	0.789550
18	Lahav_23160_RandomForest	0.796537	0.796537	0.813460	0.797518
19	Lahav_23160_LightGBM	0.806105	0.806105	0.816239	0.803156
20	Lahav_23160_XGBoost	0.813737	0.813737	0.811151	0.796719

Figure 10. Model name refers to the type of meteorological station, and the hydraulic station code, along with the model itself. nocvF1Scores refers to the F1 scores method 1), cvF1Scores refers to the F1 scores method 2), and mutual refers to the F1 scores method 3), and F-Value refers to method 4). As can be seen, barely any difference between the models.



Model Name	Accuracy_test	Precision_test	Recall_test	F1 Score_test	ROC_AUC_test
Besor Farm_23150_most-frequent	0.942834325	0.888936565	0.942834325	0.915092505	0.5
Dorot_23160_most-frequent	0.833749636	0.695138455	0.833749636	0.758160701	0.5
Sede Boqer_23105_most-frequent	0.949189336	0.900960395	0.949189336	0.92444626	0.5
Sede Boqer_23150_most-frequent	0.927141985	0.85959226	0.927141985	0.892090222	0.5
Beer sheva_23150_most-frequent	0.957178463	0.91619061	0.957178463	0.936236146	0.5
Beer sheva_23160_most-frequent	0.815140686	0.664454339	0.815140686	0.73212434	0.5
Lahav_23160_most-frequent	0.82175	0.675273063	0.82175	0.741345478	0.5

Figure 11. The results of the dummy classifiers, which only guess the occurrence of normal flow.

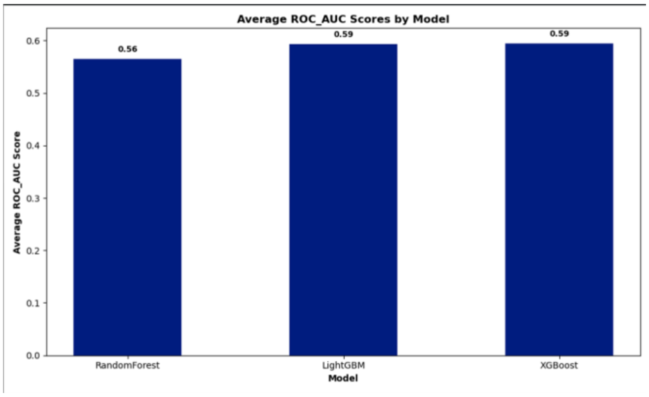


Figure 12. The AUC-ROC score of the three tree models on average across all the experiments without cross validation.

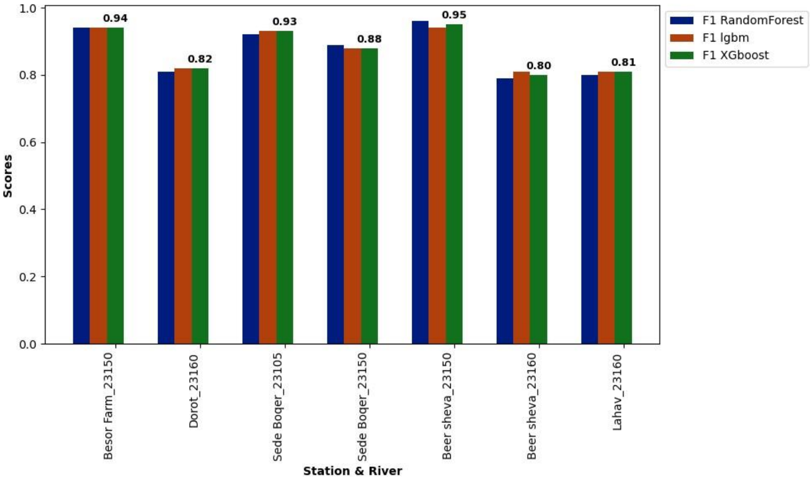


Figure 13. The F1 score of the three tree models across all the experiments without cross validation.

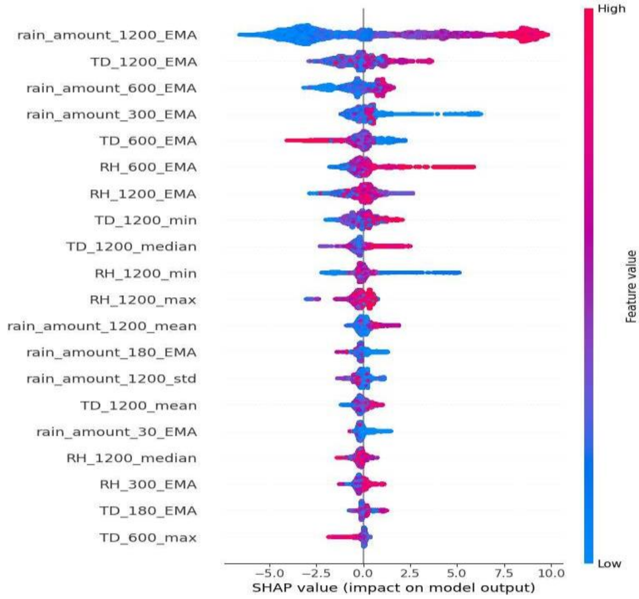
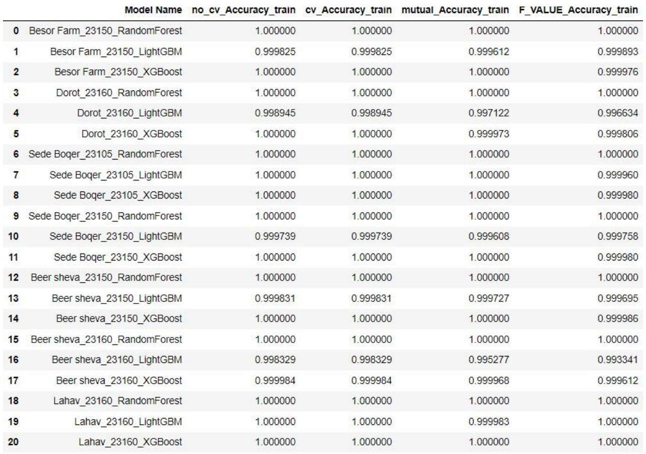


Figure 14. A shap plot of the Xgboost model on Beer Sheva and station code 23150. Can be seen the The EMA of Rainfall in the windows of 5, 10 and 20 hours are very impactful.



	Model Name	no_cv_Accuracy_train	cv_Accuracy_train	mutual_Accuracy_train	F_VALUE_Accuracy_train
0	Besor Farm_23150_RandomForest	1.000000	1.000000	1.000000	1.000000
1	Besor Farm_23150_LightGBM	0.999825	0.999825	0.999612	0.999893
2	Besor Farm_23150_XGBoost	1.000000	1.000000	1.000000	0.999976
3	Dorot_23160_RandomForest	1.000000	1.000000	1.000000	1.000000
4	Dorot_23160_LightGBM	0.998945	0.998945	0.997122	0.996634
5	Dorot_23160_XGBoost	1.000000	1.000000	0.999973	0.999806
6	Sede Boqer_23105_RandomForest	1.000000	1.000000	1.000000	1.000000
7	Sede Boqer_23105_LightGBM	1.000000	1.000000	1.000000	0.999960
8	Sede Boqer_23105_XGBoost	1.000000	1.000000	1.000000	0.999980
9	Sede Boqer_23150_RandomForest	1.000000	1.000000	1.000000	1.000000
10	Sede Boqer_23150_LightGBM	0.999739	0.999739	0.999608	0.999758
11	Sede Boqer_23150_XGBoost	1.000000	1.000000	1.000000	0.999980
12	Beer sheva_23150_RandomForest	1.000000	1.000000	1.000000	1.000000
13	Beer sheva_23150_LightGBM	0.999831	0.999831	0.999727	0.999695
14	Beer sheva_23150_XGBoost	1.000000	1.000000	1.000000	0.999986
15	Beer sheva_23160_RandomForest	1.000000	1.000000	1.000000	1.000000
16	Beer sheva_23160_LightGBM	0.998329	0.998329	0.995277	0.993341
17	Beer sheva_23160_XGBoost	0.999984	0.999984	0.999968	0.999612
18	Lahav_23160_RandomForest	1.000000	1.000000	1.000000	1.000000
19	Lahav_23160_LightGBM	1.000000	1.000000	0.999983	1.000000
20	Lahav_23160_XGBoost	1.000000	1.000000	1.000000	1.000000

Figure 15. Can clearly see the accuracy is extremely high for all models. The attempt to select features and limit the trees depth had very minor success.

Bibliography:

We would like to thank Dr Merav and Prof Avshalom from the Dead Sea and Arava center for their guidance and help throughout the whole process.

Also, Thanks for ChatGPT, the Open AI tool, for helping in writing the article.

article.