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

Ido Hersko*, Elran Oren*
Department of Software and Information Systems Engineering
Ben-Gurion University of the Negev
Beer-Sheva, Israel
{hersko, elranor}@post.bgu.ac.il

***Abstract*—Southern Israel faces a crucial challenge in water resource management due to the absence of historical hydraulic flow data beyond the last 20 years. While meteorological data spanning decades is available, the lack of knowledge about past river flows directly impacts the region's aquifer systems, vital for water sustainability. This study presents a machine learning approach to predict historical river flows using high-resolution 10-minute meteorological data. We constructed a feature-rich dataframe and trained multiple models using available hydraulic flow data. SHAP analysis was applied to identify key predictive features. As part of the project, we trained 21 models for 21 links between meteorological stations and hydraulic stations (that measure flow rates) and created a GUI application that enables scientists to use those models in further research. Our most promising model, an XGBoost model, achieved a precision and recall rate of 0.98 when trained on meteorological data from the Mitzpe Ramon station on the Arava hydraulic station. This research offers a crucial insight into river flow patterns in Southern Israel. It enables a better understanding of historical aquifer changes since the establishment of the State of Israel. All the relevant code and ressources are available in our Github repository [1].**

***Index Terms*—Machine learning, Meteorological data, Hydrology, Xgboost, SHAP**

I. Introduction

Water resource management is a critical concern in arid regions, and Southern Israel is no exception. With a history of meteorological data spanning several decades, it is well-documented that rainfall, temperature, pressure, and humidity patterns have shaped the environmental dynamics of the region. However, a substantial gap in our knowledge exists when it comes to historical hydraulic flow data, which represents a vital component of Southern Israel's water resource landscape. Specifically, hydraulic flow data, which characterizes river flows over time, is only available and organized for approximately the past two decades, thus limiting our understanding of river flows over time.

This deficiency in historical hydraulic data presents a pressing challenge for water resource managers, scientists, and policymakers. The reason for this urgency lies in the direct connection between river flows and the water levels of aquifers. This link plays a central role in maintaining and understanding water resources in the region. Generally, when river flows occur, they significantly influence the aquifers' water levels, which, in turn, impact the overall water resource picture. Consequently, the absence of historical flow data hampers our ability to gain a comprehensive insight into the changes these aquifers have undergone since humans started drawing water from them for irrigation and everyday 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. Leveraging this 10-minute dataset, we construct time windows of 30-minute and calculate several features for each timestamp, resulting in building a dataframe with many meteorological features.

We deploy multiple machine learning techniques, such as XGBoost, Random Forest, and Decision Tree, to create models for simulating river flows. Following this, we apply SHAP analysis to uncover the most influential factors within our models, shedding light on the key drivers behind past flow patterns.

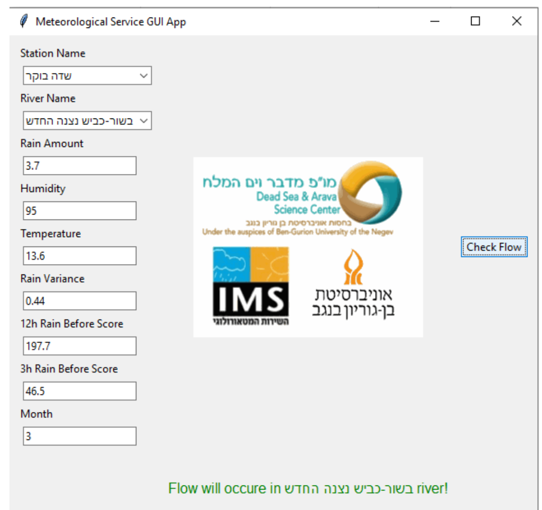


Fig. 1. Screenshot of a success query in the GUI application

The standout result of our study is the remarkable performance of the XGBoost model, which we trained using our features driven from the meteorological data of the Mitzpe Ramon station. This model achieves an impressive precision and recall rate of 95%. Our remarkable metrics are a result of our distinctive features.

This research represents a significant advancement in understanding past river flow dynamics in Southern Israel and, by extension, 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 focuses on 21 unique links between meteorological and hydraulic stations, with a dedicated machine learning model for each. To facilitate further research, we've developed a user-friendly GUI application showcasing and making these 21 models accessible, depicted in Figure 1.

II. Background and Related Work

Machine Learning (ML) has emerged as a pivotal field in computer science and artificial intelligence, revolutionizing our ability to extract knowledge from data. [2] Ensemble learning methods such as **XGBoost** [3] and **Random Forest** [4] have garnered considerable attention in ML research and practice. XGBoost, an implementation of gradient boosting, is renowned for its efficiency and predictive power, making it a popular choice in regression and classification

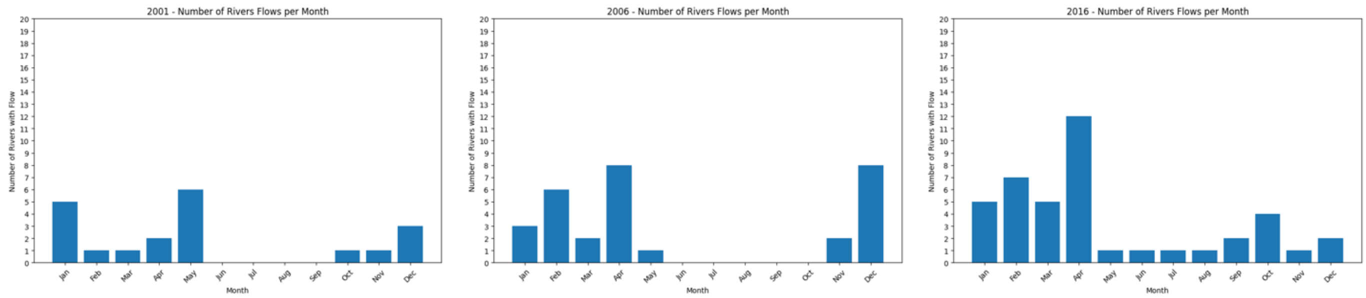


Fig. 2. A visualization of three years - the score for each month

tasks. Random Forest, on the other hand, is an ensemble of decision trees that offers robustness to overfitting and excels in handling high-dimensional data. To interpret and understand the complex models generated by these algorithms, the **Shapley Additive Explanations (SHAP)** [5] framework has become indispensable. SHAP values provide a comprehensive approach to model explainability, allowing practitioners and researchers to ascertain the contributions of individual features to model predictions, enhancing transparency and trust in ML models.

Flash floods are rapid and high-intensity flooding events, which are mainly caused by heavy rainfall. Since flash floods have a short response time of several hours, they are difficult to predict and cause damages and even casualties [6]. As such, the importance of predicting flash flood events is reflected in numerous studies that have been conducted on this topic in recent years. For example, the FLASH project [7] used lightning data to improve flash flood predictions in the Mediterranean Basin. Recently, Ziskin and Reuveni [8] investigated the use of precipitable water vapor (PWV) derived from ground-based global navigation satellite system (GNSS) stations to assist in predicting flash floods in an arid region of the eastern Mediterranean (EM).

III. Methodology and Experimental Setup

The goal of this section is to provide an understanding of the flow of our experiment and to discuss its main aspects.

A. Datasets

IMS The IMS data refer to the data from the israeli meteorological service (IMS) [9]. We connect by API and downloaded 130 csv of meteorological stations. For each stations we downloaded 21/22 csv (one for each year from 2000 until 2021).

Flows The Flows data refers to the data from the water authority in Israel [10]. We downloaded three hydrographs. The first is flows data from 2000 until 2009, the second is from 2010 until 2019 and the third is from 2020 until 2022.

Link The Link data refers to a dataset that we preprocess from the dead sea and arava science center. The dataset can be found on our github [1]. This dataset helps us to set up a time window and set up the links between meteorological stations to hydraulic stations.

Hydraulic stations The hydraulic stations are taken from the dataset of the water authority [11]. It can be found on our github [1].

B. Visualization

We employ data visualization to familiarize ourselves with the dataset's attributes and explore how this data can be of value. Our initial focus centers on visualizing the Flows datasets, with the goal of gaining valuable insights to understand what is pertinent for successful flow prediction. Our analysis encompasses a total of 20 rivers, each assessed monthly. For each river in a given month, we assign a score of 1 for recorded flow. This approach aims to identify months that offer little to no meaningful data across most rivers. Figure 2 clearly indicates that, by and large, the months of June, July, and August consistently exhibit minimal flow. Consequently, we can reduce our dataset by one-quarter by excluding these months.

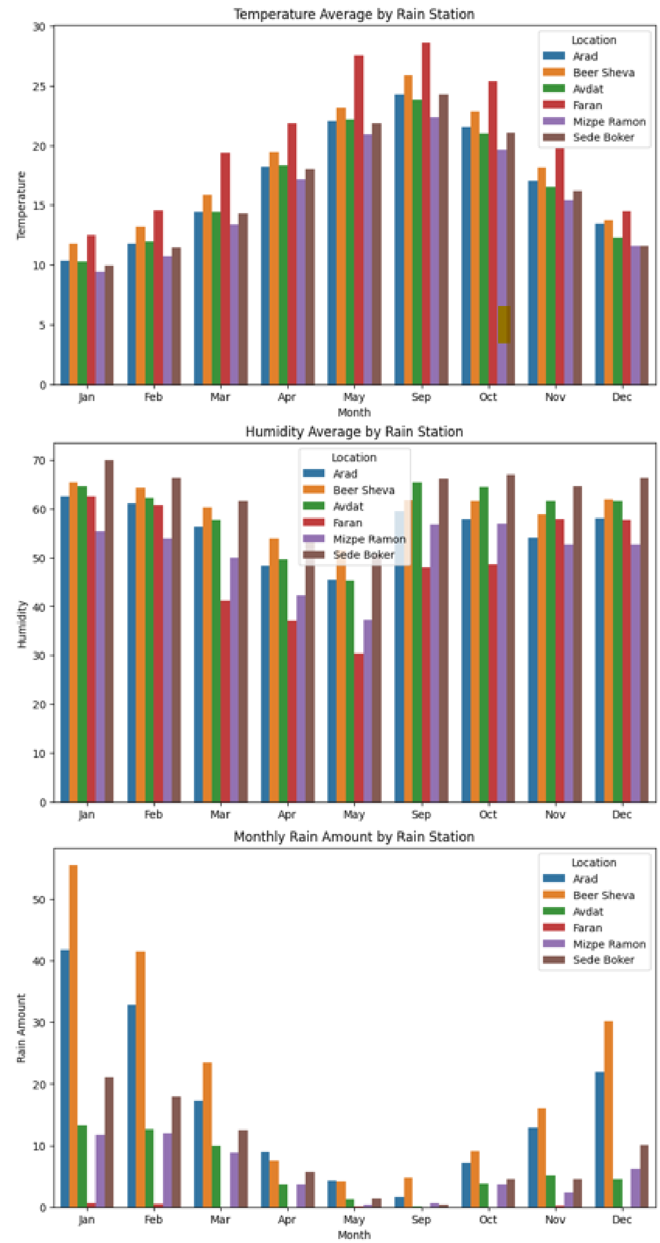


Fig. 3. Average Temperature, Humidity and Rain for Years 2000-2021, per Month

Figure 4 shows the average flow strength when a flow occurs in a given river. In figure 5 we can visualize the number of flows that occurs in a 20 year scope. It seems that over the years, the number of flows increase.

Next, we investigate the IMS data, we build from the 130 csv one

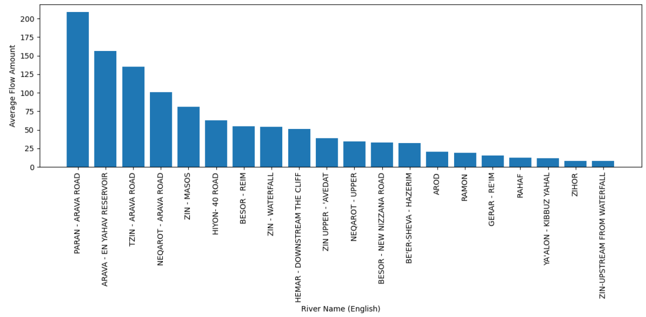


Fig. 4. A visualization of average flow strength through 20 years for each river

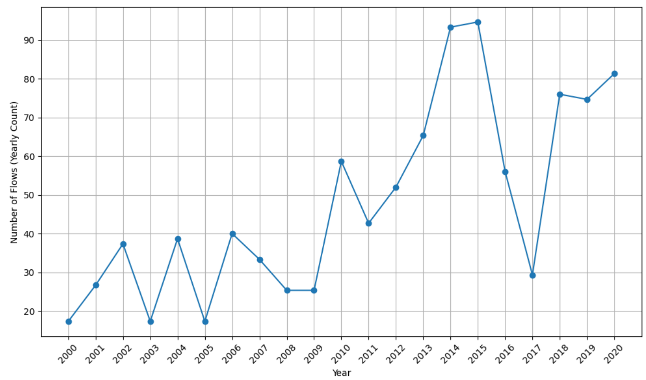


Fig. 5. Number of Flows by Year (2000-2021)

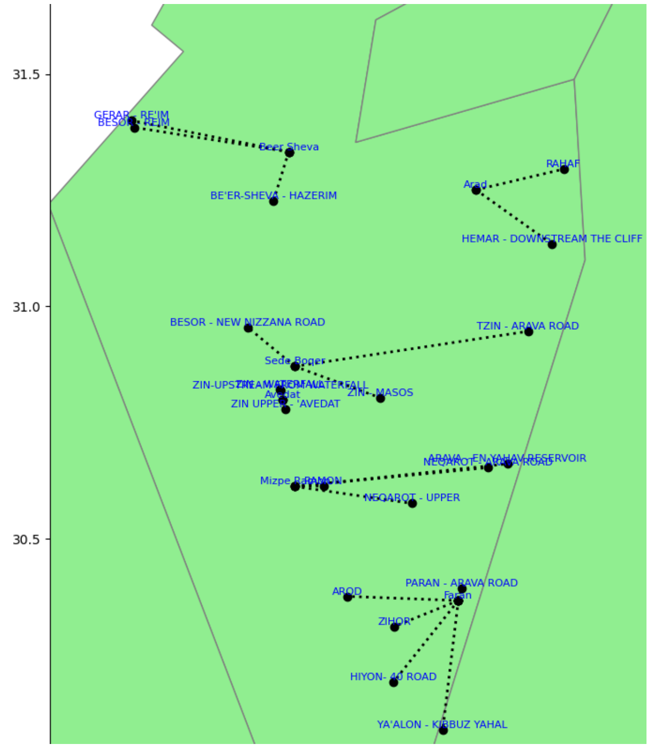


Fig. 6. Southern Israel Meteorological Stations and Hydraulic Stations

big dataframe and analyze it in order to gain insight over the features and understand a general distribution of the data. Figure 3 shows the average temperature, average humidity and the average rain sum for each month, over 20 years.

Finally, we located the meteorological stations and the hydraulic stations in a map, to asses the time window in relation to the distance and the time of the water to arrive from the meteorological stations to the hydraulic stations during rainfall. Figure 6 shows the map of Southern Israel with the locations.

C. Features Engineering

In this section, we comprehensively describe the features we integrated into our dataset that capture the intricate relationship between rainfall and river flow. These features have been carefully engineered to provide insights into the environmental conditions influencing river flow following a rainfall event.

- **Rain_Timestamp_begin/end**: These temporal attributes define the 30-minute interval during which rainfall and associated environmental parameters are recorded, serving as the temporal context for all other features.
- **Temperature_Avg**: Represents the average temperature within the 30-minute interval. Temperature influences snowmelt and evaporation, impacting river flow.
- **Air_Moisture_Avg**: Captures the average air moisture or humidity during the 30-minute interval, affecting evaporation and overall moisture content, thus influencing river flow.
- **Rain_Accumulation_30min**: Quantifies the total rainfall accumulation in millimeters during the 30-minute interval, providing a direct measure of precipitation, a key driver of river flow.
- **Rain_Dispersion_Score**: Assesses the equal distribution of rainfall within the 30-minute window. Calculated as the variance of rainfall across three 10-minute intervals within the half-hour period, a lower score indicates more uniform rainfall dispersion over the 30 minute time window.
- **12_hours_Rain_Score** and **3_hours_Rain_Score** Captures the historical impact of rainfall events in the 12/3 hours leading up to the current interval. It assigns higher weights to more recent events, and less weights as the time window is far from the current time window. It reflect and pronounce the influence of the wetness and possible saturation of the ground. It can be calculated as in equation 1. $X\_hours\_Rain\_Score$ is the feature capturing the historical impact of rainfall events in the X hours leading up to the current interval. $rain\_amount(t)$ is the rainfall amount at time step t. N is the total number of time steps within the X-hour window.

$$X\_hours\_Rain\_Score = \sum_{t=1}^{N-1} rain\_amount(t) \cdot (N - t) \quad (1)$$

- **Month_of_the_Rain**: This newly added feature represents the month in which the rainfall event occurred. Understanding the seasonality of rainfall is crucial, as different months may exhibit distinct hydrological behaviors.
- **Is_Flow**: Our target label, 1 if there is flow, 0 otherwise.

D. Preprocess and dataframe building

Our datasets came very "unclean", we had to make a lot of preprocess, delete rows that were obiously anomalies (e.g. temperature of 300 degree, humidity of 450%). For each feature we set an upward threshold and a minimum threshold and change the anomalies to the value that is upward to the given anomaly row. Additionaly, we had rows that were blank, therefore we needed to fill it as mentioned before.

Dataframe

In this study, we employed the IMS dataset as our primary data source. To ensure data quality and consistency, we processed this dataset to a preprocessing phase. During preprocessing, we performed various data cleanup and transformation tasks to prepare it for further analysis.

As a key step in our data preparation pipeline, we generated a set of features based on the IMS dataset. These features are crucial for subsequent analyses and are discussed in detail in Section III-C. Additionally, to facilitate time-based analyses, we put the data into

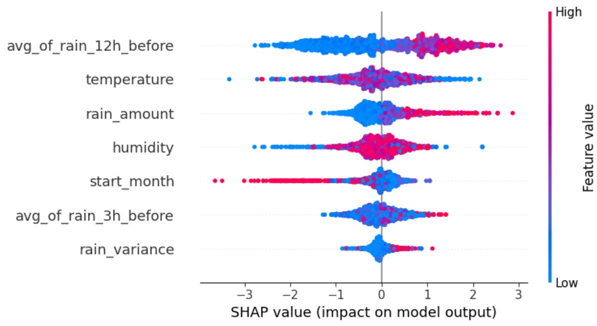


Fig. 7. XGboost - SHAP Summary Plot

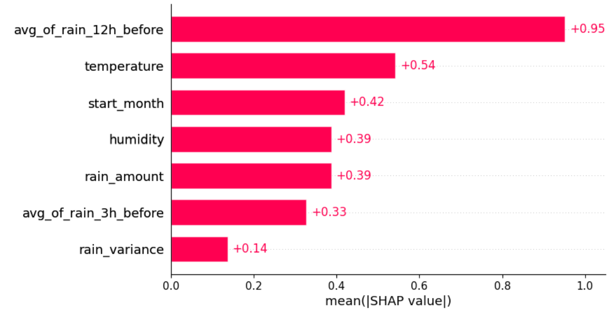


Fig. 8. XGboost - SHAP Bars Plot

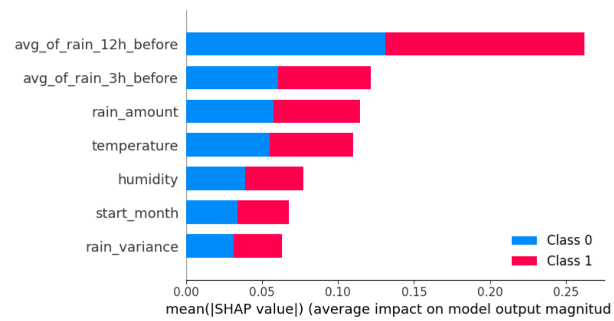


Fig. 9. Random Forest - SHAP Bars Plot

30-minute time windows, a choice driven by the dead sea and arava science center requirements.

To establish temporal relationships and cross-reference information across different datasets, we integrated the processed IMS dataset with the Link dataset. This integration enabled us to associate time windows, which served as temporal anchors for subsequent queries in the flow dataset.

Utilizing the time windows obtained from the Link dataset, we successfully integrated the Flows dataset into our analysis pipeline. The Flows dataset serves as our target label, where we aim to predict a binary outcome: 1 denotes the presence of a flow instance, while 0 is assigned as the label for rows where no corresponding flow instance was identified.

IV. Results

In this section, we present our findings and conduct a detailed exploration of the features that exerted the most significant influence on our models.

A. Machine Learning Models

Our constructed dataframe exhibited an imbalance, characterized by an excessive number of instances labeled as '0' (representing no flows) and a lack of instances labeled as '1' (indicating the presence of flows). To facilitate effective machine learning model training, our initial step involved addressing this data imbalance.

Subsequently, we initiated the process of training machine learning models on the balanced dataset. Our selection of models included Random Forest, Decision Tree, and XGBoost. Each model underwent training on a 60/40 stratified data split, ensuring a representative distribution of data points in both the training and testing sets.

For model implementation, we imported these algorithms from the scikit-learn library with their default hyperparameters. To evaluate the performance of our models, we employed key metrics such as accuracy, precision, recall, and the F1 score. These metrics served as crucial indicators of the models' effectiveness in capturing patterns and making predictions based on the data. We trained 21 models, every link between a meteorological stations and hydraulic stations is fitted to a specific suited model. The links can be shown on the map of southern israel in figure 6. Our results can be seen in table 1 and a visualization of the accuracy of all the 21 models can be seen in 10.

B. SHAP - Feature Importance

We utilize the SHAP library to gain valuable insights into the impact of features on our model and the degree of their influence. Our primary focus lies on the Random Forest and XGBoost models due to their consistently high accuracy performance. We will present findings related to the Avdat meteorological station and the Zin-Upstream hydraulic station. In the case of XGBoost, two visualizations offer a comprehensive understanding of feature contributions to output predictions: the SHAP Summary Plot (Figure 7) and mean SHAP values (Figure 8). These visuals provide a global view of feature importance. For the Random Forest model, the SHAP Bars Plot (Figure 9) provides a detailed perspective on how each feature influences predictions. This plot employs horizontal bars to depict the magnitude and direction of their impact.

As observed in our analysis, the most influential feature for predicting the occurrence of a flow is the feature we developed, $avg\_of\_rain\_12h\_before$. As detailed in Section III-C, this feature represents a weighted average of rainfall amounts in the 12 hours leading up to a specific 30-minute time frame.

What's intriguing is that our results indicate that the temperature feature also plays a significant role in flow prediction. Surprisingly, the amount of rainfall within the current 30-minute period does not emerge as the most critical factor, which defies the expectation that it would be the primary predictor.

V. Conclusion and Future Work

This study addressed a critical knowledge gap in Southern Israel's water resource management by predicting historical river flows using machine learning and high-resolution meteorological data. The absence of hydraulic flow data beyond the last two decades hindered aquifer understanding and sustainability. We integrated data from various sources, engineered rich features, and achieved a 98% precision and recall rate with the XGBoost model in our best case.

Future research directions may be extending the research to cover all rivers in Israel to assess the robustness and reliability of the method, a development of a real-time model for security forces to monitor river flows in real-time for emergency response. Lastly, further feature engineering efforts should be made to enhance predictive capabilities. These steps will contribute to more comprehensive water resource management and decision-making across the country.

VI. Acknowledgment

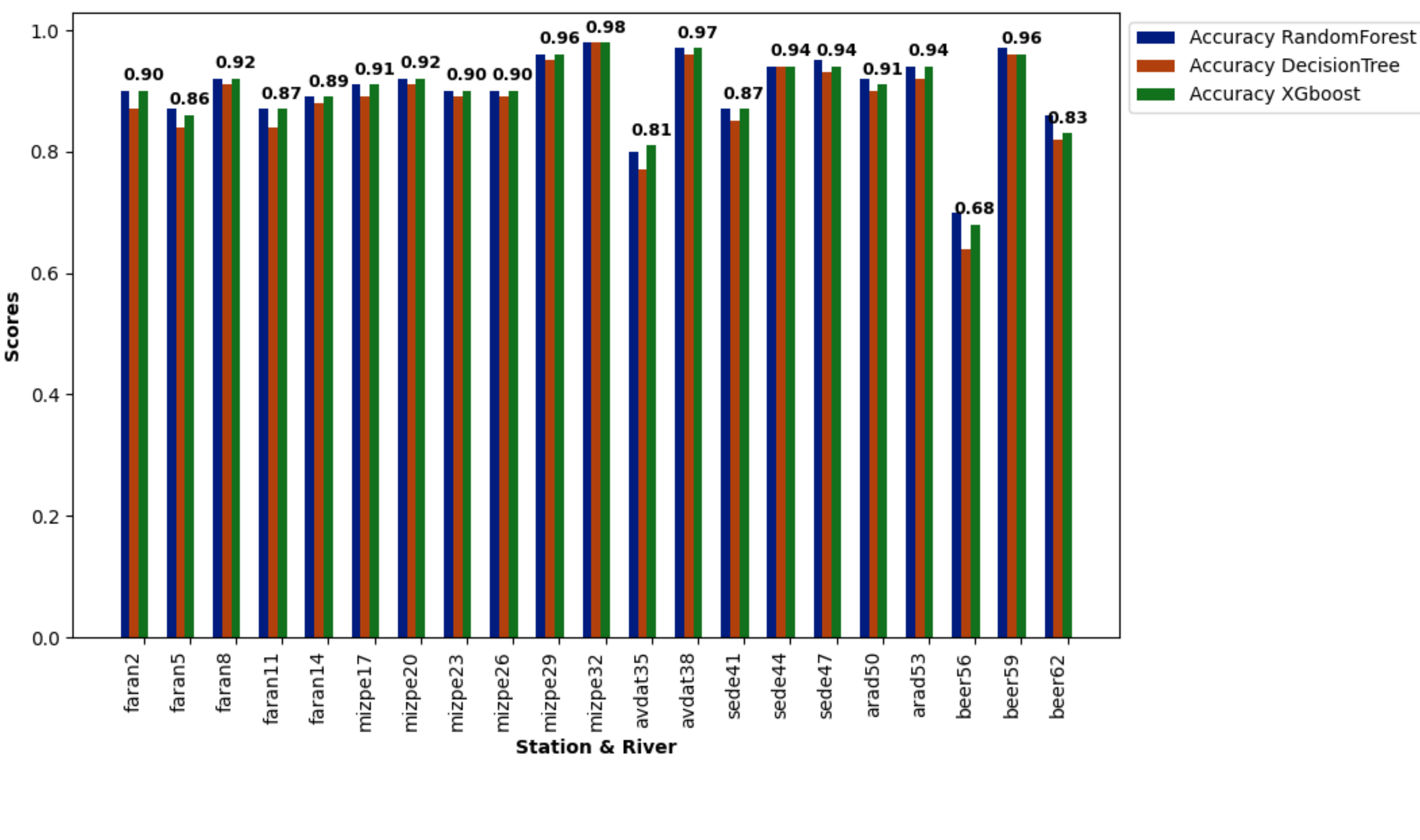
We would like to deeply thank Dr. Avshalom Babad and Dr. Merav Cohen from the dead sea and the arava science center for their guidance and help througout the research. As well, we would like to thank OpenAI for creating ChatGPT, which assisted us in writing this paper.

APPENDIX

TABLE I
SUMMARY OF THE 21 LINKS AND THEIR ACCORDING 3 MODELS

station	River	Random Forest				Decision Tree				XGBoost			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Faran	Zihor	0.897	0.899	0.897	0.897	0.873	0.876	0.873	0.873	0.896	0.896	0.899	0.896
	Arava Road	0.868	0.875	0.868	0.867	0.835	0.840	0.835	0.835	0.862	0.869	0.862	0.862
	Hiyon	0.919	0.924	0.919	0.919	0.907	0.911	0.907	0.906	0.921	0.926	0.921	0.921
	Arod	0.871	0.881	0.871	0.870	0.841	0.849	0.841	0.840	0.868	0.878	0.868	0.867
	Ya'alon	0.894	0.901	0.894	0.893	0.875	0.883	0.875	0.875	0.895	0.902	0.895	0.895
Sede Boker	Zin - Masos	0.870	0.876	0.870	0.869	0.850	0.856	0.850	0.849	0.871	0.877	0.871	0.871
	Zin - Arava	0.944	0.947	0.944	0.944	0.936	0.938	0.963	0.935	0.944	0.946	0.944	0.943
	Basor	0.946	0.948	0.946	0.946	0.934	0.936	0.934	0.934	0.942	0.944	0.942	0.942
Mizpe Ramon	Ramon	0.908	0.914	0.908	0.907	0.887	0.893	0.887	0.887	0.905	0.912	0.905	0.905
	Neqarot - Up	0.920	0.926	0.920	0.920	0.909	0.914	0.909	0.909	0.922	0.927	0.922	0.921
	Avedat	0.905	0.912	0.905	0.904	0.886	0.893	0.886	0.885	0.898	0.905	0.898	0.897
	Arod	0.889	0.906	0.889	0.889	0.890	0.896	0.890	0.890	0.903	0.911	0.903	0.902
	Neqarot	0.955	0.957	0.955	0.955	0.949	0.951	0.949	0.949	0.956	0.958	0.956	0.955
	Arava	0.980	0.981	0.980	0.980	0.978	0.979	0.978	0.978	0.981	0.982	0.981	0.981
Beer Sheva	Hazerim	0.698	0.703	0.698	0.697	0.639	0.640	0.639	0.638	0.682	0.685	0.682	0.680
	Basor - Reim	0.972	0.974	0.972	0.972	0.964	0.966	0.964	0.964	0.965	0.967	0.965	0.965
	Gerar - Reim	0.856	0.865	0.856	0.855	0.820	0.829	0.820	0.819	0.825	0.832	0.825	0.824
Avedat	Zin - Up	0.803	0.816	0.803	0.802	0.766	0.776	0.766	0.764	0.806	0.821	0.806	0.805
	Zin	0.970	0.971	0.970	0.970	0.962	0.964	0.962	0.962	0.969	0.971	0.969	0.969
Arad	Hemar	0.917	0.922	0.917	0.917	0.899	0.905	0.899	0.898	0.907	0.911	0.907	0.906
	Rahaf	0.940	0.942	0.940	0.940	0.925	0.928	0.925	0.925	0.937	0.940	0.937	0.937

Fig. 10. Visualization of the accuracy of the 21 models



References

[1] I. Hersko and E. Oren, "Our github repo can be found at:."

[2] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[3] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[4] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[6] M. Borga, E. Anagnostou, G. Blöschl, and J.-D. Creutin, "Flash flood forecasting, warning and risk management: the hydrate project," *Environmental Science & Policy*, vol. 14, no. 7, pp. 834–844, 2011.

[7] C. Price, Y. Yair, A. Mugnai, K. Lagouvardos, M. Llasat, S. Michaelides, U. Dayan, S. Dietrich, F. Di Paola, E. Galanti, *et al.*, "Using lightning data to better understand and predict flash floods in the mediterranean," *Surveys in geophysics*, vol. 32, pp. 733–751, 2011.

[8] S. Z. Ziv and Y. Reuveni, "Flash floods prediction using precipitable water vapor derived from gps tropospheric path delays over the eastern mediterranean," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[9] IMS, "You can find more information about the ims dataset at:." https://ims.gov.il/en.

[10] W. Authority, "You can find more information about the water authority dataset at:." https://data.gov.il/dataset/hidrograph".

[11] W. Authority. https://data.gov.il/dataset/hydro$_{s}tation$".