WaterFinder

An Interactive Visualizer based on Big-Data Framework for Location of Ground Water Wells

Vishvaditya Luhach
University of California, Riverside
SID: 86254741
Group WaterFinders
vluha001@ucr.edu

Seokha Kang
University of California, Riverside
SID: 862306301
Group WaterFinders
skang121@ucr.edu

Jiahua Xue
University of California, Riverside
SID:
Group WaterFinders
jxue041@ucr.edu

ABSTRACT

"When the well is dry, we know the worth of water." Benjamin Franklin wrote these words in his 'Poor Richard's Almanack' in 1746 [1]. He meant the quote in a more metaphorical sense about the human condition of taking our surroundings for granted but, in today's day and age these words can be taken literally.

With the advent of climate change, water – a renewable resource – has become scarce. A lot of places across the globe are suffering from severe droughts and rainfall has lessened in significant parts. This has led to a decrease in ground water levels, a major source of drinking water in large parts of the world. This has led to a vast population of humans to not have access to safe drinking water. With our project, we want to provide an interactive and easy to use web-based application for visualizing the location of ground-water wells and display their relevant information like ground water levels, water quality and so on. We also plan to integrate a ML/AI model to forecast the ground water levels around the wells. This can help governments and individuals of the region to make informed decisions about their water usage and take preventive measures if a severe depletion in the water level is forecasted.

1 Introduction

1.1 Motivation

Groundwater is a critical resource for drinking water, agriculture, and industrial use worldwide. With increasing demand and climate change impacting water availability [2], there is an urgent need to monitor and manage groundwater resources effectively. However, the data on groundwater wells is often scattered across various sources, making it challenging to obtain a comprehensive view of groundwater status in different regions.

This project aims to address these challenges by developing a website that aggregates data on groundwater wells from various online sources and visualizes it on an interactive map. By leveraging Big Data technologies, the platform will fetch, process, and display up-to-date information on well locations, water levels, and quality indicators. Users will be able to explore trends in groundwater availability across regions, track changes over time, and make data-driven decisions for resource management.

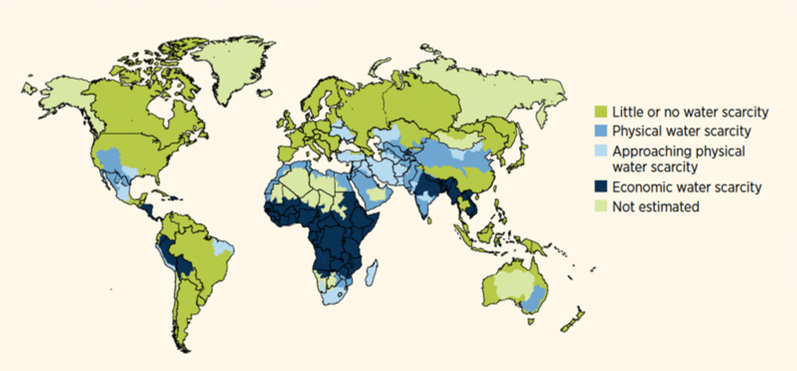

Little or no water scarcity
Physical water scarcity
Approaching physical water scarcity
Economic water scarcity
Not estimated

Figure 1: Physical and economic water scarcity in the world (credits: United Nations, un.org)

The project also aims to enhance transparency and accessibility, providing vital data to researchers, policymakers, and the general public. By presenting the data in an intuitive and user-friendly format, the platform seeks to bridge the gap between data complexity and public understanding, ultimately contributing to better-informed water conservation efforts and sustainable groundwater management. The project's motivation lies in its potential to improve transparency and accessibility in groundwater data, contributing to a more responsible approach to water resource management in the face of increasing water scarcity and climate-related challenges.

1.2 Dataset

The project will use the 'California Water Supply Well Completion Report Data' [3] which is aggregated by the California Department of Water Resources. The data will be collected through the use of United States Geological Survey's (USGS) ScienceBase API [4] to gather information about groundwater wells, including their locations, depth, and attributes such as the years they were built etc. For visualizing the data on an interactive map, we will be using the OpenStreetMap Overpass API [5] to fetch the geospatial data and display the map on an interactive webpage. To showcase that the project can also be integrated into a machine learning task, we also plan to utilize the above dataset in conjunction with the 'California Ground Water Level data' [6] to forecast the ground water level in the region of the selected well. For the current scope of the project, we intend to use the data only for the Southern California region, but we will showcase that the current implementation can be scaled to handle much larger datasets without significant changes in the underlying system.

The rest of the report consists of five other sections. Firstly, we go over the literature review to understand previous work that has been done in this field. Next, we go over the methodology, which describes the implementation of our data processing pipeline, prediction model framework and web application interface. In the next section – results and evaluations – we evaluate the performance of our data processing pipeline as well as the prediction model that we have designed. We conclude the report by mentioning the contributions of each group member and provide our concluding remarks in the final section.

2 Literature Review

2.1 Background

Water resources have been greatly affected due to the advent of climate change and overpopulation [7]. In regions across the globe groundwater is a primary source of fresh water for industrial and everyday use [8], but the groundwater reservoirs have also been steadily depleting [9]. This poses a need for better management of these resources, which involves keeping a track of the data collected from various reservoirs and groundwater wells, and to actively predict, with a certain degree of accuracy, the future level of these resources. This poses a big-data management problem as this data is collected for a large number of wells across a huge geographic region and necessitates the processing of this data at scale. Our project aims to showcase a working solution, by implementing a Spark based web application which shows the pertinent information for groundwater well in the Southern California region and also aims to predict groundwater level in the region.

2.2 Big-Data Tools

This involves data cleaning, validation, integration, and preparation processes that are essential for accurate analysis and modeling. In big data projects, efficient data processing is crucial to handle large volumes of data [10], ensure data quality [11], reduce errors, and enhance the performance of analytical models and visualizations.

Our project uses PySpark, an API (Application Programming Interface) developed to use Spark in Python [12]. It allows for faster processing of large datasets in Python that can also be scaled to distributed systems. The processed data will be stored in JSON (JavaScript Object Notation) format on MongoDB [13] which utilizes a NoSQL database to store data.

2.3 Big-Data and Ground water

Much work has been done in order to leverage big-data tools and techniques to measure and analyze groundwater data. Gaffoor et al. leveraged big-data analysis for effective management of groundwater resources in the Southern African region [14]. They conceptualized a framework that can be utilized to collect, process and analyze large datasets in order to fill gaps in the knowledge regarding the groundwater resources in the region. We aim to leverage similar techniques and showcase a working web application which can deliver these insights to the general public and policymakers effectively. Cheng et al. used big-data analysis for assessing the quality of the groundwater in multiple districts in Shanghai, China [15]. Martinez-Santos and Renard utilized a collection of big-data methodologies to identify the groundwater potential in the sub-Saharan region [16]. These studies showcase the potential of leveraging big-data techniques to solve real-world problems and their applications.

2.4 Geospatial Visualization

Geospatial visualization and GIS are a crucial part of our project as the main goal of the project is facilitation of information to end users in an easy-to-understand and accessible manner through visualization on an interactive map. The GIS data is not only useful for data visualization but also data mining as evident in the paper by Naghibi et al. [17]. Lopez et al. used satellite and geospatial data to effectively map groundwater resources using land with active irrigation agriculture [18]. These works showcase the importance of GIS data and satellite data for management of groundwater.

In our project, geospatial visualization is a key component that allows users to interact with groundwater well data on an interactive map. By processing raw data into GeoJSON (Geographic JavaScript Object Notation) format containing longitude and latitude coordinates, well numbers, and county information, we can render the wells accurately based on their locations. We utilized Leaflet, an open-source JavaScript library for mobile-friendly interactive maps [19], to display this data on the web application.

2.5 Prediction Model

Accurate prediction models are crucial for sustainable groundwater management, enabling stakeholders to anticipate changes, plan resource allocation, and implement measures to mitigate adverse effects such as water scarcity or overexploitation. Tao et al. in their recent review study published a highly comprehensive collection of state-of-the-art models used in various applications of groundwater level [20]. It showcased the application of 10 broad range techniques for groundwater level prediction and the gaps in the future work potential in the domain. Daliakopoulos et al. proposed a deep-learning model for the forecasting of groundwater level in Messara Valley in Greece. In their study, a Feed-Forward Neural Network (FNN) combined with the Levenberg-Marquardt method performed the best with a root mean squared error (RMSE) of 2.11 meters [21].

Our project aims to utilize newly proposed methodologies such as transformer models for higher accuracy over a long-term prediction to allow for planned optimization of water with a clear understanding of the future implications.

3 Methodology

Our project is primarily split into three parts – data processing, machine learning and web application. The data processing aspect of the project deals with gathering the required datasets and processing them using PySpark, a python implementation of Spark which can be used for processing large datasets on distributed systems. We process the raw datasets into two parts – data for web application, which is the data that will be displayed on the web page for the end user and dataset for model training, which is the dataset we will be using for training our machine learning model for predicting future ground water levels in a county. Although the datasets we are using are smaller in size, the PySpark implementation ensures that this methodology can be scaled to much larger datasets as well.

For our prediction model, we have decided to use Temporal Fusion Transformer (TFT) [22]. Apart from robust performance, TFT offers some unique advantages as compared to other models such as support for static covariates and explainability for the trained model. The last part of our implementation is the web application which displays the location of the ground water wells on an interactive map as well as visualizes the output of the machine learning model in easy-to-understand graphs that show the predicted ground water level in a county that the user can filter. The web site uses MongoDB for fetching and managing data. The following sections of this report delve deeper into the implementation of these parts.

3.1 Data Gathering

The raw data for this project was fetched from various sources and collated together. The first dataset is the "Well Completion Report" dataset, fetched from USGS's website (filename: 'WCR_v4_2024.txt'). Figure 2 shows the distribution of these wells by county in the state of California and what percentage of wells are used for which activity. This dataset consists of fields such as the 'WCRNumber' and 'WCRN' which are unique identifiers for each well, as well as corresponding information regarding the well such as the latitude, longitude, construction date etc. This data is processed in order to remove unnecessary columns and standardize column headers etc. The following section will provide a more in-depth view of our data processing pipeline.

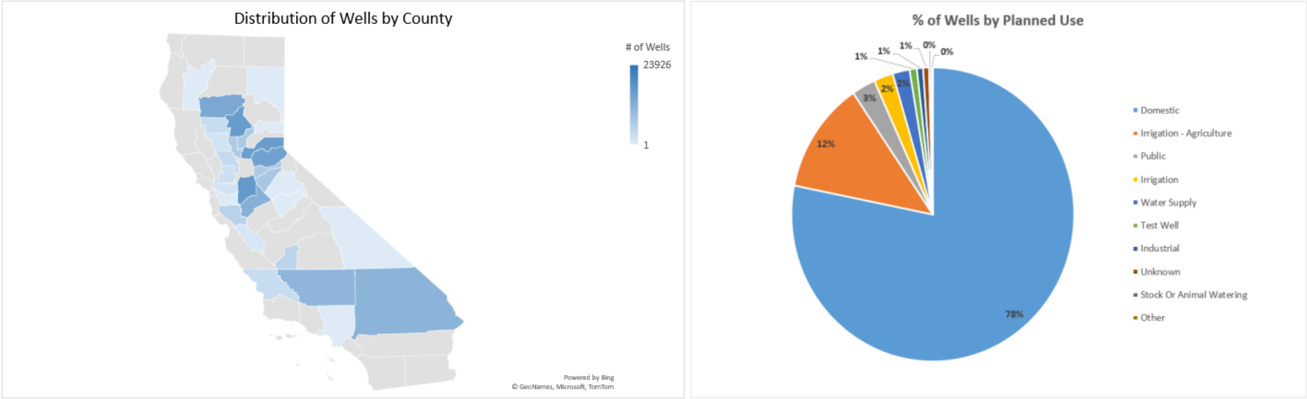


Figure 2: Distribution of wells by county in our dataset. Percentage of well for a particular usage

The second dataset that we used is the Ground Water Level (monthly) dataset (filename: 'gwl-monthly.csv') and the Ground Water Level stations (filename: 'gwl-stations.csv') dataset to gather the information regarding ground water levels measured at various stations. The stations file gives the corresponding identifiers for the ground water monitoring stations present in the monthly dataset. Figure 3 shows the groundwater level over the last decade in Sacramento County. It clearly depicts a continuous decline in the ground water level which has been exacerbated in the last five years.

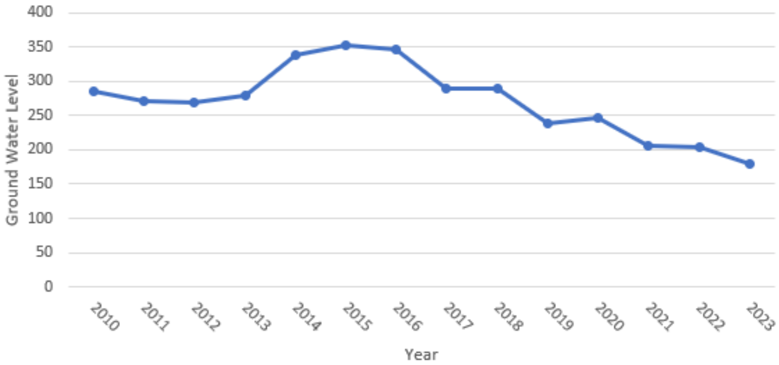


Figure 3: Groundwater levels in Sacramento County

3.2 Data Pre-Processing

Data pre-processing involved significant steps as the data was sparse and consisted of redundant information. First, we will discuss the processing steps for the well completion report dataset which will be used for displaying the relevant information on the web site. This file involved relatively less processing steps as compared to processing the dataset for the prediction model.
The first step after loading the raw dataset into a Spark dataframe involved dropping those records whose latitude and longitudinal information was not present as it is redundant since these wells will not be marked on the interactive map. Subsequently, columns with unnecessary or redundant information were dropped and post this we standardize the column headers. For the final step, we filled the NULL values in the records with relevant values and cast into appropriate data types.

Preprocessing the model training dataset required more significant steps. Firstly, the 'gwl-monthly' and 'gwl-station' files were loaded. Relevant columns were selected from both the files and the rest of the columns were dropped. An inner join was performed on the two files to get corresponding static information across all measurement stations. The output dataframe was grouped at 'COUNTY_NAME' and 'MSMT_DATE' level and the mean of continuous variable was taken. A cross join was performed between a list of unique counties and measurement dates. Subsequently, a left join was performed on this dataframe with the aggregated dataset to get a continuous timeseries across all counties. The dataset was then sorted by county name and measurement date in ascending order and filtered for relevant counties and timeframe. All these operations were performed in PySpark on our local system.

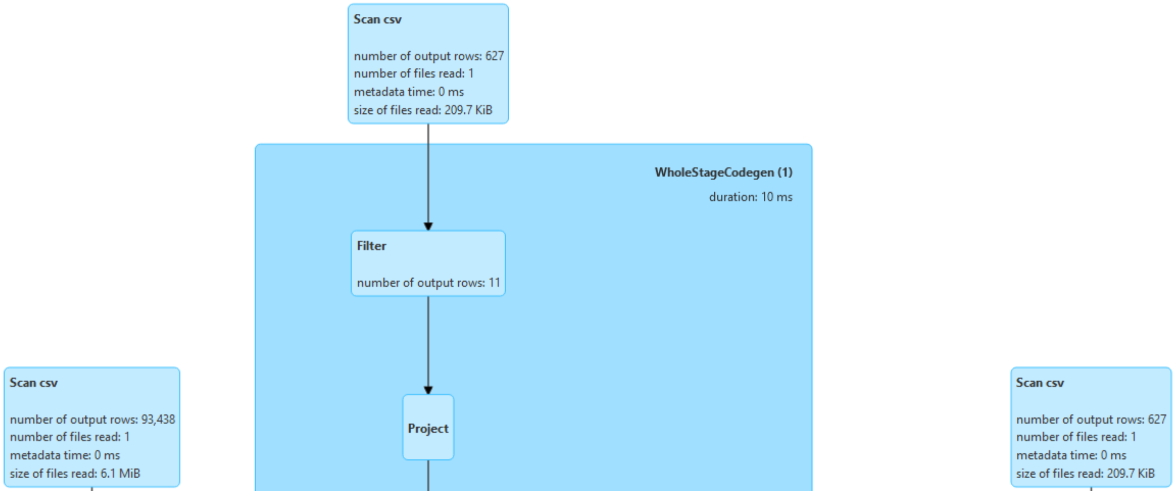


Figure 4: Snapshot of a DAG of our Spark data processing pipeline

3.3 Model Training and Prediction Framework

For our model training framework, we have decided to implement Temporal Fusion Transformers (TFTs) for long time series forecasting. TFT integrate attention mechanisms to capture both short and long-term temporal dependencies while addressing static and dynamic features. TFT combines static variable encoders, temporal variable encoders, and interpretable attention layers to highlight significant time-dependent patterns. It uses gating mechanisms to suppress irrelevant features and handles missing data effectively.

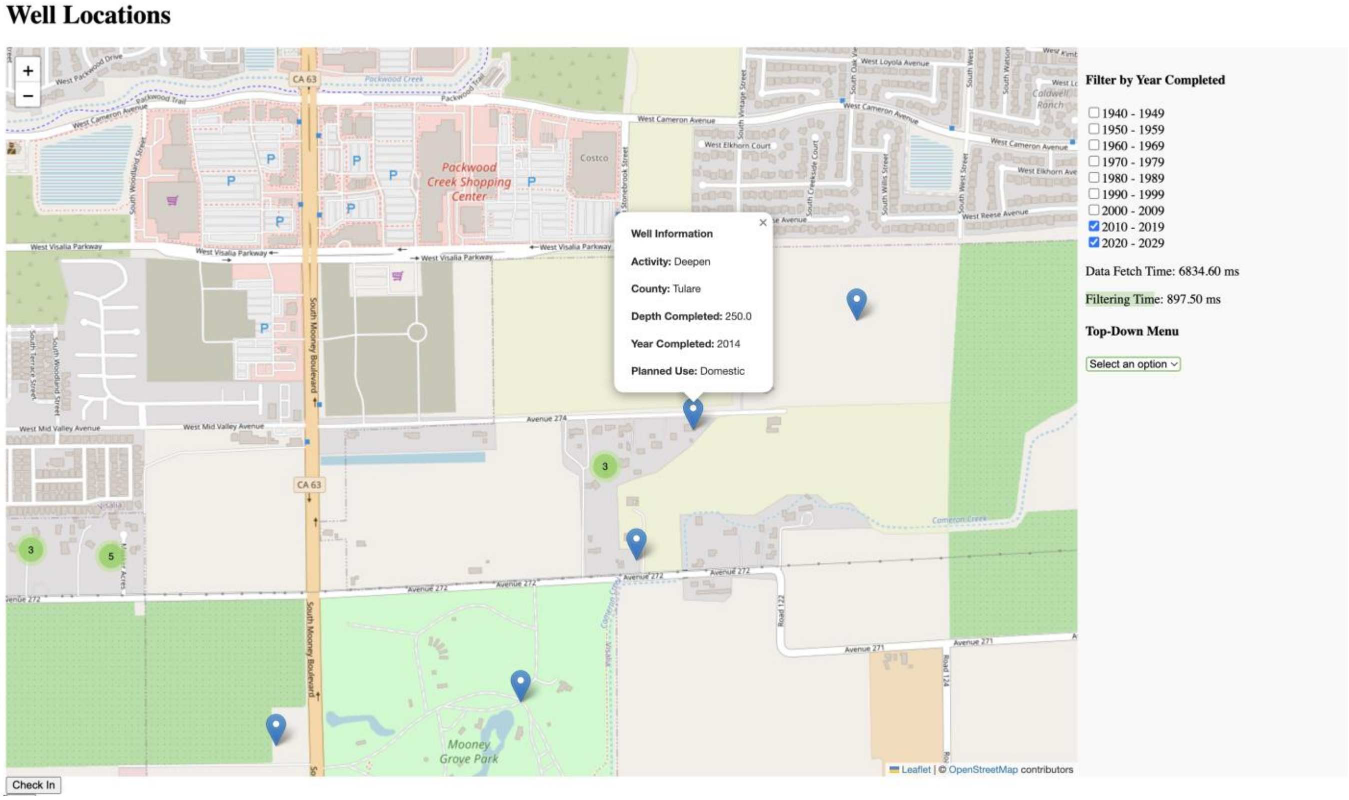
Our model is trained on the monthly ground water level data for 12 counties, for a time period starting from January 2010 to December 2023. The model is split into train-test split by splitting the last 12 months (January 2023- December 2023) to be used as the test set. To implement TFT, we used the Darts library [23] which has a collection of state-of-the-art timeseries forecasting algorithms. A very basic implementation of TFT was performed without much hyperparameter optimization as TFT is a heavy model to train on a desktop CPU. The number of hidden layers in the model was set to 64 and number of LSTM (Long Short-Term Memory) [24] layers was 1 along with 4 attention heads [25]. The batch size was set at 12, to correspond with the number of time steps in one year as TFT performs the training on temporally sequential batches.

Despite the lack of hyperparameter optimization and the model being relatively smaller, we saw that the TFT model's performance was satisfactory across multiple counties. These results are further discussed in Section 4 of this report. This showcases an opportunity to finetune such state of the art of models for long-term ground water prediction tasks to achieve much better and robust results.

3.4 Web Application

We developed a web application using the MERN stack (MongoDB, ExpressJS, React, and NodeJS) to visualize and interact with well data on a map interface. The application integrates a seamless data flow between the backend and frontend, enabling users to parse and display data from a CSV file, efficiently handle map markers, and view detailed information for each well. The project also addresses performance challenges to ensure a smooth user experience. Figure 5 shows a sample view of the landing page of our website.

The CSV file containing well data was parsed using the *papaparse* library, which efficiently extracts and organizes the data for rendering. To establish communication between the backend and frontend, we resolved CORS-related errors, ensuring secure and efficient data exchange across the application layers. For map visualization, we utilized the Leaflet framework to render interactive maps and place markers representing individual wells. Each marker is clickable, providing users with detailed information about the selected well in a user-friendly pop-up interface. Initially, the application experienced performance issues when loading a large number of markers, causing the webpage to freeze. To address this, we implemented marker clustering, significantly enhancing performance by grouping markers dynamically based on the zoom level. Additionally, we measured the latency from data upload to the rendering of markers on the map to ensure the system meets performance standards.



Well Locations

Well Information

Activity: Deepen

County: Tulare

Depth Completed: 250.0

Year Completed: 2014

Planned Use: Domestic

Filter by Year Completed

- □ 1940 - 1949
- □ 1950 - 1959
- □ 1960 - 1969
- □ 1970 - 1979
- □ 1980 - 1989
- □ 1990 - 1999
- □ 2000 - 2009
- ☑ 2010 - 2019
- ☑ 2020 - 2029

Data Fetch Time: 6834.60 ms

Filtering Time: 897.50 ms

Top-Down Menu

Select an option

Figure 5: A sample view of our web application

4 Evaluation and Results

The evaluation and results section is divided into two parts. The first part evaluates the performance of the big data application by measuring the run times of the data processing pipeline and comparing them with more traditional data manipulation methods such as Pandas in Python. The second part focuses on the prediction model and its accuracy for predicting ground water level for the next year.

4.1 Big-Data Tasks

Our big-data framework is implemented in PySpark, and we have measured its performance and compared it with that of traditional data processing frameworks such as Pandas on python. We have measured the execution time of the pipeline along with the memory usage and averaged out the results over three runs. Table 1 shows the time taken to execute the script in seconds and peak memory usage (in kB).

	PySpark (Web Data Processing)	Pandas (Web Data Processing)	PySpark (ML Data Processing)	Pandas (ML Data Processing)
Peak Memory Usage (kilobytes)	582.68	496,355.28	24,781.38	37,694.84
Execution Time (seconds)	18	11.37	81.02	15.6

Table 1: Memory Usage and Execution time comparison of PySpark and Pandas

Although it seems that the execution time of PySpark is significantly higher, we should note that it includes the time for the initialization of the Spark server. The evaluation clearly shows that PySpark is more memory efficient than Pandas and the execution time is also comparable if we exclude the time taken by the Spark server to initialize.

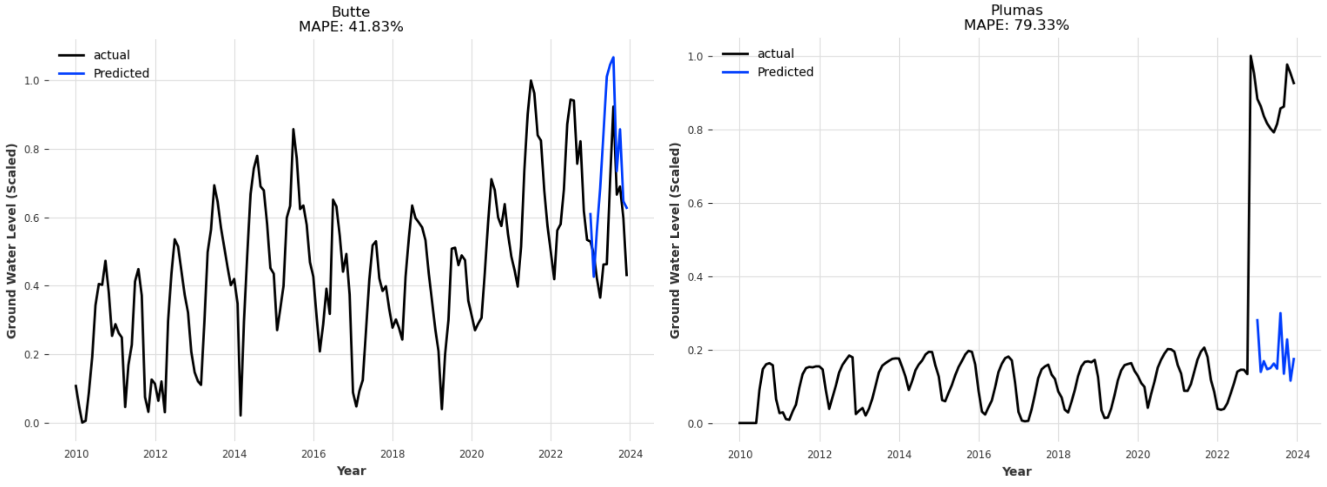
4.2 Prediction Model

Our implementation of the Temporal Fusion Transformer was relatively simple, as it did not involve complex static covariates, neither did we have any future covariates (apart from date) which can provide relevant information to the model for fine tuning. We were also bound in terms of how many model iterations we could run because even though our dataset is relatively small, we still had 132 timesteps for each county. The model was also trained on a desktop CPU without any acceleration hardware like a GPU to speed up model training. Despite these limitations, our model still delivers satisfactory results. Table 2 shows the MAPE and MSE values on the train and test set. The results are averaged over the 12 counties on which the model was trained.

	Train	**Test**
MAPE	16.14%	43.64%
MSE	0.09	0.015

Table 2: Model Results on train and test set.

The results might indicate that the model is slightly overfitted but that can be attributed to particular high noise in the dataset for some counties which skew the MAPE values. Figure 6(b) gives an example of this. Figure 6(a) shows the graph of predicted versus actual values for Butte County, which had a MAPE of 41.83%, which is close to our overall MAPE on the test set. The graph clearly shows that the model accurately predicted the rise and descent of the ground water level, although the values might not have been entirely correct.



(a) (b)

Figure 6: (a) Actual vs predicted graph of Butte County which shows that the model captures variations accurately. (b) Actual vs Predicted graph of Plumas County which showcases how a few counties can skew the results due to non-uniformity in the data

The results indicate that TFTs are very strong performers in such tasks, this opens up possibility for future work in finetuning and optimizing these models to predict ground water level. Also, with proper hardware, more complex and larger models can be trained which can yield better result.

5 Conclusion

The **WaterFinder** project successfully demonstrated the integration of Big Data management techniques, geospatial visualization, and machine learning for addressing the critical issue of groundwater management. Through the development of an interactive web application, we provided users with access to groundwater well data and predictions for future groundwater levels in a user-friendly format.

Our evaluation highlighted the efficiency of PySpark for handling and processing large datasets, showcasing its scalability and memory efficiency compared to traditional data processing frameworks like Pandas. Additionally, the Temporal Fusion Transformer (TFT) model, though implemented with minimal optimization, achieved promising results in forecasting groundwater levels, laying the groundwork for further fine-tuning and optimization in future iterations. Despite the project's achievements, certain limitations were encountered. For instance, the dataset's sparsity and variability across regions introduced challenges in achieving uniform prediction accuracy. Additionally, hardware constraints restricted our ability to fully leverage the capabilities of complex machine learning models. Addressing these challenges by employing more robust datasets and optimized hardware infrastructure would likely result in significant performance improvements.

Future work could explore incorporating additional features such as real-time data integration, advanced prediction models, and broader geographic scalability. Expanding the user base to include policymakers, researchers, and local communities can further enhance the impact of the platform by fostering informed decision-making and proactive water resource management.

In conclusion, this project underscores the importance of innovative technological solutions in addressing environmental challenges. The results and insights gained pave the way for future advancements in sustainable groundwater management, emphasizing the potential of interdisciplinary approaches combining data science, environmental studies, and geospatial analysis.

6 Author Contributions

BIBLIOGRAPHY

[2] Hanjra, Munir A., and M. Ejaz Qureshi. "Global water crisis and future food security in an era of climate change." Food policy 35, no. 5 (2010): 365-377

[3] USGS. U.S. Geological Survey: Attributed California Water Supply Well Completion Report Data for Selected Areas, Derived from CA WCR OSCWR Data (ver. 4.0, September 2024). Retrieved from https://catalog.data.gov/dataset/attributed-california-water-supply-well-completion-report-data-for-selected-areas-derived--784ef.

[4] Long, J.L., Viger, R.J., Raja, K., Enns, K.D., Sheflin, J.R., Ignizio, D.A.,2023, sciencebasepy: A Python library for programmatic interaction with the USGS ScienceBase platform (Version): U.S. Geological Survey software release, https://doi.org/10.5066/P9X4BIPR.

[5] OpenStreetMaps. Overpass API. Retrieved from http://overpass-api.de/.

[6] DWR. Department of Water Resources: Continuous Groundwater Level Measurements. Retrieved from https://catalog.data.gov/dataset/continuous-groundwater-level-measurements-5481b.

[7] Wada, Yoshihide, Ludovicus PH Van Beek, Cheryl M. Van Kempen, Josef WTM Reckman, Slavek Vasak, and Marc FP Bierkens. "Global depletion of groundwater resources." *Geophysical research letters* 37, no. 20 (2010).

[8] Siebert, Stefan, Jacob Burke, Jean-Marc Faures, Karen Frenken, Jippe Hoogeveen, Petra Döll, and Felix Theodor Portmann. "Groundwater use for irrigation–a global inventory." *Hydrology and earth system sciences* 14, no. 10 (2010): 1863-1880.

[9] Giordano, Mark. "Global groundwater? Issues and solutions." *Annual review of Environment and Resources* 34, no. 1 (2009): 153-178.

[10] Dittrich, Jens, and Jorge-Arnulfo Quiané-Ruiz. "Efficient big data processing in Hadoop MapReduce." *Proceedings of the VLDB Endowment* 5, no. 12 (2012): 2014-2015.

[11] Cai, Li, and Yangyong Zhu. "The challenges of data quality and data quality assessment in the big data era." *Data science journal* 14 (2015): 2-2.

[12] Apache Spark, PySpark API. Retrieved from https://spark.apache.org/docs/latest/api/python/index.html.

[13] MongoDB. Retrieved from https://www.mongodb.com.

[14] Gaffoor, Zaheed, Kevin Pietersen, Nebo Jovanovic, Antoine Bagula, and Thokozani Kanyerere. "Big data analytics and its role to support groundwater management in the southern African development community." *Water* 12, no. 10 (2020): 2796.

[15] Cheng, Hongxia, and Zhang Minghui. "Groundwater quality evaluation model based on multi-scale fuzzy comprehensive evaluation and big data analysis method." *Journal of Water and Climate Change* 12, no. 7 (2021): 2908-2919.

[16] Martínez-Santos, Pedro, and Philippe Renard. "Mapping groundwater potential through an ensemble of big data methods." *Groundwater* 58, no. 4 (2020): 583-597.

[17] Naghibi, Seyed Amir, Davood Davoodi Moghaddam, Bahareh Kalantar, Biswajeet Pradhan, and Ozgur Kisi. "A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping." *Journal of Hydrology* 548 (2017): 471-483.

[18] Lopez, Oliver, Kasper Johansen, Bruno Aragon, Ting Li, Rasmus Houborg, Yoann Malbeteau, Samer AlMashharawi et al. "Mapping groundwater abstractions from irrigated agriculture: big data, inverse modeling and a satellite-model fusion approach." *Hydrology and Earth System Sciences Discussions* 2020 (2020): 1-41.

[19] Volodymyr Agafonkin. Leaflet: an open-source JavaScript library for mobile-friendly interactive maps. Retrieved from https://www.leafletjs.com/index.html.

[20] Tao, Hai, Mohammed Majeed Hameed, Haydar Abdulameer Marhoon, Mohammad Zounemat-Kermani, Salim Heddam, Sungwon Kim, Sadeq Oleiwi Sulaiman et al. "Groundwater level prediction using machine learning models: A comprehensive review." *Neurocomputing* 489 (2022): 271-308.

[21] Daliakopoulos, Ioannis N., Paulin Coulibaly, and Ioannis K. Tsanis. "Groundwater level forecasting using artificial neural networks." Journal of hydrology 309, no. 1-4 (2005): 229-240.

[22] Lim, Bryan, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. "Temporal fusion transformers for interpretable multi-horizon time series forecasting." *International Journal of Forecasting* 37, no. 4 (2021): 1748-1764.

[23] Herzen, Julien, Francesco Lässig, Samuele Giuliano Piazzetta, Thomas Neuer, Léo Tafti, Guillaume Raille, Tomas Van Pottelbergh et al. "Darts: User-friendly modern machine learning for time series." *Journal of Machine Learning Research* 23, no. 124 (2022): 1-6.

[24] Hochreiter, S. "Long Short-term Memory." *Neural Computation MIT-Press* (1997).

[25] Vaswani, A. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017).