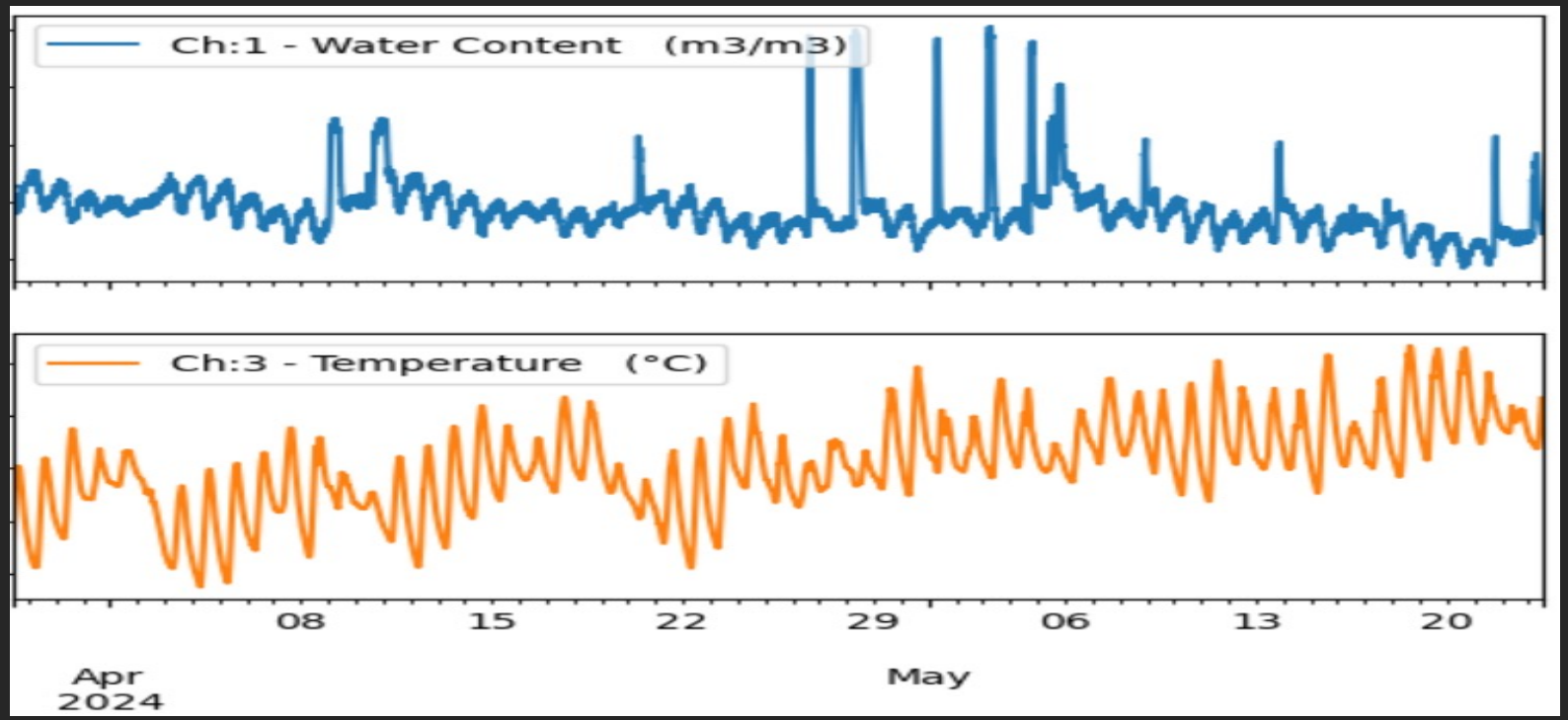
# Topic: High spatial and temporal resolution model of soil moisture and temperature time series using machine learning approach.




Advisor: Prof. Su
Research Scientist: Dr. Nieman

Intern Name: Anne Ekong
Date: October 25th, 2024.

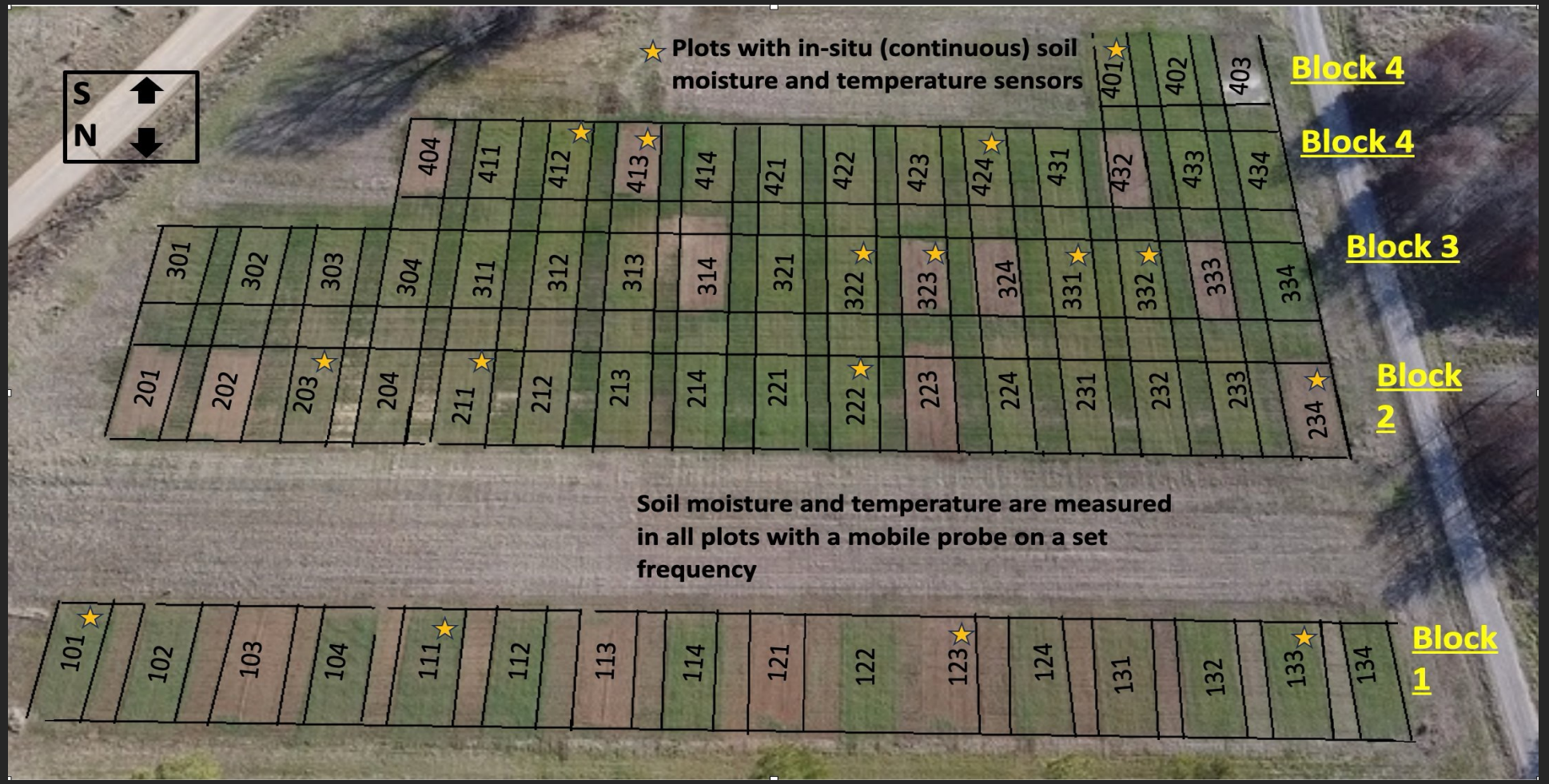
# content

- Introduction
- Data Collection/Data Preprocessing
- Machine Learning Approach
- Model Training
- Evaluation Metrics
- Result/Discussion
- Acknowledgements
- Questions & Answers

# Introduction

- Soil moisture and temperature are critically important factors affecting seed germination and monitoring these conditions are essential for understanding environmental dynamics.
- The in-situ soil temperature and soil moisture data loggers is used to capture continuous soil moisture and temperature time series data, located on 16 of 64 research site. Due to cost, it could be installed on the 64 research sites.
- The portable probe is used to collect additional data point on the research site, but they lack the temporal resolution needed to build continuous time series.
- The aim is to integrate the situ soil temperature and moisture data loggers with the portable probes using machine learning to estimate soil conditions in the unmonitored research sites.

An Aerial view of USDA-ARS Dale Bumpers Small Farms Research Center in Booneville, AR.



# Data Collection

- The in-situ soil moisture and temperature data loggers recorded a high-resolution temporal data from February 28th,2024 to June 5th, 2024.
- The portable probe recorded soil moisture and temperature with limited temporal resolution  from February 23rd,2024 to June 5th , 2024.
- A few of the data collected in various research site is shown below.

**Logger data**

Plot 211: 55040 observations

| Date-Time (CDT) | Ch:1 - Water Content (m3/m3) | Ch:3 - Temperature (°C) |
|---|---|---|
| 2024-02-28 08:40:28 | -0.056 | 17.200 |
| 2024-02-28 08:41:28 | -0.055 | 17.329 |
| 2024-02-28 08:42:28 | -0.060 | 17.372 |
| 2024-02-28 08:43:28 | -0.054 | 17.372 |
| 2024-02-28 08:44:28 | -0.055 | 17.415 |
| ... | ... | ... |
| 2024-06-05 08:06:09 | 0.412 | 23.507 |
| 2024-06-05 08:21:09 | 0.422 | 23.507 |
| 2024-06-05 08:36:09 | 0.423 | 23.464 |
| 2024-06-05 08:51:09 | 0.420 | 23.507 |
| 2024-06-05 09:06:09 | 0.413 | 23.635 |

55040 rows × 2 columns

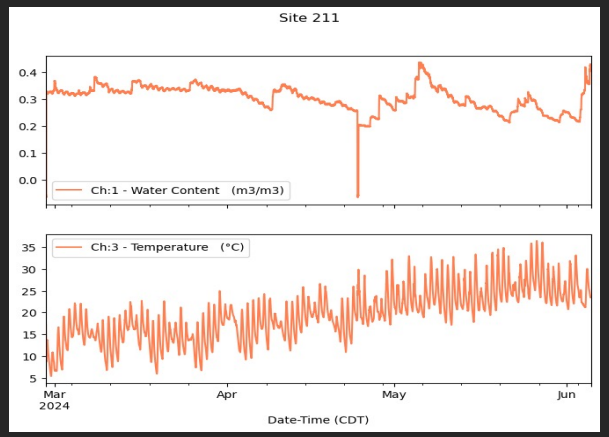
Logger data

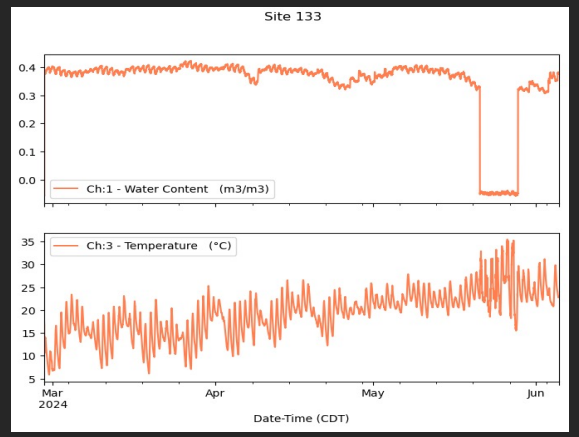
Plot 133: 48387 observations

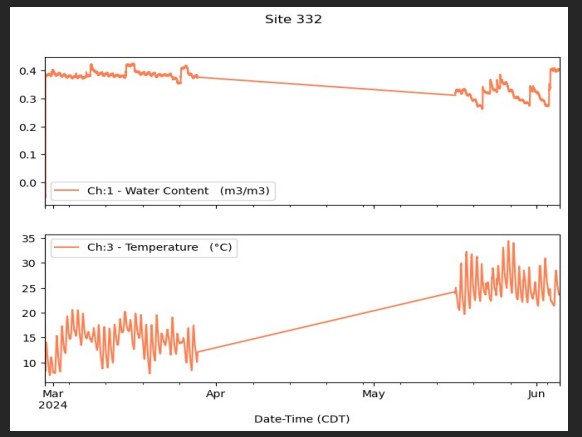
| Date-Time (CDT) | Ch:1 - Water Content (m3/m3) | Ch:3 - Temperature (°C) |
|---|---|---|
| 2024-02-28 08:34:47 | -0.054 | 17.458 |
| 2024-02-28 08:35:47 | -0.056 | 17.501 |
| 2024-02-28 08:36:47 | -0.057 | 17.543 |
| 2024-02-28 08:37:47 | -0.056 | 17.586 |
| 2024-02-28 08:38:47 | -0.058 | 17.586 |
| ... | ... | ... |
| 2024-06-05 07:57:46 | 0.380 | 22.820 |
| 2024-06-05 08:12:46 | 0.381 | 22.820 |
| 2024-06-05 08:27:46 | 0.380 | 22.820 |
| 2024-06-05 08:42:46 | 0.378 | 22.906 |
| 2024-06-05 08:57:46 | 0.378 | 22.949 |

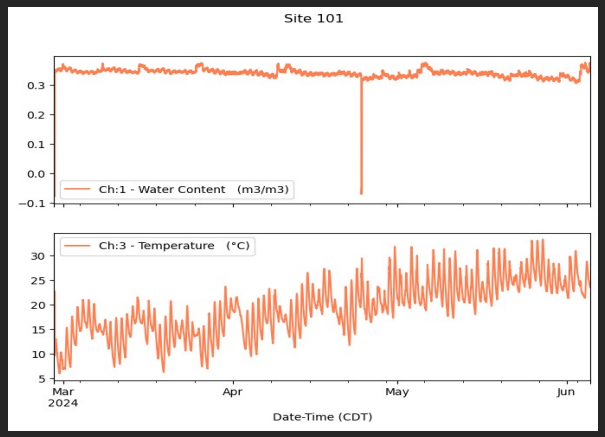
48387 rows × 2 columns

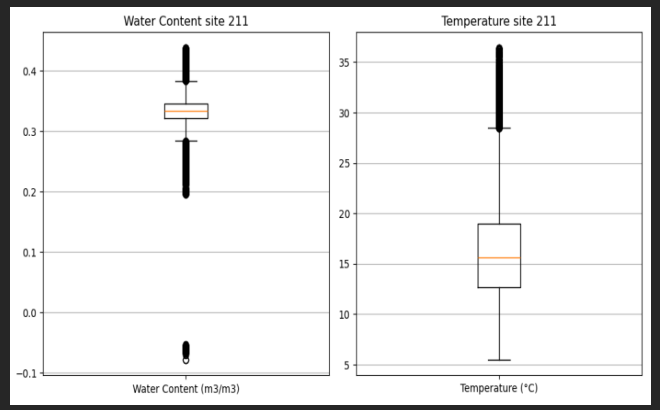
# Exploratory Data Analysis

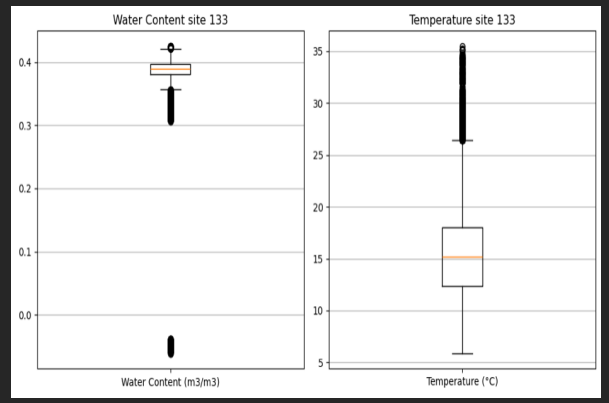


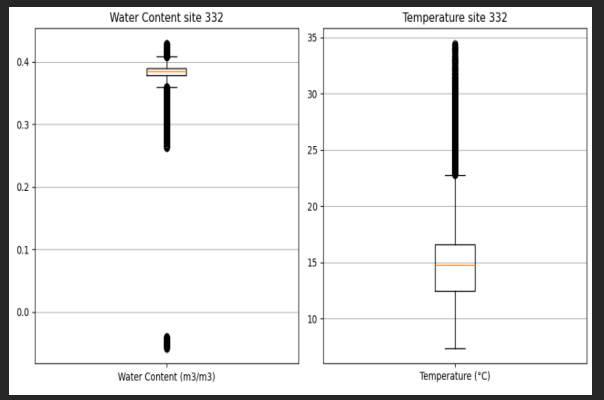



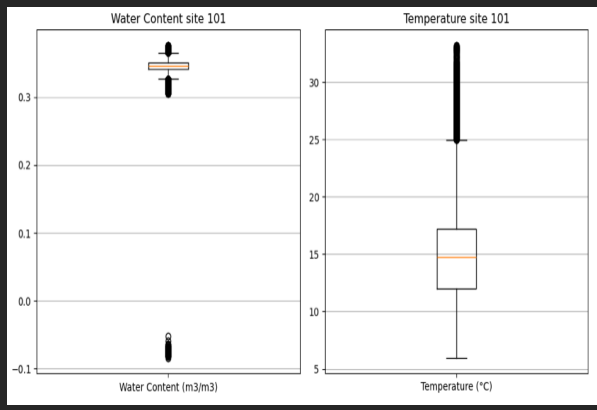



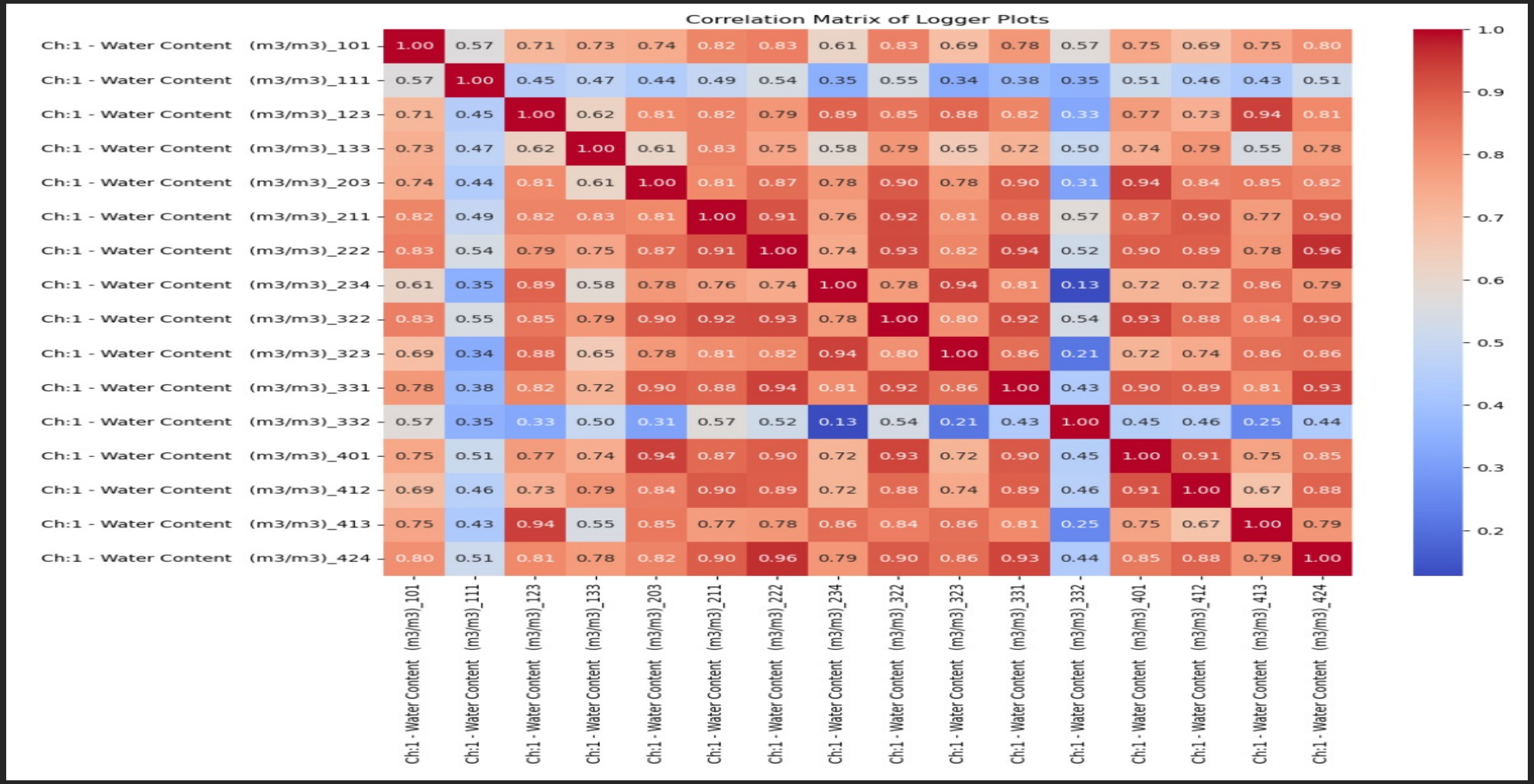



# Data Preprocessing

- The negative values of soil moisture/ outliers was replaced with NaN.
- Resampled the logger data to a minute frequency to align with the portable probe data.
- Ensured Date time column is in chronological order.
- Checked for missing values.
- Convert the datetime column to a format that  handles date time object.

# Correlation matrix/Heat Map



# Machine Learning Techniques for gap filling

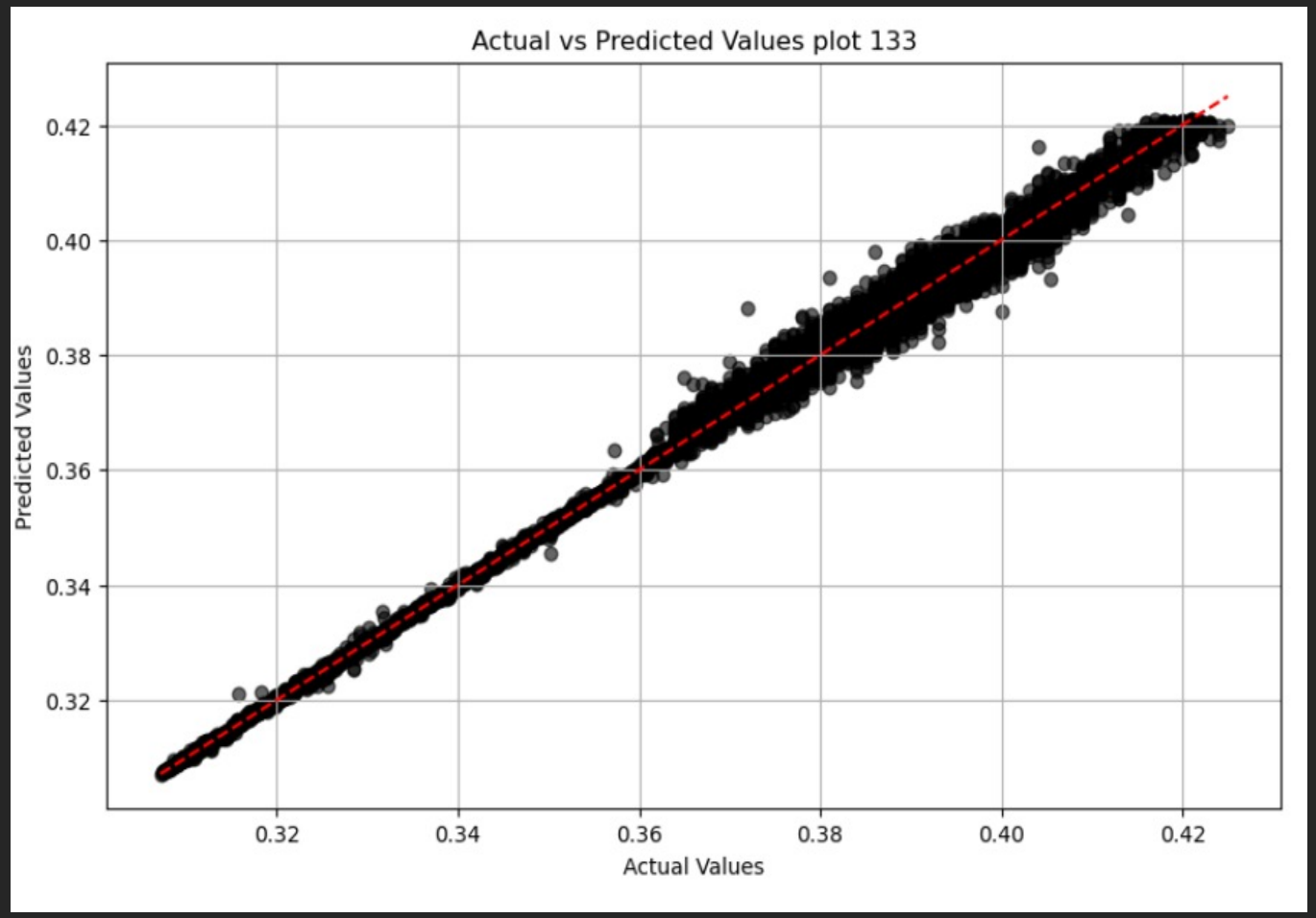
- Separate nan and non-nan rows.
- Select target variable and predictors
- Split the data into training and testing sets
- Train the RF /Lasso model
- Hyperparameter tuning using grid search cross validation.
- Evaluate the model performance
- Validate the model performance.

# Model performance metrics for continuous loggers

| Lasso Regression | | |
|---|---|---|
| Site | R2 | RMSE |
| 101 | 0.7093 | 0.00636 |
| 111 | 0.248 | 0.0105 |
| 123 | 0.907 | 0.022 |
| 133 | 0.6917 | 0.146 |
| 203 | 0.906 | 0.022 |
| 211 | 0.7093 | 0.0064 |
| 222 | 0.9503 | 0.0101 |
| 234 | 0.8974 | 0.02817 |
| 322 | 0.8972 | 0.0093 |
| 323 | 0.9160 | 0.0218 |
| 331 | 0.9278 | 0.0132 |
| 332 | 0.59 | 0.020 |
| 401 | 0.9345 | 0.0117 |
| 412 | 0.8503 | 0.0137 |
| 413 | 0.9075 | 0.0219 |
| 424 | 0.9357 | 0.0127 |

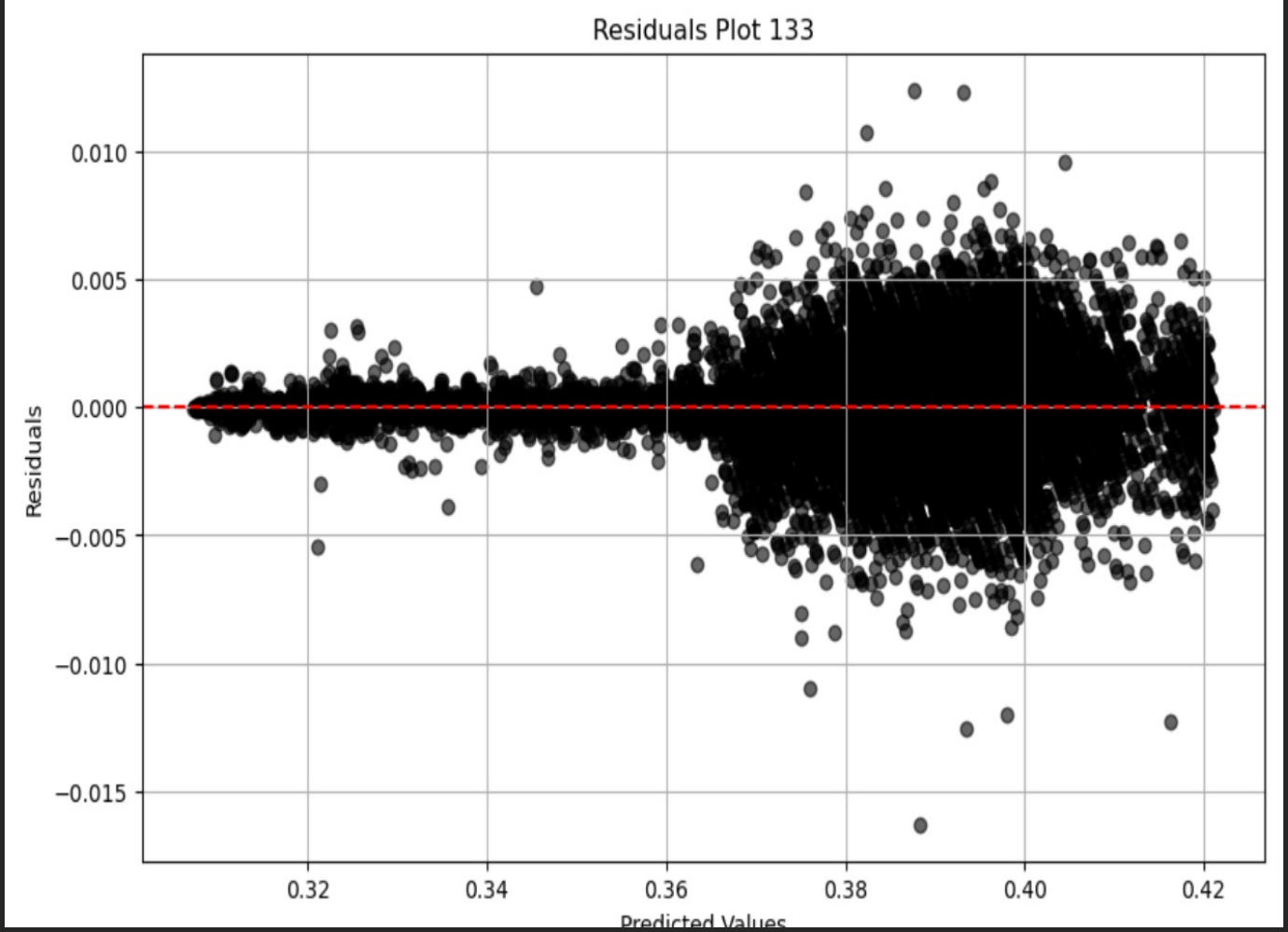
| Random Forest | | |
|---|---|---|
| site | R2 | RMSE |
| 101 | 0.98148 | 0.0016 |
| 111 | 0.98545 | 0.00146 |
| 123 | 0.9997 | 0.00129 |
| 133 | 0.99789 | 0.00121 |
| 203 | 0.9992 | 0.00119 |
| 211 | 0.9990 | 0.0013 |
| 222 | 0.9994 | 0.00107 |
| 234 | 0.9781 | 0.0130 |
| 322 | 0.9974 | 0.00149 |
| 323 | 0.9997 | 0.0013 |
| 331 | 0.9992 | 0.00138 |
| 332 | 0.9977 | 0.0015 |
| 401 | 0.9992 | 0.00126 |
| 412 | 0.9978 | 0.00168 |
| 413 | 0.9996 | 0.00138 |
| 424 | 0.9995 | 0.0011 |

# Validating metrics performance/Model interpretation




- **The point are clustered around the red dashed lines. this indicates that the model predictions are quite close to the actual values.**
- **There are some points that deviate from the line which represent prediction errors.**
- **The linear pattern also suggest the linear relationship that exist between the predictors and the target variable.**

# Validating metrics performance




- Points that lie far from the zero line indicates outliers as result of external factors like precipitation.
- The residuals is centered around zero, indicating the model is unbiased.
- The clustering in the upper predicted values indicate the data points.

```
logger_data
```

| | Date-Time (CDT) | Plot Number | Moisture_logger |
|---|---|---|---|
| **0** | 2024-02-28 08:27:00 | 101 | 0.338168 |
| **1** | 2024-02-28 08:28:00 | 101 | 0.338168 |
| **2** | 2024-02-28 08:29:00 | 101 | 0.338168 |
| **3** | 2024-02-28 08:30:00 | 101 | 0.338168 |
| **4** | 2024-02-28 08:31:00 | 101 | 0.338168 |
| **...** | ... | ... | ... |
| **2258923** | 2024-06-05 09:25:00 | 424 | 0.422867 |
| **2258924** | 2024-06-05 09:26:00 | 424 | 0.423000 |
| **2258925** | 2024-06-05 09:27:00 | 424 | 0.417507 |
| **2258926** | 2024-06-05 09:28:00 | 424 | 0.417507 |
| **2258927** | 2024-06-05 09:29:00 | 424 | 0.417507 |

2258928 rows × 3 columns

Portable probe data

| | DateTime | Plot Number | Moisture (m3/m3) | Temp (°F) | E.C. (M.S.) |
|---|---|---|---|---|---|
| **0** | 2024-02-23 07:52:00 | 134 | 0.429 | 51.0 | 0.62 |
| **1** | 2024-02-23 07:52:00 | 134 | 0.429 | 51.0 | 0.62 |
| **2** | 2024-02-23 07:53:00 | 132 | 0.421 | 50.0 | 0.53 |
| **3** | 2024-02-23 07:53:00 | 133 | 0.422 | 50.0 | 0.54 |
| **4** | 2024-02-23 07:53:00 | 132 | 0.421 | 50.0 | 0.53 |
| **...** | ... | ... | ... | ... | ... |
| **3771** | NaN | 423 | NaN | NaN | NaN |
| **3772** | NaN | 431 | NaN | NaN | NaN |
| **3773** | NaN | 432 | NaN | NaN | NaN |
| **3774** | NaN | 433 | NaN | NaN | NaN |
| **3775** | NaN | 434 | NaN | NaN | NaN |

3776 rows × 5 columns

# Data Preprocessing to gap fill the portable probe

- Use the complete logger data from the monitored plots and portable probe data as input features.
- Organize the logger and probe data to align time points.
- Ensured Date time is  in chronological order .
- Logger data from the monitored plots is aligned with portable probe data from unmonitored plots based on site numbers and time stamp.
- For time points where probe data is unavailable, the model will use the aligned time series data from the logger plots to predict soil moisture and temperature for the unmonitored plots

# Steps to merge the portable probe data and logger data

- Merges the portable probe data and logger data on Date-Time (CDT) and Plot Number.
- Trains a Random Forest model on rows where both Moisture (m3/m3) and Moisture_logger are present.
- Use the model to predict and fill in missing moisture values in the portable dataset.

# Gap filled soil moisture data

| | Date-Time (CDT) | Plot Number | Moisture (m3/m3) |
|---|---|---|---|
| **0** | 2024-02-23 07:52:00 | 134 | 0.429 |
| **1** | 2024-02-23 07:52:00 | 134 | 0.429 |
| **2** | 2024-02-23 07:53:00 | 132 | 0.421 |
| **3** | 2024-02-23 07:53:00 | 133 | 0.422 |
| **4** | 2024-02-23 07:53:00 | 132 | 0.421 |
| **...** | ... | ... | ... |
| **3376** | 2024-05-31 10:09:00 | 434 | 0.520 |
| **3377** | 2024-05-31 10:10:00 | 401 | 0.609 |
| **3378** | 2024-05-31 10:10:00 | 402 | 0.507 |
| **3577** | 2024-05-28 13:30:00 | 322 | 0.530 |
| **3775** | 2024-05-28 13:30:00 | 321 | 0.338 |

2864 rows × 3 columns

# Conclusion

- The in -situ logger,  hand -held portable soil moisture and temperature probe and machine learning approaches  has been used  create  time series data of soil moisture and temperature.
- The Random Forest algorithm appears to perform well for predicting soil moisture with a good prediction accuracy for the in- situ measurements which occurred because of  sensor malfunction.
- The low correlation that exist between covariates(in-situ measurements) indicate that in-situ may not be directly predictive of the portable soil moisture and temperature probe.
- Improve Model Performance by integrating external data and augmenting the data.

# Questions