*STATS\_517*

*12/11/18*

*Final Report*

DALYN MCCAULEY

Predicting Ground Heat Flux from Meteorological data & Vapor Pressure Gradient Error Predictions

by Dalyn McCauley

## Problem Statement, Motivation, Research Goals:

An essential component in the water balance is evapotranspiration (ET) which is the processes of evaporation from open water and bare soil and transpiration from vegetation happening simultaneously. ET can be difficult to measure directly, unlike other components like runoff and precipitation, and so it is often estimated by meteorological data. Net radiation, air temperature, and relative humidity are driving factors for ET, but unfortunately the resolution required for precise ET estimates are often beyond the feasibility of affordable sensing equipment. Of particular concern is the error of relative humidity sensors, as the vapor pressure gradient is a vital component of the ET equation. Additionally, the amount of energy available for evaporation is dependent on the difference between net radiation and the ground heat flux, which is the amount of energy absorbed by the earth. Ground heat flux can be a cumbersome quantity to measure because it requires the use of heat flux plates. Heat flux plates are simply a thermopile that can determine the heat flux by measuring the temperature difference from the bottom and top of the plate. The problem is, in order for the assumptions of the heat diffusion equation to hold true, the plate must be fully submerged in the soil. The soil just above the plate has heat holding capacity, and not counting for this can cause result in measurement errors. H eat flux plates are also expensive, and it would be cost prohibitive to equip every micro weather station with one.

I propose to analyze meteorological data and use machine learning models to predict reference ground heat flux as well as predict when the data from the humidity sensors is unreliable, which is when the vapor pressure gradient is smaller than the bounds of uncertainty of the humidity sensors. The goal of this analysis is to expediate the data pre-processing of meteorological data by flagging data that might be unreliable and ultimately develop an automated and low-cost approach to determining ET. If the algorithms designed in this project prove useful and robust, they will be incorporated into the programming of  Arduino based micro weather stations. These low energy low, cost stations can be placed in vineyards or orchards to obtain a high spatial-temporal resolution map of ET across the landscape in hopes of fine-tuning crop water demand and limiting consumptive use of water in agriculture.

## Data Source and Description:

Two different datasets were used in this project. Both datasets are from the same location and over the same time period, but they differ slightly in the sensor set up and duration of sampling. The first is the dataset that will be used for the ground flux estimation. This dataset consists of 12 variables and 17,188 samples, equating to a size of 206,256. This dataset was chosen because it contains actual ground heat flux measurements, a necessary variable for training the models to predict the target variable, ground flux. This dataset also differs in that it only covers 12 days of data, and is average over 1 minute sampling interval. These variable can be found in table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Measured/**  **Calculated** | **Data type** |
| TIMESTAMP | Date and time [YYYY:MM:DD hh:mm:ss] | Measured | datetime64[ns] |
| ApogeeWm2 | net radiation using quantum technology [W/m2] | Measured | float64 |
| HMP3m\_T\_avg | Averaged temperature at 2m [°C] | Measured | float64 |
| HMP3m\_RH\_avg | Averaged relative humidity at 2m [%] | Measured | float64 |
| Wspd | Horizontal windspeed from a sonic anemometer [m/s] | Measured | float64 |
| Wdir | Wind direction [degrees] | Measured | float64 |
| Potential | Soil matric potential [kPa] | Measured | float64 |
| Temp | Soil temp at 5cm [C] | Measured | float64 |
| Temp.1 | Soil temp at 15cm [C] | Measured | float64 |
| EC | Soil electrical conductivity [mS] | Measured | float64 |
| GrFlux\_Avg | Ground heat flux [W/m2] | Measured | float64 |

Table 1

The second dataset was used for the VPD error clustering. This dataset consists of nine variables, and 24,585 samples. Ten intermediate columns were created in order to calculate the reference ET, making the final dataset a 19 by 24585 matrix, equating to a size of 467,115. The variable descriptions and datatypes can be found in Table 2. There are two radiation measurements, one from an Apogee quantum sensor and one from a Q7 thermopile sensors. The two humidity and temperature measurements at different heights provide the temperature and vapor gradients that drive ET. The wind measurement is from a sonic anemometer and is used to determine the aerodynamic effects on ET so the vapor being carried away from the measurement control volume can be determined.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Measured/**  **Calculated** | **Data type** |
| TIMESTAMP | Date and time [YYYY:MM:DD hh:mm:ss] | Measured | datetime64[ns] |
| Q7corr | Corrected net radiation using thermopile technology [W/m2] | Measured | float64 |
| ApogeeWm2 | net radiation using quantum technology [W/m2] | Measured | float64 |
| HMP2m\_T\_avg | Averaged temperature at 2m [°C] | Measured | float64 |
| HMP5m\_T\_avg | Averaged temperature at 5m [°C] | Measured | float64 |
| HMP2m\_RH\_avg | Averaged relative humidity at 2m [%] | Measured | float64 |
| HMP5m\_RH\_avg | Averaged relative humidity at 5m [%] | Measured | float64 |
| Wspd | Horizontal windspeed from a sonic anemometer [m/s] | Measured | float64 |
| P | Air pressure [kPa] | Measured | float64 |
| air\_d | Air density [kg/m3] | Calculated | int64 |
| aero | Aerodynamic resistance [m/s] | Calculated | int64 |
| delta | Slope of the saturation vapor pressure -temperature curve [kPa/°C] | Calculated | int64 |
| e\_0 | Saturation vapor pressure [kPa] | Calculated | int64 |
| e\_a | Vapor pressure [kPa], | Calculated | int64 |
| gamma | Psychrometric constant [kPa/°C] | Calculated | int64 |
| Gflux | Ground heat flux [W/m2] | Calculated | int64 |
| VPG | Vapor pressure gradient [kPa], | Calculated | int64 |
| VPGError | Indicates when the vapor pressure gradient is too small to yield reliable ET estimate (0 when VPG > = 0.2, 1 when VPG < 0.2 \*error) | Calculated | int64 |
| ET\_ref | Calculated reference ET [mm/5min] | Calculated | int64 |

Table 2

## Literature Review and References:

Evapotranspiration can be measured in many different ways. I am using the energy balance approach in this project. The energy balance approach initially stemmed from H.L. Penman’s 1948 paper *Natural Evaporation from open water, bare soil and turf* whichdetails the assumptions and equations necessary to derive an equation for natural evaporation using only meteorological data. Penman defines two approaches, an aerodynamic approach and an energy balance approach. The only measured parameters necessary to calculate evaporation using Penman’s combination equation are mean air temperature, mean dewpoint (or relative humidity), mean wind velocity and mean duration of sunshine (radiation) (Penman, 1948) . Many assumptions are made throughout the derivation:

* Ideally restricted to a field after thorough wetting
* There is a zero-temperature gradient between Tsurface and Tair
* The changes in stored heat and heating of the test material surroundings is negligible over the period of several days

Monteith later revised the original Penman equation to build a fundamental equation for calculating reference evapotranspiration (“ASCE Manual 70 – Second Edition,” 2015). Reference ET represents the upper limit of ET given the available energy and meteorological measurements applied over a reference surface, often clipped grass or alfalfa. A crop coefficient factor is applied to the reference ET value to obtain the estimated ET from the vegetation of interest (R. G. Allen, Pereira, Howell, & Jensen, 2011).

## Preliminary EDA:

The data were preprocessed by applying a 1 minute (Dataset 1) or 5 minute (Dataset 2) moving average to the 10Hz samples. This moving average smooths the data and reduces the sensor noise. Additional preprocessing work was done in deriving the value for evapotranspiration. The calculation method is outlined by Allen et al in the FAO 56 irrigation manual (R. Allen, 1998). The intermediate calculations and variables for reference ET are shown and briefly described below. The aerodynamic resistance is a function of the measurement height and the roughness length, and the surface resistance is a function of the stomatal resistance the leaf, rl, and the amount of vegetation area, known as the Leaf Area Index (LAI).

Air density is calculated from the ideal gas law:

Saturation vapor pressure is the pressure at which water vapor in the air condenses and is a function of temperature:

The actual vapor pressure is the pressure from water vapor in air and is a function of the saturation vapor pressure and the humidity, q:

The vapor pressure gradient is the difference between saturation vapor pressure and actual vapor pressure:

The slope of the saturation vapor pressure – temperature curve is defines as Δ and is a function of temperature and saturation vapor pressure:

The psychrometric constant relates the partial pressure water in the air to air pressure, where Mw is the molecular weight of water and L is the latent heat of evaporation:

Finally, the available energy (R – G), vapor pressure gradient (VPG) and resistance terms are combined to calculate the reference evapotranspiration rate for the given meteorological conditions:

The results of the data preprocessing for the ground flux data are shown in the Figure 1. The image shows the daily cycle of radiation, ground heat flux and volumetric water content (VWC) over the course of 12 days during the summer. The places where the radiation data looks noisy indicates a cloudy day. A few selected processed data for the VPD error analysis are shown in Figure 2.

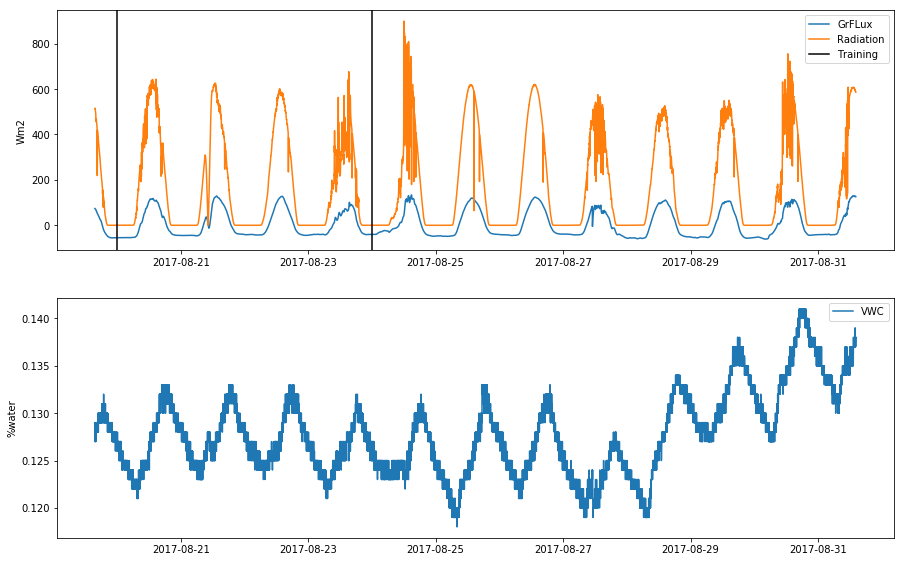
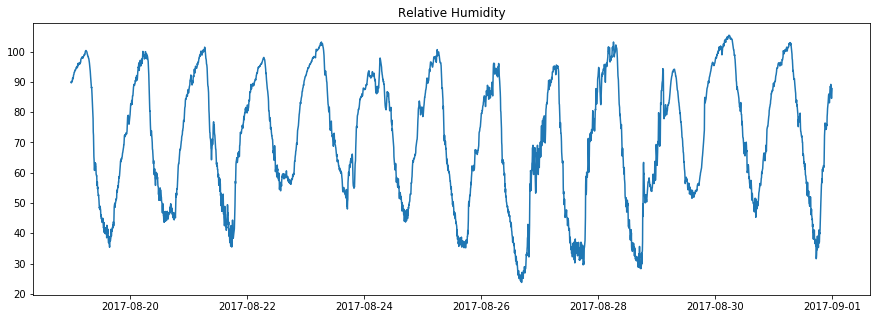
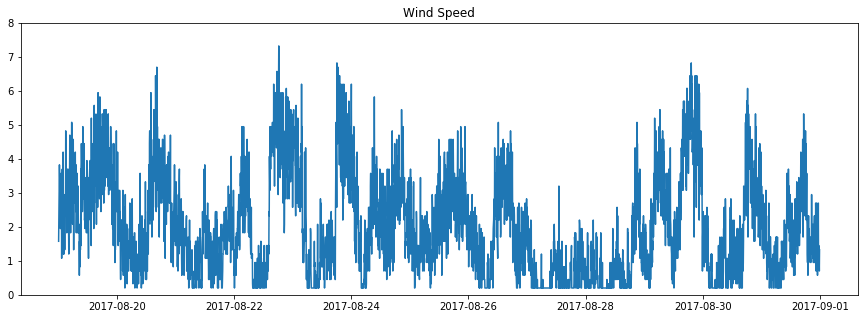
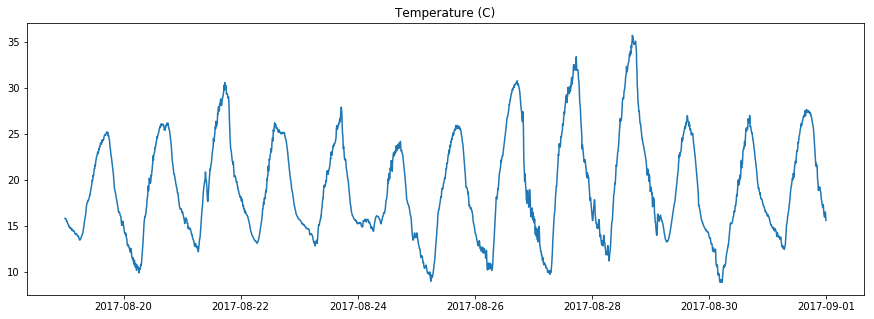


Figure 1







Figure

Modeling Process:

### Supervised Regression Results

Regression models were used to estimating ground flux. A training period of four days was used, from August 20, 2017 to August 24, 2018. This period was selected because it had representative ranges in weather conditions, such as cloudy and sunny days. The models were fitted using only the measured meteorological data, which did not include the calculated components. Figure 3 shows the results of three of the regression models, Linear Regression, Adaboost on Decision Tree and Random Forest.

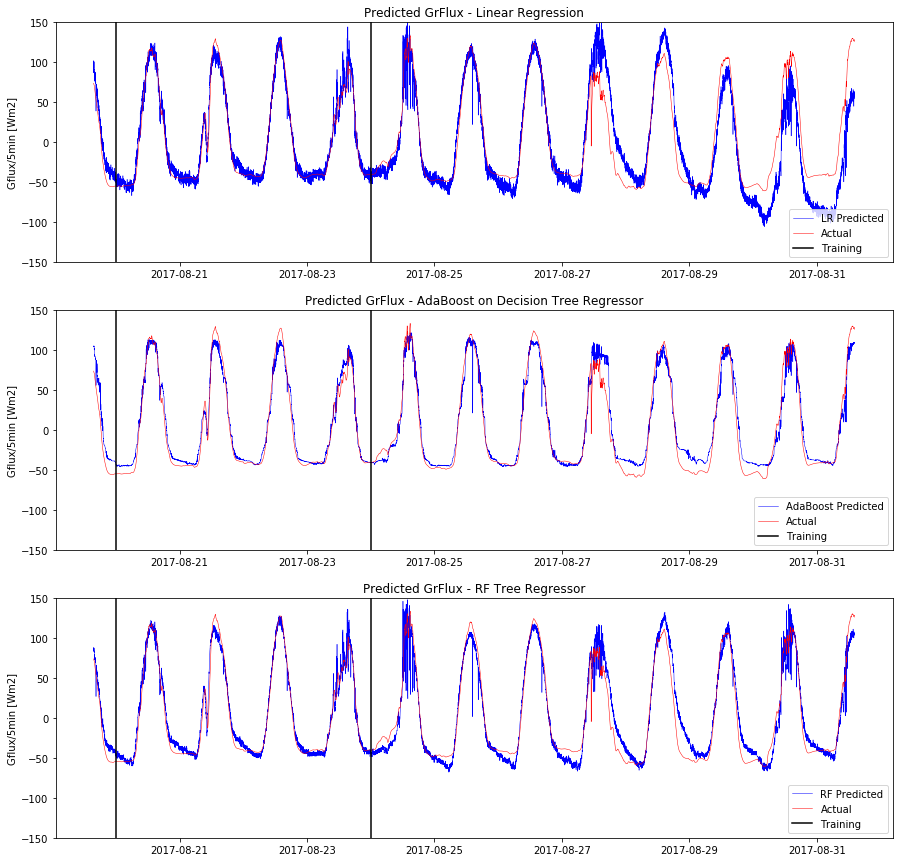


Figure 3

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Model Parameters | Train | Test |
| Linear Regression | - | 0.957 | 0.845 |
| Ridge Regression | Alpha = 10 | 0.952 | 0.945 |
| Lasso Regression | Alpha = 1.0 | 0.949 | 0.950 |
| Decision Tree Regressor | Depth = 4 | 0.962 | 0.952 |
| AdaBoost | Decision Tree | 0.973 | 0.942 |
| Random Forest Regressor | Depth = 4 | 0.962 | 0.952 |

Table 3

The decision tree regressor and random forest regressor performed the best, both with an R2 value of 0.952, Table 2. All models performed poorly at night and during cloudy days. This is most likely due to the lack of informative radiation data at these times, because as suggested by the feature importance model, radiation data is the most important variable in predicting ground flux, Figure 4.



Figure 4

The successful results of the regression models indicate that with a training period of four days the model would be able to take in meteorological data and accurately predict the ground heat flux at that location.

### Unsupervised Clustering Results

Predicting ground flux was not the only goal of this project, there was also some success in using unsupervised learning to classify when the data was unreliable due to the vapor pressure gradient. Three unsupervised models were used to classify whether or not the vapor pressure gradient is too small to give reliable data, K-means clustering, Hierarchal clustering and gaussian mixture model based clustering. The figure below shows a schematic of the instances where the VPD value should not be trusted because is falls within the measurement uncertainty, black horizontal line.

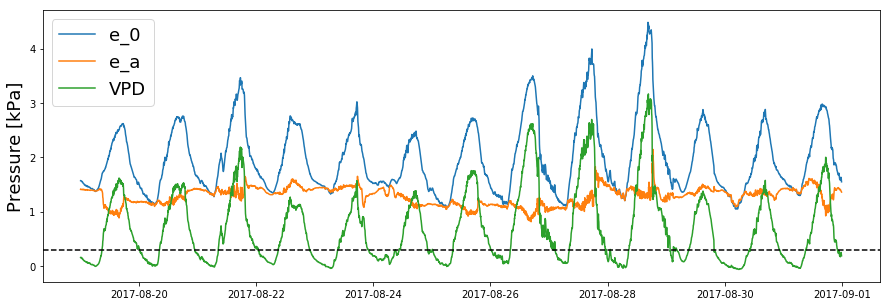


Figure 5

Principal component analysis was used first, and the first three principle components were kept, Figure 6. All three models were set to create two clusters, in hopes of classifying the data into two groups, good data and bad data.

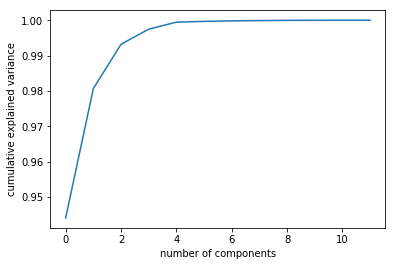


Figure 6

All three models performed very poorly. The best result was from k-means clustering where there was a 47% success rate in identifying the unreliable data, this of course means that it also falsely identified around 53% of the data. The code and results are shown in Figure 7 and 8, respectively.

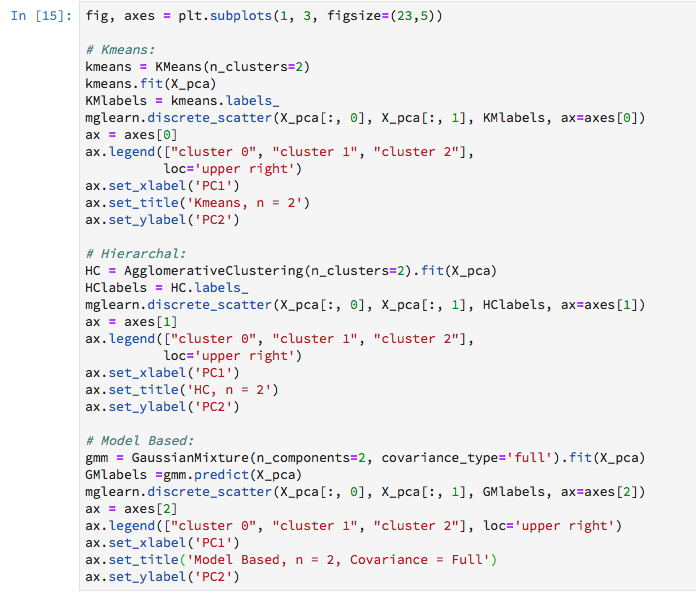
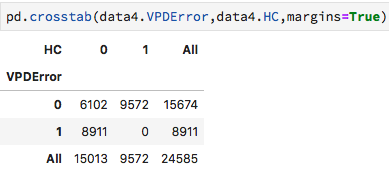
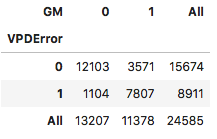
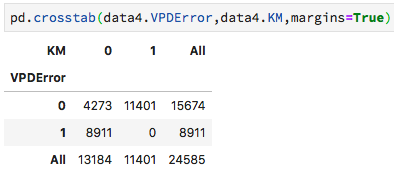


Figure 7



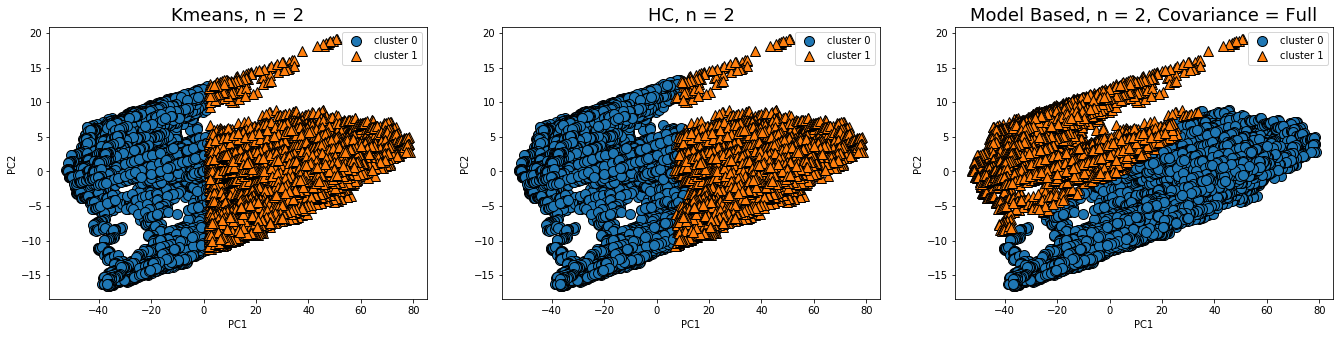


Figure 8

### Association Results

After clustering, association analysis was done on the dataset to come up with likely “itemsets”, or in this case, coupled meteorological events. Because the datasets are all continuous variables, new categorical variables were created. The categorical variables represent a range. The first set of categorical values are night, day, sunset and sunrise. A value of 1 was given to the sample if it fell into the category. The other categorical values are ranges of relative humidity , temperature and wind speed values, denoted low\_RH, mid\_RH, high\_RH, low\_T, mid\_T, high\_T, low\_W, mid\_W, high\_W. The naming should intuitively indicate the relative value of the data point. The mxltend apriori python library was used to create association rules. The minimum support was set to 0.1, and only itemsets of two or more items were recorded. Table 4 shows the association results.

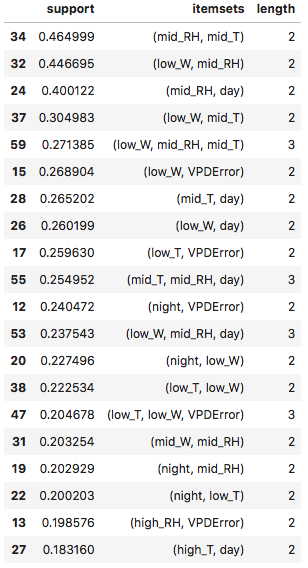


Table 4

The association rules were quite interesting. Valuable information can be learned from the itemsets containing VPD Error (15,17,12,47,13). For example, in the event of low wind speeds, (less than 3 m/s), there is a 26% chance a VPD error will occur. Even more interesting is that in the event the relative humidity is high, greater than 90%, there is about a 20% chance the VPD Error will occur. These are useful relationships to understand. I theorize that they could possibly be used to intelligently optimize a sensing system. Such as if a low wind or high humidity event is occurring, a control system could apply more voltage to the humidity sensor in an attempt to increase it’s measurement resolution. Or conversly, a farmer looking at his data manually in excel could pay extra attention to areas that are at high risk of measurement errors.

## Project Progress, Timeline, and Achievement:

The project progressed as expected. The initial data preprocessing and reference ET calculations took longer than expected, around two weeks. The supervised learning models were developed quite easily, but the unsupervised clustering has been challenging. The association models took some data manipulation that was time consuming. Once the data was successfully imputed, the association analysis was a breeze.

The significance of my work thus far is that I can successfully predict ground flux from a training period of only four days. This will be valuable when integrating the model with the micro Arduino based weather stations. This means that I only need to purchase and deploy one ground heat flux sensors, and can train the rest of the stations based on meteorological data that is simple, accurate, and easy to obtain.

## Conclusions and Possible Future Work:

In conclusion, the results of this project are promising. There is still a lot of room for improvements especially in the unsupervised modeling. In the future, it would be excellent to only have to purchase a few pyranometers as well. I wonder if I could replicate the training done with the ground flux data to predict radiation from only temperature, humidity and wind speed. If that estimate was accurate enough, I could then use the predicted radiation data to predict the ground flux, because it is clear the ground flux prediction models need a radiation input.

## References:

Allen, R. (1998). FAO Irrigation and Drainage Paper No. 56.

Allen, R. G., Pereira, L. S., Howell, T. A., & Jensen, M. E. (2011). Evapotranspiration information reporting: I. Factors governing measurement accuracy. *Agricultural Water Management*, *98*(6), 899–920. https://doi.org/10.1016/j.agwat.2010.12.015

ASCE Manual 70 – Second Edition: Evaporation, Evapotranspiration and Irrigation Requirements. (2015). In *2015 ASABE / IA Irrigation Symposium: Emerging Technologies for Sustainable Irrigation - A Tribute to the Career of Terry Howell, Sr. Conference Proceedings* (pp. 1–16). American Society of Agricultural and Biological Engineers. https://doi.org/10.13031/irrig.20152143358

Penman, H. L. (1948). Natural Evaporation from Open Water, Bare Soil, and Grass. *JSTOR*, 120–145.