Prediction of Evaporation using Linear Regression Algorithm: IT 586 Machine Learning

Mahederemariam Bayleyegn Dagne
Marymount University, 2807 North Glebe Rd, Virginia

Abstract

Evaporation is an important climactic factor that affects water availability in sectors such as irrigation and hydroelectric power. However, measurements of evaporation are not always readily available like other meteorological datasets. Due to this, several empirical models are developed to estimate evaporation using meteorological variables that are known to control evaporation. In this project, linear regression based machine learning algorithm is implemented to predict potential evaporation of a certain climatic area using meteorological variables temperature, humidity, radiation and wind. The model was trained and implemented using 42years of reanalysis data from ERA-5 in a single location in Ethiopia. The results showed that the selected variables are capable of predicting evaporation and the model was able to predict evaporation with very high skill. The model was later tested on locations in Kenya and United Kingdom. The results showed variable implications with relatively low predictions observed in Kenya while better scores are achieved in the United Kingdom. Based on this, it can be concluded the implemented machine learning is highly skillful but dependant on training data.

1 Introduction

Ethiopia is a developing nation known to be heavily reliant on rain-fed agriculture to sustain its economy. According to Mera (2018), rain availability in Ethiopia affects electricity, agricultural production, domestic and irrigation water use that cumulatively also affect the entire GDP of the country. According to the World Bank, the GDP of the country shows very clear correlation with rainfall availability.

In order to minimize the control of rainfall on several services and economy, authorities in Ethiopia resort to hydropower electric generation that rely on accumulated water over long periods of time. Fortunately, the country is known for its significant resources that allow for such developments. In fact van der Zwaan et al. (2018) project the country will be able to produce hydroelectric power north of 70TWH/year.

One of the most important climactic factors that affect the efficiency of hydropower generation is Evaporation. Especially in warm conditions, evaporation takes up significant amounts of water from hydroelectric reservoirs affecting the amount of water available for energy production. According to Scherer and Pfister (2016), the mean water foot print of hydroelectric reservoirs can reach upto 55 cubic meter per gigajoule of energy produced. Thus, Engineers who design hydropower structures in such conditions need up-to-date estimates of evaporation in the area.

Despite the need for evaporation datasets, it's not as easily available in most countries due to the lack of stations that provide fine-scale measurements. However, the physical basis of evaporation depends on climate factors such as temperature, radiation, humidity, and wind speed. Such variables are usually available at basic weather stations that are abundant in many places. Therefore, empirical methods such as the Penman-Monteith equation are used to estimate evaporation based on such parameters. This is a common practice in fields such as agriculture, which estimate Evapotranspiration using such equations as prescribed by FAO.

In this project, some of the common meteorological variables are used to predict evaporation. Being located in a tropical region, Ethiopia is exposed to high evaporation demand due to warm climate conditions. This makes the availability of evaporation data in the area proposed for dam building necessary. Accordingly, the model was

trained and tested in an arbitrary small area located in Ethiopia. The model was also later tested if it can be used in other regions.

2 Data and Methods

In this project, a statistical machine learning model was developed to predict evaporation in a 25kmx25km area in Ethiopia using net radiation, wind speed, dew point temperature, and air temperature. Dew point temperature is used as a measure of humidity in the atmosphere.

The data used for this project is the ERA 5 reanalysis data that provides climate data produced using observation-corrected models at a 25kmx 25km grid (Hersbach et al., 2020). The ERA 5 dataset consists of data from 1979 until the present and is consistently updated. ERA5 contains near-surface temperature, net surface radiation, wind speed at 10m height, and dew point temperature as required by the proposed model as predictors of evaporation. The data also contains evaporation datasets for the same years which are used to train and validate the model.

Initially, all datasets were acquired from the Copernicus website that contains all ERA-5 datasets. Then, since the last three months of ERA-5 data is not evaluated, the months of 2022 were dropped to only use timeseries from 1979-2021. Then an arbitrary grid point was selected so that the input data we have only consists of a timeseries. Based on this, first the predicting ability of the proposed variables was assessed by performing a correlation test. All variables showed high correlation with evaporation with very low p-test scores except for wind speed that showed a p-test score of around 0.3

Once the importance of all variables was established, the data was trained by using 70% of the time series. The remaining 30% was used to test the prediction capabilities of the model. The model trained in Ethiopia was also later used to predict evaporation in arbitrary areas located in Kenya and United Kingdom. This was done to test the possible universal use of a machine learning model trained using limited available data.

The evaluation of the prediction was done by using different metrics. The first one is by simply plotting the predicted data against the observed data and comparing the correlation. The second metric is the Mean Squared Error (MSE) that is commonly used to assess bias. Finally, the Nash-Sutcliffe efficiency (NSE) that scores the models ability to minimize bias and replicate the observed variability of data. The equations of MSE and NSE are given by equations 1 and 2 respectively where N represents number of data points M and O represent modelled and observed data and indices i and m represent single time and time mean data at each point.

$$MSE = \frac{\sum_{i=1}^{N}(M_i - O_i)^2}{N} \tag{1}$$

$$NSE = \frac{\sum_{i=1}^{N}(M_i - O_i)^2}{(O_i - O_m)^2} \tag{2}$$

3 Results and Discussions

Once the data was split into 70% training and 30% test data, linear regression test is performed on the training data against the target data. The result of this test is presented in Figure 1. The figure shows that there is very high correlation between the proposed data with R squared value of 0.999. The p-test indicates that all input variables correlate with the target data at a significant level except for wind speed. Since all variables correlated very well in accordance with the underlying physical science, all the variables are retained to build the model further.

After training the model using the training data, the model was used to predict evaporation for the rest of the timesteps. The predicted data was plotted against the observed data (figure 2). As can be seen in the plots, the modelled data follows the observed data at a very high correlation value. The other metrics also show very high prediction skills with MSE at the order of 10e-10 and NSE around 0.95 out of 1. This shows that the model is highly capable of predicting evaporation in the study area based on the selected meteorological variables. The full time series of the model output and the observed data are shown in figure **??**.

The model that was trained using data located in Ethiopia was applied to arbitrary locations in Kenya and the United Kingdom. The results

OLS Regression Results

Dep. Variable:	pev	R-squared (uncentered):	0.999
Model:	OLS	Adj. R-squared (uncentered):	0.999
Method:	Least Squares	F-statistic:	1.097e+05
Date:	Thu, 28 Apr 2022	Prob (F-statistic):	0.00
Time:	18:58:34	Log-Likelihood:	3459.6
No. Observations:	361	AIC:	-6911.
Df Residuals:	357	BIC:	-6896.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
u10	-6.114e-07	6.13e-07	-0.997	0.320	-1.82e-06	5.95e-07
d2m	7.165e-06	3.35e-07	21.410	0.000	6.51e-06	7.82e-06
t2m	-7.495e-06	3.41e-07	-21.975	0.000	-8.17e-06	-6.82e-06
ssr	-1.841e-10	5.93e-12	-31.054	0.000	-1.96e-10	-1.72e-10

Omnibus:	3.064	Durbin-Watson:	1.859
Prob(Omnibus):	0.216	Jarque-Bera (JB):	3.143
Skew:	0.216	Prob(JB):	0.208
Kurtosis:	2.848	Cond. No.	1.41e+06

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 1.41e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 1: Linear regression test between input data variables temperature (t2m), dew point temperature(d2m), net surface radiation (ssr) , wind speed (u10) and target data evaporation.

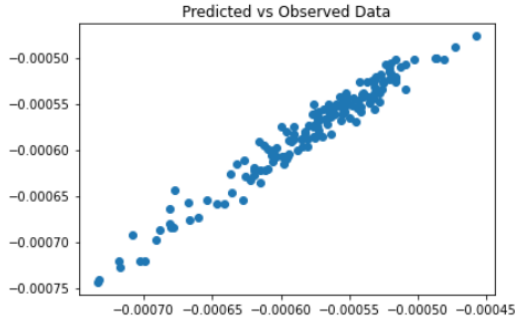

Figure 2: Modelled versus observed evaporation data in Ethiopia.

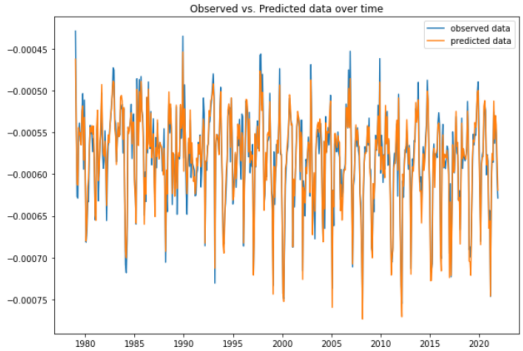

Figure 3: Time series of modelled and observed evaporation data in Ethiopia.

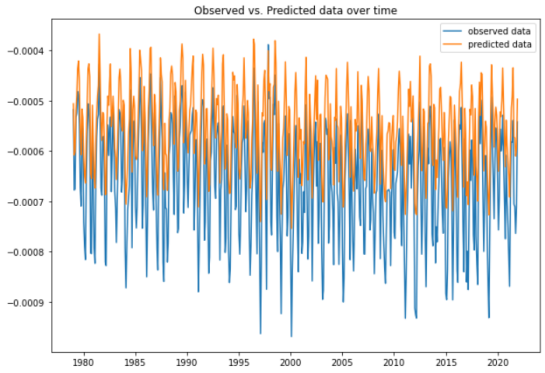

Figure 4: Time series of modelled and observed evaporation data in Kenya.

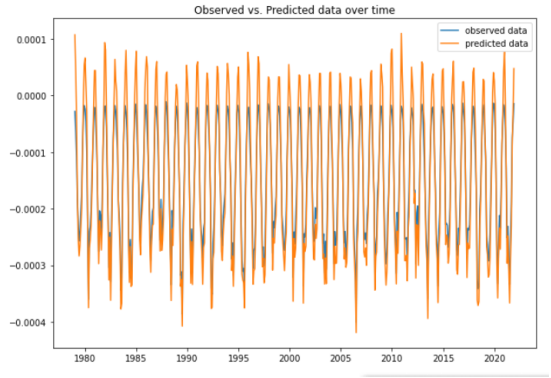

Figure 5: Time series of modelled and observed evaporation data in UK.

indicate variable skill with NSE values of 0.03 in Kenya and 0.84 in UK. However, the MSE metrics has stayed low in both cases with values swing in the range of 10e-8 and 10e-9. The full time series of the modelled and observed datas in Kenya and UK are shown in figures 4 and 5. It can be seen that the model underestimates evaporation in Kenya with most of its values showing a negative bias. On the other hand, the model fails to capture extreme values of evaporation (both low and high) and stays in the medium range of the observed data in the UK.

This indicates that the model may not be fully compatible to predict evaporation at different location from which it was trained in. However, the fact that the model bias as indicated by MSE stays very low in all the regions implies that such models can be used to get a general estimate of evaporation to get a sense of the orders of magnitudes in which it normally stays.

4 Conclusions

In this project, a linear regression model was developed in Ethiopia to predict evaporation based on meteorological variables. This is aimed to fill the gap of evaporation data that is needed to assess water availability for sectors such as hydroelectric power generation. The results yielded indicate that the developed model is very well capable of predicting evaporation with very small bias and excellent replication of data variability. Application of the model developed in Ethiopia in different locations (Kenya and UK) both near and far from Ethiopia show varying skill. The model tends to sometimes underestimate and other times fail to identify extremes. However, the model bias is generally very low. Based on this, it can be concluded that such a model can be used to get a general gist of what evaporation would look like based on other meteorological datasets.

5 References

References

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803.

Getachew Mera. 2018. Drought and its impacts in Ethiopia. *Weather and Climate Extremes*, 22.

Laura Scherer and Stephan Pfister. 2016. Global water footprint assessment of hydropower. *Renewable Energy*, 99:711–720.

Bob van der Zwaan, Agnese Boccalon, and Francesco Dalla Longa. 2018. Prospects for hydropower in Ethiopia: An energy-water nexus analysis. *Energy Strategy Reviews*, 19:19–30.