Predicting proglacial lake discharge using machine learning algorithms

Faye Perchanok[1]

Abstract

Greenland is home to some of the most valuable climate change information. The ice sheet and surrounding watershed systems provide insight into the future of sediment transport, discharge, and the consequences of a rapid increase in melting. Predicting discharge from the outflow of a pro-glacial river could provide further and valuable comprehension into river dynamics in a climatically important area. Using Python to train and test a yearly discharge dataset using regression and a decision tree predicted a steady annual increase in discharge, consequently elevating sediment transport.

1. Introduction

Sediment Transport in proglacial rivers provides vital information for predicting morphodynamics. Greenland glaciers, ice sheet, and linked watersheds hold some of the most important information on sediment transport properties. Proglacial streams, such as those found at the Greenland Ice Sheet (GIS), have distinct characteristics that affect data collection, measurements, analysis, and predictions. Firstly, their sediment transport flows have a stronger link to air temperature than to precipitation events, which favours predictability (Mao et al. 2018). Alternatively, due to the high turbidity and bedload transport, the tracking of sediment transport is only possible in a short summer period where melt flows are inhibited by lower temperatures and runoff is groundwater-dominated(Mao et al. 2018).

Transport dynamics in proglacial rivers are complex and hitherto , are not fully understood. Understanding and predicting river discharge is a parameter that could be particularly useful in furthering sediment transport research, as discharge is measureable in the Greenland proglacial area, and is a well-developed aspect of hydrology with many applications. It is expected that the sediment transport from the GIS will be accelerated due to climate change, heavily altering global oceanic sediment configuration and impact. Discharge data will provide valuable information regarding ice sheet surface mass balance, hydrology, and sediment release (Noel et al. 2018).

My masters thesis will focus on dating sediment transport along a proglacial river in Greenland to gain insight into its transport dynamics. The dataset used contains yearly discharge along Watson River (Qinnguata Kuussua). Machine learning algorithms were used to create forecasting models to simulate future discharge by observing trends and using linear regression and decision trees.

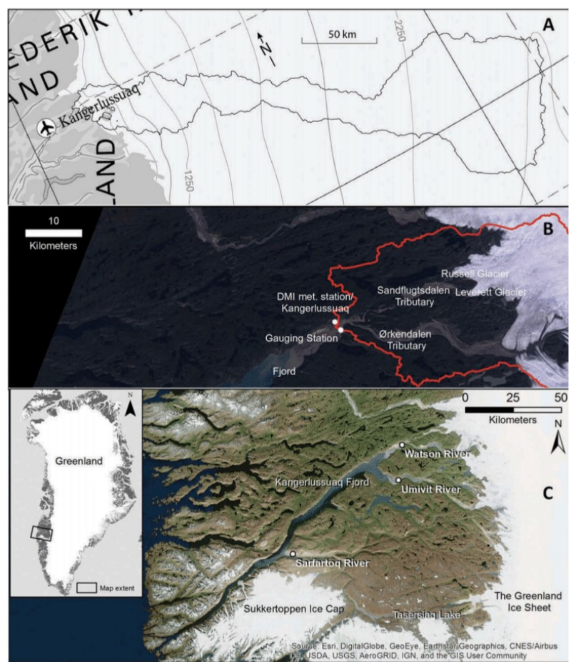

Figure 1. (A) Greenland Ice Sheet (GrIS) catchment after Lindbäck et al. (2015) (Hasholt *et al*, 2018);(B) proglacial area (Hasholt *et al*, 2018); and (C) Watson River and the fjord Kangerlussuaq (Hasholt *et al*, 2018).

2. Data

Beginning in 2006, a hydrometric station has operated in the settlement of Kangerlussuaq, located on the Watson River in southern west Greenland (Noel et al. 2018). The hydro-

[1]University of Lausanne, Lausanne, Switzerland.

metric station collects water stage measurements, which are converted into hourly discharge (Hasholt et al.2018). The station was firstly established by the University of Copenhagen, Department of Geosciences and Natural Resource Management, and was taken over by the Geological Survey of Denmark and Greenland in 2013, who have continued the monitoring as part of the Programme for Monitoring of the GIS (Hasholt et al. 2018).

This data set was chosen because of the range of dates available with the discharge. The years 1949 - 2021 have recorded or calculated discharge, which is the biggest time frame available for discharge in this river.

3. Methods

To compare different machine learning algorithms, linear regression and a decision tree were used to compare their results for predicitng future discharge of Watson River. The first step in creating the code for this project was to organize the data in a table in python and replacing all the empty values with NaN values. Then, the years with NaN values were removed to "clean" up the data.

Linear Regression is based on supervised learning to perform a regression task (Maulud and Abdulazeez, 2020). The model targets a specific prediction value based on independent variables (Maulud and Abdulazeez, 2020).In this project, it is used as a forecasting tool.Linear regression was firstly used to model the data points and create a regression line. The data set was split into a test and training set. The years 1949 - 1999 were chosen as training years, and 2000 - 2021 were chosen as test years, which is approximately 28% data in the test set and 72% data in the training set. The training set was trained and plotted, with a set of predictions produced from the linear space array.

A decision tree was then created for this data set, to train and adjust hyperparameters. A decision tree is a form of supervised learning wherein predictions can be done based on a previous data set. This was done as a comparison tool with the linear regression. The decision tree was created by scattering all of the data, then the training and test data, which was split using a test size of 20%, number of samples at 1000, random state set to 42, and noise at 0.4.

Finally, the mean squared error (MSE) and the root mean square error (RMSE) were calculated to show the accuracy of the results. This was done using the "r2_score" and "mean_squared_error" metrics from sklearn.

4. Results

What was observed with this regression prediction, was a slight, and consistent overall increase with time. Therefore, predictions for the future have increased values from the training and testing set. Conversely, the data was very noisy and inconsistent, and therefore the predictions found seem to under fit the data and are perhaps not an appropriate gauge of future discharge.

Train Error	Value
RMSE	0.48
MSE	0.13

Table 1. Root mean squared error (RMSE) and mean squared error (MSE) from the decision tree training data

Test Error	Value
RMSE	0.51
MSE	0.12

Table 2. RMSE and MSE from the decision tree testing data

Figure displays the 3 graphs produced using linear regression. The first graph is all of the data plotted, where an increasing trend with the data is observed, although the data points are quite scattered. The second graph shows the training set, from the years 1949-1999, and a regression line from 2020-2040 plotted to view how it was trained. The last plot shows the test set and the regression line produced with the data from 2000-2020. The data is quite scattered and the regression line is quite under fitted, however it still displays the broad trend of general increase.

	Year	Discharge_(km3)	Uncertainty_(km3)	Data_origin	Discharge
0	1949	4.14	1.16	3	4.14
1	1950	4.42	1.18	3	4.42
3	1952	4.57	1.20	3	4.57
4	1953	5.02	1.24	3	5.02
5	1954	3.49	1.10	3	3.49

Figure 2. Sample of data set, with yearly discharge, uncertainty, and location of collection (as a number). Discharge data in the last column is the cleaned up data, with NaN values removed.

Table 1 shows the root mean squared error (RMSE) and the mean squared error (MSE) of the training set. The root mean squared error:

$$\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(\frac{x_i - y_i}{\sigma_i}\Big)^2}$$

,is the proportion of the variance in the dependent variable that is predictable from the independent variable as a percentage. As it is calculated to be 48%, this shows heavy variance, with 100% being no variance between the variables. The mean square error is calculated as the average of the square of the errors:

$$\sum_{i=1}^{D}(x_i - y_i)^2$$

,a larger number indicating a higher error. A 13% error in the training set is a low number, indicating less error. Table 2 shows the RMSE and MSE for the test data, with similar findings of a low RMSE (high variance) and low MSE (low error).

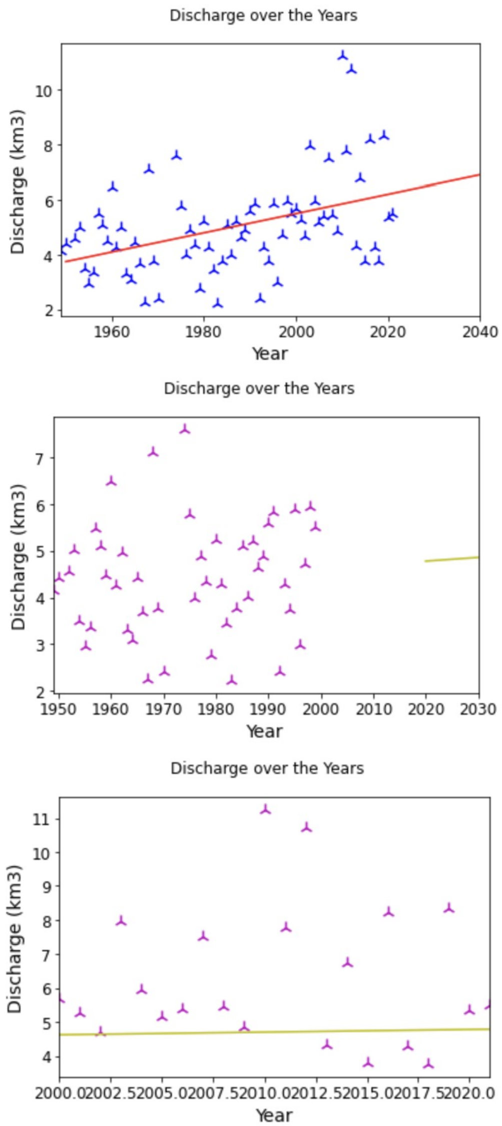

Figure 3. Annual Discharge plots, the top plot is from all of data to create a regression line. The middle plot is with training data, and a regression line from 2020-2040. The bottom plot is the test data, from 2000-2020. As can be seen in all of the plots, the data is highly scattered and the regression lines do not fit the data that factors in the fluctuations.

Predicting discharge of the Watson River is important in understanding sediment transport processes. The regression line produced suggests a steady and linear increase in annual discharge. This result does not correlate with the study done by Hasholt et al. (2018), which found significant seasonal and annual fluctuation in discharge - no significant trend could be detected from the 11 year observation period (Hasholt et al.2018).

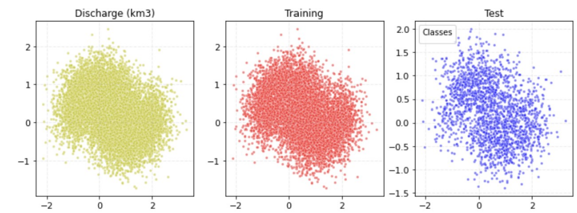

Figure 4. Decision tree scatter plots of A) all data B) training set C) test set.

5. Discussion

There is evidence of accelerated ice loss on the GIS, which is one of the largest sources of contemporary global sea rise (Box et al. 2022). Although it is a major climate change indicator, many factors involved in GIS hydrology are difficult to quantify. Gaining insight into the discharge history and potential forecast can help bridge knowledge gaps in GIS information. While the errors calculated for the decision tree showed little error, the linear regression curves plotted initially, visually show under fitting. With this under fitting taken into consideration, and the statement of further research required to properly analyze the discharge - sediment transport relationships, the increasing trend supports the assumption that sediment transport and sediment discharge alter the erosive capacity of the ice sheet (Hasholt et al. 2018). The increase in sediment transport discharge has, and will continue to have consequences in the surrounding deltas, with expansion predicted to continue (Hasholt et al. 2018).

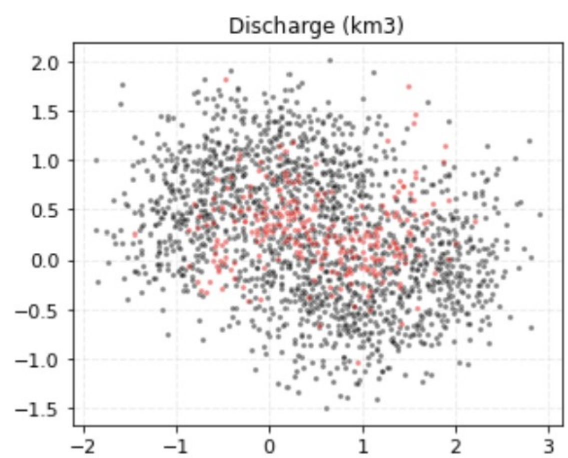

Figure 5. Scatter plot of predictions. Black points indicate correct predictions, red points indicate incorrect predicitons.

6. Conclusion

The decision tree model has shown to be more accurate than the linear regression model, with the linear regression plots showing under fitted curves. The discharge from the outflow of the proglacial river in southern west Greenland shows promising insight on river dynamics in a rapidly changing and climatically significant area. Machine learning algorithms are promising tools in forecast predictions. Further research in other areas of river dynamics are required to continue to understand the consequences of accelerated and changing melting and water fluxes.

7. Other resources

The code used for this project is at this link.

The link for the data set used is at this link.

8. References

Box, J. E., Hubbard, A., Bahr, D. B., Colgan, W. T., Fettweis, X., Mankoff, K. D., Wehrlé, A., Noël, B., van den Broeke, M. R., Wouters, B., Bjørk, A. A., amp; Fausto, R. S. (2022). Greenland ice sheet climate disequilibrium and committed sea-level rise. Nature Climate Change, 12(9), 808–813. https://doi.org/10.1038/s41558-022-01441-2

Hasholt, B., van As, D., Mikkelsen, A. B., Mernild, S. H., amp; Yde, J. C. (2018). Observed sediment and solute transport from the kangerlussuaq sector of the Greenland Ice Sheet (2006–2016). Arctic, Antarctic, and Alpine Research, 50(1). https://doi.org/10.1080/15230430.2018.1433789

Mao, J., Carlton, A., Cohen, R. C., Brune, W. H., Brown, S. S., Wolfe, G. M., Jimenez, J. L., Pye, H. O., Lee Ng, N., Xu, L., McNeill, V. F., Tsigaridis, K., McDonald, B. C., Warneke, C., Guenther, A., Alvarado, M. J., de Gouw, J., Mickley, L. J., Leibensperger, E. M., … Horowitz, L. W. (2018). Southeast Atmosphere Studies: Learning from model-observation syntheses. Atmospheric Chemistry and Physics, 18(4), 2615–2651. https://doi.org/10.5194/acp-18-2615-2018

Maulud, D., amp; Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), 140–147. https://doi.org/10.38094/jastt1457

Noël, B., Van Den Broeke, M., Van De Wal, R., H Van Ulft, L., Smeets, C. J. P., Munneke, P. K., Lhermitte, S., Lenaerts, J., Van As, D., Van Meijgaard, E., Van Wessem, J. M., amp; van de Berg, W. J. (2017). Review: "modelling the climate and surface mass balance of polar ice sheets using RACMO2, part 1: Greenland (1958-2016)" by noël et al.. The Cryosphere, 12(3), 811–831. https://doi.org/10.5194/tc-2017-201-rc2