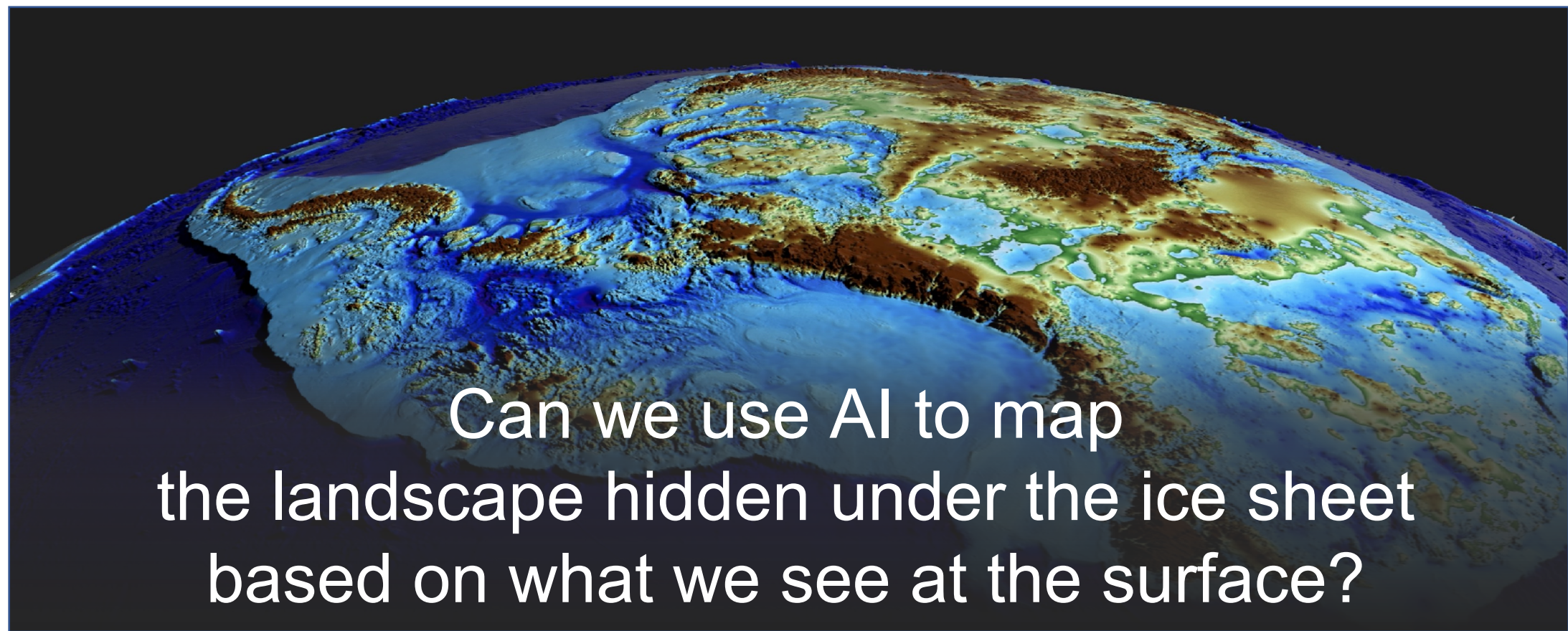




Can we use AI to map the landscape hidden under the ice sheet based on what we see at the surface?

Introduction

Predicting subglacial bed topography accurately is crucial for understanding how ice sheets move and respond to climate change. In this study, we enhance the capabilities of machine learning models using Apache Spark's distributed computing framework. By applying Spark, we accelerate the training and evaluation processes of models that estimate subglacial bed elevation from essential surface attributes such as ice thickness, flow velocity, and surface elevation. This approach not only improves the handling and analysis of large-scale radar track datasets but also boosts the computational efficiency and accuracy of predictions. Our objective is to utilize Spark's robust framework to facilitate scalable and faster machine learning computations, thereby providing deeper insights into ice sheet dynamics crucial for climate change mitigation efforts.

Data and Methods

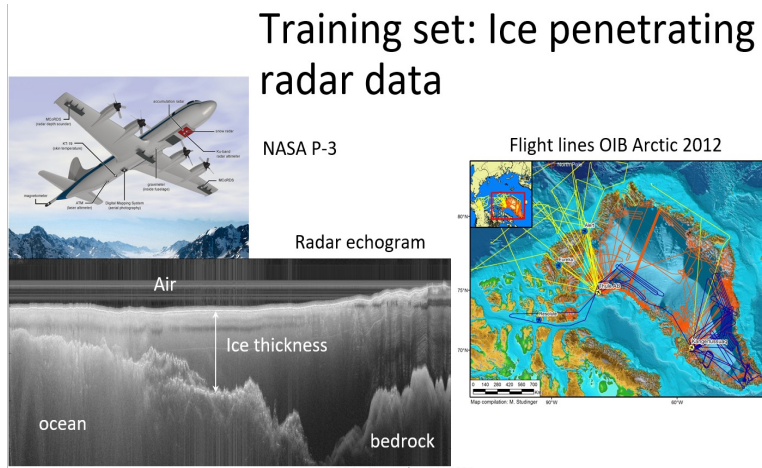
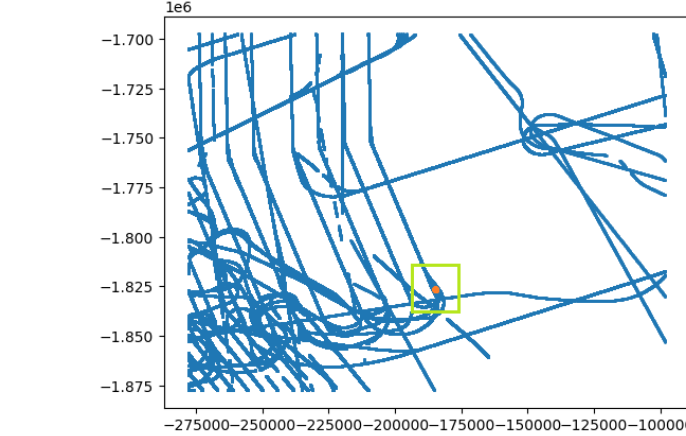
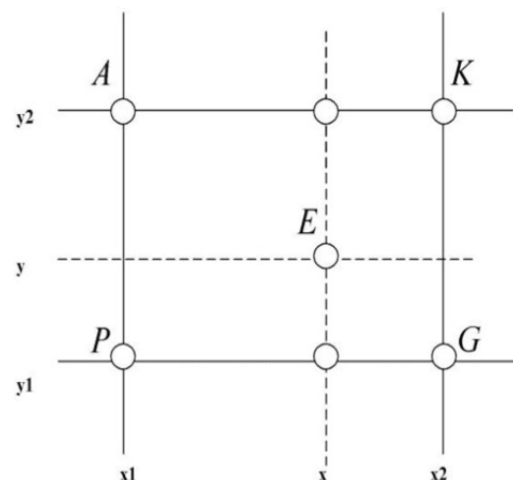
Data and Preprocessing:

- We choose a 2D square area from Upernavik in West Greenland
- The area is split into 1200x1200 square grids with grid size of 150x150 m², which are 1201x1201 data points
- 632,706 known data points, were collected via ice penetrating radar, from NASA IceBridge mission
- Surface elevation from Greenland Ice Mapping Project (GIMP/GrIMP) (2012) (Surf_elv),
- Rate of ice thinning from ICESat-2 (Surf_dhdt),
- Surface mass balance from RACMO. (Surf_SMB),
- Ice flow surface velocity from satellite interferometry (Surf_vx, Surf_vy)
- Because coordinates for the five surface variables are very different from the ones for target variable, we need to first match the two sets of data points based on their geolocations.

Variable Name	Description
surf_x, surf_y	Coordinates of grids (m)
surf_vx, surf_vy	Ice flow velocity (m/yr)
surf_elv	Ice surface elevation (m)
surf_dhdt	Ice thinning rates (m/yr)
surf_SMB	Surface mass balance (m/yr)
track_bed_x, track_bed_y	Coordinates of radar bed points (m)
track_bed_target	Subglacial bed elevation along flight lines (m)

Interpolation - Universal Kriging:

- Use the autocorrelation among these known surface observations within that area to assign appropriate weights to each contributing location's value.



Machine Learning Models:

1. Random Forest

- Uses bagging to build trees in parallel from bootstrap samples.
- Averages predictions from all trees for final results.

2. Gradient Boosting (GBoost)

- Combines weak models iteratively using gradient descent to minimize errors.
- Final prediction from weighted sum of all trees.

3. Extreme Gradient Boosting (XGB)

- Advanced gradient boosting with parallel tree construction.
- Uses level-wise growth and optimized split evaluation for final sum of predictions.

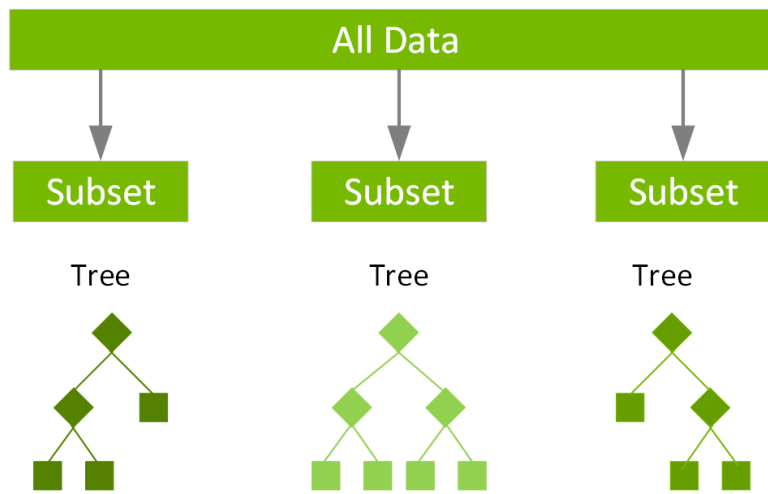


Figure from Nvidia Developer

Results

Evaluation Metrics:

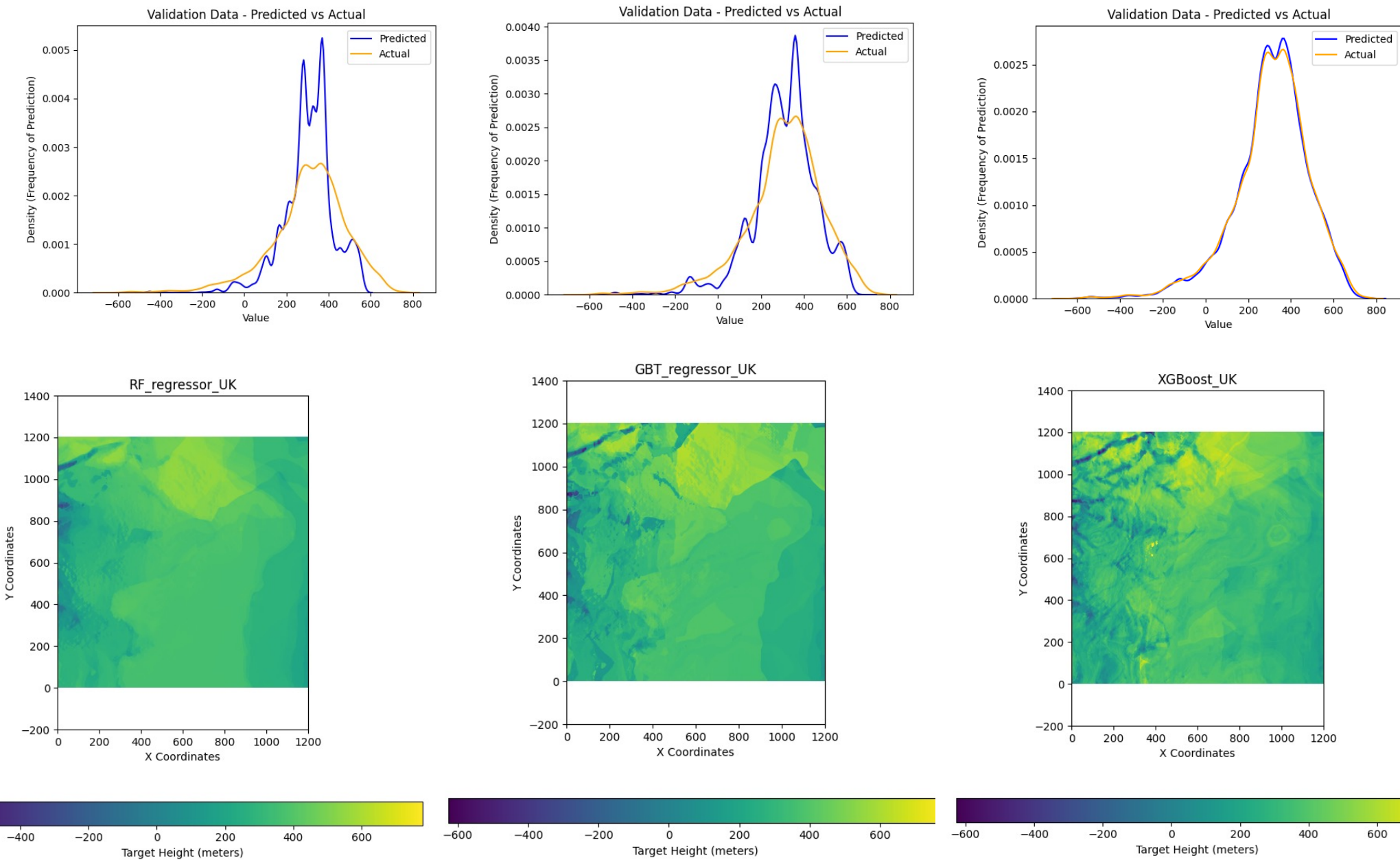
1. **Root Mean Squared Error (RMSE)**
2. **Mean Absolute Error (MAE)**
3. **Coefficient of Determination (R²)**

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\text{predicted}_i - \text{actual}_i)^2}{N}}$$
$$MAE = \frac{\sum_{i=1}^N |\text{actual}_i - \text{predicted}_i|}{N}$$
$$R^2 = \frac{SSR}{SST} = \frac{\sum (\text{predicted}_i - \bar{y})^2}{\sum (\text{actual}_i - \bar{y})^2}$$

Model	RMSE	MAE	R ²	Training Time
Random Forest	112.495	81.657	0.609	1m 16s 208ms
GBoost	87.633	61.781	0.763	1m 3s 218ms
XGBoost	29.762	19.795	0.972	2m 45s 459ms

Table 1: Comparative Performance of Machine Learning Models for Subglacial Bed Elevation Prediction on single node CPU

Validation Data Density Plots: Predicted vs. Actual for Random Forest, Gradient Boosting, Extreme Gradient boosting respectively



Scalability Test:

- Given significant outperformance of XGBoost relative to other models, we have exclusively presented the scalability test results for XGBoost.
- Increasing the number of workers per node generally reduces training time, with minor fluctuations in RMSE and MAE suggest that model accuracy remains stable across different worker counts.
- As the number of nodes increases, the training time decreases significantly, especially noticeable when going from fewer to more workers per node.
- The R² value remains consistently high across all configurations, indicating robust model performance irrespective of parallelization scale.

# Nodes	# Workers	RMSE	MAE	R ²	Training Time
1	2	29.809	19.776	0.972	2m 38s 658ms
1	6	29.762	19.795	0.972	2m 45s 459ms
1	16	29.642	19.663	0.972	3m 15s 969ms
2	2	29.809	19.776	0.972	2m 36s 599ms
2	6	29.571	19.616	0.973	2m 44s 255ms
2	16	29.678	19.663	0.972	2m 58s 748ms
3	2	29.809	19.776	0.972	0m 56s 637ms
3	6	29.771	19.811	0.972	1m 2s 848ms
3	16	29.631	19.788	0.972	1m 26s 282ms
4	2	29.809	19.776	0.972	0m 54s 517ms
4	6	29.808	19.892	0.972	1m 10s 103ms
4	16	29.488	19.652	0.973	1m 46s 061ms

Table 2: Results from a scalability test using XGBoost across different configurations of computing nodes and workers on CPU based HPCF.

CONCLUSIONS

- ❖ Enhanced predictions of subglacial bed elevation improve our understanding of ice sheet dynamics, essential for forecasting climate responses and supporting global climate change mitigation.
- ❖ Across various metrics such as RMSE, MAE, and R², XGBoost consistently outperformed other models, showcasing its robustness in predictive accuracy and computational performance.
- ❖ XGBoost demonstrated superior efficiency and scalability when enhanced with Spark's distributed computing, making it highly suitable for processing large-scale subglacial datasets.
- ❖ The integration of Spark parallelization significantly reduced training times, underscoring the benefits of distributed computing in environmental data analysis.
- ❖ **Further research will explore the application of deep neural networks and the use of multi-GPU architectures to enhance model training and performance on even larger datasets.**

ACKNOWLEDGMENT

- ❖ The authors would like to thank NSF (Big Data REU), UMBC, HPCF, and IHARP.
- ❖ The hardware used in the computational studies is part of the UMBC High-Performance Computing Facility (HPCF). The facility is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS-0821258, CNS-1228778, OAC-1726023, and CNS-1920079).

REFERENCES

- ❖ Katherine Yi, Angelina Dewar, Tartela Tabassum, Jason Lu, Ray Chen, Homayra Alam, Omar Faruque, Sikan Li, Mathieu Morlighem, and Jianwu Wang. Evaluating Machine Learning and Statistical Models for Greenland Subglacial Bed Topography. In 22nd IEEE International Conference on Machine Learning and Applications (ICMLA 2023)
- ❖ A. N. Howat, I. and B. Smith., "Measures greenland ice mapping project (gimp) digital elevation model from geoeye and worldview imagery, version 1," 2017. [Online]. Available: <https://nsidc.org/data/NSIDC-0715/versions/1>
- ❖ B. Smith, S. Adusumilli, B. M. Csatho, D. Felikson, H. A. Fricker, A. Gardner, N. Holschuh, J. Lee, J. Nilsson, F. S. Paolo, M. R. Siegfried, T. Sutterley, and the ICESat-2 Science Team, "ATLAS/ICESat-2 L3A Land Ice Height, Version 5," Online Data, 2021. [Online]. Available: <https://nsidc.org/data/ATL06/versions/5>
- ❖ J. Mouginot, E. Rignot, B. Scheuchl, and R. Millan, "Comprehensive annual ice sheet velocity mapping using landsat-8, sentinel-1, and radarsat-2 data," Remote Sensing, vol. 9, no. 4, p. 364, 2017.
- ❖ M. Morlighem, C. N. Williams, E. Rignot, L. An, J. E. Arndt, J. L. Bamber, G. Catania, N. Chauche, J. A. Dowdeswell, B. Dorschel et al., "Bedmachine v3: Complete bed topography and ocean bathymetry mapping of greenland from multibeam echo sounding combined with mass conservation," Geophysical research letters, vol. 44, no. 21, pp. 11-051, 2017.