Evaluating Machine Learning and Statistical Models for Greenland Subglacial Bed Topography

Katherine Yi[1], Angelina Dewar[2], Tartela Tabassum[3], Jason Lu[4], Ray Chen[5],
Homayra Alam[3], Omar Faruque[3], Sikan Li[6], Mathieu Morlighem[7], Jianwu Wang[3]
[1]Department of Computer Science, Purdue University, West Lafayette, IN, USA
[2]Department of Physics, University of Oregon, Eugene, OR, USA
[3]Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD, USA
[4]Department of Information Studies, University of Maryland, College Park, College Park, MD, USA
[5]Marriotts Ridge High School, Maryland, Ellicott City, MD, USA
[6]Texas Advanced Computing Center, University of Texas at Austin, Austin, TX, USA
[7]Department of Earth Sciences, Dartmouth College, Hanover, NH, USA
Emails: [1]klyi@purdue.edu, [2]adewar@uoregon.edu, [3]{tartelt1, halam3, omarf1, jianwu}@umbc.edu,
[4]lu422839@terpmail.umd.edu, [5]rchen6@umbc.edu, [6]sli@tacc.utexas.edu, [7]mathieu.morlighem@dartmouth.edu

***Abstract*—The purpose of this research is to study how different machine learning and statistical models can be used to predict bedrock topography under the Greenland ice sheet using ice-penetrating radar and satellite imagery data. Accurate bed topography representations are crucial for understanding ice sheet stability and vulnerability to climate change. We explore nine predictive models including dense neural network, long-short term memory, variational auto-encoder, extreme gradient boosting (XGBoost), gaussian process regression, and kriging based residual learning. Model performance is evaluated with mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), and terrain ruggedness index (TRI). In addition to testing various models, different interpolation methods, including nearest neighbor, bilinear, and kriging, are also applied in preprocessing. The XGBoost model with kriging interpolation exhibit strong predictive capabilities but demands extensive resources. Alternatively, the XGBoost model with bilinear interpolation shows robust predictive capabilities and requires fewer resources. These models effectively capture the complexity of the terrain hidden under the Greenland ice sheet with precision and efficiency, making them valuable tools for representing spatial patterns in diverse landscapes.**

***Index Terms*—Machine Learning Applications, Model Evaluation, Interpolation, Greenland, Subglacial Bed Topography**

I. Introduction

Accurately mapping the topography of the bedrock under the ice sheets is critical to improve our understanding of their response to climate change and reduce the uncertainty in sea level rise projections. The most effective way to measure the height of the bedrock under a glacier is through airborne ice penetrating radar. However, acquiring such data entails significant costs, and it only provides data directly underneath the aircraft, leaving significant data gaps. In contrast, surface glacier data can be readily obtained through satellite observations, which researchers are using to explore model capabilities of estimating glacial bedrock topography based on surface features. Prior efforts in this domain include the physics-based model BedMachine by Morlighem et al. [1], which predicts bed topography in Greenland based on mass conservation, and the deep neural network-based DeepBedMap by Leong and Horgan [2], designed to reconstruct a more realistic bed topography roughness of the Antarctic bed based on existing maps.

To comprehensively understand the efficacy of diverse machine learning and statistical models in predicting subglacial bed topography, we evaluate nine distinct methods for Greenland. Our exploration includes statistical techniques such as gaussian process regression and universal kriging, as well as machine learning approaches such as variational autoencoder (VAE), extreme gradient boosting (XGBoost), dense layer-based neural network, and long-short term memory (LSTM) based neural network. Additionally, we also attempt three novel approaches by combining two models into a single hybrid model, e.g., dense + LSTM, VAE + XGBoost, and universal kriging + XGBoost.

Our contributions are summarized as follows with our implementation open source at the Big Data REU GitHub repository [3].

- Our XGBoost model is the best-performing model overall and demonstrates favorable performance compared to the commonly used topography product BedMachine [1], indicating the potential of machine learning for accurate subglacial bed topography prediction.
- These experiments also show that the universal kriging model achieves the best MAE score due to its ability to leverage neighboring data points effectively. However, its poor RMSE performance can be attributed to challenges in handling sparse target variable observations, leading to outliers that impact overall accuracy. A notable disadvantage of universal kriging is its extensive resource requirement for convergence compared to other models.
- To improve data integration and leverage physics-guided

This work is supported by the grants OAC–2050943 and OAC–2118285 from the National Science Foundation.

knowledge in preprocessing, we conducted an ablation study on different interpolation approaches. The results demonstrate that different interpolation techniques have varying effects on machine learning model performance, and the incorporation of physics knowledge positively impacts machine learning performance.
- Besides evaluating the models with ground-truth track data, we apply them to a large region spanning 32,400 km^2. Visualizing the predictions for the entire region provides a comprehensive understanding of the performance differences among the models. Furthermore, we find the terrain ruggedness index (TRI) [4] to be a valuable metric for quantifying these differences.

The subsequent sections of this paper are structured as follows. Section II, provides an overview of related works in the field. Section III presents the background and data sources utilized in this study. Section IV delves into the detailed exploration of the predictive models employed. In Section V, we outline the experimental metrics, present the results, ablation studies, and corresponding descriptions. Section VI entails a comprehensive discussion of our findings. Finally, in Section VII, we draw conclusive remarks from our investigation.

II. Related Work

The work of Morlighem and colleagues [5] resulted in a physics-based model for inferring bedrock topography beneath the Greenland ice sheet using a mass conservation optimization scheme. This model, constructed from a dataset comprising of radar-derived bed topography, where available, and satellite-based ice flow velocity, serves as a reference for visualizing the extensive dataset utilized in our study, considering the absence of a ground truth.

In the context of Antarctic bed topography, Leong and Horgan [2] introduced a GAN-based approach to minimize per-pixel elevation error. Their work was motivated by the compilation of Bedmap1 [6] and Bedmap2 [7]. Notably, mass conservation was only applied to fast-flowing regions, while ordinary kriging or streamline diffusion was used in slow-moving ice flow regions [8]. However, this method faced challenges in capturing the intrinsic anisotropy of the ice thickness data where mass conservation is essential.

Furthermore, the works of Leong and Horgan [2] and Liu-Schiaffini et al. [9] have provided valuable insights into the application of deep learning, specifically convolutional neural networks (CNNs), for predicting bed topography roughness. These prior studies have informed and guided our team in the implementation of deep learning models to advance the state-of-the-art in this field.

Our work encompasses extensive implementation of both statistical and machine learning models on our dataset building on previous research, thereby enhancing the understanding of their respective performances.

III. Background and Data Preprocessing

Addressing bed topography is crucial due to its significance in predicting future sea level rise. It has been shown that the shape of the bedrock can slow down the retreat of ice sheet, or accelerate it through the Marine Ice Sheet Instability [10]. Typically, glaciers tend to be stable on pro-grade bed slopes (when the bed rises as we move inland), and unstable on retro-grade bed slopes. Small scale roughness can also affect the overall rate of retreat. Measuring the landscape hidden beneath thousands of meters of ice, however, remains challenging. Scientists have developed numerical models to predict sea level rise based on existing bedrock topography maps. However, uncertainties persist in forecasting the future response of the ice sheets to climate change over the coming decades and centuries because of our incomplete and sometimes erroneous representation of subglacial bed topography. Identifying retrograde regions with deepening beds and areas with bumps and ridges that may impede rapid retreat is critical. Without accurate bed representations, precise sea level rise predictions are hindered. Understanding climate change requires crucial insights into the underlying bed, even as thick ice poses significant obstacles to direct observation.

A. Problem Definition

Our main objective is to develop and train machine learning models to infer bed elevation across a vast area using solely surface data as input because the available ground truth data for the area's actual bed elevation is spatially limited. Our analysis is centered on predicting bed topographic elevation at new locations, where direct measurement is unavailable, thereby enhancing the depiction of Greenland's subglacial topography. Below, we explain the datasets in our study and data preprocessing tasks.

B. Data Sources

We selected a 32,400 km^2 square area in the Upernavik glacier system, in West Greenland. By splitting this area into 1200×1200 uniform grids (each grid's size is 150 m $\times$ 150 m) we obtain a dataset with 1,442,401 (1201×1201) total data points. Each data point provides values for five surface variables (see rows 1-5 in Table I), which are original features to be leveraged by our predictive models. Because these are uniform grids, we refer to them as *Grid Data*. Meanwhile, we also obtain 632,706 data points of radar-derived bed elevation measurements as target values, which are rows 6-7 in Table I. These data points have ground truth that are used to measure the accuracy of our predictive models. Because these data points are only along the tracks of the airborne radar sensor, we refer to them as *Track Data*.

1) Ice Sheet Surface Measurement: We utilize five ice sheet surface measurements from four sources: 1) Surface Elevation is obtained from the Greenland Ice Mapping Project (GIMP) [11]; 2) Ice flow surface velocity data on both longitudinal and latitudinal directions is generated by integrating multiple satellite interferometry data products including Landsat-8, Sentinel-1, and RADARSAT-2 via the approach at [12]; 3) Ice

TABLE I
DESCRIPTION OF DATA VARIABLES

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)

thinning rates are provided by ICESat-2: (Ice, Cloud, and land Elevation Satellite-2) [13]; 4) Surface mass balance indicating annual snow accumulation and ice surface ablation is derived from RACMO (Regional Atmospheric Climate Model) [14]. Examples of surface measurement variables in our study are illustrated in the left part of Figure 1.

2) Bed Elevation Measurement: Bed topography measurements are acquired through ice-penetrating radar from NASA's Operation IceBridge [15]. A radar system is mounted beneath an aircraft's wings and emits an electromagnetic signal that penetrates the ice. This signal is reflected at the ice-bed interface directly underneath the aircraft and the travel time is converted to ice thickness. The bed topography is calculated by subtracting the ice thickness from the ice surface elevation. The right part of Figure 1 presents a visual representation of the dataset. This target variable data, that is irregularly distributed, is contrasted with the uniformly distributed surface variable data previously discussed.

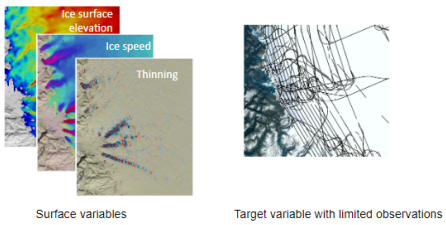


Fig. 1. Surface variable grid data and target variable track data.

C. Data Preprocessing

The first data preprocessing step conducted is to estimate the surface variables for each track data point based on their geolocations, so every track data point has five feature variables and a target variable. To achieve this goal, we employ three distinct interpolation techniques: nearest neighbor, bilinear, and universal kriging.

The first two interpolation approaches are straightforward. Nearest neighbor, involves associating each radar bedrock observation with its spatially-nearest surface observation. The second approach, bilinear interpolation, predicts the value of each surface variable by calculating the weighted average of the four nearest neighbor observations of that variable. The weights of each neighbor are determined based on their distances from the prediction location.

The third interpolation technique in our study is universal kriging. Similar to bilinear interpolation, it estimates the values of surface variables using nearby observations. However, unlike bilinear interpolation, universal kriging incorporates all observations in the dataset (or a specified neighborhood due to memory constraints) to calculate a weighted average. The weights are assigned based on the auto correlation of the observations, utilizing a variogram model. For each batch, universal kriging is fit using different variogram models, selecting the model with the lowest capability ratio (CR).

The application of all three interpolation methods to the original region of interest dataset yield three distinct interpolated datasets. Each interpolated dataset comprises a total of 632,706 examples, with each example containing the interpolated values of the five surface variables, the true observed value of the bed elevation at the corresponding location, and the respective coordinates.

Using our three interpolated datasets, we conduct further preprocessing steps to prepare the data for our models. This involves dropping non-numerical values, scaling the data with Standard Scaler to account for outliers in real-world data, and calculating the magnitude of the ice flow velocity vector for each surface observation. Additionally, for the universal kriging and universal kriging + XGB models, duplicates are removed from the dataset to ensure convergence during the training process. Finally, the dataset is randomly split into 60% training, 40% testing, and 20% validation datasets with a constant randomization seed to ensure repeatability. Next, models are developed and trained.

IV. STATISTICAL AND MACHINE LEARNING MODELS EVALUATED

A. Statistical Models

Due to both models' popularity in problems involving spatial data, we apply gaussian process regression (GPR) and universal kriging to our dataset as predictive models.

1) Gaussian Process Regression (GPR): GPR is a powerful probabilistic regression technique that leverages Bayesian principles to compute predictive distributions [16]. Despite its promising properties, GPR presents challenges in terms of computational resources and storage requirements. During development using a subset of data, 90% training and 10% testing split, the model demonstrates impressive metrics, but has poor computation time. Due to the high computational cost of training GPR on large datasets, we employ mini-batch processing with 10,000 points and one epoch, which demonstrates sufficient metrics during development to manage the significant time and financial investments. Unfortunately, the model continues to demand an incredible number of resources for computation, therefore, continuing the challenge of applying GPR to large-scale geographical datasets. Because of the challenge to engage the entire topography, we explore other models.

2) Universal Kriging: We select universal kriging [17] as our preferred kriging variant due to its robustness in handling trends, outliers, and skewed data distributions. Our application

of kriging adopts a nontraditional approach to account for the large size and high degree of variance across different regions of our dataset. Instead of employing a train-test split, we use all of the labeled data in approach more akin to cross-validation. We fit five different variogram models (exponential, power, linear, gaussian, and spherical) on each prediction point using a specific neighborhood of the prediction point. All covariance models are fit with the same settings: an anisotropy scale of 3, an anisotropy angle of the arc-tangent of ice flow velocity, and using the number of fitting points as the number of lags. The best covariance model is chosen for each prediction point as the model with the lowest capacity ratio (CR). These adjustments are crucial to manage the considerable volume of data points in our dataset, reduce computational complexity, and accommodate the substantial variance in correlation falloff across the entire region of interest. Even with these adjustments, similar to GPR, universal kriging requires a large number of resources during computation time and other models are explored.

B. Machine Learning Models

1) Extreme Gradient Boosting (XGB/XGBoost): XGBoost is a popular supervised machine learning algorithm. Known for its effectiveness in regression tasks and computational efficiency, XGB is well-suited for processing large datasets [18] with a focus on learning from smaller, diverse regions. The model combines multiple decision trees, optimizing them to minimize errors, which enhances flexibility and makes it adaptable to diverse topographic data. XGB incorporates techniques to prevent overfitting, ensuring reliable and accurate predictions, while its ability to aggregate predictions from individual trees solidifies its standing as the preferred choice for our objectives.

In justifying our choice of the XGBoost algorithm, we consider the importance of finding the right balance among key parameters. To achieve optimal results, we employ a reverse-engineering approach, overfitting the data and fine-tuning parameters through extensive experimentation and emphasizing domain-specific characteristics to enhance the model's performance and adaptability. These efforts significantly impact the XGBoost model's accuracy in predicting Greenland's bedrock elevation.

Parameter settings are carefully chosen to optimize the model's performance while avoiding overfitting. The depth of the decision trees is set to a balanced value of seven to capture intricate patterns. The number of boosting rounds and XGBoost trees is set at 350 to ensure comprehensive learning while maintaining training efficiency. The minimum child weight, set at 0.25, controls overfitting by requiring a minimum number of samples to split a node. The subsample parameter, set at 0.8, balances incorporating diverse samples and providing sufficient dataset coverage. Finally, the learning rate of 0.25 facilitates stable performance and convergence without overshooting.

By precisely tuning the parameters of the XGBoost algorithm, we gain the ability to discern intricate details across the entire area, with a particular focus on emphasizing deep and narrow glacial fjords that channelize the ice flow from the interior of the ice sheet to the ocean. Through this fine-tuning process, we achieve noteworthy results shown in Table II, showcasing the model's ability to capture nuanced patterns and achieve accurate predictions tailored to our research objectives.

2) Multilayer Perceptron (MLP) Dense Network: Deep learning models are gaining widespread popularity across various domains, proving effective in tasks such as classification, regression, and detection, among others motivating their exploration in this study. In this study, we utilize a dense layer-based neural network, known for its simplicity and feature aggregation capabilities, to process the labeled dataset. The sequential network is comprised of eight layers, including five dense layers of sizes [128, 64, 32, 32, 16], two dropout layers, and a final dense layer with one output unit. A dropout coefficient of 50% is employed to enhance generalizability and mitigate overfitting. ReLU serves as the activation function for all layers, with a linear activation function being used for the output layer. Training utilizes the Adam optimizer with a mean squared error loss function over 200 epochs without early stopping. These choices are to take advantage of the network's ability to effectively capture features and patterns in our dataset.

3) Long Short-Term Memory (LSTM) Deep Neural Network: LSTM, a specialized type of recurrent neural network, excels at learning features from sequential data, thanks to its long-term and short-term memory units that address the vanishing gradient problem and retain patterns from distant sequences [19]. The LSTM model is explored to capitalize on LSTM's ability to learn from sequential data. Leveraging these advantages, our model comprises three LSTM layers of sizes [64, 32, 16], followed by a 50% dropout layer and a single-neuron dense layer for output generation. Training uses the Adam optimizer with mean squared error as the loss metric and 100 epochs without early stopping. This choice of LSTM enables us to effectively capture hidden patterns in our data, making it a reasonable approach for the prediction task.

4) Variational Autoencoder (VAE): VAE models have gained significant prominence in the field of machine learning. By employing feature reconstruction and incorporating KL regularization, VAE excels at learning probabilistic representations of geographical landscapes which motivates exploration of the model in this research [20]. The decoder's goal is to accurately reconstruct the original input from the latent space, leveraging existing knowledge to enhance performance. The introduction of a cyclical training schedule further improves the model's capabilities, making VAE a valuable tool for analyzing and predicting geographical features, not limited to Greenland but potentially applicable to other regions as well. Nevertheless, continued research and fine-tuning are essential to achieve even higher levels of accuracy and reliability in this challenging domain.

C. Hybrid Models

1) Dense + LSTM: In this study, we aim to leverage the strengths of both dense layers and LSTM layers by integrating them into a hybrid deep-learning model. The motivation behind this approach lies in the limited number of features available for the regression task, as we want to extract complex patterns from the dataset. The dense layer component of this hybrid model generates latent representations of the input features as a long sequence, which are then fed into the LSTM layers to generate high-level features, treating each data observation as an independent sequence. The dense layer part of the hybrid model comprises three dense layer blocks, each consisting of two dense layers, one batch normalization layer, and one dropout layer with a 50% dropout coefficient. On the other hand, the LSTM part of the model is two LSTM layers with 64 and 32 cells, with a batch normalization layer and one dense layer. The final regression result is generated using a dense layer with a single neuron and a linear activation function. This hybrid model is trained for 200 epochs using the Adam optimization function with mean squared loss. By integrating dense layers and LSTM layers, our hybrid deep-learning model captures both combinatorial and sequential patterns from the dataset, enabling us to obtain more comprehensive and accurate predictions for the regression task.

2) Variational Autoencoder + XGBoost: The rationale behind developing the VAE + XGB model is to leverage the respective strengths of VAE and XGBoost. The VAE component excels at capturing essential features and patterns within the dataset through its latent space representation. By compressing and encoding the input data into a lower-dimensional latent space, it creates a more concise and meaningful representation. Subsequently, the XGBoost model demonstrates its capability for producing high-quality and rapid predictions for the dataset. In the hybrid model, we utilize the trained VAE encoder to capture patterns within our dataset. A subset comprising 80% of the compressed patterns from the encoder is employed as input for the XGBoost model, which is fine-tuned with 350 trees, a maximum depth of seven levels, and a minimum child weight of 0.25. The model performs predictions for the target elevations on all testing and validation data, thereby enabling comprehensive analysis and predictions with improved accuracy and efficiency.

3) Universal Kriging + XGBoost: Inspired by residual learning [21], we study how to leverage the strengths of universal kriging and XGBoost in predicting bedrock elevation. In this pipeline, universal kriging makes an initial prediction for the target variable. Subsequently, the XGBoost model is trained using the same features, with the target variable being the residuals from the kriging prediction from the true observed values. During testing, the XGBoost predictions of residuals are combined with the kriging guess. This approach aims to enhance the accuracy of the overall prediction by allowing XGBoost to learn and correct any errors or inconsistencies present in the kriging estimation.

V. Evaluation Results

We conduct three types of experiments to understand how the models in the previous section perform for Greenland bedrock elevation prediction. First, we compare all nine predictive models to understand their performance difference using the same interpolated track data because it has ground truth values. Second, we complete an ablation study for our best performing model to understand how our preprocessing affects the performance. Third, we use grid data to understand the performance of the model for the whole area. Because BedMachine [1] is a popular Greenland bedrock topography data product is used in the ice sheet modeling community, we use this physics-based estimate as one baseline.

A. Metrics for Model Evaluation

To evaluate the predictive performance of our models on the track data, we employ root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) as evaluation metrics for the test sets.

Besides the common statistical metrics above, for well-performing models, we also compute the terrain ruggedness index (TRI) on the grid data. TRI is a valuable measure for quantifying terrain characteristics by capturing the elevation differences between adjacent cells in a digital elevation model (DEM). It quantifies topographic heterogeneity by computing the squared and averaged differences between the center cell and its eight surrounding cells, followed by taking the square root to yield the TRI value. Higher TRI values indicate greater terrain ruggedness and complexity, while lower values suggest smoother terrain. Incorporating TRI into our analysis enables us to assess how effectively our predictions capture the study area's terrain variability and roughness, thereby ensuring an accurate representation of its complex topographic features.

B. Statistics Evaluation Results

Table II presents test data evaluation results for nine models in Section IV trained with data interpolated using the nearest neighbor approach. The results are ordered by RMSE with the best results emboldened for each metric. Because our baseline, BedMachine, has the same geolocation with our grid data, we did bilinear interpolation to get its values for the track data first in order to calculate its metrics using the same test data. Below, the results are discussed for each of the models employed.

TABLE II
Summary of Results (Sorted by Best Metrics)

Model	RMSE	MAE	R^2
XGBoost	**32.680**	22.273	**0.967**
BedMachine [1]	71.554	50.422	0.842
Dense + LSTM	81.174	57.993	0.797
LSTM	101.630	74.522	0.682
Dense	104.475	74.381	0.663
Variational AutoEncoder	106.884	83.798	0.648
VAE + XGBoost	129.760	100.035	0.481
XGBoost on Kriging Residual	136.650	8.122	0.424
Kriging Only	136.637	**7.617**	0.424
Gaussian Process Regression	150.086	97.855	0.293

1) XGB: Overall, XGB performs the best. The impressive metrics of XGBoost come from the combination of multiple decision trees in XGBoost, which are optimized to minimize errors and provide flexibility in capturing diverse topographic data. This architecture allows us to break down complex topography, capturing unique sections, and then combine the results for more precise predictions.

2) Dense + LSTM: The dense + LSTM model performs better than dense model alone and LSTM alone. It shows this hybrid approach is generally successful due to its unique architecture that combines the strengths of both dense layers and LSTM cells. The dense layers are effective in learning complex relationships and patterns within the data, capturing high-level features that contribute to predictive accuracy. On the other hand, LSTM cells excel in capturing temporal dependencies and long-term patterns, making them well-suited for sequential data such as time-series or spatial-temporal datasets. Combining dense and LSTM layers, the model can effectively capture both spatial and temporal dependencies in the topographical data. The dense layers process the input data and extract essential features, while the LSTM cells process the sequential information and capture temporal patterns. This combination enables the model to understand the complex interactions between geographical features over time, resulting in improved predictive performance. Additional tuning may result in further improved metrics.

3) LSTM: The LSTM model performs similarly with metrics falling in the middle. LSTM alone might not fully exploit the spatial relationships and interactions between geographical features. The model's architecture focuses primarily on capturing temporal patterns, but it may not be as effective in handling the spatial complexities of the topographical data.

4) MLP Dense: The MLP dense model metrics are moderate. While MLP is capable of capturing some patterns in the data, it is not as effective in handling the complex spatial dependencies and sequential nature of the topographical data in Greenland. MLP dense model lacks the specialized architectures of models like LSTM, which captures long-term dependencies in sequential data, or XGBoost, which efficiently handles complex interactions between geographical features.

5) Variational Autoencoder (VAE): In this study, VAE demonstrates moderate predictive capabilities compared to other models. The model is known for its capability to capture essential features and patterns in the data, but it is not as effective in capturing complex spatial relationships and variations present in the topographical data of Greenland.

6) Variational Autoencoder + XGBoost: The VAE + XGB hybrid model did not perform better than VAE model itself. The VAE component is adept at capturing essential features and patterns in the data through its latent space representation. It effectively compresses and encodes the input data into a lower-dimensional latent space, allowing for a more concise and meaningful representation. XGBoost, tries to take advantage of the compressed and meaningful representations from the VAE, but it does not have enough opportunity to learn with the compressed data.

7) Universal Kriging + XGBoost: The XGBoost model, trained for predicting kriging's first guess residuals, exhibits notably poor performance compared to the other models. The large difference in RMSE to MAE can be explained by the lack of outliers present in the dataset. Hoping to improve the RMSE and R^2 scores, XGB is implemented to predict residuals and correct the first pass prediction. The XGB struggles to capture the relationship between surface features and computed kriging residual with an RMSE of 142.204 and R^2 of -0.083. Full implementation of the poorly performing XGB with residual correction shows little difference in final results of kriging residual learning seen in Table II.

Alternatively, to predict residuals accurately the XGB model would need to understand the decisions made during kriging that allowed the computation of residuals XGB is expected to predict. In an ablation study, adding the initial kriging prediction into the features of XGB showed an improvement in predicting residuals (RMSE 34.188 and R^2 of 0.937), but still results in an extremely poor final score when the predicted residuals are combined with the kriging predictions. Again, the XGB model struggles to understand the implicit decisions made during kriging.

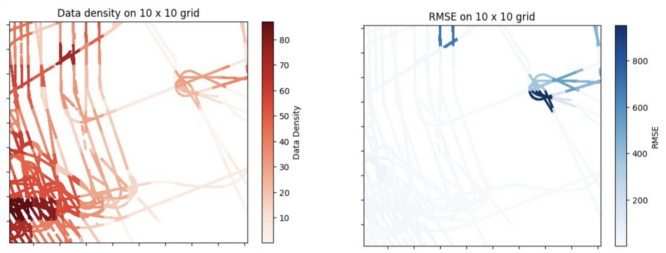
Because the hybrid model is unable to accurately correct initial kriging predictions with or without a well trained XGB model predicting residuals, the hybrid model is unsuitable for practical use.

8) Gaussian Process Regression (GPR): The GPR model's poor performance can be attributed to its computational demands, which hinder extensive hyperparameter searches and fine-tuning. The complex and diverse geographical features of Greenland's bed topography challenge GPR to accurately capture underlying patterns, as it relies on the normal distribution assumption that does not appear to hold in the tested regions. Additionally, GPR's relatively low R^2 value indicates its struggle to capture variability. As a result of poor metrics, the focus shifts towards alternative modeling approaches that could provide more efficient and cost-effective solutions for predicting topography data in Greenland.

9) Universal Kriging: Interestingly, our kriging model achieves the best MAE score with very poor R^2 and RMSE. Upon further investigation, it becomes clear that this is because of a small number of outliers with extremely poor predictions. Kriging actually performs very well on the majority of the dataset. By our calculation, 83.4% of predictions have a residual of 10 or less, 81.4% had an RMSE of 25 or less, and 79.6% have an R^2 score of 0.97 or higher. Because RMSE is more sensitive to outliers than MAE, kriging's RMSE is poor.

Hoping to clarify why predictions are so poor for some data points, we visualize the distribution data points and the RMSE metric for grid regions of the kriging input dataset and predictions. We split the input dataset and kriging prediction set into spatial batches along an even 10×10 grid and plot the data density (number of radar observations present in grid square/area of grid square in m^2) and RMSE for each grid square as a heatmap on the original pixel points contained within each grid square. The results are shown below in

Figure 2.



(a) Data density in 10×10 grid (b) RMSE in 10×10 grid

Fig. 2. Heatmaps of data density and RMSE on each batch in 10x10 grid.

Figure 2 shows a weak correlation between data density and the RMSE metrics (Pearson correlation coefficient of -0.104). This indicates some higher dense regions have lower RMSE scores. But this factor alone is not enough to explain why kriging results are so poor for certain regions. For instance, the batch in the upper right (dark blue lines on the RMSE plot) has very large error, even though its data density is not low. This leads us to conclude that additional factors, such as low spatial correlation or high variability, may be the reason why kriging is unable to make good predictions in these particular regions.

C. Ablation Study Results

1) Interpolation Study Results: We use our best performing model, XGBoost for ablation studies to understand how different preprocessing approaches affect model performance. The metrics for XGBoost trained on each of the three interpolated datasets are shown in Table III.

TABLE III
INTERPOLATION EVALUATION METRICS

Interpolation	RMSE	MAE	R^2
Kriging	**27.099**	**17.947**	**0.977**
Bilinear	28.085	18.549	0.976
Nearest Neighbor	32.680	22.273	0.967

We form three conclusions about interpolation methods with XGBoost. First, XGBoost produces the best metrics with universal kriging interpolation, but knowingly uses extensive resources during preprocessing. Second, XGBoost with bilinear interpolation is the most efficient in time and memory required and produces high metrics when compared to universal kriging. Finally, XGBoost with nearest neighbor interpolation produces competitive metrics and captures subglacial topography well, but ranks third in the metrics tested.

2) Derived Feature Selection Study Results: In addition to studying interpolation methods, feature selection is also explored. By adding the additional feature, namely ice velocity magnitude, according to domain knowledge, our model slightly improved. This can be seen in Table IV.

D. Topographic Evaluation

In addition to the metrics and ablation studies, predictions for validation data are visualized and quantified using the popular topographic metric terrain ruggedness index (TRI). First, the TRI is calculated for known validation data where XGBoost performs with over 93% R^2 for all interpolation methods. Then, the TRI is calculated for the grid data and is compared to the TRI of BedMachine, which is known to be overly smoothed compared to radar data. Average TRIs for XGB with nearest neighbor interpolation, XGB with kriging interpolation, XGB with bilinear interpolation and BedMachine for the grid data are given in Table V.

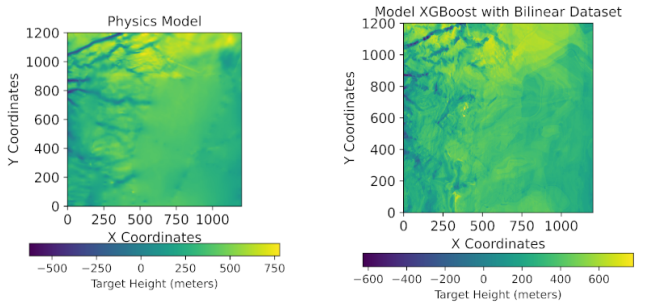
TABLE IV
EVALUATION METRICS WITH XGBOOST AND NEAREST NEIGHBOR INTERPOLATION

Velocity Magnitude	RMSE	MAE	R^2
With Velocity Magnitude	**32.680**	**22.273**	**0.967**
Without Velocity Magnitude	33.016	22.663	0.966

TABLE V
TERRAIN RUGGEDNESS INDEX FOR XGBOOST ON VALIDATION DATA

Interpolation	TRI Mean
Kriging	15.926
Bilinear	14.682
Nearest neighbors	14.178
BedMachine [1]	3.302

Figure 3 shows the trained XGBoost model produces improved details and roughness in the terrain when compared to the baseline result which is known to be overly smooth. Meanwhile, XGB based predictions show some discontinuity especially for the dark colored fjords, which might differ from actual topography.



(a) BedMachine (b) XGB with bilinear interpolation

Fig. 3. Comparison of interpolation ablation study results.

VI. DISCUSSION

The comprehensive evaluation of predictive models in this study yields valuable insights into the most effective approach for capturing and representing complex topography. To enhance predictive accuracy and efficiency in modeling topographical features in diverse landscapes, there are promising avenues for future research.

Firstly, implementing the terrain ruggedness index (TRI) as an early stopping method could be beneficial. By utilizing TRI

as a metric during the training process, we can effectively determine when the model adequately captures the terrain's complexity and variability, potentially saving computational resources and time.

Secondly, exploring tuning options for universal kriging presents another opportunity for improvement. While universal kriging demonstrates favorable performance in certain regions, it exhibits subpar results in others. Investigating potential flaws in the resizing method or identifying regions with poor spatial correlations can contribute to enhancing universal kriging's overall performance.

This study significantly advances our understanding of predictive capabilities for complex topography representation using various models. By continually refining and advancing our predictive approaches, we can deepen our understanding of complex topographical features and their implications, ultimately contributing to diverse applications, including environmental modeling and climate change research.

VII. Conclusions

Driven by the importance of comprehending complex topography and its implications in diverse fields like climate change research, environmental modeling, and glaciology, this study's objective is to identify the most effective predictive models for accurately capturing and representing spatial patterns in complex terrain. Leveraging insights from past literature and a comprehensive understanding and preprocessing of our data, we embark on a rigorous evaluation of several predictive models, employing well-established evaluation metrics like RMSE, MAE, R^2, and TRI with a specific focus on predictive accuracy and topography representation.

The results reveal distinct performance characteristics among the evaluated models. The XGBoost model with kriging interpolated data exhibits the best metrics, achieving an RMSE of 27.099, a MAE of 17.947, and an R^2 value of 0.977, indicating its strong predictive capabilities. However, the method's computational requirements, taking much longer time even with distributed resources, limit its practicality.

The XGBoost model with bilinear interpolation also demonstrates excellent predictive capabilities, achieving an RMSE of 28.085, a MAE of 18.549, and an R^2 value of 0.976. The TRI analysis further supports its effectiveness. Notably, this model requires fewer resources, showcasing both efficiency and accuracy in its predictions.

The other evaluated models exhibit varying degrees of predictive performance, with some showing promising results and potential for further refinement. However, the XGBoost models with kriging and bilinear interpolated data stand out as the top performers, effectively capturing the complexity of the Greenland ice sheet terrain with precision and efficiency.

In conclusion, the XGBoost models with kriging and bilinear interpolated data prove to be highly effective in predicting topographical features in complex terrains for our dataset. These models demonstrate exceptional predictive accuracy while efficiently processing large datasets, making them valuable tools for capturing and representing spatial patterns in the bed topography under the Greenland ice sheet and potentially other diverse landscapes.

References

[1] M. Morlighem, C. N. Williams, E. Rignot, L. An, J. E. Arndt, J. L. Bamber, G. Catania, N. Chauché, 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.

[2] W. J. Leong and H. J. Horgan, "Deepbedmap: A deep neural network for resolving the bed topography of antarctica," *The Cryosphere*, vol. 14, no. 11, pp. 3687–3705, 2020.

[3] "Github repository for predicting ice-bed topography using predictive modeling." https://github.com/big-data-lab-umbc/big-data-reu/tree/main/2023-projects/team-1, [Online; Accessed: 2023-07-30].

[4] S. J. Riley, S. D. DeGloria, and R. Elliot, "Index that quantifies topographic heterogeneity," *intermountain Journal of sciences*, vol. 5, no. 1-4, pp. 23–27, 1999.

[5] M. Morlighem, E. Rignot, J. Mouginot, H. Seroussi, and E. Larour, "Deeply incised submarine glacial valleys beneath the greenland ice sheet," *Nature Geoscience*, vol. 7, no. 6, pp. 418–422, 2014.

[6] M. B. Lythe and D. G. Vaughan, "BEDMAP: Anew ice thickness and subglacial topographic model of Antarctica." *J. Geophys. Res.*, vol. 106 (B6), pp. 11,335–11,351, 2001.

[7] P. e. a. Fretwell, "Bedmap2: improved ice bed, surface and thickness datasets for Antarctica," *Cryosphere*, vol. 7, no. 1, pp. 375–393, 2013.

[8] P. Goovaerts, "Geostatistical software," in *Handbook of applied spatial analysis: Software tools, methods and applications*. Springer, 2009, pp. 125–134.

[9] M. Liu-Schiaffini, G. Ng, C. Grima, and D. Young, "Ice thickness from deep learning and conditional random fields: application to ice-penetrating radar data with radiometric validation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[10] C. Schoof, "Ice sheet grounding line dynamics: Steady states, stability, and hysteresis," *J. Geophys. Res.*, vol. 112, no. F03S28, pp. 1–19, JUL 14 2007.

[11] I. Howat, A. Negrete, 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

[12] 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.

[13] 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. (2021) Atlas/icesat-2 l3a land ice height, version 5. [Online]. Available: https://nsidc.org/data/ATL06/versions/5

[14] J. M. V. Wessem and M. K. Laffin. (2020, Feb) Regional atmospheric climate model (racmo2), version 2.3p2. [Online]. Available: https://doi.org/10.5281/zenodo.3677642

[15] J. A. MacGregor, L. N. Boisvert, B. Medley, A. A. Petty, J. P. Harbeck, R. E. Bell, J. B. Blair, E. Blanchard-Wrigglesworth, E. M. Buckley, M. S. Christoffersen *et al.*, "The scientific legacy of nasa's operation icebridge," 2021.

[16] C. Williams and C. Rasmussen, "Gaussian processes for regression," *Advances in neural information processing systems*, vol. 8, 1995.

[17] R. Webster and M. A. Oliver, *Geostatistics for environmental scientists*. John Wiley & Sons, 2007.

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.