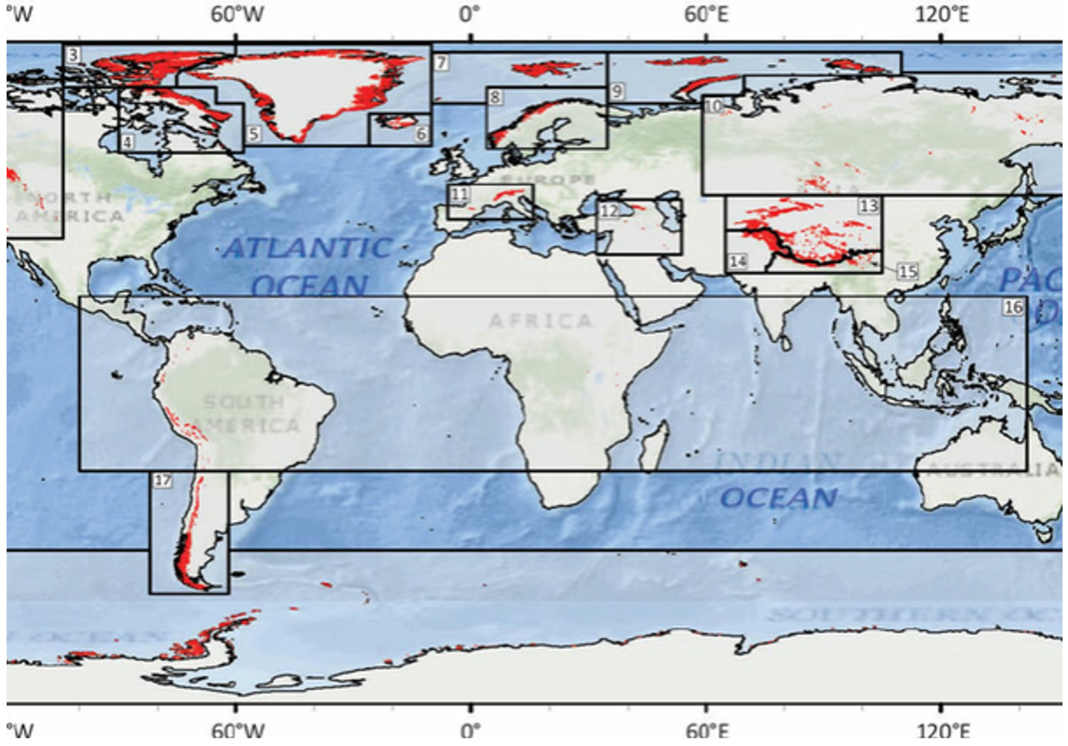
Predicting Glacier Thickness using Machine Learning

•Emma Hvid Møller, Marcus Benjamin Newmann & Cerina von Bruhn

•Course: Applied Machine Learning

•Dato: 12.06.2024

- **Goal:** Predict the thickness of glaciers
- **Motivation:** Image analysis and something completely different from what we are used to work with
- **What we have done**: Generated images of of glaciers, created input variable from a CAE, combined it with the tabular data, and used it to make an regression model.
- **Main focus**: Cleaning data, generating images and developing a Convalutional Auto Encoder



Data

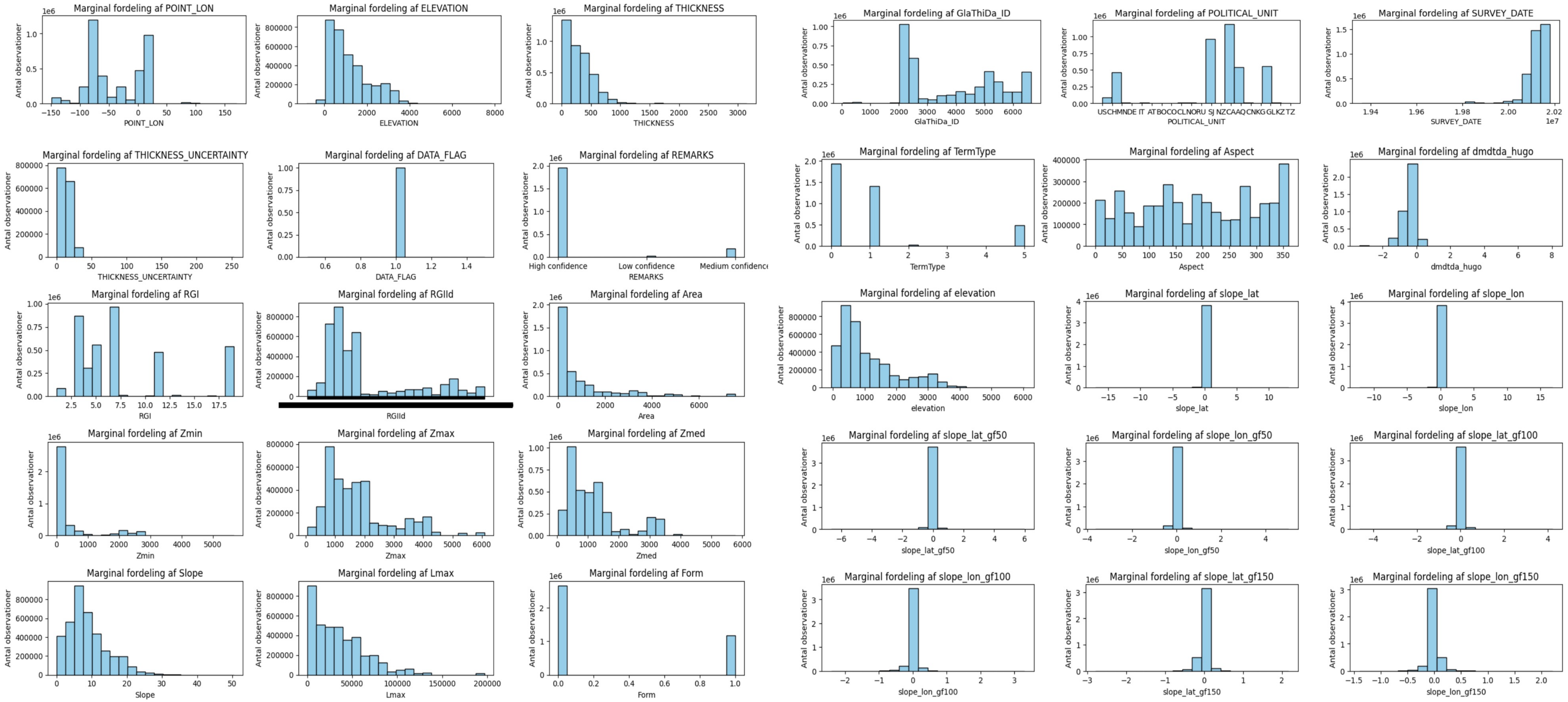
- **Dimensions:** Our data consist of 3,854,279 rows and 68 columns describing different measurements of 4,681 glaciers and ice capes.

- **Target:** Thickness

- **Features:** Survey identifier, survey date, country code, the min, max end mean elevation of the glacier, velocity, mean glacier slope, the area (km^2), the term type (e.g. land- and marine-termination) etc.

- **Missing values:** There is NA/NAN in 3,854,279 rows and 40 columns.

Marginal distributions of features



Investigation of missing values

```
Column 'GlaThiDa_ID':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'POLITICAL_UNIT':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'GLACIER_NAME':
Occurrences of 'na': 0
Occurrences of 'nan': 1402056

Column 'SURVEY_DATE':
Occurrences of 'na': 0
Occurrences of 'nan': 44

Column 'PROFILE_ID':
Occurrences of 'na': 0
Occurrences of 'nan': 1463774

Column 'POINT_ID':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'POINT_LAT':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'POINT_LON':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'ELEVATION':
Occurrences of 'na': 0
Occurrences of 'nan': 481445

Column 'THICKNESS':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'THICKNESS_UNCERTAINTY':
Occurrences of 'na': 0
Occurrences of 'nan': 2329692

Column 'DATA_FLAG':
Occurrences of 'na': 0
Occurrences of 'nan': 3854278

Column 'REMARKS':
Occurrences of 'na': 0
Occurrences of 'nan': 1686043

Column 'RGI':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'RGIId':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Area':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Zmin':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Zmax':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Zmed':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Slope':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Lmax':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Form':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'TermType':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Aspect':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'dmdtda_hugo':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'elevation':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lat':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'curv_300':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'curv_gfa':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'aspect_50':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'aspect_300':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'aspect_gfa':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'smb':
Occurrences of 'na': 0
Occurrences of 'nan': 2950

Column 'ith_m':
Occurrences of 'na': 0
Occurrences of 'nan': 291380

Column 'vx':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gf50':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gf100':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gf150':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gf300':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gf450':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gfa':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gf50':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gf100':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gf150':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gf300':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gf450':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gfa':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'dvx_dx':
Occurrences of 'na': 0
Occurrences of 'nan': 27409

Column 'dvx_dy':
Occurrences of 'na': 0
Occurrences of 'nan': 26968

Column 'dvy_dx':
Occurrences of 'na': 0
Occurrences of 'nan': 27409

Column 'dvy_dy':
Occurrences of 'na': 0
Occurrences of 'nan': 26968

Column 'dist_from_border_km_geom':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'ith_f':
Occurrences of 'na': 0
Occurrences of 'nan': 282471
```

Link between missing values and how we dealt with them

- **DATA_FLAG:** Missing not at random.
  " *Erroneous data will have non-Nan here".*
  Removed rows with non-Nan and afterwards removed the column.
- **ELEVATION:** Missing completely at random.
  Use another variable "elevation" with interpolated values for the missing ELEVATION values.
- **THICKNESS = 0**: It can seems odd that a glacier or ice cape has a thickness of 0. We observe a correlation with "REMARKS" and "THICKNESS_UNCERTAINTY". We erased the rows with thickness=0.
- **Surface velocity and velocity derivatives:** (vx, vy, …, dvy_dx, dvy_dy) all missing the same 20,608. Couldn't find a link to other features.
- And many more…

THICKNESS	0	1	2	3	4	5	6	7	8	3147	3150	3151
THICKNESS_UNCERTAINTY												
0.0	55	117	131	210	239	242	290	275	337	0	0	0
1.0	0	0	6	12	3	114	105	39	32	0	0	0
2.0	467	0	10	11	105	175	489	538	506	0	0	0
3.0	14	125	102	88	85	84	109	68	85	0	0	0
4.0	0	0	18	45	42	15	14	11	16	0	0	0
5.0	19	36	66	108	109	146	86	217	189	0	0	0
6.0	0	0	0	0	0	0	1	0	0	0	0	0
7.0	0	0	0	0	0	0	0	0	0	0	0	0
8.0	1756	197	59	31	30	38	39	36	35	1	1	1
9.0	1	0	0	0	0	0	0	0	0	0	0	0
10.0	0	0	0	0	0	0	0	0	0	0	0	0
11.0	0	0	0	0	0	0	0	0	0	0	0	0
12.0	3	1	0	0	1	0	0	0	0	0	0	0
13.0	0	0	0	0	0	0	0	0	0	0	0	0
14.0	0	0	0	0	0	0	0	0	0	0	0	0
15.0	0	0	0	0	0	0	0	0	3	0	0	0
16.0	0	0	0	0	0	0	0	0	0	0	0	0
17.0	0	0	0	0	0	0	0	0	0	0	0	0
18.0	28366	55	51	35	45	48	49	42	73	0	0	0
19.0	0	0	0	0	0	0	0	0	0	0	0	0
20.0	0	0	0	0	0	0	0	0	0	0	0	0

REMARKS	High confidence	Low confidence	Medium confidence
THICKNESS			
0	159228	1422	11215
1	881	6	96
2	865	9	49
3	763	7	55
4	695	12	35
...	...	...	...
2255	1	0	0
2259	1	0	0
2263	1	0	0
2265	1	0	0
2268	1	0	0

New variables and impact encoding

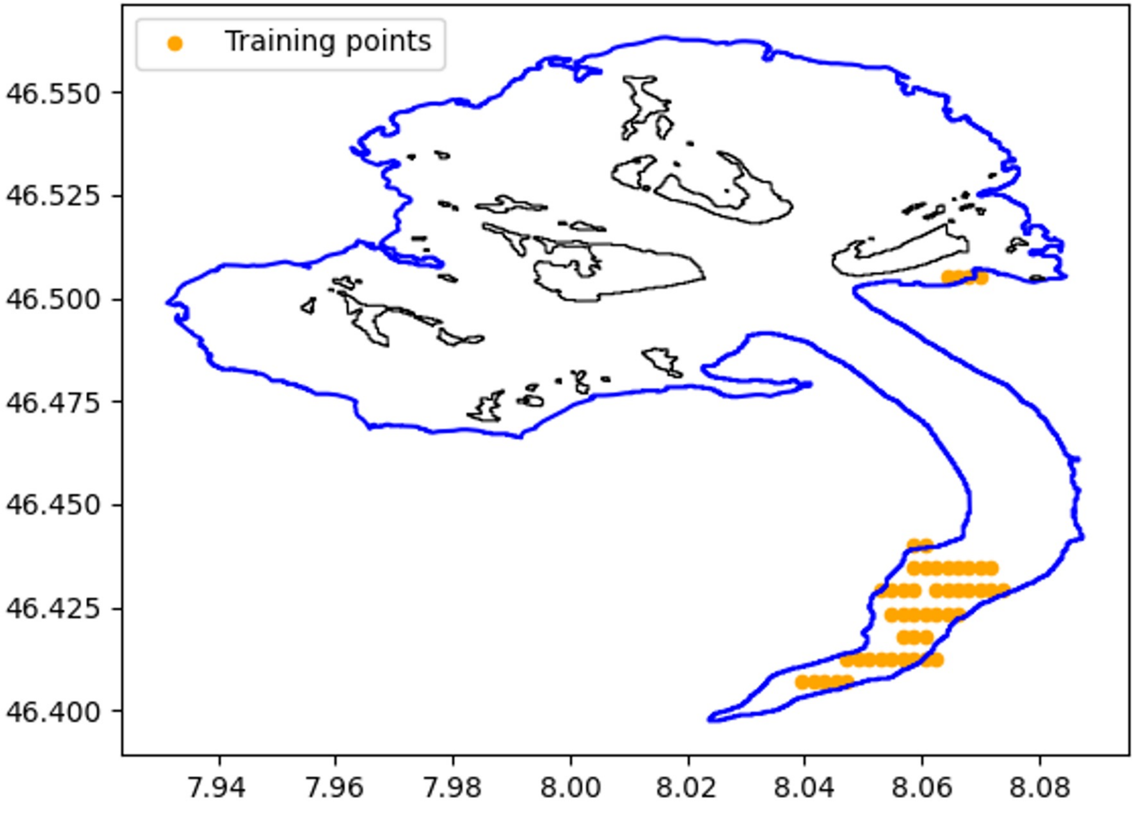
- **The module of the slop:** $$||s|| = \sqrt{slope_{lat}^2 + slope_{lon}^2}$$
- **Local longitude :** $$local_{lat} = \left|point_{lat} - \mu(point_{lat})\right|$$
- **Loca latitude:** $$local_{lon} = \left|point_{lon} - \mu(point_{lon})\right|$$
- **Impact encoding of character variables:**

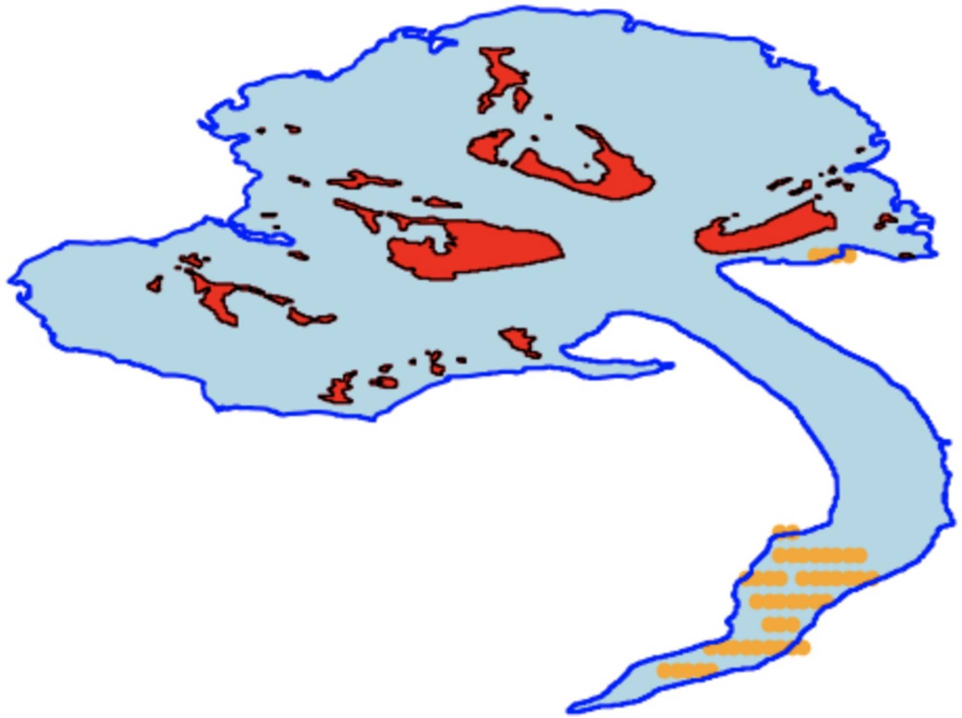
  A technique for converting categorical variables into numerical values based on the target variable we are trying to predict.

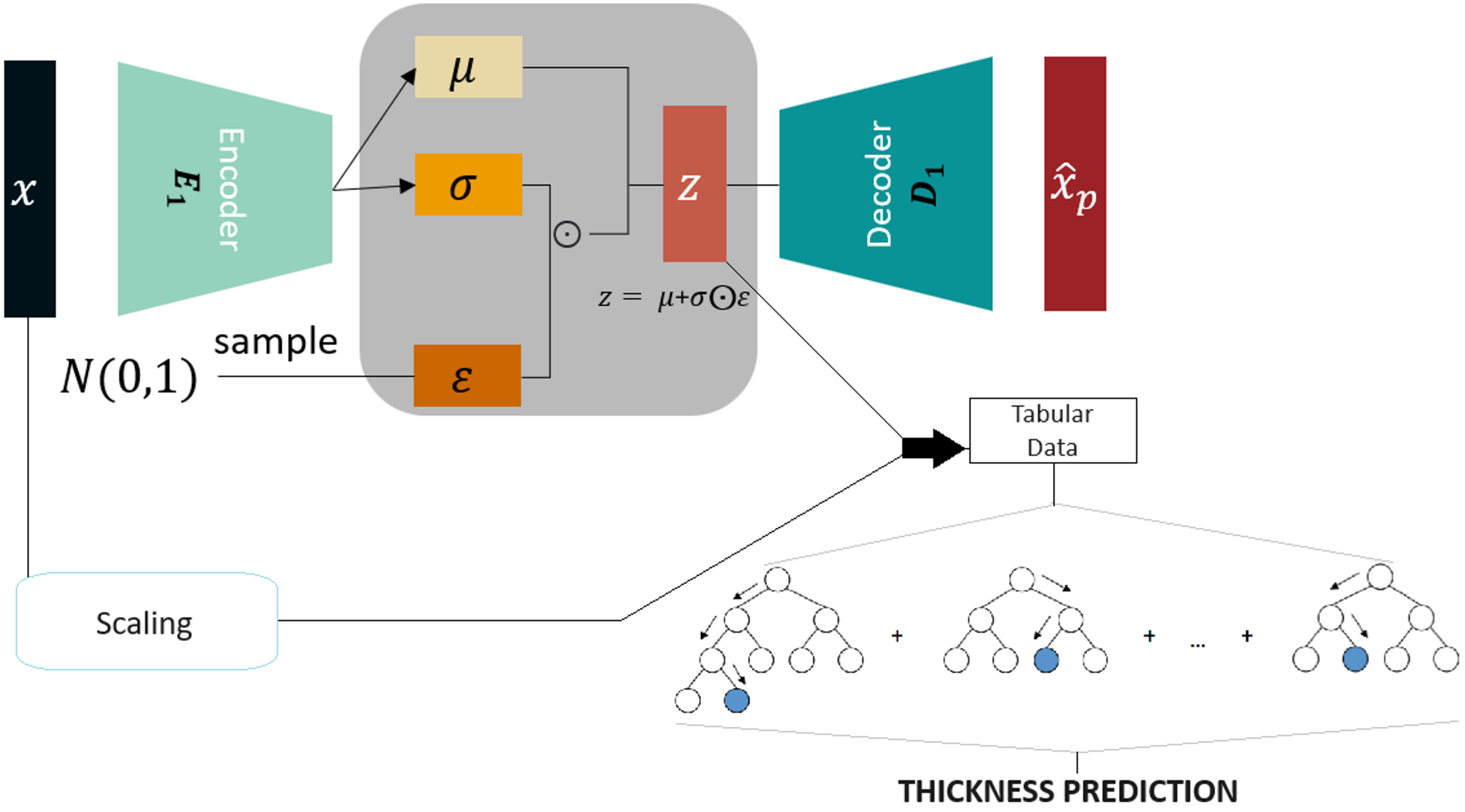
  Calculates the mean of thickness for each category in the characteristic variable and replace it with the calculated mean.

Generating images

- Based on the dataset, one can generate plots of the glaciers
- Images created by removing axes, color nunataks red, glacier blue and scaled to 64x64 pixels

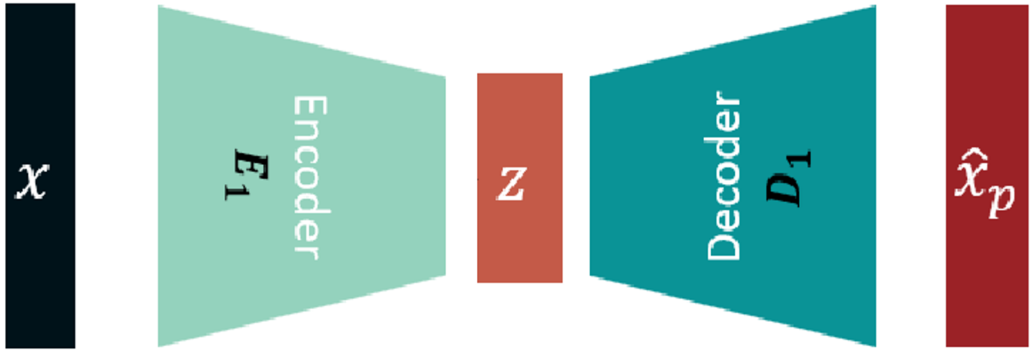


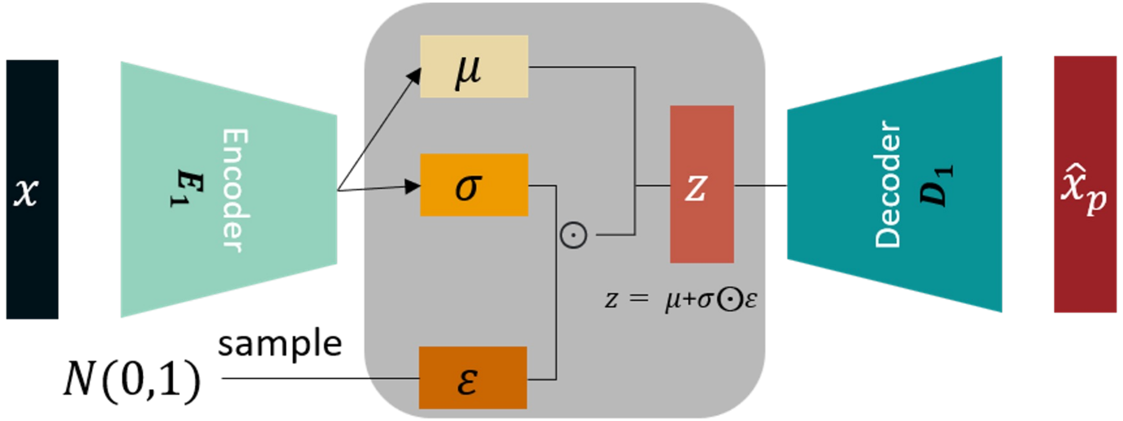




x
Encoder
E1
μ
σ
ε
z
z = μ+σ⊙ε
Decoder
D1
x̂p
sample
N(0,1)
Tabular Data
Scaling
+
...
THICKNESS PREDICTION

Model



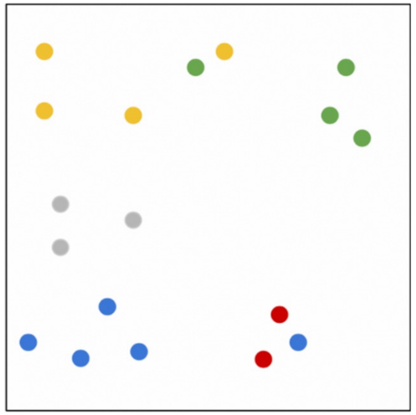


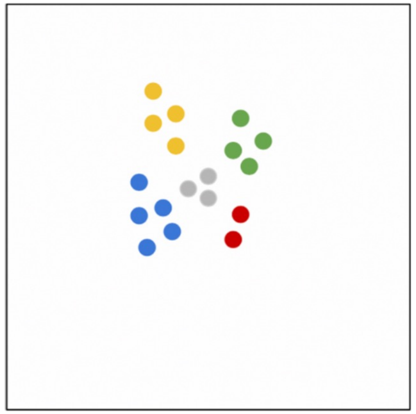
Loss

Loss=
Reconstruction

Loss=
Reconstruction + **w** * KL_div

Latent Space





Hyperparameters for VAE

To Tune

- Learning Rate
- Conv/MaxPool Layers
- Dense Layers
- Filters
- Activation Functions

Not To Tune

- KL Weight
- Latent Space Dimension (24 / 32 / 48 / 64)

Autoencoder training times and CUDA

CPU:

```
Epoch 1/30
  3/276 [..............................] - ETA: 45:23
```

GPU:

Autoencoder training times and CUDA

CPU:

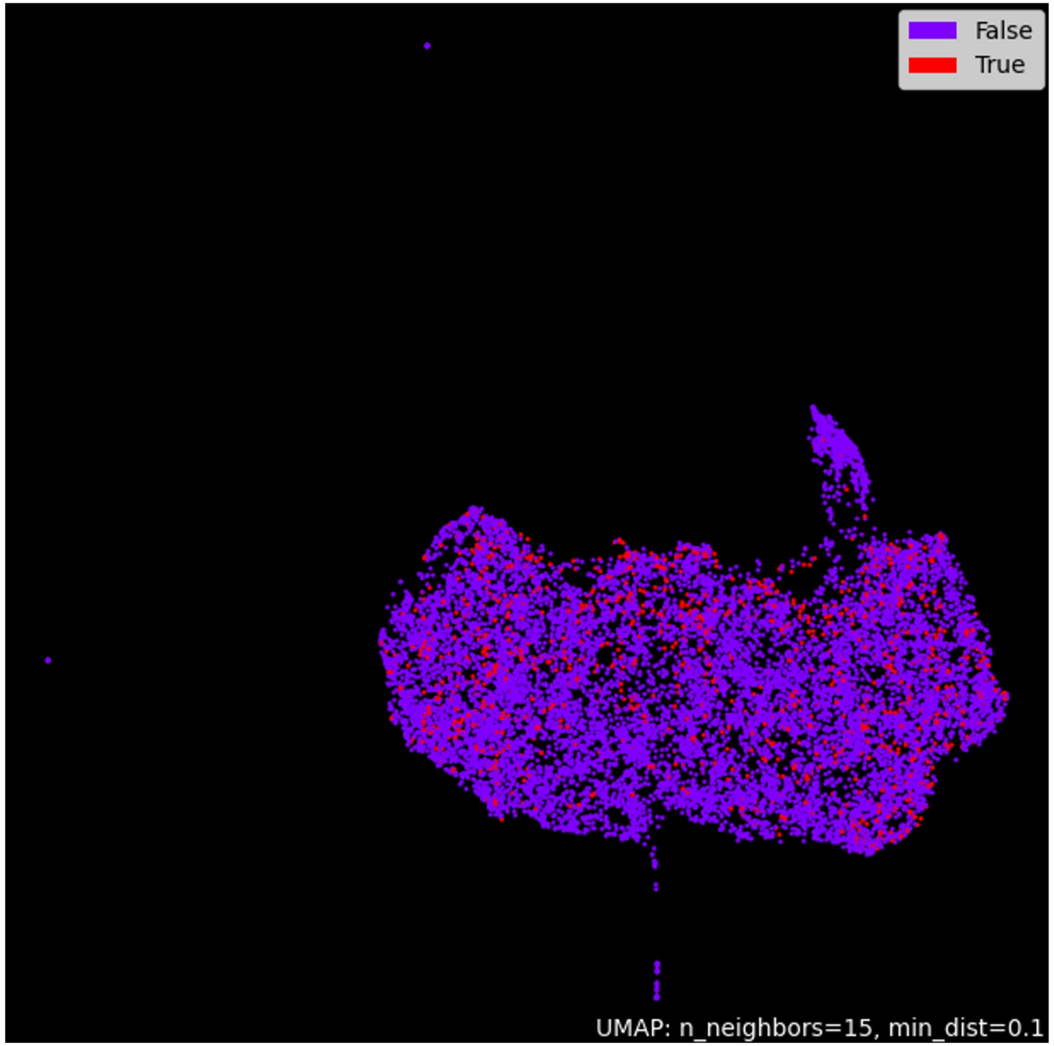
```
Epoch 1/30
  3/276 [..............................] - ETA: 45:23
```

GPU:

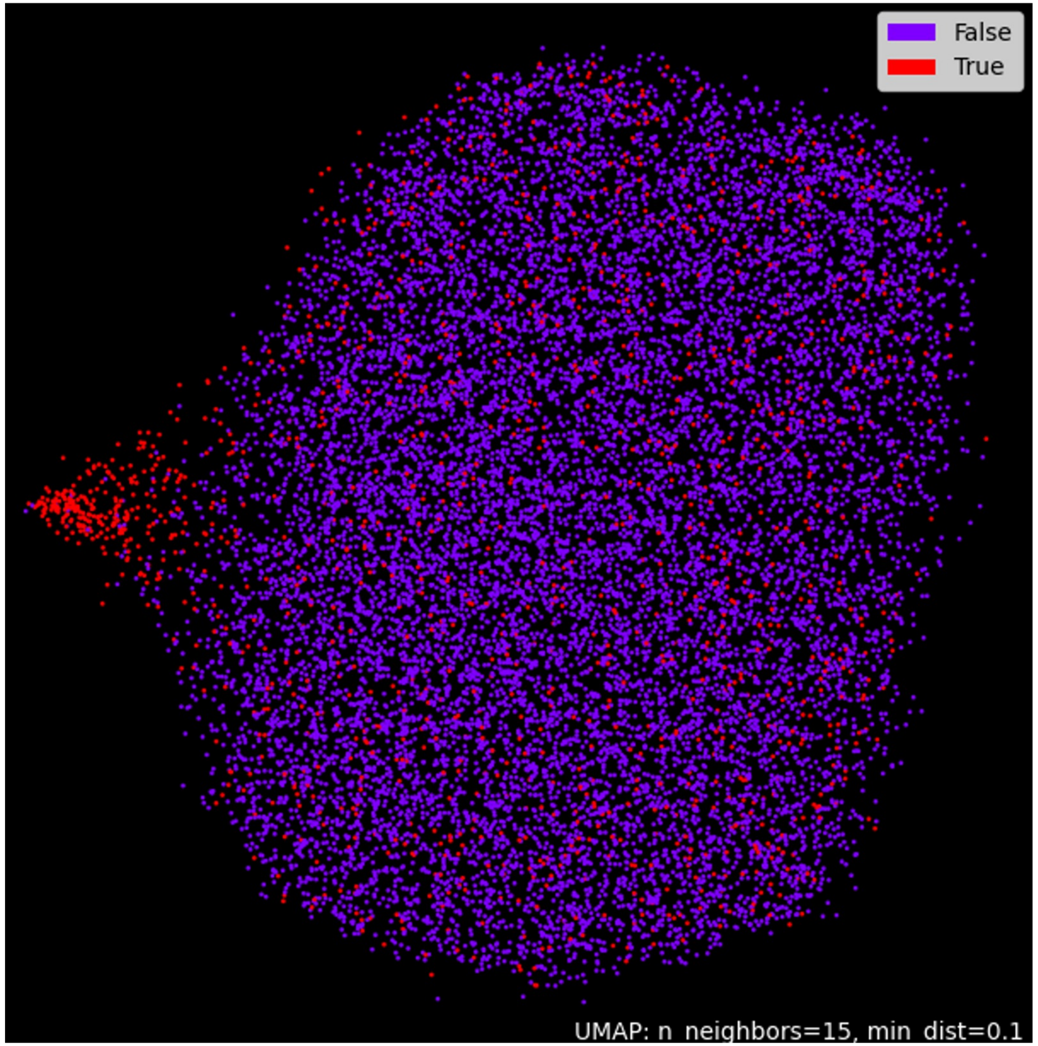
```
Epoch 1/30
2264/2264 [==============================] - 17s
```

2 dimensional UMAP representation of LP

Autoencoder

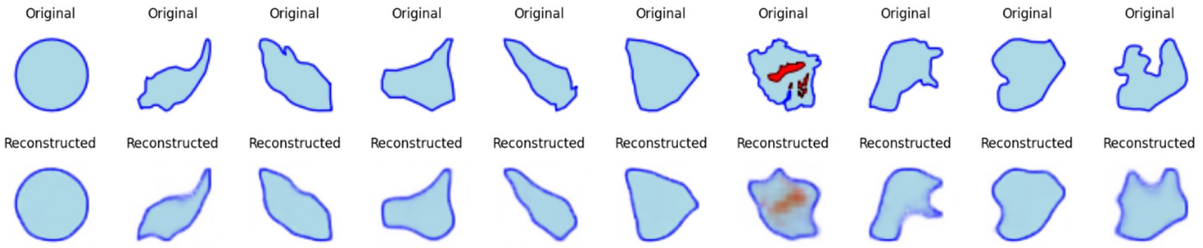


Variational Autoencoder

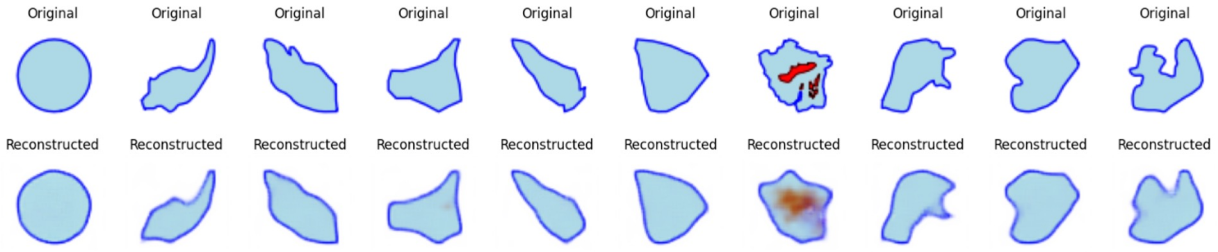


Latent space dimensions and reconstruction

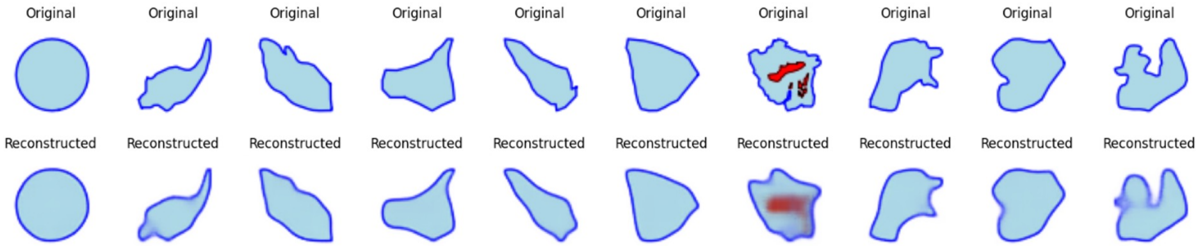
24 Dimensions



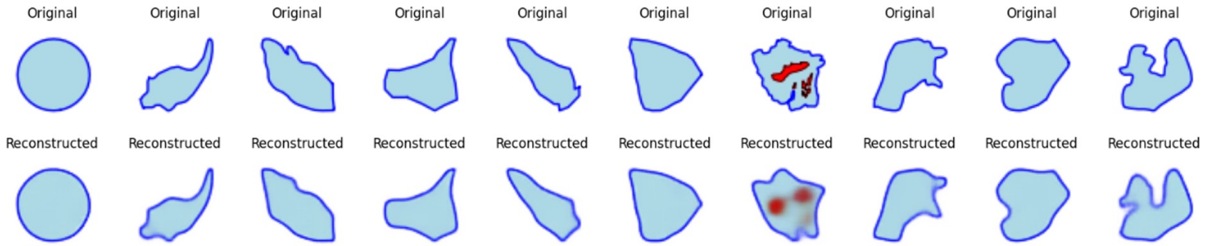
32 Dimensions



48 Dimensions



64 Dimensions



Regression models

01 Random Forest

- Robustness and Stability
  - Average the results of many decision trees, reducing the risk of overfitting
- Handling of Missing Data
  - By using the median imputation strategy and by building trees based on different subsets of data

02 Gradient Boosted Decision Tree

- High Predictive Accuracy
  - Builds trees sequentially, each correcting the errors of the previous ones
- Control Overfitting
  - Regularization parameters (learning rate and tree depth)

03 XGBoost

- Efficiency and Speed
  - Parallel processing
- Robustness to Overfitting
  - Tree pruning and cross-validation

Regression models

Train/Test

Split data up into test and train

Optuna

Optimization of hyperparameters

Cross-validation

K-fold with early stopping rounds to prevent overfitting

Minimize MAE

Choosing loss function: Mean absolute error

Challenges

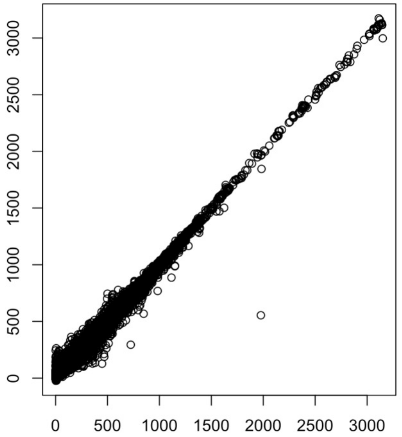
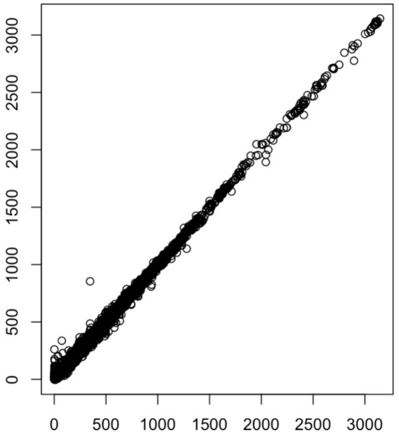
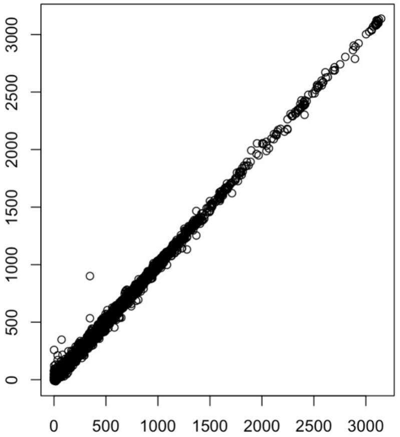
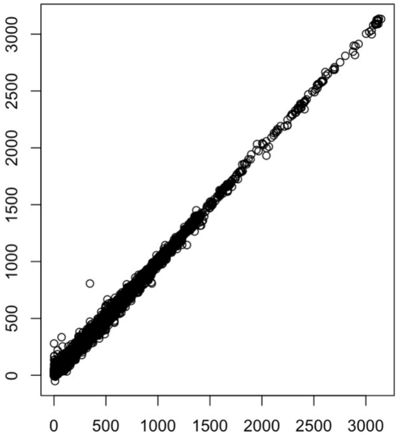
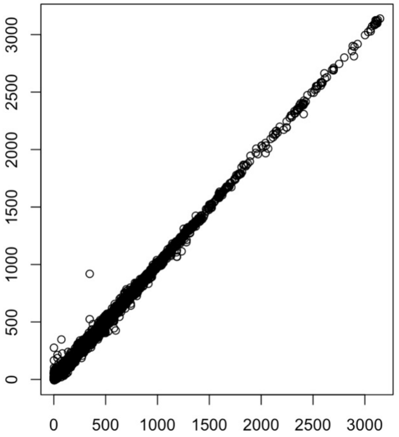
Lack of computer power

Lack of storage

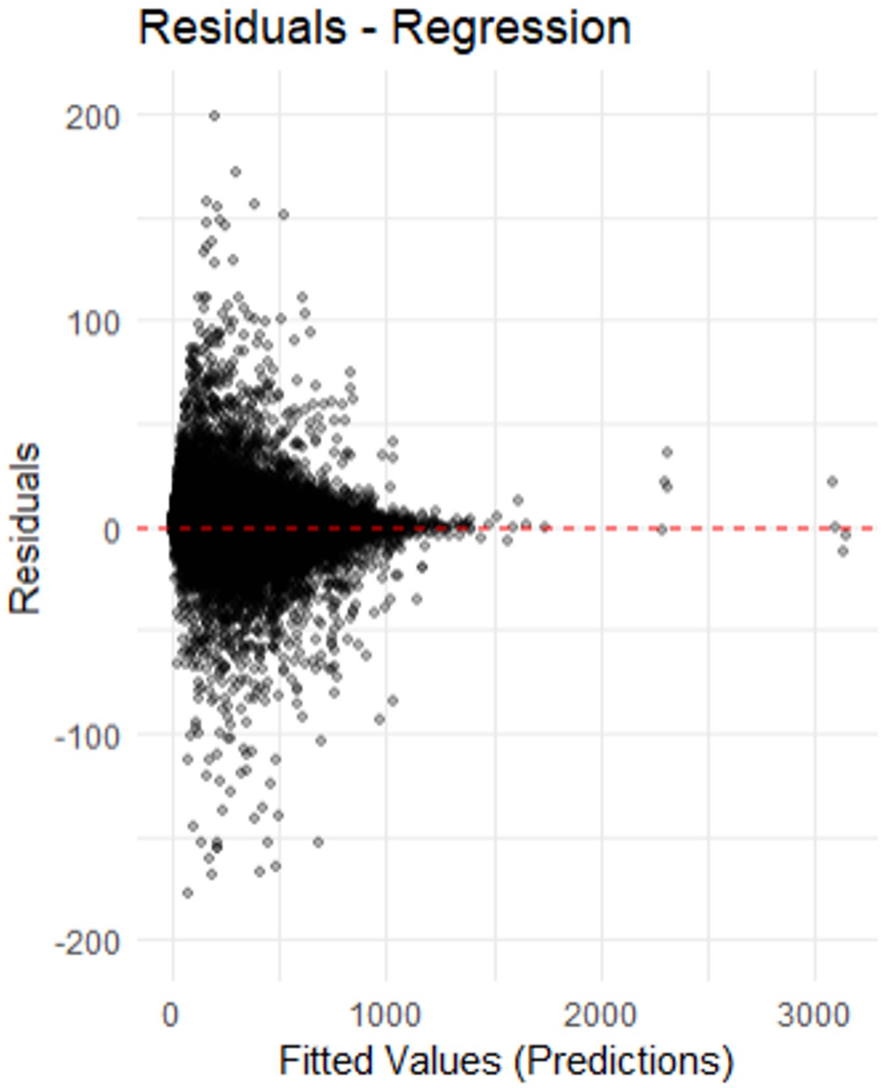
$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} \left| Y_i - \hat{Y}_i \right|$$

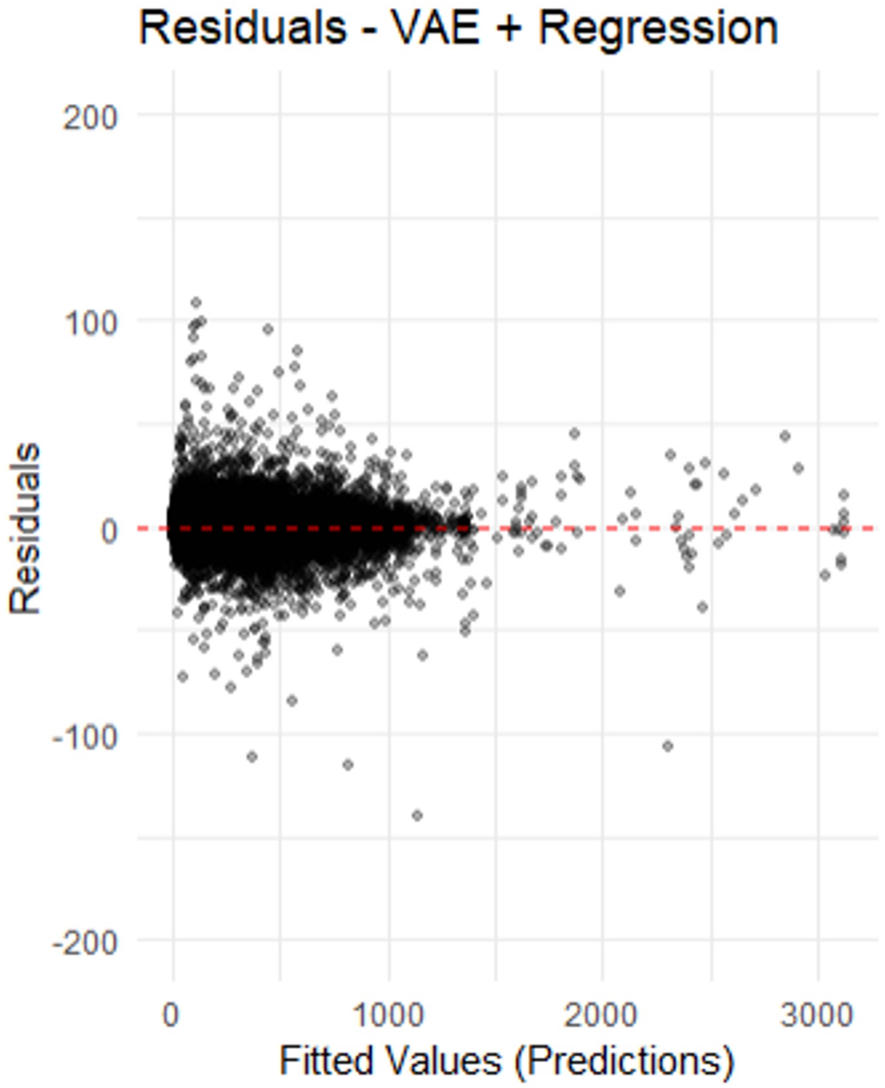
Model:	**Without latent space**	**Latent space w. 24 dimensions**	**Latent space w. 32 dimensions**	**Latent space w. 48 dimensions**	**Latent space w. 64 dimensions**
Random Forest	3.1400	2.3574	2.3562	2.6222	2.3436
Gradient Boosted Decision Tree	12.4767	7.3917	6.1224	7.6633	7.2676
XGBoost	5.5575	<u>2.3376</u>	2.4447	11.1353	2.6211

Predictions vs. target variable for XGBoost

Model:	Without latent space	Latent space w. 24 dimensions	Latent space w. 32 dimensions	Latent space w. 48 dimensions	Latent space w. 64 dimensions
XGBoost					

Residuals for XGBoost





Comparisons

181.6

MAE for ice thickness predicted by a featureless model (only an intercept).

- We calculated the MAE for the featureless model, which always uses the mean of the target variable THICKNESS as its prediction, and compared it to the actual target variable THICKNESS.

88.0

MAE for ice thickness predicted by Farinotti et al. (2019).

- Calculating the MAE for the predictions from our XGBoost model compared to the predicted thickness values by Farinotti.

77.8

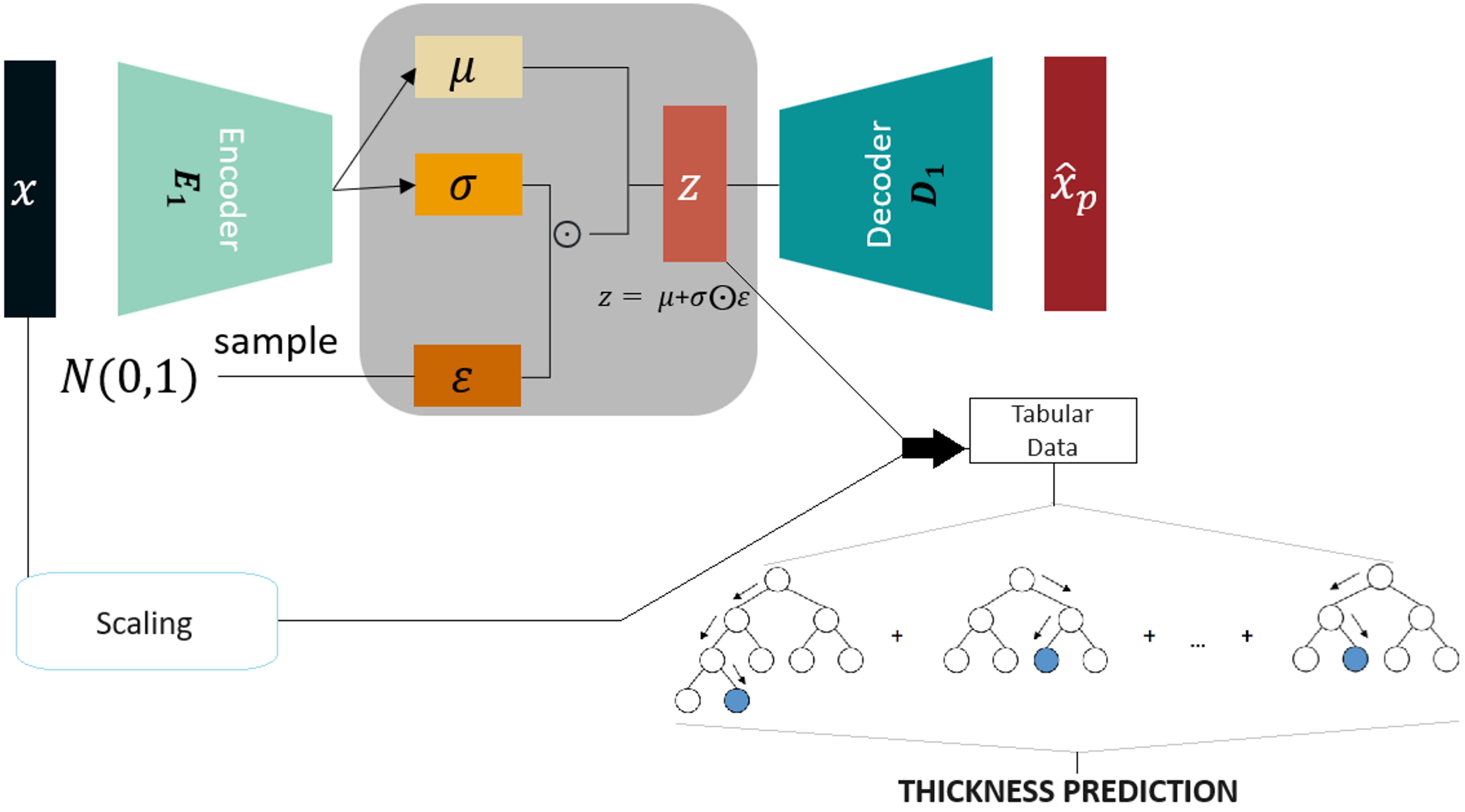
MAE for ice thickness predicted by Millan et al. (2022).

- Calculating the MAE for the predictions from our XGBoost model compared to the predicted thickness values by Millan.

Improvement ideas

- Instead of removing the feature survey_date, then keep the information about the month. Furthermore potentially add time-series data of the temperature.

- Instead of tuning the autoencoder and regression models independently, running 1 complete optimization on the entire model, including latent space dimensions and KL-weight as hyperparameters, would be optimal.


x
Encoder
E_1
μ
σ
$\odot$
z
$z = \mu + \sigma \odot \varepsilon$
Decoder
D_1
$\hat{x}_p$
sample
$N(0,1)$
ε
Tabular Data
Scaling
+
+ ... +
THICKNESS PREDICTION

- **To sum up:**
  - Generated images of of glaciers, created input variable from a CNN, combined it with the tabular data, and used it to make an regression model.

- **Conclusion:** Better result with latent space data from the CAE

- **Relevance:**
  - It can be difficult to measure the thickness of glaciers due to the rough environment. Climate research, sea level rise etc.

And lastly, a huge thanks to Niccolo!

Questions

Appendix I

	Random forest	Gradient boosted decision tree	XGBoost
Input features	See appendix VII.	All except the ones with least corr.	All except the ones with least corr.
HP Optimization	Naive approach	Optuna	Optuna
Hyperparameters	n_trees: 100	'lambda_l1': 0.3872671475587192, 'lambda_l2': 0.1347211582956847, 'num_leaves': 208, 'feature_fraction': 0.42307038814, 'bagging_fraction': 0.9960229027, 'bagging_freq': 7, 'min_child_samples': 37, 'learning_rate': 0.09414611653869	'lambda': 0.0832200463578115, 'alpha': 6.296986987802592, 'colsample_bytree': 0.6995472160, 'subsample': 0.7687278029948124, 'learning_rate': 0.0323799185617, 'n_estimators': 537, 'max_depth': 15, 'min_child_weight': 10, 'gamma': 0.0803458919901354
MAE	2.3436	7.2676	2.6211
Run time (HP optim. time + training time)	~42 min	~ 65 min	~ 85 min

Appendix II - Feature selection

- Many feature in the metadata-dataset was describing the same, but measured in different width (in meters) of a gaussian filter. e.g. ‘aspect’, ‘aspect_50’, ’aspect_300’, ‘aspect_gfa’. We pick we only used one of each

- Many of the features where ID features. We dident used that for the model

- Removed features with high uncertainty (e.g. survey date with a lot of 99-99-9999 values)
- Fill in NAs as ‘unknown/non’ in ‘REMARKS’

Appendix III - Input features - Metadata

- x_scale (Scales from images)
- y_scale (Scales from images)

- RGI
- Area
- Zmin
- Zmax
- Zmed
- form
- remarks_encoded
- Slope
- Lmax
- Termtype_encoded

- dmdtda_hugo
- elevation
- slope_total
- aspect
- curv_gfa
- smb
- vx
- vy
- dvx_dx
- dvy_dy
- dvx_dy
- dvy_dx
- dist_from_border_km_geom

Appendix IV - Input features - Latent space w. 24 dimensions

- x_scale (Scales from images)
- y_scale (Scales from images)

- RGI
- Area
- Zmin
- Zmax
- Zmed
- form
- remarks_encoded
- Slope
- Lmax
- Termtype_encoded
- dmdtda_hugo
- elevation
- slope_total
- aspect

- curv_gfa
- smb
- vx
- vy
- dvx_dx
- dvy_dy
- dvx_dy
- dvy_dx
- dist_from_border_km_geom

- dim_0
- dim_1
- …
- dim_23
- dim_24

Appendix V - Input features - Latent space w. 32 dimensions

- x_scale (Scales from images)
- y_scale (Scales from images)

- RGI
- Area
- Zmin
- Zmax
- Zmed
- form
- remarks_encoded
- Slope
- Lmax
- Termtype_encoded
- dmdtda_hugo
- elevation
- slope_total
- aspect

- curv_gfa
- smb
- vx
- vy
- dvx_dx
- dvy_dy
- dvx_dy
- dvy_dx
- dist_from_border_km_geom

- dim_0
- dim_1
- …
- dim_31
- dim_32

Appendix VI - Input features - Latent space w. 48 dimensions

- x_scale (Scales from images)
- y_scale (Scales from images)

- RGI
- Area
- Zmin
- Zmax
- Zmed
- form
- remarks_encoded
- Slope
- Lmax
- Termtype_encoded
- dmdtda_hugo
- elevation
- slope_total
- aspect

- curv_gfa
- smb
- vx
- vy
- dvx_dx
- dvy_dy
- dvx_dy
- dvy_dx
- dist_from_border_km_geom

- dim_0
- dim_1
- …
- dim_47
- dim_48

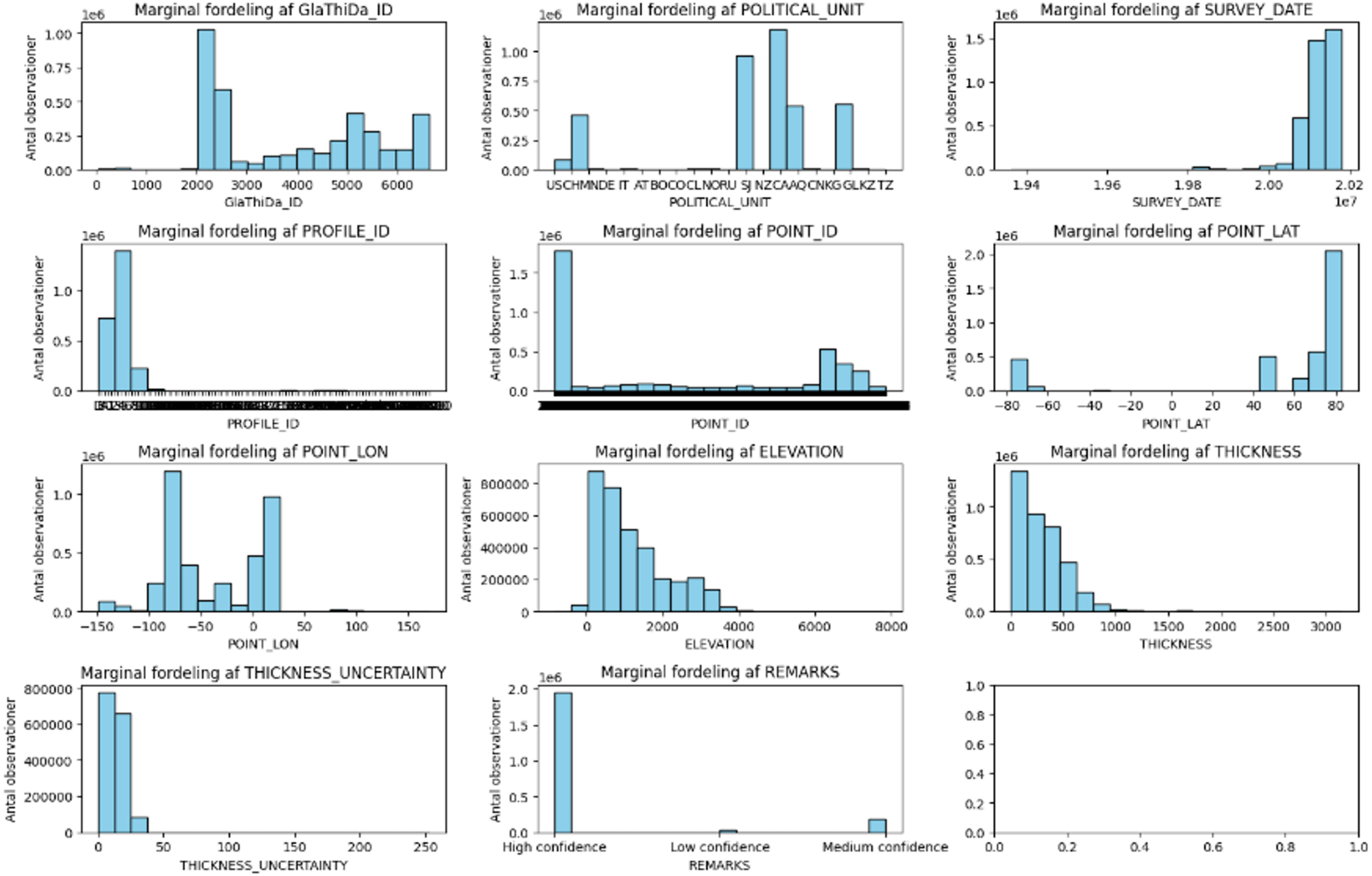
Appendix VII - Input features - Latent space w. 64 dimensions

- x_scale (Scales from images)
- y_scale (Scales from images)

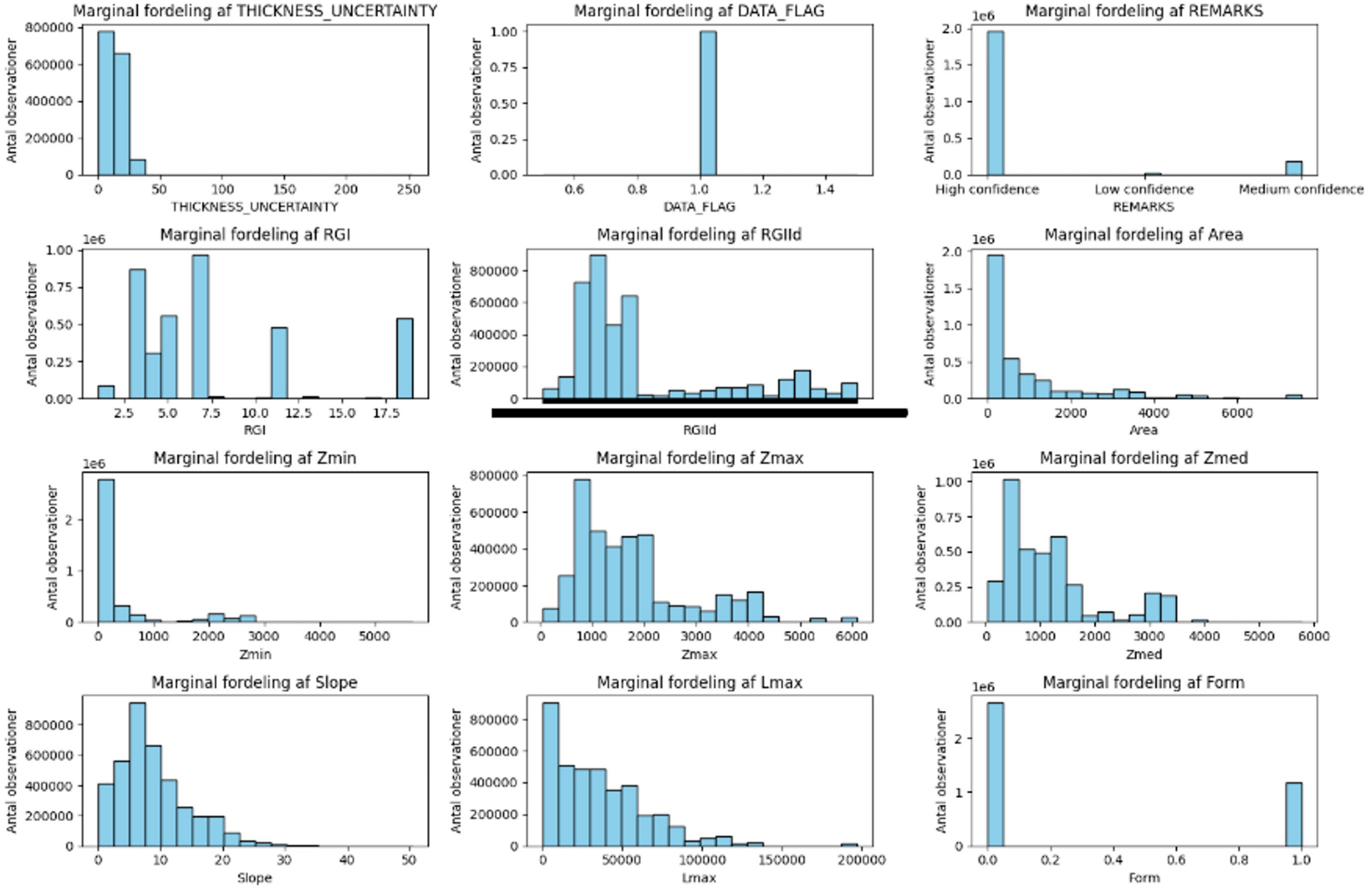
- RGI
- Area
- Zmin
- Zmax
- Zmed
- form
- remarks_encoded
- Slope
- Lmax
- Termtype_encoded
- dmdtda_hugo
- elevation
- slope_total
- aspect
- curv_gfa
- smb
- vx
- vy
- dvx_dx
- dvy_dy
- dvx_dy
- dvy_dx
- dist_from_border_km_geom

- dim_0
- dim_1
- …
- dim_63
- dim_64

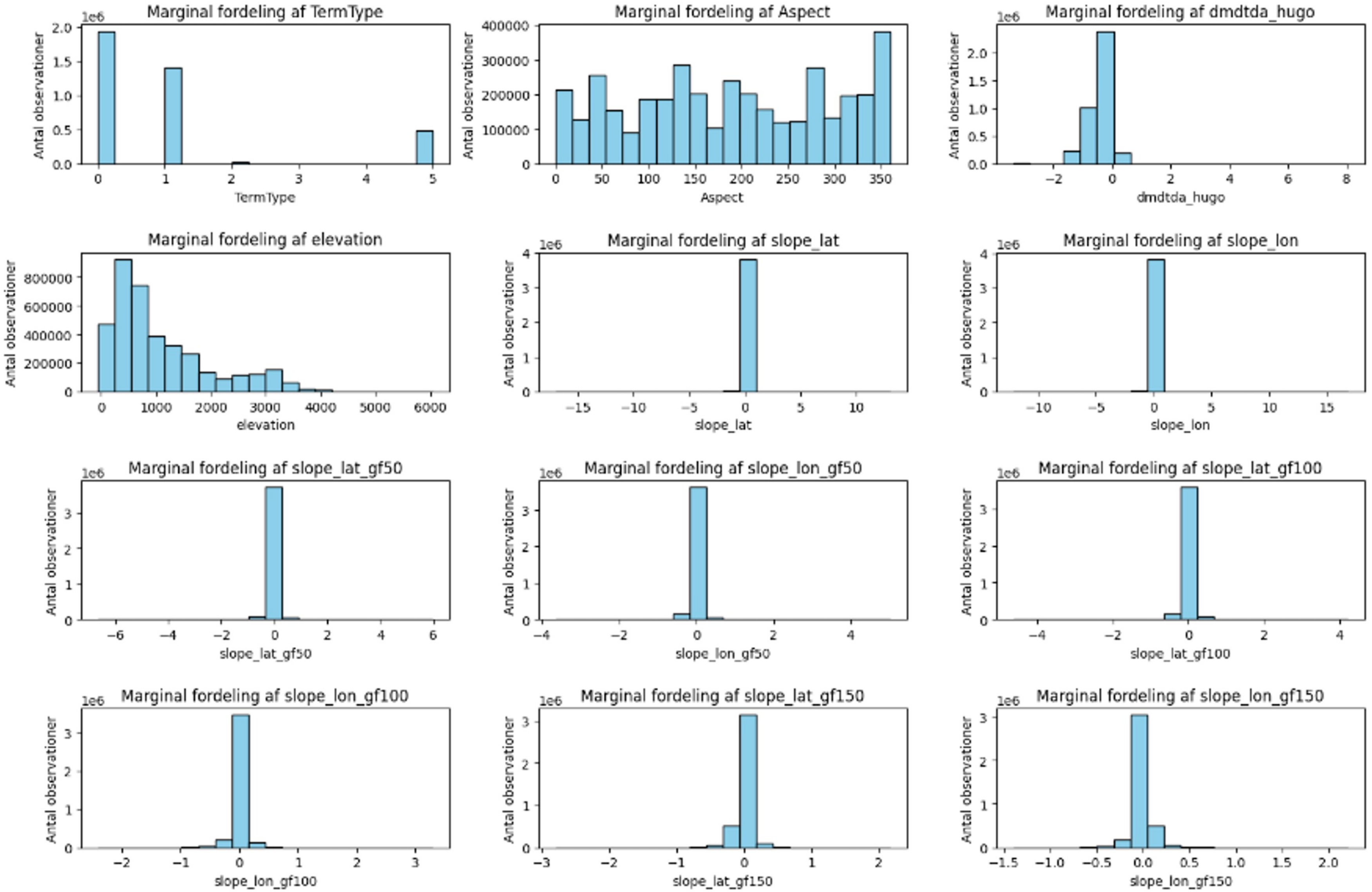
Appendix VIII - Marginal distributions



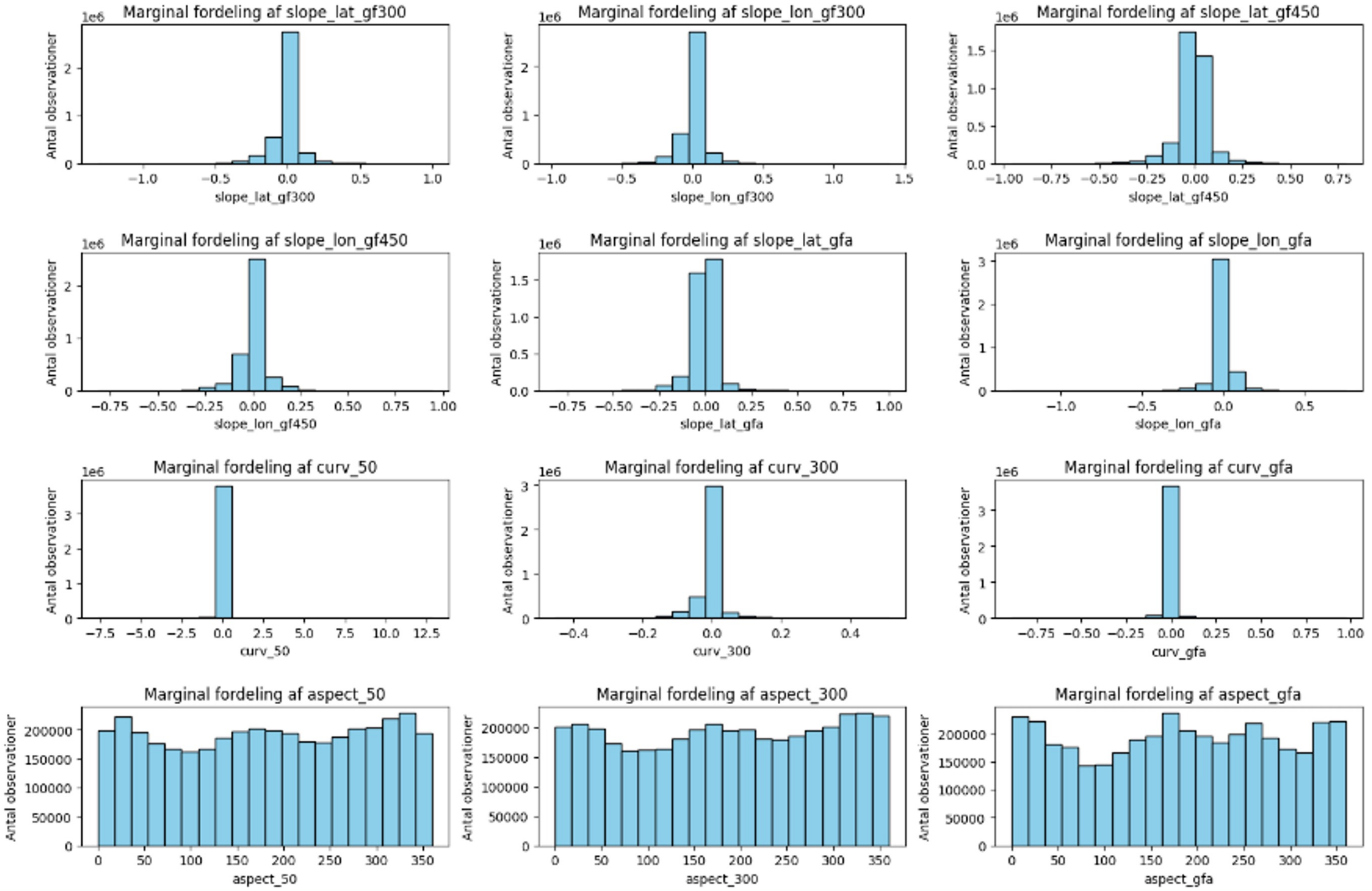
Appendix IX - Marginal distributions



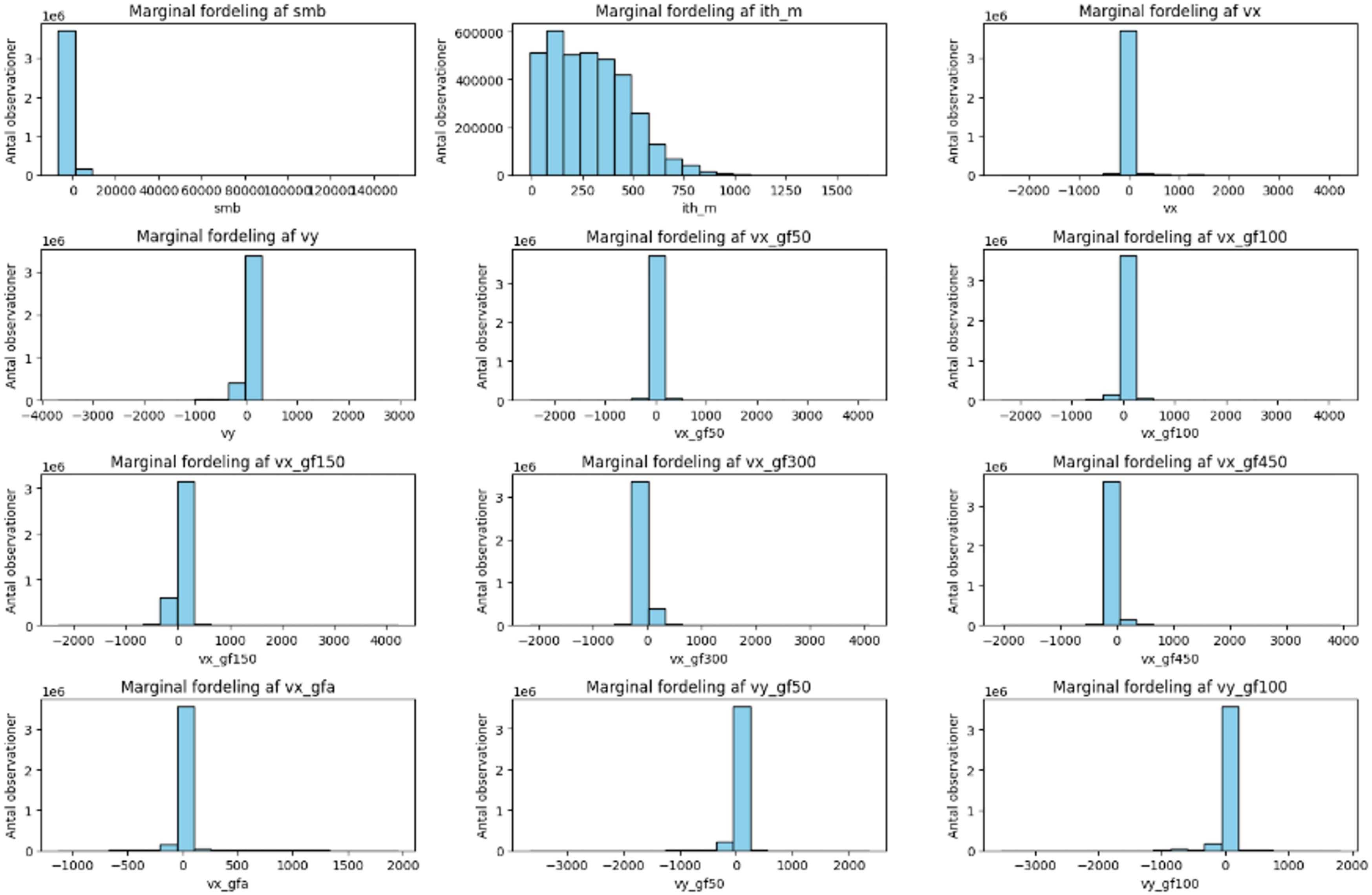
Appendix X - Marginal distributions



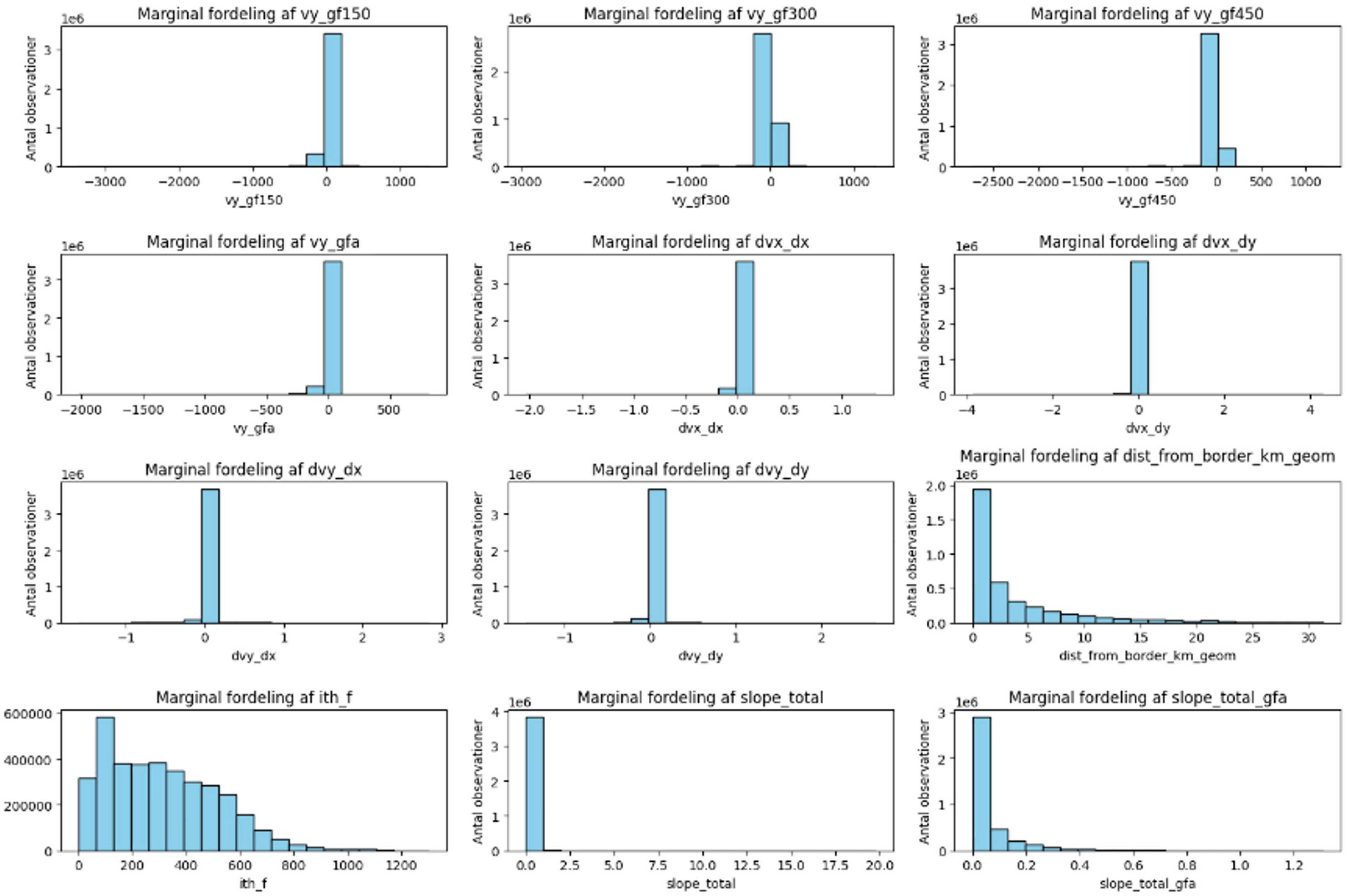
Appendix XI - Marginal distributions



Appendix XII - Marginal distributions



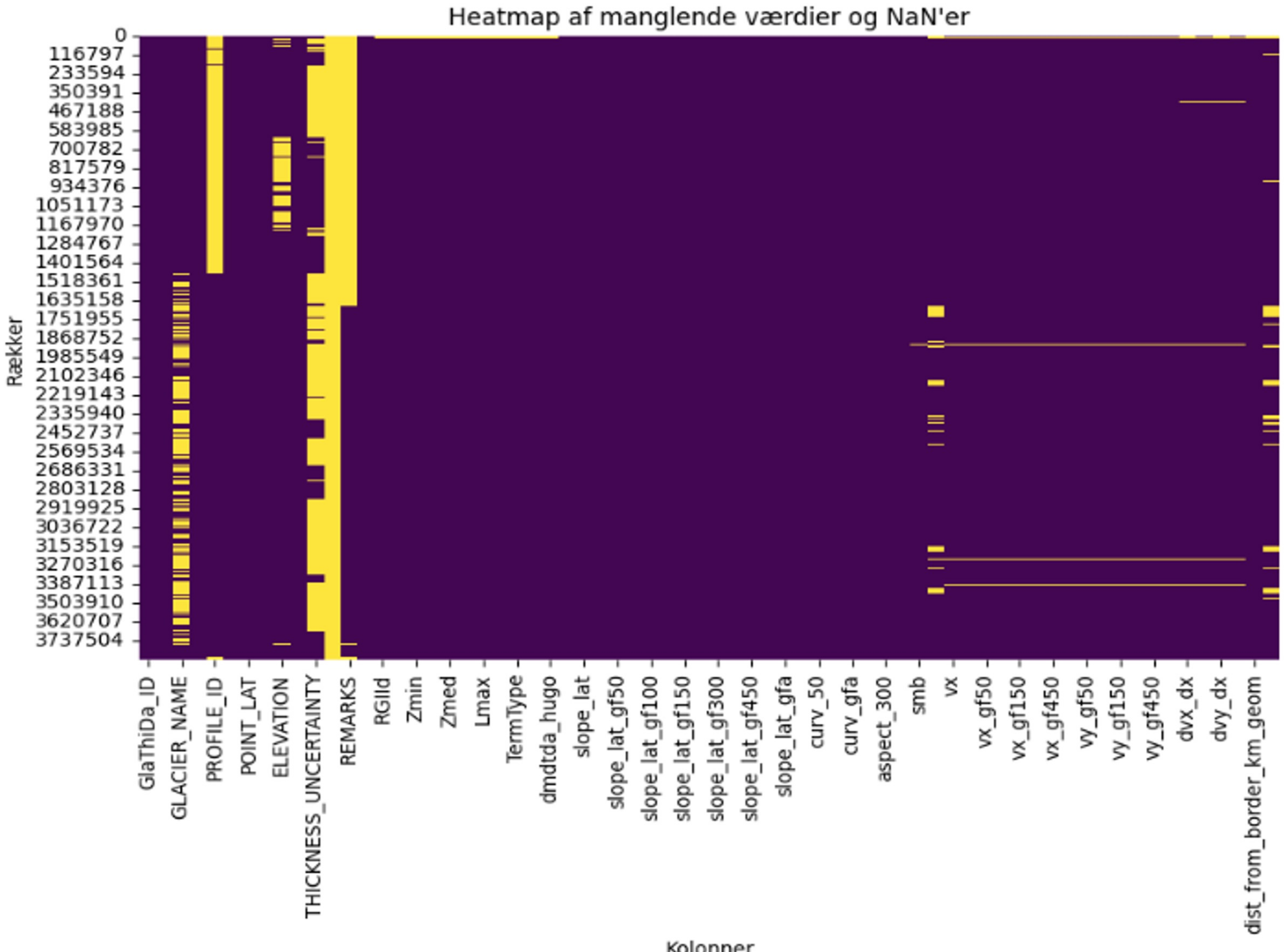
Appendix XIII - Marginal distributions



Appendix XIV - Missing values

```
Column 'GlaThiDa_ID':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'POLITICAL_UNIT':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'GLACIER_NAME':
Occurrences of 'na': 0
Occurrences of 'nan': 1402056

Column 'SURVEY_DATE':
Occurrences of 'na': 0
Occurrences of 'nan': 44

Column 'PROFILE_ID':
Occurrences of 'na': 0
Occurrences of 'nan': 1463774

Column 'POINT_ID':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'POINT_LAT':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'POINT_LON':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'ELEVATION':
Occurrences of 'na': 0
Occurrences of 'nan': 481445

Column 'THICKNESS':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'THICKNESS_UNCERTAINTY':
Occurrences of 'na': 0
Occurrences of 'nan': 2329692

Column 'DATA_FLAG':
Occurrences of 'na': 0
Occurrences of 'nan': 3854278

Column 'REMARKS':
Occurrences of 'na': 0
Occurrences of 'nan': 1686043

Column 'RGI':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'RGIId':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Area':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Zmin':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Zmax':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Zmed':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Slope':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Lmax':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Form':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'TermType':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'Aspect':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'dmdtda_hugo':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'elevation':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lat':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lon':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lat_gf50':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lon_gf50':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lat_gf100':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lon_gf100':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lat_gf150':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lat_gf300':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lon_gf300':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lat_gf450':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lon_gf450':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lat_gfa':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'slope_lon_gfa':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'curv_50':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'curv_300':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'curv_gfa':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'aspect_50':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'aspect_300':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'aspect_gfa':
Occurrences of 'na': 0
Occurrences of 'nan': 0

Column 'smb':
Occurrences of 'na': 0
Occurrences of 'nan': 2950

Column 'ith_m':
Occurrences of 'na': 0
Occurrences of 'nan': 291380

Column 'vx':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gf50':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gf100':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gf150':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gf300':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gf450':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vx_gfa':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gf50':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gf100':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gf150':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gf300':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gf450':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'vy_gfa':
Occurrences of 'na': 0
Occurrences of 'nan': 20608

Column 'dvx_dx':
Occurrences of 'na': 0
Occurrences of 'nan': 27409

Column 'dvx_dy':
Occurrences of 'na': 0
Occurrences of 'nan': 26968

Column 'dvy_dx':
Occurrences of 'na': 0
Occurrences of 'nan': 27409

Column 'dvy_dy':
Occurrences of 'na': 0
Occurrences of 'nan': 26968

Column 'dist_from_border_km_geom':
Occurrences of 'na': 0
Occurrences of 'nan': 17045

Column 'ith_f':
Occurrences of 'na': 0
Occurrences of 'nan': 282471
```

Appendix XV - Correlation between missing values



Appendix XVI - Cross tabulation - missing values

```
is_na_glacier_name    False     True
is_na_remarks
False                866563  1301673
True                1585659   100383
is_na_thickness   False     True
is_na_remarks
False            565179  1603057
True             959407   726635
is_na_elevation     False    True
is_na_remarks
False            2168236       0
True             1204597  481445
is_na_profile_id     False     True
is_na_remarks
False              2168236        0
True                222269  1463773
```

```
is_na_glacier_name    False     True
is_na_profile_id
False                989037  1401468
True                1463186      588
is_na_thickness     False     True
is_na_profile_id
False              596579  1793926
True               928008   535766
is_na_data_flag   False     True
is_na_profile_id
False                 0  2390505
True                  1  1463773
is_na_elevation     False    True
is_na_profile_id
False              2387207    3298
True                985627  478147
is_na_remarks       False     True
is_na_profile_id
False              2168236   222269
True                     0  1463774
```

```
is_na_profile_id      False     True
is_na_glacier_name
False                989037  1463186
True                1401468      588
is_na_thickness       False     True
is_na_glacier_name
False               1297637  1154586
True                 226950  1175106
is_na_data_flag     False     True
is_na_glacier_name
False                   1  2452222
True                    0  1402056
is_na_elevation       False    True
is_na_glacier_name
False               1974076  478147
True                1398758    3298
is_na_remarks         False     True
is_na_glacier_name
False                866563  1585660
True                1301673   100383
```

```
is_na_profile_id     False    True
is_na_thickness
False               596579  928008
True               1793926  535766
is_na_glacier_name    False     True
is_na_thickness
False                1297637   226950
True                 1154586  1175106
is_na_data_flag  False     True
is_na_thickness
False                1  1524586
True                 0  2329692
is_na_elevation    False    True
is_na_thickness
False            1053453  471134
True             2319381   10311
is_na_remarks      False    True
is_na_thickness
False             565179  959408
True             1603057  726635
```

Appendix XVII - (Variational) Autoencoder Models

Latent Space Dim / Model Type	64 Dim / AE	24 Dim / VAE	32 Dim / VAE	48 Dim / VAE	64 Dim / VAE	Note
Input	64x64 Pixels	64x64 Pixels	64x64 Pixels	64x64 Pixels	64x64 Pixels	
HP Optimization	Random Search	Bayesian	Bayesian	Bayesian	Bayesian	
Hyper-parameters	Learning Rate: 0.0012 Convolutions/Maxpooling Layers: 4 MaxPooling: 2x2 Filters: 32 / 64 / 128 / 256 Dense Fully Connected Layers: 3 Neurons in dense layer: 4096 / 256 / 128 / Latent Space Batch Size: 32 Epochs: 30	Learning Rate: 0.001 Convolutions/Maxpooling Layers: 2 MaxPooling: 2x2 Filters: 16 / 32 Dense Fully Connected Layers: 1 Neurons in dense layer: 8192 / Latent Space Batch Size: 32 Epochs: 30 KL Weight: 0.011	Learning Rate: 0.001 Convolutions/Maxpooling Layers: 2 MaxPooling: 2x2 Filters: 32 / 64 Dense Fully Connected Layers: 1 Neurons in dense layer: 8192 / Latent Space Batch Size: 32 Epochs: 30 KL Weight: 0.011	Learning Rate: 0.0011 Convolutions/Maxpooling Layers: 4 MaxPooling: 2x2 Filters: 32 / 64 / 128 / 256 Dense Fully Connected Layers: 1 Neurons in dense layer: 4096 / Latent Space Batch Size: 32 Epochs: 30 KL Weight: 0.01	Learning Rate: 0.0012 Convolutions/Maxpooling Layers: 4 MaxPooling: 2x2 Filters: 32 / 64 / 128 / 256 Dense Fully Connected Layers: 2 Neurons in dense layer: 4096 / 512 / Latent Space Batch Size: 32 Epochs: 30 KL Weight: 0.01	After optimal learning rate was found with optimization, the models are trained with a flexible learning rate, reducing after hitting plateu. Since the KL Weight changes the loss function, it was not tuned upon. It was instead set "optimally" by analysing latent space structure and reconstruction loss.
Binary Crossentropy Loss	0,1415	0,1692	0,1616	0,1598	0,1509	Note that this is only the reconstruction loss, and NOT the KL Divergence loss. As the loss functions are different, comparing them does not make sense.
Run Time (Tuning + Training)	5 Hours	4.5 Hours	4.5 Hours	5 Hours	5 Hours	
Comments	Best reconstructions, but unpleasant latent space, giving worse regression predictions.	Worst reconstruction, but pleasant latent space, and potentially fewer high-dimensionality problems for regression.			As expected, second best reconstruction. Pleasant latent space, but potentially high-dimensionality problems for regression.	

Appendix XVIII - Autoencoder Comments

- All autoencoder models were symmetric. Meaning the decoder structure is a direct mirror of the encoder. 4 convolutions and maxpoolings in the decoder, meant 4 upscalings and convolutions in the decoder.

- For HP Tuning, we found that choosing Learning Rate from a log-uniform distribution, instead of standard uniform gave good results. The idea is simple. Say you want to search for the optimal learning rate between 0.001 and 0.1. There is almost a 10 times larger numerical difference between 0.1 and 0.01, than there is between 0.01 and 0.001. But it is quite unlikely that you want to try 10 times more models between 0.01 and 0.1 than between 0.001 and 0.01. Log transforming the intervals makes the probability of trying a learning rate from each interval, equally likely.

- The KL-Weight of all models were chosen by the analysing the structure of the latent space, and comparing it to how much performance we lost in terms of reconstruction.