Explaining the spatial pattern of debris flow and flood hazard in High Mountain Asia

Varvara Bazilova, Tjalling de Haas, Walter Immerzeel

Dorje Dolma lama
@DolmaLama444

Over 50 people are missing in the Melamchi and Indrawati rivers' flooding. The floods have also caused damages to the dam in Melamchi drinking water project, Timbu Bazaar, Chanaute Bazaar, Talamarang Bazaar and Melamchi Bazar.

Kaushal Gnyawali
@KaushalGnyawali

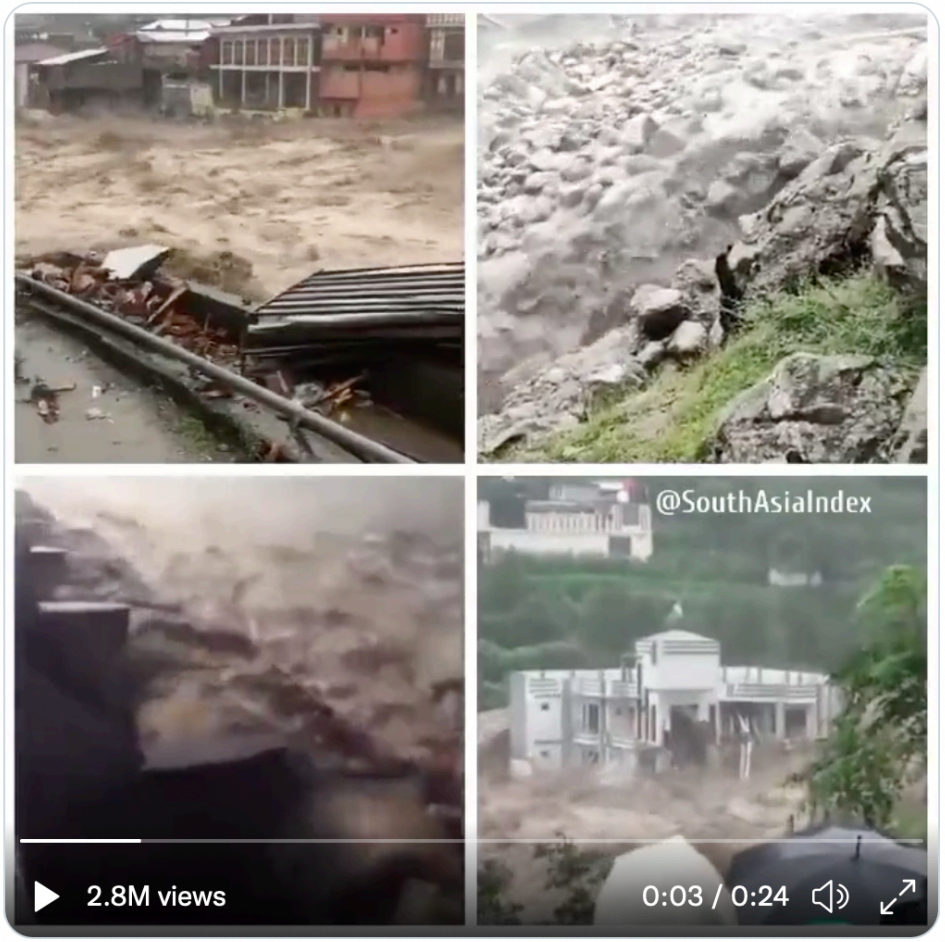
Landslide dam outburst seems to have Melamchi flooding in Sindhupalchowk, Nepal

Colin McCarthy @US_Stormwatch · Aug 30

Hard to comprehend the scale of the **flood** disaster in Pakistan, the 5th most populated nation in the world.

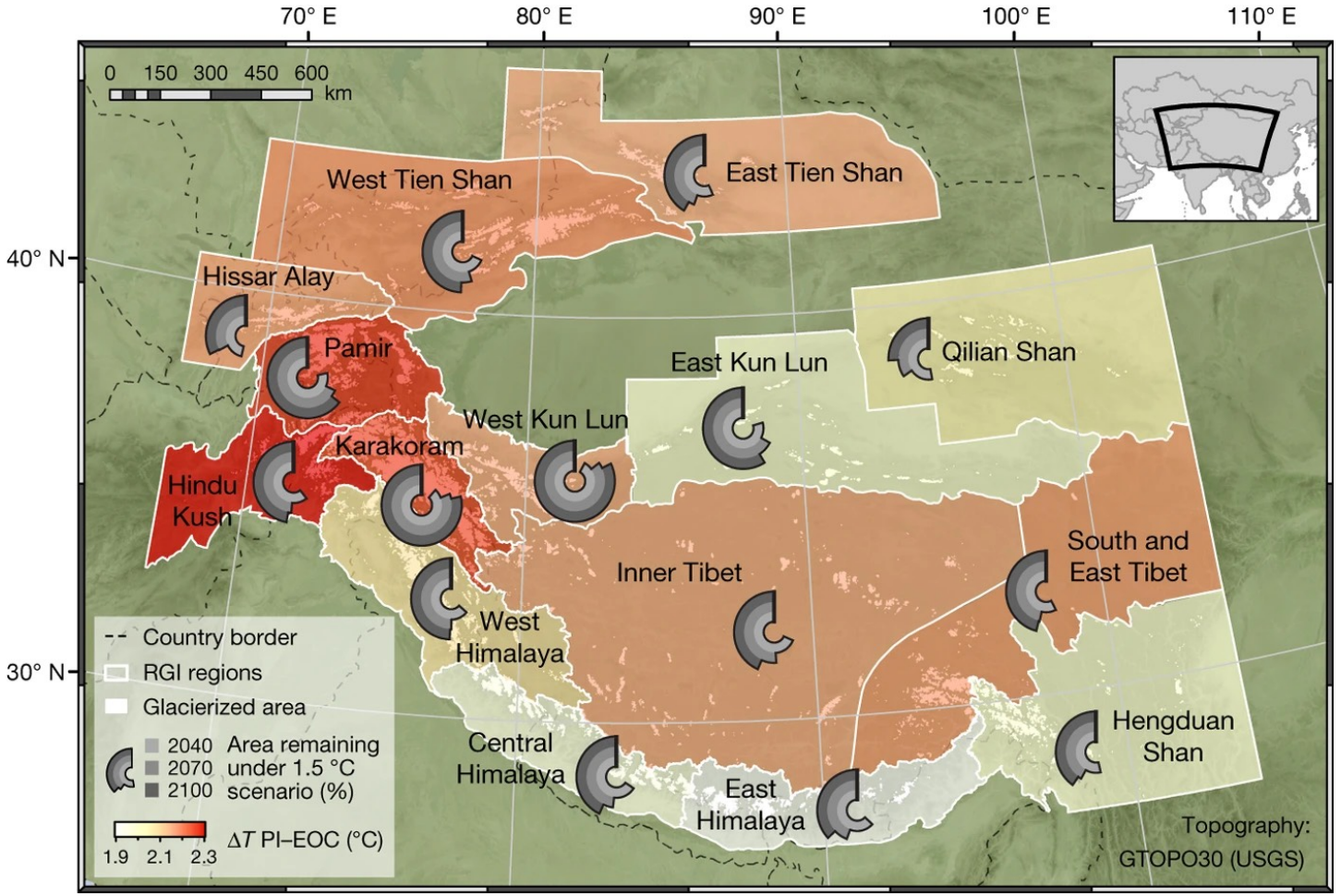
Nearly 1400 dead, 1 million houses damaged or destroyed, and 50,000,000 people displaced.

1/3 of the country is underwater.

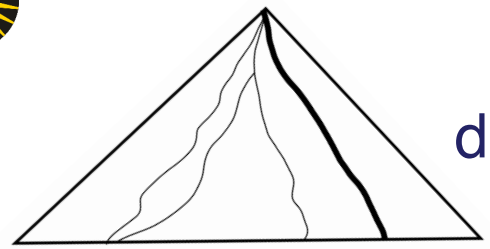


What is happening in HMA?

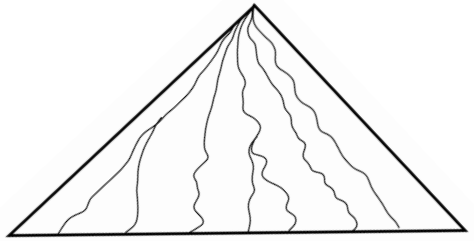
(Glaciers are projected to change (for this amount) (Climate is projected to change (to this amount)



Impact of a global temperature rise of 1.5 degrees Celsius on Asia's glaciers (P. Kraaijenbrink et al., 2017)



debris flow dominated



fluvial flow dominated

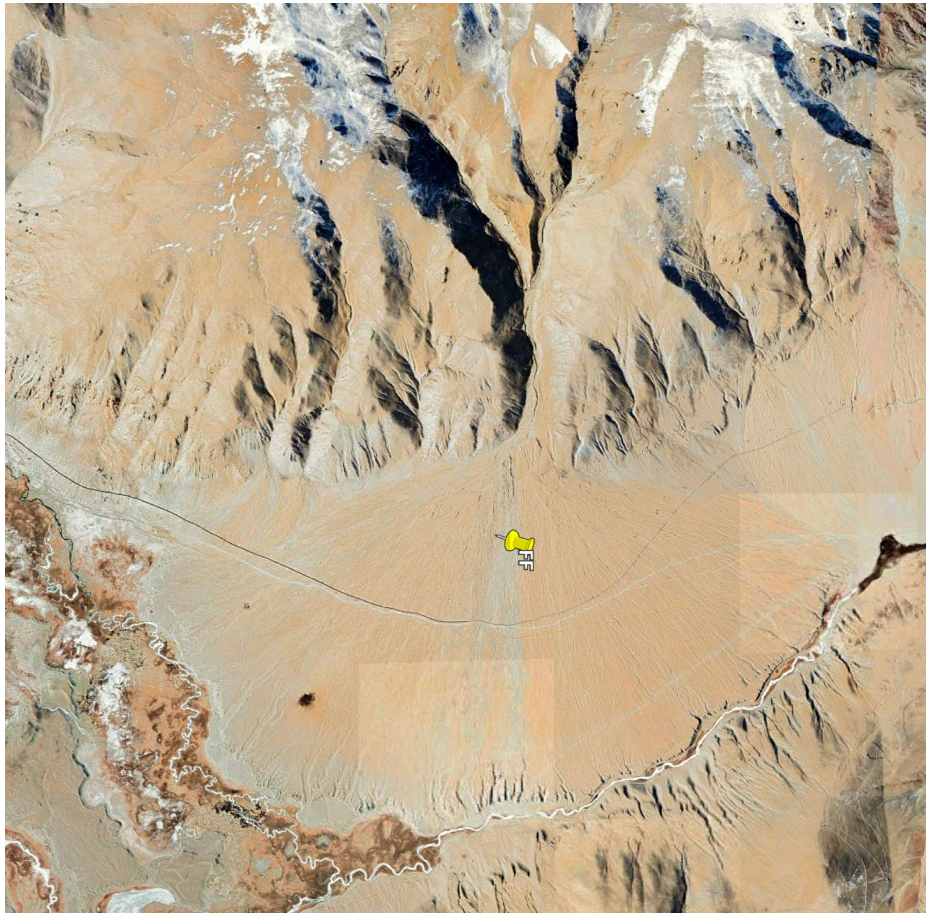


Image source: Google Earth

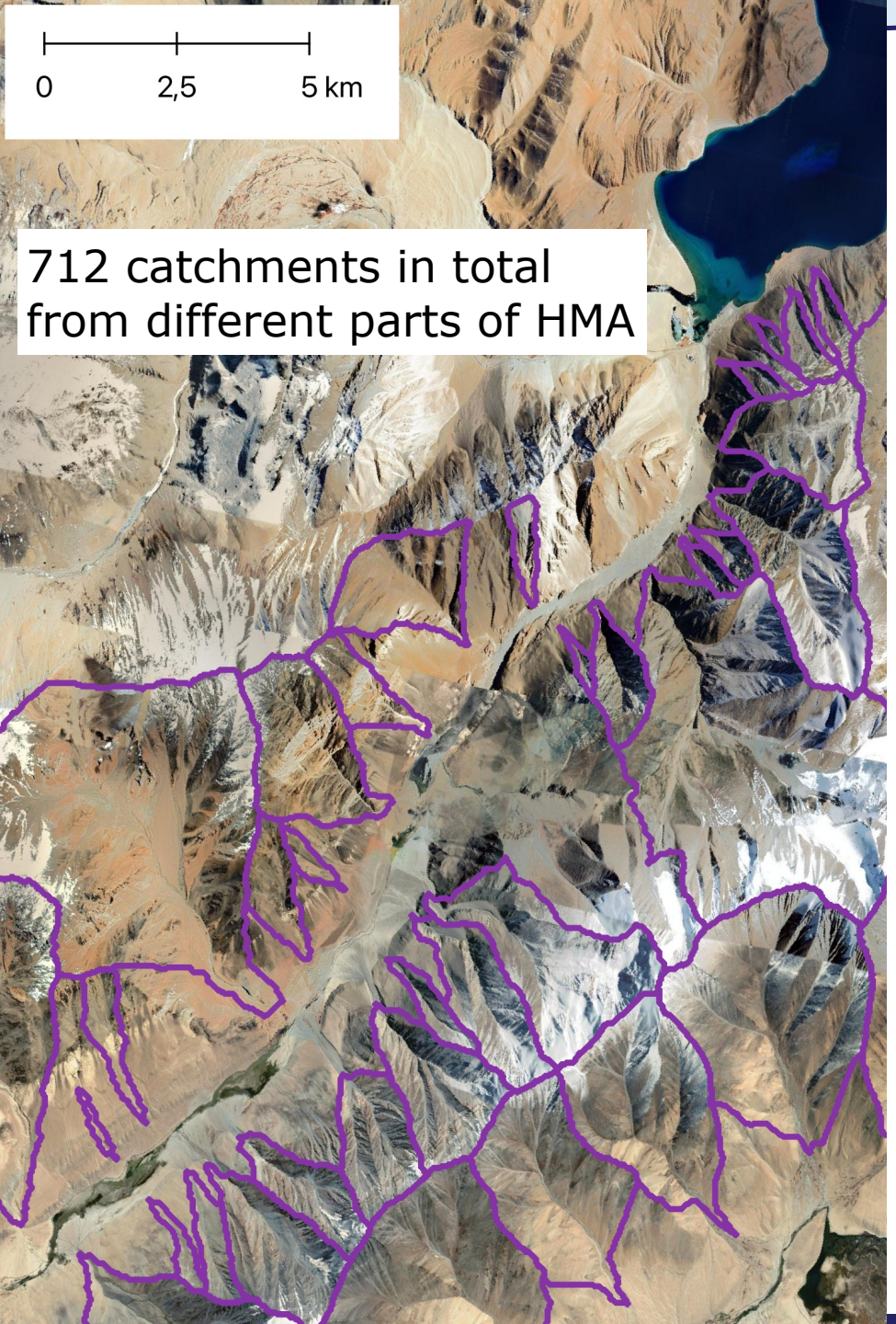
Research questions

- to use ML classifier to estimate probabilities of debris-flow vs flood dominated system
- to identify the parameters, that matter for the classification
- to see if adding climatic features affects the classification
- to find out if there are any regional differences

- to make projections based on climate scenarios

“Pipeline” of this ML project - methods overview

- 1. Get the data (1.
- 2. Build the model (split data to train and test)
- 3. Evaluate how good the model is
- 4. Apply the model to the data



Morphometric + climate

Morphometric

- x_centroid
- y_centroid
- area_m
- perimeter
- mean_elevation
- median_elevation
- std_elevation
- min_elevation
- max_elevation
- range_elevation (relief)
- variance_elevation
- mean_slope
- median_slope
- std_slope
- min_slope
- max_slope
- range_slope
- variance_slope
- Melton_ratio (relief*area$^{0.5}$)
- circularity_ratio
- compactness_coefficient
- **region**

- mean_annual_temp
- mean_jan_temp
- mean_jury_temp
- mean_monsoon_temp
- mean_outside_monsoon_temp
- temp_crosses_zero (frost cracking)
- belowzero_fraction_of_year
- mean_daily_precipitation
- mean_annual_sum_precipitation
- mean_daylymonsoon_precipitation
- mean_monsoon_sum_precipitation
- monsoon_precipitation_fraction
- n_rainy_days (>10mm)
- rainy_days_fraction
- **avgtemp_belowzero**

- glacier_area_sum
- glacier_area_fraction
- **glacier**

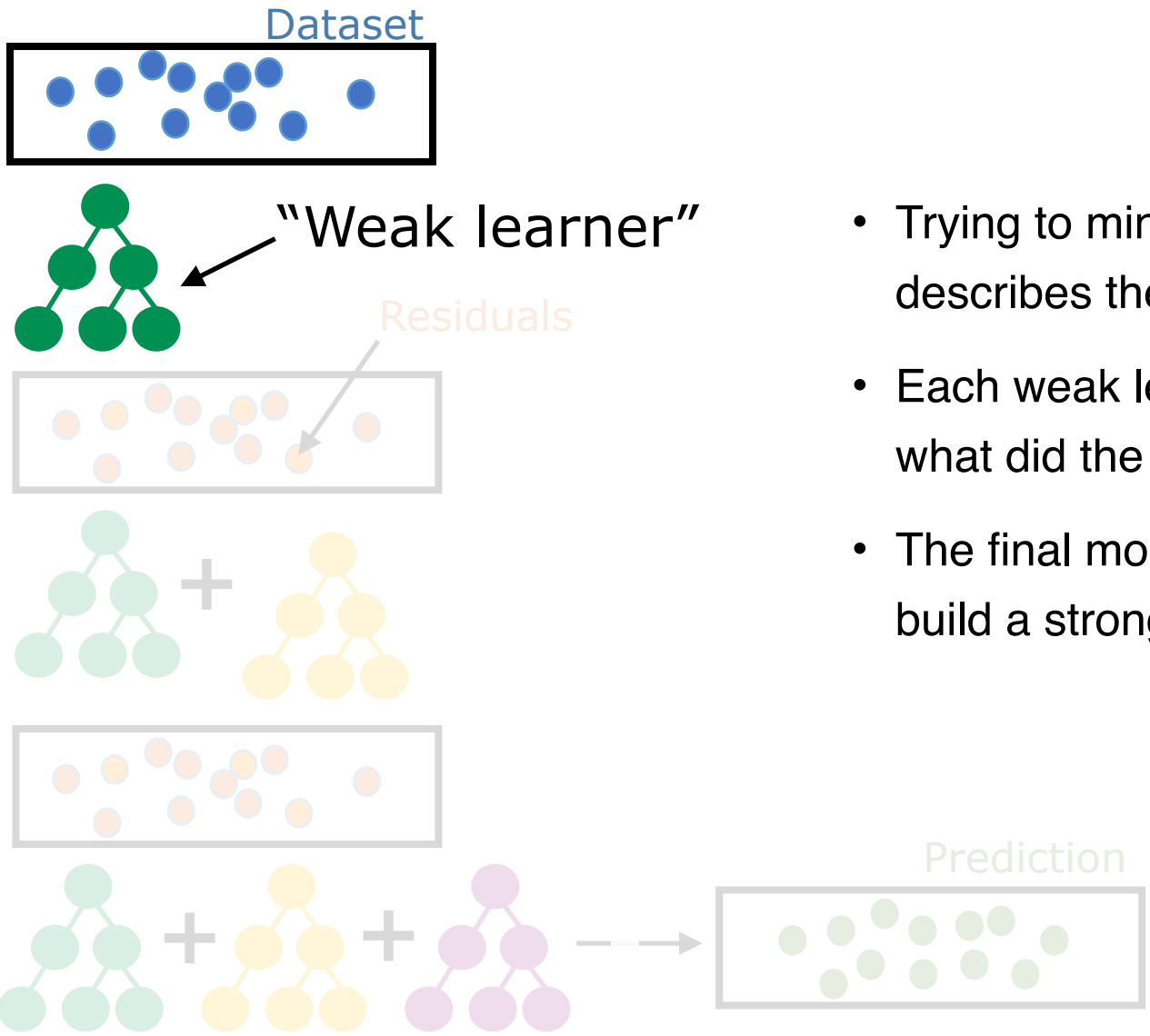
- isolated_permafrost_area
- sporadic_permafrost_area
- discontinious_permafrost_area
- continuous_permafrost_area
- sporadic_permafrost_frac
- discontinious_permafrost_frac
- isolated_permafrost_frac
- continious_permafrost_frac
- all_permafrost_frac
- **cont_permafrost_frac > 50%**
- **any_permafrost**

Gael Varoquaux
@GaelVaroquaux

For thousands of data points and moderate dimensionality (99% of cases), gradient-boosted trees provide the necessary regression model scikit-learn.org/stable/modules...
They are robust to data distribution and support missing values (even outside MAR* settings arxiv.org/abs/1902.06931)

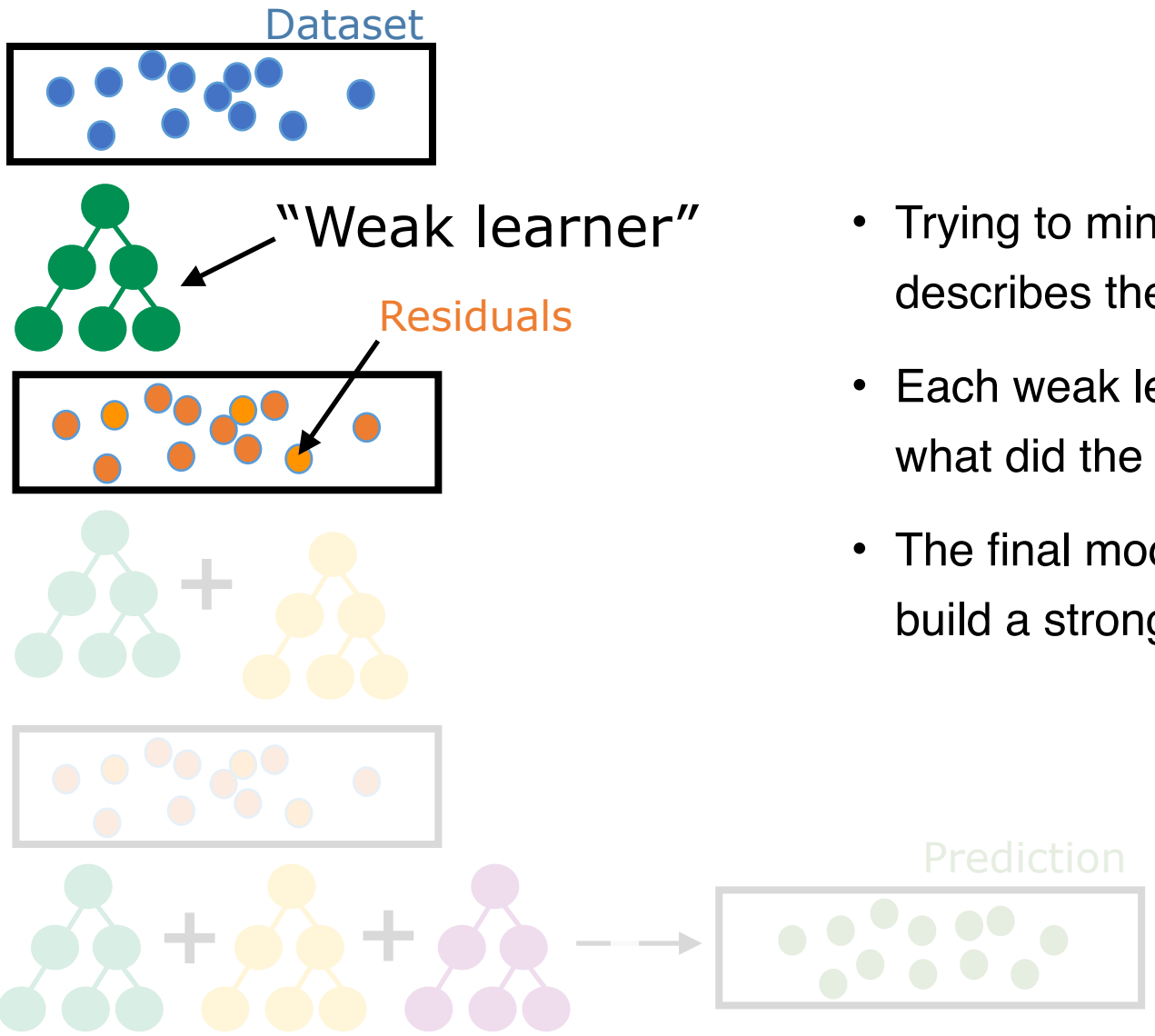
* MAR = Missing At Random

Gradient boosted decision trees



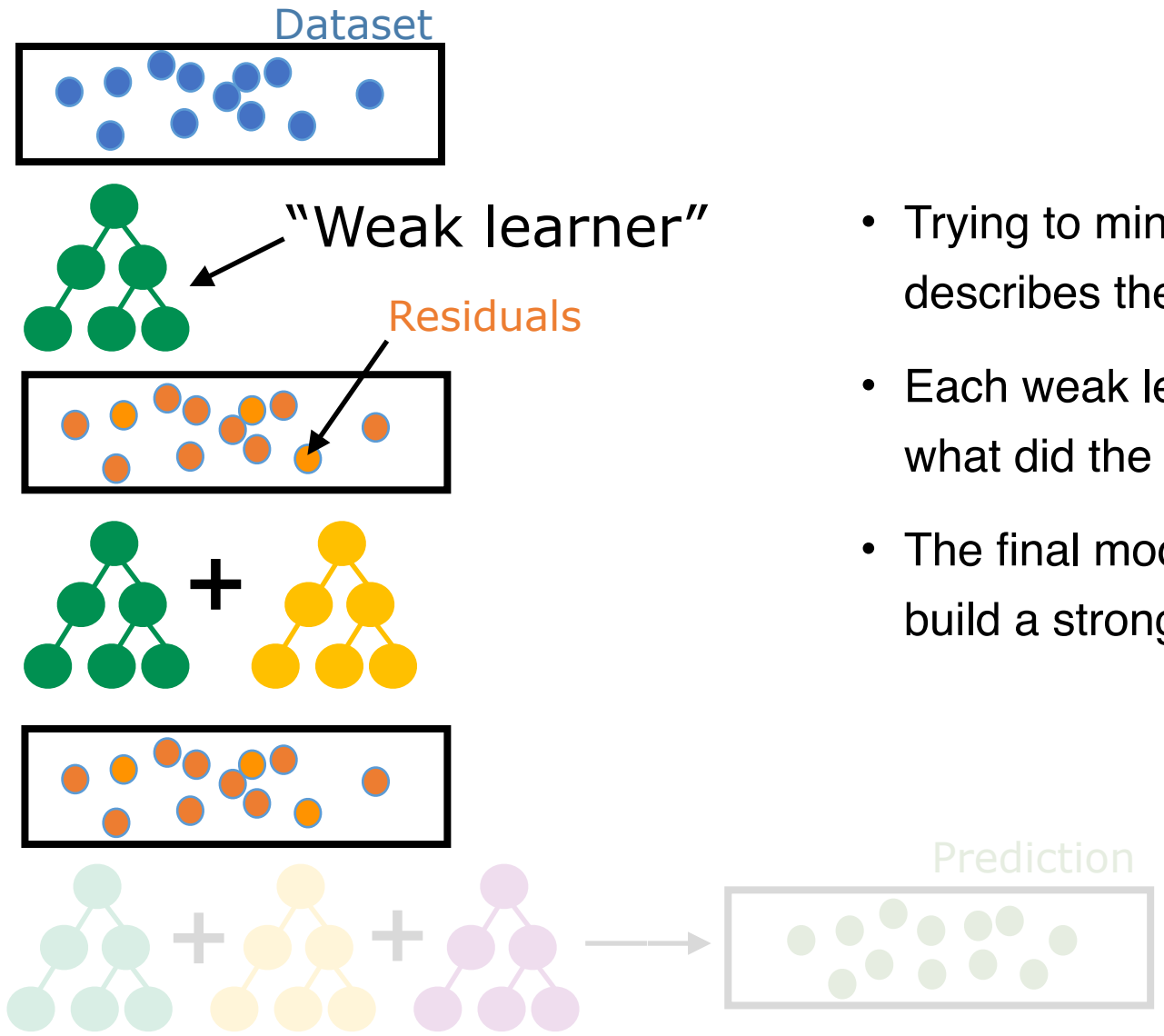
- Trying to minimize the "loss function" (function, that describes the error) on every iteration
- Each weak learner (i.e. tree/iteration) is trying to learn what did the previous one did "wrong" and do better
- The final model is the "combination" of all weak trees to build a strong classifier

Gradient boosted decision trees



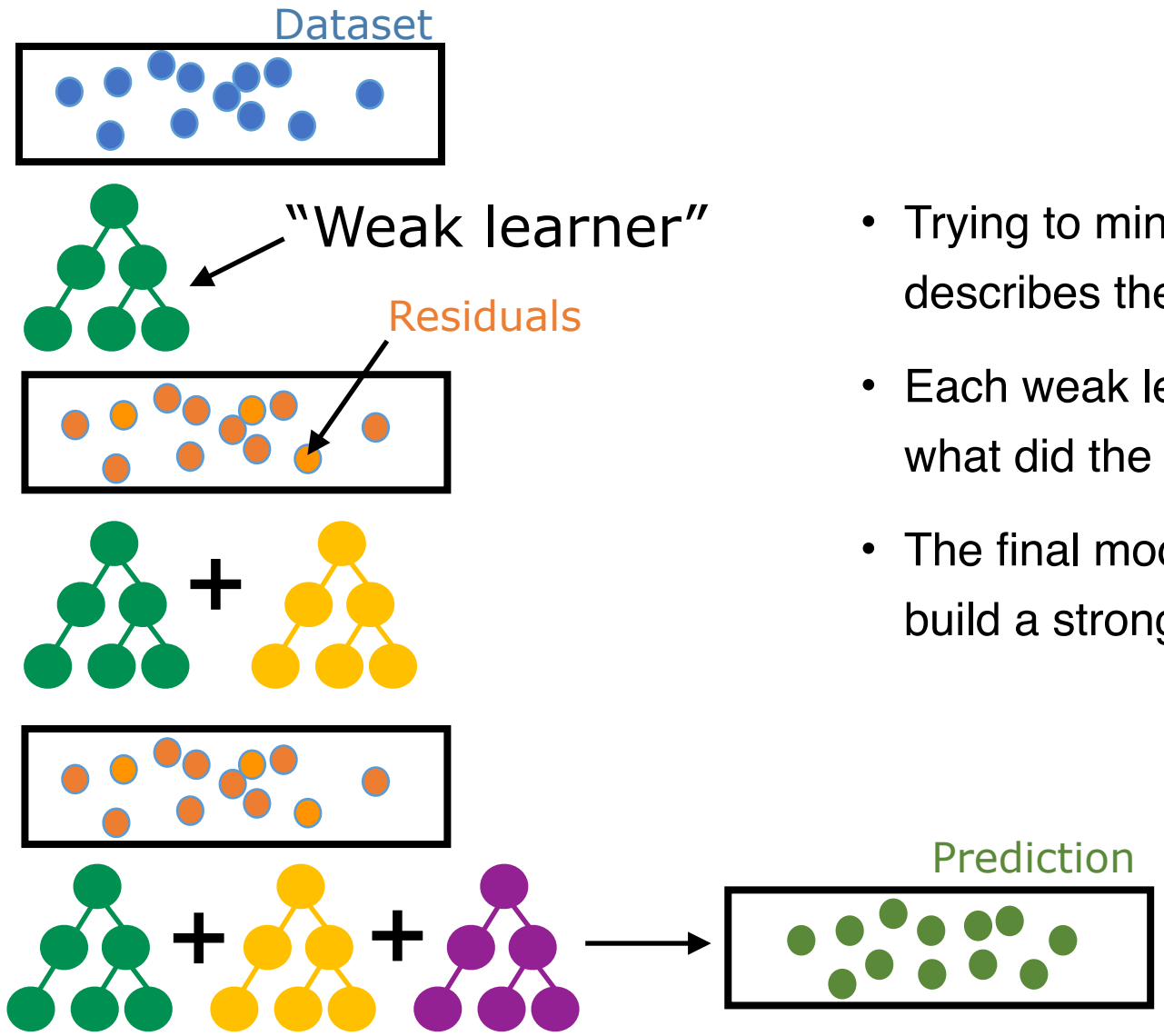
- Trying to minimize the "loss function" (function, that describes the error) on every iteration
- Each weak learner (i.e. tree/iteration) is trying to learn what did the previous one did "wrong" and do better
- The final model is the "combination" of all weak trees to build a strong classifier

Gradient boosted decision trees



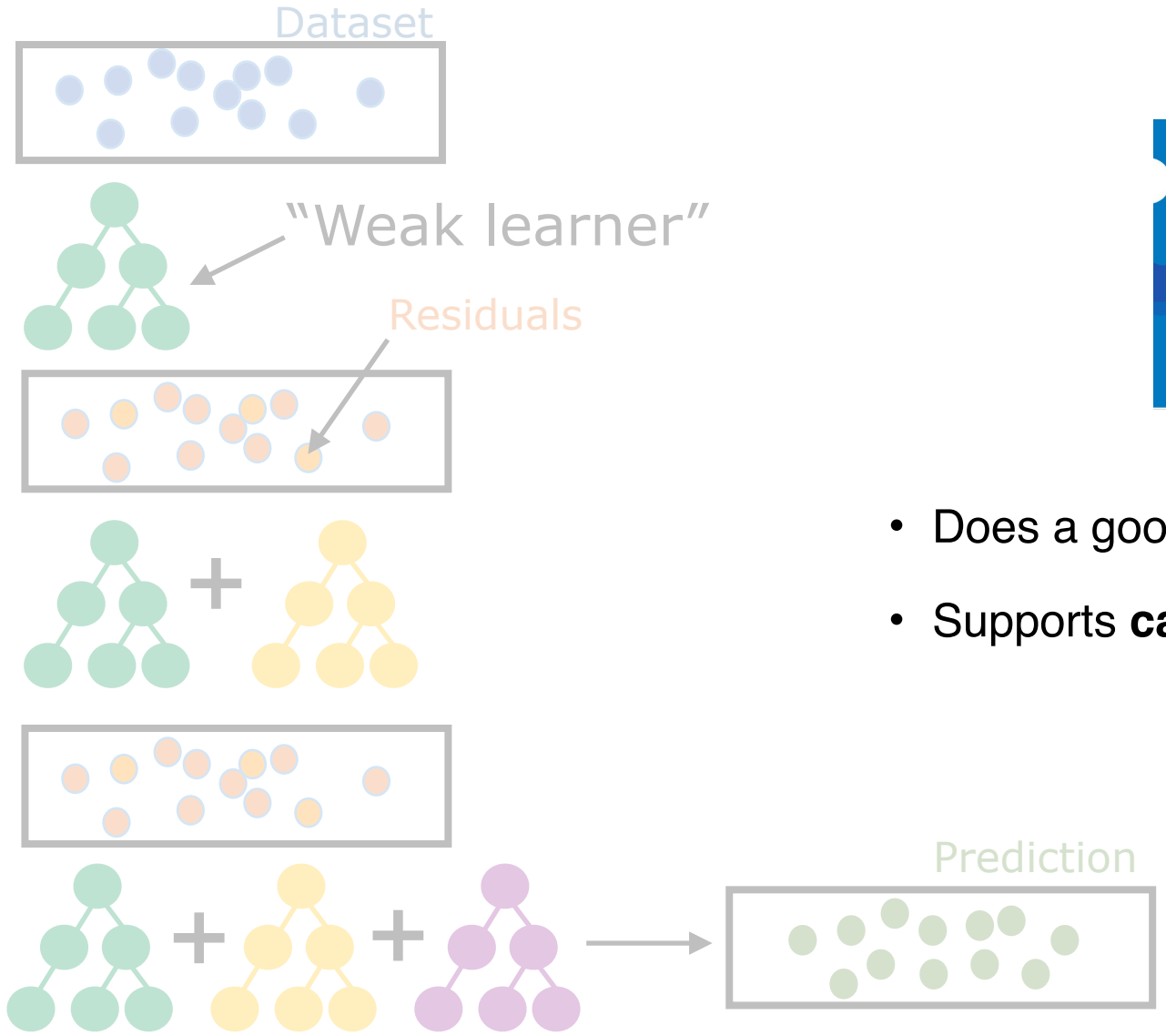
- Trying to minimize the "loss function" (function, that describes the error) on every iteration
- Each weak learner (i.e. tree/iteration) is trying to learn what did the previous one did "wrong" and do better
- The final model is the "combination" of all weak trees to build a strong classifier

Gradient boosted decision trees



- Trying to minimize the "loss function" (function, that describes the error) on every iteration
- Each weak learner (i.e. tree/iteration) is trying to learn what did the previous one did "wrong" and do better
- The final model is the "combination" of all weak trees to build a strong classifier

Implementation: Catboost



- Does a good job as an "out of the box" tool
- Supports **categorical** features (predictors) as an input

Building the model: how good is it?

Morphometric

- Accuracy : 91 %

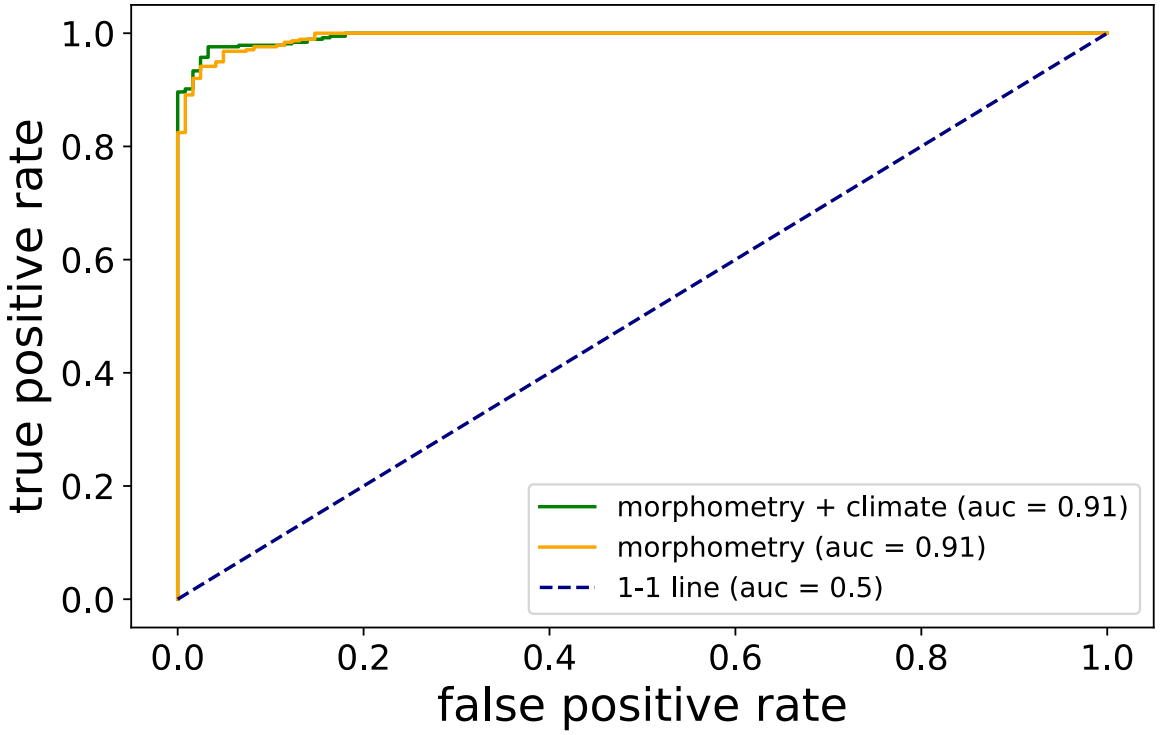
(fraction of correct predictions)

- Confusion matrix: [145., 28.]
  [17., 522.]

Morphometric + climate

- Accuracy: 92 %
- Confusion matrix: [148., 25.]
  [14., 525.]

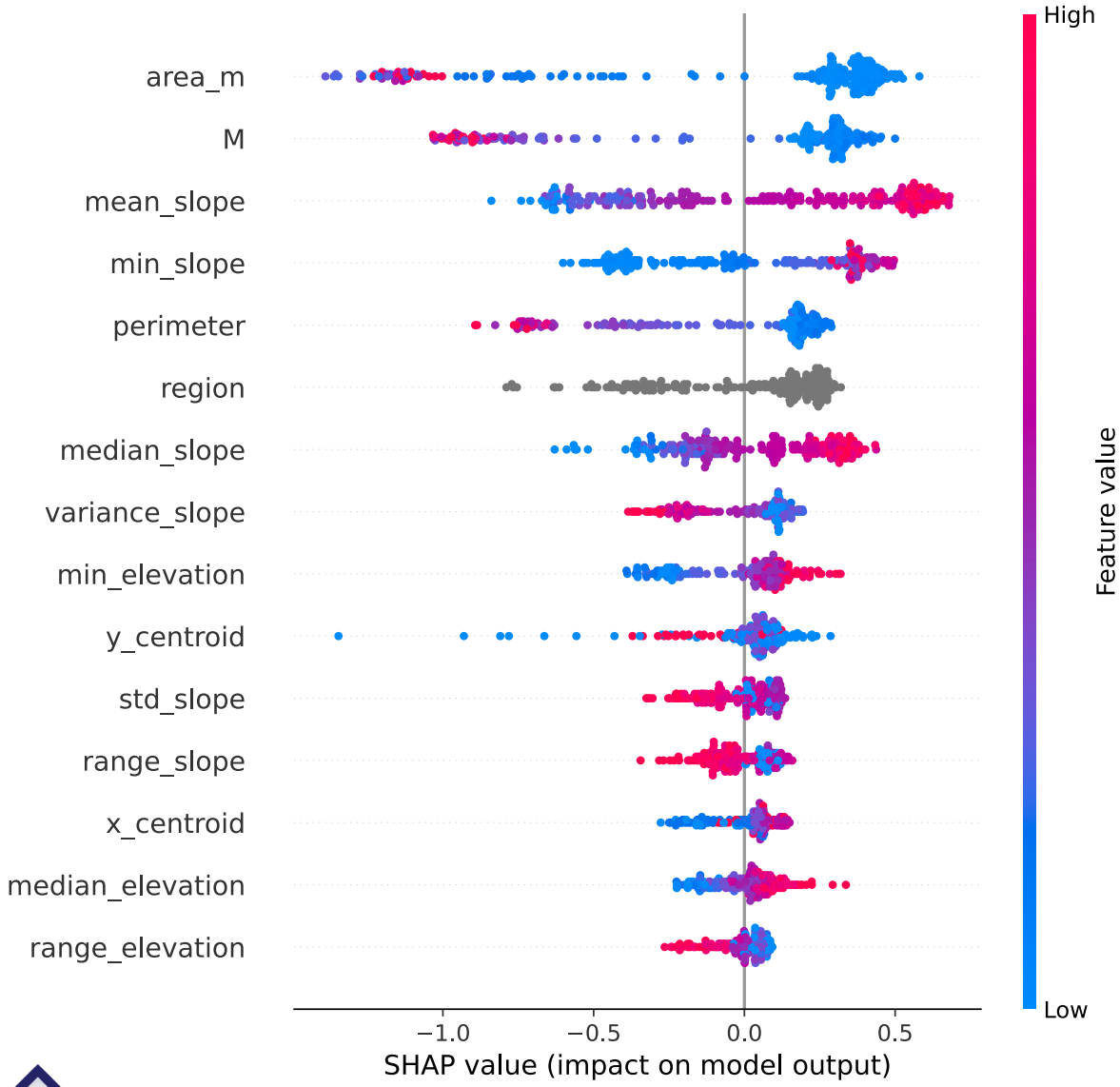
TP	FN
FP	TN



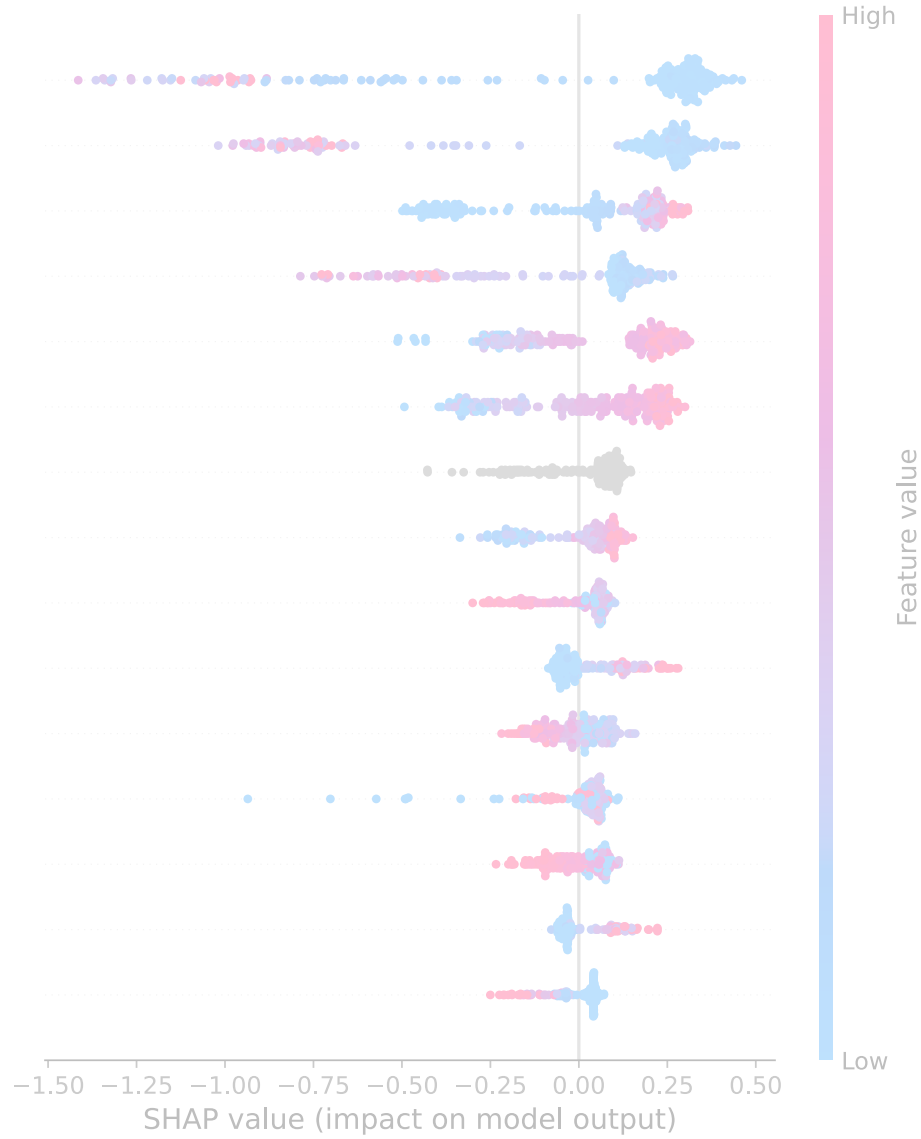
- auc = area under curve
- only “tuned” parameter: number of trees (iterations)
- debris flow (1): 539, flood (0): 173
- accuracy, when guessing randomly: 75.7 %

(Here example maps with “predicted” catchments): same area but 2 different “models” we built

Why does Catboost model make this predictions?

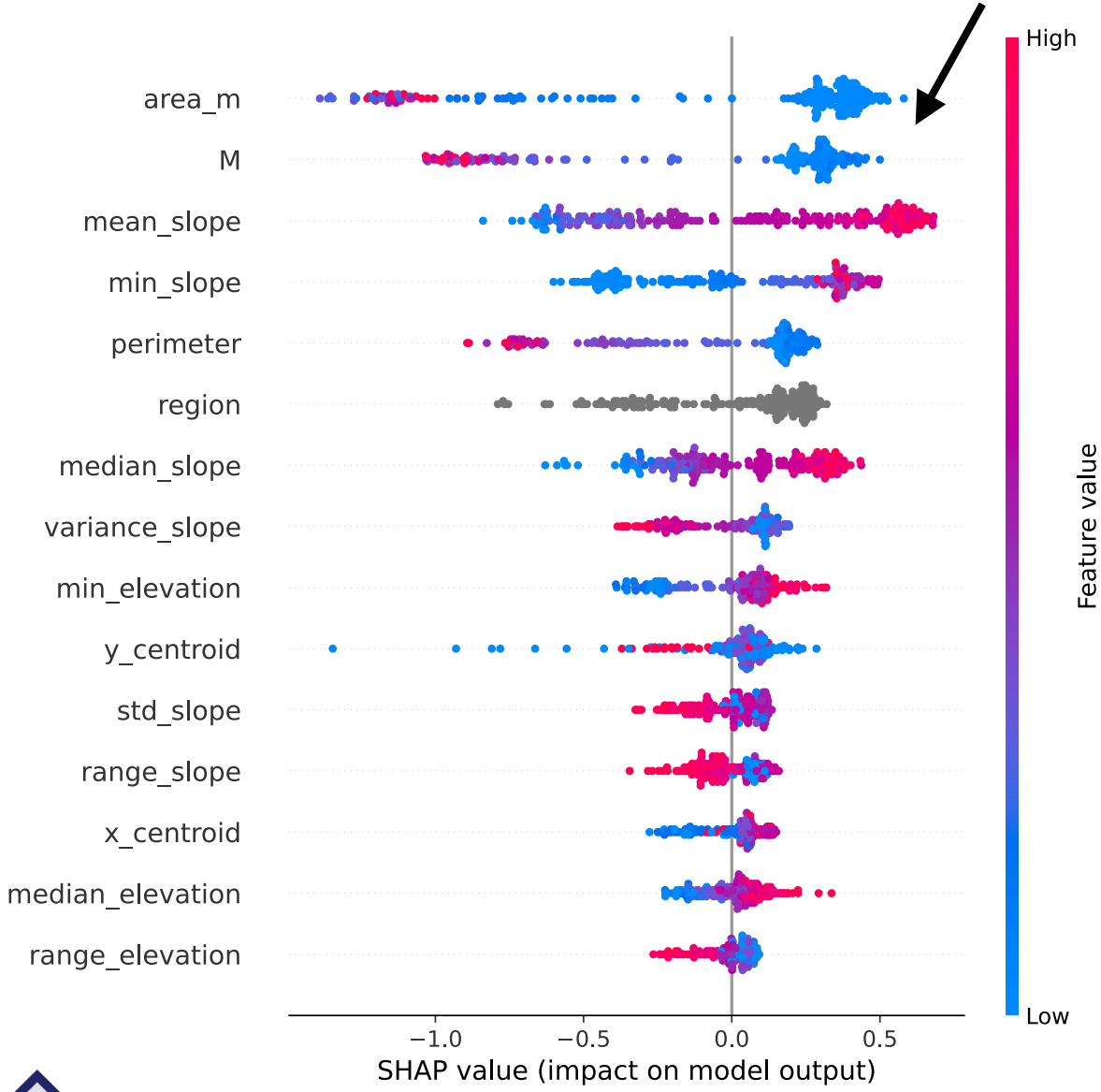


Morphometric

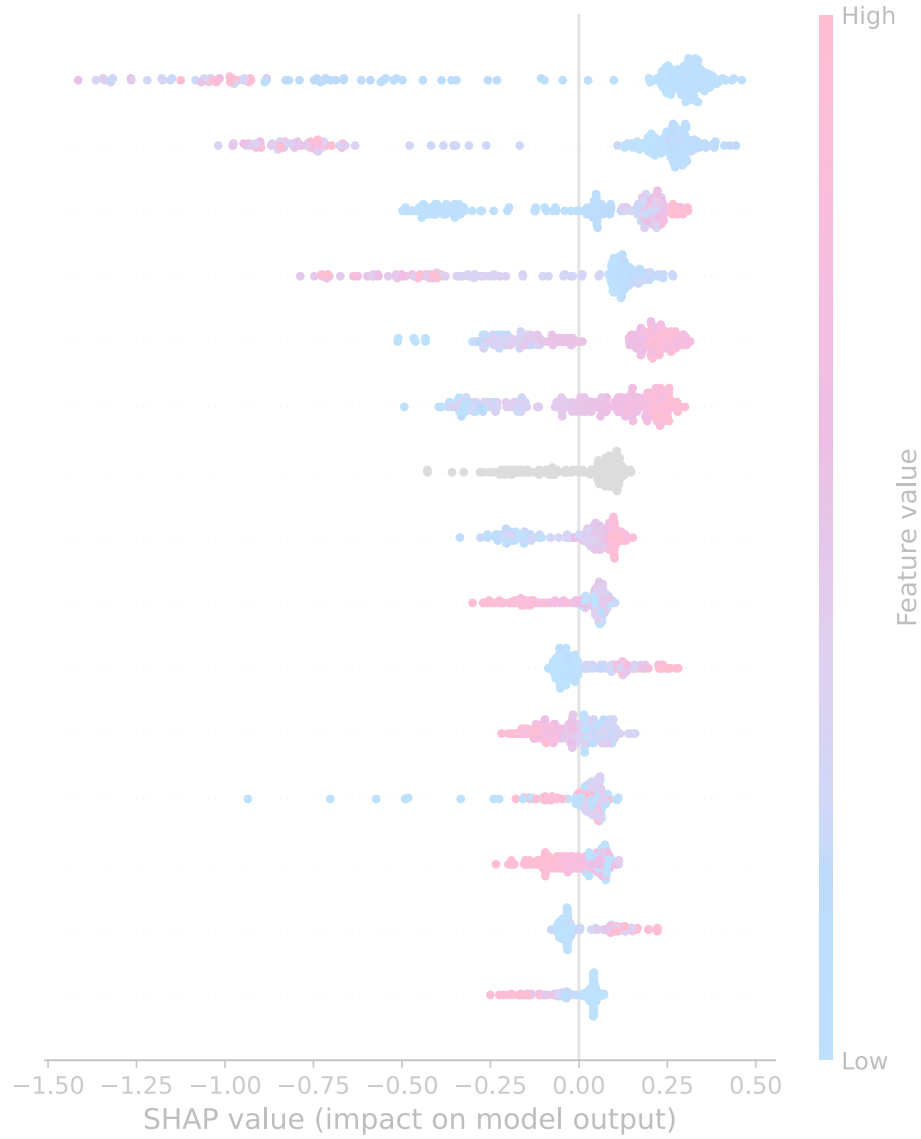


Morphometric + climate

Why does Catboost model make this predictions?

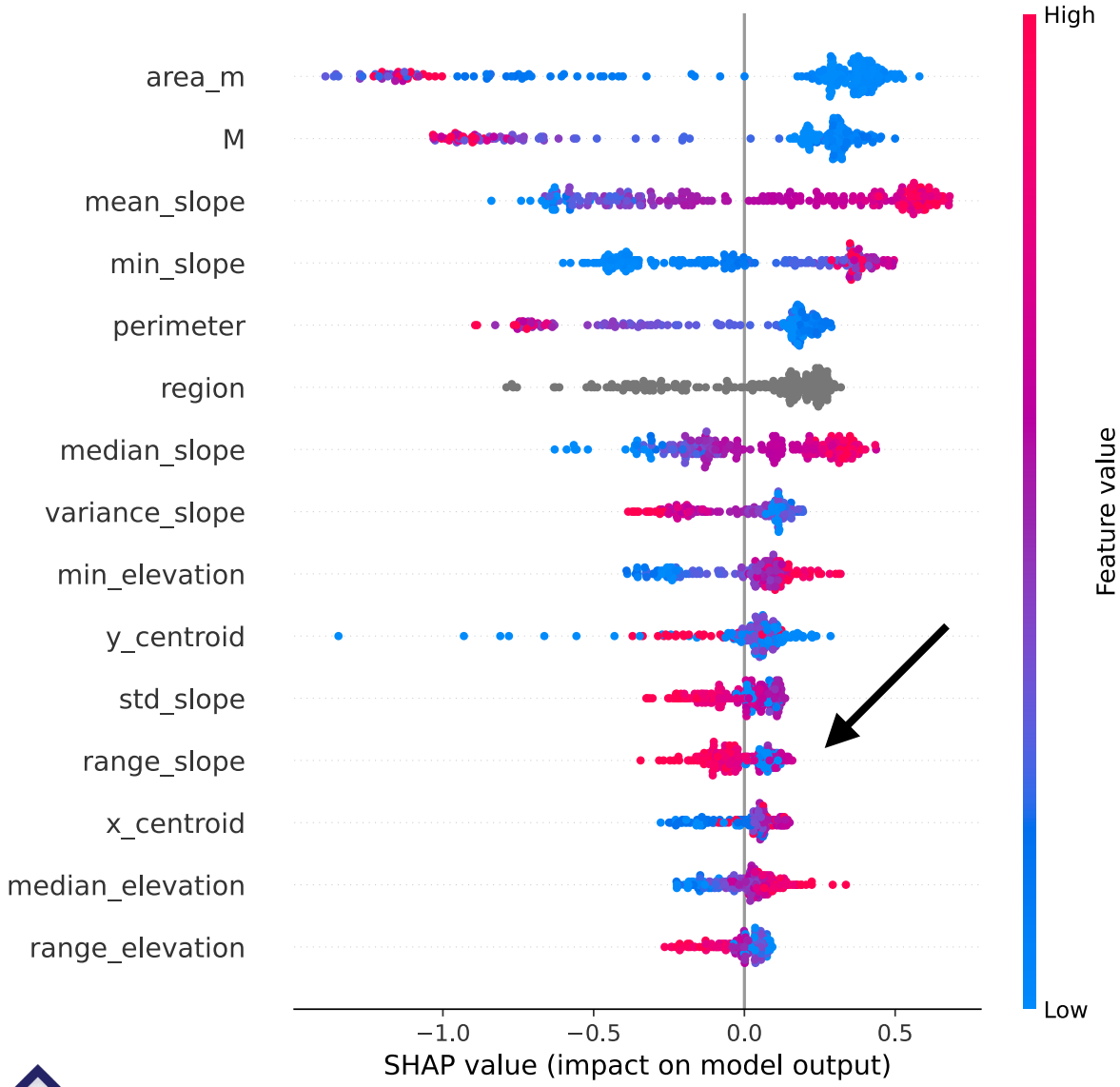


Morphometric

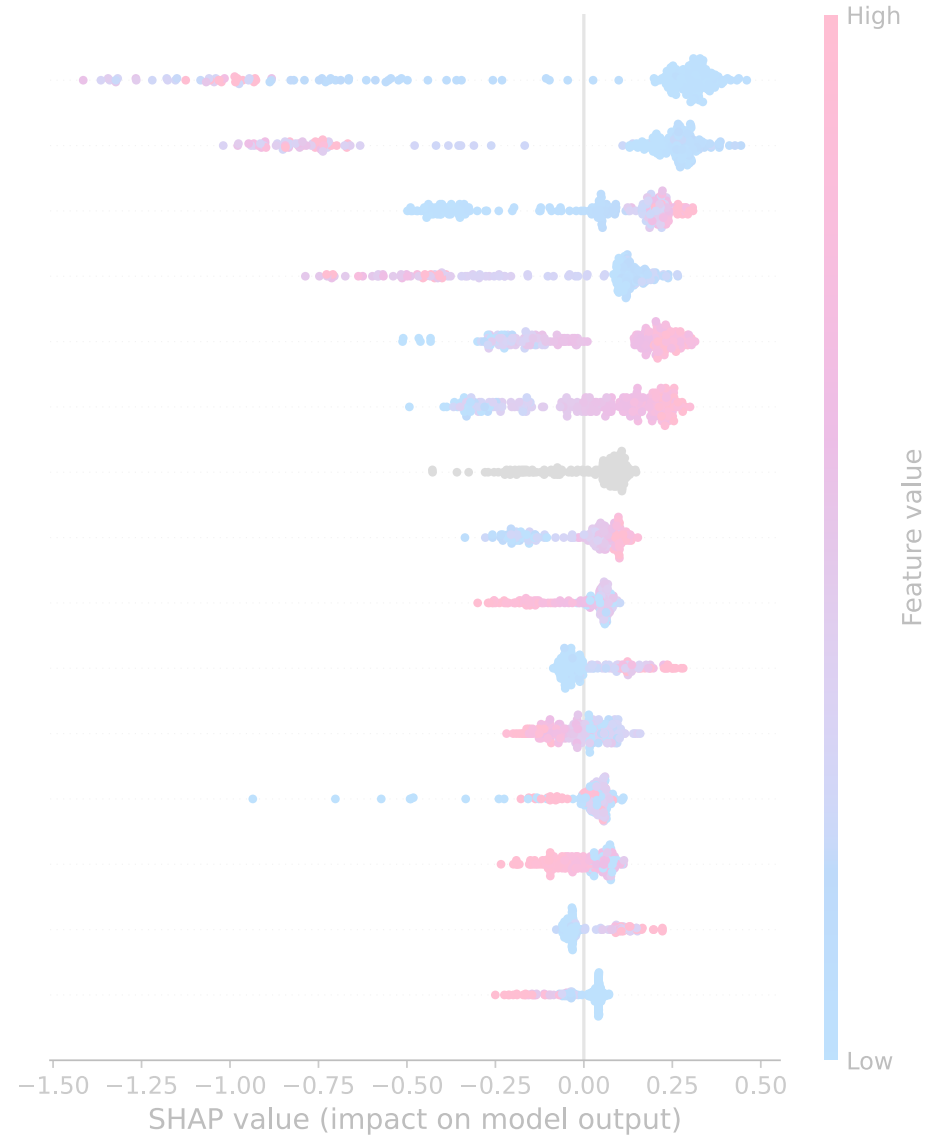


Morphometric + climate

Why does Catboost model make this predictions?

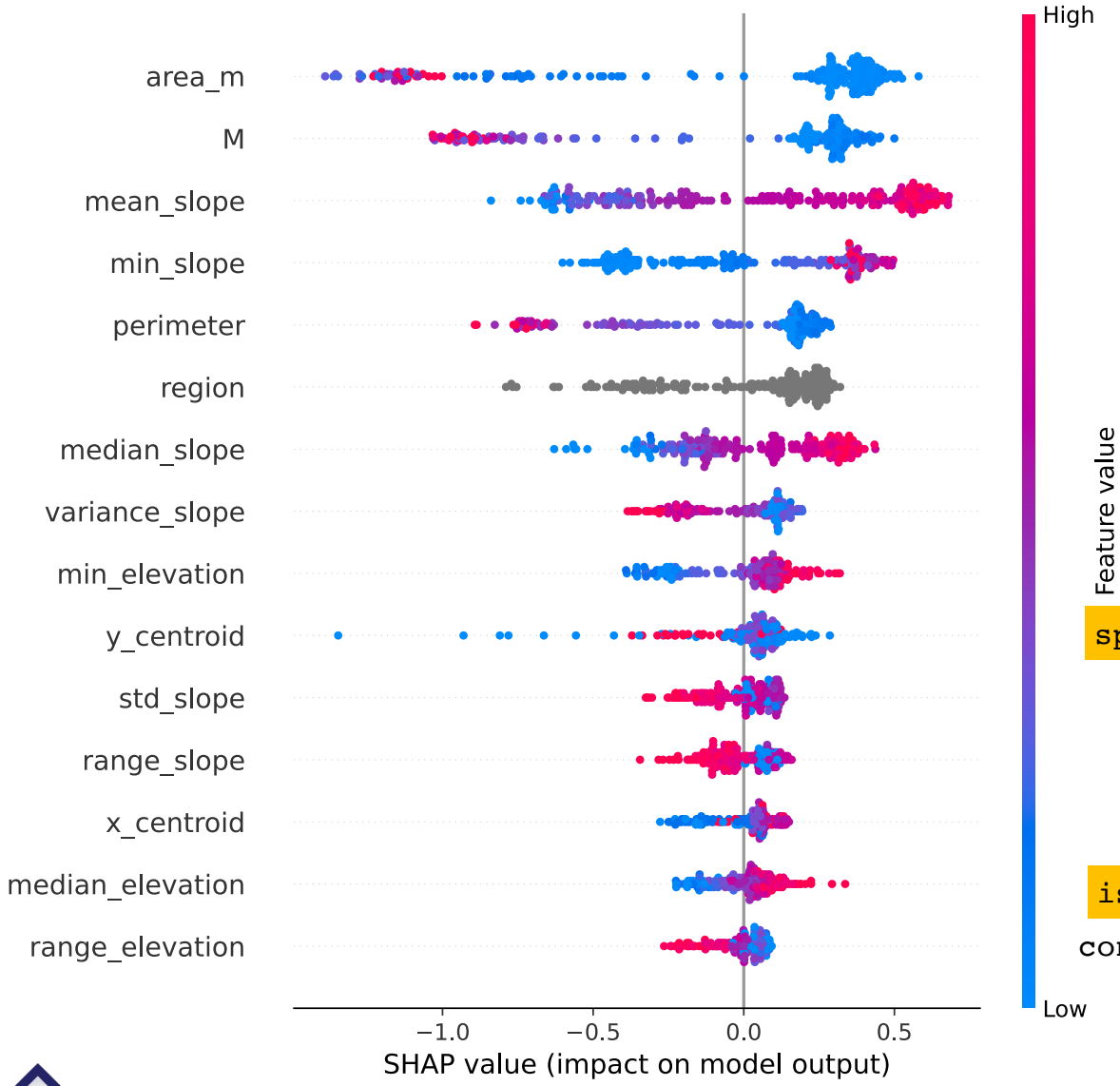


Morphometric

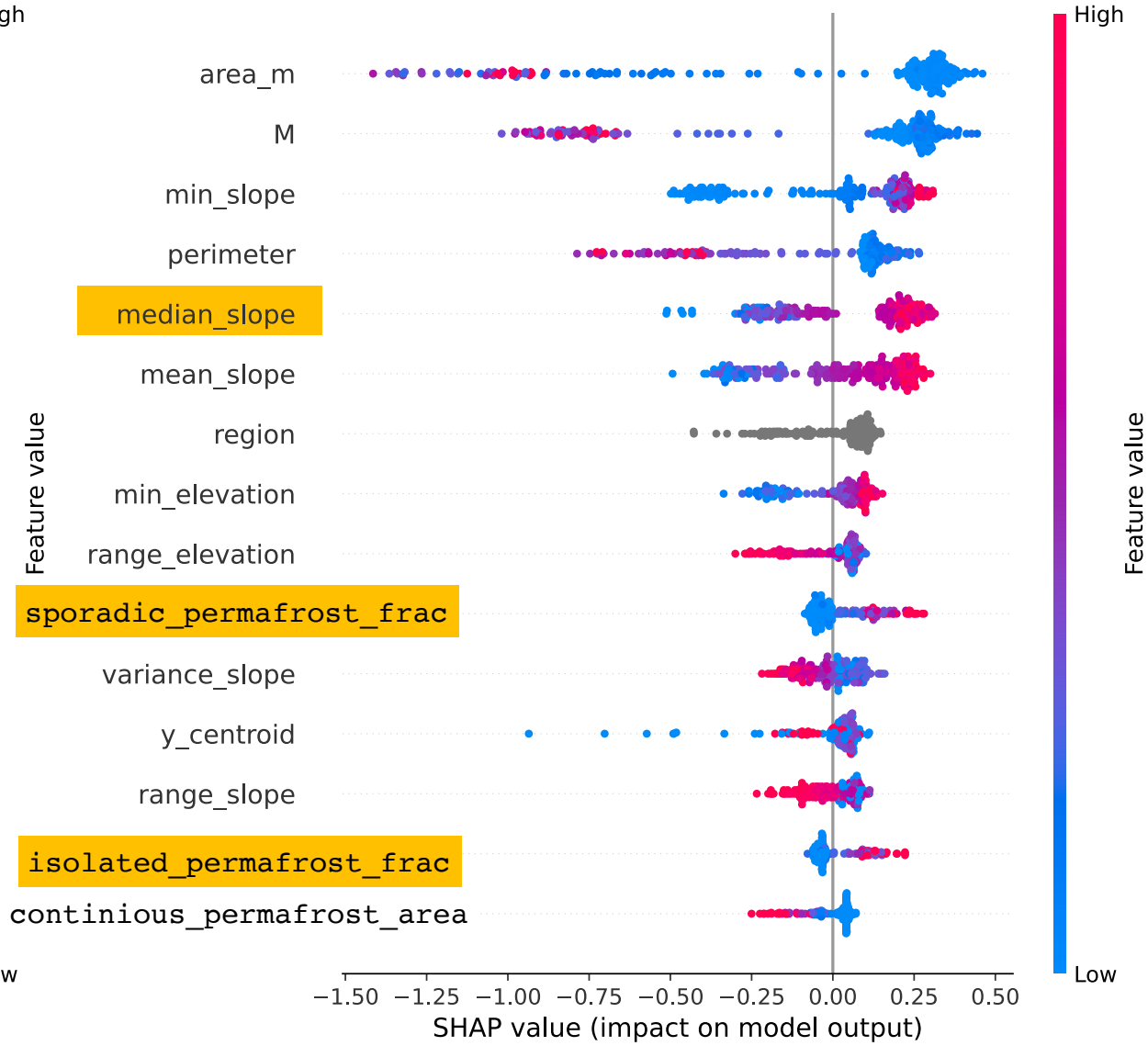


Morphometric + climate

Why does Catboost model make this predictions?



Morphometric



Morphometric + climate

Can we make predictions?

- (Stuff about partial dependence)

- Plot and explanation

- Slide 11: Next steps
- Extend the dataset covering more regions
- Add vegetation coverage as predictor
- Apply the model beyond the training dataset
- Analyse spatial patterns
- Assess the climate change impact (more debris flows and floods, shifts from flood to debris flow, s

Conclusions

- We can build a machine learning classifier for distinguishing debris-flow dominated systems from flood dominated ones
- Climate data adds a lot of information to the model, but (all other things being equal) does not improve model performance
-

Outlook

- Extend the dataset for “creating” the model by covering more diverse regions
- Add vegetation cover to the feature list
- Apply the model to the “new” areas (i.e. catchments without alluvial fan)
- To see the effect of the climate change - use RCP scenarios as a climate information