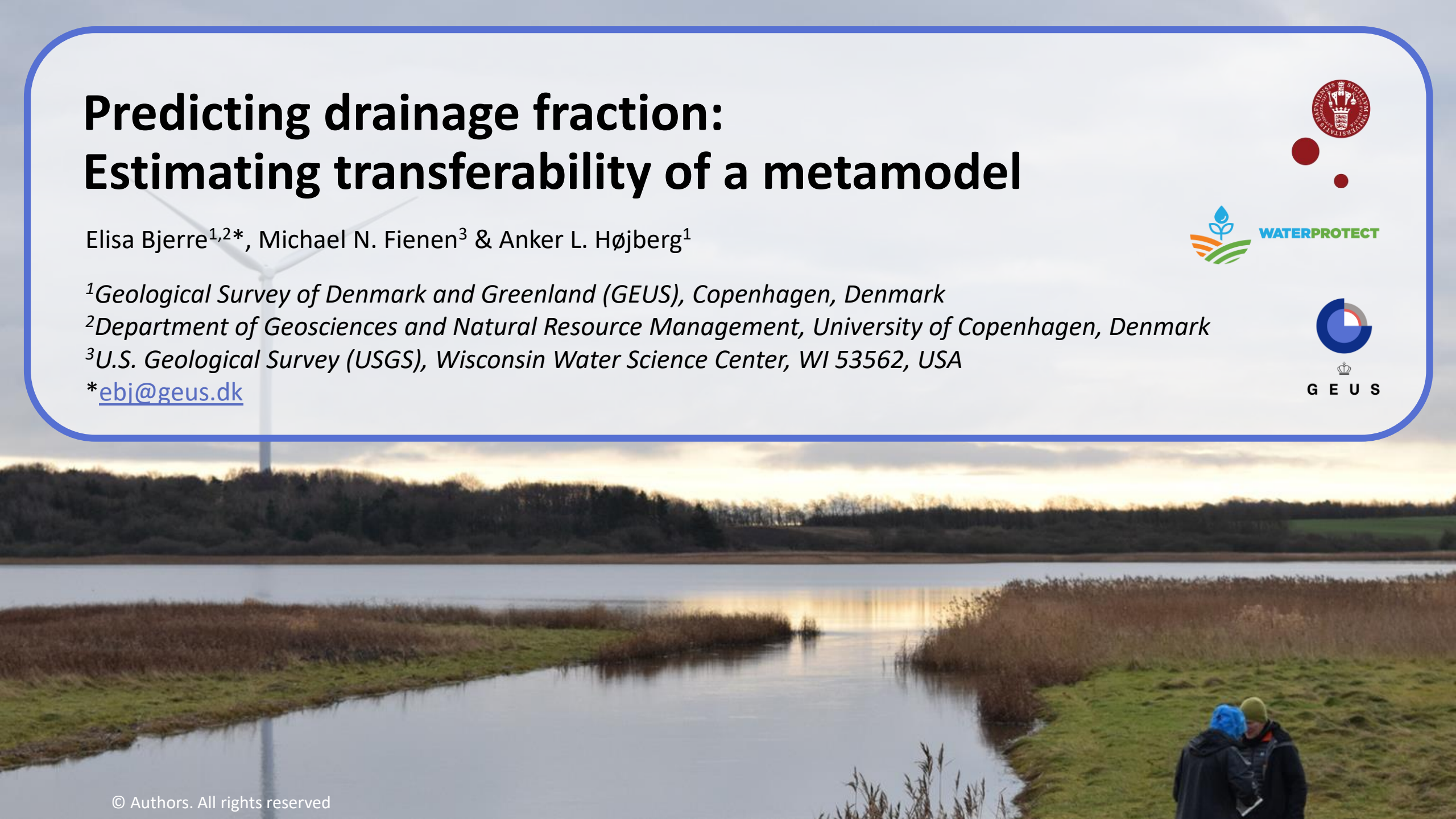
Predicting drainage fraction: Estimating transferability of a metamodel

Elisa Bjerre[1,2]*, Michael N. Fienen[3] & Anker L. Højberg[1]

[1]*Geological Survey of Denmark and Greenland (GEUS), Copenhagen, Denmark*
[2]*Department of Geosciences and Natural Resource Management, University of Copenhagen, Denmark*
[3]*U.S. Geological Survey (USGS), Wisconsin Water Science Center, WI 53562, USA*
*ebj@geus.dk



Why drainage fraction?

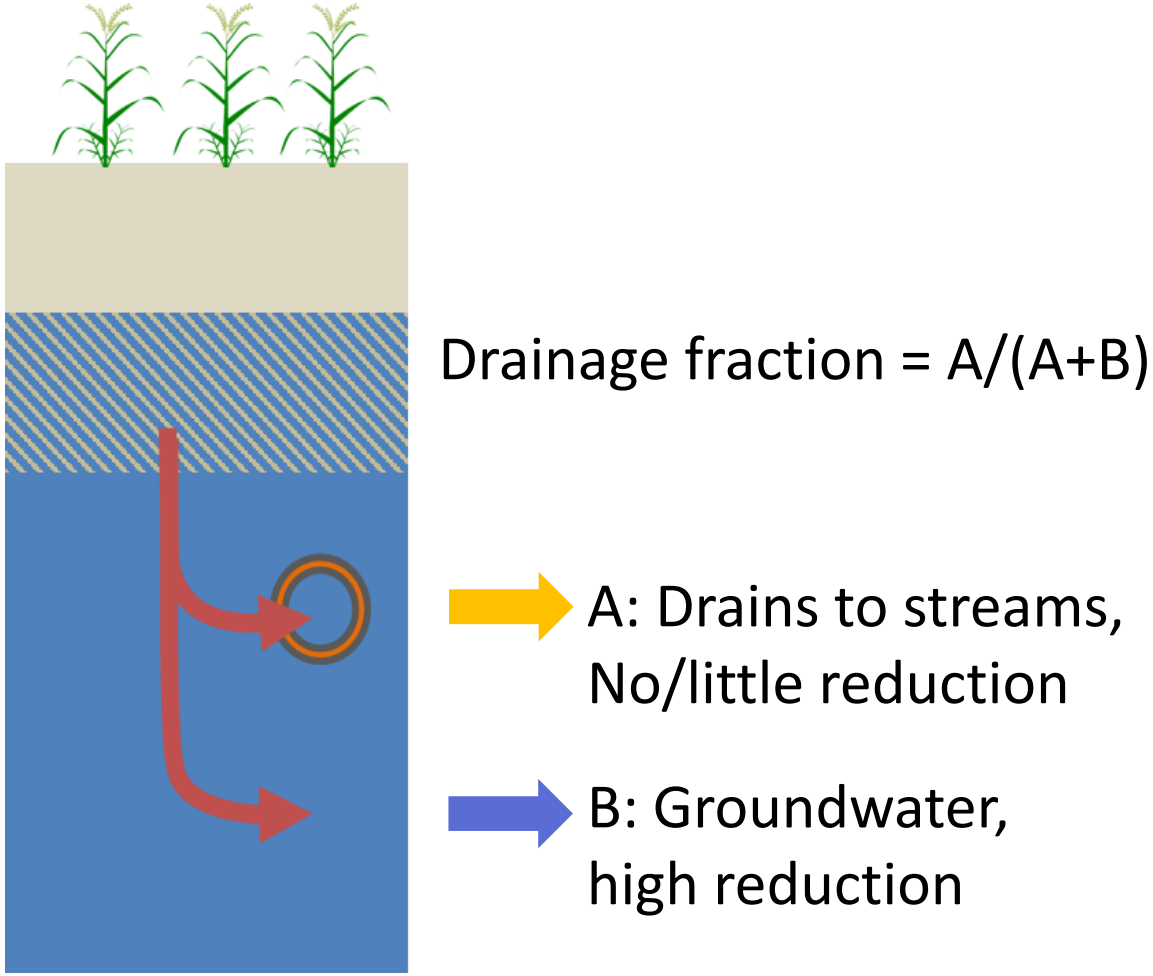


Figure 1. Partitioning of the flux of infiltrating water (red arrows) to the saturated zone below a field, between drains (A) and groundwater (B).

Diffuse nitrogen pollution is a major cause of degraded water quality in rivers and groundwater across Europe [1]. In artificially drained agricultural catchments, nitrate leaching from the root zone is either transmitted directly to streams by tile drains or transported to the groundwater system.

Thus, the partitioning of the water flux to drains, the drainage fraction [0,1] (Fig. 1), can be used as an **indicator of surface-water/groundwater vulnerability** to nitrogen application.

A **decision support tool** for fast predictions and mapping of drainage fraction could potentially support decision making on spatially differentiated regulation of nitrate emissions.

Drainage fraction can be estimated using a hydrological model, **however**…

Problem 1: Hydrological models for decision-making

- Running models for predictions is computationally very time-consuming
- Models are not always available at the site of interest or in the required resolution.

Solution → Developing a drainage fraction **metamodel** for faster predictions and predictions beyond model domain.

A metamodel is a computationally efficient surrogate for a more detailed numerical model [2]

Problem 2: Predicting beyond training data with metamodels

In areas beyond the training data, e.g. in different spatial regions/catchments that might be biased compared to the training data, the model has no information about the uncertainty of the predictions.

Thus, estimating the metamodel **transferability** to new (data) regions, and mapping the area to which the metamodel can be reliably applied, is highly relevant.

These two problems are the motivation of the study.

Methodology

1) Develop metamodel for drainage fraction prediction in the Storaa catchment (Fig. 2) and analyse transferability by within-catchment spatial cross-validation (Fig. 3).

2) Rank predictor variables by the variable importance scores of the trained models.

3) Compute distance measures in the predictor space between train and test data as a proxy for metamodel transferability.

4) Develop Transferability Index and map area to which the metamodel can be reliably applied.

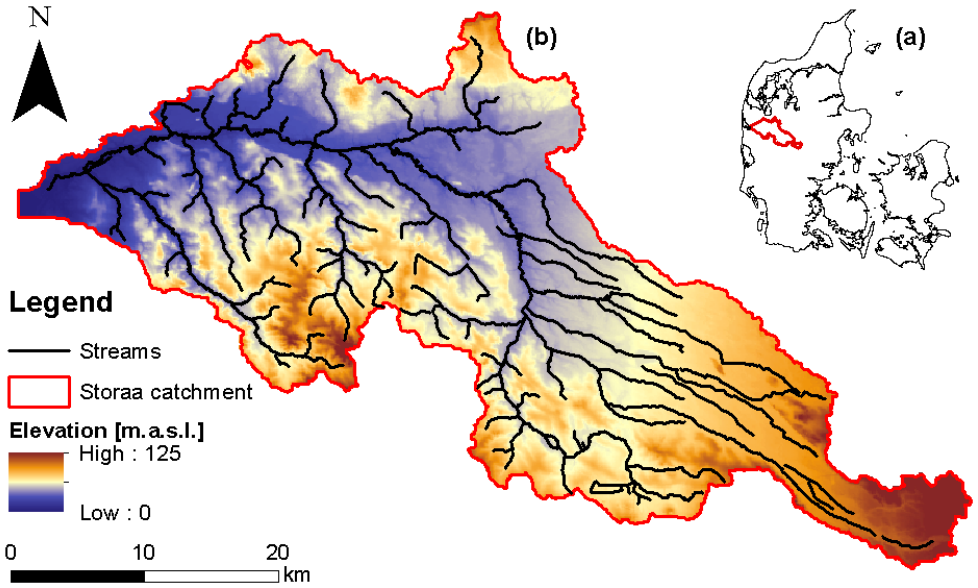


Figure 2 (top right): a) Location of the Storaa catchment in Denmark, b) the Storaa catchment, topography and streams.

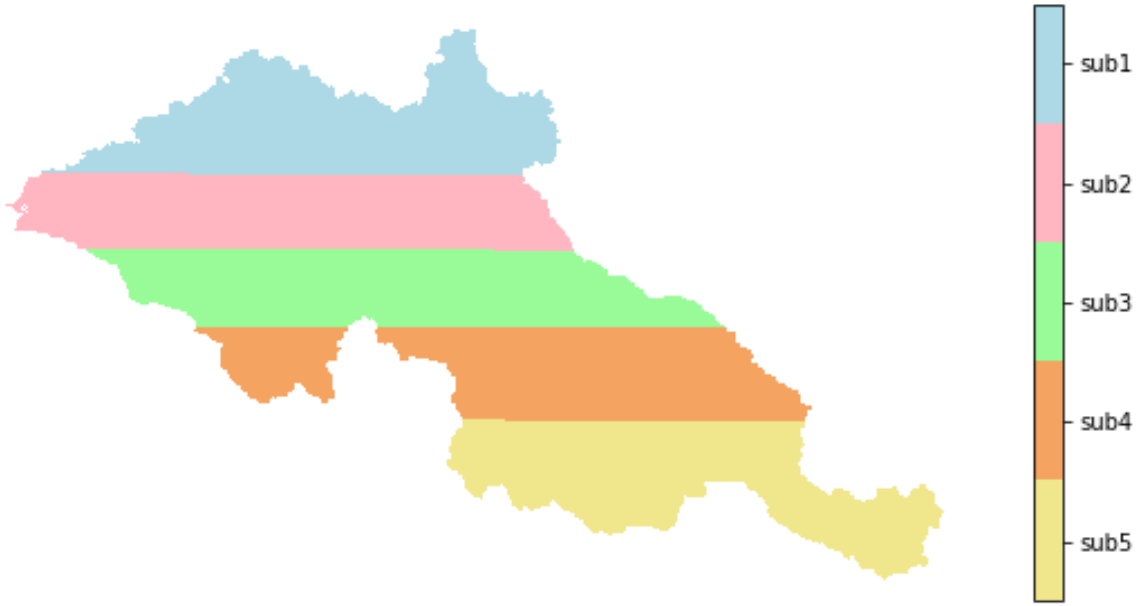


Figure 3 (bottom right): Five spatial subsets (sub1-5) of the Storaa catchment, each representing 20% of the data for spatial cross-validation.

Metamodel setup

- Random Forest algorithm applied on a dataset with random selection of 20% hold-out for validation (80-20 data set).
- 18 mappable predictor variables (Fig. 4).
- Target variable: drainage fraction estimates from a fully coupled surface water–groundwater Mike-SHE model for the Storaa catchment in 100m resolution (98,946 datapoints) (Fig. 5).

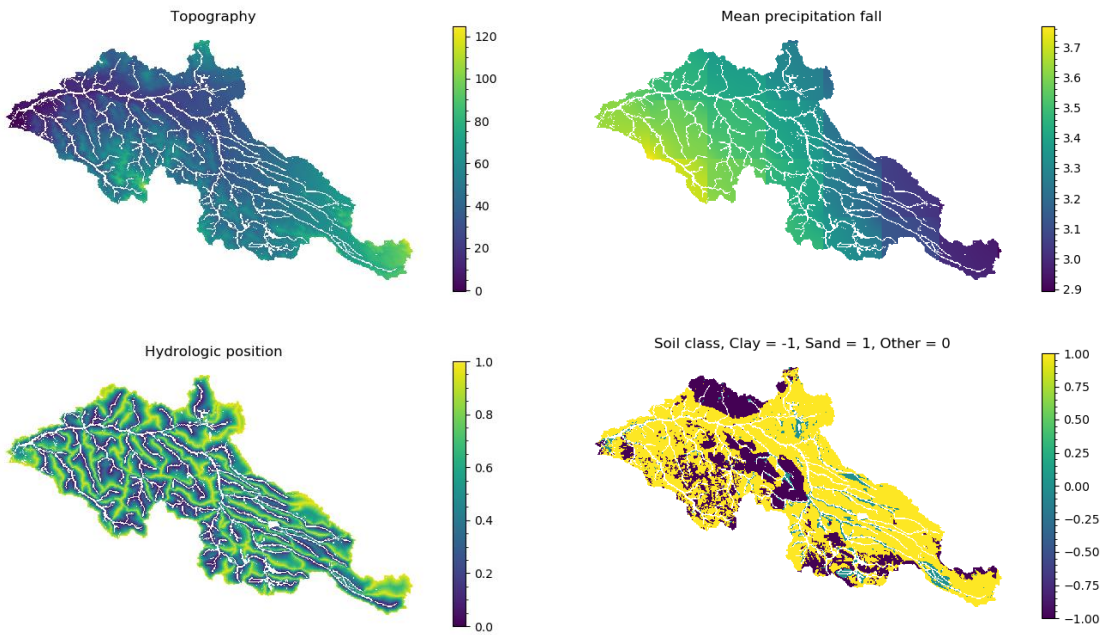


Figure 4. Examples of predictor variables: topography [m], mean precipitation fall [mm/day], hydrologic position and soil class.

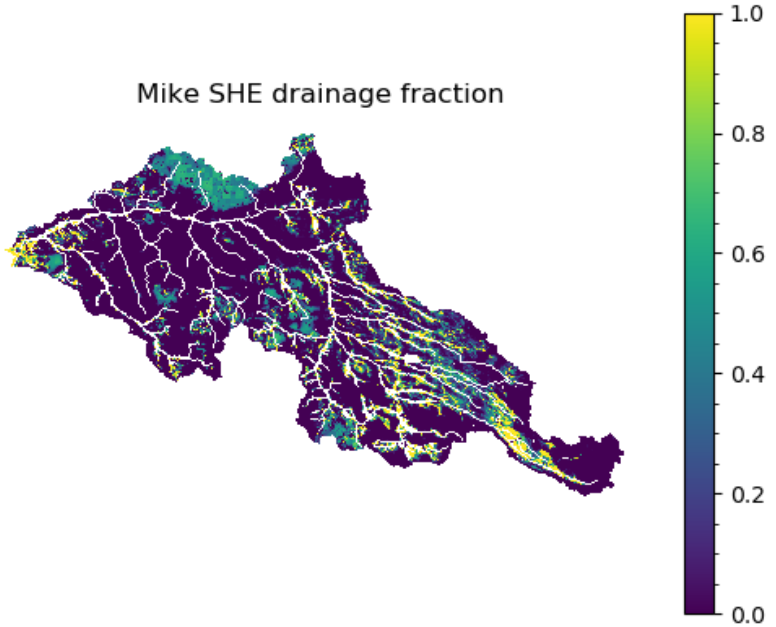


Figure 5. Drainage fraction [0,1] target variable derived from Mike-SHE.

Metamodel results

The metamodel is able to explain 78% of the variability of the drainage fraction in the 20% hold-out dataset (Fig. 6).

The most important predictor variable groups (Fig. 7) are:
1) topography
2) precipitation
3) soil texture

This finding corresponds with our physical understanding of the drainage system.

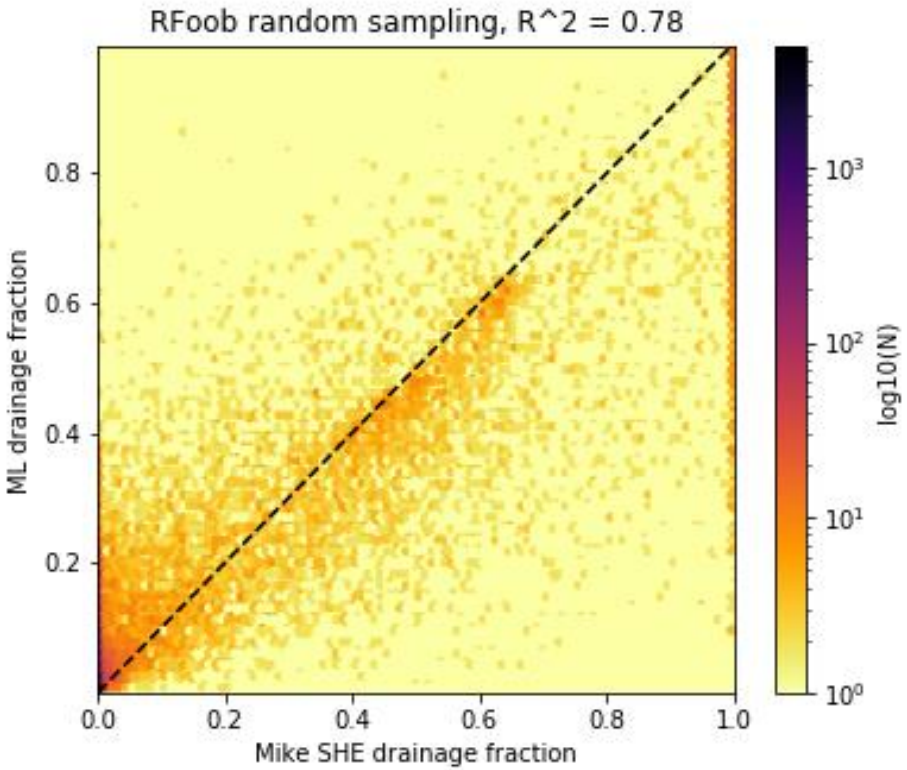


Figure 6. Density plot of Mike-SHE drainage fraction plotted against predicted metamodel (ML) drainage fraction for the randomly sampled test dataset. Dotted line is 1:1.

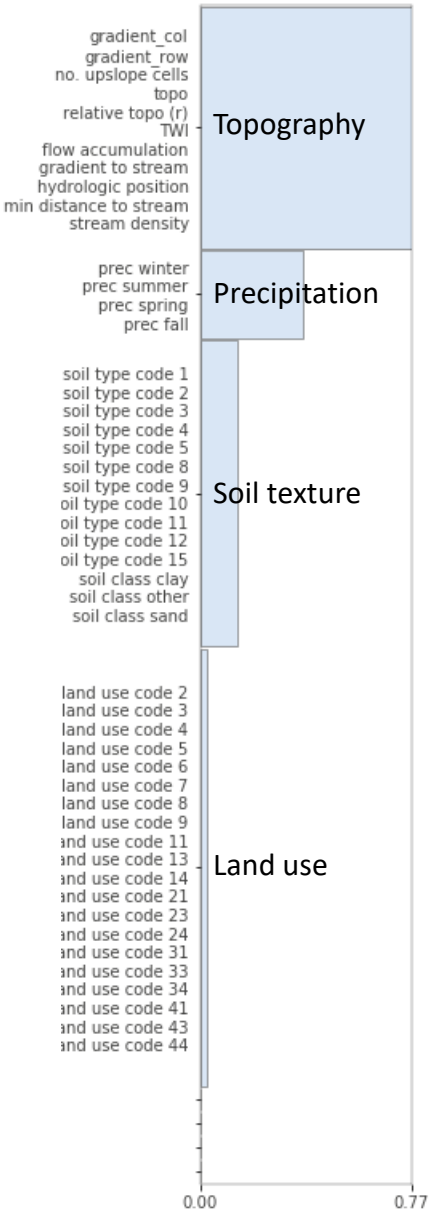


Figure 7. Grouped predictor variables ranked by importance (decrease in R2 after permutation).

5-fold spatial cross-validation

In addition to the random selection of the 80-20 dataset, the Random Forest algorithm was trained and tested on five spatially defined 80-20 datasets, where the 20% hold-out is a spatial subset of the catchment (Fig. 8).

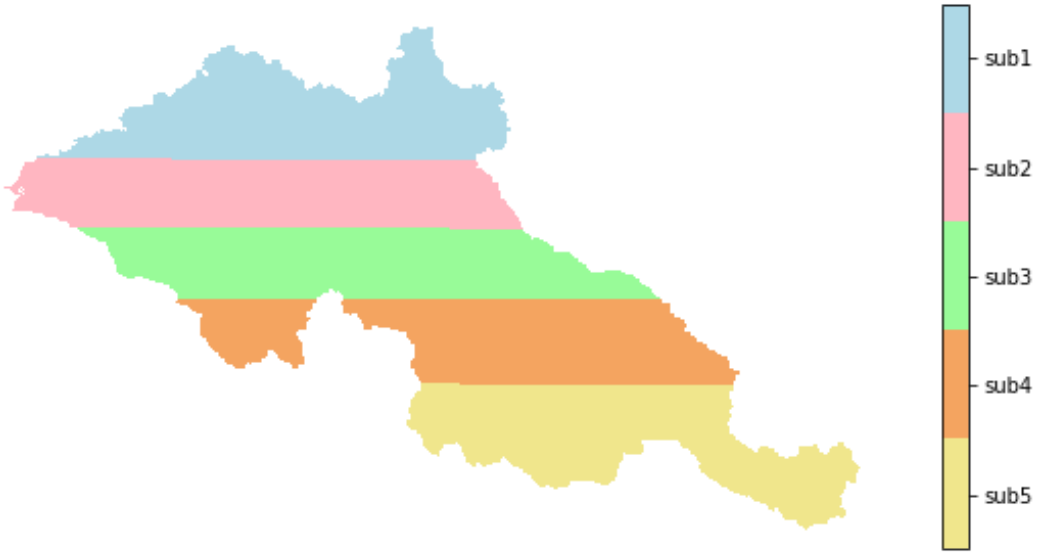


Figure 8. Spatial subsets (sub1-5) used for spatial cross-validation. Each subset (20% of the data) is used for testing the metamodel, trained on the remaning 80% of the data.

Model performance varies considerably between the spatial hold-outs and decreases significantly compared to the randomly sampled dataset (Fig. 9). The OOB R2 is only equal to the hold-out R2 for the random dataset.

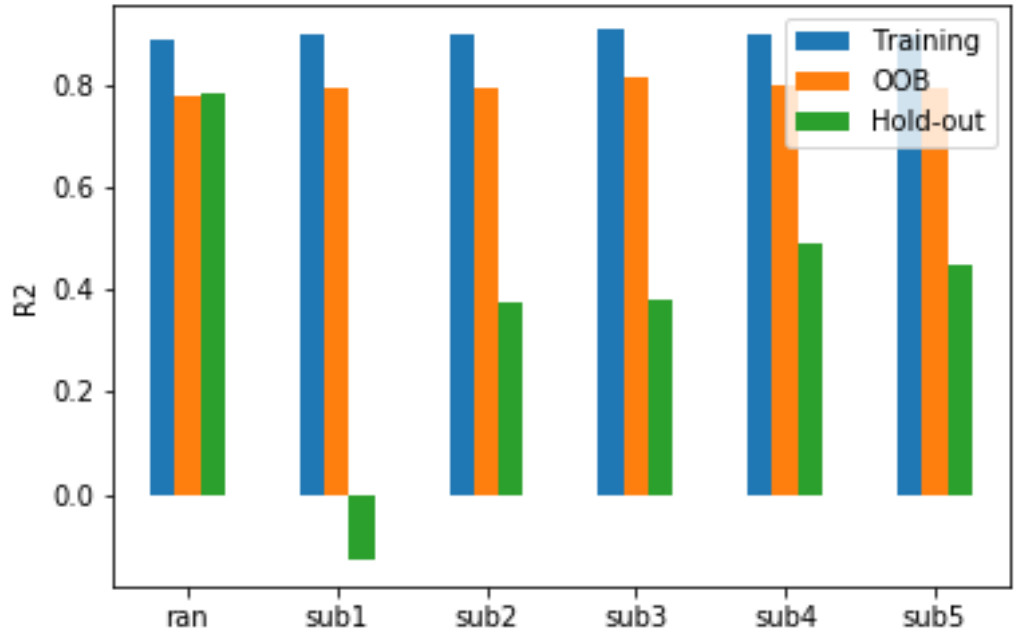
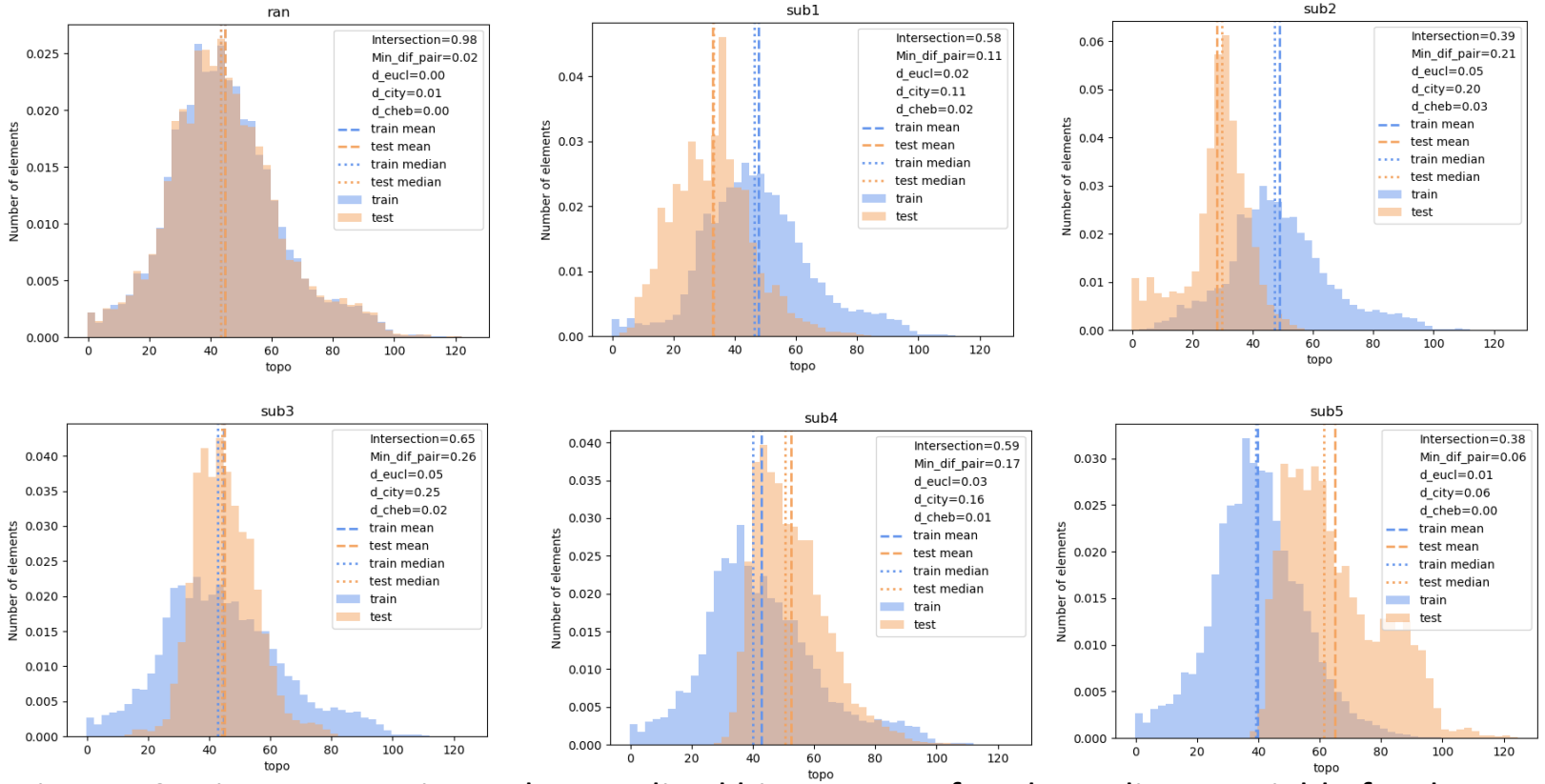


Figure 9. Model performance (R2) for the spatial subsets and randomly sampled dataset, for the training data (blue), out-of-bag (OOB) estimate (orange) and the 20% hold-out data.

Distance metrics

A set of histogram distance metrics (equations to the right [3]) are calculated for each predictor variable for each hold-out test set.
The train-test histograms for the random train-test dataset are near-identical whereas the spatial subsets vary in range and shape (Fig. 10).



$$D_{\text{mindifpair}}(A,B) = \min_{A,B}\left(\sum_{i,j=0}^{n-1} d(a_i, b_j\right)$$

$$D_{\text{Euclid}ian}(A,B) = \sum_{i=0}^{b-1} |H_i(A) - H_i(B)|$$

$$D_{\text{City}}(A,B) = \sqrt{\sum_{i=0}^{b-1} (H_i(A) - H_i(B))^2}$$

$$\text{Intersection}(A,B) = \frac{\sum_{i=0}^{b-1} \min(H_i(A), H_i(B))}{\sum_{i=0}^{b-1} H_i(A)}$$

Figure 10. Distance metrics and normalized histograms of each predictor variable for the six 80-20 datsets are calculated, here exemplified by the topography variable.

Metamodel transferability

Preliminary results:
The positive correlation between RMSE and distance metrics and the negative correlation between R2 and distance metrics (Fig. 11) indicates that distance metrics could be used as a first step towards a metamodel transferability index.

Work in progress:
1) Compute Transferability Index based on mean distance metrics weighted by predictor variable importance.
2) Use Transferability Index to map area to which the metamodel can be reliably applied and locate areas where additional training data is required to increase it.

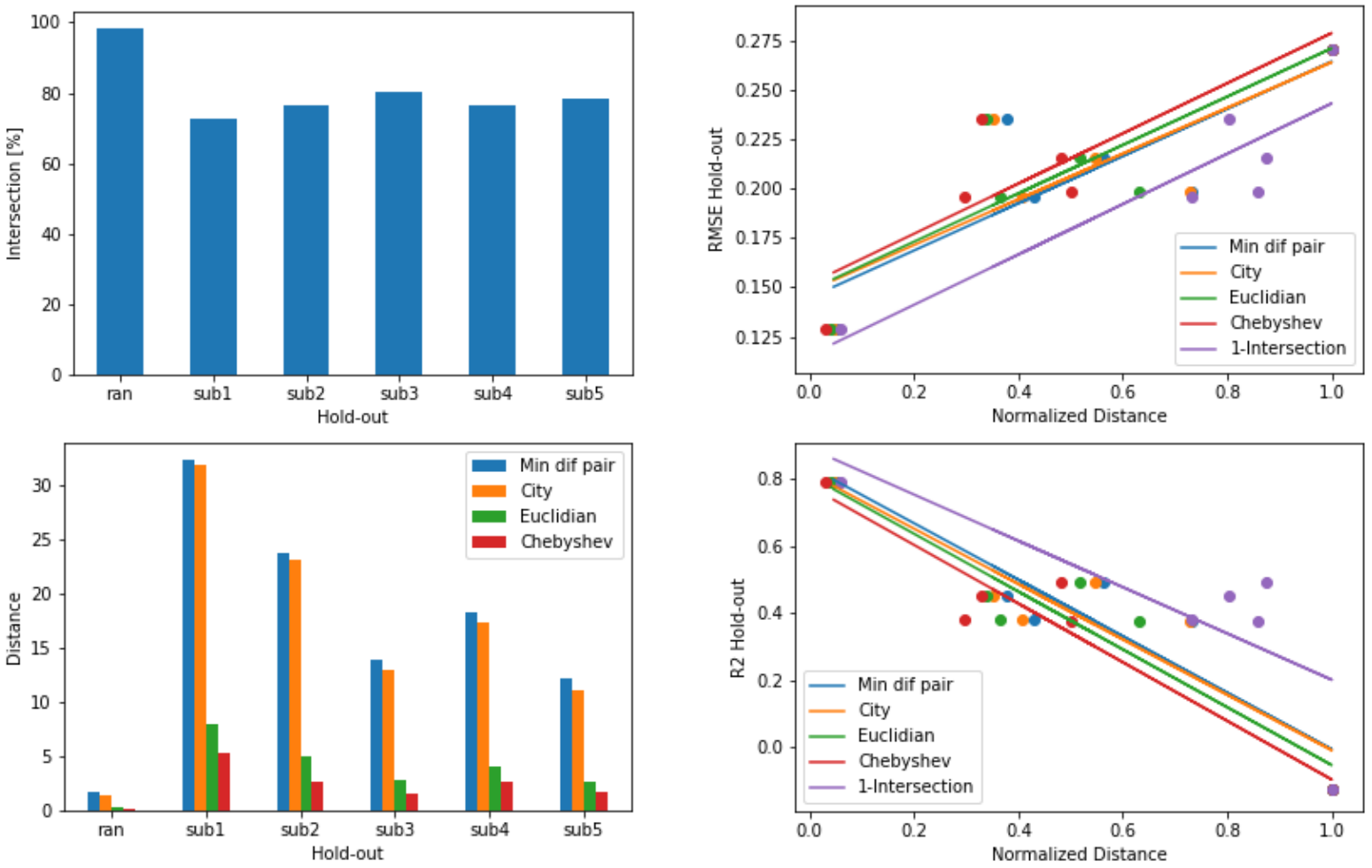


Figure 11. Top left: mean histogram intersection across predictor varibles for each train-test dataset. Bottom left: mean distance metrics across predictor variables for each train-test dataset. Normalized distance metrics plotted against RMSE (top right) and R2 (bottom right), 6 points of similar color represent the 6 train-test datasets.

Discussion and Conclusions

Machine learning based metamodels have enjoyed increasing attention in the context of hydrology [4], due to their capacity of making fast predictions and their seemingly high performance. We found that:

- The Random Forest metamodel was capable of mapping drainage fraction with an R2 of 0.78 for a 80-20% randomly sampled train-test dataset.
- Metamodel performance varied considerably between spatial subsets of the catchment and was significantly lower than the Random Rorest out-of-bag-estimates and the metamodel performance based on random sampling.
- Histogram distance metrics in the predictor space could be used to calculate a metamodel Transferability Index.
- Improved understanding of the conditions for which metamodels provide reliable predictions, would enable decision-makers to evaluate whether the model used is fit for the predictive task at hand.

References

[1] European Commission. (2010). The EU Nitrates Directive. Retrieved December 18, 2019, from https://publications.europa.eu/en/publication-detail/-/publication/b2f78dad-e7cb-41c7-8f31-c71653f95631/language-en/format-PDF/source-100865477

[2] Blanning, R. W. (1975). The construction and implementation of metamodels. Simulation, 24(6), 177–184. https://doi.org/10.1177/003754977502400606

[3] Cha, S. H., & Srihari, S. N. (2002). On measuring the distance between histograms. *Pattern Recognition*, *35*(6), 1355–1370. https://doi.org/10.1016/S0031-3203(01)00118-2

[4] Asher, M. J., Croke, B. F. W., Jakeman, A. J., & Peeters, L. J. M. (2015). A review of surrogate models and their application to groundwater modeling. *Water Resources Research*, *51*(8), 5957–5973. https://doi.org/10.1002/2015WR017464