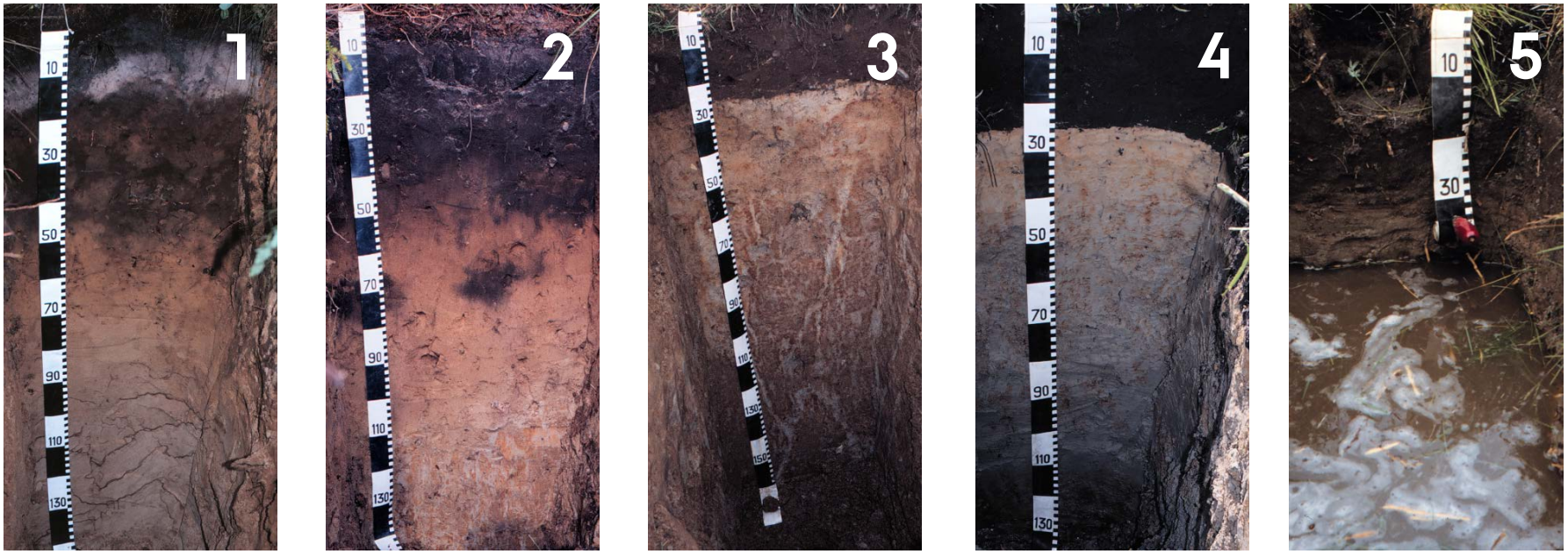
MAPPING DRAINAGE CLASSES IN DENMARK BY MEANS OF DECISION TREE CLASSIFICATION

Anders Bjørn Møller, Bo Vangsø Iversen, Amélie Beucher and Mogens Humlekrog Greve



METHODS

Decision tree classification (C5.0)

1702 soil profiles; 31 predictor variables

Boosting vs. bagging

Differentiated costs vs. equal costs for misclassifcation

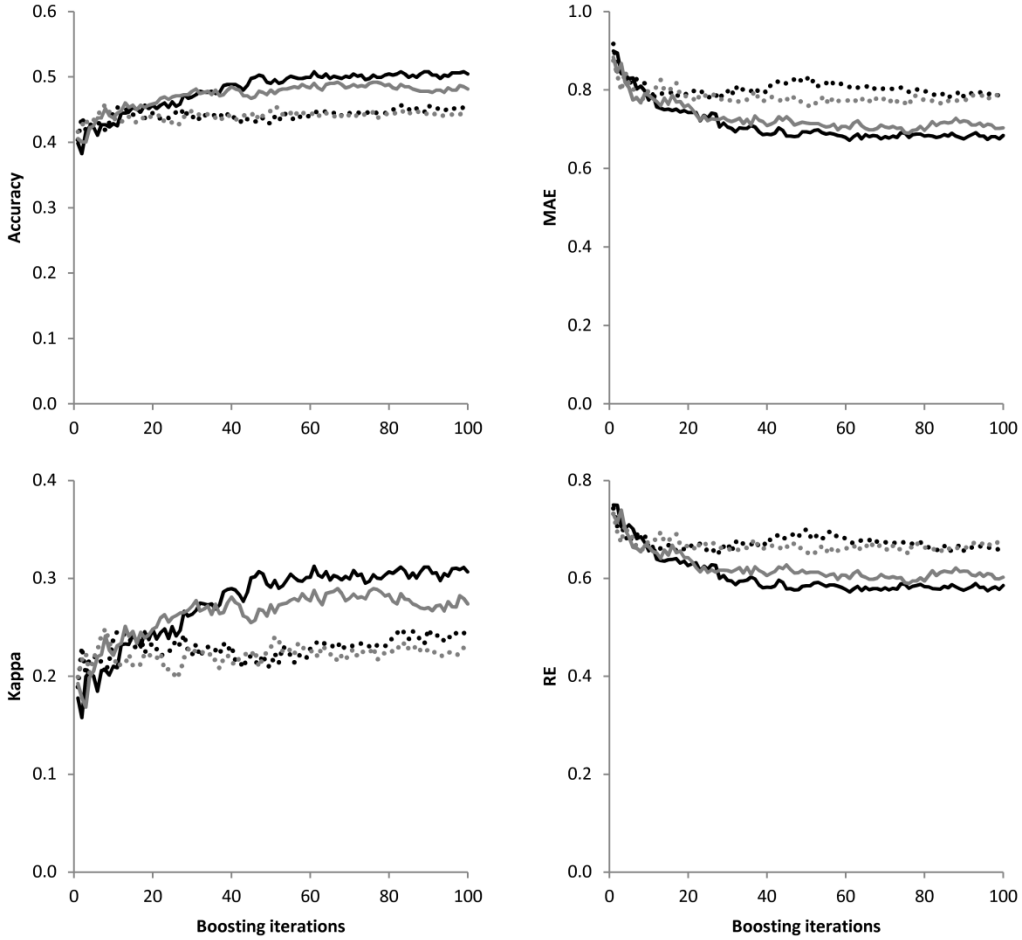
All predictor variables vs. selected predictor variables

Probability bagging vs. ordinary bagging

Cost matrix for misclassification

		Predicted DC				
		1	2	3	4	5
Reference DC	1	0	1	2	3	4
	2	1	0	1	2	3
	3	2	1	0	1	2
	4	3	2	1	0	1
	5	4	3	2	1	0

BOOSTING OPTIMIZATION



Black: Equal costs

Grey: Differentiated costs

Solid line: All predictors

Dots: Selected predictors

BAGGING OPTIMIZATION

30 repetitions.

Times when each model had the best performance as measured by…

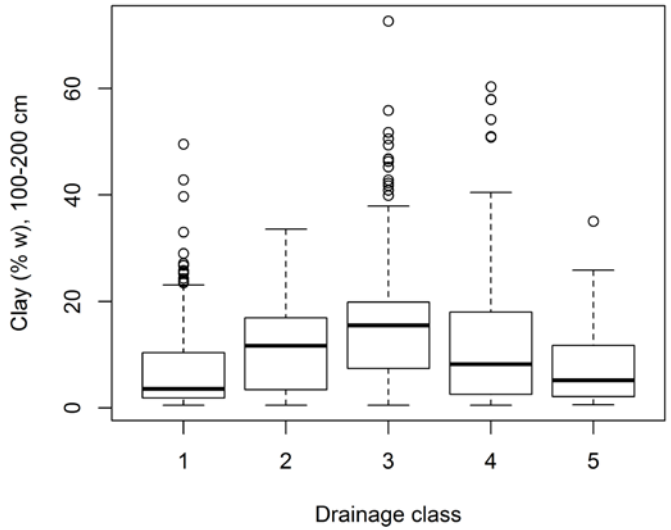
Model	Best Accuracy	Best Kappa	Best MAE	Best RE
Undifferentiated costs, all variables	9	5	6	5
Undifferentiated costs, selected variables	0	0	0	0
Differentiated costs, all variables	12	14	7	15
Differentiated costs, selected variables	8	9	8	6
Probability bagging, all variables	7	2	15	4
Probability bagging, selected variables	0	0	0	0

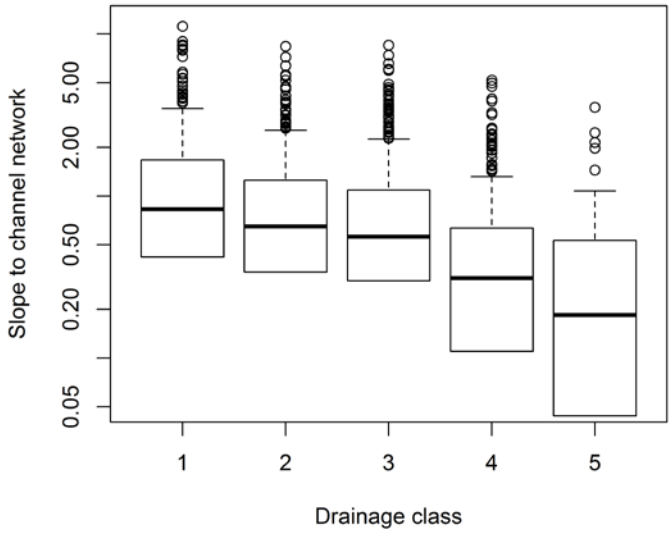
MODEL PERFORMANCE

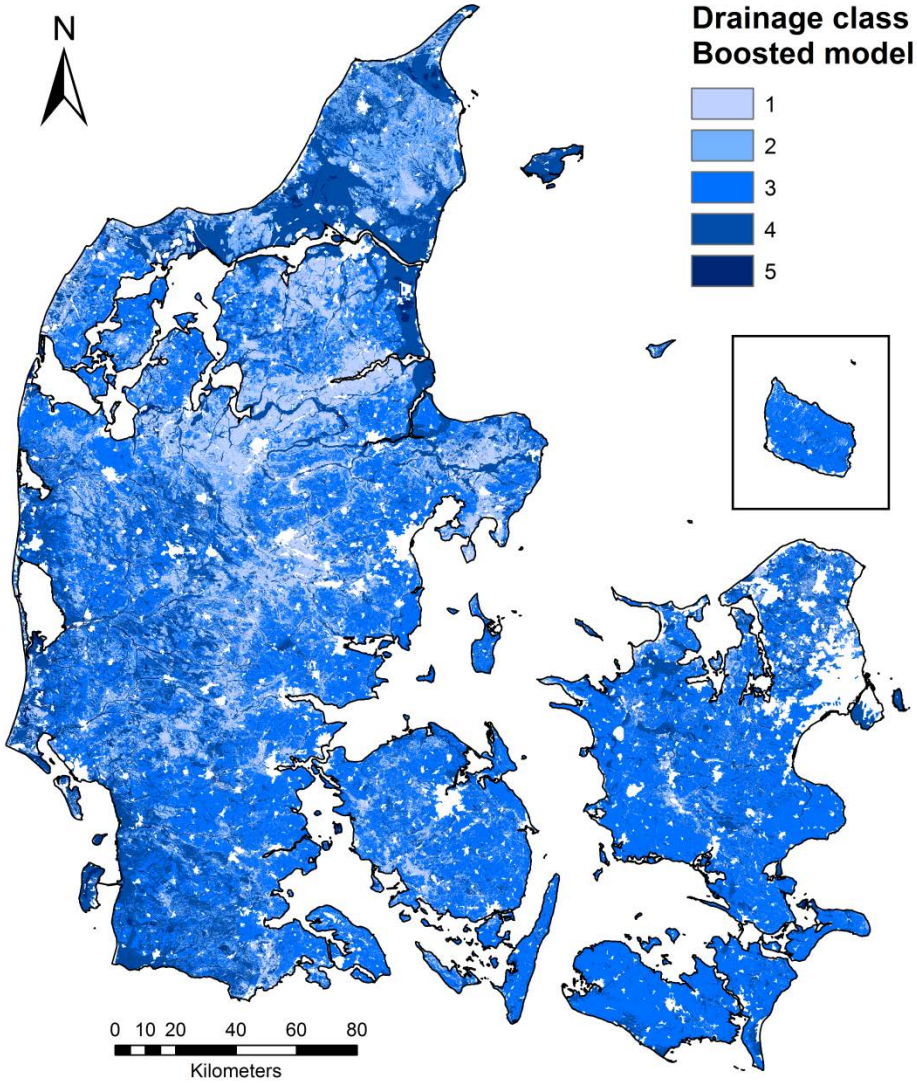
	Boosting	Bagging	Boosting	Bagging
	Validation sample	Validation sample	Cross validation	Cross validation
Accuracy	0.508	0.520	0.471	0.488
Kappa	0.313	0.330	0.259	0.281
MAE	0.672	0.656	0.757	0.749
RE	0.572	0.553	0.642	0.636

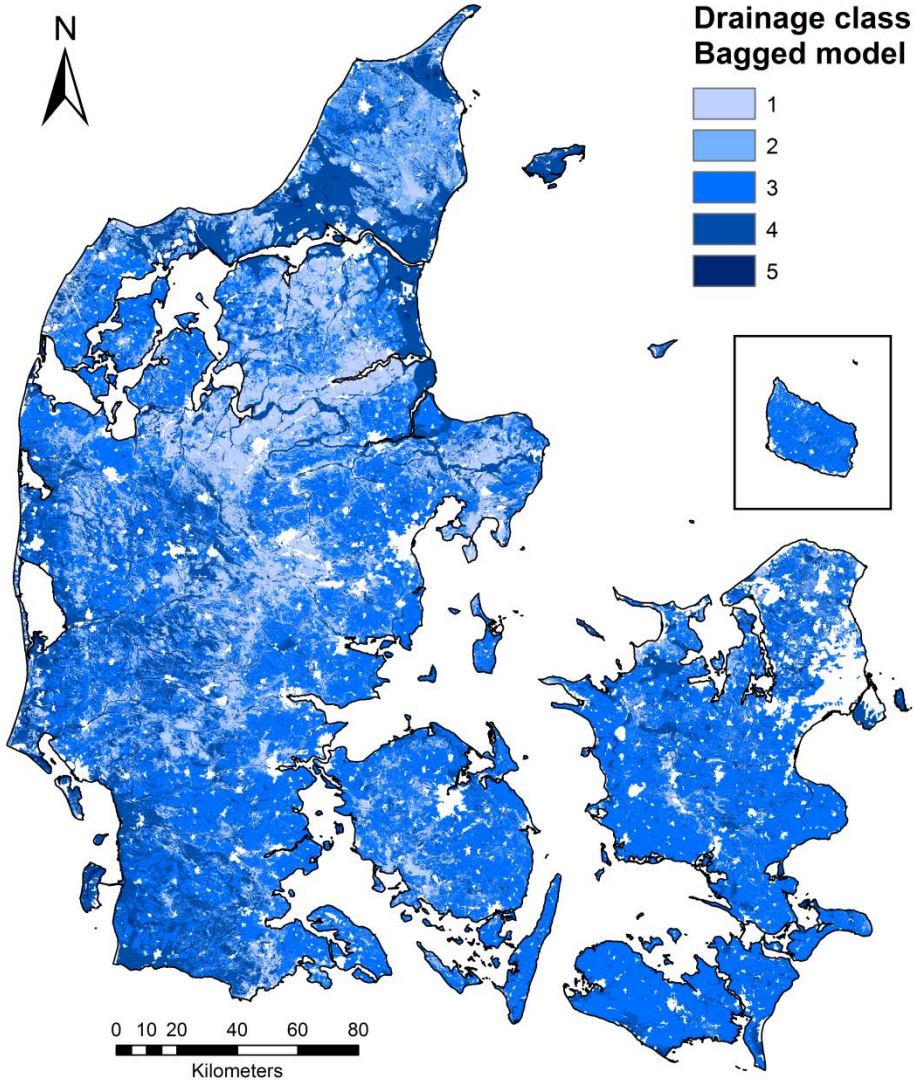
PREDICTOR USAGE

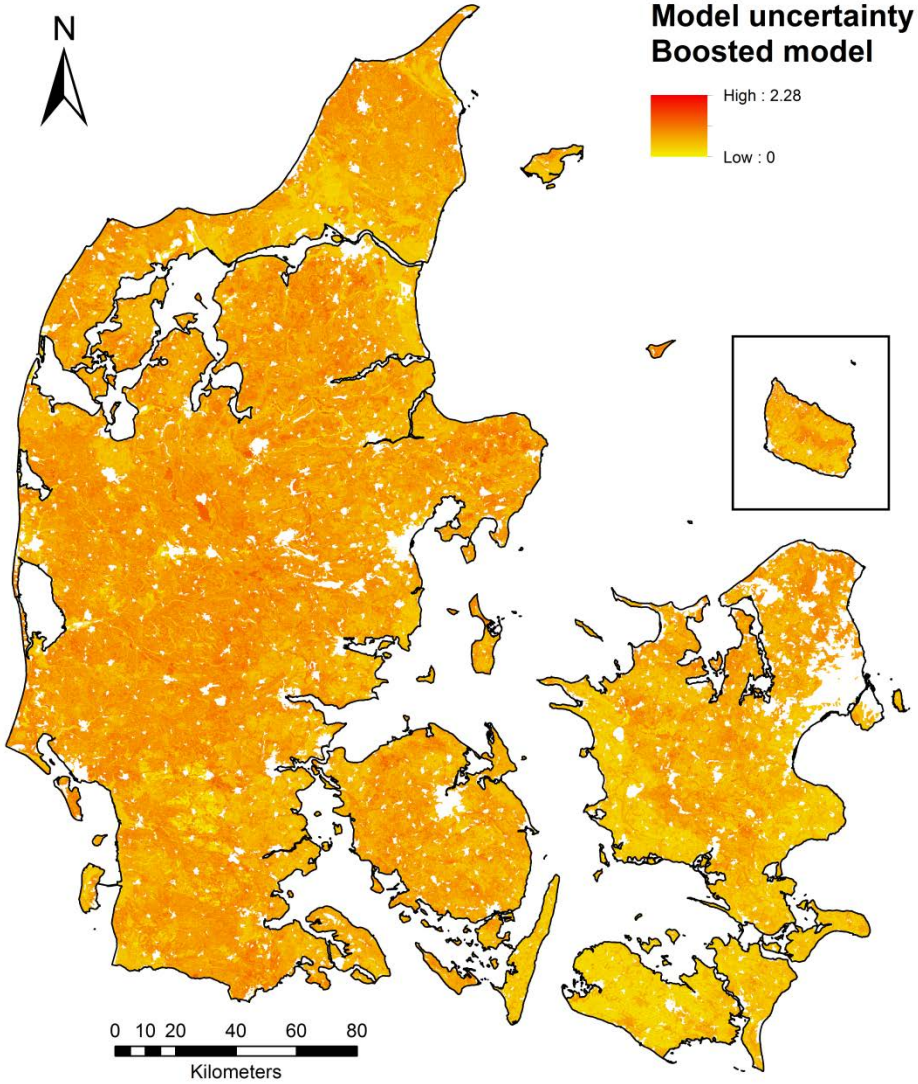
Top 5		
	Boosting	Bagging
Predictor	Mean (Range)	Mean (Range)
Wetlands	71.2 (2.1 - 98.4)	68.9 (1.2 - 100.0)
Slope to channel network	72.7 (5.6 - 100.0)	65.8 (3.4 - 100.0)
Clay content (100 – 200 cm)	69.5 (16.3 - 92.5)	65.7 (4.9 - 100.0)
Land use	93.6 (14.7 - 100.0)	63.2 (3.3 - 100.0)
Geology	76.0 (36.6 - 100.0)	63.1 (11.1 - 100.0)

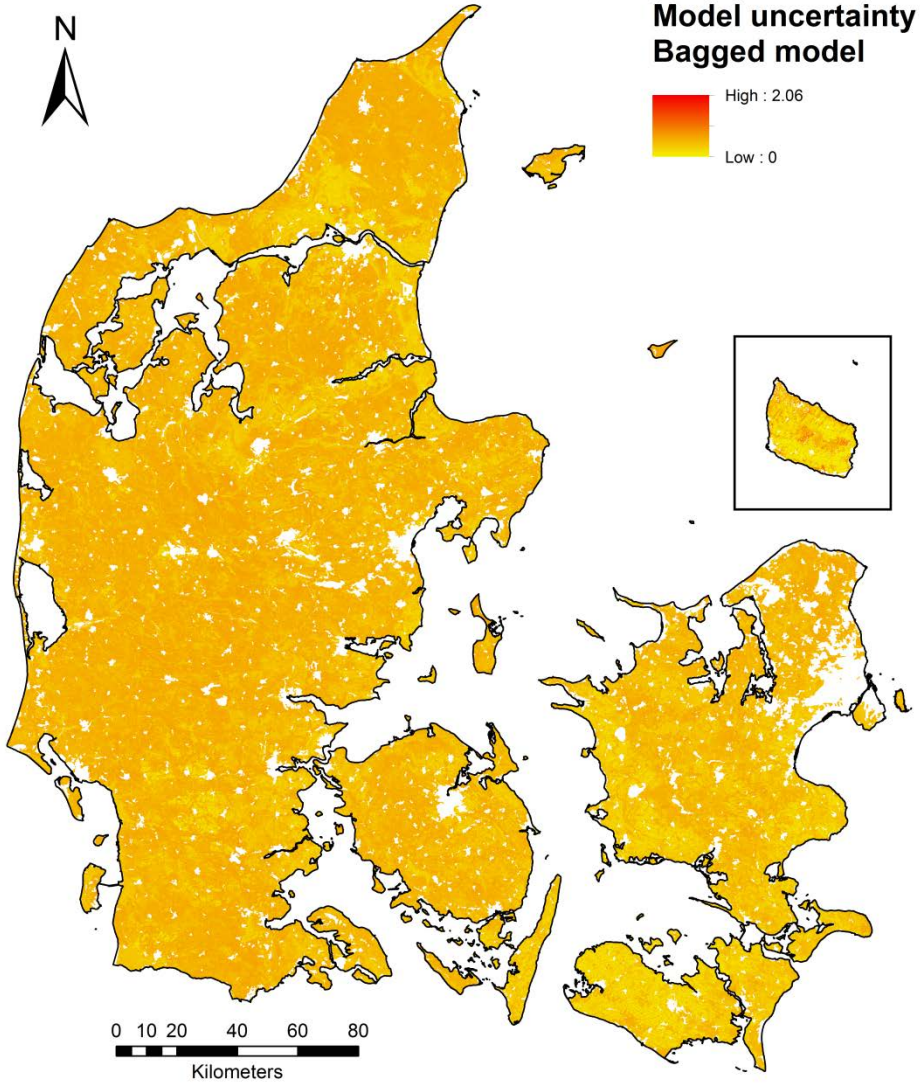




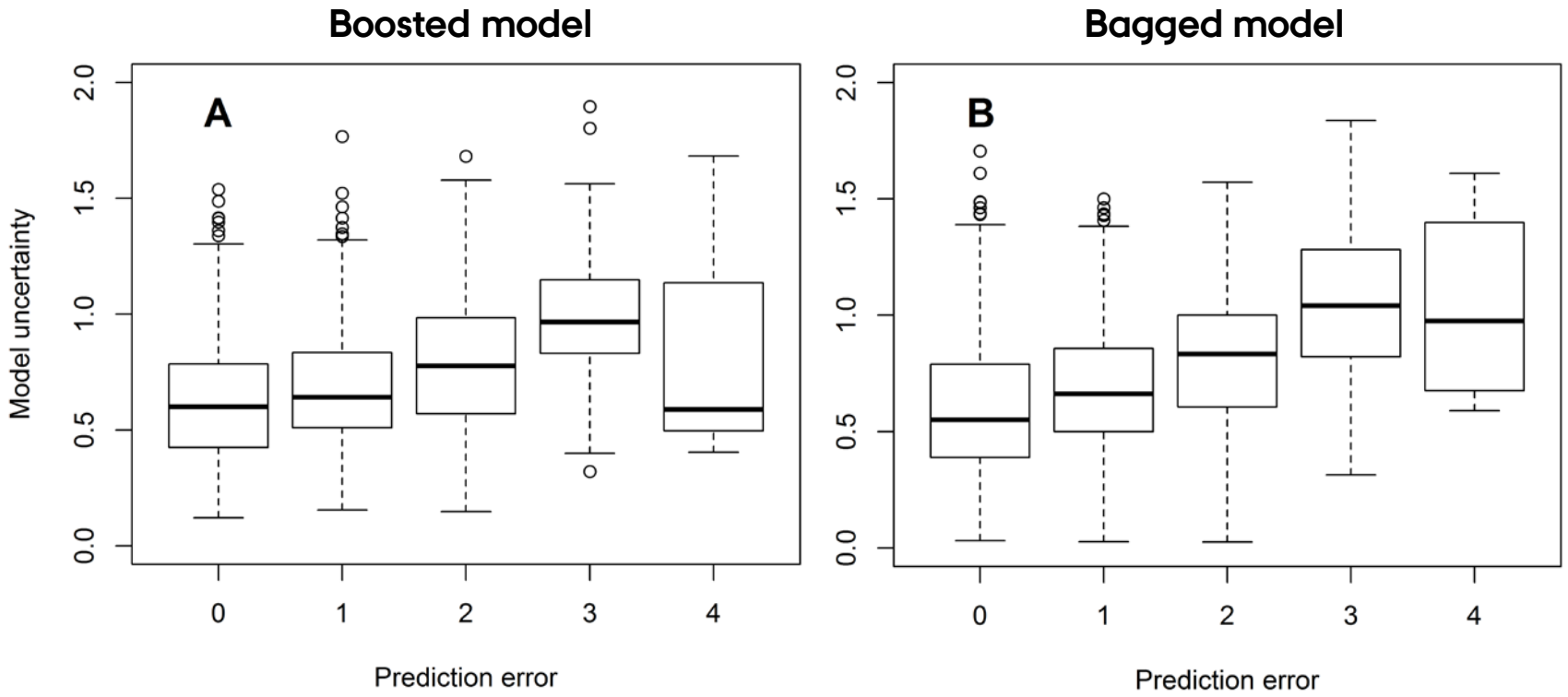

N
Drainage class
Boosted model
1
2
3
4
5
0 10 20 40 60 80
Kilometers


N
Drainage class
Bagged model
1
2
3
4
5
0 10 20 40 60 80
Kilometers


N
Model uncertainty
Boosted model
High : 2.28
Low : 0
0 10 20 40 60 80
Kilometers


N
Model uncertainty
Bagged model
High : 2.06
Low : 0
0 10 20 40 60 80
Kilometers

MODEL UNCERTAINTY VS. PREDICTION ERROR



CONCLUSIONS

Best model:

- Bagging.
- All predictior variables.
- Differentiated costs for misclassification.

Boosting not improved by differentiated costs for misclassification.

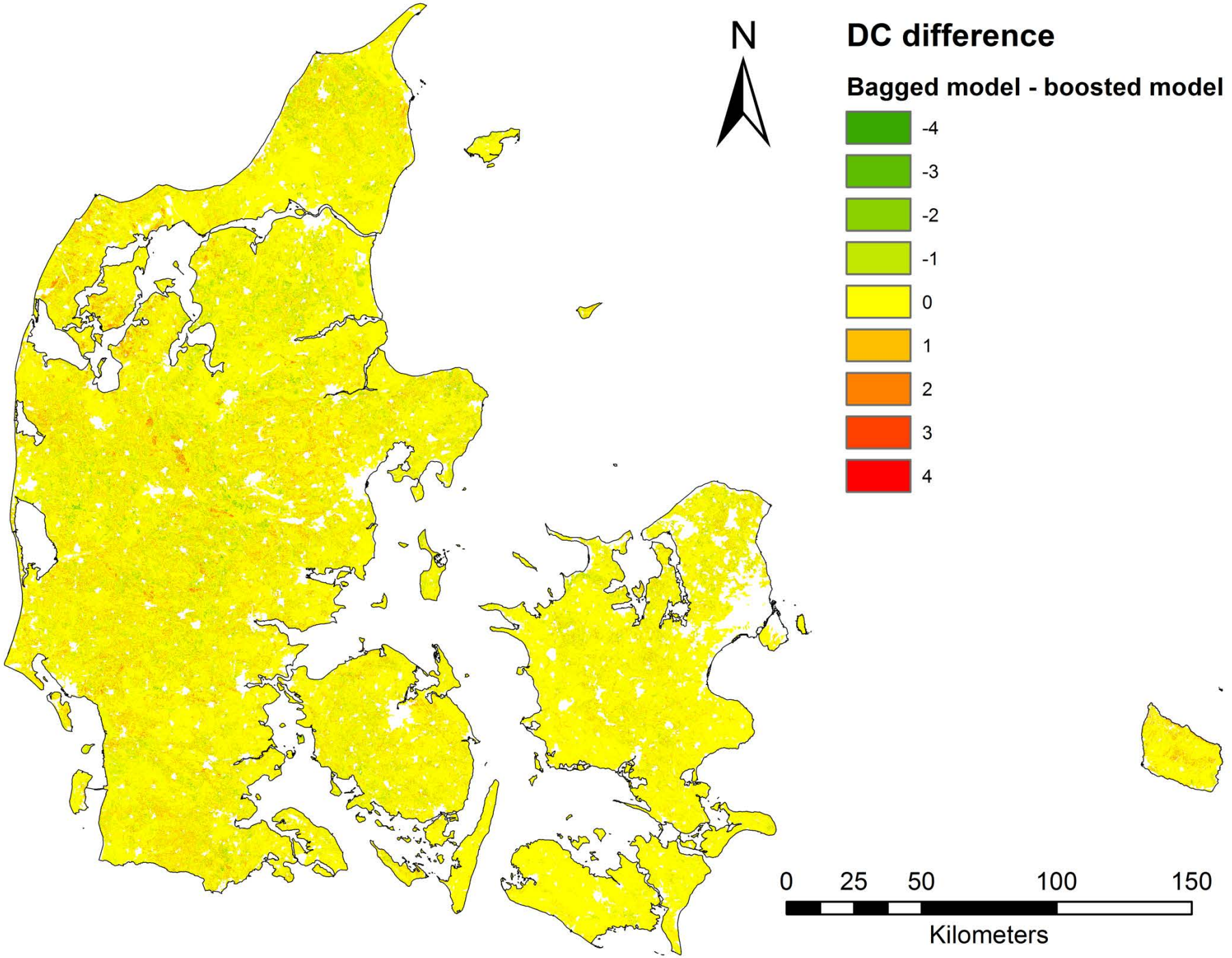
Similar results from boosting and bagging.

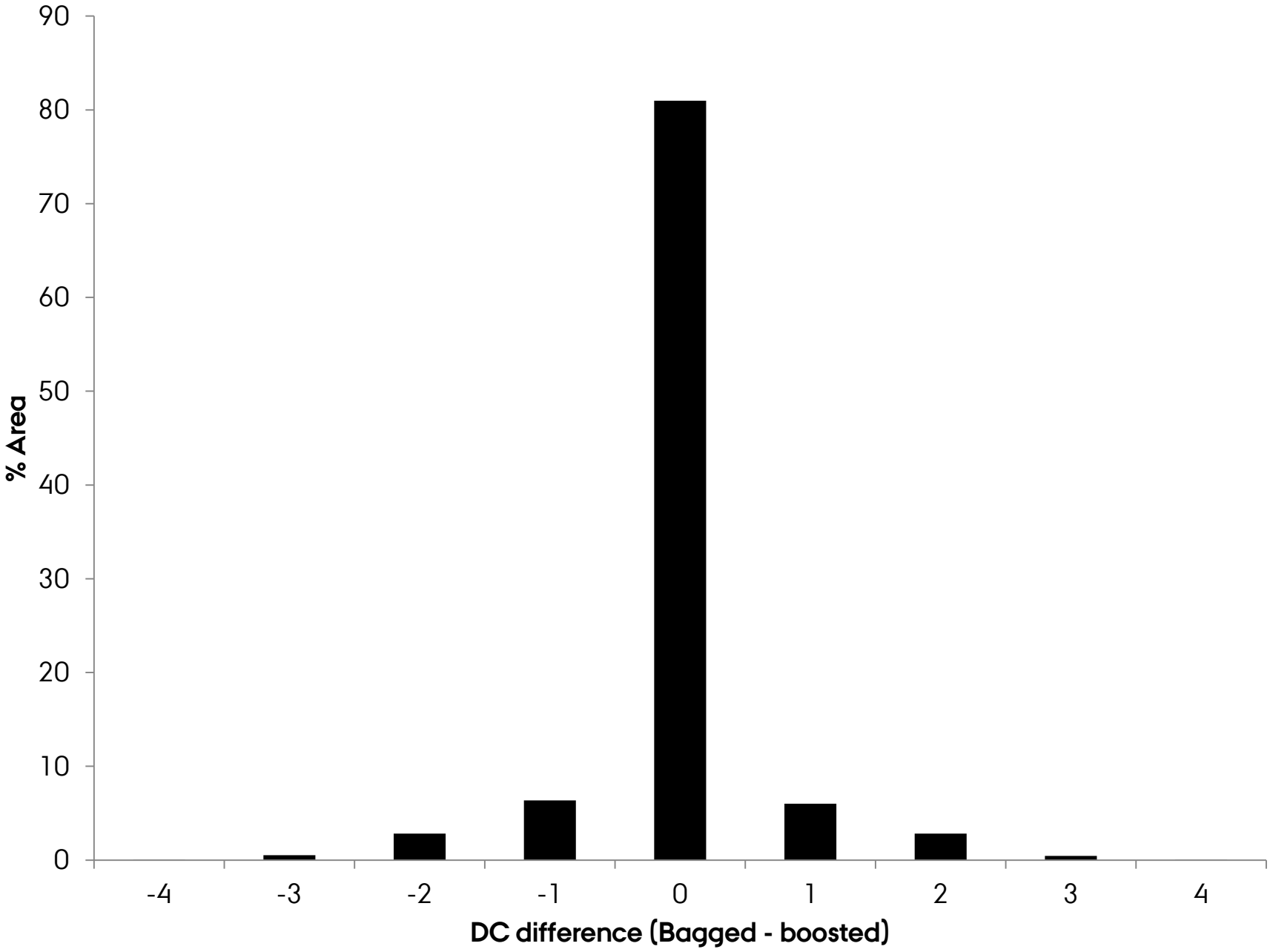
Poorer performance with reduced number of predictors.

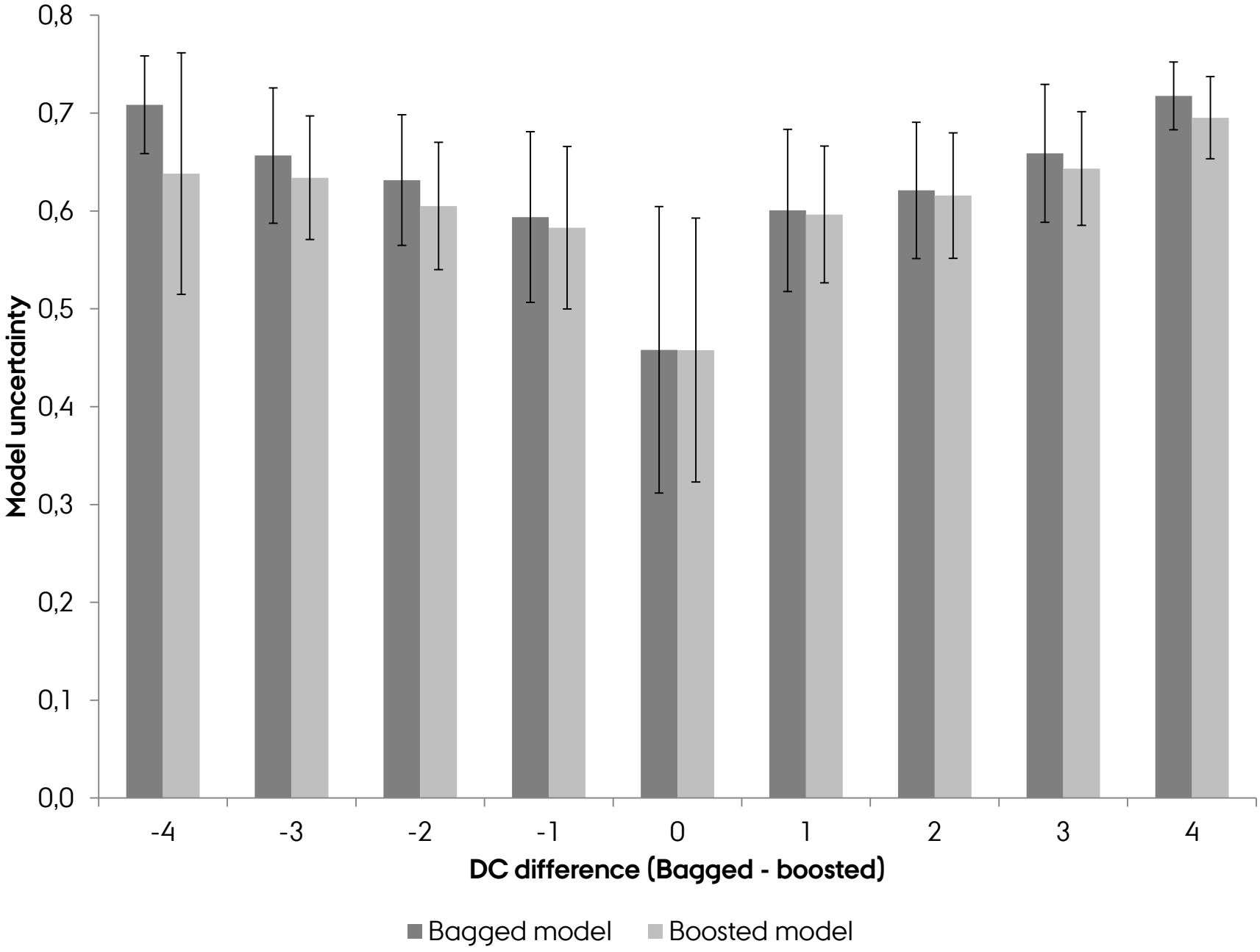
Top 5 variables: Wetlands, Slope to the channel network, Clay content (100 – 200 cm), Land use and Geology.

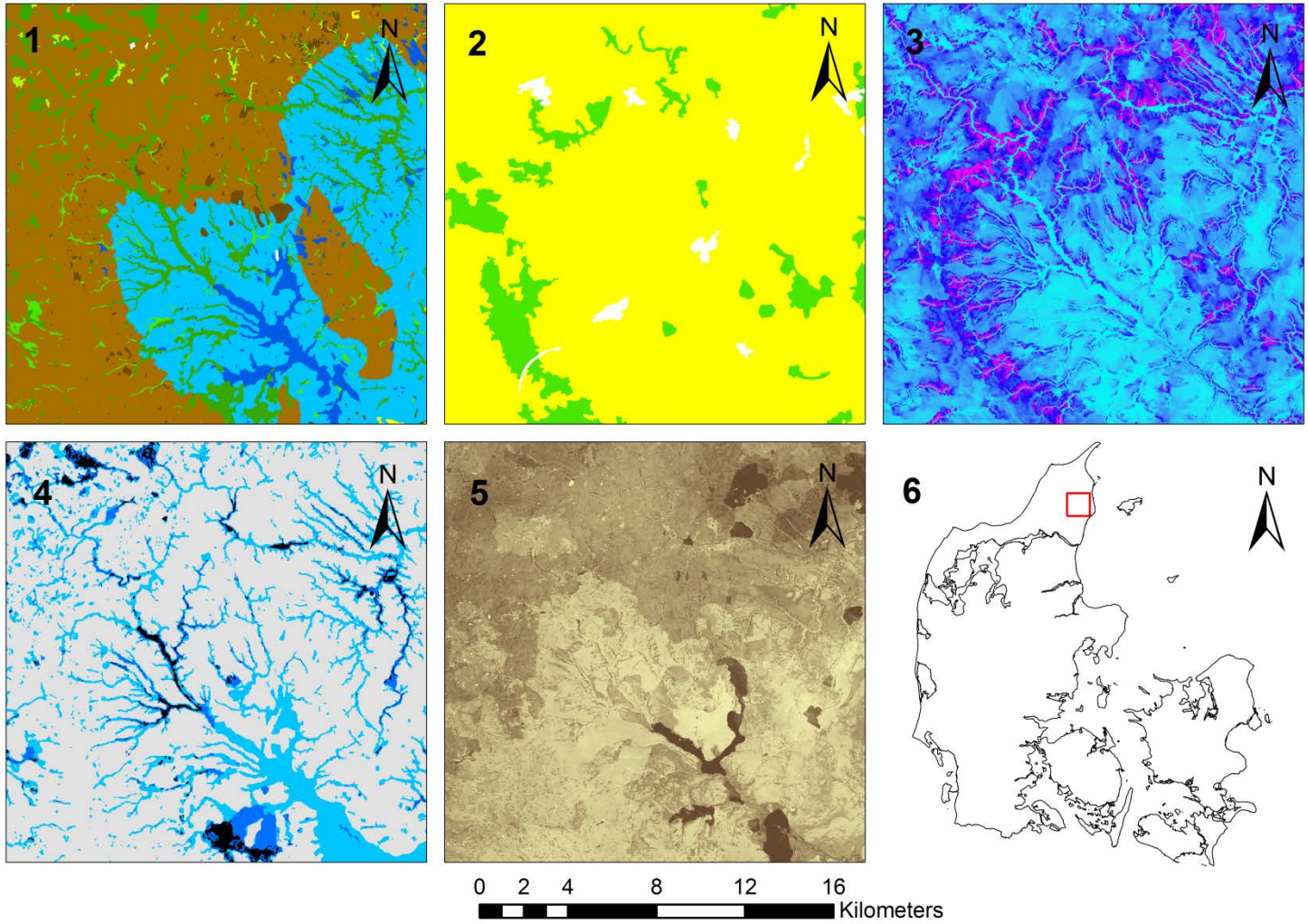
Model uncertainty related to prediction error.

Thank you ☺


N
DC difference
Bagged model - boosted model
-4
-3
-2
-1
0
1
2
3
4
0
25
50
100
150
Kilometers


90
80
70
60
50
40
30
20
10
0
% Area
-4
-3
-2
-1
0
1
2
3
4
DC difference (Bagged - boosted)


0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
Model uncertainty
-4
-3
-2
-1
0
1
2
3
4
DC difference (Bagged - boosted)
Bagged model
Boosted model


1
N
2
N
3
N
4
N
5
N
6
N
0 2 4 8 12 16
Kilometers
1: Geology
Aeolian sand
Freshwater clay
Freshwater sand
Freshwater peat
Marine clay
Marine sand
Glacial clay
Glacial sand
Meltwater clay
Meltwater sand
2: Land use
Agriculture
Natural vegetation
Wetland vegetation
3: Slope to channel
High : 27
Low : 0
4: Wetlands
Non-wetlands
Wetlands
Central wetlands
Peat
5: Clay 100 - 200 cm (%)
High : 28.7
Low : 0
6: Example area
Extent

Boosted model

Bagged model