Expanding Models of Lake Trophic State to Predict Cyanobacteria in Lakes:

A Data Mining Approach

Jeffrey W. Hollister, W. Bryan Milstead, and Betty J. Kreakie

U.S. Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, Narragansett, RI 02882

Introduction

Cyanobacteria are an important taxonomic group associated with harmful algal blooms in lakes. Understanding the drivers of cyanobacteria presence has important implications for lake management and for the protection of human and ecosystem health. Chlorophyll a concentration, a measure of the biological productivity of a lake, is one such driver and is largely, although not exclusively, determined by nutrient inputs. As nutrient inputs increase, productivity increases and lakes transition from low trophic state (e.g. oligotrophic) to higher trophic states (e.g. hypereutrophic). These broad trophic state classifications are associated with ecosystem health and ecosystem services/disservices (e.g. recreation, aesthetics, fisheries, and harmful algal blooms). Thus, models of trophic state might be used to predict things like cyanobacteria.

We have three goals for this preliminary research:

1. Build and assess models of lake trophic state
2. Assess ability to predict trophic state in lakes without available *in situ* water quality data
3. Explore association between cyanobacteria and trophic in order to expand models.

Data and Modeling Methods

Data We utilize four primary sources of data for this study. These are outlined below and in Table 1.

1. National Lakes Assessment (NLA) 2007: The NLA data were collected during the summer of 2007 and the final data were released in 2009. With consistent methods and metrics collected at 1056 locations across the conterminous United States (Map 1), the NLA provides a unique opportunity to examine broad scale patterns in lake productivity. The NLA collected data on biophysical measures of lake water quality and habitat. For this analysis we primarily examined the water quality measurements from the NLA (USEPA 2009).
2. National Land Cover Dataset (NLCD) 2006: The NLCD is a nationally collected land use land cover dataset. We collected total land use land cover and total percent impervious surface within a 3 kilometer buffer surrounding the lake to examine larger landscape-level effects (Homer et al. 2004; Xian, Homer, and Fry 2009).

3. Modeled lake morphometry: Various measures of lake morphometry (i.e. depth, volume, fetch, etc.) are important in understanding lake productivity, yet many of these data are difficult to obtain for large numbers of lakes over broad regions. To add this information we modeled lake morphometry (J. Hollister and Milstead 2010; Jeffrey W. Hollister, Milstead, and Urrutia 2011; J. Hollister 2013; Jeffrey W Hollister and Milstead).

4. Estimated Cyanobacteria Biovolumes: Cyanobacteria biovolumes is a truer measure of Cyanobacteria dominance than abundance as there is great variability in the size within and between species. To account for this, Beaulieu *et al.* (2013) used literature values to estimate biovolumes for the taxa in the NLA. They shared this data and we have summed that information on a per-lake basis.

```
## Error: object 'predictors_all' not found
```

```
## Error: object 'predictors_all' not found
```

```
## Error: object 'type' not found
```

```
## Error: object 'type' not found
```

```
## Error: object 'predictors_all' not found
```

```
## Error: object 'hkm2014Data' not found
```

```
## Error: error in evaluating the argument 'obj' in selecting a method for functi
##   error in evaluating the argument 'obj' in selecting a method for function '
```

```
## Error: error in evaluating the argument 'x' in selecting a method for functior
```

```
## Error: error in evaluating the argument 'obj' in selecting a method for functi
```

```
## Error: object 'lakes_dd' not found
```

```
## Error: object 'lakes_dd' not found
```

Predicting Trophic State with Random Forests

Random forest is a machine learning algorithm that aggregates numerous decision trees in order to obtain a consensus prediction of the response categories (Breiman 2001). Bootstrapped sample data is recursively partitioned according to a given random subset of predictor variables and completely grown without pruning. With each new tree, both the sample data and predictor variable subset is randomly selected.

While random forests are able to handle numerous correlated variables without a decrease in prediction accuracy, unusually large numbers of related variables can reduce accuracy and increase the chances of over-fitting the model. This is a problem often faced in gene selection and in that field, a variable selection method based on random forest has been succesfully applied (Díaz-Uriarte and De Andres 2006). We use varselRF in R to initially examine the importance of the water quality and GIS derived variables and select a subset, the reduced model, to then pass to random forest(Diaz-Uriarte 2010).

Using R's randomForest package, we pass the reduced models selected with varSelRF and calculate confusion matrices, overall accuracy and kappa coeffecient (Liaw and Wiener 2002). From the reduced model random forests we collect a consensus prediction and calculate a confusion matrix and summary stats.

Model Details

Using a combination of the `varSelRF` and `randomForest` we ran models for six combinations of variables and trophic state classifications. These combinations included different combinations of the Chlorphyll *a* trophic states (Table 2) along with all variables and the GIS only variables (i.e. no *in situ* infromation). The six model combinations were:

1. Chlorophyll *a* trophic state - 4 class = All variables (*in situ* water quality, lake morphometry, and landscape)
2. Chlorophyll *a* trophic state - 3 class = All variables (*in situ* water quality, lake morphometry, and landscape)
3. Chlorophyll *a* trophic state - 2 class = All variables (*in situ* water quality, lake morphometry, and landscape)
4. Chlorophyll *a* trophic state - 4 class = All variables (lake morphometry, and landscape)
5. Chlorophyll *a* trophic state - 3 class = All variables (lake morphometry, and landscape)
6. Chlorophyll *a* trophic state - 2 class = All variables (lake morphometry, and landscape)

Trophic State (4)	Trophic State (3)	Trophic State (2)	Cut-off
oligo	oligo	oligo/meso	<= 0.2
meso	meso/eu	oligo/meso	>2-7
eu	meso/eu	eu/hyper	>7-30
hyper	hyper	eu/hyper	>30

Results

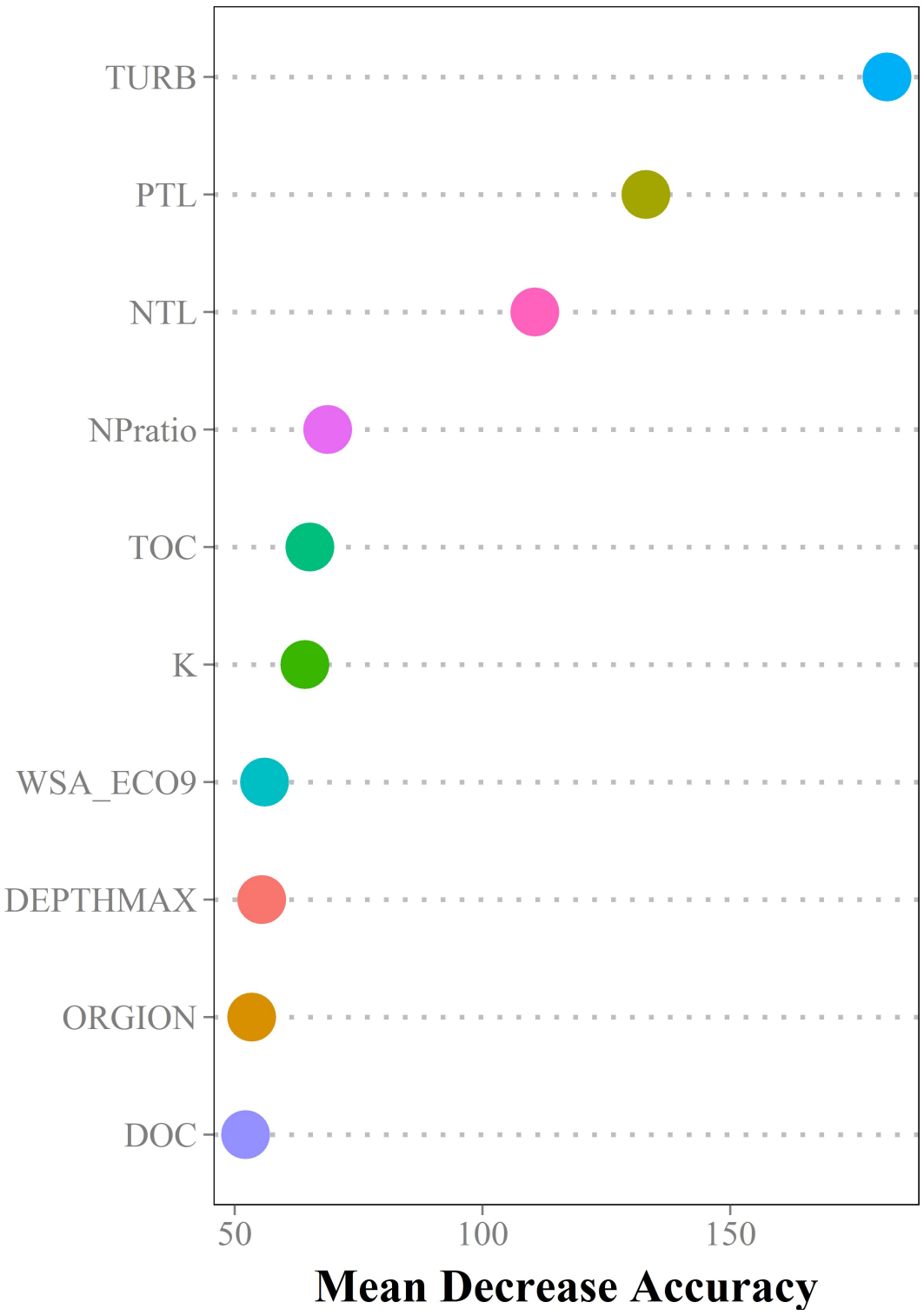
Model 1: 4 Trophic States ~ All Variables

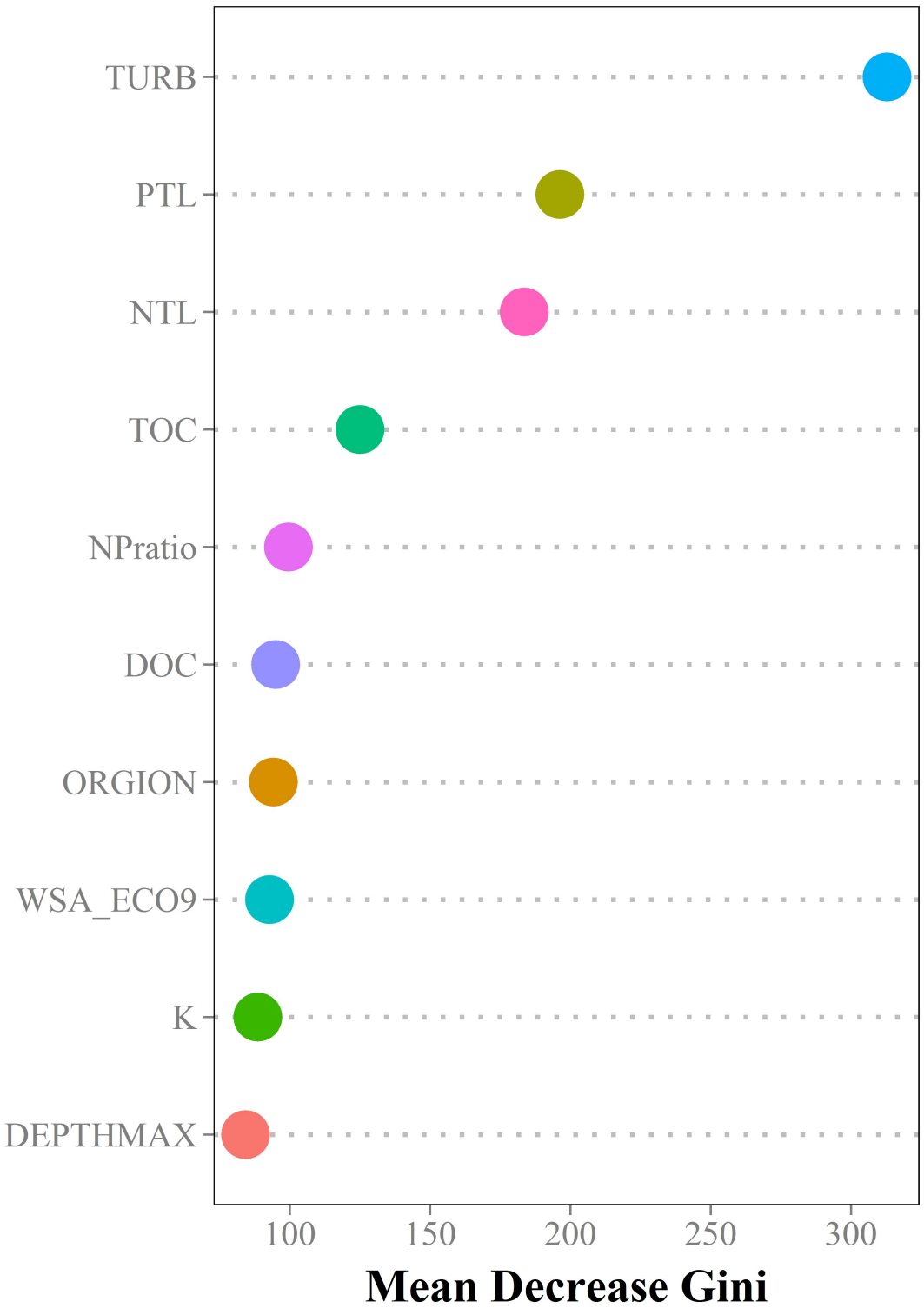
```
## Saving 6 x 8 in image

## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database

## Saving 6 x 8 in image

## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database
```


TURB
PTL
NTL
NPratio
TOC
K
WSA_ECO9
DEPTHMAX
ORGION
DOC
50
100
150
Mean Decrease Accuracy


TURB
PTL
NTL
TOC
NPratio
DOC
ORGION
WSA_ECO9
K
DEPTHMAX
100
150
200
250
300
Mean Decrease Gini

VARIABLE	PERCENT
K	1.00
NPratio	1.00
NTL	1.00
PTL	1.00
TOC	1.00
TURB	1.00
WSA_ECO9	1.00
ORGION	0.29
DOC	0.18
DEPTHMAX	0.03

	Oligo	Meso	Eu	Hyper	class.error
Oligo	135	58	4	1	0.32
Meso	42	235	76	9	0.35
Eu	2	70	217	47	0.35
Hyper	0	3	68	175	0.29

Total accuracy for Model 1 is 0.667% and the Cohen's Kappa is 0.546.

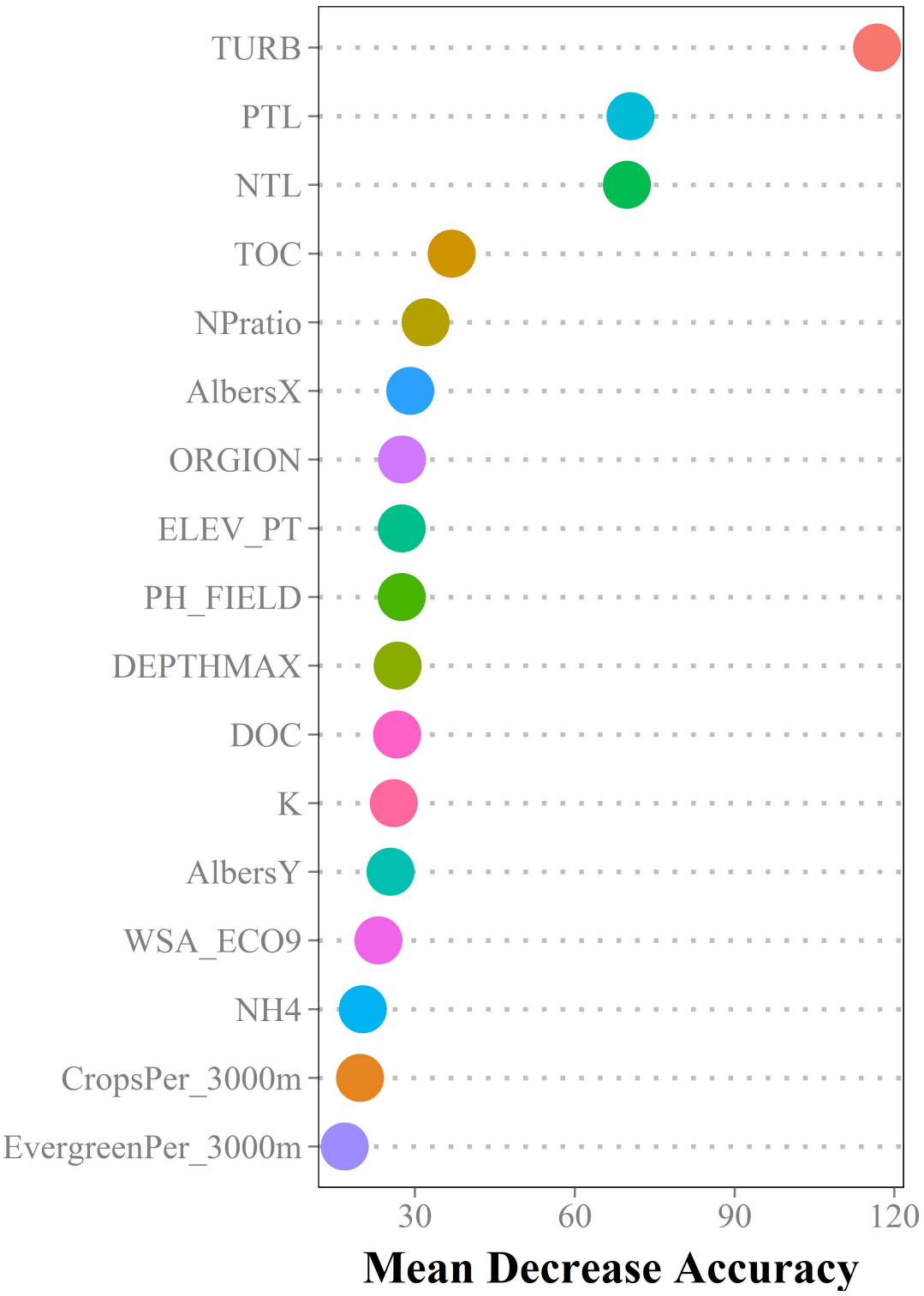
Model 2: 3 Trophic States ~ All Variables

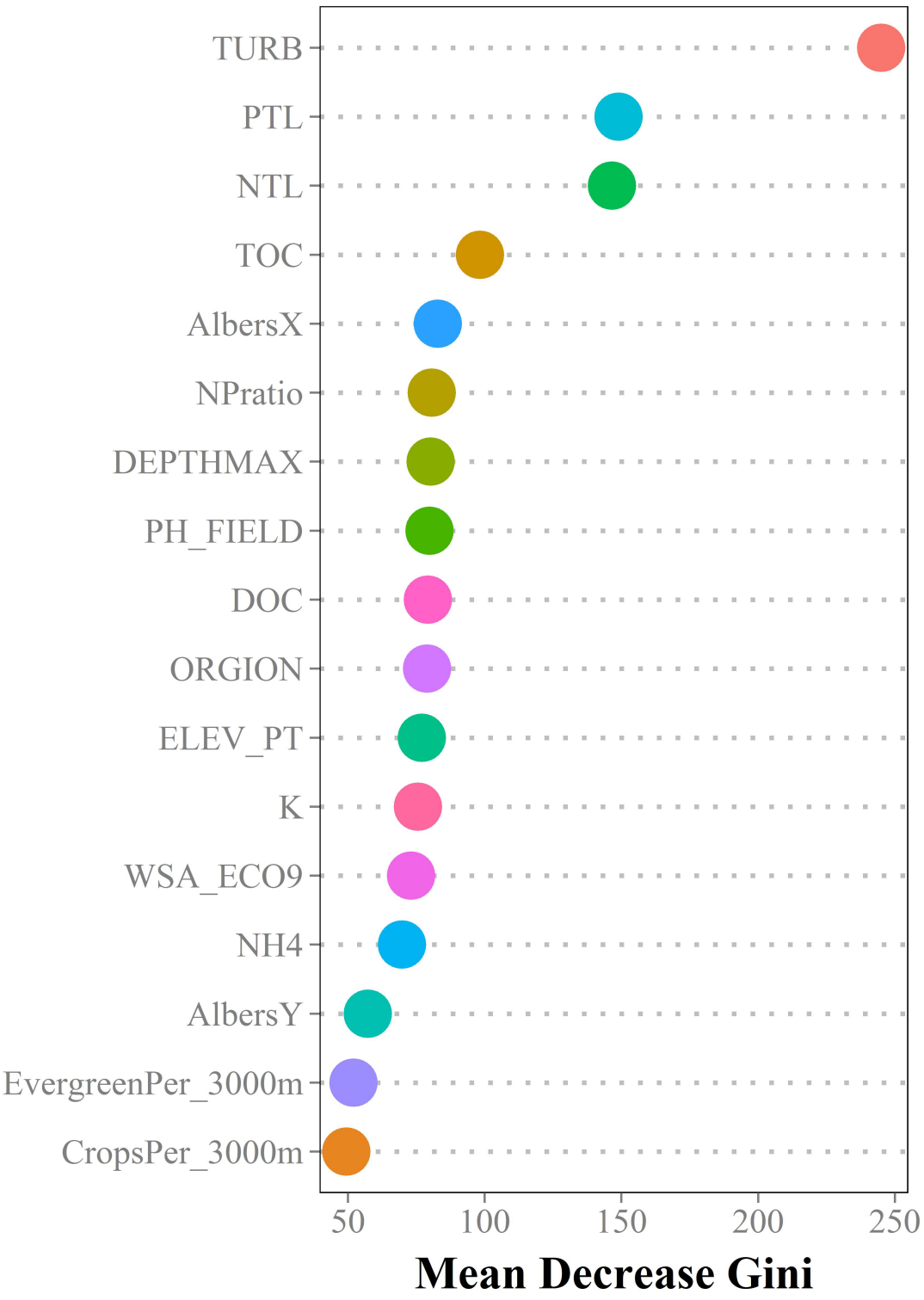
```
## Saving 6 x 8 in image
```

```
## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database
```

```
## Saving 6 x 8 in image
```

```
## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database
```


TURB
PTL
NTL
TOC
NPratio
AlbersX
ORGION
ELEV_PT
PH_FIELD
DEPTHMAX
DOC
K
AlbersY
WSA_ECO9
NH4
CropsPer_3000m
EvergreenPer_3000m
30
60
90
120
Mean Decrease Accuracy


TURB
PTL
NTL
TOC
AlbersX
NPratio
DEPTHMAX
PH_FIELD
DOC
ORGION
ELEV_PT
K
WSA_ECO9
NH4
AlbersY
EvergreenPer_3000m
CropsPer_3000m
50
100
150
200
250
Mean Decrease Gini

VARIABLE	PERCENT
DOC	1.00
K	1.00
NTL	1.00
ORGION	1.00
PTL	1.00
TOC	1.00
TURB	1.00
WSA_ECO9	1.00
DEPTHMAX	0.98
NPratio	0.76
AlbersX	0.48
CropsPer_3000m	0.27
ELEV_PT	0.16
AlbersY	0.05
NH4	0.05
PH_FIELD	0.01
EvergreenPer_3000m	0.01

	Oligo	Meso/Eu	Hyper	class.error
Oligo	121	75	0	0.38
Meso/Eu	40	609	40	0.12
Hyper	0	72	173	0.29

Total accuracy for Model 2 is 0.799% and the Cohen's Kappa is 0.618.

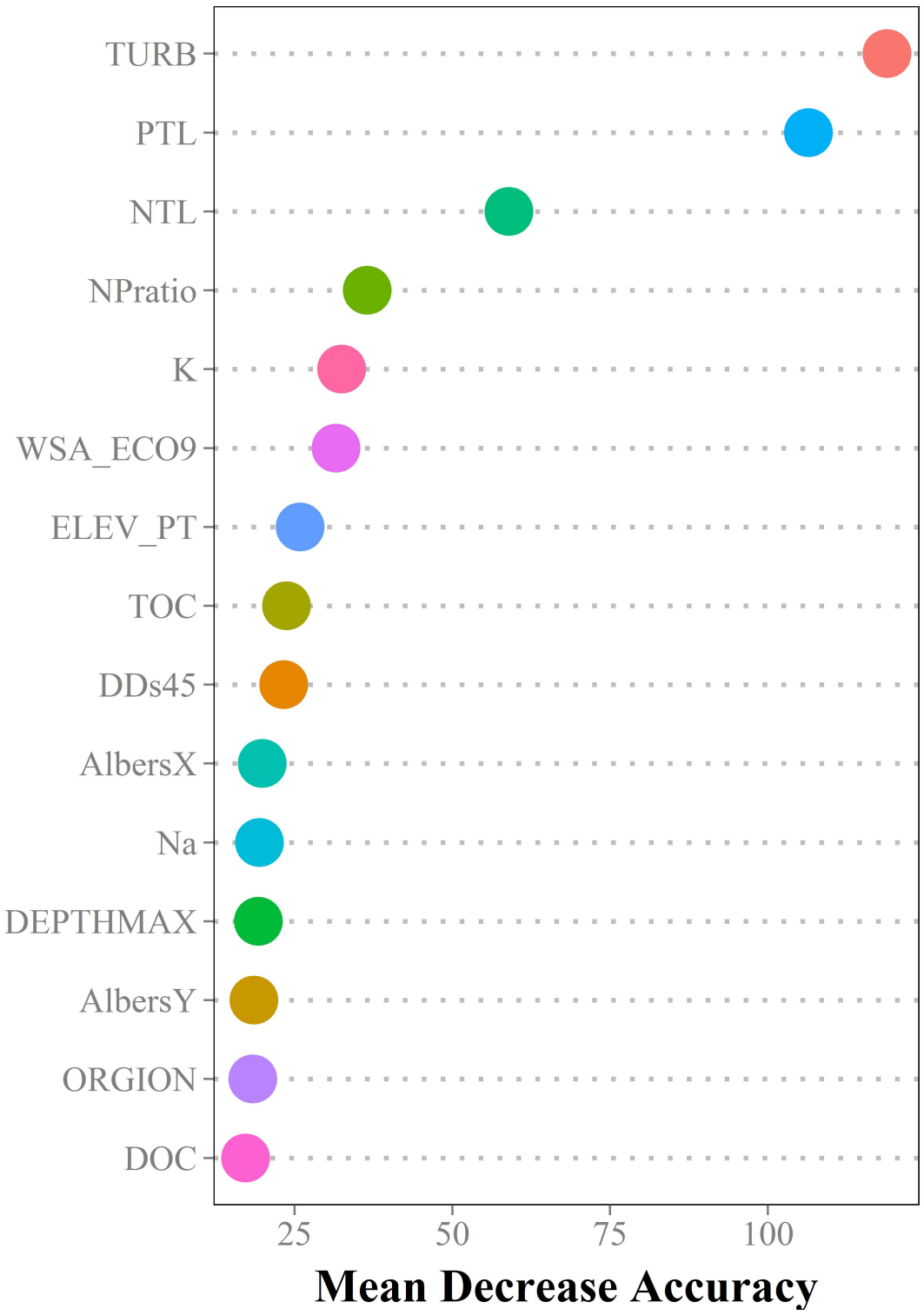
Model 3: 2 Trophic States ~ All Variables

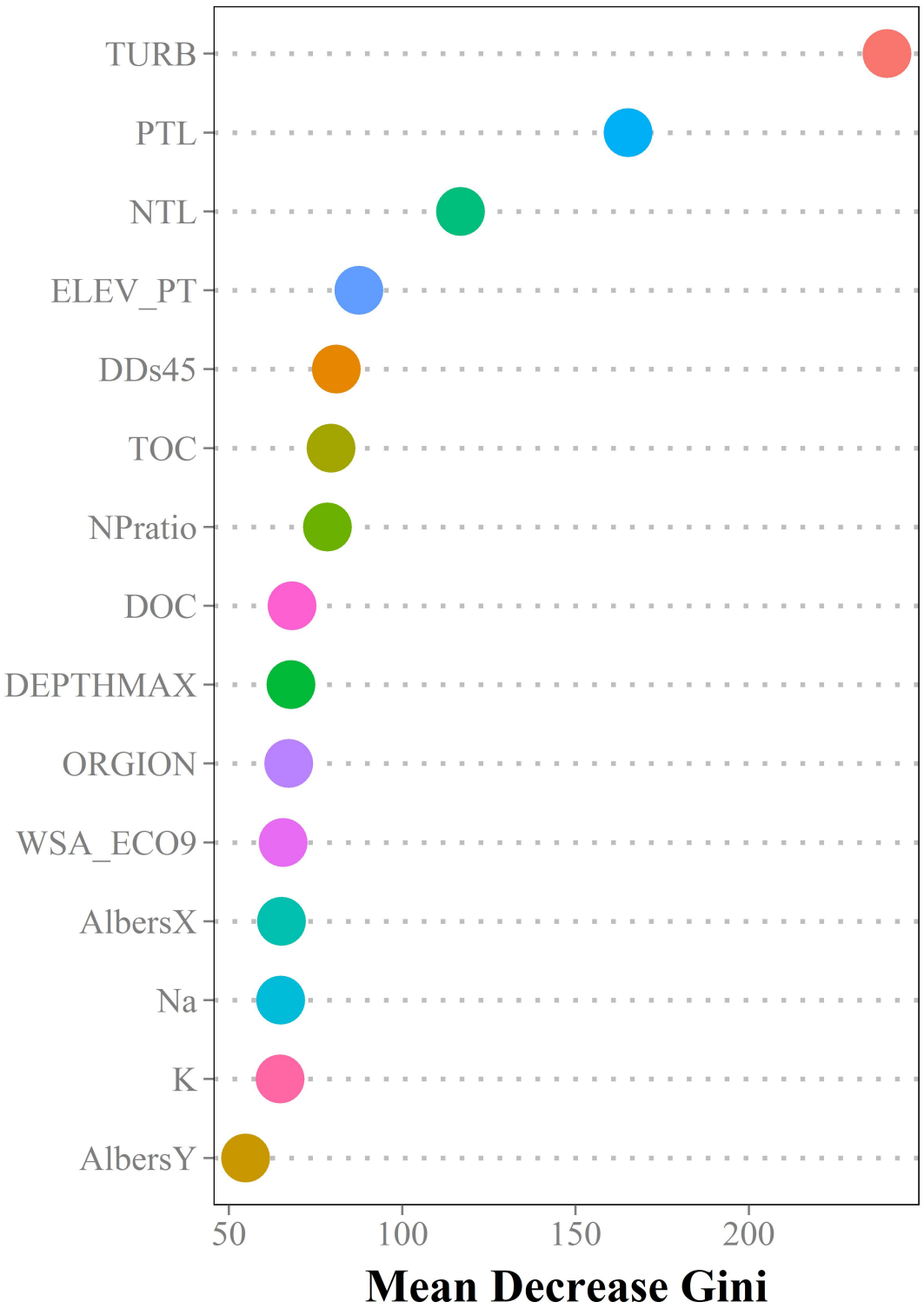
```
## Saving 6 x 8 in image
```

```
## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database
```

```
## Saving 6 x 8 in image
```

```
## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database
```


TURB
PTL
NTL
NPratio
K
WSA_ECO9
ELEV_PT
TOC
DDs45
AlbersX
Na
DEPTHMAX
AlbersY
ORGION
DOC
25
50
75
100
Mean Decrease Accuracy


TURB
PTL
NTL
ELEV_PT
DDs45
TOC
NPratio
DOC
DEPTHMAX
ORGION
WSA_ECO9
AlbersX
Na
K
AlbersY
50
100
150
200
Mean Decrease Gini

VARIABLE	PERCENT
K	1.00
NPratio	1.00
NTL	1.00
PTL	1.00
TOC	1.00
TURB	1.00
WSA_ECO9	1.00
ORGION	0.99
DEPTHMAX	0.96
DDs45	0.90
ELEV_PT	0.85
DOC	0.58
AlbersX	0.06
AlbersY	0.03
Na	0.03

	Oligo/Meso	Eu/Hyper	class.error
Oligo/Meso	489	71	0.13
Eu/Hyper	77	505	0.13

Total accuracy for Model 3 is 0.87% and the Cohen's Kappa is 0.741.

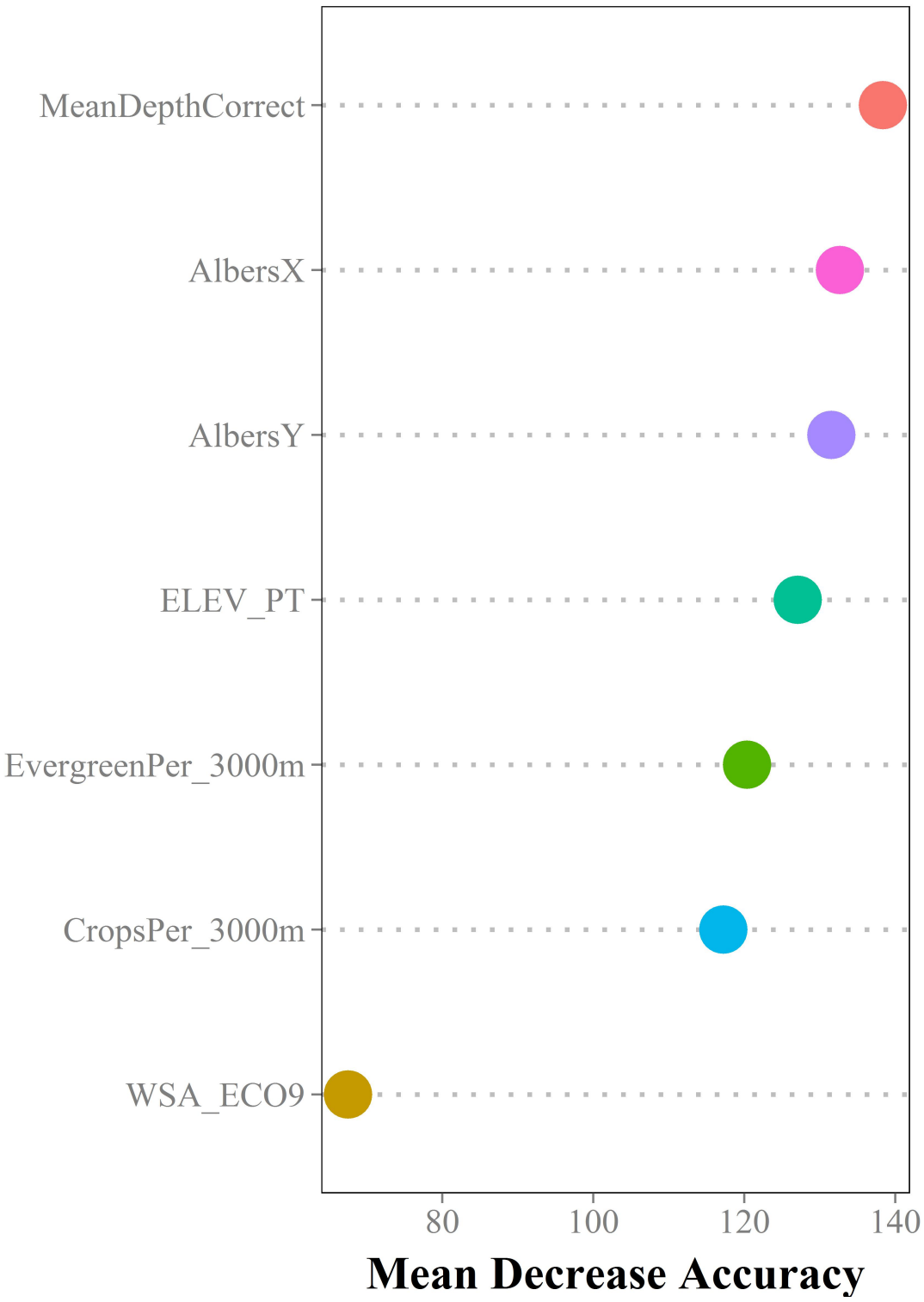
Model 4: 4 Trophic States ~ GIS Only Variables

```
## Saving 6 x 8 in image
```

```
## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database
```

```
## Saving 6 x 8 in image
```

```
## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database
```


MeanDepthCorrect
AlbersX
AlbersY
ELEV_PT
EvergreenPer_3000m
CropsPer_3000m
WSA_ECO9
80
100
120
140
Mean Decrease Accuracy


EvergreenPer_3000m
WSA_ECO9
CropsPer_3000m
AlbersX
AlbersY
ELEV_PT
MeanDepthCorrect
100
120
140
Mean Decrease Gini

VARIABLE	PERCENT
AlbersX	1.00
CropsPer_3000m	1.00
EvergreenPer_3000m	1.00
MeanDepthCorrect	1.00
WSA_ECO9	1.00
AlbersY	0.35
ELEV_PT	0.02

	Oligo	Meso	Eu	Hyper	class.error
Oligo	95	73	27	2	0.52
Meso	48	201	80	32	0.44
Eu	20	114	124	77	0.63
Hyper	2	36	79	129	0.48

Total accuracy for Model 4 is 0.482% and the Cohen's Kappa is 0.292.

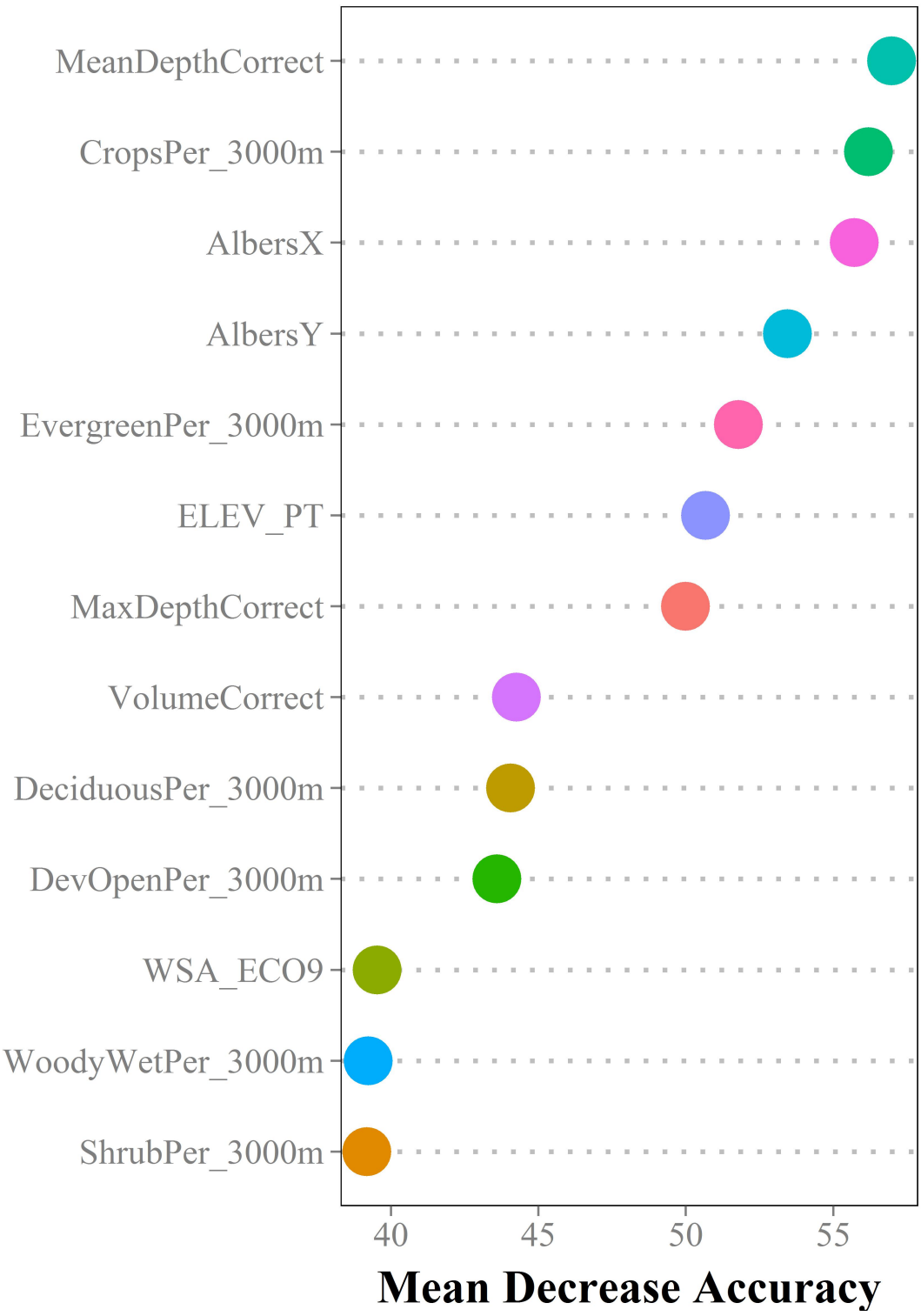
Model 5: 3 Trophic States ~ GIS Only Variables

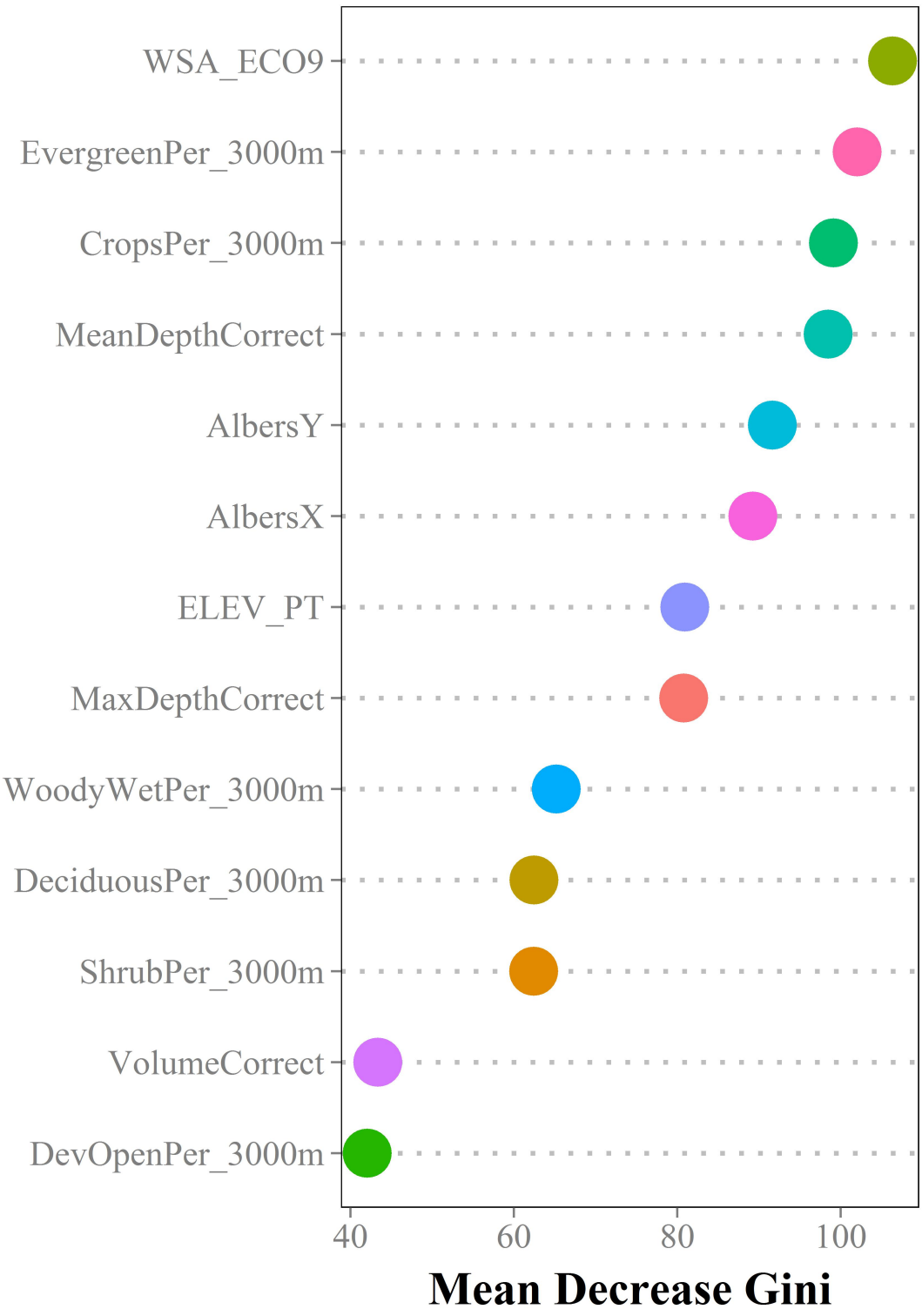
```
## Saving 6 x 8 in image
```

```
## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database
```

```
## Saving 6 x 8 in image
```

```
## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database
```


MeanDepthCorrect
CropsPer_3000m
AlbersX
AlbersY
EvergreenPer_3000m
ELEV_PT
MaxDepthCorrect
VolumeCorrect
DeciduousPer_3000m
DevOpenPer_3000m
WSA_ECO9
WoodyWetPer_3000m
ShrubPer_3000m
40
45
50
55
Mean Decrease Accuracy


WSA_ECO9
EvergreenPer_3000m
CropsPer_3000m
MeanDepthCorrect
AlbersY
AlbersX
ELEV_PT
MaxDepthCorrect
WoodyWetPer_3000m
DeciduousPer_3000m
ShrubPer_3000m
VolumeCorrect
DevOpenPer_3000m
40
60
80
100
Mean Decrease Gini

VARIABLE	PERCENT
AlbersX	1.00
AlbersY	1.00
CropsPer_3000m	1.00
EvergreenPer_3000m	1.00
MaxDepthCorrect	1.00
MeanDepthCorrect	1.00
WSA_ECO9	1.00
ELEV_PT	0.97
DeciduousPer_3000m	0.94
ShrubPer_3000m	0.21
WoodyWetPer_3000m	0.11
DevOpenPer_3000m	0.10
VolumeCorrect	0.04

	Oligo	Meso/Eu	Hyper	class.error
Oligo	79	116	1	0.6
Meso/Eu	48	582	66	0.16
Hyper	0	141	105	0.57

Total accuracy for Model 5 is 0.673% and the Cohen's Kappa is 0.343.

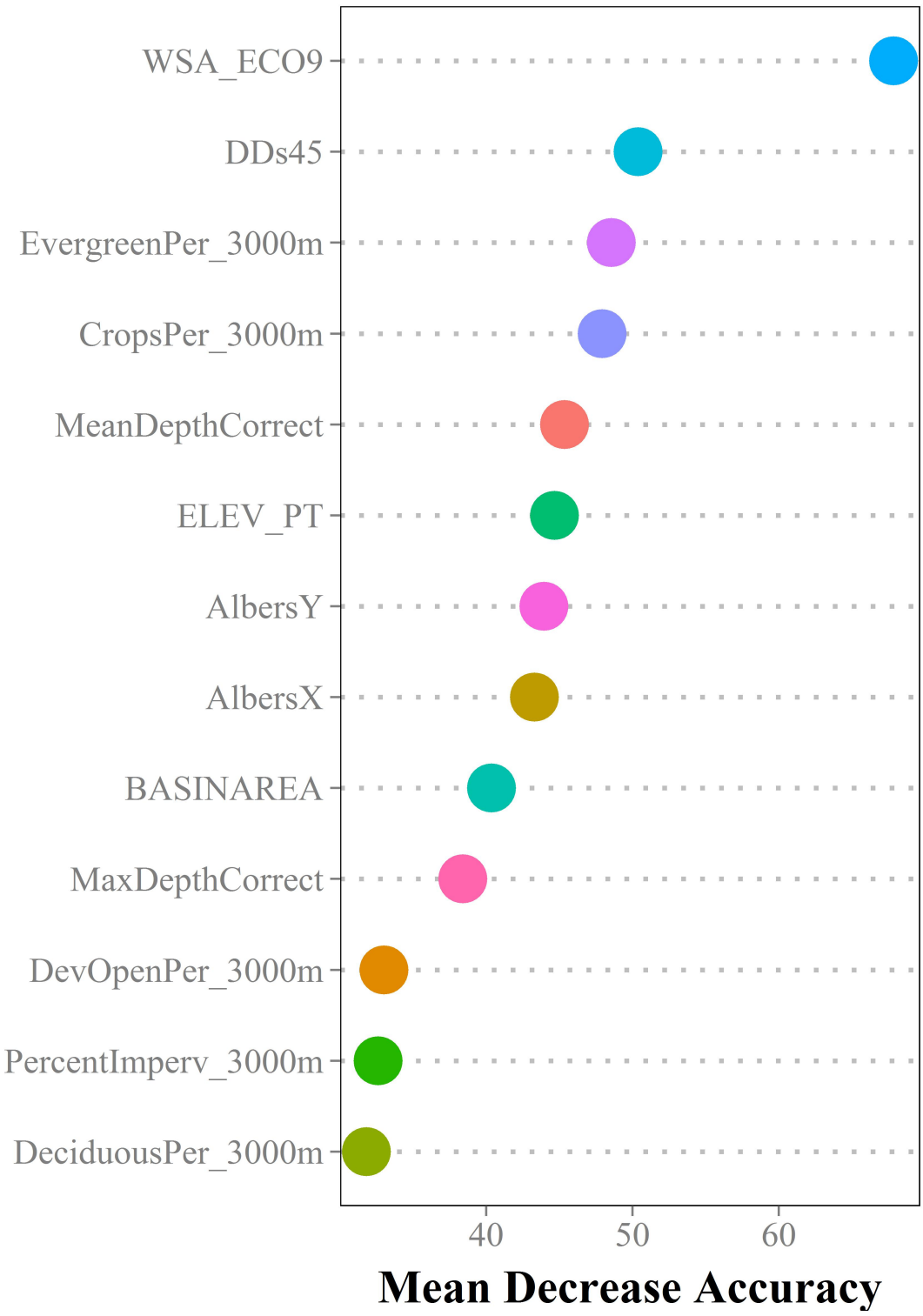
Model 6: 2 Trophic States ~ GIS Only Variables

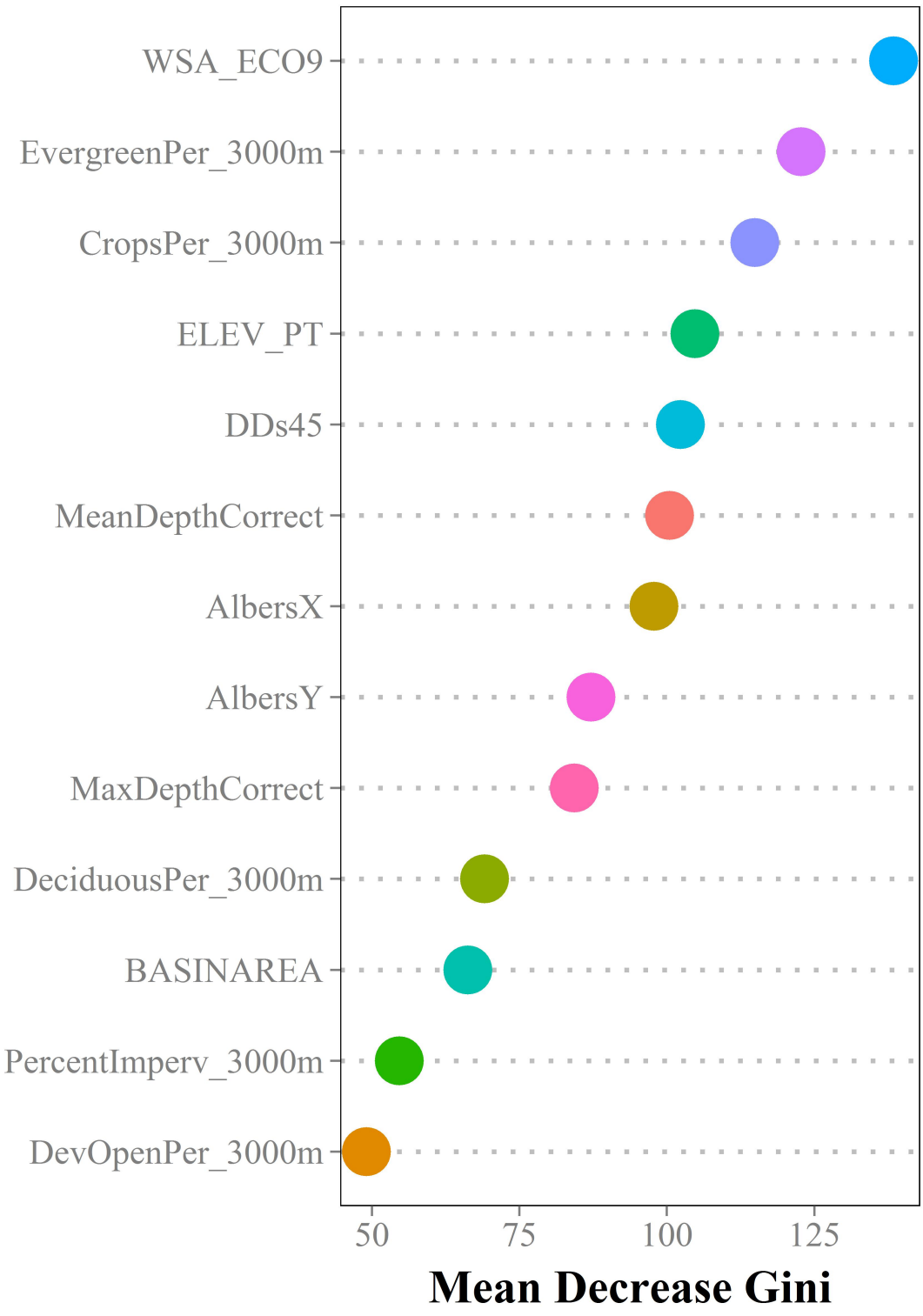
```
## Saving 6 x 8 in image

## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database

## Saving 6 x 8 in image

## Warning: font family not found in Windows font database
## Warning: font family not found in Windows font database
```


WSA_ECO9
DDs45
EvergreenPer_3000m
CropsPer_3000m
MeanDepthCorrect
ELEV_PT
AlbersY
AlbersX
BASINAREA
MaxDepthCorrect
DevOpenPer_3000m
PercentImperv_3000m
DeciduousPer_3000m
40
50
60
Mean Decrease Accuracy


WSA_ECO9
EvergreenPer_3000m
CropsPer_3000m
ELEV_PT
DDs45
MeanDepthCorrect
AlbersX
AlbersY
MaxDepthCorrect
DeciduousPer_3000m
BASINAREA
PercentImperv_3000m
DevOpenPer_3000m
50
75
100
125
Mean Decrease Gini

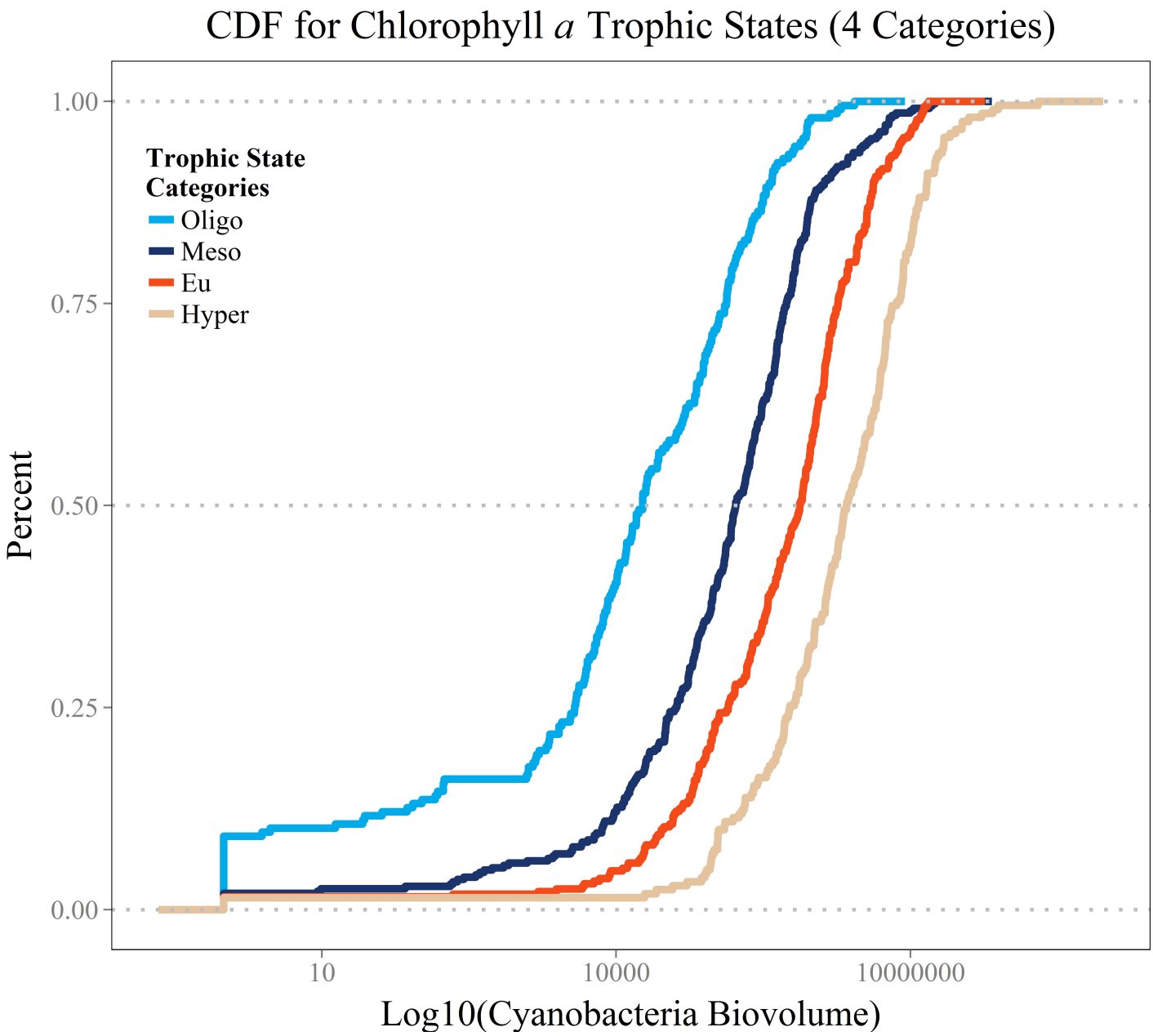
VARIABLE	PERCENT
AlbersX	1.00
CropsPer_3000m	1.00
DDs45	1.00
ELEV_PT	1.00
EvergreenPer_3000m	1.00
MeanDepthCorrect	1.00
WSA_ECO9	1.00
AlbersY	0.98
MaxDepthCorrect	0.98
DeciduousPer_3000m	0.92
DevOpenPer_3000m	0.67
BASINAREA	0.31
PercentImperv_3000m	0.01

	Oligo/Meso	Eu/Hyper	class.error
Oligo/Meso	428	129	0.23
Eu/Hyper	146	435	0.25

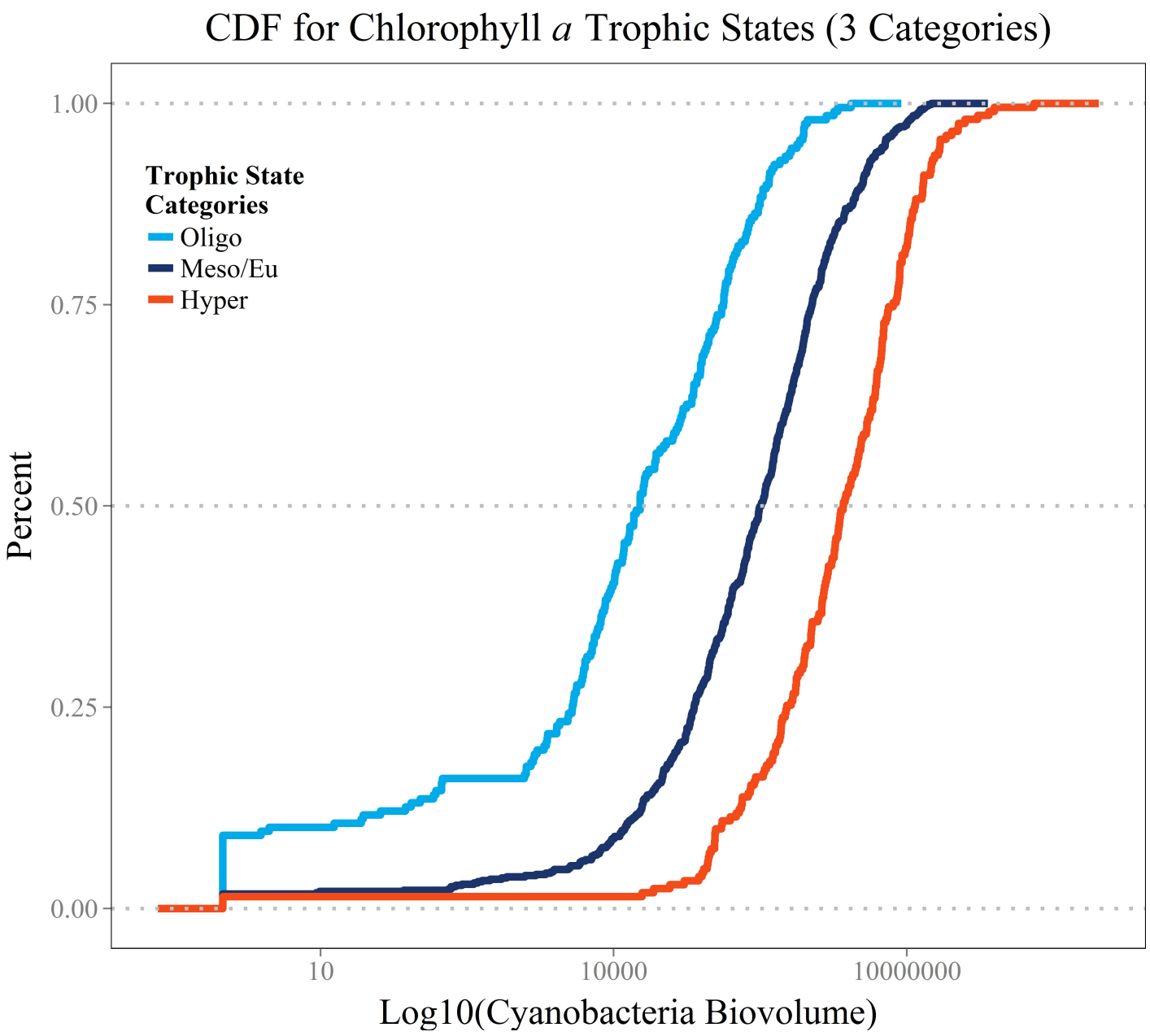
Total accuracy for Model 6 0.758% and the Cohen's Kappa is 0.517.

Associating Trophic State and Cyanobacteria

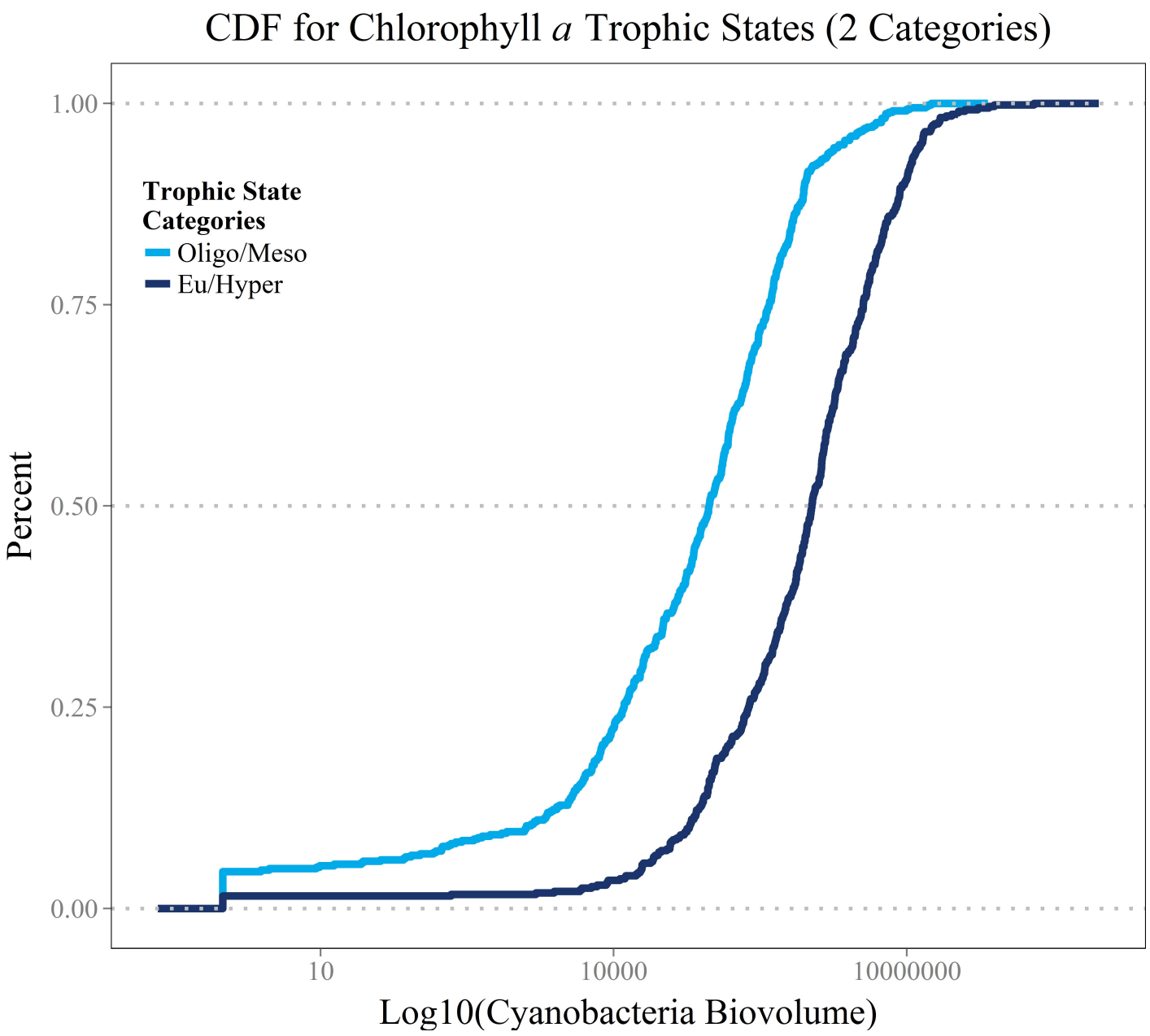
```
## Error: object 'hkm2014Data' not found
```



```
## Error: object 'hkm2014Data' not found
```



```
## Error: object 'hkm2014Data' not found
```



```
## Error: object 'hkm2014Data' not found
```

```
## Error: object 'scp_df' not found
```

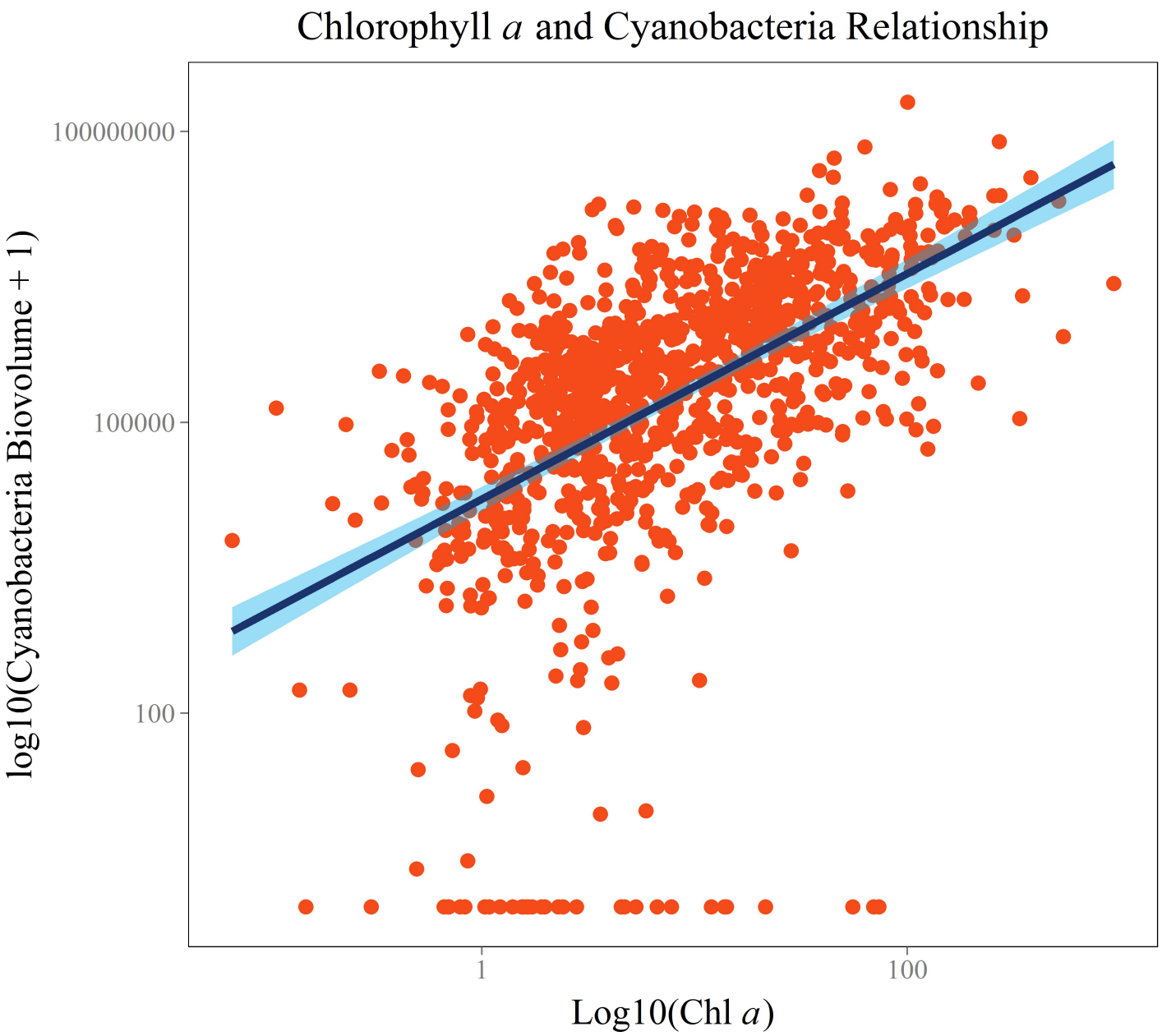


Figure 1: Chla/BioV Scatterplot

Poster Source on GitHub

All of the The materials that make up this poster are available via GitHub. Included in this repository are an R Markdown document, and R Package with data, and the final poster layout as .svg or .pdf. Please use the QR Code to access this repository.

References

Beaulieu, Marieke, Frances Pick, and Irene Gregory-Eaves. 2013. "Nutrients and Water Temperature Are Significant Predictors of Cyanobacterial Biomass in a 1147 Lakes Data Set." *Limnol. Oceanogr* 58 (5): 1736–1746.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Diaz-Uriarte, Ramon. 2010. *VarSelRF: Variable Selection Using Random Forests*. http://CRAN.R-project.org/package=varSelRF.

Díaz-Uriarte, Ramón, and Sara Alvarez De Andres. 2006. "Gene Selection and Classification of Microarray Data Using Random Forest." *BMC Bioinformatics* 7 (1): 3.

Hollister, Jeffrey. 2013. *Lakemorpho: Lake Morphometry in R*. http://www.github.com/USEPA/lakemorpho.

Hollister, Jeffrey W, and W Bryan Milstead. "National Lake Morphometry Dataset V1.0."

Hollister, Jeffrey W., W. Bryan Milstead, and M. Andrea Urrutia. 2011. "Predicting Maximum Lake Depth from Surrounding Topography." *PLoS ONE* 6 (9) (September): e25764. doi:10.1371/journal.pone.0025764. http://dx.doi.org/10.1371/journal.pone.0025764.

Hollister, Jeffrey, and W Bryan Milstead. 2010. "Using GIS to Estimate Lake Volume from Limited Data." *Lake and Reservoir Management* 26 (3): 194–199.

Homer, Collin, Chengquan Huang, Limin Yang, Bruce Wylie, and Michael Coan. 2004. "Development of a 2001 National Land-cover Database for the United States." *Photogrammetric Engineering & Remote Sensing* 70 (7): 829–840.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by RandomForest." *R News* 2 (3): 18–22. http://CRAN.R-project.org/doc/Rnews/.

USEPA. 2009. "National Lakes Assessment: a Collaborative Survey of the Nation's Lakes. EPA 841-r-09-001." Office of Water; Office of Research; Development, US Environmental Protection Agency Washington, DC.

Xian, George, Collin Homer, and Joyce Fry. 2009. "Updating the 2001 National Land Cover Database Land Cover Classification to 2006 by Using Landsat Imagery Change Detection Methods." *Remote Sensing of Environment* 113 (6): 1133–1147.