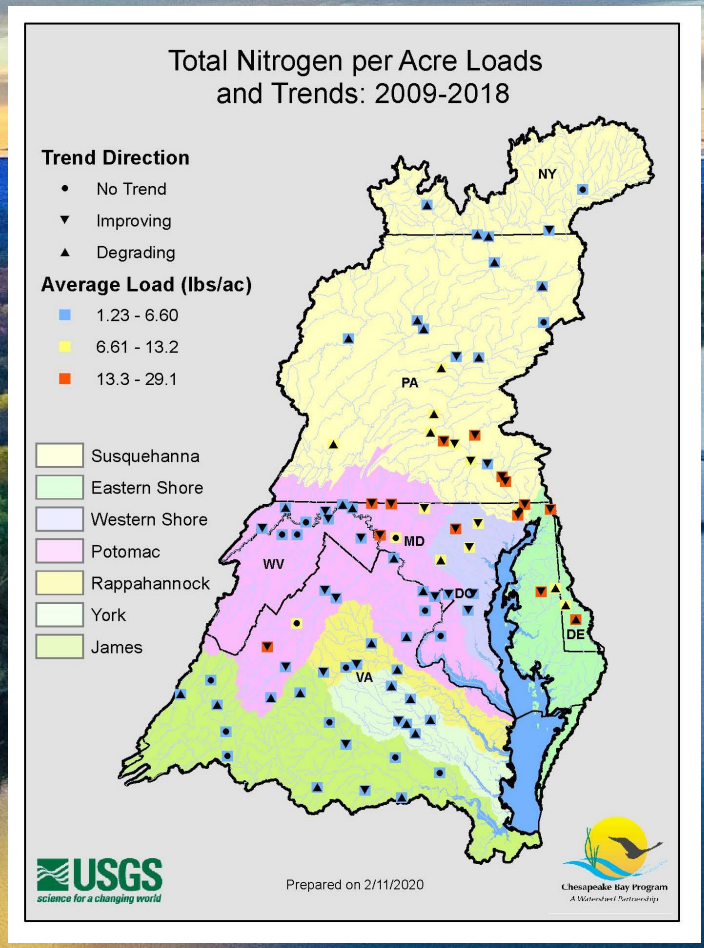
# Regional patterns and drivers of total nitrogen trends in the Chesapeake Bay watershed: Insights from machine learning approaches and management implications


Qian Zhang[1], Joel Bostic[2], Robert Sabo[3]

[1] University of Maryland Center for Environmental Science / USEPA Chesapeake Bay Program

[2] University of Maryland Center for Environmental Science, Appalachian Laboratory

[3] U.S. Environmental Protection Agency


Any opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Environmental Protection Agency.

# Motivations

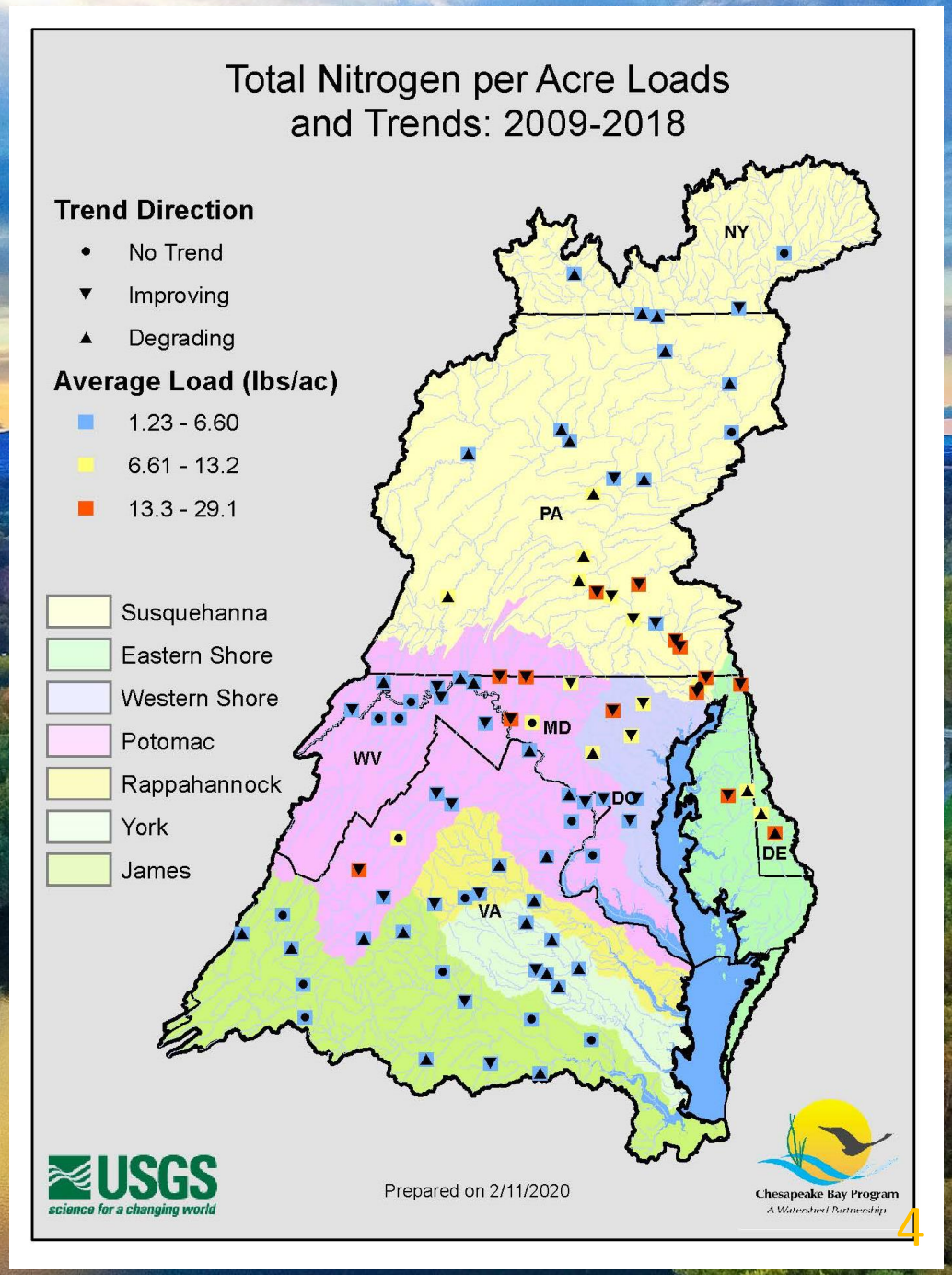
- River water-quality (WQ) trend studies often focus on one or a few monitoring locations, making conclusions difficult to generalize.
- Much can be learned from the similarity in WQ signals and the similarity in WQ responses to natural and anthropogenic drivers, which is made possible by data from regional monitoring networks.
- While many studies are aimed at the long-term scale (~30 years), short-term analysis can leverage data from newly established stations and provide relatively current information.
- Monitoring networks (i.e., CBNTN) do not often cover the entire watershed, leading to missing information in certain regions.
- Prior analyses of drivers do not always evaluate all major input sources, leading to potentially inaccurate or even contradicting inferences.

# Objective

To reveal regional patterns and drivers of total nitrogen (TN) trends using machine learning approaches -- combined use of hierarchical clustering and random forest (RF).

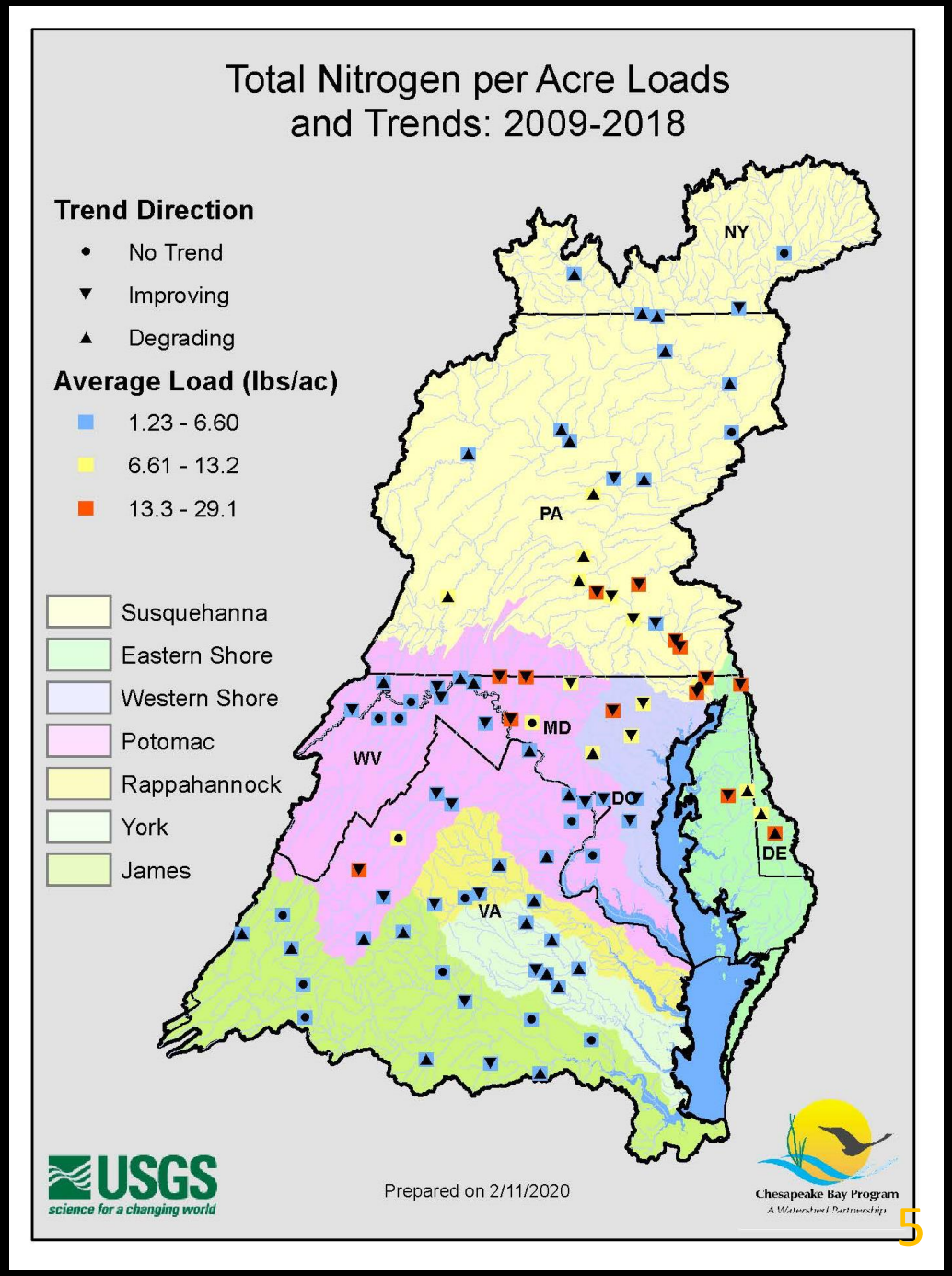
1. **Clustering**: Categorize the short-term (2007-2018) TN trends at the Chesapeake NTN stations (84) into distinct clusters,
2. **Classification**: Develop random forest (RF) models to identify the most influential drivers for the cluster assignment, and
3. **Prediction**: Use the RF model to predict short-term trend clusters for the entire watershed at a fine spatial resolution.

# 1. Regional patterns of nitrogen trends in the Bay watershed (Clustering)



# CBNTN stations and TN data

- CBNTN watersheds (n = 84)
- 2007-2018 TN flow-normalized (FN) loads
- Standardized for each station (mean = 0, sd = 1)

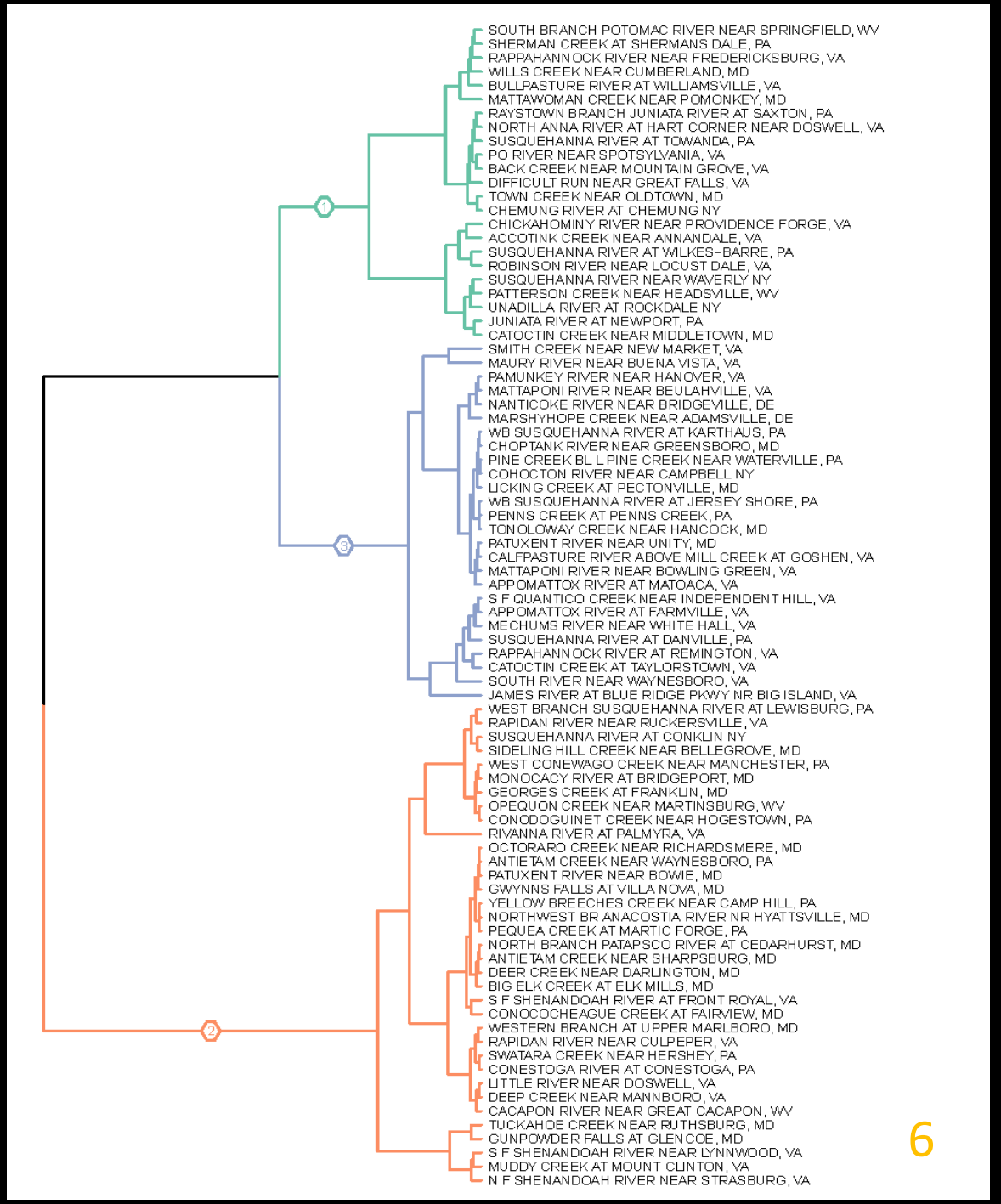


# Hierarchical cluster analysis

Dissimilarity method:
Euclidean distance

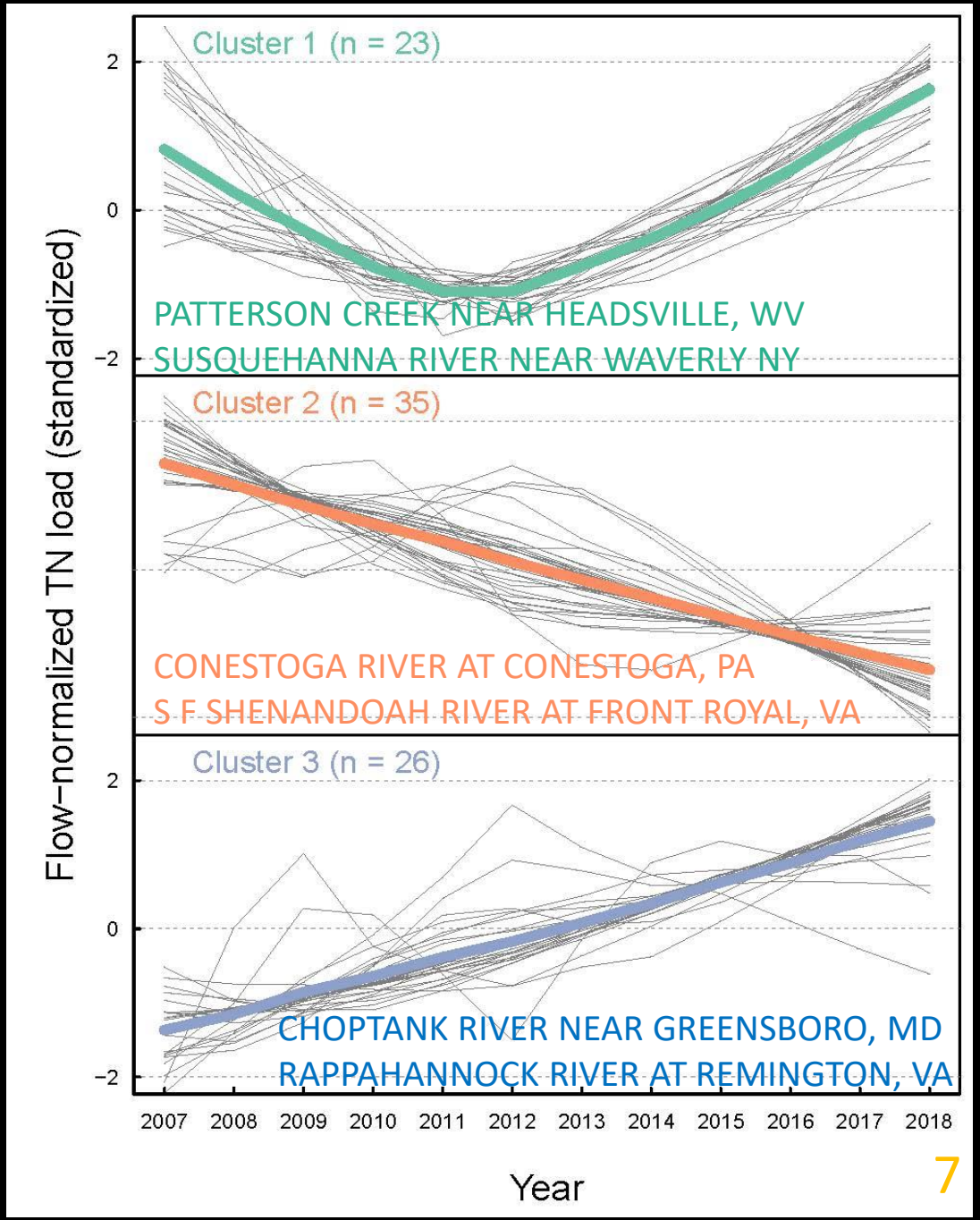
Linkage method:
Ward's minimum variance method

Optimal cluster number:
Total Within Sum of Square



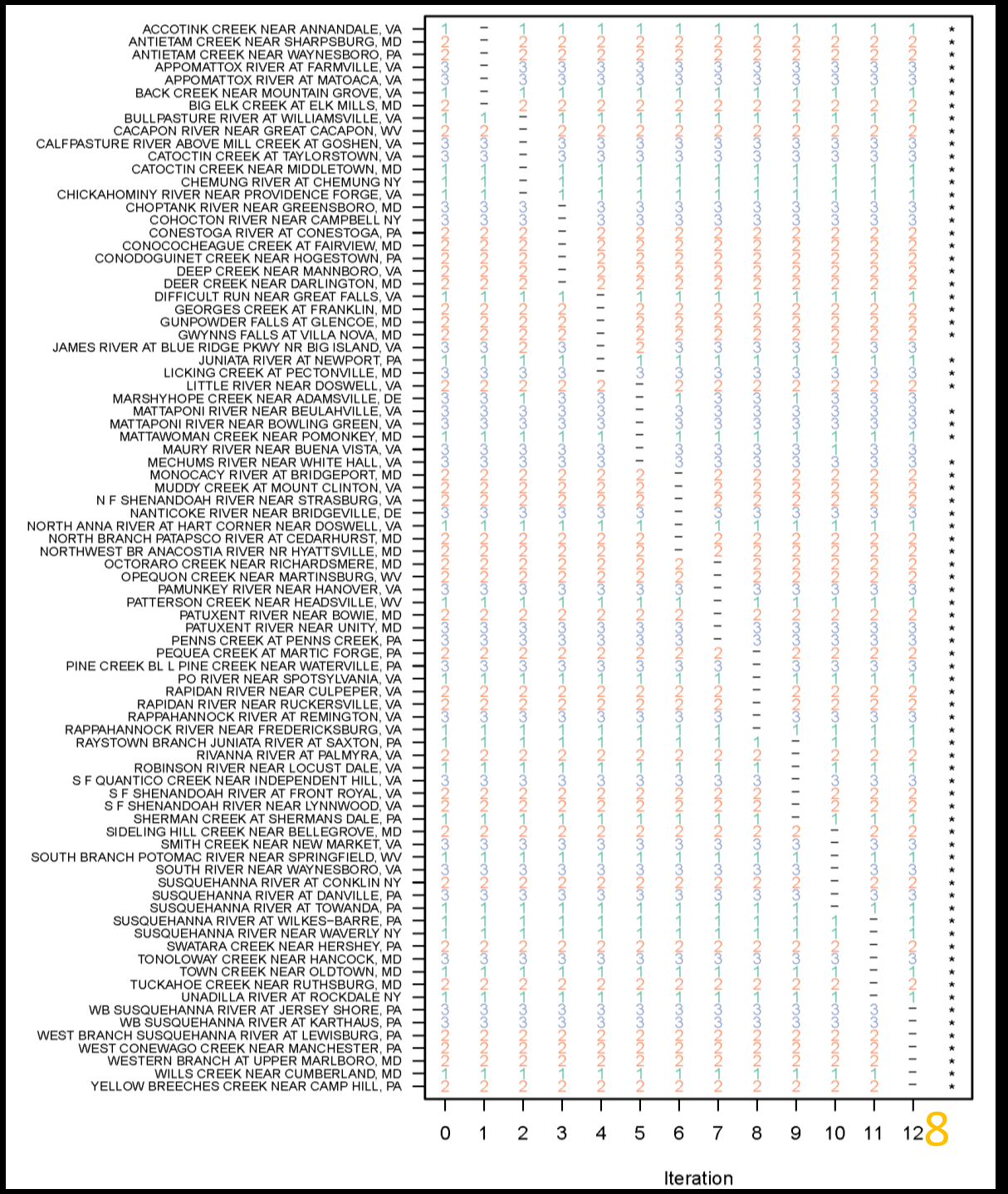
# Hierarchical cluster analysis

- Cluster 1 (n = 23):
  *a V-shape trajectory.*
- Cluster 2 (n = 35):
  *a monotonic decline.*
- Cluster 3 (n = 26):
  *a monotonic increase.*

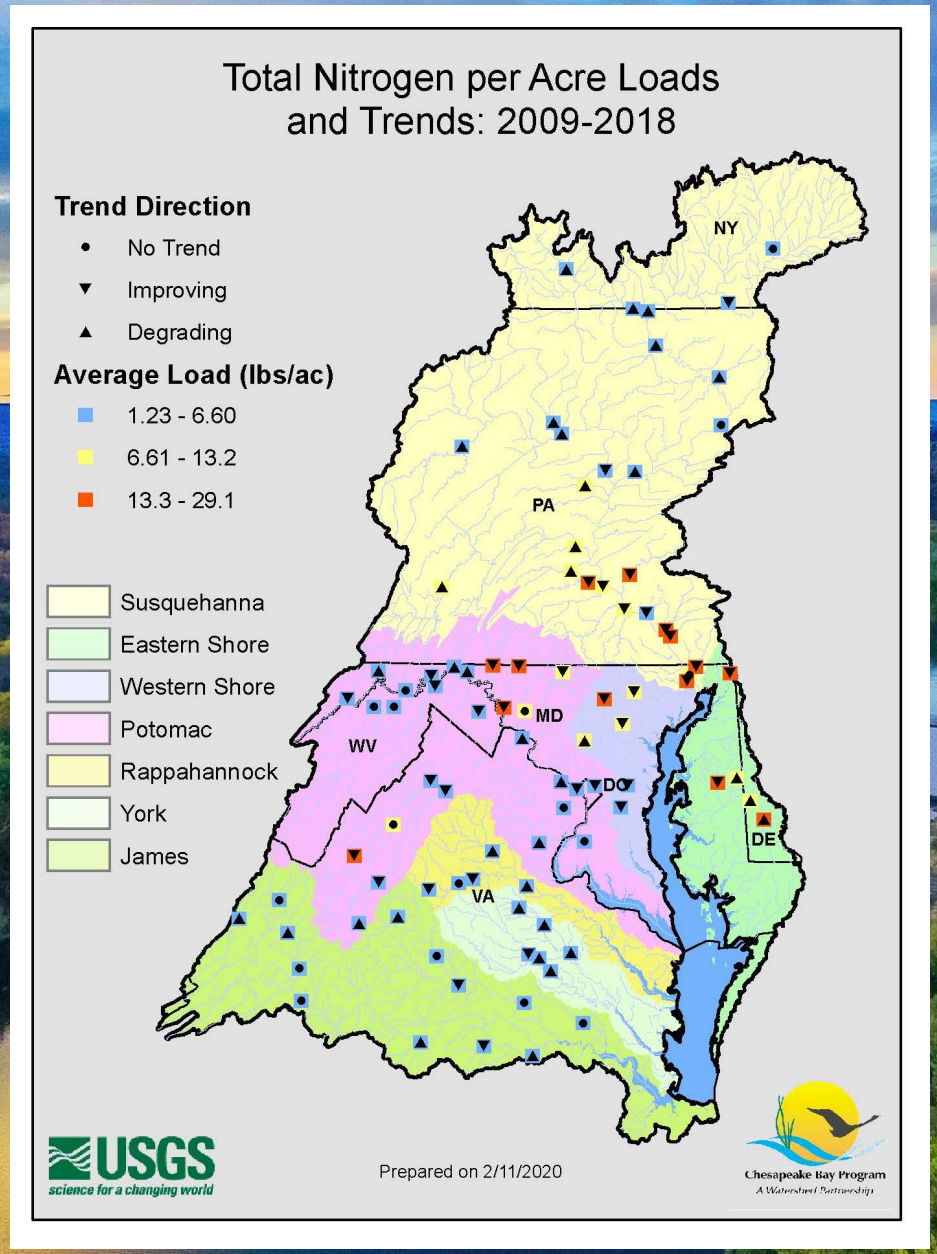


# Sensitivity Analysis

- 1/12 of the stations (n = 7) were removed without replacement.
- The remaining stations (n = 77) were reanalyzed using the same procedure.
- The number of clusters was set at three to be consistent.
- Cluster assignments are almost always consistent among the iterations.



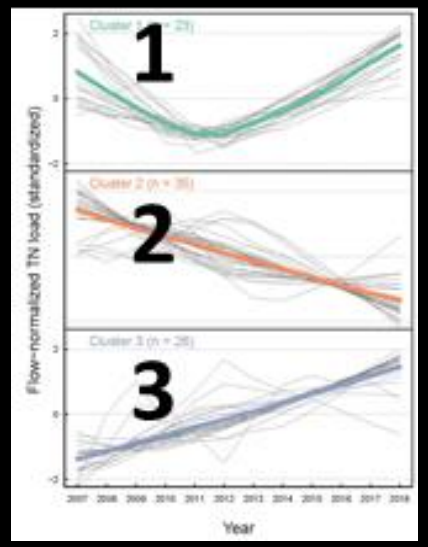
# 2. Regional drivers of nitrogen trend clusters (Classification)

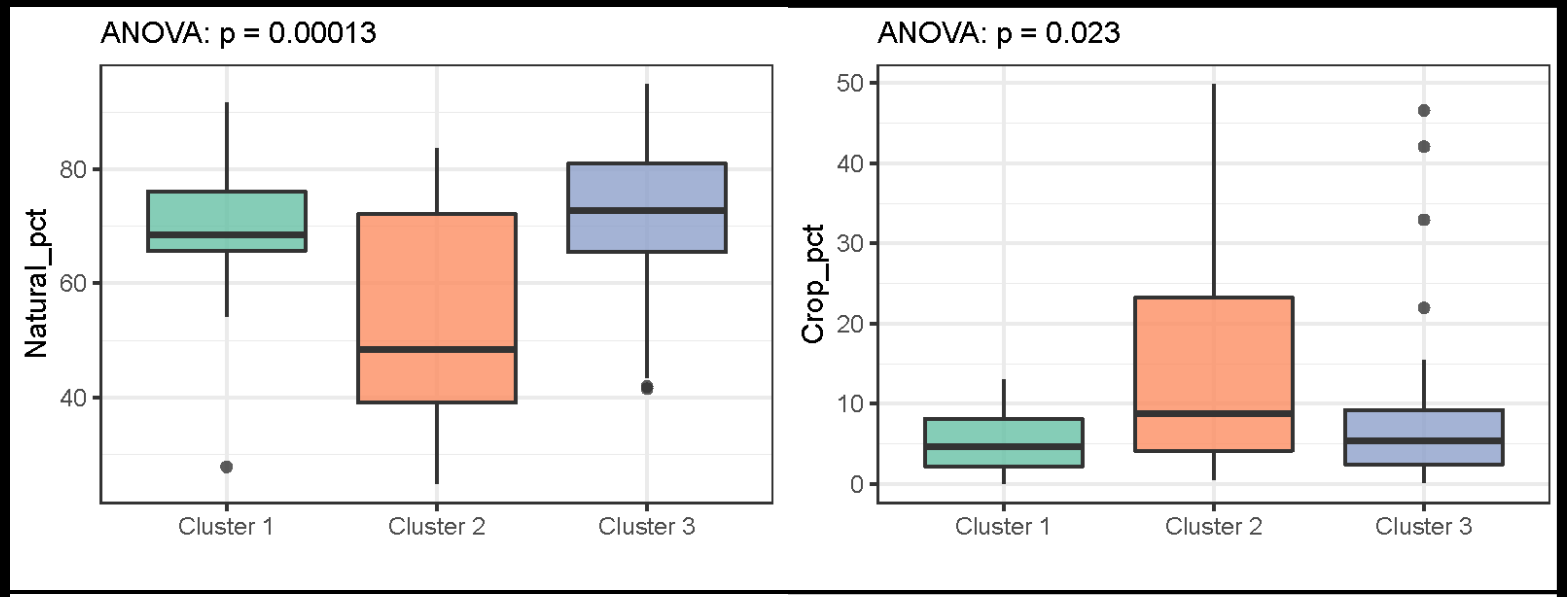


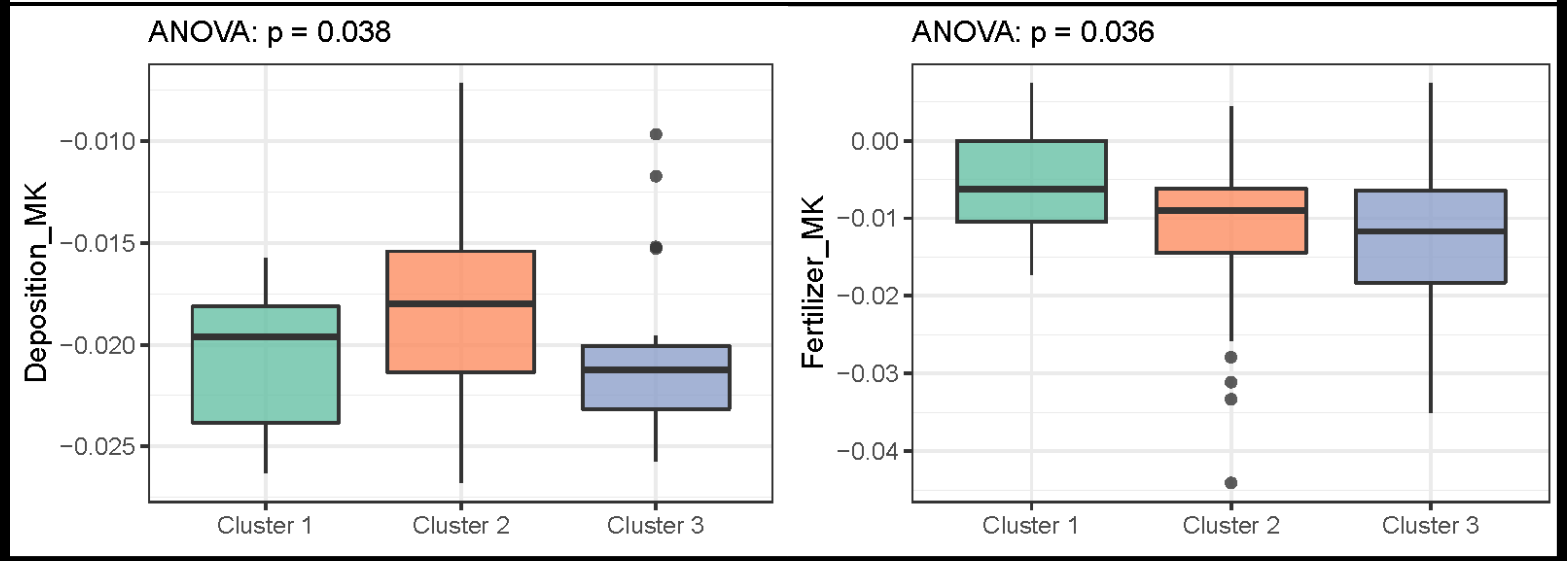
# Explanatory Variables (Features)

- Watershed size (n = 1) - Area_km2
- Land uses, in % (n = 4) - Natural_pct, Crop_pct, Pasture_pct, Hay_pct
- Geology, in % (n = 1) - Carb_pct
- Physiography, in % (n = 5) - Appalachian_pct, BlueRidge_pct, ValleyRidge_pct, Piedmont_pct, Coastal_pct
- N input source trends (n = 6) - PointSource_MK, Deposition_MK, Fertilizer_MK, Manure_MK, AgInput_MK, AgSurplus_MK
    1. CAST data aggregated for each NTN watershed – 2007-2018 for point sources; 1997-2018 for nonpoint sources.
    2. Annual time series scaled by respective period-of-record medians.
    3. Mann-Kendall trend and Sen's slopes computed.

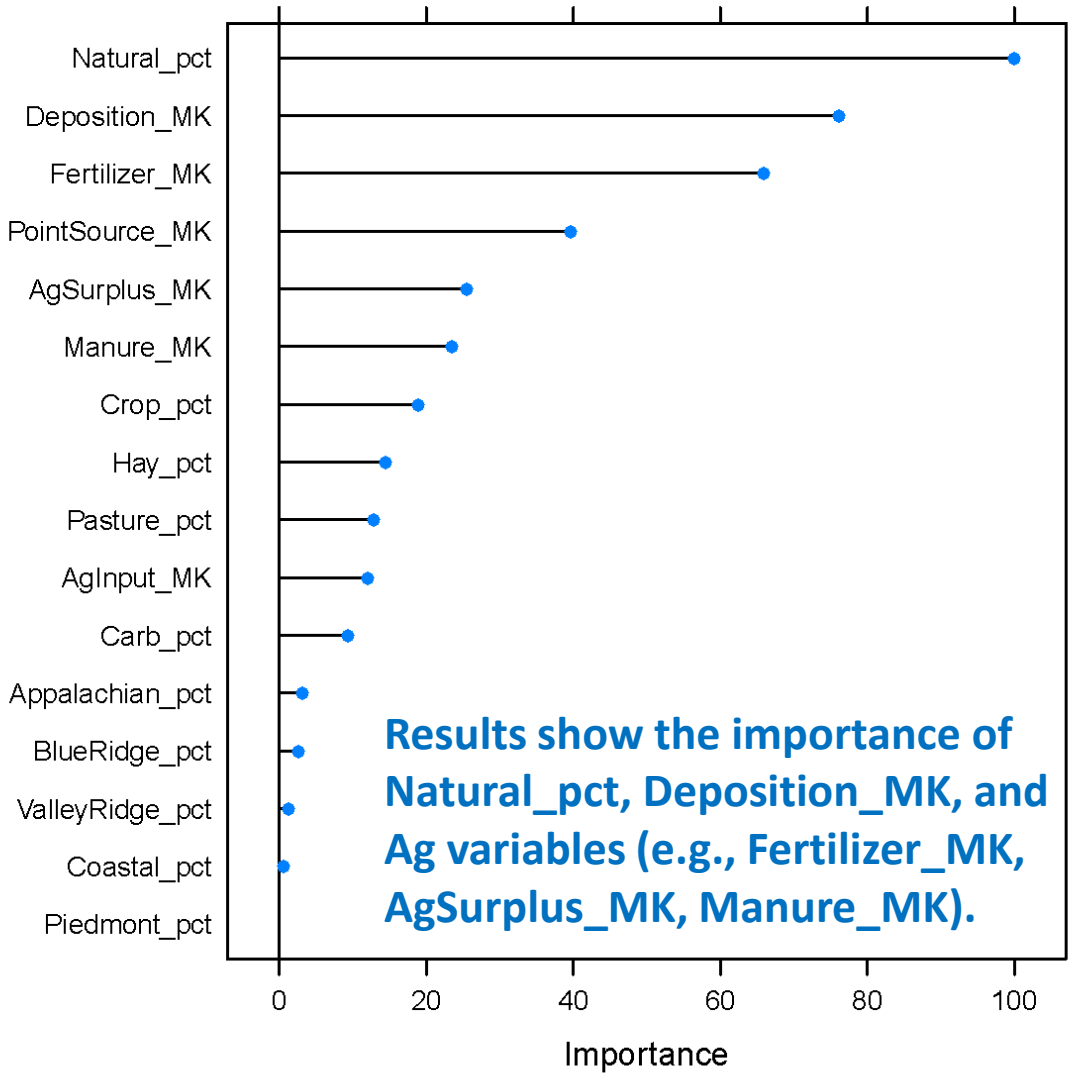
# Explanatory Variables (Features)

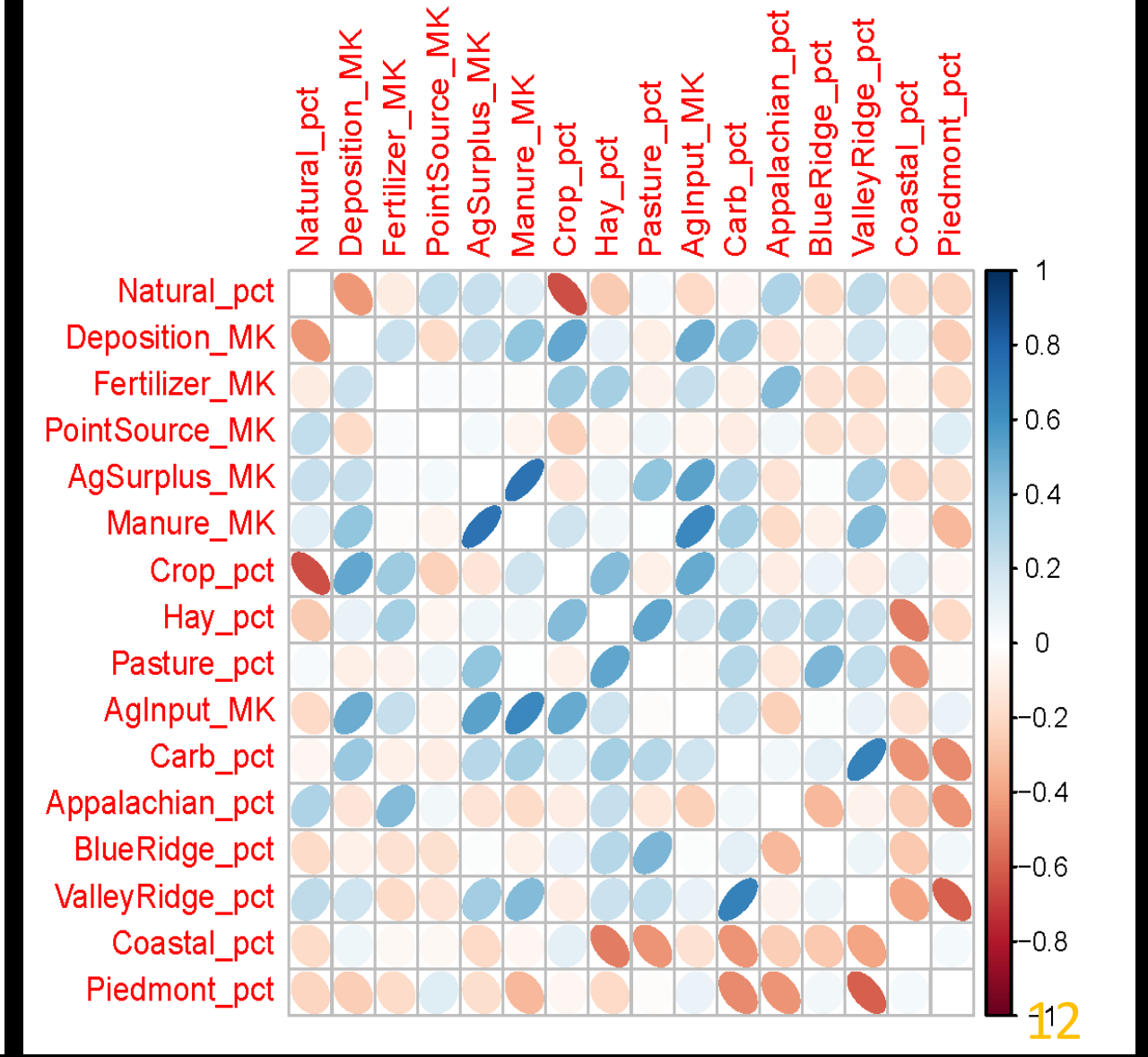


# Random Forest (Base Model)



# Exhaustive Search for Optimal Models (n ≤ 6)

| Model | Model form | OOB accuracy, percent | | | |
|---|---|---|---|---|---|
| | | *Overall* | *Cluster1* | *Cluster2* | *Cluster3* |
| **A** | Class ~ Natural_pct + Fertilizer_MK + ValleyRidge_pct + Deposition_MK + Carb_pct | 70.5 | 66.7 | 68.8 | **76.0** |
| **B** | Class ~ AgSurplus_MK + Fertilizer_MK + Deposition_MK + Natural_pct | 70.5 | 66.7 | **75.0** | 68.0 |
| **C** | Class ~ BlueRidge_pct + Deposition_MK + Coastal_pct + Crop_pct + Fertilizer_MK + Natural_pct | 69.2 | **81.0** | 65.6 | 64.0 |

The selected models have varying accuracies for each cluster, indicating that each model settled on a specific set of features that are most useful to explain a specific cluster. To make predictions, an ensemble model approach was adopted to combine the strengths of these three models – i.e., choosing the prediction with the highest probability from the three models.

# Regional Drivers

## Message 1 (AgSurplus_MK, Fertilizer_MK):

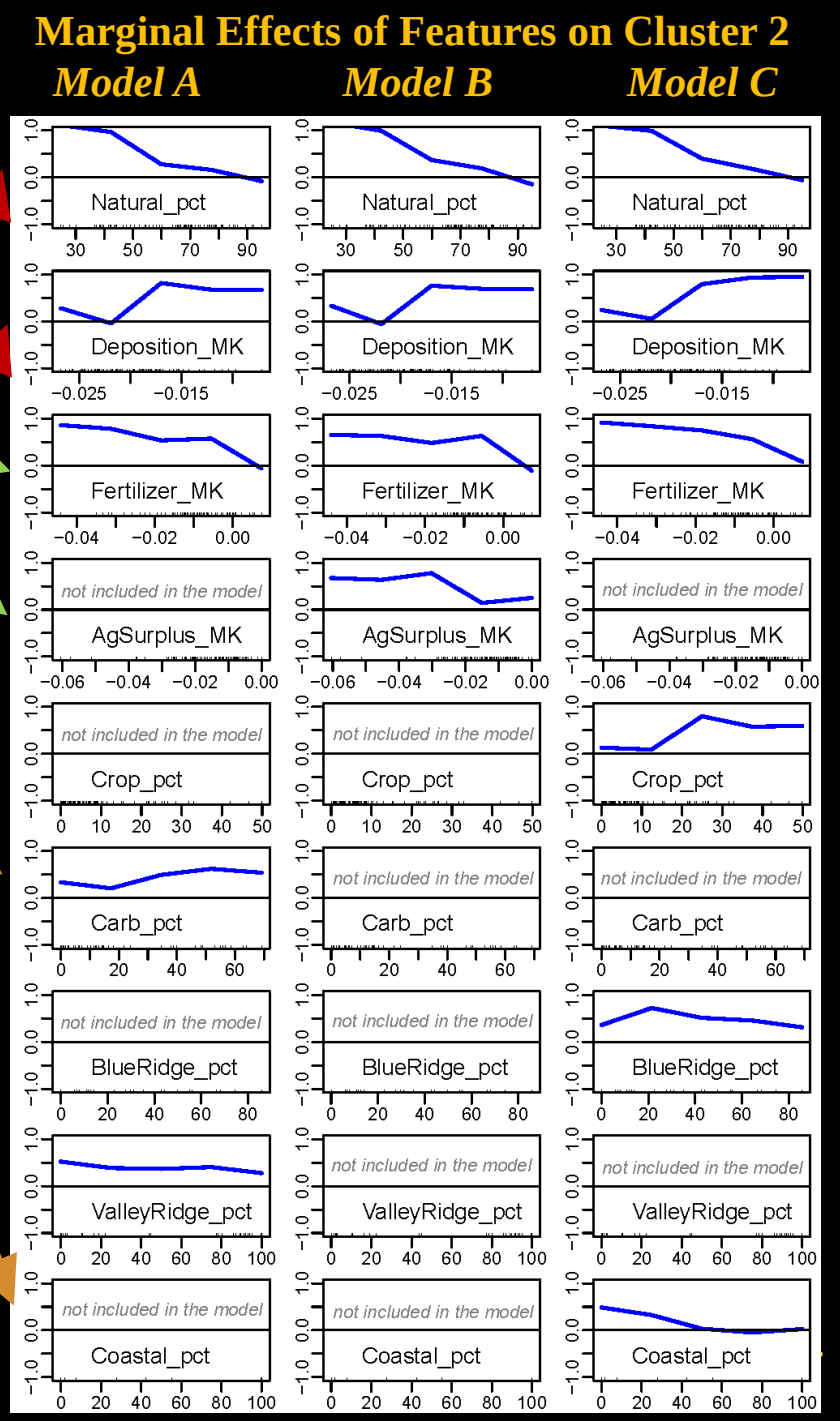
- Agricultural nutrient management contributed to water-quality improvement.

## Message 2 (Carb_pct, Coastal_pct):

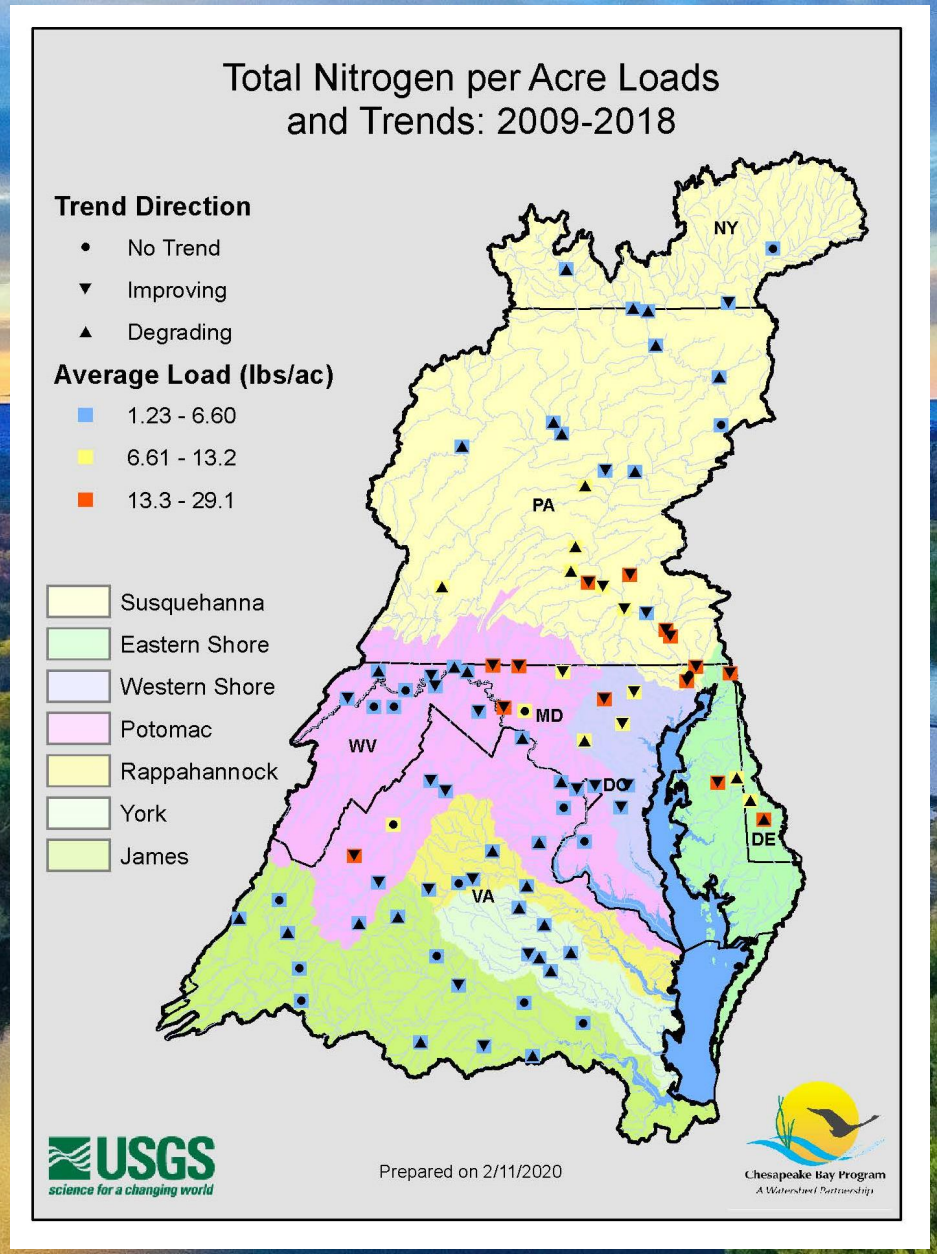
- Water-quality improvements are more likely in carbonate areas (relatively quick infiltration and faster groundwater transport) but less likely in Coastal Plain areas (accumulations of legacy N in the groundwater).

## Message 3 (Natural_pct, Deposition_MK):

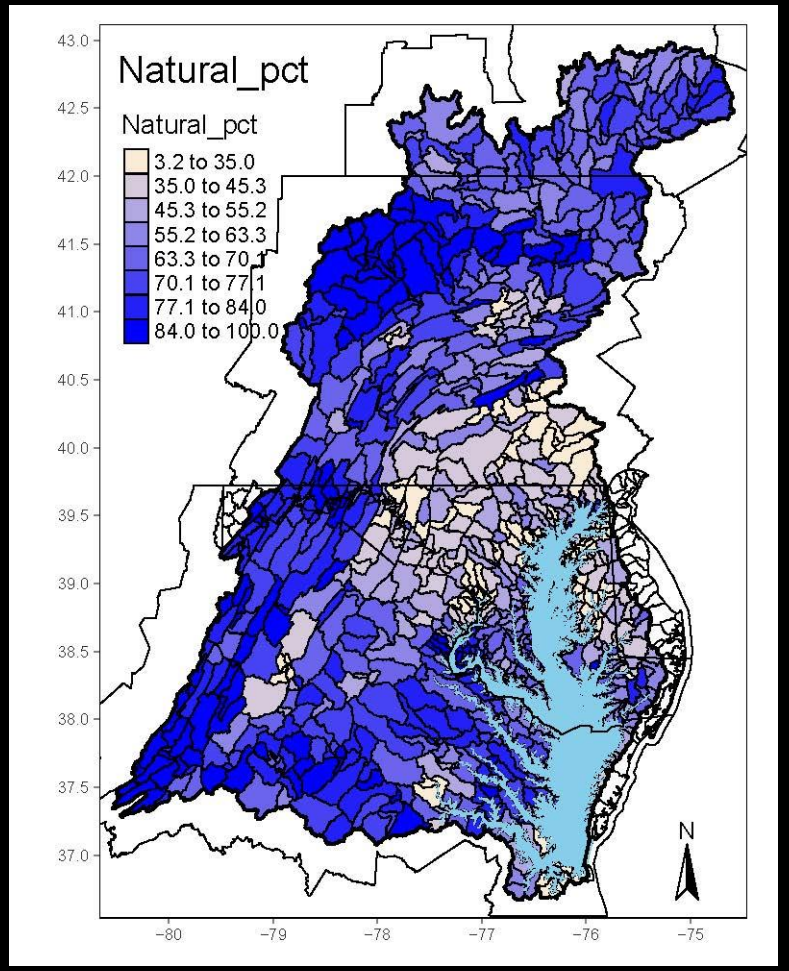
- We speculate that recent trends of increased TN in forested watersheds are attributed to: (1) increasing N inputs to non-forest regions and (2) mobilization of N from internal pools possibly due to deacidification.

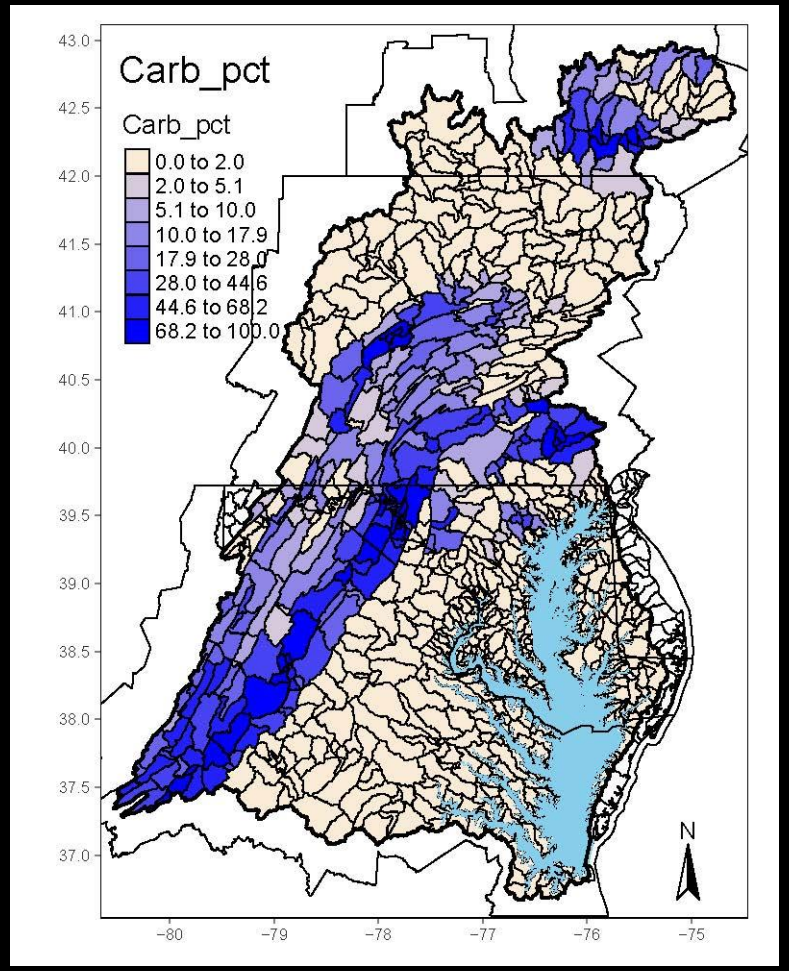


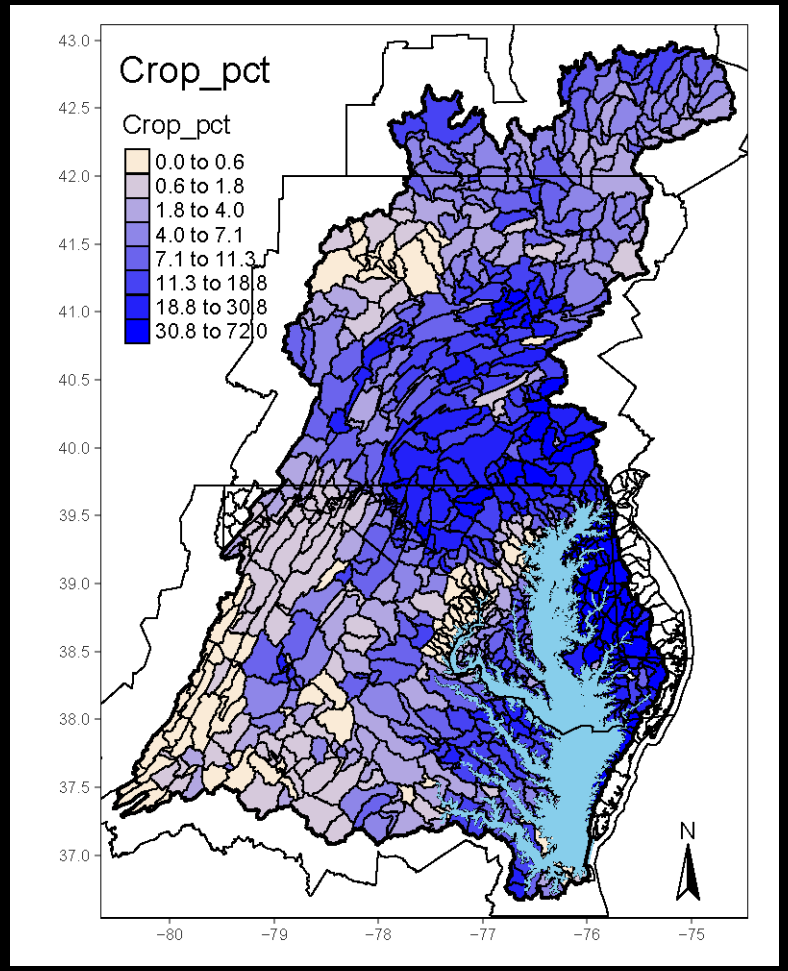
# 3. Prediction of nitrogen trend clusters for the entire watershed (Prediction)



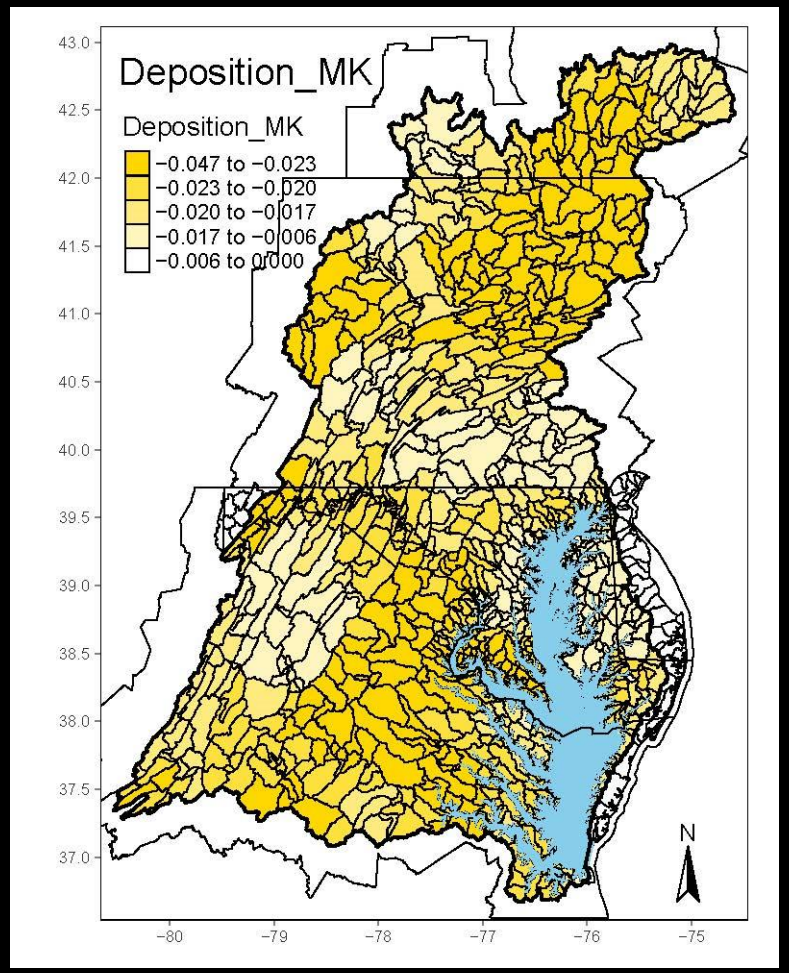
# Explanatory Variables for River Segments

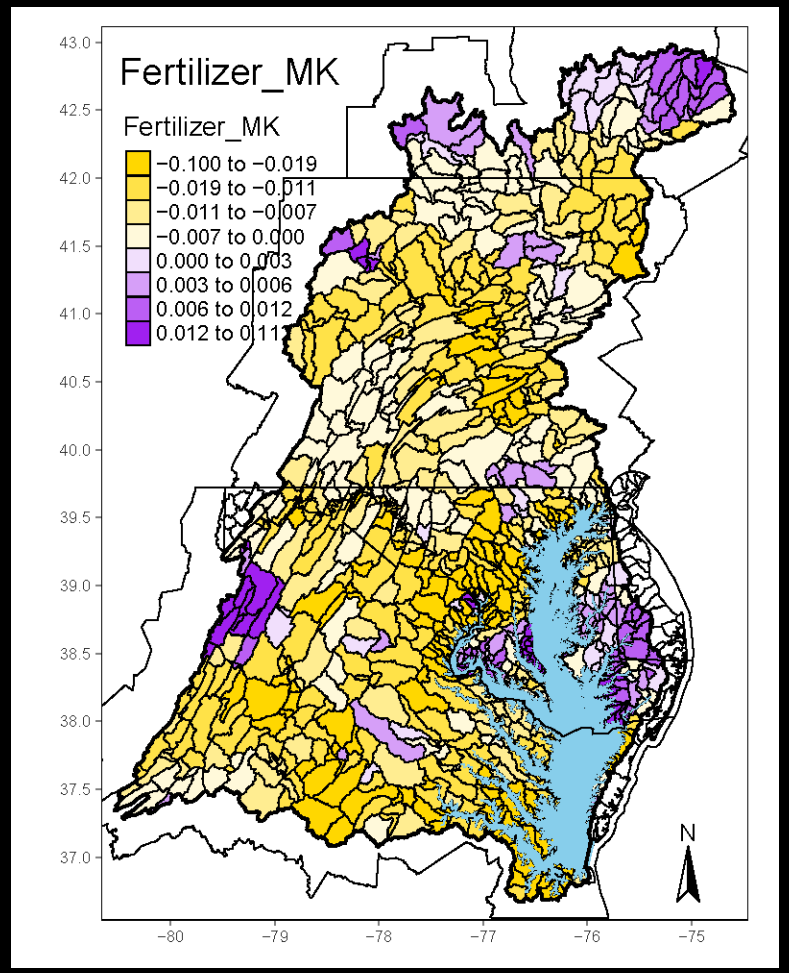


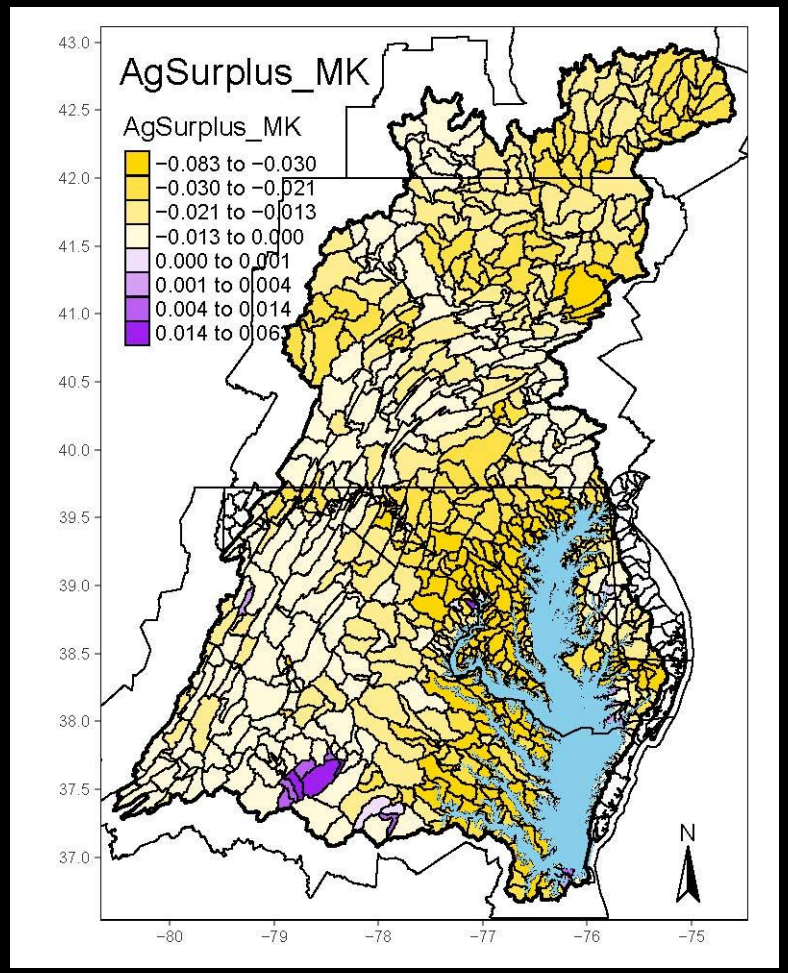



# Explanatory Variables for River Segments



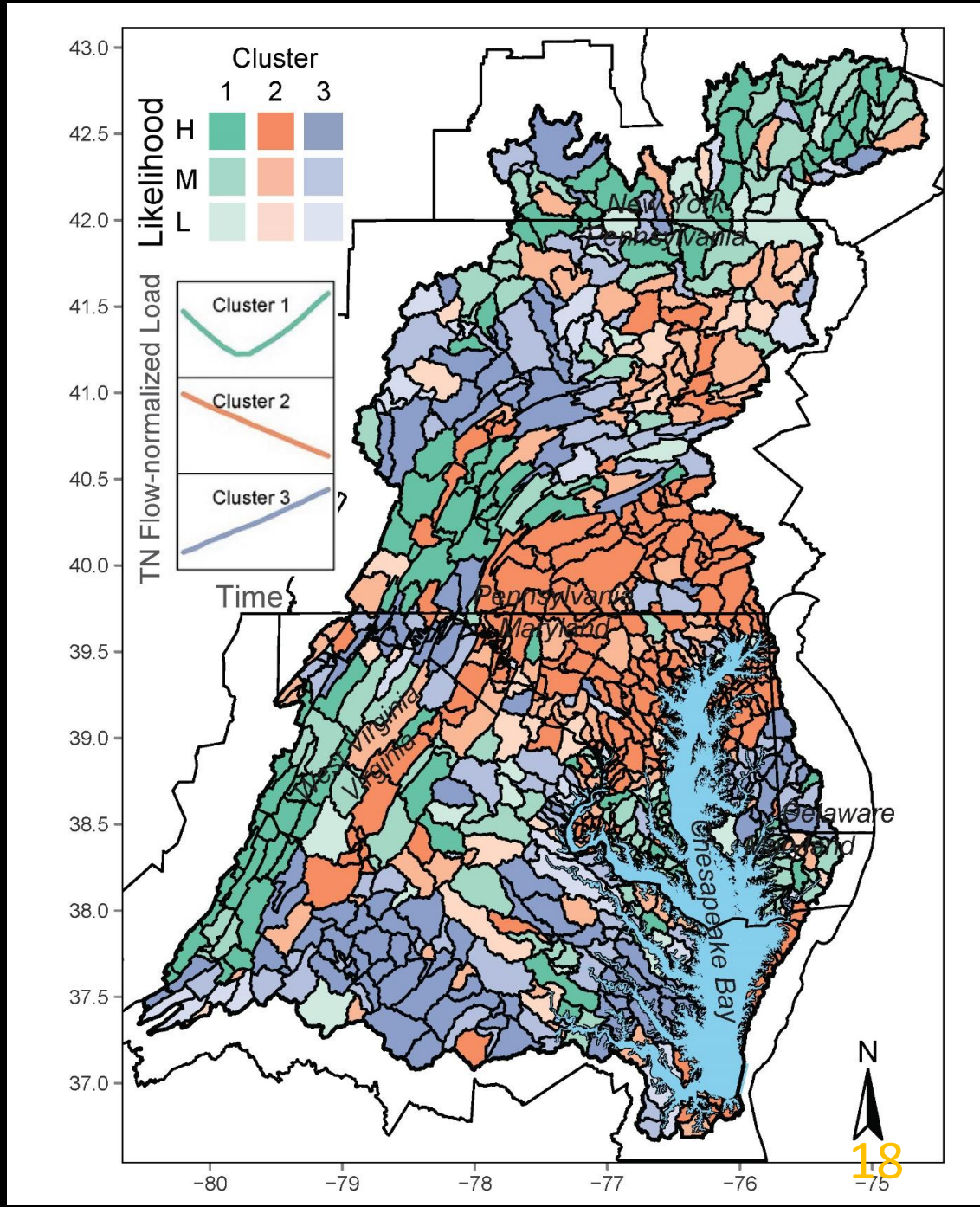



# Predictions for River Segments

| Cluster | No. of Segments | High Likelihood | Medium Likelihood | Low Likelihood |
|---|---|---|---|---|
| Cluster 1 | 292 (30%) | 103 | 138 | 51 |
| Cluster 2 | 392 (40%) | 227 | 122 | 43 |
| Cluster 3 | 295 (30%) | 128 | 117 | 50 |

- These predictions are useful for watershed managers to understand trends across the watershed, including unmonitored areas.
- Combined with the effects of the model features, these predictions may inform managers on choosing priority watersheds toward water-quality improvement.



# Conclusions

- Machine learning approaches – i.e., hierarchical clustering and random forest – can be combined to better understand the regional patterns and drivers of TN trends in large river monitoring networks.
- We explicitly incorporated temporal trends in agricultural fertilizer, manure, and agricultural input as well as agricultural surplus, providing evidence that improved nutrient management has resulted in declines in agricultural nonpoint sources, which in turn contributed to water-quality improvement.
- Water-quality improvements are more likely in watersheds underlain by carbonate rocks but less likely in watersheds in the Coastal Plain.
- Results show degrading trends in forested watersheds, suggesting new and/or remobilized sources of N.
- Although we aimed for parsimony, models may be improved with additional features, e.g., management practice, legacy N, and riparian buffers.

# Regional patterns and drivers of total nitrogen trends in the Chesapeake Bay watershed: Insights from machine learning approaches and management implications

For more information of this work, check out our latest publication:

Zhang, Q., Bostic, J. T. & Sabo, R. D. 2022. *Water Research 218, 118443, doi:10.1016/j.watres.2022.118443.*

EGU General Assembly 2022

*Thank you!*

*Qian Zhang*
*UMCES / USEPA Chesapeake Bay Program*
*qzhang@chesapeakebay.net*