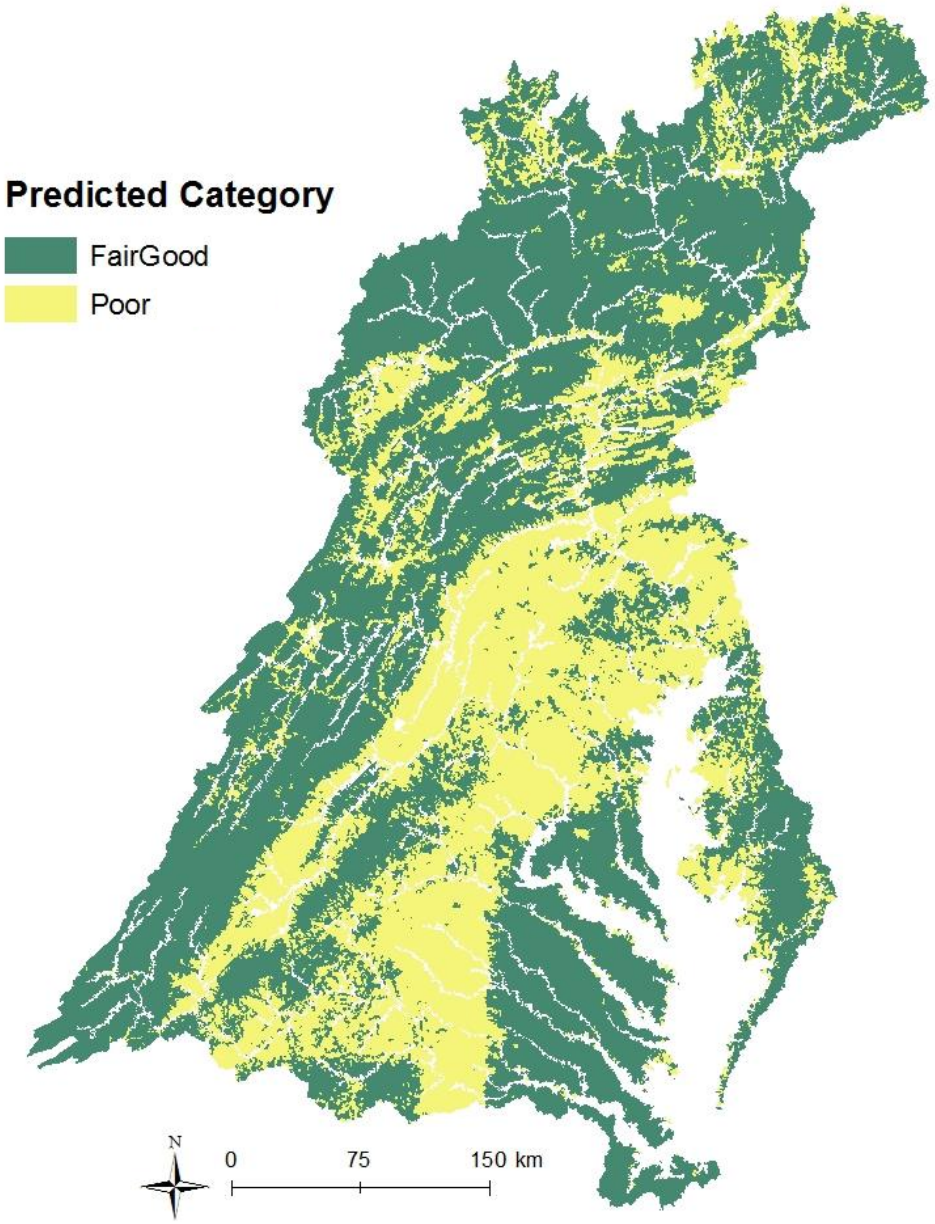



# Determining the Status of Unsurveyed Streams in the Chesapeake Bay Watershed

Kelly O. Maloney & John A. Young, USGS
Zachary M. Smith & Claire Buchanan, ICPRB

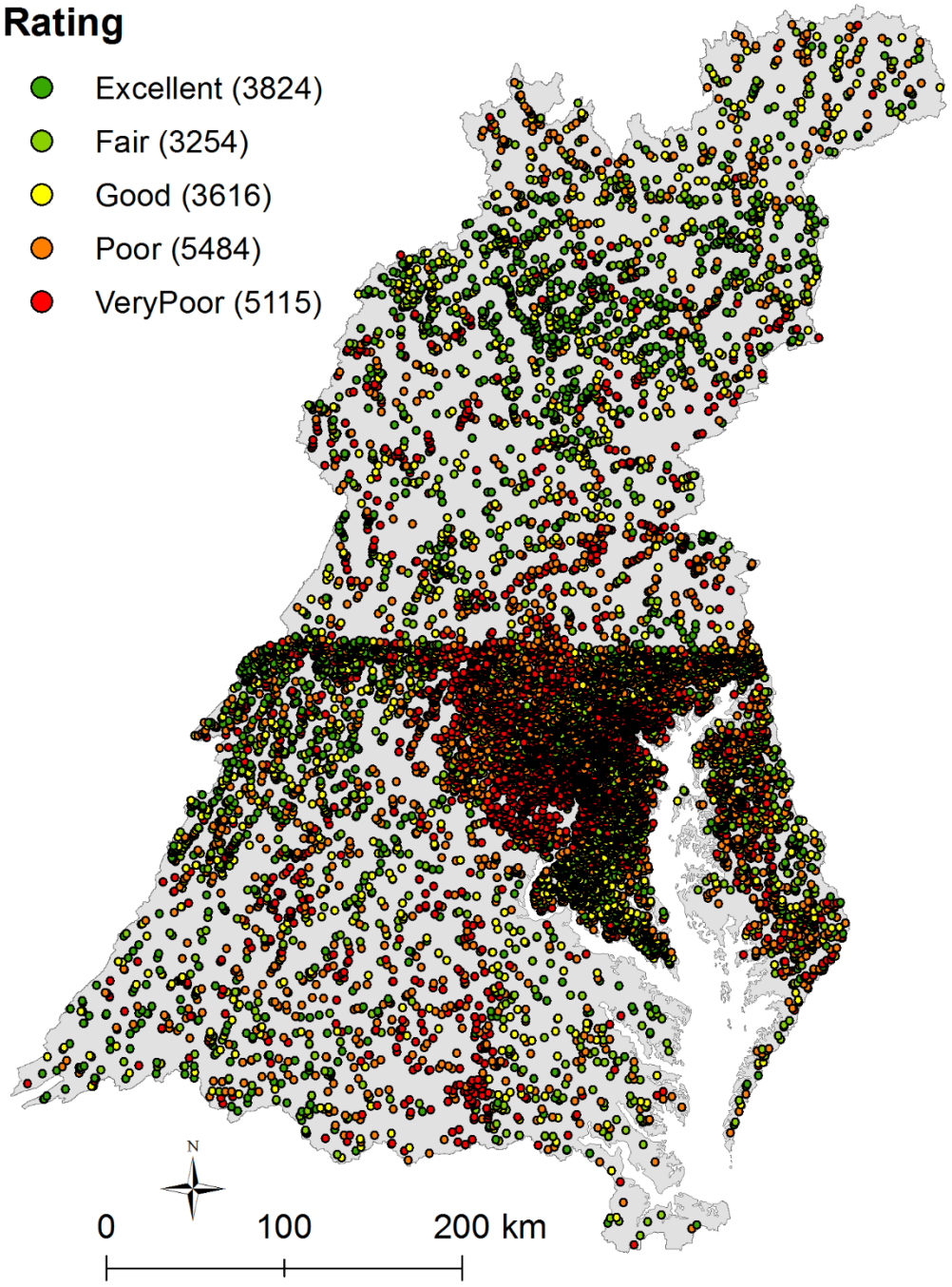
Developing a 2008 Baseline for the CBP Stream Health Indicator Workshop
5-6 April 2018
Cacapon Resort State Park, Berkeley Springs WV.

USGS science for a changing world

Interstate Commission on the Potomac River Basin
Established by Act of Congress 1940

# Chessie BIBI
# Family Level - Bioregion

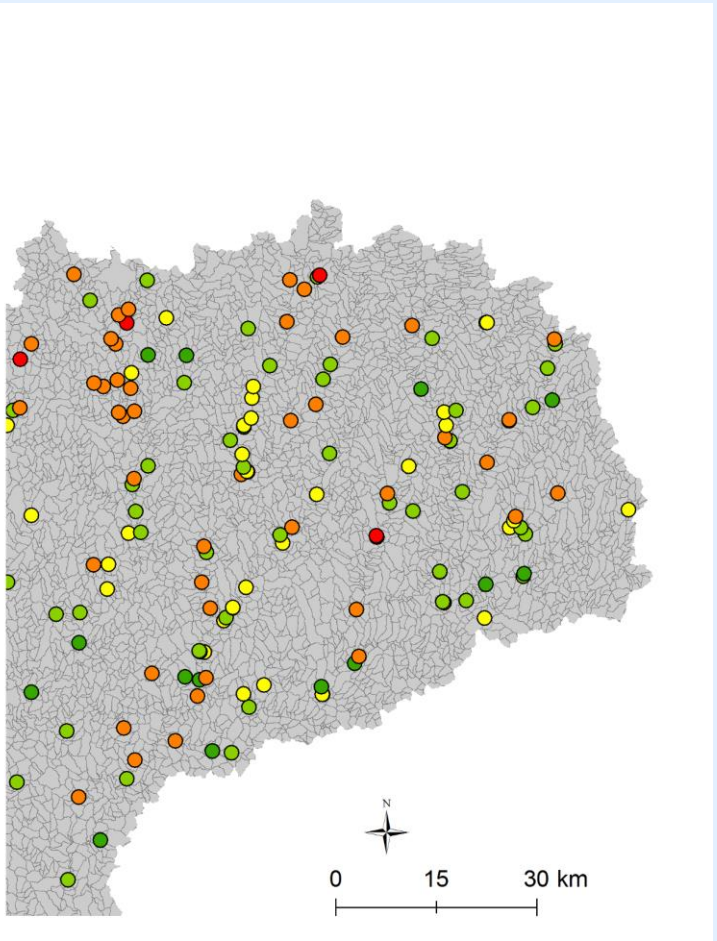
- 21,293 samples with latitude and longitude.
- Samples from 1992 through 2015.
- Maryland and Fairfax County have a disproportionate number of samples.

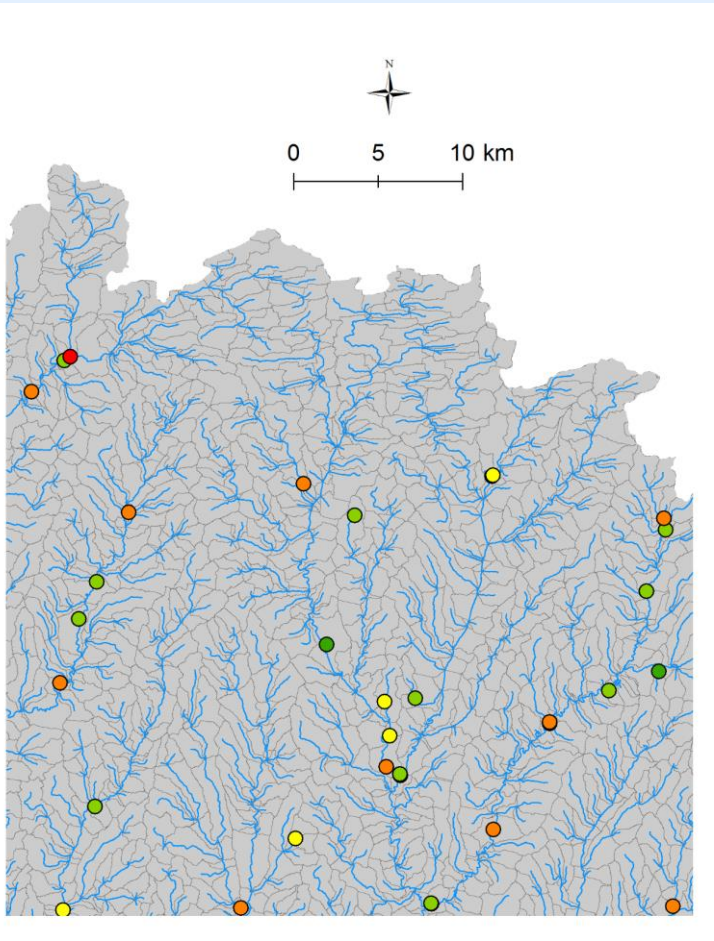


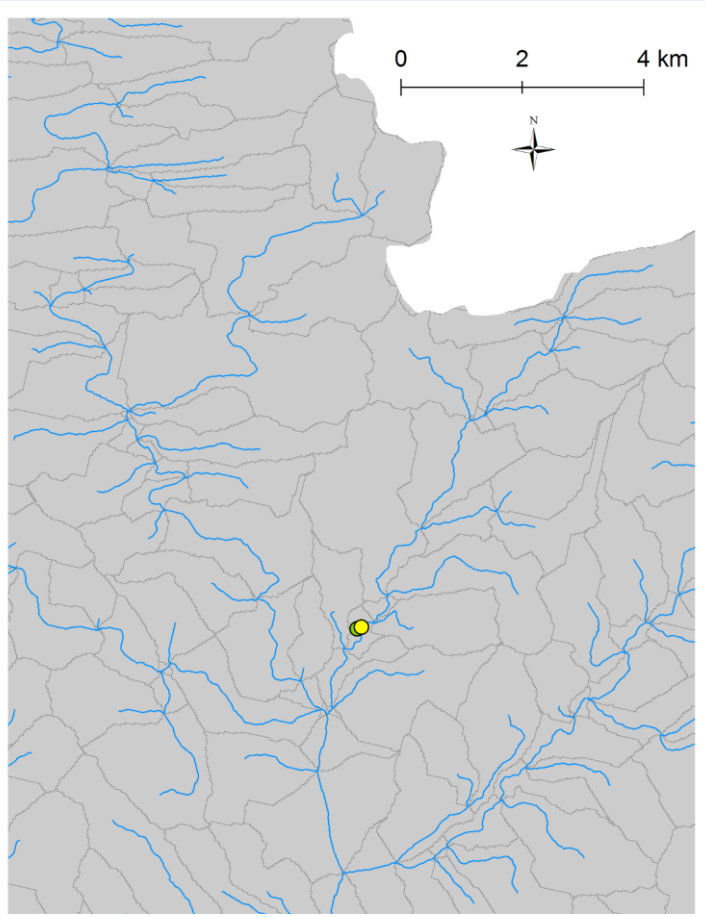

USGS

# Chessie BIBI
# Family Level - Bioregion

- Sampled/assessed streams are a small overall component of actual streams.

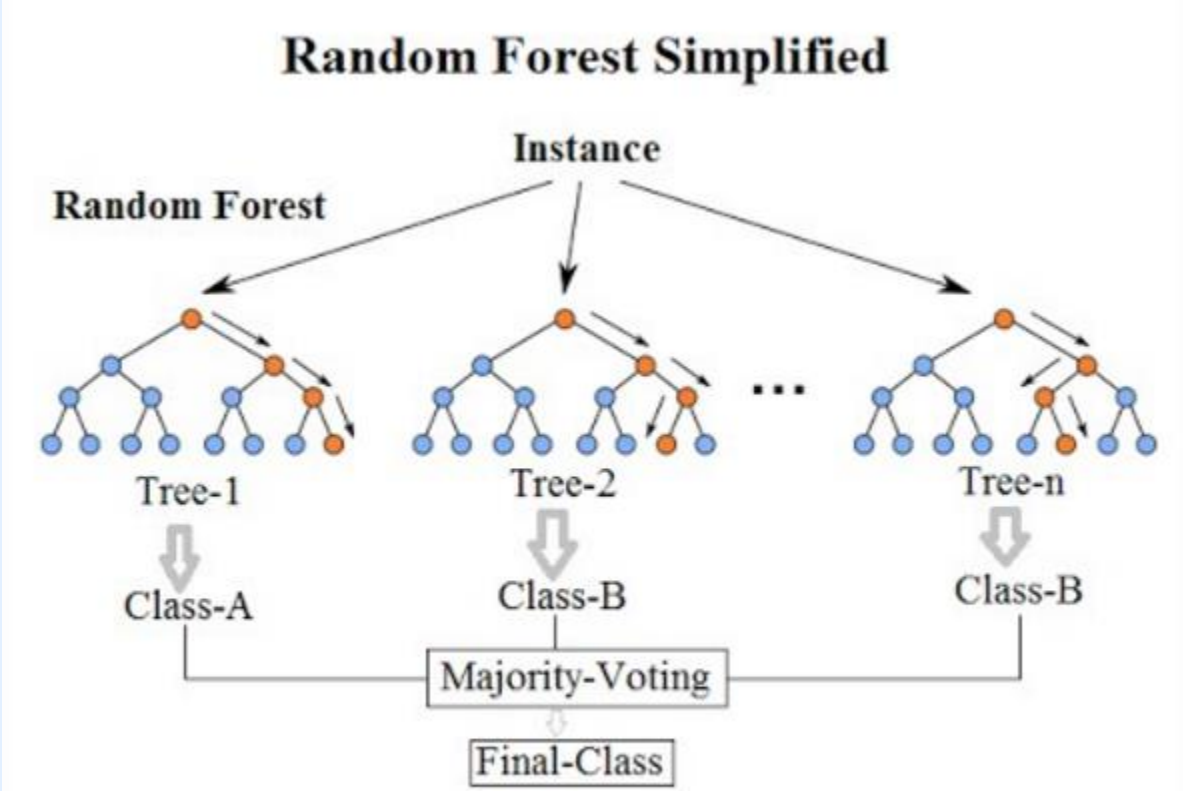


# Solution – Predictive Modeling Random Forest

- Combines many classification trees to increase classification accuracy (Cutler et al. 2007).
- Handles non-linear relationships and interactive effects.
- Top performer in classification accuracy (Fernández-Delgado et al. 2014).
- Variable importance plots and partial dependence plots can be used to examine predictors.

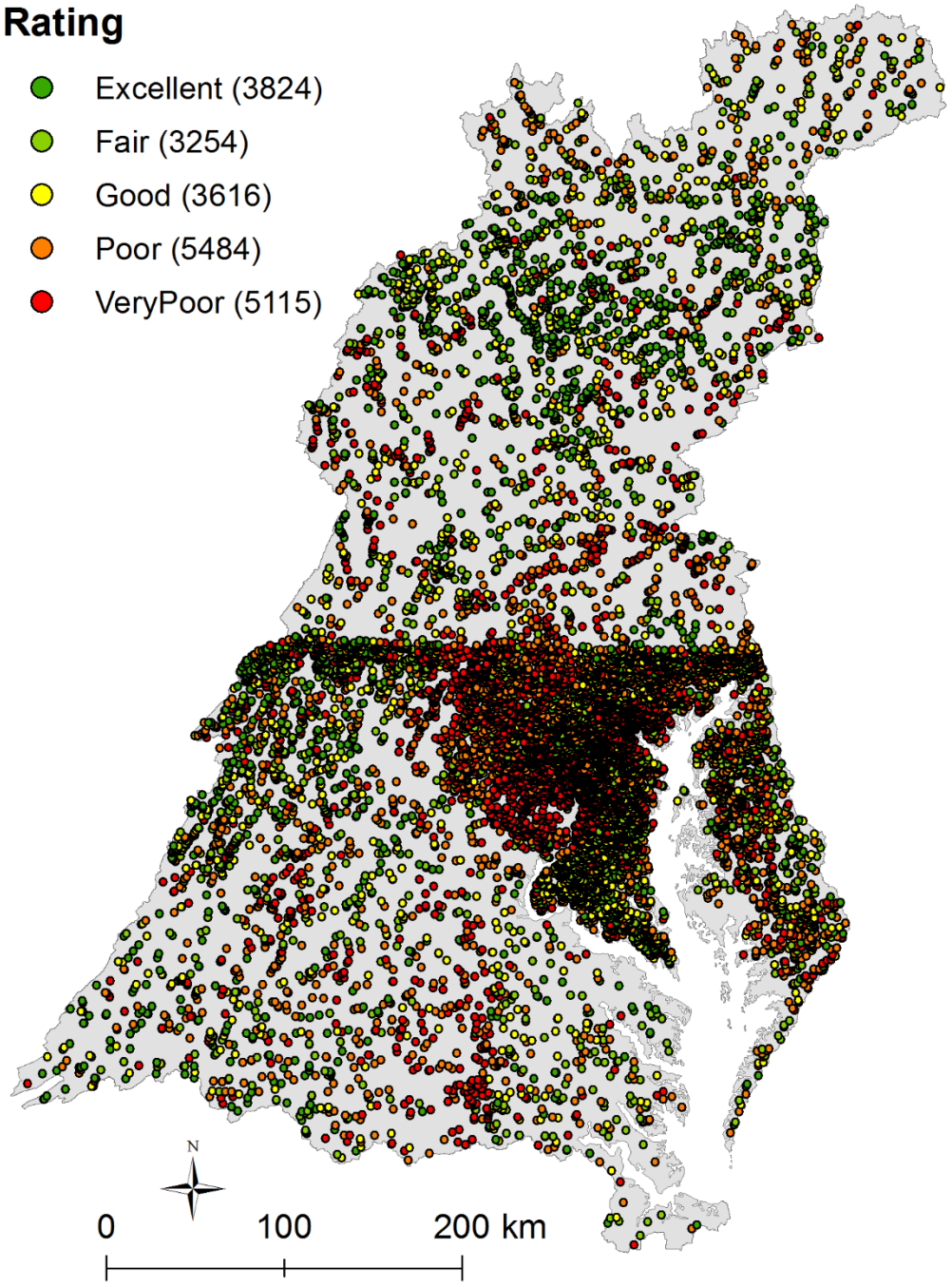



https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

USGS

# Chessie BIBI
# Family Level - Bioregion

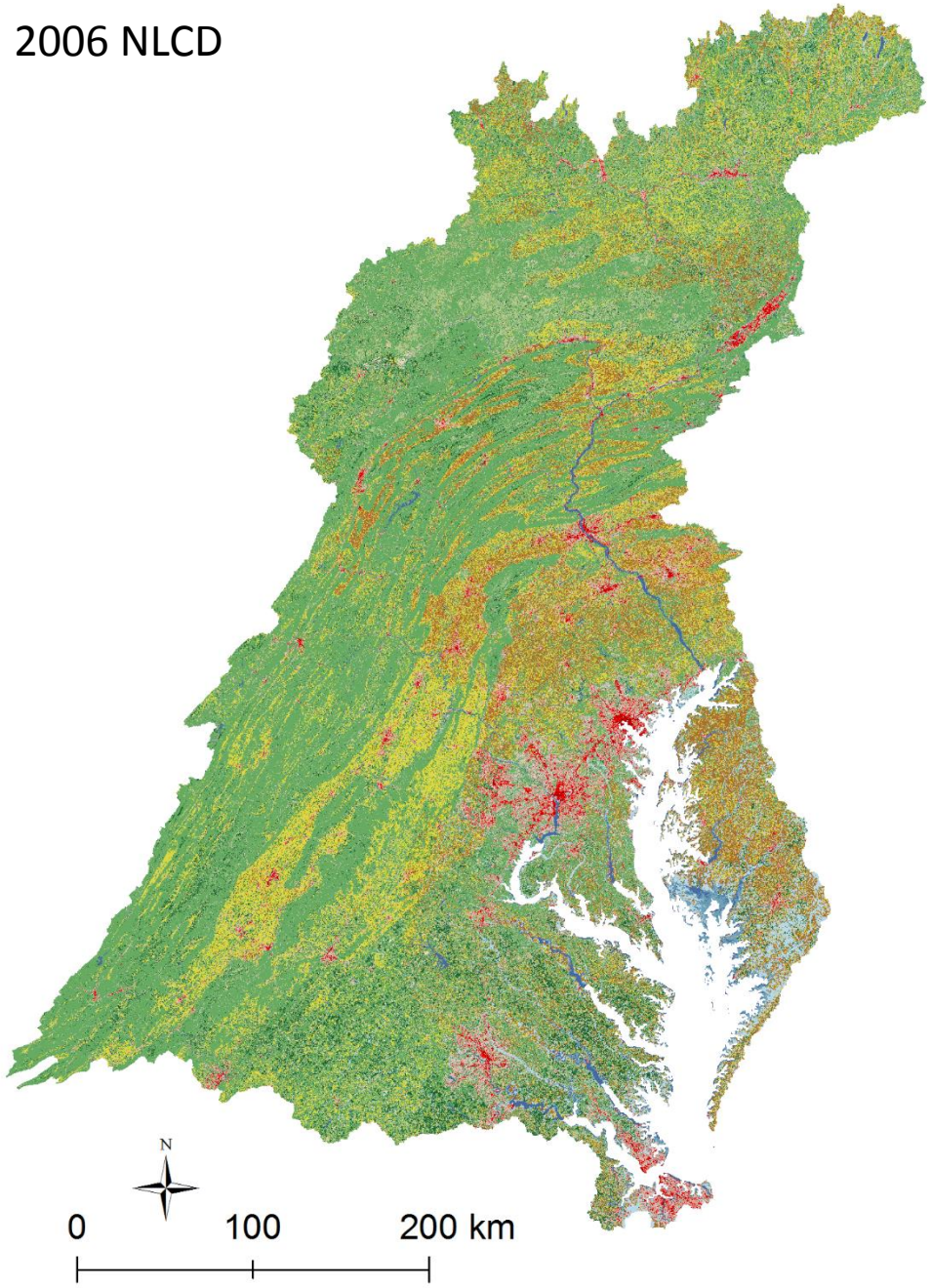
- 21,293 samples with latitude and longitude.
- Samples from 1992 through 2015.
- Maryland and Fairfax County have a disproportionate number of samples.




USGS

# Chessie BIBI
# Family Level - Bioregion

- Need to learn more about samples – drainage sizes, repeated sampling, etc.
- For modeling, landscape predictors largely available from the 2006 NLCD.
- Subset Chessie BIBI data from 2004-2008.
- 8,118 samples.




USGS

# Base Layer – Spatial Hydro-Ecological Decision System (SHEDS)

www.ecosheds.org

- 1:24,000 scale.
- Local catchment, upstream catchment and riparian summaries of 2006 NLCD, dams, soils, precipitation, elevation, slope.
- Spatially linked the 8,118 Chessie BIBI samples to the SHEDS data set.
- 82 Chessie BIBI samples did not have complete ecosheds data and were removed (n = 8,036).

USGS

# Base Layer – Spatial Hydro-Ecological Decision System (SHEDS)

- Catchment sizes range from 0.8 to 16,345 $km^2$.
- Focused modeling on small, headwaters < 200 $km^2$.
- Removed 576 samples that had upstream drainages areas equal to or above 200 $km^2$ (n = 7,460).

- Identify and remove repeated samples.
- Removed 3,084 duplicates (n = 4,376).

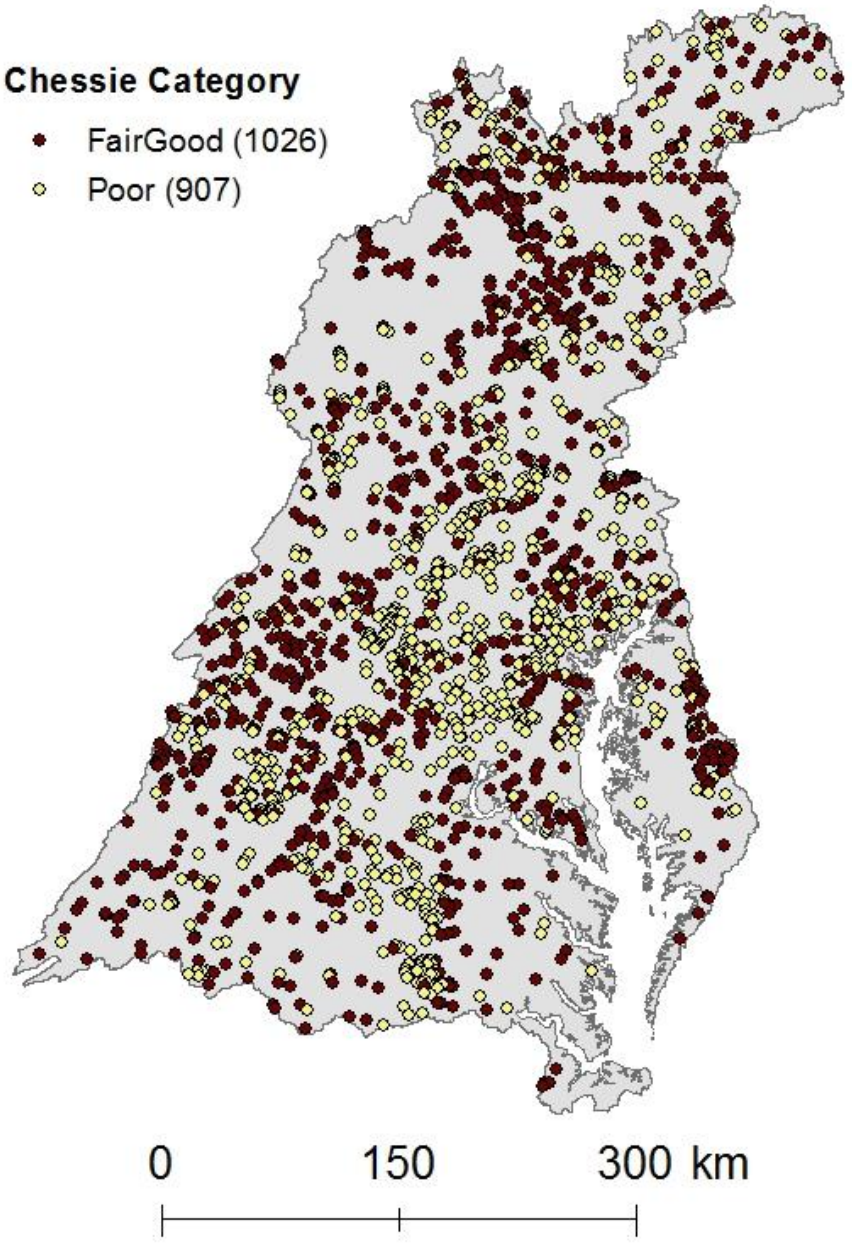
# Chessie BIBI
# Family Level - Bioregion

- 4,376 samples.
- Maryland and Fairfax County have a disproportionate number of samples.
- Resampled MD & Fairfax data to be consistent with density of samples across rest of watershed – 1,649 unique samples (~0.011586 samples $km^{-2}$).

USGS

# Chessie BIBI
# Family Level - Bioregion

- The area of Maryland and Fairfax county is ~24,604 $km^2$, therefore 24,604*0.011586= 285 samples; about 10.5% of the samples.
- Following removal of one erroneous point left 1,933 samples to build models.
- Reclassified Chessie BIBI rating to FairGood and Poor.
- 1,026 samples were rated FairGood (53.1%); 907 were rated Poor.



# Building the Random Forest Model

- Final model included dominant upstream bioregion, Latitude, Longitude and 9 predictors from the SHEDS data set.
- 75% of reduced Chessie BIBI dataset (n = 1,449) was used as a training data set and 25% (n = 484) was used as an independent validation data set.
- Predicted FairGood or Poor.
- Used the R package randomForest with 1,000 trees and an mtry (tested # candidates at each split) = 4).

SHEDS data used in Random Forest model.

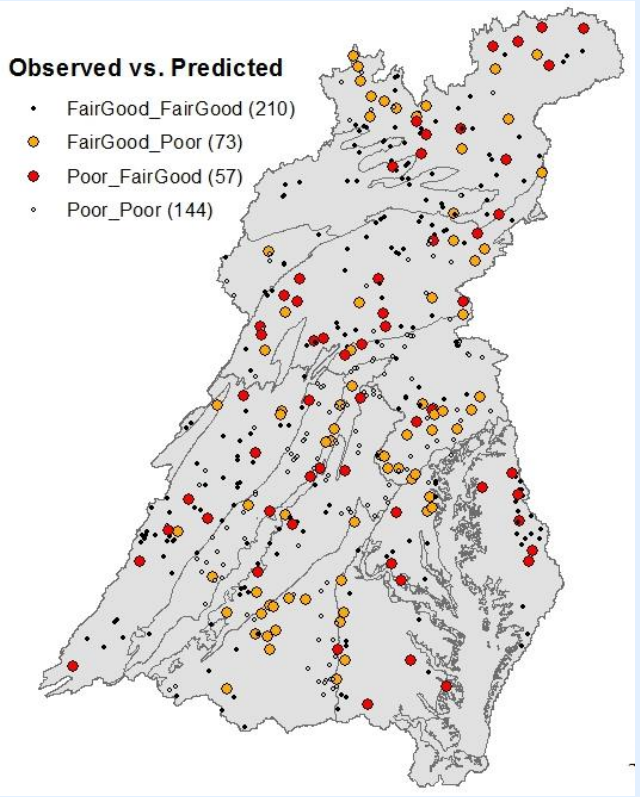
| Predictor (units) | Description |
|---|---|
| upstream.AreaSqKM ($km^2$) | Catchment area. |
| upstream.elevation (m) | Elevation spatially averaged over the upstream catchment area. |
| upstream.percent_sandy (%) | Percentage upstream catchment covered by soil material with texture described as "sandy". |
| upstream.hydrogroup_d4 (%) | Percentage of upstream catchment area with soils that have high runoff potential and very low infiltration and water transmission rates. |
| upstream.total_precip (mm) | 30 year (1981-2010) mean annual precipitation. |
| upstream.dep_so4_2011 (kg/ha) | Total wet atmospheric deposition of sulfate ion in 2011 spatially averaged over the catchment area. |
| upstream.impervious (%) | Percentage of upstream drainage area as impervious surface. |
| upstream.tree_canopy (%) | Percentage of upstream catchment area covered by tree canopy. |
| local.deg_barr_all (count) | Total number of dams in the local catchment. |

USGS

# Random Forest - Results

- Optimized cutoff that maximized probability of a Poor rating was identified at 0.4829.
- The model correctly predicted 73.1% of Poor samples and 71.7% of FairGood samples in the training data set.
- It predicted 71.6% of Poor samples and 74.2% of FairGood samples in the validation data set.
- No spatial pattern in misclassified sites evident.

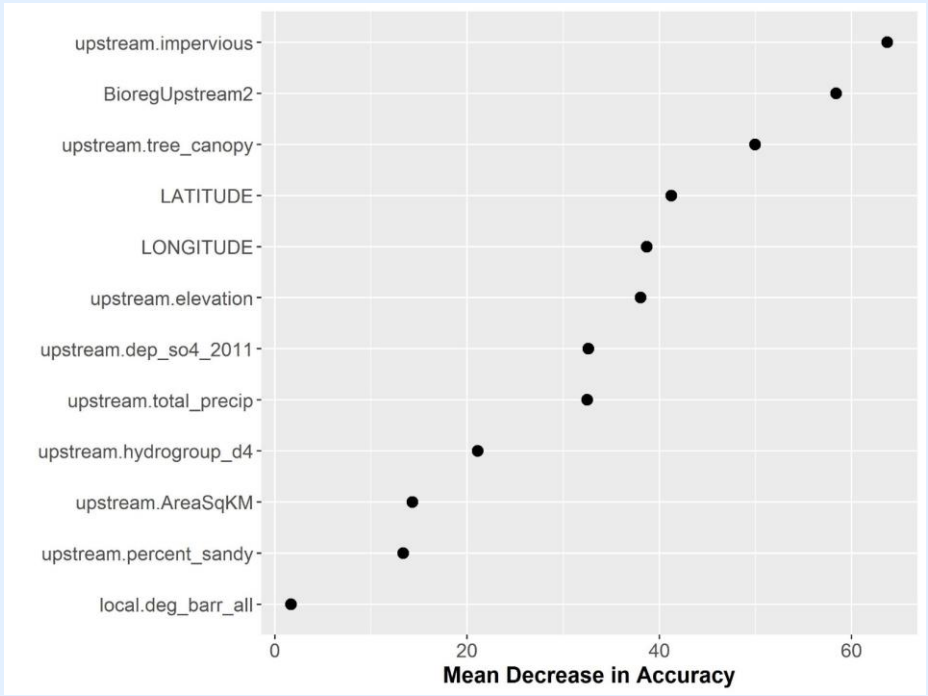
| | Training | | | Test | | |
|---|---|---|---|---|---|---|
| | Predicted category | | | Predicted category | | |
| Observed category | Poor | Fair/ Good | PCC | Poor | Fair/ Good | PCC |
| Poor | 516 | 190 | 73.1 | 144 | 57 | 71.6 |
| FairGood | 210 | 533 | 71.7 | 73 | 210 | 74.2 |

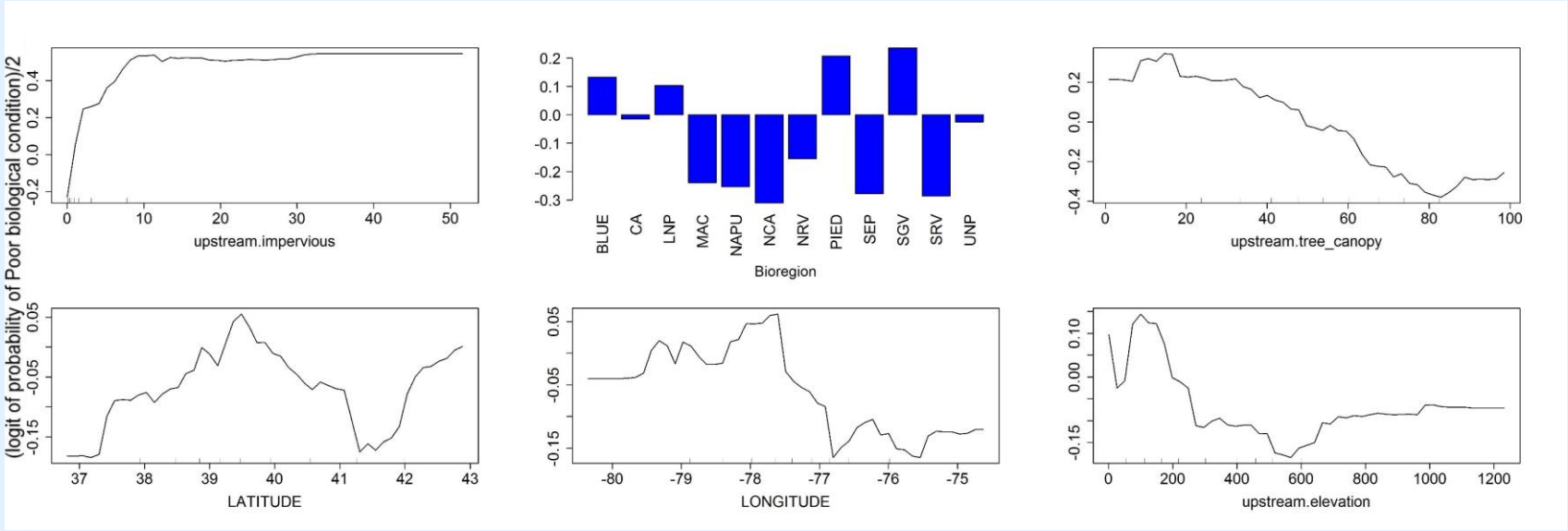



*Preliminary Information-Subject to Revision. Not for Citation or Distribution

# Random Forest - Results

- Upstream impervious surface, dominant bioregion and tree canopy were the top three important variables.
- Functional relationships agreed with literature.



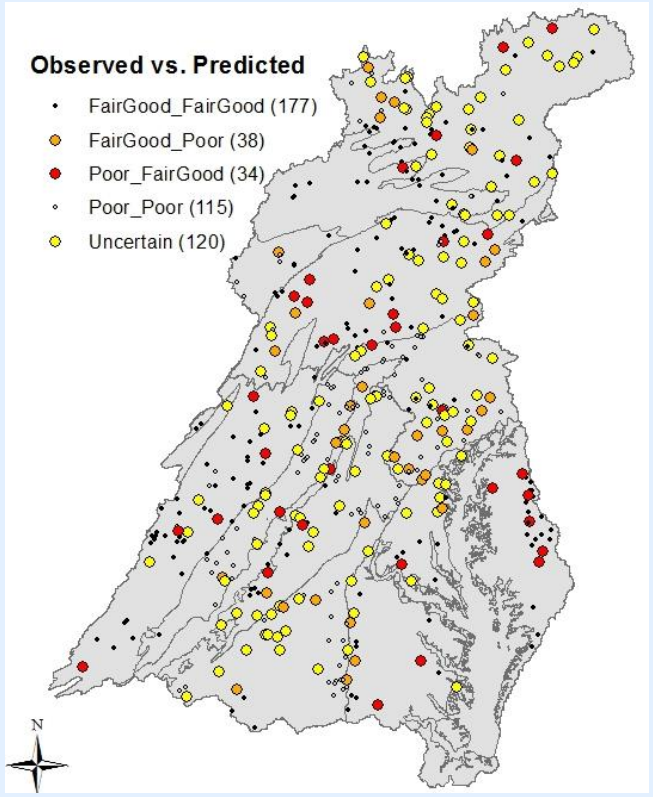




*Preliminary Information-Subject to Revision. Not for Citation or Distribution

# Random Forest - Results

- Uncertainty - ± 0.10 of the 0.4829 cutoff considered "Uncertain"
- Increased PCC for validation test data, but 120 samples (24.8%) were in the uncertain band.
- No spatial pattern in uncertain sites was evident.

| | Test Buffered Cutoff | | | |
|---|---|---|---|---|
| | Predicted category | | | |
| Observed category | Poor | Fair/ Good | Uncer-tain | PCC |
| Poor | 115 | 34 | 52 | 77.2 |
| FairGood | 38 | 177 | 68 | 82.3 |




*Preliminary Information-Subject to Revision. Not for Citation or Distribution

# Random Forest - Results

- Also summarized by bioregion.
- For Training data BLUE, CA and NRV Poor and FairGood were predicted reasonably well.
- For remaining bioregions the model predicted one rating better that the other and for some accuracies were less than 40%.

| | | Training | | |
|---|---|---|---|---|
| | | Predicted category | | |
| Bioregion | Observed category | Poor | Fair/ Good | PCC |
| BLUE | Poor | 40 | 10 | 80.0 |
| | FairGood | 15 | 32 | 68.1 |
| CA | Poor | 11 | 6 | 64.7 |
| | FairGood | 3 | 19 | 86.4 |
| LNP | Poor | 70 | 12 | 85.4 |
| | **FairGood** | **25** | **8** | **24.2** |
| MAC | **Poor** | **7** | **16** | **30.4** |
| | FairGood | 7 | 39 | 84.8 |
| NAPU | Poor | 45 | 39 | 53.6 |
| | FairGood | 37 | 86 | 69.9 |
| NCA | Poor | 9 | 11 | 45.0 |
| | FairGood | 8 | 55 | 87.3 |
| NRV | Poor | 45 | 29 | 60.8 |
| | FairGood | 19 | 87 | 82.1 |
| PIED | Poor | 87 | 7 | 92.6 |
| | FairGood | 23 | 16 | 41.0 |
| SEP | Poor | 20 | 15 | 57.1 |
| | FairGood | 9 | 57 | 86.4 |
| SGV | Poor | 84 | 4 | 95.5 |
| | **FairGood** | **15** | **6** | **28.6** |
| SRV | **Poor** | **15** | **27** | **35.7** |
| | FairGood | 12 | 104 | 89.7 |
| UNP | Poor | 82 | 15 | 84.5 |
| | **FairGood** | **37** | **24** | **39.3** |

*Preliminary Information-Subject to Revision. Not for Citation or Distribution

# Random Forest - Results

- For Test data, the model predicted similarly except for the LNP – both categories predicted reasonably well and for NRV where Poor were now predicted better than FairGood.
- The ±0.10 uncertainty results showed similar patterns but reduced sample sizes became an issue.

| | | Training | | | Test | | | Test Buffered Cutoff | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Predicted category | | | Predicted category | | | Predicted category | | | |
| Bioregion | Observed category | Poor | Fair/ Good | PCC | Poor | Fair/ Good | PCC | Poor | Fair/ Good | Uncer-tain | PCC |
| BLUE | Poor | 40 | 10 | 80.0 | 11 | 2 | 84.6 | 8 | 1 | 4 | 88.9 |
| | FairGood | 15 | 32 | 68.1 | 5 | 8 | 61.5 | 4 | 6 | 3 | 60.0 |
| CA | Poor | 11 | 6 | 64.7 | 4 | 0 | 100.0 | 4 | 0 | 0 | 100.0 |
| | FairGood | 3 | 19 | 86.4 | 2 | 7 | 77.8 | 1 | 7 | 1 | 87.5 |
| LNP | Poor | 70 | 12 | 85.4 | 20 | 4 | 83.3 | 16 | 2 | 6 | 88.9 |
| | **FairGood** | **25** | **8** | **24.2** | 2 | 9 | 81.8 | 1 | 6 | 4 | 85.7 |
| MAC | **Poor** | **7** | **16** | **30.4** | **0** | **7** | **0.0** | **0** | **6** | **1** | **0.0** |
| | FairGood | 7 | 39 | 84.8 | 0 | 18 | 100.0 | 0 | 18 | 0 | 100.0 |
| NAPU | Poor | 45 | 39 | 53.6 | 8 | 11 | 42.1 | 6 | 4 | 9 | 60.0 |
| | FairGood | 37 | 86 | 69.9 | 14 | 31 | 68.9 | 5 | 19 | 21 | 79.2 |
| NCA | Poor | 9 | 11 | 45.0 | 1 | 1 | 50.0 | 1 | 1 | 0 | 50.0 |
| | FairGood | 8 | 55 | 87.3 | 1 | 24 | 96.0 | 1 | 21 | 3 | 95.5 |
| NRV | Poor | 45 | 29 | 60.8 | 12 | 13 | 48.0 | 6 | 9 | 10 | 40.0 |
| | FairGood | 19 | 87 | 82.1 | 7 | 32 | 82.1 | 5 | 28 | 6 | 84.8 |
| PIED | Poor | 87 | 7 | 92.6 | 21 | 0 | 100.0 | 20 | 0 | 1 | 100.0 |
| | FairGood | 23 | 16 | 41.0 | **15** | **8** | **34.8** | **7** | **3** | **13** | **30.0** |
| SEP | Poor | 20 | 15 | 57.1 | 4 | 6 | 40.0 | 2 | 3 | 5 | 40.0 |
| | FairGood | 9 | 57 | 86.4 | 6 | 16 | 72.7 | 3 | 15 | 4 | 83.3 |
| SGV | Poor | 84 | 4 | 95.5 | 32 | 1 | 97.0 | 29 | 1 | 3 | 96.7 |
| | **FairGood** | **15** | **6** | **28.6** | **3** | **0** | **0.0** | **2** | **0** | **1** | **0.0** |
| SRV | **Poor** | **15** | **27** | **35.7** | 6 | 8 | 42.9 | 4 | 5 | 5 | 44.4 |
| | FairGood | 12 | 104 | 89.7 | 3 | 44 | 93.6 | 0 | 44 | 3 | 100.0 |
| UNP | Poor | 82 | 15 | 84.5 | 25 | 4 | 86.2 | 19 | 2 | 8 | 90.5 |
| | **FairGood** | **37** | **24** | **39.3** | 15 | 13 | 46.4 | 9 | 10 | 9 | 52.6 |

*Preliminary Information-Subject to Revision. Not for Citation or Distribution

# Random Forest Model – Summary & Next Steps

- Built using a subset of Chessie BIBI data, 2004-2008, small catchments, no repeated samples using the Family level Bioregion data.
- Model predicted probability of Poor and FairGood, which were composites of the 5 category rating in the Chessie BIBI.
- Model overall performed well; however, accuracy varied across bioregions.
- Next steps:
  - Move to the NHDplus High Resolution data set and test additional predictors that are available for the Chesapeake and more management relevant (e.g., riparian measures, N, P, herbicide, pesticide application amounts, septic systems, land use change).
  - Reproduce with 2011 NLCD and 2009-2013 Chessie BIBI data and other taxonomic and scale data?

# Acknowledgements


- Ecosheds Team: www.ecosheds.org
- R: A Language and Environment for Statistical Computing.
- Support was provided by the USGS Fisheries Program and the Chesapeake Bay Program.
- Andrea Nagel, ICPRB

- The following programs that provided data for the Chessie BIBI:
  - Anne Arundel County Maryland - Department of Public Works
  - City of Baltimore - Department of Public Works
  - Baltimore County Department of Environmental Protection
  - District of Columbia Department of Energy and Environment
  - Delaware Department of Natural Resources and Environmental Control
  - Frederick County Department of Public Works
  - Fairfax County Department of Public Works and Environmental Services
  - Howard County Department of Public Works
  - Loudoun County Department of Building and Development
  - Montgomery County Department of Environmental Protection
  - Maryland Department of Natural Resources
  - New York Department of Environmental Conservation
  - Pennsylvania Department of Environmental Protection
  - Prince George's County Department of the Environment
  - Susquehanna River Basin Commission
  - United States Environmental Protection Agency
  - USDA Forest Service
  - United States Geological Survey
  - Virginia Department of Environmental Quality
  - Virginia Commonwealth University
  - West Virginia Department of Environmental Protection


USGS science for a changing world

INTERSTATE COMMISSION ON THE POTOMAC RIVER BASIN ESTABLISHED BY ACT OF CONGRESS 1940