Network Analysis: Data-Driven Optimization

Tools & Insights

Matthew Cashman

Leah Staub, Joel Blomquist,

Zach Clifton, John Hammond

MD-DE-DC Water Science Center
US Geological Survey

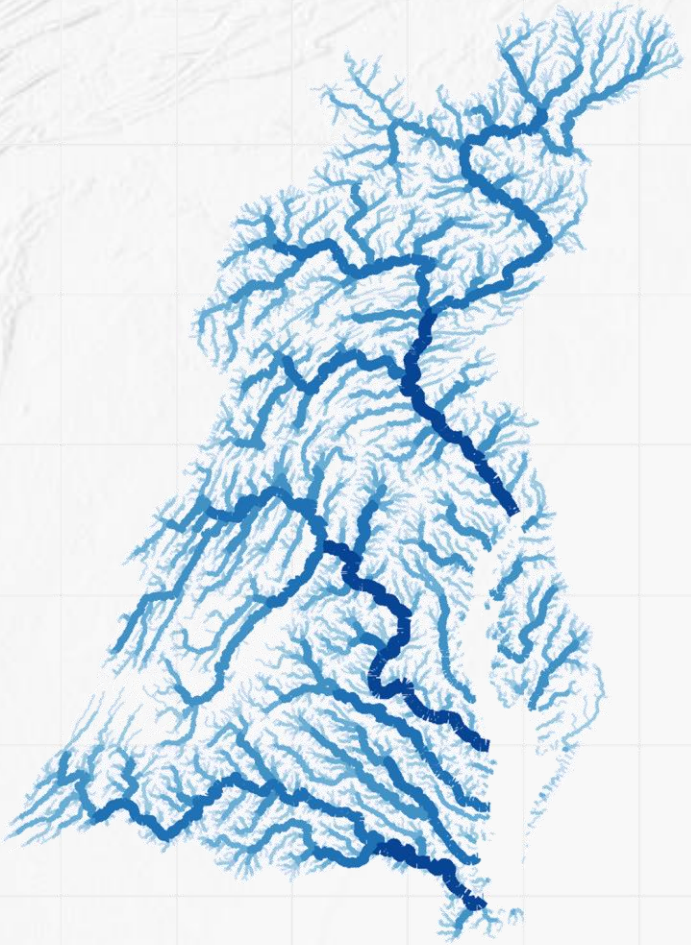
Outline

- Introduction
- A “Reusable” Workflow
- Example USGS Use-Cases
- Quantifying Gage Loss

Network Analysis

- A comparison of the representativeness of a monitoring network (“sample” or “target”) versus the underlying population it is supposed to represent (e.g. the river network)

- Useful for:
  - Predictive modeling
  - Identifying monitoring gaps to fill
  - Minimizing the impacts of monitoring cuts
  - Others…?

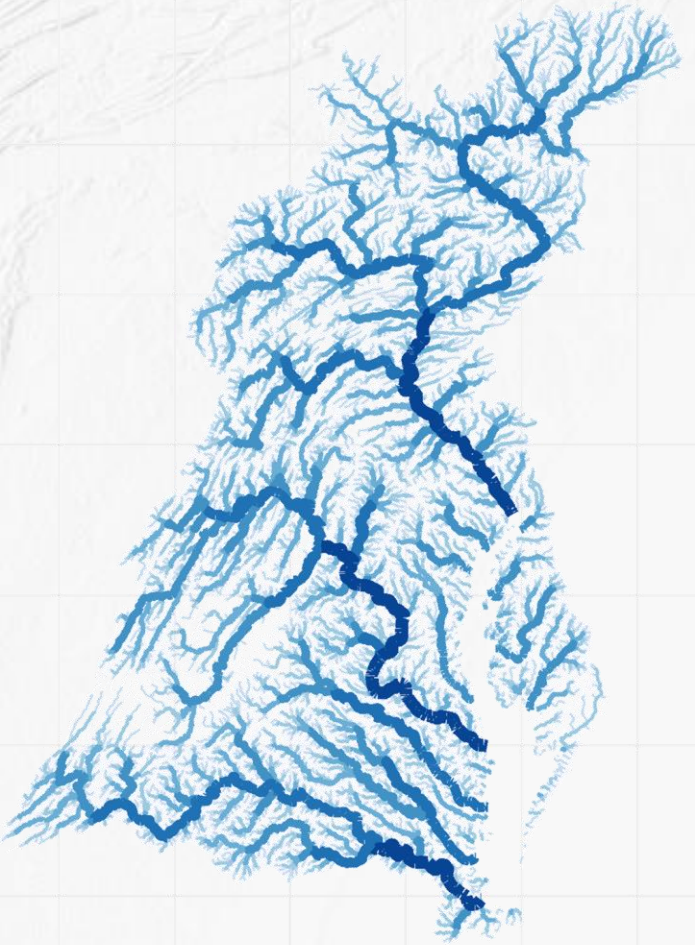


Network Analysis

A basic question: “Are the samples representative of the population?”

- But challenges:
  - Unwieldy “big” data (large sizes, different sources)
  - Large GIS preprocessing *before* any analysis
  - Analysis methods (“how”)
  - Method reproducibility and iterative changes

➢“Representative of ***what****?”*

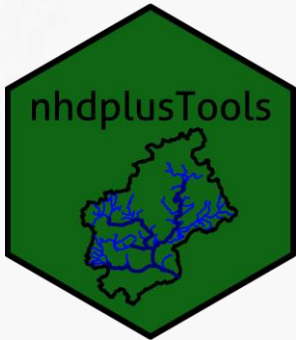


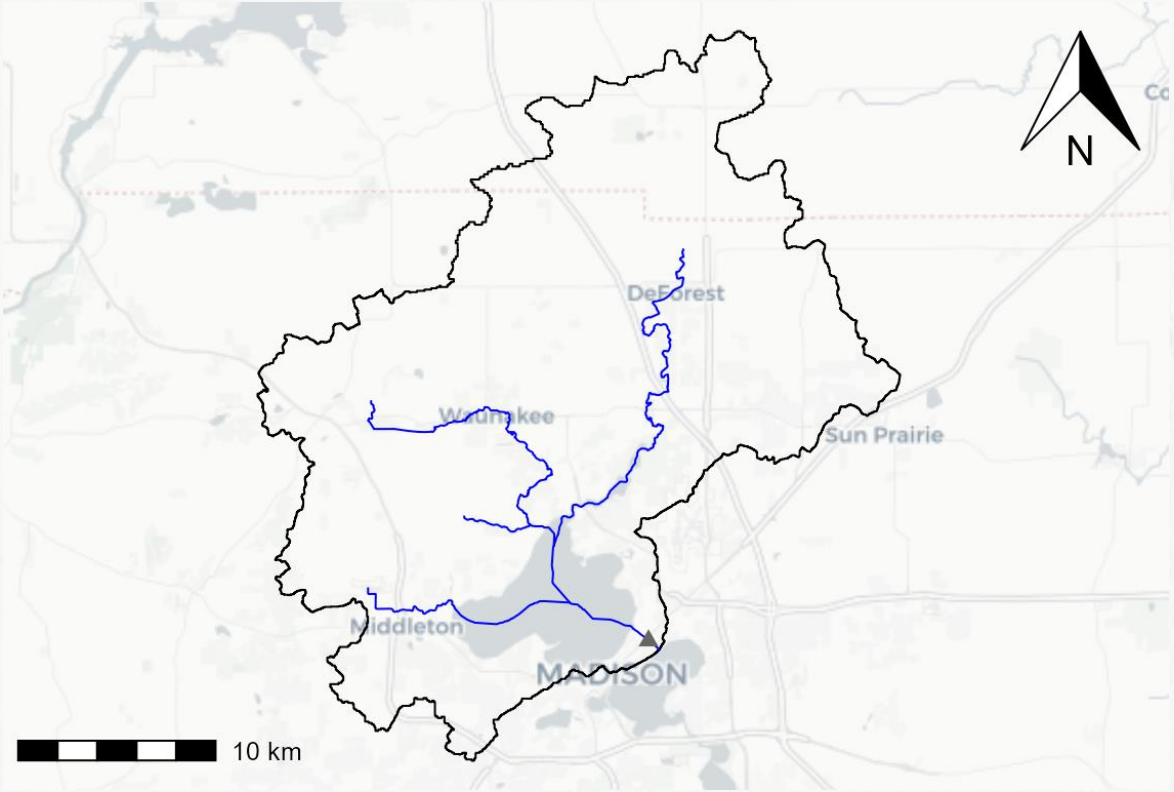
Making Network Analysis “easier”

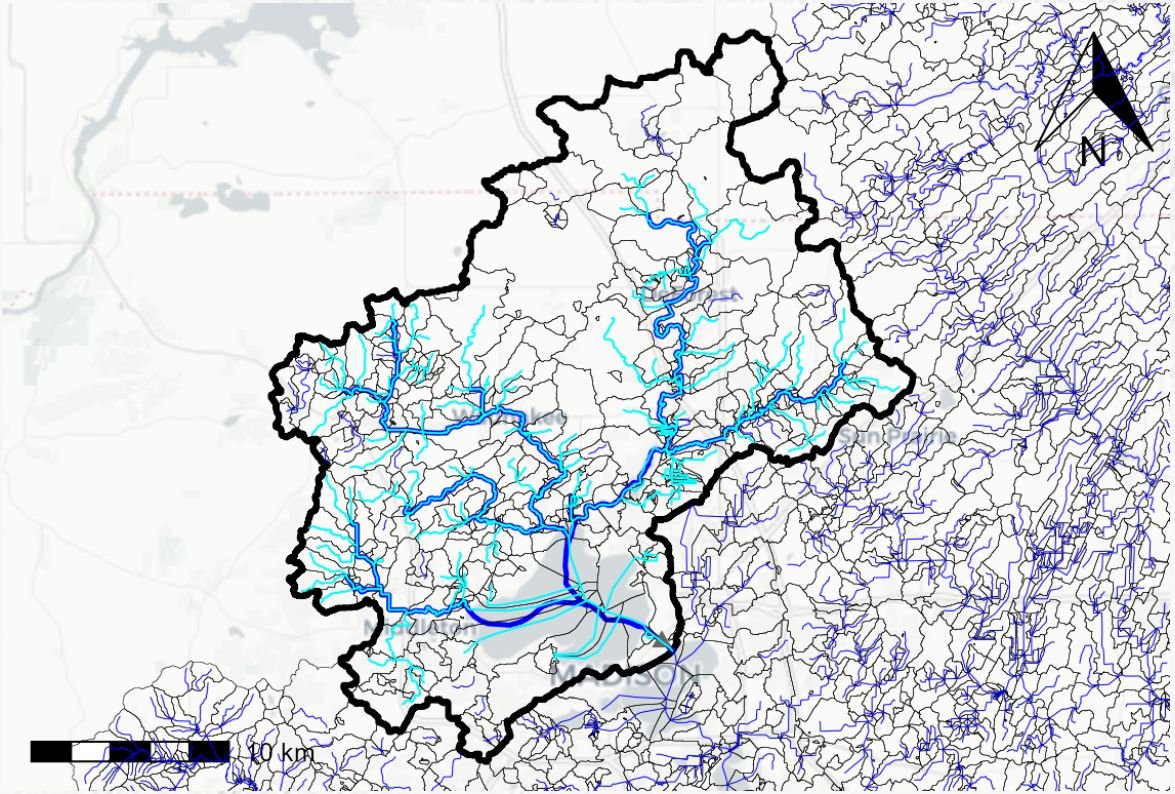
- *In-development* R code repository at www.code.usgs.gov

- Tools to streamline network analyses (some R skills involved)
  - Setup and download data
  - Custom queries (landscape variables and monitoring locations)
  - Reproducible for different networks/focal areas

- Leverages David Blodgett’s nhdplusTools package

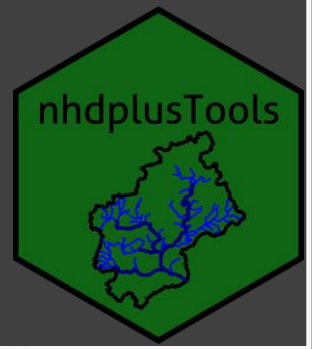






Using NHDPlus v2.1 1:100k Framework

- Using linked data framework
  - Flowlines
  - Catchments



Gathering required data

1. NHDPlus v2.1 CONUS seamless dataset directly from EPA[1]
   - *get_nhdplus2.1_seamless_l48()*

2. 1,000s of upstream variables for 2.7 mil catchments (CONUS) from USGS[2]
   - *get_mikew_nhdplus_accum_attributes()*

3. Reach COMID/HUC crosswalk table from USGS[3]
   - *get_comid_crosswalk()*

- [1]https://www.epa.gov/waterdata/nhdplus-national-data
- [2]Wieczorek, M.E., Jackson, S.E., and Schwarz, G.E., 2018, Select Attributes for NHDPlus Version 2.1 Reach Catchments and Modified Network Routed Upstream Watersheds for the Conterminous United States (ver. 3.0, January 2021): U.S. Geological Survey data release, https://doi.org/10.5066/F7765D7V. https://www.sciencebase.gov/catalog/item/5669a79ee4b08895842a1d47
- [3]Moore, R.B., Johnston, C.M., and Hayes, L., 2019, Crosswalk Table Between NHDPlus V2.1 and its Accompanying WBD Snapshot of 12-Digit Hydrologic Units: U.S. Geological Survey data release, https://doi.org/10.5066/P9CFXHGT. https://www.sciencebase.gov/catalog/item/5c86a747e4b09388244b3da1

Create local GIS river network

- Subset NHDPlus v2.1 to area of interest and save as smaller geodatabase for future use

  - subsetNHD()
    - Takes any HUC combination: HUC02 → HUC12
    - Single or in combination

  - Run once
    - Repeat analysis

  - Avoids RAM limitations via SQL queries

Landscape queries

17.0 GB
134 folders
772 files
.zip

1. **Best Management Practices -** characteristics such as agricultural management practices and land in conservation practices.
2. **Chemical -** characteristics such as nitrogen application or toxicity weighted use.
3. **Climate and Water Balance Model -** characteristics such as model outputs of runoff, actual evapotranspiration or ground water storage.
4. **Climate -**characteristics such as mean precipitation, temperature, relative humidity, or evapotranspiration.
5. **Geology -** characteristics such as Hunt or Soller surficial geologies.
6. **Hydrologic -** characteristics such as base flow or infiltration excess overland flow. Hydrologic Modifications, characteristics such as dam storage or tile drains.
7. **Hydrologic Modifications -** characteristics such as dam storage or tile drains.
8. **Landscape -** characteristics such as NLCD, CDL or NWALT.
9. **Population Infrastructure -** characteristics such as population, housing, and road densities.
10. **Regions -** characteristics such as EcoRegions, Physiography or Hydrologic Landscapes.
11. **Soils -** characteristics such as STATSGO, soil salinity, and soil restrictive layer.
12. **Topographic Characteristics -** characteristics such as basin area, slope and elevation.
13. **Water use -** characteristics such as estimated freshwater withdrawls and estimated freshwater consumption by thermo-electric power plants

Wieczorek, M.E., Jackson, S.E., and Schwarz, G.E., 2018, Select Attributes for NHDPlus Version 2.1 Reach Catchments and Modified Network Routed Upstream Watersheds for the Conterminous United States (ver. 3.0, January 2021): U.S. Geological Survey data release, https://doi.org/10.5066/F7765D7V.
https://www.sciencebase.gov/catalog/item/5669a79ee4b08895842a1d47

Landscape queries

what_attributes()

- Best_Management
- Chemical_Attributes
- Climate_Attributes
- Climate_Water
- Geologic_Attributes
- Hydrologic_Attributes
- Hydrologic_Modification
- Land_Cover
- Population_Infrastructure
- Regional_Attributes
- Soil_Attributes
- Topgraphic_Attributes
- Water_Use

Landscape queries

what_attributes(directory = “Climate_Water”)

Climate_Water

- Annual_Average_Actual_Evapotranspiration_millimeters_from_2014__2015
- Annual_Average_Potential_Evapotranspiration_millimeters_from_2014__2015
- Annual_Average_Precipitation_millimeters_from_1945_2015
- Annual_Average_Runoff_millimeters_from_1945__2015
- Annual_Average_Temperature_Celsius_from_1945__2015
- an_Annual_Water_Balance_Variables_for_the_Period_of_Record_2000_2014_and_Detrended_for_the_Year_2012
- Average_Annual_Water_Balance_Variables_over_the_Period_of_Record_2000_2014
- Average_Monthly_Runoff_mm_from_McCabe_and_Wolock_s_Runoff_Model_Over_Period_of_Record_1951_2000
- Monthly_Average_Annual_Runoff_mm_from_McCabe_and_Wolock_s_Runoff_Model_1945__2015
- Monthly_Average_Precipitation_millimeters_from_1945__2015
- Monthly_Average_Temperature_Celsius_from_1945__2015

Landscape queries

what_attributes(directory = “Geologic_Attributes”)

Geologic_Attributes

- Attributes_for_NHDPlus_Version_2_1_Flowlines_for_the_Conterminous_United_States_Lithology
- Bedrock_Permeability_Classes
- Combination_of_landuse_and_geology_
- Generalized_Geology_Type_Reed_and_Bush_2001_
- Hunt_Geology_1999
- Principal_Aquifers_and_Rock_Types
- Select_Geochemical_Characteristics_Based_on_Olson_Geology_Types
- Soller_Surficial_Materials_2009

Landscape queries

what_attributes(directory = “Hydrologic_Modification_Attributes”)

Hydrologic_Modification

- Attributes_for_Surface_Water_Impoundments
- Estimated_Area_of_Agricultural_Land_Drained_by_Field_Ditches_1992
- Estimated_Area_of_Subsurface_Drainage_on_Agricultural_Land_1992
- Estimates_of_Subsurface_Tile_Drainage_Extent_for_the_Conterminous_United_States_Early_1990s
- Major_Sites_of_the_National_Pollutant_Discharge_Elimination_System_NPDES_
- National_Inventory_of_Dams_NID_Storage_and_Construction_by_Decade_1930_to_2010
- Percent_of_Irrigated_Agriculture_2002_MIrAD_data
- Percent_of_Irrigated_Agriculture_2007_MIrAD_data
- Percent_of_Irrigated_Agriculture_2012_MIrAD_data

Landscape queries

what_attributes(directory = “Land_Cover”)

Land_Cover

- 2012_Conservation_Reserve_Program_land_allocated_to_2012_NWALT_landuse
- Crop_Land_Data_Layer_2012
- Crop_Land_Data_Layer_2013
- Crop_Land_Data_Layer_2014
- he_conterminous_United_States_seasonal_enhanced_vegetation_index_fall_of_2011_through_summer_of_2012
- Modeled_historical_land_use_and_land_cover_for_the_conterminous_United_States_1992_2002
- National_Land_Cover_Database_2001_NLCD_2001_
- National_Land_Cover_Database_2006_NLCD_2006_
- National_Land_Cover_Database_2011_NLCD_2011_
- National_Land_Cover_Database_2016_Versions_for_the_Years_2001_2004_2006_2008_2011_2013_and_2016
- NAWQA_Wall_to_Wall_Anthropogenic_Land_Use_Trends_NWALT_Timber_1999_2012
- NLCD_2001_Percent_Imperviousness
- NLCD_2001_Percent_Imperviousness_in_100_Meter_Riparian_Buffer
- NLCD_2006_Percent_Imperviousness
- NLCD_2006_Percent_Imperviousness_in_100_Meter_Riparian_Buffer
- NLCD_2011_Percent_Imperviousness
- Percent_NLCD_2011_in_50_Meter_Riparian_Buffer
- Percent_NLCD_2011_Tree_Canopy
- Percent_of_Watershed_Covered_by_Waterbodies_from_NHD_High_Resolution
- Wildfire_2000_2012

Landscape queries

what_attributes(Subdirectory: National_Land_Cover_Database_2016_re-release)

- NHDV2_NLCD2016s_Revised.xml
- NLCD01_ACC_CONUS.zip
- NLCD01_CAT_CONUS.zip
- NLCD01_TOT_CONUS.zip
- NLCD04_ACC_CONUS.zip
- NLCD04_CAT_CONUS.zip
- NLCD04_TOT_CONUS.zip
- NLCD06_ACC_CONUS.zip
- NLCD06_CAT_CONUS.zip
- NLCD06_TOT_CONUS.zip
- NLCD08_ACC_CONUS.zip
- NLCD08_CAT_CONUS.zip
- NLCD08_TOT_CONUS.zip
- NLCD11_ACC_CONUS.zip
- NLCD11_CAT_CONUS.zip
- NLCD11_TOT_CONUS.zip
- NLCD13_ACC_CONUS.zip
- NLCD13_CAT_CONUS.zip
- NLCD13_TOT_CONUS.zip
- NLCD16_ACC_CONUS.zip
- NLCD16_CAT_CONUS.zip
- NLCD16_TOT_CONUS.zip
- Version_History.txt

Batch load directly from zips

load_attributes(Subdirectory: National_Land_Cover_Database_2016_re-release)

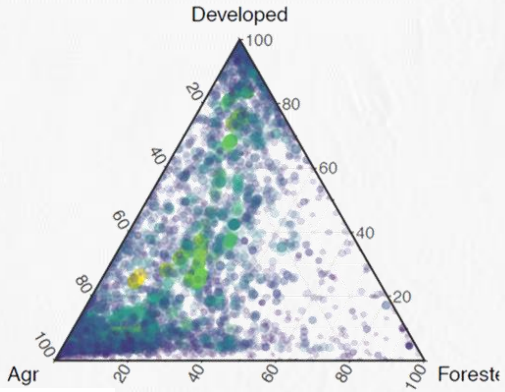
- Batch import of subdirectory of attributes
- Reads zip files directly
- Fast
  - Benchmark: 2 min for ~1billion NLCD values (2.8 million records x 133 variables)
- Delivers subset table for Area of Interest
- Saves CSV to disk for future easy reference

Gage queries

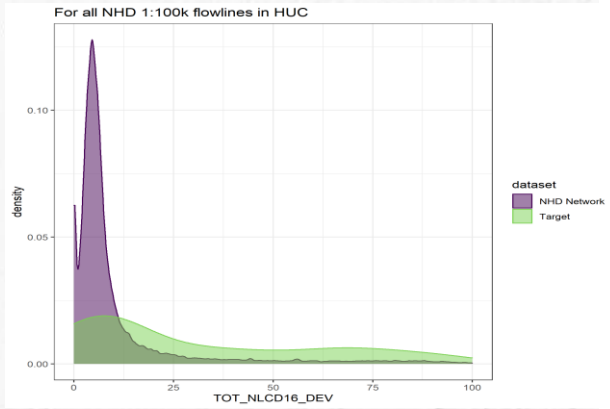
GetNWISGages()

- HUCs: single or combination
- Parameters: USGS code (e.g. 00010: Water Temperature)
- Data type: Real-time or discrete
- Active: Currently active or historical

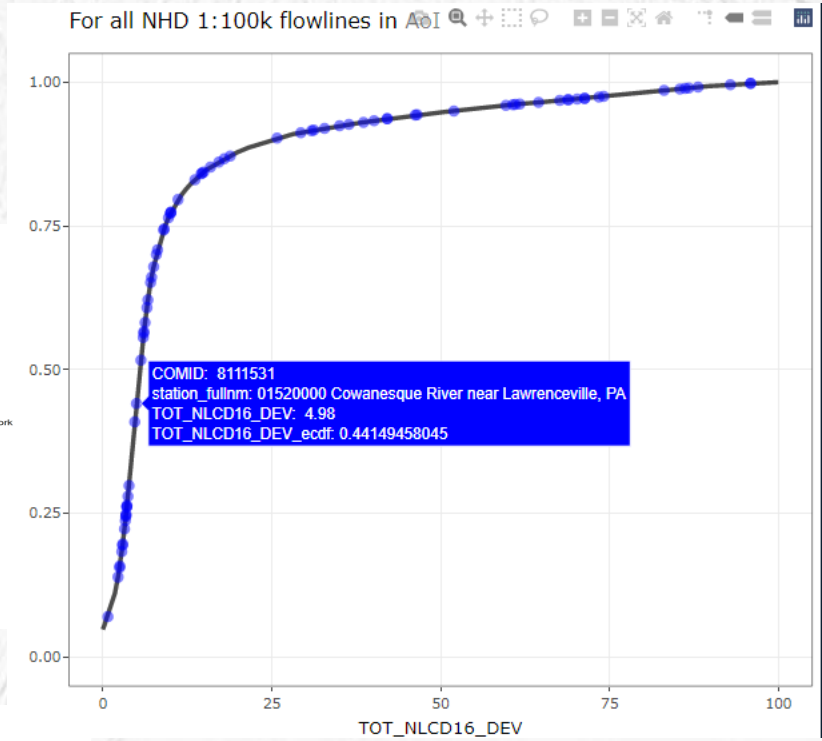
Static & Interactive Plotting

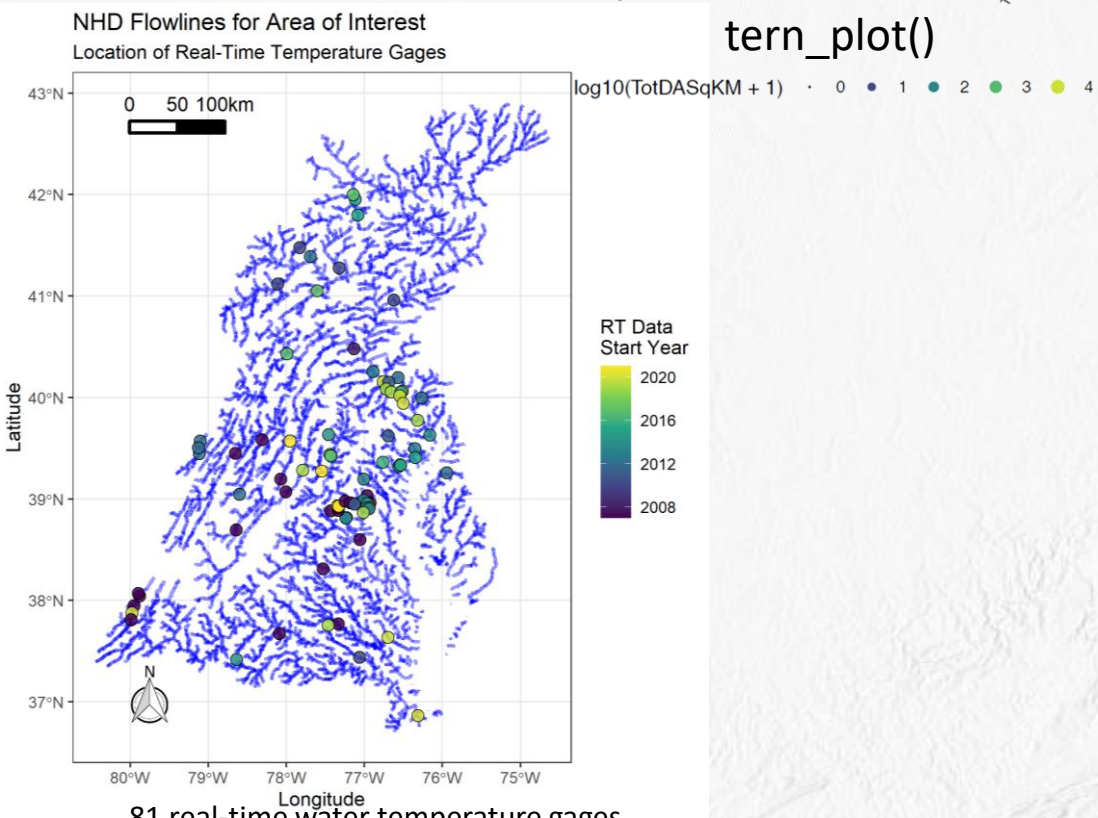


tern_plot()

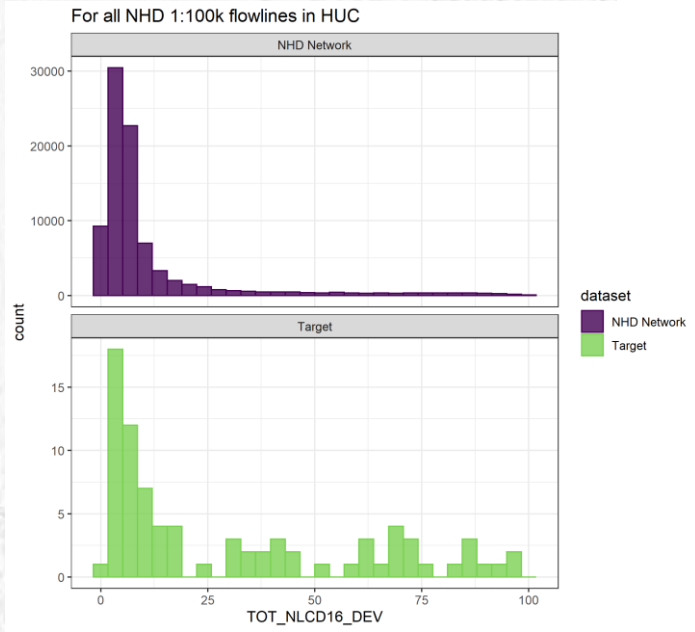


Density_analysis()



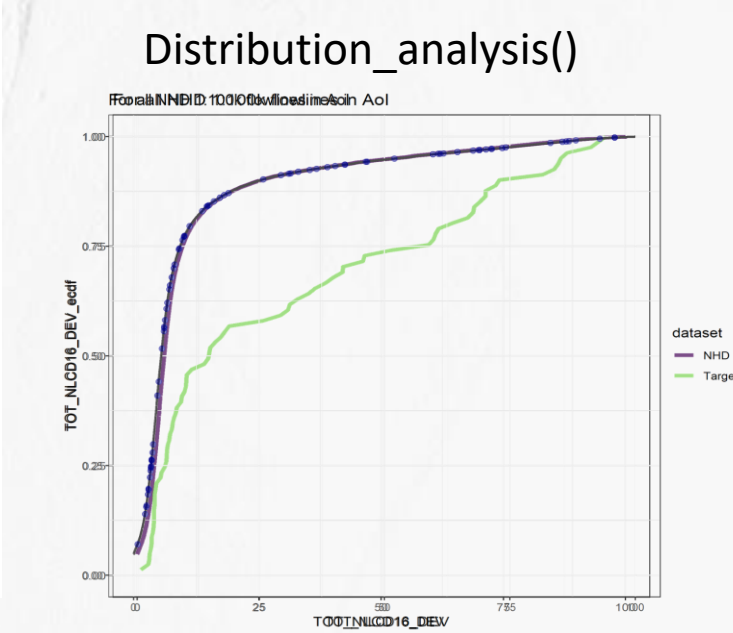


81 real-time water temperature gages



Histogram_analysis()

Distribution_analysis()



Data are provisional and subject to revision

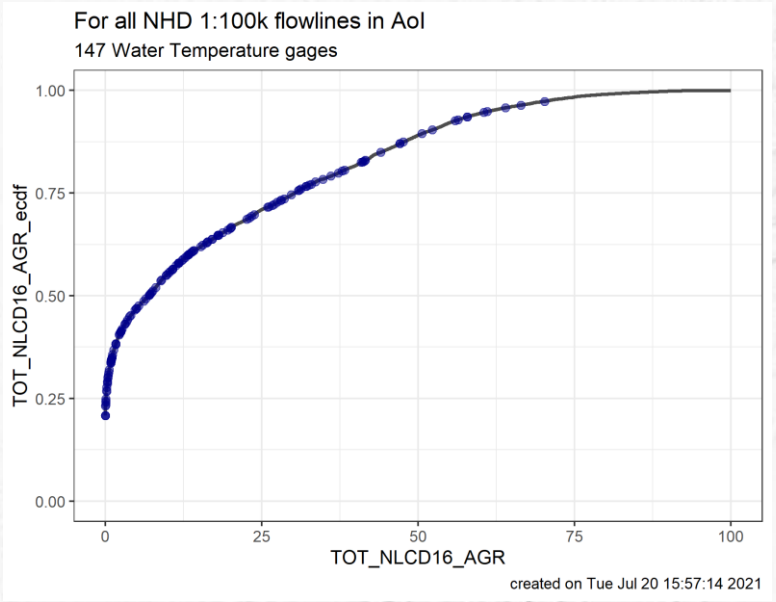
Continued analysis

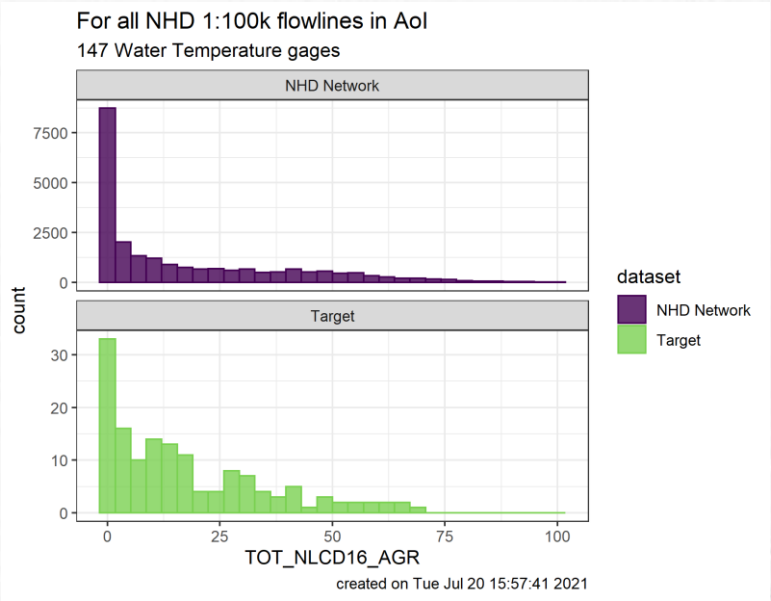
- Data are all in R for further or custom analysis
  - Spatial data in *sf* (simple features) format

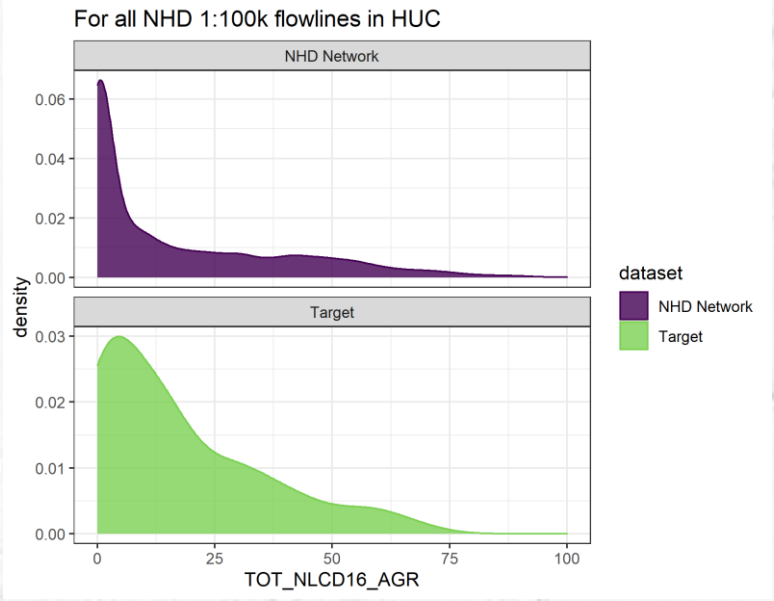
- Export tables to CSV
- Also easy export to ESRI/Arc friendly formats

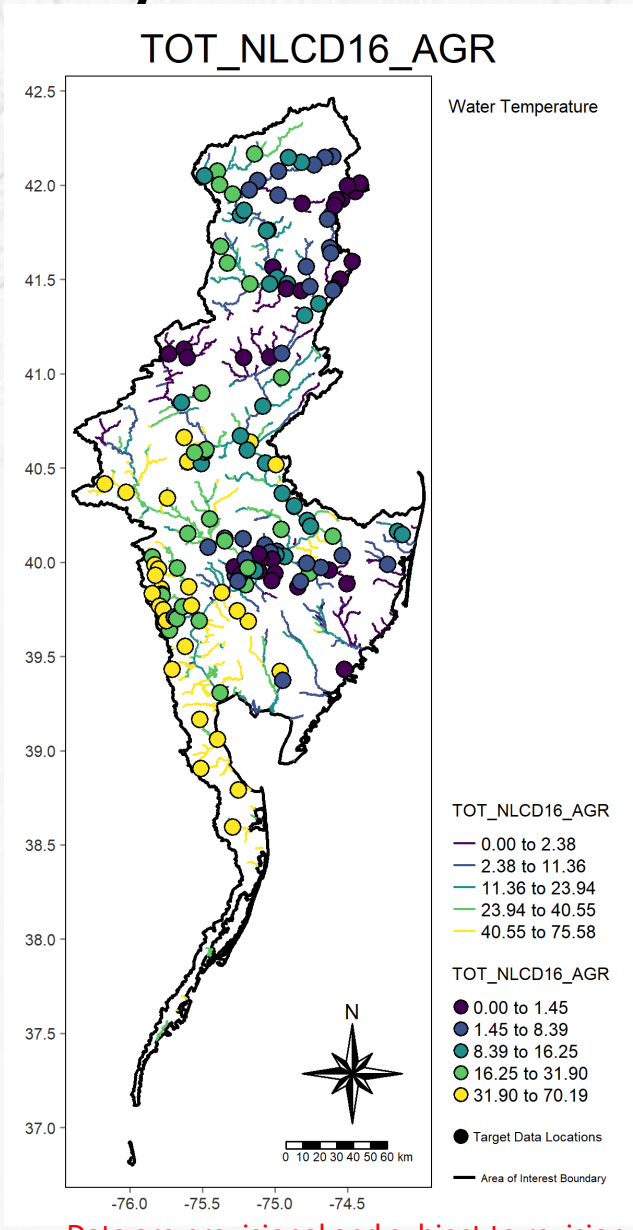
Use Case: Next Generation Water Observation System

- Delaware River Basin
- Objective: Temperature
  - Assess before/after 2019 gage additions







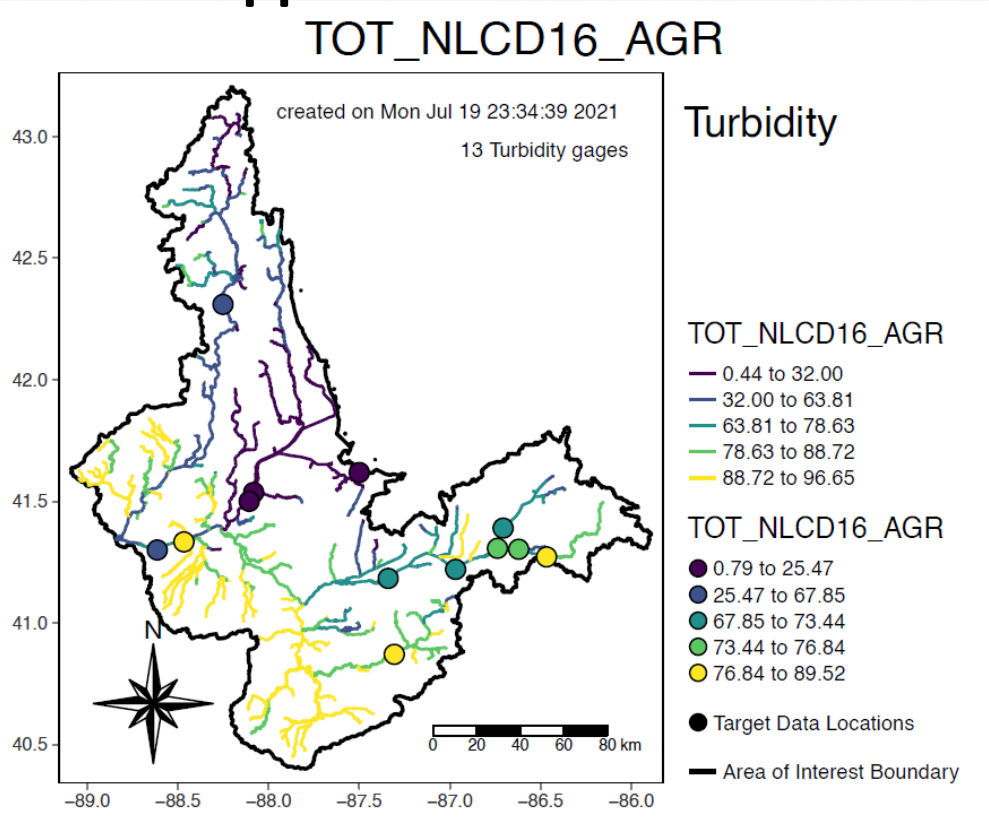


Data are provisional and subject to revision

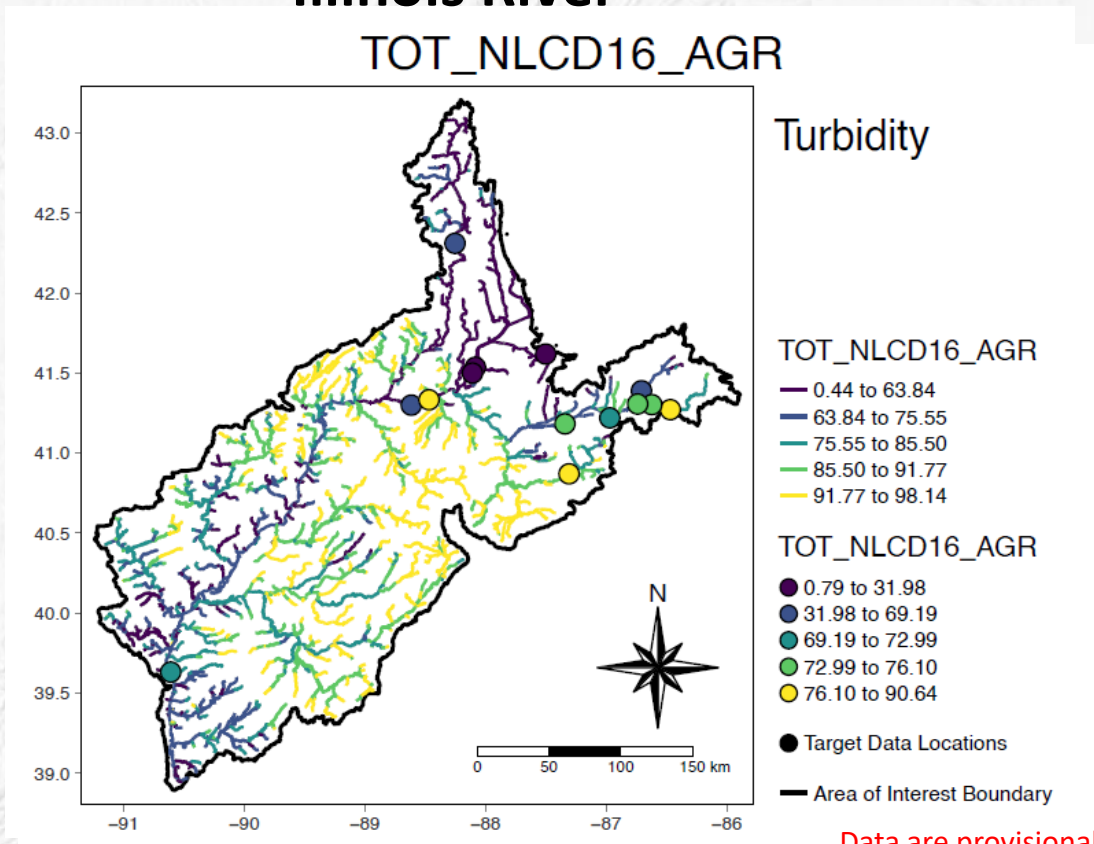
Use Case: Next Generation Water Observation System

- Upper Illinois River
- Objective: Temp, SpC, DO, pH, Turbidity

Upper Illinois River



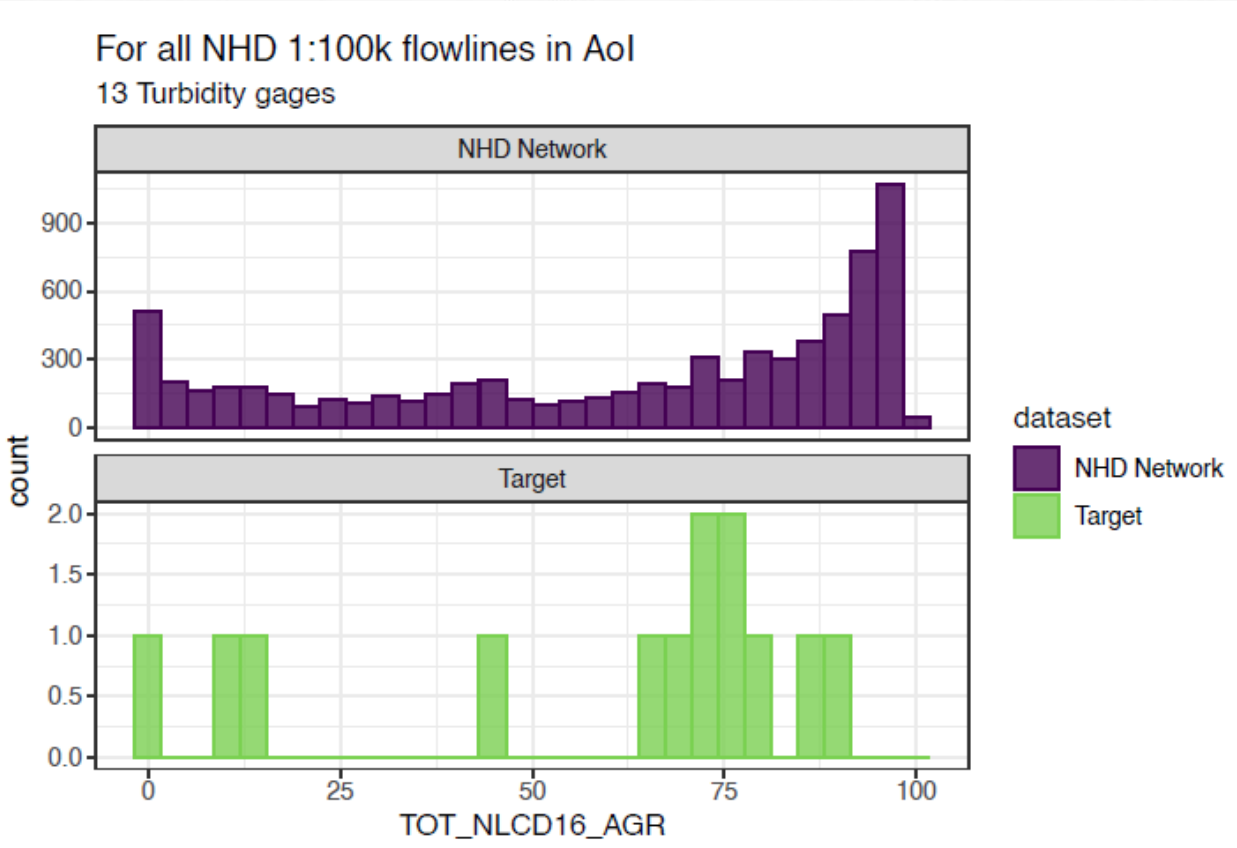
Illinois River



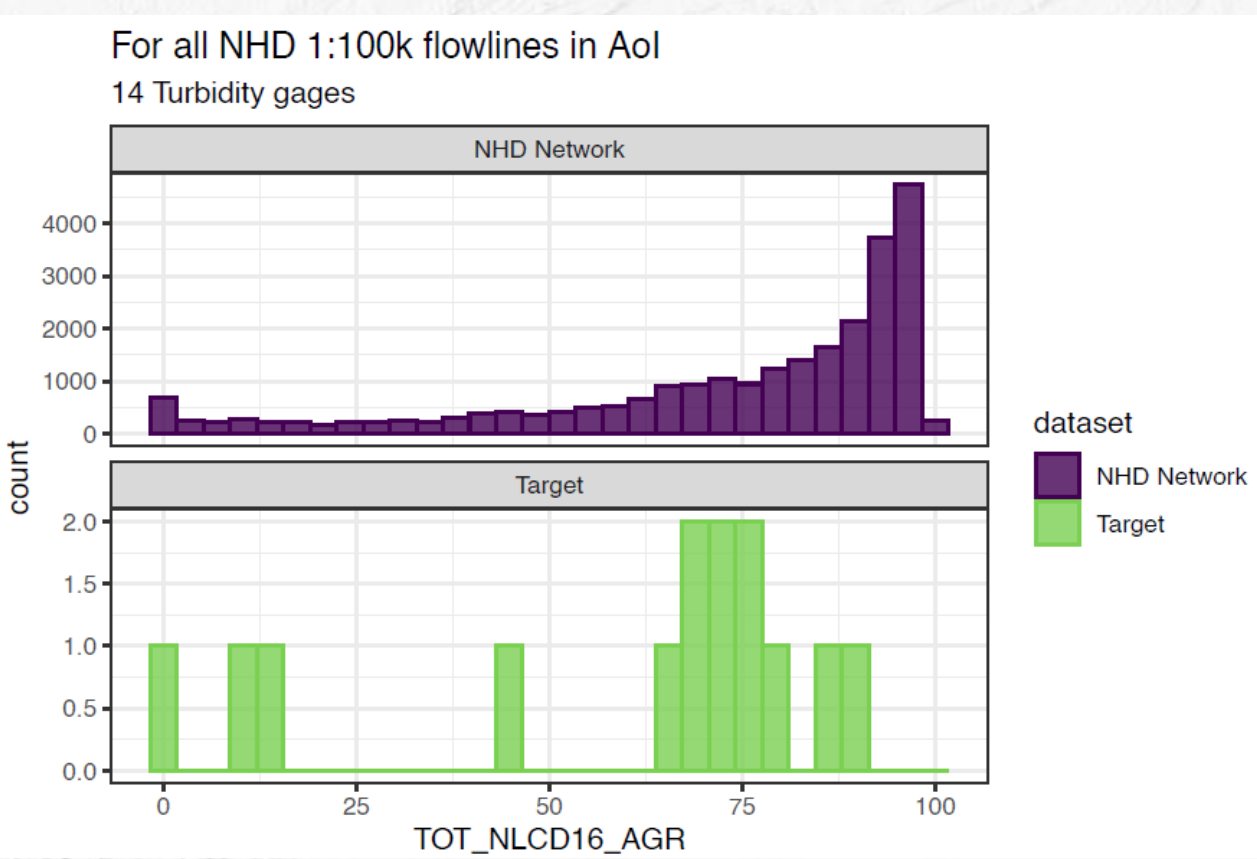
Data are provisional and subject to revision

Use Case: Next Generation Water Observation System

Upper Illinois River



Illinois River



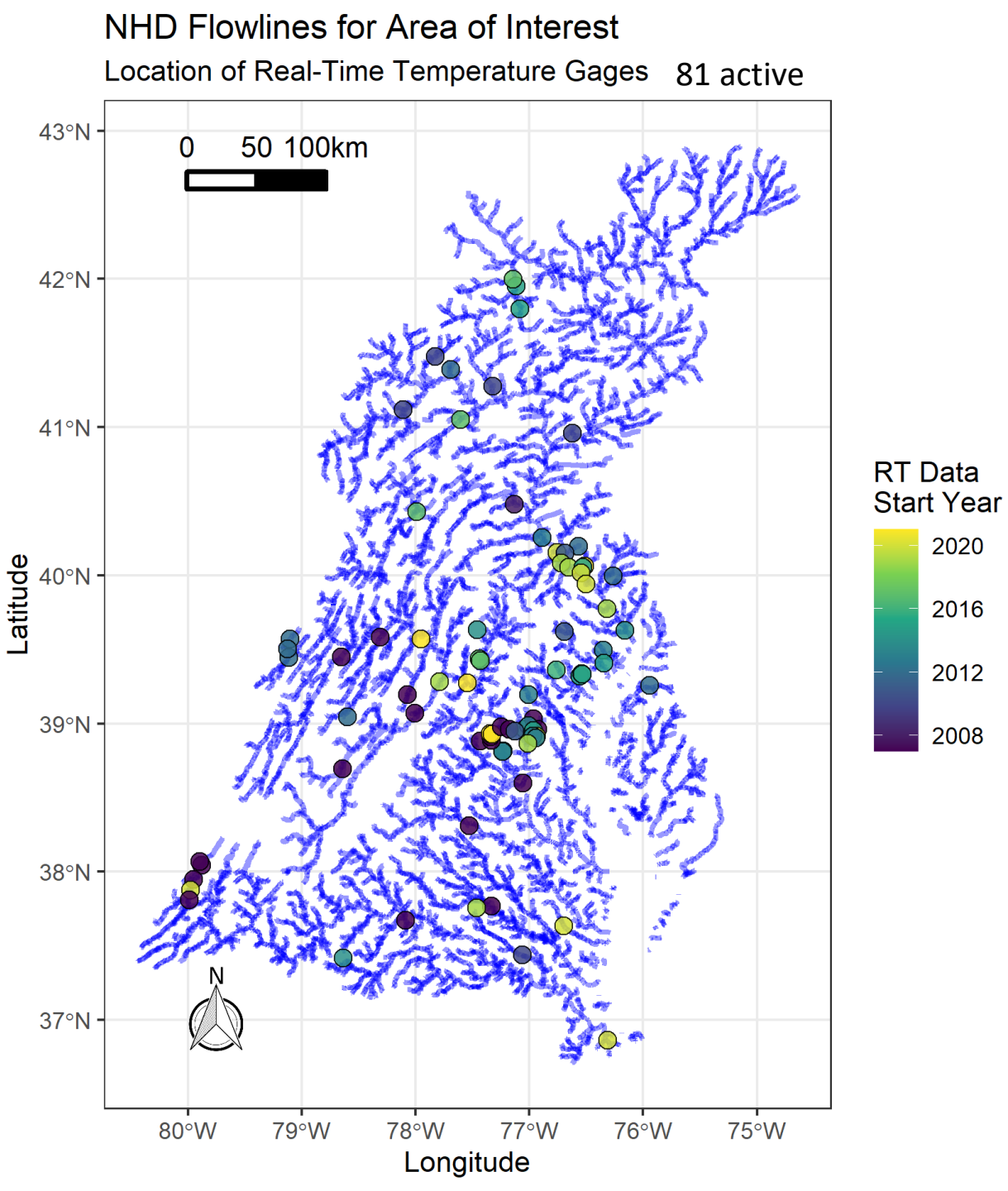
Making analyses actionable

- These analyses so far have been largely qualitative:
  - Different graphical ways to compare networks and variable distributions

- How do you use these for decision-making?
  1. Gap identification and filling (network enhancements - adding)
  2. Quantifying change in bias (network reductions - subtractions)

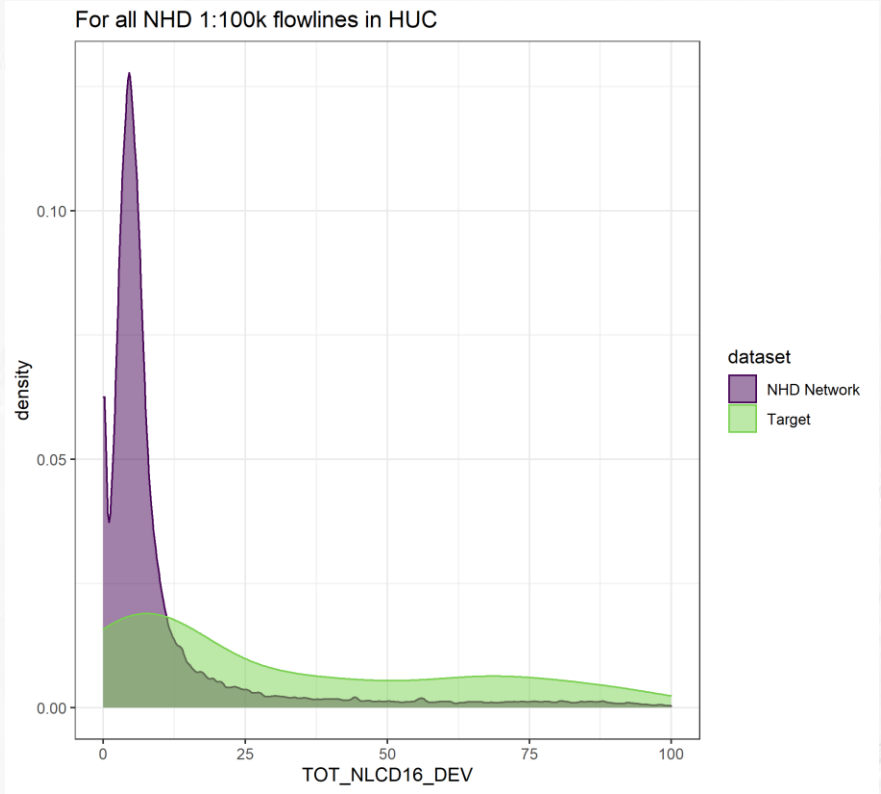
Chesapeake Bay Demo:

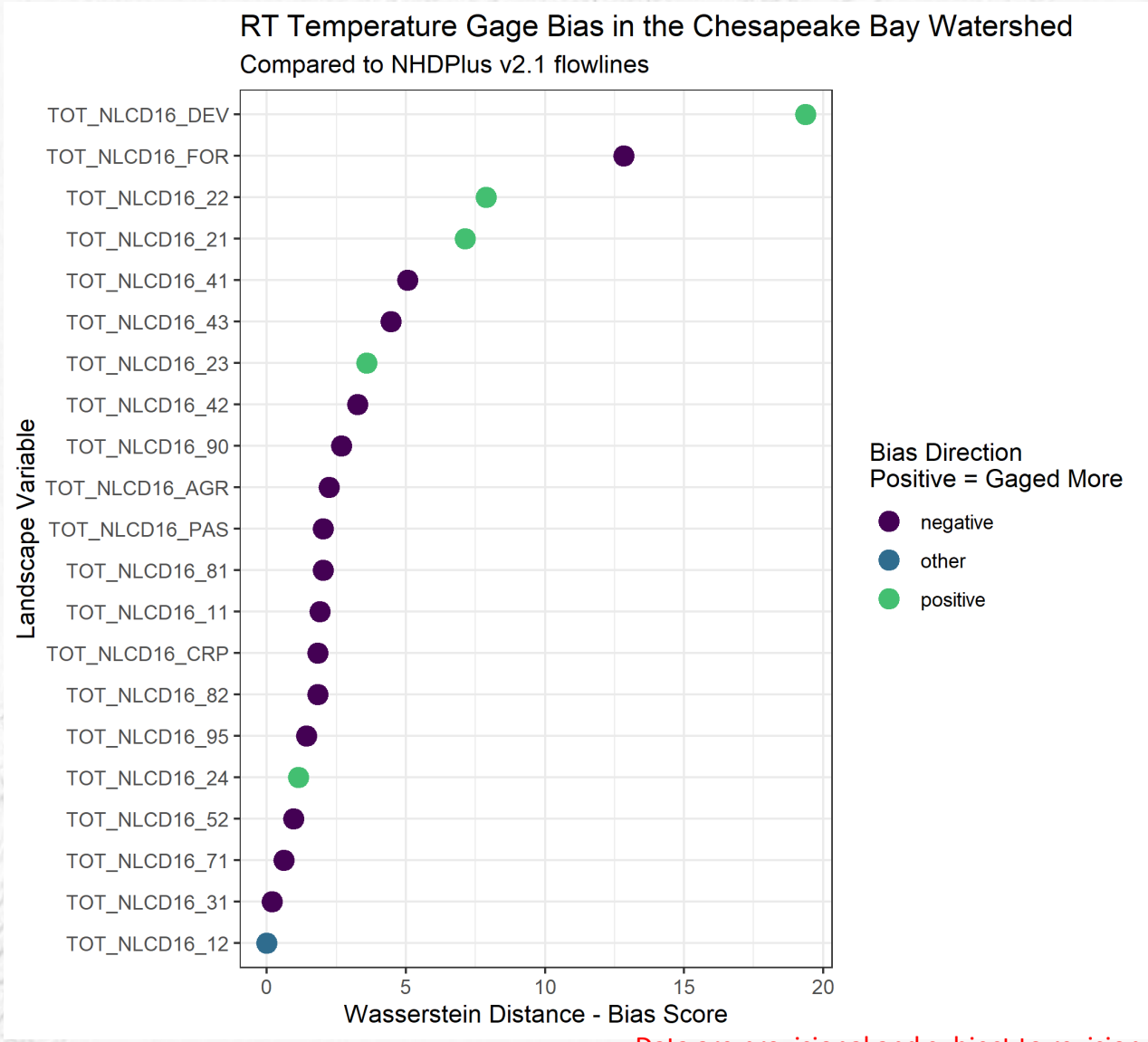
- Queried active, RT temperature gages (81)
- Want to examine ALL NLCD2016 land-uses
- But, want a “quick” number to show which variables are most/least representative!
  - “Bias”



Chesapeake Bay Demo: Bias Quantification

- Bias quantified using Wasserstein Distance ('earthmover score') for all NLCD2016 variables
  - ❖ "How much work is needed to "bulldoze" distributions to be the same?"



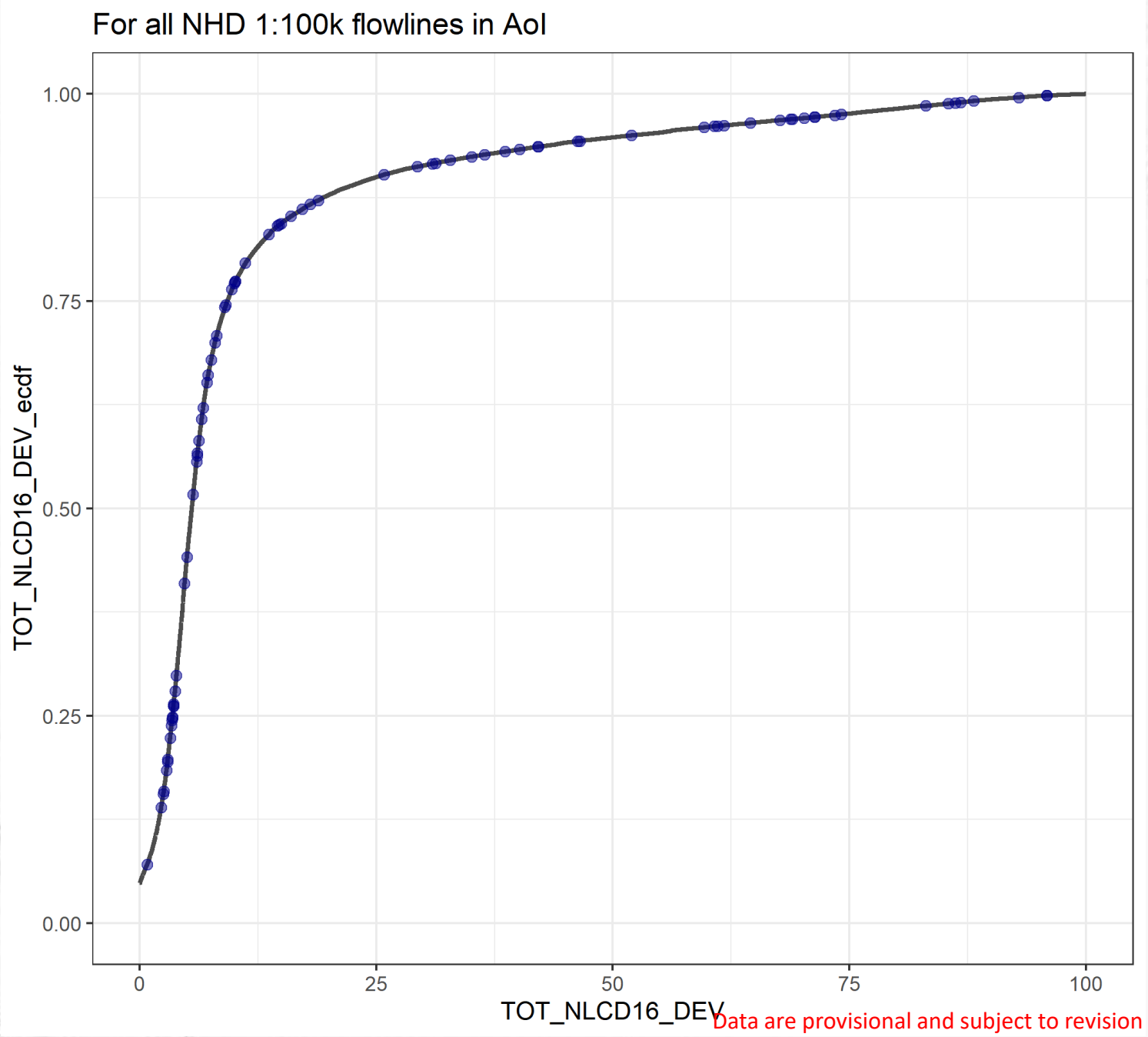


Data are provisional and subject to revision

Gap Filling

- What if we're in the position to add gages?

- Distribution_analysis()
  - Type: POINT



Gap Filling

- Distribution_analysis()
  - Type: POINT
    - Target GAP: 0.3-0.5

- First 100 Candidate COMIDs

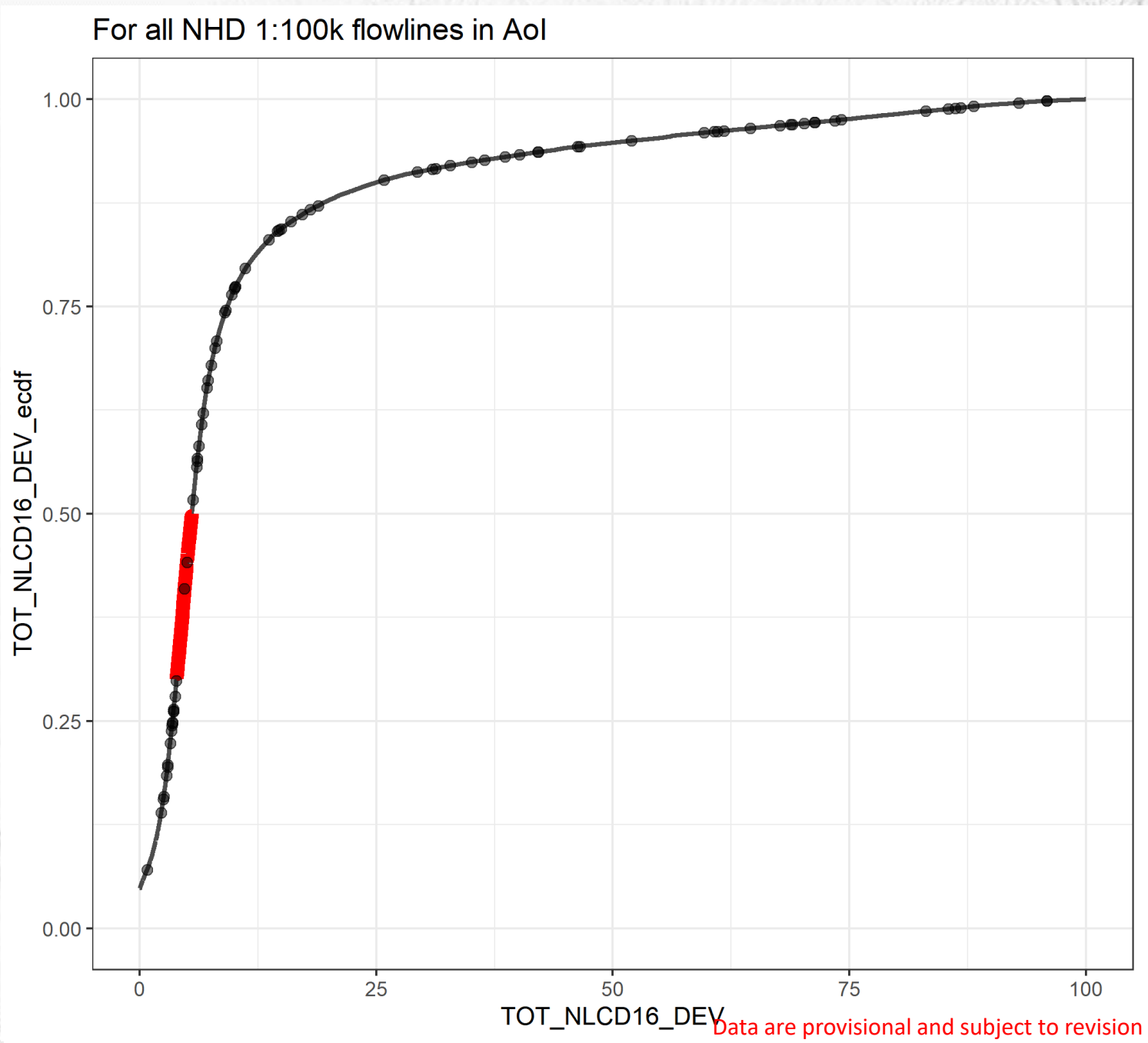
- 4196310 4507182 4673153 4689163 4697207 4697237 4697673 4697675 4699853 4710730 4764944 5890986 5892890 5893356 5893470 8101199 8101975 8102617 8111937 8118019

- 8124969 8127489 8143572 8143584 8143586 8153211 8403479 8431706 8431922 8431928 8436674 8456894 8457374 8477944 8478264 8501386 8507458 8539513 8539919 8566657

- 8612147 9423769 10234695 2602267 4196050 4672055 4673205 4698827 4711598 4763910 8100799 8110763 8112413 8112657 8124821 8126315 8126407 8139372 8144200 8385881

- 8385899 8401523 8420616 8440195 8440839 8448752 8456514 8464839 8468791 8490950 8493354 8493572 8501662 8503150 8526133 8539845 8552283 8574677 8608481 8608793

- 9407856 9420343 14363248 14365368 8523575 8523579 8523595 8523749 8124983 8143622 8151729 8457600 8457616 8479156 8479558 8547253 8609055 9423075 9424555 22340163

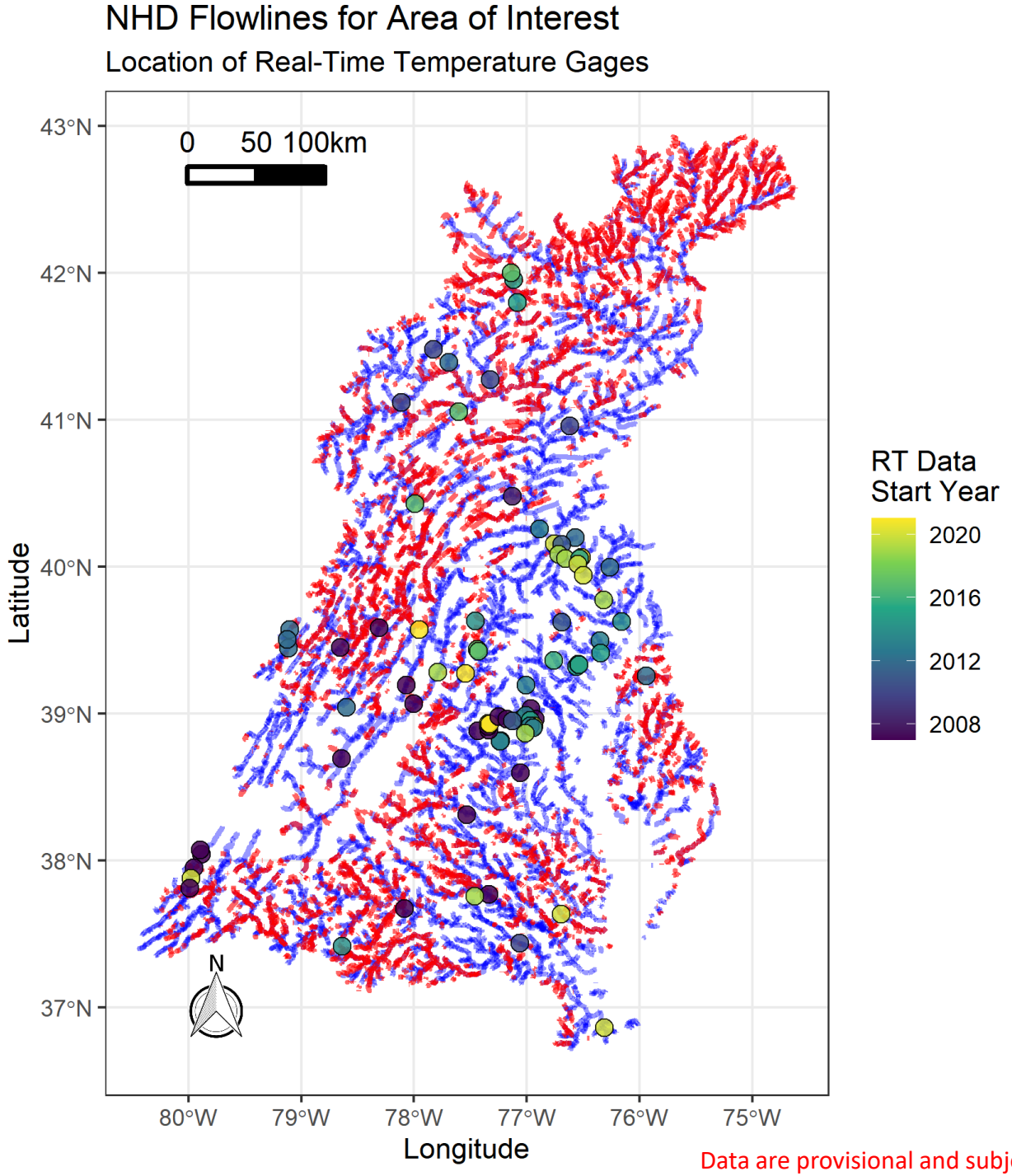


Gap Filling

- Distribution_analysis()
  - Type: POINT
    - Target GAP: 0.3-0.5

- First 100 Candidate COMIDs

- 4196310 4507182 4673153 4689163 4697207 4697237 4697673 4697675 4699853 4710730 4764944 5890986 5892890 5893356 5893470 8101199 8101975 8102617 8111937 8118019
- 8124969 8127489 8143572 8143584 8143586 8153211 8403479 8431706 8431922 8431928 8436674 8456894 8457374 8477944 8478264 8501386 8507458 8539513 8539919 8566657
- 8612147 9423769 10234695 2602267 4196050 4672055 4673205 4698827 4711598 4763910 8100799 8110763 8112413 8112657 8124821 8126315 8126407 8139372 8144200 8385881
- 8385899 8401523 8420616 8440195 8440839 8448752 8456514 8464839 8468791 8490950 8493354 8493572 8501662 8503150 8526133 8539845 8552283 8574677 8608481 8608793
- 9407856 9420343 14363248 14365368 8523575 8523579 8523595 8523749 8124983 8143622 8151729 8457600 8457616 8479156 8479558 8547253 8609055 9423075 9424555 22340163



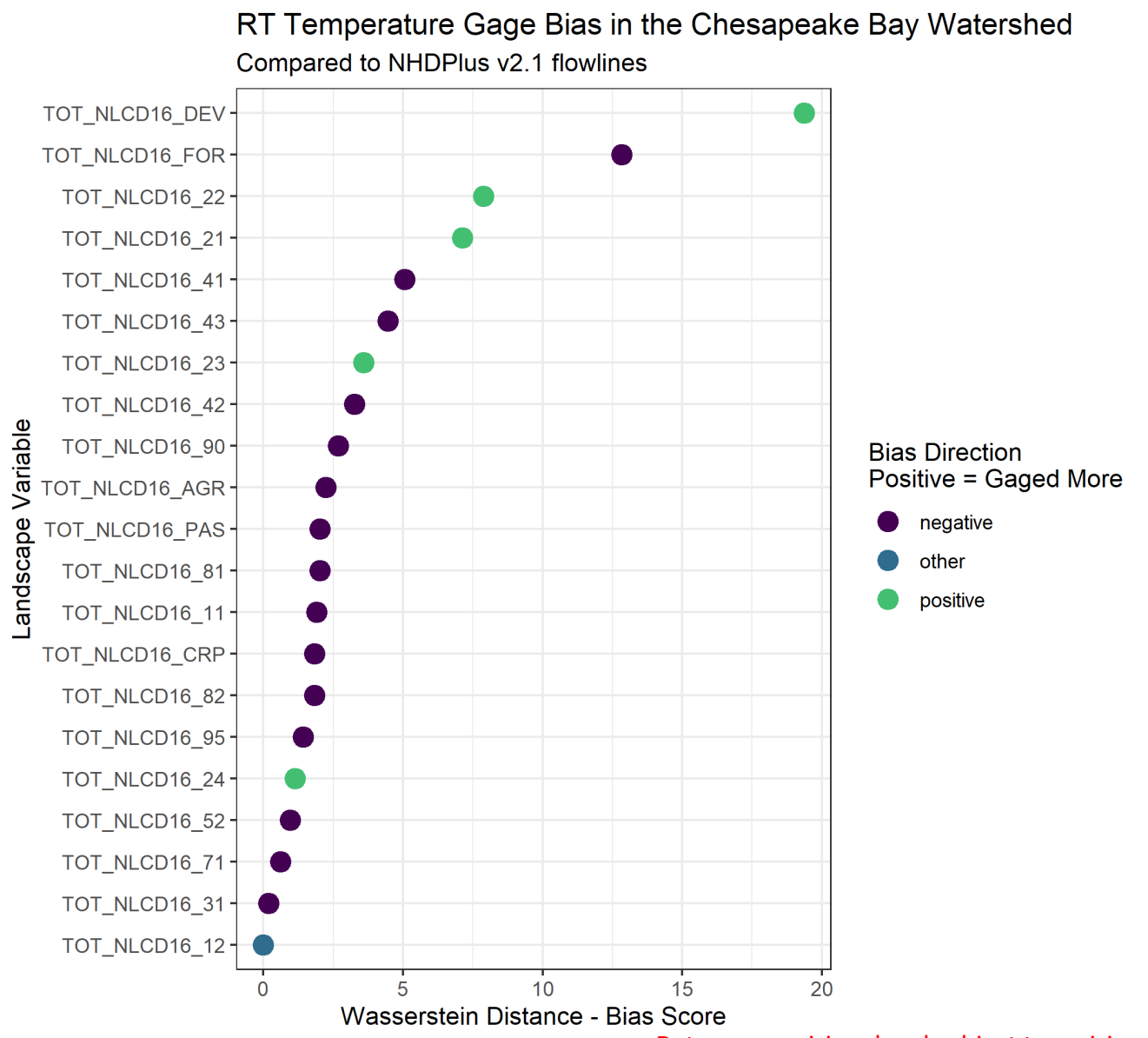
Data are provisional and subject to revision

Gage Removal Analysis

- Minimizing the impacts of monitoring cuts
- Can bias scores evaluate impact of gage removal?

Simulations:

- Remove gage from network → recalculate bias
- Calculate change in bias from baseline
- Iterate for each gage in network

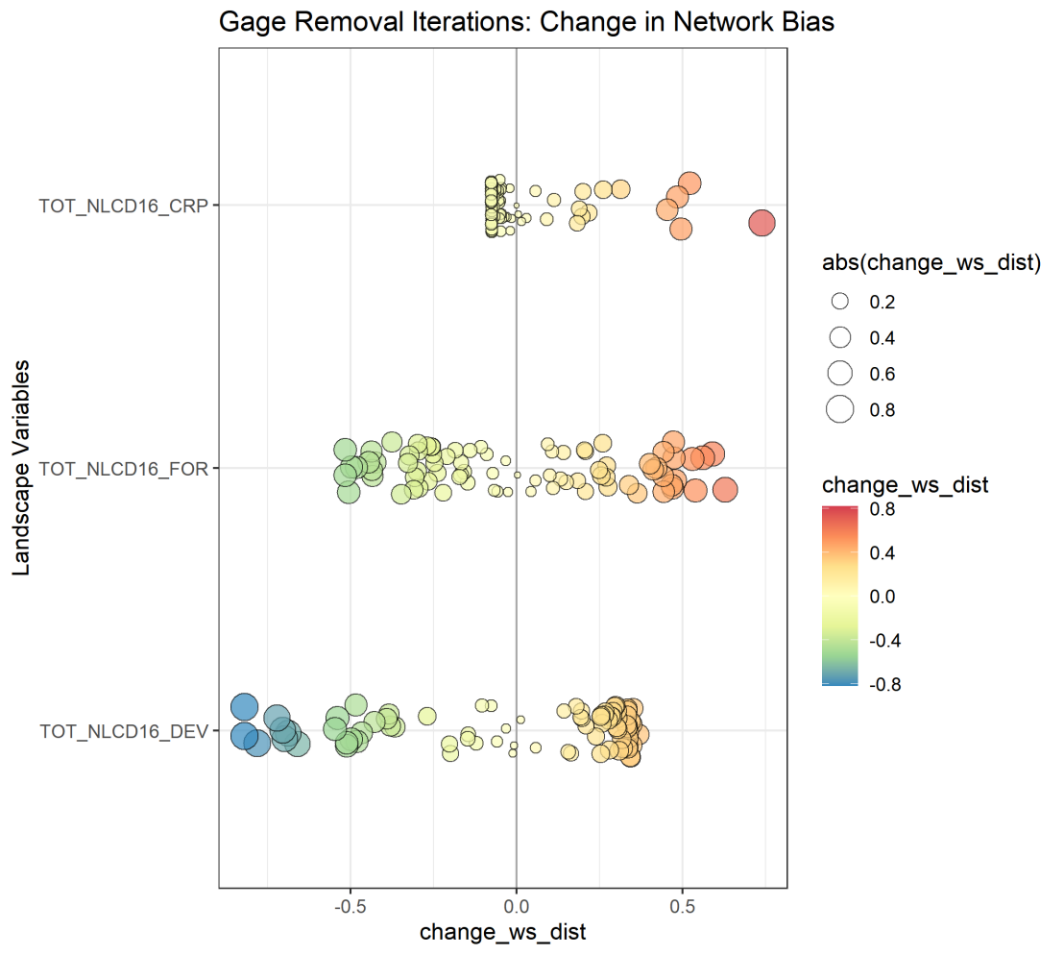


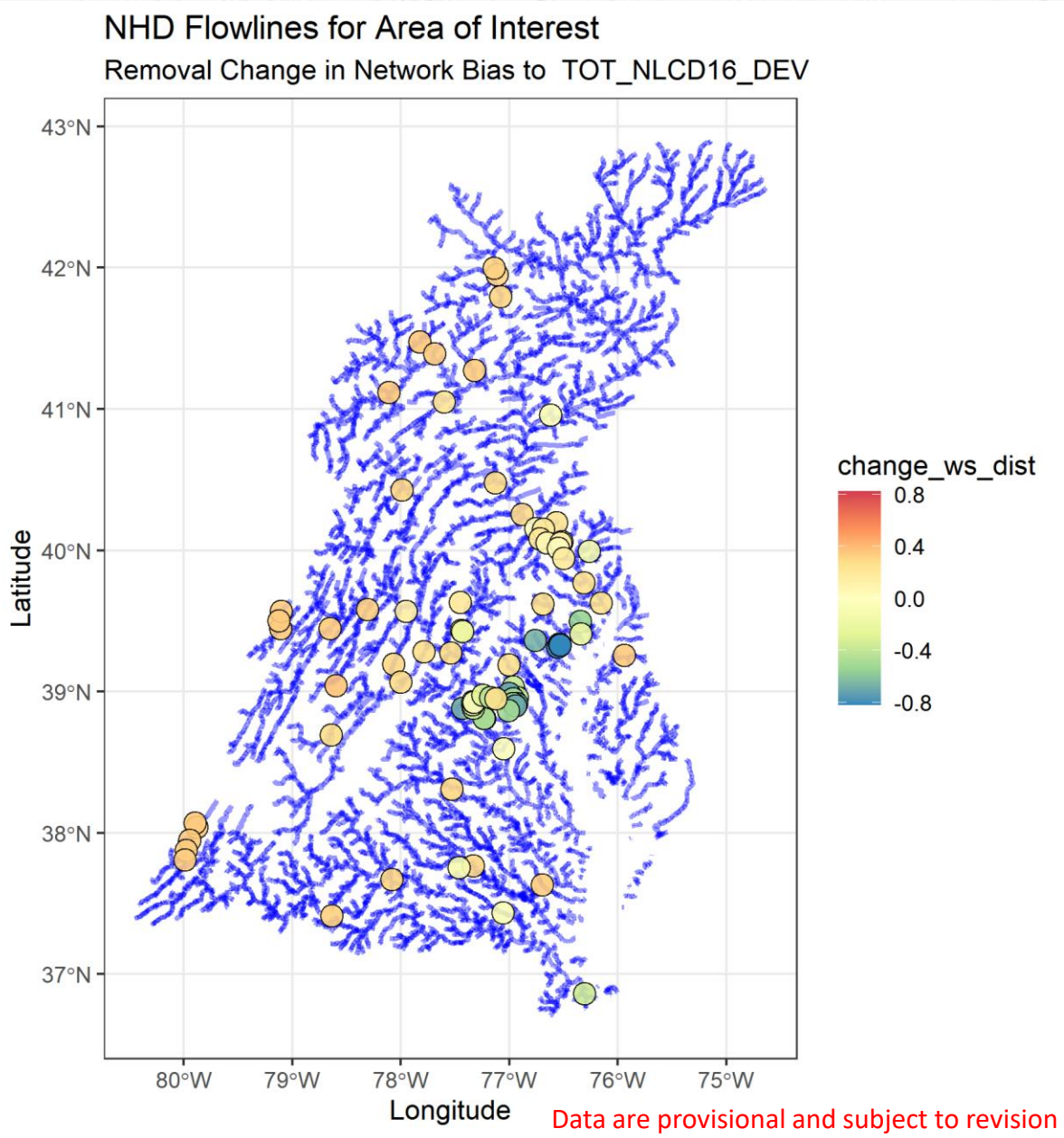
Data are provisional and subject to revision

Gage Removal Analysis

How does variable bias change if it was lost in the network?

Results for every gage:



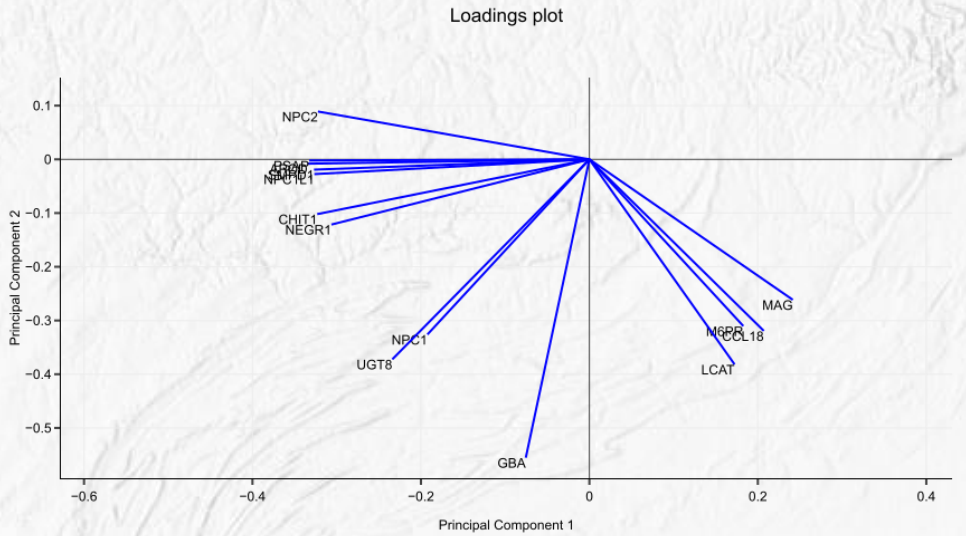


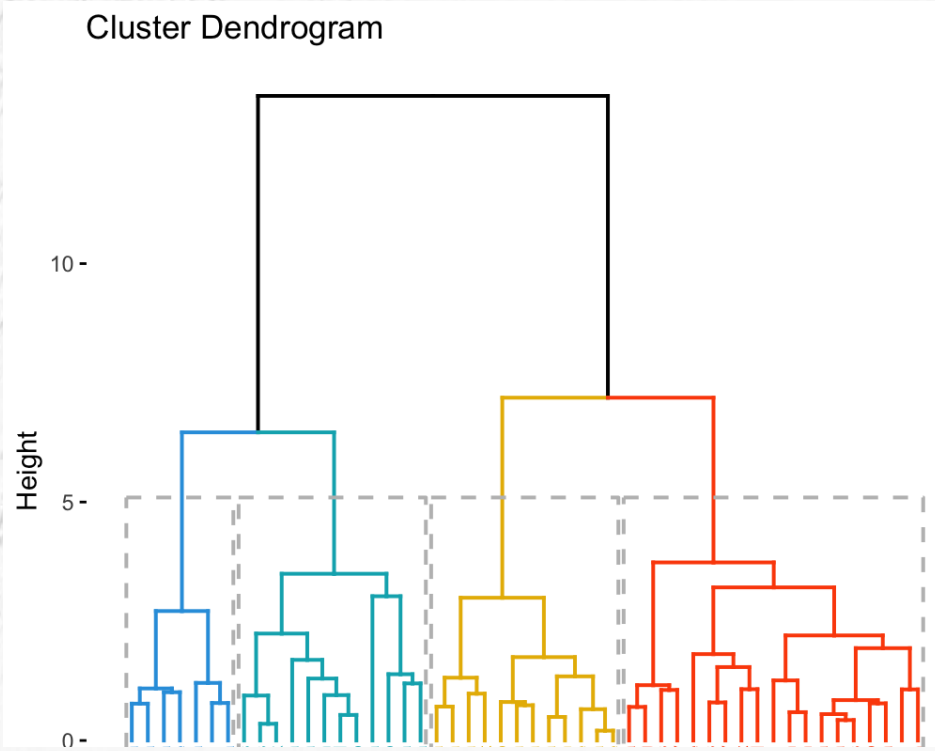
Data are provisional and subject to revision

Next steps

- Scripts are still in development
  - Bugs do exist (but it now works on mac)
  - Happy to have collaborators
  - New features continue to be added
    - ➢ Bias simulations with gage additions

- A way to evaluate multivariate bias
  - Equal weighting of considered variables?
  - Variable ‘reduction’ approach
    - PCA or hierarchical clustering?

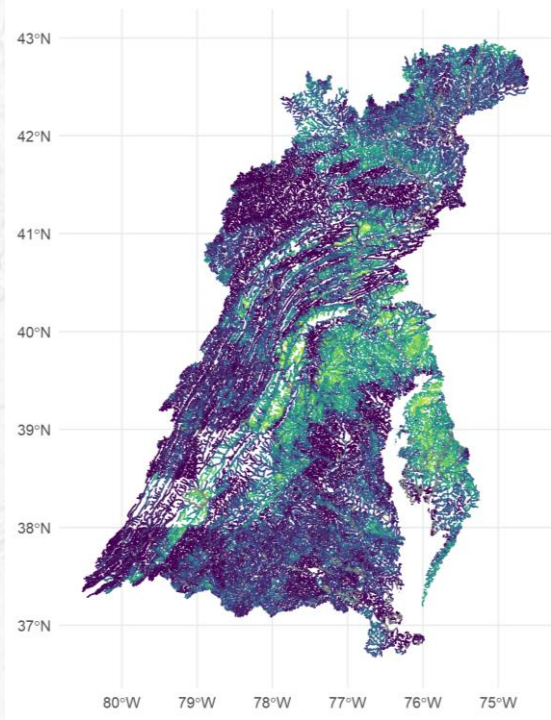


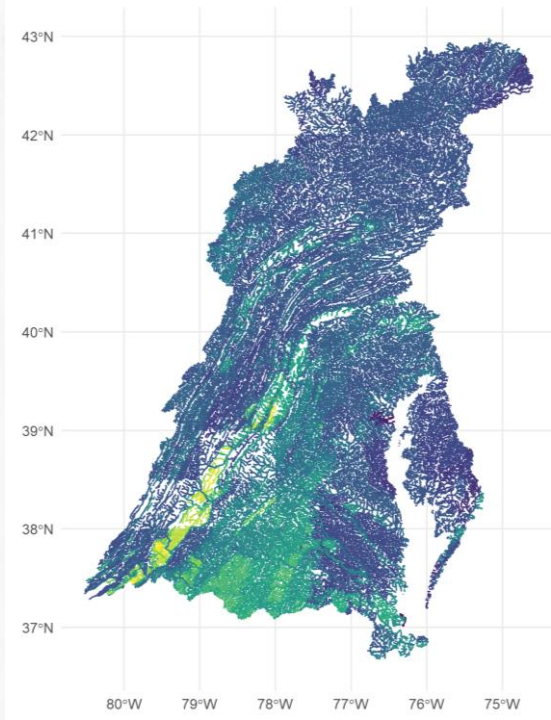


Tools are only tools

- These tools make network analyses “easier”
  - Handle “big” data
  - Avoid GIS pre-processing
  - Increase reproducibility
  - Built-in analysis methods
    - ➢ But always looking for more/improved methods!

- Yet cannot answer the fundamental issue:
  - Is the network representative?
    - *– Representative of “what”?*





What are your objectives and values?

- A network might be representative for one objective or goal, but not another

- With thousands of variables to evaluate, knowing what you are interested in and **why** is more important than ever
  - Decision-point is variable selection, not data availability

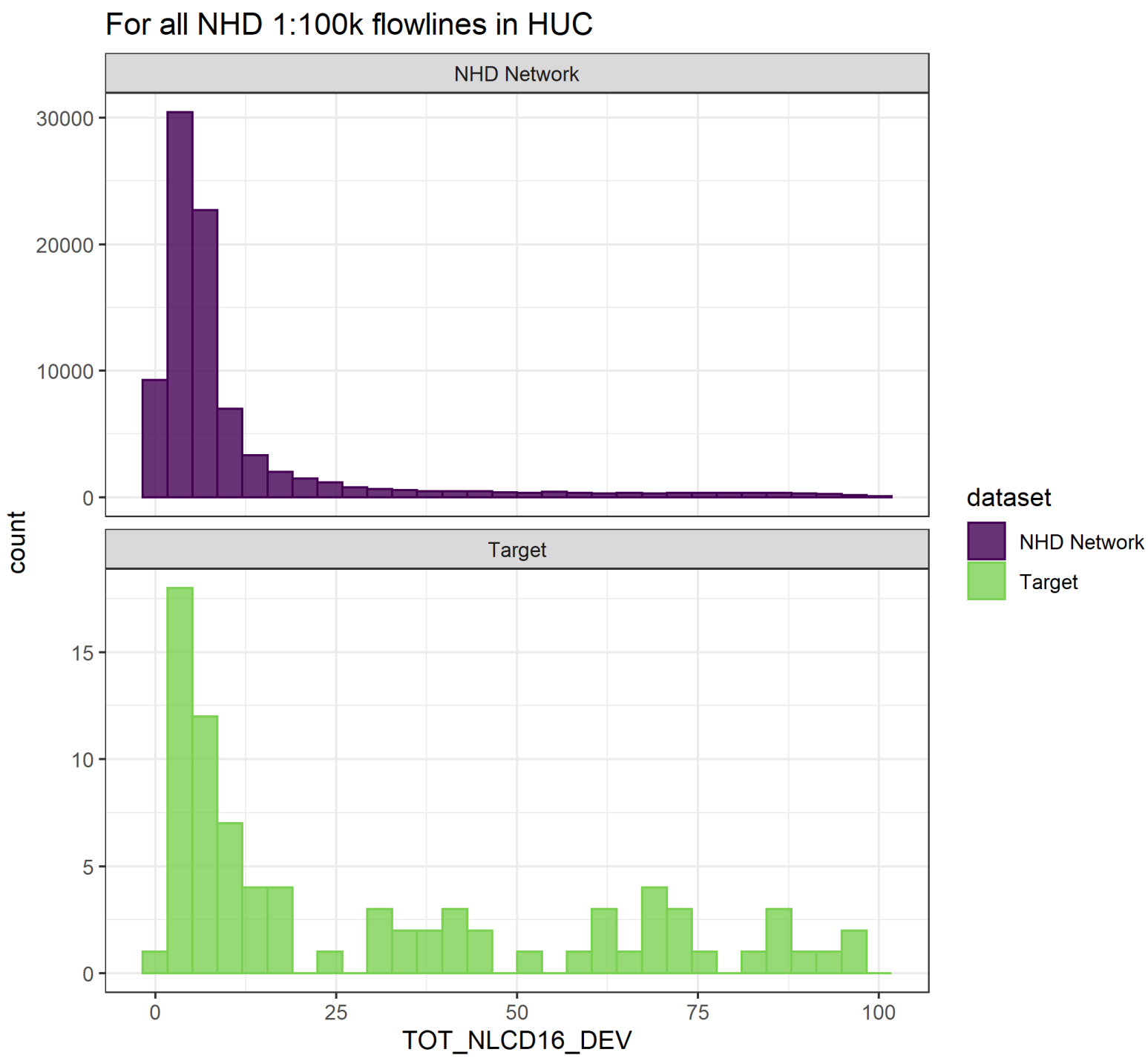
- Need a strong sense of priorities to evaluate inevitable tradeoffs
  - Network changes might decrease bias in one variable (or sensor type) but increase it in another

via Tok Talk

Thank you!

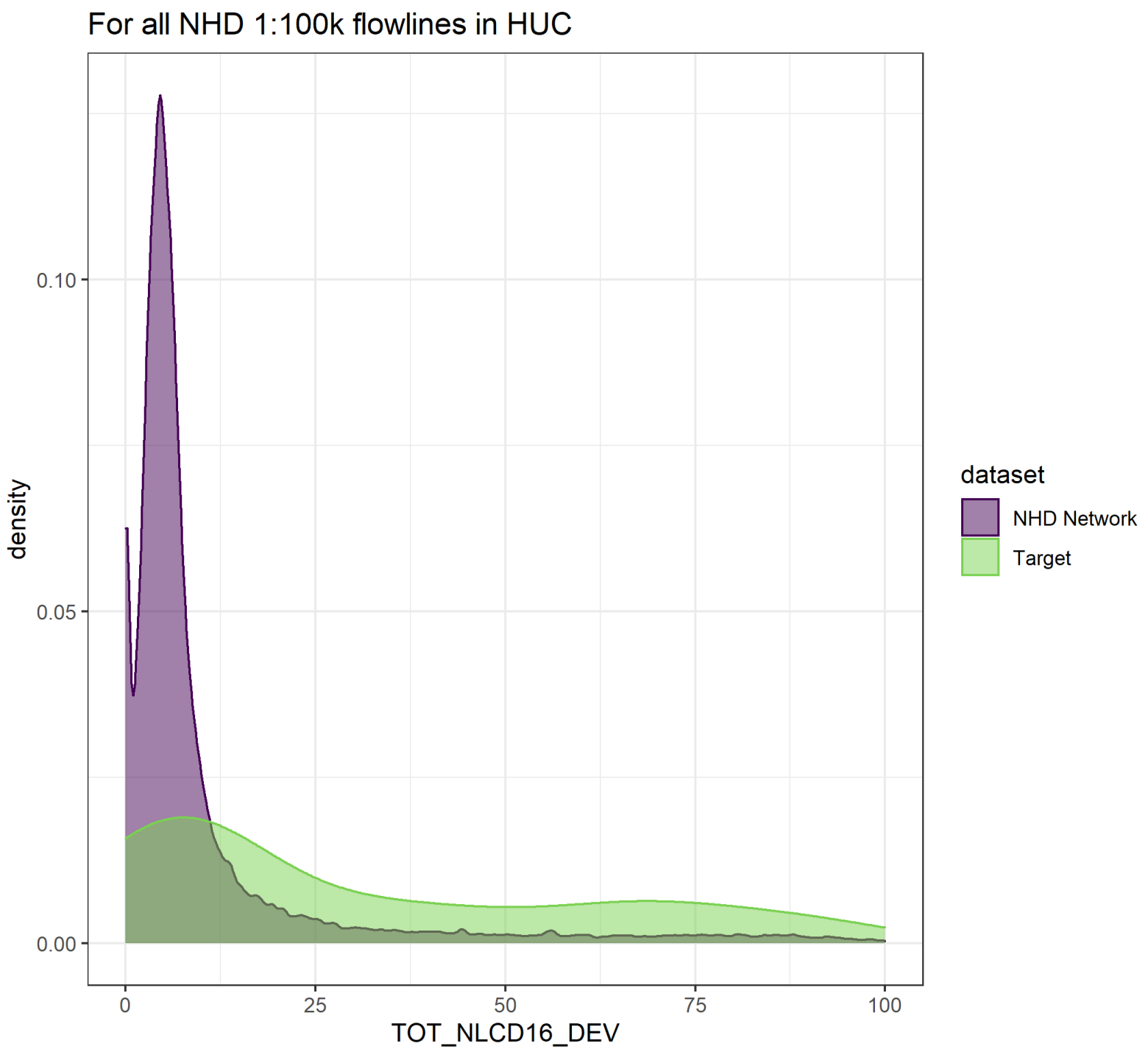
Plotting

- Histogram_analysis()



Plotting

- Density_analysis()



Plotting

- Distribution_analysis()
  - Type: line