**NEW…2024 modeling**

**--------there were significant errors associated with the One Rule ALL table with reference designations**

**--------we decided to scrap the old efforts and start new. This time we :**

1. fixed the reference designations in One Rule All,
2. then requested data from USU
3. uploaded USU data to AWQMS
4. Queried missing data from Biomon\_Phoenix and uploaded to AWQMS
5. Pulled raw bug data from AWQMS to start the process over

* **Selecting REFERENCE samples to build models**
  + AWQMS queries were not working. So Lesley built a query from SQL (?) and created a new project
    - C:\Users\shubler\Oregon\DEQ - Biomonitoring is Fun! - General\Data Management\BioMon\_Dataexplorer\BioMon\_Dataexplorer.Rproj
  + I used this to filter all raw bug data by station table reference designation.
    - Reference sites\_bug samples\_total and OTU abundances.xlsx
    - 881 reference samples
  + Next I designated which samples to drop: pools, early year data with incompatible protocols, and very low abundances, finally selecting a single sample from each possible reference site (n=316).
  + FINAL ref samples: here I took all ref samples (stations) and plotted them in Google Earth. I used site name, GNIS name, and Assessment Unit to find potential spatial duplicates. Where two or more sites feel on the same stream/segment, I chose one of them. If there was a choice between sites with less than 250 or more than 250, I usually chose the higher abundance. I also favored DEQ data when possible, and older samples that were likely part of the PREDATOR models.
    - This resulted in 265 available reference samples for modeling
    - There were a few instances of multiple sample son the same stream, but only when there was considerable distance between sites, or in SE OR I allowed for a couple of sites a bit closer together
    - Again, these samples were identified from the ‘raw\_bugs’ associated with Lesley’s new R project on One Drive
* **PREDICTORS**
  + Need to assign StreamCat and NHD slope values to all 265 ref sites used to build the models
  + Some sites do not fall on NHD medium resolution, thus they do not have COMIDs and subsequently need to be matched to StreamCat by hand.
  + ALT took our stations list with matching bug samples, found sites without COMID, then through GIS matched them to the nearest COMID.
  + For our 265 reference sites used to build the 2024 model(s), SLH observed 7 ref sites that were associated to COMID by hand.
  + Through conversations with Ryan Hill (USEPA, StreamCat), he recommended we use WS metrics for sites with COMIDs; but then use CAT metrics for these other sites not on NHD med res. The CAT mets would be more representative of the smaller streams.
  + SLH then combined the WS and CAT metrics into a single predictor moterics table (but kept a column to show whether the mets were WS or CAT for each site).
  + Dropping predictors:
    - MAST was dropped because 37 ref sites were missing this metric
    - 10 other geology metrics were dropped because they had minimal information (no spread/variability across ref sites).
    - CORRELATIONS: SLH used a correlations matrix to identify highly correlated variables (r > 0.89)
      * TMAX8110: highly correlated with TMIN and TMEAN
        + Drop TMIN and TMEAN
      * PRECIP8110: highly correlated with PRECIP08, PRECIP09
        + Drop PRECIP08, PRECIP09
      * PCTBL2004: highly correlated with PCTBL2001
        + Drop PCTBL2001
      * INORGNWETDEP\_2008:
        + Drop NO3\_2008, NH4\_2008
      * AL2O3 & NA20
        + Drop AL2O3

**OLD….2022/2023 Oregon O/E models**

* Reference sites
* Bug data
  + We shipped ref data to Trip and he sent it back to us, subsampled to 300 count.
* Predictors
  + Explored potential issues with StreamCat values for small (< 5km2) watersheds (see 1/26/23 email)
    - Watershed areas compared to StreamStats (USGS): >5km2 r = 0.99, < 5km2 r = -0.38
    - StreamCat as predictors for small watersheds = erroneous/problematic?
  + Limited to StreamCat variables
  + Used PCA as a first guide to selecting variables
    - Balance between high axes loadings and interpretability
* Modeling
  + Build full RF model with all predictors, then use VIP to narrow down the list
  + Poor performance, generally: null > predictive (this was true for WCCP area, but not MWCF—later determined to be outliers, several glacial)
    - True with SE OR included
    - Chuck suggested removing 4 (?) outlier samples—2 of which were high elevation and glacial
  + Explored running “new” ref sites (based on DEQ’s updated ref screening protocols) through our old (2005) models. Saw similar issues: poor predictive SD of O/E, null SD better.
    - FYI to Jen: Trip discovered errors on USU end where some elevations were being divided by 10, then taking sqrt. Trip said he fixed this in the USU database (see 1/19/23 email)
  + Explored potential for Lab effect
    - Noticed USU samples and DEQ samples in Cascades ecoregion were somewhat consistently found in separate cluster groups.
    - Trip ran NMDS on lab, large scale ecoregions (Coastal, SE OR, other), Agency. Good spatial agreement across all levels.
      * Species Indicator Analysis also run and explored.
      * PCA across CA/OR/USU samples also looked good.
    - Chuck signed off on no lab/source effect.
  + Trip/Chuck identified 6 outlier sites, due to very low O/E values in model attempts. (see email 1/24/23)
    - 1 in CA, 1 in NBR, 4 in Cascades/E. Cascades.
    - Scanned in GE for missed human disturbance = no. Some fire. Some highly glacial.
  + Removing SEOR and outliers improved predictive performance