**2022/2023 Oregon O/E models**

* Reference sites
* Bug data
* 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