**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
  + Missing predictors: 4 sites had no StreamCat metrics due to incorrect COMIDs. SLH used the StreamCat map interface and site lat/longs to find the correct COMID. There were changed in the Bio\_MlocIDs\_AWQMS.xlsx file.
    - Details in Team chat between Lesley, Adam, Shannon
  + Missing ecoregions: there were multiple ref sites (5?) missing ecoregion info. All from CA. SLH changed them by hand, in the Bio\_MLocIDs file.
  + Predictors explored: all from StreamCat, except NHD slope
    - See r-code for ASSEMBLE DATA
* Model builds
  + First run = all 265 sites
    - X = 0.97, SD = 0.196 🡪 POOR
  + Second run = remove NBR (35 samples dropped, new n = 230)
    - X = 1.0, SD = 0.188
    - A bit better, but not good.
  + Third run = remove potential outlier samples
    - 35752-ORDEQ:19990707:R:SR USU(NOSTORETID) 35752-ORDEQ HORSE SIGN CR AT RM 4.3
    - 21814-ORDEQ:20010912:R:SR OREGONDEQ 21814-ORDEQ Chetco River at River Mile 56.09
    - 30343-ORDEQ:20030814:R:SR OREGONDEQ 30343-ORDEQ Polallie Creek tributary
    - ALL predictors, 12 ref groups: X = 1.03, SD = 0.175 … Not bad, could be better.
    - Reduced (9) predictors, 12 groups: X = 1.02, SD = 0.173
  + 4th run = remove 3 more outlier sites (n = 224)
    - 32555-ORDEQ:20050801:R:SR OREGONDEQ 32555-ORDEQ Twin Lakes Creek
    - 35813-ORDEQ:20000901:R:SR USU(NOSTORETID) 35813-ORDEQ LOSTINE R EF AT RM 2.8
    - 123437 USU(NOSTORETID) PIBO:0892 Freezout Creek
    - All predictors, 12 groups: X = 1.03, SD = 0.163
    - Reduced predictors (9), 12 groups: X = 1.02, SD = 0.161
  + 5th run = No NBR, 6 outliers removed, 9 ref groups (n=224)
    - Full: X = 1.036, SD = 0.167
    - Reduced (9): X = 1.025, SD = 0.169
  + 6th run: same, except add MSST and MWST back in
    - Reduced (11): X = 1.026, SD = 0.169
    - No benefit
  + 7th run: 8 groups, 9 preds (diff, based on top varImp with 8 grps)
    - FULL: X = 1.035, SD = 0.167
    - REDUCED: X = 1.027, SD = 0.172

**Outliers**

* **35752 = Horse Sign.** SW OR. Used USU site, but dropped DEQ site (12092). 35752 = low diversity (OTU richness = 17), switch back to 12092 (OTU rich = 46)
* **21814 = Chetco R**. 2001 Total rich ~ 18.
  + 1999 richness ~ 17
  + 2000 = 24
  + These low values are an artifact of where we sampled in the reach. The first half of reach is fast flowing boulder field. Very few bugs. Would sample differently given what I know now.
  + Drop this site.
* **30343 = Pollalie Creek.** Off Mt Hood. OTU rich = 24. Highest PCTICE—glacial influenced, lots of glacial till. DROP SITE.
* **32555 = Twin Lakes.** Central Cascades. OTU rich = 18. Close to lake edge, more of a wetland complex than a creek in this reach. DROP SITE.
* **35813 = EF Lostine.** Wallowas. Poor taxonomic resolution. 23 out of 58 records are DNI due to poor resolution—including 127 “Plecoptera”. Looks glacially influenced, based on sat imagery (water color), but PCTICE is 0…? DROP SITE.
* **123437 = Freezeout.** Near Hells Canyon. Low OTU rich—mostly due to extreme dominance by Simulium taking up most of counts. Use sample 152645 instead.
* **35618 = Miller**. Downstream of a lake with several campgrounds, roads. Passes ref criteria, but looks close. OTU rich = 24. Drop due to disturbances.
* **24044 = Mill Creek**. Relatively short stretch of creek between two lakes, near upper watershed. Pretty good richness, but 2/3 of the sample is tied up in one midge taxon. This seems to be common in near-lake settings. Certainly not impaired based on the taxa list. Not enough reasons to exclude, but lake effect sites might be considered outliers?
* **21812 SF Salmon**. Multiple samples from USU (1998) and DEQ (1999). Initial sample used () was USU. These samples have more DNI taxa, use DEQ where possible. Switch to ‘21821-ORDEQ:19990803:R:SR’.
* **13200 = Trib to Salmon**. Very low total abundance, dominated by B. tri. (80%). No disturbance, no outlier conditions.
* **12894 Cultus**. 1998 DEQ sample, could remove because of that--methods slightly different (300 ct target). DROP SAMPLE.
* **35633 = NF Whychus**. Glacial dominated. Outlier. DROP SITE.
* **PIBO:0886 = Cottonwood (NE OR).** 3 samples from 04 – 14. Sat imagery suggests this is intermittent at best and some years completely dry. The point looks to be the bottom end of water for most years, with interrupted flow above most years. **DROP SITE?**
* **134530 Dutch Flat.** Bad sample, really low richness. **Replace with 140016.**
* **121071 EF Annie.** EXTREME dominance, 500 Baetis. DROP SITE.
* **24426 Trib to Strawberry.** Lots of DNI. **Replace with 24426-ORDEQ:20010719:R:SR.**

**Start over modeling after removing outliers**

* 257 ref sites brought in
* Remove NBR
* Remove Miller 35618
* 221 ref sites total in model
* 1st run:
  + 8 ref groups
    - Full X = 1/034, SD = 0.161 (Null SD = 0.173), rep SD = 0.1311
    - Reduced (6 preds: TMAX8110,BFI,ELEV,MWST\_mean08.14,CLAY,PRECIP8110)
      * X = 1.012, SD = 0.162 (Null 0.173), rep SD = 0.1288
      * GOOD!!!!!
* grps.8
* 1 2 3 4 5 6 7 8
* 25 26 16 38 16 38 28 34
  + 9 ref groups, same 6 predictors
    - Reduced: 1.011, SD 0.159, rep SD = 0.1288
    - Pretty much the same but a little better.
    - Stick with 8, as 9 ref groups has one group with 9 sites.

**Model validation**

* Compare 2024 O/E to 2005 O/E
* Sensitivity: % Most Disturbed actually score poor
  + Use 10th percentile of ref O/E to start

**OLD attempts….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