GAM Write Up

Background on GAMs? Or should this belong in introduction?

Steps:

-Dataset Selection

1. Ensure dataset has balance between containing all predictor vars (engineered soils) while maintaining sufficiently large sample size

-Variable Selection

1. Hand select important vars

2. Run PCA/PLSR to reduce dimensionality

3. Run random forest to select top x vars

4. Run GAM and remove vars until all have p-value below 0.2

-Running GAM

1. Model configuration—base with select = TRUE

2. Enter all non-categorical vars as smooth, all categorical vars as parameter

3. If edf is below 1, remove smoothing function from that var

-Model performance metrics

1. Adjusted R2 to assess model performance

2. VIF scores to assess multicollinearity

3. Concurvity to assess smoothed multicollinearity (not sure what correct word is for that)

4. Other GAM output graphs to make sure residuals follow normal distribution and are evenly distributed around 0

5. GCV/AIC to compare to other models

The obtained GCV score in GAM can be used as the AIC, and smaller values indicate better fitting models ([Wood, 2017](https://www.sciencedirect.com/science/article/pii/S0022169422006643" \l "b0375))

-Verify the sample size is sufficiently large

1. Follow GAM process with full sample size

2. Remove one sample and repeat GAM process

3. Continue removing one sample at a time, and repeating GAM process

4. Identify when model starts to break down