**GAM Methods Write Up**

**Paragraph 1: Background on GAM**

Generalized Additive Models (GAMs) are widely-used statistical tools that extend the framework of ordinary least squares (OLS) regression by allowing an arbitrary link function of the response variable to vary linearly with predictors instead of assuming response variable itself must vary linearly. GAMs are a generalized form of Generalized Linear Models (GLMs) that incorporate spline smoothing functions, allowing it to capture complex, non-linear relationships between predictor and response variables. This flexibility makes GAMs particularly valuable for modeling data with non-linear relationships, offering a balance between simpler models, such as linear regression, and more complex approaches, like neural networks. In our study, GAMs are especially well-suited as our distribution of infiltration rates do not follow a normal distribution, and relationships between infiltration and other variables, such as soil type composition, are widely recognized as non-linear (cite).

-gam is good blend of simple models (like linear regression) yet can account for complex relationships (like neural networks)

-generalized form of OLS but uses spline smoothing functions to capture non-linear relationships

-both GLMs and GAMs allow response variables to have non-normal distributions and uses link function that varies linearly with predictors (link function is arbitrary function of response variable) instead of assuming response variable itself must vary linearly

-good for our case where infiltration rates do not follow a normal distribution and it is widely known the relationships between infiltration and other variables (% soil type for example) is not linear (need citation)

**Paragraph 2: Dataset Introduction/Selection**

Our dataset consists of 182 biofilters across seven states in the Great Lakes region (IL, IN, MI, MN, NY, OH, WI). For all 182 sites, we initially collected 35 predictor variables related to the size and land use type of the surrounding drainage area, and the size, age, condition, vegetation cover, underlying soil type, inlet type, and outlet type of the biofilters themselves. For 102 of the 182 sites, we were also able to collect information on the engineered soil depths and types at each biofilter. From preliminary analysis using linear regression models, we determined that the engineered soil variables often had strong relationships with the infiltration rates. Thus, the dataset we used for most of the statistical analysis was the trimmed down dataset (n=102) to balance maintaining a sufficiently large sample size while keeping the often-important engineered soil variables. We also ran analyses on the full dataset (n=182) and removed the engineered soil variables and on a smaller subset of the data for the samples that had at least four total infiltration tests and at least one SATURO infiltration test (n=51).

-dataset consists of 182 biofilters across 7 states in the great lakes region

-of the 182 sites, we were able to get engineered soil information for 102 of the sites

-from preliminary analysis (including SHAP, PLSR, GAM), determined engineered soil vars were often important

-used trimmed down dataset (n=102) to balance maintaining a sufficiently large sample size while keeping the engineered soil vars in

-also ran analyses on full dataset and smaller dataset where we were more confident in infiltration measurements

**Paragraph 3: Subjective Variable Selection**

With the trimmed down dataset (n=102), we initially started 41 predictor variables, many of which were categorical or compositional. While GAMs can deal with categorical variables, they are sensitive to multicollinearity. After performing a correlation analysis of our predictor variables (figure?), we determined that many groups of variables had issues of multicollinearity, including land type percentage, vegetation cover percentage, underlying soil percentage, engineered soil percentage, and inlet and outlet type. To reduce multicollinearity, we removed all categorical variables related to inlet/outlet type and condition and removed at least one from each group of compositional variables so that the percentages no longer added up to one for each biofilter. For the land type percentages in the drainage area, we used only percent non-residential which was a combination of the percent commercial and percent institutional variables. For percent vegetation, we removed the percent vegetation bare predictor variable and kept percent vegetation tree, grass, shrub, prairie, and other (mulch). For the engineered soil variables, we used only engineered soil depth, engineered soil percent sand, and engineered soil percent compost. Finally, for the six underlying soil variables, we ran a Principal Component Analysis (PCA), and used the first two principal components to reduce dimensionality. After this paring down of predictor variables, we had a set of 17 variables that were initially inputted to our statistical models.

-Originally started with 41 predictor variables, many of which were categorical, compositional, or otherwise highly collinear

-Removed categorical vars related to inlet/outlet type and condition as we did not think they would be important predictors

-for compositional variables, removed at least one from each set so they no longer added up to 1

-% land type: added up % commercial and % industrial as % non-residential

-% vegetation: removed % vegetation bare

-% engineered soil: only used engineered soil % sand and engineered soil % compost

-underlying soil vars: ran PCA and used first two PCs, which explained 62% of variance, to reduce dimensionality

-reduced initial variables to 17

**Paragraph 4: Objective Variable Selection**

To further reduce the predictor variables, we inputted the 17 selected variables into a random forest model, which can be useful for preliminary exploration of data as they handle collinearity, capture non-linear relationships, do not require any assumptions about the distribution of predictor variables or their relationship to the response variable, and provide a ranking of variable importance. Using the variable importance rankings, we selected the top 12 most important predictor variables to use in the GAMs. Then, we initially ran the GAM with these 12 variables and used backwards selection to determine the final model, where we iteratively removed the variable with the largest p-value until all variables had a p-value below 0.2.

-Ran random forest on all 17 vars. Random forests can be beneficial for preliminary exploration of data as they handle collinearity, capture non-linear relationships, do not require any assumptions about the distribution of predictor variables or their relationship to response variable, and provide ranking of variable importance

-Selected the top 12 important variables from RF to input into GAM

-Used backwards selection (check this terminology) to select final model, where we would run GAM and remove variable with highest p value until all p values were below 0.2

**Paragraph 5: Running GAM**

When running the GAM, we used the mgcv package in R (cite), the generalized cross-validation (GCV) method for selecting the smoothing parameters, and set select equal to true to provide automatic smoothing parameter selection and to further penalize each smoothing term. We began by applying a smoothing term to all non-categorical variables and adding the categorical variables as linear parameters. After each run of the model, we removed the smooth term from each variable with an effective degrees of freedom (edf) below one, which indicates that there is either a linear or no relationship between that predictor variable and the response variable.

-Used base method for smoothing: generalized cross-validation (GCV)

-Set select equal to TRUE which is recommended to simplify model and help with overfitting (cite). Provides automatic smoothing parameter selection with additional penalties for smooth term

-Started by adding smoothing term to all non-categorical variables—included categorical vars as parameters

-After each model run, checked effective degrees of freedom. If edf = 1, indicates variable has linear relationship and removes smoothing term. Also removed smoothing term if edf was less than 1, as indicates none to linear relationship between predictor and response

**Paragraph 6: Assessing Model Performance**

To assess overall model performance and variable importance, we used a variety of statistical metrics. We used R2 and percent variance explained to assess general model performance. We used gam.check to assess residual plots, where Q-Q plots should

-Used R2 (% variance explained) and cross-validated R2 to assess general model performance

-Used VIF scores and concurvity to assess multicollinearity. Concurvity tells you if smooth term can be approximated by one or more other smooth terms.

-Used gam.check to assess residual plots (make sure Q-Q plot follows line, make sure residuals are randomly distributed around 0 and follow Gaussian distribution)

-Can also use GCV/AIC to compare models to others (not sure if I should bring this up bc I didn’t really do it). 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))

-Used the estimate of the parametric coefficients to identify sign of relationship between parametric predictors and infiltration rate. Also plotted results of GAM to assess relationship between smoothed vars and infiltration rate

**Paragraph 7: Verifying Sample Size is Sufficiently Large**

One limitation of this study is the relatively small sample size compared to the number of predictor variables. The general recommendation is that for GAMs, there should be roughly 20 samples per predictor variable included in the final model (cite). To verify that our sample size was sufficiently large, we followed our GAM methods outlined above, but iteratively removed one sample from the pared down dataset (n=102) and continued running the model algorithm until 20% of the samples were removed. We found that the R2 remained between 0.404 and 0.431 and the deviance explained ranged from 50.1% to 52.7%. Each final model always contained the same variables except the engineered soil percent sand was removed in the last three runs as the p-value crept above 0.2. The engineered soil percent compost variable was sometimes typically kept as a smoothed term but sometimes the edf dropped below 1 so the variable was converted to a linear parameter. The order of variable significance based on the p-values was slightly altered throughout, but the percent vegetation other variable remained as the most significant for each run.

-Something about how various people suggest having like 20 samples per predictor variable in GAM (need citations)

-Because our sample size was relatively small compared to our number of predictors, we wanted to test whether the sample size was sufficiently large to create robust models.

-To test this, we followed the GAM methodology outlined above, but iteratively removed one sample from the pared down dataset (n = 102) and continued rerunning the model until 20% of the samples were removed (n = 82).

-R2 remained between 0.404 and 0.431, deviance explained remained between 50.1% and 52.7%, always had the same variables although engineered soil % compost was sometimes kept as a smoothed term and sometimes changed to a parametric term. Towards the end, p value of engineered soil % sand would sometimes creep above 0.2 and have to be removed (happened 3 times). Order of significance sometimes changed slightly, but % vegetation other was always most significant var