WDA Final - Assesing AIM Water Quality Monitoring Data and Dams in WA State

Andrew Friedman-Herring, Cat Otero, Jackie Van der Hout

Spring 2022

Setup

Purpose of our Research Project:

We chose this project because water quality is an incredibly important issue, and we wanted to see how physical and biological water quality indicators collected by the BLM correlate with each other. Dams also have a very large impact on streams and rivers in the US, as they increase stream temperatures and provide conditions for algal blooms which can reduce water quality. Dams are also a very overlooked piece of infrastructure. By collecting the number of dams upstream of each monitoring site, we are able to add that value to the AIM data set and incorporate it with our analysis of how water quality variables correlate.

Data Used: For our analysis, we used the BLM Assessment, Inventory, and Monitoring (AIM) Data, which we found and data scraped from online in a GeoJSON format.

We also used the Washington State Department of Ecology Inventory of Statewide Dams, which was also found and data scraped from online in a GeoJSON format. This information included latitudes and longitudes for all 57,000 dams in Washington State, as well as their fish barrier classification, which ranged from “Not a barrier” to “Complete Fish Barrier”.

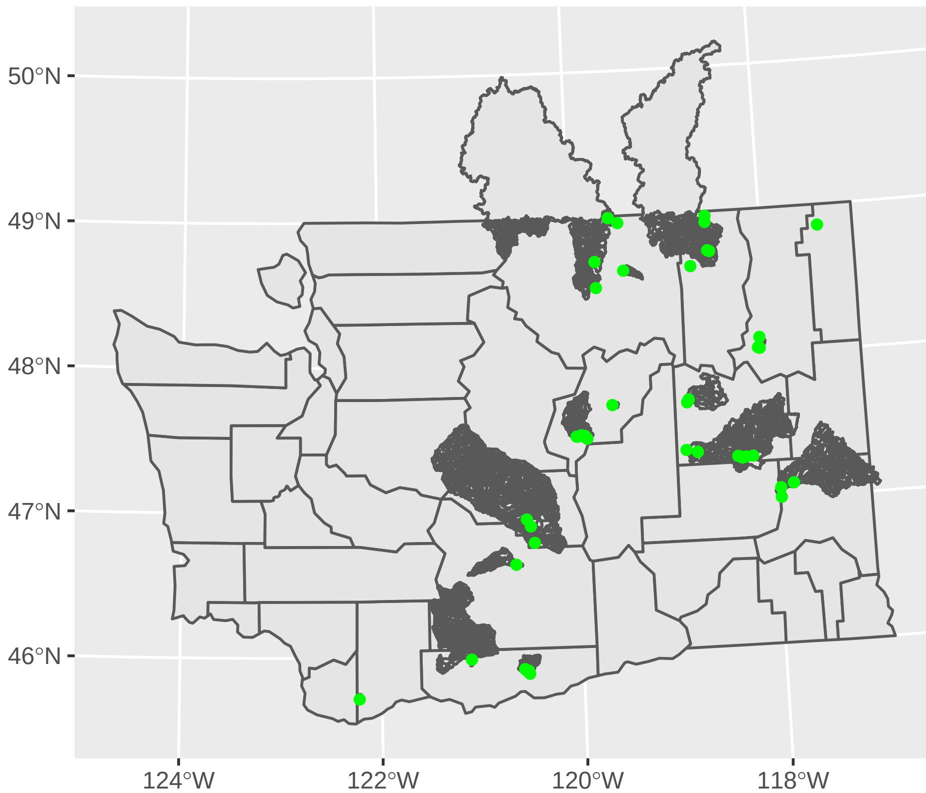
Finally, we used the National Hydrography Dataset of Flowlines and Catchments developed by the US EPA and USGS to connect first two data set spatially. This dataset was used in the form of the NHDplusTools R package to find the catchment and flowline data upstream of each monitoring site.

Data Wrangling:

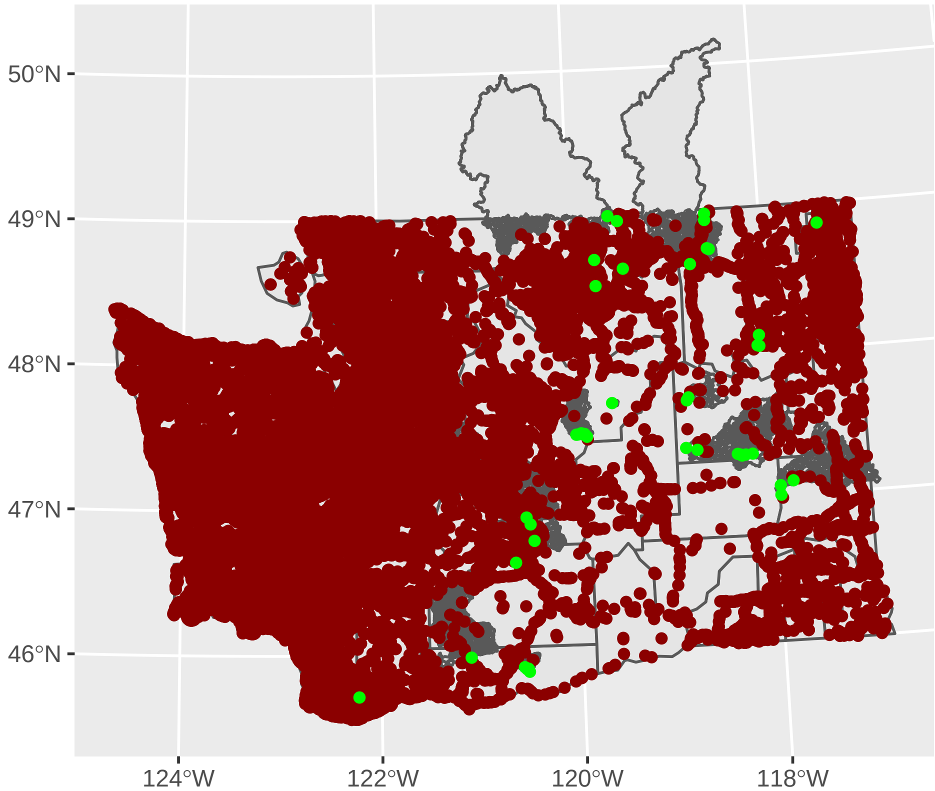
The data wrangling phase of our project was the most time and code intensive, as we were trying to intersect two very large data sets spatially to create a new variable to add to the AIM data set. Because there were 45 AIM sites in Washington, we had to repeat this intersection for all 45 sites. We accomplished this by writing a custom function that extracted the catchment above a monitoring site and saved it as a a spatial polygon object. We then wrote a ‘for’ loop that would apply this catchment extraction function for all 45 sites and save them as one combined spatial dataframe. We repeated this process for the flowlines above each monitoring site as well to be used in a future analysis. We then wrote another set of functions that takes the catchment data, intersects it with the inventory of dams and returns the number of dams contained in that catchment, the area of the catchment in square kilometers, and the dams per square kilometer. We then wrote another ‘for’ loop to apply this function to all 45 AIM sites, and return a data frame containing the number of dams above each AIM site. This was then joined to the original AIM dataframe, effectively adding a column to the AIM data set counting how many dams were above each monitoring site.

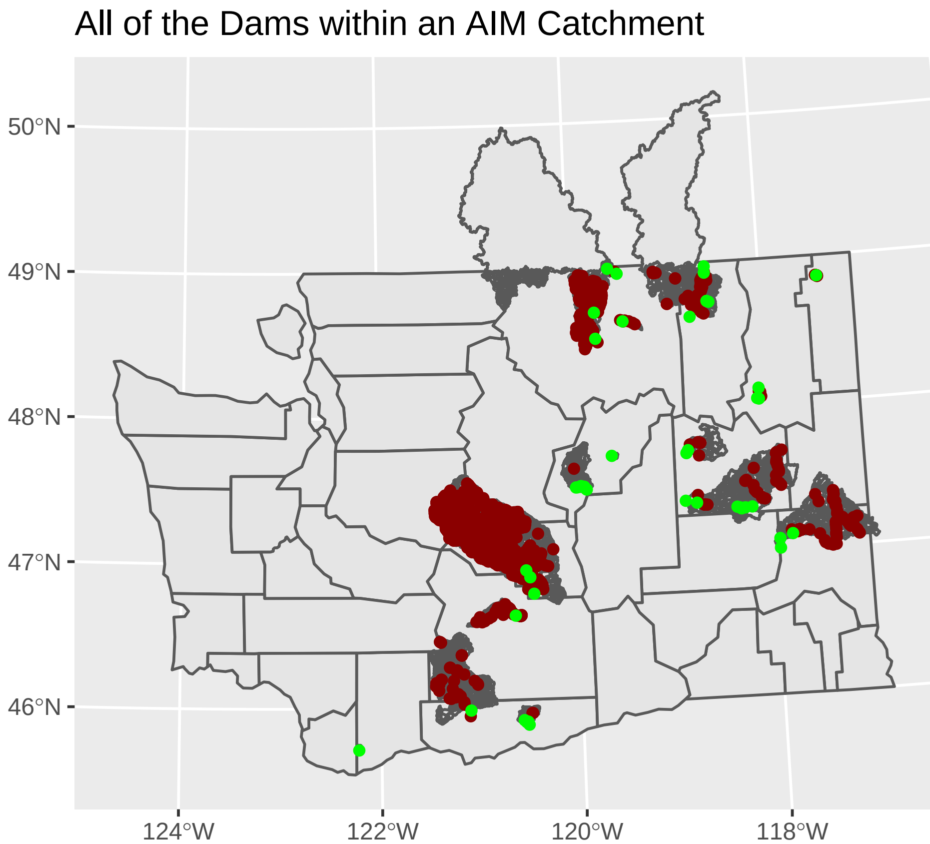
Spatial Visualization:

AIM Sites and Upstream Catchements



AIM Catchements and Every Dams in WA





Data Scraping and Wrangling done for WDA

```
catchment_extract <- function(site_id, waaim_data){
  selectingsites <- filter(waaim_data, SITE_CD == site_id) #takes site ID and filters dataset
  MIDLAT <- selectingsites$MIDLAT
  MIDLONG <- selectingsites$MIDLONG
  start_point <- st_sfc(st_point(c(MIDLONG, MIDLAT)), crs = 4269)
  start_comid <- discover_nhdplus_id(start_point)
  NLDI <- navigate_nldi(list(featureSource = "comid", featureID = start_comid),
                        mode = "upstreamTributaries",
                        distance_km = 1000)
  subset_file <- tempfile(fileext = ".gpkg")
  subset <- subset_nhdplus(comids = as.integer(NLDI$UT$nhdplus_comid),
                         output_file = subset_file, #the above temporary file
                         nhdplus_data = "download",
                         flowline_only = FALSE,
                         return_data = TRUE, overwrite = TRUE)
  catchment <- subset$CatchmentSP
  return(catchment)
}

Dam_catchment_intersect <- function(CATCHMENT, DAM_DATA){
  DAM_DATA_UTM <- st_transform(DAM_DATA, crs = 26910)
  CATCHMENT_UTM <- st_transform(CATCHMENT, crs = 26910)
  DAM_SUBSET <- DAM_DATA_UTM[lengths(
  st_intersects(DAM_DATA_UTM, CATCHMENT_UTM)) > 0,]
  return(DAM_SUBSET)
```

```
}

Dam_WAAIM_Join_function <- function(WAAIM_ROW_NUM, WAAIM_DATA, DAM_DATA){
  site_id <- WAAIM_DATA$SITE_CD[WAAIM_ROW_NUM]
  site_catchment <- catchment_extract(site_id, WAAIM_DATA)
  site_dams <- Dam_catchment_intersect(site_catchment, DAM_DATA)
  dams_number <- dim(site_dams)[1]
  return(dams_number)
}

result <- data.frame(matrix(nrow = 45, ncol = 4))
for(i in 1:dim(WAAIM_tidy_utm)[1] ){
result[i,1] <- WAAIM_tidy_utm$SITE_CD[i]
site_catchment <- catchment_extract(result[i,1],WAAIM_tidy_utm)
site_dams <- Dam_catchment_intersect(site_catchment, dam_complete_raw_utm)
result[i,2] <- dim(site_dams)[1]
result[i,3] <- sum(site_catchment$areasqkm)
result[i,4] <- ((result[i,2])/(result[i,3]))
}
result <- result %>%
  rename("Site_CD" = X1,
         "Number_of_dams_upstream" = X2,
         "Catchment_area_sqkm" = X3,
         "Dams_per_sqkm" = X4)

write.csv(result, "./New Data/Dam_catchment_data.csv")

#check class of columns, we want numeric for analysis
sapply(WAAIM_tidy_dams, class)

#Specify columns want to change to numeric for WAAIM data
i <- c(5:58)

#convert the columns indicated in i above to numeric
WAAIM_tidy_dams[ , i] <- apply(WAAIM_tidy_dams[ , i], 2,
                    function(x) as.numeric(as.character(x)))

#check that all columns we want are numeric
sapply(WAAIM_tidy_dams, class)

#_____________________________________________________

#check class of catchment_data variables
sapply(catchment_data, class)

#Specify columns want to change to numeric for WAAIM data
j <- c(1,3)

#convert the columns indicated in i above to numeric
catchment_data[ , j] <- apply(catchment_data[ , j], 2,
                    function(x) as.numeric(as.character(x)))

#check class of catchment_data variables
```

```r
sapply(catchment_data, class)

#rename Site_CD to SITE_CD so can join
names(catchment_data)[names(catchment_data)=="Site_CD"] <- "SITE_CD"
names(catchment_data)

#merge the dfs, must be the same class to merge
dam_data_merge = WAAIM_tidy_dams %>% left_join(catchment_data,by="SITE_CD")

#check which columns have NAs, only works on numeric columns
colSums(is.na(dam_data_clean))

#calculate a channel geometry column and add it onto df
#select columns of interest, gets rid of geometry so can save as a csv file
#drop rows for sites w/ NA for TotalP&N, invert variables, and D50
dam_data_clean <- dam_data_merge %>%
  mutate(WDRatio = round(BNKFLL_WT / BNKFLL_HT)) %>%
  select(OBJECTID, ReachLength, StreamOrder, VEG_CMPLXTY, RPRN_VEG_CNPY_CVR, RPRN_VEG_UNDRSTRY_CVR, RPRN
  slice(-c(7,9,10,13,14,30,33,39,42,43,44))


#check data is complete, no NAs
colSums(is.na(dam_data_clean))

#save cleaned data as a csv
#write.csv(dam_data_clean, "./New Data/dam_data_clean.csv", row.names=FALSE)
```

Analysis: AIC Models and Regressions

We used AIC models to select which explanatory variables to use in a linear model. It can be tempting to throw all variables with available data into a model, but that can get confusing and not actually increase the explanatory power of a model. By using AIC to select variables, we can minimize covariates and end up with a simpler model that is informative and easier to interpret.

We performed regressions to see if and to what extent explanatory variables influence our dependent variables. We used the common p-value of below 0.05 to indicate significant relationships.

Table 1: Total Phospohrus AIC Results

Variables_Correlated	Direction_of_Correlation	P-Value
StreamOrder	Positive	<0.001
BNK_STBLTY	Negative	<0.001
D50	Negative	<0.001
ReachLength	Positive	<0.01
INSTNT_TEMP	Negative	<0.01
LWD_FREQ	Negative	<0.01
MACROINVRTBRTE_CNT	Negative	<0.01
WDRatio	Negative	<0.05

Table 2: Total Nitrogen AIC Results

Variables_Correlated	Direction_of_Correlation	P-Value
OBSRVD_INVRT_RCHNSS	Negative	<0.01
WDRatio	Negative	<0.01
RPRN_VEG_CNPY_CVR	Negative	<0.05
RPRN_VEG_GC	Positive	<0.05
NON_NTVE_HRB	Negative	<0.05
LWD_FREQ	Negative	<0.05
INCSN_HT	Negative	<0.05
CHN_INCSN	Positive	<0.05
WTTD_WT	Positive	<0.05
Dams_per_sqkm	Positive	<0.05

Table 3: Invertibrate Counts AIC Results

Variables_Correlated	Direction_of_Correlation	P-Value
NON_NTVE_HRB	Negative	<0.001
TotalN	Negative	<0.001
LWD_FREQ	Negative	<0.01
BNK_STBLTY	Negative	<0.01
INCSN_HT	Negative	<0.01
CHN_INCSN	Positive	<0.01
WTTD_WT	Positive	<0.01
WDRatio	Negative	<0.01
Catchment_area_sqkm	Negative	<0.01
StreamOrder	Positive	<0.05
RPRN_VEG_CNPY_CVR	Negative	<0.05
RPRN_VEG_UNDRSTRY_CVR	Negative	<0.05
RPRN_VEG_GC	Positive	<0.05
NTVE_HRB	Negative	<0.05
Number_of_dams_upstream	Negative	<0.05
D50	Negative	<0.05
Dams_per_sqkm	Positive	<0.05

Table 4: Bankfull Height AIC Results

Variables_Correlated	Direction_of_Correlation	P-Value
NON_NTVE_HRB	Negative	<0.01
LWD_FREQ	Negative	<0.01
OBSRVD_INVRT_RCHNSS	Negative	<0.01
RPRN_VEG_CNPY_CVR	Negative	<0.05
RPRN_VEG_GC	Positive	<0.05
NTVE_HRB	Negative	<0.05
SDGE_RSH	Positive	<0.05
BNK_STBLTY	Negative	<0.05
TotalN	Negative	<0.05
INCSN_HT	Negative	<0.05
CHN_INCSN	Positive	<0.05
WTTD_WT(+)	Positive	<0.05

Variables_Correlated	Direction_of_Correlation	P-Value
Number_of_dams_upstream	Negative	<0.05
D50	Negative	<0.05
Dams_per_sqkm	Positive	<0.05

Table 5: D50 Particle Size AIC Results

Variables_Correlated	Direction_of_Correlation	P-Value
WTTD_WT	Positive	<0.001
WDRatio	Negative	<0.001
VEG_CMPLXTY	Negative	<0.01
RPRN_VEG_GC	Positive	<0.01
SDGE_RSH	Positive	<0.01
LWD_FREQ	Negative	<0.01
Number_of_dams_upstream	Negative	<0.01
TotalN	Negative	<0.01
StreamOrder	Positive	<0.05
RPRN_VEG_CNPY_CVR	Negative	<0.05
NON_NTVE_HRB	Negative	<0.05
pH	Positive	<0.05
INCSN_HT	Negative	<0.05
OBSRVD_INVRT_RCHNSS	Negative	<0.05
Dams_per_sqkm	Positive	<0.05

Table 6: Dams/sqkm AIC Results

Variables_Correlated	Direction_of_Correlation	P-Value
RPRN_VEG_GC	Negative	<0.01
LWD_FREQ	Positive	<0.01
WTTD_WT	Negative	<0.01
Number_of_dams_upstream.x	Positive	<0.01
WDRatio	Positive	<0.01
VEG_CMPLXTY	Positive	<0.05
RPRN_VEG_CNPY_CVR	Positive	<0.05
NON_NTVE_HRB	Positive	<0.05
SDGE_RSH	Negative	<0.05
pH	Negative	<0.05
LWDVolume	Negative	<0.05
INCSN_HT	Positive	<0.05
OBSRVD_INVRT_RCHNSS	Positive	<0.05
TotalN	Positive	<0.05
D50	Positive	<0.05

Conclusions:

Overall, either Total number of dams or dams per sq km was a statistically significant variable in all of the AIC linear regressions that we ran. In addition, they were typically associated with metrics that indicate lower water quality, such as increased notrogen levels and decreased vegitation cover. This indicates that dam are either a driver of reduced water quality, or are an correlated indicator of highly impacted watersheds affecting water quality and number of dams in the landscape both.