Analysis for bw project:

1. Get data into R
   1. Presence data: Environmental data from locations where defoliated
   2. Background data: Environmental data from random locations
   3. Remove NAs
   4. Create same number of samples for defoliation and background data (600 = closest centennial once NAs removed)
2. Correlation structure
   1. correlation plot with numbers
   2. dendrogram of distance between values (clustering)
   3. Variance inflation factor
   4. Combine info from a, b, and c to remove MaxT0607, MaxT0708, sg\_date, sg\_\_gdd\_s2, minT0607, minT0708
   5. Retry a, b, and c
   6. Remove pcp0708 and pcp0506
   7. Retry a, b, and c
3. Fit model

full: r\_d ~ MaxT0506 + MinT0506+ pcp0607 + Sg\_\_gdd\_p1 + DistanceRo + DistanceRi + SpRasterNA

\* made sure testing not sequential

1. Fit residuals against fitted values and evaluate model

Using DHARMa

* 1. if straight line appropriate – depends on sample run
  2. error homogenous – depends on sample run
  3. normal – yes totally fine dispersion and no outliers
  4. Test for spatial autocorrelation – patterns depend on sample run

1. Increase sample size
   1. Increased max number of random points to 2400 once NAs removed (change way random points extracted from QGIS so that the 3000 points placed land only on NL and not in the ocean lol)

Correlation structure same with more samples

* 1. the stats done by the package say that the residuals aren't homogenous but I think the graphs all look quite straight \*I think the power of my sample size is so large the type ii error is huge
  2. Can’t tell if spatial autocorrelation a problem or how to incorporate

1. Looked at lag.plots – expanded to 5000 and the pattern emerged
   1. Found strong structure on pcp0607, MaxT0506 and DistanceRo

Smaller structure for MinT0506 and DistanceRi

Little/no structure for sg\_\_gdd\_p1, lat, and long

* 1. Find real world distance represented by lagplot (I think this would be 5000/furthest distance = 0cross/real distance) so

But for cross of threshold it is 175 km. ~ 15 km window where they are below the threshold

1. Created gam with smooth functions on X and Y
   1. Significance of smooth terms very significant
   2. Different significant values:

\* testing interaction terms would likely increase deviance explained but I have so many variables

\* decided not to use PCA because it cannot handle categorical data which is something important I am looking at