P-SHEC data and modeling steps:

1. Extract all available yield data for all target crops in all possible counties from NASS survey (2000-2023). (**Done: 12/13/2024**)
   1. De-trend yield for each crop in each county
   2. Save as separate files by crops
   3. Extract an all-county list
2. Extract all available irrigation area data for all crops in all possible counties from NASS census (2002-2022) (**Done: 12/20/2024**)
   1. Combine irrigation data with yield data
3. Extract all available crop phenology data for all crops in all states from NASS survey (2000-2023).
   1. Crop phenology data was only extracted for Tobacco (KY, NC, TN)
      1. NC contains transplant and harvest data for different types: used the type 31
      2. KY and TN don’t have specified types
   2. For the other crops, extracted the Crop Calendar data (<https://sage.nelson.wisc.edu/data-and-models/datasets/crop-calendar-dataset/>) Accessed on 1/17/2024.
   3. Save in a single document (**Done: 1/23/2025**)
4. Extract all weather data from PRISM for 1999-2023
   1. Aggregate the daily weather data to county levels and get the county mean total precipitation and mean daily temperature from plant to harvest using the Crop Calendar data **(Done: 1/25/2025)**
   2. Calculate growing season weather data for each crop in available counties from 2000-2023 **(Done: 2/24/2025)**
      1. Process crop progress data from plant to mature: state mean for each crop
      2. Calculate Growing degree days
   3. Since using county average across years didn’t improve the model performance, need to extract annual weather data. Keep both annual data and county average in the same file.
5. Nitrogen input data **(Done 2/26/2025)**:
   1. If available, use it.
   2. For oats, tobacco, and sugarbeets use the county average Nugis data; check if using averages improve model performance.
6. AFT generated crop frequency layer:
   1. Unzip and store all frequency layer together (the frequency layer is so big, maybe because all cells have a value instead of just NULL?). Unzip will occupy too much disk space
      1. Unzip when needed and delete after the function
7. gSSURGO data:
   1. Extract all states and filter counties from the all-county list obtained from the NASS yield step. **(Done: 1/25/2025)**
   2. Extract all soil variables including drainage class, soil order, and slope **(Done: 1/25/2025)**
   3. Save files by crop

Modeling part:

1. Merge all data to prepare for modeling (**Done: 3/3/2025**):
   1. Some counties are missing soil order
   2. Get the finalized county list, save in a file
2. Modeling steps:
   1. For corn:
      1. Use rainfed models: test the pd curves for all variables vs. correlated variables removed curves.
      2. Check changes in performance metrics and curves
      3. Decide which approach to use **(choose the best model)**
      4. Check overfitting: use more grids and see if that reduce overfitting
   2. For corn, soybean, and winter wheat:
      1. Run separate irrigation and rainfed models
         1. RMSE, MAE, R-square, and overfitting
      2. Run a single model
         1. RMSE, MAE, R-square, and overfitting
   3. For the other crops:
      1. Run separate models for individual crops
   4. Combine all crops separate and all model metrics **(Done: 3/12/2025)**
   5. Compare results from rainfed model/ irrigation model vs. all model for all crops. Decide which model to go (separate vs. all) for which crops.
   6. Run a model for grouped crops:
      * 1. Other small grains: barley, oats
        2. Other cereals: barley, sorghum, oats
        3. All small grains: wheat, barley, oats
        4. Warm-season cereals: corn and sorghum

Note: 1) for combined models, split data proportionally; 2) check the final results: for crops, irrigation, RMSE, MAE, R-square, and overfitting