Mitigating Flood Hazards on Tribal Lands with Data Science

# Problem Statement

Floods are natural hazards whose impacts are intensified or mitigated through direct and indirect human actions.

Floodplain designation is a hazard mitigation approach to reduce the risk that a rare flood leads to the loss of human life or damage to homes or businesses.

Floodplain designation has not been completed northern Great Plains Reservation communities.

Flood frequency analysis is a first step towards floodplain designation. Recently updated Federal guidance on flow frequency analysis recommends the use of data science methods to increase the accuracy of flow frequency analysis. However, the updated guidance does not identify how data science should be used to accurately determine the hundred-year flood magnitude.

Our research aim is to increase the accuracy and robustness of flood magnitude predictions by identifying groups stream gauges with similar flood behaviors from a large group of stream gauges across a wide geographic area with a wide variance of elevation, climate, and geography. The purpose of the research is to better estimate flood risk in Great Plains Reservation communities.

# Approach

We will use a statistical technique called **clustering**, to group stations based on their flood characteristics. The **Gaussian Mixture Model** is a clustering method that creates flexible groupings of stations to ensure that we capture the diversity of flood behaviors across the landscape.

A first step in the analysis is to remove the influence of scale, ensuring that differences between regions aren’t just due to the size of the river or stream. Next, we will select the most likely number of clusters using a criterion called **BIC**. BIC prefers a smaller number of groups to avoid unnecessary complexity.

We will use both visualization and quality metrics to evaluate how well the groups are defined. A technique called **t-SNE** will allow us to visualize complex data in a simple, easy-to-understand form. Additionally, we will calculate the **Silhouette Score** and **Calinski-Harabasz Index** to confirm that the groups are meaningful and distinct.

**Visualization and Evaluation:**

**Understanding Key Drivers:**

Once we’ve identified these regions, we want to understand what factors are driving the differences in flood behavior. We focus on variables like **elevation, climate, and geography**. To do this, we use a modeling technique called **Elastic Net**, which helps us select the most important factors without overwhelming the analysis with too many details.

**Outcome:**

This approach not only allows us to identify regional patterns in flood behavior but also helps pinpoint which environmental factors are most important for predicting extreme floods. This can be especially useful for flood risk management in areas like Tribal Lands, where understanding local flood risks is essential for community planning and safety.

Floodplain designation is a two-step approach involving flow frequency analysis to determine the peak discharge magnitude of a hundred-year flood, then apply a computer model to incorporate flood frequency analysis results with highly accurate topographic data to identify the land surface area that would be inundated by the floodwater.

This project aims to designate the hundred-year floodplain along portions of White Clay Creek, SD using publicly available LiDAR data and a FEMA-recommended computer model. The achievement of this aim requires the prior approval from the OLC Institutional Review Board and the OST Research Review Board, and the development of a replicable data science approach to accurately determine hundred-year flood magnitudes at gaging stations across the northern Great Plains.

The approach to accurately determine hundred-year flood magnitudes at gaging stations across the northern Great Plains is to apply a machine learning approach to identify streams

of statistically similar gages

to peak discharge observations from stream discharge recording gages from Montana to Texas to identify

Data science incorporates statistics, machine learning, programming, scientific expertise, and large datasets to address complex technical problems. Ethical data science requires FAIR (Findable, Accessible, Interoperable, Reusable) principles are addressed through the lens of CARE (Collective Benefit, Authority to Control, Responsibility, Ethics) principles.

# Background

Floods are natural hazards whose impacts are intensified or mitigated through direct and indirect human actions. Floodplain designation is a hazard mitigation approach involves determining the discharge of a rare flood, a flood with a 1-percent likelihood of occurring in a given year, then modeling the land surface area that would be inundated by the floodwater.

Flow frequency analysis is the process of determining the statistical likelihood a flood of a particular magnitude occurring in a given year.

Flow frequency analysis involves fitting a statistical distribution to the peak discharge series, the largest annual instantaneous discharge over the period of record, the time over which stream discharge was measured.

The statistical distribution is fitted to the peak discharge series using the skewness coefficient of the peak discharge series at a particular gage, a stream discharge recording station and a regional skewness coefficient, an average of the skewness coefficients of similar gages.

Two major sources of uncertainty in flow frequency analysis are: 1) the peak discharge of a rare flood is substantially larger than a majority of the peak discharge over the period of record, and 2) because the available period of record is short, most often 20 to 30 years, a rare flood might not occur within the period of record

To reduce uncertainty

A regional skew coefficient is added to

# approach,

# discussion.

# Next Steps

https://www.google.com/books/edition/Natural\_Hazards/\_TpbEAAAQBAJ?hl=en&gbpv=1&dq=floods+as+hazards&pg=PA21&printsec=frontcover

## Overview

Regional skew coefficients are necessary to determine flood frequency following [Bulletin 17C](https://pubs.usgs.gov/publication/tm4B5) guidelines. The goal of FFA\_regional-skew is to estimate a regional skew coefficients for stream gages in the Northwestern Great Plains and northern portion of the High Plains sections of the Great Plains within southeastern Montana, eastern Wyoming, northeastern Colorado, and the western portions of Nebraska and the Dakotas.

<!--

the [level-3 ecoregions](https://www.epa.gov/eco-research/level-iii-and-iv-ecoregions-continental-united-states) for peak flow gages in

West-central Prairie ecoregion \*\*add link to level 3 ecoregion maps\*\*.

-->

## Study Area

<!--

Bounding box needs to extend further south than the Nebraska border

-->

The study area encompasses the Northwestern Great Plains and the northern portion of the High Plains ecoregion. Northwestern Great Plains ecoregion encompasses the Missouri Plateau section of the Great Plains in southeastern Montana, northeastern Wyoming, and the western portion of the Dakotas. The northern portion of the High Plains ecoregion encompasses southeastern Wyoming, western Nebraska, eastern Colorado, and western Kansas (http://www.cec.org/files/documents/publications/10415-north-american-terrestrial-ecoregionslevel-iii-en.pdf).

The study area is a dry mid-latitude steppe climate marked by hot summers and cold winters, and a mean annual temperature varying by latitude.

<!--

The mean annual temperature of the Northwestern Great Plains ranges from of approximately $5C$ in some northern areas to 8.5C in the south.

The mean annual temperature of the High Plains ranges from approximately 8C in the north to 17C in the far south.

Northwestern Great Plains--The frost-free period ranges from 90 days to 155 days. The mean annual precipitation is 393 mm, ranging from 250 to 510 mm."

High Plains--The frost-free period ranges from 120 to 230 days. The mean annual precipitation is 433 mm, and ranges from 305 to 530 mm.

NGP Vegetation: Grasslands persist in rangeland areas, especially on broken topography, but have been replaced by cropland on some areas of level ground. Shortgrass and mixedgrass prairies contain blue grama, western wheatgrass, green needlegrass, prairie sandreed, and buffalograss. There are areas of sagebrush steppe with fringed sage, Wyoming big sagebrush, rabbitbrush, and sand sagebrush; some areas have scattered ponderosa pine and Rocky Mountain juniper.

HP Vegetation: Historically, the region had mostly short and midgrass prairie vegetation; much of it is now greatly altered. Shortgrass prairie featured blue grama, buffalograss, and fringed sage, and mixed grass areas had sideoats grama, western wheatgrass, and little bluestem. Sandsage prairies had sand sagebrush, sand bluestem, prairie sandreed, little bluestem, Indian ricegrass, and sand dropseed. Shinnery sands areas in the south featured Havard shin oak, fourwing saltbush, sand sagebrush, yucca, and mid- and shortgrasses.

NGP Hydrology: Mostly ephemeral and intermittent streams are found here, with a few larger perennial rivers that cross the region from the western mountains. Many small impoundments occur, and there are some large reservoirs on the Missouri River.

HP Hydrology: Mostly intermittent and ephemeral streams prevail here. A few larger rivers that originate in the Southern Rockies (6.2.14) cross the region, such as the Platte, Arkansas, and Cimarron. The southern portion has few to no streams. Surface water there occurs in numerous ephemeral pools or playas. These serve as recharge areas for the important Ogallala Aquifer. Water withdrawals from the aquifer usually exceed recharge, however.

NGP Terrain: The region is an unglaciated, rolling plain of shale and sandstone punctuated by occasional buttes. Some areas are of dissected, badland terrain and river breaks. Entisols, Mollisols, Aridisols, and Inceptisols occur. Frigid and mesic soil temperature regimes and ustic and aridic soil moisture regimes are typical.

NGP Land Use/Human Activities: The region's grassland and shrubland are used for livestock grazing, mostly of cattle and sheep. Agriculture is restricted by the erratic precipitation and limited opportunities for irrigation. Some areas grow wheat, alfalfa, and barley. A few areas are used for coal mining. Larger settlements include Billings, Lewiston, Livingston, Miles City, Dickinson, Mandan, Belle Fourche, Pierre, Rapid City, Sheridan, Gillette, and Casper.

-->

```{r map-of-study-area, eval=FALSE}

# Make a map

# Here is some example code:

# https://www.r4wrds.com/intro/m\_intro\_mapmaking

```

# Introduction

Ecoregion membership was identified as a key predictor of hydrologic similarity in south-central South Dakota (Tinant, PhD dissertation).