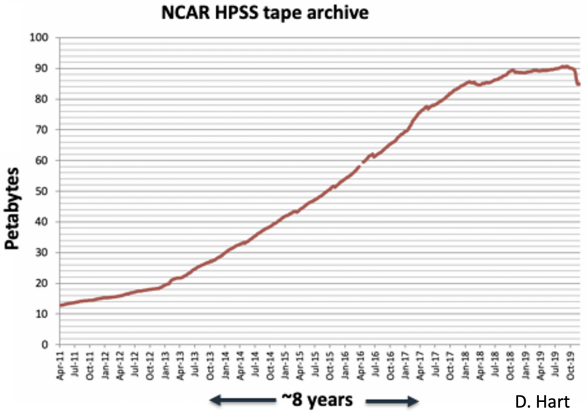
MATH 561 Final Project

Colorado School of Mines
Applied Mathematics & Statistics

November 18, 2021

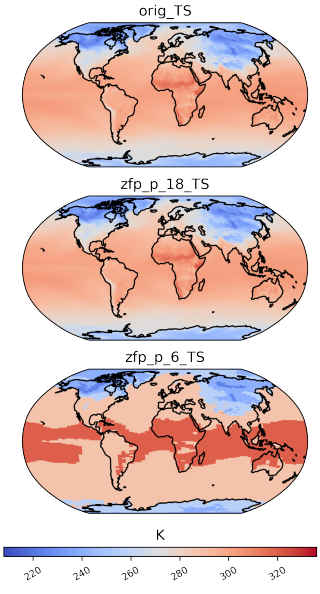
Background - Data Storage Problem

- Climate model simulation can produce extremely large output data files.
- Community Earth System Model - Large Ensemble (CESM-LE) climate model created by the National Center for Atmospheric Research is estimated to produce 10 petabytes of climate model data.
- Scientists at NCAR are interested in applying lossy compression to the data to reduce file size.



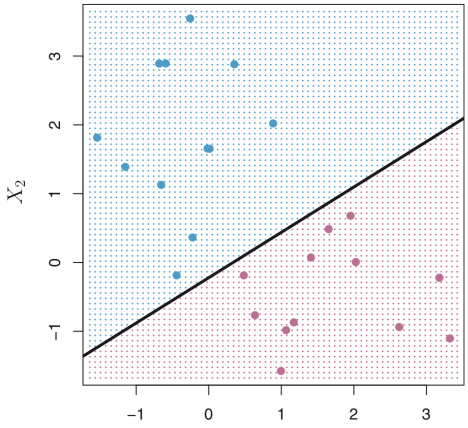
Background - Lossy Compression

- Traditional lossless compression techniques are not effective on floating-point data.
- Instead, use lossy compression to achieve larger reduction in storage size.
- Optimal compression for a particular dataset means that the data size has been reduced by as much as possible without affecting scientific conclusions.



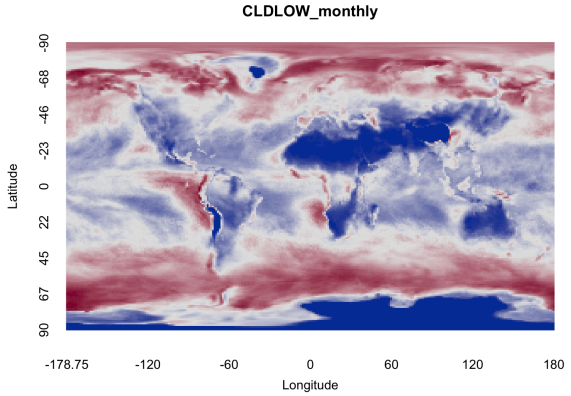
Problem Statement

- You will be given climate model datasets and the corresponding optimal compression levels.
- Your job is to first identify features of the data that may be useful in predicting the optimal level.
- Then use these features as predictors in classification models to predict the optimal compression level.



Data

- The data is located in Canvas under Files -> Final Project
- These data files contain ten time steps each from a selection of CESM-LE variables.
- The datasets are stored as R matrices, where the columns represent equally spaced longitudes and the rows represent equally spaced latitude coordinates.



Using the Data

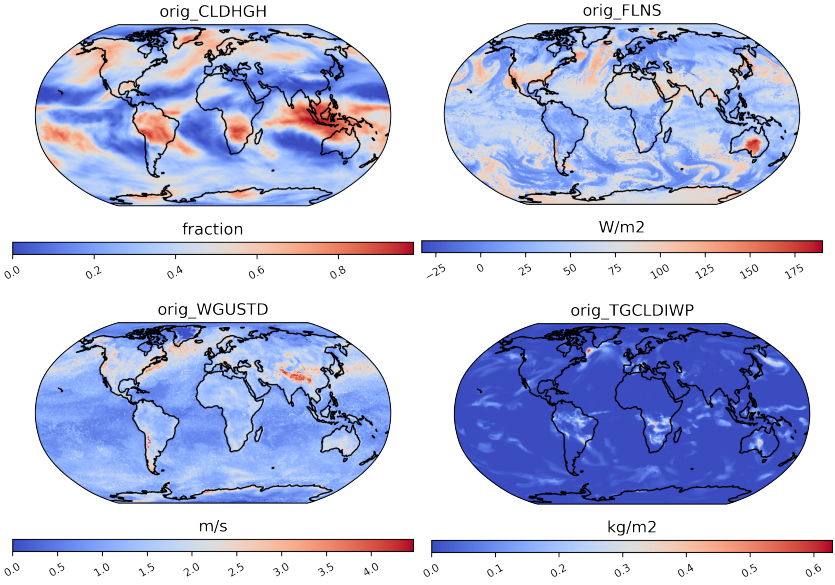
- Plotting matrices can be done using the built-in R function image()
- Features and classifications should be placed into a data frame before applying statistical techniques.
- A tutorial script is available in Files -> Final Project, showing how to load, plot, and create features from the data.

```
load("~/Desktop/high_train.RData")

# Select data and axis labels
n=170
selected_dataset <- high_train
l <- length(selected_dataset$mat)
latitudes <- round(as.numeric(colnames(selected_dataset$mat[[n]])))
longitudes <- as.numeric(rownames(selected_dataset$mat[[n]]))
dataset <- selected_dataset$mat[[n]]
varname <- selected_dataset$var[n]

# Plot
image(dataset, main=varname, col = hcl.colors(100, "Blue-Red"), axes
=FALSE, xlab="Longitude", ylab="Latitude")
axis(3, at=seq(0,1, length=7), labels=longitudes[seq(1, 288, length.out
=7)], lwd=0, pos=-0.2, outer=T)
axis(2, at=seq(1,0, length=9), labels=latitudes[seq(1, 192, length.out=9
)], lwd=0, pos=0)
```

Variable Plots

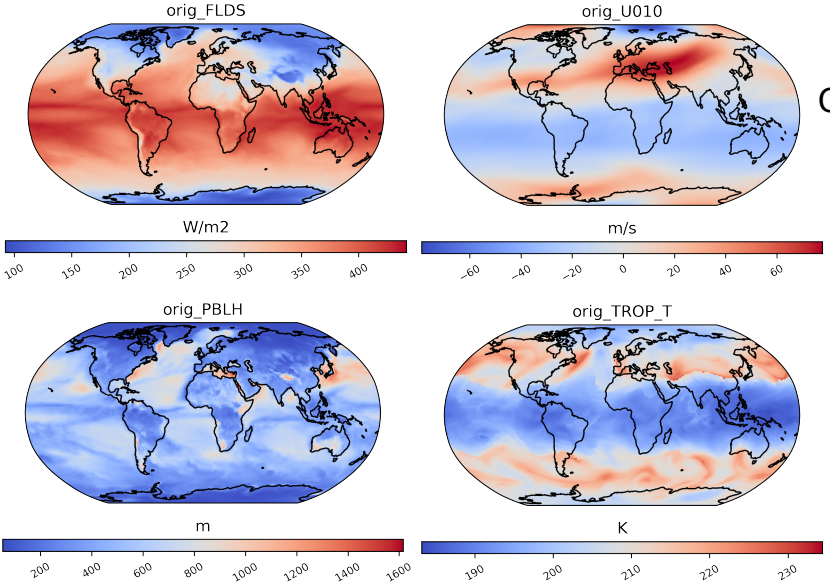
- Visual inspection of the data may help inspire ideas for features to use in your models. The next three slides show variables that are optimally compressed at high, medium, and low levels, respectively.

Variables - High Compression



These datasets can be highly compressed without altering the data quality.

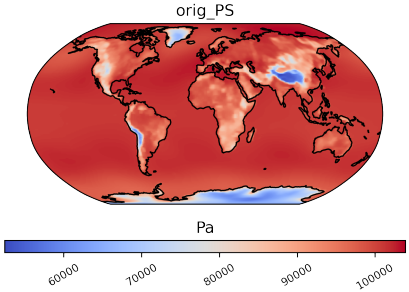
Variables - Medium Compression

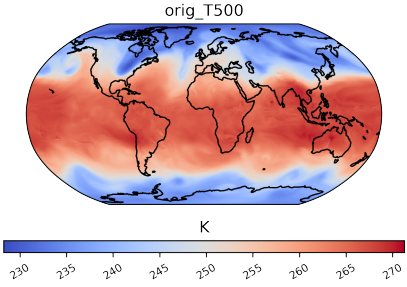


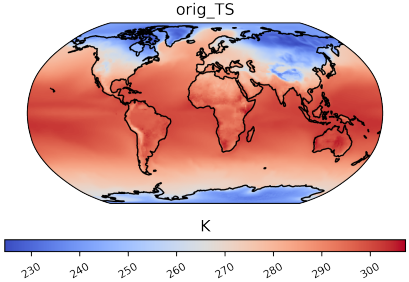
Columns are longitude
Rows are latitude

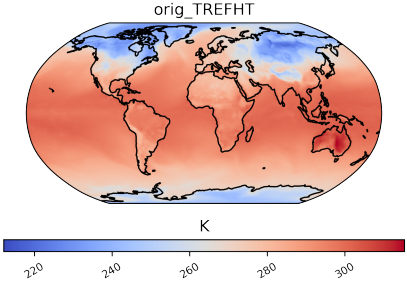
These datasets can be moderately compressed with lossy compression.

Variables - Low Compression









These datasets can only be lossily compressed a little (or not at all) without significantly affecting the data.

Requirements

- You are required to apply at least three classification models to the data to come up with predictions of the optimal compression level for a separate test set of data.
- Two of these should include methods discussed in the class such as LDA or regression trees. Your other model(s) may include other methods such as neural networks.
- Your grade will partly be based on the classification accuracy of your methods on the test data, so using appropriate methods and selecting useful features are both essential.

Training, Validating, Testing Models

- Each variable has been preassigned to be part of either the training, validation, or testing data.
- These data are available through Canvas.
- Training data should be used to build your models, and the validation data can be used to assess the classification accuracy of each model.
- Testing data is provided without optimal classifications. You will make predictions for these data and include them in your final report.

Questions

Any questions?