Today’s Texas Might be Tomorrow’s Ohio:

Building a Geographic Climate Change Predictor

Alison Duck, Jack Neustadt,
David Pochik, and Tawny Sit

The Team

Jack Neustadt, PhD
Astronomy Postdoc
JHU (OSU grad)

David Pochik
Final Year PhD Candidate at OSU
Physics

Tawny Sit
3rd year PhD Candidate
OSU, Astronomy

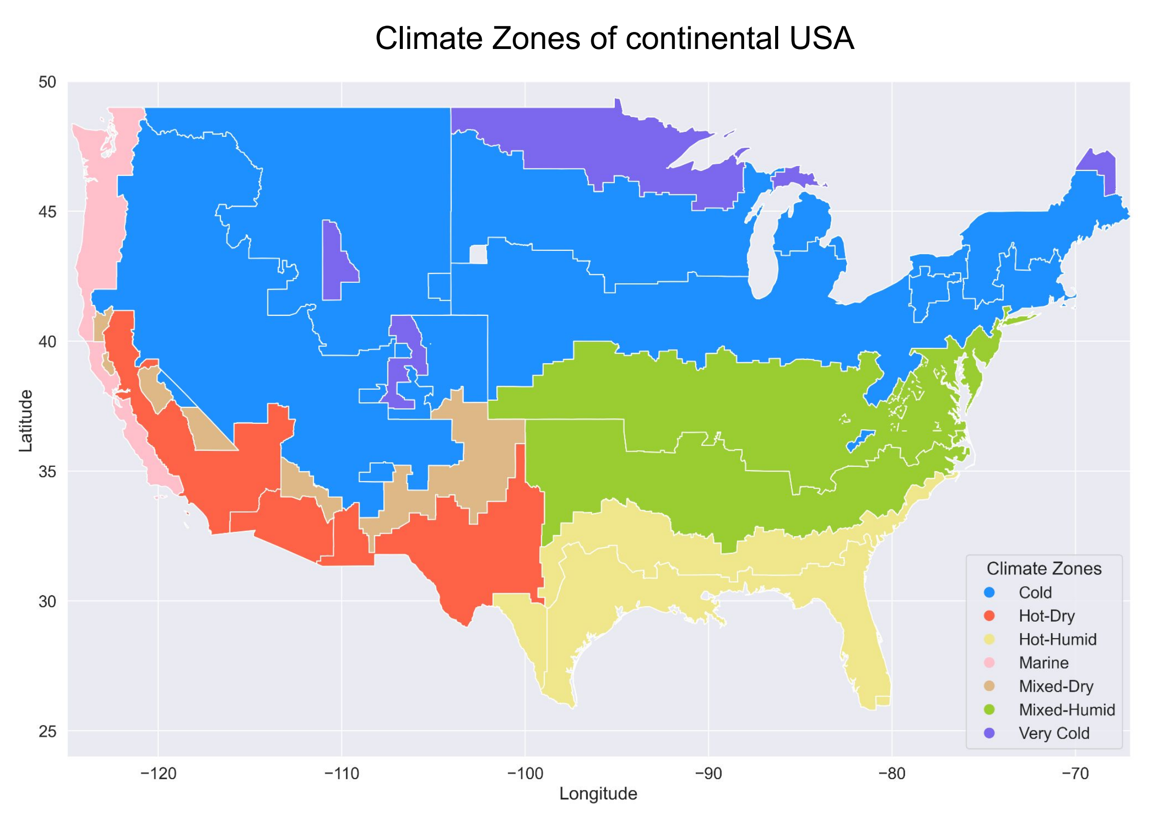
Alison Duck
Final Year PhD Candidate
OSU, Astronomy

Project Goals

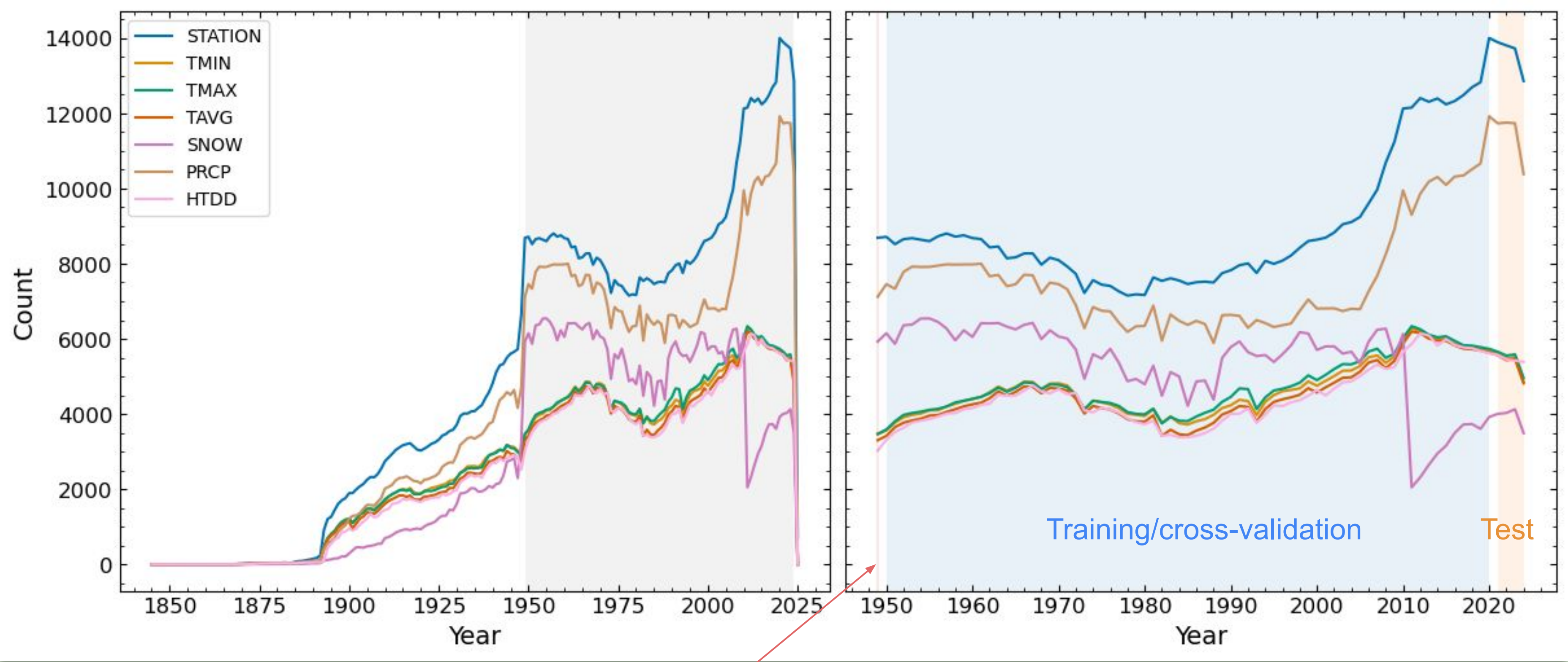
"Where should I move to in the future to experience the same climate that I enjoy today?"

Where are you *now*?

- We employed the Building America Climate Regions
- Developed by U.S. Department of Energy Researchers at the Pacific Northwest National Laboratory
- Delineates regions on the county level
- Temperature and moisture information required to understand the evolution of these regions



The NOAA Global Summary of the Year



1949 only used for K-means clustering training

Method 1: Latitude/Longitude Structured Gridding

Inputs

- Number of latitude and longitude grid lines
- Minimum number of weather stations per grid cell
- Timeline

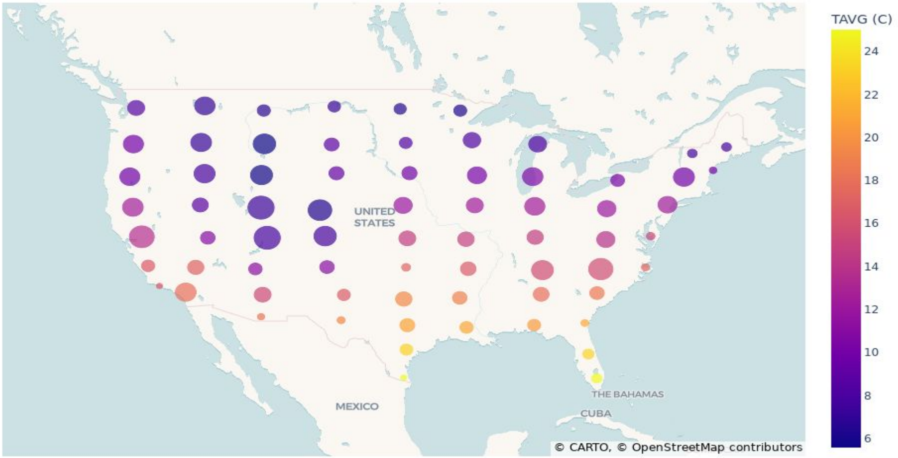
Outputs

- Returns feature average for all grid cells for a given year that meet the minimum weather station requirement

Purpose

- Time series analysis: form train/test split with averaged data
- Perform n-fold cross validation in train set
- Compute linear regression model at each cross validation step.

Stations from 2024 in a 10x10 structured grid with cell-averaged coordinates (size = # of stations)



Method 2: Binning with K-Means Clustering and KNN

**K-Means Clustering:
Identify geographic regions**

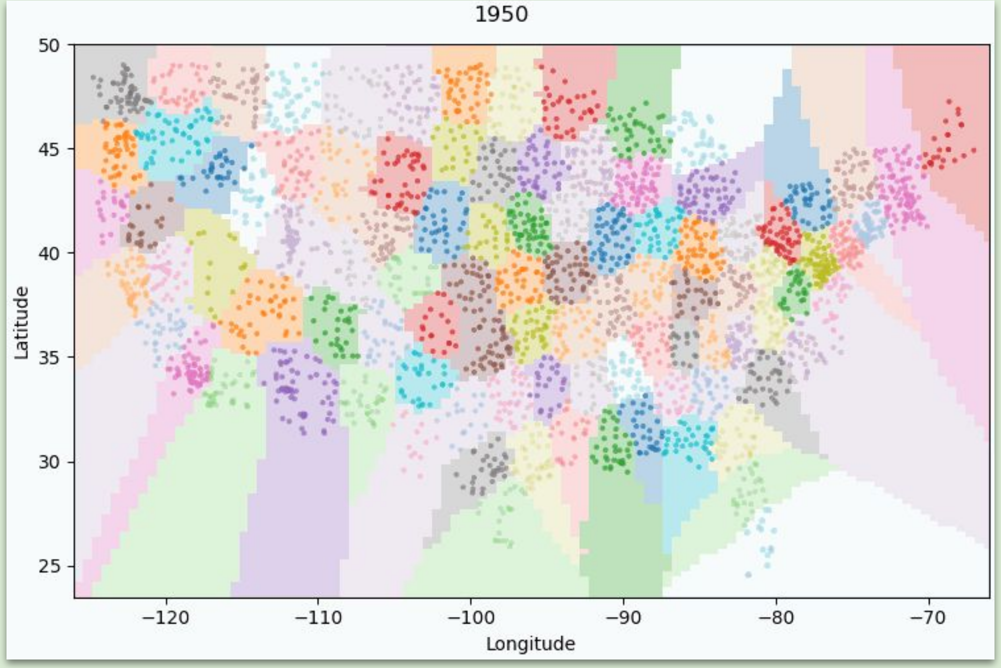
- *k* = 100
- Use station coordinates from 1949

**KNN Classification:
Place stations in regions**

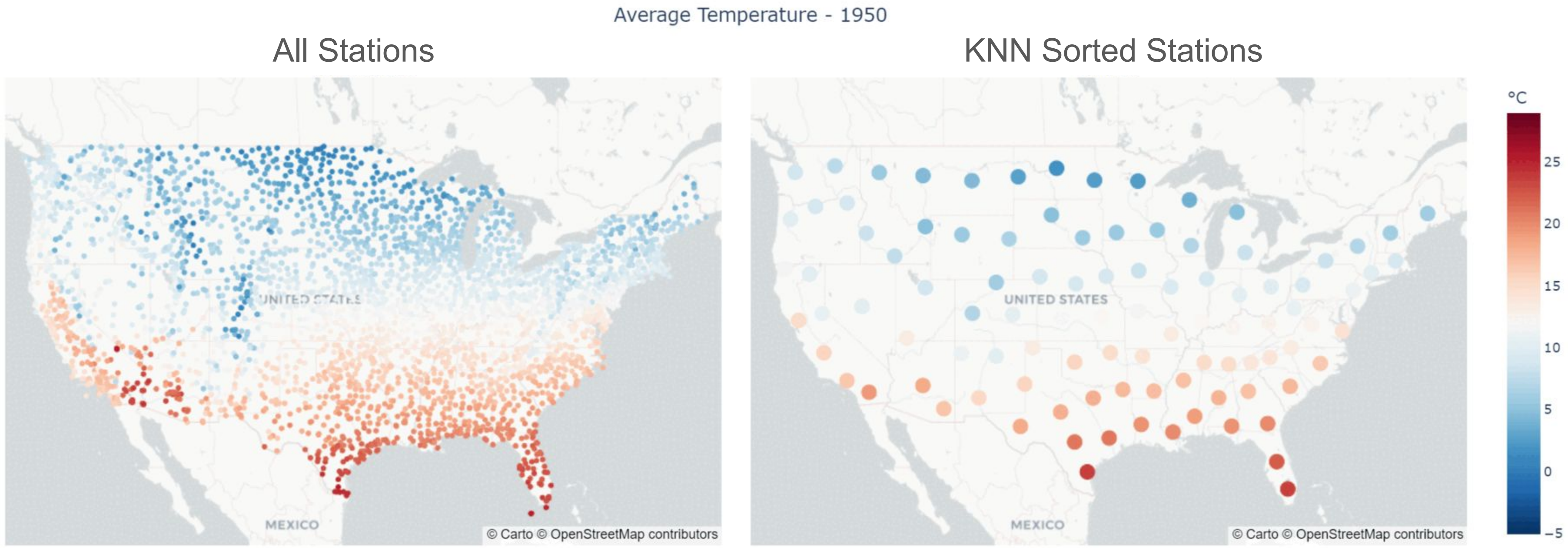
- Train classifier on K-means clusters
- From 1950 onward, predict region each station belongs to
- Decision boundaries consistent from year-to-year: only tiny deviations at US geographical boundaries

Calculate feature averages over all stations in a region for each year

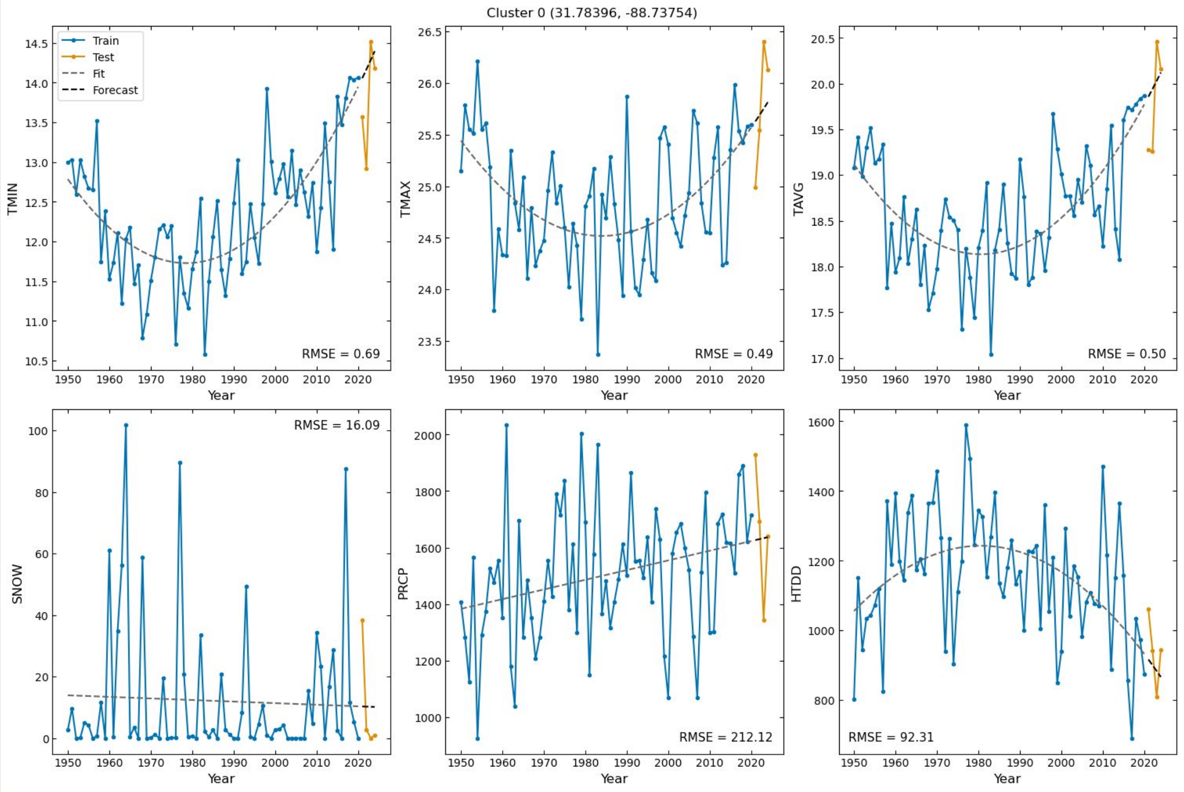
- Account for missing data in any given year
- Regress region-by-region on yearly averages



Binning Preserves Geographic Temperature Gradients



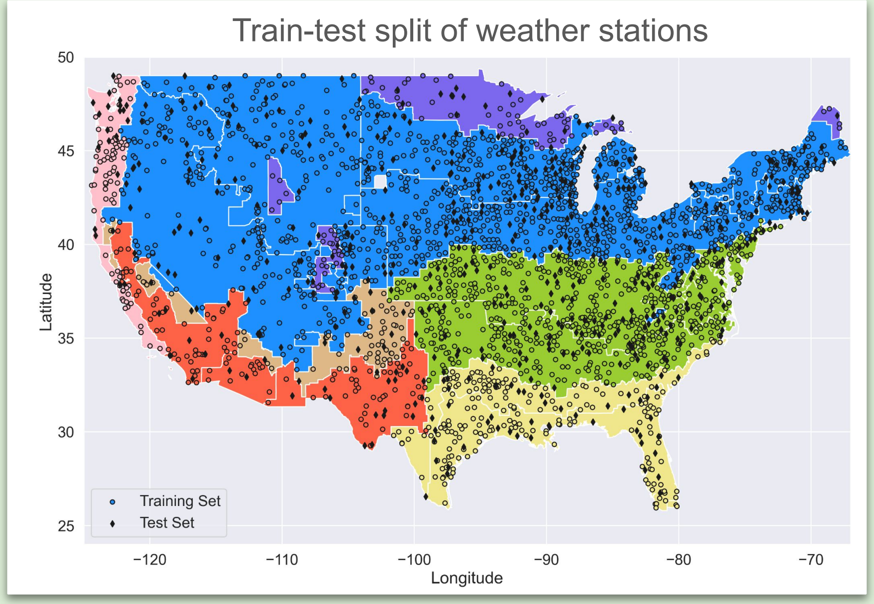
Feature Regression Models For Each Bin



- Binned data is just noisy, no clear seasonality
- Two models: **linear** and **quadratic** trends
  - Fit each feature independently
- Model selection: RMSE comparison of final cross-validation set (2011-2020)
  - Preferred model changed with choice of validation set: we need the most recent data possible!

Classification Methodology

- Individual weather stations** used to capture geography of climate zones
- 80/20 (zone-stratified) train-test split, then 10 k-fold validation sets



Mean scores of validation sets

Model	Accuracy	F1	Precision	Log-loss
Logistic Regression	78.8	80.5	85.5	0.59
LDA	84.4	83.9	84.5	0.478
kNN	88.3	87.6	88.5	0.742
Random Forest	89.8	89.5	89.8	0.386

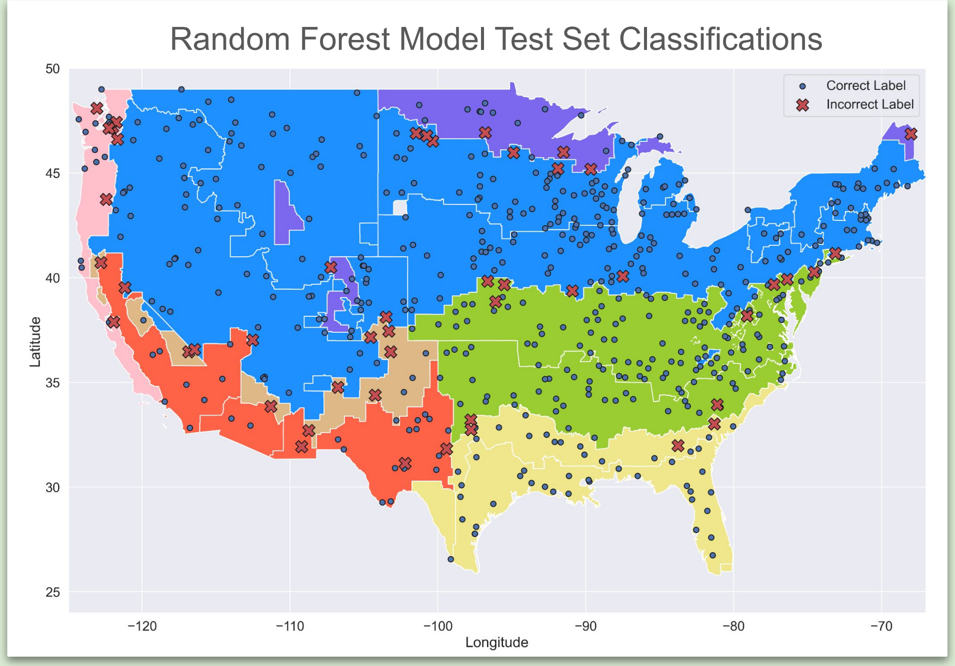
Scores for test set

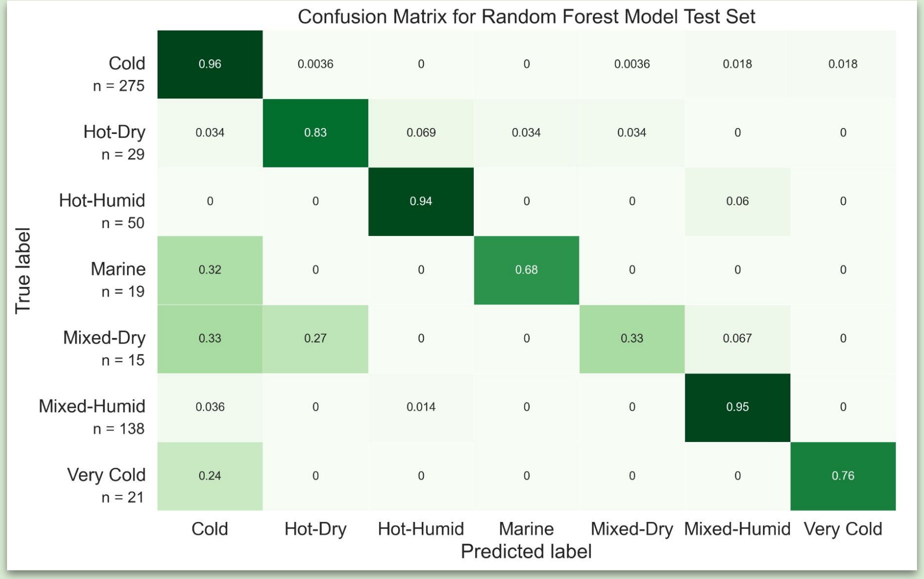
Random Forest	91.2	90.8	90.9	0.435

- Random Forest performs best!

**from 2010, the year with the most complete data

Results with Climate Zones





- Marine and Mixed-Dry Climates defined by seasonal data (which we don’t have)
- Classification fails in border regions (where geography is important)

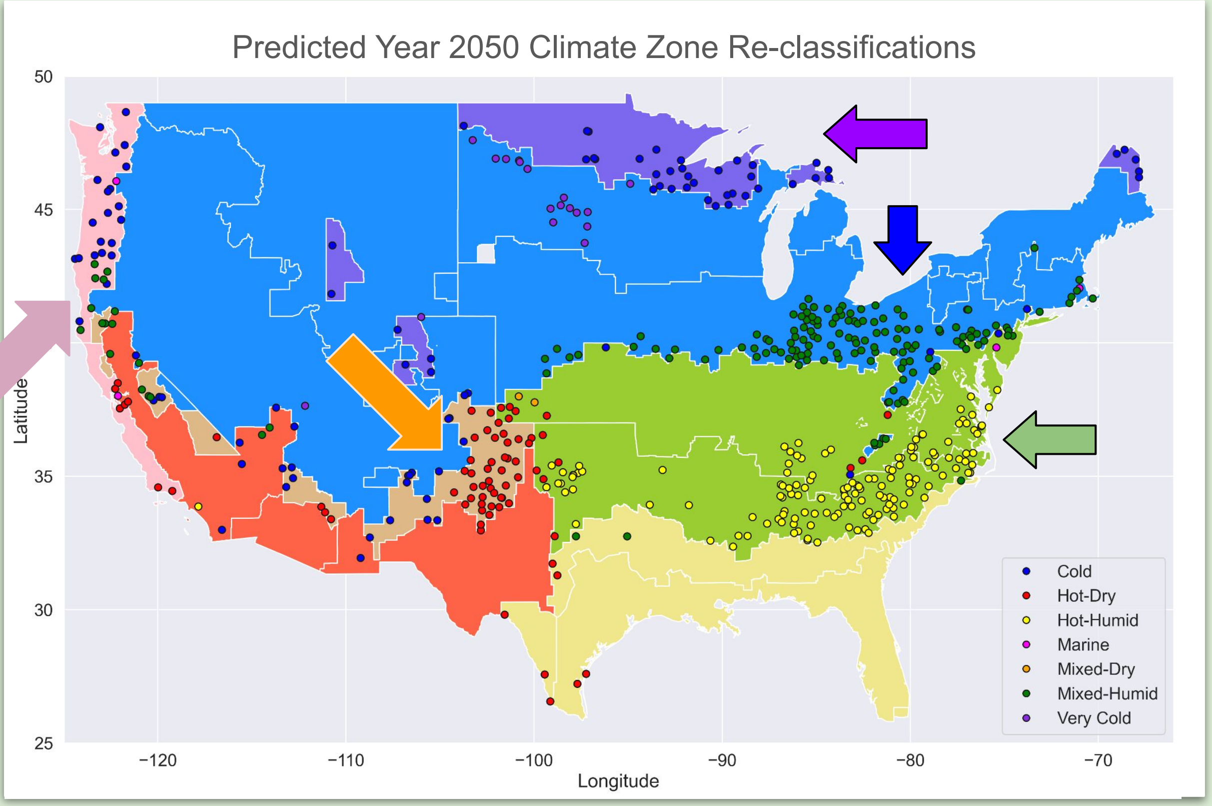
Where Will You Be *in 25 Years*?

Northeast & Midwest
Cold to **Mixed-Humid**
Very Cold to **Cold**

Coastal South
Mixed-Humid to **Hot-Humid**

Great Plains
Mixed-Dry to **Hot-Dry**

Marine to **Cold** is probably classifier error (based on confusion matrix of test set)



Future Work

- Matching our predicted climate change trends to individual weather stations required certain assumptions that we could have probed more to see how/if results changed
- Latitude and longitude gridding showed greater errors for snowfall and precipitation, which may be improved upon by further input parameter optimization.
- One future extension would be to use our climate change models to build a predictive model for household energy costs.