# REDUCED ORDER THERMOKARST INITIATION MODELS WITH RANDOM FORESTS

Rawser Spicer

# Committee

- Dr. Orion Lawlor
- Dr. Bob Bolton
- Dr. Glenn Chappell

# Goal

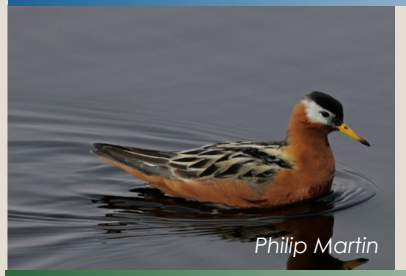
Can a reduced order model be built from the Alaska Thermokarst motel's (ATM) thermokarst initiation model the has similar results to the full model, but with fewer inputs.

# Background

Understanding thermokarst initiation and resulting changes to arctic landscape are important to understanding changes in vegetation, wildlife, drainage patterns, and the carbon & energy regimes. These changes are at the intersection of climate, ecosystem, hydrology and permafrost sciences, and are important to wildlife managers and development projects amongst others.

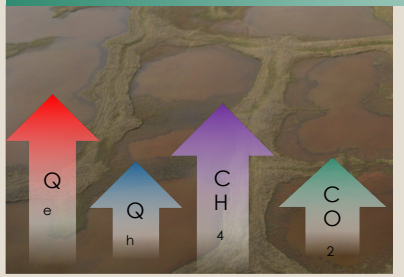
*A: Vegetation*

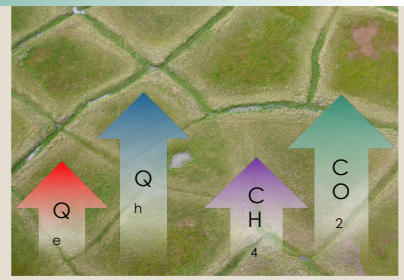
*B: Wildlife*



*C: Drainage*

*D: Energy Regime*

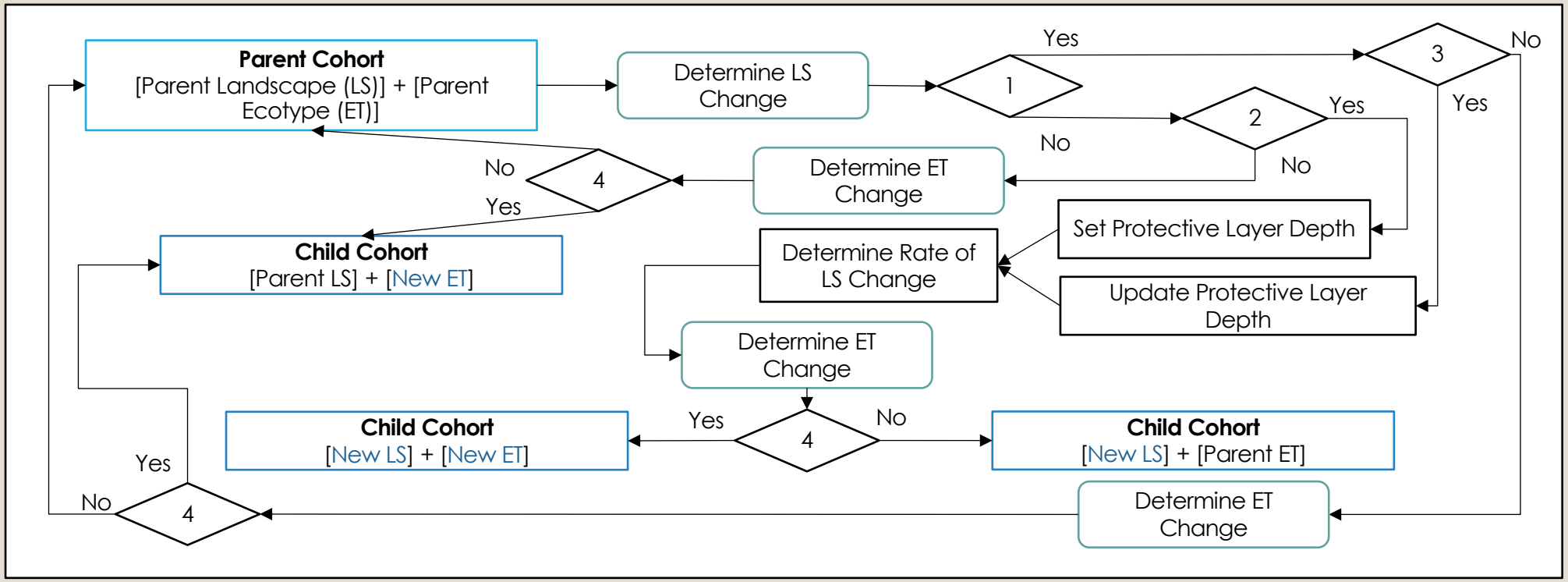


# The Alaska Thermokarst Model

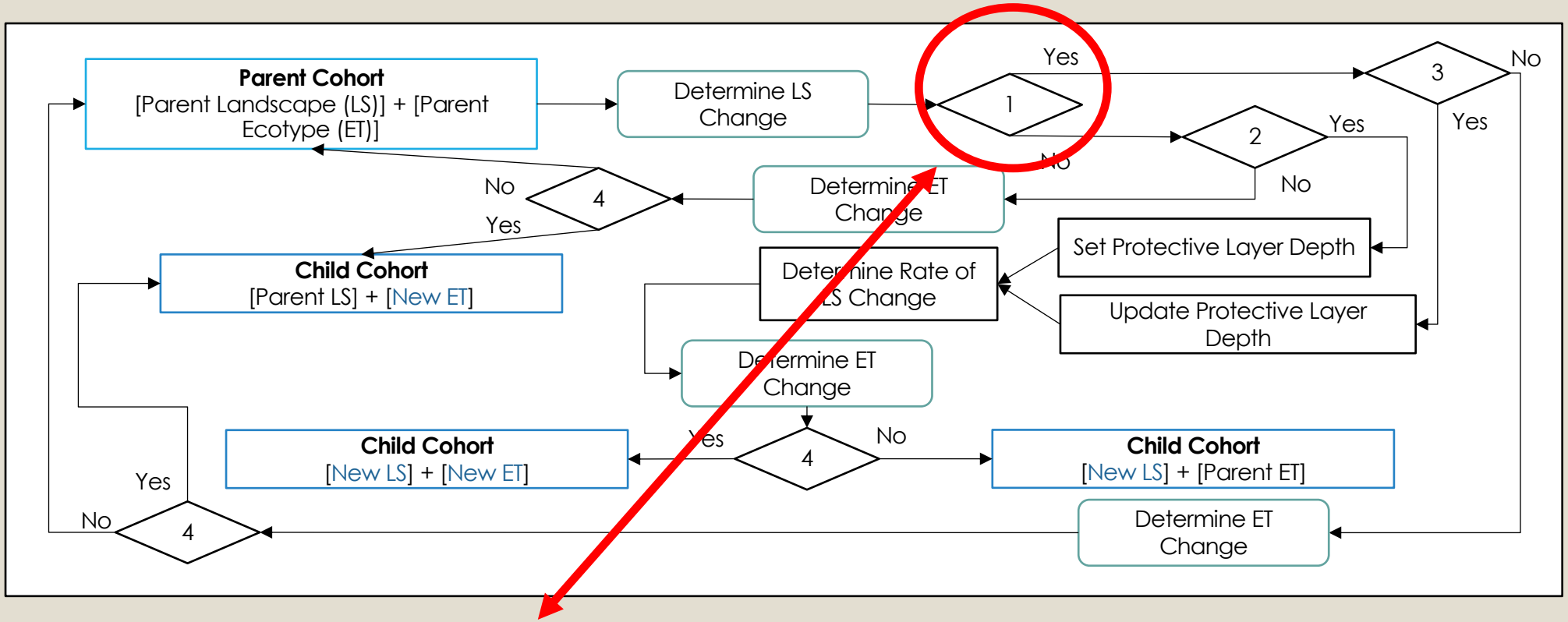
- Models landscape change due to changes caused by thermokarst cycle
- Landscapes transition from meadows to low center polygon covered to high center polygonal cover to ponds and lakes
- Landscape units (cohorts) are tracked on a gridded subdivision of the model area
- Logic-based rules are used to determine when the landscape transitions from its current cohort to the next cohort

# Alaska Thermokarst Model Flow Chart

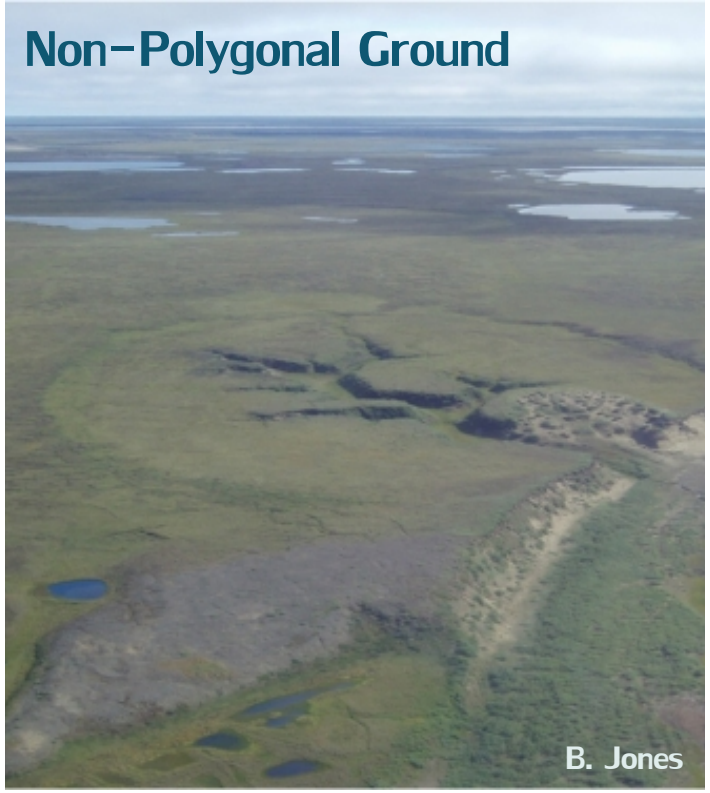


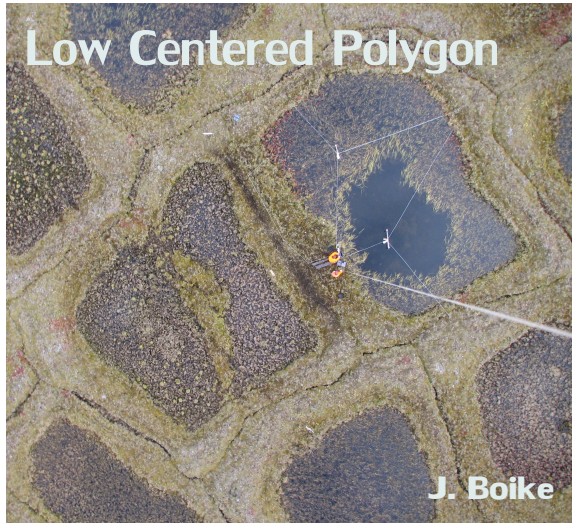

1. Is the thermokarst active?
2. Has the thermokarst been initiated by a extreme event?
3. Is the active layer > protective layer?
4. Does climate support new ecotype?

# Project Focus




1. Is the thermokarst active?
2. **Has the thermokarst been initiated by a extreme event?**
3. Is the active layer > protective layer?
4. Does climate support new ecotype?



# Permafrost and Thermokarst

- Permafrost occurs when soils remain continuously frozen for many years.
- During summer the top layer of permafrost thaws this is called the active layer.
  - active layer depth is the depth to the top of the permafrost
- Thermokarst features are caused by thawing ground Ice.
- Thermokarst features include polygonal patterned ground, and thermokarst lakes

# Thermokarst Initiation

- What causes the thermokarst process to begin?
  - It is thought that thermokarst initiation is caused by extreme climate conditions.
- Once thermokarst process has begun it takes 20-30 years for a landscape to stabilize

# Current Method

- Look for regions where the temperature and precipitation have been extremely greater than the average for a winter and into the next summer.
- Based on field observations on the Seward Peninsula
- Use the 1901-1950 average
- Look at:
  - Summer temperature
  - Early winter precipitation (Oct-Nov)
  - Total winter precipitation (Oct-Mar)
  - Winter temperature
  - The following summer temperature
- Data are in raster format for years 1901-2015
- Combine these results with map of thermokarst predisposition.

# Current Method

Let $Y$ be years from $y_s$ to $y_e$

$y_{em} \in Y$ is the end of the period used for calulating means

Let $t$ be thermokarst predispositioin value [0,1], or set as one if not applying this

Let $V$ be set of variables with values for $y \in Y$

Let $M_v$ be $\frac{1}{Y_{em} - Y_S}\left(\sum_{y_i = y_s}^{y_{em}} v_{y_i}\right)$ for each $v$ in $V$

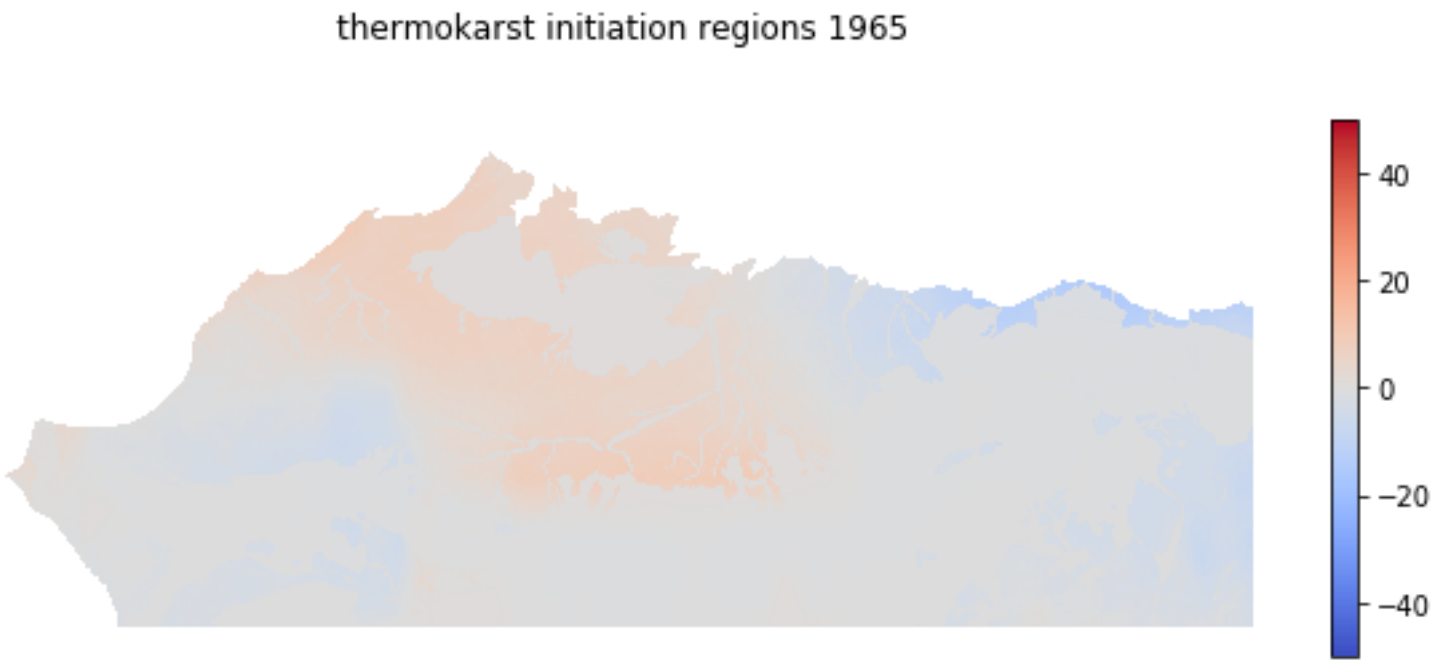
Let $PD_v$ be $\left\{\left(\left(V_y - M_v\right)/|M_v|\right) * 100 \text{ for } y \in Y\right\}$ for $v \in V$

Let $TKI_y$ be $\frac{t}{n_V}\sum PD_{v_y}$

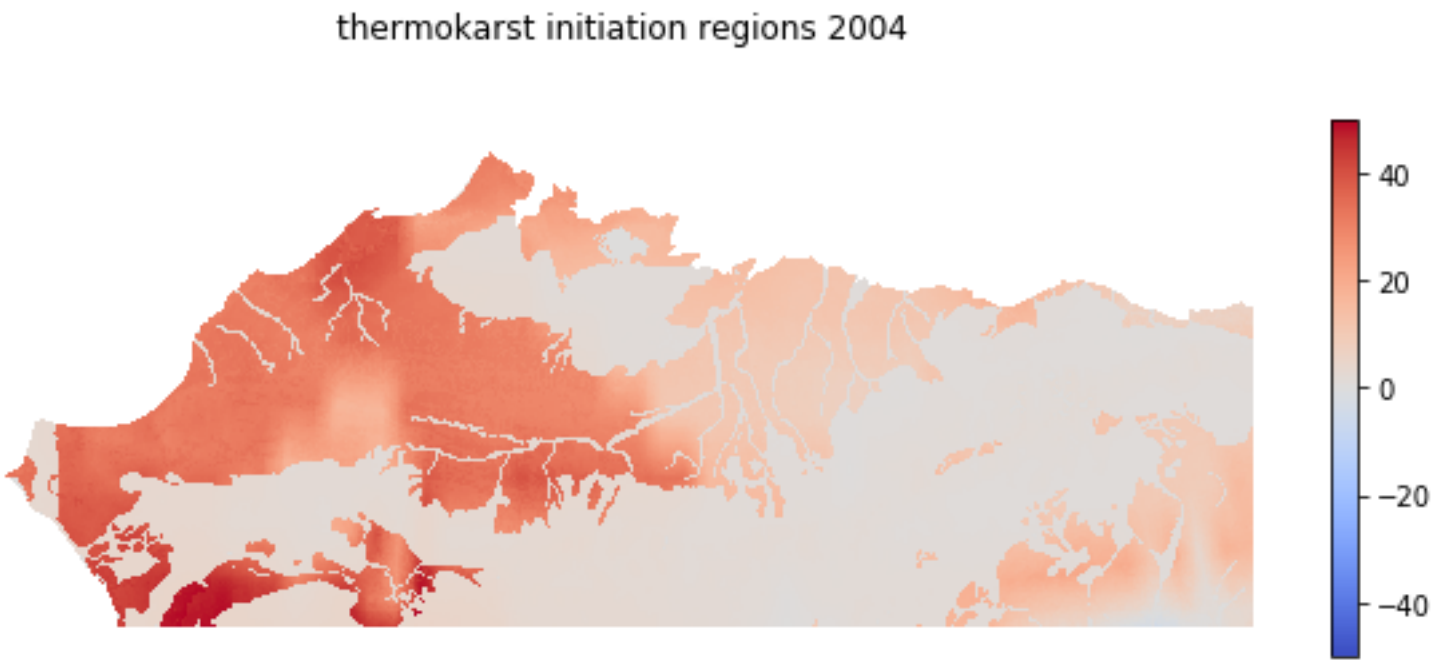
# Current Method

Algorithm gives a numerical value were higher values mean higher likely hood of thermokarst initiation.

A threshold is set as the minimum thermokarst initiation value




# CURRENT METHOD EXAMPLE 1965




# CURRENT METHOD EXAMPLE 2004

# Project Goals

- Use random forests to determine which features are the most important variables to thermokarst Initiation, and develop a reduced order model from those
- Compare new model to original model

# Random Forests

- Ensemble learning method
- The first algorithm for random decision forests was created by Tin Kam Ho
- Further developed by Leo Breiman and Adele Cutler
- Designed to correct the overfitting that may occur in regular decision tree models
- Constructs multiple decision trees during training, and when testing averages the result from each tree to find the final result
  - Average is used when doing regression, mode when classifying

# Methodology for Random Forests

## Train:

- Train Random forest model using a subset of features that may influence prediction and predictions from the existing model.

## Test:

- Use the trained model to make predictions based on full set of features

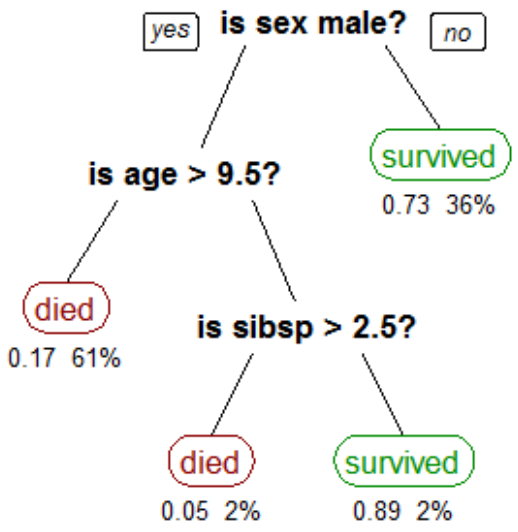
## Verify:

- Compare random forest predictions to existing model predictions

# Decision Trees

- The goal is to create a model that predicts the value of a target variable based on several input variables
- A tree is built by splitting the source set, constituting the root node of the tree, into subsets - which constitute the successor children. The splitting is based on a set of splitting rules based on classification features

**Probability of Survival on Titanic**



# Bootstrap Aggregating (Bagging)

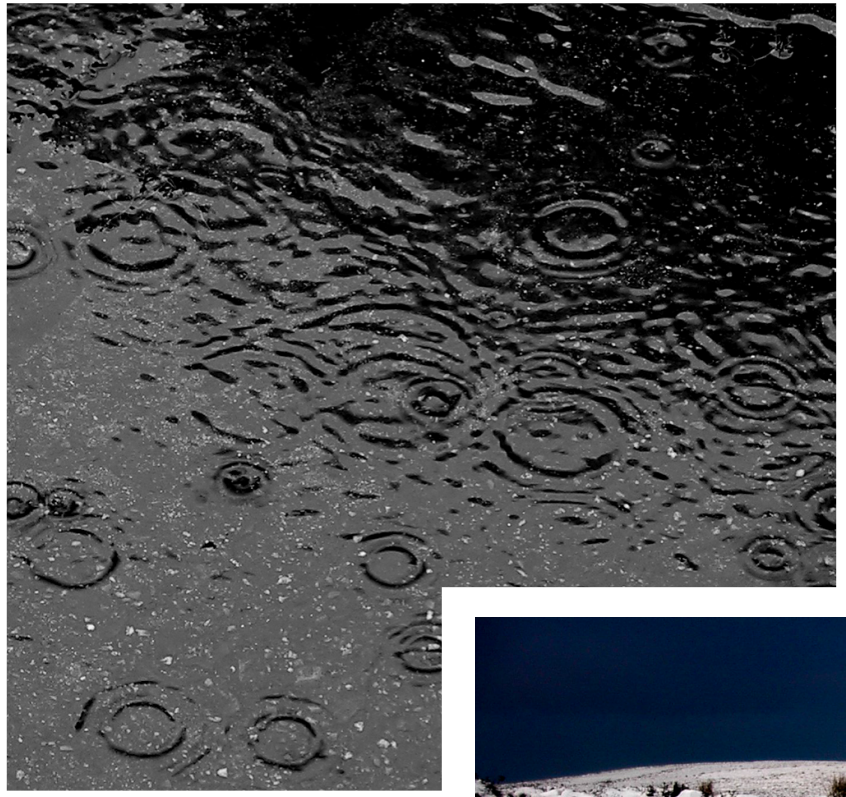
- Given a standard training set D of size n, bagging generates m new training sets $D_i$, each of size n′, by sampling from D uniformly and with replacement.
- This is used in random forests to select a random subset of features at each split in the learning process
  - sometimes called feature bagging

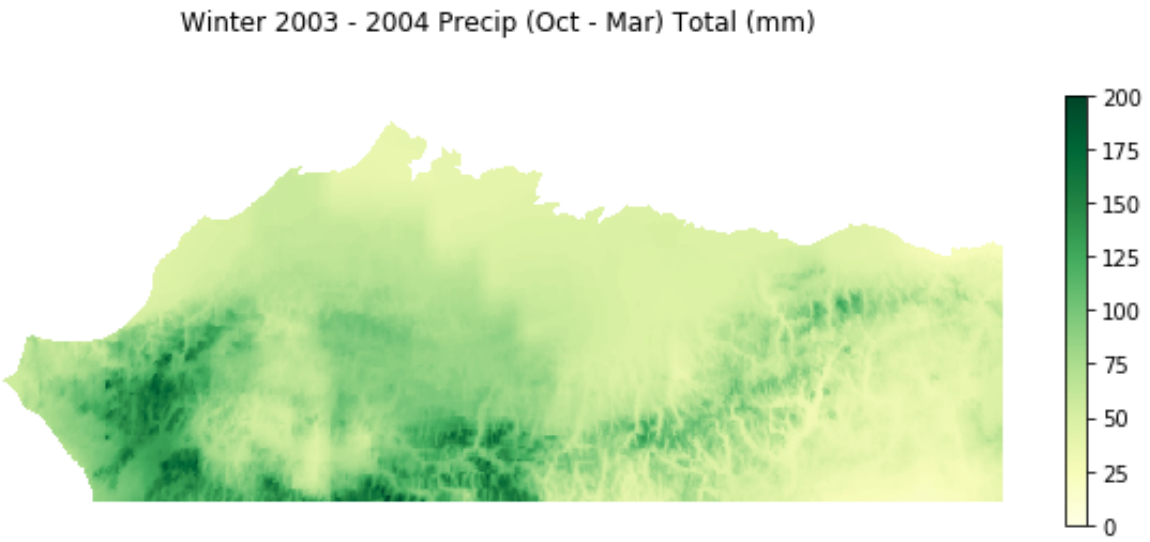
# Thermokarst Initiation: features

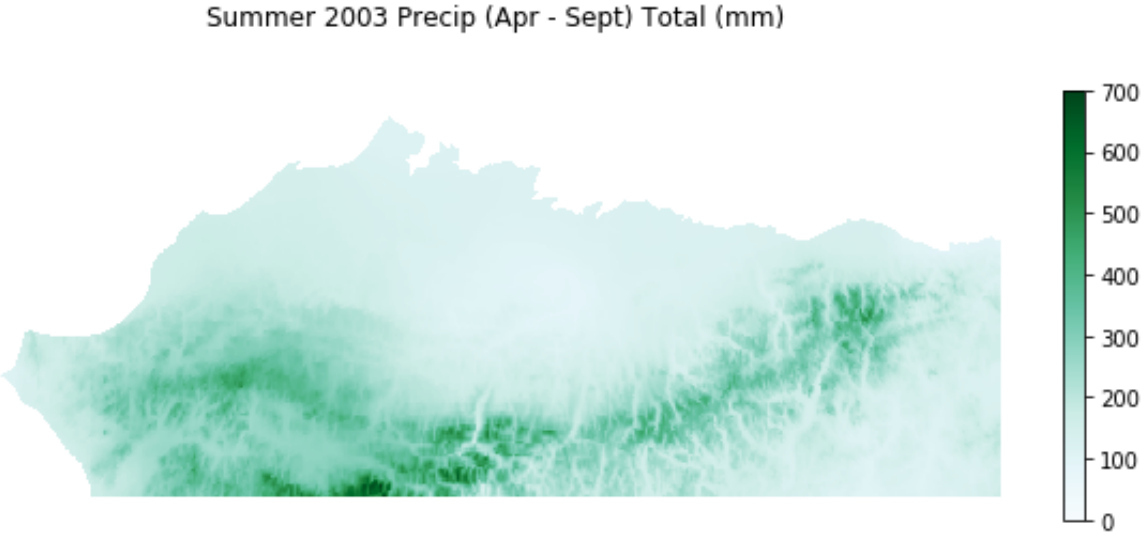
- Features selected for training thermokarst initiation random forests
- Two types of features
  - climate based
  - physically based

# Precipitation Features

- Precipitation is measured in mm
- We will look at early winter precipitation for year N, full winter precipitation spanning yean N and N+1 , summer precipitation in year N and N+1, and late summer precipitation in year N.
- It is thought that high winter precipitation as snowfall insulates the ground which causes a slowing in the refreezing of the active layer
- Summer precipitation can increase thawing of permafrost, and add heat into soil.

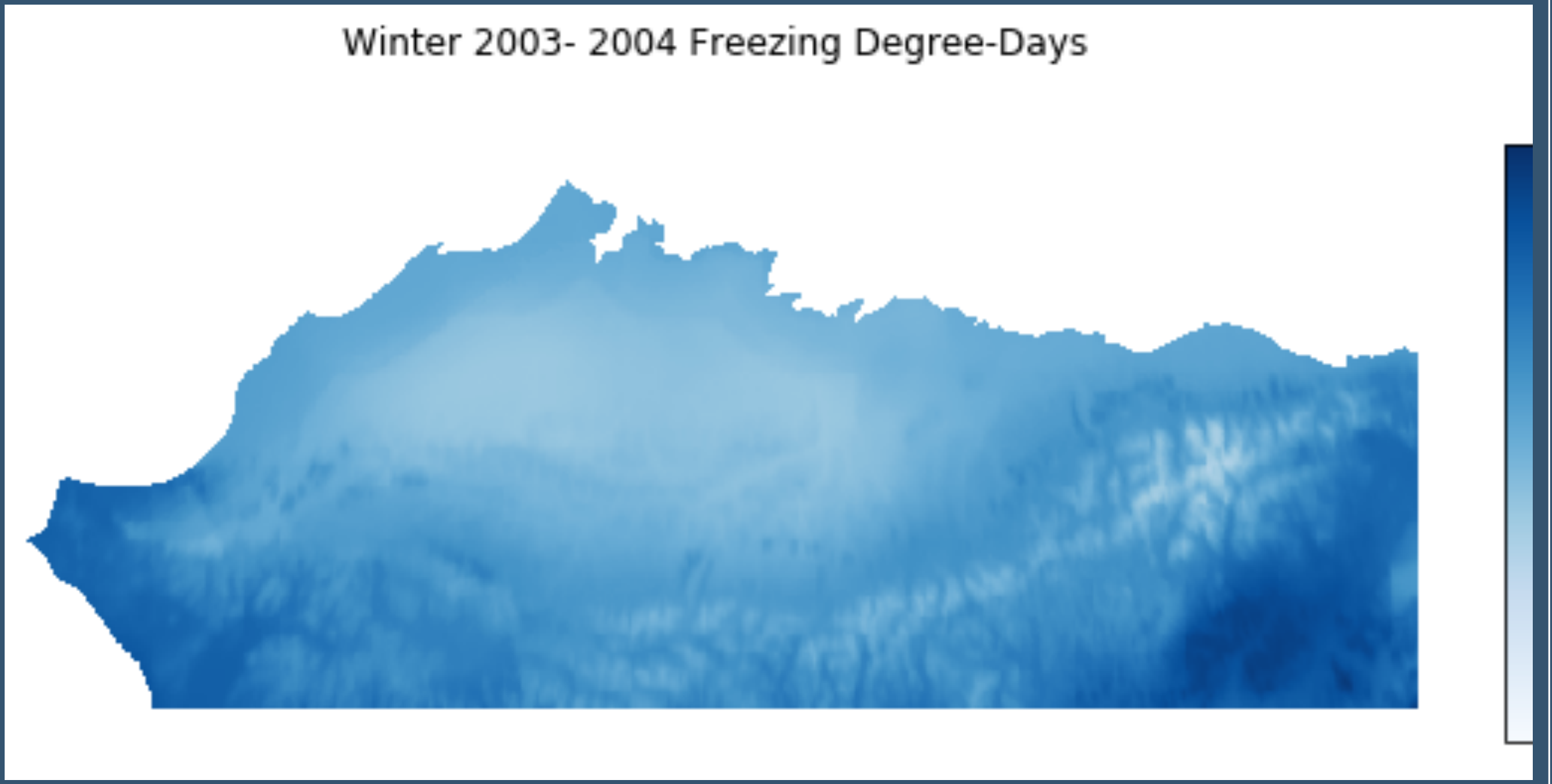


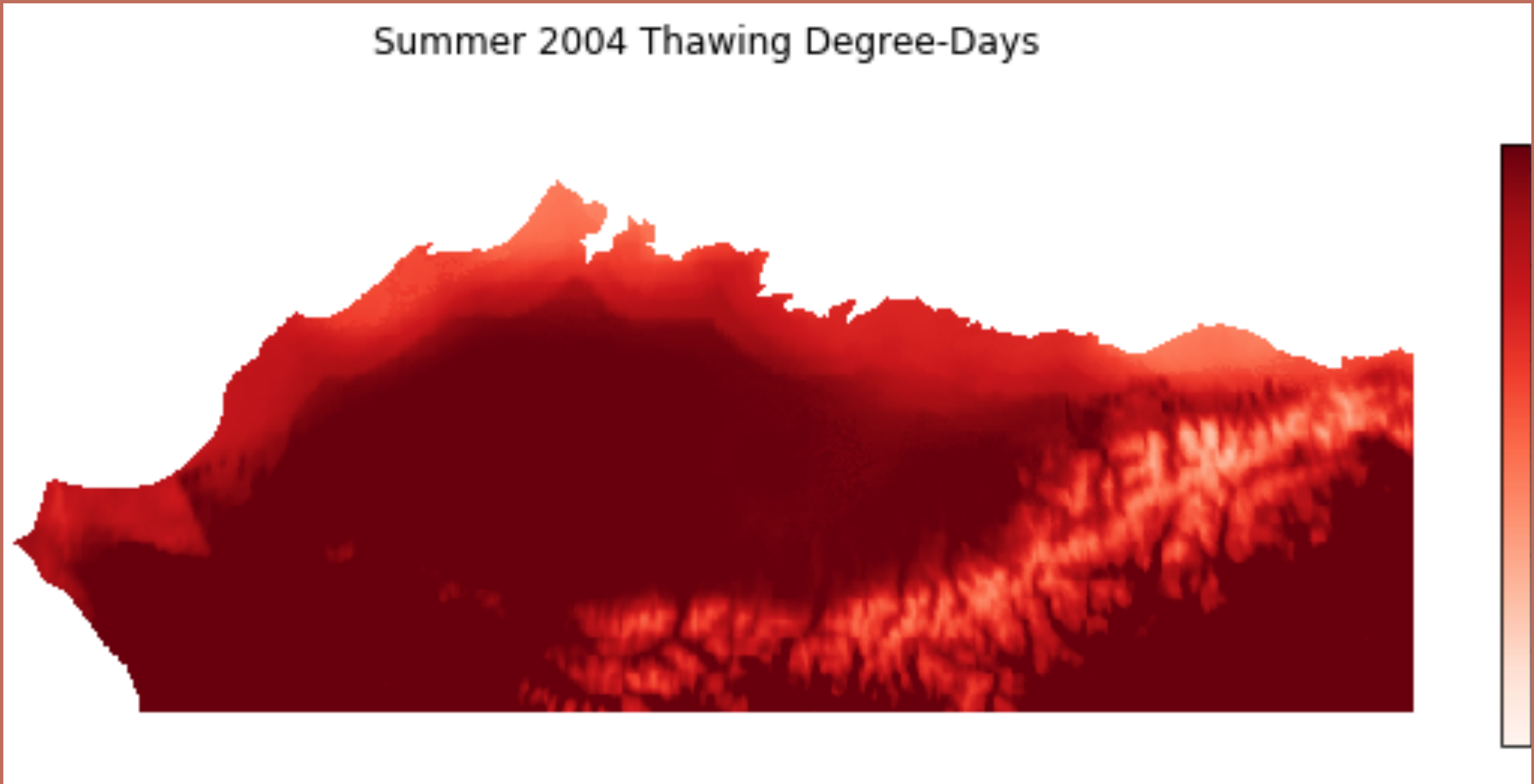

Winter 2003 - 2004 Precip (Oct - Mar) Total (mm)
200
175
150
125
100
75
50
25
0


Summer 2003 Precip (Apr - Sept) Total (mm)
700
600
500
400
300
200
100
0

# Features - Air Temperature

- Air temperature expressed as freezing degree days (FDD) or thawing degree days (TDD)
  - Thawing degree days is used to measure summer temperatures
    - Higher numbers indicate warmer summers (values are > 0)
  - Freezing degree days is used to measure winter temperatures
    - Higher numbers indicate warmer winters (values are < 0)
- High temperatures can indicate increased thawing in permafrost
- We will look at summer temperatures for the year N and year N+1
- We will look at winter temperatures for the winter between year N and year N+1


Winter 2003- 2004 Freezing Degree-Days


Summer 2004 Thawing Degree-Days

# Physical Features

- Features of the location of a model grid cell. Theses may effect the result
  - Latitude, longitude
  - Elevation
  - Slope, Aspect

# Features - Summary

- Previous summer air temperature
- Early winter precipitation (Oct-Nov)
- Full winter precipitation  (Oct-Mar)
- Winter air temperature
- Summer precipitation (Apr-Sept)
- Summer air temperature
- Latitude
- Longitude
- Elevation
- Slope
- Aspect

# Experiment

- Develop random forest models using all features described previously, and original thermokarst initiation model.
- Develop new model using variables indicated as most important by random forest analysis.
- Compare results from original model, random forest results, and new model on a year by year basis to determine how close as a percentage of area new model is to original model
  - Subtract original model from new model to determine how similar the model predictions are

# Tools Used

- Python
  - Numpy
  - Scipy
  - Scikit-learn
    - Random forest regressor.
- GDAL and QGIS
  - Georeferencing and data preparation

# Scikit-learn Random forest parameters

- n_estimators
  - number of trees in forest
- criterion
  - function to use to measure quality of a split
- random_state
  - seed value for random number generator
- max_depth, min_samples_leaf. max_features, max_leaf_nodes
  - used to control tree size

# Reduced Order Model From Random Forest Analysis

Plug the features indicated by random forest model into original method

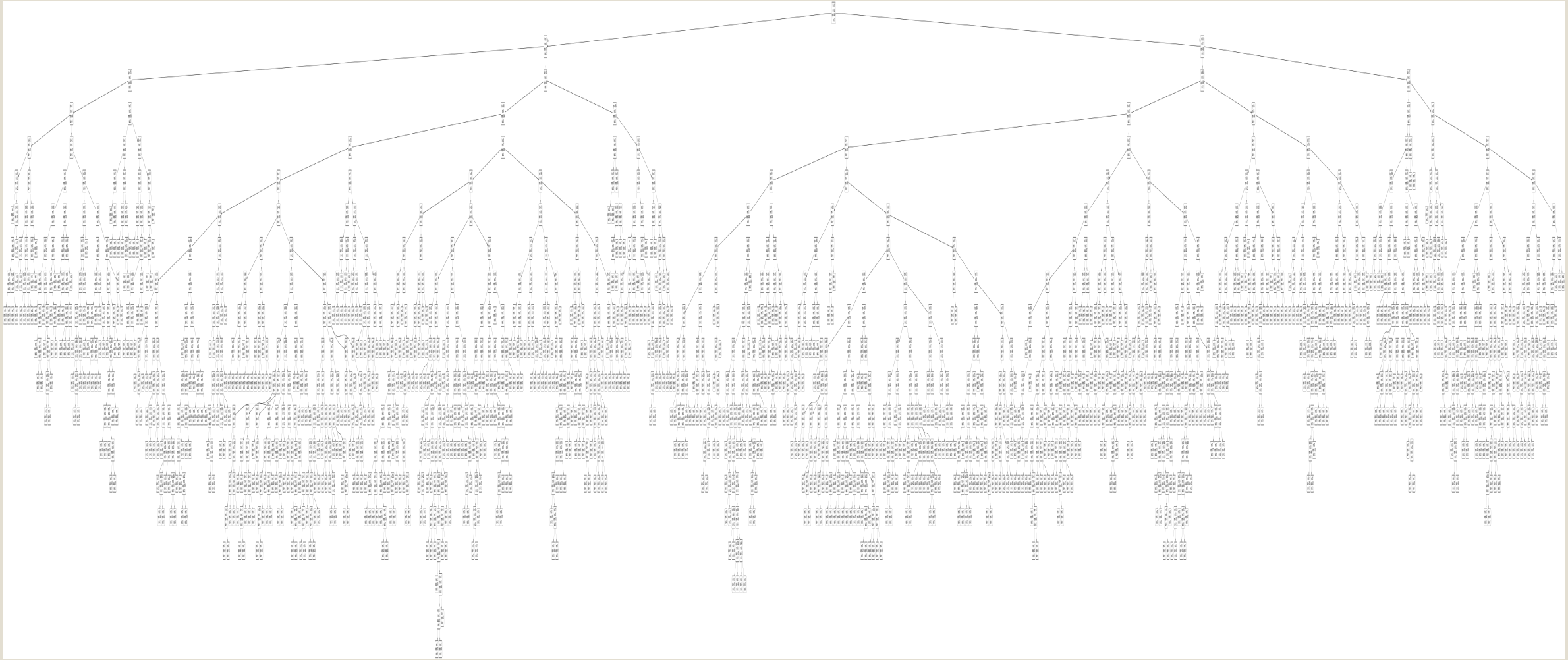
OR

Iterate on random forest model excluding features that do not contribute to the overall variance

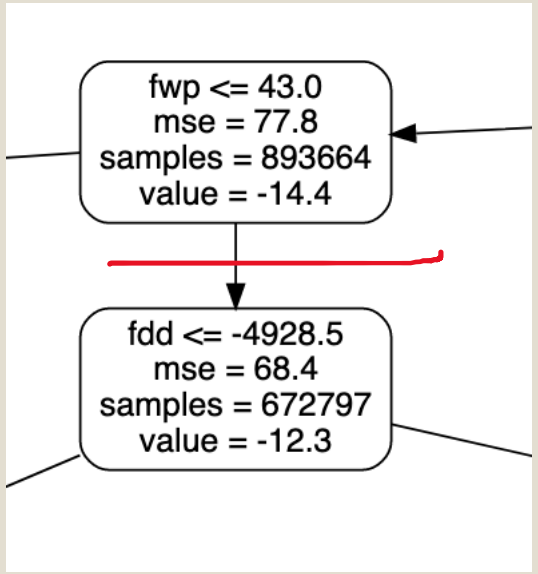
# Early Example

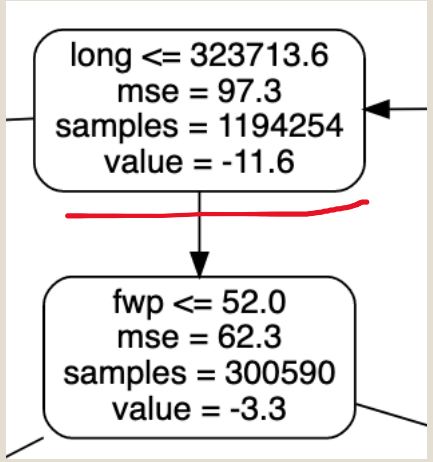
- Random forest model trained with features described earlier and predictions from original model without predisposition applied
- 10 estimators
- Max leaf nodes set to1000
- Used ½ of training data

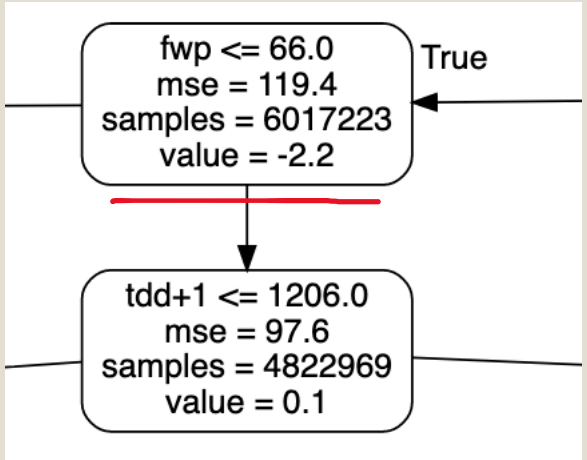
# A Tree in the Forest

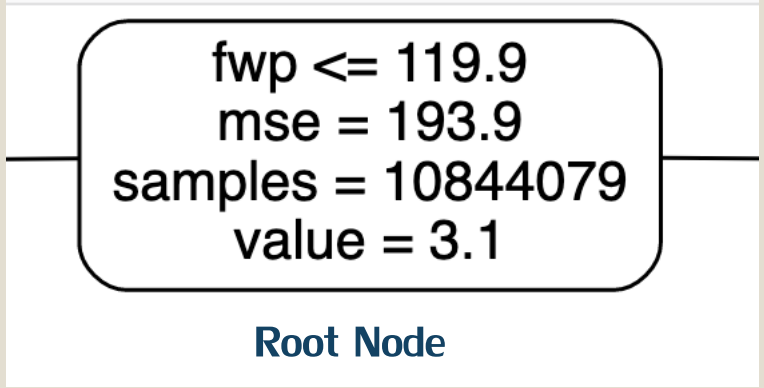


# A Tree in the Forest

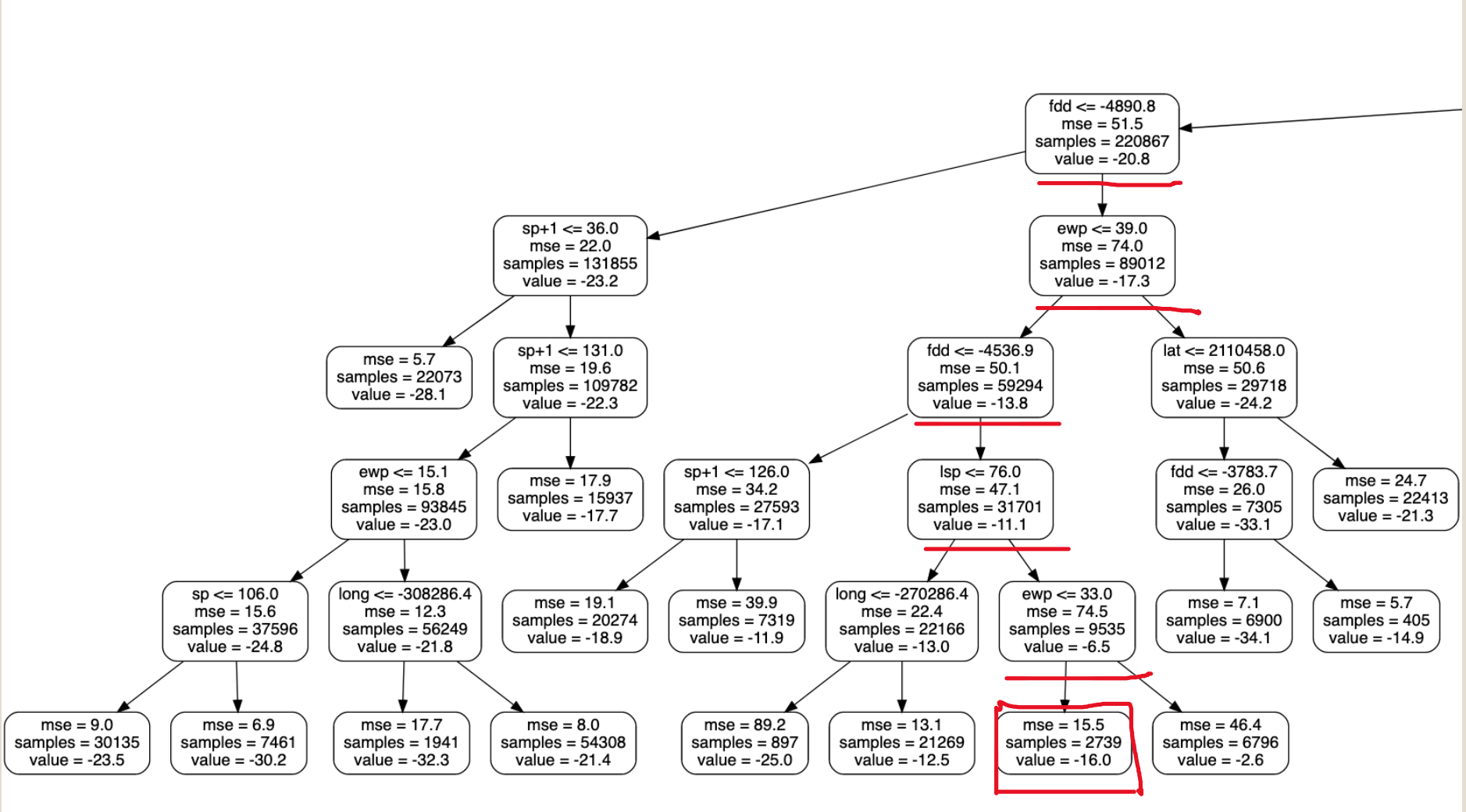


# A Tree in the Forest



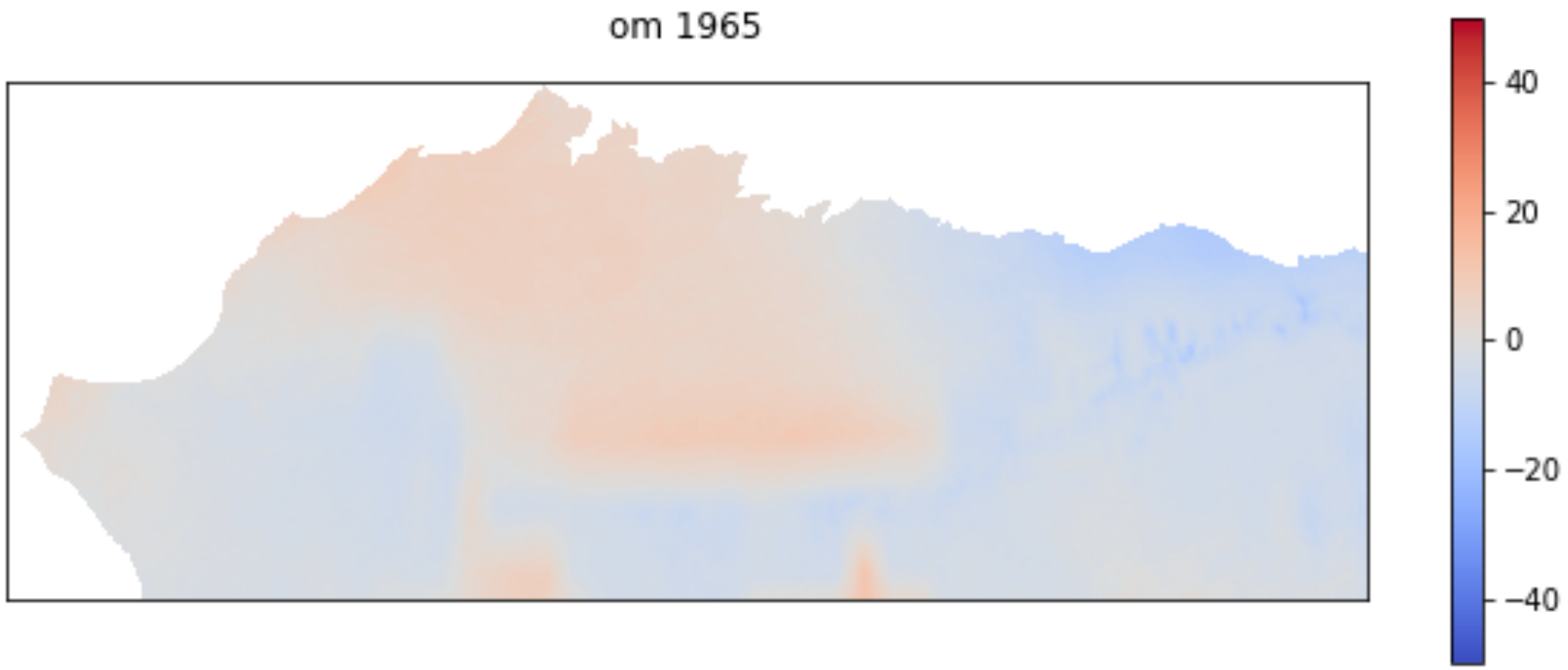
# A Tree in the Forest

# Early Example

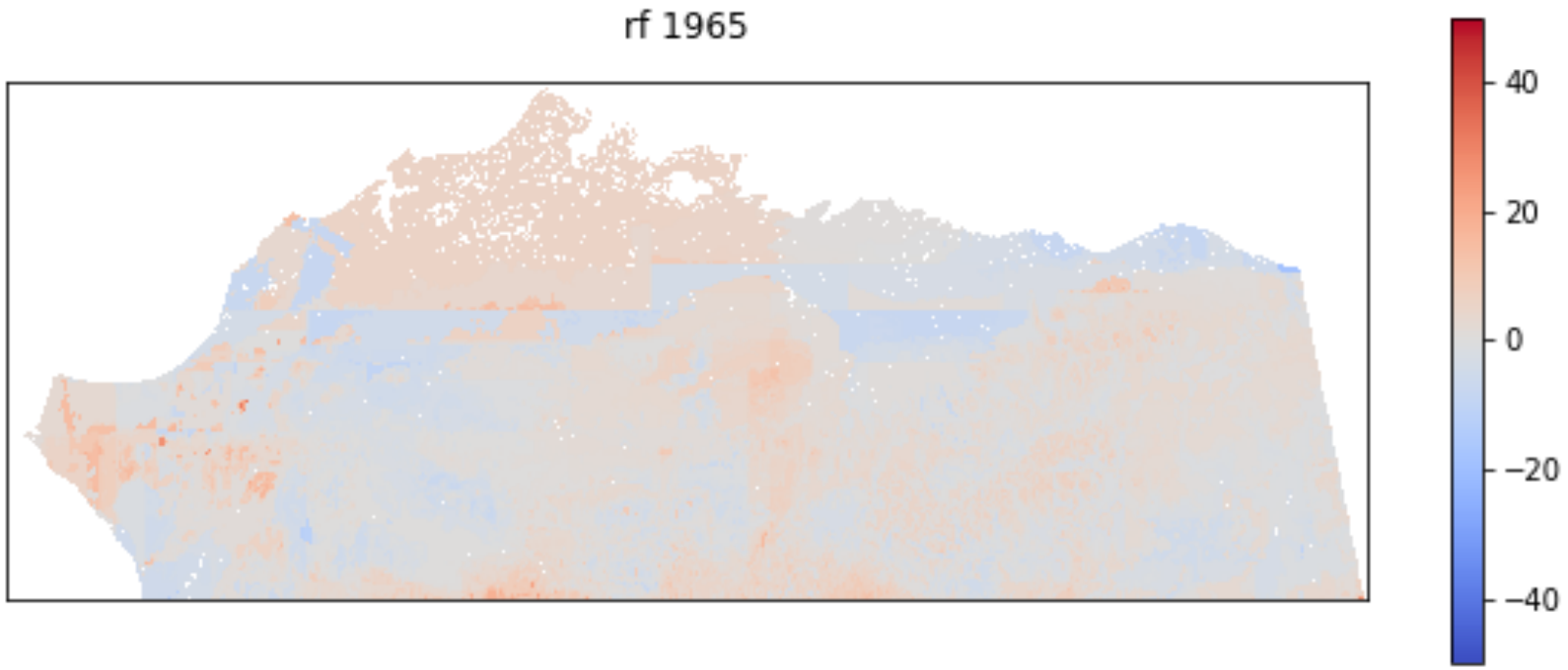
# Feature Importance

- Winter temperature: 3.616%
- First summer temperature: 4.563%
- Second summer temperature: 10.249%
- Early winter precipitation: 3.314%
- Total winter precipitation: 45.813%
- First summer precipitation: 2.271%
- Late summer precipitation: 4.254%
- Next summer precipitation: 2.481%
- Latitude: 11.462%
- Longitude: 11.939%
- Aspect: 0.0%
- Slope: 0.038%
- Elevation: 0.0%

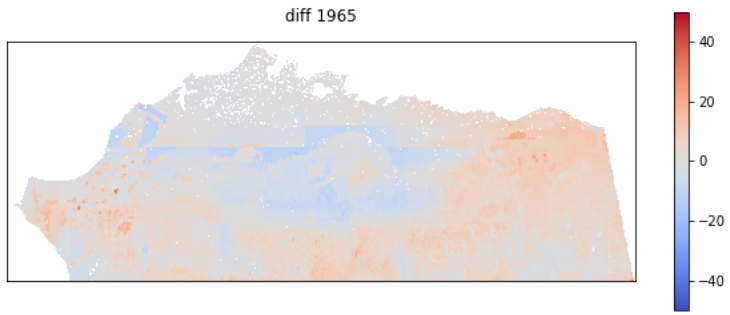
# Original Model Prediction for 1965



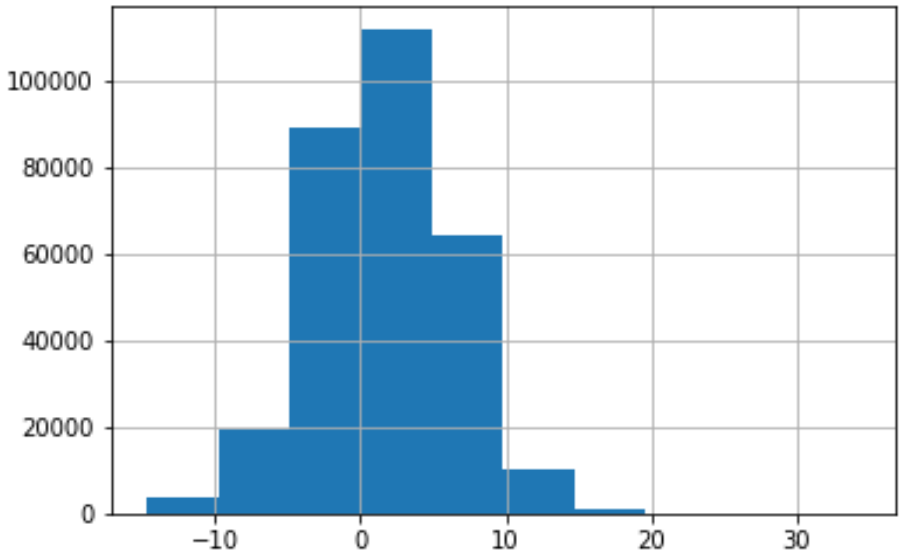
# Random Forest Model Prediction for 1965



# Model Differences for 1965

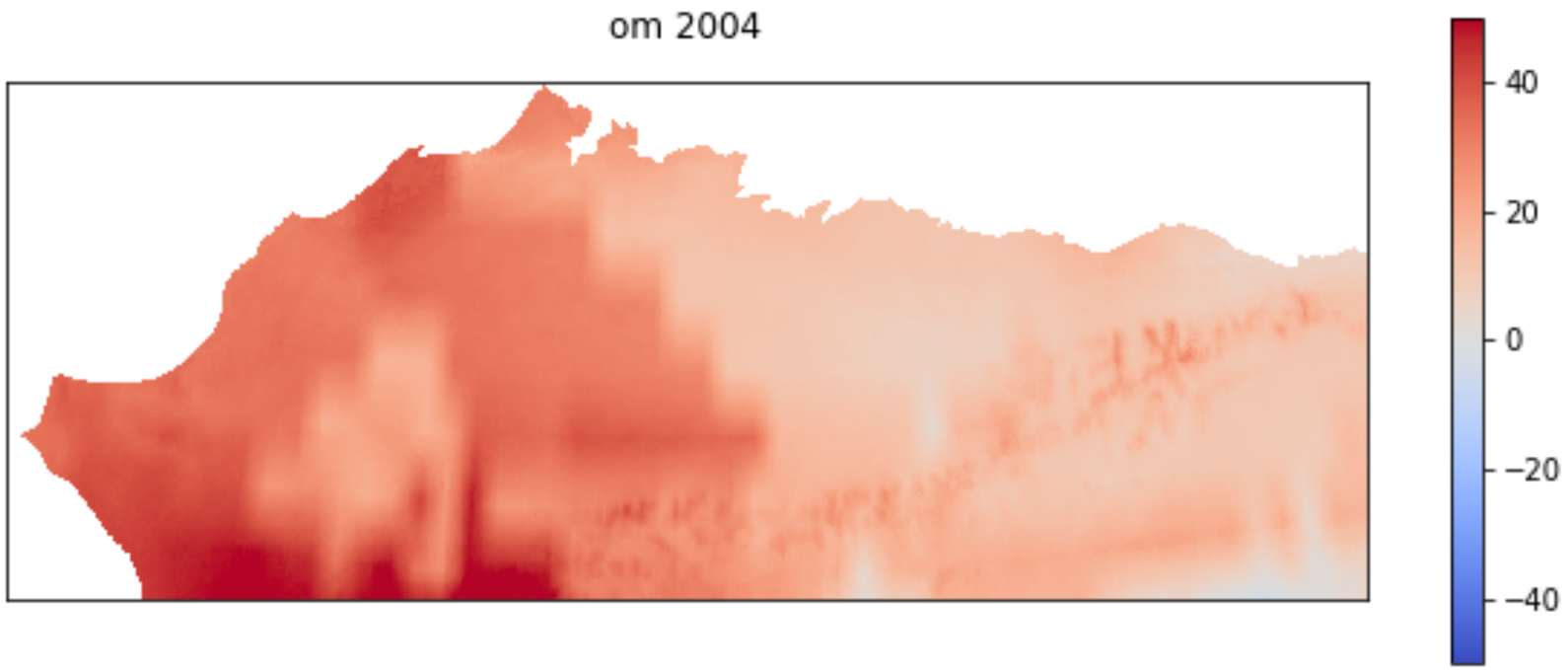



Map

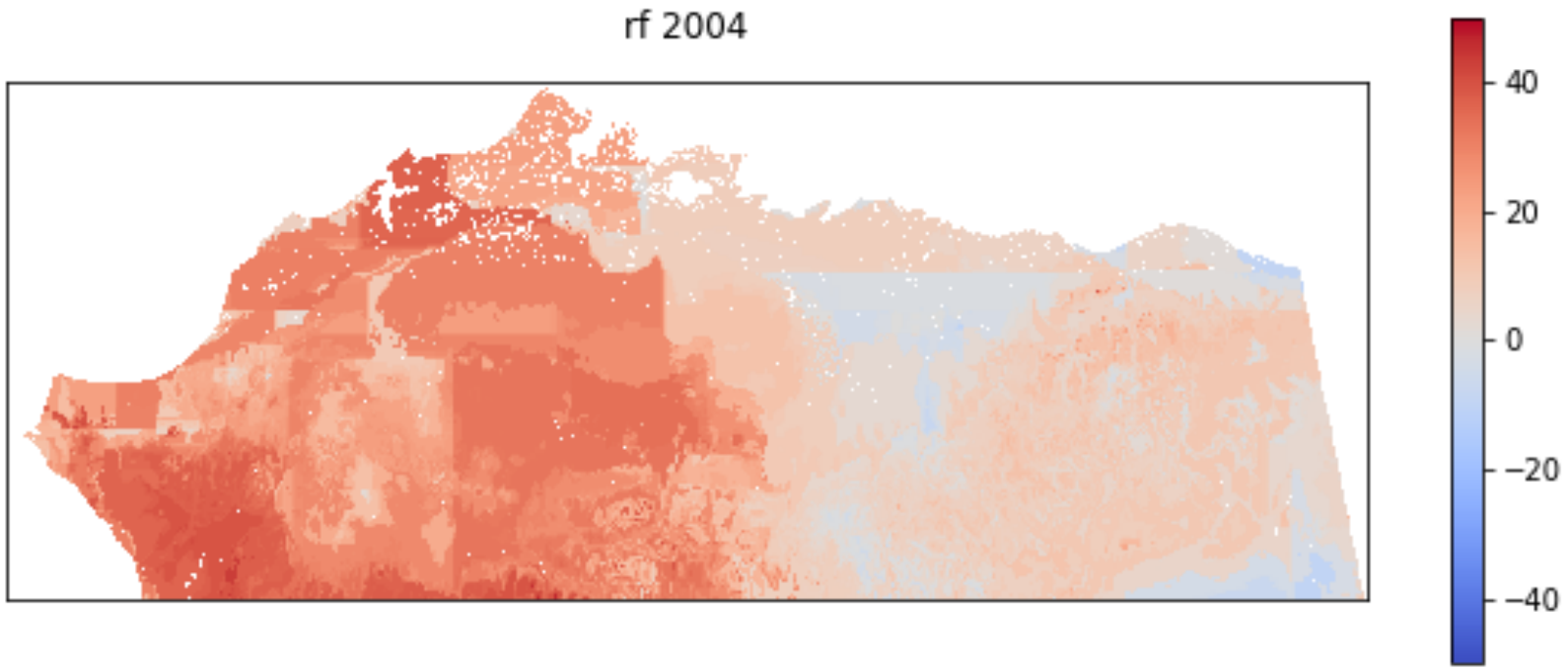



Histogram

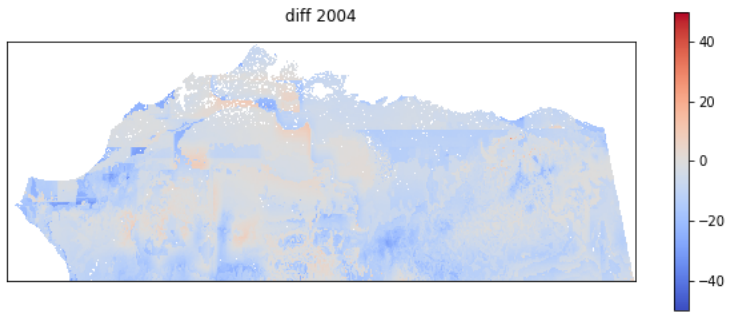
# Original Model Prediction for 2004



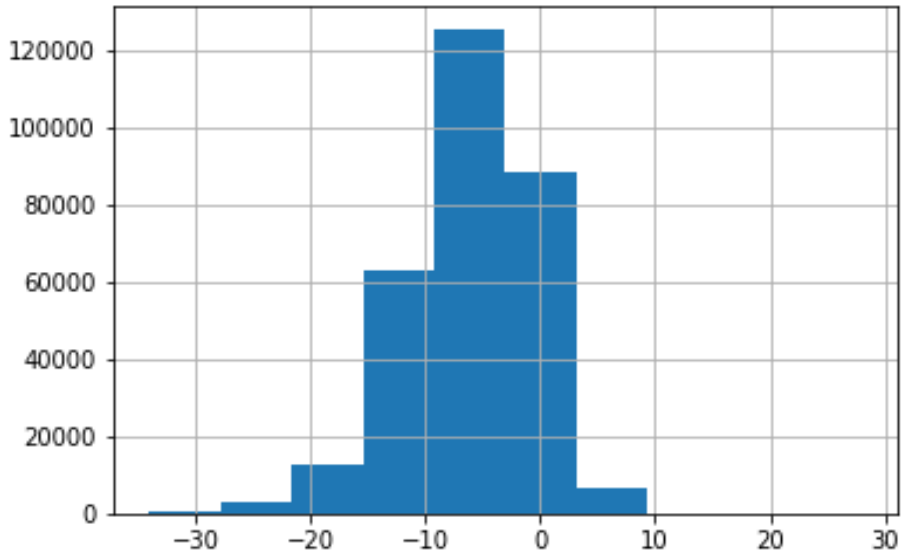
# Random Forest Model Prediction for 2004



# Model Differences for 2004




Map




Histogram

# Timeline

- Nov 2019
  - This presentation
- Dec 2019 – Jan 2020
  - Finish project work
  - Start paper
- Feb 2020 – Mar 2020
  - Finish paper
  - Final Presentation
- Apr 2020
  - Comprehensive exam
  - Submit paperwork
- May 2020
  - graduate

# Sources


- Brown,J., O.J.Ferrians,Jr., J.A. Heginboom, and E.S. Melnikov. 1998, revised February 2001. *Circum-arc map of permafrost and ground ice conditions.* Boulder, CO: National Snow and Ice Data Center. Digital media.
- Jorgenson,T.,K.Yoshikawa, M.Kanevskyi,Y.Shur ,V. Romanovsky, S. Marchenko, G. Grosse, J. Brown, B. Jones. 2008. Permafrost Characteristics of Alaska. Institute of Northern Engineering, University of Alaska Fairbanks, NICOP.
- Herman, G.R., Schumacher, R.S. 2018. Money Doesn't Grow on trees but Forecasts do: Forecasting Extreme Precipitation with Random Forests. American Meteorological Society
- Hugelius,G.,C. Tarnocai, G.Broll,J.G.Canadell, P.Kuhry, D.K. Swanson. 2013. The Northern Circumpolar Soil Carbon Database: spatially distributed datasets of soil coverage and soil carbon storage in the northern permafrost regions. *Earth System Science Data 5*: 3–13
- Scenarios Network for Alaska and Arctic Planning, University of Alaska. 2019 Historical Monthly Precipitation - 1 km CRU TS - Datasets - SNAP Data Portal . Retrieved 2019-11-04 from http://ckan.snap.uaf.edu/dataset/historical-monthly-precipitation-1-km-cru-ts
- Scenarios Network for Alaska and Arctic Planning, University of Alaska. 2019 Historical Monthly Temperature - 1 km CRU TS - Datasets - SNAP Data Portal. Retrieved 2019-11-04 from http://ckan.snap.uaf.edu/dataset/historical-monthly-temperature-1-km-cru-ts
- Scenarios Network for Alaska and Arctic Planning, University of Alaska. 2019. Thermokarst Formation - Datasets - SNAP Data Portal. Retrieved 2019-06-21 from http://ckan.snap.uaf.edu/dataset/thermokarst-formation.

# Other sources

- https://www.sciencedirect.com/topics/earth-and-planetary-sciences/thermokarst
- https://en.wikipedia.org/wiki/Random_forest
- https://en.wikipedia.org/wiki/Bootstrap_aggregating
- https://scikit-learn.org/stable/modules/ensemble.html#forest
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#sklearn.ensemble.RandomForestRegressor
- https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
- https://en.wikipedia.org/wiki/Decision_tree_learning
- Other images from wikipedia