EXAMINING THERMOKARST INITATION WITH RANOM FOREST MODELS

By

Rawser W. Spicer B.S.

A Project Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Computer Science

University of Alaska Fairbanks

May 2020

APPROVED:

W. Robert Bolton, Committee Chair

Orion Lawlor, Co-Chair

Glenn Chappell, Committee Member

Chris Hartman, Chair

*Department of Computer Science*

**Abstract**

This project examines thermokarst initiation through the application of random forest models. Thermokarst initiation marks the start of the formation of thermokarst features and is increasingly occurring due to warming climate conditions. Random forests are an ensemble learning technique that combines the results of many independent decision trees to create results that avoid the overfitting in regular decision trees. Random forests were trained against an existing thermokarst initiation model. Results showed that random forests were useful in this context. Random forest hyperparameters were also examined through a series of sensitivity analyzes.

Contents

[1. Introduction 3](#_Toc36296888)

[2. Background 4](#_Toc36296889)

[2.1. Permafrost 4](#_Toc36296890)

[2.2. The Active Layer 4](#_Toc36296891)

[2.3. Ground Ice 5](#_Toc36296892)

[2.4. Thermokarst 5](#_Toc36296893)

[2.5. Alaska Thermokarst Model 6](#_Toc36296894)

[2.6. Random Forests 7](#_Toc36296895)

[2.6.1. Decision Trees 7](#_Toc36296896)

[2.6.2. Bagging 8](#_Toc36296897)

[2.6.3. Applications 9](#_Toc36296898)

[3. Data 9](#_Toc36296899)

[3.1. Study Area and Period 9](#_Toc36296900)

[3.2. Air Temperature 11](#_Toc36296901)

[3.3. Precipitation 12](#_Toc36296902)

[3.4. Location 12](#_Toc36296903)

[3.5. Elevation, Slope, and Aspect 12](#_Toc36296904)

[3.6. Summary of Features 13](#_Toc36296905)

[4. Methods 14](#_Toc36296906)

[4.1. The Original Model 14](#_Toc36296907)

[4.2. Software 16](#_Toc36296908)

[4.3. Finding Baseline Hyperparameters 16](#_Toc36296909)

[4.4. Comparing Models 19](#_Toc36296910)

[4.5. Scenarios for Sensitivity Analysis 21](#_Toc36296911)

[5. Results 23](#_Toc36296912)

[5.1. The Baseline Random Forest 23](#_Toc36296913)

[5.2. Feature changes 28](#_Toc36296914)

[5.2.1. Add Thermokarst Initiation Model 29](#_Toc36296915)

[5.2.2. Add Random Data 30](#_Toc36296916)

[5.2.3. Remove Latitude and Longitude 31](#_Toc36296917)

[5.2.4. Remove Slope, Elevation, and Aspect 32](#_Toc36296918)

[5.2.5. Remove Partial Season Precipitation 33](#_Toc36296919)

[5.2.6. Remove Features with Low Importance 34](#_Toc36296920)

[5.2.7. Remove Features with High Importance 35](#_Toc36296921)

[5.3. Hyperparameter changes 35](#_Toc36296922)

[5.3.1. Vary Estimators 35](#_Toc36296923)

[5.3.2. Vary Maximum Tree Depth 36](#_Toc36296924)

[5.3.3. Vary Features Used for Decisions 37](#_Toc36296925)

[5.3.4. Vary Maximum Leaf Nodes 38](#_Toc36296926)

[5.3.5. Vary Minimum Samples to Spit Node 39](#_Toc36296927)

[5.3.6. Vary Minimum Samples per Leaf 40](#_Toc36296928)

[5.3.7. Vary Percent Data Used for Training 41](#_Toc36296929)

[5.4. A Final Model 42](#_Toc36296930)

[6. eDiscussion 43](#_Toc36296931)

[7. Future Work 45](#_Toc36296932)

[8. Conclusion 46](#_Toc36296933)

[9. Acknowledgments 47](#_Toc36296934)

[10. References 47](#_Toc36296935)

# Introduction

Arctic landscapes are changing at an increasing rate. One process driving these changes is thermokarst. This process that occurs as permafrost and ground ice thaw due to warming surface temperatures (Lara, et al. 2014) (Jorgenson, Shur and Pullman 2006). Modeling these changes is important to understanding how arctic environments will react in the future which has direct applications in infrastructure improvement, and scientific understanding of the arctic. This research looks at using random forests (RF) to model an existing method of thermokarst initiation (TKI) which looks for regions where ground ice thaw may begin, the start of the thermokarst process. The sensitivity of random forest models to changes in input features, and changes in hyperparameters are examined.

Random forests are an ensemble of decision trees that have been used for classification and regression (McGovern, et al. 2017) (Breimen 2001). Random forests use a set of input variables, called features, to find some output value, usually called labels. RF models are generally trained a on subset of known features, and labels. These models can also be used to determine how much each input features affects the decision process, and therefor how important they are in relation to other variables (Kanevskiy, et al. 2017). This importance ranking can be applied to the development of reduced order models. Reduced order models (ROMs) aim to create models that are less complex than existing models in order to increases speed. These models should have a small approximation error, conserve the properties of the original model, and be computationally efficient (Antoulas 2004).

Changing arctic landscapes affect vegetation, wildlife, drainage, and the carbon & energy regimes (Jorgenson, Shur and Pullman 2006) (Lara, et al. 2014) (Gandodamage, et al. 2014) (Kanevskiy, et al. 2017) (Jorgenson, et al. 2015) (Raynolds, et al. 2014). Understanding these changes is important for ecosystem, hydrology and permafrost sciences. Knowing were and how landscapes will change can inform wildlife managers in deciding what areas to protect as important habitats change, and development in the arctic region continue. As the landscape changes surface water distribution will change. Finding how this will occur can increase our understanding of feedbacks to the climate system (Jorgenson, Shur and Pullman 2006) (Lara, et al. 2014) (Jorgenson, et al. 2015) (Liljedahl, et al. 2016).

This paper examines background related to thermokarst, and random forests in section 2. Sections 3 and 4 look at the data and methods used to perform the analysis. Section 5 looks at finding a random forest to use, and how random forest models respond to changes in features, and hyperparameters. A final model is also presented. The remainder of the paper discusses these results, and future improvements this process and applications of random forests to other problems.

# Background

# Permafrost

Permafrost is soil that is permanently frozen, and occurs primarily in polar and alpine environments. Soils are considered to be continuously frozen if the mean soil temperature remains at or below 0 °C, the freezing point of water, for two or more consecutive years (Davis 2001) (Pollard 2018) (Rowley, et al. 2015). Permafrost has its greatest extent in areas where it has existed for thousands of years or longer. Permafrost generally occurs terrestrially, but also exists in the seabed in polar regions (Davis 2001) (Pollard 2018). Permafrost affects approximately 25% of the earths land area mostly in the arctic and sub-arctic (Pollard 2018) (Rowley, et al. 2015). Permafrost covers approximately 80% of Alaska, 50% of Canada, and 60% of Russia (Pollard 2018).

Permafrost is classified into three groups: continuous, discontinuous, and sporadic. Continuous permafrost covers 90% or more of a specific landscape with mean annual soil temperatures around -8 °C. Discontinuous permafrost occupies 50-90% of landscapes in permafrost areas with mean soil temperatures around -5 °C. Sporadic permafrost covers less than 50% of these environments, and the annual soil temperatures near 0 °C (Pollard 2018) (Rowley, et al. 2015). Stable mean annual temperature is important to permafrost, as low heat flux is important for maintaining permafrost. Soil moisture, air temperature, snow cover, aspect, and elevation also influence the local heat fluxes. (Pollard 2018) (Rowley, et al. 2015)

Permafrost depth ranges from more than 1000m to only a few meters at its southern limits. The deepest known permafrost occurs in Siberia and is around 1400m thick (Davis 2001) (Pollard 2018) (Rowley, et al. 2015). The maximum depth of permafrost is limited by heat in the earth’s mantle (Rowley, et al. 2015). Most permafrost is thousands to millions of years old, but some has formed more recently (Davis 2001) (Pollard 2018).

# The Active Layer

During summer, as temperatures rise above 0 °C, the top layer of permafrost seasonally thaws. This seasonal thaw is known as the active layer (Davis 2001) (Pollard 2018) (Rowley, et al. 2015). The active layer depth ranges from a few decimeters in the far north to several meters in the discontinuous permafrost zone (Davis 2001). The active layer buffers permafrost from warm summer temperatures (Pollard 2018). The conditions at the surface of the soil determine the influence the of air temperature on frozen soils. These conditions include slope, aspect, soil moisture, and snow cover (Rowley, et al. 2015). Warmer summer temperatures are increasing the active layer depth and increasing destabilization of near surface permafrost (Pollard 2018) (Rowley, et al. 2015). `

# Ground Ice

Ground ice is any ice in freezing or frozen soils and occurs extensively in permafrost zones (Rowley, et al. 2015). There are many types of ground ice ranging from massive ice deposits to small ice crystals in soil called pore ice. Ground ice can be ice that has formed in the soil or ice that has become covered as soil is deposited (Pollard 2018). Ice wedge arrays are a type of ground ice that occur in an estimated 10% of the permafrost in Alaska’s arctic coastal plain (Davis 2001).

Ice wedge ground ice occurs as moisture seeps into cracks in permafrost. These wedges grow as the ice freezes and thaws allowing more water in. These wedges reach 10m in depth and range from 2-3m wide (Davis 2001). Networks of ice wedge features come together in distinctive polygonal patterns called ice wedge polygons that range in size from 8 to 18 m across (Davis 2001) (Rowley, et al. 2015). Ice wedges, along with other ground ice features, are prone to completely melt as temperatures warm (Pollard 2018). This leaves distinctive depressions in landscapes through a process called thermokarst (Davis 2001).

# Thermokarst (Lawlor’s note)

Thermokarst describes the melting of ground ice, and the features this process creates (Farquharson, et al. 2016) (Kanevskiy, et al. 2017). This process is caused by disturbances to the ground surface and accelerated by increasing thaw in permafrost (Jorgenson, Shur and Pullman 2006) (Farquharson, et al. 2016) (Kanevskiy, et al. 2017) (Jorgenson, et al. 2015) (Raynolds, et al. 2014). One type of thermokarst is polygonal patterned ground formed by the melting of ice wedge features. Polygonal features consist of high center polygons, and low center polygons amongst others (Lara, et al. 2014). The melting of these wedges can eventually lead to the formations of ponds and lakes (Jorgenson, Shur and Pullman 2006) (Kanevskiy, et al. 2017) (Liljedahl, et al. 2016) (Raynolds, et al. 2014). Simulating ice wedge polygons transitions, and their effects on landscapes are the main goal of the Alaska Thermokarst model (Robert, et al. 2018).

# Alaska Thermokarst Model

The Alaska Thermokarst model (ATM) aims to predict how landscapes change in response to the thermokarst process. It is a pixel-based state and transition model that tracks each percentage of each landscape cohort in all pixels and changes these in response to thermokarst activity. Each time step is a single year, and model is run over 100-year time periods. Landscape cohorts describe the characteristics of the land scape. These include meadows, different types of polygonal patterned ground, lakes and ponds (Robert, et al. 2018). This model is being developed.

The ATM asks for main questions at each timestep: Is thermokarst active? Has thermokarst been initiated by extreme event? Is active layer > protective layer? Does climate support new ecotype? Figure 1 shows how these questions fit into each time step. The second question, has thermokarst been initiated by an extreme event, is related to the work being done here. In fact, the thermokarst initiation model presented here can be considered a subprocess with in the ATM (Robert, et al. 2018). The current method for modeling TKI is further described in section 4.1.

Yes

No

No

No

No

No

No

Yes

Yes

Yes

Yes

Yes

4

4

**Child Cohort**

[New LS] + [Parent ET]

Determine ET Change

4

**Parent Cohort**

[Parent Landscape (LS)] + [Parent Ecotype (ET)]

Determine ET Change

Update Protective Layer Depth

Set Protective Layer Depth

2

3

1

Determine LS Change

Determine ET Change

**Child Cohort**

[Parent LS] + [New ET]

**Child Cohort**

[New LS] + [New ET]

Determine Rate of LS Change

Figure 1: Flow Chart of ATM Logic. 1. Is thermokarst Active? 2. Has thermokarst been initiated by an extreme event? 3. Is Active layer > protective layer? 4 Does climate support new ecotype? Question 2 is the is focus of the Thermokarst Initiation model.

# Random Forests

Random forest models are an ensemble machine learning technique that combines the results of many decision trees. Random forest models can be used as classification or regression tools. Each tree in a random forest is constructed using a randomized subset of the training data, and random subset of questions at each node. Final results are calculated by taking the mode (for classification) or mean (for regression) of all the trees. This corrects for possible overfitting that a single tree may have (McGovern, et al. 2017) (Breimen 2001).The mechanics of each concept used in random forests are described in the following sections.

# Decision Trees

Decision trees are a network of decision nodes and leaf nodes. Each decision node has exactly two child nodes which may be either another decision node or a leaf node. The split at each decision node is based on comparing the value of a single input feature. Predictions travel left or right based on whether this comparison is true or false. For a prediction, the tree is traversed for a set of inputs through each node, starting at the root, until a leaf node is reached. Leaf nodes contain the final result of the prediction. For classification this is a label, while for regression this is a numerical value (McGovern, et al. 2017) (Herman and Schumacher 2018).

Figure 2 shows a decision tree based off the TKI problem presented in this paper. The tree is three levels deep which results in eight leaf nodes. Each decision node shows the question asked (for example the root asks is thawing degree days (TDD) <= 1239.1), mean square error for the split, the percentage of samples left for further nodes, and the value that the node would return. The leaf nodes contain the same information minus the question. This tree does not give very accurate results due to its limited depth.

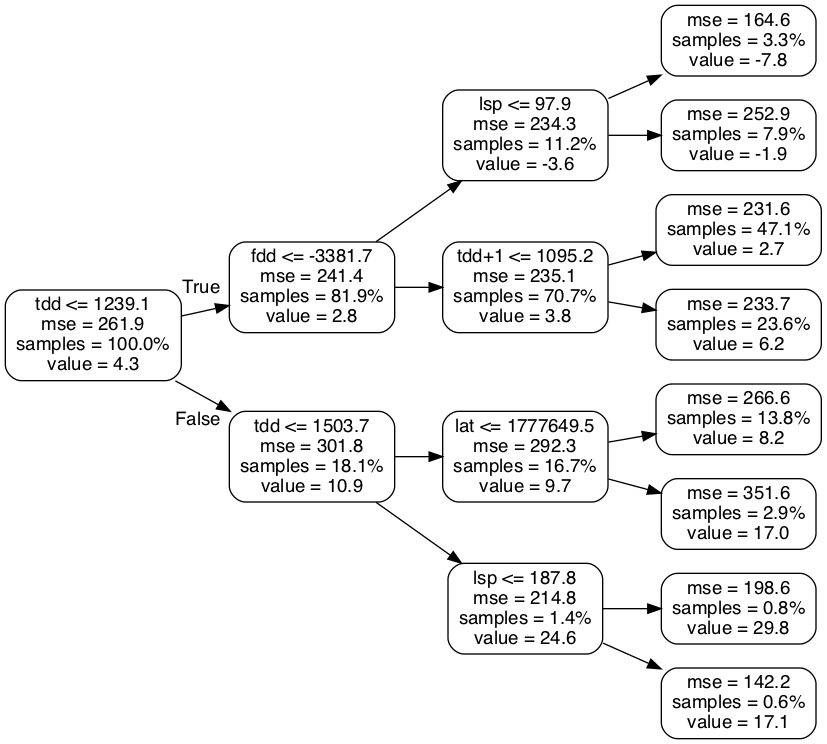


Figure : Example Decision Tree

# Bagging

Bootstrap aggregating, or bagging, is a method used to randomize data used in statistics. In general, Bootstrapping refers to selecting a random subset of samples from a given set of data with replacement. Giving a training set, size *n*, bagging builds modified training data, also size *n*, for each tree by selecting *n* samples with replacement from the original training data. Sampling with replacement selects from the full original data for each new, so the same item may appear more than once. This method is called tree bagging. Feature bagging can also occur, and only selects a subset of features at each decision node. Using these bagging techniques results in largely uncorrelated trees (McGovern, et al. 2017) (Herman and Schumacher 2018) (Hastie, Tibshirani and Jerome 2017) (Breimen 2001).

# Applications

Random forest models have been used as precipitation forecasting models. Using 11 years of data from NOAA’s Second-Generation Global Ensemble Forecast System Reforecast, random forest models were created that could predict accumulated precipitation for 2 and 3 day periods for regions across the contiguous United States. It was found that random forest models were capable of predicting extreme precipitation events over other methods examined (Herman and Schumacher 2018)

Random forests have also been used in ecological sciences. RF techniques were used to classify landscapes containing invasive species in Lava Beds National Park, locating areas containing rare lichen species in the Pacific North West, and finding cavity nesting bird habitats in the Uinta Mountains. For these applications, RF methods were compared to linear discriminant analysis, logistic regression, and classification trees. RF techniques were found to be superior to the other methods especially for the invasive species example (Cutler, et al. 2007). RF methods were also used to predict soil organic carbon (SOC) stocks in Kenya. This study compared RF to other modern machine learning methods: artificial neural networks (ANN), and support vector machines (SVM) It was found that RF had a higher tendency to overpredict than the other methods (Were, et al. 2015).

RF methods have found use in remote sensing as well. Estimations of wheat biomass from satellite images in China was completed using RF, SVM, and ANN methods. It was found that estimation accuracy of the RF out preformed SVM, and ANN methods for this problem (Wang, et al. 2016). RF methods were also used to classify different crop types in the UK from satellite data. This study found that RF results were comparable to SVM methods, noting that RF methods had significantly fewer parameters to adjust and thus are easier to implement and use (Pal 2007).

# Data

# Study Area and Period

The study area main consists of Alaska’s Arctic Coastal Plain(ACP). This area mostly consists of tundra north of the Brooks Range in Alaska. All data used is in the form of geotiff raster files to ensure consistency of the locations of each cell. The data have all been converted to a 1km square scale. This gives the rasters an area of 415 rows by 1096 column. This area is shown in Figure 3.

A black and white photo of a person

Description automatically generated

Figure : ACP area, elevation.

The secondary study area consists of the Seward peninsula (SP). This area is considerably smaller and is where the observations that the original TKI model is based on were made. The raster has the same resolution as the ACP data, with an area of 270 row by 384 columns.

A close up of a mountain

Description automatically generated

Figure : Seward Peninsula area, elevation

The climatological data exists for a period of 115 years starting in 1901. This is also the time period used to create the training data for the Random forest models described later. For the Random Forest models the data is subsampled into 25, 50, and 75 percent portions by taking random raster cells from across the full 115 year period. For data were there is no change, or very little, from year to year like geolocation the rasters are assumed to be constant over time. The raster projection used is the NAD83 Alaska Albers equal-area conic projection (EPSG:3338) (spatialreference.org 2006).

# Air Temperature

Air temperature is considered because it is a factor that controls active layer depth which determines if ground ice experiences warming (Kanevskiy, et al. 2017). Air temperature data is used in a degree day format which represent heating or cooling. Here freezing Degree-days (FDD) represent winter temperatures while thawing degree days(TDD) represent summer temperatures. Degree Days are the sum of all daily average temperatures in given period in this case summer or winter. The Degree-day data used was derived from the Scenarios Network for Alaska and Arctic Planning (SNAP) Historical Monthly Temperature - 1 km CRU TS data which contain downscaled estimates of monthly mean temperatures raster data at a 1 km square resolution for 1901 to 2015 (Scenarios Network for Alaska and Arctic Planing 2019). Data for the ACP was clipped to the extent described in section 3.1.

To calculate FDD and TDD from the monthly data spline interpolation provided by Scipy was used. A univariate spline function was fit to the data, producing a periodic function with 2 roots per year, representing the change from summer to winter or winter to summer. Counting the number of roots and comparing to the number of years times two was used to verify the correct number of roots. These number are required to be equal, and it was assumed that these roots occurred 2 per year due to the sparse nature of the data. This function was then integrated for all periods from root N to N+1 for all roots of the function in the given data range. This produced a set of alternating negative (FDD) and positive (TDD) values for each year. This process was done for each pixel in the rasters. Any pixels where the method failed were interpolated from the surrounding pixels. This occurred in less than ten locations. FDD values were labeled according to the year where the winter starts. Examples of freezing degree day and thawing degree day data are shown in Figures 5 and 6.

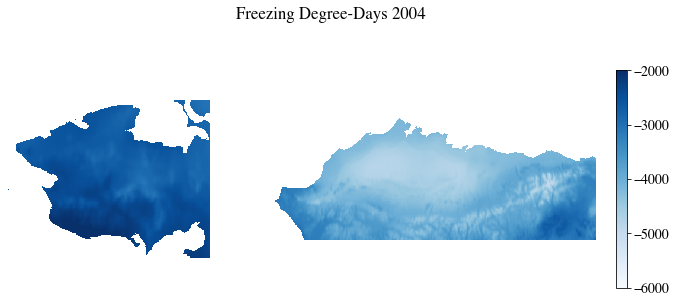


Figure : Freezing Degree Days for 2004 (Seward Peninsula Left, ACP Right)

A picture containing food

Description automatically generated

Figure : Thawing Degree Days 2004 (Seward Peninsula Left, ACP Right)

# Precipitation

Precipitation is another factor that effects active layer depth (Kanevskiy, et al. 2017). Precipitation data are in total millimeters and were derived from SNAP Historical Monthly Precipitation - 1 km CRU TS data which contain data downscaled to the same parameters as the SNAP air temperature data (Scenarios Network for Alaska and Arctic Planning 2019). The monthly data are summed into the desired time periods used in the TKI and random forest models. Values from April to September are summed for summer precipitation (SP). Values for August and September are summed for late summer precipitation (LSP). October and November are summed for early winter precipitation (EWP) while October through March is summed for full winter precipitation (FWP). Example precipitation figures are in appendix A.

# Location

Location data is derived from the raster data for FDD, though any georeferenced raster for the proper area of interest would work. The affine transform was used to calculate the latitude and longitude for the study area. These values were then saved as raster files themselves. The values in these rasters represent northing and easting for the center of the pixel, in meters, in NAD83 Alaska Albers equal-area conic projection (EPSG:3338). Location data is used only in the RF models

# Elevation, Slope, and Aspect

Elevation data is derived from the national elevation dataset 60 m hill shade product provided via the state of Alaska elevation data portal (DGGS n.d.). The 60 m data was rescaled to a 1 km resolution, for use in the RF models. Slope and aspect were calculated from the 1 km version of the data using QGIS.

# Summary of Features

Table 1 summarizes the data used as features in the initial Random Forest models. The table notes if the data are used in the original TKI model and if features are climate related or physical location based. The source of the data is also listed.

Table 1: Summary of Features

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Used in Original TKI Model** | **Type** | **Source** |
| Freezing Degree Day (FDD) | Yes | Climate | Derived from SNAP monthly air temperature data |
| Thawing Degree Day - First Summer (TDD) | Yes | Climate | Derived from SNAP monthly air temperature data |
| Thawing Degree Day - Second Summer (TDD+1) | Yes | Climate | Derived from SNAP monthly air temperature data |
| Summer Precipitation - First Summer (SP) | No | Climate | Derived from SNAP monthly precipitation data |
| Late Summer Precipitation - First Summer (LSP) | No | Climate | Derived from SNAP monthly precipitation data |
| Early Winter Precipitation (EWP) | Yes | Climate | Derived from SNAP monthly precipitation data |
| Full Winter Precipitation (FWP) | Yes | Climate | Derived from SNAP monthly precipitation data |
| Summer Precipitation - second Summer (SP+1) | No | Climate | Derived from SNAP monthly precipitation data |
| Latitude (Lat) | No | Physical | Derived from raster affine transform |
| Longitude (Long) | No | Physical | Derived from raster affine transform |
| Elevation (Elev) | No | Physical | Derived from AK DGGS dataset |
| Slope | No | Physical | Derived from elevation |
| Aspect | No | Physical | Derived from elevation |

# Methods

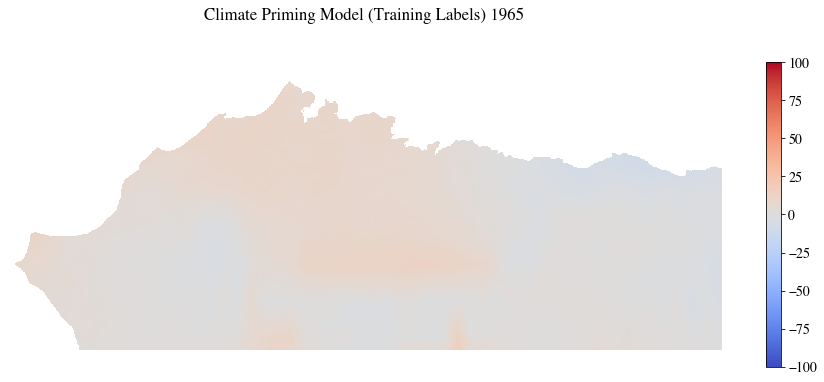
# The Original Model

The original thermokarst initiation model is based on the idea that consecutive extreme climate conditions lead to the start of the thermokarst process. This climate priming process is based on observations on the Seward Peninsula were consecutive seasons with warm and wet conditions lead to thermokarst initiation. These features also control active layer depth which influences Thermokarst (Kanevskiy, et al. 2017). Based on this idea the TKI model looks for years where the summer temperature, and the preceding summer and winter temperatures, and winter precipitation were greater than average. For each of these values the percent difference from the average was calculated. These presented differences were then averaged to find the TKI value.

This method is presented in Algorithm 1. The variables that were used for the TKI model presented here were TDD, EWP, FWP, FDD and TDD+1. Figure 7 shows examples of the results of this process for 1965 a regular year, and 2004 a year with high TKI values. Further examples are in appendix B.

7:

Algorithm : method for finding TKI



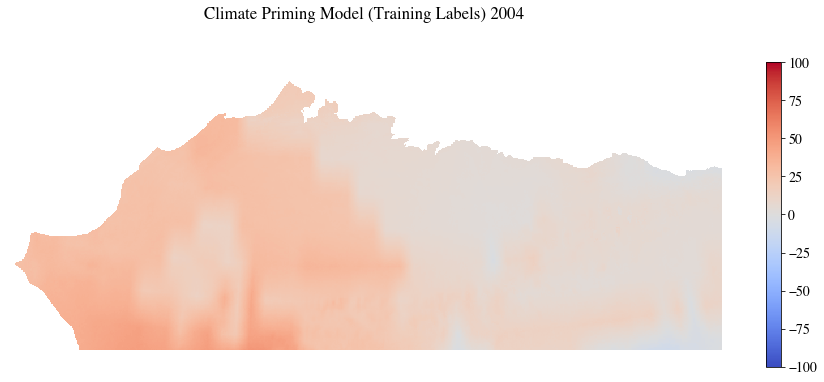


Figure : ACP TKI for 1965, a normal year, and 2004, an extreme year.

# Software

Model implementation and analysis were done in Python (version 3.7.3). The Conda package manager was used to create the Python environments used. Conda is the package manager used and created for the Anaconda platform. It allows for the creation of isolated python environments that can be shared between systems (Conda n.d.). The environment file used can be found at the GitHub repository for this project.

The Random Forest regressor in Scikit-learn (version 0.20.3) was used to create the random forest models. This Random Forest implementation has many useful features including providing the feature importance, and ability to provide a decision path. There are also several hyperparameter available to control tree growth. If these parameters are unset their defaults create fully grown trees which can be very large (scikit-learn-developers n.d.).

# Finding Baseline Hyperparameters

Random Forest models have several hyperparameters that can be adjusted to affect the accuracy of the model being trained. Determining what values of these hyperparameters optimizes the model results can be difficult. Scikit-learn provides access to several hyperparameters. These include the number of trees, the number of decision features at each node, and the maximum depth of each tree. Other parameters influence how the trees are split at each decision node (Pedregosa, et al. 2011). The Scikit-learn hyperparameters considered are summarized in Table 2.

Table 2: Summary of Random Forest Hyperparameters

|  |  |
| --- | --- |
| **Hyperparameter** | **Description** |
| n\_estimators (e) | Number of trees in the forest. |
| max\_depth (md) | Maximum depth of each tree. |
| min\_samples\_split (mss) | Number of samples required to split a node. |
| min\_samples\_leaf (msl) | Number of samples required at each leaf. |
| max\_featuers (mf) | Maximum number of features considered at each split. |
| max\_leaf\_nodes (mln) | Maximum leaf nodes of each tree. |
| Percent Training. Data (tdp) | Percentage of data used for training model |

In order to determine which combination of hyperparameters create strong random forest models we looked at various combinations of parameters. The values of each hyperparameter considered are shown in Table 3. The combinations of these hyperparameters were examined in a brute force manner, all for the ACP region. Each variation of a hyperparameter was combined with each possibility of other the hyperparameters. With this method there were 5,184 models to train

Table 3: Hyperparameter Values

|  |  |
| --- | --- |
| **Hyperparameter** | **Values** |
| n\_estimators | 10, 50, 100 |
| max\_depth | 12, 25, 60, 100 |
| min\_samples\_split | 2, 5, 10 |
| min\_samples\_leaf | 1, 2, 4, 8 |
| max\_featuers | AUTO, SQRT, LOG2 |
| max\_leaf\_nodes | 1,000, 5,000, 10,000, 50,000 |
| percent of training data | 25, 50, 75 |

Training the 5,184 possible models was accomplished using four systems with varying capabilities, and existing workloads. They ran either Mac OSX, Debian, or Ubuntu. Table 4 summarizes the systems used. The Random Forest regressors were configured in a way that four parallel jobs were used to build trees in the forest. Statistics were collected on each of the models trained, but the models themselves were not saved at this time. The statistics recorded were the time to train each model, the time to predict the values for a given year, the mean difference and its variance, the mean absolute difference and its variance, and the median.

The work done by each computer was tracked via a csv file located in a git repository. This list initially contained the parameters to be used, a column to track the progress of the training, and empty columns for the desired statistics to be collected. Examples of this are in the projects GitHub repository, see the readme for the exact location. After each model was trained, the csv data was updated with the model’s statistics, and checked in with the git repository. During this step, the next set of Random Forest parameters was also acquired.

A problem arose with merge conflicts in the csv file after the first few days of running, and a step to correct this had to be developed. Merge conflicts were corrected by removing any duplicate rows, keeping completed rows over unrun or in progress rows with the same parameters, and deleting any of the auto generated merge conflict text. This step occurred right before the final resync that push new stats to the remote repository. This updated utility was deployed to only one of the systems, and the problem was corrected. Example code for this process can be found on GitHub at (URL).

Table 4: Training Systems Summary

|  |  |  |  |
| --- | --- | --- | --- |
| **Computer Name** | **OS** | **CPU** | **Memory** |
| Bristlecone | Ubuntu 18.04 | 3.2 GHz  Intel Core i7 (12) | 16 GB  1333 MHz  DDR3 |
| Finwhale | Debian 10 | 3.2 GHz  Intel Core i7 (12) | 16 GB  1333 MHz  DDR3 |
| Ocotal | Ubuntu 18.04 | 2.4 GHz  Intel Xeon (16) | 24 GB  1067 MHZ |
| Chickadee | Mac OSX 10.12 | 2.5 GHz Intel Core i7 (4, or 8 with hyperthreading) | 16 GB  1600 MHz DDR3 |

The bulk training of the random forest models was run over the period of about a month before the training programs were externally stopped. A combined training time of 125 days 17:38:38 was completed during this period. The order for training took the number of estimators into account before any other hyperparameter. Therefor all 10 and 50 estimator models were run during the testing period while only ~83% of all 100 estimator models were completed before the training was stopped. Figures 1 and 2 show how the model training times vary color coded by number of estimators. Timing data was largely inconsistent, and it was difficult to find how changing hyperparameters affected timing from this data. This could be due to the nature of the many changing hyperparameters, and the fact that the systems were not equally powerful because of this timing was largely not considered as a performance measure.

# Comparing Models

Statistics that can be used to measure the accuracy of random forest models include the mean error (ME), the mean absolute error (MAE), and the coefficient of determination () for each model. The equations are shown in Equation 1, 2, and 3 where RF refers to the predictions of the random forest model and TKI refers to the thermokarst initiation model predictions.

Equation 1: Mean Error

Equation 2: Mean Absolute Error

Equation 3: Coefficient of Determination

Both ME, and MAE were used because ME is inadequate for showing how far from the TKI model predictions are. This is because many of the error values are above and below zero, the expected value, and when the average is taken the ME is always close to zero. The MAE corrects this by taking the absolute value of the differences before calculating the mean there for giving a measure of how far off the random forest model is predicting. Figures 4 and 5 show this phenomenon graphically for the models that were trained during the brute force testing.

The coefficient of determination, , is another measure of how well observed outcomes are modeled by each random forest. The maximum value is one, which would indicate a perfect model, and the value can be negative if the models poorly simulate the observations. was calculated using the Scikit learn random forest regressors score method. A random forest model that performs well should have a high score, and a low MAE. was primary used to validate that the MAE.

A screenshot of a computer

Description automatically generated

Figure 8: Mean Error

A screen shot of a computer

Description automatically generated

Figure 9: Mean Absolute Error

A best model, called the baseline model, was chosen from the brute force examples by sorting the models by their MAE. High accuracy models have low MAE values. A selection of top preforming models is presented in Table 6 and discussed in Section 5.1.

# Scenarios for Sensitivity Analysis

To determine the accuracy of the baseline model, and to look at how small changes in hyperparameters effect random forest accuracy several scenarios for sensitivity analysis were performed. These can be broken down in to two categories feature changes, and hyperparameter changes. The feature changes were designed to look at how the model results, and feature importance scores change when various features are removed from the training feature set. This should help to verify their importance or not. The hyperparameter changes are designed to look at how small changes to the hyperparameters effect the model accuracy, as the method for choosing the baseline only looked at relatively large changes for many parameters. The scenarios for sensitivity are presented in Table 5.

Table 5: Scenarios for Sensitivity Analysis

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Change** |
| Thermokarst Initiation Model as Feature | feature change | Add Thermokarst Initiation Model (training labels) as input feature |
| Random Data as Feature | feature change | Add randomly generated data as an input feature |
| Remove Lat/Long | feature change | Remove latitude and longitude from input features |
| Remove Elevation, Slope and Aspect | feature change | Remove features with very low feature importance |
| Remove Seasonal Precipitations | feature change | Remove features that may be included as part of other features |
| Remove Other Low Impotence Features | feature change | Remove other features that may not be important |
| Remove High Impotence Features | feature change | Remove Highly important features to see how model importance is effected |
| Remove Other Low Impotence Features and Lat/Long | feature change | Remove Highly important features, and geolocation, to see how model importance is effected |
| Vary estimators | Hyperparameter  change | Vary Estimators based on Baseline Random Forest |
| Vary Maximum Tree Depth | Hyperparameter  change | Vary Tree depth around the value based on Baseline Random Forest |
| Vary Maximum decision features | Hyperparameter  change | Vary number of features considered at each decision node based on Baseline Random Forest |
| Vary Maximum Leaf Nodes | Hyperparameter  change | Vary maximum leaf nodes based on Baseline Random Forest |
| Vary minimum samples to split node | Hyperparameter  change | Vary minimum samples to split node based on Baseline Random Forest |
| Vary minimum samples per leaf | Hyperparameter  change | Vary minimum samples per leaf based on Baseline Random Forest |
| Vary Percent of training data used | Hyperparameter  change | Vary Percent of training data used based on Baseline Random Forest |

# Results

# The Baseline Random Forest

Choosing a baseline model from the many trained in the brute force period was done by sorting the models by their MAE. The model with the lowest MAE was selected, along with 11 other semi-randomly picked models with low MAE. The semi random nature of the selection aimed to select models that had a wide range of hyperparameter variation. These models were examined in more detail to determine which model was best to use as a baseline model for further testing.

While selecting the baseline random forest model it was discovered that the FDD data used for training the brute force models was mislabeled, making it off by a year. This created inaccurate results, which was especially apparent from the original feature analysis scores which did not equally score all climate features as expected, though it did not affect the MAE or scores. To fix this, the 12 baseline candidate models were retrained with the corrected data. These retrained models had similar performance to the incorrect models, but with more equally scored all climate features. Because of the similar performance these new models were used as the baseline candidates. One thing that this mistake was able to show was that there may be some universality to the hyperparameter choices.

The accuracy of all baseline options was examined over time, and all models had similar performance year to year(Figure 9). The baseline model could be chosen from the candidates based on the highest , or the lowest MAE. Table 6 summarizes the accuracy of these models. All models scored greater than .95 for their , and all had a MAE less than 2.0. Three options for a baseline model became apparent from these scores as indicated by the first 3 rows of Table 6. Two of these options had the same scores, and the lowest MAE (1.2866). The other option had the highest (0.9802). MAE was selected as the criteria to choose the model because the scores were very similar for the all options. For the two remaining options all hyperparameters were the same except for max depth which was 60 or 100. The model with a max depth of 100 was selected as the baseline model. The baseline models hypermeters are summarized in Table 7.

A picture containing shirt

Description automatically generated

Figure : Comparison of candidate model accuracy over time. Each model predicts similarly for all years tested.

Table 6: Accuracy of baseline candidates. Model names encode the values used by each hyperparameter. The abbreviations are summarized in Table 2.

|  |  |  |
| --- | --- | --- |
| **Name** | **MAE** |  |
| rfm\_e50\_md100\_mfAUTO\_mln50000\_msl8\_mss5\_tdp75 | **1.2866** | 0.9801 |
| rfm\_e50\_md60\_mfAUTO\_mln50000\_msl8\_mss5\_tdp75 | **1.2866** | 0.9801 |
| rfm\_e50\_md60\_mfAUTO\_mln50000\_msl2\_mss10\_tdp75 | 1.2929 | **0.9802** |
| rfm\_e50\_md60\_mfAUTO\_mln50000\_msl4\_mss5\_tdp25 | 1.3393 | 0.9771 |
| rfm\_e50\_md60\_mfAUTO\_mln50000\_msl8\_mss5\_tdp25 | 1.3415 | 0.9759 |
| rfm\_e10\_md60\_mfAUTO\_mln50000\_msl8\_mss2\_tdp50 | 1.3418 | 0.9773 |
| rfm\_e10\_md60\_mfAUTO\_mln50000\_msl8\_mss5\_tdp50 | 1.3418 | 0.9773 |
| rfm\_e10\_md100\_mfAUTO\_mln50000\_msl4\_mss10\_tdp50 | 1.3437 | 0.9777 |
| rfm\_e100\_md25\_mfAUTO\_mln50000\_msl2\_mss5\_tdp25 | 1.8124 | 0.9499 |
| rfm\_e100\_md25\_mfAUTO\_mln50000\_msl2\_mss2\_tdp25 | 1.8127 | 0.9499 |
| rfm\_e100\_md25\_mfAUTO\_mln50000\_msl4\_mss10\_tdp25 | 1.8139 | 0.9493 |

Table 7: Baseline Hyperparameters

|  |  |
| --- | --- |
| Hyperparameter | Value |
| Estimators | 50 |
| Maximum Tree Depth | 100 |
| Number Features Considered for decisions | All (13) |
| Maximum Leaf Node | 50,000 |
| Minimum Samples Needed to Split Node | 5 |
| Minimum Samples per Leaf | 8 |
| Training Data Percent | 75% |

Once the final baseline model was selected it was trained on the Seward Peninsula, as well. The feature importance scores for both regions are shown in Table 8. The numbers vary slightly for each region, but they have the same basic order with the exception of latitude and next summer’s TDD. For the ACP the MAE was 1.29 while it was 0.36 for the SP. was 0.980 for the ACP and 0.997 for the SP. These values indicate the model trained for the SP were more accurate with the same hyperparameters. These models were used as a baseline for all sensitivity analysis performed. Examples of the baseline model results, and the per pixel error are shown in Figure 11 and Figure 12.

A close up of an animal

Description automatically generated

A close up of an animal

Description automatically generated

Figure : Baseline ACP results 2004, and the pixel error from the Original TKI model

A close up of text on a white background

Description automatically generatedA close up of a map

Description automatically generated

Figure : Baseline SP results 2004 and the pixel error from the Original TKI model

Table 8: Baseline feature importance scores

|  |  |  |
| --- | --- | --- |
| Feature | ACP Baseline RF model | SP Baseline RF model |
| Late Summer Precipitation | 13.34 | 14.61 |
| Next Summer Precipitation | 12.75 | 13.12 |
| FDD | 11.70 | 11.49 |
| Early Winter Precipitation | 9.91 | 11.75 |
| TDD | 9.71 | 10.50 |
| Winter Precipitation | 9.58 | 11.20 |
| Summer Precipitation | 9.21 | 9.09 |
| Latitude | 8.40 | 5.14 |
| Next Summer TDD | 8.13 | 7.98 |
| Longitude | 7.24 | 5.11 |
| Slope | 0.02 | 0.02 |
| Elevation | 0.00 | 0.00 |
| Aspect | 0.00 | 0.00 |

# Feature changes

The sensitivity analysis looking at how changes to the input features affects model accuracy was designed to look at random forest robustness, and how random forest models can be used to develop reduced order models. The first two scenarios added input features at training, and were primary looked at random forest model capabilities. The third scenario, removing latitude and longitude was intended to find if a location agnostic model trained on another was as accurate as a location-based model. The remaining scenarios looked at the robustness of random forest models as various features are removed from the training data. All scenarios are summarized in the first half of Table 5. Table 9 summarizes the MAE and scores for each sensitivity analysis scenario.

Table 9: Summary of Feature Based Model Accuracy (check theses numbers)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Seward Peninsula** | | **ACP** | |
| **MAE** |  | **MAE** |  |
| Baseline | 0.3598 | 0.9973 | 1.2866 | 0.9801 |
| Added Original Results | 0.0001 | 1.0000 | 0.0000 | 1.0000 |
| Added Random Data | 0.3619 | 0.9972 | 0.8685 | 0.9781 |
| Removed Latitude and Longitude | 0.4672 | 0.9963 | 1.2685 | 0.9481 |
| Removed Elevation, Slope, and Aspect | 0.3607 | 0.9972 | 0.8399 | 0.9797 |
| Removed Partial Precipitations | 0.5710 | 0.9936 | 1.6709 | 0.9646 |
| Removed Least Important Features | 0.5660 | 0.9941 | 2.6200 | 0.9137 |
| Removed Most Important Features | 0.7298 | 0.9898 | 1.9718 | 0.9506 |
| Removed Most Important Features, and Latitude, Longitude | 1.6535 | 0.9622 | 3.6571 | 0.8334 |

# Add Thermokarst Initiation Model

The thermokarst initiation model was used as the training data for the baseline random forest model, and the purpose of adding this data to the training features was to determine if random forest models are capable of selecting features that are truly important. If this was in fact the case, then the new model would use the TKI model to make decisions at most decision nodes, as the TKI model is also used as the training data. The TKI model was added to the input features and a random forest was retrained using the baseline hyperparameters for the ACP, and the SP. These models almost always used the TKI model at each decision node, as shown in Table 10, Additionally, the MAE, and were near 0 and near 1 respectively which are the best scores for those values. These results indicate that Random forest models are capable of selecting the most important features.

Table 10: Adding TKI model to features, feature importance percentages

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | ACP Baseline RF model | ACP Adding TKI RF Model | SP Baseline RF model | SP Adding TKI RF Model |
| Summer Precipitation | 9.21 | 0.00 | 9.09 | 0.00 |
| Late Summer Precipitation | 13.34 | 0.00 | 14.61 | 0.00 |
| Early Winter Precipitation | 9.91 | 0.00 | 11.75 | 0.00 |
| Winter Precipitation | 9.58 | 0.00 | 11.20 | 0.00 |
| Next Summer Precipitation | 12.75 | 0.00 | 13.12 | 0.00 |
| TDD | 9.71 | 0.00 | 10.50 | 0.00 |
| FDD | 11.70 | 0.00 | 11.49 | 0.00 |
| Next Summer TDD | 8.13 | 0.00 | 7.98 | 0.00 |
| Latitude | 8.40 | 0.00 | 5.14 | 0.00 |
| Longitude | 7.24 | 0.00 | 5.11 | 0.00 |
| Elevation | 0.00 | 0.00 | 0.00 | 0.00 |
| Slope | 0.02 | 0.00 | 0.02 | 0.00 |
| Aspect | 0.00 | 0.00 | 0.00 | 0.00 |
| TKI model | N/A | 100.00 | N/A | 100.00 |

# Add Random Data

The robustness of random forest models indicate that unimportant features should be ignored. To verify if this is an accurate a feature was added consisting fully of random data, for both the ACP and Seward Peninsula. In both cases the new random feature was fully ignored by the decision nodes. There were small changes in the other feature importance scores. The MAE and remained close to the baseline values. These results show that RF models can successfully ignore unimportant features in the training process.

Table 11: Adding random data to features, feature importance percentages

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | ACP Baseline RF model | ACP Adding Random Data | SP Baseline RF model | SP Adding Random Data |
| Summer Precipitation | 9.21 | 9.19 | 9.09 | 9.24 |
| Late Summer Precipitation | 13.34 | 13.19 | 14.61 | 14.70 |
| Early Winter Precipitation | 9.91 | 9.92 | 11.75 | 11.80 |
| Winter Precipitation | 9.58 | 9.69 | 11.20 | 11.04 |
| Next Summer Precipitation | 12.75 | 12.80 | 13.12 | 13.29 |
| TDD | 9.71 | 9.46 | 10.50 | 10.41 |
| FDD | 11.70 | 11.72 | 11.49 | 11.42 |
| Next Summer TDD | 8.13 | 8.33 | 7.98 | 7.97 |
| Latitude | 8.40 | 8.35 | 5.14 | 5.20 |
| Longitude | 7.24 | 7.32 | 5.11 | 4.89 |
| Elevation | 0.00 | 0.00 | 0.00 | 0.00 |
| Slope | 0.02 | 0.02 | 0.02 | 0.02 |
| Aspect | 0.00 | 0.00 | 0.00 | 0.00 |
| Random Data | N/A | 0.00 | N/A | 0.00 |

# Remove Latitude and Longitude

Removing Latitude and Longitude was intended to show if a location agnostic version of a random forest model could be created for determining TKI values. Applying these new models to the region they were trained on produced results with MAE scores comparable to the baseline RF scores. Also, the scores for these models changed very little. This shows that Latitude and Longitude can likely be dropped during the training as long as the model remains restricted to its original training region.

On the other hand, when these models were applied to the other region (ACP model with SP training data, and vice versa) the results were very poor. The MEA increased greatly and scores decreased greatly. The increase in the error is shown in Figure 14 for the ACP were it can be seen that the by pixel error is greater than values shown in

Figure 13. These results show that despite location not being needed to train a successful model to replicate the TKI process, the random forest models are restricted to the region for which they are trained.

A picture containing bird

Description automatically generated

Figure 13: ACP random forest model (no latitude/longitude) on ACP

A close up of an animal

Description automatically generated

Figure 14: Seward Peninsula random forest model (no latitude/longitude) on ACP

# Remove Slope, Elevation, and Aspect

Slope, elevation, and aspect all had near zero feature importance scores for the baseline random forest model, so removing them should have no effect on the accuracy of a model trained without them. When this test was performed the new models the MAE and remained largely unaffected. As seen in the feature importance scores also remained close to the bassline values for the Seward Peninsula, but varied to a greater extent for the ACP. Despite the unexplained changes in the ACP feature importance scores the similar accuracies as indicated by the MAE and show that features with very low importance scores can be removed from further iterations of the random forest models used here.

Table 12: Remove Slope, Elevation, and aspect, feature importance percentages.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | ACP Baseline RF model | ACP Remove TKI RF Model | SP Baseline RF model | SP Remove TKI RF Model |
| Summer Precipitation | 9.21 | 14.16 | 9.09 | 11.62 |
| Late Summer Precipitation | 13.34 | 12.61 | 14.61 | 17.20 |
| Early Winter Precipitation | 9.91 | 14.19 | 11.75 | 12.52 |
| Winter Precipitation | 9.58 | 10.89 | 11.20 | 12.56 |
| Next Summer Precipitation | 12.75 | 15.34 | 13.12 | 14.05 |
| TDD | 9.71 | 10.23 | 10.50 | 11.11 |
| FDD | 11.70 | 12.11 | 11.49 | 12.06 |
| Next Summer TDD | 8.13 | 9.42 | 7.98 | 8.72 |
| Latitude | 8.40 | N/A | 5.14 | N/A |
| Longitude | 7.24 | N/A | 5.11 | N/A |
| Elevation | 0.00 | 0.00 | 0.00 | 0.01 |
| Slope | 0.02 | 1.04 | 0.02 | 0.15 |
| Aspect | 0.00 | 0.00 | 0.00 | 0.01 |

# Remove Partial Season Precipitation

Removing late summer precipitation and early winter precipitation was designed to determine if the random forest model could find that the information they provide, as it was also included in the full season summer and winter precipitation measures. There was a large change in the MAE for the new random forest models in this test while the score stayed close to the baseline value. The remaining feature importance scores, shown in Table 13, also all increased, as opposed to just the full season summer and winter precipitation scores, as would be expected if the RF model was capable of detecting this type of data differentiation. These results show that removing features that may be included twice is not advisable.

Table 13: Remove partial season precipitation features: feature importance percentages

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | ACP Baseline RF model | ACP Remove TKI RF Model | SP Baseline RF model | SP Remove TKI RF Model |
| Summer Precipitation | 9.21 | 16.77 | 9.09 | 14.86 |
| Late Summer Precipitation | 13.34 | N/A | 14.61 | N/A |
| Early Winter Precipitation | 9.91 | N/A | 11.75 | N/A |
| Winter Precipitation | 9.58 | 11.87 | 11.20 | 15.00 |
| Next Summer Precipitation | 12.75 | 16.54 | 13.12 | 15.86 |
| TDD | 9.71 | 10.76 | 10.50 | 13.77 |
| FDD | 11.70 | 12.25 | 11.49 | 15.27 |
| Next Summer TDD | 8.13 | 12.36 | 7.98 | 11.15 |
| Latitude | 8.40 | 7.85 | 5.14 | 6.87 |
| Longitude | 7.24 | 11.60 | 5.11 | 7.01 |
| Elevation | 0.00 | 0.00 | 0.00 | 0.02 |
| Slope | 0.02 | 0.07 | 0.02 | 0.19 |
| Aspect | 0.00 | 0.00 | 0.00 | 0.01 |

# Remove Features with Low Importance

Being able to remove features with low feature importance scores would be an important step in developing reduced order models using random forests. For this test all features with a baseline feature importance score less than 10 were removed from the training data. These were the TDD, next summer’s TDD, latitude, longitude, elevation, slope, and aspect. scores indicated that these new models were largely still successful at modeling the original TKI model. In contrast the MAE for these new models almost doubled indicating these features still contribute to the outcome.

# Remove Features with High Importance

Removing features with high feature importance scores was expected to negatively affect model accuracy greatly. Despite this a test removing the features with baseline feature importance scores greater than 12 (summer precipitation, winter precipitation, and next summer precipitation) was performed to examine if the random forest models could use less important features to successfully predict the TKI model. This test was not successful at producing accurate models, as expected. The MAE increased greatly, and the scores decreased more than for any of the previously described scenarios.

Despite the findings that removing the most important features did not produces useful models, the same scenario was performed again well also removing latitude and longitude. This produced the worst results of all of the scenarios with the MAE tripling, and the scores dropping to the lowest values seen. This adds further to the conclusion that each model must be retrained for new regions to be effective.

# Hyperparameter changes

# Vary Estimators

The number of estimators used in a random forest is considered to be one of the most important hyperparameters. As the value increases, the results of the random forest model should approach the actual value of what is being modeled. The baseline random forest model for thermokarst initiation had 50 estimators in its forest. To examine how changing this value affects the accuracy of the model values between 1 and 100 were used. As expected, low values decreased the accuracy of the model while higher values yield better values. However, increasing the number of estimators beyond 50 did not noticeably increase the accuracy of the model, and for the arctic coastal plain created a slight decrease in accuracy.

Figure 15 shows the changes in and MAE for varying the number of estimators graphicly. Notice that the MAE decrease as the score increases. This is expected and true for all of the hyperparameter sensitivity analysis results, so only will be shown in the remainder of section 5.3. The remaining MAE plots are provided in the appendix.

A picture containing light, traffic, sitting, red

Description automatically generatedA picture containing light, sitting, traffic, red

Description automatically generated

Figure 15: Estimator Sensitivity Analysis

# Vary Maximum Tree Depth

The maximum tree depth controls the size of each tree in the forest by limiting their depth. Because theses trees are binary in nature, this creates maximum possible maximum possible leaf nodes where *n* is the trees depth. The baseline value was 100 which would create a tree with possible leaf nodes. This value is very high and could not possibly be reached, as the value set for maximum leaf nodes was 50,000. Because of the already large value and the maximum leaf node cutoff only values up to the baseline were used for this sensitivity analysis.

Maximum depth values between 1 and 100 were used for the sensitivity analysis with a greater density being used for tests between 1 and 20. Increasing the value increased the accuracy in an almost linear fashion until reaching an upper bound as shown in Figure 16 with the scores. The value increases steadily from near 0 with a maximum depth or 1 until somewhere between 20 and 30 when the value reaches a ceiling.

A picture containing light, traffic, sitting, hanging

Description automatically generated

Figure 16: Max Tree Depth Sensitivity analysis

# Vary Features Used for Decisions

The number of features used in decisions is also considered an important hyperparameter, as it helps to create more randomness in the trees created for the random forest. The baseline value found in the brute force testing was AUTO which has each tree consider the full number of features at each decision node in this case 13. The other values tested were SQRT and LOG2 which consider the square root and log2 of the full number of features at each decision node. This was 3 for each, as such values between 3 and 13 were used for the sensitivity analysis. The sensitivity analysis shows that using the default value was not optimal for the number of decision features, as the score was actually slightly less then its maximum which was achieved with around 11 decision features, as shown in Figure 17.

A close up of a red light at night

Description automatically generated

Figure 17: Maximum Decision Features Sensitivity Analyis

# Vary Maximum Leaf Nodes

Varying the maximum leaf nodes is another possible option for controlling the maximum size of each tree in the forest. The baseline value was 50,000 which likely controlled the size of each tree as it is much less than . Values between 12,500, and 100,000 were tested during the sensitivity analysis. As with the Maximum Tree Depth, increasing maximum leaf nodes increases the accuracy of each random forest. Using a value higher than the baseline value will likely create better models, as seen in Figure 18 though the score is already extremely high.

A picture containing outdoor, light, traffic, sitting

Description automatically generated

Figure 18: Maximum Leaf Node Sensitivity Analysis

# Vary Minimum Samples to Spit Node

The minimum samples to split a decision node controls the size of each tree to a much lesser extent. This value indicates how many samples are required to split a decision node. The default value is 2 which is also the lowest possible value. The value used in the baseline model was 5. For the sensitivity analysis values between 2 and 15 were used. For each region the score only changed with in a range of .005. This indicates that any small value for this hyperparameter would create a good model.

A picture containing outdoor, light, traffic, sitting

Description automatically generated

Figure 19: Minimum samples for Splitting decision node Sensitivity analysis

# Vary Minimum Samples per Leaf

The Minimum samples per leaf controls the how many samples are allowed to be at a leaf node in each decision tree. The default value is 1, and the value found for the baseline value was 8. Increasing this value slowly decreased the model accuracy for the Seward Peninsula, but not for the ACP where the decrees only appeared to occur after the baseline value. This is demonstrated in Figure 20. Despite this as with the Minimum samples to split each node, any small value is likely adequate for this model.

A picture containing sitting, traffic, light, red

Description automatically generated

Figure 20: Minimum Samples per Leaf Sensitivity Analysis

# Vary Percent Data Used for Training

The Percentage of total data used for training effects model accuracy in a linear fashion. The bassline value used was 75% of the full data. For the sensitivity analysis 25% and 50% were also tested. These results are shown in Figure 21. Increasing the percent data was one sensitivity analysis where good timing data was collected at least for the Seward Peninsula. This showed a linear increase were doubling the data used doubled the time taken (Table 14). All values tested created accurate models with high scores. The value used for this hyperparameter can likely be selected based on the time available to run the model.

Table 14: Seward peninsula percent data sensitivity analysis timing

|  |  |
| --- | --- |
| Percent Data | Time Minuets |
| 25% | 03:00.0 |
| 50% | 06:22.3 |
| 75% | 09:43.4 |

A picture containing outdoor, light, star, road

Description automatically generated

Figure 21: Training Data Percent Sensitivity Analysis

# A Final Model

Using the insight gained in the feature based, and hyperparameter sensitivity analyses a final set of models was developed. These models were trained for each region separately. As indicated by the feature-based analysis was used to train the models. The hyperparameters used were very similar to the base line model. Key differences include only using maximum leaf nodes to control tree size and using 11 as the number of decision features considered. Max leaf nodes was selected as it was more apparent where the increasing value became less useful than adjusting Max depth. A value of 75,000 was used for this parameter. 11 was selected for the decision feature consideration because it was between the two values that produced the best results during the sensitivity analysis. The hyperparameters used in the final model are summarized in Table 15.

Table 15: Summary of Final Hyperparameters

|  |  |  |
| --- | --- | --- |
| Hyperparameter | Baseline | Final |
| Estimators | 50 | 50 |
| Maximum Tree Depth | 100 | None |
| Number Features Considered for decisions | Auto (13) | 11 |
| Maximum Leaf Node | 50,000 | 75,000 |
| Minimum Samples Needed to Split Node | 5 | 5 |
| Minimum Samples per Leaf | 8 | 8 |
| Training Data Percent | 75% | 75% |

The final model yielded better results than the baseline mode for both regions. There was a significant decrease ( for SP, for the ACP) in the MAE showing that predictions were closer to the expected value. The increase in the score was less apparent, but those values are already very high. Table 16 summarize the final accuracy results.

Table 16: Final Accuracy Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SP | | ACP | |
| Measure | Baseline | Final | Baseline | Final |
| MAE | 0.3598 | 0.1790 | 1.2866 | 0.9662 |
|  | 0.9973 | 0.9983 | 0.9801 | 0.9885 |

# Discussion

The feature-based sensitivity analysis yielded three important results. The first is that random forest models are able to successfully find important features and use them to inform their decision process. The second is that models are not a useful tool for elimination input features if all features are equally important to the decision process. Finally, this analysis showed that the models must be retrained for each region they are required to be run on.

Random forest models are a robust tool for finding relationships between many possible inputs and a given output. This is especially supported by the sensitivity analysis where the original TKI model, and random data were added. When the TKI model was added the RF process the model used it as the main driver of the decision process. While when the random data was added it was ignored. The robustness of random forest models is supported to a lesser extent by the fact that the random forest models assigned equal weight to features used to train the original TKI model.

One of the purposes of using random forest models to replicate the thermokarst imitation model was to determine if they are a useful tool to develop reduced order models. From the results of the feature-based analysis it was not immediately apparent that this is possible. All of the features from the original model were indicated as being equally important to the random forest decision process, as indicated by the feature importance values. Additionally, removing any of these features from the training process reduced model accuracy. Conversely removing features that were not important (Slope, elevation, and aspect) did not affect model accuracy. This shows that using random forests to develop a reduced order model may be possible if any of the inputs to the original model were shown to be considerably less important than the other inputs. Performing these analyses with other models would need to be completed to determine if this is truly the case.

As well as being less complex, it is also desired that reduced order models be more efficient in time. While useful timing data for comparing was not collected, all random forest models trained took less than 8 hours to train (Find this Number for ACP brute force analysis, 1.77 hours for final ACP model). This may not seem efficient given that the original TKI model takes a few minutes to generate predictions, but the prediction time for the random forest models took 45 seconds for the final ACP model. The benefit would come from applying the random forest model to different data sets (for example using projected data to create future forecasts) though as other results have shown this is limited to single regions for this project.

An attempt to develop a location agnostic model was attempted during the feature-based sensitivity analysis. When latitude and longitude were removed model, accuracy was not significantly reduced, but this only applied to the region for which the model was trained. These models did not perform well for the other sample region. This shows that random forest models should be limited to the regions for which they are trained. This could also lead to limitations to this RF process if continued climate change causes changes to temperature or precipitation that are outside of the ranges of values seen in the training data.

The hyperparameter sensitivity analysis showed that the most important hyperparameters are indeed the number of estimators, and number of decision features. The analysis showed that increasing the values for these features increased model accuracy in the quickest fashion. The results also indicated that using only one hyperparameter to control tree size is advisable either choose maximum leaf nodes or maximum tree depth. Finally changing training data percentages increases the training time in a linear way, but for this problem accuracy did not increase in an extreme way as more data was added.

One thing that using a random forest model may be useful for is determining if assumptions about inputs made in an existing model are accurate. A related use would be using a random forest model of a complex model, to determine if running the full model is worthwhile. For the TKI process all of the original inputs were indicated as being equally important during the analysis. This was expected as the original TKI climate priming model weighed all inputs equally. If this had not been the case it would indicate that the original model would need to be reworked. Again, preforming this type of analysis on other, perhaps more complex models, would verify the usefulness of this process.

# Future Work

Designing a testing framework that give useful timing data, for training and evaluation, would be a good future project that would give another point of comparison for random forest models. Random forest models can accurately simulate other models, but from the data gathered here it is difficult to determine how long training will take when a hyperparameter is changed. Better timing data would allow for more informed decisions when choosing models if time efficiency is important. This could be important for more complex models than the TKI model examined here.

Further improving the underlying TKI model and retraining the random forest models can give new data to show the validity of random forest models. The original TKI model is currently uncalibrated and lacks meaningful verification from real world observations. These observations are hard to come by due to lack in existing data for the regions being compared. When the original model’s results are verified more concrete statements can be made about random forest accuracy. Additionally, using the random forest models developed here to predict future TKI values would help to determine the robustness of the random forest process. Values predicted from projected data could be compared to values with similar input features to determine validity.

Random forest models are a good tool for simulating the TKI process examined here, but they could also be a useful tool for simulating other parts of the full ATM. Many parts of the ATM are more complex than the TKI component, and a random forest model that could be pretrained and integrated into the ATM could save time when running the full ATM. They may also be useful in determining which inputs to the ATM, or other complex model, are truly important through the feature importance scores. This may aid in the development of ROMs. Random forests could also be used to create quick initial results from new data sets to see if they may be useful if run in the full model.

Another improvement to these random forest models may come from applying them to smaller regions. The SP models performed better than the ACP model in these tests. To determine if smaller regions are truly effective the ACP could be split into smaller regions, and a random forest could be trained for each. The ACP and SP data could be combined. Based on this hypothesis the smaller regions should have better accuracy while the larger should have poorer performance.

Random Forest models can have similar performance to other machine learning methods like SVMs and ANNs, but best accuracy is application dependent (Cutler, et al. 2007) (Were, et al. 2015) (Pal 2007) (Wang, et al. 2016). Applying those methods this problem could determine if RF was the best choice for this project. Random forests do have two advantages the other methods do not. They are simpler to use requiring less parameters to be set (Pal 2007) (Wang, et al. 2016), and they provide a variable importance measure that the other methods do not (Cutler, et al. 2007) .

# Conclusion

Random forest models were able to simulate the results of an existing model as shown by their application to TKI modeling. It was also shown that random forest models are able to find important features and ignore unimportant ones. Sensitivity to changes in hyperparameters was examined. From this, it was determined that the number of estimators, number of features considered at each node are important to the process which is in agreement with future work. Finally, the interaction between hyperparameters was examined which showed that only one of these should be set.

Looking at random forests showed their value for verifying assumptions in existing models. They can also be used in and for the development of reduced order models. The feature importance scores can be used to demine which inputs may be eliminated, and for very complex models they may be used as a ROM, as they only need to be trained once, and prediction times are generally quick. Though this project only looked a applying random forests to the TKI part of the ATM they would also be useful in other portions of the ATM.

# Acknowledgments

The research presented in this work was funded by The Alaska Climate Adaptation Science Center and the Department of Energy Office of Biological and Environmental Research Next Generation Ecosystem Experiment (NGEE-Arctic) project.

# References

Antoulas, A.C. 2004. "Approximation of Large-Scale Dynamical Systems: An Overview." *IFAC Proceedings Volumes* 37 (11): 19-28.

Breimen, Leo. 2001. "Random Forests." *Machine Learning* 45: 5-32.

n.d. *Conda.* Accessed 02 05, 2019. https://conda.io/projects/conda/en/latest/.

Cutler, Richard D., Thomas C. Edwards, Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson, and Joshua J. Lawler. 2007. "Random Forests for Classifaction in Ecology." *Ecology* 2783-2792.

Davis, Neil. 2001. *Permafrost.* Fairbanks, Alaska: University of Alaska Press.

DGGS. n.d. *DGGS elevation portal.* Accessed 11 02, 2019. https://elevation.alaska.gov/.

Farquharson, L.M., D.H. Mann, Grosse G., Jones B.M., and Romanovsky V.E. 2016. "Spatial distobution of thermokarst terrain in Arctic Alaska." *Geomprphology* 273: 116-133.

Gandodamage, Chandana, Rowland Joel C., Susan S. Hubbard, Steven P. Brumby, Anna K. Liljedahl, Harukdo Wainwright, Cathy J. Wilson, et al. 2014. "Extapolating active layer thickness measurements across Arctic poolygoonal terrain using LiDAR and NDVI data sets." *Water Resurces Research* 50: 6339-6357.

Hastie, Trevor, Robert Tibshirani, and Friedman Jerome. 2017. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction Second Edition.* Springer.

Herman, Gregorgy R., and Russ J. Schumacher. 2018. "Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation WIth Randm Forests." *America Meteorlogical Society .*

Jorgenson, M. T., Kenevskiy M., Shur Y., Moskalenko N., Brown D. N. R., Wickland K., R. Striegl, and Koch J. 2015. "Role of ground ice dynamics and eclgical feedbacks in recent ice wedge degradtion and stabilization." *Joournal of Geophysical Research: Earth Surface* 2280-2297.

Jorgenson, M. Torre, Yuri L. Shur, and Erik R. Pullman. 2006. "Abrupt Increase in permafrost degradion in Arctic Alaska." *Geophysical Research Letters* 33.

Kanevskiy, Mikhail, Yuri Shur, Torre Jorgensoon, Dana R.N. Brown, Nataliya Moskalenkom, Jerry Brown, Dnald A. Walker, Martha K. Raynolds, and Marcel Buchhorn. 2017. "Defradation and Stabilization of Ice Wedges Implications for Assessing Risk of Thermokarst in Northern Alaska." *Geomorphology* 297: 20-42.

Lara, Mark J., McGuire A. David, Eugenie Euskirchen, Tweedie Craig E., and Kenneth M., Skurikhin Alexei N., Romanivsky Valdimir E., Grisse Guido, Bolton, W. Robert, Genet, Helene Hinkel. 2014. "Polygonal Tundra Geomrpholoogical Change in Respnse to Warming Alters Future CO2 and CH4 Flux on The Barrow Peninsula." *Golbal Change Biology.*

Liljedahl, A.K., J. Boije, R.P. Dannen, Fedorov A.N., G.V. Frost, G. Grosse, L.D. Hinzman, et al. 2016. "Pan-Arctic ice-wedge degradtion in warming permafrost and its influence on tundra hydroloogy." *Nature Geoscience* 9.

McGovern, Amy, Kimberly L. Elmore, David Jonh II Gagne, Sue Ellen Haupt, Christopher D. Karstens, Ryan Lagerquist, Travis Smith, and John K. Williams. 2017. "Using Artifical Intelligence to Improve Real-Time Decision-Making for High-Impact Weather." *American Meteorological Society.*

Pal, M. 2007. "Random forest Classifier for remote sensing classification." *Internatinal Journal f Remote Sensing* 26: 217-222.

Pedregosa, F., G Varoquaux, A Gramfrt, V Michel, B Thirin, O Grisel, M Blondel, et al. 2011. "Scikit-learn: Machine Learning in Pythn." *Journal of Machine Learning Research* 2825-2830.

Pollard, W. 2018. "Chapter 15 - Periglacial Processes in Glacial Environments." In *Past Glacial Environments*, edited by John Menzies and Jaap J.M. van der Meer, 537-564. Elsevier.

Raynolds, Martha K., Donald A. Walker, Kenneth J. Ambrosius, Brown Jerry, Kayer Everett, Kanevskiy Mikhail, Gary P. Koofinas, Vladimr E. Romanovsky, Yuri Shur, and Patrick J. Webber. 2014. "Cumulative geoecological effects of 62 years of infrastructure and climate change in ice-rich permafrost landscapes, Prudoe Bay Oilfeild, Alaska." *Global Change Biolgy* 1211-1224.

Robert, Bolton W., Hélène Genet, Mark Lara, Valdimir Romanovsky, and William Riley. 2018. "Identifying Regions Susceptible to Thermokarst Initiation on the Alaskan Arctic Coastal Plain for the Alaska Thermokarst Model."

Rowley, Taylor, John R. Giardino, Raquel Granados-Aguilar, and John D. Vitek. 2015. "Chapter 13 - Periglacial Processes and Landforms in the Critical Zone." Edited by John R. Giardino and Chris Houser. *Developments in Earth Surface Processes* (Elsevier) 19: 397-447.

Scenarios Network for Alaska and Arctic Planing. 2019. "Historical Monthly Temperature - 1 km CRU TS." University of Alaska. Accessed 11 04, 2019. http://ckan.snap.uaf.edu/dataset/historical-monthly-temperature-1-km-cru-ts.

Scenarios Network for Alaska and Arctic Planning. 2019. "Historical Monthly Precipitation - 1 km CRU TS." University of Alaska. Accessed 11 04, 2019. http://ckan.snap.uaf.edu/dataset/historical-monthly-precipitation-1-km-cru-ts.

scikit-learn-developers. n.d. *3.2.4.3.2. sklearn.ensemble.RandomForestRegressor - scikit-learn 0.21.3 documentation.* Accessed 02 05, 2020. https://scikit-learn.org/0.21/modules/generated/sklearn.ensemble.RandomForestRegressor.html#sklearn.ensemble.RandomForestRegressor.

spatialreference.org. 2006. *Spatial Reference epsg projection 3338.* Accessed 02 12, 2020. https://spatialreference.org/ref/epsg/nad83-alaska-albers/.

Wang, Li'ai, Xudong Zhou, Xinkai Zhu, Zhadodi Dong, and Wenshan Guo. 2016. "Estimation of biomass in wheat using randm forest regressin algorithm and remotesensing data." *The Crop Journal* 4: 212-291.

Were, Kennedy, Dieu Tien Bui, Oystein B. Dick, and Bal Ram Singh. 2015. "A comparitive assesment of support vector regression, artifical neural networks, and random forests for predicting and mapping soil orgianc carbon stoocks across Afromontane landscape." *Ecooloogical Indicators* 394-493.

# Appendix A: Example Precipitation Maps

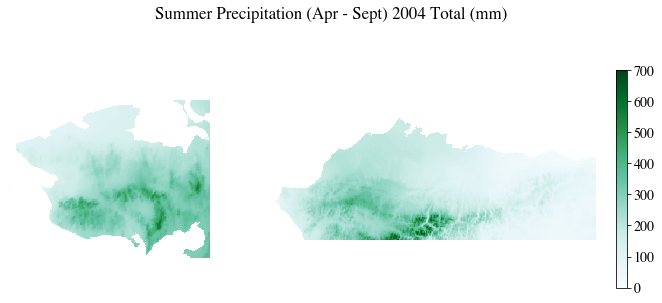


Figure : Summer Precipitation for 2004 (Seward Peninsula Left, ACP Right

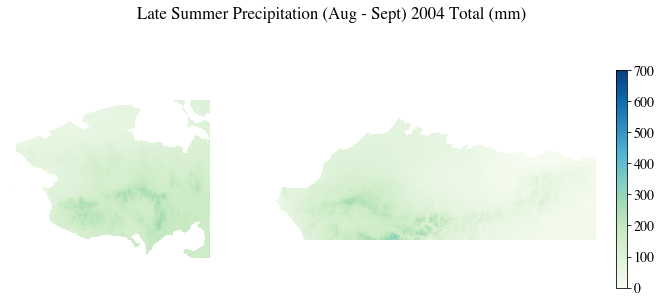


Figure : Late Summer Precipitation for 2004 (Seward Peninsula Left, ACP Right)

A close up of a map

Description automatically generated

Figure : Early Winter Precipitation for 2004-2005 (Seward Peninsula Left, ACP Right)

A close up of a logo

Description automatically generated

Figure :Total Winter Precipitation for 2004 - 2005 (Seward Peninsula Left, ACP Right)

**Appendix B: Climate Priming Model TKI predictions**

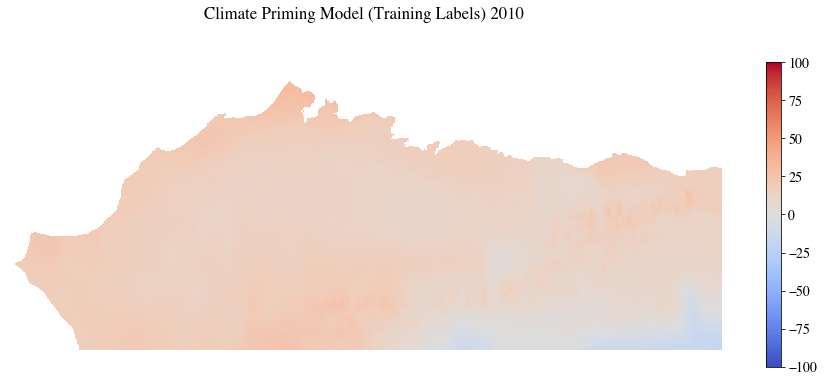
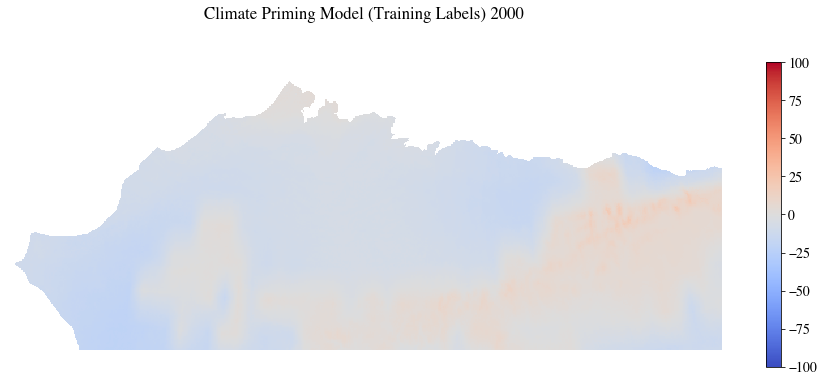
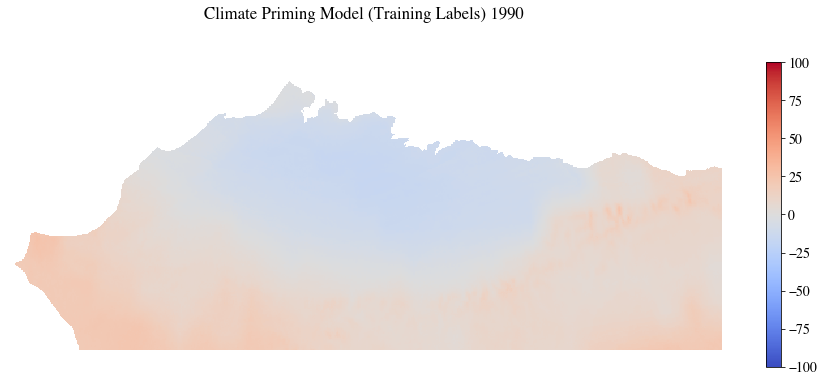
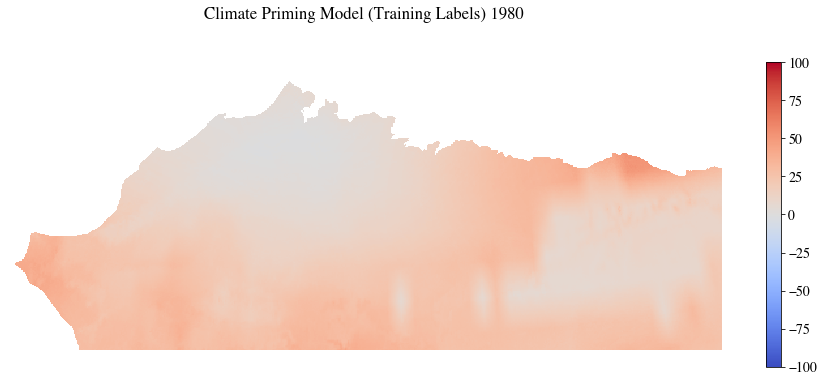
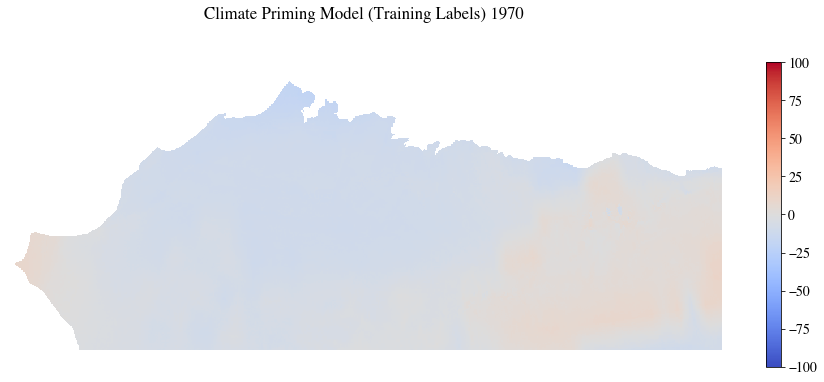
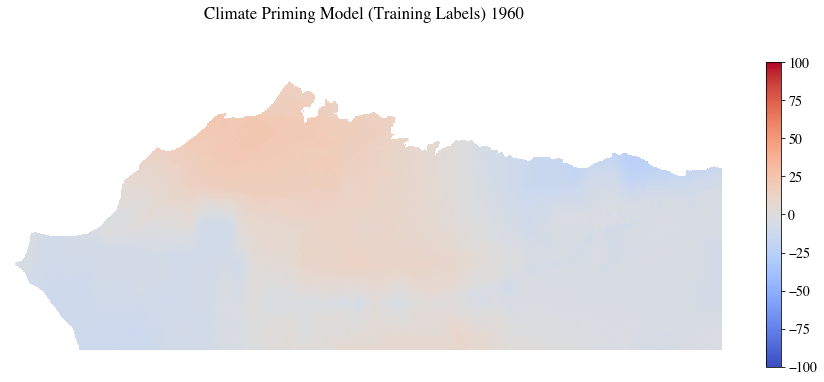
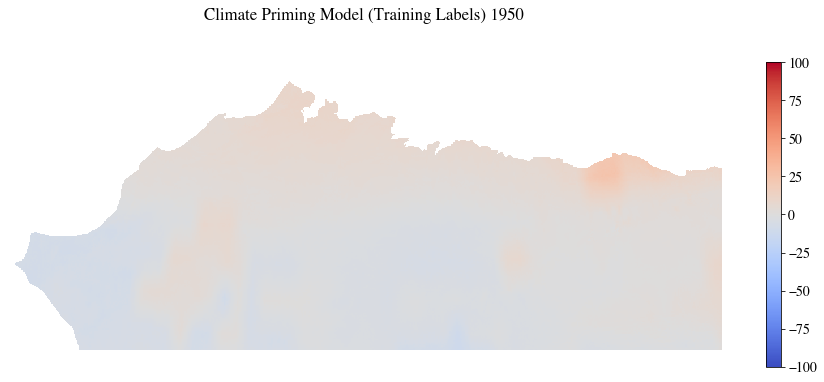
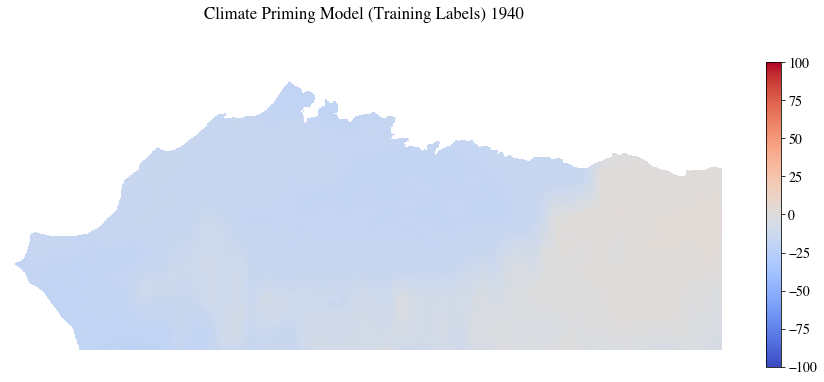
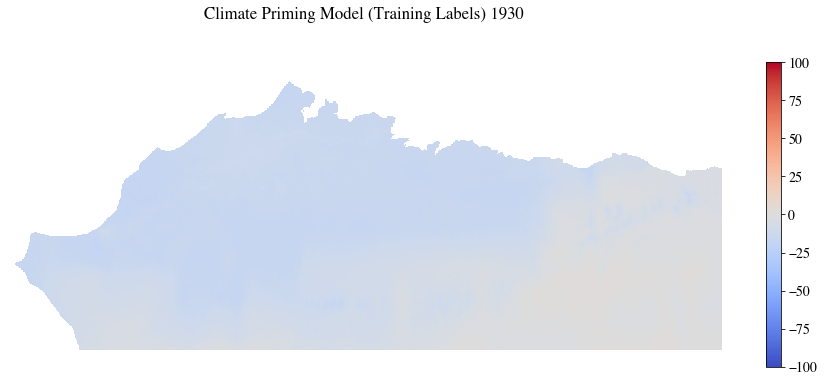
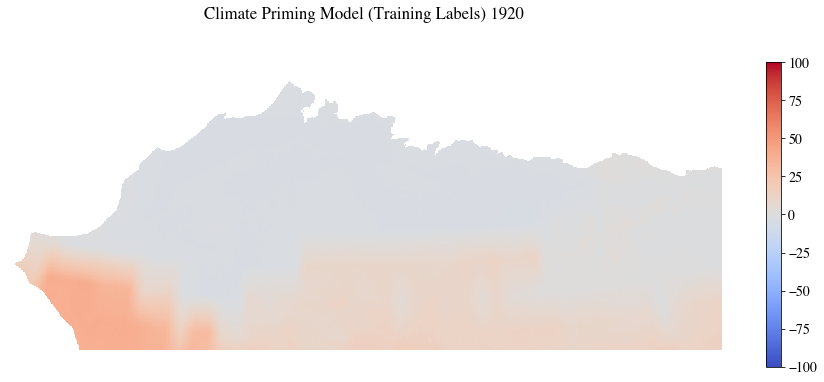


Figure : ACP TKI examples

A screenshot of a cell phone

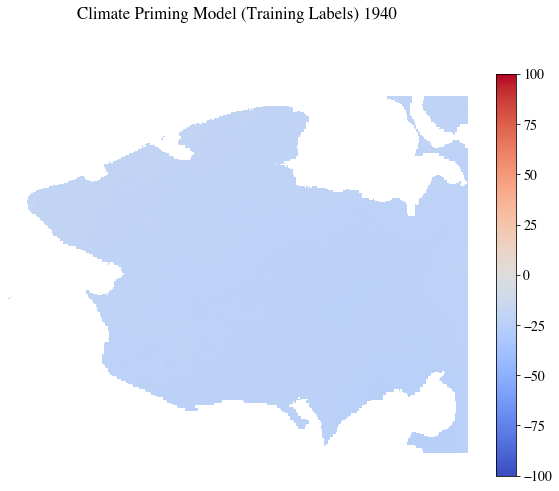
Description automatically generatedA close up of a map

Description automatically generated

Figure : Seward Peninisula TKI for 1965 and 2004

A picture containing text

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure : Seward Peninsula TKI examples

**Appendix C: MAE hyperparameter sensitivity analysis charts**

**A picture containing light, outdoor, traffic, sitting

Description automatically generated**

Figure : MAE for maximum tree depth sensitivity analysis

**A picture containing light, outdoor, traffic, sitting

Description automatically generated**

Figure :MAE for maximum decision feature sensitivity analysis

**A picture containing light, black, sitting, traffic

Description automatically generated**

Figure : MAE for maximum leaf nodes sensitivity analysis

**A picture containing sitting, red, light, traffic

Description automatically generated**

Figure : MAE for minimum samples for node split sensitivity analysis