Abstract

The Nenana ice classic is an annual betting competition where people can put in guesses on when the ice of the nearby Tanaka river will break. Because of this, the exact ice breakup date and time of the Tanaka river has been recorded for the last 105 years. Previous research has shown a relation between the breakup dates and climate related variables. These relations where used to forecast the breakup dates. The aim of this report is to improve the existing forecasting models. To do so, some of the used variables are extensively analysed and a new variable is introduced. The machine learning technique of artificial neural network modelling will be used to create two models. The first model will relate the variables to the breakup dates, without taking the betting deadline of April 1st into account. The results of this model were an improvement of the previous models with a root mean squared error of 2.73 days and a mean absolute error of 2.28 days. The second model was bounded by the betting deadline of April 1st and showed far less performance with a root mean squared error of 5.98 days and a mean absolute error of 4.83 days. The main reason for this poor performance is the lack of relevant variables that could be included in the model. Although further research is required, this report gives some interesting insights in the effect certain variables have on the ice breakup dates as well as the use of artificial neural networks for capturing these relations.

Introduction

The Nenana Ice Classic is an annual betting competition which takes place in the small town of Nenana, Alaska. It started In 1917, when a group of railroad engineers started putting bets on when the ice of the nearby Tanaka river would break. Since then, the guessing game has turned into an annual tradition with thousands of participants and a jackpot as high as $300.000 (“Nenana Ice Classic”, 2022). Every year during freeze up in November, a tripod is placed on the ice. The moment of break up is defined as the moment when the tripod falls over. This triggers a clock to stop and this way, the exact ice breakup date and time of the river for the last 105 years is recorded, resulting in a nontraditional data set.

The ice normally breaks up in late April or early May. However, previous research on the data set of the Nenana ice classic by Van Asselt, 2020 and Terwogt, 2021 showed the existence of relations between the breakup dates and certain climate related variables. They both made models that used those relations in order to forecast the ice breakup dates. This report will build upon their work and tries to to improve the previous made models by extensively analysing some of the variables used and introducing new ones. Also, the report introduces a new machine learning approach called artificial neural network modelling to relate the variables to the breakup dates. The main questions that will be investigated in this report are: *How can the current climate variables data be improved?* and *Are better results obtained when using the new data and modelling technique?*.

In order to constructively substantiate answers for the research questions, this report is structured in five chapters. First, the context of the research is given. This consists of theoretical background about the climate of Nenana and its surroundings, ice breakup, previous research on the same topic and an introduction to artificial neural networks. The second chapter provides an analysis on the temperature, ice thickness and discharge data. Chapter two also introduces a new variable that will be used in the model. Next, the method of obtaining the results will be explained. In the fourth chapter, the results will provided and be discussed. The final chapter answers the research questions and draws conclusions out of the findings. Also, new research recommendations are given.

Contents

1 Context

This chapter introduces the context of this research. First, a synopsis of the climate and environmental subjects related to the Nenana Ice Classic is given. Next, previous research on the Nenana Ice Classic is summarized and the following research gaps are discussed. Last, the concept of artificial neural networks is introduced and its use is motivated.

1.1. Climate and Environment

This section provides an overview of the climate and environmental facets related to the Nenana Ice Classic. The section will work from small scale (Nenana) to larger scale (Alaska). This section is mainly based on the analysis of Terwogt, 2021 and Van Asselt, 2020. However, some new insights are presented as well.

1.1.1. Nenana

The city of Nenana is a small city, situated in the centre of Alaska, United States. It is located near the intersection of the Nenana river (South) and the the Tanana river (East). The city was incorporated in 1921 and has a population of 363 (“City of Nenana”, 2021). The existence of the city has its roots in the gold mining industry and just after it was founded, in 1917, the start of the tradition of the Nenana Ice Classic began. The tripod used for the Nenana Ice Classic is located on the Tanana river, between the Mears Memorial Bridge and the George Parks Highway bridge. The tripod is placed at about 90 meters from the shore (“Nenana Ice Classic”, 2022). The marker in figure 1.1 gives a visual indication of the location of the tripod.

Figure 1.1: Approximate location of Nenana Ice Classic Tripod, taken from Apple Maps.

The air temperature in Nenana can differ between -25°C in January and +22°C in July. The summers are often wetter than the winters with a mean precipitation of about 58 mm in July and 5 mm in March (Weatherspark, 2022).

1.1.2. Tanana River Basin

The Tanana river basin is located in the central east of Alaska. It covers about 115.500 km^2 and is bordered on the north by the Yukon-Tanana Highlands and on the south by the Alaska range. Its tributaries are either glacially fed rivers coming from the Alaska ranges or non-glacially fed rivers coming from the Yukon-Tanana Highlands. Eighty-five percent of its annual discharge comes from the Alaska range (Collins, 1990).

The climate of the Tanana river basin can be classified as continental. It has long, cold winters and short, but hot summers. The annual mean temperature is about -3.5°C, but extremes of -52°C and +35°C do occur. The precipitation varies from 250 to 560 mm/year with snowfall of 76-150 cm/year (Collins, 1990).

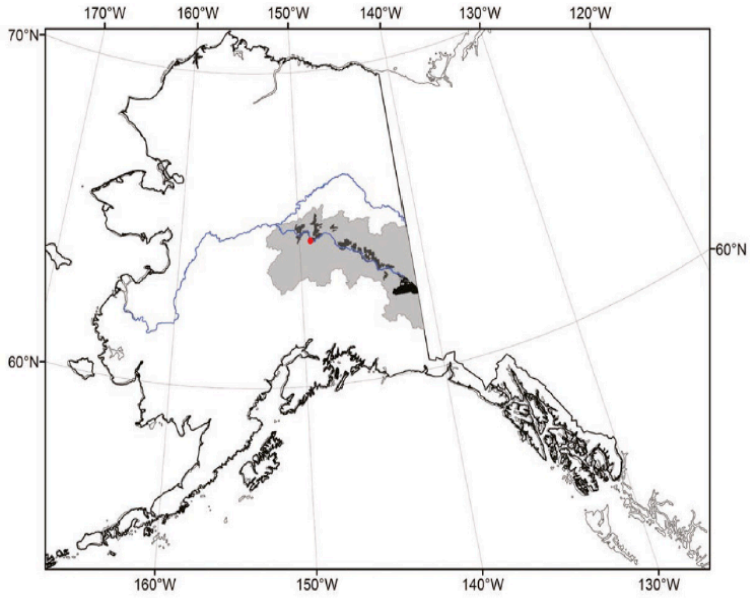


Figure 1.2: Location of Tanana river basin (grey), Yukon and Tanana river (blue) and Nenana (red dot) (edit by Terwogt, 2021, from original figure by Pattison et al., 2018).

1.1.3. Alaska

Alaska is Northernmost state of the United States. It is bordered by the Pacific Ocean, Bering sea, Arctic Ocean and Canada. It has a subarctic climate (Kottek et al., 2006) and because of unsteady nature and correlation between climate forcings, the temperature and precipitation in Alaska can change significantly. It is concluded by Papineau, 2016 that large scale climate forcings can be linked to the climate patterns of Alaska. The two most important climate forcings for Alaska are described below.

- Pacific Decadal Oscillation (PDO)
  PDO is the fluctuation of the sea surface temperature (SST) of the Central North Pacific Ocean. It has a phase of about 20 to 30 years and a 2 to 5 year reversal occurrence. When a positive phase occurs, the sea surface temperatures are higher than normal, resulting in warmer SST's on near the shore of Alaska.

- El Nino and Southern Oscillation (ENSO)
  ENSO refers to the fluctuations of SST's in equatorial and subtropical regions. It can be divided in two phenomena: El Nino and El Nina, which have opposite effects. El Nino causes water to warm up while El Nina causes the water to cool. Both phenomena last for about 12-18 months and can

be measured by water temperature and air pressure. During El Nina, polar jet air streams spend more time at lower latitudes, this leads to cooler temperatures in Alaska. (Papineau, 2016, Van Asselt, 2020).

The large scale climate forcings described above are linked to ice breakup by Bieniek et al., 2011. They conclude that during El Nino, fewer storms will occur in Alaska, and air temperatures will be higher. This will lead to an early ice breakup. During El Nina, the opposite happens, meaning more storms occur and air temperature is lower. This leads to a late breakup date. In the research of Van Asselt, 2020 and Terwogt, 2021, they based the PDO and ENSO variables on Bieniek et al., 2011. The same will be done here.

1.2. Ice breakup

This section gives an insight in the process of ice breakup.

1.2.1. Definition

River ice breakup is an annual event that takes place when milder temperature conditions or increasing discharge cause ice covers on frozen rivers to break. It is a crucial moment that triggers many other biological processes (such as the migratory behaviour of fish) as the ice cover disintegrates and the river opens up (Beltaos, 1997). Also, flooding caused by ice jams can have severe consequences (Beltaos, 2003). Beltaos, 2003 describes the process of ice breakup as a succession of four distinct phases: pre-breakup, onset, drive and wash. The pre-breakup phase is characterized by thermally induced reductions in thickness and strength. This will make the ice more vulnerable for fracture and movement. The onset is defined by the ice fracturing in smaller blocks. This is caused by an increasing discharge caused by ice and snow melting and higher precipitation rates. The drive is the movement of the ice blocks and slabs by the current and the wash is when all ice parts have been washed away.

1.2.2. Ice Dynamics

According to Beltaos, 2003 ice breakup can be demarcated into two different extremes. *Thermal* decay of ice occurs when mild and warm temperatures are combined with low run-off. The ice will melt slowly until it disintegrates into the water under the limited forces of the current. In case of *mechanical* ice breakup, river run off is often very high and is caused by high precipitation rates and ice and snow melting. The higher run-off results in high hydrodynamic forces on the ice, causing the ice to lift and break into smaller blocks. This form of ice breakup has a high risk for ice jams. Normally, a breakup will be somewhere in between these two extremes, meaning that the breaking of the ice is both caused by thermal and mechanical influences.

1.3. Previous Research

As described in section 1.2, ice breakup can have severe consequences. Because of this, many researchers have tried to relate and forecast the breakup dates of river ice with climate variables. This report will build upon the work of Van Asselt, 2020 and Terwogt, 2021. Their objective was to relate the ice breakup process of the Tanana river with chosen climate related variables and to construct a forecasting model to predict future ice breakup dates. The ice breakup output variable is defined as the number of days after the equinox (i.e.: the number of days after the day where the center of the sun is directly above Earth's equator). This variable will be called *DE* in the rest of the report. In this section the linear regression model and the Random Forest regression model made by Van Asselt, 2020 and Terwogt, 2021, will be discussed. This section first provides an overview of the variables used in the previous models. Next, a summary of their work and the models that they made is given. Last, research gaps and other potentially interesting aspects that were not included are discussed.

1.3.1. Variables

This section provides an overview of the variables used by Van Asselt, 2020 and Terwogt, 2021. Most variables were introduced by Van Asselt. Terwogt subsequently added the number of heatwave days per month. A day is considered as a heatwave day when it is part of at least three consecutive days where the temperature is higher than the 95th percentile of the temperature data between 1917 and 1947 (Terwogt, 2021). The Accumulated degree-days thaw (ADDT) and accumulated degree-days

frost (ADDF) are based on the reference point of -5°C and can be defined by the sum of all mean daily temperatures below (ADDF) or above (ADDT) the reference point in a certain period. The ice thickness is measured by the organisation of the Nenana Ice Classic itself and the amount of measurements as well as the dates on which they are done differ per year. Therefore, to create a consistent data set, Van Asselt, 2020 chose to use four reference points (February 20th, March 1st, March 15th and April 1st). The measurement made closest to one of the reference points is assigned to that reference point. This resulted in a dataset with four measurements per year. The variables used, along with the symbol that will be used further on in the report, the units, the period of observation and the source of the data, are presented in table 1.1 below.

Ice breakup aspect	Variable	Symbol	Unit	Period	Source
Thermal ice deformation	Monthly average temperature	T_{month}	°C	1917-2018	BE
	Accumulated degree-days thaw (base = -5°C)	ADDT	°C	1917-2018	BE
	Heat wave days per month	HWd_{month}	d	1947-2018	BE
Solar radiation	Cloud coverage	CC_{month}	%	1917-2018	BE
	Average winter precipitation	P_{DJFM}	mm	1917- 2018	CRU
River discharge	Monthly average discharge	Q_{month}	m^3/s	1963-2018	USGS
Ice thickness	Average winter temperature	T_{DJFM}	°C	1917-2018	BE
	Accumulated degree-days frost (base = -5°C)	ADDF	°C	1917-2018	BE
	Ice thickness	t	”	1989-2018	NIC
Large scale SST data	ENSO effect in February - May	$ENSO_{FMAM}$	°C	1951-2018	NOAA
	PDO effect in February - May	PDO_{FMAM}	°C	1951-2018	NOAA

Table 1.1: Variables used by Terwogt and Van Asselt. Including the symbol that will be used further on in the report, the units, the period of observation and the source of the data. Sources are: BerkeleyEarth, 2021, CRU, 2021, USGS, 2021, “Nenana Ice Classic”, 2022 and NOAA, 2020.

1.3.2. Linear regression model

In this paragraph, the potentially linear relationships between the initially chosen variables and the breakup dates are discussed. To do so, a linear multi regression model, heavily based on Van Asselt’s model, is constructed and the process is briefly explained below.

A linear multi regression model is a strong statistical tool which can be used for determining correlation between a response variable and explanatory variables (Chatterjee et al., 1979). It will result in the following model:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_p * x_p + \epsilon$$

Where:
y = The response variable (ice breakup date)
β = Partial regression coefficients
x = Explanatory variables (climate related variables)
ϵ = Random disturbance

First, for all variables, an ordinary least squares (OLS) regression is held. To determine whether the model is a good fit, the results of the OLS regression need to be examined. To do so, two output values are taken and qualified: The R^2 value is an indication for the proportion of the variance in the dependent variable that is approached by the regression and is calculated by dividing the sum of the regression total by the sum of the total squares total. R^2 values vary between 0 and 1 where 0 means no correlation and 1 means full correlation. The F-statistic value shows how much the fit is statistically meaningful. It is calculated by comparing the variance with and without the input of independent variables. (Dekking et al., 2005). Only the variables that are qualified as significantly linear related with the breakup date (meaning they have a R^2 value of at least 0.1 and a minimum F-statistic of 7), are used.

The scatterplots with fitted line of these variables are presented in appendix A.

The next step in conducting a linear regression model is to test the remaining variables on their normality and multicollinearity. This can be done with conducting a variable matrix. On the diagonal of the matrix (see appendix B), it can be seen that most variables are normally distributed. It can also be noticed that the mean temperature and the accumulated degree days thaw (ADDT) are highly correlated. Therefore, they cannot be used in the same model.

After making multiple models with the remaining variables and testing them for their performance and significance, two final models are constructed which are presented below:

$$DE = 57.52 - 0.045 * ADDT + 0.25 * t_{FEB20}$$

$$DE = 55.24 - 1.32 * T_{Apr} - 1.57 * T_{May}$$

The ADDT model has a R^2 value of 0.84 and an F-statistic value of 65.48. The mean temperature model has a R^2 value of 0.776 and an F-statistic value of 171.9. These results show that, when making a linear model, only very simple models with two variables, produce somewhat good results.

Van Asselt, 2020 made two models: One model to predict the ice breakup dates for the coming 80 years based on IPCC climate scenarios and a model that is time bounded by the betting deadline of April 1st used as a tool to forecast the ice breakup date of that year. His forecasting model predicted an ice breakup date which diverted about twelve days from the actual breakup date of that year. He concluded that the relatively poor performance of his model could be attributed to the fact that the breakup dates are mostly influenced by the variables close to the breakup dates, such as the temperature in April and May and the ice thickness close to the breakup date. He also suggested looking at a non-linear model.

1.3.3. Random forest regression model

In this paragraph the results of the random forest regression model of Terwogt, 2021 are discussed. The variables and data used in this model are the same as in the linear model.

Random forest (RF) regression is a machine learning technique and is introduced by Breiman, 2001. The algorithm uses multiple decision trees or 'tree predictors' that all give an estimate for the output variable of the model. The mean of all tree predictors is then taken to give a better prediction than just a single model. The tree predictors are functions that relate the input to the output variable. These functions are formed by training the model. To do so, the input data is split into two groups: One dataset to train the model and one to test it. The readily available Random forest regression module of the package scikitlearn (Pedregosa et al., 2011) is used to implement and validate the RF regression model.

Terwogt chose to use 75 percent of the data for training and 25 percent for testing. He also made two models: A descriptive model, which wasn't bounded by the betting deadline, and a forecasting model to predict the ice breakup date of that specific year. His descriptive model has a Root Mean Squared Error (RMSE) of 2.8 and a Mean Absolute Error (MAE) of 2.26. The forecasting model showed less performance with a RMSE of 5.98 and an MAE of 5.07. Terwogt concludes that the relatively poor performance of his forecasting model is mostly influenced by the lack of available data and the fact that RF regression models are not well suited for extrapolation cases (Hastie et al., 2009). In the figure below, the relative importance of all variables can be seen.

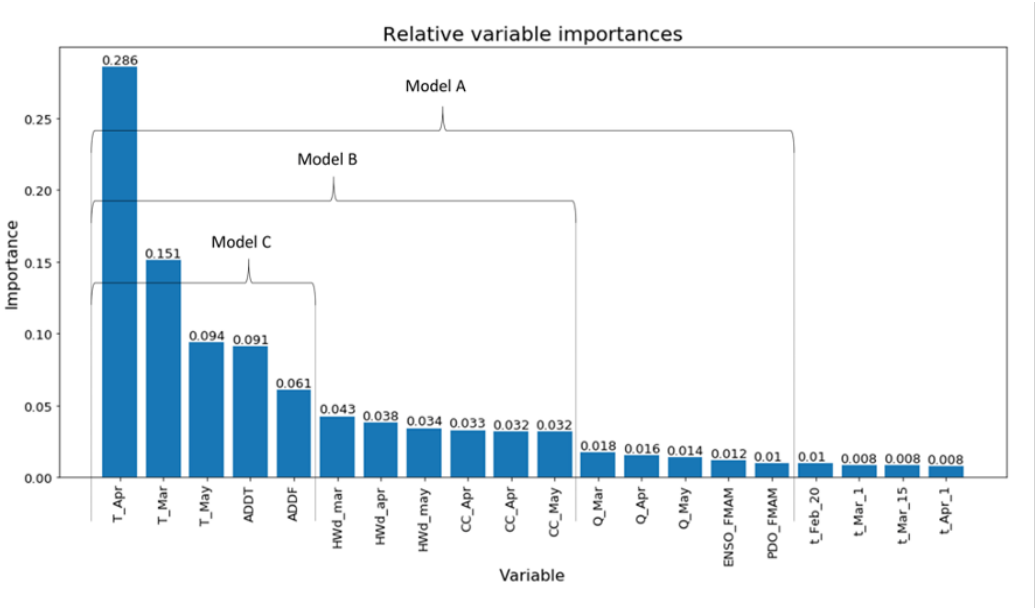


Figure 1.3: Relative importance of all variables for a Random Forest model with 100 tree predictors. In the figure the different models tested by Terwogt are indicated. The final model contained the variables of model B (source: Terwogt, 2021)

1.3.4. Research gaps

The recommendations made by Van Asselt, such as using the river's discharge data and creating a non-linear model are mostly discussed by Terwogt. Terwogt further mentions that his RF regression model can still be significantly improved by gathering more input data, performing a hyper-parameter sensitivity analysis or a further, detailed analysis of the data by using different computations and smaller time intervals. He also suggests looking at the possibilities of using an artificial neural network (ANN) model for the forecasting of the ice breakup.

At last, some, more general, research directions suggested by Van Asselt and Terwogt are the following:

- Exploring relationships between betting behaviour of participants and their view on climate (change).
- Performing a broader analysis on climate change in general and the effect it has on organisms in Alaska.

1.4. Artificial neural networks

In line with the suggestions of Terwogt, 2021, this report will describe the use of an artificial neural network (ANN) in forecasting of the ice breakup. In this section, a brief introduction to artificial neural networks will be given. Next, the choice of using ANN is motivated.

1.4.1. Introduction to artificial neural networks

The fundamental theory behind artificial neural network modeling, a machine learning technique inspired by the way neurons signal to each other in the human brain, was first described by McCulloch and Pitts in 1943 McCulloch and Pitts, 1943. Since then, many researchers have investigated the use of it, but it wasn't until the 1980's when interest in ANN re-emerged due to technical developments which increased the processing capacity.

An artificial neural network consists of a group of processing units called nodes that receive input from other nodes or external sources. This input is then used to compute an output signal which is propagated to other nodes. Each node k has a state of activation y_k and each connection between node k and node j has a weight w_{jk}. ANN models consist of three different types of layers with nodes: One *input* layer, which receives external data from outside the neural network, one or multiple *hidden* layers, whose input and output stay within the neural network, and one *output* layer, which sends the

data out of the neural network. The input of node k (s_k) at time t can be described as the weighted sum of the outputs of the connecting nodes plus a bias term, θ. This input s_k is then transformed by activation function F_k to a new activation state y_k for node k (Kröse et al., 1993). This process is visually described in figure 2.1

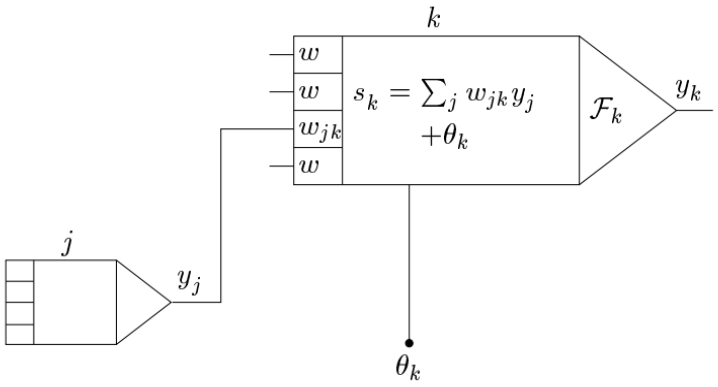


Figure 1.4: Propagation rule with weighted summation for an artificial neural network. The input of node k (s_k) at time t can be described as the weighted sum of the outputs of the connecting nodes plus a bias term, θ. This input s_k is then transformed by activation function F_k to a new activation state y_k for node k (source: Kröse et al., 1993)

When input as well as output data is known, an ANN model can train itself by constantly adjusting the weights between two nodes. In this way, a final model is conducted which can be used to predict the values of output data by providing it input data.

1.4.2. Motivation

It is expected that making use of an ANN will improve the forecasting (bounded to betting deadline) as well as the descriptive (not bounded to betting deadline) model with respect to the random forest regression models by Terwogt, 2021. This is because ANNs have more parameters to tune and it is concluded by Płoński, 2019 that a well tuned neural network model will perform better than a random forest regression model. One downside of using ANN over random forest is the fact that ANN will give less insight in the decision making process. Where computing the exact influences of each variable is possible with RF, it is not possible to do so with ANN. Therefore, the ANN model can be seen as a black box. However, this does not influence the results. One other unfavorable feature of ANN is the amount of data it needs to come up with good results. Yet, this is also the case for RF and moreover, (among others) Zhao et al., 2012 proved that using ANN with a dataset of similar size and variables can still produce good forecasting results.

2

Variables

Apart from using a new modelling technique, it is expected that introducing new variables and critically analysing old ones will increase the performance of the model. This chapter investigates the temperature, discharge and ice thickness data.

2.1. Local and regional temperature

As concluded by Van Asselt, 2020 and Terwogt, 2021, the variables that are most important when making a model to forecast ice breakup are related to the temperature. There are two temperature data sets available: The first one represents the average regional temperature data of a large area around Nenana (area with borders lon = -148.000, lon = -149.000 and lat = 64.000, lat = 65.000). The other data are measurements of a weather station inside Nenana (local temperature data). Because of the high correlation between temperature and ice breakup, it is important to determine what data set is best to use when relating the temperature data to the breakup dates. To do so, first, the two data sets are compared in the figure below:

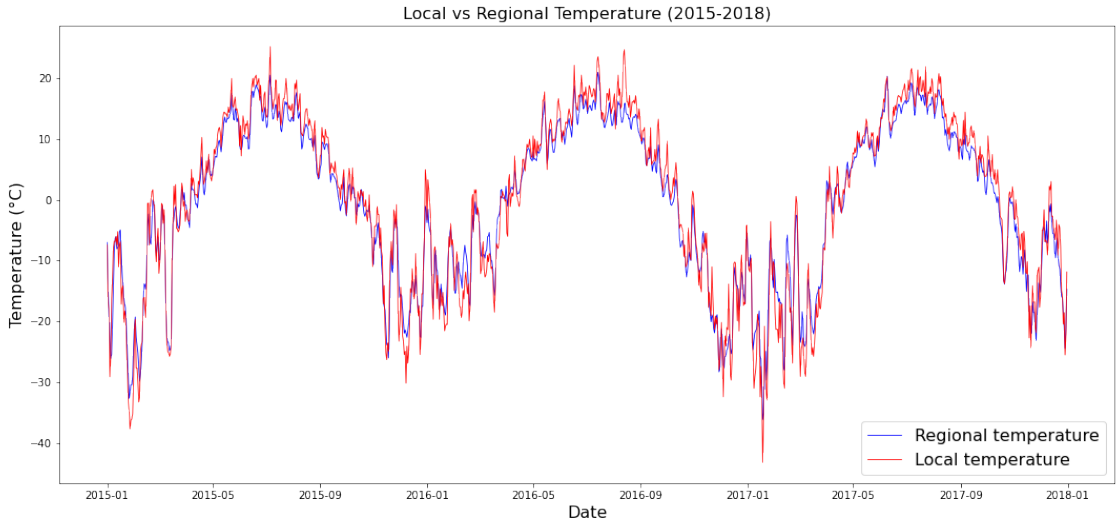
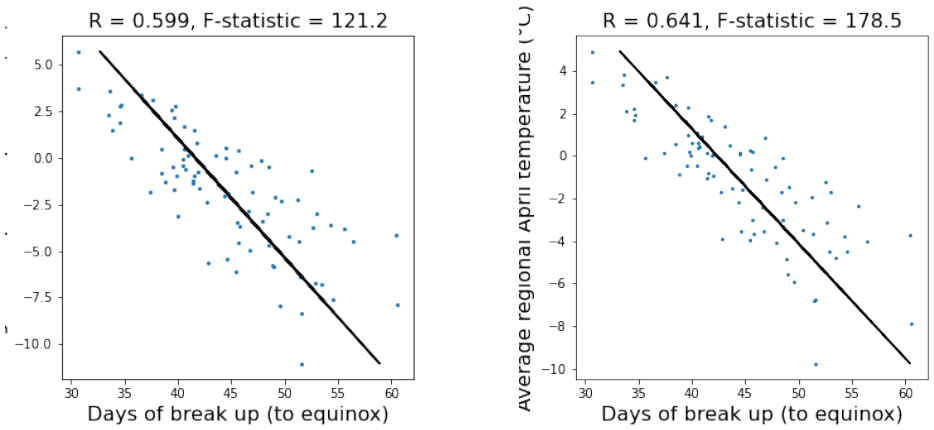


Figure 2.1: Comparison of local and regional temperature data. For readability, only the data between 2015 and 2018 is shown

The regional and the local data show almost the same trend. However, it can be seen that the extremes of the local data are bigger than those of the regional data. This means that in Nenana, the winters are often colder and the summers warmer than the regional average. To see what this means in

relation to the ice breakup dates, an ordinary least squares (OLS) regression analysis is held with both data sets. The two figures below show the relation between the mean, local and regional, temperature in April with the breakup dates.



(a) Local (b) Regional

Figure 2.2: Local and regional April temperature OLS fit. The figures show a higher R^2 value as well as a higher F-statistic value for the regional temperature data.

The figures show that the mean regional temperature in April has a stronger relationship with the ice breakup dates than the local temperature. The same process has been repeated with other months and they showed similar results. This could be attributed to the fact that the regional temperature influences not only the local temperature, but also other driving factors such as discharge and water temperature. These findings are in line with the work of Williams et al., 2004.

2.2. Ice thickness

One of the main things that stands out when comparing the linear- with the random forest regression model is the influence of the ice thickness. In the linear model it is concluded that ice thickness has a significant relationship with the breakup dates. According to the RF model, the ice thicknesses are the variables with the least importance to the ice breakup date. This can be explained by the fact that the individual ice thicknesses are strongly correlated. This is proven with the variable matrix in the appendix. Because of their high correlation they all show the same trend making their relative importance lower than when just one ice thickness is considered in the model. In the linear model it is concluded that the ice thickness close to February 20th is the ice thickness variable with the greatest correlation with the breakup dates. According to Beltaos and Bonsal, 2021, the primary variable of interest related to ice thickness is the maximum measured ice thickness. One other way of interpreting the ice thickness data is to look at the last known (April 1st) ice thickness measurement. Therefore, to test which ice thickness variable is most important in relation to the Nenana Ice Classic data, all three variables are tested for their linear relation with the ice breakup dates. Also, since ice growth and decay is proportional to the square root of ADDF and ADDT (Michel, 1971; Murfitt et al., 2018), the ice thickness variables are also tested for their quadratic relation with the ice breakup dates. The result can be seen in figure 2.4a,b,c.

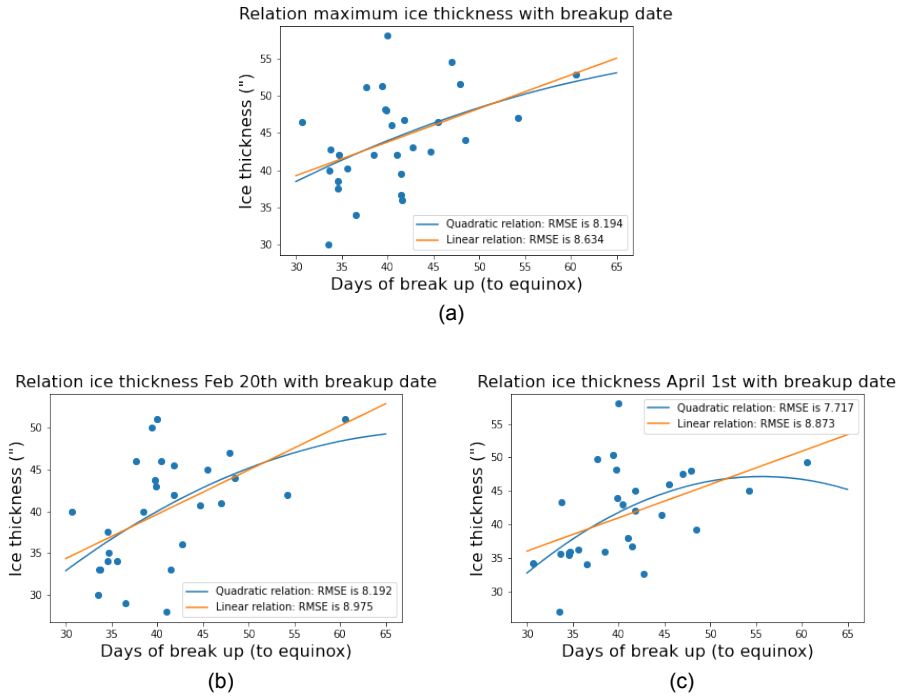


Figure 2.3: Linear (orange line) and quadratic (blue line) fit for maximum ice thickness (a), ice thickness measurement closest to February 20th (b) and ice thickness measurement closest to April 1st (c)

The figures show that the maximum measured ice thickness has the strongest linear relation with the breakup dates and the ice thickness closest to April 1st has the strongest quadratic relation. However, the differences in RMSE are very small so no ice thickness measurement can be pointed out as most important. Since it was expected that the maximum ice thickness would have gained the best results, these findings are not in line with Beltaos and Bonsal, 2021. However, it could be explained by the fact that the ice thickness is not measured every day, and therefore the maximum ice thickness measured might not be the actual maximum ice thickness of that year. Because of the uncertainty of the maximum ice thickness measurement as well as the fact that the ice thickness measurement closest to April 1st has the strongest quadratic relation, the ice thickness measurement closest to April 1st will be used in the remainder of this report.

2.3. River Discharge

In this section, the Tanaka river's discharge is related to the temperature and the breakup dates. The discharge data is obtained from a gauging station 200 meters downstream of the location were the tripod is placed. The figure shows that the discharge and temperature are strongly correlated. This can be seen when looking at the peaks and the initial discharge (meaning the first discharge measurements after the ice breakup) of each year. This is because higher temperatures cause more snow and ice to melt, resulting in a higher discharge (Bieniek et al., 2011). The figure shows that the discharge is only recorded after the breakup date. The reason for this is that from a frozen river no discharge measurements can be made. This trivial observation is, however quite important for the final model. Since the discharge measurements are always made after the ice breakup, the discharge cannot be used in the forecasting model.

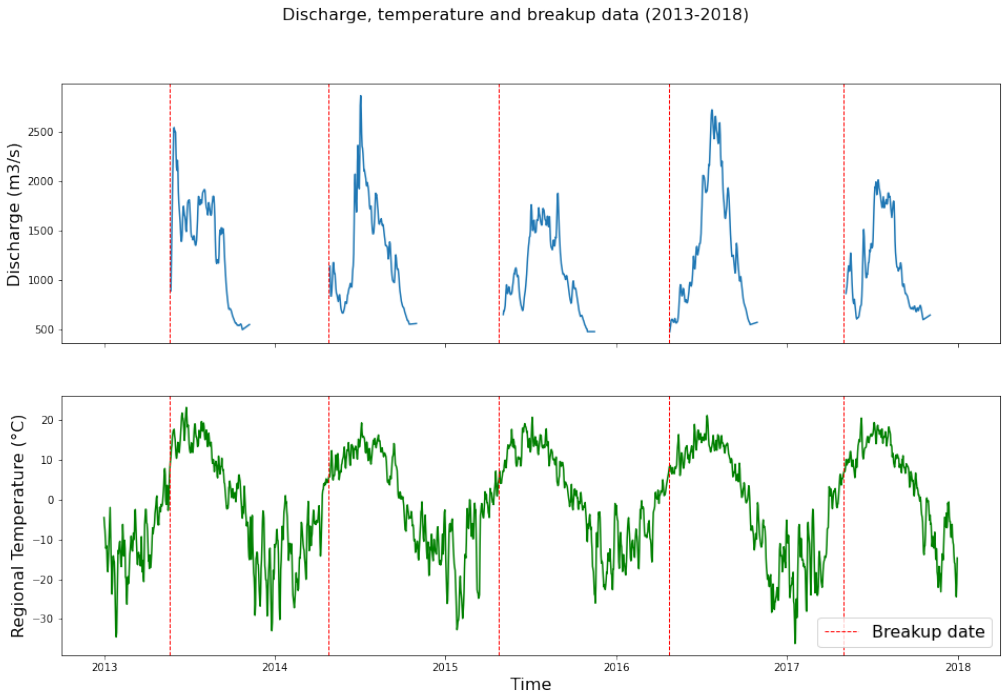


Figure 2.4: Time series of the the regional temperature data (shown in green green) and river discharge (blue). For all variables, the daily mean is used. For readability, only the the data of 2015 till 2018 is shown.

2.4. Thaw onset

This section introduces the thaw onset variable. The thaw onset is defined as the last day before the ice breaks with a minimum temperature of -5°C. It is expected to have a strong linear relationship with the breakup dates because it indicates a starting point of the decay of the ice thickness. In the figure below, the temperature data of 2015-2018 is given. Indicated are the breakup and the thaw onset dates.

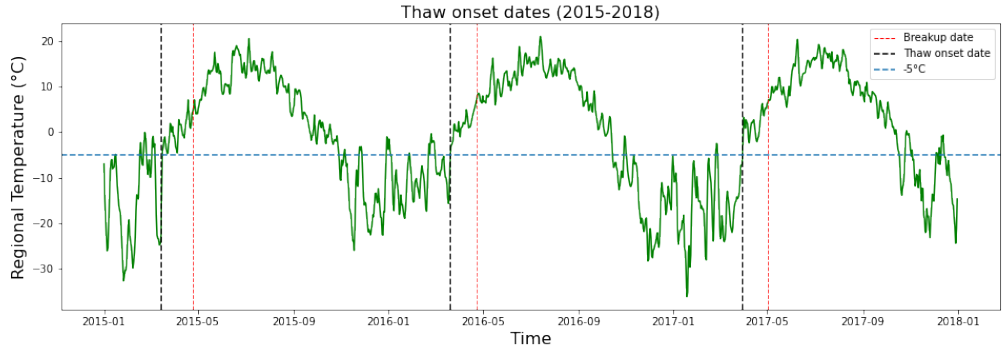


Figure 2.5: Time series of the the regional temperature data (shown in green green), indicated are the breakup and thaw onset dates (red and black respectively). For readability, only the the data of 2013 till 2018 is shown.

From figure 2.5 it can be observed that the distances between the thaw onset date and the breakup dates all have a similar length. Because of this a strong linear relation ship is expected. To test this, an OLS regression is held. The results can be seen in figure 2.6.

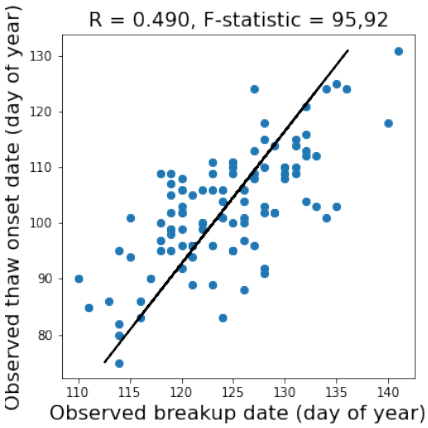


Figure 2.6: Ordinary least squares regression analysis of thaw onset dates.

Figure 2.6 shows a strong linear relation between the thaw onset and breakup date. Therefore, it is expected that using the thaw onset variable will improve the model. However, more than two thirds of the thaw onset dates were after the betting deadline. Therefore the thaw onset data will probably not be used in the forecasting model. To overcome this problem, the thaw onset threshold has been tested for lower temperatures (-10°C, -15°C) as well. These thresholds resulted in more data points that could be used for the forecasting model. However, their relation with the breakup dates was far less significant and therefore, they will not be used.

3

Methodology

In this chapter, the artificial neural network model is applied on the case of the Nenana ice classic. The model's setup as well as the validation and configuration procedures are explained.

3.1. Python package

The mathematical background and process of an ANN is explained in paragraph 1.4.1. This process will not be programmed by hand. Instead, the python package *scikit-learn* (Pedregosa et al., 2011) will be used. From the scikit-learn package, the multi-layered perceptron regressor (MLPRegressor) will be used. An MLP can be trained to approximate any measurable function without making prior assumptions concerning the data distribution (Gardner and Dorling, 1998).

3.2. Model set up

In this section the model's set up is explained. The model is based around the assumption of a relation between the response- and explanatory variables. The response variable is the breakup date and the explanatory variables are the climate related variables introduced in section 1.3.1 and chapter 2. The objective is to find a function f which relates the explanatory variables X to the response variable y:

$$\mathbf{y} = f(\mathbf{X})$$

Where:

- $\mathbf{y}$ is a $[n \times j]$ matrix with n test patterns and j response variables
- $\mathbf{X}$ is a $[n \times k]$ matrix with n test patterns and k explanatory variables
- f is formulated by the trained artificial neural network's structure and weights.

The function f is determined by training the neural network. To do so, the data is split into two groups: Training data and testing data. The training data consists of the explanatory as well as the response variable. By providing both the input ($\mathbf{X}$) and the output ($\mathbf{y}$) of the model, the function f is found by the MLPRegressor. Next, explanatory variables of the test data are used as input in the now known function f. This way, output variables are obtained.

3.3. Validation process

This section provides the validation process to quantify a certain model's performance.

3.3.1. Metrics

When validating the model, the obtained output data from the model has to be compared with the observed values. To quantify the performance of the model the root mean squared error (RMSE) and

the mean absolute error (MAE) are calculated. A low RMSE and MAE indicate a good performing model.

$$RMSE = \sqrt{1/n \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

$$MAE = 1/n \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

Where:

- $\hat{y}$ = The predicted value
- y = The observed value

3.3.2. Cross validation

For each time the model is validated, the data is distributed into train and test data randomly. Since some distributions perform better than others, each time the model is validated, a different RMSE and MAE is obtained. This is problematic since the inconsistency of the validation metrics makes it hard to compare different models. To overcome this problem, for each model, the K-fold cross validation will be applied. K-fold cross validation splits the data into $k = 1, 2, ...K$ equal sized groups. First, for $k = 1$, the model is trained with the $K - 1$ groups and tested with the $k = 1$ group. This process is repeated for $k = 2, 3, ...K$. The validation metrics of all validations are then averaged, resulting in a constant RMSE and MAE that can be compared with other models. In this case, a K value of 5 is chosen.

3.4. Model configurations

In order to determine the best model configurations, different models have to be created and compared with each other. This section discusses the process of finding these configurations.

3.4.1. Data organisation

To use the MLPRegressor function, the data has to be sorted in the right way. The input of the function f is a $[n \times k]$ matrix, meaning that all variables must have n data points. Since certain variables have less data points than others, not all variables can be used in the same model without shortening some of the variables. E.g., When the ice thickness data (data available from 1989 till 2018) and the temperature data (data available from 1917 till 2018) are used in the same model, only the temperature data of 1989 till 2018 can be used. Therefore, the first step in conducting the model is to determine the best performing combination of n and k.

3.4.2. Overfitting

When variables with a relative low importance and weak relation with the ice breakup dates are left inside the model, they can cause the model to overfit. Since it is not possible to see the relative importances of all variables in the ANN model, this process has to be done by trial and error. One by one, variables will be left out, resulting in different models. This way, by looking at the cross validated RMSE and MAE, the model which overfits the least can be obtained.

3.4.3. Hyperparameters grid search

The MLPRegressor module from scikit-learn has several parameters that can be tweaked. These parameters greatly influence the outcome of the model and therefore, choosing the right hyperparameters is crucial. To find the right parameters, a grid search will be held. For this, the scikit-learn package GridsearchCV will be used. GridsearchCV takes an estimator (the MLPRegressor) and the parameters that need to be optimised as input. It returns the values of the parameters which cause the estimator to have the best performance. The parameters that will be investigated are: *Hidden layer size, activation function (F_k)* and *alpha (regularization term)*. Other parameters will either be set on default or another fixed value.

4

Results and discussion

In this chapter, the results and findings of the ANN model described in chapter 3 will be given. First, the findings of the model's configuration process are discussed and the results presented. Next, the results of the two final models are analyzed.

4.1. Model configuration

To construct a good performing model, the model's configurations have to be determined. This will be done by performing the steps described in 3.4. The different models performances will be quantified by the validation process described in section 3.3.

4.1.1. Data

The first step in determining which variables will be used in the final model, four different datasets are created: One dataset with data available from 1917-2018, one with data available from 1951-2018, one with data available from 1963-2018, and one with data available from 1990-2018. These models were chosen in such a way that every explanatory variable has all its data points present in at least one model. All datasets were tested for their cross validated RMSE and MAE. The results can be seen in the table below.

Model	**Variables (k)**	**Data points (n)**	**RMSE**	**MAE**
A (1917-2018)	11	102	6.65	4.05
B (1951-2018)	15	68	2.96	2.45
C (1963-2018)	18	56	13.53	6,66
D (1990-2018)	19	29	6.93	5.86

Table 4.1: First step of configuration process: Determining the cross validated RMSE and MAE of four models with different sizes

Table 4.1 shows a clear winner. The model with datapoints from 1951-2018 (Model B) results in the lowest RMSE and MAE. This model contains the variables: $T_{Mar}, T_{Apr}, T_{May}, ADDT, ADDF, HWd_{Mar}, HWd_{Apr}, HWd_{May}, CC_{Mar}, CC_{Apr}, CC_{May}, ENSO_{FMAM}, PDO_{FMAM}, ThawOnset$ and Pre_{FMA}. Despite the smaller number of datapoints, model B shows significantly better results than model A. This indicates that the variables added in model B have a strong relationship with the breakup dates. These variables are the Heatwave day variables ($HWd_{Mar}, HWd_{Apr}, HWd_{May}$) and $ENSO_{FMAM}$.

Another interesting observation that can be made from table 4.1 is the difference in RMSE and MAE of model C and model D. Model D has almost half the amount of datapoints, yet its RMSE and MAE is significantly lower. The only variable added in model D is the ice thickness measurement closest to April 1st (t_{Apr1}). In section 2.2 it was predicted that the ice thickness variable has a strong relation with the breakup dates. The findings in this section prove this prediction.

4.1.2. Overfitting and relative importance

The next step in determining the variables that will be used in the final model, consists of a trial and error process of removing variables one by one and comparing the different models. This way, overfitting is reduced. After testing 15 different models, it is observed that none of the models result in a lower RMSE and MAE than the model with all variables in it. The metrics of all models can be seen in the table in appendix C. The table also gives an indication for the relative importance of the variables. If a model without a certain variable has a high RMSE or MAE, that certain variable has a higher importance in the model. With this in mind, it is confirmed that the heatwave day variables have a high importance, as stated in section 4.1.1. Also the newly introduced variable $ThawOnset$ has a high relative importance. The results indicate that no variable should be left out of the final model. However, the relative importances also depend on the hyperparameters. After the hyperparameters are tweaked, the trial and error process will be redone.

4.1.3. Hyperparameters

The parameters that were investigated using the grid search were *hidden layer sizes* (sizes tested are: 1,2,3,5,7,9,11,13,15,17,19), *Activation function* (F_k) (options are: *identity, logistic, tanh, relu*) and *alpha (regularization term)* (alphas tested are $5 * 10^{-5}, 1 * 10^{-4}, 2.5 * 10^{-4}, 5 * 10^{-4}$). The results of the grid search indicated that the best performing model has the parameters: $hiddenlayersize = 1$, $F_k = identity$, $alpha = 5 * 10^{-5}$. The parameters *Maximum iterations* (2500) and *solver* (lbfsg) were set on a fixed value. The rest of the parameters were set on default. After redoing the trial and error process with the new parameters, it is found that removing the variables CC_{Apr} and Pre_{FMA} will result in a better performing model. Since the MLPregressor as well as the GridsearchCV modules don't give any insight in the decision making process, the exact reason of these results remains unknown.

4.2. Final model

The steps described in section 4.1 have led to the final model. The final model consists of the variables $T_{Mar}, T_{Apr}, T_{May}, ADDT, ADDF, HWd_{Mar}, HWd_{Apr}, HWd_{May}, CC_{Mar}, CC_{May}, ENSO_{FMAM}, PDO_{FMAM}$ and $ThawOnset$. The model has a RMSE of 2.73 and a MAE of 2.28. The results are visualized in figure 4.1

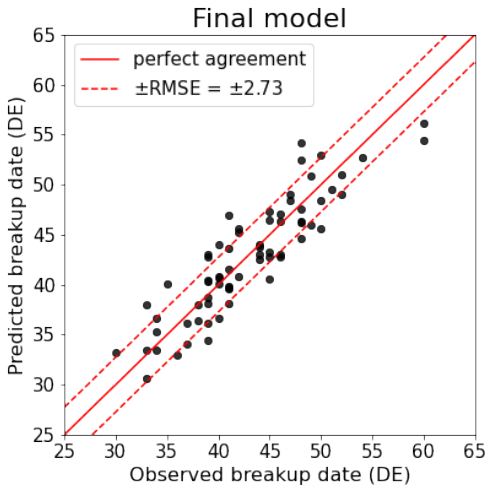


Figure 4.1: Observed and predicted breakup dates of final model. The figure shows a relative good performance, with both a RMSE and MAE of 2-3 days.

The figure shows that the model can relate the climate variables to the breakup dates quite accurate. Most predictions were less than 3 days off of the actual breakup date. The error of the model remains almost constant, meaning that no significant difference in error between very early and average breakup dates can be seen. Relatively late breakup dates are predicted a little bit worse. The results show a slight improvement in respect tot the work of Terwogt, 2021 (RMSE = 2.91, MAE= 2.26). The question

arises whether this improvement is the result of the added variable $ThawOnset$ or the use of ANN instead of RF. To check this, the thaw onset is removed from the model. The results of this model are an RMSE of 2.75 and an MAE of 2.21. This is in line with the hypothesis that making use of ANN instead of RF will improve the model slightly, as stated in section 1.4.2. The reason for the small influence the thaw onset date seems to have, could be the great correlation between the thaw onset date and the accumulated degree days thaw.

4.3. Forecasting model

To obtain a model that is bounded by the betting deadline of April 1st, first, the start variables have to be heavily reduced due to the betting deadline. All the variables which only have data points in April or May ($T_{Apr}, T_{May}, HWd_{Apr}, HWd_{May}, CC_{Apr}$ and CC_{May}) will be removed. The variables $ENSO_{FMAM}, PDO_{FMAM}$, $Pre_{FMA}, ADDT$ and $ADDF$ will be adjusted in such a way that only data from before April 1st is left over. The discharge variable Q as well as the $ThawOnset$ will also be left out of the forecasting model. The reasoning behind this is described in sections 2.3 and 2.4, respectively. Next, the same steps as for the descriptive model have to be taken. From this it is concluded that the best prediction model also uses the data from 1951-2018. The trial and error procedure showed that a model without the PDO_{FM} and Pre_{FM} variables will result in a better model. Also the $ADDT$ will be left out. This is because there are not enough years with thaw days before April 1st, which drastically decreases the amount of data points. Therefore, the final forecasting model consists of the variables: $T_{Mar}, ADDF, HWd_{mar}, CC_{Mar}$ and $ENSO_{FM}$. The gridsearch showed that the model performed best with parameters: $hiddenlayersize = 15$, $F_k = logistic$, $alpha = 2.5 * 10^{-4}$. The RMSE is 5.98 and the MAE is 4.83. The results are shown in figure 4.2

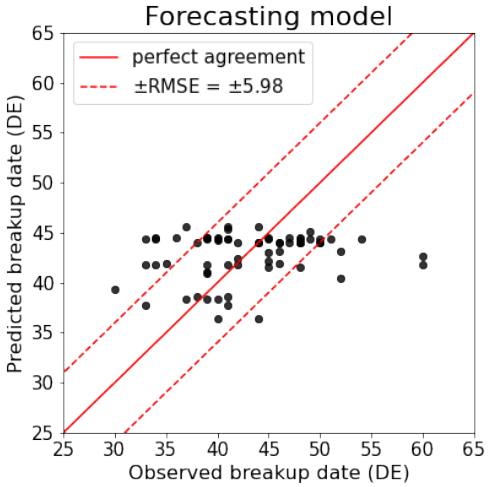


Figure 4.2: Observed and predicted breakup dates of forecasting model. The models shows that it works quite good for breakup dates between 35 and 45 days after equinox, but a bad performance for relatively early or late breakup dates..

The first thing that it noticed when looking at figure 4.2 is the fact that the model performs relatively well for breakup dates between 36 and 46 days after equinox. However, the model predicts rather poorly for breakup dates relatively early or late in the year. The observed breakup dates range between 30 and 60 days after equinox while the predicted breakup dates range between 36 and 46 days. This could be attributed to the fact that in the forecasting model, apart from the variables with data after april 1st, many other very important variables, such as the ice thickness, thaw onset and ADDT were excluded. Because of this, it can be expected that the model had great difficulties finding significant relations between the remaining variables and the breakup dates. Therefore, the model could have trained itself to make predictions around the mean of the observed breakup dates in order to result in the lowest RMSE and MAE. To check whether this assumption is right. A horizontal line indicating the mean observed breakup date (42.96 days after equinox) is drawn.

Figure 4.3 confirms that most predictions were made very close to the mean observed breakup

date. In fact, 51 out of 68 predictions (75%) were made within 2 days of the mean observed breakup date. Although it seems plausible that this is the reason for the model's poor performance, definite prove of this assumption can not be given since an ANN does not give insight in its decision making process.

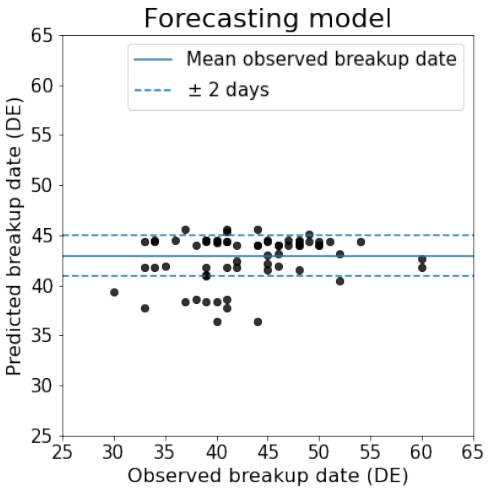


Figure 4.3: Observed and predicted breakup dates of forecasting model. The figure shows that most predictions were made very close tot the mean observed breakup date. This could indicate that the model assigned a heavy weight to the mean observed breakup date and lighter weights to the explanatory variables.

5

Conclusion and recommendations

The main objective of this report was to conduct an artificial neural network model that predicts the ice breakup dates of the Tanaka river based on climate related variables. In order to do so, first, previous research of Van Asselt, 2020 and Terwogt, 2021 was extensively analysed. This analysis concluded that the variables related to ice thickness, discharge and temperature needed to be revisited. The analysis on the ice thickness variables showed that using only one ice thickness variable gives a better representation of its influence than when multiple ice thickness variables are used in the same model. The maximum ice thickness has been pointed out by Beltaos and Bonsal, 2021 as the most important ice thickness variable. However, due to inconsistent measurements, this variable was unreliable to use. The discharge variable is noted as a variable which is strongly correlated with the breakup dates, but since discharge measurements are only made after the ice breakup, it could not be used in the forecasting model. Concerning temperature data, a new variable was introduced. The thaw onset date showed great potential, since it had a relatively high correlation with the breakup dates. These variables, including the ones introduced by Van Asselt, 2020 and Terwogt, 2021 were then used to develop two models. The descriptive model, which was not bounded by the betting deadline of April 1st, showed promising results. A RMSE and MAE of 2.73 and 2.28 days, respectively, were obtained. This is a slight improvement in respect to the random forest regression model made by Terwogt, 2021 (RMSE of 2.91 and MAE of 2.26). It is concluded that making use of ANN instead of random forest had more influence on the results than the use of new variables.

The forecasting model, bounded by the deadline of April 1st showed far less performance than the descriptive model. The RMSE of 5.98 days was exactly equal to the RMSE found by Terwogt, 2021. The MAE of the forecasting model was 4.83 days, a slight improvement of Terwogt's forecasting model (MAE = 5.07). This poor performance is mainly due to the decrease in variables that could be used. Variables that showed high potential such as the ice thickness, ADDT, and the thaw onset date had to be excluded because of the small amount of measurement points. Because of the absence of these high performing variables, the model probably trained itself to give predictions around the mean of the observed breakup dates. However, definite prove of this assumption can not be given since an ANN does not give insight in its decision making process.

The main problem with the forecasting model is that there are not enough relevant input variables. This problem is due to the small and inconsistent sample sizes of some relevant variables. For future research, it would be wise to investigate a way of implementing variables with different sizes into the same model. *Keras* (Chollet, 2015) provides ways of doing this. This way more relevant variables can be added into the same model. Another interesting approach would be to take the models and findings of this and previous research and use it for more serious problems, such as predicting flood and ice jam risks.

Bibliography

Beltaos, S. . (1997). Onset of river ice breakup. *Cold Regions Science and Technology*, *25*(3), 183–196. https://doi.org/10.1016/s0165-232x(96)00011-0

Beltaos, S. . (2003). Threshold between mechanical and thermal breakup of river ice cover. *Cold Regions Science and Technology*, *37*(1), 1–13. https://doi.org/10.1016/s0165-232x(03)00010-7

Beltaos, S. ., & Bonsal, B. . (2021). Climate change impacts on peace river ice thickness and implications to ice-jam flooding of peace-athabasca delta, canada. *Cold Regions Science and Technology*, *186*, 103279. https://doi.org/10.1016/j.coldregions.2021.103279

BerkeleyEarth. (2021). *Berkeley temperature data*. http://berkeleyearth.org/data/

Bieniek, P. A., Bhatt, U. S., Rundquist, L. A., Lindsey, S. D., Zhang, X. ., & Thoman, R. L. (2011). Large-scale climate controls of interior alaska river ice breakup. *Journal of Climate*, *24*(1), 286–297. https://doi.org/10.1175/2010jcli3809.1

Breiman, L. . (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Chatterjee, S., Hadi, & Price. (1979). Regression analysis handbook regression analysis by example samprit chatterjee bertram price. *BioScience*, *29*(8), 484–484. https://doi.org/10.2307/1307545

Chollet, F. (2015). *Keras: The python deep learning api*. https://keras.io/

City of nenana. (2021). https://www.cityofnenana.com/about-1

Collins. (1990). Morphometric analyses of recent channel changes on the tanana river in the vincinity of fairbanks, alaska. *U.S. Army Corps of Engineers*. https://apps.dtic.mil/dtic/tr/fulltext/u2/a229511.pdf

CRU. (2021). *Cru data*. http://www.cru.uea.ac.uk/data/

Dekking, Kraaikamp, Lopuhaä, & Meester. (2005). *A modern introduction to probability and statistics* (1st ed.). Springer.

Gardner, M., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron),äîa review of applications in the atmospheric sciences. *Atmospheric Environment*, *32*(14), 2627–2636. https://doi.org/https://doi.org/10.1016/S1352-2310(97)00447-0

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random forests. *The elements of statistical learning: Data mining, inference, and prediction* (pp. 587–604). Springer New York. https://doi.org/10.1007/978-0-387-84858-7_15

Kottek, M. ., Grieser, J. ., Beck, C. ., Rudolf, B. ., & Rubel, F. . (2006). World map of the köppen-geiger climate classification updated. *Meteorologische Zeitschrift*, *15*(3), 259–263. https://doi.org/10.1127/0941-2948/2006/0130

Kröse, B., Krose, B., van der Smagt, P., & Smagt, P. (1993). An introduction to neural networks.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133.

Nenana ice classic. (2022). https://www.nenanaakiceclassic.com/brochures

NOAA, N. (2020). *Flooding in alaska*.

Papineau. (2016). *Understanding alaska's climate variation*.

Pattison, R. ., Andersen, H. E., Gray, A. ., Schulz, B. ., Smith, R. J., & Jovan, S. . (2018). Forests of the tanana valley state forest and tetlin national wildlife refuge, alaska: Results of the 2014 pilot inventory. *U.S. Department of Agriculture*. https://doi.org/10.2737/pnw-gtr-967

Pedregosa, Varoquaux, Vanderplas, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Passos, Cournapeau, Brucher, Perrot, & Duschesnay. (2011). *Scikit-learn: Machine learning in python.* Journal of Machine Learning Research. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Płoński, P. (2019). Random Forest vs Neural Network (classification, tabular data). https://mljar.com/blog/random-forest-vs-neural-network-classification/

Terwogt. (2021). Nenana ice classic: Gambling with river ice (bsc thesis). *TU Delft*.

USGS. (2021). *Tanana river at nenana: Discharge data*. https://waterdata.usgs.gov/monitoring-%20location/%2015515500/#parameterCode=00060&period=P365D&compare=true

Van Asselt. (2020). Nenana ice classic: Gambling with river ice (bsc thesis). *TU Delft*.

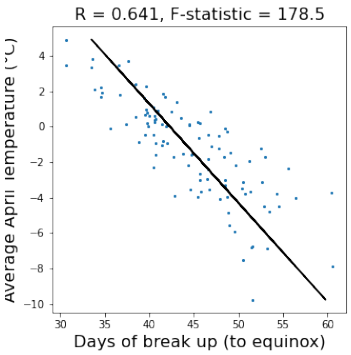
Weatherspark. (2022). *Nenana municipal airport climate, weather by month, average temperature (alaska, united states) - weather spark*. https://weatherspark.com/y/145082/Average-Weather-at-Nenana-Municipal-Airport-Alaska-United-States-Year-Round#Sections-Temperature

Williams, G. ., Layman, K. L., & Stefan, H. G. (2004). Dependence of lake ice covers on climatic, geographic and bathymetric variables. *Cold Regions Science and Technology*, *40*(3), 145–164. https://doi.org/10.1016/j.coldregions.2004.06.010

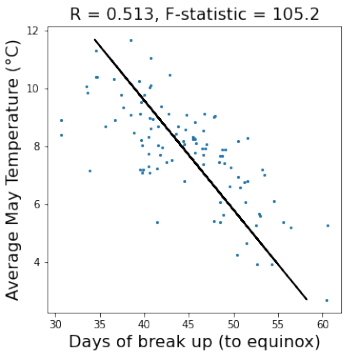
Zhao, L., Hicks, F., & Fayek, A. R. (2012). Applicability of multilayer feed-forward neural networks to model the onset of river breakup. *Cold regions science and technology*, *70*, 32–42.

A

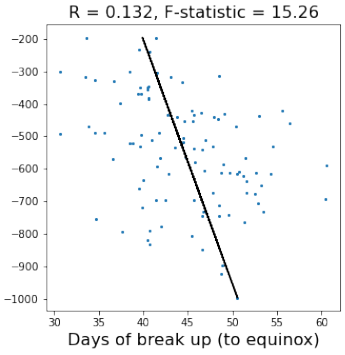
Linear relationships with OLS fit



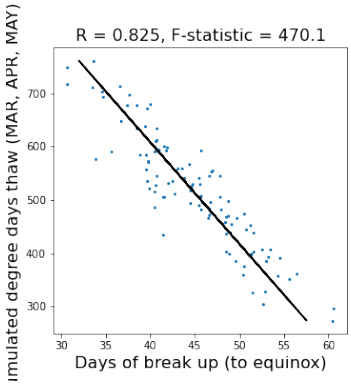
(a) April Temperature



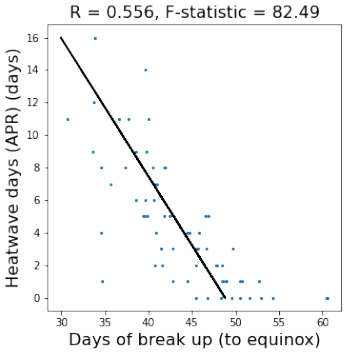
(b) May Temperature



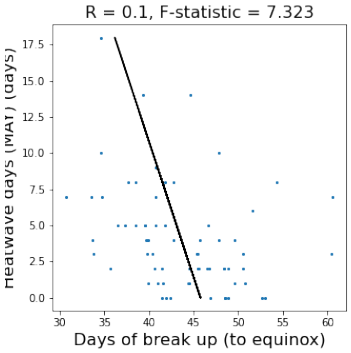
(c) ADDF



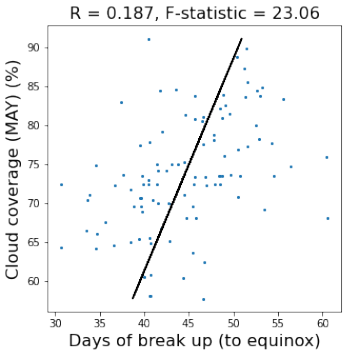
(d) ADDT



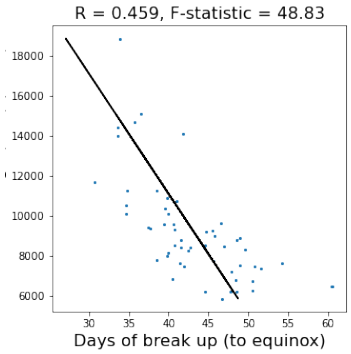
(a) Heatwave days April



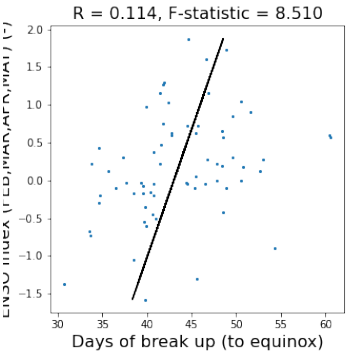
(b) Heatwave days May



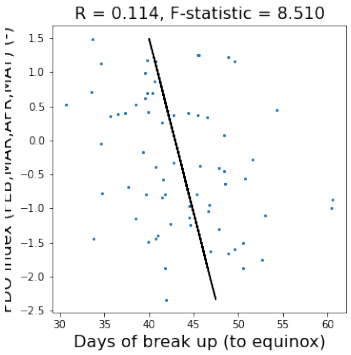
(c) Cloud coverage May



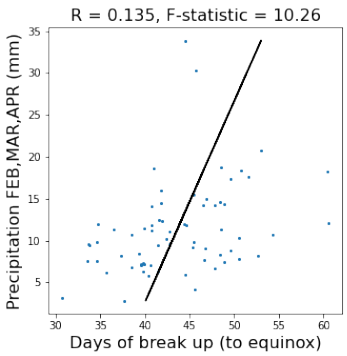
(d) Discharge April



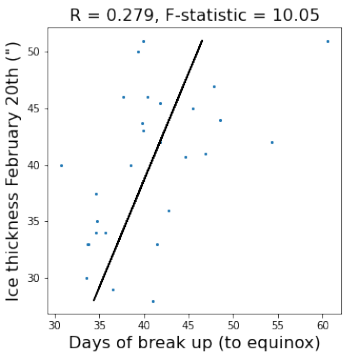
(e) ENSO index



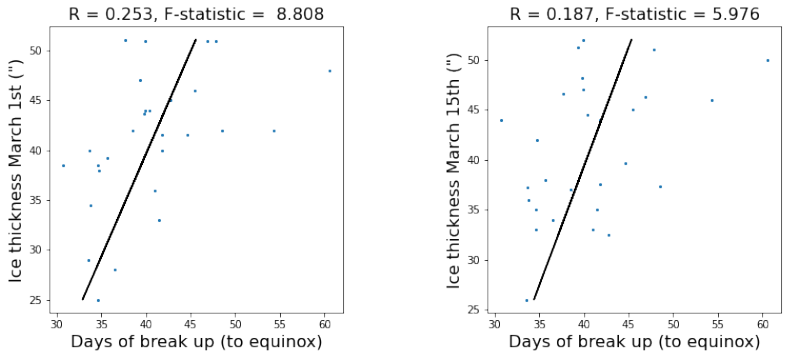
(f) PDO index



(g) Precipitation (Feb, Mar Apr)

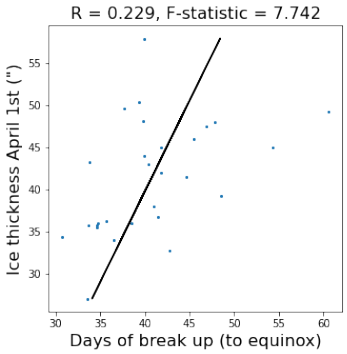


(h) Ice thickness Feb 20th



(a) Ice thickness Mar 1st

(b) Ice thickness Mar 15th



(c) Ice thickness Apr 1st

Figure A.1: OLS fit of different variables with break up dates

B

Variable matrices

Matrices on next pages

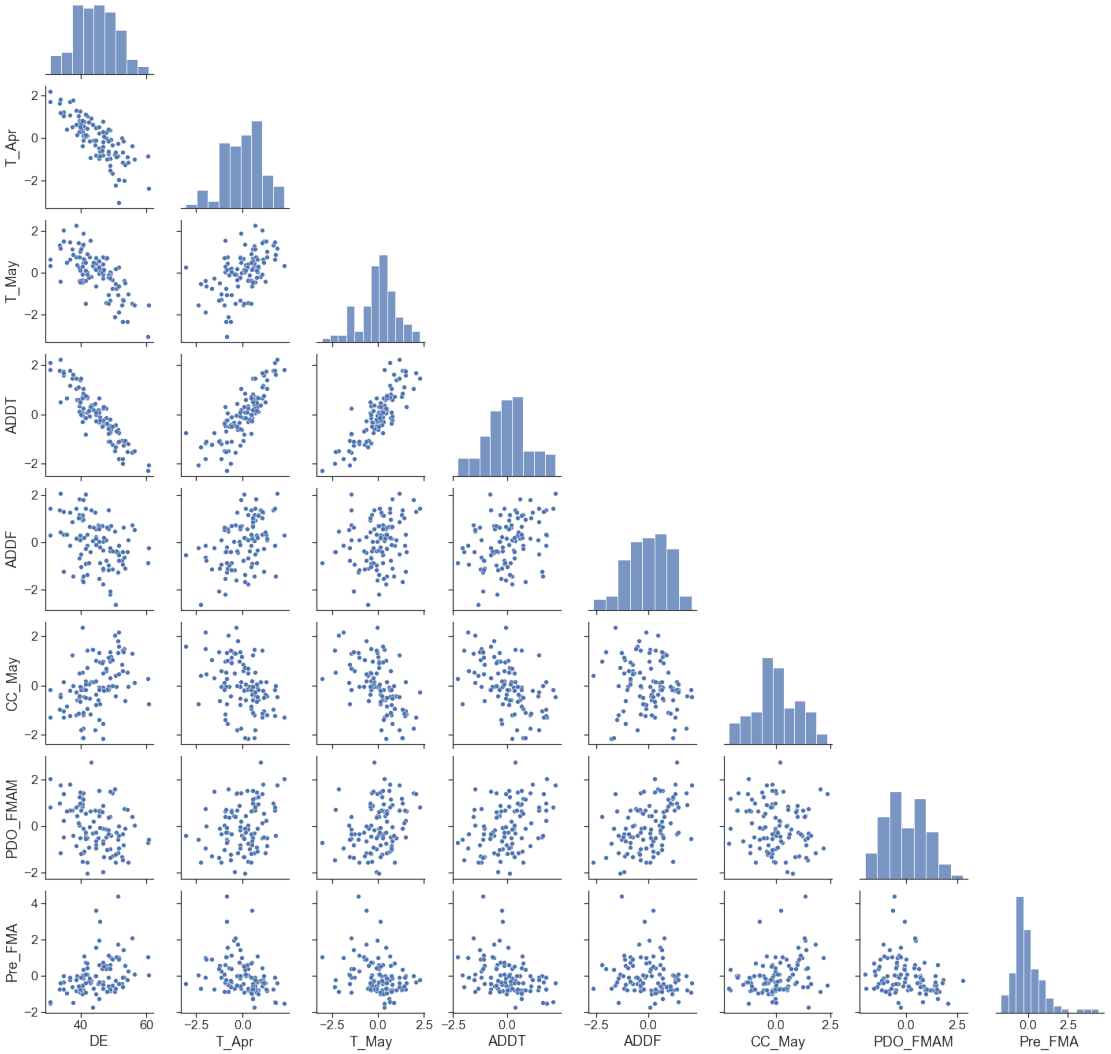

T_Apr
T_May
ADDT
ADDF
CC_May
PDO_FMAM
Pre_FMA
DE
T_Apr
T_May
ADDT
ADDF
CC_May
PDO_FMAM
Pre_FMA

Figure B.1

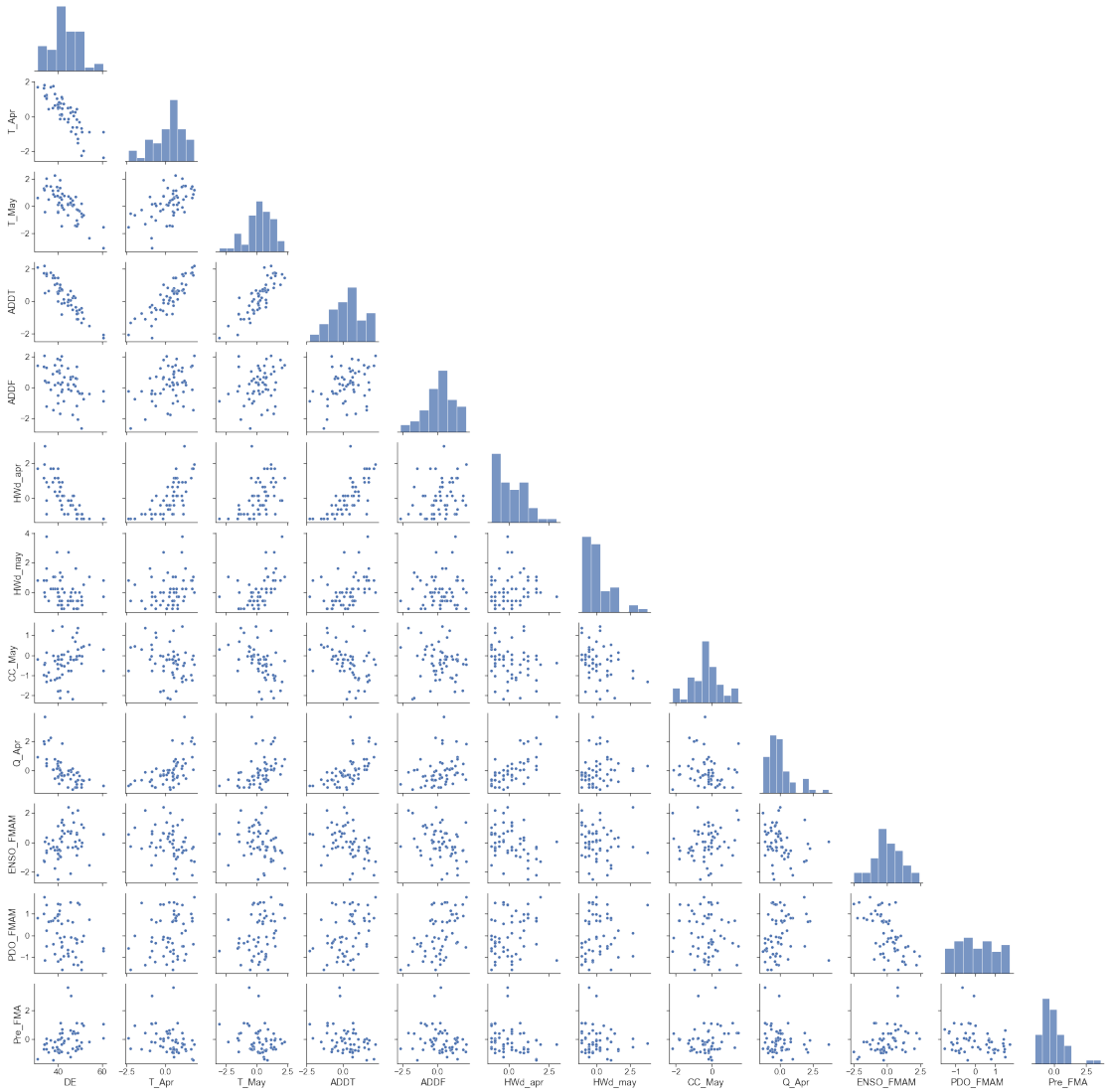

DE
T_Apr
T_May
ADDT
ADDF
HWd_apr
HWd_may
CC_May
Q_Apr
ENSO_FMAM
PDO_FMAM
Pre_FMA

Figure B.2

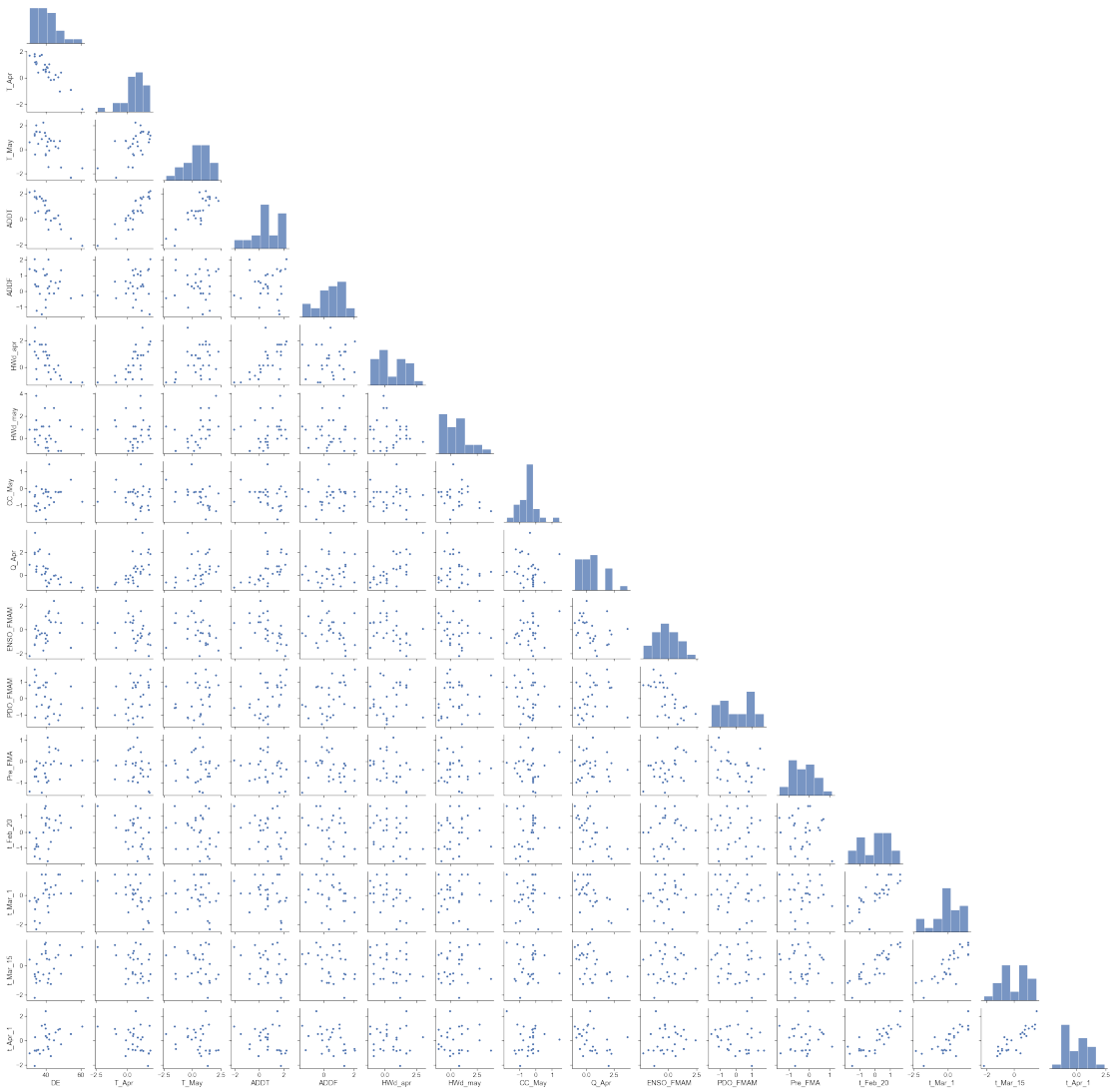

DE
T_Apr
T_May
ADDT
ADDF
HWd_spr
HWd_may
CC_May
Q_Apr
ENSO_FMAM
PDO_FMAM
Pre_FMA
t_Feb_20
t_Mar_1
t_Mar_15
t_Apr_1

Figure B.3

C

Results of trial and error process

Model without	RMSE	MAE
T_{Mar}	3.91	3.21
T_{Apr}	3.83	3.24
T_{May}	3.97	3.22
$ADDT$	3.70	2.95
$ADDF$	3.69	2.87
HWd_{mar}	4.07	3.16
HWd_{apr}	4.09	3.41
HWd_{may}	3.66	2.99
CC_{mar}	3.58	2.99
CC_{apr}	3.88	3.18
CC_{may}	3.74	3.10
$ENSO_{FMAM}$	3.84	3.24
PDO_{FMAM}	3.91	3.05
$ThawOnset$	4.40	3.48
Pre_{FMA}	3.47	2.90

Table C.1