MSc Thesis Meteorology and Air Quality
Wageningen University

Improving the predictive skill of Arctic climate predictions

Daan Gommers
Supervision: Yang Liu & Wilco Hazeleger

MSc Thesis Report
Meteorology and Air Quality

Improving the predictive skill of Arctic climate predictions

Daan Gommers

April 2019

Wageningen University
Supervised by Yang Liu & Wilco Hazeleger

Abstract

Arctic Sea Ice has been decreasing for many years, due to climate change. However, climate models and especially initialized climate predictions, are unable to accurately capture this. Due to errors and biases with the models, the amount of sea ice drops significantly in the first year in the model. This study focused on correcting for this model drift, during post-processing using statistical methods.

We applied two conventional methods, Quantile Mapping (QM) and Ensemble Model Output Statistics (EMOS), and a novel correction method (Recurrent Neural Networks or RNNs) to the first year predictions for anomalies of sea ice extent, of an initialized large ensemble forecast (CESM-DP-LE dataset). Using QM, we were able to reduce the error up to some extent. EMOS resulted in a larger reduction, and a better representation of the distribution, compared with Quantile Mapping. However, there was much to be desired. RNNs, with Long Short Term Memory nodes, were also used. This type of machine learning network is able to use previous steps in predictions for next steps, and therefore ideal for time series analysis. These RNNs resulted in a reduction of the error, but the distribution of the RNN was largely under-dispersive. However, using more or different data, the correction might be improved.

Acknowledgements I would like to thank Stephen Yeager and Gary Strand for providing the CESM-LE-DP dataset, and SURFsara for providing the computational resources for this study. I would also like to express thanks to my supervisors, Yang Liu and Wilco Hazeleger for the fruitful discussions we have had over the past few months. Finally, I would like to thank my girlfriend and the Thesis-ring, for their (nearly) endless stream of comments regarding structure, spelling and grammar.

Contents

1. Introduction

The Arctic sea ice has been slowly melting away for several decades (Walsh and Chapman, 2001) due to climate change. A decrease in Arctic sea ice will have an impact on the rest of the earth (Budikova, 2009), since the Arctic sea ice influences many global processes. For example, sea ice reflects a high fraction of incoming solar radiation and plays an important role in the global ocean currents (When forming, it increases the salt content of surrounding seawater. This seawater densifies and sinks)

Furthermore, in the Arctic region, this poses both threats and opportunities. For example, the Arctic houses a huge amount of natural resources that can be mined with decreasing sea ice (Lindholt, 2006). Shipping also benefits, by using shorter routes between the Atlantic and Pacific oceans through the Arctic region (Khon et al., 2010). However, the retreat of sea ice poses a threat to the infrastructure, health and safety of the indigenous people in the region (Cochran et al., 2014). Wildlife will also be affected by the decrease in sea ice (Larsen et al., 2014): some species might go extinct while others migrate from the south.

Because of the consequences of Arctic sea ice melting, the Arctic has received a lot of interest from researchers. In climate models, sea ice change has been intensively studied. However, when the model results are compared to observations, the model generally overestimates the amount of sea ice (Stroeve et al., 2007). These models do not use observations as a starting point (uninitialized) and follow their own (imperfect) climatology. Later on, models that do use observations as a starting point have been developed (initialized forecasts). We see that these initialized forecasts develop a bias over time: they drift away from the observations, towards the climatology it had, when it was not initialized (see Figure 1.1). Although different strategies to remove this bias have been studied (e.g. Kharin et al., 2012; Krikken et al., 2016), there is still room for improvement.

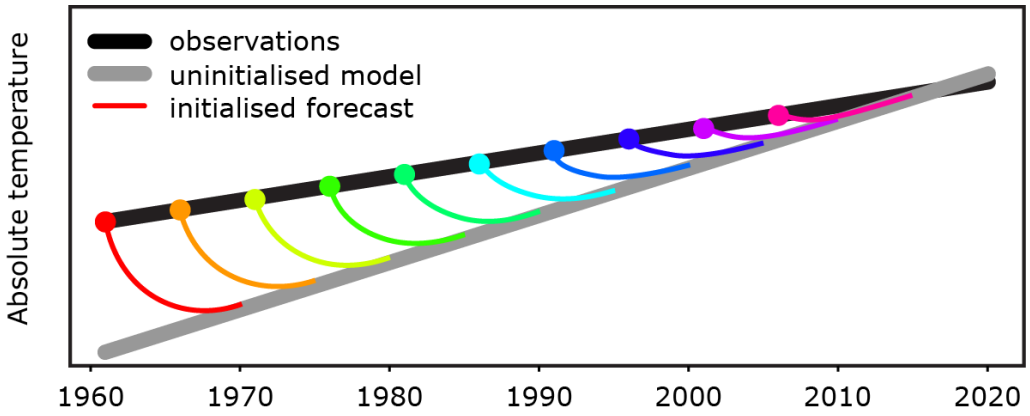

Figure 1.1.: Schematic illustration showing how climate predictions drift away from the initialization to their own long-term climatology (Kharin et al., 2012). The thick black line shows the observations, from which the forecasts (colored thin lines) are initialized. The thick grey line shows the uninitialized model.

Ensemble models have been used to represent the uncertainty in initial fields in both weather and climate models. In an ensemble model, different ensemble members are calculated, each with a randomly perturbated initial state, thus giving other results. The spread of the (entire) ensemble represents then the uncertainty in the forecast. However, due to different model errors, these ensemble members may contain biases. The entire ensemble may also not represent the spread in uncertainty accurately (Wilks and Hamill, 2007).

Therefore, ensemble calibration techniques need to be applied to correct for these biases and to improve the forecast (Wilks and Hamill, 2007; Kharin et al., 2012; Krikken et al., 2016). In weather modelling, these calibration techniques are referred to as Ensemble Model Output Statistics (Ensemble MOS or EMOS). Ensemble calibration techniques generally take either all ensemble members, or the ensemble mean and spread as an input. From those inputs, they calculate a distribution of the variables and adapt the entire ensemble to fit this distribution.

Recently, Yeager et al. (2018) developed a new dataset: CESM-DP-LE. They computed a set of 62 decadal predictions (DP), using the Community Earth System Model (CESM). For each prediction, a 40 member ensemble was run, each with slightly randomly perturbated atmospheric initial conditions. This dataset is thus the output of an initialized forecast, in which the ensemble of all members represents the uncertainty in the initial conditions. It represents the initialized counterpart of the CESM Large Ensemble (CESM-LE; Kay et al., 2015), which has been studied.

1.1. Research Objective

In this study, we want to improve climate predictions for sea ice extent in the Arctic region. Using different ensemble calibration techniques we will analyze the output of a high resolution, fully coupled, initialized, large ensemble climate model: CESM-DP-LE. To obtain this goal, we defined the following research question.

How can the predictive skill of the outcome of a climate prediction, in the Arctic region, be improved?

To answer this question, we will correct the output of the climate model using three different techniques: Quantile Mapping (QM), Ensemble Model Output Statistics (EMOS), and Recurrent Neural Networks (RNN). We will assess the predictive skill of the corrected data, and compare them against each other, and to the uncorrected model output.

1.2. Background

In this section, we will dive further into the Arctic region and the physical processes involved (Section 1.2.1). We will also discuss briefly the history of Ensemble Calibration Techniques, and why we used Quantile Mapping, Ensemble Model Output Statistics and Recurrent Neural Networks (Section 1.2.2). Afterwards, in Chapter 2, we describe how these methods are implemented, which pre-processing steps we take and how we define and measure predictive skill.

1.2.1. The Arctic Region

The Arctic Region is the northernmost part of the Earth (see Figure 1.2). It is home to the Arctic Ocean as well as parts of the American and Eurasian continents. The region consists of the Arctic Ocean, surrounded by land masses. Sea ice covers the ocean through the year, but shows a strong seasonal cycle (Fetterer et al., 2017). In winter, the amount of Arctic sea ice grows, and at its maximum extent (in March) it covers the Arctic ocean completely, as well as parts of surrounding seas (e.g. the Bering Strait, and the Atlantic Ocean) (Fetterer et al., 2017). Since the Arctic Ocean is bordered by land masses, the lateral development of Arctic sea ice is limited by these. In summer, around half of the sea ice melts away, and the sea ice retreats back into the Arctic Ocean (Parkinson et al., 1999). At it's minimum extent (in September) shipping through the Arctic ocean is possible (Khon et al., 2010).

Figure 1.2.: Map showing the Arctic Region, with Arctic Circle (blue, dashed line), and names of the smaller sea's in the region.

Sea ice forms when the surface layer has been cooled to -1.8°C. It first forms as small crystals, which are broken up due to wind and turbulence. When freezing together, these crystals form a thin (up to 10 cm), transparent sheet of young ice (called *nilas*). Over time, these nilas thicken to dark *young ice* (between 10 and 30 cm) and towards *first-year ice* (thicker than 30 cm). When it has survived an ablation (melting) season, it forms *old ice* (World Meteorological Organization, 2015).

The amount of sea ice in the Arctic can be measured in many different ways. Historically, sea ice is measured or estimated in situ: sea ice thickness was measured by drilling ice cores and sea ice concentration by making estimates from ships or aircraft. However, these observations do not provide any information about the spatial distribution of sea ice (Divine and Dick, 2006).

Remote sensing techniques (e.g. satellite observations, submarine sonar, radar) made it possible to measure the sea ice concentration per pixel, or data cell. *Sea ice extent*, which reports the area at which (at least some) sea ice is present, can be calculated from these concentrations (by using a threshold sea ice concentration or fraction). *Sea ice area* data is less abundantly available (Meier et al., 2014). It only measures the area of the sea ice itself. It can be approximated by using the sea ice concentration as weight when summing the areas of the data cells.

Spatial *sea ice thickness* measurements are also available (Meier et al., 2014), since the launch of the ICESat (2003) and the Cryosat-2 (in 2010). In theoretical sea ice models (like PIOMAS), sea ice volume is often calculated (Schweiger et al., 2011). This parameter represents sea ice formation the best, however, it is hard to verify with direct observations.

1.2.2. Ensemble Calibration Techniques

Over the past decades, many different techniques have been developed, to remove biases from model predictions. Especially in weather forecasting, these techniques are often used on an operational basis. However, climate predictions are not often corrected using these techniques.

In this study, we focus on two different types of Ensemble Correction Methods. The first category focuses on correcting the distribution of the variable of interest and thereby minimizes the systematic errors in the prediction of this variable. The other category uses a machine learning approach, in which the correction is calculated based on many other variables, by a computer.

Correction techniques, based on the distribution

Koenker and Bassett (1978) proposed to use Quantile Regression, or Quantile Mapping (QM) to resolve this problem. Quantile mapping is a relatively simple bias calibration method and has been used extensively to correct for the biasses in climate and hydrology forecasts (Verkade et al., 2013). This technique compares the distributions of both the observation and prediction time series. For each quantile, a correction is defined, based on the difference between the distributions (see also Section 2.4.1). Since it approaches the distribution of the observation directly, it does not require that the parameter is distributed normally. It does, however, require a large amount of data. In this study, we try to correct our ensemble using Quantile Mapping, since QM is a direct way to map our predictions towards the observations, and make sure the distributions are equal.

Gneiting et al. (2005) developed a framework, called Ensemble Model Output Statistics (EMOS) or non-homogeneous Gaussian regression. This framework parameterizes Quantile Mapping, by defining a distribution to map towards. It uses a normal distribution (or Gaussian distribution), which is fitted to the observations, by applying linear regression to the ensemble mean and variance (see Section 2.4.2). Therefore, it does not require a large dataset, but is also unable to correct for parameters that are not distributed normally (for example precipitation). Since the introduction of EMOS, it has been implemented in many operational weather forecasting systems and developed itself as the baseline Ensemble Calibration Technique. Therefore EMOS is also used in this study.

Krikken et al. (2016) used two different types of logistic regression to make an ensemble correction for a climate model for the Arctic region. With this type of regression, an S-shaped line is fitted through the cumulative distribution figure (CDF) of the data. Both extended logistic regression (ELR) and heteroscedastic extended logistic regression (HELR) were used in that study, but no clear distinction in skill between the two methods was found. Since this method also tries to parameterize the distribution of the variables of interest, like EMOS, we did not use (H)ELR in our study.

Machine learned Ensemble Correction Techniques

Machine learning is a data analysis method, in which statistical models are built and adjusted by a computer to perform a specific task. These models are created from the data and do not contain any pre-programmed relations about the data. Machine-learned models are able to solve classification problems or analyze non-linear relationships between many variables, but do require huge amounts of training data.

Machine learning algorithms have existed for quite some time (e.g. McCulloch and Pitts, 1943; Quinlan, 1986; Meinshausen, 2006), but computing power was too limited to train these models and outperform regular statistical methods. Recently, these issues have been resolved, and Machine Learning became a feasible method to analyze a wide range of problems, including atmospheric problems (e.g. Taillardat et al., 2016; Rasp and Lerch, 2018).

An Artificial Neural Network (ANN), is such a machine learning technique and is able to model non-linear relationships quite effectively. First proposed by McCulloch and Pitts (1943), these networks exist of multiple layers of interconnected linear regression nodes (see Figure 1.3). Each node calculates the weighted sum of the inputs, passes the result through an activation function. This function maps the result of the weighted sum to another result (for regression problems, a sigmoid or $\tanh$ curve is often used). By adjusting the weights and activation functions of each node, regression problems can be solved.

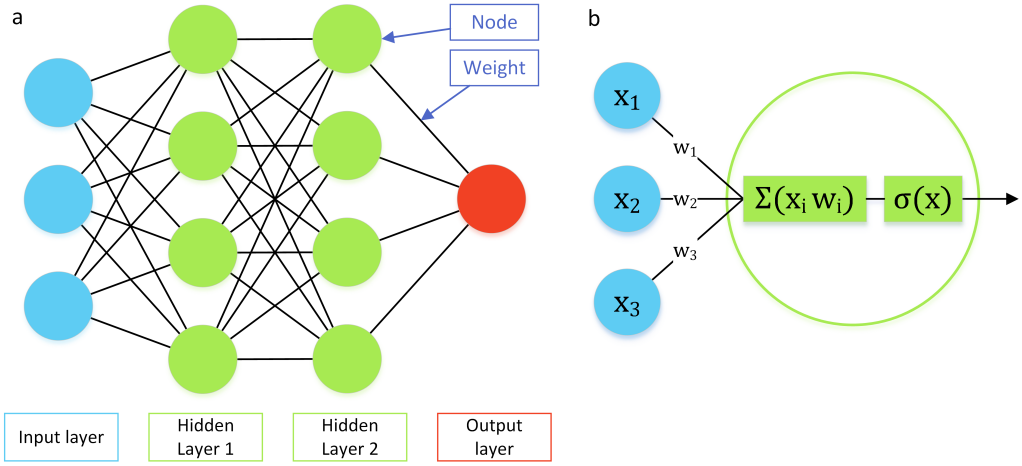

Figure 1.3.: a) Schematic overview of an artificial neural network (here a feed-forward network). Information flows in from the left and passes to 2 layers with 4 hidden nodes. Names of the components and layers are shown. b) Schematic overview of a simple node in a neural network. The inputs x_i are summed, using their weights w_i. The result of the summation is passed to the activation function $\sigma(x)$.

Rasp and Lerch (2018) used an ANN to post-process an ensemble weather model for different stations in Germany. They used a relatively simple network with two layers (one input layer, and one output layer), and a three-layer network (one input layer, one "hidden layer", and one output layer). Their three-layer network out-preformed EMOS for many stations.

In this study we will use Recurrent Neural Networks (RNN). This type of ANN uses nodes in which results and node-weights from a previous (time)step can be used in the next step (see Figure 2.2). These networks are therefore able to store or remember events from previous time steps and use those event to adjust the predicted value in future steps. Regular ANNs do not take any info across (time)steps.

2. Methods

As discussed in Chapter 1, if we aim to improve climate predictions for the Arctic region, we first needed to retrieve our dataset. The dataset was then preprocessed, in order to obtain time series of sea ice extent and average surface temperature at 2 m. These time series were corrected using three different Ensemble Calibration Techniques. Finally, we compared the predictive skill of corrected and uncorrected datasets, to see which technique was performing the best. These steps are visualized in the flowchart below (see Figure 2.1)

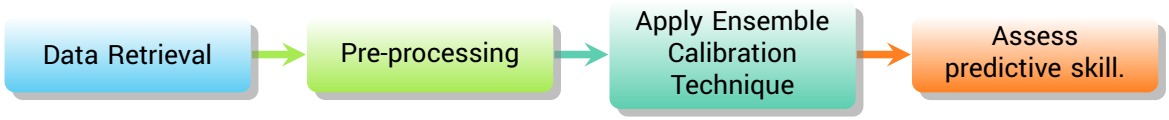

Figure 2.1.: Flowchart illustrating the steps in this project.

In this chapter we will introduce our dataset, preprocessing steps, skill assessment and the different Ensemble Calibration Techniques. The mathematical notation used in this chapter, and the rest of the report, are listed in Appendix A.1.

2.1. Data

As mentioned in the introduction, we used the CESM-DP-LE dataset developed by Yeager et al. (2018). They computed a set of 62 Decadal Predictions (DP), using the Community Earth System Model (CESM). Each prediction contains a Large Ensemble (LE) set of 40 members, with slightly perturbated initialization conditions. The predictions in this dataset started on November 1st of each year between 1954 and 2015, and predicted up to 10 in the future. They had time steps of 1 day, and a horizontal resolution of 1×1°.

We used two different variables: sea ice extent and surface temperature (at 2 m above the surface). To remove some of the larger biases the predictions develop over time, we looked only at the first year of each prediction.

Observations of sea ice extent and surface temperature were used to train the ensemble calibration methods and to validate the corrected forecast. For sea ice extent, we used the NSIDC (National Snow & Ice Data Center) Sea Ice Index (Fetterer et al., 2017). The Sea Ice Index represents the sea ice extent, since 1979 and is calculated from many different instruments (aboard different satellites) over time. Observations from 1979 to 2015 were used for comparison with the results. For surface temperature, NCEP/NCAR 40-year reanalysis was used (Kalnay et al., 1996).

2.2. Preprocessing

To reduce the size and dimensionality of our dataset, we integrated our variables over the entire Arctic region, to create simple time series of our variables. This still allowed us to validate our Ensemble Correction Methods. Sea ice extent is often defined as the area with a sea ice fraction higher than 15% (Vaughan et al., 2013; Sigmond et al., 2013, e.g.). Therefore, we summed the area of all cells with a sea ice fraction higher than 15 %. The same approach is used for the observations we use, made by Fetterer et al. (2017). For the average temperature, we used the average of all cells above 80°N, weighted by cell area. The Danish Meteorological

Institute (2019) uses this definition in their reports on arctic weather.

In our study, we were not interested in forecasting the long-term climatological trend or seasonal cycle. Krikken et al. (2016) showed that there is also a large monthly component in the long-term climatological trend (see also Figure 3.4b & e). To make sure our ensemble correction techniques do not fit towards these, we needed to remove these components. Therefore, we first averaged our data to monthly data and afterwards removed the long-term climatological trend, for each month separately (thereby we also removed the seasonal cycle). The long-term climatological trend was approached linearly, to keep as much variability between members as possible. We did this with the following formula:

$$E' = E - (a * year + b) \tag{2.1}$$

in which E is the sea ice extent, E' are the anomalies of the sea ice extent, $year \in [1954, 2015]$ is the year and a and b are the linear regression model parameters for slope and intercept (respectively). The parameters a and b were defined for each of the 12 months, separately. By using the equation above, we obtained monthly-averaged sea ice extent anomalies. We used the same approach for the surface temperature (T), and the observations for sea ice extent and surface temperature. Since observations for sea ice extent are not available between 1954 and 1979, the parameters (a and b) were defined using the data between 1980 and 2015.

2.3. Skill assessment

To assess the skill of the ensemble calibration methods, we used three different measures. The root mean square error (RMSE) was used to assess the absolute error of the ensemble mean. The continuous ranked probability score (CRPS; Hersbach, 2000) was used to compare the distribution of the forecast to the observations. The CRPS is defined as

$$CRPS = \int_{-\infty}^{\infty} \Big(F_Y(y) - \mathbb{1}(y \leq x) \Big)^2 \mathrm{d}y \tag{2.2}$$

where F is the cumulative distribution function (CDF), x and y denote the observed and forecasted variables respectively, and $\mathbb{1}$ is a Heaviside step function which results to 1 if $y \leq x$, and to 0 otherwise. Both the RMSE and CRPS values have the same unit as the variable of interest and both are negatively oriented (a lower score indicates a higher skill).

To compare the different ensemble calibration techniques, we used the continuous ranked probability skill score (CRPSS), which is defined as

$$CRPSS = 1 - \frac{CRPS}{CRPS_{ref}} \tag{2.3}$$

where $CRPS$ is the CRPS of a forecast, and $CRPS_{ref}$ is the CRPS of a reference forecast (uncorrected ensemble). The CRPSS values range between $-\infty$ and 1. Above 0, they denote that the forecast performs better than the reference forecast. We used the `R`-package `easyVerification` (R Core Team, 2018; MeteoSwiss, 2017) to calculate the CRPS, CRPSS and RMSE.

Since our ensemble calibration techniques were fitted to correct the forecast and bring them closer to the observations, they are prone to overfitting [1]. Therefore, we divided the dataset

[1] Process in which a model or technique is fitted almost exactly to a particular set of data and fails to generalise its results to other sets of data. An overfitted model or technique will show no per formance with any dataset, other than the dataset used to train it.

into two parts: a training and a validation period. Predictions that started in 1954 until 2002, and the corresponding observations, have been used to train the ensemble calibration techniques (e.g. parameter fitting). These trained techniques were then applied to the predictions starting in 2003 until 2015 for validation. The RMSE, CPRS and CRPSS were calculated from the performance of the trained techniques in this validation period. In that way, we were able to exclude overfitting and assess the ensemble calibration techniques fairly.

2.4. Ensemble Calibration Techniques

Using three different correction methods, we tried to improve the ensemble of the climate predictions. In this section we will discus the workings of Quantile Mapping, Ensemble Model Output Statistics, and Recurrent Neural Networks.

2.4.1. Quantile Mapping

Quantile Mapping (QM), or Quantile regression, is a method to align the distribution of the forecast to the distribution of an observation series, assuming that a perfect forecast has the same distribution as the observations. To implement this, cumulative distribution figures (CDFs; denoted F in our equations) were created for the observations and forecasts. A CDF is defined as

$$F_A(\alpha) = \mathrm{P}(A \leq \alpha) \tag{2.4}$$

For every value α in a series A, the CDF F_A results in the probability that a random value in A will be less than or equal to α. The result is called a quantile (θ) and ranges between 0 and 1.

$$\theta = F_X(x) \tag{2.5}$$

With Quantile Mapping, we use the forecast CDF (F_X) to calculate the quantile (θ) of a forecast (x). This quantile is then put into the reverse CDF of the observation (F_Y^{-1}), to obtain a corrected forecast (x_{corr}).

$$x_{corr} = F_Y^{-1}(\theta) \tag{2.6}$$

Since sufficient data points were available to create CDFs empiricaly, we used a simple empirical approach to quantile mapping in this study.

2.4.2. Ensemble Model Output Statistics

Ensemble Model Output Statistics (EMOS), or non-homogeneous Gaussian regression, is a regression technique, in which the distribution of the observations (Y), is adjusted. This was done by modelling the distribution as a normal distribution ($\mathcal{N}$), based the forecast ensemble (X).

$$Y \mid X \sim \mathcal{N}(\mu, \sigma^2) \tag{2.7}$$

In the equation above, $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with a mean μ and variance σ^2. Within the EMOS framework, the mean was modelled linearly from the ensemble mean ($\overline{X}$). The variance was also modelled linearly from the ensemble variance (s_X^2) as shown below.

$$\begin{aligned} \mu &= a + b \cdot \overline{X} \\ \sigma^2 &= c + d \cdot s_X^2 \\ Y|X &\sim \mathcal{N}(a + b \cdot \overline{X}, c + d \cdot s_X^2) \end{aligned} \tag{2.8}$$

The regression parameters a, b, c, and d were adjusted to minimize the CRPS score on the training dataset. Validation was done with the same regression parameters, on the validation

dataset (as discussed in Section 2.3).

When applying EMOS, one often uses all ensemble members in the equation for μ, such that,

$$\mu = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_n \cdot X_n \tag{2.9}$$

in which $X_1..X_n$ represent the ensemble members, and $b_1..b_n$ represent their regression parameters. However, this requires the ensemble members to have individually distinguishable characteristics (Gneiting et al., 2005). Here, our members came from small perturbations in the initial conditions in the model and therefore did not have individually distinguishable characteristics.

2.4.3. Recurrent Neural Networks

As described in the Introduction, we used Recurrent Neural Networks (RNN), a type of neural network which is able to use results and node-weights from a previous (time)step, to improve the performance in the next time step. In our study, we used `Tensorflow` (Abadi et al., 2016) and `Keras` (Chollet, 2015) to implement our RNN.

In this section, we will first describe how the model is designed, which components it uses and how it was trained. Afterwards, we describe three different experiments.

RNN design

Our RNN model consisted of one to four hidden layers with Long Short Term Memory (LSTM) nodes. Each layer contained 16, 32, 64, 125, or 256 nodes. We assessed the performance of different setups, to see which setup gives us the best results. A list of all different setups can be found in Appendix A.3.

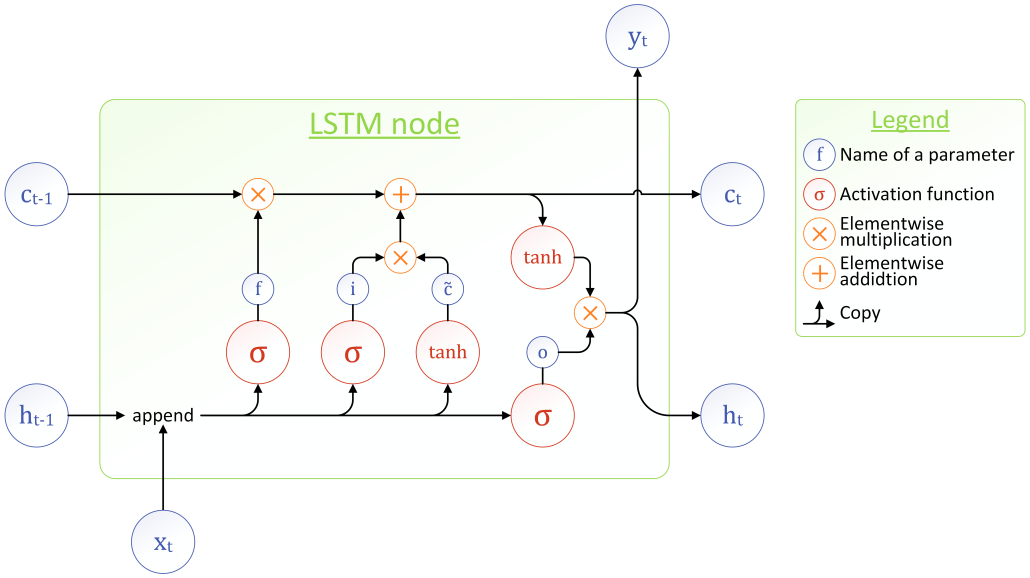

Figure 2.2.: A schematic overview of an LSTM node. x_t and y_t denote the input and output vectors of this node. The cell and hidden states are denoted by c_t and h_t (or c_{t-1} and h_{t-1} when they are calculated in a previous step). The gates f, i, and o use a sigmoid function (σ). A $\tanh$-activation function is used to calculate the candidate cell state ($\tilde{c}$) and the result of this node (h_t and y_t).

Long Short Term Memory (LSTM) nodes are able to use events, that happened, a few or many, time steps ago, in the calculation of the current step. Developed by Hochreiter and Schmid-

huber (1997), each LSTM node passes their result (called the hidden state h), and a cell state (c) along to the next step (see Figure 2.2). Three gates (f, i, and o) determine which components are included in these states. The forget gate (f) regulates which components of the cell state are forgotten. The input gate (i) determines which parts of the candidate cell state ($\tilde{c}$), are passed along to the cell state. Finally, the output gate (o) regulates which components of the hidden state and the input are used to compute the final result. The gates (f, i, and o) and candidate cell state ($\tilde{c}$) all use their own set of weights ($\boldsymbol{w_f}$, $\boldsymbol{w_i}$, $\boldsymbol{w_o}$, and $\boldsymbol{w_{\tilde{c}}}$) and biases ($b_f$, b_i, b_o, and $b_{\tilde{c}}$) in the activation functions. Equation 2.10 shows how these parameters are calculated for a certain (time)step t. Note that the weights ($\boldsymbol{w}$) and biasses (b) are equal for all time steps.

$$
\begin{aligned}
\boldsymbol{z_t} &= \begin{pmatrix} \boldsymbol{h_{t-1}} \\ \boldsymbol{x_t} \end{pmatrix} \\
\boldsymbol{f_t} &= \sigma(\boldsymbol{w_f} \cdot \boldsymbol{z_t} + b_f) \\
\boldsymbol{i_t} &= \sigma(\boldsymbol{w_i} \cdot \boldsymbol{z_t} + b_i) \\
\boldsymbol{o_t} &= \sigma(\boldsymbol{w_o} \cdot \boldsymbol{z_t} + b_o)
\end{aligned}
\qquad\qquad
\begin{aligned}
\boldsymbol{\tilde{c}_t} &= \tanh(\boldsymbol{w_{\tilde{c}}} \cdot \boldsymbol{z_t} + b_{\tilde{c}}) \\
\boldsymbol{c_t} &= \boldsymbol{f_t} \odot \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \odot \boldsymbol{\tilde{c}_t} \\
\boldsymbol{y_t} = \boldsymbol{h_t} &= \boldsymbol{o_t} \odot \tanh(\boldsymbol{c_t})
\end{aligned}
\tag{2.10}
$$

Training process

The model was trained in chunks, or batches of a certain number of time steps. In each training step one batch was passed through the model and the parameters (weights, biasses and activation functions) were adjusted to minimize the loss function, using the optimizer. After an *epoch* (in our study, 100 training batches), the loss function was calculated over the validation dataset. If the network did not improve the validation loss over five epochs, the training was terminated.

As loss function (L), many different variables can be used (e.g. Root Mean Square Error or Mean Absolute Error), as long as they have a negative orientation (a lower loss indicates a higher predictive value). By changing the loss function, one can train the neural network to perform better at certain tasks. We used the following loss function:

$$
L \equiv 1 - \mathrm{R}^2 = \frac{\sum (Y - x_{corr})^2}{\sum (Y - \overline{Y})^2} \tag{2.11}
$$

This loss function motivated the neural network to approach the distribution of the observations as closely as possible. The loss function was not calculated over all time steps within a batch, since the hidden and cell states of the RNN require some initialization. Therefore we defined a warm-up period of 12-time steps, which were ignored when calculating the loss.

The loss function was minimized by an optimizer. An optimizer calculates the gradient of the loss, as a function of the parameters (weights, biasses and activation functions). Calculating the gradient over all samples in the dataset is, however, computationally challenging. Therefore, one often uses Stochastic Gradient Descent (SGD). SGD is a method to step down the gradient of a loss function, in small steps. It calculates the gradient over a small subset of the dataset, assuming that this represents the gradient over the entire dataset. By changing a *learning rate* (η) one can adjust the sensitivity of the optimizer to the gradient.

In this study, Root Mean Square Propagation (RMSprop; Hinton et al., 2012) was used as an optimizer. RMSprop is an advanced form of SGD, which uses a moving mean square gradient

(v), calculated over the last 10 batches (s). The parameters of the neural network (p; weights and biases) were then adjusted using the square root of v:

$$\begin{aligned} \boldsymbol{v_s} &\equiv 0.9 \cdot \boldsymbol{v_{s-1}} + 0.1 \cdot (\boldsymbol{\nabla} L)^2 \\ \boldsymbol{p_s} &= \boldsymbol{p_{s-1}} - \frac{\eta}{\sqrt{\boldsymbol{v_s}}} \boldsymbol{\nabla} L \end{aligned} \tag{2.12}$$

By using a moving mean, RMSprop is quite resilient against rapid changes in the gradient, and it ensures that the descent steps for a certain batch are (approximately) in the same direction as the descent steps of the last batches.

By tweaking our hyper-parameters (e.g. the number of epochs, batches per epoch, data-points per batch, and ignored (time)steps when calculating the loss function, and learning rate), we selected the best preforming model. The resulting values of hyper-parameters can be found in Table 2.1.

Experiments

Table 2.1.: RNN settings and hyper-parameters

Setting	Value
Node type	LSTM
Layers	1, 2, 3, or 4
Nodes per layer	16, 32, 64, 128, or 256
Training Epochs	20
Training steps per epoch	100
Data points per step	240
Loss function	$1 - R^2$
Warm-up period	12
Optimizer	RMSprop
Learning rate	10^{-3} to 10^{-5}

With our RNN setup, we ran three different experiments. In this section, we will discuss their overlapping aspects, and their induvidual setups, where they are different.

For our three experiments, we used both sea ice extent and surface temperature data. However, our activation functions (sigmoid and $\tanh$) required that the input variables lay between 0 and 1. Therefore we normalized these, using these equations:

$$\hat{E}' = \frac{E' + 4.35 \times 10^6}{6.70 \times 10^6} \qquad \hat{T}' = \frac{T' + 11}{22} \tag{2.13}$$

where E' and T' denote the climatological anomalies of sea ice extend and surface temperature (resp.), and E'' and T'' represent their normalized versions. 0 and 1 are also the limits of the output range, but we did not want to limit the RNN to the minimum and maximum of the input. Therefore, we picked the values in the equations in such a way, that we allow small margins around the minima and maxima of E' and T' to still fall within the $[0, 1]$-range.

In contrast to Quantile mapping and Ensemble Model Output Statistics, we did not use the observations of sea ice extent, to validate our RNN. To train a Neural Network a large dataset is needed. Since the observations were not available for the first 25 years of the CESM-LE-

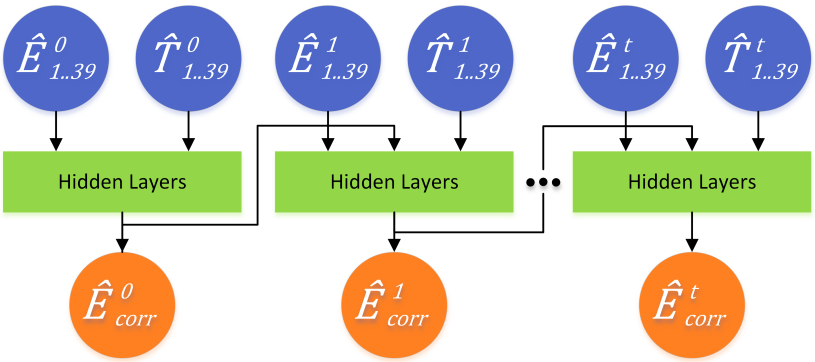

Figure 2.3.: Representation of an RNN with recurrence over time.

DP dataset (1954-1978), they limit the amount of data available. Therefore, instead of using observations, we picked one of the ensemble members which will act as pseudo-observation, per batch.

Predicting seasonal cycle As a first experiment, we tried to predict the full seasonal cycle of sea ice extend, to see how well our model was able to learn from these (regular) signals. We did not use $\hat{E}'$ and $\hat{T}'$ in our RNN, but instead we used $\hat{E}$ and $\hat{T}$ (without applying Equation 2.1). All members (except the pseudo-observation member) were used as input in the RNN model. For the setup, see Figure 2.3

Anomaly correction with recurrence over Time Our second experiment focussed on correcting the anomalies (as was done, with Quantile Mapping and Ensemble Model Output statistics). Again, all members (except the pseudo-observation member) were used as input in the RNN model, together with some information about the time (month and year). with the recurrency steps over time (see Figure 2.3).

Anomaly correction with recurrence over Members Our third and last RNN experiment focused also on correcting the anomalies. However in this experiment, we did use the different ensemble members as an input, but used a year of time steps as input. At each following step in the RNN, we provided it with the following ensemble member. Thereby not recurring over time, but over the ensemble members. The output was also defined as 12 timesteps of (normalized) sea ice anomalies (see Figure 2.4).

We only wanted to assess the RNN on it's final output (after analyzing all 39 members). Therefore we used 39 data points per batch (instead of 240), and a warm-up period for the loss function of 38 (instead of 12).

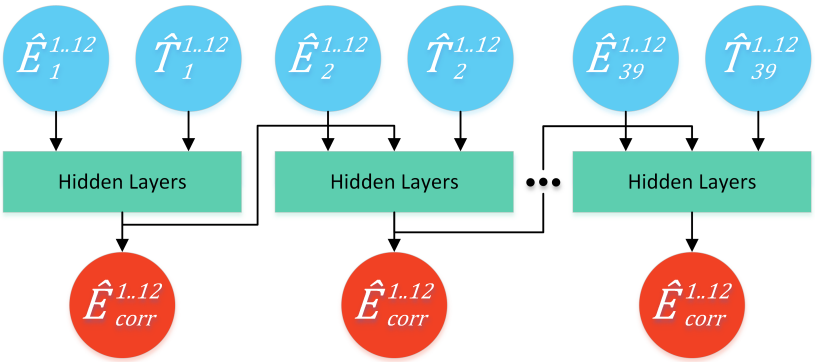

Figure 2.4.: Representation of a RNN with recurrence over members.

3. Results

In this study, we seek to improve the predictive skill of the outcome of a climate prediction in the Arctic Region, using different Ensemble Calibration Techniques.

In this chapter, we will first explore the data we have (Section 3.1), and show the results of the preprocessing steps taken (Section 3.2). We assess its quality before any correction is made (Section 3.3). Afterwards we apply our three different Ensemble Calibration Techniques to improve this data. In Sections 3.4 to 3.6 we will discuss how Quantile Mapping, Ensemble Model Output Statistics and Recurrent Neural Networks improved our forecast. Finally, we will compare different methods with each other, in Section 3.7.

3.1. Data Exploration

The data of the CESM-DP-LE model (Yeager et al., 2018), predicts sea ice concentration spatially. Since we calculate sea ice extent, from the sea ice concentration, we will first take a look at the spatial distribution of sea ice concentration and surface temperature in the Arctic region.

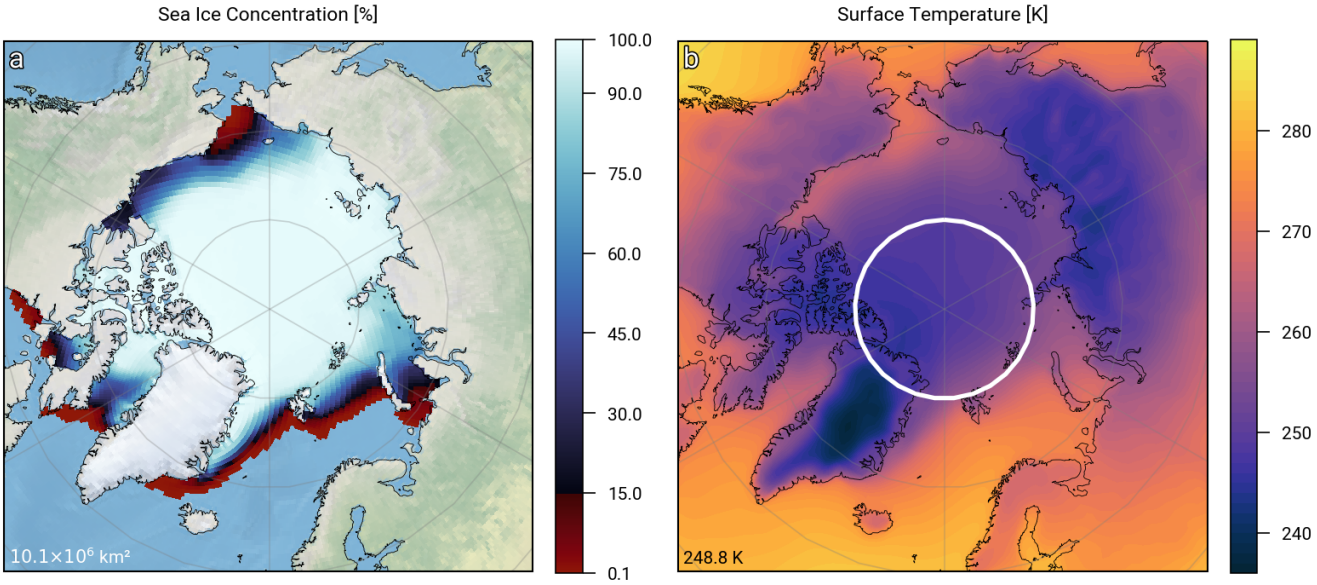

Figure 3.1.: Spatial distribution of the climate prediction for 1980-11-01 (40-member ensemble average). This climate prediction was initialized at 1979-11-01. a) Sea ice concentration. Cells with a concentration higher than 15% where used to calculate the sea ice extent, other cells were ignored. The corresponding sea ice extent for this date is shown in the lower-left corner. b) Surface temperature. The white circle indicates the 80°N parallel. The average surface temperature north of the 80°N parallel is shown in the lower-left corner.

In general, sea ice concentration is quite homogeneous through space (see Figure 3.1a): large area's contain a sea ice concentration higher than 90%. But sea ice does not spread evenly across all longitudes. The Arctic basin is fully covered by sea ice (see Figure 1.2 for the location of the seas within the Arctic Ocean). However, south of the basin (between 80°N and 70°N) sea ice does not fully cover the ocean. The Kara, Barents and Greenland seas, do not contain as much sea ice.

Especially in Baffin Bay the concentration is lower farther from the islands. The funnel-shaped Chukchi Sea seems to limit the growth of sea ice; due to the landmasses sea ice floats into

more southern regions, where it will melt due to the higher temperatures.

The surface temperature distribution in the Arctic region (Figure 3.1b), seems to follow the sea ice extent, however over icey landmasses (like Greenland, the Canadian Archipelago and Siberia) it shows even lower temperatures.

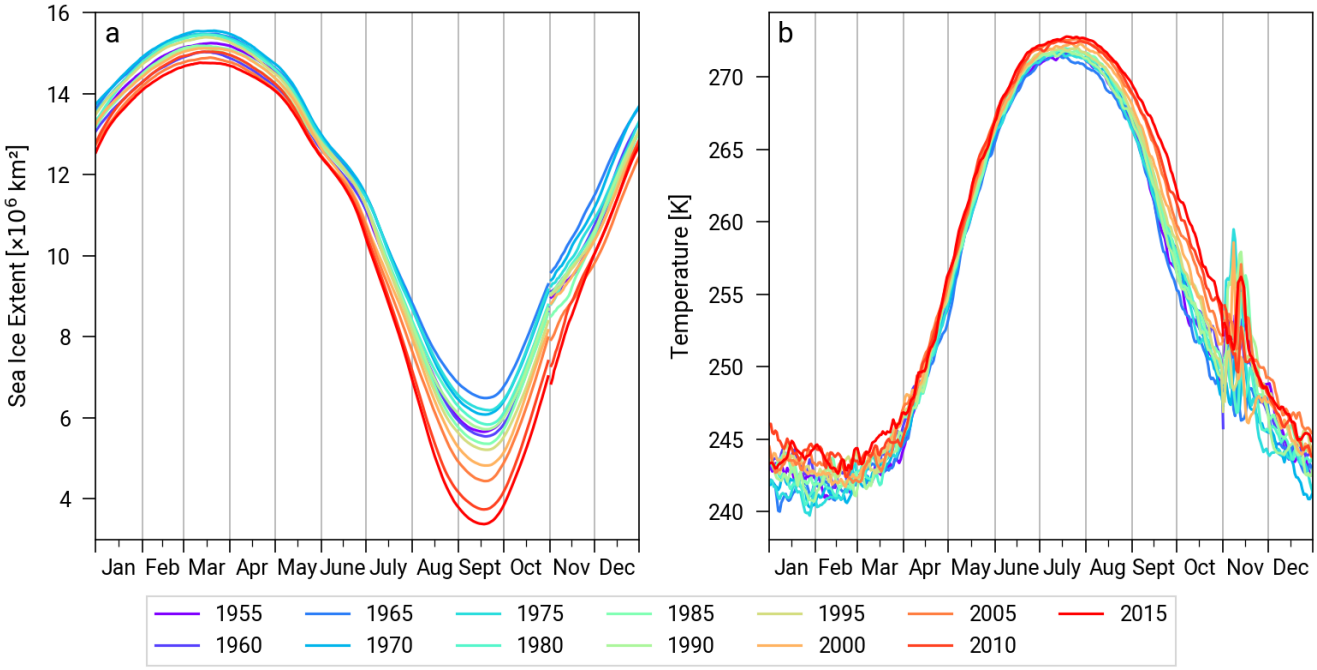

Figure 3.2.: Seasonal trend of the climate predictions (40-member ensemble average; daily data). Only the first year of a prediction is plotted. Since the predictions start on the 1st of November, a small jump is visible there. The seasonal trend is plotted for different years, at 5 year intervals, showing the long term trends. a) Sea ice extent. b) Surface temperature, spatial average north of 80°N.

In Figure 3.2, the seasonal trends of sea ice and surface temperature are shown. The seasonal trend of the surface temperature is quite variable, especially in November until March. The seasonal trend in sea ice extent however, shows less variability, due to the slower processes involved in melting and formation of sea ice. The seasonal cycle of sea ice is thus highly dominant in the signal, and this figure shows the need for the seasonal signal to be removed before any other corrections can be made.

Sea ice extent minima fall in September, a few months after the temperature maximum. This is expected since even in September, it is warm enough to melt sea ice away. Especially at more southern latitudes. In April temperatures above 80°N begin to rise, coinciding with the sea ice maxima.

The effects of climate change, are also visible in Figure 3.2. Temperatures are slowly rising and sea ice extend is overall decreasing. From 1990 onwards, this is especially visible. The minimal extend of sea ice (in September) is strongly effected by this change. In March, however, the differences are not that pronounced, but still distinguishable. Therefore we need to remove this trend, lead-time dependently.

Please note, that our predictions where initialized in November. Therefore there is a small gap in the lines between October 31st, and November 1st. The spikes in surface temperature in November, are probably due to an initialization shock of the CESM-DP-LE model.

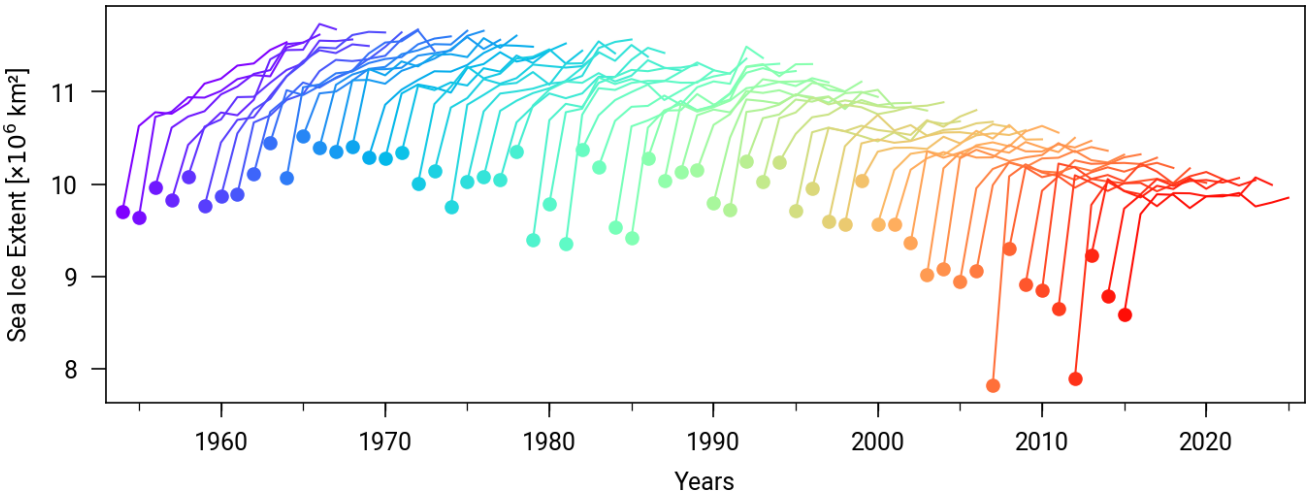

Figure 3.3.: Predictions (40-member ensemble average) for sea ice extent in November; monthly averaged. The dot (at the beginning of each line) indicates the initialization moment of each prediction, the line connects the following (10) Novembers in that prediction.

To illustrate the bias of the predictions developing over time, we show in Figure 3.3 the predictions of sea ice extent for November. Predictions are initialized at the November 1st, with observations (displayed by a dot). The lines connect the other years in the prediction (only November values). A striking feature of this figure is, the strong increase in sea ice extent ($0.8 \times 10^6\,\mathrm{km}^2$ on average) within the first year, similar to the schematic drawing in Figure 1.1. In following years, the bias develops further, but at a lower rate. Over time this development also slows (form a difference of $1.3 \times 10^6\,\mathrm{km}^2$ to a difference of $0.1 \times 10^6\,\mathrm{km}^2$ over the last 9 years). The figure also clearly shows the years 2007 and 2012, in which the sea ice extent reached historic minima. Overall, this indicates that the observations are not correctly followed by the prediction, and that a correction is needed.

3.2. Preprocessing

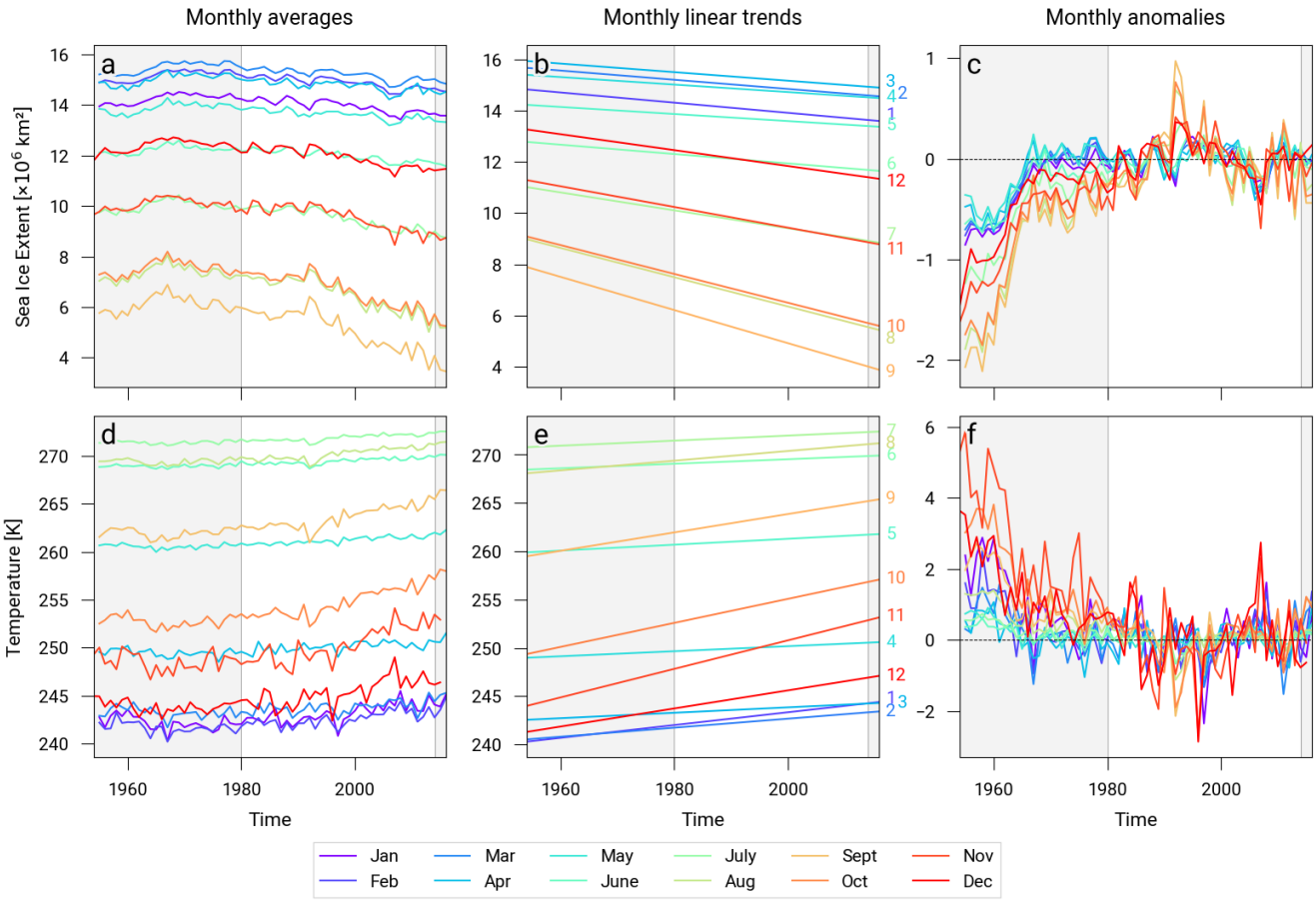

Figure 3.4.: Variations of the monthly averaged sea ice extent (a, b and c) and surface temperature (d, e and f), illustrating how the seasonal cycle and long term climate was removed and the anomalies were created. a & d) Monthly averaged ensemble averages for the first year of each prediction. Each line indicates a different month. b & e) Linear trends of the ensemble averages. These trends were created over the period 1980-2015 (white). The other periods (in grey; 1954-1980 and 2015-2016) were not taken into account. c & f) Anomalies in sea ice extent and temperature. Created by subtracting the linear trends (from figures b and e) from the monthly averages (figures a and d).

Figure 3.4 shows the pre-proccessing process. By subtracting the linear trends (3.4b and 3.4e) from the monthly averages (3.4a and 3.4d), we obtained monthly anomalies to the climate (3.4c and 3.4f). As mentioned in Section 2.2, only the period between 1980 and 2015 was used, to create the monthly linear trends (in white), and the other years (1954–1979 and 2016; in grey) were ignored to create these. The anomalies in 2016 do seem consistent with those between 1980 and 2015. However, the anomalies between 1954 and 1979 show large systematic deviations, ranging between −2 and 0×10^6 km^2.

From Figure 3.4, one may conclude that the climatological trend is a non linear process, and subtracting the linear trends from the monthly averages is not an accurate way to remove it. Another way, would be by computing 30-year moving averages for separate month. In this study, however, we choose to approach the trend linearly, to retain as much data as possible.

3.3. Uncorrected Predictions

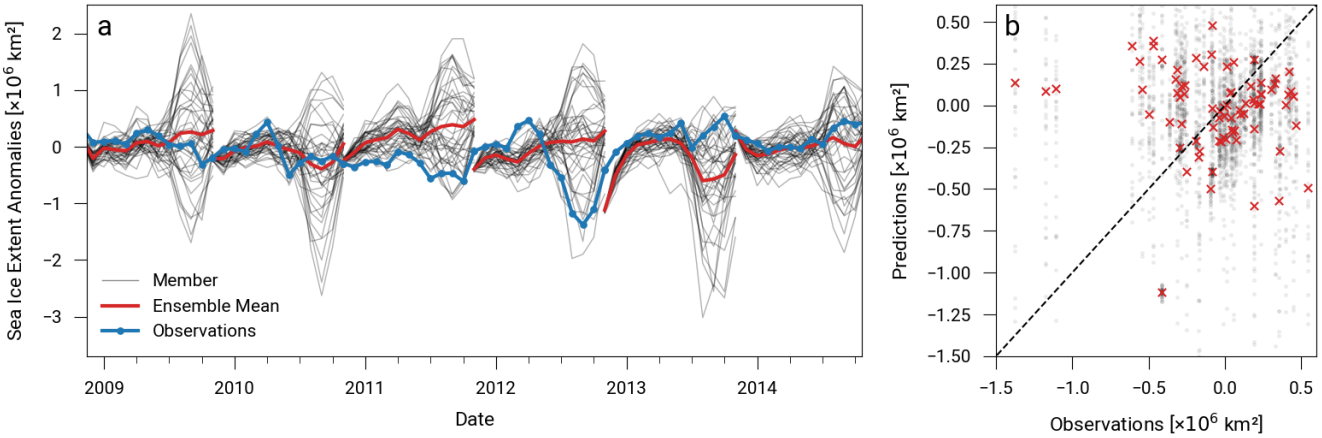

Figure 3.5.: Anomalies in sea ice extent; observations are compared with the first years of the uncorrected predictions. a) Sea Ice Extent anomalies over time. The grey, thin lines represent the different ensemble members, the red, thick line the ensemble average and the blue line with dots the observations. b) Cross plot between the observations and uncorrected predictions. Members are displayed in grey-semitransparent dots, ensemble means are displayed with red crosses.

To illustrate what our starting point is for our Ensemble Correction methods, we will first dive into the uncorrected predictions and the observations. In Figure 3.5a, we see the uncorrected ensemble prediction for sea ice extent, and the corresponding observations. The time range displayed is part of our validation dataset (see Section 2.3). In Figure 3.5b however, the entire validation dataset is plotted.

In Figure 3.5a, we see that the ensemble members show little variations between November and May. From June onward, the ensemble members diverge up to September, after which they converge again. One would expect that ensemble members would keep diverging from the ensemble, due to an increase in uncertainty, when making predictions in a chaotic system. However, sea ice has a clear seasonal cycle, and the period which shows the diverging ensemble members is the ablation period (see also Figure 3.2). Here the divergence comes from uncertainties in the ablation rate, and the magnitude of the minimal sea ice extent. Around September, the minimum sea ice extent is reached, and the members start to converge again. The ensemble mean (in Figure 3.5), does not deviate much from a sea ice extent anomaly of 0×10^6 km^2, except where the CESM-LE-DP model was initialized with a large anomaly (like in 2013). The observations fall often within the range of the ensemble, but not always.

To illustrate the accuracy of the prediction ensemble, we show in Figure 3.5b, the observations against the predictions. This figure highlights that many of the predictions and observations lay between -0.5 and 0.5×10^6 km^2, but there is not a high correlation between them. This is also shown in Table A.1 (in Appendix A.3).

3.4. Quantile Mapping

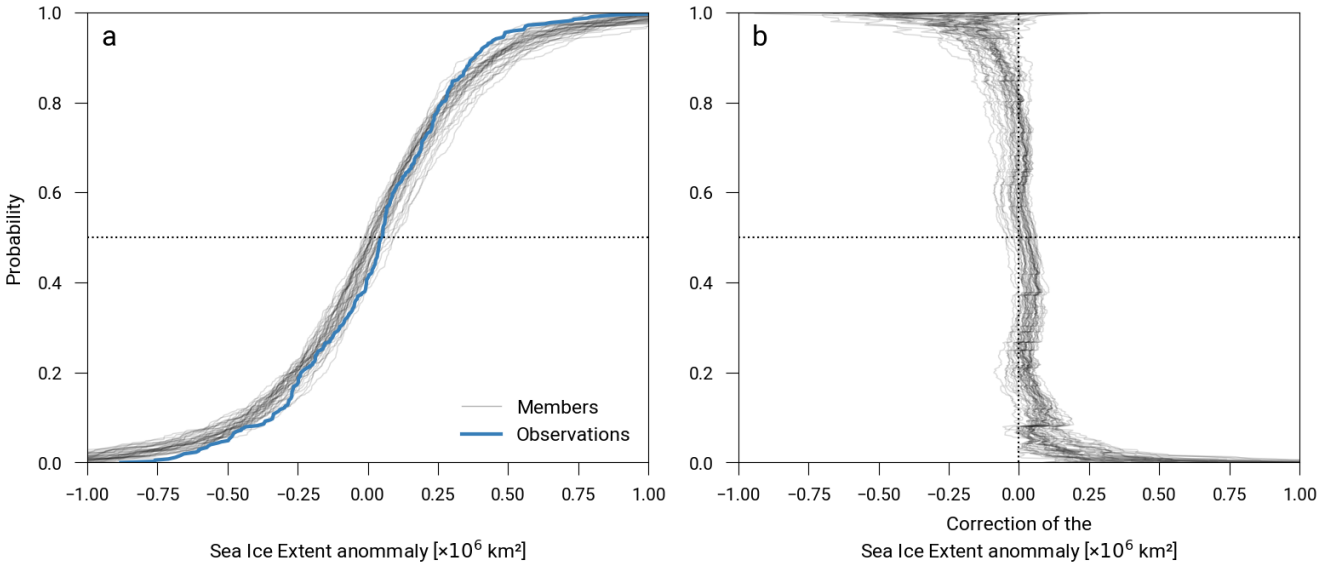

Figure 3.6.: a) Cumulative distribution figures (CDFs) of all members (in grey) and the observations (in blue), for the training period. The dotted line indicates the 50% quantiles or medians. b) Correction applied to each quantile, during Quantile Mapping, for each member. Calculated from the difference between the CDF of a member and the CDF of the observations.

To illustrate how Quantile Mapping was preformed, we show the cumulative distribution figures (CDFs) of all members and the observations during the training period (in Figure 3.6a). When comparing these CDFs, we see that the CDF of almost all members lie lower than the CDF of the observations, above the median. Below the median, almost all members show a higher probability than the observations. This indicates that the ensemble members have a higher spread than the observations, and are therefore overdispersive. This can also be seen in the correction that gets applied to the members (in Figure 3.6b). Negative sea ice anomalies are corrected positively, and positive sea ice anomalies are corrected negatively. Between 0.2 and 0.8, the correction is minimal.

The resulting ensemble looks quite squashed (see 3.7a): the peaks of the uncorrected ensemble (in July, August and September) have been limited at a maximum of 1.25 and a minimum of $-0.88 \times 10^6\,\mathrm{km}^2$. In November to May, the ensemble became also slightly more condensed and moved closer to $0 \times 10^6\,\mathrm{km}^2$. As a result, the ensemble average of November 2013 moved closer to the observations. When we compare the corrected ensemble to the observations overall, they seem to overlap slightly more. Figure 3.7b illustrates the same features. The ensemble means are more concentrated around the 1:1-line and the peaks are at $-0.88 \times 10^6\,\mathrm{km}^2$.

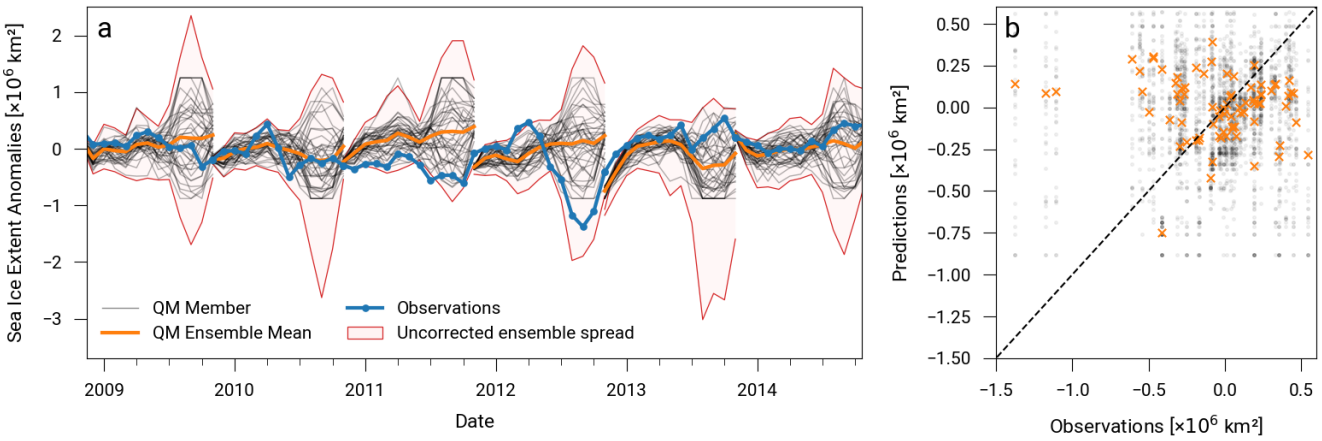

Figure 3.7.: Anomalies in sea ice ex tent; observations are compared with the first years of the predictions. Same as Figure 3.5, displaying here the Quantile mapping-corrected predictions. a) Sea Ice Extent anomalies over time. b) Cross plot between the observations and uncorrected predictions.

3.5. Ensemble Model Output Statistics

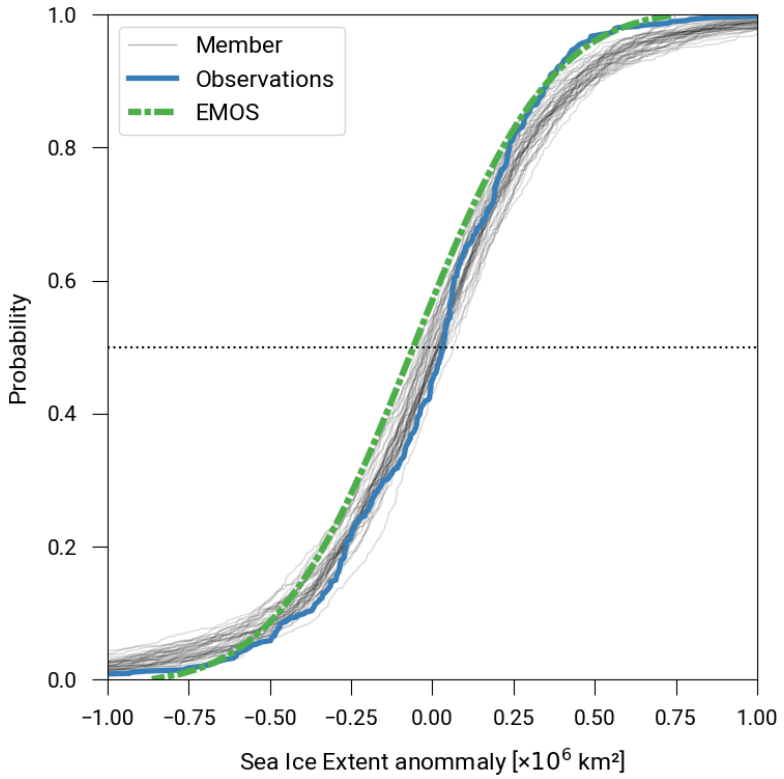

Figure 3.8.: Cumulative distribution figures (CDFs) of all members (in grey, thin lines) and the observations (in blue, continuous line) for the training period. The CDF, that follows from the normal distribution of the observations, approximated by the EMOS technique is displayed in green (dashed-dotted line).

Where Quantile Mapping uses the CDF of the observations directly, Ensemble Model Output Statistics (EMOS) approaches the distribution of the observations using a mathematical model (see also Sections 2.4.1 and 2.4.2). This is illustrated in Figure 3.8, which shows the CDFs for the observations and the EMOS approximated distribution, for our training period. Since EMOS assumes a normal distribution (which is adapted using the ensemble mean and

standard deviation), it is very smooth. In contrast with the CDFs of the members and observations, which are calculated empirically

When we comparing the CDF of EMOS with the CDF of the observations, we see that they deviate from each other, between a probability of 0.1 and 0.6. A possible explanation for this deviation is the loss function: the parameters of EMOS are optimized using the CRPS as a loss function. It seems here that the CRPS awards EMOS primaraly on improving the extremities, and does not motivate EMOS enough to also match the mean of the observations (which differs 0.1×10^6 km^2). The minimum of the EMOS distribution overlaps nicely with the minimum of the observations (during the training period), however, the maximum has decreased by 0.5×10^6 km^2. The kurtosis of the observations is thus larger than the kurtosis of the EMOS distribution.

In Table A.1, we see the effect of this difference in distribution. The mean of the ensemble deviates farther form the observations, than with the uncorrected ensemble. However, the standard deviation was lowed by 0.2×10^6 km^2, by the EMOS correction process. For all months, the standard deviation is lower than the standard deviation of the observations, indicating that the corrected ensemble is under dispersive

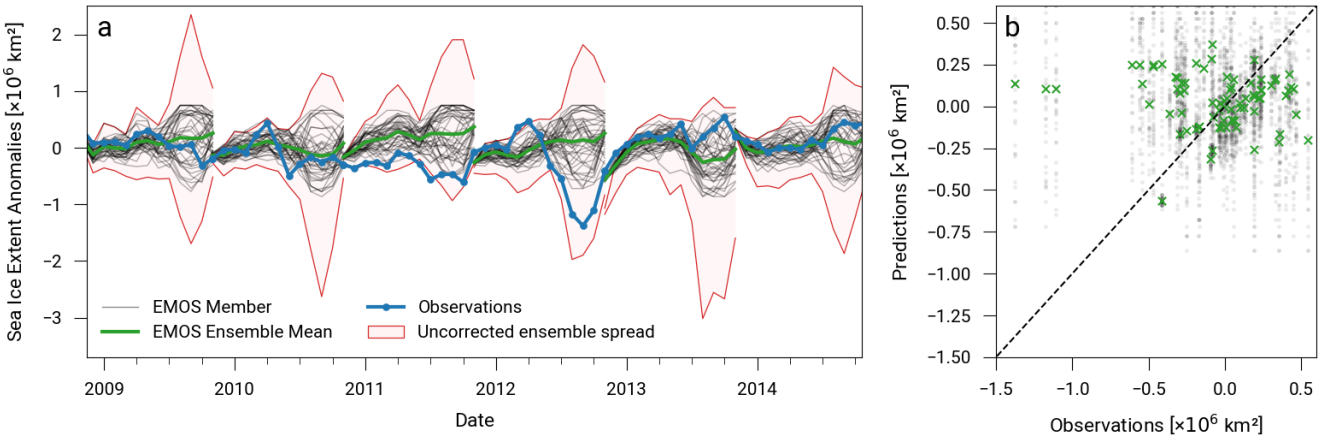

Figure 3.9.: Anomalies in sea ice extent; observations are compared with the first years of the predictions. Same as Figures 3.5 and 3.7, displaying here the EMOS-corrected predictions. a) Sea Ice Extent anomalies over time. b) Cross plot between the observations and uncorrected predictions.

When looking at the results of EMOS correction technique (in Figure 3.9a), we see that the main correction took place in the months July, August, and September, like with Quantile Mapping. However, the ensemble members here not squashed as with quantile mapping; the members are not limited to a maximum, but rather show a curve. This is also visiable in Figure 3.9b. Figure 3.9b also shows that the ensemble members are concentrated more on the 1:1 line, indicating a higher preciesion of the predictions.

In Table A.1, we see that the RMSE and CRPS scores improved compared with the uncorrected ensemble forecast. Both the RMSE and CRPS for the EMOS corrected ensemble scored slightly lower, compared with the uncorrected ensemble. However, for May, the EMOS skill scores are higher than those for the uncorrected data. Compared with Quantile mapping, EMOS scored only very slightly better.

3.6. Recurrent Neural Networks

In this section, we will discuss the results of our RNN experiments. Firstly, we tried to use RNNs to correct the seasonal cycle. Afterwards, we corrected the anomalies in the ensemble (just like with Quantile Mapping and Ensemble Model Output Statistics). We used two different setups for this: recurrency over time and recurrency over members (see Section 2.4.3).

3.6.1. Learning the Seasonal Cycle

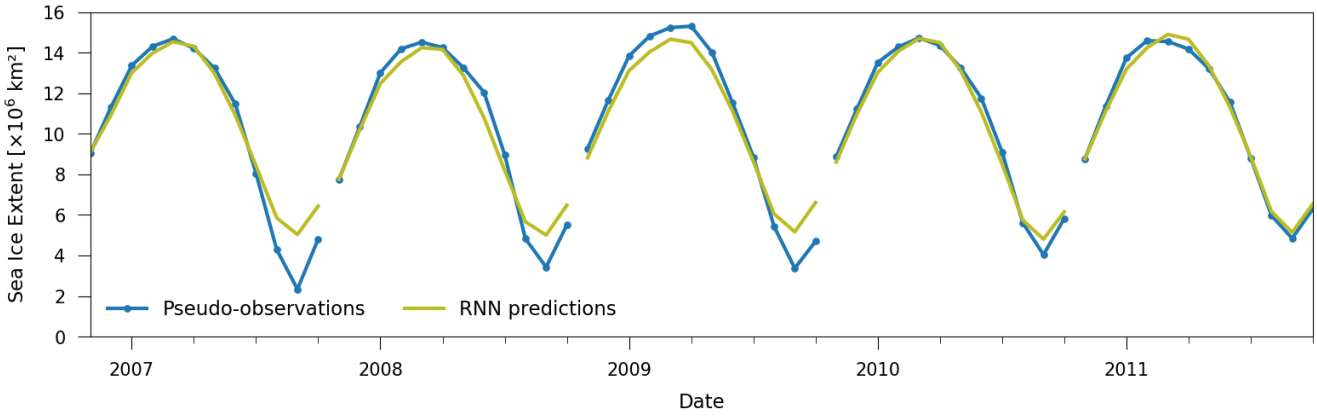

Figure 3.10.: Sea ice extent over time; the first year of the pseudo-observations and RNN predictions are compared with each other for part of the validation period (2006-2011). For this figure, the uncorrected ensemble member 23 represented the pseudo-observations.

When we look at the predictions our first RNN made (see Figure 3.10), they follow quite nicely the seasonal cycle of sea ice extent. The predictions start in November and are able to follow the seasonal cycle almost perfectly untill June. For the months August, September, and October, the RNN overestimates the amount of sea ice. Since we devided our training and validation dataset over time (with the later years as validation), the model was not able to learn from the period after 2003. Therefore it might have missed the seasonal trend of climate change when the sea ice minimum occurs (see also Figure 3.2). Another striking feature of the predictions is the lack of variation across years. The RNN seems to predict the general seasonal trend. Overall the RNN was able to achieve a R^2 of 0.96, for the validation set

The seasonal and climatological trends are dominant features in the sea ice cycle. However, the RNN was not able to predict both of them accurately. Therefore, this set up is not optimal to use as an ensemble correction method. To solve this issue, we removed these trends from the input data (using Equation 2.1, just like with Quantile mapping and EMOS). The RNN predictions of those experiments are discussed in the following two sections.

3.6.2. Anomaly correction with recurrence over Time

To find the best RNN to correct the anomalies we tried different setups in the number of nodes (between 16 and 256) and the number of layers (between 1 and 4). In Table 3.1, the R^2 scores for the best and worst preforming RNNs are shown. All R^2 scores are shown in Appendix A.2 (Table A.2).

The table shows high R^2 scores for November, but they quickly drop. For September the R^2 score never exceeds 0.1. Due to the divergence in the uncorrected ensemble in the ablation season, it appears harder for the RNN to predict accurately The loss-function might not moti-

Table 3.1.: R^2-scores for the entire validation period, for the Recurrent Neural Networks, with recurrence over Time. The 5 best and 5 worst preforming RNNs are displayed. All scores are displayed in Table A.2.

Nodes in hidden layer				R^2-score				
1st	2nd	3th	4th	Overall	Nov	Feb	May	Sept
256	–	–	–	0.209	0.942	0.413	0.233	0.048
256	256	–	–	0.209	0.946	0.405	0.229	0.046
256	256	256	–	0.204	0.963	0.410	0.202	0.044
16	–	–	–	0.203	0.907	0.331	0.188	0.056
32	–	–	–	0.199	0.922	0.397	0.215	0.046
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
64	64	64	64	-0.008	-0.183	-0.070	-0.297	-0.002
16	16	16	16	-0.009	-0.278	-0.048	-0.123	-0.013
16	32	64	128	-0.009	-0.240	-0.056	-0.185	-0.010
128	64	32	16	-0.020	-0.122	-0.137	-0.447	-0.002
256	128	64	32	-0.022	-0.112	-0.152	-0.468	-0.002

vate the RNN enough to improve its predictions here. The overall score for the RNNs is also quite low, due to the lack of accuracy in the last months. The table also shows that all 4-layer RNNs achieve R^2 scores below 0.13, often even below 0. In this case, more layers do not improve the predictions. This could indicate that the dataset used is to small to train these complex models efficiently.

Figure 3.11 shows the predictions for the 1-layer RNN with 256 nodes (denoted 1×256), and the 3-layer RNN with 256 nodes (denoted 3×256). We show these results since the 1×256 RNN preformed best overall, while the 3×256 RNN performed even better in November.

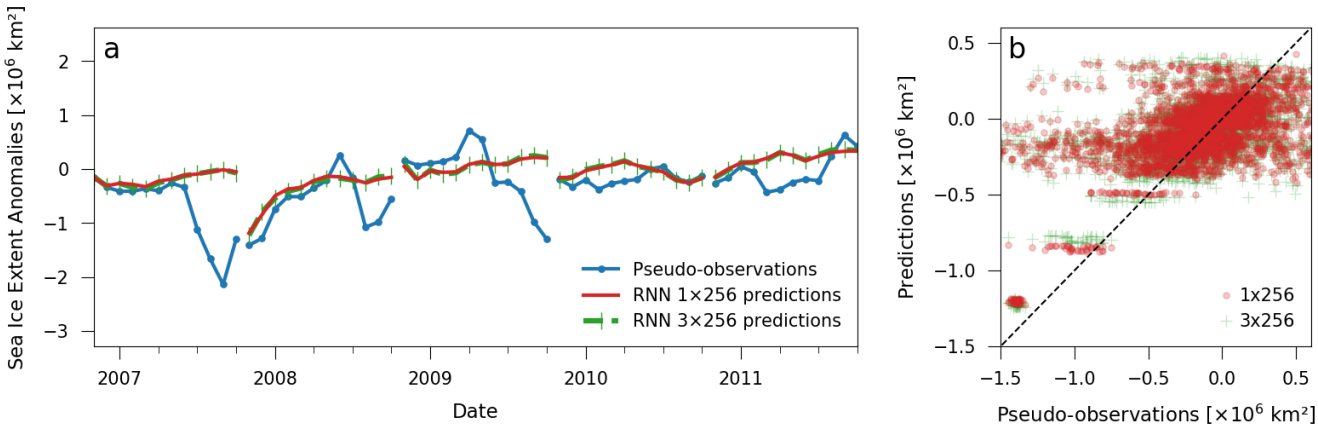

Figure 3.11.: Anomalies in sea ice extent; first year of the pseudo-observations and RNN predictions are compared with each other for part of the validation period (2006-2011). The RNNs here used the recurrence over time setup. The red line and dots indicate the 1-layer RNN predictions. The green line and crosses display the 3-layer RNN predictions. Both displayed RNNs had 256 LTSM nodes per layer. a) Sea Ice Extent anomalies over time. The blue line with dots represent the pseudo-observations (here, uncorrected member 23). b) Cross plot between the pseudo-observations and predictions.

As shown in Table 3.1, Figure 3.11a also shows that the predictions of our RNNs are quite accurate in November, but worsen over time. Furthermore, both RNNs predict close to the mean of the pseudo-observations: the lines are quite flat and do not deviate much from a sea ice extent anomaly of 0×10^6 km^2. This indicates that the RNNs are not able to accurately forecast the pseudo-observations for a certain time step (based on the uncorrected time step, and previous steps), or they are not motivated by the loss-function to learn that behavior.

In Figure 3.11b, we see the RNN predictions for the 1-layer and 3 layer RNN models (with 256 nodes per layer) plotted against the observations. Both RNNs show (almost) the same predictions. The bulk of the predictions is around a sea ice extent anomaly of 0.0×10^6 km^2. The RNN predictions clearly do not have the same distribution as the observations. Even though, there is some correlation visible

The RNNs also show a grouping behavior There are groups of predictions visible around −1.2, −0.8, and -0.5×10^6 km^2. To illustrate this behavior, we show the same plot for 4 different months (see Figure 3.12). In November, the predictions made by the RNN seem to fall within three discreet ranges. The years 2007 and 2012 (in which history sea ice extend minima occurred) are clearly visible in November. However sea ice growth, quickly diminshes these characteristic years. In later months we see mainly that the groups get wider; the observations spread out, but the predictions do not follow. For Septemer the neural network mainly predicts an ensemble average for each year, without much variation across the ensemble.

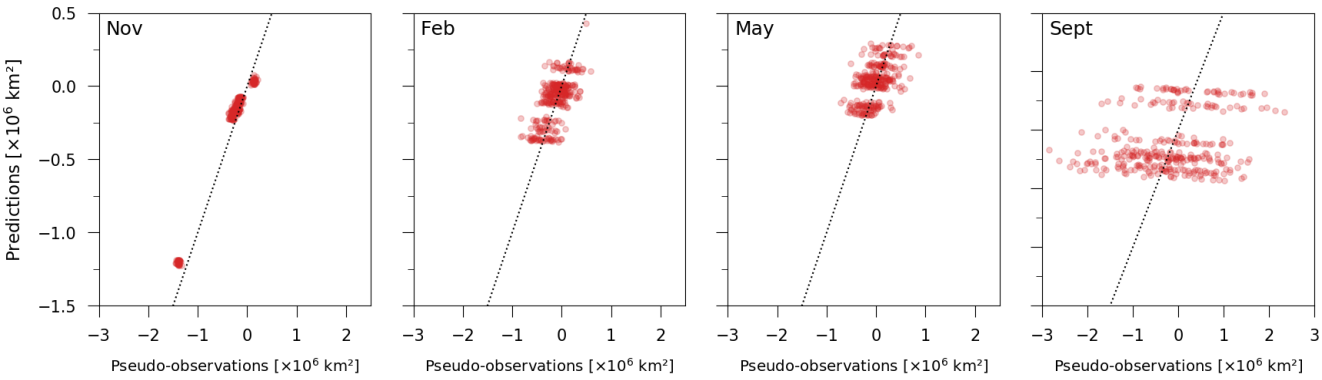

Figure 3.12.: Cross plot for the anomalies in sea ice extent, for different months; first year of the pseudo-observations and 1-layer 256 nodes RNN predictions, with recurrence over time.

3.6.3. Anomaly correction with recurrence over Members

In the following experiment, we tried correcting the anomalies, with the RNN recurrency step over members (instead of time). Since there are only 39 members to recurse over (the 40th member is used as pseudo-observation), the time in which these RNNs were trained was shorter, than when we recurred over time.

When we look at the R^2 scores for the different layer-node setups of this type of RNN (see Tables 3.2 and A.3), we notice the same features as in our last experiments: In November a high R^2 score is reached, but after November the score drops quickly; In September a very low score is achieved (with a maximum of 0.07); and due to the lack of accuracy in the last months, a low overall score for all RNNs is reached. In contrast with our last experiment, we see here that more layers and more nodes per layer generally lead to higher R^2 score. Also, no overall R^2 scores below 0.1 are generated. By using this setup, we seem to generate enough data to train these models well.

Table 3.2.: R^2-scores for the entire validation period, for the Recurrent Neural Networks, with recurrence over Members. The 5 best and 5 worst preforming RNNs are displayed. All scores are displayed in Table A.3.

Nodes in hidden layer				R^2-score				
1st	2nd	3th	4th	Overall	Nov	Feb	May	Sept
256	256	256	–	0.224	0.974	0.413	0.228	0.063
64	64	64	–	0.219	0.952	0.400	0.158	0.062
256	128	64	–	0.215	0.966	0.370	0.197	0.060
256	–	–	–	0.215	0.950	0.413	0.154	0.056
128	–	–	–	0.213	0.945	0.408	0.141	0.056
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
16	16	16	–	0.130	0.427	0.232	0.062	0.036
16	16	–	–	0.113	0.216	0.044	-0.078	0.058

In Table 3.2, we see that the best preforming model in this case is the 3-layer 256-nodes model (3×256). When looking at the predictions of this model over time (see Figure 3.13a), we again see many of the same features as when we recurred over time. The predictions for November are quite accurate, but with increasing time, the RNN seems to predict more and more the towards the ensemble average (around 0×10^6 km^2). The resulting lines are therefore relatively flat. In the cross plot (Figure 3.13b) we see, again, that many predictions are around a sea ice extent anomaly of 0×10^6 km^2. Also groups of predictions are still visible.

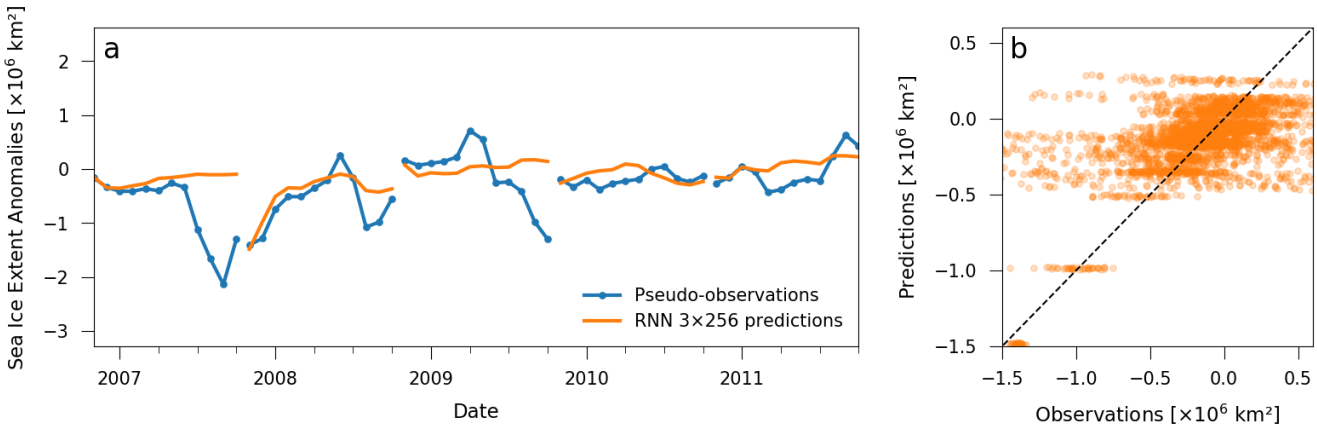

Figure 3.13.: Anomalies in sea ice extent; first year of the pseudo-observations and RNN predictions are compared with each other for part of the validation period (2006-2011). The RNN used here used the recurrence over members setup. The orange line (in a) and dots (in b) indicate the 3×256 RNN predictions. a) Sea Ice Extent anomalies over time. The blue line with dots represent the pseudo-observations (here, uncorrected member 23). b) Cross plot between the pseudo-observations and predictions.

3.7. Comparison and skill assessment

After looking at the uncorrected predictions and the corrected sea ice extent predictions (by Quantile Mapping, Ensemble Model Output Statistics and Recurrent Neural Networks), we want to compare these methods. To determine which method preformed best, we devised 3 skill scores (see also Section 2.3). In Table 3.3, we show the skill scores achieved by the (un)corrected sea ice extent predictions, for the validation period. As described in Section 2.3, both RMSE and CRPS are negatively oriented (a lower score, indicates a higher skill). The CRPSS looks at the difference between the uncorrected ensemble and a corrected ensemble. This score ranges between $-\infty$ and 1, where a score larger than 0 indicates that the corrected ensemble preformed better.

Since we used real observations to correct the ensemble with Quantile Mapping and Ensemble Model Output Statistics, and pseudo-observations when correcting with Recurrent Neural Networks, we can not directly compare the skill scores between them. However, we can look at the relative reduction of the RMSE and CRPS compared with uncorrected forecasts.

Table 3.3.: Skill scores of the predictions, by correction method. (Root Mean Square Error [RMSE], Continuous Ranked Probability Score [CRPS] and the Continuous Ranked Probability Skill Score [CRPSS]). The CRPSS was calculated with reference to the Uncorrected Predictions.

Correction method	RMSE $[\times 10^6\,\mathrm{km}^2]$	CRPS $[\times 10^6\,\mathrm{km}^2]$	CRPSS [–]
Compared with observations			
Uncorrected	0.4284	0.2305	–
Quantile Mapping	0.3973	0.2162	0.0624
Ensemble Model Output Statistics	0.3929	0.2166	0.0605
Compared with pseudo-observations			
Uncorrected	0.5146	0.2491	–
RNN, Recurrence over Time (1×256)	0.4467	0.2894	-0.1619
RNN, Recurrence over Members (3×256)	0.4470	0.2887	-0.1591

In Table 3.3, we see that both Quantile Mapping and EMOS result in a similar overall reduction in RMSE. The correction made by EMOS resulted in a slightly lower RMSE. The CRPS scores show the opposite result, indicating that the Quantile Mapping corrected ensemble shows a distribution closer to the observations than the EMOS corrected ensemble. From Figures 3.7 and 3.9, we see that EMOS limits the ensemble smoother, but slightly stronger. This results in a distribution that lies further from the observational distribution, resulting in a higher CRPS and lower CRPSS (compared with Quantile mapping).

Both RNN setups (recurrence over time, and recurrence over members) preform very similarly. Both show a reduction in RMSE, but an increase in the CRPS. When looking at the Figures 3.11 and 3.13, this was to be expected. The distribution of the RNN predictions is not comparable with the distribution of the pseudo-observations. When comparing the RNN scores with Quantile Mapping and EMOS, we see that the RNN models achieved a higher relative reduction in RMSE.

4. Discussion & Conclusion

In this study, we tried to improve climate predictions for sea ice extent, using different ensemble correction techniques. For this study, however, was a limited amount of time available and therefore some simplifications needed to be made. This study tries to show the benefit of post-processing these climate projections, but due to these simplifications, it can not be applied immediately into operational climate predictions. Instead, our study tries to introduce quantile mapping, ensemble model output statistics and recurrent neural networks, to the scientific field of post-processing climate predictions.

One of the main questions regarding this study could be: "Why would we focus on statistical post-processing the model if we know that there are small physical errors in our models, that create the model drift and biasses?". However, as stated, these decadal climate predictions have a high sociatal and scientific relevance. Users of these models (e.g. water resource managers and climate policy makers) depend on these predictions to be accurate These initialised climate predictions also allow us, to give users an estimation for the accuracy of our predictions and models. It is, therefore, necessary to post-process these climate predictions using statistical methods, at this point.

One of the simplifications we made, was using sea ice extent to represent all Arctic sea ice. This allowed us to preform simple time serie analysis on the data and verify the data with long series of observations. However sea ice forms in three dimensions. Taking the area with a sea ice concentration higher than 15% (using a threshold value), is quite a rough approximation for sea ice growth. Possible solutions for this issue include taking the sea ice area instead. since it takes the sea ice concentration as a weight when calculating the area. More three dimensional measures for sea ice could also be used (e.g. sea ice volume), however, there are no long observational time series for these measures available.

The choice for sea ice extent, also removed all spatial information about sea ice in the Northern hemisphere from our data. However, as was shown in figure 3.1, there is quite some variation visible. By training our ensemble correction techniques on different regions within the arctic ocean, one might be able to more accurately correct the biases within the model. Krikken et al. (2016) showed that another correction technique (Extended Logistic Regression) achieves different skill scores in different regions in the Arctic Ocean. One could also use the high resolution output directly, for example using a convolutional neural network (CNN). These CNNs have been used for image recognition and excel at pattern detection in spatial grids (Krizhevsky et al., 2012).

Another way to improve our ensemble correction techniques is by using more data. Using more (or only) predictor variables to train the RNN might result in better predictions. Also using different ensemble forcasts together might improve both our EMOS and RNN results. Quantile Mapping is insensitive to this since it simply fits one variable onto the distribution of another.

In this study, we used one CDF for quantile mapping for the entire dataset. In the same way, we used fitted EMOS once for the entire dataset and were the RNNs trained on the entire dataset. However, as we have shown on multiple occasions, do different months show different properties and distributions. It might, therefore, be advisable to use different distributions for different months (or seasons), given that there is enough data available to create good representations for these distributions.

Conclusion

During the course of this study, we have successfully implemented and used Quantile Mapping, Ensemble Model Output Statistics and Recurrent Neural Networks (with Long Short Term Memory nodes) to improve the predictive skill of the outcome of a climate prediction in the Arctic region, for sea ice extent. However, the resulting predictions leave much to be desired. EMOS and Quantile Mapping show some improvement in the ensemble, but no large correction of the bias was found. Our Recurrent Neural Networks also show promising results, but the distribution of those predictions is largely under-dispersive, compared with the observations and ensemble members.

A. Appendices

A.1. Notation

This list describes the symbols and their meaning, which are used in this report.

Variables

E	Sea ice extent
T	Surface temperature at 2 m
E'	Climatological anomalies of Sea ice extent
$\hat{E}'$	Normalized climatological anomalies of Sea ice extent

Forecasts

X	Forecast series
x	A forecast $(x \in X)$
X_n^t	Ensemble member n at timestep t of the forecast
$\overline{X}$	Mean of the ensemble members
s_X^2	Variance of the ensemble members
x_{corr}	Corrected forecast

Observations

Y	Observation series
y	A observation $(y \in Y)$

Distributions

F_A	Cumulatice distribution function (CDF) of the series A
F_A^{-1}	Inverse of the CDF of the series A
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
θ	Quantile

Neural networks

$\boldsymbol{b}$	Biasses
$\boldsymbol{c}$	Cell state of a LSTM node
$\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$	Canidate cell state of a LSTM node
$\boldsymbol{f}$	Forget gate (of a LSTM node)
$\boldsymbol{h}$	Hidden state of a LSTM node
$\boldsymbol{i}$	Input gate (of a LSTM node)

L	Loss function
$\boldsymbol{o}$	Output gate (of a LSTM node)
$\boldsymbol{p}$	Parameters (weights and biases)
$\boldsymbol{v_s}$	Moving mean square gradient at sample s
$\boldsymbol{w}$	Weights
$\boldsymbol{x_t}$	Input vector at time t
$\boldsymbol{y_t}$	Output vector at time t
$\boldsymbol{z_t}$	Input vector and hidden state appended together
$\sigma(x)$	Sigmoid function ($\sigma(x) = {}^{1}/_{1+e^{-x}}$)
$\tanh(x)$	Hyperbolic tangent function

Symbols & Operators

$\boldsymbol{a}$	Vector
P	Probability
$=$	is equal to
$\equiv$	is defined as
$\in$	is an element of
$\mid$	is such that
$\sim$	is distributed as
∇	gradient
$\odot$	element-wise multiplication

A.2. Statistical values of the different predictions, and observations

Table A.1.: Statistical values showing the distribution, range, correlation and skill of the (corrected) predictions and observations (observat.) of sea ice extent.

				Per Month			
ECT	Measure	Unit	Overall	Nov	Feb	May	Sept
Observat.	Mean	[$\times 10^6$ km^2]	-0.1117	-0.0614	-0.0502	-0.0484	-0.3083
	Std.Dev	[$\times 10^6$ km^2]	0.3712	0.3243	0.1721	0.3044	0.5930
	Minimum	[$\times 10^6$ km^2]	-1.3773	-0.4686	-0.2838	-0.7454	-1.3773
	Maximum	[$\times 10^6$ km^2]	0.5422	0.4232	0.1997	0.3030	0.4507
Uncorrected	Mean	[$\times 10^6$ km^2]	-0.1250	-0.3623	-0.1232	-0.0060	-0.1447
	Std.Dev	[$\times 10^6$ km^2]	0.5043	0.4909	0.2331	0.2581	0.9411
	Minimum	[$\times 10^6$ km^2]	-3.0173	-1.4507	-0.8344	-0.8299	-2.8524
	Maximum	[$\times 10^6$ km^2]	2.3493	0.3273	0.5856	0.8514	2.3493
	R^2	[–]	0.0001	0.3451	0.0244	0.0684	0.0676
	RMSE	[$\times 10^6$ km^2]	0.4586	0.4979	0.2531	0.2877	0.7319
	CRPS	[$\times 10^6$ km^2]	0.2486	0.3652	0.1322	0.1553	0.4106
	CRPSS	[–]	–	–	–	–	–
Quantile Mapping	Mean	[$\times 10^6$ km^2]	-0.0763	-0.2485	-0.0938	0.0075	-0.0350
	Std.Dev	[$\times 10^6$ km^2]	0.3618	0.3200	0.2098	0.2228	0.6061
	Minimum	[$\times 10^6$ km^2]	-0.8799	-0.8799	-0.6846	-0.6024	-0.8799
	Maximum	[$\times 10^6$ km^2]	1.2498	0.3277	0.5593	0.7935	1.2498
	R^2	[–]	0.0009	0.3988	0.0269	0.0761	0.0750
	RMSE	[$\times 10^6$ km^2]	0.4237	0.3267	0.2347	0.2855	0.6989
	CRPS	[$\times 10^6$ km^2]	0.2307	0.2475	0.1248	0.1543	0.3965
	CRPSS	[–]	0.0720	0.3222	0.0559	0.0065	0.0342
EMOS	Mean	[$\times 10^6$ km^2]	-0.0322	-0.1747	-0.0518	0.0400	-0.0062
	Std.Dev	[$\times 10^6$ km^2]	0.2993	0.2688	0.1769	0.2046	0.4853
	Minimum	[$\times 10^6$ km^2]	-0.8631	-0.7195	-0.4808	-0.4529	-0.8631
	Maximum	[$\times 10^6$ km^2]	0.7472	0.3506	0.4996	0.6036	0.7472
	R^2	[–]	0.0007	0.3843	0.0213	0.0828	0.0833
	RMSE	[$\times 10^6$ km^2]	0.4149	0.2774	0.2140	0.2914	0.6920
	CRPS	[$\times 10^6$ km^2]	0.2278	0.2221	0.1171	0.1548	0.3981
	CRPSS	[–]	0.0834	0.3917	0.1144	0.0031	0.0302

A.3. RNN Results

A.3.1. Anomaly correction with recurrence over Time

Table A.2.: R^2-scores for the entire validation period, for the Recurrent Neural Networks, with recurrence over Time.

Nodes in hidden layer				R^2-score				
1st	2nd	3th	4th	Overall	Nov	Feb	May	Sept
16	—	—	—	0.203	0.907	0.331	0.188	0.056
32	—	—	—	0.199	0.922	0.397	0.215	0.046
64	—	—	—	0.198	0.904	0.366	0.176	0.045
128	—	—	—	0.176	0.877	0.318	0.128	0.036
256	—	—	—	0.209	0.942	0.413	0.233	0.048
16	16	—	—	0.180	0.826	0.243	0.139	0.045
32	32	—	—	0.184	0.818	0.324	0.185	0.039
64	64	—	—	0.193	0.885	0.355	0.168	0.045
128	128	—	—	0.190	0.813	0.365	0.129	0.056
256	256	—	—	0.209	0.946	0.405	0.229	0.046
16	16	16	—	-0.003	-0.289	-0.056	-0.079	-0.006
32	32	32	—	0.105	0.121	0.192	0.168	0.020
64	64	64	—	0.130	0.259	0.166	0.146	0.030
128	128	128	—	0.168	0.802	0.305	0.145	0.029
256	256	256	—	0.204	0.963	0.410	0.202	0.044
16	16	16	16	-0.009	-0.278	-0.048	-0.123	-0.013
32	32	32	32	-0.005	-0.197	-0.051	-0.249	-0.005
64	64	64	64	-0.008	-0.183	-0.070	-0.297	-0.002
128	128	128	128	0.008	-0.184	-0.026	-0.084	-0.004
256	256	256	256	0.126	0.345	0.247	0.087	0.025
16	32	64	—	0.127	0.258	0.335	0.157	0.030
32	64	128	—	0.186	0.829	0.302	0.179	0.050
64	128	256	—	0.152	0.683	0.242	0.044	0.044
16	32	64	128	-0.009	-0.240	-0.056	-0.185	-0.010
32	64	128	256	-0.006	-0.232	0.133	-0.112	-0.030
64	32	16	—	0.004	-0.240	-0.026	-0.119	-0.002
128	64	32	—	0.143	0.694	0.161	-0.021	0.036
256	128	64	—	0.171	0.831	0.349	0.124	0.038
128	64	32	16	-0.020	-0.122	-0.137	-0.447	-0.002
256	128	64	32	-0.022	-0.112	-0.152	-0.468	-0.002

A.3.2. Anomaly correction with recurrence over Members

Table A.3.: R^2-scores for the entire validation period, for the Recurrent Neural Networks, with recurrence over Members.

Nodes in hidden layer				R^2-score				
1st	2nd	3th	4th	Overall	Nov	Feb	May	Sept
16	–	–	–	0.152	0.589	0.273	0.011	0.041
32	–	–	–	0.192	0.932	0.190	0.062	0.054
64	–	–	–	0.203	0.875	0.275	0.098	0.061
128	–	–	–	0.213	0.945	0.408	0.141	0.056
256	–	–	–	0.215	0.950	0.413	0.154	0.056
16	16	–	–	0.113	0.216	0.044	-0.078	0.058
32	32	–	–	0.169	0.777	0.396	0.099	0.023
64	64	–	–	0.208	0.937	0.323	0.122	0.059
128	128	–	–	0.208	0.966	0.395	0.218	0.046
256	256	–	–	0.204	0.972	0.323	0.173	0.054
16	16	16	–	0.130	0.427	0.232	0.062	0.036
32	32	32	–	0.170	0.894	0.197	-0.076	0.046
64	64	64	–	0.219	0.952	0.400	0.158	0.062
128	128	128	–	0.204	0.972	0.233	0.087	0.062
256	256	256	–	0.224	0.974	0.413	0.228	0.063
16	16	16	16	0.147	0.524	0.154	-0.065	0.053
32	32	32	32	0.188	0.804	0.217	0.218	0.050
64	64	64	64	0.202	0.815	0.299	0.071	0.071
128	128	128	128	0.204	0.848	0.292	0.169	0.065
256	256	256	256	0.185	0.853	0.155	-0.070	0.062
16	32	64	–	0.198	0.634	0.439	0.158	0.065
32	64	128	–	0.209	0.935	0.296	0.154	0.066
64	128	256	–	0.184	0.891	0.219	0.015	0.057
16	32	64	128	0.188	0.931	0.196	0.091	0.065
32	64	128	256	0.198	0.763	0.293	0.027	0.071
64	32	16	–	0.177	0.762	0.267	-0.108	0.059
128	64	32	–	0.196	0.938	0.173	0.065	0.060
256	128	64	–	0.215	0.966	0.370	0.197	0.060
128	64	32	16	0.178	0.846	0.202	-0.058	0.062
256	128	64	32	0.205	0.814	0.417	0.186	0.064

Bibliography

Abadi, M., and Coauthors, 2016: TensorFlow : A System for Large-Scale Machine Learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, USENIX Association, Savannah, GA, USA, 265–283, URL https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

Budikova, D., 2009: Role of Arctic sea ice in global atmospheric circulation: A review. *Global and Planetary Change*, **68 (3)**, 149–163, doi: 10.1016/j.gloplacha.2009.04.001.

Chollet, F., 2015: Keras. URL https://keras.io.

Cochran, P., O. H. Huntington, C. Pungowiyi, S. Tom, F. S. Chapin, H. P. Huntington, N. G. Maynard, and S. F. Trainor, 2014: Indigenous frameworks for observing and responding to climate change in Alaska. *Climate Change and Indigenous Peoples in the United States: Impacts, Experiences and Actions*, 49–59, doi: 10.1007/978-3-319-05266-3_5.

Danish Meteorological Institute, 2019: Daily average temperatures in the Arctic 1958-2019. Copenhagen, URL http://ocean.dmi.dk/arctic/meant80n.php.

Divine, D. V., and C. Dick, 2006: Historical variability of sea ice edge position in the Nordic Seas. **111**, 1–14, doi: 10.1029/2004JC002851.

Fetterer, F., W. Knowles, W. Meier, M. Savoie, and A. Windnagel, 2017: Northern hemisphere, daily sea ice extent. *Sea Ice Index*, version 3 ed., NSIDC: National Snow and Ice Data Center, Boulder, Colorado USA, doi: 10.7265/N5K072F8.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, **133 (5)**, 1098–1118, doi: 10.1175/MWR2904.1.

Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, **15 (5)**, 559–570, doi: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hinton, G., N. Srivastava, and K. Swersky, 2012: rmsprop: Divide the gradient by a running average of its recent magnitude. Computer Science, University of Toronto, Toronto, URL https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, 26–31 pp.

Hochreiter, S., and J. Schmidhuber, 1997: Long Short-Term Memory. *Neural Computation*, **9 (8)**, 1735–1980, doi: 10.1162/neco.1997.9.8.1735, URL https://www.bioinf.jku.at/publications/older/2604.pdf.

Kalnay, E., and Coauthors, 1996: The ncep/ncar 40-year reanalysis project. *Bulletin of the American Meteorological Society*, **77 (3)**, 437–472, doi: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kay, J. E., and Coauthors, 2015: The community earth system model (CESM) large ensemble project : A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, **96 (8)**, 1333–1349, doi: 10.1175/BAMS-D-13-00255.1.

Kharin, V. V., G. J. Boer, W. J. Merryfield, J. F. Scinocca, and W. S. Lee, 2012: Statistical adjustment of decadal predictions in a changing climate. *Geophysical Research Letters*, **39 (19)**, 1–6, doi: 10.1029/2012GL052647.

Khon, V. C., I. I. Mokhov, M. Latif, V. A. Semenov, and W. Park, 2010: Perspectives of Northern Sea Route and Northwest Passage in the twenty-first century. *Climatic Change*, **100 (3)**, 757–768, doi: 10.1007/s10584-009-9683-2.

Koenker, R., and G. Bassett, 1978: Regression Quantiles. *Econometrica*, **46 (1)**, 33, doi: 10.2307/1913643.

Krikken, F., M. Schmeits, W. Vlot, V. Guemas, and W. Hazeleger, 2016: Skill improvement of dynamical seasonal Arctic sea ice forecasts. *Geophysical Research Letters*, **43 (10)**, 5124–5132, doi: 10.1002/2016GL068462.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012: ImageNet Classification with Deep Convolutional Neural Networks. *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, Lake Tahoe, US, 1097–1105.

Larsen, J. N., O. A. Anisimov, A. Constable, A. B. Hollowed, N. Maynard, P. Prestrud, T. D. Prowse, and J. M. R. Stone, 2014: Polar regions. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel of Climate Change*, V. R. Barros, C. B. Field, D. J. Dokken, M. D. Mastrandrea, K. J. Mach, T. E. Bilir, M. Chatterjee, K. L. Ebi, Y. O. Estrada, R. C. Genova, B. Girma, E. S. Kissel, A. N. Levy, S. MacCracken, P. R. Mastrandrea, and L. L. White, Eds., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1567–1612, URL http://www.ipcc.ch/pdf/assessment-report/ar5/wg2/WGIIAR5-Chap28_FINAL.pdf.

Lindholt, L., 2006: Arctic natural resources in a global perspective. *The Economy of the North*, Statistics Norway, Oslo, chap. 3, 27–40.

McCulloch, W. S., and W. Pitts, 1943: A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, **5 (4)**, 115–133, doi: 10.1007/BF02478259, arXiv:1011.1669v3.

Meier, W. N., and Coauthors, 2014: Arctic sea ice in transformation: A review of recent observed changes and impacts on biology and human activity. *Reviews of Geophysics*, **52 (3)**, 185–217, doi: 10.1002/2013RG000431.

Meinshausen, N., 2006: Quantile Regression Forests. *Journal of Machine Learning Research*, **7**, 983–999, doi: 10.1111/j.1541-0420.2010.01521.x, NIHMS150003.

MeteoSwiss, 2017: *easyVerification: Ensemble Forecast Verification for Large Data Sets*. URL https://cran.r-project.org/package=easyVerification.

Parkinson, C. L., D. J. Cavalieri, P. Gloersen, H. J. Zwally, and J. C. Comiso, 1999: Arctic sea ice extents, areas, and trends, 1978–1996. *Journal of Geophysical Research: Oceans*, **104 (C9)**, 1978–1996, doi: 10.1029/1999JC900082.

Quinlan, 1986: Induction of Decision Trees. Tech. rep., 81–106 pp. URL https://link.springer.com/content/pdf/10.1023%2FA%3A1022643204877.pdf.

R Core Team, 2018: *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing, URL https://www.r-project.org/.

Rasp, S., and S. Lerch, 2018: Neural networks for post-processing ensemble weather forecasts. URL http://arxiv.org/abs/1805.09091, 1805.09091.

Schweiger, A., R. Lindsay, J. Zhang, M. Steele, H. Stern, and R. Kwok, 2011: Uncertainty in modeled Arctic sea ice volume. **116 (June)**, 1–21, doi: 10.1029/2011JC007084.

Sigmond, M., J. C. Fyfe, G. M. Flato, V. V. Kharin, and W. J. Merryfield, 2013: Seasonal forecast skill of Arctic sea ice area in a dynamical forecast system. *Geophysical Research Letters*, **40 (3)**, 529–534, doi: 10.1002/grl.50129.

Stroeve, J., M. M. Holland, W. Meier, T. Scambos, and M. Serreze, 2007: Arctic sea ice decline: Faster than forecast. *Geophysical Research Letters*, **34 (9)**, 1–5, doi: 10.1029/2007GL029703.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics. *Monthly Weather Review*, **144 (6)**, 2375–2393, doi: 10.1175/MWR-D-15-0260.1, URL http://journals.ametsoc.org/doi/10.1175/MWR-D-15-0260.1.

Vaughan, D., and Coauthors, 2013: Observations: Cryosphere. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley, Eds., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 317–382, URL https://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_Chapter04_FINAL.pdf.

Verkade, J. S., J. D. Brown, P. Reggiani, and A. H. Weerts, 2013: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, **501**, 73–91, doi: 10.1016/j.jhydrol.2013.07.039.

Walsh, W. L., and J. E. Chapman, 2001: 20Th Century Sea-Ice Variations From Observational Data. *Annals of Glaciology*, **33 (1)**, 444–448, doi: 10.3189/172756401781818671.

Wilks, D. S., and T. M. Hamill, 2007: Comparison of Ensemble-MOS Methods Using GFS Reforecasts. *Monthly Weather Review*, **135 (6)**, 2379–2390, doi: 10.1175/MWR3402.1.

World Meteorological Organization, 2015: Sea ice nomenclature. Tech. Rep. 259, WMO, Geneva, Switzerland, 1–147 pp.

Yeager, S. G., and Coauthors, 2018: Predicting near-term changes in the Earth System: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. *Bulletin of the American Meteorological Society*, BAMS–D–17–0098.1, doi: 10.1175/BAMS-D-17-0098.1.