Norwegian School of Economics
Bergen, Fall 2021

Freight pricing and climate change: The case of grain cargoes in the Paraná River

Kaja Berg Henie and Hanne Utheim

Supervisor: Roar Os Ådland

Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Contents

Abstract

Climate change is affecting the shipping industry in Paraná River in Argentina. Lower levels of precipitation cause the water levels to decrease, which limits the amount of cargo a ship can load. Accurate predictions of water level in shallow rivers are therefore crucial in reducing the risk of under- or overestimating the cargo during trades.

This thesis aims to create a predictive model for the maximum allowable draft for vessels to pass through Paraná River, using historical levels of precipitation and precipitation forecasts. The maximum allowable draft levels are obtained using reported draft from official AIS data. In the process, we have studied the seasonal pattern of precipitation and tested different types of moving averages to calculate the precipitation variable. Further on, we investigated different statistical methods, including generalized additive models and linear and multiple regression models. Using the generalized additive model, the precipitation variable explained ~40% of the variance in maximum allowable draft. A time-based cross-validation method was utilized to predict future maximum draft levels, based upon precipitation forecasts. Finally, the modelling accuracy using AIS data is compared to the modelling accuracy of using monthly reported water levels in Paraná River.

The result suggests that the high-frequency AIS reported draft can perform better than monthly reported water levels, and therefore provide a better estimation basis for pricing of cargoes. The model provides estimated water level margins 1-12 weeks ahead of time and may be used by shipping companies to improve cargo estimation and reduce risk.

Acknowledgements

This thesis was written as a concluding part of our master's degree at the Norwegian School of Economics. Due to little experience and knowledge within the shipping field, it has been an educational process with a steep learning curve. We would like to express our gratitude for all the support we have received along the way. We want to thank our supervisor Roar Os Ådland for sharing his valuable network with us and for always providing prompt and accurate responses to all our inquires. We would also like to give a big thanks to Alex Lenkoski, who spent much time collecting, structuring, and forecasting weather data for us, and teaching us about various meteorological statistical approaches in R. Furthermore, we must thank Gabriel Fuentes who provided access to AIS data, as well as giving us advice for our code. We would also like to express our gratitude to Western Bulk for providing us with data, as well as bringing us to their office to learn more about the financial and theoretical aspects of the shipping industry. Finally, we thank *The Norwegian Shipowners Associations Fund* at NHH for their grants, and we hope our work will be of interest.

Norwegian School of Economics

Bergen, December 2021

Kaja Berg Henie

Hanne Utheim

1. Introduction

Argentina produces 5% of the total world grain, according to the U.S. Department of Agriculture (2021). During the last year, world prices of various grains and especially wheat have risen. This is due to the climate-related issues, such as low precipitation levels and drought (Wei, et. al., 2019). For example, drought has during the past 1-2 years lowered Australia's supply of grain, which has increased the demand of Argentinian grain (U.S. Department of Agriculture, 2021).

This master thesis will concentrate on the Paraná River in Argentina. The 1,477-km waterway is essential for the growth of the Argentinian economy as it links the Atlantic Ocean with major ports inland. The river transports approximately 80% of national exports, and 81% of the volume of agricultural exports (Portal oficial del Estado argentine, n.d.a). However, the river is currently facing conditions of extreme drought. This has reduced the water level in the river to record low levels, forcing grain exporters to load less cargo than ordered. This influences the pricing of cargoes due to lower supply than demand, which then makes the voyage cost sensitive to water levels.

Due to such factors and their consequences, Western Bulk has requested a tool that can reduce the risk of uncertainties in Paraná River, where they have a high volume of trades. Currently, shipping companies receive daily or weekly reports from local pilot companies on river conditions, as a basis for estimation of cargo load. In this master thesis, we aim to create a prediction model that can be used to forecast the maximum allowable draft for vessels to pass through the river. The prediction of maximum allowable draft should reduce the risk of under- or overestimating the cargo during trades, but at the same time portray good enough margins to not run aground. This model needs to be able to predict water levels from 2 weeks to 3 months ahead in order to be useful for the company.

In our approach to predict the maximum allowable draft, we will use the reported draft from all ships in the area from official Automatic Identification System (AIS) data. Since 2002, all new ships and seagoing vessels with a gross tonnage of over 300 tons, including passenger ships, have been obliged to install an Automatic Identification System (Wu, et. al., 2020. The system contains high-frequency data of reported draft and can provide an alternative approach of receiving information of the water levels.

The contributions of this thesis are twofold: Firstly, we investigate whether AIS data is a good data source to measure the constrained water levels. Secondly, this thesis provides an alternative source to predictions of water levels. This thesis is structured as follows: Section 2 covers the theoretical aspects of the thesis. Section 3 studies the existing literature around AIS data and water level prediction. Section 4 presents our data sources, assumptions, and data transformation. Section 5 presents our final model results, accuracy measures and a profit estimation example. Finally, section 6 summarizes our findings and limitations to our research.

2. Literature review

In accordance with technological development, the Automatic Identification System's substantial information on vessels has grown in popularity over the last years. Many recent studies have focused on how AIS data can be utilized to increase profits, by either creating new opportunities or reducing uncertainties. There are several articles that discuss how AIS data is used for estimation or forecasting, however, little research has measured river depth by using a combination of reported draft from AIS data and weather data.

2.1 AIS as a data source

As mentioned, AIS is a large database of digitalized vessel data, created to strengthen safety and control in maritime traffic. However, several input variables in the system must be entered manually by the crew, such as draft. This creates a base for human error. Research by Harati-Mokhtari et. al. (2007) defines the errors as "Frontline Operator failure", which means the crew can forget, are inattentive or can omit actions to enter the data. The study finds that 17% of AIS reported draft are either not available or contain zeros. Furthermore, it was observed that 14% of the reported draft was greater than the length of the vessels. Thus, 69.5% of the data remains. However, they were unable to verify whether the rest of the data was inaccurate or not, which indicates that the data can be imprecise (Harati-Mokhtari, et. al., 2007).

On the other hand, this issue was also discussed by Jia et. al. (2019), in trying to use AIS reported draft as a basis for estimating cargo payload. The research compared the data with other sources, such as port agent reports, which may be more precise in terms of manually recorded vessel information. However, the reports are less appealing in comparison to AIS data in terms of availability, scope, and timeliness. As a result of the discussion, the AIS reported draft was determined to be useful, even though it contained errors (Jia, et. al., 2019).

As we want to create a tool that can influence the profitability of a ship voyage, it can be interesting to consider various variables in the AIS data that could have a significant impact. A study by (Wu, et. al., 2020) explores eleven months of AIS data to estimate the voyage duration in the HESC channel, as it affects the profitability of a ship voyage. The study discovered that neither the time of day nor month had significant influence on the travel time. However, the result showed that the draft in particular had a significant impact on the voyage duration (Wu, et. al., 2020). Thus, it seems that calculating draft using a high-frequency source of data, such as AIS, can contribute to the calculation of cargo trades in shallow waters such as Paraná River.

2.2 Predication of water levels

The water level on a ship voyage can be a critical feature for shipping companies that pass through shallow waters to reach their destination. If it is too shallow, the ship will not be able to load optimally or even pass the water level barrier, resulting in wrongful pricing of cargo trades. Therefore, critically shallow points in rivers and channels are crucial to observe, but difficult to predict due to factors such as lack of data. However, several studies have attempted to estimate water levels in rivers and channels using complex predictive methods and different types of variables.

In a similar study, Lejeune (2020) used Long Short-Term Memory (LSTM) to forecast next-day water levels at Kaub, a critical point in the Rhine River in Germany. The dataset included 15 different independent variables, where 12 of them were weather-related. There were 3 precipitation variables that represented the average amount of rain in 3 influential geographical areas. The remaining variables were water flow connected to the different areas and the associated minimum, maximum, and median values of temperature. The fitted model obtained a root mean squared error (RMSE) of 6.2 centimeters after fitting, in other words, an average prediction error of 6.2 cm in water level (Lejeune, 2020). This result is very precise and shows that the chosen variables can explain a high percentage of the variance in water levels. However, as the time frame of the forecast is day-to-day, the performance measures should not be compared directly with long-term models, such as ours.

There are also studies that attempt to predict water levels with other variables than weather-related variables. Bazartseren et. al. (2003) aimed to find a short-term prediction of water levels in two different rivers in Germany, by avoiding the use of historical or future values of precipitation depth. The study rather focused on measuring the change in water levels upstream, and the approximate travel time it takes for the change to reach the mouth of the river. To accomplish this, he used artificial neural networks and neuro-fuzzy systems. Bazartseren et. al. (2003) suggests this method only when precipitation depth data and forecasts are scarce or not available.

However, none of the studies described above focuses on utilizing weather data to predict water levels, indirectly by using AIS reported draft. The goal of this research is to create a robust long-term model, based on a large scale of available, high-frequency data, that can be used to derive a future draft constraint. A model such as this would enable shipping companies to accomplish more accurate cargo pricing.

3. Theory

In this section, we will explain time series applied to our data and how the final model can be trained by using a time-based cross-validation. Secondly, we intend to explain how different simple and interpretable prediction methods, such as linear and multiple regression and generalized additive models, can be applied to the predictive model. Finally, we explain how various accuracy measures can be used and compare them to find the best prediction model.

3.1 Time series and cross-validation

3.1.1 Time series

As both weather variables and water levels fluctuate over time, we use time series to observe changes. A time series is defined as a series of data that occur in successive time order. The objective of using time series is to discover statistical features and find a suitable statistical model of the raw data, to estimate future values of the series. There are four different components of variations in time series: seasonal variation, trend, cyclical variation, and irregular fluctuations. An important feature in using time series for our data set is to remove seasonal fluctuations to find trends. This makes it easier to view the relationship between the variables (Chatfield, 2000). Due to these conditions, we would like to focus on these three components, described below:

(1) *Seasonal variation*. The type of variation that occurs in annual periods for different series, for a specific time of the year. This variation can be observed in a series measured weekly, monthly, or quarterly (Chatfield, 2000).
(2) *Trend*. We can observe a trend when a series shows a steady decline or incline of observations over several time periods. The perception of the trend will often be determined by the duration of the time series, as there is no fully satisfying mathematic definition (Chatfield, 2000).
(3) *Irregular fluctuations.* This is often described as any variation left after removing seasonality, trend, and cyclical variation. Variation of such type could often be completely random and is therefore impossible to forecast (Chatfield, 2000).

As mentioned, we would like to find the trend in the data. A moving average can act as a filtering and smoothing of the data, by removing random variation. They are commonly used to identify the trend direction of a stock but can also be used as a trend-following or lagging indicator for instance for precipitation, as it is based on past data (Fernando, 2021). A moving average can also estimate and remove seasonality patterns to reveal trends, by setting the length of the moving average equal to the length of the seasonal pattern. As revealed in section 4.1.3, the weather variables reveal a yearly seasonality pattern. In addition, the impact of random or short-term changes in precipitation will be mitigated over a longer time frame. In other words, the higher the moving average time frame, the less sensitive it will be to changes, such as longer periods of low rainfall or high concentration of rain. By using a moving average on the independent weather variables, we can account for the annual seasonality and its effect on water levels.

3.1.2 Time-based cross-validation

The data set is divided into a training and test set, which will be used to evaluate the forecasting model. We will use a time-based splitting to consider the time-related characteristics in the data, making the time-based cross-validation method useful. This form of model validation requires the training of the data to be earlier in time than the testing of the data. In other words, the test set cannot go back in time and the training set can never include future observations. In our model, the test set will always contain a default number of observations, while the training set will increase in size as the time-series moves forward. Both the training and test set are therefore based on a rolling forecasting window (Herman-Saffar, 2020). How this applies to our model is further discussed in 5.1.1.

3.2 Regression models

3.2.1 Linear and multiple regression model

Linear regression models estimate the relationship between a dependent and independent variable and are often used for prediction and forecasting. The formula is structured as follows:

$$Y_i = \beta_0 + \beta_1 X_1$$

where Y_1 is the dependent variable, and X_1 is the independent variable (Yale University, 1993). The sign before the β_1 signifies whether the relationship between the variables is positive or negative, and β_0 indicates where the regression line crosses the y-axis.

Multiple regression is an extension of linear regression, including multiple independent variables. The objective is to use significant independent variables, with known values, predicting the unknown value of the single dependent value (Wang, et. al., 2020). The formula is defined as:

$$Y_i = \beta_0 + \beta_1 X_{i1} \ldots + \beta_n X_{in} + \epsilon$$

where β_n are the slope coefficients for each independent variable, and ϵ equals the model's error term (Hayes, 2021).

3.2.2 Generalized Additive Model (GAM)

Hastie & Tibshirani (1986) created the generalized additive model (GAM), which is a model that does not assume any specific form of relationship between the dependent and independent variables. Unlike traditional generalized linear models (GLM), GAM does not require a relationship to be a simple weighted sum. Instead, it assumes the relationship is a sum of arbitrary functions of each variable (Shafi, 2021). The model can be structured as:

$$g(\mu_i) = A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i})$$

where $g(\mu_i)$ is defined as the expected value of the response variable Y_i, and $Y_i \sim EF(\mu_i, \phi)$, meaning that the model has an exponential family distribution by an average of μ_i and a scale parameter ϕ. The first vector, A_i, denotes the explanatory variables for strictly parametric components, the corresponding parameter vector is θ, and the smooth functions of the covariates x_k is defined as f_j (Hastie & Tibshirani, 1986).

The model uses flexible functions called splines, and the sum of these splines form the curve of a GAM. This makes the model able to form a non-linear relationship. A GAM model can fit a large number of splines in order to fit the curve, however, one must be aware of the danger of overfitting the problem. If the curve is overfitted, it will not perform well with forecasting (Shafi, 2021).

3.3 Evalution indices

The evaluation of a prediction model is an important part in its development as it determines which model is most accurate and reliable (González-Sopeña, et. al, 2021). To examine the relative performance of the chosen model, prediction values are compared to actual values. Four different measures are used to assess the forecast accuracy: R^2, mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). It is advantageous to examine different measures when comparing the models in order to obtain a more nuanced basis for model evaluation.

3.3.1 R²

The parameter R^2 is often known as the "coefficient of determination", as it represents how much of the variance in the dependent model that is explained by the independent variables. The function consists of the sum of squares of regression (SSR) to the total sum of squares (SST). In other words, it takes the portion of variation explained by the regression, to the total variation in the dependent variable (Naghshpour, 2016).

$$R^2 = \frac{SSR}{SST}$$

The R^2 output ranges between zero and one, where higher values indicate higher explanatory power of the model, keeping other things equal.

However, in some cases, a higher R^2 does not always imply a better fit. R^2 has one flaw: it does not account for the number of variables used to calculate it. Hence R^2 rises as the number of variables in the model increases. Even if certain factors do not contribute to the explanation of a variable, it can contribute to increasing the measure value (Naghshpour, 2016).

3.3.2 MSE and RMSE

The mean squared error (MSE) and root mean squared error (RMSE) are defined as:

$$MSE = \frac{1}{N}\Sigma\left(Y_i - \widehat{Y}_i\right)^2$$

$$RMSE = \sqrt{\frac{1}{N}\Sigma\left(Y_i - \widehat{Y}_i\right)^2}$$

MSE subtracts the observed value from the predicted value, squares it, and finally calculates the mean. In other words, it measures the prediction error by the model. RMSE measures the average of the squares of the errors. By squaring the root, the function penalizes large values more than small. This is desirable in our thesis as high differences between prediction and actual value is critical, as wrong measures could result in a higher economic cost for the shipping company (Chai & Draxler, 2014). The lower the MSE and RMSE value, the more accurate the model performance is.

3.3.3 MAPE

Another way to illustrate the accuracy of the model, is to use the mean absolute percentage error (MAPE).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right|$$

This can be calculated by subtracting the actual value from the predicted value and dividing it by the predicted value. Then, the absolute value is taken of the result, and divided by the number of observations. MAPE is commonly used in regression problems, as the result is intuitive in terms of relative error. The accuracy measure has been suggested to be well-suited for forecasting applications, particularly when large amounts of data is available. The smaller the mean absolute percentage error, the better the forecast (De Myttenaere, et. al., 2016).

The combination of these four measures creates an accuracy basis for model evaluation.

3.4 Correlation

Correlation is a measure of dependency between variables, or whether past values are related with present values within same variable. Test of correlation can assess the credibility of variables in the same model by finding what influences change the variables, making it useful when choosing the composition of variables in a model.

3.4.1 ACF

The autocorrelation function (ACF) is a measure of how data in a time series are related, on average, to the prior data points (Song & Esogbue, 2006). The function makes a plot with k numbers of lags, visualizing a series of correlation in the data that changes over time. An autocorrelation function at lag k, expressed as ρ_k, can be structed as follows:

$$\rho_k = \frac{\Upsilon_k}{\Upsilon_0}$$

where

$$\Upsilon_k = cov(Y_i, Y_{i+k})$$

for any i

Υ_0 denotes the variance of the stochastic process. We can use the ACF plot to tentatively identify seasonal variation by detecting whether past data are related to present data. The relationship can be identified by looking at the spikes at each lag k to decide whether they are significant. A spike is significant if it extends the visualized significant limit on the graph (Song & Esogbue, 2006).

3.4.2 Pearson correlation test

The Pearson correlation test measures a linear dependence between two continuous variables and can only be used when the variables are normally distributed. The test is known as the most common method for measuring the relationship between two variables. The correlation coefficient, often denoted as r, can be calculated for two quantities X and Y on each of N individuals as

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

$$r = \frac{\sum_{i=1}^{N}\{(X_i - \overline{X})(Y_i - \overline{Y})\}}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}}$$

The Pearson correlation coefficient is also known to provide information on the intensity of correlation, as well as if the relationship is positive or negative. If the correlation coefficient is zero, there is no correlation. There is perfect correlation if the coefficient is one (Puth, et. al., 2014).

4. Data and methods

In this section, we will start by presenting our data sources and assumptions. Then, the final selection of input variables from the different sources is shown. At last, pre-processing and transformation of the data is explained.

4.1 Data sources and assumptions

The first section is structured into four separate parts, depending on the data source. Our sources of data are reported water levels from Western Bulk, AIS data, external weather data and weather forecasts.

4.1.1 Reported draft

Western Bulk has provided draft statistics from January 2009 to August 2021 obtained through a Paraná pilot company. The data shows the minimum depths in the river from San Lorenzo to San Pedro for the 15th of each month, i.e. monthly draft data. Statistics for previous years are not available, due to the fact that there was no digitization of the information in the official authorities or pilot companies.

The data has 152 observations (months) of reported water levels. Although this is a good place to start for prediction, water levels in both rivers and lakes may fluctuate naturally in the short term, seasonally and long term (Gibb, 2015). In a similar research study, Lejeune (2020) used 20 years of daily data to predict water levels in the Rhine River. In order to find any potential water level fluctuations or cycles in the Paraná River, it would be interesting to look at a higher resolution data set, as it is likely this could generate a better outcome. In this research, we want to look at the possibility of using reported draft from AIS data along with weather data from external sources to predict future water levels.

Descriptive statistics

Figure 1 illustrates the fluctuation in river water levels in the Paraná River over time. River levels were at their highest in 2016, and at their lowest in 2009 and 2021. Depth seems to change throughout the year. The red trend line is calculated using a two-sided 12-month moving average, where each side is 6 consecutive months. We can see that the water levels have been decreasing the past few years, in accordance with on-going drought in the Paraná River.

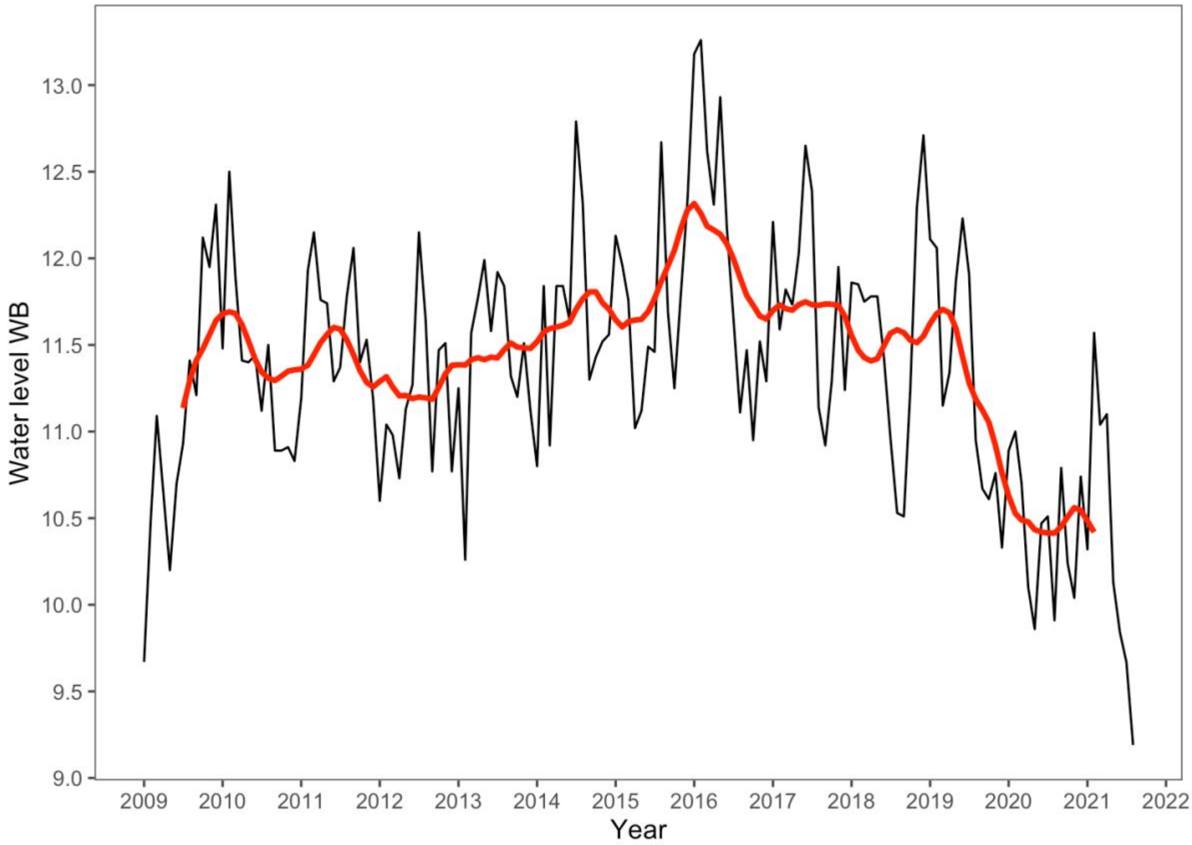

Figure 1: Data points for monthly river levels in the Paraná River (2009-2021), including a two sided 12-month moving average trend component (in red).

4.1.2 AIS data

In order to achieve a higher resolution of water levels over time, we have retrieved AIS data from Vesseltracker.com. AIS provides information about a ships route, such as unique identification number, position, destination, course, speed, draft, and much more.

The idea is to extract the reported draft from every ship sailing in the defined fairway and find the maximum reported value for each week. The maximum value will be the maximum reported draft any ship has had in order to pass the water level barrier. The maximum value will signify two things:

(1) When the river water levels are low, the safety margins for the ships will be small and the risks higher for ship grounding. The value will therefore provide a picture of the maximum recommended draft by the pilot companies.

(2) When the river water levels are high, the safety margins will be greater, and it will be safer to pass through the water level barrier. Therefore, the reported draft by each ship will not be representative of the actual water level, solely the maximum draft reached for each time frame.

The goal is to create a thorough and trustworthy function to calculate draft over time, with a prediction time frame of 2 weeks to 3 months ahead. However, one must have in mind the economic aspect of our model and the attitude of other shipping companies, which will affect the maximum reported draft each week.

- Some loads have a high stowage factor so it will be fully loaded per cubic meter before it is limited in depth. In other words, the ship is not necessarily fully loaded up to the maximum allowable draft in the river. If all ships during a week have a high stowage factor, the maximum draft reported will not represent the actual allowable draft level, but instead the maximum draft reached the relevant week.

- Sometimes the market is such that the larger ships (Supramax / Ultramax 50-63,000 dwt) get lower market rates than the smaller ones (Handysize 28-40,000 dwt). This means that it may pay off to bring in larger ships to load a smaller cargo. Then, the ship is not loaded to maximum load capacity, and may not maximize the draft restrictions.

- Vessels occasionally go upriver to load at a terminal before sailing to a port outside the area, to make a "top-off" of cargo. This is usually Necochea or Bahia Blanca, where there are no depth restrictions. There it is possible to load the ships all the way up to maximum capacity. Again, if all ships during a week do not load up to maximum draft that is allowed in the river, the maximum reported draft they report will not be representative for the actual water level.

If some shipping companies have an incentive to not load up to maximum draft in the river, either by having larger ships than necessary, larger stowage factor, or an opportunity to top-off at a location downriver of the shallowest points, it may lead to the reported draft the relevant week to be misleading.

Data assumptions

The data provided contains dates from 01.01.2013 to 31.12.2020. The area we are interested in gathering information about, is the mouth of the sailing fairway up to San Lorenzo port terminals. This is the area Western Bulk sends their ships to load and transport grain, along with many other shipping companies. Therefore, we have filtered the AIS data to only include an area of a specific set of coordinates: **longitude = (-60.89, -58.56); latitude = (-34.32, -32.57)**. The defined area is portrayed in Figure 2.

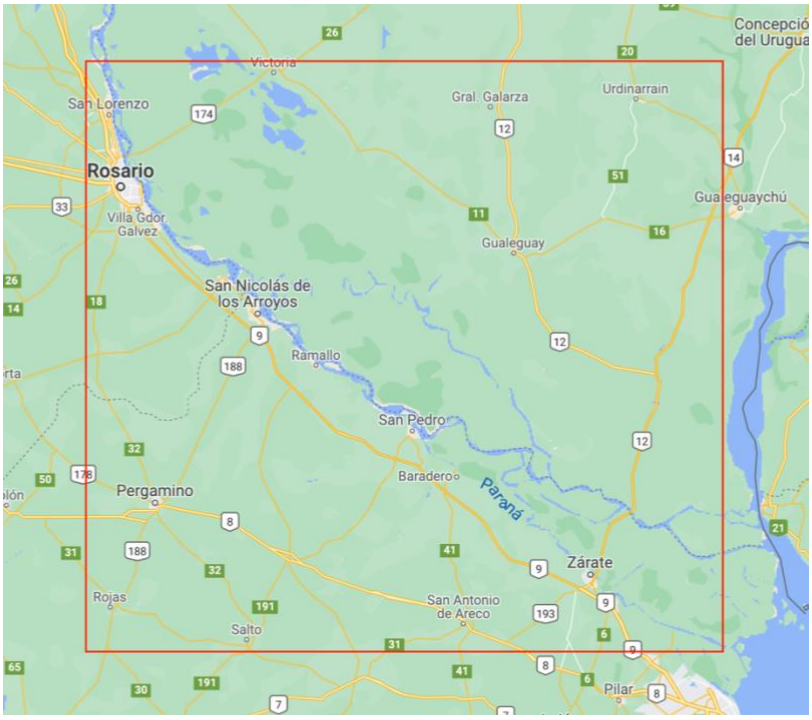

Figure 2: Map of defined area of the river.

Using AIS data does not come without potential sources of error. The data registered by each ship can be uncertain or erratic. The system can easily be turned off, manipulated, spoofed, etc. In areas with heavy traffic, the signal of the ship may disappear. AIS data is not standardized, and each captain may have their own way of registering information on their whereabouts. This makes the AIS system prone to human error.

In order to account for some of the human errors and faulty components of the AIS system, we have taken the following assumptions:

(1) Some captains do not update their draft level continuously or after loading the cargo into the ship. This can affect our model in the sense that the water level is presented as lower than it actually is after loading the ship. To account for this, we register all drafts reported in AIS from the ship leaves the terminal upriver until their next destination. Then, we find the maximum reported draft on the ship voyage.

(2) Some dates or weeks have very few or no observations, which either cause holes in the data or lead to a wrong output of the data. For example, if a smaller ship with a lower maximum draft limit than the typical Supramax or Handysize is the only observation, the reported draft would be misleading relative to the actual water level. To account for this, we are looking at the maximum draft on a weekly level, as it is not guaranteed that daily data would be sufficient. If outliers (relatively low or high reported levels) are detected, these are adjusted accordingly.

(3) In addition, any draft levels in the defined area that are below 8 and above 14 meters are considered faulty observations and are therefore filtered out of the data. This margin is based on the minimum and maximum values of the benchmark data. This will remove any observations of ships with a design draft that will never reach the constraint even at low river water levels.

(4) All ships that sail across the water level barrier are included, regardless of whether they are going upriver with an empty cargo, their stowage factor, ship size or incentives to top-off at other destinations. The maximum draft variable will filter out all vessels that are going upriver with mainly empty cargoes.

Data transformation

From 2013 to 2020, the AIS data consisted of 399.6 GB, which was filtered in order to obtain a final data set consisting of 394.6 MB and 2 411 541 observations, containing data on 3953 unique vessels. In order to download the data into RStudio, we used an application called SQLite. This program permits a single database connection to access several files simultaneously, as well as it handles large data amounts. The data was limited by first identifying vessels that have appeared in the defined area of the river at any point during the period. As previously stated, draft may not be adjusted or reported after cargo loading or unloading. To adjust for this inaccuracy, the unique IMO number of vessels in the relevant area was utilized as an indicator to extract solely the relevant vessels from the data set. Further on, when a vessel departed the river, a binary algorithm marked the end of the voyage. All rows including NAs or reported draft of 0 were eliminated to filter out wrong information.

The filtering through SQLite and RStudio leaves a complete data frame of the following observations, shown in Figure 3 to the left below as the black dots. In this map, we observe that there are some outlier observations that are registered outside of the Paraná River. These outliers may exist because a vessel has sailed through another river. Outlier data may alter or influence the outcome, which can make prediction models inaccurate. As we are only interested in the ships that are in the Paraná River, and no other rivers or lakes, we must filter out any outliers. We identified the outliers by creating a separate data set, and rounded the reported longitudes and latitudes by 0.01, followed by creating indexes for each position. Then, by using the indexes, we find the frequency of vessels per area. We assume that if there have been less than 10 ships in an area, the area is considered an outlier, and the related observations are therefore removed. There were 100 locations that had less than a total of ten vessels visited since the observations were started in 2013, resulting in 309 observations defined as outliers. Figure 3 to the right below shows the defined area with no outliers, as well as the other filtering actions mentioned above.

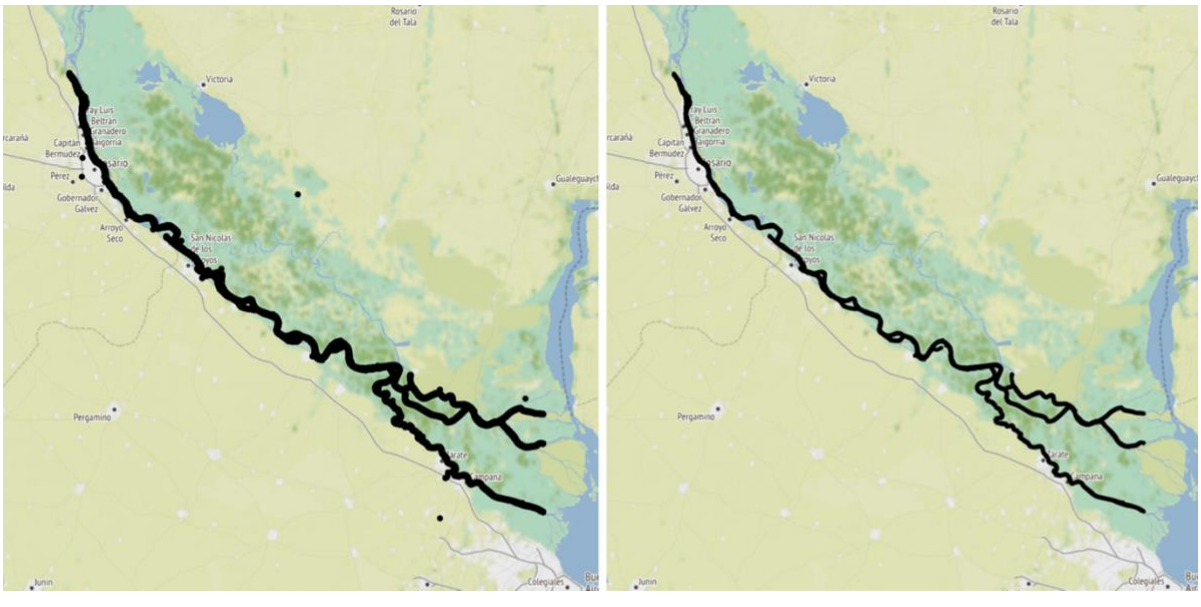

Figure 3: The filtering through SQL and R and observations shown by the defined area in the map, with and without outliers.

After deleting the outliers, the maximum draft was found per IMO number, time frame, and destination. The correct maximum draft was now possible to calculate by only including relevant voyages, as well as including updated draft from crew who submitted the data at the end of a voyage in a location outside the relevant area. All coordinates that did not fall within the prescribed area were removed. At last, we have a data set where all week numbers from 2013 to 2020 contain a maximum allowable draft.

As mentioned, even if reported draft levels below 8 and above 14 are removed, there may be some faulty reports caused by human errors. 21 out of 415 weekly observations from AIS are considered outliers by an automatic time series function and are adjusted accordingly. The difference between reported draft in AIS with and without outliers are shown in Figure 4. The red trend line is calculated using a two-sided 52-week moving average, where each side is 26 consecutive weeks.

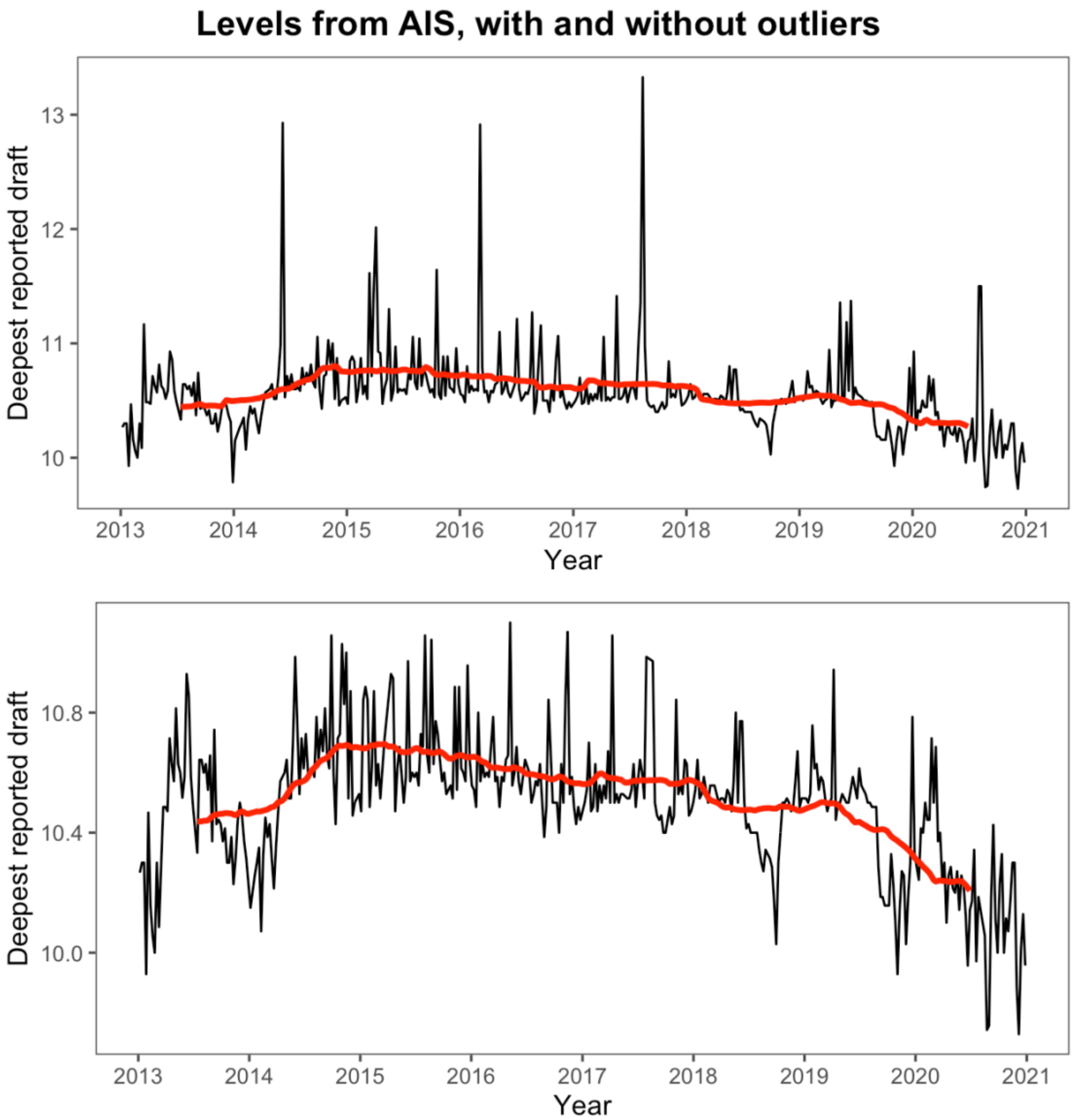

Figure 4: Data points for maximum draft reported per week in AIS (2013-2020), with and without adjusted potential outliers. Data points for draft over time, including a two sided 52-week moving average trend component (in red).

At last, the difference between the reported deepest draft per week from AIS and the reported draft levels from Western Bulk are presented in Figure 5. When the AIS levels are at their highest, there is usually a higher margin between the two variables, and when the AIS levels are at their lowest, the lines seem to overlap.

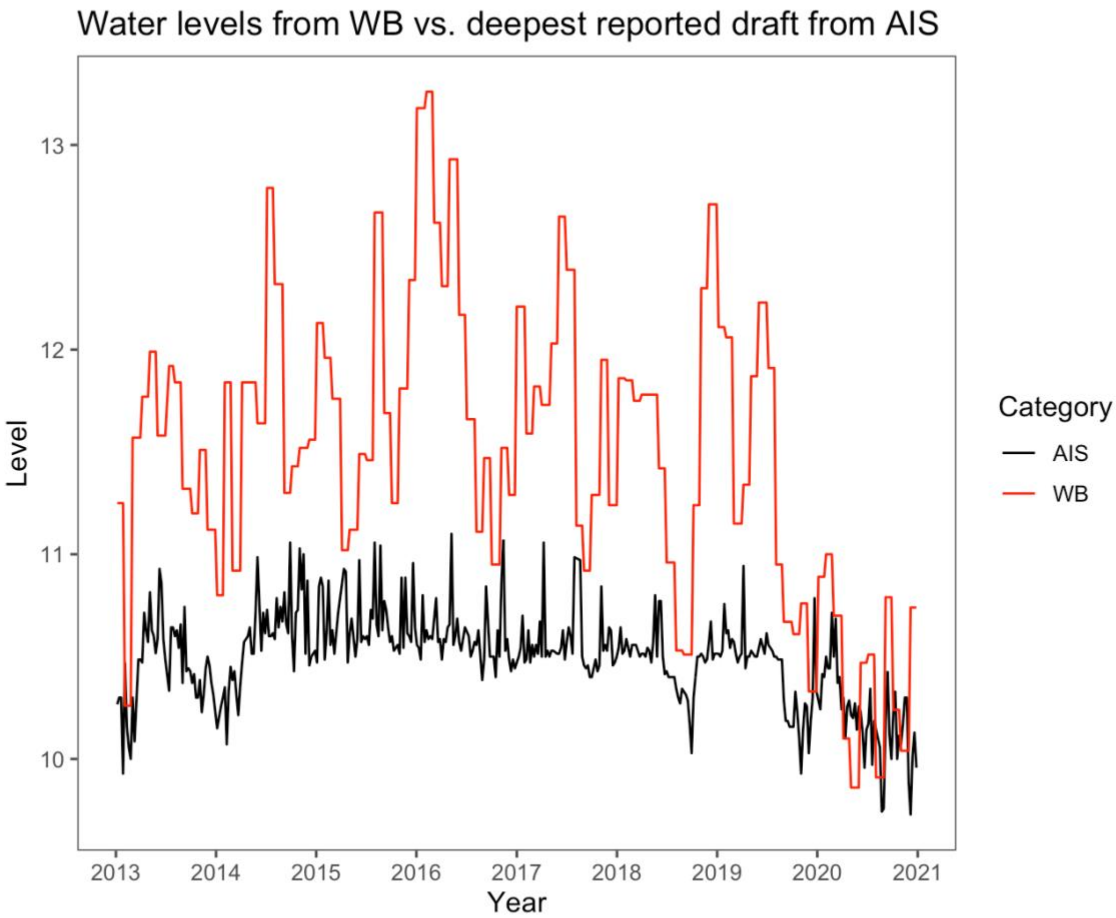

Figure 5: Water levels from Western Bulk (in red) versus the deepest reported draft from AIS data (in black) from 2013 to 2020.

4.1.3 Weather data

In our paper, we would like to research how weather variables affect water levels. Table 1 summarizes the variables we have investigated and their data source.

Dataset (Source)	Temporal resolution (Resolution transformed to)	Spatial resolution (Resolution transformed to)	Temporal Coverage (Period extracted)	Variables Used	Units
ERA5 hourly data on single levels from 1979 to present	1 hour (1 week)	0.25° x 0.25° (1.00° x 1.00°)	1979-01-01 – Present (1993-01-01 – 2021-15-11)	Total precipitation (tp)	m
				Surface runoff (sro)	m

Table 1: The variables are extracted from Copernicus Climate Data Store (CDS).

Data assumptions

The weather variables include total precipitation and surface runoff. Total precipitation is the accumulated frozen and liquid water, rain, and snow, that falls on the surface to Earth. Surface runoff is when the precipitation or melting snow that is stored in the soil, drains out. The unit of the variables is the depth in meters the water would have if it were spread evenly over a model grid box.

The data of the weather variables were collected for the region: **longitude = (-65, -50); latitude = (-35, -20)**, on a 1-degree spatial resolution level. This region includes the mouth of the river, which is Río de la Plata, and the source of the river, the confluence of Paranaíba River and Río Grande. It also includes the western and eastern borders of the river. The defined area for the weather variables is shown in Figure 6.

Figure 6: Map of defined area for the weather variables.

We have used the original temporal resolution of hourly accumulations, limited to 1993-2021. However, as we want the data on a weekly level to fit the draft data from AIS, we have taken the mean hourly value across all coordinates, then across all weeks. In order to compare the results from AIS water levels with WB water levels, we create another data set on a *monthly* level to fit the draft data from Western Bulk, and take the mean hourly value across all coordinates, then across all *months*.

Weather variables have been limited to the variable's precipitation and surface runoff, due to data size limitations. However, one should have in mind that these variables could potentially be highly correlated with one another. We will check and discuss the correlations between our chosen weather variables in the next section.

Data exploration and transformation

As mentioned above, our weather variables are potentially highly correlated with one another, which could have an impact on the resulting prediction model. We would therefore like to check how large the correlation is.

From Figure 7, we can see that weekly precipitation and runoff are highly correlated with a correlation value of R=0.91. This means that the relative change in weekly precipitation is more or likely the same as the relative change in weekly runoff.

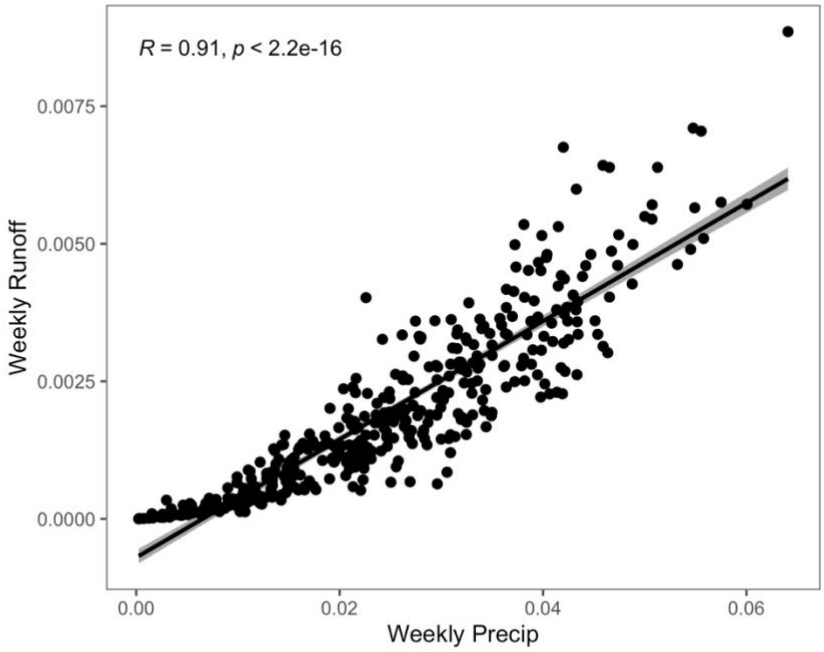

Figure 7: Correlation between weekly precipitation and runoff.

Exploring the weather variables can be useful to reveal seasonal patterns and show changes in these patterns over time. Since precipitation and runoff have the same movements, precipitation will be the focus in this part. All graphs are shown for runoff in appendix A. A useful way of showing a seasonal pattern in precipitation data is to illustrate the data points and mean for each month. Figure 8 presents a seasonal pattern in precipitation data, clearly showing a higher level of rain between October and February, and lower levels of rain between March and September. The lowest levels are observed in July and August, which could mean these months are more predisposed to lower water levels or drought.

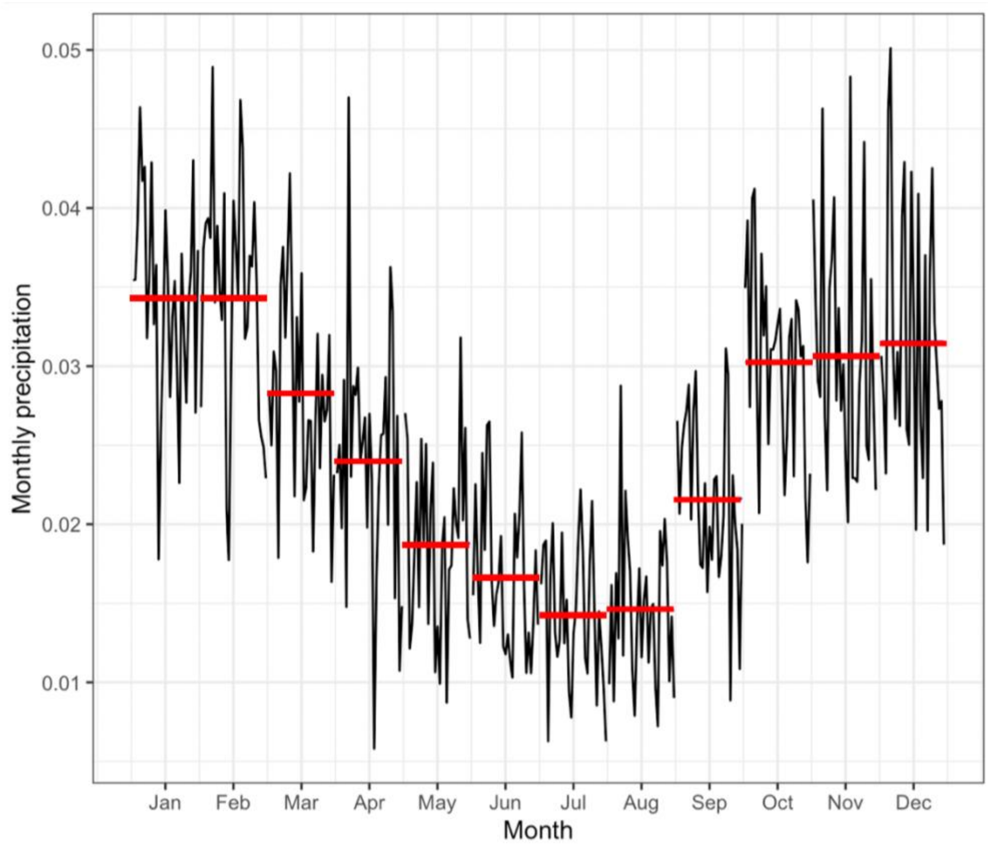

Figure 8: Seasonal pattern of monthly precipitation mean (from 1993 to 2021), where the red horizontal line represents the accumulated mean for each month.

We can observe the same pattern in the polar seasonal plot in Figure 9. Both plots illustrate a sharp decrease in precipitation during the drier months (March to August). These months are controlled at the lower levels of precipitation, with some extreme observations of rainfall some years.

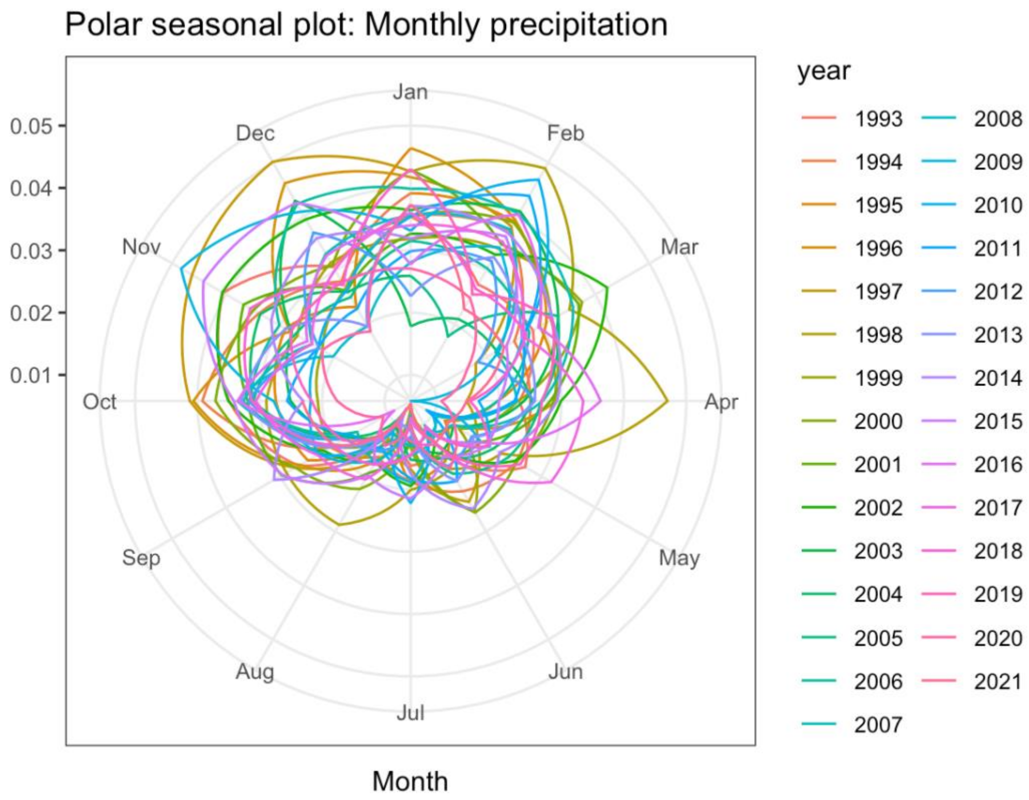

Figure 9: Polar seasonal plot for monthly precipitation data (from 1993 to 2021).

An autocorrelation plot can prove these yearly trends, shown in Figure 10. We can see that the maximal autocorrelation occurs at a lag of 12, which makes sense as this is monthly data. The strongest correlated value for a particular month will be the same month in the previous year. Also, the stronger a trend is, the closer the values will be to one another in more recent observations, which explains why the significance for each 12th lag (i.e. year) sinks.

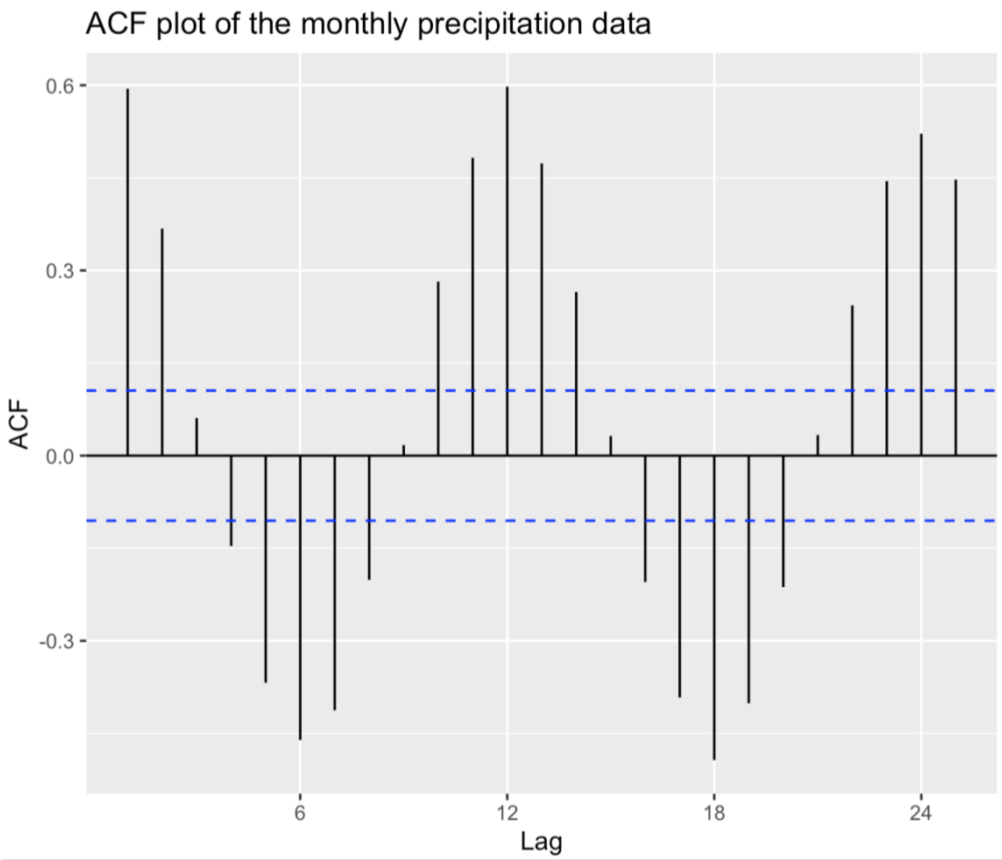

Figure 10: A compressed plot of the autocorrelation values for the first 25 lags.

Further, as there is a yearly trend present in the precipitation data, we will introduce a 52-week simple moving average precipitation variable. This independent variable will remove annual seasonality patterns to reveal the trends in the data. Figure 11 show a 52-week simple moving average from 1993-2021. We can see that the precipitation index is currently at its lowest since 1993, and the last period of lower relative levels of precipitation was between 2004 and 2009, which we do not have draft levels for.

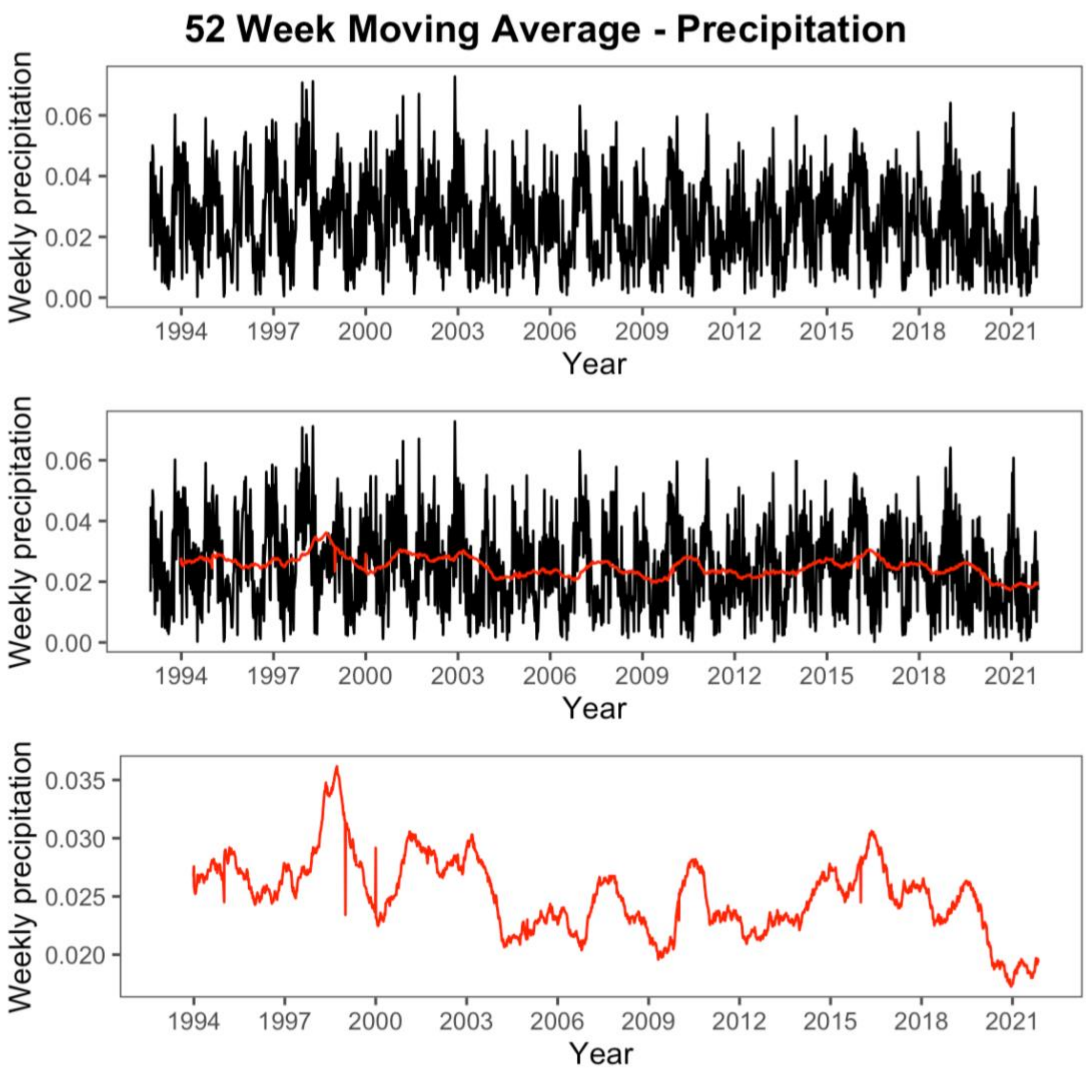

Figure 11: 52-week simple moving average of precipitation from 1993-2021.

4.1.4 Precipitation forecast

In order to add forecasts for the weekly level of precipitation to the model, we have obtained data from The European Centre for Medium Range Weather Forecasts (ECMWF). Each forecast contains 25-51 realizations of weather, which act as a probability distribution. The distribution aims to represent the uncertainty in weather. These realizations are usually postprocessed to get representative outputs. To match the forecasted data with the observed data, one needs to correct the output of the seasonal forecast to put it on the same scale as the output of the ERA5 data. For example, ERA5 contains the amount of water that falls on an exact grid point, while the systems forecast return the volume of water that fell around the grid point. We obtained the forecasts with the correct scale to match the weather data and used the forecasts in predicting water levels.

4.2 Final selection of input parameters

Table 2 shows the selection of all variables that are used in each data source.

	Parameter	Units
Draft reports		
	Month	-
	Year	-
	Minimum depths	m
AIS data		
	Imo number	-
	Timestamp position	datetime
	Destination	-
	Latitude	degrees
	Longitude	degrees
	Draft	m
Weather data		
	Total precipitation	m
	Surface runoff	m
	Time	datetime
	Latitude	degrees
	Longitude	degrees
Forecast data		
	received externally	

Table 2: Selection of variables in each data source.

4.3 Pre-processing and transformation

In this section we will explain the relationship between AIS data and the monthly draft data provided by Western Bulk, and the relationship between AIS data and weather variables.

4.3.1 Relationship between AIS data and monthly-data

In order to check if the movements of the reported water levels from Western Bulk and the reported maximum draft from AIS are associated, we use the Pearson correlation test. From Figure 12, we can see that the correlation between the variables is positive at 0.59. As the correlation is far from 1, the variables are not perfectly correlated, meaning the relative change in water levels is not the same. The correlation is shown as a linear function in the figure. The relationship between the variables is shown using a LOESS smoothing function (in red). We can see a kink in the curve between reported water levels of 11 and 12. This makes sense as at very high-water levels, no ships of the size that travel through the river would be constrained by the draft. Hence, the observed maximum draft from AIS will no longer change with the actual water depth.

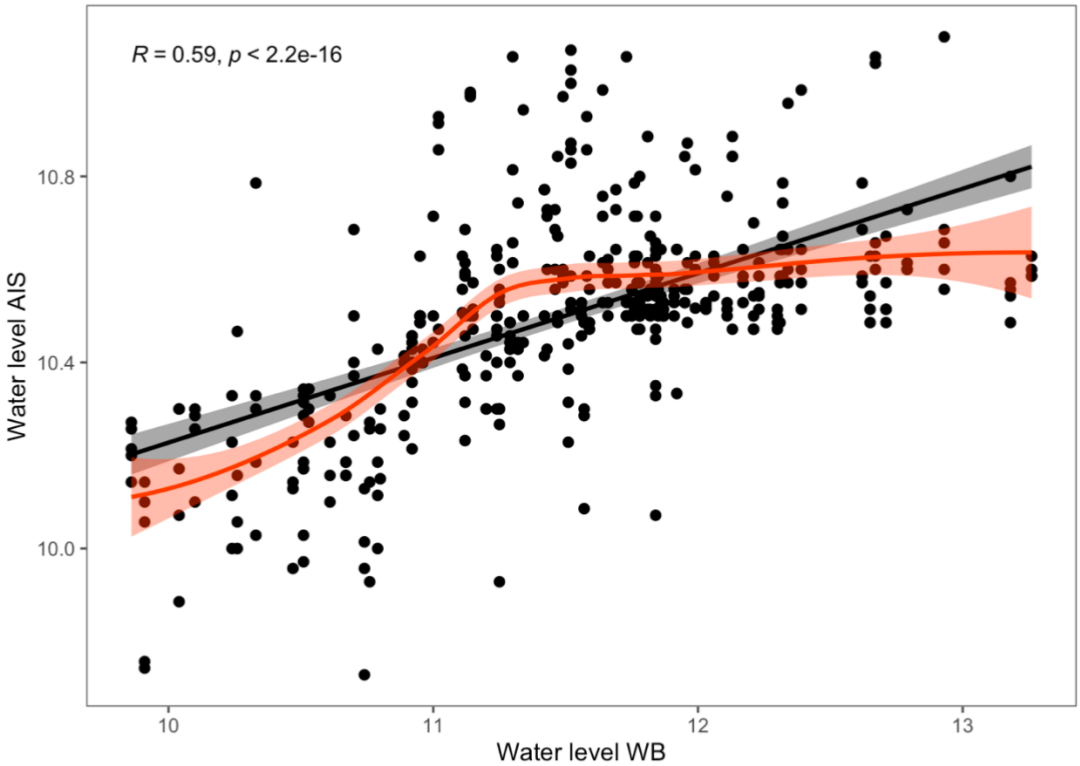

Figure 12: Correlation (in grey) and relationship (in red) between reported water levels from Western Bulk and reported maximum drafts from AIS between 2013 to 2020.

4.3.2 Relationship between AIS data and weather variables

Next, we would like to see if there is a relationship between weather data and reported maximum draft. According to the Pearson correlation test, there is no significant linear relationship between the weekly precipitation levels and maximum draft ($R = 0.083$). However, it is not given that maximum draft will increase at the exact moment of rainfall in one area.

As mentioned in section 3.1.1, moving averages are useful to estimate long-term trends in data. As the autocorrelation output proves, there is a yearly seasonal cycle in precipitation. In order to account for seasonal fluctuations throughout the year, we use a 52-week moving average. A 52-week simple moving average (SMA) has the highest correlation ($R = 0.6$), while the 52-week exponential moving average (EMA), where the nearby observations are given more importance, has a slightly lower correlation ($R = 0.55$). If we look at the correlation between AIS reported maximum draft and runoff, the correlation is lower than with precipitation with an SMA ($R = 0.49$) and with an EMA ($R = 0.52$). All the correlations with the dependent variable are summarized in Table 3, and shown graphically in appendix B:

Correlation	**Weekly Precipitation**	**52-week SMA precip**	**52-week EMA precip**	**52-week SMA runoff**	**52-week EMA runoff**
Maximum reported draft	0.083	0.6	0.55	0.49	0.52

Table 3: Pearson correlation test between potential independent weather variables and the dependent maximum reported draft variable.

Further, we will use a linear regression model and a generalized additive model to explore the relationship between water levels and the 52-week moving averages of precipitation. It is a straightforward model with one feature. For simplicity, the visual graphs will be included for the 52-week simple moving average for precipitation, shown in Figure 13.

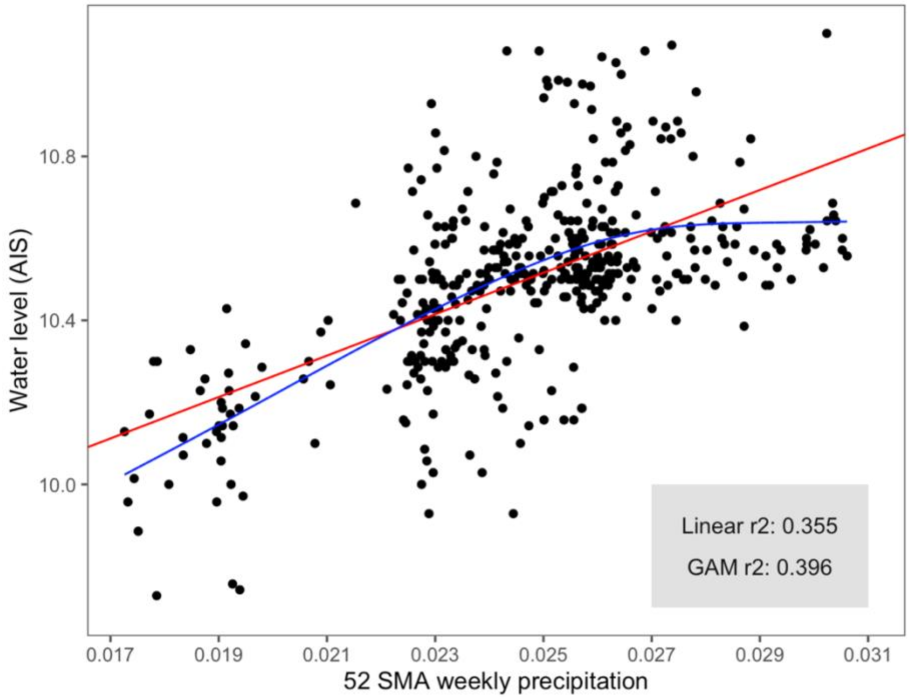

Figure 13: Relationship between 52-week moving average of precipitation for AIS reported water levels.

The linear regression model has an R^2 value of 0.355, which means that the 52-week moving average of precipitation explains 35.5% of the variance in water levels. Compared to the linear regression model, the R^2 value of the generalized additive model is improved by 4.1%. Table 4 is a result table for all the different models using AIS reported maximum draft as the dependent variable. The results are shown graphically in appendix C.

Result table using AIS reported maximum draft as the dependent variable							
		52-week SMA			**52-week EMA**		
		Precipitation	**Runoff**	**Both**	**Precipitation**	**Runoff**	**Both**
Model	Linear R^2	0.355	0.237	0.415	0.3	0.274	0.301
	GAM R^2	0.396	0.407	0.493	0.407	0.447	0.491

Table 4: Result table with the relationship between 52-week simple and exponential moving averages of precipitation and runoff for AIS reported maximum draft.

The same models are performed on water levels reported by Western Bulk, shown in appendix C. The weather data seem to perform slightly better on monthly data from Western Bulk, than reported maximum draft from AIS.

The GAM model performs better than the linear model in all cases. The EMA performs slightly better than the SMA on both weather variables, however, we will continue to perform analyses using both approaches.

Including both weather variables will increase the amount of variance that the model explains by 4.6%-9.7% and might make the prediction more accurate. However, as the GAMs R^2 is still only 50%, we will not include runoff in the final model. Having correlated variables often leads to more complications, and as runoff is less correlated with water levels, it is natural to look further at precipitation. We want to make the final model as simple as possible and will therefore observe the relationship between precipitation and water levels further.

5. Results and discussion

5.1 Results of model

The following section will describe the concept of model training and how it applies to our model, followed by a description of the model and its results. Thereafter we analyze the prediction accuracy of the model. Lastly, the model is compared to a prediction model using monthly water level data from Western Bulk.

5.1.1 Model training

To evaluate our forecasting model, we have used a time-based cross-validation with a training set and a test set, as described in section 3.1.2. The training set will continuously gain one more week for each new time step, and the test set will continuously remove the subsequent week and add a new week 12 time steps ahead of the current time step. The model is trained starting on week 1 in 2016, leaving 4 years in the first training set. The training ends in week 52 in 2020 as there is no longer any AIS reported draft data to make forecasts on or to compare accuracy measures to. The test set always contains 12 time steps (weeks), in a rolling window starting with weeks 2-13 in 2016. The training-test split is portrayed in the Figure 14.

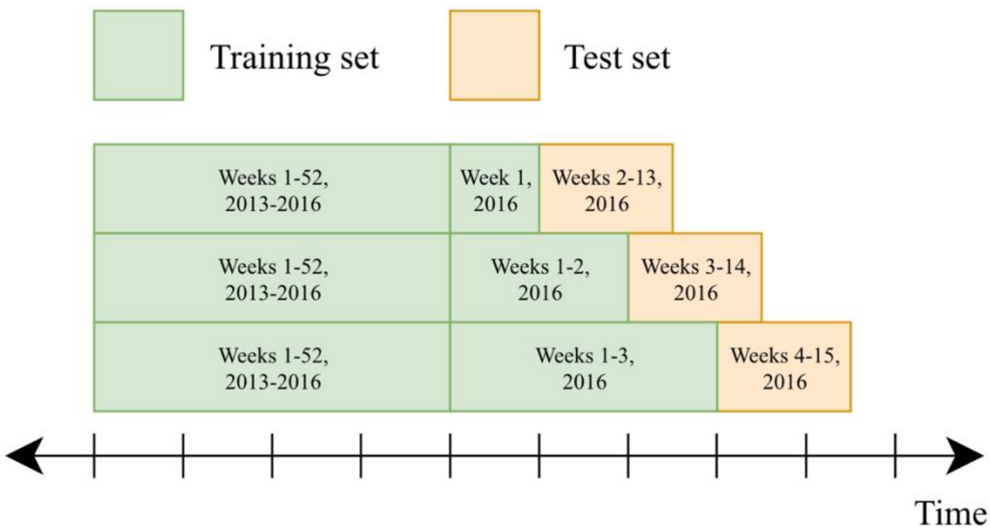

Figure 14: Train-test split for the first 3 time forecasts.

5.1.2 Model description and results

The final model uses weekly levels of AIS data from 2013 and moving averages of weekly levels of precipitation from 1993, as explained in section 4. We have used the generalized additive model with only one variable. The dependent variable is the maximum allowable draft, and the independent variable is a smooth function of the 52-week moving average of weekly levels of precipitation. We have two ways of calculating the moving average: a simple moving average (SMA) or an exponential moving average (EMA) and will check the accuracy measures for both scenarios.

The model predicts based on a data set of precipitation forecasts, whereas these are calculated as moving averages ahead of time, to match the independent variable. The model can predict up to 6 months ahead of time, which depends on how far ahead of time one has precipitation forecasts. As we have several realizations of weather forecasts, there are about 25 to 51 scenarios for each time step forecast, which leaves a distribution of potential draft levels for the applicable time step.

The forecast distribution can be portrayed using quantiles. Quantiles divides the range of distribution into subgroups. For example, by using interquartile range, you allow a distribution of the middle 50% of values. This range is less influenced by extreme values. If we assume a normal distribution, we can use a 90% prediction interval, which is defined by the 5th and 95th quantiles of the forecast distribution. The forecast distribution has also been adjusted for standard deviation, using their respective quantiles. In Figure 15, the full distribution, the 90% distribution and the interquartile distribution is shown using the simple moving average. The figures illustrate that you are in week 35 in year 2020, predicting 12 time steps (weeks) ahead. The dark purple illustrates the forecast distribution, the lighter purple illustrates the standard deviation distribution, and the black line represents the observed points for maximum allowable draft.

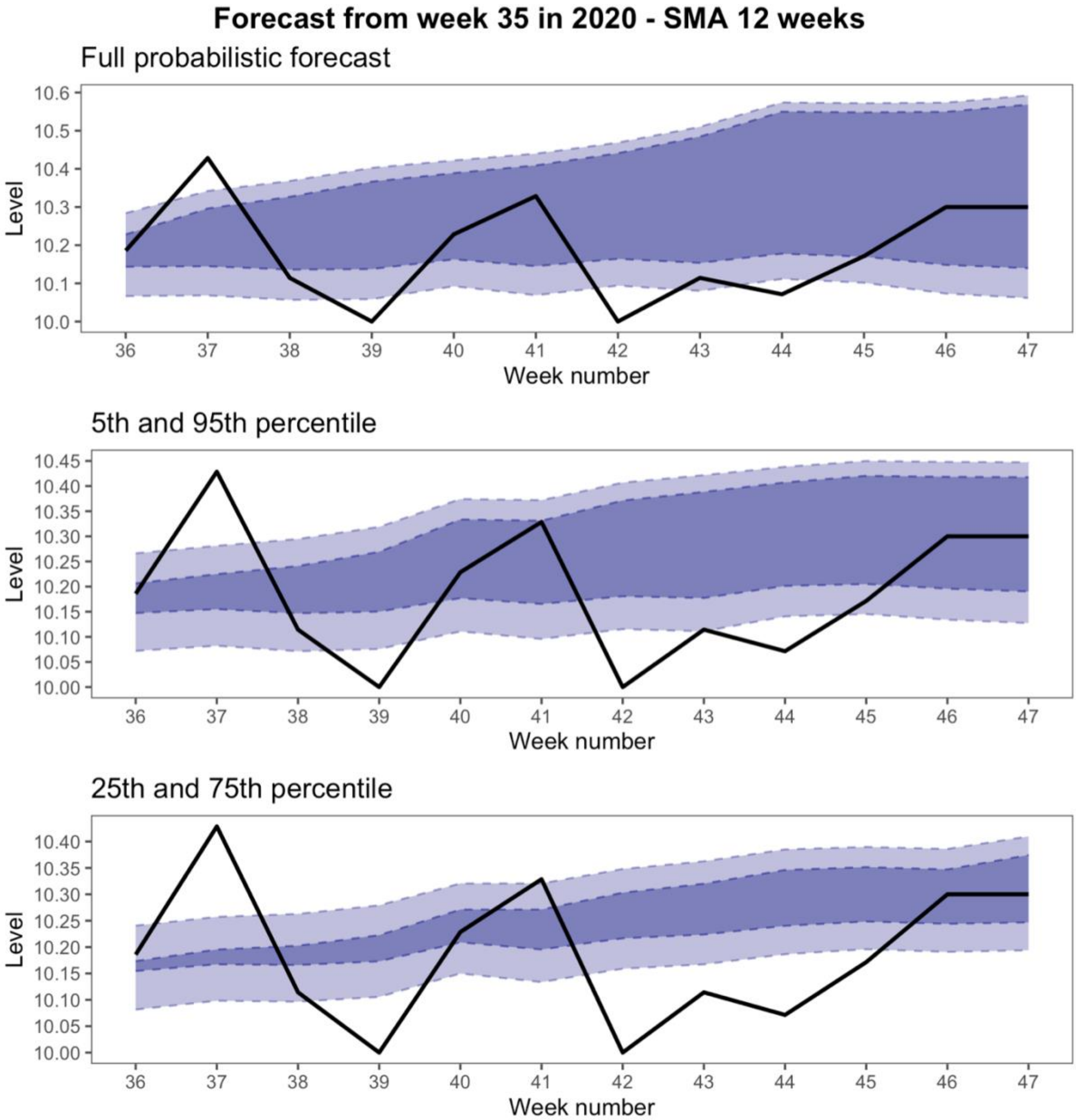

Figure 15: Forecast from week 35 in year 2020, predicting 12 weeks ahead, using the simple moving average of precipitation. Forecast distribution (in dark purple), standard deviation distribution (in light purple) and observed levels for draft (black line).

In appendix D, the full distribution, the 90% distribution and the interquartile distribution is shown using the exponential moving average.

5.1.3 Prediction accuracy analysis

After fitting the two models, we want to review the belonging prediction accuracy. In our analysis, we will utilize four different accuracy estimations on the selected models to choose the best model: R^2, MSE, RMSE and MAPE, described in section 3.3.

Table 5 shows the final result of the evaluation measures for the last available prediction. Both the model using the single moving average (SMA) variable and the model using the exponential moving average (EMA) variable have been reviewed on the measures. The goal is to find the best fitted and accurate prediction model.

Final values of accuracy measures for the AIS-based model		
Accuracy measure	**52 SMA**	**52 EMA**
R^2	0.389	0.397
MSE	0.077	0.321
RMSE	-0.265	-0.559
MAPE	0.026	0.053

Table 5: Final values of accuracy measures (rounded) of AIS-based model.

First, the table displays that the R^2 for the EMA slightly outperforms the SMA by having the highest value. In other words, the EMA model explains 39.7% of the variance in maximum draft, which is around 0.8% more than the SMA model. However, we can see the MSE is respectively 0.077 and 0.321 for SMA and EMA. The closer to 0 the MSE is, the closer the predictions are to the real values. The RMSE displays how concentrated the data is around the line of best fit, whereas the SMA model on average has a prediction error of 0.265 meters, while the EMA model on average has a prediction error of 0.559 meters. Finally, the MAPE value is respectively 0.026 and 0.053, which means the average difference between the forecasted value and actual value is 2.6% and 5.3%. The results from the MSE, RMSE and MAPE measurers indicate that the SMA model outperforms the EMA model.

As the model is trained from 2016 to 2020, we can see the development of the accuracy measures over time. Figure 16 shows the development for the model using the SMA variable, while the development for the model using the EMA variable is shown in appendix E.

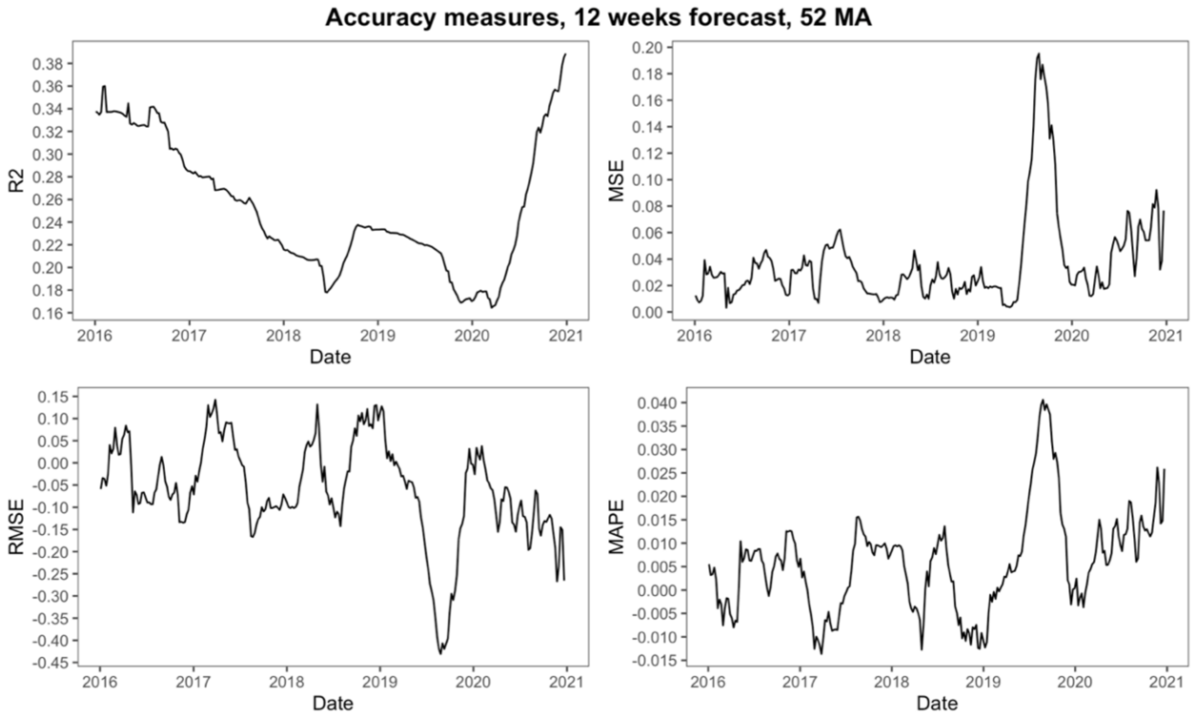

Figure 16: Accuracy measures for the model using the 52-week simple moving average (SMA) variable based on AIS reported draft.

The training of the model began in 2016, in which the reported maximum drafts had been steadily increasing since 2013. Precipitation levels were also relatively high compared to previous levels since 1993. The further variation in the measure accuracy could be due to variables that this thesis has not evaluated. For example, some sections of the river have been dredged, which can clarify why the variance is not explained by the relationship between precipitation and water level (Portal oficial del Estado argentine, n.d.b). Another example is the study by Lejeune (2020), which included water flow from three different areas to explain the river level.

5.1.4 Model comparison

Further, we will compare the AIS-based model with the model based on monthly reported draft data provided by Western bulk. Both models use the same forecast and historical data of precipitation over 12 weeks. However, the Western Bulk data contains monthly reported water levels, and not weekly reported maximum draft levels from vessels. The final values of the accuracy measures are displayed in Table 6.

Final values of accuracy measures (rounded)				
	Independent variable			
	Western Bulk (monthly)		**AIS (weekly)**	
Accuracy measure	**52 SMA**	**52 EMA**	**52 SMA**	**52 EMA**
R^2	0.318	0.399	0.389	0.397
MSE	0.499	0.228	0.077	0.321
RMSE	0.502	-0.191	-0.265	-0.559
MAPE	-0.048	0.017	0.026	0.053

Table 6: Final values of accuracy measures (rounded) of monthly data-based model.

First, in considering the monthly data from Western Bulk, we can see that the accuracy measures of model with the EMA precipitation variable dominates all accuracy measures of the model with the SMA precipitation variable. Thus, we will compare the two best models based on different draft data: The EMA-based WB model and the SMA-based AIS model. The measures are slightly better for the WB model compared to the AIS model, while the MSE is better for the AIS model. As the measures only slightly differ, we can conclude that both models are useful in forecasting water levels.

The monthly data has a training set starting from 2009-2016, while the weekly data has a training set starting from 2013-2016. In other words, the monthly data set has 4 more years of data to base its forecasts on. The monthly-based data can predict monthly future values, while the weekly-based data can predict weekly future values. The monthly data contains the reported water levels of the 15th each month, meaning the model would predict water levels based on the same monthly value, using four different precipitation forecasts. Using weekly precipitation forecasts for all weeks in a month, to predict a monthly level, could train the model inaccurately. It would be optimal to gather weekly water levels earlier than 2009 to create a more accurate model and observe the relationship between precipitation and water levels further. However, the AIS-based model with its high-frequency provides good results of accuracy measures, indicating a useful prediction model.

The development of the accuracy measures over time for the monthly-based model is shown in appendix F.

5.2 Profit estimation example

This section will investigate the sensitivity in the profit function when draft changes, to illustrate how crucial it is to have an accurate prediction model. To get a realistic example for the case study, Western Bulk provided us with representative numbers that can be used to calculate profit.

The example is based on standard travel with an Ultramax (63k dwt) vessel, departing from San Lorenzo port to El Dhekelia in Egypt. This voyage has a duration of 41 days, with a time charter equivalent (TCE) of 40 000. The calculations are based on market rates and bunker prices from late November 2021, and the maximum draft is predicted to be 9.75 meters in the Paraná River. The selected draft allows the ship to have an intake of 38 500 mts.

The total freight and profit change function for an intake of 38 500 mts and a total freight cost of \$2 781 625 is given as follows:

$$\underbrace{\$2\ 781\ 625}_{Total\ Freight} = \underbrace{\$72{,}25}_{Freight\ USD/pmt} * \underbrace{38\ 500}_{Intake}$$

$$Profit\ Change = \left(\underbrace{\$72{,}25}_{Freight\ USD/pmt} * \underbrace{\pm 38\ 500}_{Intake}\right) - \underbrace{\$2\ 781\ 625}_{Total\ Freight}$$

Table 7 summarizes the assumptions, where the prices are given in USD.

Assumptions	
Load port	San Lorenzo, Argentina
Discharge port	El Dekheila, Egypt
- VLSFO USD/pmt	575
Duration (days)	41
Vessel	Ultramax (63k dwt)
TCE	40 000
Draft (m)	9,75
Freight USD/pmt	72,25
Total Freight	2 781 625

Table 7: Assumptions taken in the profit estimation example.

Table 8 displays the numbers used in Figure 17, where both illustrate the relationship between draft, profit, and intake of cargo.

Draft (m)	Intake (mts)	Profit Change
8,5	31000	- 541 875
8,75	32500	- 433 500
9,00	34000	- 325 125
9,25	35500	- 216 750
9,50	37000	- 108 375
9,75	38500	0
10,00	40000	108 375
10,25	41500	216 750
10,50	43000	325 125

Table 8: Relationship between draft, profit, and intake of cargo, tablewise.

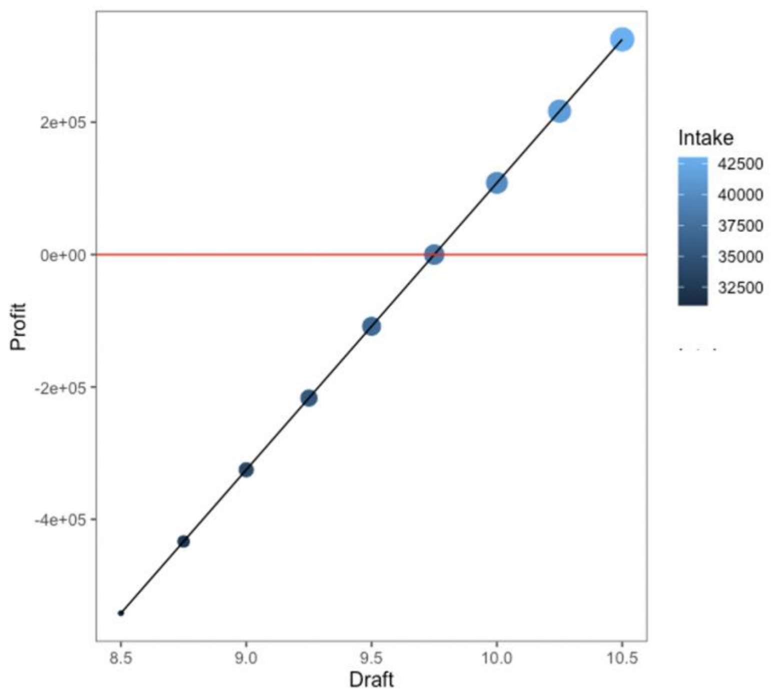

Figure 17: Relationship between draft, profit, and intake of cargo, graphically.

By wrongly over- or underestimating 0.25 meters in draft, the profit can result in a range of $\pm$ \$108 375. The calculation demonstrates the large impact draft has on Western Bulk's profitability, indicating the importance and need of an accurate prediction model.

6. Conclusion

The purpose of this thesis is to reduce uncertainty in under- or overestimation of cargo in the Paraná River in Argentina, by predicting the allowable maximum draft for ships to pass the water level barrier. Accurate predictions can improve profitability of ships voyage. The study is based on reported draft from AIS (Vesseltracker), observed precipitation from Copernicus (CDS) and precipitation forecasts from ECMWF.

The first part of the study focused on transforming the large amount of AIS data, to solely include relevant information related to Paraná River. A crucial limitation in the study is the case of "Frontline Operator failure" in AIS reported draft. From our and other studies' experience, the manually reported AIS data are not always filled in, updated or reliable. To tackle this problem, we aggregated the data on a weekly level and removed potential errors in reported draft. However, the data transformation could have removed large amounts of valuable data, which could be crucial if the true value of maximum draft for a given week was removed.

The second part of the study focused on exploring seasonality and trends in precipitation and surface runoff levels and transforming it to an appropriate independent variable to predict water levels. Both a simple and exponential 52-week moving average were explored, using Pearson correlation test, as well as the linear and GAM model. Solely precipitation, using a 52-week simple moving average, was included in the final model, as it gave the highest correlation ($R = 0.6$), and explained ~40% of the variance in AIS reported draft levels using the GAM model. However, to find a model that explains a larger part of the variance, variables such as measurement of lag time from water levels upriver to downriver, frequency and measurement of dredging, flow velocity of nearby dams (Itaipu Dam and Yacyretá Dam), water flow and other hydrological variables should be considered and explored in order to improve the model.

The third part of the study included transforming precipitation forecasts to match the independent variable as a moving average and predict water levels forward. Using precipitation forecasts accounted for the uncertainty in weather and created a distribution of potential water levels.

At last, to predict the allowable draft level, we have used the GAM model, as it performed better than linear regression. The final GAM model used a 52-week SMA to predict reported maximum draft levels from AIS. The accuracy measures R^2, MSE, RMSE and MAPE resulted in respectively 0.389, 0.077, -0.265 and 0.026. A limitation in the model is the lack of use of penalties in the GAM model, which allows complexity of the model to increase (i.e. overfitting). Earlier studies have shown good results by using complex models such as ANN, as it catches non-linear relationships better compared to regression models (Bazartseren et. al., 2003). Further work should therefore investigate other methods of modelling to achieve more accurate predictions.

We believe AIS reported draft to become more useful in a couple of years, as the use of it is increasing and longer timeframes in general of observations usually contribute to a higher accuracy in prediction models. However, the prediction model shows promising results, and using it with as a complementary model along with data points from the pilot company can be useful.

Conclusively, further development of the model should include adding more variables, reviewing other potential modelling methods, and achieving a higher-frequency or longer time-frame data set for water levels.

References

Bazartseren, B., Hildebrandt, G., & Holz, K. (2003). *Short-term water level prediction using neural networks and neuro-fuzzy approach*. Neurocomputing (Amsterdam), 55(3), 439-450.

Chai, T. & Draxler, R. (2014). *Root mean square error (RMSE) or mean absolute error (MAE)? –Arguments against avoiding RMSE in the literature*. Geoscientific Model Development. 7. 1247-1250. 10.5194/gmd-7-1247-2014.

Chatfield, C. (2000). *Time-series forecasting*. Boca Raton, Fla: Chapman & Hall/CRC.

De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. Neurocomputing (Amsterdam), 192, 38-48. https://doi.org/10.1016/j.neucom.2015.12.114.

Fernando, J. (2021). *Moving Average (MA)*. Retrieved 15.11.2021 from https://www.investopedia.com/terms/m/movingaverage.asp.

Gibb, T. (2015). *Understanding Great Lakes water levels*. Retrieved from https://www.canr.msu.edu/news/understanding_great_lakes_water_levels.

González-Sopeña, J., Pakrashi, V., & Ghosh, B. (2021). *An overview of performance evaluation metrics for short-term statistical wind power forecasting*. Renewable & Sustainable Energy Reviews, 138, 110515.

Harati-Mokhtari, A., Wall, A., Brooks, P., & Wang, J. (2007). Automatic Identification System (AIS): A Human Factors Approach. Journal of Navigation, 60(3), 373-389.

Hastie, T., & Tibshirani, R. (1986). *Generalized Additive Models*. Statistical Science, 1(3), 297-310.

Hayes, A. (2021). *Multiple Linear Regression (MLR)*. Retrieved 08.11.2021 from https://www.investopedia.com/terms/m/mlr.asp.

Herman-Saffar, O. (2020). *Time Based Cross Validation*. Retrieved from https://towardsdatascience.com/time-based-cross-validation-d259b13d42b8.

Jia, H., Prakash, V. & Smith, T. (2019). *Estimating vessel payloads in bulk shipping using AIS data*. International Journal of Shipping and Transport Logistics, 11 (1), 25-40. https://doi.org/10.1504/IJSTL.2019.096864.

Lejeune, O. (2020). *Using machine learning to predict Rhine water levels*. Retrieved from https://towardsdatascience.com/using-machine-learning-to-predict-rhine-water-levels-44afce697074.

Naghshpour, S. (2016). *Regression for economics* (Second ed., Economics collection). New York, New York (222 East 46th Street, New York, NY 10017): Business Expert Press.

Portal oficial del Estado argentino. (n.d.a). *Vía Navegable Troncal*. Retrieved 10.12.2021 from https://www.argentina.gob.ar/transporte/administracion-general-puertos-se/via-navegable-troncal.

Portal oficial del Estado argentino. (n.d.b). *¿Qué es la Vía Navegable Troncal?* Retrieved 10.12.2021 from https://www.argentina.gob.ar/transporte/via-navegable-troncal/sobre-la-hidrovia.

Puth, M., Neuhäuser, M., & Ruxton, G. (2014). *Effective use of Pearson's product–moment correlation coefficient*. Animal Behaviour, 93, 183-189.

Schober, P., Boer, C. & Schwarte, L. (2018). *Correlation Coefficients: Appropriate Use and Interpretation*. Anesthesia & Analgesia, 126 (5), 1763-1768.

Shafi, A. (2021). *What is a Generalised Additive Model?* Retrieved from https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a.

Song, Q. & Esogbue, A. O. (2006), A New Algorithm for Automated Box-Jenkins ARMA Time Series Modeling Using Residual Autocorrelation/Partial Autocorrelation Functions, International Journal of Industrial Engineering and Management Systems, 5(2), 116-125.

U.S. Department of Agriculture. (2021). *Argentina: Grain and Feed Annual*. Retrieved from https://www.fas.usda.gov/data/argentina-grain-and-feed-annual-7.

Wang, C., Yuan, J., Zhang, J., Deng, N., Zhou, Z., & Gao, F. (2020). *Multi-criteria comprehensive study on predictive algorithm of heating energy consumption of district heating station based on timeseries processing*. Energy (Oxford), 202, 117714.

Wei, T., Zhang, T., Cui, X., Glomsrød, S., & Liu, Y. (2019). *Potential influence of climate change on grain self-sufficiency at the country level considering adaptation measures*. Earth's Future, 7, 1152-1166. https://doi.org/10.1029/2019EF001213.

Wu, X., Roy, U., Hamidi, M., & Craig, B. (2020). *Estimate travel time of ships in narrow channel based on AIS data*. Ocean Engineering, 202, 106790.

Yale University: Department of Statistics and Data Science (1993). *Linear Regression*. Retrieved from http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.

Appendices

Appendix A: Data exploration of the runoff variable

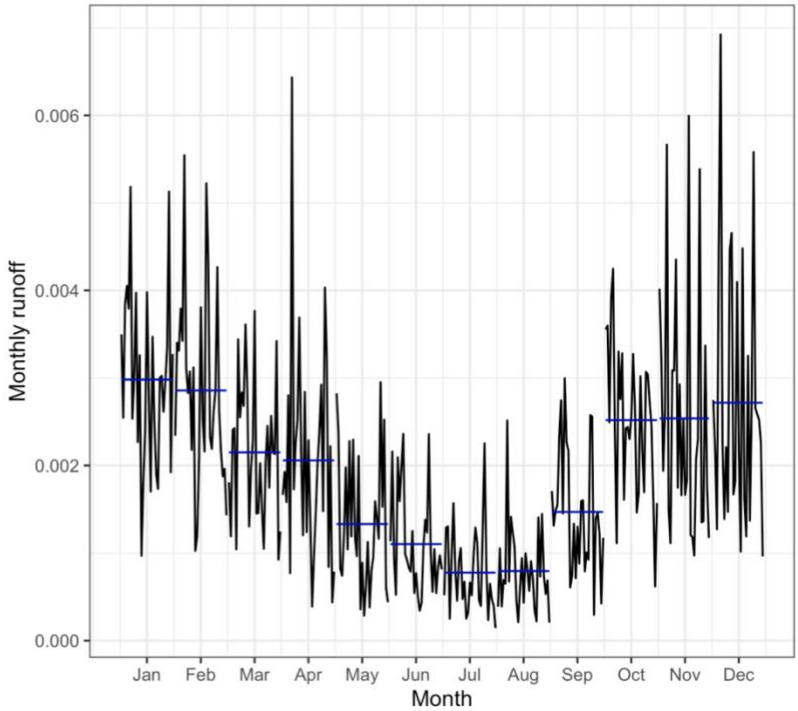

Figure 19: Seasonal pattern of monthly runoff mean (from 1993 to 2021), where the red horizontal line represents the accumulated mean for each month.

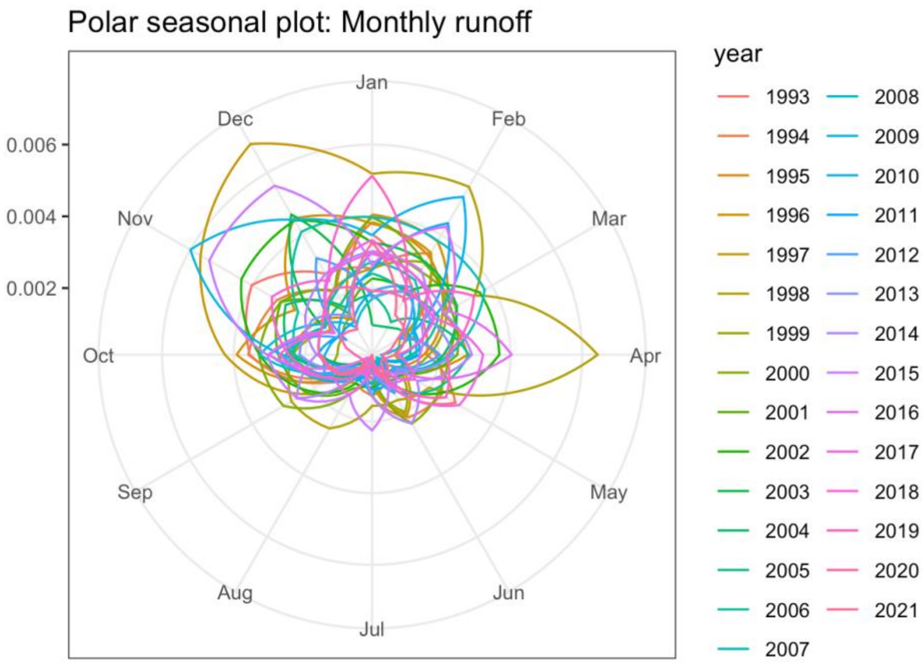

Figure 18: Polar seasonal plot for monthly precipitation data (from 1993 to 2021).

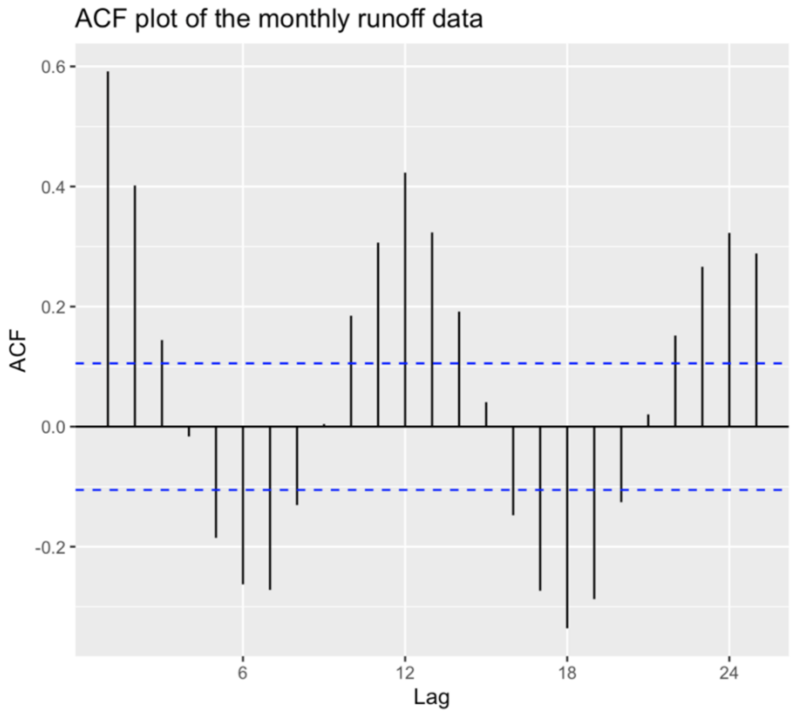

Figure 21: A compressed plot of the autocorrelation values for the first 25 lags.

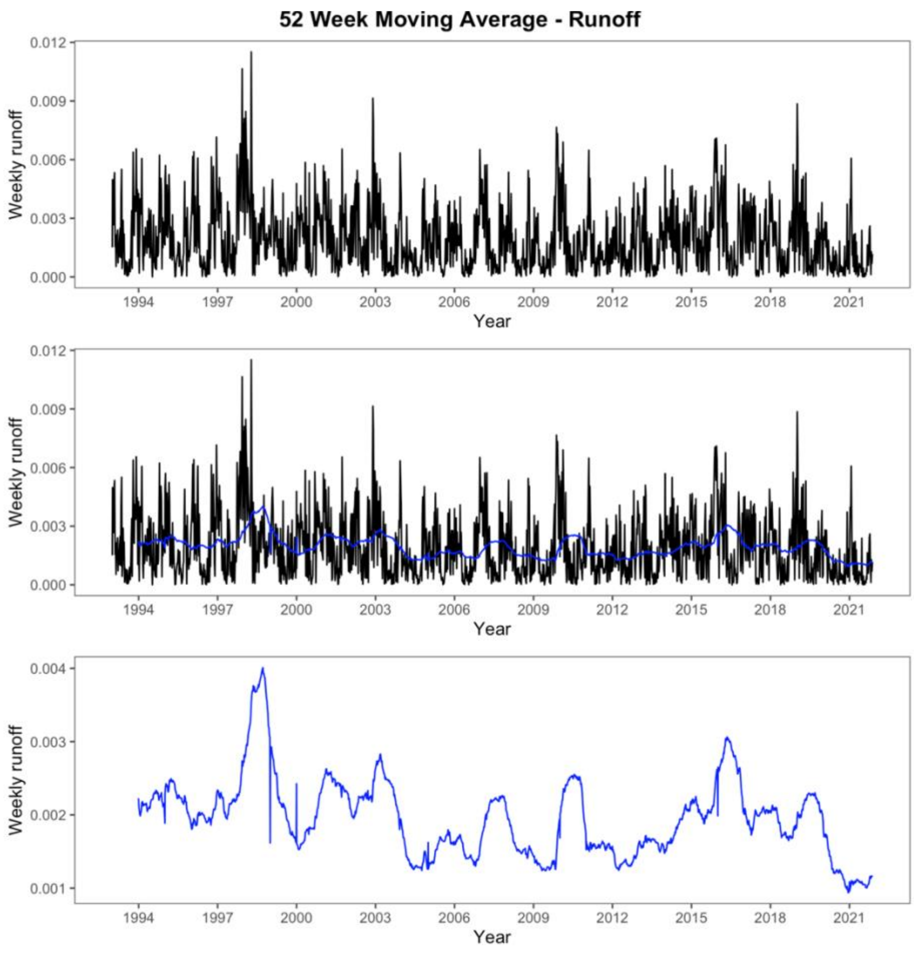

Figure 20: 52-week simple moving average of runoff from 1993-2021.

Appendix B: Correlation between weather variables and the maximum reported draft level

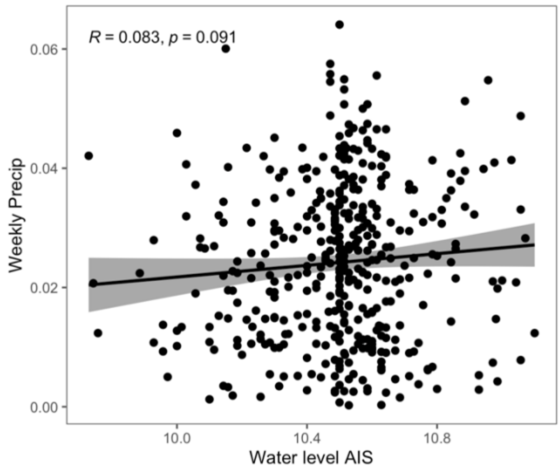

Figure 22: Correlation between AIS level and weekly precipitation.

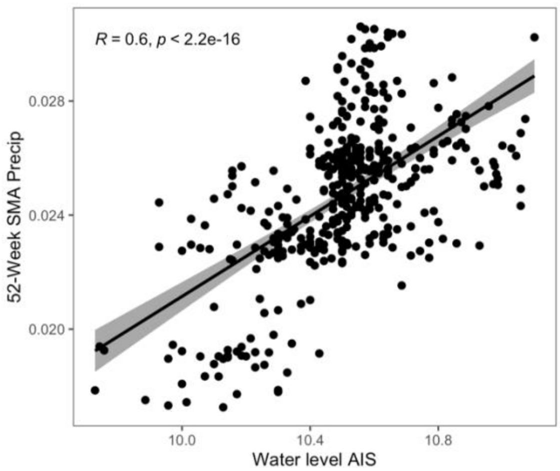

Figure 24: Correlation between AIS water levels and 52-week simple moving average of precipitation.

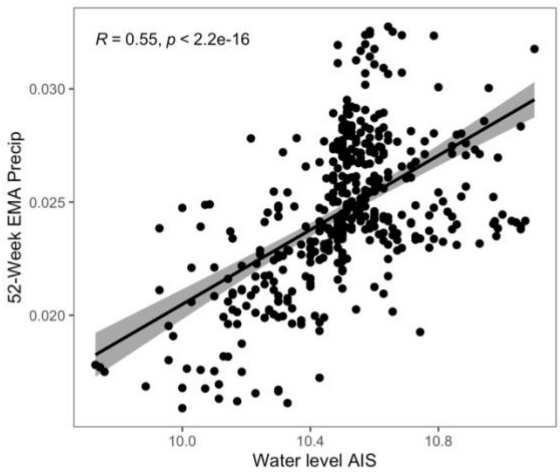

Figure 23: Correlation between AIS water levels and 52-week exponential moving average of precipitation.

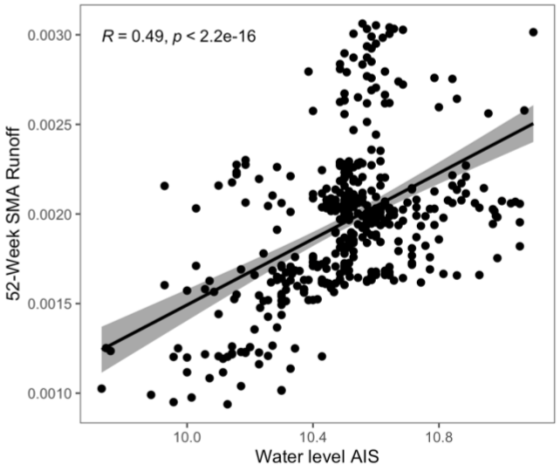

Figure 25: Correlation between AIS water levels and 52-week simple moving average of precipitation.

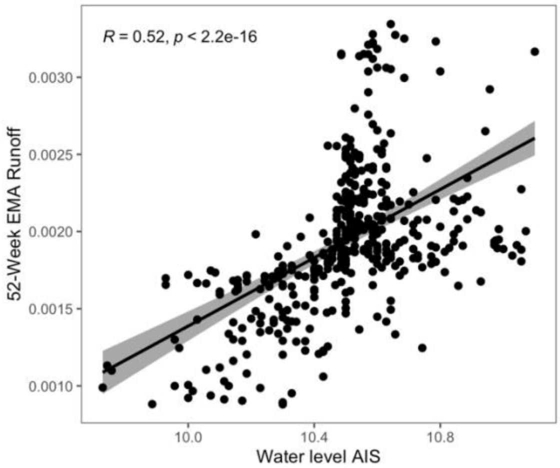

Figure 26: Correlation between AIS water levels and 52-week exponential moving average of precipitation.

Appendix C: Linear and GAM models for weekly-based model and monthly-based model.

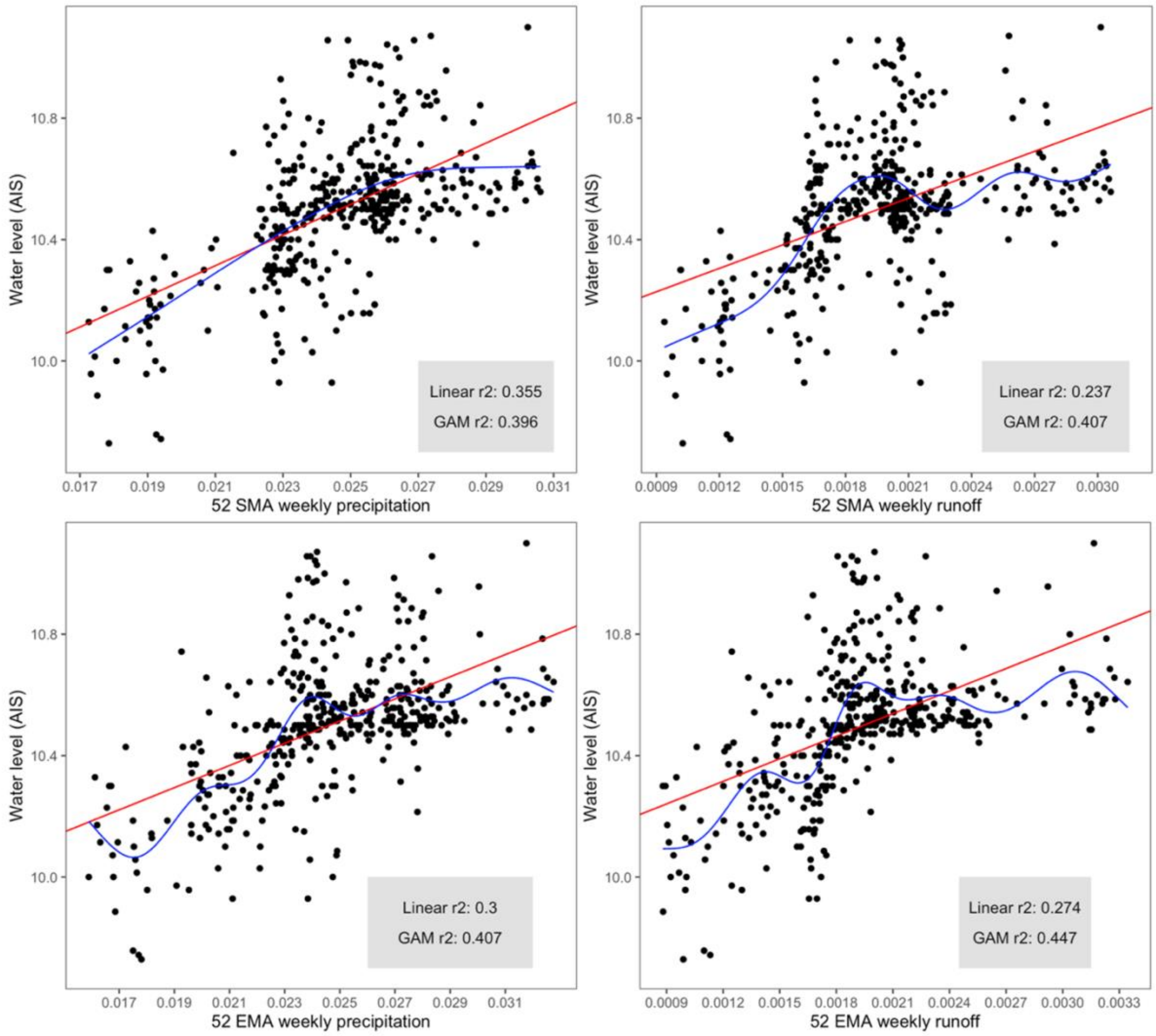

Figure 27: Relationship between 52-week simple and exponential moving average of precipitation and runoff for AIS reported water levels.

The result table from models performed on water levels reported by Western Bulk is shown in Table 9, and graphically in Figure 28.

Result table using monthly reported water levels as the dependent variable							
		SMA			EMA		
		Precip	Runoff	Both	Precip	Runoff	Both
Model	Linear R^2	0.355	0.268	0.399	0.457	0.445	0.463
	GAM R^2	0.377	0.337	0.402	0.453	0.5	0.51

Table 9: Result table with the relationship between 52-week simple and exponential moving averages of precipitation and runoff for reported water levels from Western Bulk.

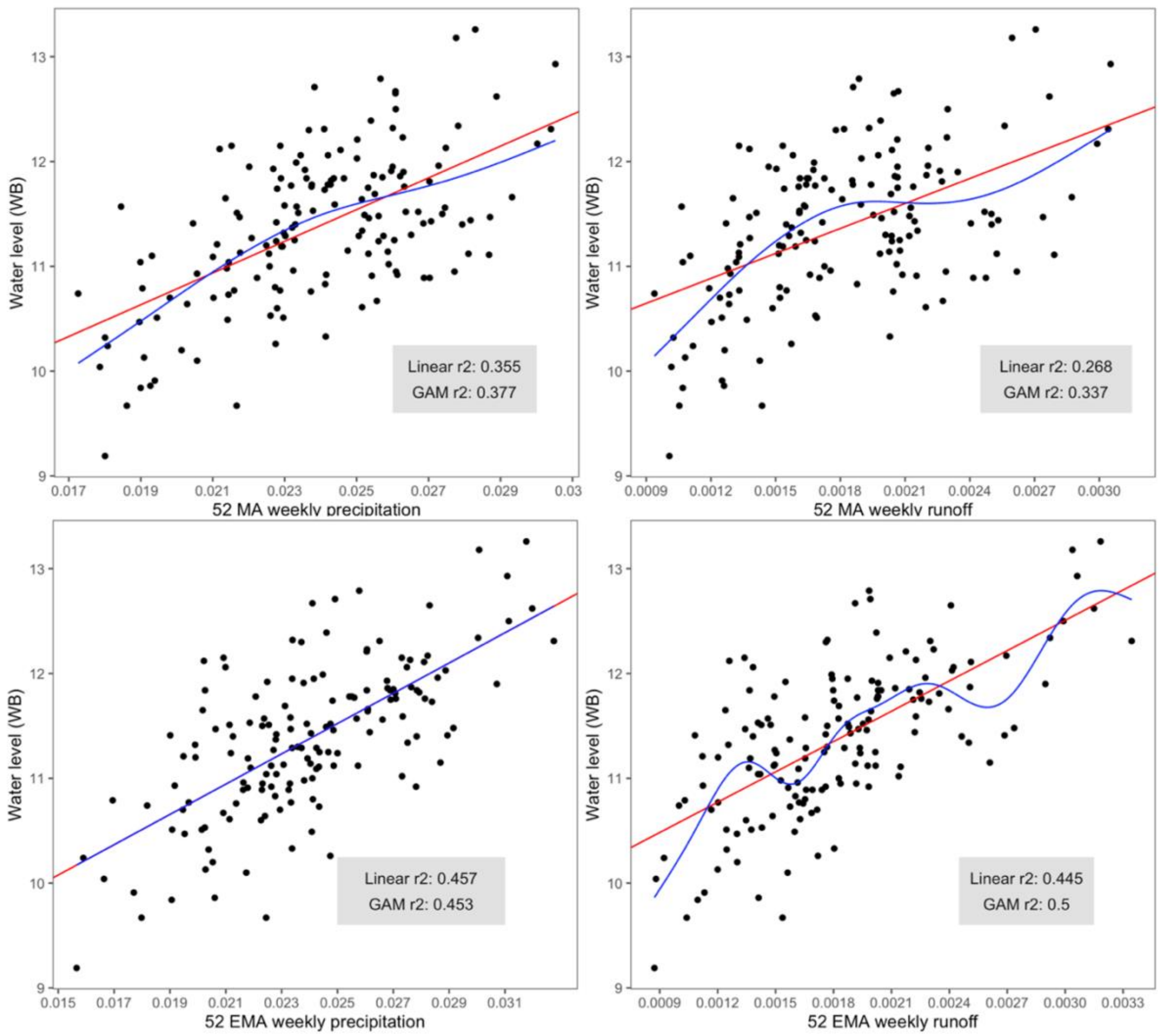

Figure 28: Relationship between 52-week simple and exponential moving average of precipitation and runoff for Western Bulks monthly reported water levels

Appendix D: Forecast model example, using EMA

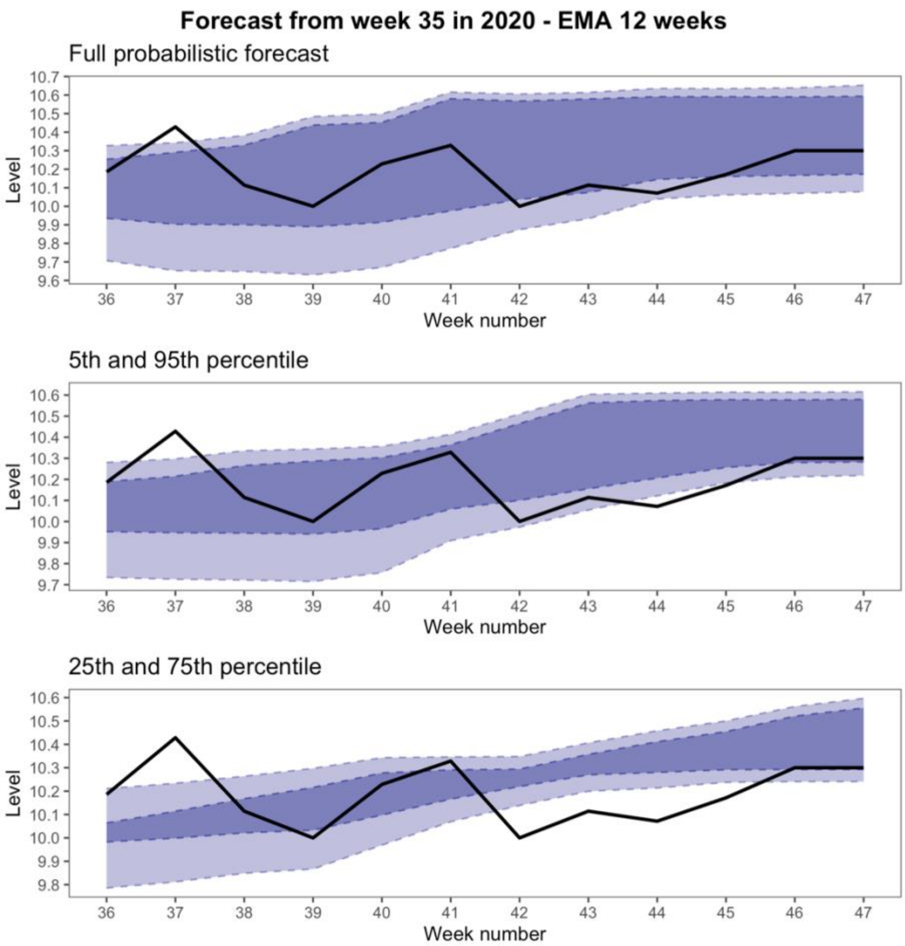

Figure 29: Forecast from week 35 in year 2020, predicting 12 weeks ahead, using the exponential moving average of precipitation. Forecast distribution (in dark purple), standard deviation distribution (in light purple) and observed levels for draft (black line).

Appendix E: Development of accuracy measures, using EMA based on AIS reported draft

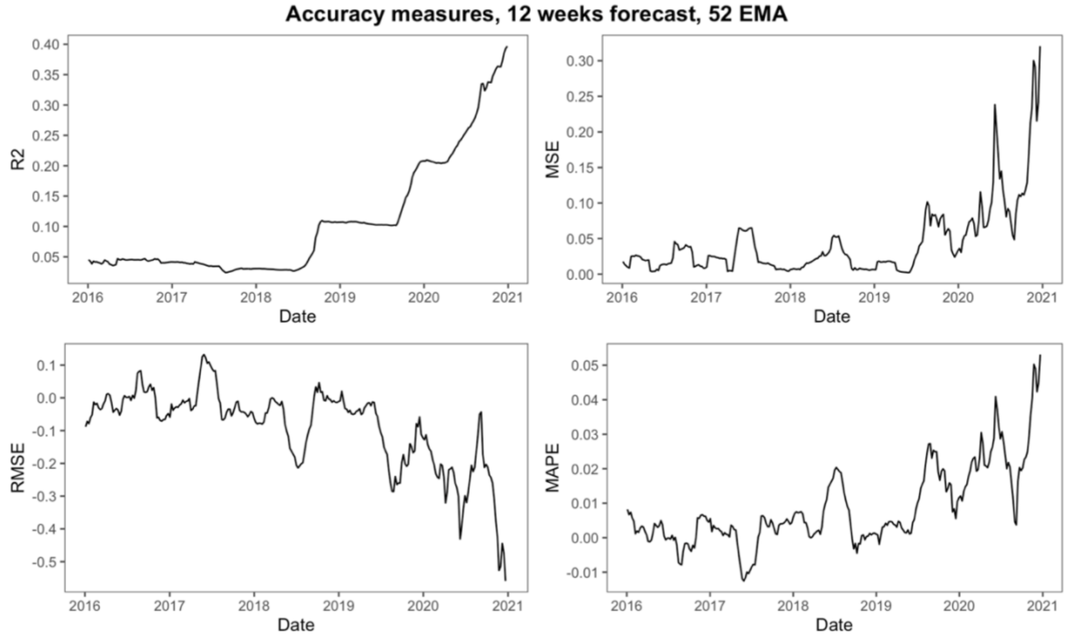

Figure 30: Accuracy measures for the model using the 52-week exponential moving average (EMA) variable based on AIS reported draft.

Appendix F: Development of the accuracy measures (monthly-based model)

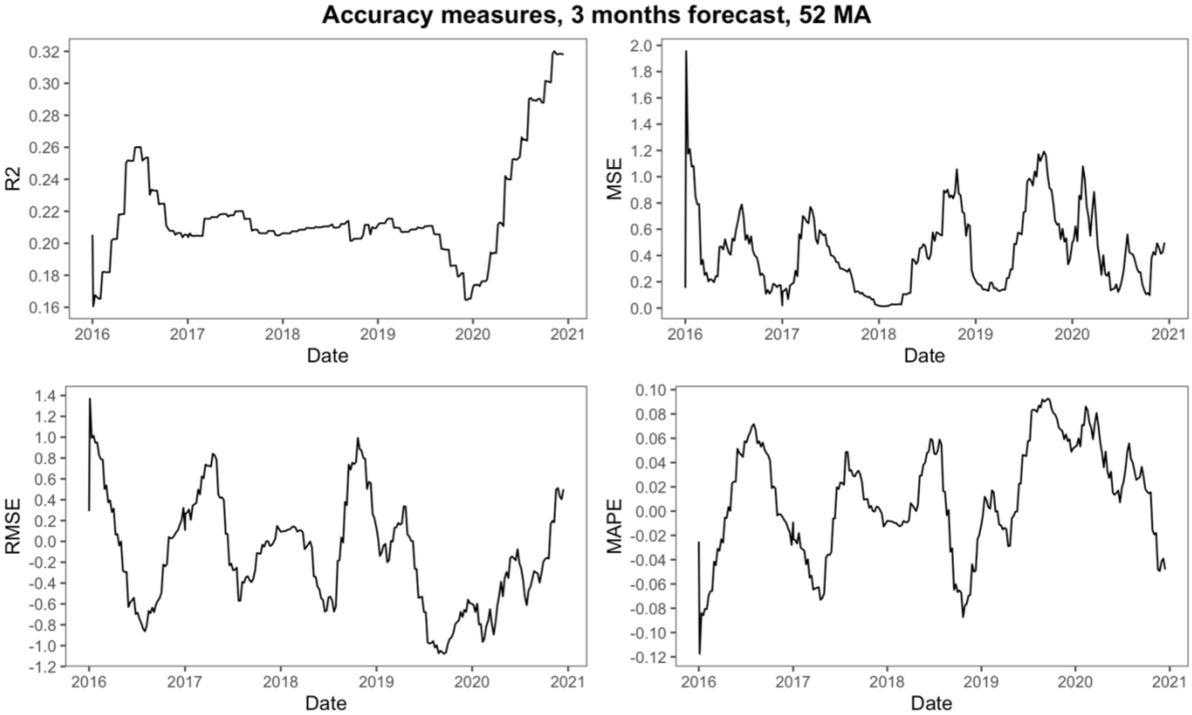

Figure 31: Accuracy measures for the model using the 52-week simple moving average (SMA) variable based on monthly reported water levels.

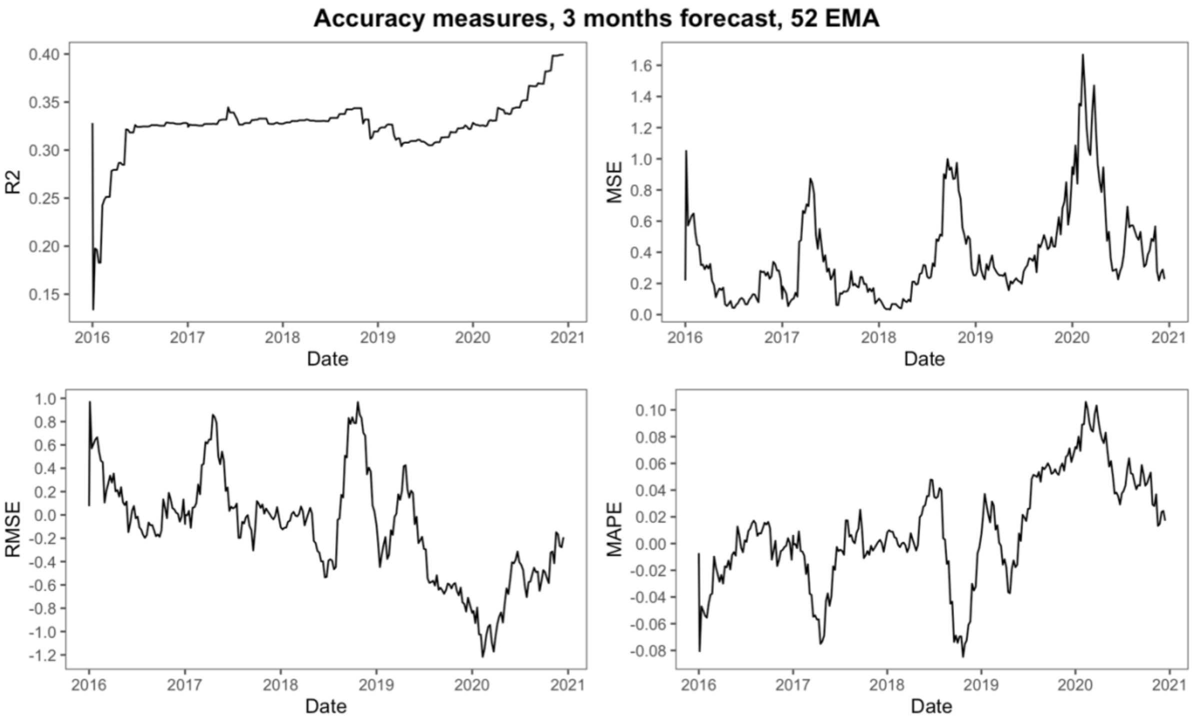

Figure 32: Accuracy measures for the model using the 52-week exponential moving average (EMA) variable based on monthly reported water levels.

An example of the forecast model using the monthly reported water level data:

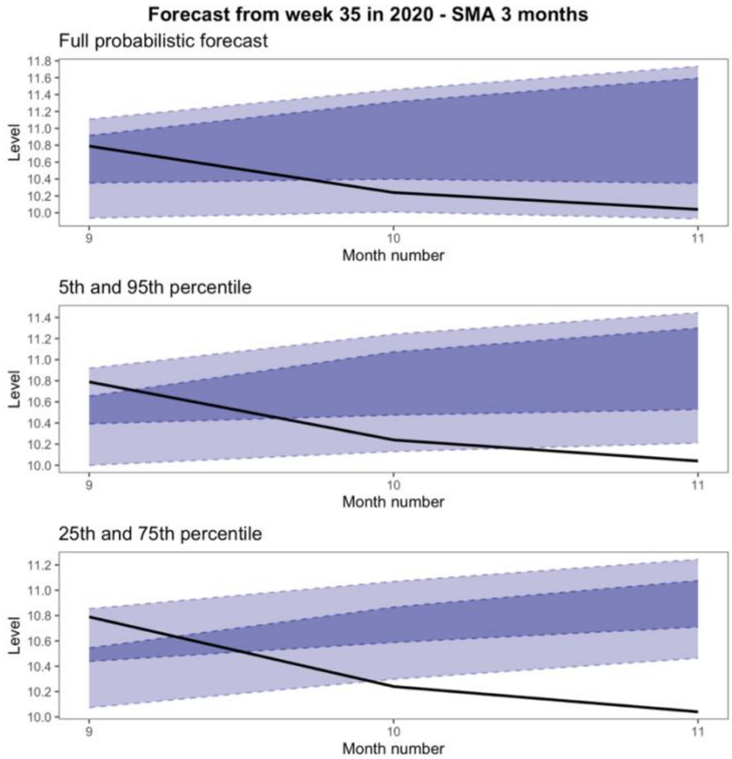

Figure 33: Forecast from week 35 in year 2020, predicting 12 weeks ahead, using the exponential moving average of precipitation

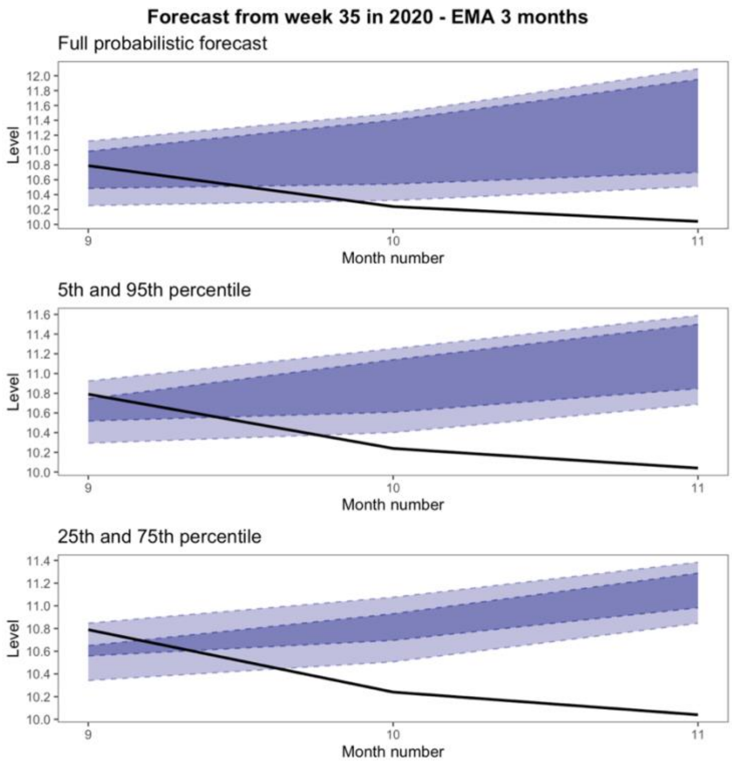

Figure 34: Forecast from week 35 in year 2020, predicting 12 weeks ahead, using the exponential moving average of precipitation